Andreas Eichler | Markus Vogel

Leitfaden Stochastik

Für Studierende und Ausübende des Lehramts

STUDIUM

VIEWEG+
TEUBNER

Andreas Eichler | Markus Vogel

Leitfaden Stochastik

Andreas Eichler | Markus Vogel

Leitfaden Stochastik

Für Studierende und Ausübende des Lehramts

STUDIUM

VIEWEG+
TEUBNER

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.d-nb.de> abrufbar.

Prof. Dr. Andreas Eichler
Pädagogische Hochschule Freiburg
IMBF Institut für Mathematische Bildung Freiburg
Kunzenweg 21
79117 Freiburg

andreas.eichler@ph-freiburg.de

Prof. Dr. Markus Vogel
Pädagogische Hochschule Heidelberg
Institut für Mathematik und Informatik
Im Neuenheimer Feld 561
69120 Heidelberg

vogel@ph-heidelberg.de

1. Auflage 2011

Lektorat: Ulrike Schmickler-Hirzebruch | Barbara Gerlach

Vieweg+Teubner Verlag ist eine Marke von Springer Fachmedien.
Springer Fachmedien ist Teil der Fachverlagsgruppe Springer Science+Business Media.
www.viewegteubner.de

Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: AZ Druck und Datentechnik GmbH, Berlin
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier
Printed in Germany

ISBN 978-3-8348-1402-9

Vorwort

> „Statistik ohne Wahrscheinlichkeitsrechnung ist blind [...], Wahrscheinlichkeitsrechnung ohne Statistik ist leer [...]."
>
> (Schupp, 1982, S. 210)

Diesem didaktisch motivierten Zitat ist dieses Buch zur Stochastik – als Überbegriff von Statistik und Wahrscheinlichkeitsrechnung – verpflichtet. Es versteht sich dabei als fachlich orientierte Betrachtung der sogenannten *Leitidee Daten und Zufall* für den Stochastikunterricht der Sekundarstufe I (Eichler & Vogel, 2009). Angesprochen sind insbesondere Lehramtsstudierende und Lehrende dieser Schulstufe, aber ebenso der Primarstufe. Das Buch ist entstanden aus Vorlesungen zur Stochastik für das Lehramt an Primar- und Sekundarstufen, die auf folgenden überinhaltlichen primären Zielen basierten, nämlich

- der Betonung eines datenorientierten Zugangs zur Stochastik, bei dem auch elementare Methoden der Datenanalyse breiten Raum einnehmen und Verbindungen zwischen der Datenanalyse und der Wahrscheinlichkeitsanalyse aufgezeigt werden,
- der Betonung des Unterschieds zwischen der empirischen Welt der Daten und der Modell-Welt der Wahrscheinlichkeiten, in der aufgrund von Modellannahmen vorgegangen wird,
- der Betonung von Simulationen, um zwischen empirischen Daten und in Wahrscheinlichkeitsverteilungen bestehenden theoretischen Modellen zukünftiger Daten zu vermitteln,
- der Betonung von Ideen, die hinter den stochastischen Methoden stehen, gegenüber einer alleinigen Betonung der elaborierten und standardmäßig verwendeten Methoden der Stochastik, ohne Letztere auszublenden, und schließlich
- der Verwendung je eines führenden Beispiels, an dem sich die Entwicklung der Methoden jeweils orientiert.

Mit diesen Zielen werden andere ausgeblendet. So ist dieses Buch sicher nicht als Handbuch statistischer Methoden gedacht. Hier gibt es eine große Anzahl zu empfehlender Werke, von denen wir Hartung, Elpelt und Klösener (2009), Sachs (1999) und, mit einem alternativen Zugang, Polasek (1994) häufiger anführen. Wir beschränken uns dagegen auf wenige Methoden der Datenanalyse, diskutieren diese aber anhand von Beispielen und punktuellen Vertiefungen ausführlich. Das Buch ist auch keine umfassende Darstellung der Wahrscheinlichkeitstheorie, die für Studierende des Fachs Mathematik oder der Stochastik anwendenden Wissenschaften vorgesehen ist. Für diese Ausrichtung gibt es ebenfalls eine kaum zu überschauende Fülle von Publikationen, von denen wir Fisz (1980), Henze (2010) und Kreyszig (1998) gerne anführen. So beschränken wir uns im Bereich der Wahrscheinlichkeitsanalyse etwa strikt auf endliche Ergebnismengen, loten dabei aber aus, wie weit die Konzepte zu endlichen Ergebnismengen tragen. Zudem skizzieren wir Aspekte der beurteilenden Statistik nur und gehen dabei einen in Lehrbüchern zur Stochastik eher unüblichen Weg. Zentral bleibt aber für uns, die (aus unserer Sicht) entscheidenden Aspekte der Stochastik, die deskriptive Analyse von Daten, das darauf basierende Aufstellen und Durcharbeiten von wahrscheinlichkeitstheoretischen Modellen und schließlich das erste Beurteilen solcher Modelle, miteinander verbunden darzustellen.

Wie die Studierenden, die unsere Vorlesungen besucht haben, führen wir auch Sie anhand eines Datensatzes zu Eigenschaften von Studierenden durch dieses Buch. Dabei unterscheiden wir zwischen einem Hauptstrang, in dem wir die stochastischen Methoden anhand von Beispielen einführen, einem Nebenstrang bestehend aus fachlichen Ergänzungen, Online-Zusatzmaterialien und schließlich einem Umsetzungsteil, in dem die vorher angesprochenen Methoden auf den Datensatz zu den Studierenden exemplarisch angewendet werden. Die Online-Zusatzmaterialien stehen zum Download bereit unter:

http://www.viewegteubner.de/Buch/978-3-8348-1402-9/Leitfaden-Stochastik.html

und unserer Homepage:

http://www.leitideedatenundzufall.de

In den Online-Materialien stellen wir zusätzlich auch Lösungen zu den Übungsaufgaben des Buches zur Verfügung.

Mit dieser Konzeption haben wir eine strikt schulorientierte Variante der Stochastik und damit eine Ergänzung der Leitfäden Arithmetik und Geometrie konzipiert. Wir hoffen, dass Sie dieser Konzeption und damit uns in die Welt von Daten und Zufall folgen wollen.

Freiburg und Heidelberg im April 2011 Andreas Eichler & Markus Vogel

Inhaltsverzeichnis

1 Erhebung statistischer Daten

Einstiegsbeispiel

Abbildung 1.1: Studierende im Hörsaal

Aufgabe 1: Planen Sie eine Erhebung zu Eigenschaften von Studierenden und führen Sie diese durch.

Worum es geht

Die Datenerhebung steht oftmals ein wenig abseits, wenn Fragen der Datenanalyse und -auswertung diskutiert werden. Zu Unrecht! Die Datenerhebung ist das Fundament, auf dem alle weiteren Schritte der Datenbearbeitung aufbauen. Sie gewährleistet von Anfang an, ob die Ergebnisse einer Datenanalyse brauchbar sein können oder nicht. Selbst die ausgefeiltesten datenanalytischen Methoden der Statistik können nicht fruchten, wenn die Daten nicht mit der nötigen Sorgfalt erhoben wurden und deshalb die Daten nicht das tun, was sie eigentlich sollten: möglichst gut einen Ausschnitt der Realität in ihrer gesamten Komplexität abbilden. Das bedeutet, „gute Daten“ können einen guten Erkenntnisgewinn zu einem Realitätsausschnitt ermöglichen, „schlechte Daten“ werden auch durch die ausgereifte Anwendung der statistischen Methoden zu keinem guten Erkenntnisgewinn führen.

Wie kann man aber sichern, dass „gute Daten“ die Grundlage für einen späteren guten Erkenntnisgewinn gewährleisten? Diese Frage birgt ein Dilemma, dem sich dieses Kapitel und der Anfang dieses Buches gegenüber sieht: Die Erhebung der Daten steht am Anfang der Datenanalyse, damit auch am Anfang dieses Buchs und ist für alle weitere Analysen ein zentral wichtiger Schritt. Wie „gute Daten“ erhoben werden beziehungsweise wie dies gewährleistet werden kann, basiert dagegen bereits auf fortgeschrittener Kenntnis der Stochastik: Erst wenn man weiß, was

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_1,

alles schiefgehen kann, weiß man auch, wie Missgeschicke und Unzulänglichkeiten bei der Datenerhebung umgangen werden können. So gesehen müsste die Datenerhebung also fast am Ende der Betrachtungen stehen.

Wir lösen dieses Dilemma hier so, dass wir anhand einer Umfrage zu Eigenschaften von Studierenden auf Grundideen der Datenerhebung eingehen, ohne diese jedoch im fachlichen Detail zu diskutieren. In diesem Zusammenhang sind folgende drei Fragen grundlegend:

Was soll erhoben werden? Hier sollen es Eigenschaften der Studierenden sein. Da könnte etwa das Geschlecht von Interesse sein, das man im Regelfall auch eindeutig beobachten oder erfragen kann. Schwieriger wird es, wenn beispielsweise eine Eigenschaft wie musikalische Freizeitaktivitäten erhoben werden soll. Dazu muss zunächst einmal ein fester Zeitrahmen gegeben werden, etwa die (geschätzte) durchschnittliche wöchentliche Anzahl von Stunden, da ohne einen festen Zeitrahmen alle Angaben nicht interpretierbar wären. Ebenso muss aber auch festgelegt werden, was mit der Freizeitaktivität Musik gemeint ist. Ist es der bloße Musikkonsum und/oder das aktive Musizieren? Legt man die interessierende Eigenschaft von Studierenden nicht fest, so wird eine Person diese Eigenschaft so, eine andere Person anders verstehen. Die Ergebnisse der Erhebung würden in diesem Fall wertlos.
Ein wichtiges Kriterium für „gute“ Daten besteht also darin, die zu erhebenden Eigenschaften möglichst eindeutig zu definieren.

Von wem sollen Eigenschaften erhoben werden? Im besten Fall erhebt man die Eigenschaften aller Studierenden. Aber wer sind eigentlich *alle Studierende*? Hier könnte man sich zunächst auf eine Hochschule beschränken. Die Konsequenz einer solchen Einschränkung hat zur Folge, dass jegliche Aussagen sich zunächst nur auf die Studierenden dieser Hochschule beziehen und nicht für alle Studierenden (in Deutschland, Europa, der Welt, ...) *verallgemeinerbar* sind. Aber selbst unter dieser Einschränkung ist genau festzulegen, was mit dem Begriff *Studierende* gemeint ist. Sind es alle tatsächlich eingeschriebenen Studierenden, werden Gasthörer mitgezählt, sind Urlaubssemester auszuschließen, ... ? Solche Fragen zeigen, dass auch ein solch vermeintlich festes Kriterium wie „eingeschrieben – ja oder nein“ nicht abschließend alle Detailfragen beantwortet. Der Versuch, alle Studierenden – wenn das denn genau geklärt ist – für eine Datenerhebung zu erfassen, erweist sich in der Praxis selbst bei kleinen Hochschulen als fast unmöglich. Man ist in aller Regel darauf angewiesen, nur von einem Teil der Studierenden (einer Hochschule) Daten zu erheben. Ziel der Datenerhebung muss sein, die Ergebnisse auf die Gesamtheit der Studierenden verallgemeinern zu können. Dies ist aber nicht einfach so gegeben: Werden etwa nur weibliche Studierende zu einer Eigenschaft befragt, so ist es sicher fragwürdig, ob man das Ergebnis der Erhebung ohne Weiteres auch auf die männlichen Studierenden übertragen kann. Vielleicht beeinflusst auch die Beschränkung auf ein bestimmtes Alter, ein Studienfach oder die Anzahl der Semester die Ergebnisse einer Umfrage.
Ein Kriterium für „gute Daten“ besteht also darin, die *Grundgesamtheit* und die *Stichprobe*[1] so festzulegen, dass sich die Ergebnisse der Stichprobe (zumindest theoretisch) für die Grundgesamtheit verallgemeinern lassen.

Wie sollen die Eigenschaften erhoben werden? Selbst wenn man die Eigenschaften von Studierenden eindeutig festgelegt hat, ist es nicht per se gewährleistet, dass tatsächlich auch diese

[1] Das Erste ist die Gesamtheit der potentiell zu erhebenden Objekte (hier der Studierenden), das Zweite ist der Anteil der tatsächlich zu erhebenden Objekte (der Studierenden), vgl. Kapitel 1.1

Eigenschaft erhoben wird oder, allgemeiner ausgedrückt, dass das gemessen wird, was gemessen werden soll. Erhebt man etwa die Eigenschaften von Studierenden per Befragung und hat vorab das Merkmal musikalische Freizeitaktivität genau festgelegt, so muss natürlich auch die entsprechende Frage möglichst eindeutig verstehbar gestellt werden. Schwieriger wird es noch, wenn nur indirekt feststellbare Eigenschaften erhoben werden sollen. So eine Eigenschaft wäre beispielsweise das pädagogische Wissen der Studierenden. Dieses kann man höchstens über eine bestimmte Anzahl von Fragen oder Testaufgaben erheben.

In der Sozialforschung wird großer Aufwand betrieben hinsichtlich der Frage, wie Eigenschaften von Personen definiert und anschließend *operationalisiert* (messbar gemacht) werden können. Hier ist die *Validität* der Messung von Personeneigenschaften ein Gütekriterium einer statistischen Untersuchung (vgl. Bortz & Döring, 2006). Mit Blick darauf werden wir die Fachbegriffe *Grundgesamtheit* und *Stichprobe*, die wir bereits in diesem einführenden Kapitel verwendet haben, im folgenden Kapitel 1.1 mit den in der Stochastik gebräuchlichen Sprach-Konventionen einführen. Eine konkrete Umsetzung der Begriffe und Konventionen zum Beispiel der Befragung von Studierenden nehmen wir in Kapitel 1.2 vor. In Kapitel 1.3 erweitern wir schließlich noch den Begriff der Erhebung.

1.1 Grundbegriffe

1.1.1 Grundgesamtheit, Stichprobe, Untersuchungseinheit

Eine der zentralen Entscheidungen einer statistischen Erhebung betrifft die Frage, *von wem* Eigenschaften erhoben werden sollen. Bei der Frage nach dem *von wem* haben sich drei Begriffe durchgesetzt, die wir als sprachliche Konvention aufnehmen und am Beispiel der Studierenden verdeutlichen:

- Die **Grundgesamtheit *G*** ist die Menge aller zu erhebender Studierenden. Allgemeiner gesprochen ist es die Menge der Objekte, die eine bestimmte, zu erhebende Eigenschaft haben. Mathematisch gesehen ist die Grundgesamtheit G eine Menge.
- Einen Teil aller Studierenden und allgemein einen Teil der Grundgesamtheit G bezeichnet man als **Stichprobe**. Die Stichprobe ist eine Teilmenge von G.
- Ein einzelner Student oder eine einzelne Studentin – allgemein ein Objekt mit einer bestimmten zu erhebenden Eigenschaft – wird neutral als **Untersuchungseinheit**, als **statistische Einheit** oder auch als **Merkmalsträger** bezeichnet. Eine Untersuchungseinheit ist ein Element der Grundgesamtheit beziehungsweise ein Element der Stichprobe.

Vordergründig ist damit alles festgelegt: die Studierenden einer bestimmten Hochschule. Wie oben beschrieben, müssen aber auch die Hochschule, die Zeit und der genaue Ort der Befragung festgelegt werden. Diese drei Aspekte bestimmen, *von wem* die interessierenden Eigenschaften erhoben werden sollen, nämlich die *sächliche, zeitliche* und *örtliche* Festlegung der Grundgesamtheit. Sie könnte im vorliegenden Fall beispielsweise so vorgenommen werden:

> Die Grundgesamtheit G besteht aus allen im Sommersemester 2010 (zeitliche Festlegung) eingeschriebenen Studierenden (sächliche Festlegung) der PH Freiburg (ausschließlich an der PH Freiburg eingeschrieben, örtliche Festlegung).

Der Sinn einer solchen Festlegung ist es, ein möglichst eindeutiges Kriterium zu erhalten, ob eine bestimmte Person (ein bestimmtes Objekt) hinsichtlich seiner Eigenschaft erhoben werden soll oder nicht. Die Stichprobe muss als Einschränkung der Grundgesamtheit in gleicher Weise sächlich, örtlich und zeitlich festgelegt werden.

1.1.2 Merkmale, Merkmalsausprägungen

Bisher haben wir stets von Eigenschaften von Studierenden gesprochen. Das können die Haarfarbe, das Geschlecht, das Alter oder auch die wöchentliche Zeit sein, die für musikalische Freizeitaktivitäten aufgewendet wird. Jede solche Eigenschaft hat in der Statistik einen festgelegten Namen und heißt **Merkmal** Ω. Mehrere Merkmale werden durch Indizes gekennzeichnet, also etwa Ω_1, Ω_2 usw.

Ein Merkmal Ω kann durch eine Umschreibung (z.B. Haarfarbe) festgelegt werden. Eindeutiger ist allerdings die Festlegung eines Merkmals über die sogenannten **Merkmalsausprägungen** ω, wobei verschiedene Merkmalsausprägungen wiederum durch einen Index gekennzeichnet werden. Nimmt man etwa das Geschlecht der Studierenden, so könnte man die Merkmalsausprägungen durch $\omega_1 = \textit{weiblich (W)}$ und $\omega_2 = \textit{männlich (M)}$ festlegen.

Das Merkmal Ω lässt sich mathematisch als Menge, festgelegt durch ihre Elemente, die Merkmalsausprägungen, fassen:

$$\text{Geschlecht: } \Omega = \{\omega_1, \omega_2\} = \{\text{weiblich, männlich}\} = \{\text{W,M}\}$$

Hinsichtlich eines Merkmals bezeichnet man das Paar, das aus einem Merkmalsträger (also einer Untersuchungseinheit) und einer Merkmalsausprägung besteht, als **statistisches Datum**. Wenn also von **statistischen Daten** gesprochen wird, ist eine Menge solcher Paare gemeint.

Wird ein Merkmal durch die gewünschten Merkmalsausprägungen festgelegt, kann dies zu einer Einschränkung der Merkmalsträger führen. Wird beispielsweise die Haarfarbe erhoben und die Merkmalsausprägungen *blond*, *braun*, *rot* und *schwarz* festgelegt, so ist die Person mit grünlila gefärbten Haaren nicht zuzuordnen. Solch ein Ausschluss kann dadurch umgangen werden, dass man die Merkmalsausprägung *Sonstige* zulässt.

Bestehen die Merkmalsausprägungen $\omega_1, \omega_2, \ldots$ aus Zahlen (z. B. Alter, Semesteranzahl, Abiturnote usw.), so kann man alle in der Stichprobe möglichen Merkmalsausprägungen über ein geeignetes Intervall erfassen. Beispielsweise könnte das Merkmal Ω: Alter der Studierenden durch $\Omega = [15;100] = \{\omega | 15 \leq \omega \leq 100\}$ festgelegt werden, also das Intervall, das alle potentiellen Merkmalsausprägungen von 15 bis 100 Jahre umfasst.

1.1.3 Eigenschaften von Merkmalen

An den bisher genannten Beispielen kann man schon erkennen, dass es unterschiedliche Merkmalstypen gibt. Zunächst können **kategoriale** von **numerischen** Merkmalen unterschieden werden: Kategoriale Merkmale haben Worte (oder Buchstaben) als Merkmalsausprägungen (z. B. das Geschlecht), numerische Merkmale werden durch Zahlen angegeben (z. B. das Alter).

Etwas systematischer werden die Merkmale durch die Art der **Skala**[2], auf der alle Merkmalsausprägungen abgebildet werden, charakterisiert. Man unterscheidet üblicherweise drei Skalierungsarten:

Nominalskalierung Sind die Merkmalsausprägungen Begriffe, die keiner Hierarchie genügen, so sind diese **nominalskaliert**.

Beispiel:
Das Merkmal Ω bezeichne das Geschlecht von Studierenden. Man betrachtet zu diesem Merkmal die Menge der Merkmalsausprägungen $\Omega = \{\text{W,M}\}$. Die Merkmalsausprägungen genügen keiner Hierarchie, da *W* für weiblich nicht über *M* für männlich steht (und umgekehrt).

In gleicher Weise können auch Religionszugehörigkeiten, Haarfarben, die Parteipräferenz oder etwa eine ausgeübte Sportart nicht in eine Hierarchie eingeordnet werden.

Ordinalskalierung Lassen sich die Merkmalsausprägungen hierarchisch anordnen, während aber die Abstände zwischen den Merkmalsausprägungen nicht definiert sind, so sind die Merkmalsausprägungen **ordinalskaliert**.

Beispiel:
Als Maß für die Leistung in Mathematik können die Zensuren in dem Punktesystem der Sekundarstufe II ($\Omega = \{0, 1, ..., 15\}$) betrachtet werden. Hier ist keine Hierarchie vorhanden, denn 10 Punkte bezeichnen zwar eine bessere Leistung als 5 Punkte, allerdings ist der Abstand zwischen den einzelnen Merkmalsausprägungen nicht einheitlich definiert. So kann der Leistungsabstand zwischen 5 und 10 Punkten sich vom Leistungsabstand zwischen 10 und 15 Punkten unterscheiden, obwohl dieser numerisch beide Male 5 Punkte beträgt. Es gibt zwar Versuche, die Leistung von Schülerinnen und Schülern *eindeutig* zu messen, aber das könnte höchstens dann funktionieren, wenn genau festgelegt werden kann, was unter Leistung zu verstehen ist. Dies ist aber bei Schulleistungen im Allgemeinen nicht sinnvoll möglich. Aussagen wie „10 Punkte sind eine doppelt so gute Leistung wie 5 Punkte“ stimmen daher in der Regel nur numerisch, nicht aber hinsichtlich der tatsächlich erbrachten Leistungen.

Noch deutlicher wird das Vorliegen einer hierarchischen Skala ohne fest definierte Abstände beispielsweise bei Spargelklassen. Auch hier kann zwar davon gesprochen werden, dass nach bestimmten Kriterien die Spargelklasse I besser ist als Spargelklasse II, aber es gibt keinen festgelegten Abstand zwischen diesen beiden Spargelklassen, der zudem identisch wäre zum Abstand zwischen den Spargelklassen II und III.

Metrische Skalierung Lassen sich die Merkmalsausprägungen hierarchisch anordnen und sind zudem die Abstände zwischen den Merkmalsausprägungen definiert, so sind die Merkmalsausprägungen **metrisch skaliert**.

[2] Wir behandeln die Skala als grundlegende Eigenschaft eines Merkmals, obwohl Merkmale, wie etwa die Wassertemperatur, auch mit unterschiedlichen Messinstrumenten „gemessen“ werden können, etwa mit einem Thermometer oder der Hand.

Beispiel:
Alle Anzahlen (z. B. Semesteranzahlen) und alle physikalisch *messbaren* Größen (z. B. Länge, Zeitdauer, Flächeninhalt etc.) sind metrisch skaliert. Physikalisch *messbar* heißt: Es wurde ein Maß so geschaffen, dass verschiedene Objekte hinsichtlich einer physikalischen Eigenschaft verglichen werden können. So gilt beispielsweise unabhängig vom Betrachter und vom Ort, dass ein Objekt mit der Länge von 2 Metern doppelt so lang ist wie ein Objekt mit der Länge von 1 Meter und dass der Unterschied zwischen den beiden Objekten genauso groß ist wie zwischen zwei Objekten mit den Längen 5 und 6 Meter.

Die metrische Skalierung lässt sich noch weiter unterteilen:

1. Intervallskala: Hat ein Merkmal (z. B. die Temperatur, gemessen in °C) numerische Merkmalsausprägungen und werden zusätzlich identische Abstände (Intervalle) gleich interpretiert, so handelt es sich um eine Intervallskalierung. Ein typisches Beispiel dafür sind Temperaturunterschiede: Beispielsweise ist der Temperaturunterschied von 0°C zu 10°C der gleiche wie von 10°C zu 20°C.
2. Verhältnisskala: Wenn ein Merkmal intervallskaliert ist und sich zusätzlich die Verhältnisse zweier Merkmalsausprägungen sinnvoll interpretieren lassen, so liegt eine Verhältnisskala vor. Das lässt sich am Beispiel von Längen veranschaulichen: So ist z. B. das Verhältnis von 20 m zu 10 m sinnvoll interpretierbar, weil die erste Länge *doppelt so lang* wie die zweite ist. Im Gegensatz dazu ist die Aussage „20°C ist doppelt so warm wie 10°C“ unsinnig. Der entscheidende Punkt ist, dass bei Längen der Nullpunkt (Länge 0, d.h. keine Längenausdehnung) unmittelbar gegeben ist, während der Nullpunkt bei der °Celsius-Temperaturskala willkürlich gewählt wurde. Die Kelvinskala, die vom absoluten Temperaturnullpunkt 0 K ausgeht[3], ist dagegen wiederum eine Verhältnisskala.
3. Absolutskala: Besteht eine Verhältnisskala und ist zusätzlich die Maßeinheit natürlich festgelegt, wie etwa bei Anzahlen (z. B. 3 Stück), so liegt eine Absolutskala vor. Bei Längen trifft dies nicht zu, da hier beispielsweise eine Maßeinheit wie der Meter willkürlich definiert wurde.

Merkmale mit numerischen Merkmalsausprägungen Wenn im Folgenden die Daten numerisch oder sogar metrisch skaliert sind, dann werden die Merkmale statt $\Omega_1, \Omega_2, \ldots$ mit $X_1, X_2, \ldots$, beziehungsweise die Merkmalsausprägungen statt mit $\omega_1, \omega_2, \ldots$ mit $x_1, x_2, \ldots$ bezeichnet.

Auch kategoriale Daten, z. B. das Geschlecht, können prinzipiell in numerische Daten transformiert werden. Dazu ordnet man jeder Merkmalsausprägung ω_i genau eine reelle Zahl zu, d.h. man wendet eine Funktion (X) auf die Menge aller Merkmalsausprägungen an:

$$X : \begin{cases} \Omega & \to & \mathbf{R} \\ \omega_i & \to & x_j \quad i = 1, \ldots, s; \quad j = 1, \ldots, r \end{cases}$$

[3] Dies entspricht auf der °Celsius-Temperaturskala $-273,15$°C

Im Prinzip erhält man durch die Funktion X (die in der Wahrscheinlichkeitsrechnung Zufallsgröße genannt wird) ein neues Merkmal mit eben numerischen Merkmalsausprägungen, repräsentiert durch die Wertemenge der Funktion X. Die Skalierung der Merkmale ändert sich durch die Transformation in Zahlen aber nicht. Im Folgenden wird auch bei der Einführung der einzelnen statistischen Methoden, wenn nicht explizit anders erforderlich, die Notation X und x_i für Merkmale und Merkmalsausprägungen verwendet.

1.1.4 Repräsentativität

Obwohl die **Repräsentativität** kein mathematisch definierter Begriff ist, so ist sie dennoch von grundlegender Bedeutung für die Statistik.

Mit der *Repräsentativität einer Stichprobe* wird zum Ausdruck gebracht, dass eine Stichprobe die gleichen Eigenschaften aufweisen sollte, wie die Grundgesamtheit, aus der sie entnommen wurde. Das bedeutet insbesondere, dass die in der Stichprobe erhobenen Untersuchungseinheiten zu allen ihren Merkmalen die gleiche *Häufigkeitsverteilung* (siehe Kap. 2.1) aufweisen sollten wie die Grundgesamtheit. Bezogen auf das Beispiel einer Studierenden-Stichprobe bedeutet das etwa, dass bei 45% männlichen Studierenden einer Hochschule auch zumindest annähernd 45% der Studierenden in der Stichprobe männlich sein sollten. Das sollte nach Möglichkeit auch für alle weiteren Merkmale gelten (wie z. B. das Alter, die Semesteranzahl, das Körpergewicht, die Haarfarbe usw.), die im Verlauf der statistischen Untersuchung bedeutsam werden.

Hier wird unmittelbar einsichtig, dass die Repräsentativität ein theoretischer Begriff ist, der zwar mit bestimmten Methoden angenähert, aber nie nachgewiesen werden kann. Es ist unmöglich, tatsächlich alle Merkmale von Studierenden zu benennen, geschweige denn diese zu erheben. Selbst wenn man sich auf offensichtlich oder theoretisch beeinflussende Merkmale beschränkt, wird man an Grenzen stoßen. So könnte beispielsweise das Merkmal Wohnung, die Eigenschaft, bei den Eltern zu wohnen oder nicht, vom Alter abhängen: Jüngere Studierende wohnen vielleicht eher bei den Eltern als ältere. Vielleicht wohnen auch Studentinnen eher bei den Eltern als Studenten (oder umgekehrt). Ebenso könnte aber auch der finanzielle Status der Eltern, die Beziehung zu den Eltern, das Verhalten von Freunden o.ä. das Wohnverhalten der Studierenden beeinflussen. So scheint es auch in diesem Beispiel weder möglich, *alle* beeinflussenden Merkmale zu benennen, noch diese zu messen. Daher sind die Ergebnisse einer Datenanalyse immer auch kritisch hinsichtlich der Repräsentativität zu hinterfragen: Beispielsweise werden die Umfrageergebnisse unter den Lesern dieses Buches zu mathematischen Kenntnissen nicht repräsentativ für die Gesamtbevölkerung sein, da bereits die Bereitschaft, dieses Buch zu lesen, eine gewisse Affinität zur Mathematik voraussetzt.

Eine wichtige Methode, mit der die Repräsentativität in einer (möglichst umfangreichen) Stichprobe angenähert werden kann, ist die Ziehung einer **Zufallsstichprobe**. Im Studierenden-Beispiel bedeutet das, dass nicht die Erhebenden selbst, sondern der Zufall bestimmt, wer von den Studierenden befragt wird. Die Festlegung solch einer Stichprobe wäre möglich, indem zufällig eine bestimmte Anzahl von Studierenden nach folgendem Verfahren festgelegt wird: Es werden wiederholt Zufallszahlen mit sieben Ziffern erzeugt. Trifft eine dieser Zufallszahlen eine eindeutig zugeordnete siebenstellige Matrikelnummer, so wird die betreffende Person ausgewählt. Dieses Verfahren wird so lange wiederholt, bis die gewünschte Anzahl von Studierenden beisammen ist. Wenn die so bestimmte Stichprobe groß genug ist, kann man einerseits davon

ausgehen, dass tatsächlich jedes beliebige Merkmal der Studierenden zufällig in der Stichprobe vorkommt und andererseits auch die Häufigkeitsverteilung eines Merkmals in der Stichprobe und der Grundgesamtheit zumindest annähernd gleich ist.[4] Dieses Verfahren ist allerdings aufwendig und im Falle von Grundgesamtheiten, die nicht systematisch aufgelistet sind (wie z. B. alle Deutschen), praktisch nicht durchführbar.

Ist eine Zufallsstichprobe auf Grund pragmatischer oder sächlicher Zwänge nicht durchführbar, so kann man bei sogenannten **Beurteilungsstichproben** zwar eine zumindest annähernde Repräsentativität einer Stichprobe postulieren, diese aber theoretisch kaum belegen. Die Konstruktion einer Beurteilungsstichprobe bedeutet, dass aufgrund des theoretischen Wissens zur Grundgesamtheit, das vor der Erhebung vorhanden ist, eine Stichprobe theoretisch konstruiert wird.[5]

1.2 Eigenschaften von Studierenden

Wir haben für die folgenden Kapitel dieses Buches einen umfangreichen Datensatz zu Grunde gelegt, der verschiedene Merkmale von Studierenden umfasst und in verschiedenen Semestern an der Universität Münster und der Pädagogischen Hochschule Freiburg (dort 2009 und 2010) erhoben wurde.[6] Die Daten wurden jeweils am Beginn eines Semesters von Studierenden mit keinen oder geringen Stochastik-Kenntnissen erhoben. Die Merkmale sind soweit wie möglich einheitlich festgelegt, ohne Interpretationsspielräume zuzulassen. Manche Merkmale enthalten dennoch diesen Interpretationsspielraum. So muss natürlich streng genommen definiert werden, ab wann etwa eine Person sich zu den Nicht-Singles oder zu den Rauchern zählt, wenn man nach dem Beziehungsstatus oder der Rauchgewohnheit fragen möchte, wie bei dieser Erhebung. Dies ist ein Beispiel dafür, dass nicht alle Merkmalsfestlegungen so vorgenommen wurden, dass sie die den strengen Kriterien von „guten Daten“ genügen. Da sich jedoch die stochastischen Methoden an diesem Beispiel gut erläutern lassen und dies für die in den folgenden Kapiteln diskutierten statistischen Methoden kaum Auswirkungen hat – die Daten in dieser Hinsicht tatsächlich also hinreichend „gut“ sind – arbeiten wir anhand ausgewählter Fragestellungen damit. Sollten kritische Stellen auftreten, werden wir direkt darauf hinweisen.

Was wird erhoben? Der Datensatz umfasst folgende Merkmale:

Merkmal	Skalierung
Geschlecht, $\Omega_1 = \{W, M\}$ für männlich und weiblich	nominal
Alter in Jahren, $\Omega_2 = \{15, 16, ..., 100\}$	metrisch
Hochschulsemester in einem Fach, $\Omega_3 = \{1, 2, ..., 50\}$	metrisch

[4] Wir werden die Frage nach Schätzungen von Häufigkeiten in der Grundgesamtheit in Kapitel 5.4 noch einmal anhand unserer Stichproben von Studierenden besprechen.

[5] Für weiterführende Fragen, z. B. wie bei einer Beurteilungsstichprobe Qualitätsstandards zu erfüllen sind, verweisen wir an dieser Stelle auf Hartung et al., 2009.

[6] Dieser Datensatz ist vollständig auf unserer Homepage http://www.leitideedatenundzufall.de sowie in den Online-Materialien des Verlags zu erhalten (siehe auch Vorwort).

Studienfach, $\Omega_4 = \{\text{Jura, BWL, Medizin, ...}\}$	nominal
Abiturnote, $\Omega_5 = \{1,0;1,1;...;4,0\}$	ordinal
BAföG, $\Omega_6 = \{J,N\}$ für ja und nein	nominal
Zufriedenheit mit dem Studium, $\Omega_7 = \{1,2,3,4,5\}$ für 1: sehr zufrieden bis 5: sehr unzufrieden	ordinal
Mensagänge durchschnittlich pro Woche (in der Vorlesungszeit), $\Omega_8 = \{0,1,...,50\}$	metrisch
Wohnung, $\Omega_9 = \{E,\overline{E}\}$ für bei den Eltern und nicht bei den Eltern	nominal
Entfernung vom Wohnort zur Hochschule in km, $\Omega_{10} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 1000\}$	metrisch
Zeit für den Weg zur Hochschule in min, $\Omega_{11} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 600\}$	metrisch
Beförderungsmittel (Wie), das überwiegend verwendet wird, $\Omega_{12} = \{P,F,B,DB,A\}$ für zu Fuß (P), Fahrrad (F), Bus/Staßenbahn (B), Zug (DB), Auto (A)	nominal
Rauchverhalten, $\Omega_{13} = \{R,\overline{R}\}$ für Raucher und Nichtraucher	nominal
Zeit, die wöchentlich mit Sporttreiben zugebracht wird (Sport) in Stunden, $\Omega_{14} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 168\}$	metrisch
Zeit, die wöchentlich mit Musizieren (aktiv) zugebracht wird (Musik) in Stunden, $\Omega_{15} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 168\}$	metrisch
Zeit, die wöchentlich mit bezahlter Arbeit zugebracht wird (Arbeit) in Stunden, $\Omega_{16} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 168\}$	metrisch
Zeit, die wöchentlich vor dem Computer zugebracht wird (Computer) in Stunden, $\Omega_{17} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 168\}$	metrisch
Zeit, die wöchentlich mit ehrenamtlicher Arbeit zugebracht wird (Ehrenamt) in Stunden, $\Omega_{18} = \{\omega \in \mathbb{Q} \mid 0 \leq \omega \leq 168\}$	metrisch
Beziehungsverhalten, $\Omega_{19} = \{S,\overline{S}\}$ für Single und Nicht-Single	nominal
Parteipräferenz, $\Omega_{20} = \{\text{CDU, FDP, Grüne, Linke, SPD, ...}\}$	nominal

Von wem wird etwas erhoben? Wie im vorangegangenen Abschnitt bereits angedeutet, beruht der Datensatz zu den Studierenden auf folgender Festlegung der Grundgesamtheiten:

> Die Grundgesamtheit G besteht aus allen im Sommer/Wintersemester x (zeitliche Festlegung) eingeschriebenen Studierenden (sächliche Festlegung) der Hochschule y (örtliche Festlegung).

Die für dieses Buch vorliegenden Stichproben basieren auf einer Beurteilungsstichprobe, die folgenden Modellannahmen und Erhebungsfestlegungen genügt:

- Alle Studierenden sind gleichmäßig an der Hochschule präsent. Das wird so sicher nicht der Realität entsprechen, da es Studierende mit unterschiedlicher Präsenz gibt und z.B. Studierende, die ihren Abschluss vorbereiten, seltener an der Hochschule anzutreffen sein könnten als solche, die einen regelmäßigem Studienbetrieb nachgehen.

- Um eine möglichst annähernd repräsentative Stichprobe zu erhalten, müssen die Orte und die Zeiten variieren. Das kann an einer Campus-Hochschule wie der Pädagogischen Hochschule Freiburg hinreichend gewährleistet werden, an Flächenhochschulen wie der Universität Münster ist dies weniger gleichmäßig möglich.

In welcher Form diese Festlegungen zumindest annähernd einer Repräsentativität (vgl. Kap. 1.1) genügen können, werden wir exemplarisch in Kapitel 8 untersuchen. Aber selbst wenn sich dabei Hinweise auf die nicht erreichte Repräsentativität ergeben sollten, muss man in Betracht ziehen, dass viele wissenschaftliche Erkenntnisse auf Beurteilungsstichproben beruhen und dennoch diese Untersuchungen Entwicklungen steuern. Das entscheidende Kriterium ist die Offenlegung der Umstände, unter denen die Daten erhoben wurden: Wir werden bei der Entwicklung der stochastischen Methoden im Folgenden mit der impliziten Annahme arbeiten, mit einer Umfrage durch Novizen annehmbar „gute Daten“ erzeugt zu haben.

1.3 Ergänzungen

Es gibt verschiedene Möglichkeiten, Erhebungsarten zu klassifizieren. Wir beschränken uns auf die Betrachtung von

- Befragungen,
- Beobachtungen und
- Experimente.

Zu jeder dieser Klassen gibt es unzählige Beispiele. Wir nennen eine sehr kleine Auswahl von Beispielen und werden die bisher genannten Kriterien für „gute Daten“ exemplarisch anwenden.

Befragung	Beobachtung	Experiment
Eigenschaften von Studierenden	Wettermessung	Freier Fall (Galileo Galilei, 1623)
Parteipräferenz („wenn am kommenden Sonntag Bundestagswahl wäre ...“)	Verkehrszählung	Wirkung eines Medikaments
Demoskopie (Volkszählung 2011)	Intelligenzmessung	Interventionsstudie in den Erziehungswissenschaften
Präferenz für ein kommerzielles Produkt	Ethogramme: Verhaltensbeobachtung von Tieren in ihrer natürlichen Umgebung	Crash-Test
...	...	...

Was wird erhoben: Die Erhebung zu allen genannten Beispielen basiert im Optimalfall auf der *eindeutigen* Festlegung der interessierenden Merkmale. Da wir die Erhebung unter Studierenden bereits in Kapitel 1.2 behandelt haben, greifen wir für die Beobachtung und das Experiment jeweils ein Beispiel heraus:

- Beobachtung: Erhebt man die Tagestemperatur, so ist zu definieren, was als Tagestemperatur zu verstehen ist. Bis März 2001 war die *Tagesmitteltemperatur* vom Deutschen Wetterdienst durch das arithmetische Mittel der Messungen um 07:30, 14:30 und 21:30 Uhr MEZ definiert, wobei der letzte Messwert doppelt in die Berechnung einging. Seither ist die Tagesmitteltemperatur durch die arithmetische Mittelwertbildung aller zur vollen Stunde gemessenen Temperaturwerte eines Tages festgelegt. Solche Festlegungen ermöglichen die Vergleichbarkeit der Messung verschiedener Tage, aber auch verschiedener Wetterstationen. In vergleichbarer Weise wären auch bei allen anderen Eigenschaften die zu erhebenden Merkmale zu definieren. Dass dies nicht immer einheitlich möglich ist, zeigt etwa das Beispiel der Intelligenzmessung, bei der unterschiedliche Konstrukte, was eigentlich Intelligenz ist, miteinander konkurrieren (Zimbardo & Gerrig, 2004).
- Experiment: Bei einer Interventionsstudie, etwa zur Wirksamkeit einer bestimmten Lehrmethode, muss genau festgelegt werden, was als Wirksamkeit verstanden wird (z.B. Punkte in einem Test). Es gibt dazu unterschiedliche Möglichkeiten der Festlegung. In der Regel wird in Experimenten, zu denen eine Interventionsstudie unter gewissen Voraussetzungen zählen kann, versucht, den Realitätsausschnitt möglichst so zu verengen, dass ein mögliches, potentiell zu veränderndes Merkmal isoliert im Zusammenhang mit Einfluss nehmenden Variablen betrachtet werden kann. Bei einfachen physikalischen Experimenten, wie etwa zum freien Fall, ist dies im Allgemeinen besser möglich als bei experimentellen Untersuchungen der Psychologie oder der Soziologie.

Von wem wird etwas erhoben: Wie in Kapitel 1.1 ausgeführt, berührt dieser Aspekt die sächliche, örtliche und zeitliche Festlegung der Grundgesamtheit wie auch der Stichprobe.

- Im Beispiel der Wetterbeobachtung ist die zeitliche Festlegung durch die mittlere Tagestemperatur (s.o.) und die örtliche durch verschiedene Wetterstationen gegeben. Die sächliche Festlegung erfolgt dadurch, dass bei der Messung der mittleren Tagestemperatur nur die Lufttemperatur in 2 m Höhe über dem Erdboden an einem Ort berücksichtigt wird, der nicht weiteren Einflüssen durch andere Wärmequellen oder direkter Sonneneinstrahlung ausgesetzt ist. Dadurch wird ein *Beobachtungsausschnitt* möglichst eindeutig festgelegt, wiederum, um die Vergleichbarkeit von Ergebnissen zu ermöglichen.
- Bei der Wirksamkeit einer Interventionsstudie müssen die Grundgesamtheit (etwa eine Klassenstufe) und die Stichprobe festgelegt werden. Bei der Stichprobe wird in der Regel die randomisierte, also zufällige Zuordnung von Teilnehmern zu einer Schulungsgruppe und einer Kontrollgruppe vorgenommen. Die Randomisierung ist ein zentraler Aspekt der Güte experimenteller Studien, die gewährleisten soll, dass alle womöglich nicht benannten oder bekannten Einflussgrößen auf das experimentelle Ergebnis zufällig aufgeteilt werden.

Wie wird etwas erhoben: Hier müssen verallgemeinernde Beschreibungen scheitern. So ist die Konstruktion von Messinstrumenten, etwa eines Thermometers, eines Intelligenzstests oder eines Fragebogens für die Ermittlung einer Produktpräferenz kaum noch miteinander zu vergleichen. Als zusammenfassender Aspekt ist zu nennen, dass mögliche Störeinflüsse theoretisch fundiert nach Möglichkeit ausgeschlossen werden. Dies ist im Allgemeinen bei einem kontrollierten Experiment weitaus effizienter möglich als bei einer Beobachtung im Feld.

1.4 Aufgaben

Aufgabe 1.1: Gegeben ist ein Teil der Abschlusstabelle der Fussball-Bundesliga aus der Saison 2009/10. Ordnen Sie jedem Merkmal eine Skalierungsart zu und begründen Sie Ihre Antwort.

Platz	Verein	Spiele	Sieg	Unentsch.	Niederlage	Tore	Gegentore	Punkte	Kommentar
1	Bayern München	34	20	10	4	72	31	70	CL
2	FC Schalke 04	34	19	8	7	53	31	65	
3	Werder Bremen	34	17	10	7	71	40	61	CL-Qual.
4	Bayer Leverkusen	34	15	14	5	65	38	59	EL
5	Borussia Dortmund	34	16	9	9	54	42	57	
6	VfB Stuttgart	34	15	10	9	51	41	55	Mittelfeld
7	Hamburger SV	34	13	13	8	56	41	52	
8	VfL Wolfsburg	34	14	8	12	64	58	50	

Aufgabe 1.2: Finden Sie zur Nominalskalierung, Ordinalskalierung und metrischen Skalierung Merkmale, die nicht in den vorangegangenen Kapiteln behandelt wurden. Begründen Sie, dass die jeweilige Skalierungsart zutrifft.

Aufgabe 1.3: Planen Sie eine statistische Erhebung zu

- den Fernsehgewohnheiten von Schülerinnen und Schülern.
- der Anziehungseigenschaft einer blühenden Pflanze auf Honigbienen.
- den Flugeigenschaften eines Papierfliegers.

Überlegen Sie sich zusätzlich, welche Abänderungen Ihrer Planung das Ergebnis erheblich beeinflussen könnten.

2 Analyse statistischer Daten zu einem Merkmal

Einstiegsbeispiel

Abbildung 2.1: Studentinnen und Studenten im Hörsaal

Aufgabe 1: Wählen Sie Merkmale der Studierenden-Stichprobe einer Hochschule aus und analysieren Sie diese. Formulieren Sie am Ende eine möglichst plakative Aussage zu Ihren Ergebnissen.

Worum es geht

Sind die Daten zu einem Merkmal gesammelt, so liegen diese als Zahlen- oder Buchstabenkolonnen vor. Um die Daten interpretieren zu können, müssen diese aufbereitet werden. Dabei werden in der Regel die in den Daten enthaltenen Einzelinformationen reduziert, so dass wesentliche Muster in den Daten deutlicher hervortreten. Der zentrale Gedanke statistischen Arbeitens ist dabei, ein mathematisches Ergebnis hinsichtlich eines solchen Musters, das durch statistische Methoden gewonnen wird, zu interpretieren und dabei in der Regel auf eine reale Problemstellung zu beziehen: Hier entscheidet sich, ob die Ausgangsfrage, die zur Datenerhebung und -aufbereitung geführt hat, eine befriedigende Antwort erhält. Wesentliche informationsreduzierende Methoden sind:

Häufigkeiten Eine grundlegende Methode ist das Auszählen von Merkmalsausprägungen, also das Bestimmen ihrer *Häufigkeit* und ihres Anteils an einer Stichprobe. Untersucht man beispielsweise die Singles in der vorliegenden Studenten-Stichprobe, so kann man bestimmen, wie viele Singles in der Stichprobe vorhanden sind oder welcher Anteil der Stichprobe Singles sind. Man vernachlässigt aber in aller Regel die Information darüber, wer eigentlich genau die einzelnen Singles (Merkmalsträger mit Merkmalsausprägung „Single“) sind.

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_2,

Grafische Darstellung Eine grafische Darstellung sollte einen Datensatz übersichtlich, aber sachgerecht repräsentieren. Eine einfache grafische Aufbereitung basiert auf Häufigkeiten von Merkmalsausprägungen. In dieser Darstellungsform wird die in den Daten enthaltene Information häufig dadurch reduziert, dass Merkmalsausprägungen *klassiert* (d.h. zusammengefasst) werden. Die grafische Darstellung zur Abiturnote der Studierenden kann beispielsweise dadurch übersichtlicher werden, dass nicht die einzelnen Nachkommastellen betrachtet werden, sondern die Kommanoten den verschiedenen Stufen in ganzen Noten zugeordnet werden.

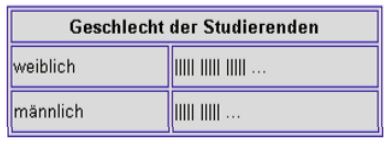
Lage- und Streuparameter Eine nichtgrafische Reduktion der Daten findet durch die Berechnung sogenannter *Lage-* und *Streuparameter* zu einem Merkmal statt. Der Datensatz wird unabhängig davon, ob es sich um 2 oder 2 Millionen Daten handelt, durch einen oder wenig mehr berechnete statistische Werte zusammengefasst und repräsentiert. Es geht um Verfahren, mit denen ein Zentrum der *Häufigkeitsverteilung* eines Merkmals identifiziert und Abweichungen von diesem Zentrum charakterisiert werden können. Diese mathematische Vorgehensweise ist nur sinnvoll bei eingipfligen Häufigkeitsverteilungen anwendbar, da nur bei diesen sinnvoll von *einem* Zentrum und den Abweichungen von diesem Zentrum gesprochen werden kann.

Diese drei Aspekte der Auswertung werden in einzelnen Teilkapiteln anhand kleiner Datensätze aus der Gesamtstichprobe zu den Studierenden getrennt erläutert (Kap. 2.1 – Kap. 2.5). In Kapitel 2.6 wird eine Stichprobe exemplarisch ausgewertet und in Kapitel 2.7 werden fachliche Ergänzungen gegeben. Das Abschlusskapitel (Kap. 2.8) enthält Übungsaufgaben. Die in diesem Kapitel diskutierten Beispiele beziehen sich stets – wenn nicht explizit anders festgelegt – auf den in Tabelle 1 enthaltenen Datensatz zu Teilstichproben der Studierenden aus Freiburg und Münster. Bei den abgebildeten Freiburg-Daten ist *jedes* Merkmal der Größe nach sortiert, wodurch eine Zeile nicht einen Studierenden, sondern mehrere unterschiedliche repräsentiert.

2.1 Erste Ordnung und Häufigkeiten

2.1.1 Erste Ordnung

Eine erste Ordnung der statistischen Daten kann bereits in der Phase der Erhebung stattfinden, indem man die Daten in einer **Urliste** aufnimmt: Die Anzahl der Merkmalsträger n (im Folgenden stets als Umfang der Stichprobe bezeichnet) zu den Ausprägungen eines Merkmals wird in einer Strichliste notiert. Werden mehrere Merkmale erhoben, wird in der Regel keine Urliste, sondern eine **Tabelle** erstellt, in der die Merkmalsausprägungen fallweise eingetragen werden. Diese Tabelle repräsentiert die Daten fallbezogen (vgl. Abb. 2.2).



Nr.	Geschlecht	Nationalität	Alter	Semester
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				

Abbildung 2.2: Ausschnitt aus der Erhebungstabelle zur Erhebung der Merkmale von Studierenden

PH Freiburg

Nr.	Alter	Semester	Abi	Entfernung	Raucher	Partei
1	19	1	1,3	1	j	CDU
2	20	2	1,4	1	j	CDU
3	21	2	1,4	1	j	CDU
4	21	2	1,6	2	j	CDU
5	22	2	1,7	2	j	FDP
6	22	2	1,8	2	j	FDP
7	22	3	1,9	2	j	Grüne
8	22	3	2	3	j	Grüne
9	22	3	2,1	3	j	Grüne
10	22	3	2,1	3,5	j	Grüne
11	22	4	2,2	3,5	j	Grüne
12	22	4	2,2	4	j	Grüne
13	22	4	2,2	4	j	Grüne
14	23	4	2,3	4	j	Grüne
15	23	4	2,3	5	j	Grüne
16	23	4	2,3	5	j	Grüne
17	23	4	2,4	5	j	Grüne
18	23	5	2,4	5,5	j	Grüne
19	23	5	2,4	7	n	Grüne
20	23	5	2,5	7	n	Grüne
21	24	5	2,5	7	n	Grüne
22	24	6	2,5	7,5	n	Grüne
23	24	6	2,5	8	n	Grüne
24	24	6	2,5	8	n	Grüne
25	24	6	2,5	10	n	Grüne
26	25	6	2,5	10	n	Grüne
27	25	6	2,6	10	n	Grüne
28	26	7	2,6	10	n	Grüne
29	26	7	2,6	15	n	Grüne
30	26	7	2,7	15	n	Grüne
31	28	7	2,7	15	n	Grüne
32	28	8	2,8	15	n	Linke
33	28	8	2,8	16	n	Linke
34	29	8	2,9	16	n	Linke
35	31	8	2,9	25	n	Linke
36	34	8	3	35	n	Sonst
37	35	8	3	40	n	Sonst
38	37	9	3,1	50	n	SPD
39	40	9	3,1	60	n	SPD
40	49	13	3,2	80	n	SPD

Uni Münster

Nr.	Alter	Raucher
1	19	n
2	19	n
3	20	n
4	21	n
5	21	n
6	21	j
7	21	n
8	21	n
9	22	n
10	22	j
11	23	n
12	23	j
13	24	j
14	25	j
15	25	n

Tabelle 1: Eigenschaften von Studierenden der PH Freiburg und der Uni Münster

2.1.2 Häufigkeiten

Die Auszählung der Urliste oder Tabelle führt zum Begriff der Häufigkeit:

Definition 1
Die Anzahl der Merkmalsträger mit der Merkmalsausprägung $x_i, (i = 1, 2, ..., s)$ zu einem Merkmal X in einer Stichprobe mit dem Umfang n heißt **absolute Häufigkeit** $H_n(x_i)$.

Daraus folgt unmittelbar:

Satz 1
Sei s die Anzahl der verschiedenen Merkmalsausprägungen zu einem Merkmal X, dann ist die Summe der absoluten Häufigkeiten aller Merkmalsausprägungen gleich dem Umfang der Stichprobe n.

$$H_n(x_1) + H_n(x_2) + ... + H_n(x_s) = \sum_{i=1}^{s} H_n(x_i) = n$$

Betrachtet man verschiedene Merkmalsausprägungen eines Merkmals in *einer* Stichprobe, so sind bereits die absoluten Häufigkeiten ausreichend für einen sinnvollen Vergleich.

Beispiel:
Unter $n = 15$ im Sommersemester 2009 an der Uni Münster erhobenen Studierenden (Teilstichprobe, Tabelle 1) befinden sich 5 Raucher und 10 Nichtraucher. Bezeichnet man mit $x_1 = R$ die Raucher und mit $x_2 = \overline{R}$ die Nichtraucher, so ist also

$$H_{15}(R) = 5 \text{ und } H_{15}(\overline{R}) = 10; \qquad H_{15}(R) + H_{15}(\overline{R}) = 5 + 10 = 15 = n$$

Mit dem absoluten Vergleich dieser beiden Häufigkeiten wird unmittelbar deutlich, dass in der Stichprobe mehr Raucher als Nichtraucher vorhanden sind, nämlich doppelt so viele.

Betrachtet man dagegen ein Merkmal bzw. Merkmalsausprägungen auf der Basis unterschiedlich großer Stichproben, so ist der Vergleich der absoluten Häufigkeiten irreführend.

Beispiel:
Unter den $n_1 = 15$ im Sommersemester 2009 an der Uni Münster erhobenen Studierenden befinden sich 5 Raucher. Unter $n_2 = 40$ im Sommersemester 2010 an der Pädagogischen Hochschule Freiburg erhobenen Studierenden befinden sich dagegen 18 Raucher. Bezeichnet man mit R_F die Raucher aus Freiburg und mit R_M die Raucher aus Münster, so gilt $H_{15}(R_M) = 5 < 18 = H_{40}(R_F)$. Eine Aussage, dass in Freiburg mehr als dreimal so viele Studierende rauchen, ist zwar bezogen auf die Stichproben (absolut betrachtet) richtig, dennoch aber irreführend, da in der Teilstichprobe aus Freiburg wesentlich mehr Studierende enthalten sind. Der Vergleich der absoluten Häufigkeiten ist in diesem Fall nicht aussagekräftig.

Der Vergleich von Merkmalen in unterschiedlich großen Stichproben führt auf die Betrachtungen von **relativen Häufigkeiten**.

Definition 2
Der Quotient der absoluten Häufigkeit $H_n(x_i)$ einer Merkmalsausprägung und dem Umfang n der Stichprobe heißt **relative Häufigkeit** $h_n(x_i)$:

$$h_n(x_i) = \frac{H_n(x_i)}{n}, \quad (i = 1, 2, ..., s)$$

Beispiel:
Unter den $n_1 = 15$ im Sommersemester 2009 an der Uni Münster befragten Studierenden befinden sich 5 Raucher. Unter den $n_2 = 40$ im Sommersemester 2010 an der Pädagogischen Hochschule Freiburg befragten Studierenden befinden sich dagegen 18 Raucher. Bezeichnet man mit R_F die Raucher aus Freiburg und mit R_M die Raucher aus Münster, so gilt

$$h_{15}(R_M) = \frac{5}{15} \approx 0,33 < h_{40}(R_F) = \frac{18}{40} = 0,45$$

Im Vergleich der beiden Stichproben ist damit der Anteil der Raucher in Freiburg deutlich höher als in Münster. Inwieweit das verallgemeinert eine Aussage zum Rauchverhalten an beiden Hochschulen ergibt, werden wir in Kapitel 3.5 noch betrachten und belassen es hier bei dem ersten deskriptiven Vergleich.

Wie bei den absoluten Häufigkeiten lässt sich auch die additive Zusammenfassung der relativen Häufigkeiten aller Merkmalsausprägungen eines Merkmals betrachten:

Satz 2
Sei s die Anzahl der verschiedenen Merkmalsausprägungen zu einem Merkmal X. Dann ist die Summe der relativen Häufigkeiten aller Merkmalsausprägungen gleich 1:

$$h_n(x_1) + h_n(x_2) + ... + h_n(x_s) = \sum_{i=1}^{s} h_n(x_i) = 1$$

Dieser Satz wird unmittelbar aus folgender Überlegung klar:

$$\sum_{i=1}^{s} h_n(x_i) = \sum_{i=1}^{s} \frac{H_n(x_i)}{n} = \frac{1}{n} \cdot \sum_{i=1}^{s} H_n(x_i) = \frac{1}{n} \cdot n = 1$$

2.1.3 Klassen

Die Anzahl der verschiedenen Merkmalsausprägungen zu einem Merkmal X macht unter Umständen eine Zusammenfassung von Merkmalsausprägungen in r Klassen k_j $(j = 1, ..., r)$ notwendig oder sinnvoll. Eine Klassierung der Merkmalsausprägungen ist beispielsweise dann sinnvoll, wenn es sehr viele unterschiedliche Merkmalsausprägungen mit jeweils kleinen absoluten Häufigkeiten gibt.

Beispiel:
Misst man die Entfernungen, die die Studierenden von ihrem Wohnort zur Hochschule zurücklegen müssen, so ist eine Klassierung der Merkmalsausprägungen zu Klassen $k_j, (j = 1, 2, ..., r)$ sinnvoll, die jeweils mehrere Kilometer umfassen.
Beim Merkmal Parteipräferenz kann es sinnvoll sein, wie bei den üblichen Wahlabend-Fernsehstatistiken die Merkmalsausprägungen Piratenpartei, Die Tierschutzpartei usw. in der Klasse *Sonstige* zusammenzufassen.

Durch eine Klasseneinteilung werden neue Merkmalsausprägungen definiert. Mathematisch betrachtet ist die Klassierung eine Funktion, bei der jede Merkmalsausprägung genau einer Klasse zugeordnet wird. Bei metrisch skalierten Merkmalsausprägungen sind die Klassen als Intervalle der reellen Zahlen darstellbar. Für viele Problemstellungen lassen sie sich in folgender Weise konstruieren:

$$k_j = [u + (j-1) \cdot b; u + j \cdot b) \qquad j = 1, 2, ..., r,$$

wobei r die Anzahl der Klassen ist, u der linke Rand der ersten Klasse k_1 und $u + r \cdot b$ der rechte (offene) Rand der letzten Klasse ist. Durch diese Konstruktionsvorschrift sind die Klassen k_j überlappungs- und lückenfrei sowie äquidistant (die Klassen haben die gleiche Breite b).

Beispiel:
X sei das Merkmal der Entfernungen, die 40 Studierende der Pädagogischen Hochschule Freiburg zur Hochschule zurücklegen. Implizit wurde bereits bei der Erhebung eine Klassierung vorgenommen, da die Entfernungen nur auf den Kilometer genau erhoben wurden. Dabei haben sich 19 unterschiedliche Merkmalsausprägungen ergeben, die der Größe nach geordnet von $x_1 = 1$ bis $x_{19} = 80$ reichen (vgl. Tabelle 1). Eine mögliche (weitere) Klassierung könnte beispielsweise dadurch erfolgen, dass $r = 17$ Klassen mit jeweils der gleichen Breite von 5 (was 5 km entspricht) gebildet werden (vgl. Abb. 2.3): $k_1 = [0;5), k_2 = [5;10), ..., k_{17} = [80;85)$.

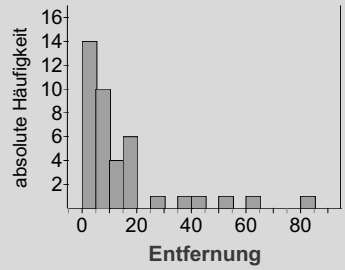


Abbildung 2.3: Klassierung des Merkmals Entfernung (zur Hochschule), durch die Merkmalsausprägungen in Klassen der Breite 5 (also zu jeweils 5 Kilometern) zusammengefasst sind

Entsprechend der oben genannten Klassenbildungsvorschrift umfasst in diesem Beispiel die erste Klasse k_1 alle Merkmalsausprägungen, die größer oder gleich 0, aber kleiner als 5 sind:

$$k_1 = [0;5) = \{x_i | 0 \leq x_i < 5\}$$

Messdaten physikalischer Größen, die zumindest theoretisch beliebige Genauigkeit zulassen, sind grundsätzlich mit einer Klasseneinteilung versehen, die auf der Messgenauigkeit basiert.

Beispiel:
Misst man eine Länge, so kann der Meter, der Zentimeter, der Millimeter etc. die Klasseneinteilung bedingen.

Hat man r Klassen von Merkmalsausprägungen zu einem Merkmal vorliegen, so kann man die absoluten bzw. relativen Häufigkeiten zu diesen Klassen angeben:

$$H_n(k_j) \quad \text{bzw.} \quad h_n(k_j) = \frac{H_n(k_j)}{n}, \quad j = 1, ..., r$$

2.1.4 Häufigkeitsverteilung

Definition 3
Die Funktion $f : x_i \to h_n(x_i)$ (oder auch $f : x_i \to H_n(x_i)$), $x_i \in X$, $i = 1, ..., s$ (n bezeichnet den Umfang der Stichprobe) heißt **Häufigkeitsverteilung** des Merkmals X.

Beispiel:
Gegeben sind die Häufigkeiten aller Merkmalsausprägungen zum Merkmal: Alter von Studierenden in einer Stichprobe an der Pädagogische Hochschule Freiburg ($n = 40$; vgl. Tabelle 1). Die relative Häufigkeitsverteilung $f : x_i \to h_{40}(x_i)$ (links) sowie die absolute Häufigkeitsverteilung $f : x_i \to H_{40}(x_i)$ (rechts) haben die in Abbildung 2.4 dargestellte grafische Gestalt:

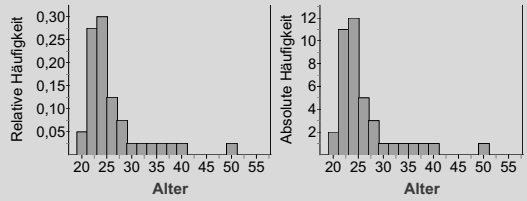


Abbildung 2.4: Häufigkeitsverteilung zum Alter von Studierenden

Man erkennt, dass die beiden Darstellungen in der Form identisch sind. Allgemein gilt, dass bei der Darstellung bloß *eines* Merkmals die Wahl von absoluten bzw. relativen Häufigkeiten unerheblich ist. Beim Vergleich *zweier* (oder mehrerer) Stichproben mit unterschiedlichen Umfängen muss dagegen auf die Darstellung relativer Häufigkeiten zurückgegriffen werden, da der absolute Vergleich in die Irre führen würde (vgl. Kap. 2.1.2).

2.1.5 Empirische Verteilungsfunktion

Definition 4
Die Funktion $F : \mathbb{R} \to \mathbb{R}$ mit

$$F : x \to \begin{cases} 0 & \text{falls } x < x_1 \\ \sum_{i=1}^{k} h_n(x_i) & \text{falls } x_k \leq x < x_{k+1} \\ 1 & \text{sonst } (x \geq x_s) \end{cases}$$

heißt empirische Verteilungsfunktion des Merkmals X, wobei n der Stichprobenumfang, s die Anzahl der verschiedenen, der Größe nach geordneten Merkmalsausprägungen ist und für $k \in \mathbb{N} : 0 < k \leq s-1$ gilt.

Statt der relativen Häufigkeiten könnten ebenso die absoluten Häufigkeiten aufsummiert werden. Die einzelnen Werte dieser Funktion umfassen damit die Kumulation (oder Summierung) der relativen bzw. absoluten Häufigkeiten einer Menge von k Merkmalsausprägungen.

Die Werte der Funktion F, die in Abbildung 2.5 für das Beispiel des Studierendenalters dargestellt ist, geben einen unmittelbaren Eindruck davon, mit welcher Altersstufe welcher Anteil der Studierenden ausgeschöpft ist. So ist z.B. unmittelbar erkennbar, dass rund 60% der Studierenden 23 Jahre oder jünger sind ($F(23) \approx 0,6$). Dieser Funktionsgraph stellt nur die Werte der Funktion F für die potentiellen Merkmalsausprägungen, also ganzzahlige Werte von x dar. Gemäß der Definition der Funktion hätte der Graph eine Treppenform mit Sprungstellen an den Stellen der empirisch auftretenden Merkmalsausprägungen. Da die gängigen Statistikprogramme allerdings die Darstellung dieses Treppengraphen nicht vorsehen und diese auch keine zusätzliche Interpretationsmöglichkeit bieten, belassen wir es bei diesem Graphen auf der Basis einer endlichen Definitionsmenge.

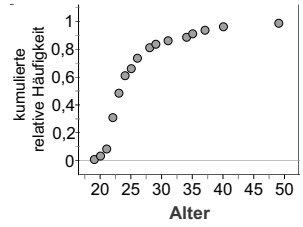


Abbildung 2.5: Empirische Verteilungsfunktion zum Alter von Studierenden

2.2 Grafische Darstellungen

„Eine Grafik sagt mehr als 1000 Worte.“In dieser Redensart steckt einige Wahrheit. So ist die Hauptmotivation, Daten oder (in der Regel) Häufigkeitsverteilungen grafisch darzustellen, die in den Daten steckenden Informationen zu kommunizieren. Sollte z.B. die Wohnentfernung von Studierenden betrachtet werden, dann ist die Aussage „Der überwiegende Anteil der Studierenden an der Pädagogischen Hochschule wohnt im Umkreis von 10 km zur Hochschule“ leichter mit der Grafik in Abbildung 2.6 zu vermitteln als mit dem Ausschnitt der daneben stehenden Zahlenkolonne der zugehörigen Roh-Daten.

Eine grafische Darstellung ermöglicht schnelle Einblicke in die Gestalt und möglicherweise Besonderheiten einer Häufigkeitsverteilung, die zu einer vertieften Untersuchung weiterführen kann. Daher eignen sich grafische Darstellungen zur *explorativen* Datenanalyse, bei der die Daten zunächst vorurteilsfrei auf verborgene Muster hin untersucht werden.

Bei den grafischen Darstellungen haben sich einige wenige Standardformen durchgesetzt, die im Folgenden vorgestellt werden. Prinzipiell ist aber in diesem Bereich der Phantasie (bei Beachtung gewisser Konventionen) keine Grenze gesetzt. Werden Konventionen (z.B. alltäglicher

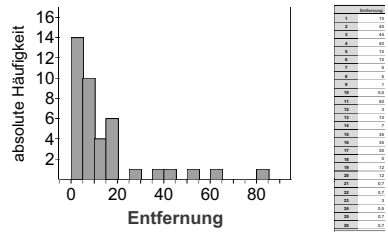


Abbildung 2.6: Aussagefähigkeit einer Grafik und der Rohdaten

Sehgewohnheiten) nicht beachtet, dann erhält man mitunter irreführende Grafiken, die die in Daten steckenden Informationen bewusst oder unbewusst manipulieren.

2.2.1 Säulen-, Balken- und Stabdiagramm

Das Säulendiagramm ist vermutlich die gebräuchlichste grafische Darstellung. Es ist für alle Skalierungsarten von Merkmalen anwendbar. Die Höhe der Säulen bzw. Stäbe (bzw. bei Balkendiagrammen die Breite) gibt die relative (absolute) Häufigkeit der Merkmalsausprägungen an (vgl.

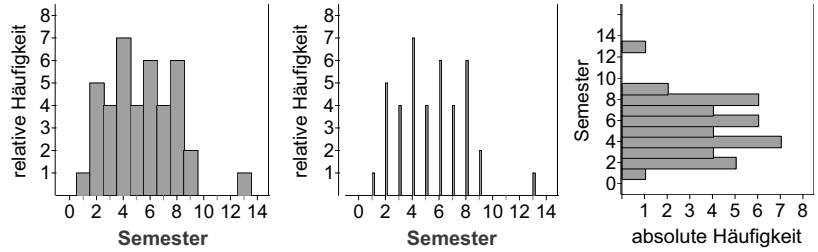


Abbildung 2.7: Beispiel zum Säulen-, Stab- und Balken-Diagramm

Abb. 2.7 zum Merkmal *Semesterzahl* der 40 Studierenden, Tabelle 1, S. 15). Die Breite der Säulen (bzw. die Höhe der Balken) ist im Säulen- bzw. Balkendiagramm eine reine Layoutfrage. Hier gilt als einzige Konvention, dass *eine* Breite (bzw. Höhe) für alle Säulen (bzw. Balken) beibehalten wird. Keine Layoutfrage ist dagegen die Wahl der Klassierung, hier die nicht vorhandene Klassierung bzw. natürliche Klassierung mit Klassen der Breite 1. In dem Studierenden-Beispiel könnte hinsichtlich der Semesteranzahl auch eine Klassierung mit Klassen der Breite 2 geeignet sein, wenn beispielsweise Studierende eines Studienjahres betrachtet werden sollen (Abb. 2.8).

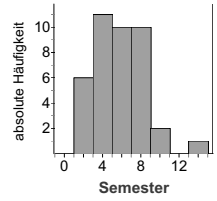


Abbildung 2.8: Säulendiagramm zur klassierten Semesteranzahl

Bei allen drei Diagrammarten kann die Reihenfolge der Merkmalsausprägungen beliebig gewählt werden. Es kann beispielsweise unter Umständen sinnvoll sein, die Merkmalsausprägungen nach dem Wert der zugehörigen relativen (absoluten) Häufigkeit wie in Abbildung 2.9 zu ordnen. Das ist unabhängig davon, ob die vorherige Ordnung numerisch, alphabetisch oder nach einem anderen eindeutigen Kriterium erfolgte.

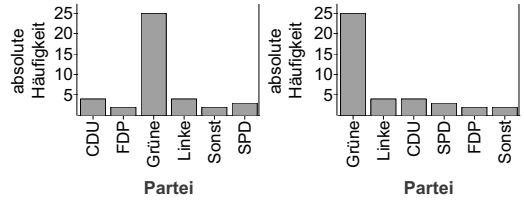


Abbildung 2.9: Freiheit der Ordnung beim Säulendiagramm

2.2.2 Histogramm

Im Gegensatz zu einem Säulendiagramm hat das Histogramm folgende Eigenschaften:

- Als Voraussetzung müssen die Merkmalsausprägungen metrisch skaliert sein. Es wird dabei angenommen, dass die Merkmalsausprägungen eine Teilmenge der reellen Zahlen sind.
- Die Merkmalsausprägungen werden nach Konvention in rechtsoffene Klassen eingeteilt.
- Die Säulen in einem Diagramm repräsentieren Häufigkeiten von Klassen. Die Säulen grenzen unmittelbar, das heißt überlappungs- und lückenfrei, aneinander.
- Der Flächeninhalt, *nicht* die Höhe einer Säule repräsentiert im Allgemeinen die Häufigkeit einer Merkmalsausprägung bzw. Klasse. Verwendet man allerdings, wie in der Praxis üblich, gleich breite (äquidistante) Klassen, so kann über die Säulenhöhe die Häufigkeit einer Merkmalsausprägung bzw. Klasse betrachtet werden (siehe Abb. 2.10, links). Ist die Einhaltung gleich breiter Klassen nicht sinnvoll oder erwünscht, so verwendet man statt eines Histogramms oft ein Säulendiagramm (siehe Abb. 2.10, rechts). Letzteres ist dann der Fall, wenn eine große Anzahl von Merkmalsausprägungen mit kleinen Häufigkeiten besteht und dadurch eine Zusammenfassung von Merkmalsausprägungen überlegenswert wird.

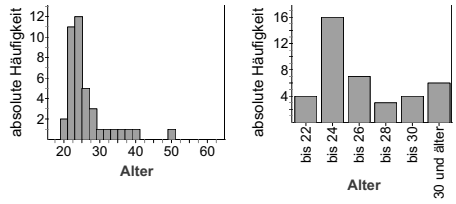


Abbildung 2.10: Beispiel eines Histogramms mit gleichbreiten (äquidistanten) Klassen (links) und der Übergang vom Histogramm zum Säulendiagramm aus Darstellungsgründen (rechts)

2.2.3 Kreisdiagramm

Das Kreisdiagramm ist für alle Skalierungsarten von Merkmalen anwendbar. Es ist für die Darstellung von Merkmalen mit wenigen Ausprägungen (bzw. Klassen) angemessen (vgl. Abb. 2.11). Das Kreisdiagramm eignet sich besonders gut, um den Anteil einer Merkmalsausprägung in einer Stichprobe (also die relative Häufigkeit) zu visualisieren.

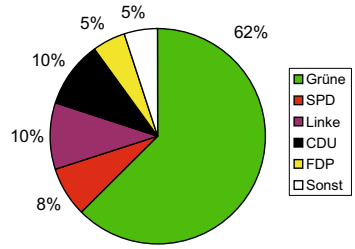


Abbildung 2.11: Beispiel Kreisdiagramm: Parteipräferenz von Studierenden der Pädagogischen Hochschule Freiburg

Der Winkel α_i des zur Merkmalsausprägung x_i gehörenden Kreisausschnitts wird über die Formel $\alpha_i = 360° \cdot h_n(x_i)$ berechnet.

Beispiel:
Von den 40 befragten Studierenden präferieren 25 die Partei der Grünen, was einem prozentualen Anteil von 62,5% entspricht. Damit ergibt sich als Winkel α des betreffenden Kreisausschnitts: $\alpha = 360° \cdot \frac{25}{40} = 225°$

2.2.4 Punktdiagramm

Bei einem Punktdiagramm, das für alle Skalierungsarten von Merkmalen anwendbar ist, werden die Punkte in gleicher Größe über der entsprechenden Merkmalsausprägung gestapelt. Wie beim Stab- oder Säulendiagramm enthält die Höhe der Punkthaufen, die sich aus der Anzahl der Punkte ergibt, die Information über die Größe der absoluten Häufigkeit. Bei kleinen Datenmengen ist das Punktdiagramm in Reinform darstellbar, bei großen Datensätzen sind in der Regel die Daten zu einer Merkmalsausprägung mit hoher absoluter Häufigkeit nur in Ansätzen abzubilden (vgl. Abb. 2.12). Die Größe der Punkte ist eine reine Layout-Frage.

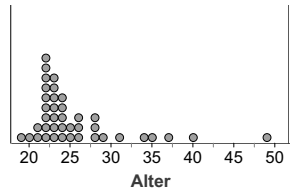


Abbildung 2.12: Beispiel Punktdiagramm: Alter von Studierenden der PH Freiburg

2.2.5 Stängel-Blatt-Diagramm

Das Stängel-Blatt-Diagramm ist eine auf kleine Datensätze beschränkte Darstellungsart. Es ist ein halbgrafisches Verfahren und umfasst stets die Klassierung der Merkmalsausprägungen. In der Grundform wird die kleinste gemessene Einheit zur Darstellung als „Blätter “ verwendet, die nächstgrößere 10er-Potenz legt in Klassen zerlegt den „Stängel“ fest.

Beispiel:
Gegeben sind 117 Daten zur Körpergröße von Studierenden (vgl. Abb. 2.13). Die kleinste gemessene Einheit ist der Zentimeter. Die Zentimeter werden als Blätter, die Dezimeter werden als Klassen verwendet. Ist beispielsweise eine Studierende 167 cm groß, so wird der Wert 7 cm als Blatt, der Wert 16 dm als Klasse des Stängels festgelegt. Als Blätter werden dann alle Zentimeter-Werte der Studierenden mit einer Körpergröße von 160 cm bis 169 cm der Größe nach geordnet eingetragen. Die entsprechende Zeile in dem Stängel-Blatt-Diagramm (Abb. 2.13) repräsentiert also Studierende mit einer Körpergröße von 160, 162, 162, 163, 163, 163, ..., 169, 169 Zentimetern.

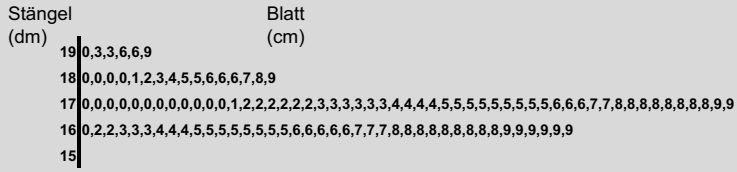
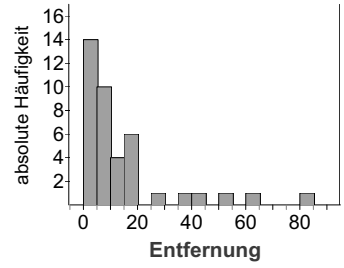


Abbildung 2.13: Beispiel zum Stängel-Blatt-Diagramm

Das Stängel-Blatt-Diagramm ist aus grafischer Hinsicht die Sonderform eines Histogramms: Sofern die Ziffern auf der Blatt-Seite mit gleicher Breite gewählt werden, repräsentiert die Breite aller Blätter die Häufigkeit einer Klasse. Zusätzlich bleiben im Gegensatz aber auch die numerischen Informationen erhalten. Man erkennt beispielsweise die Häufung der Messungen mit dem Ende 5.

2.3 Lageparameter

„Jede Studentin und jeder Student legt einen Weg von durchschnittlich 15 Kilometern zur Hochschule zurück.“
„50% der Studierenden haben einen Weg zwischen 3 und 15 Kilometern zur Hochschule.“
„Die meisten Studierenden wohnen im Umkreis bis 5 km zur Hochschule.“



Jede dieser Aussagen bezieht sich auf ein Muster in den Daten und fasst sie knapp zusammen. Im Folgenden werden mathematische Methoden vorgestellt, mit denen solche Muster und später auch die Abweichungen davon bestimmt werden. Wir gehen dabei getrennt nach sogenannten **Lageparametern** und später **Streuparametern** (Kap. 2.4) vor und ziehen zur Konkretisierung den Datensatz aus Tabelle 1 heran.

Durch die Berechnung von Lageparametern wird die Häufigkeitsverteilung auf einen Wert bzw. eine reelle Zahl reduziert. Dieser eine Wert soll für den Datensatz bzw. für die Häufigkeitsverteilung der Daten möglichst charakteristisch sein.[1]

2.3.1 Der Modalwert

Der **Modalwert** beschreibt die Merkmalsausprägung x_i (die Klasse k_i) mit der größten relativen Häufigkeit und antwortet damit auf die Frage „Wo sind am meisten?“. Der Modalwert kann auf alle Skalierungsarten angewendet werden. Er beschreibt ein *Zentrum* der Häufigkeitsverteilung. In einem Säulendiagramm repräsentiert die höchste Säule den Modalwert in einer Häufigkeitsverteilung (vgl. Abb. 2.14).

- „Die meisten Studierenden wohnen im Umkreis bis 5 km zur Hochschule.“
- „Die Grünen sind die am stärksten präferierte Partei bei den Studierenden.“

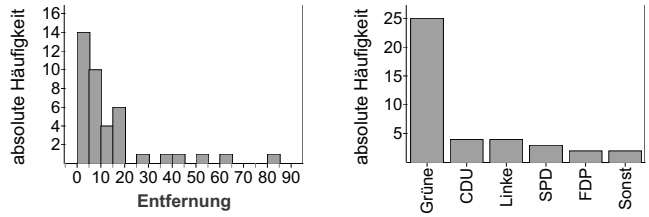


Abbildung 2.14: Modalwerte zur Entfernung (Klasse 0-5 km) und Parteipräferenz (Grüne)

Definition 5
Der **Modalwert** x_{Mod} bezeichnet diejenige(n) der s Merkmalsausprägungen (der r Klassen von Merkmalsausprägungen) mit der größten relativen (absoluten) Häufigkeit:

$$x_{Mod} = \{x_i | h_n(x_i) \geq h_n(x_j)\}, \quad (j = 1,2,...,s)$$

An der Definition erkennt man, dass der Modalwert aus mehreren Werten bestehen kann. Beispielsweise könnte die stärkste Parteipräferenz auf zwei Parteien genau gleich verteilt sein. Der Modalwert zu einem Datenzentrum ist aber nur dann sinnvoll zu interpretieren, wenn er eindeutig ist, d. h. die Häufigkeitsverteilung **eingipflig** (**unimodal**) ist (vgl. Kap. 2, S. 14).

[1] Man spricht auch von *charakteristischen Kennzahlen* einer Häufigkeitsverteilung, zu denen die Lageparameter gehören.

2.3.2 Quantile, Quartile, Median

„Bei der Erhebung an der Pädagogischen Hochschule Freiburg waren über 65% der Studierenden höchstens 25 Jahre alt“ , „Mehr als 30% der Studierenden wohnen höchstens 5 Kilometer von der Hochschule entfernt“ , ... Diese Art von Aussagen verwendet statistisch betrachtet ein Klasse von Lageparametern, die allgemein als **Quantile** bezeichnet werden.

Sind die Daten nach der Größe geordnet,[2] dann kann man die Daten in Gruppen sinnvoll so einteilen, dass jede Gruppe einen gewissen Prozentsatz aller Daten enthält (z. B. „Mehr als 30% der Studierenden wohnen höchstens 5 Kilometer von der Hochschule entfernt“). Betrachtet man die Skala der Merkmalsausprägungen, so kann man stets eine Stelle bestimmen, vor der (die Stelle selbst eingeschlossen) mindestens $p \cdot 100$ Prozent und nach der (die Stelle selbst eingeschlossen) mindestens $(1-p) \cdot 100$ Prozent der Daten liegen. Das ist eine Umschreibung von Quantilen, die wir zunächst an einem Beispiel behandeln wollen.

Beispiel:
Gegeben ist der Datensatz zu den 15 Studierenden der Universität Münster zum Merkmal Alter. In der nachfolgenden Tabelle sind die Daten bereits nach der Größe der Merkmalsausprägungen sortiert.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
19	19	20	21	21	21	21	21	22	22	23	23	24	24	25

Nun lässt sich beispielsweise fragen, wo die Stelle $x_{0,25}$ in dem geordneten Datensatz liegt, für die gilt, dass mindestens 25 Prozent ($p = 0,25$) der Daten (die Stelle selbst eingeschlossen) davor liegen und mindestens 75 Prozent ($(1-p) = 0,75$) der Daten (die Stelle selbst eingeschlossen) dahinter liegen. In diesem Beispiel ergibt 25 Prozent von 15 einen Dezimalbruch: $0,25 \cdot 15 = 3,75$. Da die Daten nicht geteilt werden können, kann man sagen, dass das vierte Datum (x_4) die gesuchte Stelle ist. Denn vor dieser Stelle (die Stelle selbst eingeschlossen) liegen vier Daten (also etwas mehr als 25 Prozent der Daten) und nach dieser Stelle (die Stelle selbst eingeschlossen) liegen 12 Daten, was einem Anteil von etwas mehr als 75 Prozent des Datensatzes entspricht.
Für $p = 0,2$ gilt sogar: *Genau* 20 % der Daten liegen vor der Stelle $x_{0,2}$ und *genau* 80 % danach (die Stelle selbst jeweils eingeschlossen). Da *genau* 20 Prozent 3 Daten und genau 80 Prozent 12 Daten entsprechen, wäre jeder Wert $x_{0,2}$ mit $20 = x_3 < x_{0,2} < x_4 = 21$ ein möglicher Wert für $x_{0,2}$. Die Antwort wäre demnach jedoch nicht eindeutig.

Fasst man das Verfahren zur Bestimmung von Lageparametern einer Häufigkeitsverteilung mathematisch und in der Berechnung eindeutig, so erhält man die **Quantile** durch folgende definierende Berechnungsformel.

[2] Das ist eine notwendige Voraussetzung.

Definition 6
Als **p-Quantil** x_p bezeichnet man diejenige reelle Zahl, für die gilt:

$$x_p = \begin{cases} x_{[n\cdot p]+1} & \text{für} \quad n\cdot p \quad \text{nicht ganzzahlig} \\ \frac{1}{2}(x_{n\cdot p} + x_{n\cdot p+1}) & \text{für} \quad n\cdot p \quad \text{ganzzahlig} \end{cases}$$

wobei $x_1, x_2, ..., x_s$ die erhobenen, mindestens ordinal skalierten Merkmalsausprägungen in einer Stichprobe mit Umfang n seien ($0 < p < 1$).

$[x]$ bezeichnet die Gaußklammer. Durch sie wird eine reelle Zahl auf ihren ganzzahligen Anteil reduziert, wie z. B. $[3,45] = 3$ oder $[18,999] = 18$. Mit der Berechnungsformel gilt auf der Basis der vorangegangenen Überlegungen:

Satz 3

> mindestens $p \cdot 100$ Prozent der metrisch-skalierten (ordinalskalierten) Merkmalsausprägungen der Daten sind kleiner oder gleich x_p und
>
> mindestens $(1-p) \cdot 100$ Prozent der Merkmalsausprägungen der Daten sind größer oder gleich x_p.

Spezielle Quantile sind der **Median** und die **Quartile**, die wir im Folgenden noch näher besprechen. Aus der Berechnungsformel geht hervor, dass das „mindestens“ sich nur dann durch ein „genau“ ersetzen lässt, wenn x_p zwischen zwei Merkmalsausprägungen mit verschiedenen Werten liegt.[3]

Wendet man die Berechnungsformel für die Quantile zu den Werten $p = 0,2, p = 0,5$ und $p = 0,75$ an, dann ergibt sich etwa:

$15 \cdot 0,2 = 3$	ganzzahlig	$x_{0,2} = \frac{1}{2}(x_3 + x_{3+1}) = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(20+21) = 20,5$
$15 \cdot 0,5 = 7,5$	nicht ganzzahlig	$x_{0,5} = x_{[7,5]+1} = x_{7+1} = x_8 = 21$
$15 \cdot 0,75 = 11,25$	nicht ganzzahlig	$x_{0,75} = x_{[11,25]+1} = x_{11+1} = x_{12} = 23$

Wie bereits angedeutet, haben einige p-Quantile eine eigene Bezeichnung.

Definition 7
Das 0,5-Quantil bezeichnet man als **Median**.

Beispiel:
Im vorausgehenden Datensatz zum Alter von 15 Studierenden der Universität Münster ergibt sich mit $n \cdot p = 15 \cdot 0,5 = 7,5$ der Median zu:

$$x_{0,5} = x_{[7,5]+1} = x_{7+1} = x_8 = 21$$

[3] Wenn die Merkmalsausprägungen nicht metrisch, sondern ordinalskaliert sind, so umfasst das p-Quantil bei $n \cdot p$ ganzzahlig statt des Mittelwertes zwischen zwei Merkmalsausprägungen beide dieser Merkmalsausprägungen, besteht also aus einer zweielementigen Menge. Wir betrachten dies aber als hier nicht weiter beachteten Sonderfall.

Im Falle des Medians ist die allgemeine Quantilsdefinition äquivalent zu:

Definition 8
Seien $x_1, x_2, ..., x_s$ die erhobenen Merkmalsausprägungen eines mindestens ordinalskalierten Datensatzes mit Umfang n, dann heißt

$$x_{0,5} = \begin{cases} x_{\frac{n+1}{2}} & \text{für } n \text{ ungerade} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{für } n \text{ gerade} \end{cases}$$

Median der Häufigkeitsverteilung.

Die „Viertelung“ eines Datensatzes wird durch die **Quartile** repräsentiert.

Definition 9
Das 0-Quantil bezeichnet man als **Minimum** der Merkmalsausprägungen (x_{Min}).
Das 0,25-Quantil bezeichnet man **1. Quartil** der Merkmalsausprägungen ($x_{0,25}$).
Das 0,5-Quantil bezeichnet man als **zweites Quartil oder Median** der Merkmalsausprägungen ($x_{0,5}$).
Das 0,75-Quantil bezeichnet man als **3. Quartil** der Merkmalsausprägungen ($x_{0,75}$).
Das 1-Quantil bezeichnet man als **Maximum** der Merkmalsausprägungen (x_{Max}).

Die beiden Definitionen zum Minimum und Maximum sind Zusätze zur Quantilsdefinition, die auf solche p mit $0 < p < 1$ beschränkt ist. Das heißt, dass Minimum und Maximum mit dieser Formel *nicht* berechnet werden können.[4]

Vier weitere Bemerkungen zur Bestimmung und Interpretation von Quantilen sind als Ergänzung wichtig:

1. Mit der Formel zur Quartilsbestimmung berechnet man zunächst auf der Ebene der Indizes eine Position im geordneten Datensatz. Erst anschließend geht man auf die Ebene der Werte der Merkmalsausprägungen, um ein Quantil, Quartil oder den Median zu bestimmen. Wie im einleitenden Beispiel angedeutet, lassen sich Quantile allerdings auch intuitiver bestimmen oder gar auszählen.
2. Bei den einleitenden Überlegungen zur Berechnungsformel ist Wert auf das Wort *mindestens* gelegt worden, da nur in bestimmten Fällen ein *genau* richtig ist. Die Bedeutsamkeit der Formulierung *mindestens* kann man an den beiden konstruierten Datensätzen zum Alter von 6 Studierenden verdeutlichen:

	x_1	x_2	x_3	x_4	x_5	x_6
Daten 1	21	21	21	22	22	22
Daten 2	21	21	21	21	21	21

Im ersten Datensatz sind tatsächlich *genau* 50% der Daten größergleich bzw. kleinergleich dem Median, die Formulierung mit *mindestens* bleibt dennoch richtig

$$n = 6; \quad n \cdot p = 6 \cdot 0,5 = 3; \quad x_{0,5} = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(21 + 22) = 21,5$$

[4] Hier wäre eine „Berechnung“ auch unnötig, da bei einem der Größe nach geordneten Datensatz ein Blick auf die Enden des Datensatzes genügt, um das Minimum und Maximum abzulesen.

Für etwa das erste Quartil

$$n = 6; \quad n \cdot p = 6 \cdot 0,25 = 1,5; \quad x_{0,5} = x_{[1,5]+1} = x_2 = 21$$

stimmt das *genau* nicht mehr, da 100% der Daten eine Merkmalsausprägung haben, die größer oder gleich 21 ist, und 50% der Daten eine Merkmalsausprägung haben, die kleiner oder gleich 21 ist. Die *mindestens*-Formulierung bleibt dagegen richtig. Noch extremer ist das Beispiel des zweiten Datensatzes, bei dem zu jedem beliebigen Quantil gilt, dass 100% der Merkmalsausprägungen größergleich und ebenso kleinergleich dem Quantil (das unabhängig von p immer den Wert 21 hat) sind.

3. Es gibt alternative Definitionen der Quantile. Beispielsweise verwendet das Tabellenkalkulationsprogramm Excel eine abweichende Definition für das 1. und 3. Quartil. Dort wird, wenn diese Quartile zwischen zwei Merkmalsausprägungen (z.B. zwischen $x_{n\cdot p} = 10$ und $x_{n\cdot p+1} = 20$) liegen, nicht das arithmetische Mittel verwendet (hier also 15), sondern die Berechnung erfolgt dadurch, dass der Abstand zwischen den betreffenden Merkmalsausprägungen geviertelt wird. Das erste Quartil würde durch $x_{0,25} = \frac{1}{4}(3 \cdot x_{n\cdot 0,25} + 1 \cdot x_{n\cdot 0,25+1}) = \frac{1}{4}(3 \cdot 10 + 20) = 12,5$ berechnet werden. Läge das 3. Quartil zwischen den beiden genannten Merkmalsausprägungen, so würde sich analog 17,5 als Wert dieses Quartils ergeben.
4. Insbesondere in der Wahrscheinlichkeitsanalyse werden nach Konvention häufig diejenigen Bereiche einer Verteilung betrachtet, die außerhalb eines Bereiches von $[x_{0,025}; x_{0,975}]$ liegen. Durch solch ein Intervall werden also 5% der Verteilung als besonders betrachtet. Das würde alternativ auch auf Werte außerhalb eines Bereiches wie $[x_{Min}; x_{0,95}]$ oder $[x_{0,05}; x_{Max}]$ zutreffen. Quantile, bei denen $p \cdot 100$ eine ganze Zahl ergibt, werden verallgemeinernd auch als **Perzentile** bezeichnet.

2.3.3 Der Boxplot

Der Boxplot ist eine Visualisierung der fünf Lageparameter Minimum, 1. Quartil, Median, 3. Quartil und Maximum (vgl. Abb. 2.15).

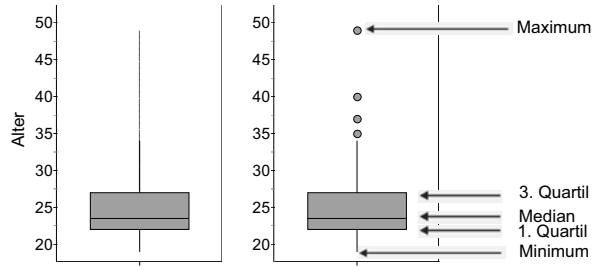


Abbildung 2.15: Boxplot zum Alter der Studierenden

Der Boxplot ist hier zum Merkmal Alter der 40 Studierenden der Pädagogischen Hochschule Freiburg dargestellt (vgl. Datensatz in Tabelle 1). Die fünf Lageparameter dazu sind:

Quartil	Minimum	1. Quartil	Median	3. Quartil	Maximum
Wert	19	22	23,5	27	49

Im Gegensatz zum Histogramm ist im Boxplot zwar die konkrete Verteilung zwischen den Lageparametern verborgen (vgl. Abb. 2.16, die 5 Lageparameter sind hier durch Pfeile gekennzeichnet), der entscheidende Vorteil des Boxplots ist jedoch die schnelle Übersicht zu wesentlichen Lageparametern der Häufigkeitsverteilung eines Merkmals. Die Verwendung des Boxplots ist bei eingipfligen Häufigkeitsverteilungen sinnvoll, bei mehreren Datenhäufungen verschleiert der Boxplot dagegen die Gestalt der Häufigkeitsverteilung. Dies kann dann besser z.B. durch ein Säulendiagramm repräsentiert werden.

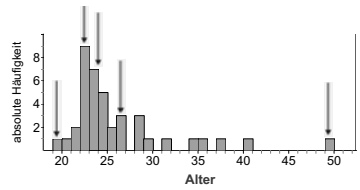


Abbildung 2.16: Histogramm zum Alter der Studierenden

In der linken Variante des Boxplots in Abbildung 2.15 ist die „Antenne“ über der Box, vom 3. Quartil zum Maximum, als Linie durchgezogen. Zum Teil wird auch die rechte Variante verwendet, um vom Zentrum (dem Median) weiter entfernte Daten zu kennzeichnen. Diese Grenze wird als **inner fence** bezeichnet und wird bei den in Abbildung 2.15 dargestellten Daten nur im oberen Teil des Boxplots überschritten. Der inner fence ist durch

$$IF = \left[x_{0,25} - \frac{3}{2} \cdot (x_{0,75} - x_{0,25}), x_{0,75} + \frac{3}{2} \cdot (x_{0,75} - x_{0,25}) \right]$$

definiert. Alle außerhalb dieses Intervalls (inner fence) liegenden Merkmalsausprägungen werden im Boxplot durch Punkte repräsentiert. Das sind hier alle Merkmalsausprägungen, die größer $x_{0,75} + 1,5 \cdot (x_{0,75} - x_{0,25}) = 27 + 1,5 \cdot (27 - 22) = 33,5$ sind. Sie werden als *Ausreißer* bezeichnet. Grafisch betrachtet sind dies alle Datenpunkte, bei denen die Boxplotantenne länger als das $1,5$-fache der Boxenlänge wäre.

2.3.4 Arithmetisches Mittel

Definition 10
Seien $x_1, x_2, ..., x_n$ die in einer Stichprobe vom Umfang n erhobenen metrisch-skalierten Merkmalsmalsausprägungen eines Merkmals X, so heißt $\bar{x}$ mit

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

arithmetisches Mittel der Häufigkeitsverteilung.

Die Berechnung des arithmetischen Mittels ist ausschließlich auf Häufigkeitsverteilungen mit metrisch-skalierten Merkmalsausprägungen bezogen und ebenfalls nur bei eingipfligen Häufigkeitsverteilungen aussagekräftig.

Beispiel:
Gegeben sind die 15 Daten zum Merkmal Alter der Studierenden der Uni Münster (vgl. Tabelle 1). Es ergibt sich ein durchschnittliches Alter von:

$$\bar{x} = \frac{19+19+20+20+21+21+21+21+22+22+23+23+24+25+25}{15} \approx 21,73$$

2.4 Streuparameter

Ein Lageparameter verweist auf ein bestimmtes Muster in einem Datensatz, etwa das durchschnittliche Alter von Studierenden. Je altershomogener die Studierenden sind, desto charakteristischer ist das arithmetische Mittel (d.h. das Durchschnittsalter) zur Beschreibung der Studierenden. Die Einschätzung der Homogenität der Daten – d.h. der **Streuung** der Daten – ist also eine wichtige Zusatzinformation, die in einem Lageparameter nicht enthalten ist. Das gilt insbesondere dann, wenn verschiedene Datensätze zu einem Merkmal verglichen werden.

Konstruiert man z.B. die in Abbildung 2.17 enthaltenen Häufigkeitsverteilungen, die alle drei identischen Werte für den Median und das arithmetische Mittel haben (hier $x_{0,5} = \bar{x} = 8$), so erkennt man, dass die drei Verteilungen unterschiedliche Homogenität bezogen auf ein *Zentrum*, also bezogen auf Median und arithmetisches Mittel aufweisen. Im Fall der unteren U-förmigen Verteilung ist zudem die Angabe eines Zentrums wie Median oder arithmetisches Mittel wenig charakteristisch. Es wird hier besonders deutlich, dass sich beide Muster – arithmetisches Mittel wie Median – insbesondere auf eingipfelige Häufigkeitsverteilungen beziehen.

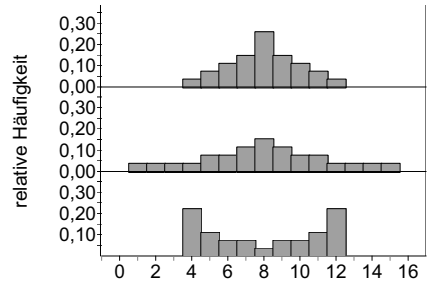


Abbildung 2.17: Drei Verteilungen mit unterschiedlicher Homogenität der Daten

Wir werden im Folgenden Methoden diskutieren, mit denen die Streuung von Daten bezogen auf einen Lageparameter charakterisiert werden kann.

2.4.1 Spannweite

Die Spannweite der Merkmalsausprägungen als Abstand von Minimum und Maximum ist ein intuitiv einsichtiges Streumaß. Es ist unabhängig von einem zentralen Lageparameter und kann für alle numerischen Merkmalsausprägungen bestimmt werden.

Definition 11
Der Abstand von Maximum und Minimum der Merkmalsausprägungen zu einem mindestens ordinalskalierten Merkmal X heißt **Spannweite** R (Range):

$$R = x_{Max} - x_{Min}$$

Beispiel:
Die Abiturnoten der 40 Studierenden der Pädagogischen Hochschule Freiburg haben in der vorliegenden Stichprobe eine Spannweite von 1,9:

$$R = x_{Max} - x_{Min} = 3,2 - 1,3 = 1,9$$

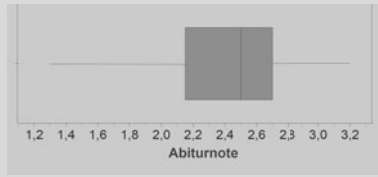


Abbildung 2.18: Spannweite hinsichtlich des Merkmals Abiturnote

Die Spannweite ermöglicht keine Aussage zu einem Zentrum der Daten und ist bei Ausreißern anfällig. Im Boxplot ist die Spannweite durch die Länge des (gesamten) Boxplots repräsentiert.

2.4.2 Quartilsabstand

Der Quartilsabstand repräsentiert die Spannweite des Zentrums des Datensatzes:

Definition 12
Der Abstand von 3. und 1. Quartil der Merkmalsausprägungen zu einem mindestens ordinalskalierten Merkmal heißt **Quartilsabstand** $Q_{0,5}$:

$$Q_{0,5} = x_{0,75} - x_{0,25}$$

Der Boxplot visualisiert (vgl. Abb. 2.18) den Quartilsabstand als Breite der Box selbst. Der Quartilsabstand beschreibt den Bereich, in dem die (mindestens) 50 Prozent der zentralen Daten liegen.

Der oben definierte inner fence ist ebenfalls ein Streumaß einer Häufigkeitsverteilung: Er umfasst den Bereich der nicht als Ausreißer geltenden Merkmalsausprägungen. Ebenso wäre die Länge des auf Seite 29 angesprochenen Intervalls $[x_{0,025}; x_{0,975}]$, also $Q_{0,95} = x_{0,975} - x_{0,025}$, ein Streumaß, das den Bereich beschreibt, in dem (mindestens) 95% der zentralen Daten liegen.

Beispiel:
Die Abiturnoten der 40 Studierenden der PH Freiburg haben in der vorliegenden Stichprobe einen Quartilsabstand von 0,55:[5] $Q_{0,5} = x_{0,75} - x_{0,25} = 2,7 - 2,15 = 0,55$ (vgl. Abb. 2.18).

2.4.3 Varianz und Standardabweichung

Im Gegensatz zum Quartilsabstand und zur Spannweite gibt es mit der empirischen Varianz und der empirischen Standardabweichung Streumaße, deren Bedeutung innerhalb der beschreibenden Analyse von Daten noch nicht unmittelbar einsichtig ist. Beide Streumaße beziehen sich auf das arithmetische Mittel und sind Maße für die mittlere Abweichung vom arithmetischen Mittel. Sie sind allerdings für viele der Methoden in der beschreibenden Datenanalyse sowie der Wahrscheinlichkeitsanalyse und der beurteilenden Datenanalyse wichtig.

Definition 13
Das arithmetische Mittel der quadratischen Abstände der Merkmalsausprägungen zum arithmetischen Mittel der Häufigkeitsverteilung heißt **empirische Varianz** s^2:

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

Die Wurzel aus der empirischen Varianz heißt **empirische Standardabweichung** $s = \sqrt{s^2}$.

Beispiel:
Gegeben sind die Daten zum Alter der 40 Studierenden der PH Freiburg (vgl. Tabelle 1). Das arithmetische Mittel ist $\bar{x} \approx 25,7$. Über die Definition ergibt sich:

$$s^2 = \frac{1}{40}\sum_{i=1}^{40}(x_i - \bar{x})^2 \approx \frac{1}{40}\left((19-25,7)^2 + ... + (49-25,7)^2\right) \approx 34,9$$

Dadurch ergibt sich unmittelbar $s = \sqrt{s^2} \approx 5,9$. Die Standardabweichung lässt sich als Abstand zum arithmetischen Mittel deuten: Analog zum Quartilsabstand betrachtet ergäbe sich das Intervall $[\bar{x} - s; \bar{x} + s] = [19,85; 23,61]$.

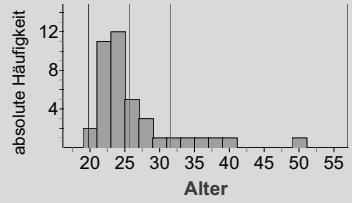


Abbildung 2.19: Häufigkeitsverteilung zum Alter der Studierenden mit arithmetischem Mittel und Visualisierung des Intervalls $[\bar{x} - s; \bar{x} + s]$

[5] Allgemein könnte man zusätzlich Quantilsabstände definieren, die den Abstand von Quantilen beschreiben, die symmetrisch um den Median liegen (die Symmetrie bezieht sich hier auf den p-Wert des Medians, also 0,5).

Die empirische Varianz kann *nicht* als Abstand innerhalb einer Häufigkeitsverteilung gedeutet werden, da sie ein zweidimensionales Maß ist, d. h. in geometrischer Perspektive einen Flächeninhalt repräsentiert. Die Bedeutung der Varianz begründet sich in der Minimalitätseigenschaft des arithmetischen Mittels hinsichtlich dieses Streumaßes (vgl. Kap. 2.7).[6] Die Standardabweichung hat die gleiche Dimension wie die Merkmalsausprägungen und kann damit als Abstand innerhalb der Häufigkeitsverteilung gedeutet werden.

2.4.4 Mittlere absolute Abweichung

Die Abweichung eines Datenpunkts von einem Muster, hier einem Lageparameter, werden wir insbesondere im folgenden Kapitel als *Residuum* r_i $(i = 1, ..., s)$ bezeichnen. Ein Residuum zum Median hat bezogen auf die Merkmalsausprägung x_i den Wert $r_i = x_i - x_{0,5}$. Die mittlere Summe der absoluten Residuen, die **mittlere absolute Abweichung**, die sich auf den Median der Häufigkeitsverteilung bezieht, ist ein weiterer Streuparameter.

Definition 14
Das arithmetische Mittel der (absoluten) Abstände der Merkmalsausprägungen eines metrischskalierten Merkmals X zum Median $x_{0,5}$ der Häufigkeitsverteilung heißt **mittlere absolute Abweichung** $d_{x_{0,5}}$:

$$d_{x_{0,5}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_{0,5}|,$$

Hinsichtlich dieses Streumaßes lassen sich Minimalitätseigenschaften des Medians zeigen (vgl. Kap. 2.7). Anders als bei der Analyse von Datensätzen mit zwei Merkmalen im folgenden Kapitel 3 betrachten wir dieses Streumaß hier als Exkurs.

2.5 Vergleich der Lage- und Streuparameter und die Form der Verteilung

2.5.1 Schiefe – Steilheit

Wir haben in den Daten zu den Studierenden unterschiedliche Formen kennengelernt (vgl. Abb. 2.20). Die Häufigkeitsverteilung zum Merkmal Entfernung (Abb. 2.20, links) hat die Eigenschaft, dass die meisten Studierenden kurze Entfernungen zur Hochschule zurücklegen und nur wenige große (zum Teil sehr große) Entfernungen zu bewältigen haben. Diese Form der Verteilung heißt *linkssteil* (sinnfällig für „links ist es steil“). Der analoge Begriff der **Schiefe** ist vom Attribut her genau entgegengesetzt: Linkssteile Verteilungen sind rechtsschief. Dagegen zeigt die Häufigkeitsverteilung von Körpergrößen einer Teilstichprobe von weiblichen Studierenden eine symmetrische Form (Abb. 2.20, rechts). Entsprechend können als weitere Formvariante rechtssteile (linksschiefe) Verteilungen identifiziert werden.

[6] Häufig wird bei der Varianz und Standardabweichung – anders als in der obigen Definition – der Vorfaktor $\frac{1}{n-1}$ angegeben. Dies hat mit der Eigenschaft eines sogenannten *erwartungstreuen Schätzers* zu tun, ein statistisches Konzept, das wir in diesem Band nicht thematisieren werden.

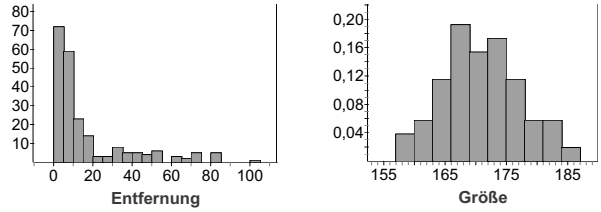


Abbildung 2.20: Verteilungen mit unterschiedlicher Form

Bei realen Datensätzen findet man

- linkssteile Häufigkeitsverteilungen häufig bei „sozialen“ Daten etwa in Bezug auf Einkommen oder allgemeiner dem Besitz irgendeines Guts.
- symmetrische Häufigkeitsverteilungen bei natürlich entstandenen Daten (Körpergröße, Anzahl der Haare etc.).
- rechtssteile Häufigkeitsverteilungen selten: Die Verteilung der Abiturnoten der Studierenden an der PH Freiburg weist beispielsweise eine leicht rechtssteile Tendenz auf.

Die Begriffe der Linkssteilheit, Symmetrie und Rechtssteilheit, die nur für eingipflige Verteilungen sinnvoll anwendbar sind, haben wir hier rein qualitativ eingeführt. Über den Begriff der Schiefe können sie auch quantifizierend betrachtet werden. Diese Betrachtung werden wir in einem Exkurs (Kap. 2.7) aufnehmen und uns hier zunächst nur mit den Auswirkungen der Verteilungsform anhand eines konstruierten Beispiels befassen (Abb. 2.21).

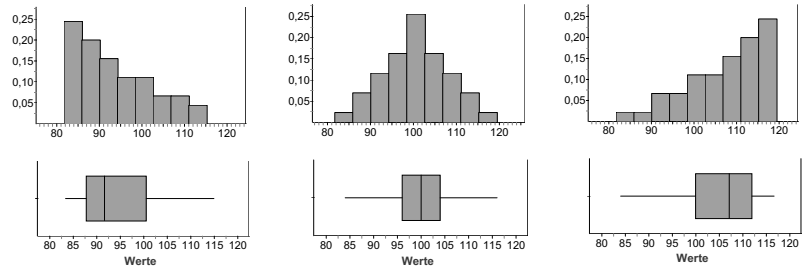


Abbildung 2.21: Linkssteile, symmetrische und rechtssteile Häufigkeitsverteilung

Der Vergleich der Lageparameter Modalwert, Median und arithmetisches Mittel ergibt bei den eingipfligen Häufigkeitsverteilungen des konstruierten Beispiels, die in Abbildung 2.21 zu sehen sind:

- Ist die Häufigkeitsverteilung symmetrisch, so gilt: $\overline{x} = x_{0,5} = x_{Mod}$
- Ist die Häufigkeitsverteilung linkssteil, also die Masse der Häufigkeitsverteilung nach links verschoben (bzw. rechtsschief), so gilt: $\overline{x} > x_{0,5} \quad (> x_{Mod})$
- Ist die Häufigkeitsverteilung rechtssteil, also die Masse der Häufigkeitsverteilung nach rechts verschoben, so gilt: $\overline{x} < x_{0,5} \quad (< x_{Mod})$

Beispiel:
Nimmt man das Merkmal Entfernung der 40 Studierenden der Pädagogischen Hochschule Freiburg als Beispiel, so ergibt sich für die linkssteile Verteilung dieses Merkmals[7]:

$$x_{Mod} = 2,5 \ < \ x_{0,5} = 7 \ < \ \bar{x} \approx 13,1$$

Deutungen:
„Die meisten Studierenden wohnen im Umkreis bis 5 km zur Hochschule“ (Modalwert).
„Der bzw. die mittlere Studierende wohnt 7 Kilometer von der Hochschule entfernt“ (Median).
„Durchschnittlich wohnen die Studierenden 13,1 Kilometer von der Hochschule entfernt“ (arithmetisches Mittel).

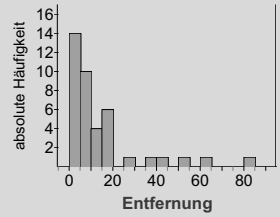


Abbildung 2.22: Linkssteile Häufigkeitsverteilung zum Merkmal Entfernung

Ist demnach eine Häufigkeitsverteilung eingipflig und symmetrisch, so repräsentieren alle drei Lageparameter das gleiche Zentrum. Ist die Verteilung dagegen (zu einer Seite) steil, so ergeben die drei Lageparameter unterschiedliche Werte und damit unterschiedliche Interpretationen des gleichen Datensatzes. Die Form der Verteilung ist wie im Histogramm auch im Boxplot in der Regel gut zu unterscheiden (vgl. Abb. 2.21). Die gegenseitigen Verhältnisse der drei Lageparameter können als Definition der Begriffe linkssteil, rechtssteil und symmetrisch dienen (vgl. Hartung et al., 2009).

Die Unterschiedlichkeit der Interpretation – insbesondere der beiden durch Median und arithmetisches Mittel repräsentierten Zentren der Häufigkeitsverteilung – werfen die Frage nach dem geeigneten Lageparameter auf, wenn die Häufigkeitsverteilung nicht symmetrisch ist. Diese Frage steht im Zusammenhang mit einer Eigenschaft von statistischen Methoden (wie Median und arithmetisches Mittel), die im folgenden Kapitel 2.5.2 diskutiert wird.

Die Form der Verteilung hat nicht nur bezogen auf die Lageparameter Bedeutung, sondern ebenso auf die Interpretation eines Streuparameters: Bei symmetrischen Verteilungen ist auch ein Streumaß wie der Quartilsabstand symmterisch interpretierbar, $x_{0,5} \pm \frac{1}{2} Q_{0,5}$ würde bezogen auf den Boxplot die Grenzen der Box beschreiben. Bei schiefen Verteilungen ist dagegen auch die Box schief, so dass der Quartilsabstand überwiegend nur in eine Richtung vom Median aus die Streuung beschreibt (Abb. 2.23). Betrachtet man analog $\bar{x} \pm s$, so erfasst die Streuung auf der einen Seite des arithmetischen Mittels einen größeren Teil der Verteilung, als auf der anderen Seite (Abb. 2.23).

[7] Als Modalwert wurde hier die Mitte der Klasse von 0 bis 5 km verwendet.

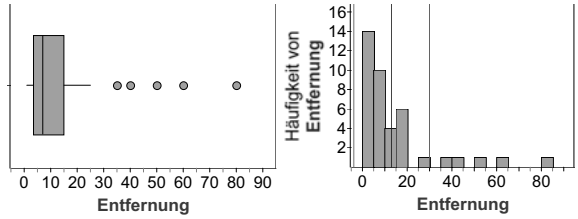


Abbildung 2.23: Quartilsabstand $Q_{0,5} = 11,5$ bei schiefer Box (links) und Intervall $[\bar{x} - s; \bar{x} + s]$ zu der Verteilung zum Merkmal Entfernung

2.5.2 Robuste und nicht robuste Methoden

Ist eine statistische Methode wenig anfällig gegen Ausreißer, d.h. von Daten, die weit von einem Zentrum der Häufigkeitsverteilung entfernt sind, dann nennt man diese Methode **robust** bzw. **resistent**. Der Median ist robust gegen Ausreißer: Da bei der Bestimmung des Medians die Position innerhalb des geordneten Datensatzes von entscheidender Bedeutung ist, nehmen die einzelnen Datenwerte nur mittelbar Einfluss auf die Berechnung. Enthält ein Datensatz einen extrem hohen Wert, so hat dies bei der Ermittlung des Medians keinen Einfluss, da die Position dieses Ausreißers im geordneten Datensatz festgelegt ist. In einem konstruierten Datensatz könnten fast 50 Prozent der Daten erhöht werden, bevor sich der Median ändert. Beim arithmetischen Mittel kann dagegen die Erhöhung eines Datums den Wert des arithmetischen Mittels beliebig erhöhen.

Beispiel:
Gegeben sind folgende Daten, für die $\bar{x} = x_{0,5} = 3$ ist:

x_1	x_2	x_3	x_4	x_5
1	2	3	4	5

Erhöht man den Wert von x_5 beliebig, so bleibt der Median gleich, das arithmetische Mittel steigt bis asymptotisch $x_5/5$. Auch das Datum x_4 könnte beliebig erhöht werden, ohne den Median zu ändern.

Ein Maß für die Robustheit ist der sogenannte *Bruchpunkt*, der den Anteil der Daten angibt, die maximal Ausreißer (nach oben) sein können, ohne den Wert der statistischen Methode zu ändern. Beim Median sind das asymptotisch 50% der Daten, beim arithmetischen Mittel kann die Erhöhung bereits eines Datums den Wert des arithmetischen Mittels verändern.[8]

Betrachtet man noch einmal das Beispiel „Entfernung zur Hochschule“ (Abb. 2.22), so sind einige der Daten „Ausreißer“, die weit vom Zentrum bzw. der Masse der Daten entfernt sind. Diese Daten erhöhen das arithmetische Mittel, während sie den Median unberührt lassen. Das arithmetische Mittel $\bar{x} = 13,1$ scheint hier nicht mehr ein Zentrum der Daten zu repräsentieren, da rund 70% der Daten kleiner sind als das arithmetische Mittel. Dennoch kann man nicht davon sprechen, dass hier das arithmetische Mittel die „falsche“ Methode wäre, es kommt vielmehr auf die Fragestellung an.

[8] Nähere Informationen zum Bruchpunkt bietet z.B. Polasek (1994).

Wichtig ist die Erkenntnis, dass bei nichtsymmetrischen Verteilungen aufgrund der Robustheit des Medians und der fehlenden Robustheit des arithmetischen Mittels diese Methoden zu unterschiedlichen Interpretationen eines Datensatzes führen. Wird etwa ein soziales Merkmal zum Besitz erhoben (wie z.B. das Einkommen), so kann der Median, wenn die Häufigkeitsverteilung linkssteil ist, Anlass geben, über den niedrigen Besitzstand zu klagen, während das arithmetische Mittel als deutlich höherer Wert ein mögliches Gegenargument ist.

Wie die Lageparameter können auch die Streuparameter mehr oder weniger robust sein. Während Varianz, Standardabweichung, mittlerer absoluter Abstand und auch die Spannweite nichtrobuste Methoden sind, ist der Quartilsabstand eine robuste Methode.

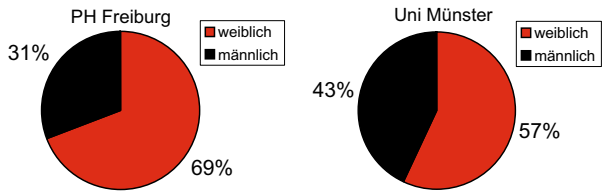
2.6 Eigenschaften von Studierenden

Wir setzen im Folgenden die Methoden der vorangegangenen Abschnitte exemplarisch auf den Datensatz zu den Studierenden aus Freiburg (abgekürzt durch F, Erhebung 2009) und Münster (abgekürzt durch M, Erhebung 2009) um. Dabei werden wir die Ergebnisse aller datenanalytischen Methoden einerseits sachgemäß interpretieren, andererseits aber den Interpretationsspielraum auch bis zur Überspitzung ausdehnen. Auf diese Weise wollen wir deutlich machen, dass ein und derselbe Datensatz zu durchaus verschiedenen Aussagen führen kann, wenn der zur Verfügung stehende Interpretationsspielraum ausgereizt wird.

Geschlechterverteilung, grafische Darstellung und Häufigkeitsverteilung

Verwendet man die Indizes W für weiblich und M für männlich, so ergibt sich in rein deskriptiver Darstellung:

	Freiburg	Münster
weiblich	$h_{178}(W_F) = 0,69$	$h_{1081}(W_M) = 0,57$
männlich	$h_{178}(M_F) = 0,31$	$h_{1081}(M_M) = 0,43$



Der Frauenanteil scheint also an beiden Hochschulen höher zu sein als der Männeranteil (die Universität Münster gibt offiziell ein Verhältnis von 53% zu 47% an). Das Ungleichgewicht ist allerdings an der Pädagogischen Hochschule Freiburg sehr viel deutlicher, was sich mit den dort beheimateten erziehungswissenschaftlichen Studienfächern und der Beschränkung der Lehramtsstudiengänge auf die Grund-, Haupt- und Realschule plausibel erklären ließe. Beide Stichproben sind dabei mit einer Ausschöpfung von etwa 3-4% aller Studierenden vergleichbar.

Interpretation:
Ist eine Männer-Quote notwendig? Umfragen an deutschen Hochschulen haben ergeben, dass der Frauenanteil erheblich größer ist als der Männeranteil der Studierenden. Besonders die Pädagogischen Hochschulen sind fest in der Hand von Frauen ...

Alter von Studierenden, Lage- und Streuparameter Die Häufigkeitsverteilung zum Alter der Studierenden in Münster und Freiburg ist recht ähnlich, wie man an den Boxplots in Abbildung 2.24 erkennen kann. Insbesondere weisen die mittleren 50% in beiden Standorten eine Symmetrie auf, zudem ist der Quartilsabstand identisch. Die Spannweite ist allerdings in Freiburg (durch Ausreißer bedingt) erheblich größer. Der wesentliche Unterschied scheint darin zu bestehen, dass die Altersverteilung in Freiburg ein Stück (ein Jahr) nach rechts verschoben ist. Das Durchschnittsalter ist durch die Ausreißer bedingt sogar um etwa 1,5 Jahre höher. Man kann also zu dem Schluss kommen, dass die Altersverteilung in Freiburg zwar ähnlich homogen ist wie in Münster, dass aber die Studierenden tendenziell älter sind.

	$\overline{x}$	$x_{0,5}$	$Q_{0,5}$	R
Münster	22,4	22	4	21
Freiburg	24,0	23	4	30

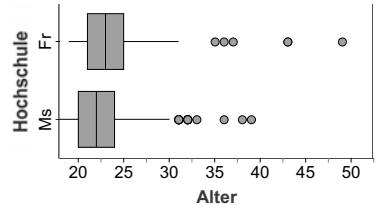


Abbildung 2.24: Alter der Studierenden in Münster und Freiburg

Interpretation:
Alte Pädagogen Umfragen an deutschen Hochschulen haben ergeben, dass die Studierenden der Erziehungswissenschaften den anderen Studierenden hinterherhinken und deutlich älter sind als ihre Kollegen anderer Hochschulen ...
...
Erfahrene Pädagogen Umfragen an deutschen Hochschulen haben ergeben, dass die Studierenden der Erziehungswissenschaften mit einem höheren Alter und damit mehr Lebenserfahrung in ein Studium einsteigen, im Gegensatz zu den blutjungen, unerfahrenen Studierenden anderer Fakultäten ...

Welchen Weg haben Studierende, Form der Verteilung Die beiden Verteilungen zum Merkmal Entfernung sind linkssteil (rechtsschief). Bei dieser Form werden also die zentralen Lageparameter unterschiedliche Werte und demnach unterschiedliche Interpretationen erzeugen.

Man erkennt die leicht höhere Entfernung, die Studierende der Pädagogischen Hochschule Freiburg zurücklegen. Insbesondere ist der mittlere Bereich (1. und 3. Quartil) im Gegensatz zu Münster verschoben. Woran kann das liegen? Hier kann nur der Kontext der Daten helfen. Die Hochschule in Freiburg liegt in einem innenstadtentfernten Bereich ohne weiteres Umland

	$\overline{x}$	$x_{0,25}$	$x_{0,5}$	$x_{0,75}$
Münster	9,6	2	3	6
Freiburg	11,5	3	5,5	15

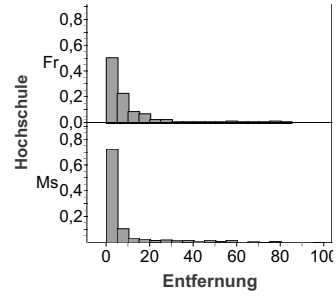


Abbildung 2.25: Entfernung zur Hochschule von Studierenden in Münster und Freiburg

(in Richtung Schwarzwald). Im Unterschied dazu ist die Universität Münster von allen Seiten zugänglich über die Innenstadt verteilt gelegen. Mit diesen Überlegungen könnte allein die Lage ein plausibles Erklärungsmuster abgeben. Beiden Verteilungen ist gemeinsam, dass jeweils die meisten (Modalwert) und über 50% der Studierenden im Umkreis von 5 km um die jeweilige Hochschule wohnen. Ebenso liegen die Mittelwerte beider Hochschulen jeweils deutlich über dem Median, so dass sie die Datensätze hinsichtlich eines Zentrums nicht repräsentieren. Kurz zusammengefasst: Viele der Studierenden wohnen relativ nah zur Hochschule, wenige wohnen weit entfernt.

Interpretation:
Uni-Studenten suchen Abstand, PH-Studenten die Nähe Umfragen an deutschen Hochschulen haben ergeben, dass die Studierenden der Uni Münster bei ihrer Wohnortsuche im Durchschnitt rund 10 km Abstand zwischen sich und der Hochschule schaffen. Über zwei Drittel der Studierenden in Freiburg suchen dagegen eine größere Nähe zu ihrer Hochschule.

In dieser sicherlich übertriebenen Interpretation wird ausgenutzt, dass beide Verteilungen linkssteil sind. So lässt sich durch die Verwendung von arithmetischem Mittel einerseits (Münster) und einem Quartil andererseits ($p_{0,67}$, Freiburg) ein interpretativer Gegensatz künstlich erzeugen.

Wie lange sitzen Studierende vor dem Rechner, Streuung Betrachtet man die Anzahl der Stunden, die die Studierenden wöchentlich vor dem Rechner sitzen (in Münster mehr als in Freiburg, vgl. Abb. 2.26), dann ist zunächst die relative Homogenität der Studierenden in Freiburg gegenüber den weiter gestreuten Zeiten in Münster auffällig.

Selbst wenn man den Spitzenwerten keinen Glauben schenken möchte, zeigt insbesondere der Quartilsabstand den deutlichen Unterschied zwischen den Studierenden beider Hochschulen (in dieser Stichprobe) auf. In diesem Fall verzichten wir auf die Umsetzung dieser Tatsache in einer plakativen Aussage.

	R	$Q_{0,5}$	s
Münster	56	10	8,5
Freiburg	49	6	7,8

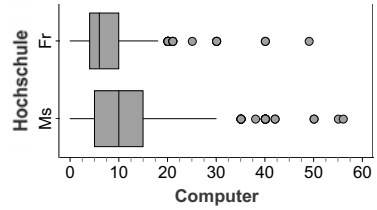


Abbildung 2.26: Zeitlicher Umfang der wöchentlichen Rechnernutzung von Studierenden in Münster und Freiburg

2.7 Ergänzungen

2.7.1 Grafische Darstellung

Im vorausgehenden Kapitel 2.6 haben wir dargestellt, dass Parameter einer Häufigkeitsverteilung überspitzt oder auch manipulierend kommuniziert werden können. Manipulationen, ob gewollt oder nicht, finden sich häufig in grafischen Darstellungen von Häufigkeitsverteilungen. Wir gehen hier auf drei Manipulationsmöglichkeiten ein.[9].

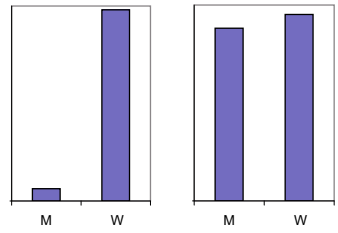


Abbildung 2.27: Verteilung der Geschlechter an der PH Freiburg (Erhebung 2010)

Manipulation der Häufigkeitsachse Je nachdem, wie die Häufigkeitsachse skaliert wird, können deutliche oder weniger deutliche Unterschiede zwischen Merkmalsausprägungen visualisiert werden. Beide Grafiken aus Abbildung 2.27 basieren auf der Verteilung $h_{218}(m) = 0,29$ und $h_{218}(w) = 0,71$ (Erhebung Freiburg, 2010). Im ersten Fall – mit visuell erheblichen Unterschieden – ist die Skalierung auf den Bereich $[0,26;0,72]$ eingeschränkt worden. Im zweiten Fall – mit visuell geringen Unterschieden – ist die Skalierung auf den Bereich $[-5;1]$ ausgeweitet worden. Das dahinterstehende Prinzip lässt sich folgendermaßen zusammenfassen: Um geringe Unterschiede sehr deutlich anzuzeigen, zeige man den entsprechenden Achsenabschnitt nur in der nahen Umgebung der Häufigkeiten, will man dagegen Unterschiede verschleiern, so wähle man einen möglichst großen Ausschnitt der zugehörigen Achse.

Manipulation der Achsen Bei zweidimensionalen Datensätzen gibt es die Möglichkeit, beide Achsen zu manipulieren. Die in Abbildung 2.28 links gezeigte Entwicklung der Arbeitslosenzahlen haben wir zweifach manipuliert. Die mittlere Grafik impliziert den Stillstand auf dem

[9] Zu einer Reihe von schönen Beispielen aus den Medien siehe z. B. Kütting, 1994.

Arbeitsmarkt („Regierungsarbeit ist erfolgslos“). Die rechte Grafik impliziert den rasanten Abfall der Arbeitslosenzahlen („Regierungsarbeit überragend“). In beiden Grafiken wurden deutlich sichtbar die Achsen manipuliert: Im ersten Fall wurden die Jahre nicht äquidistant auf der Abszissenachse abgetragen, im zweiten Fall führt die Skalierung der Ordinatenachse ohne Nullpunkt dazu, dass die Unterschiede grafisch vergrößert werden.

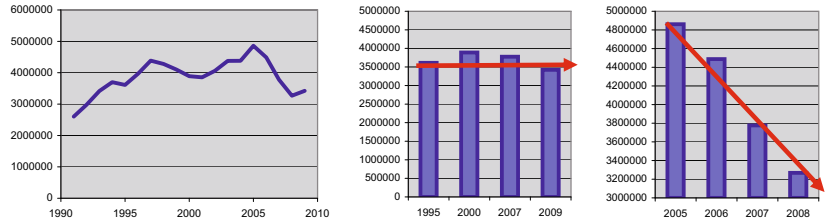


Abbildung 2.28: Entwicklung der Arbeitslosenzahlen in Deutschland, Quelle: DESTATIS; Stagnierende oder fallende Arbeitslosenzahlen in Deutschland?

Manipulation der Fläche Bis auf das Histogramm wird in grafischen Darstellungen eine Dimension (Höhe, Breite, Winkel) zur Repräsentation des Anteils einer Merkmalsausprägung an der Stichprobe verwendet. Eine Manipulation besteht dann, wenn man z.B. bei Balken den betreffenden Anteil sowohl in der Höhe als auch in der Breite repräsentiert. Dadurch wird das Verhältnis zwischen der Häufigkeit zweier Merkmalsausprägungen quadriert (Abb. 2.29).

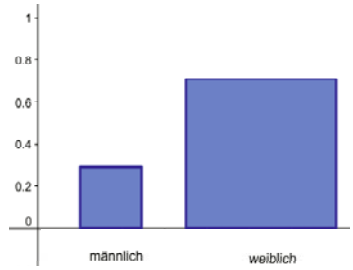


Abbildung 2.29: Verteilung der Geschlechter in der Pädagogischen Hochschule Freiburg

Bei den Zahlen, die der Abbildung 2.29 zu Grunde liegen, ist das Verhältnis der Häufigkeiten ($h_{218}(m) = 0,29$ und $h_{218}(w) = 0,71$) etwa 1:2,45. Das Verhältnis der in Abbildung 2.29 dargestellten Flächen ist dagegen $0,29^2 : 0,71^2$ oder etwa 1:6.

2.7.2 Lage- und Streuparameter

Minimalität Der Mittelwert ist derjenige Wert, für den die Summe der quadratischen Abweichungen minimal ist, für jeden anderen Wert c ist diese Summe größer.

Satz 4
Seien $x_1, ..., x_n$ die Merkmalsausprägungen eines Merkmals in einer Stichprobe vom Umfang n und $c \in \mathbb{R}$, dann gilt:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - c)^2$$

Der Beweis dieses Satzes funktioniert rein algebraisch. Es gilt zunächst:

Hilfssatz:
Seien $x_1, ..., x_n$ die Merkmalsausprägungen eines Merkmals in einer Stichprobe vom Umfang n und $c \in \mathbb{R}$, dann gilt:

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

Durch Umformung beweisen wir diesen Hilfssatz als Erstes:

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

Diese Erkenntnis wenden wir im Beweis an:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - c)^2 &= \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - c)^2 = \sum_{i=1}^{n}\left((x_i - \overline{x}) + (\overline{x} - c)\right)^2 \\
&= \sum_{i=1}^{n}\left((x_i - \overline{x})^2 + 2(x_i - \overline{x})(\overline{x} - c) + (\overline{x} - c)^2\right) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + 2(\overline{x} - c) \cdot \underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} + n(\overline{x} - c)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - c)^2 \\
&\geq \sum_{i=1}^{n}(x_i - \overline{x})^2, \text{da } n(\overline{x} - c)^2 > 0 \text{ für } c \neq \overline{x}
\end{aligned}$$

Damit ist der Satz bewiesen.

Im Hauptteil dieses Kapitels hatten wir weiterhin behauptet, dass der Median minimal zur Summe der absoluten Residuen ist (s. Abschnitt 2.4.4, S. 34).

Satz 5
Seien $x_1, ..., x_n$ die Merkmalsausprägungen eines Merkmals in einer Stichprobe vom Umfang n und $c \in \mathbb{R}$, dann gilt:

$$\sum_{i=1}^{n}|x_i - x_{0,5}| \leq \sum_{i=1}^{n}|x_i - c|$$

Die Beweisidee machen wir an einer Skizze zu einem Datensatz plausibel, bei dem der Median identisch mit einer Merkmalsausprägung x_i ist (vgl. Abb. 2.30):

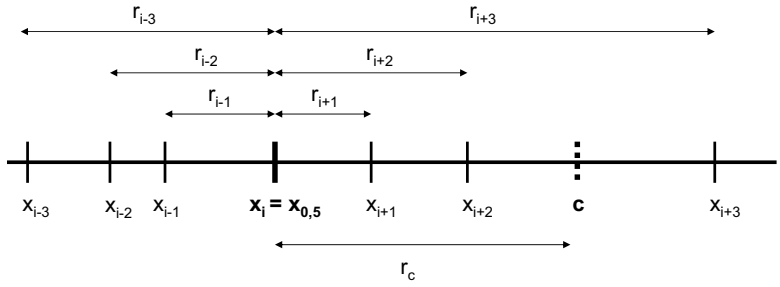


Abbildung 2.30: Beweisidee zur Minimalitätseigenschaft des Median

Misst man die Residuen im Vergleich der Verwendung von $x_{0,5}$ und c und berechnet man zusätzlich die Differenz der Abstände, so erhält man:

Datum	$\lvert x_i - x_{0,5}\rvert$	$\lvert x_i - c\rvert$	$\lvert x_i - c\rvert - \lvert x_i - x_{0,5}\rvert$
$i-3$	r_{i-3}	$r_{i-3}+r_c$	r_c
$i-2$	r_{i-2}	$r_{i-2}+r_c$	r_c
$i-1$	r_{i-1}	$r_{i-1}+r_c$	r_c
i	0	r_c	r_c
$i+1$	r_{i+1}	$r_c - r_{i+1}$	$r_c - 2r_{i+1}$
$i+2$	r_{i+2}	$r_c - r_{i+2}$	$r_c - 2r_{i+2}$
$i+3$	r_{i+3}	$r_{i+3} - r_c$	$-r_c$
Summe	—	—	$5r_c - 2r_{i+1} - 2r_{i+2}$

Mit Beachtung der Lage von c in diesem Fall ergibt sich die Abschätzung $5r_c - 2r_{i+1} - 2r_{i+2} \geq 5r_c - 2r_c - 2r_c = r_c (> 0)$. Für diesen Fallausschnitt ergibt sich damit die Behauptung. Um den Satz allgemein zu beweisen, müsste man die Beweisidee auf beliebige Lagen des Median sowie beliebig viele Daten erweitern, wobei sich stets eine analoge Abschätzung ergibt. Wir belassen es hier mit diesem Beweisansatz.

Mittelwert von Änderungen Wir führen als Ergänzung einen Lageparameter für Steigungen (Änderungen) ein. Will man eine durchschnittliche Steigung, z.B. den Durchschnitt jährlich variabler Zinssätze $(x_1, ..., x_n)$, bestimmen, so verwendet man das geometrische Mittel der (Steigungs-)Daten:

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Variationskoeffizient Die Streuung gibt die Homogenität eines Merkmals an. Dabei zeigt sich das empirische Phänomen, dass Merkmale mit einem numerisch hohen arithmetischen Mittel

häufig eine höhere Streuung und damit Standardabweichung haben als solche Merkmale mit numerisch kleinem arithmetischen Mittel. Man könnte dies so formulieren: „Bei einem numerisch hohen arithmetischen Mittel haben die Daten mehr Platz zu streuen." Möchte man die Homogenität verschiedener Merkmale mit möglicherweise unterschiedlichen arithmetischen Mitteln vergleichen, so bietet es sich an, die Streuung mit dem arithmetischen Mittel zu normieren. Dadurch ergibt sich der **Variationskoeffizient** $v = \frac{s}{\bar{x}}$.

Beispiel:
Gegeben ist die Streuung zum Alter A und Semester S von Studierenden (gesamter Datensatz): $s(A) \approx 3,2;\ \ s(S) \approx 2,7$.

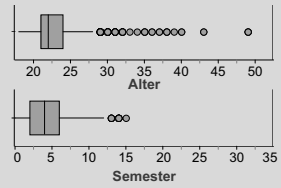


Abbildung 2.31: Verteilung von Alter und Semesteranzahl

Nimmt man allein diese Werte, so scheint das Merkmal Alter inhomogener verteilt zu sein als die Semesteranzahl. Beide folgenden Verteilungen sind in Abbildung 2.31 auf einer Skalenspannweite von 35 dargestellt. Normiert man diese Maßzahlen mit dem arithmetischen Mittel, so ergibt sich mit Abbildung 2.32 ein deutlich anderes Bild, das auch durch den Vergleich der Variationskoeffizienten ($v(A) = 0,14; v(S) = 0,63$) gestützt wird.

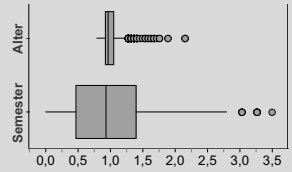


Abbildung 2.32: Normierte Verteilung von Alter und Semesteranzahl

2.7.3 Form der Verteilung

Die Form der Verteilung war bisher durch das Verhältnis der Lageparameter $x_{Mod}, x_{0,5}$ und $\bar{x}$ festgelegt worden (vgl. Abschnitt 2.5.1, S. 34 ff.). Die Steilheit bzw. Schiefe lässt sich aber auch quantifizieren. Ein Maß dafür ist die sogenannte **Schiefe** g mit

$$g = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^3}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Es wird also die mittlere kubische Abweichung vom arithmetischen Mittel mit dem Kubik der Standardabweichung normiert. Je negativer (positiver) die Werte von g sind, umso rechtssteiler (linkssteiler) ist die Verteilung. Werte um 0 deuten auf eine symmetrische Verteilung hin.

Beispiel:
Gegeben sind die Entfernungen, die die 40 Freiburger Studierenden zur Hochschule zurücklegen (vgl. Abb. 2.33).

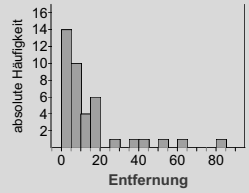


Abbildung 2.33: Entfernung zur Hochschule im Histogramm dargestellt

Das Schiefemaß hat hier den Wert $g = 2,4$, die Verteilung ist also (relativ) stark linkssteil. Eine konkrete Einordnung dieses Werts kann über den Vergleich mit anderen Verteilungen erfolgen.

Ein weiteres, aus der explorativen Datenanalyse stammendes Maß für die Schiefe basiert auf der Quotientenbildung von Quartilen (Quartilskoeffizient) und ist auf die mittleren (rund) 50% der Daten beschränkt:

$$QS_{0,25} = \frac{(x_{0,75} - x_{0,5}) - (x_{0,5} - x_{0,25})}{x_{0,75} - x_{0,25}}, \qquad \text{wobei gilt: } x_{0,75} \neq x_{0,25}$$

Bei diesem Schiefemaß wird (mit Bezug auf einen Boxplot) die Differenz der Abstände der Quartile zum Median mit dem Quartilsabstand normiert. Die Werte des Quartilskoeffizienten liegen zwischen -1 und 1. Bei dem Wert 0 ist die Verteilung symmetrisch. Je negativer (positiver) der Wert des Quartilskoeffizienten ist, desto rechtssteiler (linkssteiler) ist die Verteilung.

Beispiel:
Wiederum werden die Daten für die Entfernung der Freiburger Studierenden betrachtet.

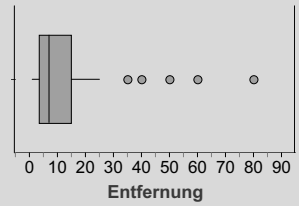


Abbildung 2.34: Entfernung zur Hochschule im Boxplot dargestellt

Das Schiefemaß hat im gleichen Beispiel wie oben den Wert $QS_{0,25} \approx 0,4$. Auch hier ergibt sich die relativ deutliche linkssteile Form der Verteilung.

Es gibt konstruierte Verteilungen, bei denen die zuerst genannte Festlegung von links- bzw. rechtssteilen Verteilungen über das Verhältnis der zentralen Lageparameter nicht mit den gerade definierten Maßzahlen korrespondiert. Im Allgemeinen sind aber die beiden Festlegungen äquivalent (vgl. Sachs, 1999).

2.8 Aufgaben

Aufgabe 2.1: In der folgenden Tabelle ist eine Teilstichprobe zu den Studierenden in Münster gegeben, die hinsichtlich ihres Studienfachs ausgewählt wurde. Vergleichen Sie die beiden Studierendengruppen. Verwenden Sie grafische Darstellungen, Lage- und Streuparameter, um Gemeinsamkeiten oder Unterschiede aufzudecken.

	Studierende des Lehramts				Studierende der Jura			
	geordnet nach Alter		geordnet nach Abi		geordnet nach Alter		geordnet nach Abi	
Nr.	Alter	Abi	Alter	Abi	Alter	Abi	Alter	Abi
1	19	2,4	25	1,2	18	1,5	19	1
2	19	2,4	25	1,2	19	2,3	21	1
3	20	1,9	26	1,3	19	1,3	20	1
4	20	2,6	22	1,7	19	1,3	21	1,1
5	20	2,5	20	1,9	19	1	21	1,2
6	20	2,7	23	1,9	19	2,2	27	1,2
7	20	2,6	23	1,9	19	2,5	19	1,3
8	20	3	29	1,9	20	1,8	19	1,3
9	20	2,7	22	2	20	2,6	18	1,5
10	20	2	25	2	20	2	27	1,6
11	20	2	22	2	20	1,9	21	1,7
12	21	2,6	20	2	20	2,7	23	1,7
13	21	2,4	20	2	20	1	25	1,7
14	21	2,2	28	2,1	21	1,7	20	1,8
15	22	2	25	2,2	21	2,2	28	1,8
16	22	2,4	21	2,2	21	2	20	1,9
17	22	2,6	22	2,4	21	1,1	29	1,9
18	22	2	24	2,4	21	1,2	23	1,9
19	22	3	21	2,4	21	2,1	21	2
20	22	1,7	19	2,4	21	1	20	2
21	23	2,8	19	2,4	21	2,3	24	2
22	23	2,7	20	2,5	23	1,7	25	2
23	23	1,9	20	2,6	23	2,6	23	2
24	23	1,9	21	2,6	23	1,9	21	2,1
25	24	2,8	22	2,6	23	2	26	2,1
26	24	3,4	20	2,6	24	2,5	21	2,2
27	24	2,4	20	2,7	24	2	19	2,2
28	25	1,2	20	2,7	25	1,7	19	2,3
29	25	2	23	2,7	25	2	21	2,3
30	25	2,2	24	2,8	26	2,1	28	2,4
31	25	1,2	23	2,8	27	1,6	24	2,5
32	26	2,9	26	2,9	27	1,2	19	2,5
33	26	1,3	20	3	28	1,8	20	2,6
34	28	2,1	22	3	28	2,4	23	2,6
35	28	3	28	3	29	1,9	20	2,7
36	29	1,9	24	3,4				

Aufgabe 2.2: Verwenden Sie den Gesamtdatensatz (Datenquelle: siehe Vorwort).

a) Untersuchen Sie anhand dieses Datensatzes Eigenschaften der Studierenden verschiedener Hochschulen mithilfe grafischer Darstellungen, Lage- und Streuparametern. Versuchen Sie, Ihre Interpretationen der Ergebnisse einerseits sachlich und andererseits plakativ zu gestalten.

b) Untersuchen Sie die Streuung verschiedener Merkmale sowie die normierte Streuung mit dem Variationskoeffizienten.

c) Untersuchen Sie, welche der Merkmale eine symmetrische bzw. asymmetrische Verteilung aufweisen. Überlegen Sie sich, ob sich die Form der Verteilung plausibel erklären lässt.

Aufgabe 2.3: In den Zusatzmaterialien ist ein Datensatz zu den Ergebnissen der Fußballbundesliga gegeben. Stellen Sie eine Regel auf, wieviele Punkte eine Mannschaft in einer Saison erhalten muss, um nicht abzusteigen bzw. um Meister zu werden. (Datenquelle: siehe Vorwort)

Aufgabe 2.4: Gegeben ist das Gehaltsgefüge einer fiktiven Firma:

		Gehalt in Euro	
Bezeichnung	**Anzahl**	**Durchschnitt**	**Gruppe**
Arbeiter	1.000	1.000	1.000.000
Arbeiter, gehobene Position	500	2.000	1.000.000
Leiter von Arbeitsgruppen	50	5.000	250.000
Management	10	10.000	100.000
Firmenbesitzer	1	1.000.000	1.000.000
Summe	1.561	—	3.350.000

a) Ein Arbeiter bekommt genau das Durchschnittsgehalt seiner Berufsgruppe, möchte aber mehr Geld bekommen. Der Firmenbesitzer möchte dagegen das Gehalt des Arbeiters nicht erhöhen. Argumentieren Sie aus der Sicht des Arbeiters und aus der Sicht des Firmenbesitzers, warum das Gehalt nicht ausreichend bzw. angemessen ist!

b) Das Gehaltsgefüge der Firma wird von außen, z. B. von einer Gewerkschaft einerseits und vom Arbeitgeberverband andererseits, begutachtet. Argumentieren Sie aus der Sicht der Gewerkschaft und aus der Sicht des Arbeitgeberverbandes, warum das Gehalt der Arbeiter angehoben werden muss bzw. beibehalten werden kann!

Hinweis: Verwenden Sie für Ihre Argumentationen insbesondere den Median und das arithmetische Mittel. Untersuchen Sie die Schiefe der Verteilung.

3 Analyse statistischer Daten zu zwei Merkmalen

Einstiegsbeispiel

Abbildung 3.1: Zusammenhänge von Merkmalen der Studierenden

Aufgabe 1: Untersuchen Sie Zusammenhänge zweier Merkmale (Eigenschaften) der Studierenden.

Worum es geht

Die Analyse der Eigenschaften von Studierenden kann wie im vorangegangenen Kapitel 2 zunächst einmal für sich selbst stehen. Jedes Ergebnis einer Datenanalyse zieht aber fast zwangsläufig die Frage nach einem Vergleich oder nach einer Ursache nach sich. Das ist bei jeder Messung so: Die Messung beispielsweise der Körpergröße von 1,85 m allein sagt noch nicht viel aus (höchstens im Vergleich zu dem Einheitsmaß Meter). Erst durch den Vergleich mit anderen Körpergrößen lassen sich die 1,85 m einordnen. Eine mögliche Ursache oder Teilerklärung einer Körpergröße ist das Geschlecht, eine andere die Körpergröße der Eltern: „je größer die Eltern, desto größer auch die Kinder“. Diese Fragen des Vergleichs und möglicher Ursachen können untersucht werden, wenn mehr als ein Merkmal erhoben wurde. Dann liegt ein zweidimensionaler (bzw. bivariater) oder mehrdimensionaler (bzw. multivariater) Datensatz vor, der auf Abhängigkeiten zwischen den erhobenen Merkmalen untersucht werden kann. Im Beispiel der Eigenschaften von Studierenden könnten mögliche Untersuchungsansätze sein:

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_3,

Merkmalsträger	Merkmale X	Merkmale Y	Beispielfragen eines Untersuchungsansatzes zu (X,Y)
Studierende	Rauchverhalten	Beziehungsverhalten	Haben Raucher häufiger einen Partner?
Studierende	Rauchverhalten	Abinote	Sind Raucher weniger leistungsfähig oder erzeugt Leistungsfähigkeit Rauch-Abstinenz?
Studierende	Zeit	Entfernung	Mit welcher Durchschnittsgeschwindigkeit kommen Studierende zur Hochschule?
Studierende	Zeit für Computer	Zeit für Musik	Sind Computerfreaks unmusikalisch?

Diese Art von Vergleichen, von Ursachenforschung und von Erklärungssuche werden wir im Folgenden diskutieren.

Arten von Zusammenhängen Unterschiedlich skalierte Merkmale lassen sich unterschiedlich bearbeiten. Ein Vergleich nach dem Geschlecht trennt beispielsweise den Datensatz nach einem nominalskalierten Merkmal, unter dessen Bedingung ein zweites, verschieden skaliertes Merkmal analysiert werden kann. Von den möglichen Paaren wollen wir folgende in diesem Kapitel behandeln:

Merkmale	nominalskaliert	metrisch skaliert
nominalskaliert	Kap. 3.1	Kap. 3.2
metrisch skaliert	Kap. 3.2	Kap. 3.3- 3.4

Clusterung Die Clusterung eines interessierenden Merkmals wie z.B. dem Alter (metrisch) oder dem Rauchverhalten (nominal) sowie die darauf basierende Analyse eines zweiten Merkmals (männlich/weiblich; gute/mittlere/schlechte Abinote etc.) wird einen Teil der Untersuchungen ausmachen. Die Clusterung ist eine elementare Methode des Vergleichs, der bei dem Vergleich der Studierenden in Freiburg und Münster in Kapitel 2.6 bereits implizit eingesetzt wurde.

Regression, funktionale Anpassung „Je weiter die Studierenden entfernt von der Hochschule wohnen, desto länger brauchen sie für den Weg.“ In Kapitel 3.3 werden wir funktionale Abhängigkeiten (insbesondere lineare) von zwei metrisch skalierten Merkmale untersuchen. Es ist das bivariate Pendant zur Ermittlung von zentralen Lageparametern (Kap. 2): Dort war der typische Wert in der Häufigkeitsverteilung eines Merkmals X gesucht, um den Datensatz darauf (und zusätzlich auf ein Streumaß) zu reduzieren. Hier wird unter der Voraussetzung des gegebenen Wertes eines Merkmals X der typische Wert der Häufigkeitsverteilung eines Merkmals Y gesucht. So wird der bivariate Datensatz auf ein Zusammenhangsmuster (auch hier wird zusätzlich die Streuung mitbedacht) reduziert.

Abhängigkeit Die Messung der Abhängigkeit zweier Merkmale, unabhängig von ihrer Skalierung, wird einen weiteren Aspekt der folgenden Analysen ausmachen. Vor dem Konstruieren eines Modells (z.B. einer Geraden, die als Modell eines zweidimensionalen Datensatzes dienen soll) stellt sich die Frage, wie gut das Modell auf die in den Daten steckende Realität passt. Daher werden wir in diesem Kapitel auch die Güte der Abhängigkeit zweier Merkmale untersuchen.

3.1 Zusammenhänge nominalskalierter Merkmale

Geht man von der Betrachtung eines eindimensionalen Merkmals X auf ein zweidimensionales Merkmal (X,Y) über, dann bedarf es einer kleinen Erweiterung des bisher verwendeten Häufigkeitsbegriffs.

- X habe die Merkmalsausprägungen x_i $(i = 1,...,s)$ mit den relativen Häufigkeiten $h_n(x_i)$ bei einer Stichprobe vom Umfang n.
- Y habe die Merkmalsausprägungen y_j $(j = 1,...,t)$ mit den relativen Häufigkeiten $h_n(y_i)$ bei derselben Stichprobe vom Umfang n.[1]
- Das zweidimensionale Merkmal (X,Y) besteht aus geordneten Paaren von Merkmalsausprägungen (x_i,y_j), also Merkmalsträgern, auf die die Merkmalsausprägung x_i *und* die Merkmalsausprägung y_i zutrifft:

  $$(X,Y) = \{(x_i,y_j)|x_i \in X \text{ und } y_j \in Y; i = 1,...,s;\ \ j = 1,...,t\}$$

  Algebraisch betrachtet handelt es sich bei dem zweidimensionalen Merkmal (X,Y) um das *kartesische Produkt* $X \times Y$ der Mengen X und Y.
- Dementsprechend ergeben sich in einer Stichprobe die Häufigkeiten $H_n(x_i,y_j)$ bzw. $h_n(x_i,y_j) = \frac{H_n(x_i,y_j)}{n}$.[2]
- Betrachtet man schließlich die Häufigkeit der verschiedenen Merkmalsausprägungen in Abhängigkeit voneinander, so werden wir das durch $H_n(x_i|y_j)$ bzw. $h_n(x_i|y_j)$ ausdrücken. In Worten: Die Häufigkeit für das Zutreffen der Merkmalsausprägung x_i unter der Bedingung, dass auch die Merkmalsausprägung y_j zutrifft.[3] Diese Häufigkeit bedeutet eine Einschränkung. Man betrachtet nicht mehr alle statistischen Einheiten hinsichtlich einer Merkmalsausprägung x_i, sondern nur noch diejenigen, auf die als Bedingung eine andere Merkmalsausprägung, nämlich y_j, zutrifft.

Bevor wir ein Beispiel betrachten, werden wir eine Strukturierungsmethode für die Merkmale X und Y mit je zwei Merkmalsausprägungen einführen, die **Vierfeldertafel**.[4] In diesem Buch werden wir stets das möglicherweise ursächliche, unabhängige Merkmal X in die Spalten aufnehmen und das abhängige Merkmal Y in die Zeilen. Was möglicherweise ursächlich, was abhängig ist, steht nicht per se fest, sondern basiert jeweils auf einer Interpretation des Sachkontexts. In den Beispielen nehmen wir zusätzlich eine passende grafische Darstellung in Form eines **gruppierten Punktdiagramms** auf.

	x_1	x_2	Summe
y_1	$H_n(x_1,y_1)$	$H_n(x_2,y_1)$	$H_n(y_1)$
y_2	$H_n(x_1,y_2)$	$H_n(x_2,y_2)$	$H_n(y_2)$
Summe	$H_n(x_1)$	$H_n(x_2)$	n

[1] Wir werden insbesondere Merkmale mit $s = t = 2$ untersuchen.

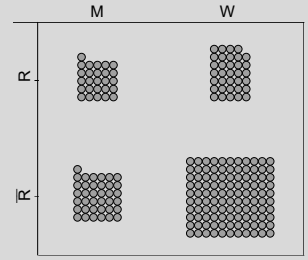
[2] Formal müsste man hier z.B. $H_n((x_i,y_j))$ schreiben, wir verzichten aber zugunsten der Lesbarkeit auf die innere Klammer zur Verdeutlichung eines Paares von Merkmalsausprägungen.

[3] Das korrespondiert mit der Definition der bedingten Wahrscheinlichkeit, die wir in Kapitel 6.1 betrachten werden.

[4] Vierfeldertafeln sind ein Spezialfall von s-t-Kontingenztafeln, die einen Zusammenhang zwischen einem Merkmal X mit x_i $(i = 1,...,s)$ Merkmalsausprägungen und einem Merkmal Y mit y_j $(j = 1,...,t)$ Merkmalsausprägungen abbilden. Wir betrachten diesen allgemeineren Fall hier nicht näher.

Beispiel:
Gegeben sind die Daten zu den Merkmalen X: Geschlecht mit $x_1 = M$ für männlich und $x_2 = W$ für weiblich sowie Y: Rauchverhalten von Studierenden der Pädagogischen Hochschule Freiburg (Erhebung 2010) mit $y_1 = R$ für Raucher und $y_2 = \overline{R}$ für Nichtraucher. Unter Verwendung von absoluten Häufigkeiten ergeben sich die nachfolgende Vierfeldertafel und das entsprechende gruppierte Punktdiagramm:

	M	W	Summe
R	$H_{218}(M,R) = 26$	$H_{218}(W,R) = 34$	$H_{218}(R) = 60$
$\overline{R}$	$H_{218}(M,\overline{R}) = 37$	$H_{218}(W,\overline{R}) = 121$	$H_{218}(\overline{R}) = 158$
Summe	$H_{218}(M) = 63$	$H_{218}(W) = 155$	$n = 218$



Die Verhältnisse lassen sich ebenso mit relativen Häufigkeiten in einer Vierfeldertafel darstellen:

	M	W	Summe
R	$h_{218}(M,R) \approx 0,12$	$h_{218}(W,R) \approx 0,16$	$h_{218}(R) \approx 0,28$
$\overline{R}$	$h_{218}(M,\overline{R}) \approx 0,17$	$h_{218}(W,\overline{R}) \approx 0,55$	$h_{218}(\overline{R}) \approx 0,72$
Summe	$h_{218}(M) \approx 0,29$	$h_{218}(W) \approx 0,71$	1

Aus der Vierfeldertafel mit den absoluten Häufigkeiten lässt sich also beispielsweise entnehmen:

- Eingeschränkt auf die 63 männlichen Studierenden gibt es $H_{218}(R|M) = 26$ Raucher.
- Eingeschränkt auf die 158 Nichtraucher sind $H_{218}(W|\overline{R}) = 121$ weibliche Studierende.

Die bedingten relativen Häufigkeiten errechnen sich durch:

$$h_{218}(R|M) = \frac{H_{218}(M,R)}{H_{218}(M)} = \frac{26}{63} = \frac{\frac{26}{218}}{\frac{63}{218}} = \frac{h_{218}(M,R)}{h_{218}(M)} \approx 0,41; \qquad h_{218}(\overline{R}|M) = 1 - h_{218}(R|M) \approx 0,59$$

und

$$h_{218}(W|\overline{R}) = \frac{H_{218}(W,\overline{R})}{H_{218}(\overline{R})} = \frac{121}{158} = \frac{\frac{121}{218}}{\frac{158}{218}} = \frac{h_{218}(W,\overline{R})}{h_{218}(\overline{R})} \approx 0,77; \qquad h_{218}(M|\overline{R}) = 1 - h_{218}(W|\overline{R}) = 0,23$$

Das bedeutet z. B., rund 41% der männlichen Studierenden sind Raucher und rund 77% der Nichtraucher sind weibliche Studierende.

Die strukturierte Darstellung in einer Vierfeldertafel schafft eine Übersicht über die Datenlage. Sie beantwortet aber noch nicht unmittelbar die in diesem Kapitel aufgeworfene Frage, ob es einen Zusammenhang zwischen zwei Merkmalen (im Beispiel hier zwischen dem Geschlecht

und dem Rauchverhalten) gibt. Dem gehen wir mit Hilfe einer gewichteten **grafischen Vierfeldertafel** sowie dem (empirischen) **Einheitsquadrat** nach. Wir erläutern die Konstruktion anhand des Beispiels des Zusammenhangs zwischen Geschlecht und Rauchverhalten (vgl. Abb. 3.2).

Beispiel:
Konstruktion von grafischer Vierfeldertafel und Einheitsquadrat:

- Ausgangsfigur ist ein Quadrat mit der Seitenlänge 1.
- Das Quadrat wird senkrecht geteilt, die Breite der beiden entstehenden Rechtecke entspricht den relativen Häufigkeiten des ersten Merkmals ($h_{218}(M)$ und $h_{218}(W)$).
- Diese beiden Rechtecke werden ein weiteres Mal waagerecht unterteilt. Die Höhen der jeweils entstehenden zwei Rechtecke entsprechen den relativen Häufigkeiten, die M bzw. W als Bedingung haben. Dadurch ergeben sich folgende Höhen der vier im Quadrat enthaltenen Rechtecke:

  Höhe des oberen linken Rechtecks: $h_{218}(R|M) = \frac{h_{218}(M,R)}{h_{218}(M)}$

  Höhe des unteren linken Rechtecks: $h_{218}(\overline{R}|M) = \frac{h_{218}(M,\overline{R})}{h_{218}(M)} = 1 - h_{218}(R|M)$

  Höhe des oberen rechten Rechtecks: $h_{218}(R|W) = \frac{h_{218}(W,R)}{h_{218}(W)}$

  Höhe des unteren rechten Rechtecks: $h_{218}(\overline{R}|W) = \frac{h_{218}(W,\overline{R})}{h_{218}(W)} = 1 - h_{218}(R|W)$
- Damit ergibt sich beispielsweise für die Fläche des rechten unteren Rechtecks:

$$h_{218}(W) \cdot h_{218}(\overline{R}|W) = h_{218}(W,\overline{R})$$

Damit ergeben sich die grafische Vierfeldertafel und das zugehörige Einheitsquadrat, die in Abbildung 3.2 zu sehen sind.

	M	W	Summe
R	0,12 (26)	0,16 (34)	0,28 (60)
$\overline{R}$	0,17 (37)	0,55 (121)	0,72 (158)
Summe	0,29 (63)	0,71 (155)	1 (218)

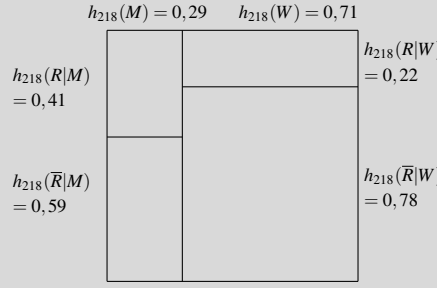


Abbildung 3.2: Grafische Vierfeldertafel und Einheitsquadrat mit den Bezeichnungen R: Raucher, $\overline{R}$: Nichtraucher, M: männlich und W: weiblich

Ein einfaches Assoziationsmaß A, d.h. ein Maß für den Zusammenhang der beiden Merkmale, erhält man geometrisch über den Abstand der waagerechten Unterteilungen des Einheitsquadrats. Je größer dieser Abstand ist, desto größer ist deskriptiv der Zusammenhang zwischen beiden Merkmalen. Algebraisch lässt sich das Assoziationsmaß A durch die Differenz der beiden bedingten Häufigkeiten $h_n(y_1|x_1)$ und $h_n(y_1|x_2)$ beschreiben

$$A = h_n(y_1|x_1) - h_n(y_1|x_2).$$

Beispiel:
Betrachtet man wie im vorausgehenden Beispiel die Merkmale Geschlecht und Rauchverhalten, so ergibt sich für das Assoziationsmaß

$$A = h_n(R|M) - h_n(R|W) = 0,41 - 0,22 = 0,19.$$

Wird die gemeinsame Verteilung zu den Merkmalen Rauch- und Beziehungsverhalten ($X = \{R, \overline{R}\}$ und $Y = \{S, \overline{S}\}$) von Studierenden analysiert, so ergeben sich aus den zugrundeliegenden Daten die grafische Vierfeldertafel und das Einheitsquadrat, die in Abbildung 3.3 zu sehen sind.

	R	$\overline{R}$	Summe
S	0,12 (27)	0,31 (67)	0,43 (94)
$\overline{S}$	0,15 (33)	0,42 (91)	0,57 (124)
Summe	0,27 (60)	0,73 (158)	1 (218)

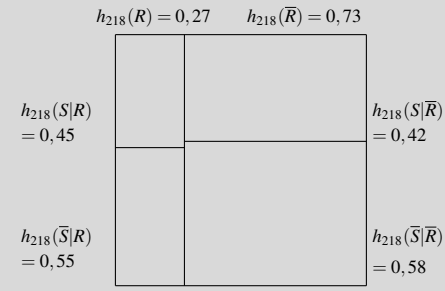


Abbildung 3.3: Grafische Vierfeldertafel und Einheitsquadrat mit den Bezeichnungen R: Raucher, $\overline{R}$: Nichtraucher, S: Single und $\overline{S}$: Nicht-Single

Analysiert man den Zusammenhang von Rauch- und Beziehungsverhalten mit diesem Assoziationsmaß, so ergibt sich der Wert:

$$A = h_{218}(S|R) - h_{218}(S|\overline{R}) = 0,45 - 0,42 = 0,03.$$

Im Gegensatz zu der vorangegangenen Analyse des Zusammenhangs zwischen den Merkmalen Geschlecht und Rauchverhalten, wo der Wert des Assoziationsmaßes mit $0,18$ deutlich von 0 verschieden ist, scheint im zweiten Fall mit einem Wert von $A = 0,03$ ein Zusammenhang zwischen Rauch- und Beziehungsverhalten eher auszuschließen zu sein.

Für A gilt $-1 \leq A \leq 1$. Die Interpretation von A unterstellt einen Zusammenhang der beiden Merkmale insbesondere bei Werten, die sich stark von 0 unterscheiden. Entsprechend wird auf die Unabhängigkeit beider Merkmale bzw. einen fehlenden Zusammenhang bei einem Wert von A nahe 0 geschlossen. Was „stark von 0 unterschieden“ bzw. „nahe 0“ heißt, ergibt sich daraus, welche Größe die Stichprobe insgesamt und die Teilstichproben haben (in Kap. 8 gehen wir darauf ein). Qualitativ lässt sich aber der Wert von A bei identischen Stichproben bzw. Teilstichproben vergleichen.

Ein anderes gebräuchliches Assoziationsmaß für zwei nominalskalierte Merkmalsausprägungen ist die **odds ratio** oder das **Kreuzprodukt** q, das ähnlich zu A aufgebaut ist. Die odds ratio beschreibt geometrisch im Einheitsquadrat betrachtet das Verhältnis der beiden waagerechten Unterteilungen durch

$$q = \frac{h_n(y_1|x_1)}{h_n(y_2|x_1)} : \frac{h_n(y_1|x_2)}{h_n(y_2|x_2)} = \frac{h_n(y_1|x_1) \cdot h_n(y_2|x_2)}{h_n(y_2|x_1) \cdot h_n(y_1|x_2)}.$$

Das Kreuzprodukt der absoluten Häufigkeiten innerhalb der Vierfeldertafel ist äquivalent zu dieser Darstellung. Für die Werte der odds ratio gilt $q \geq 0$. Sind die beiden Seiten-Verhältnisse im Einheitsquadrat gleich, so besteht kein Zusammenhang und $q = 1$. Vertauscht man Zeilen und Spalten in der Vierfeldertafel, so erhält man statt q den reziproken Wert $1/q$, der zwischen 0 und 1 liegt bzw. der 1 oder größer ist.

Beispiel:
Gegeben sind die gemeinsamen Verteilungen zu dem Geschlecht und dem Rauchverhalten der Studierenden. Es ergibt sich:

$$q = \frac{26 \cdot 121}{37 \cdot 34} \approx 2{,}5; \quad 1/q \approx 0{,}4$$

Dieser Wert der odds ratio weist auf einen Zusammenhang der beiden Merkmale hin, nämlich darauf, dass die männlichen Studierenden einen größeren Hang zum Rauchen haben als ihre Kommilitoninnen.

Betrachtet man entsprechend die gemeinsamen Verteilungen zu dem Rauch- und Beziehungsverhalten der Studierenden, so ergibt sich:

$$q = \frac{27 \cdot 91}{33 \cdot 67} \approx 1{,}11 \quad 1/q \approx 0{,}9$$

Der Wert der odds ratio von $1/q = 0{,}9$ weist darauf hin, dass es keinen (oder bestenfalls nur sehr schwachen) Zusammenhang zwischen den beiden Merkmalen des Rauch- und Beziehungsverhaltens der Studierenden gibt.

Während wir den Einfluss des Stichprobenumfangs in Kapitel 8 durch Simulation untersuchen, kann der Einfluss der freien Parameter im Einheitsquadrat anhand einer Datei explorativ untersucht werden, die im Zusatzmaterial zu diesem Buch[5] enthalten ist.

[5] Die Bezugsquelle ist im Vorwort angegeben.

3.2 Nominalskaliertes X – metrisch skaliertes Y

Zweidimensionale Merkmale (X,Y), bei denen ein Merkmal nominalskaliert, das andere metrisch skaliert ist, haben wir implizit bereits in Kapitel 2.6 behandelt. Dort war das nominalskalierte Merkmal X die Hochschule mit $x_1 =$ Pädagogische Hochschule Freiburg und $x_2 =$ Universität Münster. Mit diesem Merkmal, das den gesamten Datensatz zu den Studierenden **clustert**, d.h. in Gruppen aufteilt, haben wir z.B. die Merkmale Entfernung oder Computer untersucht.

Beispiel:
Gegeben ist die Verteilung zu den Abiturnoten der PH-Studierenden (wir betrachten die Durchschnittsnoten hier als metrisch skaliert). Als nominalskaliertes Merkmal zur Clusterung des Merkmals Abiturnote verwenden wir das Geschlecht (Abb. 3.4).

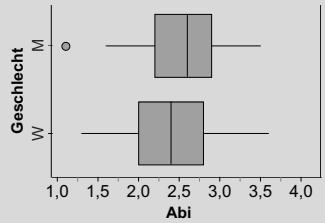


Abbildung 3.4: Abiturnoten der PH-Studierenden, geclustert nach dem Geschlecht

Der Vergleich ergibt hier eine tendenziell bessere Abiturnote für die weiblichen Studierenden, da die Quartile und der Median jeweils kleiner sind als bei den männlichen Studierenden. Die weiblichen Studierenden sind zudem etwas inhomogener hinsichtlich ihrer Abiturleistungen verteilt. Beide Verteilungen sind annähernd symmetrisch.

Beispiel:
Wir clustern das metrisch skalierte Merkmal Alter in zwei Gruppen, nämlich *jung* (23 Jahre und jünger) und *alt* (älter als 23). Für beide Cluster stellen wir das Merkmal Entfernung zur Hochschule dar.

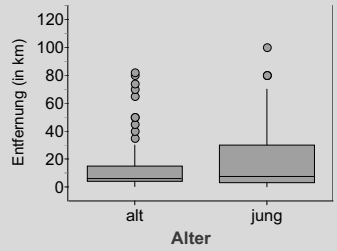


Abbildung 3.5: Entfernung zur Hochschule, geclustert nach dem Alter (jung/alt)

Mit Blick auf Abbildung 3.5 ist erkennbar, dass die älteren Studierenden tendenziell homogener (Quartilsabstand und Median) in der Nähe der Hochschule wohnen.

Für eine Clusterung ist jedes nominalskalierte Merkmal in einem Datensatz mit mehreren Merkmalen geeignet. Es ist zusätzlich wie im zweiten Beispiel gezeigt möglich, ein metrisch skaliertes Merkmal in ein nominalskaliertes zu transformieren und für eine Clusterung zu verwenden, um dadurch einen schnellen qualitativen Zugriff zu möglichen Zusammenhängen zweier Merkmale zu bekommen. Im zweiten Beispiel wurde durch die Clusterung das metrisch skalierte Merkmal Alter künstlich in ein nominalskaliertes Merkmal (jung – alt) transformiert. Bei der Interpretation ist Vorsicht geboten: eine mögliche einfache Aussage wie „je älter die Studierenden, desto näher wohnen sie an der Hochschule“ weist bereits über das vorangehende Beispiel hinaus, weil in der Aussage ein vermeintlicher Zusammenhang zwischen zwei metrisch skalierten Merkmalen zum Ausdruck gebracht wird. Auf die Untersuchung des Zusammenhangs zweier metrisch skalierter Merkmale gehen wir im folgenden Abschnitt ausführlicher ein.

3.3 Zusammenhänge metrisch skalierter Merkmale

Der Zusammenhang zweier metrisch skalierter Merkmale lässt sich am einfachsten beschreiben, wenn dieser linear ist. Ein linearer Zusammenhang zweier Merkmale lässt sich z.B. durch „je größer die Merkmalsausprägungen von X, desto größer die Merkmalsausprägungen von Y“ (positiver Zusammenhang) oder „je größer die Merkmalsausprägungen von X, desto kleiner die Merkmalsausprägungen von Y“ (negativer Zusammenhang) ausdrücken. Soll ein solcher Zusammenhang linear modelliert werden, so lassen sich sämtliche (beliebig viele) Paare von Merkmalsausprägungen (x,y) auf eine Geradengleichung $y = ax + b$ reduzieren. Analog zur Reduktion eines univariaten Datensatzes auf einen eindimensionalen Lageparameter (wie z.B. das arithmetisches Mittel) ergibt sich hier die Reduktion eines bivariaten Datensatzes auf einen zweidimensionalen Lageparameter, repräsentiert durch die Geradengleichung.

Fragt man nach der Güte eines linearen Zusammenhangs zweier metrisch skalierter Merkmale, dann lässt sich wie bei den in Abschnitt 3.1 diskutierten Assoziationsmaßen zum Zusammenhang zweier nominalskalierter Merkmale ein Maß bestimmen, der sogenannte **Korrelationskoeffizient.**

Beide Methoden zur Analyse zweidimensionaler, metrisch skalierter Merkmale (X,Y) werden wir im Folgenden stets an einem Datensatz diskutieren, der den Zusammenhang zwischen der Entfernung zur Hochschule und der dafür benötigten Zeit umfasst[6] – beschränkt auf den kleinen Anteil der Studierenden, die mit dem Auto zur Hochschule kommen:

Zeit in min (X)	5	15	20	20	25	30	40	50	60	60	75	120
Entfernung in km (Y)	1,5	7	10	15	12	10	16	50	60	74	80	70

3.3.1 Punktwolke, Gerade und Residuen

Eine einfache grafische Darstellung eines zweidimensionalen Merkmals (X,Y), bei dem X und Y metrisch skaliert sind, ist eine **Punktwolke**, die aus den Punkten mit den Koordinaten (x_i, y_i) $(i = 1, ..., n)$ besteht.

[6] Die Daten stammen aus der Erhebung, die im Jahr 2010 an der Pädagogischen Hochschule Freiburg durchgeführt wurde.

Modelliert man den Zusammenhang zweier metrisch skalierter Merkmale linear, so wird der Datensatz durch eine Gerade ersetzt. Die einfachste (und im Allgemeinen auch robuste) Methode ist das Einpassen einer solchen Geraden in die Punktwolke durch „Augenmaß“. Für die Bestimmung der Parameter einer solchen Geraden identifiziert man zwei auf dieser Geraden liegende Punkte $P_1(x_{f,1};y_{f,1})$ und $P_2(x_{f,2};y_{f,2})$. Der Index f steht für **Fit**, die Bezeichnung eines Anpassungsmodells an Daten (hier die Gerade). Diese beiden Punkte können, müssen aber nicht mit Punkten der Punktwolke übereinstimmen. Die Parameter a und b der Gleichung der Anpassungsgeraden $g_f : y_f = a \cdot x_f + b$ erhält man durch[7]:

$$a = \frac{y_{f,2} - y_{f,1}}{x_{f,2} - x_{f,1}}; \quad b = y_{f,1} - a \cdot x_{f,1}$$

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y), die sich in der Punktwolke von Abbildung 3.6 zeigt. Es wird eine Gerade g_f eingefügt und zwei Punkte $P_1(0;-1)$ und $P_2(100;79)$ auf dieser Geraden identifiziert. Es ergibt sich:

$$a = \frac{79-(-1)}{100} = 0,8; \quad b = -1 - 0,8 \cdot 0 = -1.$$

b war in diesem speziellen Fall durch die Wahl von P_1 unmittelbar gegeben. Die eingepasste Gerade g_f hat also die Gleichung $y_f = 0,8x_f - 1$

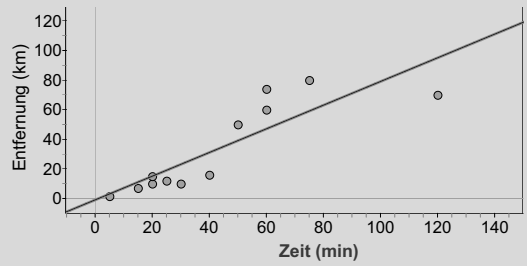


Abbildung 3.6: Zeit (X) und Entfernung (Y) zur Hochschule mit Anpassungsgerade

Die eingepasste Gerade g_f ersetzt die Punktwolke. Die Geradeneinpassung wird dadurch motiviert, dass zu jeder potentiellen Merkmalsausprägung x (bzw. y) eine mittlere Merkmalsausprägung y_f (bzw. x_f) abgeschätzt werden kann. Das ist insbesondere im Intervall $[x_{Min};x_{Max}]$ möglich (Extrapolationen über die Grenzen des Datensatzes hinaus sind im Allgemeinen mit Schwierigkeiten verbunden). Solch eine Schätzung ist umso sinnvoller, je besser die Anpassung einer Geraden möglich ist.

[7] Gängige schulrelevante Programme bieten bereits das Einfügen einer beweglichen Gerade in eine Punktwolke an. Dort entfällt bei Softwareunterstützung die Berechnung einer Geradengleichung.

Ein Gütekriterium für ein Anpassungsmodell ist darin zu sehen, dass dieses als Ganzes im Datenkontext sinnvoll interpretiert werden kann.

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Entfernung (Y) und Zeit (X). Passt man die Gerade $g_f : y_f = 0,8x_f - 1$ in die Punktwolke beider Merkmale ein, so ergibt sich nach diesem Modell, dass eine Person, die 10 km von der Hochschule entfernt wohnt, im Mittel etwa 15 Minuten für den Weg zur Hochschule benötigt:

$$10 = 0,8x_f - 1 \Leftrightarrow x_f = \frac{11}{\frac{4}{5}} = 13,75 \approx 15$$

Der Parameter $a = 0,8$ repräsentiert die mittlere Geschwindigkeit, mit der die Studierenden zur Hochschule fahren (in Kilometer pro Minute). Rechnet man diese in die üblichen Einheiten um, so ergibt sich eine Geschwindigkeit von 48 km/h.
Der Parameter $b = -1$ ist zunächst schwerer zu interpretieren. Im Sachkontext bedeutet er, dass schon vor dem ersten zurückgelegten Meter Zeit verbraucht wurde. Das ist aber plausibel, da etwa auch für das Ein- und Aussteigen Zeit benötigt wird. Bezogen auf das Merkmal Entfernung (Y) bedeutet dies genauer: Durch z.B. das Ein- und Aussteigen verlängert sich der Weg ausgehend von der gegebenen Geschwindigkeit um virtuelle 1 km. Bezogen auf das Merkmal Zeit (X) bedeutet dies genauer: Für z.B. das Ein- und Aussteigen gehen den Studierenden (bezogen auf die gegebene Durchschnittsgeschwindigkeit) rund 1,5 Minuten verloren.

Ein weiteres Gütekriterium für die Geradenanpassung ergibt sich aus der Frage, wie gut eine Gerade (oder allgemeiner ein Fit) die Punktwolke repräsentiert. Eine Methode, um die Anpassungsgüte zu beurteilen und zu verbessern, ist die Bestimmung der **Residuen** r_i ($i = 1, ..., n$). Die Differenz $y_i - y_f(x_i)$ zwischen der Merkmalsausprägung y_i eines Datenpunktes mit den Koordinaten $(x_i; y_i)$ und dem Funktionswert $y_{f,i}$ eines Modellpunktes $(x_i; y_{f,i}(x_i))$, der sich aus dem Fit (hier der Geraden) ergibt, wird als Residuum r_i bezeichnet. Geometrisch betrachtet entspricht dies der vertikalen Entfernung zwischen einem Datenpunkt und dem zugehörigen Modellpunkt. Bei Punkten oberhalb des Fits wird die Entfernung positiv, bei Punkten unterhalb des Fits wird die Entfernung negativ genommen (vgl. Abb. 3.7). Ein numerisches Kriterium für die Güte der

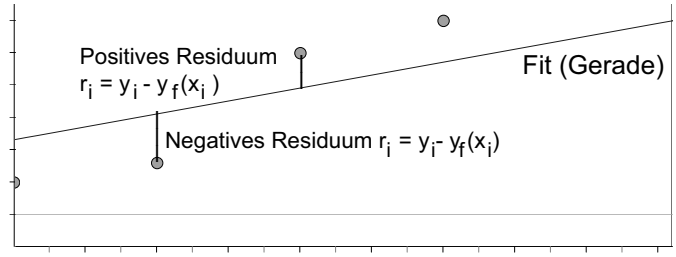


Abbildung 3.7: Positives und negatives Residuum hinsichtlich eines Fits

Anpassung eines bestimmten Modells (Fits) ist die Summe der Residuenbeträge: Sie sollte (möglichst) minimal sein unter der Voraussetzung, dass der Fit im Datenkontext (s. o.) sinnvoll interpretierbar bleibt. Für das numerische Kriterium werden wir allerdings keine algorithmische Methode verwenden, um das tatsächliche Minimum der absoluten Residuensumme zu bestimmen.[8]

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y) sowie die als Modell eingepasste Gerade $g_f : y_f = 0,8x_f - 1$. Es ergeben sich die Residuen:

Entfernung (Y)	Zeit (X)	Modell-Entfernung Y_f	Residuenbeträge
1,5	5	3	|-1,5|
7	15	11	|-4|
10	20	15	|-5|
10	30	23	|-13|
12	25	19	|-7|
15	20	15	|0|
16	40	31	|-15|
50	50	39	|11|
60	60	47	|13|
70	120	95	|-25|
74	60	47	|27|
80	75	59	|21|
		Summe:	142,5

Hinsichtlich des Kriteriums der Residuenbetragssumme kann man andere, auch bessere Modelle finden. So ergibt beispielsweise die Anpassungsgerade $y_f = 0,85 \cdot x - 4$ eine kleinere Residuensumme von 138,5. Im Kontext interpretiert entspricht diese Anpassungsgerade einer durchschnittlichen Geschwindigkeit von 51 km/h und einem Verlust durch Ein- und Ausstieg von etwa 5 Minuten. Diese Anpassungsgerade stellt also sowohl in numerischer wie auch interpretativer Hinsicht eine Verbesserung dar.

Betrachtet man die Form der Residuen als Punktwolke, so ergibt sich daraus ein weiteres Gütekriterium für die Anpassung der Daten. Dieses Kriterium basiert auf folgender Modellstrukturgleichung:

$$\begin{array}{lllll} \text{Daten} & = & \text{Fit} & + & \text{Residuen} \quad \text{(oder)} \\ \text{Daten} & = & \text{Muster} & + & \text{Zufall} \end{array}$$

Die (zweidimensionalen) Daten sollen durch einen Fit (ein Modell, ein Muster) so beschrieben werden, dass die Punktwolke der Residuen möglichst musterlos und zufällig aussieht. Das bedeutet, dass die Anpassung eines Fits verbessert werden kann, wenn die Residuen noch ein Muster aufweisen. Im einfachen Fall kann das zu einer Veränderung der Parameter eines Fits führen (wie im Beispiel oben die Veränderung der Parameter a und b), im komplexeren Fall zu einer Änderung des im Fit enthaltenen Modells selbst (also der Gerade). Neben der Frage, wie

[8] Da es sich hier um die Bestimmung des Minimums einer Abstandsfunktion $f(r) = \sum_{i=1}^{n} |r_i|$ handelt, ist das Kalkül der Differentialrechnung nicht anwendbar.

passend ein lineares Modell überhaupt ist, sind auch noch die Grenzen einer Geradenanpassung nach Augenmaß zu diskutieren. Zumindest die Summe der Residuenbeträge, die algorithmisch, nach festen Regeln bestimmt werden kann, ermöglicht einen „objektiven" Blick auf die Modellierung. Die Verwendung des Augenmaßes ist dagegen nicht objektiv kontrollierbar, erzeugt aber bei einfachen Datensätzen kaum schlechtere Modellierungen als mathematisch elaboriertere Methoden. Nachteilig oder unbrauchbar wird das nicht-algorithmisierte Verfahren allerdings bei komplexeren Datensätzen. Erzeugen beispielsweise verschiedene Daten den gleichen Punkt im Streudiagramm, so wird das in der Punktwolke, in die die Gerade nach Augenmaß eingepasst wird, nicht sichtbar, weil die beiden Punkte übereinander liegen – es ist nur ein Punkt statt zwei Punkte sichtbar. Dennoch sollte natürlich ein solcher Mehrfachpunkt stärker bei der Modellanpassung gewichtet sein als ein einfacher Punkt.

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y). In der Abbildung 3.8 links ist die Gerade eingepasst, die hinsichtlich der Summe der absoluten Residuen annähernd als optimal erachtet wurde. Das zugehörige Residuendiagramm ist der Skalierung der x-Achse entsprechend darunter gezeichnet. Rechts ist eine offensichtlich weniger gut passende Gerade eingezeichnet. Hier zeigen die Residuen das (banale) *Muster*, nahezu ausschließlich negativ zu sein.

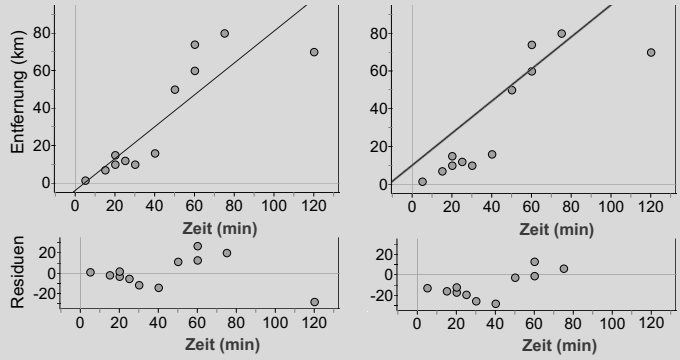


Abbildung 3.8: Annähernd optimale Geradenanpassung und wenig optimale Geradenanpassung, bei der die Residuen ein Muster aufweisen

Betrachtet man allerdings die linke Darstellung in Abbildung 3.8 genauer, so kann man auch daran zweifeln, ob das lineare Modell angemessen ist. So sind die Residuen für kleine x_i überwiegend negativ, danach für größere x_i überwiegend positiv (mit Ausnahme des Datums $(120;70)$, das aber als Sonderfall betrachtet werden könnte). Auch im Sachkontext erscheint eine verfeinerte Modellierung plausibel, da der Stadtverkehr (geringe Geschwindigkeit bei kürzeren Entfernungen) nicht unbedingt mit dem Überlandverkehr (höhere Geschwindigkeit bei längeren Entfernungen) zu einem linearen Modell passt. Wir nehmen die Diskussion anderer Modellklassen als die der linearen Funktionen im folgenden Kapitel 3.4 wieder auf.

3.3.2 Geradenanpassung mit der Median-Median-Geraden

Eine algorithmische Methode ist die Anpassung einer **Median-Median-Geraden**. Diese robuste Methode basiert auf der Clusterung des Merkmals X in drei Cluster und der Berechnung von Medianen in diesen Clustern zum Merkmal Y als Trägerpunkte einer Geraden.

Die Konstruktion der Median-Median-Geraden umfasst folgende Schritte:

1. Clusterung des Merkmals X in drei annähernd gleich große Cluster C_1, C_2 und C_3. Ist der Stichprobenumfang n durch 3 teilbar, sind die Cluster gleich groß, bleibt beim Teilen durch 3 ein Rest 1, so wird das mittlere Cluster um ein Datum größer, bei einem Rest 2 die beiden äußeren Cluster.[9]
2. Für die drei Cluster ergeben sich drei verschiedene Medianpunkte M_1, M_2 und M_3 der Merkmale X und Y innerhalb der Cluster C_1, C_2 und C_3:
   $C_1 : M_1\left(x_{0,5}(C_1); y_{0,5}(C_1)\right)$; $\quad C_2 : M_2\left(x_{0,5}(C_2); y_{0,5}(C_2)\right)$; $\quad C_3 : M_3\left(x_{0,5}(C_3); y_{0,5}(C_3)\right)$
3. Durch M_1 und M_3 legt man eine Hilfsgerade $g_h : y = a_h \cdot x + b_h$ mit

$$a_h = \frac{y_{0,5}(C_3) - y_{0,5}(C_1)}{x_{0,5}(C_3) - x_{0,5}(C_1)} \text{ und } b_h = y_{0,5}(C_1) - a_h \cdot x_{0,5}(C_1).$$

4. Man bestimmt das Residuum r_{M_2} von M_2 zu der Geraden g_h und verschiebt die Gerade g_h um 1/3 des Werts von r_{M_2} parallel und vertikal in Richtung von M_2.[10]
5. Es ergibt sich als Median-Median-Gerade $g : y = a \cdot x + b$ durch

$$a = a_h; \quad b = b_h + \frac{1}{3} r_{M_2} = b_h + \frac{1}{3}\left(y_{0,5}(C_2) - (a_h \cdot x_{0,5}(C_2) + b_h)\right).$$

Beispiel:
Gegeben sind die gemeinsame Verteilung zu den Merkmalen Entfernung (Y) und Zeit (X) und die Zuordnung der Daten zu einem von drei Clustern:

Zeit in min (X)	5	15	20	20	25	30	40	50	60	60	75	120
Entfernung in km (Y)	1,5	7	10	15	12	10	16	50	60	74	80	70
Cluster	C_1	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_3	C_3	C_3	C_3

[9] Sind durch diese Clusterung in zwei verschiedenen Clustern Daten mit derselben Merkmalsausprägung x_i, so werden die Daten mit kleineren Merkmalsausprägungen y_i dem jeweils linken Cluster zugeordnet. Sind in diesem Fall zusätzlich auch die Merkmalsausprägungen y_i identisch, so gehören identische Daten sowohl zu dem einen als auch dem anderen Cluster.

[10] Es werden also faktisch durch jeden der Punkte M_1, M_2 und M_3 Geraden mit der bereits bestimmten Steigung gelegt und als Achsenabschnitt das arithmetische Mittel der drei Achsenabschnitte der drei Hilfsgeraden (von denen zwei identisch sind) gebildet.

Es ergibt sich $M_1(17,5;8,5), M_2(35;14)$ und $M_3(67,5;72)$ und damit

$$a = \frac{72-8,5}{67,5-17,5} = 1,27; \quad b_h = 8,5 - 1,27 \cdot 17,5 = -13,725;$$
$$b = -13,725 + \frac{1}{3}(14 - (1,27 \cdot 35 - 13,725)) = -19,3$$

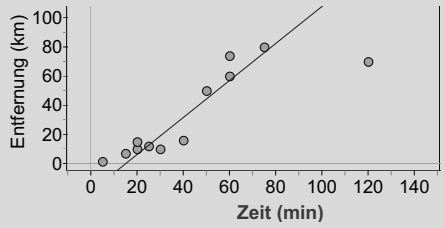


Abbildung 3.9: Geradenanpassung mit der Median-Median-Geraden

Man erkennt hier die Robustheit der Median-Median-Geraden, die nicht von dem Wert des Datums $(120;70)$ beeinflusst wird. Diese Gerade impliziert eine Durchschnittsgeschwindigkeit von 76,2 km/h bei einem Zeitverlust (z.B. durch den Ein- und Ausstieg) von etwa 15 Minuten. Während die Durchschnittsgeschwindigkeit relativ hoch zu sein scheint, könnte der Zeitverlust noch plausibel sein, wenn man zusätzlich den Weg zum Auto und die Parkplatzsuche in die Überlegungen mit einbezieht.

3.3.3 Regressionsgerade

Die Ausgleichsgerade, welche die Summe der quadrierten Residuen minimiert, ist die **Regressionsgerade**.[11] Es gilt dabei folgender Satz:

Satz 6
Es sei ein zweidimensionales Merkmal (X,Y) gegeben. Die Gerade $g: y = a \cdot x + b$ mit

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{s_X^2} \quad \text{und} \quad b = \overline{y} - a \cdot \overline{x},$$

wobei s_X^2 die Varianz (siehe Abschnitt 2.4.3, S. 33) des Merkmals X ist, minimiert die Summe der quadratischen Residuen.

Der Term

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

wird als **empirische Kovarianz** der metrisch skalierten Merkmale X und Y bezeichnet. Den Beweis dieses Satzes führen wir in Kapitel 3.6.

[11] Bei der Berechnung lässt sich – im Gegensatz zur Betrachtung der Residuenbeträge – das Kalkül der Differentialrechnung anwenden.

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y). Für die Berechnung der Parameter der Regressionsgeraden erhalten wir:

Zeit (X)	Entfernung (Y)	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
5	1,5	1469,4	1237,9
15	7	802,8	759,1
20	10	544,4	555,1
20	15	544,4	438,5
25	12	336,1	399,5
30	10	177,8	317,2
40	16	11,1	59,3
50	50	44,4	108,1
60	60	277,8	436,8
60	74	277,8	670,1
75	80	1002,8	1463,3
120	70	5877,8	2776,0
$\bar{x}=43,3$	$\bar{y}=33,8$	$s_X=947,2$	$s_{XY}=768,4$

Daraus ergibt sich

$$a=\frac{s_{XY}}{s_X^2}=\frac{768,4}{947,2}\approx 0,81; \quad b=\bar{y}-a\cdot\bar{x}=33,8-0,81\cdot 43,3\approx -1,4$$

sowie insgesamt als Regressionsgerade $g: y=0,81\cdot x-1,4$.

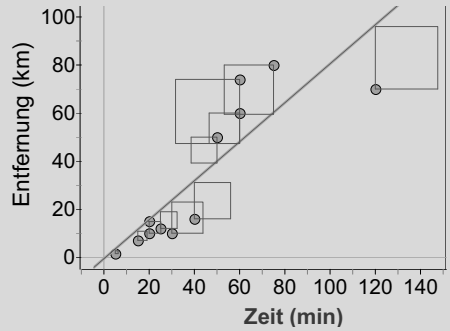


Abbildung 3.10: Regressionsgerade mit Abweichungsquadraten, deren Summe hinsichtlich dieser Geraden minimal ist

Es ergibt sich ähnlich wie bei der Einpassung einer Geraden nach Augenmaß eine Durchschnittsgeschwindigkeit von 49 km/h sowie ein Zeitverlust (durch z.B. Ein- und Ausparken) von etwa 2 Minuten.

Anhand der Definition und Abbildung 3.10 erkennt man, dass diese Methode nicht robust ist. Einerseits beeinflussen Ausreißer die Berechnung der Parameter der Regressionsgeraden stark, andererseits zieht jede Änderung eines Datums eine Parameteränderung nach sich.

3.3.4 Vergleich der Anpassungsgeraden

Wir beziehen den Vergleich auf das Beispiel des Zusammenhangs von Zeit (X) und Entfernung (Y), das wir bis hierher durchweg verwendet haben. In einem Schaubild ergibt sich der visuelle Vergleich der Anpassungsgeraden:

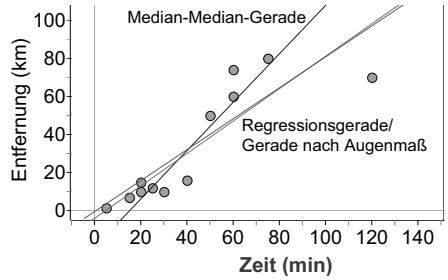


Abbildung 3.11: Anpassungsgerade (Augenmaß), Median-Median-Gerade, Regressionsgerade

Die Anpassung nach Augenmaß (mit Hilfe der Residuensumme) hat eine Gerade ergeben, die sehr ähnlich wie die Regressionsgerade liegt. Die robuste Median-Median-Gerade, die durch das Datum $(120;70)$ nicht beeinflusst wird, unterscheidet sich von diesen beiden Geraden deutlich. Der numerische Vergleich ergibt:

Typ	Gleichung	$\sum_{i=1}^{n} \lvert r_i \rvert$	$\sum_{i=1}^{n} (r_i)^2$
Gerade – Augenmaß	$y = 0,85x - 4$	138,5	2598
Median-Median-Gerade	$y = 1,27x - 19,3$	152,4	5007
Regressionsgerade	$y = 0,81x - 1,4$	141,6	2570

Hinsichtlich der Residuensumme ist die nach Augenmaß angepasste Gerade optimal, die Regressionsgerade hinsichtlich der Summe der quadratischen Residuen.[12] Die Median-Median-Gerade scheint hier die schlechteste Variante darzustellen. Andererseits repräsentiert sie den Datensatz ohne das Datum $(120;70)$ besser als die anderen beiden Geraden.[13] In dem Zusatzmaterial für dieses Buch haben wir eine Datei bereitgestellt, um anhand einer veränderbaren Punktwolke mit den drei verschiedenen Anpassungsgeraden experimentieren zu können.[14]

3.3.5 Korrelation

Wir haben verschiedene Möglichkeiten und Gütekriterien besprochen, um Geraden möglichst gut in eine Punktwolke einzupassen. Solche Kriterien geben aber keinen Hinweis darauf, ob es

[12] Es gibt offensichtlich verschiedene Kriterien, um zu definieren, was optimal ist.
[13] Ohne das Datum $(120;70)$ wäre die Regressionsgerade nahezu identisch mit der Median-Median-Geraden.
[14] Die Bezugsquelle ist im Vorwort dieses Buches angegeben.

überhaupt sinnvoll ist, eine Punktwolke durch eine Gerade zu ersetzen. Möglicherweise wird schon allein bei der Betrachtung einer Punktwolke deutlich, dass eine Geradenanpassung nicht geeignet ist, um die Punktwolke angemessen zu repräsentieren. Das lässt sich an den beiden Beispielen aus Abbildung 3.12 verdeutlichen, bei denen trotz offensichtlicher Unangemessenheit dennoch Geraden annähernd optimal eingepasst wurden.

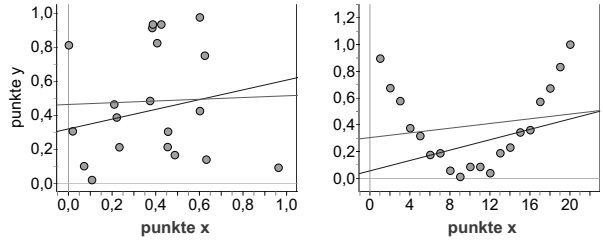


Abbildung 3.12: Anpassungsgeraden (Augenmaß) und Regressionsgeraden

Während in der linken Punktwolke offenbar kein Zusammenhang zwischen den beiden dargestellten Merkmalen besteht, ist bei der rechten Punktwolke zwar ein Zusammenhang zu erkennen, dieser ist aber nicht linear. Zusammenhänge wie in der rechten Punktwolke werden wir im Folgekapitel 3.4 noch ausführlicher behandeln. Zunächst führen wir ein Kriterium ein, das vor jeder Geradenanpassung die Güte eines linearen Zusammenhangs zweier Merkmale beschreibt. Dadurch kann die Suche nach einer möglichst optimalen Gerade motiviert werden.

Ein Maß für die Güte des linearen Zusammenhangs zwischen metrisch skalierten Merkmale besteht im **Korrelationskoeffizienten** r. Dieser Koeffizient einer zweidimensionalen Häufigkeitsverteilung soll folgende Eigenschaften haben:

1. Es gilt $-1 \leq r \leq 1$.
2. Ist $r = 0$, so besteht kein linearer Zusammenhang.
3. Liegen alle Punkte einer Punktwolke auf einer Geraden mit positiver Steigung, so ist $r = 1$ (perfekter positiver linearer Zusammenhang).
4. Liegen alle Punkte einer Punktwolke auf einer Geraden mit negativer Steigung, so ergibt sich $r = -1$ (perfekter negativer linearer Zusammenhang).

Hier sind nur drei Werte des Korrelationskoeffizienten angegeben worden, die allgemeiner Konsens sind. Inwieweit bei Zwischenwerten von einem mehr oder weniger starken oder schwachen linearen Zusammenhang gesprochen werden kann, ist vom Datenkontext, aber auch von der konkreten Konstruktion eines Korrelationskoeffizienten abhängig. In der Mathematik und den Naturwissenschaften (mit gut kontrollierbaren Variablen in einem Experiment) gelten Korrelationskoeffizienten ab 0,5 und ab 0,8 als Hinweise auf einen linearen bzw. starken linearen Zusammenhang, in den Sozialwissenschaften mit weniger gut kontrollierbaren sozialen Variablen (Merkmalen) gelten schon wesentlich kleinere Werte des Korrelationskoeffizienten als Hinweis auf einen (linearen) Zusammenhang.

Der eingeschränkte Wertebereich zwischen -1 und 1 ist nicht natürlich gegeben, sondern durch die Definitionen verschiedener Korrelationskoeffizienten konstruiert worden. Jede Neukonstruktion eines Korrelationskoeffizienten genügt daher nach Konvention diesem Wertebereich.

Wir werden im Folgenden vier verschiedene Varianten von Korrelationskoeffizienten für verschiedene Situationen einführen (drei in diesem Abschnitt, einen in Kap. 3.6). Alle Konstruktionen dieser Korrelationskoeffizienten basieren auf einer Transformation der Daten durch einen Lageparameter und in einem Fall auf der **Standardisierung** der Daten durch einen Lage- und einen Streuparameter. Daher werden wir vor der Diskussion der Korrelationskoeffizienten einen Abstecher in die Transformation bzw. Standardisierung von Daten machen.

3.3.5.1 Transformation und Standardisierung von Daten

Ziel einer ersten Transformation von Daten ist es, deren Lage in Beziehung zu einem der zentralen Lageparameter – dem arithmetischen Mittel oder dem Median – deutlich zu machen.

Das kann mittels einer algorithmischen Transformation der Merkmalsausprägungen eines metrisch skalierten Merkmals X durch

$$\tilde{x}_i = x_i - \bar{x} \quad \text{bzw.} \quad \tilde{x}_i = x_i - x_{0,5} \quad (i = 1, ..., s)$$

geschehen. Die so transformierten Daten geben die Differenz zu dem betrachteten Lageparameter an. Für diese Transformationen gilt:

$$\bar{\tilde{x}} = 0 \quad \text{bzw.} \quad \tilde{x}_{0,5} = 0$$

Die Begründung des ersten Teils der Behauptung haben wir bereits in Kapitel 2.7 gegeben, der zweite Teil der Behauptung ist unmittelbar aus der Definition des Medians einleuchtend.

Rein grafisch lässt sich diese Transformation in einer Punktwolke dadurch darstellen, dass ein Mittelkreuz eingezeichnet wird, dessen Lage durch die Mediane bzw. arithmetischen Mittelwerte der zugrunde liegenden Merkmale X und Y bestimmt wird. In Abbildung 3.13 ist der transformierte Datensatz $(\tilde{X}, \tilde{Y})$ zu den Merkmalen Entfernung und Zeit mit $\tilde{x}_i = x_i - \bar{x}$ und $\tilde{y}_i = y_i - \bar{y}$ links dargestellt. Auf der rechten Seite ist in die Punktwolke zu diesen beiden Merkmalen ein Mediankreuz eingezeichnet. Letzteres basiert auf einer Parallelen zur senkrechten Achse durch den Punkt $(x_{0,5}, 0)$ sowie einer Parallelen zur waagerechten Achse durch den Punkt $(0, y_{0,5})$.

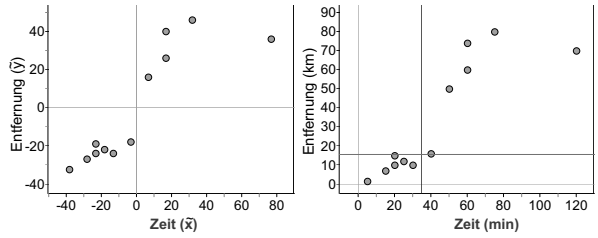


Abbildung 3.13: Transformierte und durch das Mediankreuz gruppierte Daten zu Zeit und Entfernung

Während man in der linken Abbildung unmittelbar die Differenzen zum jeweiligen arithmetischen Mittel ablesen kann, ist in der rechten Abbildung die Eingruppierung wie „größer als der Median $x_{0,5}$ und kleiner als der Median $y_{0,5}$“ möglich.

Hier setzt die prinzipielle Idee der Korrelationskoeffizienten an. So sollten die Punkte bei einer positiven linearen Abhängigkeit der beiden Merkmale zumindest überwiegend im 1. und 3. Quadranten des Koordinatensystems (bzw. des durch das Mediankreuz entstehenden Quadranten), also rechts oben und links unten liegen. Bei einer negativen linearen Abhängigkeit sollten die Punkte im 2. und 4. Quadranten (rechts unten und links oben) liegen und bei keiner linearen Abhängigkeit in allen vier Quadranten gleichmäßig verteilt sein. Wir nehmen das nach einem weiteren Zwischenschritt wieder auf.

Durch die Datentransformation schafft man willkürlich einen 0-Punkt, der ein Datenzentrum repräsentiert. Ein nächster Schritt kann darin bestehen, für beide Merkmale ein gemeinsames Maß zu schaffen, mit dem der Abstand zum 0-Punkt gemessen wird. Dieses Maß kann bei statistischen Daten in *Standardabweichungen* bestehen. Das heißt, die Entfernung eines Datums zum 0-Punkt wird nicht in den ursprünglichen Maßeinheiten wie Minuten (Zeit) und Kilometer (Entfernung) gemessen, sondern in einer Einheit, die durch die Datensätze selbst erzeugt wird. Um das zu ermöglichen, werden die Daten so transformiert, dass die Standardabweichung der Merkmale selbst 1 beträgt (und das arithmetische Mittel der transformierten Daten 0 ist). Dies gelingt durch folgende Transformation:

$$\hat{x} = \frac{x - \overline{x}}{s_X}$$

Den Beweis dafür, dass durch diese Transformation $s_{\hat{X}} = 1$ ist, werden wir in Kapitel 3.6 erbringen.

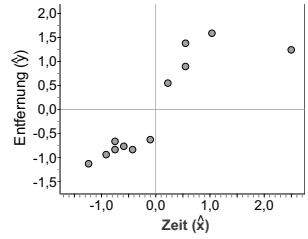


Abbildung 3.14: Standardisierte Daten zu Zeit und Entfernung

In der Abbildung 3.14 ist bezogen auf den 0-Punkt demnach eine Standardabweichung – das entspricht etwa $s_X \approx 26,4$ Minuten bzw. $s_Y \approx 15,5$ Kilometer – als einheitliches Maß zu verstehen.

Mit der Transformation und der Standardisierung der Daten werden wir im Folgenden die verschiedenen Korrelationskoeffizienten diskutieren.

3.3.5.2 Ausgezählter Korrelationskoeffizient

Für die Ermittlung des **ausgezählten Korrelationskoeffizienten** wird wie im vorausgegangenen Abschnitt 3.3.5.1 beschrieben ein Mittelkreuz in die Punktwolke gelegt. Die Ermittlung dieses Korrelationskoeffizienten basiert auf dem Auszählen von Punkten in folgenden zwei Schritten:

1. Man bestimmt die Anzahl n^+ der n Punkte in der Punktwolke des zweidimensionalen metrisch skalierten Merkmals (X,Y), die im 1. und 3. Quadranten des ermittelten Mittelkreuzes liegen. Liegen Punkte auf den Achsen des Koordinatensystems bzw. dem Mittelkreuz, so werden diese mit der Gewichtung $\frac{1}{2}$ zu n^+ gezählt.[15]
2. Man ermittelt n^- (das sind entsprechend die Punkte, die im zweiten und vierten Quadranten liegen) als Differenz von $n-n^+$ und bestimmt den ausgezählten Korrelationskoeffizienten r_z durch:

$$r_z = \frac{n^+ - n^-}{n} = \frac{n^+ - (n-n^+)}{n} = \frac{2n^+ - n}{n}$$

Da $0 \leq n^+ \leq n$ gilt, ergibt sich unmittelbar $-1 \leq r \leq 1$. Sind die Punkte gleichmäßig auf die Quadranten verteilt (also ohne linearen Zusammenhang), so ist $n^+ - n^- = 0$ und damit $r = 0$. Liegen alle Punkte auf einer Geraden mit positiver (negativer) Steigung, so ist – aufgrund möglicher auf den Achsen bzw. dem Mittelkreuz liegender Punkte – der Korrelationskoeffizient asymptotisch 1 bzw. -1 (für $n \to \infty$).

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y). In Abbildung 3.15 ist links die Transformation hinsichtlich des arithmetischen Mittels vorgenommen, in die Punktwolke rechts ist das Mediankreuz eingezeichnet.

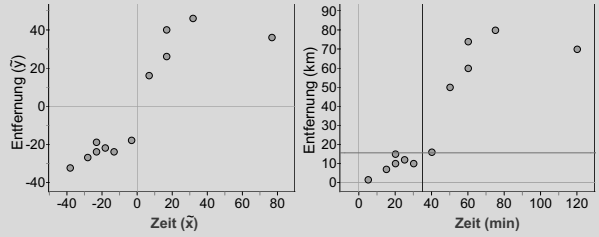


Abbildung 3.15: Transformierte und gruppierte Daten zu Zeit und Entfernung

Es ergibt sich, wenn man im Index noch als Bezugsgröße den entsprechenden Lageparameter aufführt:

$$r_{z,\bar{x}} = \frac{12-0}{12} = 1 \quad \text{und} \quad r_{z,x_{0,5}} = \frac{12-0}{12} = 1$$

[15] Es gilt $n^+ = |\{(x_i,y_i)|x_i < x_{0,5} \wedge y_i < y_{0,5} \vee x_i > x_{0,5} \wedge y_i > y_{0,5}\}| + \frac{1}{2}|\{(x_i,y_i)|x_i = x_{0,5} \vee y_i = y_{0,5}\}|$, wobei statt $x_{0,5}$ auch $\bar{x}$ gewählt werden kann. Dieser Korrelationskoeffizient ist in leicht abgewandelter Form auch Basis der sogenannten *Schnelltests* auf Unabhängigkeit von Blomqvist (vgl. Sachs, 1999).

In beiden Fällen ergibt sich ein starker positiver linearer Zusammenhang zwischen den Merkmalen Zeit und Entfernung. Außerdem erkennt man, dass der ausgezählte Korrelationskoeffizient r_z den Wert 1 annehmen kann, auch wenn die Punkte nicht auf einer Geraden liegen.

Der ausgezählte Korrelationskoeffizient ist robust hinsichtlich der Verwendung des Medians. Hinsichtlich des arithmetischen Mittels ist dieser Korrelationskoeffizient nicht robust.[16]

3.3.5.3 Resistenter Korrelationskoeffizient

Eine Variante des ausgezählten Korrelationskoeffizienten schlägt Polasek (1994) vor, den **resistenten Korrelationskoeffizienten** r_{rst}

$$r_{rst} = \sin\left[\left(\frac{n^+}{n} - \frac{1}{2}\right) \cdot \pi\right],$$

wobei n^+ hinsichtlich des Mediankreuzes wie im vorausgehenden Abschnitt 3.3.5.2 beschrieben gebildet wird. Die Analogie zum ausgezählten Korrelationskoeffizienten wird durch Umformen des Arguments des Sinus unmittelbar ersichtlich.

$$r_{rst} = \sin\left[\left(\frac{n^+}{n} - \frac{1}{2}\right) \cdot \pi\right] = \sin\left[\left(\frac{2n^+ - n}{2n}\right) \cdot \pi\right] = \sin\left[\left(\frac{n^+ - n^-}{n}\right) \cdot \frac{1}{2}\pi\right]$$

Das bedeutet, dass die Werte des resistenten Korrelationskoeffizienten durch eine Transformation der Werte des ausgezählten Korrelationskoeffizienten mittels des Sinus entstehen. Diese Transformation hat zur Folge, dass im Betrag leicht höhere Werte des Korrelationskoeffizienten auftreten (vgl. Abb. 3.16).

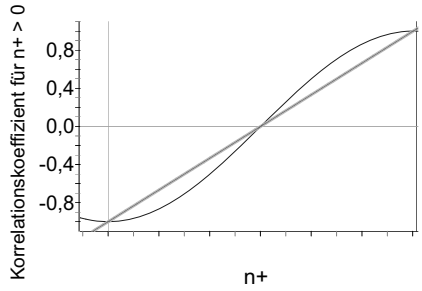


Abbildung 3.16: Unterschiede der Werte des ausgezählten und resistenten Korrelationskoeffizienten in Abhängigkeit von n^+

[16] Im Falle des Mittelkreuzes, das auf der Verwendung der arithmetischen Mittel basiert, kann die extreme Änderung eines Punktes aufgrund der Ausreißeranfälligkeit des arithmetischen Mittels dazu führen, dass das Mittelkreuz so stark verschoben wird, dass andere Datenpunkte in einem anderen, „gegensinnigen“ Quadranten zu liegen kommen, wodurch sich der ausgezählte Korrelationskoeffizient stark ändern würde.

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y). In Abbildung 3.15 ist rechts bereits ein Mediankreuz eingezeichnet. Da n^+ als 12 bestimmt war, ergibt sich mit dem resistenten Korrelationskoeffizienten ein Wert von 1:

$$r_{rst} = \sin\left[\left(\frac{n^+ - n^-}{n}\right) \cdot \frac{1}{2}\pi\right] = \sin\left[\frac{12}{12} \cdot \frac{1}{2}\pi\right] = 1$$

Auch dieser Korrelationskoeffizient ist wie der ausgezählte Korrelationskoeffizient hinsichtlich des Mediankreuzes robust.

3.3.5.4 Der Korrelationskoeffizient nach Bravais und Pearson

Der **Korrelationskoeffizient nach Bravais und Pearson** ist vielleicht als *der* Korrelationskoeffizient zu bezeichnen. Wird ein Korrelationskoeffizient ohne weitere präzisierende Bezeichnung genannt, so ist der von Bravais und Pearson gemeint. Er ist in jeglicher statistischer Software implementiert und wird von den Statistik anwendenden Wissenschaften nahezu standardmäßig benutzt, um die Güte linearer Zusammenhänge zu analysieren.

Dieser Korrelationskoeffizient basiert auf Überlegungen, die im Zusammenhang mit der Standardisierung zweier metrisch skalierter Merkmale X und Y stehen (vgl. Abschnitt 3.3.5.1). Misst man beide Merkmale mit dem gemeinsamen Maß der Standardabweichung[17], so sollte bei einem linearen Zusammenhang die Erhöhung einer Merkmalsausprägung x_i um t Standardabweichungen ($t \in \mathbb{R}$) die Erhöhung (bzw. Verringerung bei negativem linearen Zusammenhang) einer Merkmalsausprägung y_i um ebenfalls t Standardabweichungen bedingen. Das bedeutet, die standardisierten Daten müssten bei einem perfekten linearen Zusammenhang (was Büchter & Henn, 2005 sinnfällig als „optimalen linearen Gleichklang“ bezeichnen) auf einer Ursprungsgeraden mit Steigung 1 (bzw. -1) liegen (vgl. Abb. 3.17).[18]

Ausgehend von dem perfekten linearen „Gleichklang“, werden nun die Flächeninhalte betrachtet, die in Abbildung 3.17 durch die Punkte mit den Koordinaten $(0;0), (\hat{x}_i;0), (\hat{x}_i;\hat{y}_i)$ und $(0;\hat{y}_i)$ bezeichnet sind. Diese sind offenbar bei einem perfekten linearen „Gleichklang“ sämtlich Quadrate, im allgemeinen Fall sind es Rechtecke. Aus der mittleren Summe dieser Flächeninhalte konstruiert man den Korrelationskoeffizienten r. Dazu werden die Flächeninhalte in dem Sinne orientiert, dass Flächeninhalte von Rechtecken im 1. und 3. Quadranten positiv sind, Flächeninhalte von Rechtecken im 2. und 4. negativ.

Die mittlere Summe dieser gerichteten Flächeninhalte wird als Korrelationskoeffizient r bezeichnet:

$$r = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \cdot \hat{y}_i; \quad i = 1,...,n$$

[17] Dabei sind die Werte der Standardabweichungen beider Merkmale natürlich im Allgemeinen verschieden.

[18] Das ist unabhängig von der Steigung einer Geraden hinsichtlich eines perfekten Zusammenhangs zweier nicht standardisierter Merkmale. Man überlege sich das z.B. anhand der auf einer Geraden liegenden Punkte $P_1(-1;0)$ und $P_2(1,4)$, die standardisiert zu den Punkten $\hat{P}_1(-1;-1)$ und $\hat{P}_1(1;1)$ transformiert werden würden.

Diese mittlere Summe ist maximal 1, wenn die Merkmalsausprägungen $(\hat{x}_i;\hat{y}_i)$ auf einer Ursprungsgeraden mit der Steigung 1 liegen (und damit Quadrate bilden). Sie ist minimal, nämlich -1, wenn die Merkmalsausprägungen $(\hat{x}_i;\hat{y}_i)$ auf einer Ursprungsgeraden mit der Steigung -1 liegen.[19]

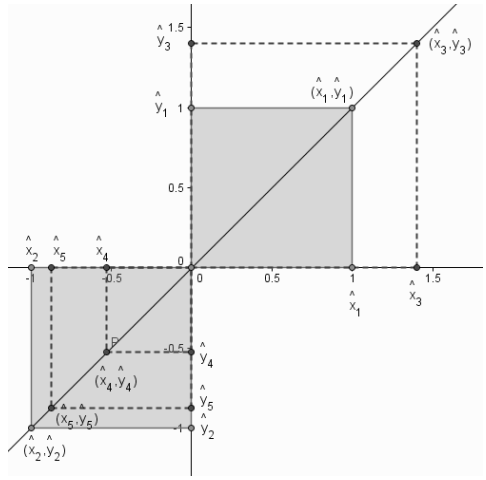


Abbildung 3.17: 5 standardisierte Daten mit perfektem linearen „Gleichklang“

Hebt man die Standardisierung der Merkmale wieder auf, so gewinnt man die übliche Darstellung der Formel des Korrelationskoeffizienten durch

$$r = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \cdot \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\overline{x}-x_i}{s_X} \cdot \frac{\overline{y}-y_i}{s_Y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\overline{x}-x_i)(\overline{y}-y_i)}{s_X s_Y} = \frac{s_{XY}}{s_X s_Y}; \quad i = 1,\ldots,n.$$

Dieser Korrelationskoeffizient ergibt sich also aus dem Verhältnis der empirischen Kovarianz (vgl. Kapitel 26) s_{XY} und dem Produkt $s_X \cdot s_Y$ der Standardabweichungen des zweidimensionalen Merkmals (X,Y). Das heißt, der Korrelationskoeffizient nach Bravais und Pearson r entspricht der durch die Standardabweichungen s_X und s_Y normierten Kovarianz s_{XY} der Merkmale X und Y (vgl. Kap. 3.3.5.1).

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y). In der folgenden Tabelle sind die Merkmalsausprägungen der Merkmale X und Y sowie der standardisierten Merkmale $\hat{X}$ und $\hat{Y}$ dargestellt.

[19] Eine Begründung für diese Behauptung bieten wir im Zusatzmaterial (die Bezugsquelle ist im Vorwort angegeben) zu diesem Buch.

Zeit (X)	Entfernung (Y)	Zeit ($\hat{X}$)	Entfernung ($\hat{Y}$)
5	1,5	-1,25	-1,12
15	7	-0,92	-0,93
20	10	-0,76	-0,82
20	15	-0,76	-0,65
25	12	-0,60	-0,75
30	10	-0,43	-0,82
40	16	-0,11	-0,62
50	50	0,22	0,56
60	60	0,54	0,91
60	74	0,54	1,39
75	80	1,03	1,60
120	70	2,49	1,25
$\overline{x} = 43,33$	$\overline{y} = 33,79$	$\overline{\hat{x}} = 0$	$\overline{\hat{y}} = 0$
$s_X = 30,78$	$s_Y = 28,94$	$s_{\hat{X}} = 1$	$s_{\hat{Y}} = 1$

In Abbildung 3.18 sind links die nicht standardisierten Daten eingezeichnet. Rechts sind die standardisierten Daten und die daraus hervorgehenden Rechtecke eingezeichnet.

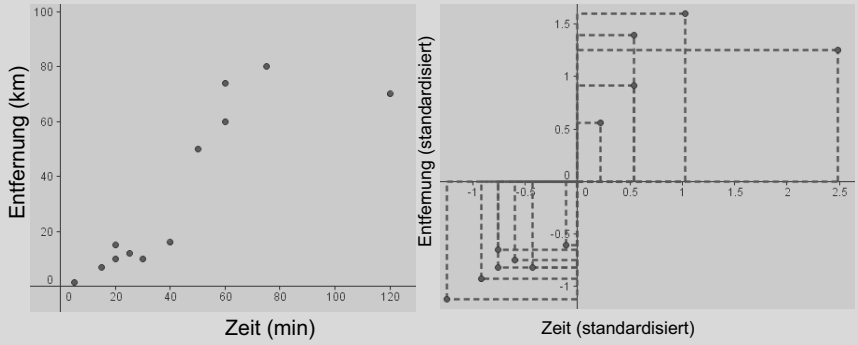


Abbildung 3.18: Linearer Zusammenhang der Merkmale Zeit und Entfernung

Es ergibt sich mit der Formel für diesen Korrelationskoeffizienten:

$$r = \frac{1}{12}\sum_{i=1}^{12} \hat{x}_i \cdot \hat{y}_i = \frac{\frac{1}{12}\sum_{i=1}^{12}(\overline{x} - x_i)(\overline{y} - y_i)}{s_X s_Y} = 0,86; \quad i = 1, ..., 12$$

In der Abbildung 3.18 zu den standardisierten Merkmalen mit den eingezeichneten Rechtecken erkennt man, dass die nahe am Ursprung liegenden Punkte wenig, die weit vom Ursprung entfernt liegenden Punkte stark die Summe der Flächeninhalte beeinflussen. Das bedeutet, dass insbesondere die entfernt liegenden Punkte den Korrelationskoeffizienten entscheidend beeinflussen. Ist solch ein Punkt wenig im linearen „Gleichklang“, so wird dieser Korrelationskoeffizient sehr stark beeinflusst, er zeigt sich wenig robust. Verändern wir den möglichen Ausreißerpunkt $(120;70)$ weiter, etwa zu $(140;50)$, so verringert sich der Korrelationskoeffizient auf

$r = 0,70$, also um fast 20%. Verändert man dagegen den Punkt $(50;50)$ auf $(70;30)$ (Erhöhung der x-Koordinate um 20, Erniedrigung der y-Koordinate um 20), so verringert sich der Korrelationskoeffizient lediglich auf $r = 0,83$, also um lediglich rund 4%.

Diese Eigenschaft des Korrelationskoeffizienten r macht ihn insbesondere dann anfällig, wenn die Datenpunkte nicht gleichmäßig um das Zentrum streuen. Das ist bei schiefen Verteilungen für X oder Y der Fall, wie es in unserem Beispiel die Verteilungen der Merkmale Entfernung und Zeit sind. Wird also der Zusammenhang schief verteilter Merkmale untersucht, so bieten sich eher die einfachen Korrelationskoeffizienten und der im Exkurs (Kap. 3.6) diskutierte an.[20] So sind die einfachen Korrelationskoeffizienten, wenn sie anhand des Mediankreuzes ermittelt werden, unabhängig von der Schiefe der Verteilung.

3.3.5.5 Vergleich der Korrelationskoeffizienten und Bezug zum Sachkontext

Wir beziehen den Vergleich auf das durchweg verwendete Beispiel des Zusammenhangs von Zeit (X) und Entfernung (Y). Im Überblick haben die Korrelationskoeffizienten folgende Werte ergeben:

Typ	Wert
r_z	1
r_{rst}	1
r	0,86

Es hat sich gezeigt, das r_z und r_{rst} im Gegensatz zu r robust sind und insbesondere bei schiefen Verteilungen der Merkmale sogar eine mögliche und elementare Alternative darstellen. In dem betrachteten Beispiel überschätzen sie aber anscheinend den linearen Zusammenhang der Merkmale. Im Zusatzmaterial für dieses Buch[21] bieten wir eine Rechneranwendung an, um für veränderbare Datensätze die Korrelationskoeffizienten explorativ untersuchen zu können.

Korrelationskoeffizient und Steigung der Regressionsgeraden Die Korrelationskoeffizienten ermöglichen eine gewisse Voraussage für eine Anpassungsgerade, die in die Punktwolke eines zweidimensionalen Merkmals (X,Y) eingepasst werden kann. So ist

- die Steigung dieser Geraden positiv, wenn der Wert des Korrelationskoeffizienten positiv ist, und
- die Steigung dieser Geraden negativ, wenn der Wert des Korrelationskoeffizienten negativ ist.

Der Wert des Korrelationskoeffizienten entspricht allerdings im Allgemeinen *nicht* der Steigung einer Anpassungsgeraden $y = ax + b$, d.h. $r \neq a$. Im Falle des Korrelationskoeffizienten von Bra-

[20] Die Voraussetzung für den Korrelationskoeffizienten r, dass nämlich das zugrunde liegende zweidimensionale Merkmal (X,Y) möglichst *binormal* verteilt sein sollte (und nicht schief), wird allerdings bei der zumeist mechanischen Anwendung des Korrelationskoeffizienten häufig nicht beachtet (vgl. auch Sachs, 1999).

[21] Die Bezugsquelle ist im Vorwort dieses Buches angegeben.

vais und Pearson r gibt es allerdings mit Blick auf den Satz 6, S. 63, den folgenden Zusammenhang:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(\overline{x}-x_i)(\overline{y}-y_i)}{s_X s_Y} = \frac{s_{XY}}{s_X s_Y}\cdot\frac{s_X}{s_X} = \frac{s_{XY}}{s_X^2}\cdot\frac{s_X}{s_Y} = a\cdot\frac{s_X}{s_Y};\quad i=1,...,n$$

Scheinkorrelationen Bei allen Beispielen zum Zusammenhang zweier Merkmale spielt der Sachkontext eine entscheidende Rolle. So kann zwar mathematisch ein linearer Zusammenhang mit dem Korrelationskoeffizient gemessen werden, ob allerdings tatsächlich ein linearer Zusammenhang im Sachkontext besteht, kann nur aus diesem heraus begründet werden. Während ein Zusammenhang zwischen Zeit und Entfernung tatsächlich plausibel ist („je größer die Entfernung, desto größer die Zeit“ bzw. „je größer die zur Verfügung stehende Zeit, desto größer die zurückgelegte Entfernung“), muss das bei der Analyse anderer zweidimensionaler Merkmale nicht der Fall sein, selbst wenn dort ein hoher Wert eines Korrelationskoeffizienten existiert. In diesem Zusammenhang wird das Beispiel der hohen Korrelation von Störchen und Geburten häufig zitiert, die im Sachkontext nicht interpretiert werden kann. Ebenso kann man im Datensatz zu den Studierenden auf die Suche nach sogenannten **Scheinkorrelationen** gehen. Faustregel könnte hier sein: Kombiniert man fortwährend numerische Merkmalsausprägungen, so wird man sicher irgendwann auf eine hohe Korrelation stoßen. Das zeigt das folgende Beispiel einer Scheinkorrelation, die hier mit einer konstruierten Auswahl von Studierenden erzeugt wurde:

> **Nachdenkliche Genies** Erhebungen unter Studierenden haben ergeben, dass im Abitur sehr gute Studierende viel länger für den Weg zur Hochschule brauchen als ihre weniger guten Kommilitonen. Möglicherweise sinnieren diese zukünftigen Genies bereits auf dem Weg zur Hochschule über komplexe Problemstellungen, während sich die anderen gedankenlos auf den Weg begeben ...

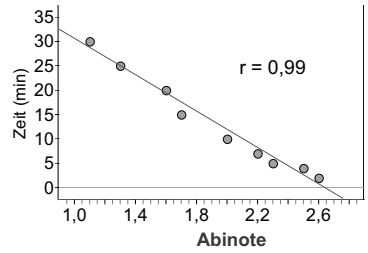


Abbildung 3.19: Scheinkorrelation von Abiturnote und Zeit

3.4 Anpassung von Funktionen in Punktwolken

Mit dem konstruierten Beispiel in Abbildung 3.12 (S. 66) ist bereits deutlich geworden, dass natürlich nicht für jedes zweidimensionale Merkmal (X,Y) ein *linearer* Zusammenhang bestehen

muss. Es kann durchaus kein Zusammenhang oder aber ein nicht-linearer Zusammenhang bestehen. Um Letztere wird es in diesem Kapitel in einem kurzen Überblick gehen. Wir werden dabei nur rein explorativ im Sinne des Einpassens nach „Augenmaß“ (vgl. Kap. 3.3) vorgehen und uns auf die Arbeit mit parametrischen Standardfunktionen beschränken.[22]

Der Vorteil der im vergangenen Kapitel in Punktwolken eingefügten linearen Funktionen ist sicher deren elementare Handhabbarkeit wie auch die in der Regel einfache Interpretation ihrer Parameter im Sachkontext. Daher kann man an komplexere Modellierungen mit nicht-linearen Funktionen folgende Anforderungen stellen (vgl. auch Stachowiak, 1973):

1. Die Verwendung einer nicht-linearen Funktion anstatt einer linearen sollte in der Modellierung einen Vorteil gegenüber der Betrachtung eines linearen Zusammenhangs aufweisen, und
2. die verwendete Funktionenklasse sollte hinsichtlich der verwendeten Parameter im Sachkontext interpretierbar sein.

Es sollten tatsächlich beide Kriterien erfüllt sein. Sind etwa n unterschiedliche Merkmalsausprägungen eines Merkmals X gegeben, so hat die Modellierung mit einem Polynom vom Grad $(n-1)$ die Eigenschaft, eine Residuensumme 0 , also eine bestmögliche Anpassung erzeugen zu können. Dagegen ist aber solch ein Modell in der Regel nicht interpretierbar. Verwendet man etwa das Beispiel der 12 Daten zu den Merkmalen Zeit und Entfernung und verwendet für die doppelten Werte 20 und 60 (Zeit) die Mittelwerte der jeweils zwei zugeordneten Merkmalsausprägungen für die Entfernung (12,5 und 67), so lässt sich ein Polynom vom Grad 9 ermitteln, das gegenüber allen anderen Funktionsmodellen die minimale Residuensumme (nämlich 19) aufweist (vgl. Abb. 3.20), aber im Sachkontext nicht interpretierbar ist.

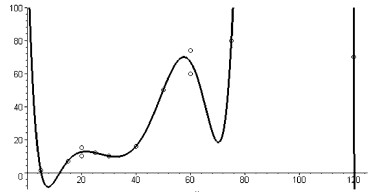


Abbildung 3.20: Bestanpassendes Polynom an die Daten zu *Entfernung* und *Zeit*

Wird bei einem zweidimensionalen Merkmal (X,Y) ein nicht-linearer Zusammenhang vermutet, so kann man folgendermaßen explorativ vorgehen:

1. Es wird nach Augenmaß eine angemessen erscheinende nicht-lineare Funktion f in eine Punktwolke eingepasst. Die Einpassung selbst besteht in der Wahl einer Funktionenklasse (etwa eine quadratische Funktion) und der nachfolgenden Justierung der in einer Funktionenklasse freien Parameter (etwa a,b und c bei einer quadratischen Funktion in der Scheitelform mit $f: y = a(x-b)^2 + c$).

[22] Deutlich vertieftere und breitere Darstellungen sind etwa in Hartung et al. (2009) oder Engel (2010) enthalten.

2. Bei der Anpassung wird versucht, Residuen ohne ein weiteres Muster zu erzeugen und die Summe der absoluten Residuen möglichst minimal zu halten.
3. Ist die eingepasste Funktion f zumindest im Bereich der Merkmalsausprägungen von X injektiv, so kann man das Merkmal Y durch die Umkehrfunktion f^{-1} transformieren und die Güte der Anpassung mit Hilfe der für lineare Funktionen zur Verfügung stehenden Methoden beschreiben (durch diese Transformation wird die eingepasste Funktion linearisiert). Die über die Linearisierung ermittelten Parameter sind anschließend in der einzupassenden Funktion entsprechend rücktransformiert zu berücksichtigen.

Beispiel:
Wir betrachten das zweidimensionale Merkmal (X,Y) zur Zeit und Entfernung, klammern aber zunächst das Datum $(120;70)$ aus den Betrachtungen aus und verwenden für das Merkmal Zeit die Angaben in Stunden. Passt man in die Punktwolke eine quadratische Funktion ein, so scheint diese gut zu dem Datensatz zu passen (vgl. Abb. 3.21) – besser zumindest als eine lineare Funktion, bei deren Anpassung im vorangegangenen Abschnitt allerdings kein Datum ausgeschlossen wurde.

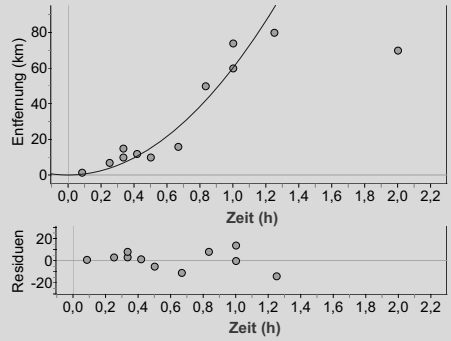


Abbildung 3.21: Quadratischer Zusammenhang der Merkmale Zeit und Entfernung mit Residuen

Die Interpretation der Funktionsgleichung $y = 60x^2$, die sich auf die Größenbeziehung km/h bezieht, könnte so lauten: Bei einer Stunde Fahrtzeit können rund 60 km zurückgelegt werden (was einer Faustformel für Autofahrten entspricht), bei längeren Fahrten wird die Zeit deutlich entfernungseffektiver genutzt, für geringere Entfernungen benötigt man sehr viel mehr Zeit. Im Residuendiagramm für die ersten 11 Daten können noch zwei Muster vermutet werden:
Zunächst ist es auffällig, dass die Residuen mit steigenden Merkmalsausprägungen x_i zunehmen. Dies ist naturgemäß zu erwarten. So ist im Sachkontext bei einer hohen Entfernung mit einer höheren Variabilität von Fahrtzeiten, umgekehrt dadurch auch bei hohen Zeiten mit einer größeren Variabilität der in dieser Zeit bewältigten Entfernungen zu rechnen als bei kleineren Zeiten. Allgemein wird man dieses Phänomen stets im Vergleich von niedrigen und im Vergleich dazu hohen Mittelwerten betrachten können, bei denen alltagssprachlich formuliert „mehr Platz zum abweichenden Streuen“ vorhanden ist (vgl. auch Kap. 2.7.2, S. 44).

Als zweites Muster könnte man in den sehr wenigen Daten eine Art Schwingung ausmachen. Da aber nicht ersichtlich ist, wie diese in Verbindung mit dem Sachkontext modelliert werden kann, wird hier nicht weiter versucht, dieses Muster zu erklären.
Als Funktion wurde $f : y = 60x^2$ in den Datensatz eingepasst. Die hier einfache Umkehrfunktion ist $f^{-1} : x = \sqrt{\frac{y}{60}}$. Wir transformieren mit Hilfe dieser Umkehrfunktion das Merkmal Y und erhalten:

X	0,08	0,25	0,33	0,33	0,42	0,50	0,67	0,83	1	1	1,25	(2)
Y	1,5	7	10	15	12	10	16	50	60	74	80	(70)
Z mit $z_i = \sqrt{\frac{y_i}{60}}$	0,16	0,34	0,41	0,50	0,45	0,41	0,52	0,91	1,00	1,11	1,15	(1,08)

Für den linearisierten Zusammenhang von X und Z ergibt sich

	mit $(x_{12};z_{12})$	ohne $(x_{12};z_{12})$
r_z	1	0,82
r_{rst}	1	0,96
r	0,87	0,96

Man erkennt, dass

- die Korrelationskoeffizienten weiterhin hoch bleiben, also auch die Einpassung der quadratischen Funktion sinnvoll scheint.
- insbesondere r mit und ohne Beachtung von $(x_{12};z_{12})$ höher ist als bei der vorausgegangenen Betrachtung ($r = 0,86$ mit Beachtung von $(x_{12};z_{12})$, vgl. Kap. 3.3.5.5, S. 74) als linearer Zusammenhang (ohne Beachtung von $(x_{12};z_{12})$ ergibt sich $r = 0,95$). Dieses Ergebnis scheint für eine quadratische Funktion als alternative anzupassende Funktion zu sprechen.
- die Korrelationskoeffizienten r_z und r_{rst} im Fall der Nichtbeachtung von $(x_{12};z_{12})$ durch die dann notwendigerweise auf dem Mediankreuz liegenden Punkte, deren eine Koordinate aus einem der Mediane von X und Z besteht ($n = 11$), hier niedriger sind. Diese auf dem Mediankreuz liegenden Punkte beeinflussen bei den hier wenigen Daten insbesondere den Wert von r_z.

3.5 Eigenschaften von Studierenden

In einer kurzen Übersicht wollen wir die in diesem Kapitel diskutierten statistischen Methoden nutzen, um Zusammenhänge zwischen je zwei Merkmalen der Studierenden, teilweise im Vergleich der Datensätze zur PH Freiburg und zur Universität Münster, analysieren.

Zusammenhang nominalskalierter Merkmale Wir beginnen mit dem Rauchverhalten und untersuchen, ob zwischen zwei Hochschulen (Münster und Freiburg) ein Unterschied besteht. In der Vierfeldertafel sind die zugehörigen relativen und absoluten (in Klammern) Häufigkeiten der analysierten 396 Studierenden aus Freiburg und 1080 Studierenden aus Münster angegeben.[23]

[23] Die Diskrepanz in den Einträgen der ersten Zeile ergeben sich aufgrund von Rundungen.

	M	F	Summe
R	0,18(270)	0,08(124)	0,27(394)
$\overline{R}$	0,55(810)	0,18(272)	0,73(1082)
Summe	0,73(1080)	0,27(396)	1(1476)

Es ergibt sich:

- $h_n(R|M) = \frac{270}{1080} \approx 0,25$; $h_n(\overline{R}|M) = 1 - h_n(R|M) = 0,75$
- $h_n(R|F) = \frac{124}{396} \approx 0,31$; $h_n(\overline{R}|F) = 1 - h_n(R|F) \approx 0,69$

Damit ergeben sich die grafische Vierfeldertafel und das zugehörige Einheitsquadrat, die in Abbildung zu sehen sind.

	M	F	Summe
R	0,18 (270)	0,08 (124)	0,27 (394)
$\overline{R}$	0,55 (810)	0,18 (272)	0,73 (1082)
Summe	0,73 (1080)	0,27 (396)	1 (1476)

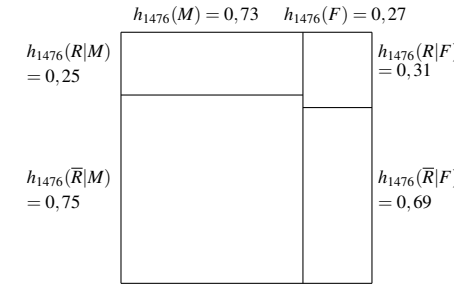


Abbildung 3.22: Grafische Vierfeldertafel und Einheitsquadrat mit den Bezeichnungen R: Raucher, $\overline{R}$: Nichtraucher, M: Münster und F: Freiburg

Der Unterschied ist absolut gesehen offenbar gering. Das einfache Assoziationsmaß A hat den Wert -0,06, ist also nahe 0. Die odds ratio q ergibt den Wert

$$\rho = \frac{270 \cdot 272}{810 \cdot 124} \approx 0,73; \qquad \frac{1}{\rho} \approx 1,37$$

und ist damit noch recht nahe an 1.

Beide Assoziationsmaße weisen also auf keinen bzw. einen höchstens schwachen Zusammenhang der beiden Merkmale hin. Unklar muss hier allerdings bleiben, inwiefern ein sehr geringer Wert beider Assoziationsmaße im Zusammenhang mit der Größe der Stichprobe zu bewerten ist. Wir werden das in Kapitel 8 wieder aufnehmen und dort Überlegungen anstellen, welche statistische Aussagekraft den hier ermittelten Assoziationsmaßen beigemessen werden kann.

Wir betrachten ein weiteres Beispiel, den Zusammenhang des Geschlechts der Studierenden mit ihrer Präferenz für die Parteien CDU (C) und Grüne (G). Dabei wird der Datensatz auf diejenigen Studierenden eingeschränkt ($n = 405$), die eine der beiden Parteien präferieren.

	M	W	Summe
C	0,21(84)	0,25(100)	0,46(184)
G	0,19(78)	0,35(143)	0,54(221)
Summe	0,40(162)	0,60(243)	1(405)

Es ergibt sich:

- $h_n(C|M) = \frac{84}{162} \approx 0,52$; $h_n(G|M) = 1 - h_n(G|M) \approx 0,48$
- $h_n(C|W) = \frac{100}{243} \approx 0,41$; $h_n(C|W) = 1 - h_n(G|W) = 0,59$

Damit ergeben sich die grafische Vierfeldertafel und das zugehörige Einheitsquadrat in Abbildung 3.23:

	M	W	Summe
C	0,21 (84)	0,25 (100)	0,46 (184)
G	0,19 (78)	0,35 (143)	0,54 (221)
Summe	0,40 (162)	0,60 (243)	1 (405)

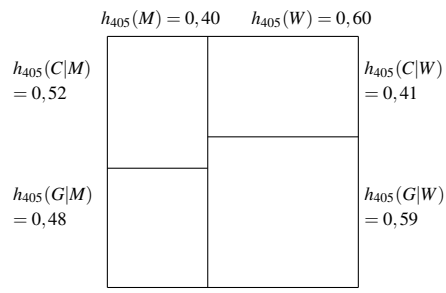


Abbildung 3.23: Grafische Vierfeldertafel und Einheitsquadrat mit den Bezeichnungen M: männlich, W: weiblich, C: CDU und G: Grüne

Der Unterschied ist absolut gesehen höher als im vorangegangenen Beispiel. Das einfache Assoziationsmaß A hat den Wert:

$$A = h_{405}(C|M) - h_{405}(C|W) \approx 0,52 - 0,41 = 0,11$$

Für die odds ratio q ergibt sich:

$$\rho = \frac{84 \cdot 143}{78 \cdot 100} \approx 1,54; \qquad \frac{1}{\rho} \approx 0,65$$

Beide Assoziationsmaße weisen also auf einen Zusammenhang der beiden Merkmale hin. Wiederum bleibt unklar, inwiefern der Wert beider Assoziationsmaße im Zusammenhang mit der Größe der Stichprobe zu bewerten ist. Wie bereits oben erwähnt, werden wir das in Kapitel 8 wieder aufnehmen. Man könnte allerdings schon hier vorsichtig betrachtet von einer höheren Präferenz von Studentinnen zur Partei der Grünen im Vergleich zu den Studenten ausgehen.

Zusammenhang nominal- und metrisch skalierter Merkmale Wir haben bereits am Ende von Kapitel 2.6 implizit nach solchen Zusammenhängen zwischen verschiedenen Merkmalen und dem Merkmal Hochschule gesucht. Hier betrachten wir nur die Frage nach einem Zusammenhang, der häufiger diskutiert wird, nämlich dem Unterschied zwischen der schulischen Leistung von Männern und Frauen anhand der Abiturnote (vgl. Abb. 3.24).

Tatsächlich sind die weiblichen Studierenden hier den männlichen Studierenden durchschnittlich um 0,1 Zensurenpunkte überlegen (deren Aussage zu hinterfragen ist, vgl. Kap. 1.1.3), der

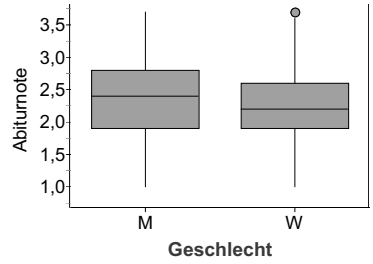


Abbildung 3.24: Zusammenhang von Geschlecht und Abiturnote

Unterschied ist allerdings gering. Auch die Unterschiede im Median und im 3. Quartil (jeweils 0,2 Zensurenpunkte) sind höchstens schwache Hinweise auf einen tatsächlichen Unterschied der Abiturleistungen der Geschlechter unter den erhobenen Studierenden.[24]

Zusammenhang metrisch skalierter Merkmale Wir betrachten zunächst die Frage, ob solche Studierende, die stark Sport treiben, weniger Zeit für andere Freizeitbeschäftigungen wie etwa das Musizieren haben. Vorab könnte man hier einen negativen linearen Zusammenhang vermuten: „je mehr Sport, desto weniger Musik“. Dieser wie auch andere vermutete Zusammenhänge lassen sich aber in den Freizeitbeschäftigungen der Studierenden (auch bei verbundenen Merkmalen) nicht nachweisen.

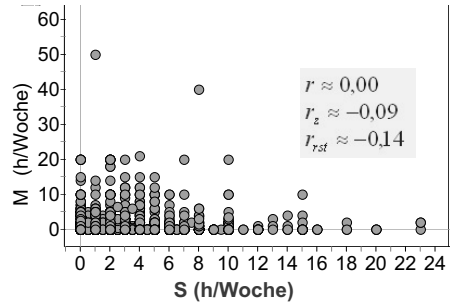


Abbildung 3.25: Zusammenhang von Sport (S) und Musik (M) (wöchentliche Zeit in h)

Man erkennt bei keinem der Korrelationskoeffizienten einen Wert, der auf einen Zusammenhang hinweist. In der Punktwolke mag zwar eine geringe negative lineare Abhängigkeit erkennbar sein (vgl. Abb. 3.25). Bei dem Umfang des Datensatzes kann aber der äußere Eindruck nicht mehr leitend sein, da viele der 1478 Datenpunkte übereinander liegen. Welche das aber sind, lässt sich nur anhand der Rohdaten erkennen. Entsprechend wurden die Korrelationskoeffizienten r, r_z und r_{rst} anhand der Rohdaten ermittelt.

Warum aber gibt es den vermuteten Zusammenhang nicht? Eine erste Vermutung könnte sein, dass die vielen Studierenden, die keinen Sport treiben, das vermutete Ergebnis nicht sichtbar

[24] Diese Aussage formulieren wir nur mit Blick auf die vorliegenden, als gegeben betrachteten Daten.

werden lassen. Clustert man allerdings die Studierenden in solche, die keinen Sport treiben bzw. solche, die Sport treiben, erkennt man ebenso die offensichtliche Unabhängigkeit der beiden Merkmale (vgl. Abb. 3.26).

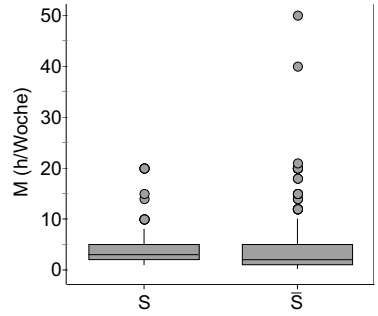


Abbildung 3.26: Zusammenhang von Sport (S) und Musik (M) (wöchentliche Zeit in h)

Eine mögliche Erklärung, dass zwischen den Merkmalen zur Freizeitbeschäftigung kein Zusammenhang zu finden ist, lässt sich durch die Transformation beider metrisch skalierter Merkmale in nominalskalierte finden (Sport in h wird in „S“ und „$\overline{S}$“ unterteilt ebenso wie Musik in h in „M“ und „$\overline{M}$“). In der Vierfeldertafel und deren Visualisierung im Punktgruppendiagramm (Abb. 3.27) erkennt man, dass von beiden Freizeitbeschäftigungen Sport offensichtlich die beliebtere ist. Selbst wenn man die Betrachtung auf eines der vier Cluster beschränkt, ergibt sich allerdings kein linearer (oder allgemein funktionaler Zusammenhang) der beiden Eigenschaften von Studierenden. Würde man das Einheitsquadrat zu diesen beiden Merkmalen zeichnen, wäre ein minimaler Wert des Assoziationsmaßes von $A \approx 0,01$ erkennbar.

	S	$\overline{S}$	Summe
M	409	91	500
$\overline{M}$	777	164	941
Summe	1186	255	1441

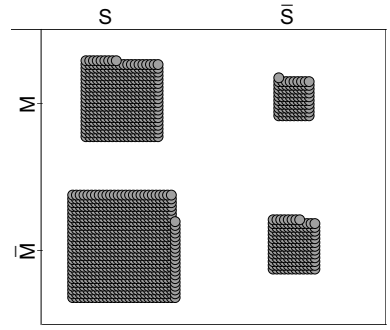


Abbildung 3.27: Zusammenhang von Sport ($S/\overline{S}$) und Musik ($M/\overline{M}$)

Einen sehr stark ausgeprägten, erwartbaren positiven linearen Zusammenhang ergibt dagegen die Analyse von Zeit und Entfernung bezogen auf die Radfahrer (Abb. 3.28). Im Gegensatz zu den im Hauptteil des Kapitels betrachteten Autofahrern ist hier ein anderer funktionaler Zusammenhang als der lineare im Sachkontext nicht plausibel, da mit dem Fahrrad bei größeren

Strecken keine höheren Geschwindigkeiten erreicht werden als bei kleineren.

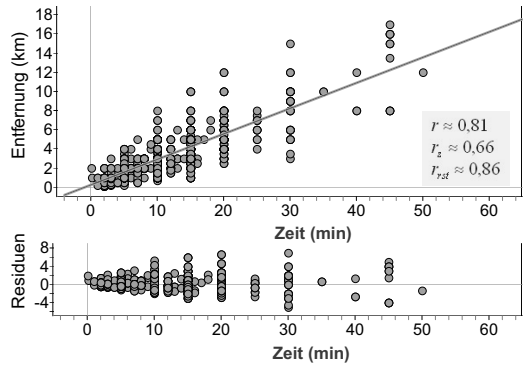


Abbildung 3.28: Zusammenhang von Zeit in Minuten und Entfernung in Kilometern (Radfahrer)

Die Korrelation ist hoch. Die Geradengleichung (Regressionsgerade) lässt sich als mittlere Geschwindigkeit von rund $0,266 \cdot 60 \approx 16$ km/h interpretieren. Das scheint plausibel zu sein. Würde man die Median-Median-Gerade einfügen, so ergäbe sich ein etwas geringerer Wert für diese Durchschnittsgeschwindigkeit.

Bei der Komplexität des Datensatzes bietet sich ein algorithmisches Verfahren (Regressionsgerade, Median-Median-Gerade) an, da nicht übersehen werden kann, welche der insgesamt vorhandenen 870 Punkte übereinander liegen. Dies müsste aber bei dem Anpassen einer Geraden nach Augenmaß beachtet werden. Lässt man sich bei der computergestützten Anpassung einer Geraden nach Augenmaß allerdings (algorithmisch) von der berechneten Summe der absoluten Residuen leiten, so ergibt sich wieder eine den anderen Anpassungen sehr ähnliche Gerade.

3.6 Ergänzungen

Im Verlauf von Kapitel 3 haben wir an einigen Stellen auf dieses ergänzende Kapitel verwiesen. Wir werden teilweise sehr gestrafft und teilweise mit erhöhtem rechnerischen Aufwand auf die Verweise eingehen.

3.6.1 Methode der kleinsten Quadrate

In Kapitel 26 haben wir behauptet, dass die Gerade $g : ax+b$ mit den Parametern

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{s_x^2} \quad \text{und} \quad b = \overline{y} - a\overline{x},$$

wobei s_x^2 die Varianz des Merkmals X ist, die Summe der quadratischen Residuen minimiert. Ein Residuum r_i für das in die Punktwolke eingepasste Modell der Geraden $g : y = ax+b$ ist durch $r_i = y_i - y_f(x_i) = y_i - (ax_i+b) = y_i - ax_i - b$ gegeben.

Für den Beweis betrachten wir die quadratischen Residuen als Funktion zweier Veränderlicher a und b ($f(a,b)$) und hier speziell zunächst als Kurvenschar in Abhängigkeit von b mit dem Parameter a. Man bestimme zuerst die Extremwerte der Funktion bei festem Parameter a:

$$\begin{aligned} f_a(b) = &\quad \textstyle\sum_{i=1}^{n}(y_i - ax_i - b)^2 && \text{Summe der quadratischen Residuen} \\ f_a'(b) = &\quad -2\textstyle\sum_{i=1}^{n}(y_i - ax_i - b) && \text{Kettenregel} \end{aligned}$$

Man bestimmt die Nullstellen der Ableitung nach b:

$$\begin{aligned} f_a'(b_0) &= -2 \cdot \textstyle\sum_{i=1}^{n}(y_i - ax_i - b_0) &&= 0 \\ \textstyle\sum_{i=1}^{n} y_i - a \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b_0 &&= 0 \\ n\overline{y} - na\overline{x} - nb_0 &&= 0 \\ b_0 &&= \overline{y} - a \cdot \overline{x} \end{aligned}$$

Die zweite Ableitung von f nach b ist eine positive Konstante ($f_a''(b) = 2n$). Das bedeutet, dass alle Extrempunkte der Parabelschar Minima sind. Man setzt nun die Extremstelle b_0 in die Funktion $f_{b_0}(a)$ ein

$$\begin{aligned} f_{b_0}(a) &= \textstyle\sum_{i=1}^{n}(y_i - ax_i - b_0)^2 \\ &= \textstyle\sum_{i=1}^{n}\left(y_i - ax_i - (\overline{y} - a\overline{x})\right)^2 \\ &= \textstyle\sum_{i=1}^{n}(y_i - a(x_i - \overline{x}) - \overline{y})^2, \end{aligned}$$

bestimmt wiederum mit Hilfe der Kettenregel die Ableitung dieser Funktion

$$f_{b_0}'(a) = -2 \cdot \sum_{i=1}^{n}(y_i - a(x_i - \overline{x}) - \overline{y}) \cdot (x_i - \overline{x})$$

und berechnet durch Ausmultiplizieren und wieder zusammenfassendes Umformen die Nullstellen der Ableitungsfunktion

$$\begin{aligned} f_{b_0}'(a_0) &= 0 \\ -2\left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i\overline{y} - \sum_{i=1}^{n} y_i\overline{x} + \sum_{i=1}^{n}\overline{x}\overline{y} - a_0\sum_{i=1}^{n}(x_i - \overline{x})^2\right) &= 0 \\ \sum_{i=1}^{n} x_i(y_i - \overline{y}) - \sum_{i=1}^{n}\overline{x}(y_i - \overline{y}) - a_0 \cdot n \cdot s_X^2 &= 0 \\ \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) - a_0 \cdot n \cdot s_X^2 &= 0 \\ a_0 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n \cdot s_X^2} \\ a_0 &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X^2} \end{aligned}$$

Damit ist der Satz bewiesen.

3.6.2 Standardisierung

Im Zusammenhang mit der Standardisierung eines Merkmals X durch die Abbildung von x_i auf $\hat{x}_i = \frac{x_i - \bar{x}}{s_X}$ hatten wir behauptet, dass $\bar{\hat{x}} = 0$ und $s_{\hat{X}} = 1$ gilt. Wir werden dies in zwei Schritten beweisen. Es gilt unter Verwendung des Hilfssatzes aus Kapitel 2.7.2, S. 43:

$$\bar{\hat{x}} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \bar{x}}{s_X} = \frac{1}{s_X}\cdot\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})}_{=0} = 0$$

Weiterhin gilt:

$$s_{\hat{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - \bar{\hat{x}})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X} - 0\right)^2 = \frac{1}{s_X^2}\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{s_X^2}{s_X^2} = 1$$

3.6.3 Bestimmtheitsmaß

In Kapitel 3.3 hatten wir als Maß für die Güte eine Anpassungsmodells zu einem zweidimensionalen Merkmal (X,Y) etwa die Summe der absoluten bzw. der quadratischen Residuen betrachtet. Wir wollen diesen Ansatz zur Analyse der Güte eines Modells vertiefen: Betrachtet man eine spezielle Lage von Datum, Fit und Residuum zu einem zentralen Lageparameter des Merkmals Y (Abb. 3.29, hier zum Lageparameter $\bar{y}$), so sollte bei einem guten Modell die mittlere Summe der absoluten bzw. quadratischen Residuen geringer sein als ein korrespondierendes Streumaß zum Merkmal Y, also die mittlere absolute Abweichung $d_{x_{0,5}} = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{0,5}|$ hinsichtlich des Medians oder die Varianz $s_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ hinsichtlich des arithmetischen Mittels von Y.

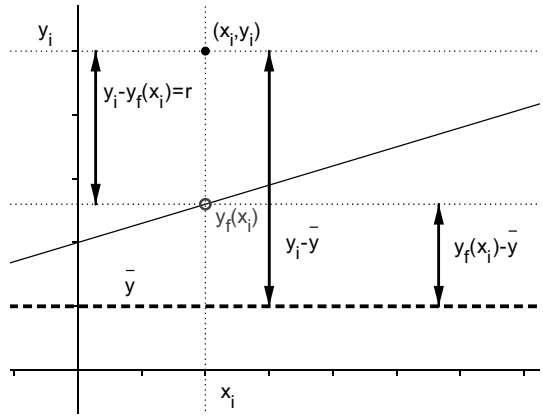


Abbildung 3.29: Verhältnis von Datum, Fit und Residuum

Wir betrachten daher folgende Verhältnisse hinsichtlich von Residuen und Streumaßen als Maß für die Güte eines Modells zu einem zweidimensionalen Merkmal (X,Y):

$$V_{y_{0,5}} := \frac{\frac{1}{n}\sum_{i=1}^{n}|(y_i - y_f(x_i))|}{\frac{1}{n}\sum_{i=1}^{n}|(y_i - \bar{y})|} \quad \text{bzw.} \quad V_{\bar{y}} := \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_f(x_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In beiden Fällen betrachten wir das Maß V als den Anteil der mittleren Summe der absoluten bzw. quadratischen Residuen an der Streuung von Y, gemessen als mittlere absolute Abweichung bzw. als Varianz. Bei einem gut passenden Modell ist dieser Anteil, also der Wert des Maßes, nahe 0, im Idealfall eines genau passenden Modells sogar genau 0.[25]

Beispiel:
Gegeben sind die Daten zur Zeit (X) und der Entfernung (Y) von Studierenden zur Hochschule sowie die drei Anpassungsgeraden, die wir in Kapitel 3.3 betrachtet haben. Wir erhalten zu diesen drei linearen Modellen folgende Werte $V_{y_{0,5}}$ bzw. $V_{\overline{y}}$:

Modell		$V_{y_{0.5}}$	$V_{\overline{y}}$
Gerade nach Augenmaß	$g_f : y = 0,85x - 4$	0,47	0,26
Median-Median	$g_f : y = 1,27x - 19,3$	0,52	0,49
Regression	$g_f : y = 0,81x - 1,36$	0,48	0,26

Jedes der Modelle (Geraden) verringert demnach bezogen auf die absoluten Residuen etwa 50% der mittleren absoluten Abweichung des Merkmals Y. Hinsichtlich der quadratischen Residuen verringern die Anpassungsgerade nach Augenmaß sowie die Regressionsgerade die Varianz von Y um etwa 75% und sind der Median-Median-Geraden überlegen, die die Varianz um etwa 50% reduziert. Rundet man den Wert von $V_{\overline{y}}$ nicht, so erkennt man, dass hier insgesamt die Regressionsgerade optimal ist.

Mit der Gleichung *Daten = Fit + Residuen* (vgl. Kap. 3.3.1, S. 60) gilt im Fall der linearen Anpassung durch die Regressionsgerade zusätzlich:

$$\begin{array}{lcll} \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 & = & \frac{1}{n}\sum_{i=1}^{n}(y_f(x_i)-\overline{y})^2 & + \quad \frac{1}{n}\sum_{i=1}^{n}(y_i-y_f(x_i))^2 \\ \text{Varianz } s_Y^2 \text{ von } Y & = & \text{erklärte Varianz von } Y & + \quad \text{unerklärte Varianz von } Y \end{array}$$

Diese in Abbildung 3.29 nicht enthaltene Beziehung ergibt sich durch[26]:

$$\begin{aligned} s_Y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_f(x_i)-\overline{y})^2+\frac{1}{n}\sum_{i=1}^{n}(y_i-y_f(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{s_{XY}}{s_X^2}x_i+\overline{y}-\frac{s_{XY}}{s_X^2}\overline{x}-\overline{y}\right)^2+\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\frac{s_{XY}}{s_X^2}x_i-\overline{y}+\frac{s_{XY}}{s_X^2}\overline{x}\right)^2 \\ &= \left(\frac{s_{XY}}{s_X^2}\right)^2\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{1}{n}\sum_{i=1}^{n}\left((y_i-\overline{y})-\frac{s_{XY}}{s_X^2}(x_i-\overline{x})\right)^2 = \left(\frac{s_{XY}}{s_X^2}\right)^2\cdot s_X^2+s_Y^2-2\frac{s_{XY}}{s_X^2}\cdot s_{XY}+\left(\frac{s_{XY}}{s_X^2}\right)^2 s_X^2 = s_Y^2 \end{aligned}$$

Wird diese Gleichung, mit der sich im Falle der Regression die Varianz s_Y^2 in einen erklärten und einen unerklärten Teil zerlegt lässt, durch die Varianz s_Y^2 dividiert, so erhält man:

$$1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_f(x_i)-\overline{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2} + \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_f(x_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

[25] Bei einem wenig passenden Modell kann der Wert V auch größer 1 sein. Ebenso kann für einzelne Indizes i ein Quotient innerhalb der Summe größer 1 sein, auch wenn V einen Wert kleiner 1 ergibt. V ist allein ein Maß für die mathematische Güte eines Anpassungsmodells, nicht aber für dessen Interpretierbarkeit (vgl. S. 76)

[26] Bei der Umformung wird außer den Varianzen s_X^2 und s_Y^2 auch die Kovarianz s_{XY} (vgl. Kap. , S. 63) verwendet.

Bezogen auf die Regressionsgerade wird $B := 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_f(x_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}$ als **Bestimmtheitsmaß** bezeichnet. Das Bestimmtheitsmaß B ist in gängiger Statistiksoftware implementiert. Im Fall der Regression ergibt sich analog zu den Umformungen oben die Beziehung:

$$B = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_f(x_i)-\bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\left(\frac{s_{XY}}{s_X^2}\right)^2 s_X^2}{s_Y^2} = \frac{s_{XY}^2}{s_X^2 \cdot s_Y^2} = r^2$$

3.6.4 Korrelationskoeffizient nach Spearman

Eine Alternative zum Korrelationskoeffizienten r, bei dem die Symmetrie der Verteilungen der Merkmale X und Y keine Voraussetzung ist, ist der **Spearmansche Rangkorrelationskoeffizient** r_S. Dieser basiert auf folgender Überlegung: Liegt ein perfekter positiver linearer „Gleichklang" vor, so müsste, wenn man die Merkmalsausprägungen der Größe nach sortiert und diesen Rangplätze von 1 bis n zuordnet, die Rangnummer der Merkmalsausprägung x_i mit der Rangnummer der Merkmalsausprägung y_i identisch sein.[27] Die Konstruktion des Korrelationskoeffizienten r_S basiert dabei auf der des Korrelationskoeffizienten r in der folgenden Umformung[28]:

$$\begin{aligned} r &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\sum_{i=1}^{n}x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[\sum_{i=1}^{n}(x_i)^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2\right]\left[\sum_{i=1}^{n}(y_i)^2 - \frac{1}{n}(\sum_{i=1}^{n}y_i)^2\right]}} \end{aligned}$$

Setzt man in diese Darstellung statt der Messwerte x_i bzw. y_i die Ränge dieser Messwerte, $R(x_i)$ und $R(y_i)$, ein, so erhält man[29]:

$$r_S = \frac{\sum_{i=1}^{n}(R(x_i)R(y_i)) - \frac{1}{n}\left(\frac{n(n+1)}{2}\right)^2}{\left(\frac{n(n+1)(2n+1)}{6} - \frac{1}{n}\cdot\frac{n(n+1)}{2}\right)^2}$$

Diese Darstellung lässt sich mit Hilfe der Rangdifferenzen d_i weiter vereinfachen. So ist[30]

$$\sum_{i=1}^{n}(Rg(x_i)Rg(y_i)) = \sum_{i=1}^{n} i^2 - \frac{1}{2}\sum_{i=1}^{n} d_i^2 = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^{n} d_i^2.$$

[27] Bei perfektem negativen linearen „Gleichklang" müsste der Rang von x_1 zum Rang von y_n, der Rang von x_2 zum Rang von y_{n-1} identisch sein, usw.

[28] Umformung im Zähler: $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum(x_i y_i) - n\overline{xy} = \sum(x_i y_i) - \frac{1}{n}\sum x_i \sum y_i$, wobei die Summen jeweils von 1 bis n laufen.
Umformung Nenner: $\sum(x_i-\bar{x})^2 = \sum(x_i^2) - 2\sum(x_i\bar{x}) + \sum(\bar{x}^2) = \sum(x_i) - 2n\bar{x}^2 + n\bar{x}^2 = \sum(x_i^2) - \sum(\bar{x}^2) = \sum(x_i^2) - \frac{1}{n}(\sum(x_i))^2$.

[29] Es gilt $\sum i = \frac{n(n+1)}{2}$ bzw. $\sum i^2 = \frac{n(n+1)(2n+1)}{6}$, wenn man die Summen von 1 bis n laufen lässt (Beweis z.B. durch vollständige Induktion).

[30] Auflösen und Umstellen der Gleichung $\sum d_i^2 = \sum(Rg(y_i) - Rg(x_i))^2 = 2\sum i^2 - 2\sum(Rg(x_i)Rg(y_i))$.

Dadurch ergibt sich schließlich, wenn man die Bruchterme vereinfacht:

$$r_S = \frac{\frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^{n} d_i^2 - \frac{1}{n}\left(\frac{n(n+1)}{2}\right)^2}{\frac{n(n+1)(2n+1)}{6} - \frac{1}{n}\left(\frac{n(n+1)}{2}\right)^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2+1)}$$

Beispiel:
Gegeben ist die gemeinsame Verteilung zu den Merkmalen Zeit (X) und Entfernung (Y). Wir erhalten die für die Berechnung des Rangkorrelationskoeffizienten r_S notwendigen Werte in folgender Tabelle:

(x_i)	$R(x_i)$	(y_i)	$R(y_i)$	$d_i^2 = (R(x_i) - R(y_i))^2$
5	1	1,5	1	0
15	2	7	2	0
20	3,5	10	3,50	0
20	3,5	15	6	6,25
25	5	12	5	0
30	6	10	3,50	6,25
40	7	16	7	0
50	8	50	8	0
60	9,5	60	9	0,25
60	9,5	74	11	2,25
75	11	80	12	1
120	12	70	10	4
Summe	78	—	78	20

Damit ergibt sich

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \approx 0,93.$$

Der Korrelationskoeffizient weist damit (wie alle bisher betrachteten auch) auf eine hohe positive lineare Abhängigkeit der Merkmale X und Y hin.

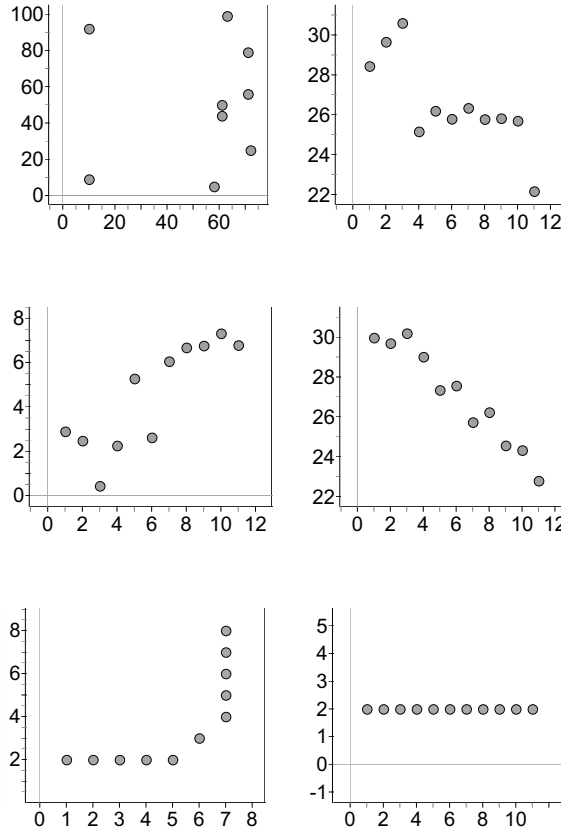
Im Gegensatz zum Korrelationskoeffizienten r ist r_S unabhängig von der Verteilungsform der Merkmale X und Y. Zu r hatten wir in Kapitel 3.3 auf die Beeinflussung des Wertes von r durch Ausreißer hingewiesen. Wir zeigen an der Veränderung von (x_{12}, y_{12}), dass dies für r_S nicht in gleichem Maße gilt, und „verbessern“ einerseits und „verschlechtern“ andererseits diesen Punkt im Sinne eines linearen Gleichklangs.

	$(x_{12}; y_{12})$	r	r_S
„Verbesserung“	(70; 70)	0,95	0,94
tatsächlich	(120; 70)	0,86	0,93
„Verschlechterung“	(180; 70)	0,73	0,93

Man erkennt die Anfälligkeit von r gegenüber den extremen Daten, die den Wert von r_S dagegen nahezu unverändert lassen. Da solche „extremen“ Werte insbesondere bei stark schief verteilten Merkmalen normal sind, ist r_S bei schiefen Verteilungen eine bessere, zumindest robustere Wahl gegenüber r.

3.7 Aufgaben

Aufgabe 3.1: Schätzen Sie den Wert eines Korrelationskoeffizienten für folgende Punktwolken und begründen Sie Ihre Schätzung:



Aufgabe 3.2: Untersuchen Sie mit Rechnerunterstützung Datensätze, die wir in den Formaten xls (Excel), ftm (fathom) und txt (Textdatei zum Einlesen in potentielle Statistiksoftware) im Zusatzmaterial für dieses Buch zur Verfügung stellen[a], auf einen funktionalen Zusammenhang. Die Datensätze betreffen:

- die Eigenschaften von Studierenden,
- die deutsche Fußball-Bundesliga und
- den weltweiten CO_2-Anstieg.

[a] Die Bezugsquelle ist im Vorwort angegeben.

Aufgabe 3.3: Gegeben sind die Datensätze zu Studierenden (Münster und Freiburg) zu der Parteipräferenz, dem bevorzugten Beförderungsmittel sowie zum Erhalt von BAföG.
Sind Grünen-Wähler umweltbewusster? Sind BAföG-Bezieher SPD-Wähler?

	Grüne	CDU/FDP	Summe
Auto	13	9	22
Fahrrad	116	132	248
Summe	129	141	270

	BAföG	kein BAföG	Summe
SPD	49	107	156
CDU/FDP	66	163	229
Summe	115	270	385

Aufgabe 3.4: Untersuchen Sie durch Clusterung und möglichst variantenreich mit einem linearen Modell ohne Rechnerunterstützung den Zusammenhang zwischen erster und zweiter Sprungweite beim Skispringen. Versuchen Sie von dem hier gegebenen Springen in Innsbruck 2009 das Modell auf die im Netz verfügbaren Springen in Innsbruck anderer Jahre sowie anderer Orte zu übertragen.

Platz	Name	Weite1	Weite2	Platz	Name	Weite1	Weite2
1	Schmitt	128,5	125,5	16	Evensen	119	119
2	Loitzl	126,5	128,5	17	Watase	118	122
3	Schlierenzauer	126	127,5	18	Eggenhofer	117,5	118
4	Amman	125,5	123,5	19	Larinto	117	120,5
5	Morgenstern	124,5	125	20	Uhrmann	116,5	125,5
6	Kasai	124	126	21	Hilde	116,5	122,5
7	Neumayer	124	126	22	Ito	116,5	121,5
8	Hautamaeki	123,5	128	23	Schoft	116,5	122
9	Rosliakow	123,5	121,5	24	Stoch	116	119,5
10	Olli	122	125	25	Hocke	115,5	124
11	Koch	122	122,5	26	Koudelka	115,5	123,5
12	Vassilev	121,5	129	27	Lackner	115,5	121,5
13	Jacobsen	121,5	126,5	28	Yumoto	115	123
14	Malysz	120,5	121,5	29	Kofler	114,5	119,5
15	Kuettel	119,5	124,5	30	Tochimoto	112,5	122

4 Datenanalyse: Rückschau

Fasst man die statistischen Tätigkeiten in den vergangenen Kapiteln zusammen, so kann man zu folgender Begriffsfestlegung für die Datenanalyse gelangen:

> Die Datenanalyse umfasst eine Sammlung von Methoden zur Erhebung, Aufbereitung und Interpretation statistischer Daten, die aufgrund von erkenntnisleitenden Fragen erhoben werden, die aus der unmittelbaren oder mittelbaren Erfahrung der natürlichen, technischen oder sozialen Umwelt resultieren.

In diesem Rückblick sollen die Methoden zur Aufbereitung ein- und zweidimensionaler Datensätze noch einmal insgesamt vernetzt und mit der Erhebung von Daten sowie der Interpretation der Ergebnisse der statistischen Methoden in Beziehung gesetzt werden.

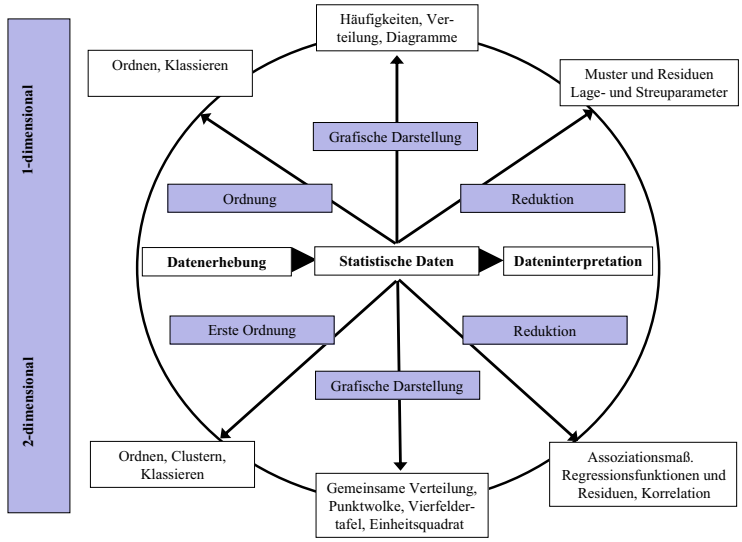


Abbildung 4.1: Struktur zur Datenanalyse

Die Erhebung und die Interpretation rahmen die Methoden zur Aufbereitung ein (vgl. Abb. 4.1). So ist das Ziel der Datenanalyse, eine statistisch fundierte Interpretation zu einem Ausschnitt der Realität, wie z.B. den Eigenschaften von Studierenden, zu erhalten. Dazu müssen zunächst als erste und unverzichtbare Voraussetzung möglichst gute Daten gewonnen werden (vgl. Kap. 1). Auf dieser Basis können nach einem ersten Ordnen die Daten in verschiedenen Repräsentationen aufbereitet oder statistische Kennwerte ermittelt werden, um Muster zu identifizieren bzw. zu beschreiben. Dabei ist die Erkenntnis wichtig, dass Daten die Muster, die vom

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_4,

Datenanalytiker „hineingelesen“ werden, nicht in Reinform enthalten, sondern stets Abweichungen zu erwarten sind. Die auf Muster und Residuen reduzierten Daten sind in der Regel die Basis für eine Dateninterpretation, bei der allerdings nicht allein die Daten selbst, sondern stets auch der Sachkontext, aus dem die Daten stammen, von Bedeutung ist (Wild & Pfannkuch, 1999).

Die Methoden als Herzstück in dem Prozess Datenerhebung – Datenauswertung – Dateninterpretation und die mit den Methoden verbundenen Tätigkeiten lassen sich weiter aufteilen und einheitlich für die Analyse ein- und mehrdimensionaler Daten (hier zweidimensionaler Daten) darstellen (vgl. Abb. 4.1).

Die Methoden für die Aufbereitung ein- wie auch zweidimensionaler Daten sind zu einem größeren Teil vergleichbar. Dies wird im folgenden tabellarischen Überblick zu den in den vergangenen Kapiteln vorgestellten Methoden deutlich:

Eindimensionale Daten	Zweidimensionale Daten	Tätigkeit
Daten	Datenpaare	Ordnen
Häufigkeitsverteilung	gemeinsame Häufigkeitsverteilung	Ordnen
Punktdiagramm (und alle anderen Diagramme)	Punktwolken (oder auch gruppierte Punktdiagramme)	grafische Darstellung
Boxplot	multipler Boxplot	grafische Darstellung
Median	Median-Median-Gerade	Reduktion
Arithmetisches Mittel	Regressionsgerade	Reduktion
Quartilsabstand	—	Reduktion
Residuensumme	Residuensumme	Reduktion
Varianz, Standardabweichung	Bestimmtheitsmaß, Korrelationskoeffizient	Reduktion

Zu manchen der eingeführten Methoden, mit denen zweidimensionale Daten aufbereitet werden, lässt sich kein Pendant zur Analyse univariater Daten finden, wie z.B. beim Assoziationsmaß. Umgekehrt verhält es sich entsprechend: So haben wir kein zweidimensionales Gegenstück zu dem eindimensionalen Streuparameter des Quartilsabstands angegeben.

An vielen Stellen haben wir in den vergangenen Kapiteln geäußert, einen beschreibend ermittelten Unterschied zwischen zwei Datensätzen könne man hinsichtlich seiner Relevanz noch nicht beurteilen. Genau hier liegt die Grenze der auf Beschreibung angelegten Datenanalyse:

- Ist z.B. eine Häufigkeit einer Merkmalsausprägung, ein Lage- oder auch Streuparameter einer Häufigkeitsverteilung ermittelt, so gilt dieser zwar hinsichtlich der erhobenen Stichprobe. Welchen Aussagewert er über die Stichprobe selbst hinaus hat, kann beschreibend jedoch nicht interpretiert werden.
- Besteht zwischen zwei verschiedenen Stichproben ein Unterschied in einer Maßzahl, so ist dieser zwar beschreibbar, ob er aber auch bei einer späteren Erhebung feststellbar sein wird, kann mit den Methoden der (beschreibenden) Datenanalyse nicht geklärt werden.

Es ist die *Variabilität* oder *Non-Uniformität* statistischer Daten (Wild & Pfannkuch, 1999), die es verbietet, Ergebnisse der beschreibenden Analyse eines Datensatzes ohne weitere Überlegun-

gen zu verallgemeinern. Jede Erhebung von Daten und ebenso jede Messung wird zu verschiedenen Zeitpunkten oder in verschiedenen Stichproben unterschiedliche Ergebnisse erzeugen. Dieses „anders sein“ bedeutet jedoch nicht völlige Beliebigkeit. Dann wäre es tatsächlich sinnlos, überhaupt Datenanalyse zu betreiben. Es ist aber nicht beliebig: Insbesondere viele Daten zu einem Vorgang, Objekt oder Phänomen zeigen oft ein bestimmtes Muster, etwas Charakteristisches, das nicht von einer speziellen Stichprobe abhängt, sondern sich in allen Stichproben (wenn sie gut gezogen wurden) zwar nicht exakt gleich, aber doch ähnlich zeigt. Betrachtet man etwa in dem bestehenden Datensatz zu den über 1000 Studierenden der Uni Münster in willkürlich ausgewählten Teilstichproben mit jeweils 100 Studierenden die Merkmale Alter und Entfernung, so erkennt man Muster (vgl. Abb. 4.2).

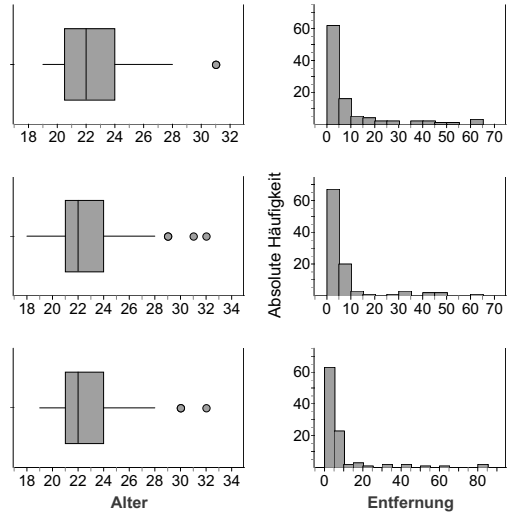


Abbildung 4.2: Drei Stichproben mit $n = 100$ zu den Eigenschaften der Studierenden der Uni Münster

Jede Teilstichprobe ist zwar anders (Variabilität), in allen Teilstichproben zeigt sich aber z.B. das Muster der nahezu identischen Quartile zum Merkmal Alter oder die immer linkssteile Häufigkeitsverteilung des Merkmals Entfernung. Es sind solche *Muster*, die in der Variabilität der statistischen Daten zu finden sind, welche die Beschäftigung mit der Datenanalyse lohnenswert erscheinen lassen: Es ist keine von vornherein sinnlose Suche, sondern im Gegenteil eine Gewinn versprechende.

Was kann alles ein näher zu untersuchendes Muster sein? Eine erste Antwort ist hier schlicht: jede beschreibbare Eigenschaft eines Datensatzes. Dazu gehören beispielsweise:

- Die Häufigkeit einer Merkmalsausprägung und die Häufigkeitsverteilung insgesamt,
- Maßzahlen wie Lage- oder Streuparameter, die eine Häufigkeitsverteilung reduzieren, oder auch
- die Form einer Häufigkeitsverteilung, die etwa durch die Schiefe gemessen werden kann.

In diesen Beispielen wurden Muster eines eindimensionalen Datensatzes genannt. Dies lässt sich ohne Weiteres auf zweidimensionale Datensätze ausweiten, also etwa auf

- bedingte (eingeschränkte) Häufigkeiten,
- Geraden- bzw. Funktionsanpassungen in eine Punktwolke oder
- Korrelationskoeffizienten.

Betrachtet man noch allgemeiner zweidimensionale Datensätze, so können auch Unterschiede zu einem Merkmal zwischen zwei verschiedenen Stichproben (bzw. Teilstichproben) ein Muster sein, z.B

- der Unterschied eines Mittelwerts in zwei Stichproben,
- der Unterschied hinsichtlich der Homogenität (Streuung) eines Datensatzes in zwei Stichproben oder auch
- Unterschiede in bedingte Häufigkeiten, Funktionsanpassungen, Korrelationen in zwei zweidimensionalen Stichproben.

Alle diese Muster können nach der beschreibenden Analyse eines Datensatzes als Modell für einen Ausschnitt aus der Realität betrachtet werden. In solchen Modellen ist es möglich, zunächst losgelöst von der Realität zu arbeiten. Das werden wir in den folgenden Kapiteln machen, indem wir Modelle mit Elementen der Wahrscheinlichkeitsanalyse bearbeiten. Dabei werden folgende Fragestellungen die Analyse von Daten vertiefen:

- Wenn ein aus der Datenanalyse stammendes Modell als sinnvoll anzunehmen ist, mit welchen Eigenschaften zukünftig erhobener Datensätze könnte man rechnen?
- Wenn man von einem theoretisch erdachten Modell ausgehen würde, inwieweit könnten empirisch ermittelte Daten das theoretische Modell stützen oder widerlegen?

Um solche Fragen beantworten zu können, muss man zunächst mit den theoretischen Modellen der Wahrscheinlichkeitsanalyse umgehen können. Da wir dabei primär auf einer theoretischen Ebene arbeiten werden, ist der Bezug zum Datensatz zu den Eigenschaften von Studierenden nicht immer unmittelbar möglich. Wir werden aber im Laufe der folgenden Kapitel sukzessive von der theoretischen Ebene zurück zu den Fragen an den realen Datensatz zurückkehren. Um die wahrscheinlichkeitstheoretischen Überlegungen nicht zusätzlich durch komplexe Modellannahmen zu erschweren, werden wir bei diesem Vorgehen einfache Situationen wie den Würfel- oder Münzwurf nutzen und damit die Komplexität des Datensatzes zu den Studierenden ersetzen, ohne auf den letzten ganz zu verzichten.

5 Elementare Wahrscheinlichkeitsanalyse

Einstiegsbeispiel

Abbildung 5.1: Studierende im Hörsaal

Aufgabe 1: Schätzen Sie die Häufigkeiten zu Merkmalsausprägungen verschiedener Merkmale in der Grundgesamtheit der Studierenden.

Worum es geht

Diese Einstiegsaufgabe ist auf der Basis der bisherigen Darlegungen bereits mit einfachem Nachdenken schnell gelöst. Denn was gibt es schon für andere Möglichkeiten, als eine relative Häufigkeit, die in einer Stichprobe zu einer bestimmten Merkmalsausprägung ermittelt wurde, zunächst einmal auch als relative Häufigkeit in der Grundgesamtheit anzunehmen (selbst wenn man um die Variabilität statistischer Daten weiß). Die absolute Häufigkeit einer Merkmalsausprägung wird dabei proportional auf die Grundgesamtheit übertragen.

Intuitiv wird man die relative Häufigkeit zu einer Merkmalsausprägung in einer größeren Stichprobe als sicherer einschätzen gegenüber der Häufigkeit in einer kleineren Stichprobe. Ist etwa die relative Häufigkeit der Singles (S) der ersten befragten 20 Studierenden der Universität Münster 0,65 ($h_{20}(S) = 0,65$) und diese relative Häufigkeit bezogen auf den gesamten Datensatz ($n = 1038$) dagegen 0,47 ($h_{1038}(S) \approx 0,47$), so wird man die zuletzt genannte relative Häufigkeit als Schätzung für den tatsächlichen Anteil in der Grundgesamtheit wählen. Wir werden uns in diesem Kapitel damit beschäftigen, warum eine relative Häufigkeit in einer Stichprobe umso besser als Schätzung der relativen Häufigkeit einer Grundgesamtheit verwendet werden kann, je größer (vorausgesetzt, die Daten sind gut) die Stichprobe ist. Auf dem Weg zu einer Begründung werden wir folgende Grundlagen der Analyse von Wahrscheinlichkeiten diskutieren:

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_5,

Zufall als Basis, Wahrscheinlichkeiten zu analysieren Was ist eigentlich an den Daten zu den Eigenschaften der Studierenden zufällig? Wird etwa ein Student nach seinem Alter, Rauchverhalten oder seiner Parteipräferenz befragt, so steht diese ja fest. Zufällig an der Antwort des Studenten ist allerdings, wie es zu dieser aktuell festliegenden Antwort gekommen ist, und gleichfalls zufällig ist im Optimalfall einer repräsentativen Erhebung (vgl. Kap. 1.1), dass gerade dieser eine Student in die Erhebung eingeschlossen wird. In diesem Kapitel geben wir einen kurzen Überblick, was allgemein mit dem Begriff Zufall verbunden wird bzw. werden kann.

Theoretische Grundgesamtheiten Im vergangenen Kapitel hatten wir den Übergang von der Empirie in die Welt der Modelle postuliert. In der *Modell-Welt* verschiebt sich im Gegensatz zur stets endlichen Empirie der *realen Welt* der Begriff der Grundgesamtheit. In der realen Welt können etwa nur *endlich* viele Studierende zu ihrem Beziehungsverhalten befragt werden. In der Modell-Welt können wir uns eine zweiseitige Münze vorstellen, die theoretisch unendlich oft geworfen werden kann und *zufällig* „Single“ oder „Nicht-Single“ anzeigt. Solche einer Münze wohnt als Grundgesamtheit eine bestimmte Wahrscheinlichkeit inne, eine ihrer beiden Seiten zu zeigen (nicht notwendigerweise 50%). Die Beziehungen zwischen der realen Welt und der Modell-Welt werden wir in diesem Kapitel immer wieder betrachten.

Wahrscheinlichkeiten Wie man von theoretischen Modellen, bei denen unendlich viele statistische Einheiten erzeugt werden können, zu einer Wahrscheinlichkeitsaussage (also zu den modellhaften theoretischen Häufigkeiten) kommt, werden wir in drei Varianten in diesem Kapitel klären, um den Zufall damit *berechenbar* zu machen. Dabei wird ein Modell allein klären, was eine Wahrscheinlichkeit ist. Die anderen beiden bieten einerseits vom theoretischen, andererseits vom empirischen Standpunkt einen Zugang zum Wahrscheinlichkeitsbegriff.

Realität oder Theorie Da wir uns in der Modell-Welt bewegen werden, werden wir zum größeren Teil einfache Situationen zur Illustration verwenden und statt den Studierenden klassische Zufallsgeneratoren wie Münze oder Würfel betrachten. Das hat den immensen Vorteil, dass hier kaum diskutiert werden muss, warum ein Modell die real existierenden Objekte wie Würfel und Münze angemessen repräsentieren kann. In Bezug auf die Studierenden müsste hier viel umfassender argumentiert werden. Dennoch bleiben wir dabei, stets auch das Studierendenbeispiel zu diskutieren und in Kapitel 5.4 die Erkenntnisse dieses Kapitels auf den Datensatz der Studierenden anzuwenden.

5.1 Grundbegriffe

Wenn Wahrscheinlichkeiten analysiert werden sollen, ist es in formaler Hinsicht hilfreich, wenn zuvor einige Grundlagen, der Zufallsbegriff und bestimmte Sprachkonventionen geklärt worden sind, so dass auf diese zurückgegriffen werden kann.

5.1.1 Zufall

Über den Begriff des Zufalls könnte man ein eigenes, philosophisch ausgerichtetes Buch schreiben, so viel ist über den Zufall nachgedacht und geschrieben worden. Prinzipiell gibt es zwei Extrempositionen:

> Die eine Seite sagt, „Alles ist Zufall“. Quasi-determistische Vorgänge spiegeln bloß verschiedene zufällige Vorgänge wider, die sich neutralisieren. Die andere Seite sagt „Gott würfelt nicht“ (z.B. Albert Einstein). Alles ist deterministisch, der Zufall beruht allein auf unserer mangelnden Kenntnis der Vorgänge (vgl. Schneider, 1988).

Beispiel:
Betrachten wir das Werfen eines Würfels, so könnte man davon ausgehen, dass der Wurf streng deterministisch physikalischen Gesetzen gehorcht. Das würde bedeuten: Wenn sich das eigentlich vorhandene Wissen vollständig nutzen ließe und dieses Wissen angewandt wird, so könnte der nächste Wurf eines beliebigen Spielers *sicher* vorhergesagt werden. Umgekehrt könnte man dann Schritt für Schritt von dem Endzustand des Wurfs den Wurfprozess zurückverfolgen (Reversibilität). Gäbe es dagegen auf der sichtbaren oder unsichtbaren Ebene eines physikalischen Vorgangs, wie dem Wurf eines Würfels, Prozesse, die nicht reversibel sind (das ist die Annahme der „Zufallsanhänger“), so wäre die Voraussage eines Wurfes selbst bei umfassendem Wissen nicht sicher möglich.

Der Gedanke lässt sich auf die Eigenschaften von Studierenden übertragen: Gäbe es ein Überwesen, dass alle menschlichen Geschicke leitet, so wäre es z.B. zu jedem Zeitpunkt festgelegt, ob jemand auf die Frage nach der Beziehungsgewohnheit mit „Single“ oder „Nicht-Single“ antwortet. Wäre aber wiederum auch nur ein kleiner Bestandteil der unüberschaubar vielen Faktoren, die zu solch einer Antwort führen, nicht gesteuert und auch nicht rückschauend zurückverfolgbar (reversibel), so wäre die Antwort nicht sicher bestimmbar.

Wir lassen den Grundlagenstreit beiseite und gehen ganz pragmatisch vor, indem wir festlegen:

> Alle Vorgänge, die sich aus pragmatischen Gründen sinnvoll als zufällige Vorgänge behandeln lassen, behandeln wir auch als solche. Sie haben die Eigenschaft, dass mehrere Ergebnisse, die allesamt bekannt sind, möglich sind.

Wir machen noch einen kleinen Unterschied zwischen den Begriffen **Zufallsexperiment** und **zufälliger Vorgang** (vgl. Sill, 1993). Der zufällige Vorgang umfasst seine mögliche Endlichkeit. So kann sowohl das theoretisch unendliche Werfen eines Würfels, aber ebenso die Genese einer Eigenschaft eines Studierenden, die nur einmal stattfinden kann, als zufälliger Vorgang aufgefasst werden. Das Zufallsexperiment ist dagegen nur theoretisch möglich, da damit beliebig oft, unter gleichen Bedingungen wiederholbare Versuche gemeint sind.

5.1.2 Ergebnis und Ereignis

Betrachtet man Wahrscheinlichkeiten, so ist zunächst zu klären, von was diese Wahrscheinlichkeiten bestimmt oder gemessen werden sollen. Man kann das zwar in einfachen Fällen noch gut beschreiben, z.B. die Wahrscheinlichkeit für die „Augenzahl 2“ beim Werfen eines Würfels oder für „der nächste befragte Student raucht“, insgesamt hilft aber eine gewisse Formalisierung im Sinne einer sprachlichen Präzisierung.

Definition 15
Die möglichen Ausgänge eines zufälligen Vorgangs (bzw. eines Zufallsexperiments) heißen **Ergebnisse** ω, die Gesamtheit aller möglichen Ausgänge heißt **Ergebnismenge** Ω**.**

Hiermit kann man Zufallsexperimente unterscheiden, die endlich viele, abzählbar unendlich viele oder überabzählbar unendlich viele Ergebnisse besitzen. Wir werden uns im Hauptteil dieses Buches aber auf endliche Ergebnismengen beschränken.

Definition 16
Jede Teilmenge A der Ergebnismenge ($A \subseteq \Omega$) heißt **Ereignis** des zufälligen Vorgangs (des Zufallsexperiments). Die einelementigen Ereignisse heißen **Elementarereignisse**.

Durch diese formalisierende Klärung wird die Mengenalgebra eine Grundlage der Analyse von Wahrscheinlichkeiten. Innerhalb der Wahrscheinlichkeitstheorie erhalten dabei Verknüpfungen von Mengen (Ereignissen) spezielle Namen.

Definition 17
Es heißt:

- $A \cup B$ das **Vereinigungsereignis** (von A und B),
- $A \cap B$ das **Schnittereignis** (gilt $A \cap B = \emptyset$, so heißen die beiden Ereignisse A und B unvereinbar oder disjunkt),
- $\overline{A} = \Omega \setminus A$ das **Gegenereignis**.

Beispiel:
Gegeben sei ein normaler Spielwürfel.

- $\Omega = \{1,2,3,4,5,6\}$ ist die Ergebnismenge. Jedes Element dieser Menge, also z.B. 1, ist ein Ergebnis.
- Die Elementarereignisse haben alle die Form $\{\omega\} \subseteq \Omega$, $\{1\}$ ist also ein Elementarereignis (in begrifflicher Abgrenzung zum Ergebnis $1 \in \Omega$).
- $A = \{1,3,5\}$ ist ein mögliches Ereignis, das mit der Aussage „Es wurde eine ungerade Zahl gewürfelt“ beschrieben werden kann. Ebenso ist $B = \{1,2,3,4\}$, „eine Augenzahl kleinergleich 4“, ein mögliches Ereignis.
- $A \cup B = \{1,2,3,4,5\} = \Omega \setminus \{6\}$ ist die Vereinigung der beiden Ereignisse A: „ungerade Zahl“ , B: „Zahl kleinergleich 4“ . Anstatt das Ereignis $A \cup B$ zu betrachten, was mühsam sein kann, könnte auch das Gegenereignis $\overline{A \cup B} = \{6\}$ betrachtet werden.
- $A \cap B = \{1,3\}$ ist das Schnittereignis der beiden Ereignisse A und B.

Man sagt, „das Ereignis A ist eingetreten“, wenn $\omega \in A$ für eine Ausführung eines zufälligen Vorgangs gilt. Nimmt man an, dass die Erhebung des Alters von Studierenden ein zufälliger Vorgang ist, kann Folgendes gesetzt werden:

- $\Omega = \{15, ..., 100\}$ ist die Ergebnismenge, also ein begriffliches Pendant aus dem Bereich der Wahrscheinlichkeitsrechnung zum statistischen Begriff Merkmal (vgl. Kapitel 1.1), das mit X: Alter der Studierenden angegeben werden kann und das durch die Menge der Merkmalsausprägungen festgelegt ist.
- $\{\omega\} \subseteq \Omega$, also z.B. $\{20\}$, ist ein Elementarereignis und stellt ein Pendant (inhaltlich, nicht mengentheoretisch) zur Merkmalsausprägung eines Merkmals X dar.
- $A = \{15, 16, 17, 18, 19, 20\}$ ist das Ereignis, dass ein Student (eine Studentin) höchstens 20 Jahre alt ist. Das Ereignis ist damit allgemein ein Pendant zu einer Klasse von Merkmalsausprägungen eines Merkmals X.

Vereinigungs-, Schnitt- oder Gegenereignis könnten ebenso analog für das Merkmal des Alters der Studierenden formuliert werden.

Fasst man alle möglichen Ereignisse (Teilmengen von Ω) zusammen, so erhält man die **Potenzmenge** $\mathscr{P}$ von Ω als Menge aller Teilmengen von Ω:

Definition 18
Die Menge aller Ereignisse A zu einer Ergebnismenge heißt **Potenzmenge** $\mathscr{P}(\Omega) = \{A | A \subseteq \Omega\}$

Bei endlichen Ergebnismengen, die wir hier betrachten, ist die Potenzmenge $\mathscr{P}(\Omega)$ diejenige Menge, zu deren Elementen Wahrscheinlichkeiten bestimmt werden sollen.[1]

5.1.3 Zufallsgrößen

Analog zur Verwendung der Bezeichnungen X für Merkmale und x_i für Merkmalsausprägungen, verwenden wir für die Ereignisse eines zufälligen Vorgangs den Begriff der **Zufallsgröße** X:[2]

Definition 19
Die Funktion X, die jedem Ergebnis $\omega \in \Omega$ genau eine reelle Zahl $x \in \mathbb{R}$ zuordnet, heißt **Zufallsgröße**:

$$X : \begin{cases} \Omega \to \mathbb{R} \\ \omega \to x \end{cases}$$

Durch die Zufallsgröße X werden also Ergebnissen ω, die nicht notwendig Zahlen sein müssen, Zahlen x zugeordnet ($X(\omega) = x$), wobei verschiedenen Ergebnissen durchaus auch identische Zahlen zugeordnet werden können. Die Werte einer Zufallsgröße, die man auch Realisierungen der Zufallsgröße nennt, werden wir, falls das notwendig ist, mit Indizes $x_1, x_2, ...$ unterscheiden. Verschiedene Zufallsgrößen unterscheiden wir entweder durch Indizes $X_1, X_2, ...$ oder durch verschiedene Großbuchstaben X, Y. Ein Ereignis, bestehend aus einer Teilmenge von Ergebnissen $\omega \in \Omega$, die einem bestimmten Wert x der Zufallsgröße X zugeordnet sind, bezeichnen wir durch:

$$\{X = x\} := \{\omega \in \Omega | X(\omega) = x\}$$

[1] Ist $|\Omega| = n$, dann ist $|\mathscr{P}(\Omega)| = 2^n$. Wichtig für den mathematischen Aufbau ist hier die mengentheoretische Setzung $\emptyset \in \mathscr{P}(\Omega)$, die aber für unsere Betrachtungen nur am Rande bedeutsam ist.

[2] Ein Synonym für den Begriff Zufallsgröße ist *Zufallsvariable.*

Durch die Verwendung einer Zufallsgröße X werden wir den formalen Aufwand zur Beschreibung von Ereignissen eines zufälligen Vorgangs zum Teil erheblich vereinfachen können. Wir werden allerdings zu Anfang der Analyse von Wahrscheinlichkeiten zum Teil noch Ereignisse in der oben beschriebenen Weise verwenden, dann allerdings Schritt für Schritt nur noch Zufallsgrößen betrachten.

Beispiel:
Gegeben ist $\Omega = \{1,2,3,4,5,6\}$ als Ergebnismenge des einfachen Wurfs eines Würfels. Für diesen Wurf soll in einem stark konstruierten Beispiel untersucht werden, wie viele Teiler die geworfene Augenzahl hat. Man erhält:

Ereignisse	Zufallsgröße X: Anzahl der Teiler
A: Ein Teiler, $A = \{1\}$	$\{X = 1\}$
B: Zwei Teiler, $B = \{2,3,5\}$	$\{X = 2\}$
C: Drei Teiler, $C = \{4\}$	$\{X = 3\}$
D: Vier Teiler, $D = \{6\}$	$\{X = 4\}$

In diesem Beispiel werden z.B. drei Ergebnisse (2, 3 und 5) einem Wert der Zufallsgröße ($X = 2$) zugeordnet. Die Ergebnismenge reduziert sich dabei auf die Werte (Realisierungen) der Zufallsgröße.

Mit diesem Beispiel ergeben sich weitere sinnvolle Notationen wie:

$$\{X > 1\} = \{X = 2\} \cup \{X = 3\} \cup \{X = 4\} \qquad \text{oder} \qquad \{1 < X < 4\} = \{X = 2\} \cup \{X = 3\}$$

Wir werden bei Bedarf die Betrachtungen zu Zufallsgrößen, insbesondere die mit Zufallsgrößen mögliche Arithmetik ausbauen.

5.2 Wahrscheinlichkeitsbegriff

5.2.1 Klassische Wahrscheinlichkeit

Beim normalen Würfel würde vermutlich jede Person, die über das Kindesalter hinausgelangt ist, direkt die Wahrscheinlichkeit für alle Elementarereignisse als $\frac{1}{6}$ angeben. Die Wahrscheinlichkeit kann man dabei als Funktion interpretieren (Die Wahrscheinlichkeit von ...) und formal schreiben als $P(\{\omega\}) = \frac{1}{6}$.[3] Warum kann man aber die Wahrscheinlichkeit für jedes Elementarereignis so angeben? Man ist offenbar vertraut mit diesem *Zufallsgenerator* und kennt die Symmetrie des Würfels, die dazu führt, dass man alle Seiten für gleichberechtigt hält. Man kann natürlich auch Zweifel hegen, dass der Würfel wirklich symmetrisch ist, aber vor dem Würfeln ist in der Regel kein handfestes Argument für diesen Zweifel vorhanden. Für die Einschätzung der Wahrscheinlichkeiten zu einem Zufallsgenerator wie dem Würfel kommt damit zur Geltung:

[3] Wir haben die Funktion statt wie üblich mit f mit P bezeichnet. Diese Bezeichnung hat sich in Tradition von Laplace durchgesetzt. P steht dabei für probabilité. Die Argumente dieser Funktion P sind Ereignisse von zufälligen Vorgängen, also Mengen.

Prinzip des unzureichenden Grundes: Dieses Prinzip besagt, dass man an dem *Modell* der Gleichwahrscheinlichkeit von Elementarereignissen festhält, wenn man keinen ausreichenden Grund hat, an diesem Modell zu zweifeln.

Warum Modell? Man kann sich vorstellen, dass ein realer Würfel auch bei höchster Präzisionsarbeit nicht vollständig symmetrisch ist. Die Symmetrie ist damit eine vereinfachende Modellannahme (aber wir befinden uns ja auch in der Modell-Welt). Zufällige Vorgänge, bei denen das Prinzip des unzureichenden Grundes, etwa aufgrund einer Symmetrie des Zufallsgenerators, verwendet wird, erhalten einen eigenen Namen:

Definition 20
Ein Zufallsexperiment, bei dem alle Elementarereignisse eines Zufallsexperiments mit endlicher Ergebnismenge gleichwahrscheinlich sind, d.h. $\forall\{\omega\} \subset \Omega : P(\{\omega\}) = \frac{1}{|\Omega|}$, heißt **Laplace-Experiment**.[4]

Diese Art Experimente existieren nur in der Theorie. Vollständige Symmetrie und ebenso vollständige Gleichwahrscheinlichkeit ist in der Realität nicht existent. Auf der Modell-Ebene ermöglichen aber Laplace-Experimente die einfache Berechnung von Wahrscheinlichkeiten:

Satz 7
Genau dann, wenn ein Lapace-Experiment vorliegt, gilt für ein beliebiges Ereignis $A \in \mathscr{P}(\Omega)$:

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

Das lässt sich für endliche Ergebnismengen Ω so begründen: Da jedes Elementarereignis die Wahrscheinlichkeit $\frac{1}{|\Omega|}$ zugeordnet bekommt und jedes Ereignis A als Vereinigung von $|A|$ Elementarereignissen aufgefasst werden kann, ist der Satz unmittelbar einsichtig.

Beispiel:
Wenn man den Wurf eines handelsüblichen Spielwürfels betrachtet ($\Omega = \{1,2,3,4,5,6\}$), dann lassen sich ohne Rechnung zunächst zwei besondere Ereignisse angeben: das sichere Ereignis ($P(\Omega) = 1$) und das unmögliche Ereignis (z. B. $P(\{\}) = 0$).[5]
Seien die Ereignisse A: „ungerade Zahl“ und B: „Zahl kleinergleich 4“ gegeben. Dann ergibt sich mit dem klassischen Ansatz:

$$\begin{aligned} P(A) &= \frac{|\{1,3,5\}|}{|\Omega|} = \frac{3}{6}; \quad P(B) = \frac{|\{1,2,3,4\}|}{|\Omega|} = \frac{4}{6} \\ P(A \cap B) &= \frac{|\{1,3\}|}{|\Omega|} = \frac{2}{6} \\ P(A \cup B) &= \frac{|\{1,2,3,4,5\}|}{|\Omega|} = \frac{5}{6} = P(A) + P(B) - P(A \cap B) \end{aligned}$$

Die letzte Gleichung ist hier einleuchtend, allgemein aber ein Satz, den wir in den Erweiterungen allgemein beweisen werden.

[4] Mit $|A|$ wird die Mächtigkeit der Menge (des Ereignisses) A bezeichnet, d.h. die bei endlichen Mengen abzählbare Anzahl der Elemente von A.

[5] Damit lassen sich alltägliche Vorstellungen, wie z. B. das Würfeln einer 7, in Verbindung bringen.

Der klassische Ansatz hat Nachteile. Er ist nur für endliche Ergebnismengen anwendbar und zudem in der Realität fast ausschließlich auf Glücksspiele beschränkt. Der Ansatz gewinnt allerdings durch die Simulation von Zufallsexperimenten. Diese basieren stets auf prinzipiell gleichwahrscheinlichen Zufallszahlen, die wir als Erweiterung später näher betrachten (Kap. 5.5). Jede Simulation beruht damit zumindest implizit auf einem Laplace-Experiment.

Die Definition des Laplace-Experiments ist keine Definition des Wahrscheinlichkeitsbegriffs, da sie auf dem Begriff der Gleichwahrscheinlichkeit basiert, der selbst einer Definition bedarf. Würde man also die Wahrscheinlichkeit auf der Basis des Laplace-Experiments definieren, so würde man sich endlos im Kreise drehen.

5.2.2 Frequentistische Wahrscheinlichkeit

Betrachtet man die Eigenschaften von Studierenden wie auch sonst reale Situationen außerhalb des Glücksspiels, so wird man erkennen, dass keine (bzw. nahezu keine) dieser Situationen eine Modellierung mit gleichwahrscheinlichen Elementarereignissen (Laplace-Experiment) angemessen erscheinen lässt. Allgemein könnte man sagen:

- Konstruierte zufällige Vorgänge können ein Modell mit gleichwahrscheinlichen Elementarereignissen implizieren,
- bei realen, nicht kontrollierten zufälligen Vorgängen ist das Modell der Gleichwahrscheinlichkeit von Elementarereignissen in der Regel nicht angemessen.

Bei den zuletzt genannten zufälligen Vorgängen kann nur die *Erfahrung* im Umgang mit diesen zufälligen Vorgängen einen Ansatz ergeben, die Wahrscheinlichkeiten für die Elementarereignisse einzuschätzen. Dabei erweist sich das empirisch erfahrbare Phänomen als nützlich, das sich bei vielen Wiederholungen eines zufälligen Vorgangs hinsichtlich der relativen Häufigkeit eines Ereignisses zeigt: Die Schwankung der relativen Häufigkeiten nimmt mit wachsenden Versuchswiederholungen ab, die relative Häufigkeit *stabilisiert* sich. Dieses Phänomen bezeichnet man als **empirisches Gesetz der großen Zahlen**.

Wir untersuchen dieses Phänomen schrittweise anhand eines simulierten Würfels.[6] Wir betrachten dabei zunächst Zusammenfassungen von 10, 100 und 1000 Würfen und bestimmen für 50 dieser Simulationen jeweils die relative Häufigkeit der Augenzahl 6 (Abb. 5.2).

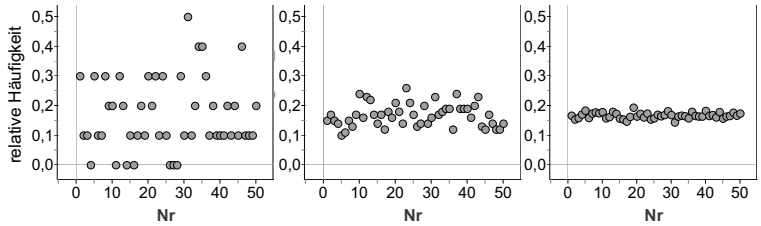


Abbildung 5.2: 50 Serien mit je 10 (links), 100 (Mitte) und 1000 Würfen (rechts), relative Häufigkeit der 6

[6] Erzeugen von gleichverteilten Zufallszahlen von 1 bis 6.

Im Boxplot-Vergleich wird die steigende Homogenität der relativen Häufigkeiten bei größeren Stichproben unmittelbar deutlich (Abb. 5.3). Je größer also die Stichprobe, desto geringer ist die Streuung – in diesem Beispiel offenbar um die „wahre“ relative Häufigkeit, also die Wahrscheinlichkeit eines Ereignisses, die hier mit $\frac{1}{6}$ in der Simulation vorgegeben wurde.

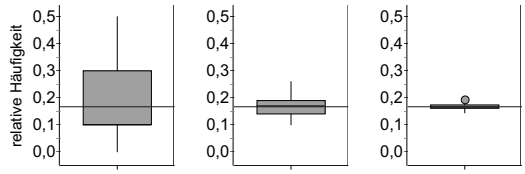


Abbildung 5.3: Streuung bei 50 Serien mit je 10 (links), 100 (Mitte) und 1000 Würfen (rechts), relative Häufigkeit der 6

Schließlich wollen wir auch die Entwicklung dieser Stabilisierung der relativen Häufigkeiten betrachten. Dazu kumulieren wir schrittweise die Anzahlen der Würfe von je 100 Simulationen sowie die Anzahlen für die 6, wie es in der folgenden Tabelle angedeutet wird:

Schritt	Anzahl Würfe	Anzahl 6	relative Häufigkeit 6	Würfe gesamt	Anzahl 6 gesamt	relative Häufigkeit gesamt
1	100	15	0,15	100	15	0,150
2	100	17	0,17	200	32	0,160
3	100	15	0,15	300	47	0,157
...	...	...	...	...	...	...

Setzt man diese Tabelle fort, so ergibt sich Abbildung 5.4, bei der links die relativen Häufigkeiten in den einzelnen Simulationen (angegeben ist die Nummer der Simulation), rechts die Entwicklung der kumulierten relativen Häufigkeiten dargestellt sind (für die Anzahl von Würfen).

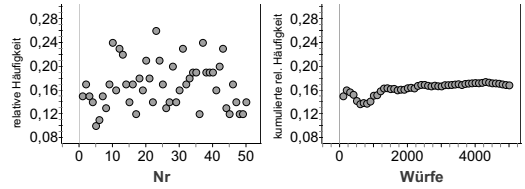


Abbildung 5.4: Relative Häufigkeiten in einzelnen Simulationen (links), Entwicklung der kumulierten relativen Häufigkeiten der 6 (rechts)

Während die relativen Häufigkeiten der einzelnen Simulationen mit je 100 Würfen offenbar schwanken, stabilisieren sich die kumulierten relativen Häufigkeiten mit Zunahme der Würfe. Es lässt sich zeigen, dass bei der Kumulation der potentiell mögliche Abstand zwischen zwei aufeinander folgenden relativen Häufigkeiten eines Ereignisses A immer geringer wird, d.h. dass die Abstände von aufeinander folgenden relativen Häufigkeiten gegen 0 gehen.

Satz 8
Sei $n \in \mathbb{N}$ mit $n > 1$, dann gilt für ein beliebiges Ereignis A eines Zufallsexperiments und $s \in \mathbb{N}$ konstant bei n Wiederholungen und $n \to \infty$:

$$|h_{n+s}(A) - h_n(A)| = 0$$

Wir beweisen diesen Satz in Kapitel 5.5.[7]

Aber auch das kann keine Definition des Wahrscheinlichkeitsbegriffs sein. Dazu müsste man zeigen, dass sich die relativen Häufigkeiten nicht nur stabilisieren, sondern stets auf einen eindeutig bestimmbaren Wert hin, nämlich einen Grenzwert, stabilisieren. Nennt man diesen Grenzwert $P(A)$ müsste man also zeigen: Für jeden beliebigen Abstand (den wir als ε bezeichnen) findet man eine fest bestimmte Anzahl von Versuchen n_0, so dass alle $h_n(A)$ mit $n > n_0$ einen kleineren Abstand zu $P(A)$ haben als ε, in Zeichen:

$$\forall \varepsilon > 0 : \exists n_0 \in \mathbb{N} : \forall_{n>n_0} |h_n(A) - P(A)| < \varepsilon$$

Existierte solch ein analytischer Grenzwert $P(A)$, so wäre damit eine Definition des Wahrscheinlichkeitsbegriffs gewonnen. Nehmen wir aber einmal an, dass solch ein ε existiert: Wir wählen ein ε aus, weil sich die relativen Häufigkeiten bei n_0 zufälligen Vorgängen stabilisiert haben, und wir nehmen an, dass der Abstand aller weiteren relativen Häufigkeiten $h_n(A)$, $n > n_0$ von der als Grenzwert betrachteten Zahl $P(A)$ immer kleiner als ε ist. Nehmen wir weiter an, das Ereignis A beschreibt das Eintreffen einer '6' beim Würfeln. Dann ist aber immer eine Folge von '6'en möglich (100, 100000, 10000000 ... mal die '6' hintereinander), so dass der Abstand $|h_n(A) - P(A)|$ größer ε wäre.[8]

Trotz dieser Schwierigkeit bei der Definition des Wahrscheinlichkeitsbegriffs bietet der frequentistische Ansatz eine plausible Möglichkeit, eine Wahrscheinlichkeit zu schätzen. Diese Schätzung bezeichnen wir durch $\hat{p}(A) \approx h_n(A)$. Mit $\hat{p}$ bezeichnen wir die geschätzte Wahrscheinlichkeit im Übergang von der realen Welt zur Modell-Welt. In der Modell-Welt werden die Wahrscheinlichkeiten festgelegt, die wir wie oben etwa mit $P(A)$ bezeichnen. Diese können auf einer Schätzung basieren, können auch mit einer relativen Häufigkeit übereinstimmen, bezeichnen aber ein nunmehr festgelegtes Modell.

Beispiel:
Wir betrachten den Wurf eines Würfels und analog die oben zu Laplace-Experimenten konstruierten Ereignisse sowie die Parteipräferenz von Studierenden (unter der Annahme einer repräsentativen Stichprobe).

- $1 = h_n(\Omega) = P(\Omega)$ (sicheres Ereignis). Bezogen auf die Umfrage zu den Studierenden wäre dieses Ereignis, einen Studierenden zu erhalten.

[7] Geht man in diesem Zusammenhang von Wiederholungen unter jeweils identischen Bedingungen aus, dann lässt sich so der Begriff der *stochastischen Unabhängigkeit* (vgl. Kap. 6.1) motivieren. Wie in diesem Satz berechnen wir im Folgenden auch relative Häufigkeiten von Ereignissen als Klassen von Merkmalsausprägungen (vgl. S. 99).

[8] Der Mathematiker van Mises hatte Anfang des vergangenen Jahrhunderts versucht, den Wahrscheinlichkeitsbegriff über relative Häufigkeiten (unter zusätzlichen Annahmen) zu definieren, war mit diesem Versuch jedoch nicht vollständig erfolgreich.

- $0 = h_n(\{\}) = P(\{\})$ (unmögliches Ereignis). Auch einen Studierenden mit Präferenz zu der unseres Wissens formal nicht existierenden Partei *Ritter der Stochastik (RdS)* zu erhalten, ist solch ein unmögliches Ereignis $0 = h_n(\{\}) = P(\{\})$.[9] Beim sicheren wie beim unmöglichen Ereignis ist die Schätzung sicher.
- A: „ungerade Zahl“ , B: „Zahl kleinergleich 4“ bzw. A: „Präferenz für die Klasse der vor 1960 existierenden Parteien (SPD, CDU, FDP)“, B: „Präferenz für die Klasse der als links von der Mitte angesehenen Parteien (SPD, Grüne, Linke)“. Die Ereignisse werden (siehe oben) als Klassen von Merkmalsausprägungen betrachtet, die nicht notwendig disjunkt sind.

$$\begin{aligned} h_n(A \cup B) &= \frac{H_n(A \cup B)}{n} = \frac{H_n(A) + H_n(B) - H_n(A \cap B)}{n} \\ &= h_n(A) + h_n(B) - h_n(A \cap B) \\ &\approx \hat{p}(A \cup B) \end{aligned}$$

  Ausgehend von dieser Schätzung kommt man zu einer Festlegung im Modell z. B. durch $P(A \cup B) := \hat{p}(A \cup B)$. Betrachtet man nur das Beispiel der Parteipräferenz, so ist hier unmittelbar einleuchtend, dass die Häufigkeit doppelt gezählter Merkmalsausprägungen (hier zur SPD) bei der Schätzung von $P(A \cup B)$ beachtet werden muss.

Dass die relative Häufigkeit eines zufälligen Ereignisses der Ausgangspunkt einer *Schätzung* ist, kann man sich anhand eines teilsymmetrischen Zufallsgenerators wie z.B. des Quaderwürfels (nach seinem Ideengeber auch *Riemer-Quader* genannt, vgl. Abb. 5.5) deutlich machen.

Abbildung 5.5: Teilsymmetrischer Quaderwürfel

Nach rund 3500 in unseren Veranstaltungen zur Stochastik ausgeführten Würfen des Quaderwürfels haben sich für die Ereignisse $\{3\}$ und $\{4\}$ die Häufigkeiten ergeben: $h_{3435}(\{3\}) = 0,344$ und $h_{3435}(\{4\}) = 0,353$. Die Tatsache, dass die gegenüberliegenden Seiten des Quaderwürfels zumindest im Modell sinnvoll als identisch anzunehmen sind $P(\{3\}) = P(\{4\})$, führt zu einer Schätzung einer identischen Wahrscheinlichkeit für beide Augenzahlen, wie z.B.:

$$h_{3435}(\{3\}) \approx \hat{p}(\{3\}) = 0,35 := P(\{3\}) \quad \text{und} \quad h_{3435}(\{4\}) \approx \hat{p}(\{4\}) = 0,35 := P(\{4\})$$

[9] Unter der gegebenen Voraussetzung, dass die Studierenden nur unter in Deutschland real existierenden Parteien auswählen durften und damit *RdS* kein Element der Ergebnismenge ist.

5.2.3 Axiomatische Wahrscheinlichkeit

Weder der klassische noch der frequentistische Wahrscheinlichkeitsbegriff geben eine Definition dessen, was mathematisch als Wahrscheinlichkeit verstanden wird. Die Klärung, was eine Wahrscheinlichkeit ist, wurde von David Hilbert auf einem Mathematikerkongress 1900 als eine der drängendsten Fragen der mathematischen Physik bezeichnet. Der russische Mathematiker A. Kolmogoroff löste diese Frage und veröffentlichte 1933 sein Axiomensystem, das heute als allgemein anerkannte Grundlage der Wahrscheinlichkeitstheorie gilt. Das Axiomensystem erklärt den Begriff der Wahrscheinlichkeit rein formal und gibt Kriterien an, wann von mathematischer Wahrscheinlichkeit gesprochen werden kann.[10] Alle bekannten Sätze der Wahrscheinlichkeitstheorie bauen auf diesen Axiomen auf. Zudem erkennt man schnell Parallelen zu den beiden behandelten Ansätzen von Wahrscheinlichkeit.

Definition 21
(Axiomatische Definition der Wahrscheinlichkeit) Ist $\mathscr{P}(\Omega)$ die Menge aller Teilmengen bzw. Ereignissen $A \subseteq \Omega$ eines Zufallsexperiments, dann heißt die Funktion $P : \mathscr{P} \to \mathbb{R}$ **Wahrscheinlichkeitsfunktion** bzw. **Wahrscheinlichkeit**, wenn folgende Axiome erfüllt sind:

K1: $P(A) \geq 0$

K2: $P(\Omega) = 1$

K3: $P(\bigcup_i A_i) = \sum_i P(A_i), \quad \text{mit} \quad \forall i \neq j : A_i \cap A_j = \emptyset$

$\overline{K3}$: $P(A \cup B) = P(A) + P(B), \quad \text{mit} \quad A \cap B = \emptyset$

Das Paar (Ω, P) wird, wenn die drei genannten Bedingungen zutreffen, auch als **Wahrscheinlichkeitsraum** bezeichnet. Für die folgenden Sätze gilt stets als Voraussetzung, dass (Ω, P) solch einen Wahrscheinlichkeitsraum bilden, ohne dass wir das explizit formulieren.

5.2.4 Wahrscheinlichkeiten und Zufallsgrößen

Ist eine Zufallsgröße X gegeben, so lässt sich eine Wahrscheinlichkeit für einen bestimmten Wert (eine Realisierung) der Zufallsgröße durch $P(\{X = x\})$ angeben. Auch hier vereinfachen wir die Notation durch

$$P(X = x) := P(\{X = x\})$$

Analog zur Häufigkeitsverteilung (Kap. 2.1.4) bezeichnen wir weiterhin die Funktion, die allen $x \in X$ eine Wahrscheinlichkeit zuordnet, als **Wahrscheinlichkeitsverteilung**:

Definition 22
Sei X eine Zufallsgröße, so heißt

$$f(x) := P(X = x)$$

Wahrscheinlichkeitsverteilung der Zufallsgröße.

[10] Nur im Umfeld der eigentlichen axiomatischen Definition nimmt Kolmogoroff Bezug auf Häufigkeiten und Zufallsexperimente. Dort fordert Kolmogoroff, dass die Wahrscheinlichkeiten, die scheinbar mehr oder weniger willkürlich festgelegt werden können, mit der Realität übereinstimmen, das heißt sich in langen Versuchsreihen bewährt oder angedeutet haben (Kolmogoroff, 1973).

Analog zur empirischen Verteilungsfunktion (Kap. 2.1.5) definieren wir:

Definition 23
Sei X eine Zufallsgröße, so heißt

$$F(x) := P(X \leq x)$$

Verteilungsfunktion der Zufallsgröße.

5.3 Modell-Welt – reale Welt

Im vorangegangenen Kapitel haben wir bezogen auf den Wahrscheinlichkeitsbegriff die Beziehungen zwischen der Modell-Welt (der Welt der Wahrscheinlichkeiten von Ereignissen) und der realen Welt (der Welt der relativen Häufigkeiten von Merkmalsausprägungen) überwiegend implizit verwendet. Wir wollen diese Beziehungen hier explizit machen.

reale Welt		Modell-Welt
Es liegt ein zufälliger Vorgang vor, bei dem die Symmetrie der möglichen Ausgänge empirisch begründet als plausibel angenommen werden kann.	$\rightarrow$	Der zufällige Vorgang wird als Laplace-Experiment modelliert. In diesem Modell werden die Wahrscheinlichkeiten der Elementarereignisse durch $P(\{\omega\}) = \frac{1}{\lvert\Omega\rvert}$ festgelegt.
Es liegt ein zufälliger Vorgang vor, bei dem keine Symmetrie vorliegt. Der zufällige Vorgang wird möglichst häufig wiederholt und die relativen Häufigkeiten von Merkmalsausprägungen bestimmt.	$\rightarrow$	Die relativen Häufigkeiten der Merkmalsausprägungen sind die Basis für die Schätzung der Wahrscheinlichkeit von Elementarereignissen $\hat{p}_i \approx h_n(x_i),\ i = 1,...,s.$ Diese Schätzungen sind wiederum die Grundlage für das Aufstellen eines Modells, in dem die Wahrscheinlichkeiten $P(\{\omega_i\})$ für die vorliegenden Elementarereignisse festgelegt werden.
In beiden Varianten ist das Ziel, die Wahrscheinlichkeiten von Elementarereignissen $\omega \in \Omega$ und damit allgemeiner für alle Ereignisse $A \in \mathscr{P}(\Omega)$ festzulegen.		

Modell-Welt		reale Welt
Es sind die Wahrscheinlichkeiten für die Elementarereignisse eines zufälligen Vorgangs gegeben.	$\rightarrow$	Die relativen Häufigkeiten einer Merkmalsausprägung x_i stabilisieren sich bei der wiederholten Ausführung (Durchführung) des zufälligen Vorgangs – wenn das Modell gut aufgestellt war – in der Nähe der Wahrscheinlichkeit $P(\{\omega_i\})$ (empirisches Gesetz der großen Zahl).

Wenn also ein realer, als zufällig erachteter Vorgang betrachtet wird, wird in der Wahrscheinlichkeitsanalyse aufgrund der Beschaffenheit des realen Vorgangs ein Modell mit Wahrscheinlichkeiten konstruiert, das wiederum für die Vorhersage zukünftiger Durchführungen des zufälligen Vorgangs dient.

Bei dem Übergang von der realen Welt in die Modell-Welt ist einsichtig geworden, dass beim frequentistischen Ansatz eine Wahrscheinlichkeit nur unzureichend aus relativen Häufigkeiten einer kleinen Stichprobe geschätzt werden kann. Entsprechend gilt dies auch beim Übergang von der Modell-Welt in die reale Welt. Eine Wahrscheinlichkeit eines Ereignisses macht keine Aussage über *eine* Durchführung eines zufälligen Vorgangs, sondern nur über *viele*.

Beispiel:
Das Modell eines Würfels mit $\Omega = \{1,2,3,4,5,6\}$ und $P(\{\omega_i\}) = \frac{1}{6}, i = 1,...,6$ ist für den (einmaligen) Wurf des Würfels gegeben. Der zufällige Vorgang des Würfel-Werfens wird nun einmal durchgeführt. Es ergibt sich die in Abbildung 5.6 zu sehende relative Häufigkeitsverteilung für die Augenzahlen:

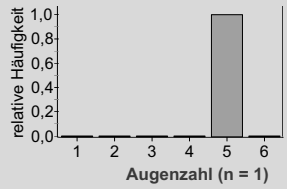


Abbildung 5.6: Ergebnis nach einmaligem Wurf

Diese Häufigkeitsverteilung hat wenig mit dem Modell der gleichwahrscheinlichen Elementarereignisse gemein. Erst wenn man den zufälligen Vorgang häufiger wiederholt (vgl. Abb. 5.7), erkennt man die Prognosekraft des Modells.

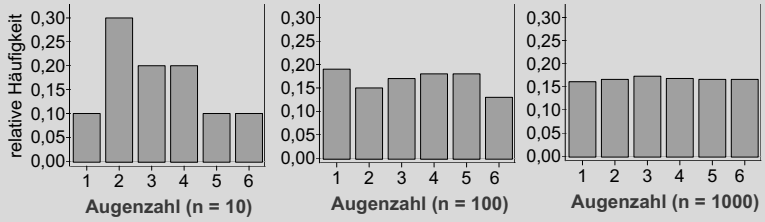


Abbildung 5.7: Ergebnis nach mehreren Würfen (10, 100, 1000)

Bleibt man vollständig in der Modell-Welt, so bedeutet die Wahrscheinlichkeit eines Ereignisses, dass bei theoretisch unendlich vielen Wiederholungen eines zufälligen Vorgangs (die es in der realen Welt nicht gibt) die relative Häufigkeit des Ereignisses der Wahrscheinlichkeit entspricht. Das ist aber eine Behauptung, die wir erst in Kapitel 7 wieder aufnehmen werden.

5.4 Eigenschaften von Studierenden: Schätzungen von Häufigkeiten in der Grundgesamtheit

Die Stichproben in Münster und Freiburg haben ungefähr 3-4% aller Studierenden der Hochschulen erfasst. Die Schätzung von den in der Grundgesamtheit der Studierenden vorherrschenden Häufigkeiten zu verschiedenen Eigenschaften ist damit mit Unsicherheit behaftet, selbst wenn die Stichproben an beiden Hochschulen als repräsentativ anzusehen wären. Eine andere Möglichkeit, als die relative Häufigkeit in der Stichprobe als Schätzung zu verwenden, bleibt uns allerdings bei dem bisherigen Stand der theoretischen Überlegungen nicht.

Wir zeigen zunächst, dass sich auch die relativen Häufigkeiten für die Merkmalsausprägungen (hier interpretiert als Elementarereignisse) stabilisieren. Dafür ist in Abbildung 5.8 exemplarisch die Stabilisierung für die Merkmalsausprägungen *weiblich* und *männlich* bei der Kumulation der Häufigkeiten der Merkmalsausprägungen in Teilstichproben von je 50 Studierenden dargestellt.

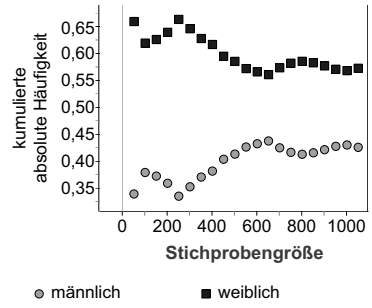


Abbildung 5.8: Stablilisierung der relativen Häufigkeiten zu den Merkmalsausprägungen *weiblich* und *männlich*

Das bedeutet, *wenn* man der Repräsentativität der Stichprobe trauen kann, so wird die Schätzung der relativen Häufigkeit einer Merkmalsausprägung, die der Grundgesamtheit zugrunde liegt, durch die entsprechenden relative Häufigkeiten größerer Teilstichproben potentiell besser.

Betrachtet man die Repräsentativität, so könnte etwa bei der Bevorzugung bestimmter Studiengänge, die einen größeren Anteil eines Geschlechts aufweisen, die Stichprobe und damit das Ergebnis der Stichprobe verfälscht werden. Möglicherweise bedingen auch andere, unvermutete Eigenschaften den Anteil der Geschlechter. Für das Merkmal Geschlecht ist es aber möglich, den *offiziellen* Anteil zu ermitteln und mit dem in der Stichprobe bestimmten zu vergleichen.

Die Universität Münster gibt für das Studienjahr 2009/2010, dem Zeitraum der Stichprobe, den Anteil der weiblichen Studierenden mit rund 53% an. In der Stichprobe ergab sich dagegen ein Anteil von rund 57%. Ob der Unterschied von ungefähr 4 Prozentpunkten auf eine Verzerrung der Stichprobe (z.B. zu viele eher weibliche Lehramtsstudierende) zurückzuführen ist, lässt sich mit dem bisherigen Wissen noch nicht entscheiden. Wir werden in Kapitel 7.3 und 8 wieder auf diesen Unterschied eingehen. In gleicher Weise gibt die PH Freiburg den Anteil der Studentinnen mit 76% an. In der Stichprobe hat sich dagegen ein Anteil von rund 71% ergeben – auch hier ein Unterschied, und auch hier wiederum die Frage, wie bedeutsam dieser ist.

Betrachtet man den offiziell angegebenen Anteil der Studentinnen etwa an der Universität Münster von rund 53%, so könnte man als – wegen der gegebenen Vollerhebung aller Studierenden – bis auf Rundungen sichere Schätzung für das Ereignis W: Studentin mit $P(W) = 0,53$ verwenden. Damit wird die Erhebung des Geschlechts von Studierenden mit dem Werfen eines zweiseitigen Würfels gleichgesetzt. Einmal geworfen, wird dieser Würfel das Geschlecht eines Studierenden ergeben und die *zufällige* Befragung eines Studierenden simulieren. Wird dieser einfache Versuch sehr häufig durchgeführt, so sollte sich die relative Häufigkeit $h_n(W)$ einerseits stabilisieren, andererseits ungefähr der Häufigkeit in der Grundgesamtheit, als Wahrscheinlichkeit $P(W)$ interpretiert, entsprechen. Was aber ist, wenn man, was ja eher der Realität entspricht, eine Stichprobe mit einem Umfang größer 1 erhebt, d.h. mehrere Studierende befragt. Die Überlegungen in der Modell-Welt zu solchen mehrstufigen zufälligen Vorgängen werden wir im folgenden Kapitel 6 aufnehmen.

5.5 Ergänzungen

5.5.1 Das empirische Gesetz der großen Zahlen

Wir hatten in Kapitel 5.2 behauptet, dass der potentielle Abstand zweier relativer Häufigkeiten $h_{n+s}(A)$ und $h_n(A)$, die sich hinsichtlich der Erhöhung der Stichprobe um s unterscheiden, für $n \to \infty$ gegen 0 geht. Für den Beweis betrachten wir zu einem zufälligen Ereignis A die beiden im Prozess der Kumulation entstehenden relativen Häufigkeiten $h_n(A)$ und $h_{n+s}(A)$, wobei s eine beliebige, aber feste Erhöhung des Stichprobenumfangs bzw. der Wiederholungen des zufälligen Vorgangs ist.

$$\begin{aligned} |h_{n+s}(A) - h_n(A)| &= \left| \frac{H_{n+s}(A)}{n+s} - \frac{H_n(A)}{n} \right| \\ &= \left| \frac{n \cdot H_{n+s}(A) - (n+s) \cdot H_n(A)}{n \cdot (n+s)} \right| \\ &= \left| \frac{n \cdot (H_{n+s}(A) - H_n(A)) - s \cdot H_n(A)}{n \cdot (n+s)} \right| \end{aligned}$$

Durch diese Umformung kann man die erste Abschätzung durch $H_{n+s}(A) - H_n(A) \leq s$ erhalten, da bei s zusätzlichen Versuchsdurchgängen maximal s mal A zutreffen kann.

$$\begin{aligned} |h_{n+s}(A) - h_n(A)| &\leq \frac{n \cdot s - s \cdot H_n(A)}{n \cdot (n+s)} = \frac{s \cdot (n - H_n(A))}{n \cdot (n+s)} \\ &= \frac{n \cdot s \cdot (1 - h_n(A))}{n \cdot (n+s)} = \frac{s}{n+s} \cdot (1 - h_n(A)) \end{aligned}$$

Im ersten Term auf der rechten Seite der voranstehenden Ungleichung können die Betragsstriche mit der Überlegung weggelassen werden, dass aus $n, s \in \mathbb{N}$ mit $n, s \geq 1$ und $n \geq H_n(A)$ (da bei n Versuchsdurchgängen maximal n mal A eintreten kann) folgt, dass der Bruch keinen

Wert kleiner als Null annehmen kann. Man erhält nach dieser Umformung unmittelbar durch Abschätzung

$$(1-h_n(A)) \leq 1 \text{ und damit } |h_{n+s}(A)-h_n(A)| \leq \frac{s}{n+s} \cdot (1-h_n(A)) \leq \frac{s}{n+s}$$

Für $n \to \infty$ geht der rechte Term und damit der Abstand zweier aufeinanderfolgenden relativen Häufigkeiten gegen 0. Der rechte Term kann dabei als maximale potentielle Spannweite des Abstandes der aufeinanderfolgenden relativen Häufigkeiten interpretiert werden.

5.5.2 Addition zweier Wahrscheinlichkeiten

Wir hatten weiterhin behauptet, dass allgemein für die Wahrscheinlichkeit zweier Ereignisse A und B gilt:

Satz 9
Seien A und B zufällige Ereignisse, so gilt: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Der Beweis dieses im Rahmen des klassischen wie frequentistischen Wahrscheinlichkeitsbegriffs plausiblen Satzes geschieht durch Umformung des Vereinigungsereignisses in disjunkte Mengen und Anwendung des dritten Axioms.

Es gilt

$$\begin{aligned} A \cup B &= (A \setminus B) \cup (A \cap B) \cup (B \setminus A) \\ A &= (A \setminus B) \cup (A \cap B) \\ B &= (B \setminus A) \cup (A \cap B) \end{aligned}$$

Für alle drei Vereinigungen disjunkter Ereingisse lässt sich das dritte Axiom anwenden:

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \\ P(A) &= P(A \setminus B) + P(A \cap B) \\ P(B) &= P(B \setminus A) + P(A \cap B) \end{aligned}$$

Formt man die zweite und dritte Gleichung um nach $P(A \setminus B)$ bzw. $P(B \setminus A)$ und setzt in die erste Gleichung ein, so erhält man:

$$\begin{aligned} P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \\ &= (P(A) - P(A \cap B)) + P(A \cap B) + (P(B) - P(A \cap B)) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Dass sich der oben genannte und auch ohne Beweis plausible Satz einfach beweisen lässt, ist keine Überraschung.

5.5.3 Zufallszahlen

Simulationen basieren auf der geplanten Erzeugung von gleichverteilten Zufallszahlen. Diese könnten beispielsweise mit einem möglichst sauber gearbeiteten Würfel erzeugt werden, was

allerdings langwierig ist. Die geplante Erzeugung vieler gleichverteilter Zufallszahlen mit dem Rechner scheint dagegen ein Widerspruch in sich zu sein, da mit dem Rechner nur deterministische Algorithmen, also Algorithmen, deren Ablauf gerade nicht zufällig ist, konstruierbar sind. Die mit dem Rechner erzeugten Zahlen werden deshalb auch Pseudo-Zufallszahlen genannt. Da wir in diesem Buch nicht diskutieren, ob und warum ein Vorgang tatsächlich zufällig ist, sondern solche Vorgänge als zufällig postulieren, bei denen es nützlich ist, werden wir auch die von einem Rechner erzeugten Zahlen dann als Zufallszahlen betrachten, wenn es als sinnvoll erscheint.

Eine Möglichkeit, Zufallszahlen mit einem deterministischen Algorithmus zu erzeugen, geschieht mit der Methode eines **linearen Kongruenzgenerators**:

- Wähle eine Startzahl $y_0 \in \mathbb{N} \cup \{0\}$ und ein $m \in \mathbb{N}$.
- Berechne den Rest y_1, der beim Teilen der Zahl $a \cdot y_0 + b$ durch m entsteht ($b \in \mathbb{N} \cup \{0\}$ und $a \in \mathbb{N}$).
- Setze $x_1 = \frac{y_1}{m}$ als erste Zufallszahl fest ($0 \leq x_1 \leq 1$).
- Wiederhole den Vorgang durch Bestimmen der Reste y_{i+1}, die beim Teilen der Zahlen $a \cdot y_i + b$ ($i = 1, 2, \ldots$) durch m entstehen, sowie durch die Bestimmung der zugehörigen Zufallszahlen $x_i = \frac{y_i}{m}$.

Die Güte dieses Algorithmus orientiert sich insbesondere an der möglichen Anzahl der Reste beim Teilen durch m, also auch der Größe von m. Gütekriterien sind die Gleichverteilung der Zahlen zwischen 0 und 1 sowie einer Regellosigkeit in dem Sinne, dass es in der notwendig vorhandenen Periode von Resten keine Abhängigkeit zwischen den verschiedenen Resten gibt. Eine Möglichkeit, diese vermeintliche Regellosigkeit zu visualisieren, geschieht dadurch, dass die Punkte mit den Koordinaten $(x_i; x_{i+1})$ im Kordinatensystem dargestellt werden.

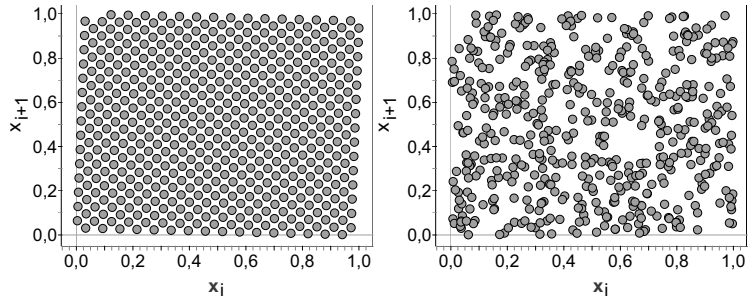


Abbildung 5.9: Zufallszahlen mit $a = 33$, $b = 1$, $m = 512$ und $y_0 = 0$ (links) und $a = 427419669081$, $b = 0$, $m = 10^{12} - 11$ und $y_0 = 1$ (rechts, zu diesem Beispiel vgl. Henze, 2010)

In Abbildung 5.9 ist ein offenbar wenig geeigneter linearer Kongruenzgenerator sowie ein offenbar geeigneter repräsentiert. Kriterien für die Wahl günstiger Parameter eines linearen Kongruenzgenerators sind etwa in Henze (2010) diskutiert. Wichtig ist hier allein zu zeigen, dass die Erzeugung von Zufallszahlen durchaus angemessen sein kann, um den Zufall zu simulieren.

5.6 Aufgaben

Aufgabe 5.1: Gegeben ist der normale Würfel.

a) Nennen Sie ein beliebiges Ereignis A für den einfachen Wurf dieses Würfels als Menge und in Worten.

b) A sei das Ereignis: es fällt eine Primzahl. B sei das Ereignis: es fällt eine gerade Zahl.

   (i) Bestimmen Sie das Schnittereignis sowie das Vereinigungsereignis von A und B.

   (ii) Bestimmen Sie das Komplementärereignis zu A.

c) Bestimmen Sie die Menge zu folgenden Ereignissen:

   (i) A: es fällt eine Augenzahl, die größer als 3 ist.

   (ii) B: es fällt eine Augenzahl, die mindestens 5 ist.

   (iii) C: es fällt eine Augenzahl, die kleiner als 1 ist.

   (iv) D: es fällt eine Augenzahl, die höchstens 2 ist.

d) Ermitteln Sie die Ergebnismenge Ω für den zweifachen Wurf dieses Würfels.

e) Nennen Sie ein beliebiges Ereignis des zweifachen Wurfs.

f) Erklären Sie anhand des zweifachen Würfelwurfs, was ein Ergebnis und was ein Ereignis ist.

Aufgabe 5.2:

a) Bestimmen Sie die Wahrscheinlichkeit dafür, dass beim zweifachen Wurf des Tetraeder-Würfels die Differenz der beiden Augenzahlen 0 oder 1 ergibt (stets: größere Augenzahl – kleinere Augenzahl).

b) Bestimmen Sie die Wahrscheinlichkeit dafür, dass beim zweifachen Wurf des normalen Würfels die Summe der beiden Augenzahlen 6, 7 oder 8 ergibt.

Aufgabe 5.3:

a) Überlegen Sie, bei welchen realen Ereignissen es Ihrer Meinung nach sinnvoll ist, von Wahrscheinlichkeiten zu sprechen.

b) Überlegen Sie weiter, wie Ihrer Meinung nach zu einem solchen Ereignis eine Wahrscheinlichkeit bestimmt werden kann.

c) Von einem *Risiko* wird gesprochen, wenn ein zufälliges Ereignis mit einer negativen Konnotation eintreten kann. Nennen Sie Situationen mit Risiken und versuchen Sie, das Eintreten von Risiken in solchen Situationen abzuschätzen.

Aufgabe 5.4: Versuchen Sie, eine Serie von 100 Zufallsziffern zu bestimmen, indem Sie den Zufall gedanklich simulieren. Notieren Sie diese Zufallszahlen (wir nehmen die Aufgabe später wieder auf).

6 Mehrstufige zufällige Vorgänge

Einstiegsbeispiel

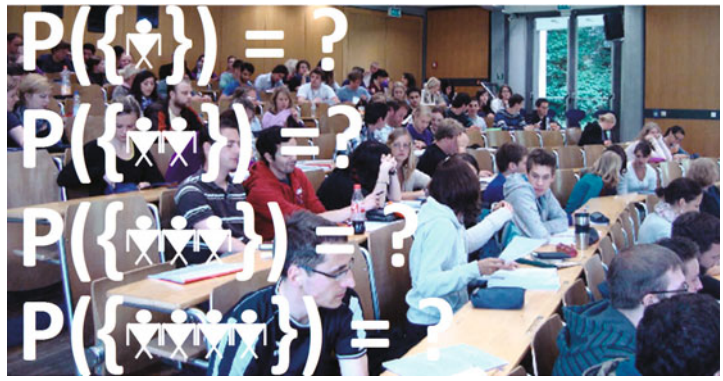


Abbildung 6.1: Prognosen schrittweise entwickeln

Aufgabe 1: Im vorangegangenen Kapitel haben wir Wahrscheinlichkeiten für Eigenschaften der Studierenden geschätzt. Bestimmen Sie auf dieser Basis modellhaft die Wahrscheinlichkeiten für verschiedene Anzahlen von Studierenden mit einer bestimmten Eigenschaft.

Worum es geht

Im vergangenen Kapitel haben wir uns damit beschäftigt, einen einzelnen zufälligen Vorgang zu modellieren. Wie fern das der Realität aber ist, haben wir bei der Betrachtung der Eigenschaften von Studierenden in Kapitel 5.4 gesehen. So betrachtet man, wenn man die Realität analysiert, in der Regel nicht einen isolierten zufälligen Vorgang, sondern die Hintereinanderschaltung mehrerer zufälliger Vorgänge, also etwa

- nicht einen Wurf eines Würfels, sondern mehrere, oder
- nicht die Erhebung eines Studierenden, sondern eine Stichprobe mit mehreren Studierenden.

Im Gegensatz zum Beginn des vergangenen Kapitels befinden wir uns dabei vollständig in einer Modell-Welt, in der Wahrscheinlichkeiten für Ereignisse (Eigenschaften der Studierenden) festgelegt werden. Wir wollen dabei wieder einen Schritt auf die Realität zugehen und im Modell bestimmen, mit welchen Anzahlen von Studierenden wir in einer bestimmten Stichprobe rechnen

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_6,

können. Ist etwa A: „Studierender ist Single“ , so wird auf der Basis der empirischen bestehenden Stichprobe $P(A)$ gesetzt. Nun sollen die Singles in einer zukünftigen Stichprobe theoretisch gezählt werden und zwar mit Hilfe einer Zufallsgröße X: Anzahl der Singles in einer Stichprobe vom Umfang n bzw. die Wahrscheinlichkeit für bestimmte Werte der Zufallsgröße bestimmt werden. Bei der theoretischen Durchdringung einer konkreten Modellierung einer zukünftigen Durchführung einer Erhebung stoßen wir dabei auf zentrale Begriffe der Wahrscheinlichkeitstheorie. Dabei werden wir im Hauptteil des Kapitels weiter mit der Ereignisschreibweise (z.B. $P(A)$) und erst in den Ergänzungen mit Zufallsgrößen arbeiten.

Abhängigkeit, Unabhängigkeit Diese beiden Begriffe sind zentral für die gesamte Analyse von Wahrscheinlichkeiten. Den Begriff der (stochastischen) Unabhängigkeit haben wir implizit bereits bei der Interpretation von Wahrscheinlichkeiten wie auch dem frequentistischen Zugang zum Wahrscheinlichkeitsbegriff verwendet, die nur dann ihre Bedeutung auf lange Sicht erhalten, wenn man von jeweils identisch ausgeführten zufälligen Vorgängen (z.B. Erhebung eines Studierenden, Werfen eines Würfels ausgeht), also nicht von Vorgängen, in denen ein Vorgang Auswirkungen auf die dann (stochastisch) abhängigen weiteren Vorgänge hat. Diese beiden implizit verwendeten Begriffe werden wir im Folgenden explizit machen.

Visualisierung Bei der Analyse von Wahrscheinlichkeiten mehrstufiger Vorgänge muss der zufällige Prozess jeder Stufe vollständig bekannt sein. Daher bieten sich bei dieser Analyse Visualisierungen des Prozesses und seiner Struktur an, die wir mit dem Baum sowie dem Einheitsquadrat veranschaulichen werden.

Kombinatorische Überlegungen Bei komplexeren oder vielstufigen zufälligen Vorgängen versagt eine Visualisierung in der Regel. Hier können Zählstrategien für bestimmte Ereignisfolgen helfen. Die dazu vorhandenen kombinatorischen Hilfsmittel werden wir in aller Kürze in diesem Kapitel bereitstellen, in dem wir ein zentrales Modell der Wahrscheinlichkeitsanalyse, den zufälligen Zug einer Kugel aus einer Urne, das sogenannte Urnenmodell, betrachten.

Lernen aus Erfahrung – subjektivistische Wahrscheinlichkeit Im Zusammenhang mit mehrstufigen zufälligen Vorgängen ist ein weiterer Wahrscheinlichkeitsbegriff, der subjektivistische, entwickelbar. Im Kern umfasst dieser Begriff die allmähliche Verbesserung einer Entscheidungsfindung zwischen zwei oder mehreren Alternativen in bestimmten Situationen auf der Basis von sukzessiv gesammelten weiteren Informationen.

6.1 Abhängigkeit – Unabhängigkeit zufälliger Vorgänge

Der Kern aller folgenden Überlegungen besteht in der Definition der **bedingten Wahrscheinlichkeit**, die allgemeine Grundlage der Analyse mehrstufiger zufälliger Vorgänge ist. Wir führen dazu zunächst einen neuen Ereignisbegriff ein:

Definition 24
Das Ereignis $A|B$ bezeichnet das Ereignis A, das unter der Bedingung, dass das Ereignis B zutrifft, betrachtet wird.

Diese Definition umfasst bereits zwei Ereignisse, die wir im Folgenden zumeist als zwei Ereignisse hintereinander ausgeführter zufälliger Vorgänge betrachten (ohne Festlegung der Chronologie). Analog zu der Einführung einer bedingten Häufigkeit in Kapitel 3 wird hier die Betrach-

tung des Ereignisses A nicht mehr auf die Ergebnismenge Ω bezogen, sondern auf ein Ereignis B eingeschränkt. Die Wahrscheinlichkeit für ein bedingtes Ereignis wird sinnfällig durch die Einschränkung auf das Ereignis B folgendermaßen definiert:

Definition 25
Sei $P(B) > 0$ und $A, B \subseteq \Omega$. Mit

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

bezeichnen wir die **bedingte Wahrscheinlichkeit** des Ereignisses A unter der Bedingung B.[1]

Aus der Definition folgt unmittelbar

Satz 10 (Multiplikationssatz):
Seien A und B zwei Ereignisse eines zufälligen Vorgangs, so gilt:

$$P(A \cap B) = P(B) \cdot P(A|B) \text{ und } P(A \cap B) = P(B \cap A) = P(A) \cdot P(B|A)$$

Man kann weiterhin zeigen, dass für $P(A|B)$ die drei Axiome der Wahrscheinlichkeitsrechnung ihre Gültigkeit behalten und stets: $P(A \cap B) \leq P(B)$ gilt.

Aus der Definition der bedingten Wahrscheinlichkeit erhält man die für die gesamte Analyse von Wahrscheinlichkeiten grundlegende mathematische Definition der (stochastischen) Unabhängigkeit zweier Ereignisse.

Definition 26
Zwei Ereignisse A und B, für die $P(A \cap B) = P(A) \cdot P(B)$ gilt, heißen **stochastisch unabhängig**.

Diese Definition hat dabei folgenden äquivalenten Ausdruck durch

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

Unvereinbare Ereignisse, für die $A \cap B = \emptyset$ gilt (vgl. Kap. 5.1.2), sind nie stochastisch unabhängig, wenn $P(A) > 0$ und $P(B) > 0$ gilt:

$$A \cap B = \emptyset \Rightarrow P(A \cap B) = 0 \neq P(A) \cdot P(B)$$

Analog zur stochastischen Unabhängigkeit zweier *Ereignisse* kann man auch von der stochastischen Unabhängigkeit zweier *zufälliger Vorgänge* sprechen, wenn alle Paare von Ereignissen aus den beiden zufälligen Vorgängen stochastisch unabhängig sind. Betrachtet man die Verkettung zweier zufälliger Vorgänge zu den Ereignissen $A_1, A_2, ..., A_n$ und $B_1, B_2, ..., B_m$, so muss also $P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$ gelten für alle (i, j) mit $i = 1, ..., n$ und $j = 1, ..., m$.

Vergleichen wir die Definition der stochastischen Unabhängigkeit mit der Untersuchung realer Daten, so erkennt man, dass die stochastische Unabhängigkeit eine Modellvorstellung ist, die sich nicht unbedingt empirisch zeigen muss.

[1] Mit dieser Definition besteht ein Pendant zur bedingte Häufigkeit (vgl. Kap. 3). Wie auch im Bereich der Datenanalyse kann und wird die Bedingung teilweise als Index notiert, also $P_B(A) := P(A|B)$.

Beispiel:
Wir betrachten dazu den doppelten Münzwurf. Im Modell wird man (zumindest weitgehend) zustimmen, dass der zweite Münzwurf nicht vom ersten Münzwurf beeinflusst wird und die beiden Münzwürfe oder zufälligen Vorgänge also stochastisch unabhängig sind.[2] Da der Münzwurf als Laplace-Experiment modelliert werden kann (Prinzip des unzureichenden Grundes), gilt in diesem Modell für die Wahrscheinlichkeit des Ereignisses W: „es fällt Wappen“ $P(W) = \frac{1}{2}$. Das ist sowohl im ersten als auch im zweiten Münzwurf der Fall. Bezeichnet man den ersten und zweiten Wurf durch einen Index, so gilt also $P(W_1) = P(W_2) = \frac{1}{2}$ und $P(W_1 \cap W_2) = P(W_1) \cdot P(W_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.
Wir betrachten nun einerseits das Modell zum doppelten Münzwurf sowie das Ergebnis eines empirischen Experiments mit 1000 doppelten Münzwürfen im Einheitsquadrat (Abb. 6.2), wobei das Einheitsquadrat für das empirische Experiment auf folgenden numerischen Ergebnissen beruht[3]:

	W_1	Z_1	Summe
W_2	254 (0,254)	257 (0,257)	511 (0,511)
Z_2	230 (0,230)	259 (0,259)	489 (0,489)
Summe	484 (0,484)	516 (0,516)	1000 (1)

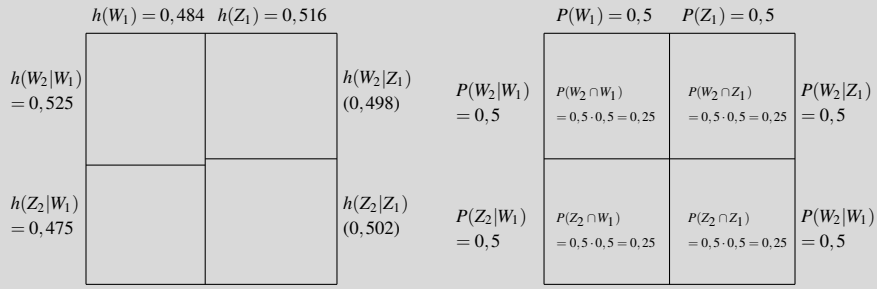


Abbildung 6.2: Einheitsquadrate empirisch (links) und Modell (rechts)

Die im Modell (sinnvoll) angenommene stochastische Unabhängigkeit der beiden Münzwürfe wird durch die „durchgezogene“ waagerechte Unterteilung des Einheitsquadrats repräsentiert. Diese ist so im empirischen Einheitsquadrat nicht vorhanden. Das Modell der stochastischen Unabhängigkeit zeigt sich empirisch lediglich durch ein geringes Assoziationsmaß (vgl. Kap. 3.1) von $A \approx 0,03$.

[2] Man muss dazu akzeptieren, dass die Münze kein Gedächtnis hat, sich also nicht sagt, „ich weiß, der erste Wurf hat Wappen gezeigt, also möchte ich das durch eine Zahl im nächsten Wurf ausgleichen“. Dass die Negierung dieses Gedächtnisses von Objekten nicht Allgemeingut ist, zeigt eine häufig anzutreffende Fehlvorstellung, die sogenannte *gamblers fallacy*, die Spieler dazu bringt, nach einer Serie von „Rot“ im Roulette daran zu glauben, dass „Schwarz“ im nächsten Durchgang bevorzugt wäre. Auch hier wird der Roulette-Kugel implizit ein Gedächtnis zugeschrieben. Wir werden den Unterschied zwischen der Betrachtung eines Einzelversuchs und einer Serie von Versuchen, die zu dieser Fehleinschätzung führt, in einer Aufgabe in Kapitel 7 noch eimal thematisieren.

[3] Entsprechend zum Ereignis W ist das Ereignis „es fällt Zahl“ mit Z beschrieben, die Indizes sind ebenfalls entsprechend verwendet.

Beispiel:
Modelliert man dagegen die Eigenschaften Geschlecht (mit den Ereignissen M: männlich und W: weiblich) sowie der Rauchgewohnheit (mit R: Raucher und $\overline{R}$: Nichtraucher) von Studierenden als zufällige Vorgänge (vgl. Kap. 3.1), so bietet sich das Modell der stochastischen Unabhängigkeit aufgrund der empirischen Ergebnisse nicht an.

	M	W	Summe
R	$h_{218}(M,R) = 0,12$	$h_{218}(W,R) = 0,16$	$h_{218}(R) = 0,28$
$\overline{R}$	$h_{218}(M,\overline{R}) = 0,17$	$h_{218}(W,\overline{R}) = 0,55$	$h_{218}(\overline{R}) = 0,72$
Summe	$h_{218}(M) = 0,29$	$h_{218}(W) = 0,71$	1

Hier ließe sich durch Schätzungen von Wahrscheinlichkeiten folgendes, auf die PH Freiburg bezogenes Modell aufstellen:

$$P(M) = 0,29 \qquad P(W) = 0,71$$
$$P(R|M) = \frac{P(M \cap R)}{P(M)} = 0,41 \qquad P(R|W) = \frac{P(W \cap R)}{P(W)} = 0,22$$
$$P(\overline{R}|M) = \frac{P(M \cap \overline{R})}{P(M)} = 0,59 \qquad P(\overline{R}|W) = \frac{P(W \cap \overline{R})}{P(W)} = 0,78$$

In diesem Modell *begünstigt* das Ereignis M (männlich) das Ereignis R (Raucher) oder, anders gesagt, das Merkmal Geschlecht beeinflusst in diesem Modell das Merkmal Rauchverhalten.

Bei der Beschreibung von mehrstufigen zufälligen Vorgängen wollen wir noch einmal zwischen Ergebnissen und Ereignissen unterscheiden:

- Im Beispiel des doppelten Münzwurfs sind die Ergebnisse Paare von Ergebnissen eines einfachen Münzwurfs. Die Ergebnismenge ist durch $\Omega = \{(W,W),(W,Z),(Z,W),(Z,Z)\}$ festgelegt, ein Ergebnis wäre etwa $\omega = (W,W)$.
- Diese oben diskutierten Ereignisse sind Teilmengen von Ω, also z.B. ist W_1: Wappen im ersten Versuch durch $W_1 = \{(W,W),(W,Z)\} \subseteq \Omega$ oder W_2: Wappen im zweiten Versuch durch $W_2 = \{(W,W),(Z,W)\} \subseteq \Omega$ festgelegt.

6.2 Visualisierung mehrstufiger zufälliger Vorgänge

Eine Möglichkeit, mehrstufige zufällige Vorgänge zu visualisieren, haben wir mit dem Einheitsquadrat bereits kennengelernt. Dieses ist besonders für zweistufige Vorgänge und eine nicht zu hohe Anzahl von Ereignissen auf jeder Stufe geeignet (in den beiden Beispielen im vorangehenden Abschnitt sind auf jeder Stufe nur jeweils zwei Ereignisse, z.B. Wappen und Zahl, betrachtet worden).

Das **Baumdiagramm** ist eine weitere, flexibel einsetzbare Visualisierung mehrstufiger zufälliger Vorgänge. Das Baumdiagramm fokussiert insbesondere den Ablauf solcher Vorgänge. Wir unterscheiden dabei im Folgenden zwei Fälle:

- das Baumdiagramm für den allgemeinen Fall von nicht notwendig (stochastisch) unabhängigen zufälligen Vorgängen
- sowie das Baumdiagramm für den Sonderfall von (stochastisch) unabhängigen zufälligen Vorgängen.

Als Beispiel nicht unabhängiger zufälliger Vorgänge betrachten wir ein Standardbeispiel, nämlich das dreifache Ziehen aus einer Urne mit 12 Kugeln, von denen 5 rot (R), 4 blau (B) und 3 grün (G) sind. Nach dem Ziehen in diesem Modell wird die Kugel nicht zurückgelegt. Das Modell umfasst zudem die Annahme, dass die Kugeln so durchmischt sind, dass der Zug tatsächlich zufällig erfolgt, also alle Kugeln mit der gleichen Wahrscheinlichkeit gezogen werden.

Das Baumdiagramm wird nun so konstruiert, dass

1. die *Ebenen* des Baums einen der zufälligen Teilvorgänge, hier also einen Zug, repräsentieren und
2. die *Äste* in jeder Ebene des Baumes disjunkte Ereignisse repräsentieren, hier den Zug einer der Farben Rot, Blau oder Grün.

Dadurch repräsentieren die verschiedenen *Pfade*, bestehend aus mehreren hintereinander hängenden Ästen, wiederum disjunkte Schnittereignisse. Wir betrachten in unserem Beispiel das (Schnitt-)Ereignis $R_1 \cap G_2 \cap B_3$. Das heißt, im ersten Zug wird eine rote, im zweiten eine grüne und im dritten eine blaue Kugel gezogen. Der hier nur partiell für das genannte Schnittereignis gezeichnete Baum hat die in Abbildung 6.3 zu sehende Gestalt.

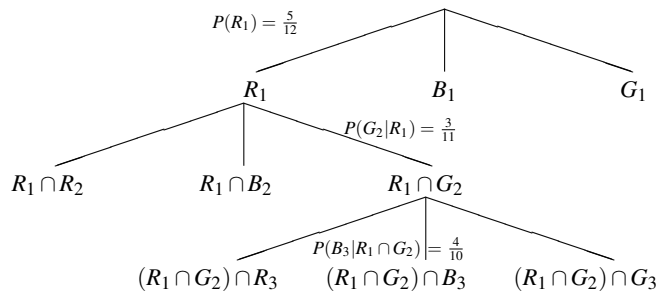


Abbildung 6.3: Baumdiagramm zum dreifachen Zug einer Kugel (ohne Zurücklegen)

Dieses Ziehen ohne Zurücklegen ist stochastisch abhängig, da das Ziehen einer grünen Kugel durch das vorherige Ziehen etwa einer roten Kugel *begünstigt* wird. Entsprechend begünstigt der Zug einer roten und einer grünen im ersten bzw. zweiten Zug das Ziehen einer blauen Kugel im dritten Zug: $P(B_1) = \frac{4}{12} < P(B_3|R_1 \cap G_2) = \frac{4}{10}$. Formal stehen an den Ästen eines solchen allgemeinen Baums bedingte Wahrscheinlichkeiten, wobei das bedingende Ereignis durch den vorangegangenen Pfad (der allgemein ein Schnittereignis repräsentiert) abgebildet wird.

Ändert man das Beispiel des Kugelziehens so ab, dass nach dem Zug die Kugel zurückgelegt wird, so erhält man ein Standardbeispiel eines zufälligen Vorgangs mit stochastisch unabhängigen Ereignissen längs jedes Pfades in einem Baum (bzw. stochastisch unabhängigen zufälligen Teilvorgängen, hier Zügen aus der Urne), da in diesem Modell der zufällige Vorgang mit exakt

gleichen Bedingungen wiederholt wird. Mit der Annahme der stochastischen Unabhängigkeit gilt bei der Betrachtung der gleichen Zugfolge rot-grün-blau etwa für den zweiten Zug:

$$P(G_2|R_1) = \frac{P(R_1 \cap G_2)}{P(R_1)} \underset{\text{Unabhängigkeit}}{\overset{\text{stochastische}}{=}} \frac{P(R_1) \cdot P(G_2)}{P(R_1)} = P(G_2) \ \ (= P(G_1))$$

Das Ergebnis lässt sich auf beliebig viele weitere folgende Äste eine Pfades erweitern, so dass auch im Diagramm die Notation der Wahrscheinlichkeiten an den einzelnen Ästen dadurch verkürzt werden kann, dass das jeweils bedingende Ereignis, weil es keine Auswirkung auf die folgenden Ereignisse hat, weggelassen werden kann. Das Baumdiagramm erhält damit für die stochastisch unabhängigen Ereignisse des dreifachen Zugs mit Zurücklegen die in Abbildung 6.4 zu sehende Gestalt. In diesem Baumdiagramm haben wir zudem Vereinfachung vorgenommen, indem wir an die Enden der Äste nur das Ereignis eines Astes anstatt das Ereignis des Pfades notiert haben.

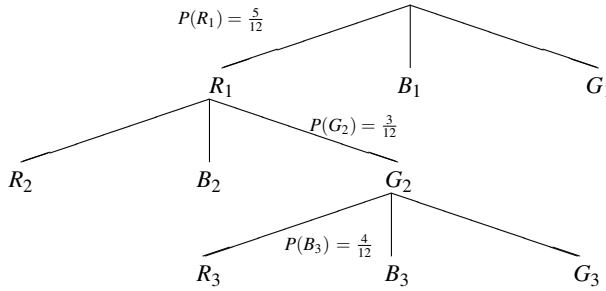


Abbildung 6.4: Baumdiagramm zum dreifachen Zug einer Kugel (mit Zurücklegen)

Für die Pfade (in beiden Beispielen bestehen diese aus drei aneinander gereihten Ästen) gilt nach dem Multiplikationssatz die sogenannte **Pfadmultiplikationsregel**, nach der die Wahrscheinlichkeit eines Ereignisses eines mehrstufigen zufälligen Vorgangs durch die Multiplikation der Wahrscheinlichkeiten längs dieses Pfades bestimmt wird. In unserem Beispiel gilt bei zweimaliger Anwendung dieser Multiplikationsregel z.B.:

$$P((R_1 \cap G_2) \cap B_3) = P(B_3|R_1 \cap G_2) \cdot P(R_1 \cap G_2) = P(B_3|R_1 \cap G_2) \cdot P(G_2|R_1) \cdot P(R_1)$$

Diese Pfadmultplikationsregel ließe sich analog auch für mehr als drei zufällige Vorgänge erweitern.

Da verschiedene Pfade disjunkte Ereignisse repräsentieren, lässt sich aus dem Baum auch die sogenannte **Pfadadditionsregel** ableiten, nach der die durch Pfadmultiplikation der Pfade bestimmten Pfadwahrscheinlichkeiten addiert werden können (3. Axiom nach Kolmogoroff, vgl. Kap. 5.2.3). Es gilt also etwa:

$$P((R_1 \cap G_2 \cap B_3) \cup (R_1 \cap B_2 \cap G_3)) = P(R_1 \cap G_2 \cap B_3) + P(R_1 \cap B_2 \cap G_3)$$

Für den Fall, dass die Kugel nach dem Zug zurückgelegt wird und damit die Ereignisse in den einzelnen Ebenen des Baums paarweise stochastisch unabhängig sind, vereinfacht sich die Pfadmultiplikationsregel zu

$$P((R_1 \cap G_2) \cap B_3) = P(R_1) \cdot P(G_2) \cdot P(B_3)$$

Eine Anwendung dieser grafischen Darstellung auf das Beispiel zu den Eigenschaften von Studierenden nehmen wir in Kapitel 6.5 auf. Mit dem Wissen um Pfad- und Multiplikationsregel betrachten wir noch einmal die *Begünstigung* des Ziehens einer grünen Kugel mit der Information, dass eine andere Kugel gezogen wurde. Bei der Betrachtung des Ziehens ohne Zurücklegen (Abb. 6.3) erhalten wir $P(G_1) = \frac{3}{12} < P(G_2|R_1) = \frac{3}{11}$. Es gilt aber auch:

$$\begin{aligned} P(G_2) &= P(R_1) \cdot P(G_2|R_1) + P(G_1) \cdot P(G_2|G_1) + P(B_1) \cdot P(G_2|B_1) \\ &= \frac{5}{12} \cdot \frac{3}{11} + \frac{3}{12} \cdot \frac{2}{11} + \frac{4}{12} \cdot \frac{3}{11} = \frac{33}{11 \cdot 12} = \frac{3}{12} \end{aligned}$$

Die Wahrscheinlichkeit, dass eine grüne Kugel im ersten oder zweiten Zug gezogen wird, ist also gleich. Damit gilt aber auch $P(G_2) < P(G_2|R_1)$, die Information erhöht in diesem Fall die Wahrscheinlichkeit für eine grüne Kugel. Die Begünstigung ist unabhängig davon, in welchem Zug sie erfolgt. So wirkt sich etwa die Information, dass im zweiten Zug eine rote Kugel gezogen wird, auf die Wahrscheinlichkeit einer grünen Kugel im ersten Wurf aus:

$$P(G_1|R_2) = \frac{P(G_1 \cap R_2)}{P(R_2)} = \frac{P(G_1 \cap R_2)}{P(R_1)} = \frac{\frac{3}{12} \cdot \frac{5}{11}}{\frac{5}{12}} = \frac{3}{11} > P(G_1)$$

Die Ergebnisse lassen sich analog erweitern etwa auf die Wahrscheinlichkeit, eine grüne Kugel im i-ten Zug zu ziehen ($i = 1,...,n$, n sei die Anzahl der Kugeln in der Urne), mit oder ohne die Information, dass im j-ten Zug ($j = 1,...,n;\ i \neq j$) eine Kugel anderer Farbe gezogen wird. Insbesondere den Gewinn durch das Verarbeiten von Informationen unabhängig von der Chronologie der Ereignisse untersuchen wir in den folgenden Abschnitten weiter.

6.3 Satz von Bayes und subjektivistischer Wahrscheinlichkeitsbegriff

6.3.1 Satz von Bayes

Wir betrachten in diesem Abschnitt insbesondere solche Verkettungen zufälliger Vorgänge, bei denen die einzelnen Vorgänge inhaltlich unterschiedlich sind, wie etwa die Ereignisse M (männlich) und W (weiblich) hinsichtlich des Merkmals Geschlecht einerseits und R (Raucher) und $\overline{R}$ (Nichtraucher) andererseits.

Auf der Basis der Erhebungen in Freiburg und Münster setzen wir folgendes Modell (für diese beiden Hochschulen) fest:

- $P(M) = 0,39; \quad P(W) = 0,61$
- $P(R|M) = 0,32;\ P(\overline{R}|M) = 0,68;\ P(R|W) = 0,23;\ P(\overline{R}|W) = 0,77$

Im Einheitsquadrat (Abb. 6.5) sind die bedingten Wahrscheinlichkeiten angegeben, die das Geschlecht als Bedingung haben. Nicht angegeben sind dagegen die Wahrscheinlichkeiten, in denen das Rauchverhalten die Bedingung ist, z.B. $P(M|R) = \frac{P(R\cap M)}{P(R)}$. Dieser *Rückschluss* wird allgemein durch den Satz von Bayes gesichert, den wir zunächst anhand des Einheitsquadrates grafisch anschaulich machen und anschließend formalisieren wollen.

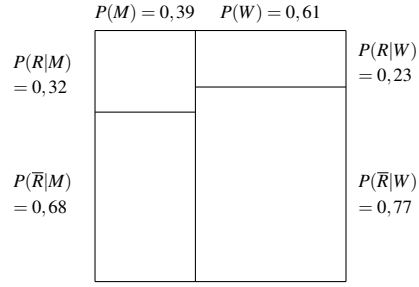


Abbildung 6.5: Geschlecht und Rauchverhalten im Modell

Um $P(M|R)$ bestimmen zu können, benötigt man die Werte von $P(R\cap M)$ und $P(R)$. Als Inhalt der linken oberen Fläche ist $P(R\cap M)$ unmittelbar gegeben:

$$P(R\cap M) = P(R|M)\cdot P(M)$$

Weiterhin sind anschaulich die Raucher durch beide obere Teilrechtecke innerhalb des Einheitsquadrats repräsentiert, also

$$P(R) = P(R\cap M) + P(R\cap W) = P(R|M)\cdot P(M) + P(R|W)\cdot P(W)$$

Die Zusammensetzung von $P(R)$ aus den im Einheitsquadrat sichtbaren Wahrscheinlichkeiten ist eine einfache Variante des sogenannten Satzes von der **totalen Wahrscheinlichkeit**. Mit den zuvor genannten Überlegungen ergibt sich (als einfache Version der Formel von Bayes):

$$\begin{aligned} P(M|R) &= \frac{P(R\cap M)}{P(R)} = \frac{P(R|M)\cdot P(M)}{P(R|M)\cdot P(M) + P(R|W)\cdot P(W)} = \frac{0,13}{0,27} \approx 0,47 \\ &= \frac{\text{Flächeninhalt des linken oberen Rechtecks}}{\text{Flächeninhalt beider oberen Rechtecke}} \end{aligned}$$

Bevor wir auf das Beispiel selber weiter eingehen, werden wir den Satz von der totalen Wahrscheinlichkeit wie auch den Satz von Bayes allgemein formulieren und beweisen.

Satz 11 (Satz von der totalen Wahrscheinlichkeit):
Seien $A_i,\ i = 1,\ldots,n$ paarweise disjunkte Teilmengen von Ω ($A_i \subseteq \Omega$) mit $A_i \cap A_j = \emptyset$ für $i \neq j$ und gelte $\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$, dann gilt für jedes Ereignis $B \subseteq \Omega$:

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

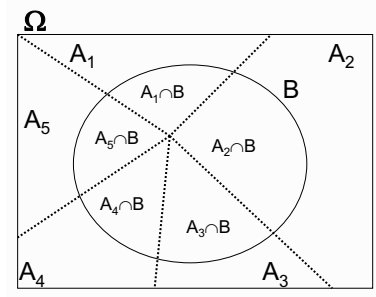


Abbildung 6.6: Totale Wahrscheinlichkeit

Wir unterteilen Ω in n (bzw. exemplarisch in 5) nichtleere disjunkte Ereignisse $A_1,...,A_n$ ($A_1,...A_5$), deren Vereinigung Ω ist, und betrachten zusätzlich ein beliebiges Ereignis $B \subseteq \Omega$. Wie in Abbildung 6.6 exemplarisch für $n = 5$ skizziert, gilt verallgemeinert stets:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup (B \cap A_3) \cup ...(B \cap A_n)$$

Daraus folgt, weil die Schnittereignisse paarweise disjunkt sind (3. Axiom nach Kolmogoroff, vgl. Kap. 5.2.3):

$$\begin{aligned} P(B) &= P(A_1) \cdot P(B|A_1) + P(A_2) \cdot P(B|A_2) + P(A_3) \cdot P(B|A_3) + ... + P(A_n) \cdot P(B|A_n) \\ &= \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i) \end{aligned}$$

Damit erhält man:

Satz 12 (Satz von Bayes):
Seien $A_i, i \in \mathbb{N}$ und B Ereignisse, die den Voraussetzungen des Satzes von der totalen Wahrscheinlichkeit genügen, dann gilt:

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

Wir gehen beim Beweis von der Definition bedingter Wahrscheinlichkeiten aus:

$$\begin{aligned} P(A_i \cap B) &= P(B \cap A_i) = P(B|A_i) \cdot P(A_i) \\ P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{P(B)} \end{aligned}$$

Betrachtet man das Beispiel des Geschlechts und des Rauchverhaltens von Studierenden, so braucht man keinen Satz, um die Wahrscheinlichkeit für das Ereignis männlich unter der Bedingung Raucher $P(M|R)$ zu bestimmen, da die Daten selbst ein Modell für diese Wahrscheinlichkeit liefern. Eine neue Bedeutung erhält aber der Satz von Bayes dann, wenn solch eine Wahrscheinlichkeit nicht aus den Daten selbst bestimmt werden kann. Diesen Fall untersuchen wir im Zusammenhang mit dem subjektivistischen Wahrscheinlichkeitsbegriff, in dem auf der Basis von zusätzlichen Informationen Gewissheiten mathematisch begründet verändert werden.

6.3.2 Subjektivistischer Wahrscheinlichkeitsbegriff

Der subjektivistische Wahrscheinlichkeitsbegriff weist fundamentale Unterschiede zu den bisher betrachteten Wahrscheinlichkeitsbegriffen auf. So geht es hier nicht primär um eine Aussage auf lange Sicht, sondern um eine Entscheidungsfindung in einer *einzelnen* Situation. Diese Entscheidung unterstützt dabei eine von mehreren Hypothesen, denen objektiv gesehen gar keine Wahrscheinlichkeit begründet zugeordnet werden kann. Wir machen das an einem Beispiel zu den Studierenden fest.

Angenommen, es ist das Profil eines Studierenden gegeben, aber ohne Geschlechtsangabe. Man soll aber entscheiden, ob es sich um einen Studenten oder eine Studentin handelt. Vor dem Ansehen des Profils, also der gegebenen Informationen, könnte man sagen: „Ich weiß nicht, ob es sich um eine Studentin oder einen Studenten handelt, ich würde persönlich beiden möglichen Hypothesen aufgrund des Prinzips des unzureichenden Grundes die Wahrscheinlichkeit $0,5$ geben, also $P(M) = P(W) = 0,5$“. Wenn man aber wüsste, dass in einer Hochschule mehr Studentinnen als Studenten sind, würde man den *subjektiven* Gewissheitsgrad vermutlich verändern, etwa in $P(W) = 0,7; P(M) = 0,3$ (oder auch anders mit $P(W) > P(M)$ und $P(W) + P(M) = 1$). Beides ist aber im eigentlichen Sinne keine „objektive“ Wahrscheinlichkeit, da in dem betrachteten *Einzelfall* de facto feststeht, ob es sich um einen Studenten oder eine Studentin handelt. Es ist zwar nicht bekannt, dennoch ist tatsächlich entweder $P(M) = 1$ oder $P(M) = 0$, je nachdem, ob hinter dem einzelnen Profil ein Student steht oder nicht. Wir bleiben aber bei den subjektiven Gewissheitsgraden und lesen im Profil:

- Information 1: es handelt sich um eine Person mit Körpergröße über 1,80 m.

Man weiß aus Erfahrung, dass Studentinnen seltener diese Körpergröße erreichen als Studenten, d. h. man hat ein Indiz dafür, dass hinter dem Profil ein Student steht. Man kann demnach die erste Einschätzung zugunsten der Hypothese männlich verändern.

- Information 2: die Person raucht.

Man weiß aus der Erhebung, dass Studenten eher rauchen als Studentinnen, wieder ein Indiz für einen Studenten. Weiter erfahren wir, dass die Person sich für Fußball interessiert, gerne Bier trinkt, gerne Dinge repariert, aber nicht gerne einen Einkaufsbummel durch die Stadt macht ... Würde man (auch wenn diese Informationen nur Klischees bedienen) jetzt $P(M) = P(W) = 0,5$ noch aufrechterhalten oder sich dafür *entscheiden*, dass das Profil eines Studenten vorliegt?

Weniger Klischee-belastet kann der subjektivistische Ansatz charakterisiert werden. Ausgehend von einer zunächst *subjektiven* Gewissheit, einer **a-priori-Wahrscheinlichkeit**, werden schrittweise zusätzliche Informationen (mit der Formel von Bayes) zu einer **a-posteriori-Wahrscheinlichkeit** verarbeitet. Ein Beispiel mit realem Bezug ist das folgende zur Mammographie (vgl. Hoffrage, 2003), das wir ausführlich behandeln.

Bezogen auf den Themenkomplex Brustkrebs und Mammographie ist (Stand 2003, vgl. ebd.) Folgendes bekannt: Etwa 1 % der Frauen im Alter von etwa 50 Jahren (frequentistische Wahrscheinlichkeit) leiden, ohne ein Symptom einer Erkrankung registriert zu haben, an Brustkrebs ($P(K) = 0,01$). Eine mögliche Vorsorge-Untersuchung durch Mammographie hat folgende Eigenschaften:

- Wenn eine Patientin krank ist (K), dann wird die Mammographie in 80% der Fälle tatsächlich die Krebsdiagnose (D) erbringen ($P(D|K) = 0,8$). Demnach schlägt die Mammographie fälschlich in 20% der Fälle bei kranken Patientinnen nicht an ($P(\overline{D}|K) = 0,2$).
- Wenn eine Patientin gesund ist ($\overline{K}$), dann schlägt die Mammographie dennoch fälschlich in 10% der Fälle an (und erzeugt eine Krebsdiagnose; $P(D|\overline{K}) = 0,1$). In 90% der Fälle erhält eine gesunde Patientin nach der Mammographie dagegen korrekterweise die Diagnose: kein Krebs ($P(\overline{D}|\overline{K}) = 0,9$).

Mit Wissen um das *Risiko*, auch ohne Symptome an Brustkrebs leiden zu können, könnte eine behandelnde Ärztin/ein behandelnder Arzt (in diesem Einzelfall) vor der Mammographie die subjektive Wahrscheinlichkeit $P(K) = 0,01$; $P(\overline{K}) = 0,99$ haben. Das ist im objektiven Sinne keine Wahrscheinlichkeit, da diese Patientin entweder definitiv Krebs hat oder eben (glücklicherweise) nicht. Die/der behandelnde Ärztin/Arzt erhält die Zusatzinformation *Mammographie ergibt Krebsdiagnose (D)*. Wie hoch ist die Wahrscheinlicheit $P(K|D)$, dass die Patientin tatsächlich Krebs hat?

Die Information lässt sich einerseits mit dem Einheitsquadrat, aber auch mit einem Baumdiagramm mit absoluten Häufigkeiten anschaulich verarbeiten. Dabei wird im Modell von einer bestimmten und fiktiven Grundmenge ausgegangen, so dass auf jeder Ebene des Baumes natürliche Zahlen als absolute Häugigkeiten eines Ereignisses erzeugt werden. In dem Mammographie-Beispiel kann man etwa von 1000 (virtuellen) Patientinnen ausgehen. Es ergibt sich dann bei der Verarbeitung der Informationen folgender Baum:

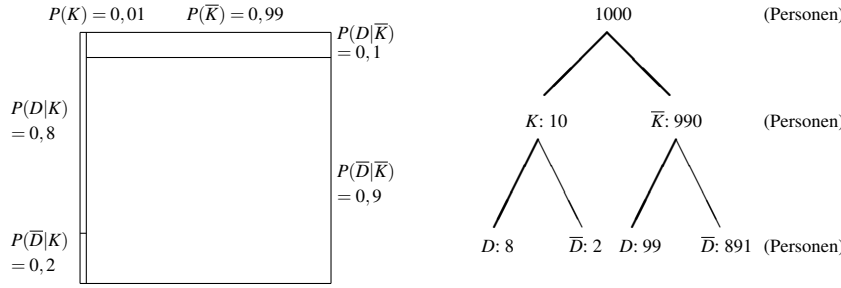


Abbildung 6.7: Einheitsquadrat und Baum mit absoluten Häufigkeiten (Bezugsgröße: 1000) zum Mammographie-Beispiel

Das Baumdiagramm mit absoluten Häufigkeiten ist geeignet, um sehr schnell eine Lösung zu erzielen. So haben in diesem Modell 107 Personen eine Krebsdiagnose, 8 davon sind tatsächlich krank, also gilt in diesem Modell $P(K|D) = \frac{8}{107} \approx 0,075$. Das ist ein überraschendes Ergebnis, da die Wahrscheinlichkeit, Krebs zu haben, trotz der „positiven“ Diagnose nur 7,5% beträgt. Tatsächlich hat sich allerdings die subjektivistische Wahrscheinlichkeit dieser Patientin, tatsächlich krank zu sein, von 1% auf 7,5% erhöht. Der Test *begünstigt* also die Gewissheit, krank zu sein: Die a-priori-Wahrscheinlichkeit beträgt 0,01, die a-posteriori-Wahrscheinlichkeit rund 0,075. Die Entstehung dieses dennoch nicht befriedigenden Ergebnisses des Tests kann, wie auch die Herleitung der Formel von Bayes, gut im Einheitsquadrat betrachtet werden.

Im Einheitsquadrat lassen sich (Abb. 6.7) die Formel von Bayes wie auch die Lösung für $P(K|D)$ durch den Vergleich der Flächeninhalte des linken oberen Teilrechtecks zu den beiden oberen Teilrechtecken bestimmen, formalisiert durch:

$$P(K|D) = \frac{P(D|K) \cdot P(K)}{P(D|K) \cdot P(K) + P(D|\overline{K}) \cdot P(\overline{K})} = \frac{0,8 \cdot 0,01}{0,8 \cdot 0,01 + 0,1 \cdot 0,99} \approx 0,075.$$

Erhöht man die Wahrscheinlichkeit $P(D|K)$, die sogenannten *Sensitivität* des Tests, so wird die Höhe des linken oberen Rechtecks größer, das Flächenverhältnis ändert sich dadurch aber kaum. Selbst wenn $P(D|K) = 1$ gelten würde, ergäbe sich nur $P(K|D) \approx 0,091$. Verkleinert man $P(D|\overline{K})$, die sogenannten Spezifität des Tests, so verringert sich die Höhe des rechten oberen Rechtecks und das Ergebnis des Tests wird zunehmend besser. Wäre etwa $P(D|\overline{K}) = 0,025$, so wäre $P(K|D) \approx 0,244$. Stark auswirken würde sich auch die nicht erwünschte Erhöhung des Anteils der Kranken, die sogenannte Basisrate. Wäre die Basisrate $P(K) = 0,1$, so ergäbe sich $P(K|D) \approx 0,471$. Die Tatsache, dass die geringe Aussagefähigkeit der Mammographie auf der glücklicherweise geringen Basisrate $P(K)$ basiert, bezeichnet man als *Phänomen der Diagnose seltener Ereignisse*.

Dieses besteht auch bei anderen seltenen Krankheiten, etwa Tests zur HIV-Infektion oder einem BSE-Test (bei Rindern). Im Gegensatz zur Diagnose von Brustkrebs gibt es etwa beim HIV-Test (vgl. auch Kap. 6.7) einen zweiten Test. Wie ein zweiter Test modelliert werden könnte, machen wir anhand der Mammographie deutlich (und gehen dabei zur Verdeutlichung aber bewusst zunächst einen Irrweg).

Die Mammographie wird nicht einmal, sondern zweimal hintereinander angewendet. Bezeichnet man mit D_1 und D_2 die positiven Diagnosen im ersten bzw. zweiten Test, so ergäbe sich:

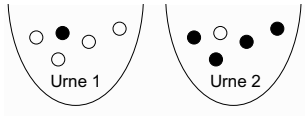
$$\begin{aligned} P(K|D_1 \cap D_2) &= \frac{P(K \cap D_1 \cap D_2)}{P(K \cap D_1 \cap D_2) + P(\overline{K} \cap D_1 \cap D_2)} \\ &= \frac{P(D_1 \cap D_2|K) \cdot P(K)}{P(D_1 \cap D_2|K) \cdot P(K) + P(D_1 \cap D_2|\overline{K}) \cdot P(\overline{K})} \\ &= \frac{0,8 \cdot 0,8 \cdot 0,01}{0,8 \cdot 0,8 \cdot 0,01 + 0,1 \cdot 0,1 \cdot 0,99} \approx 0,61 \end{aligned}$$

Das wäre ein deutlich verbessertes Ergebnis, birgt aber leider einen Fehler (!) in der Modellierung. Würde man von $P(D_1 \cap D_2|K) = 0,8 \cdot 0,8$ ausgehen, so würde man annehmen, dass das Ergebnis des ersten Tests keinen Einfluss auf das Ergebnis des zweiten hat, diese also stochastisch unabhängig sind. Das scheint aber keine sinnvolle Modellierung sein. So könnte man schließlich annehmen, dass ein Fehler im ersten Test genauso oder zumindest vermehrt auch im zweiten Test auftaucht. Eine rechnerische Kumulation von Informationen ist also nur dann problemlos, wenn die Informationen (als Ereignis interpretiert) stochastisch unabhängig sind.[4]

Ein konstruiertes Beispiel für stochastisch unabhängige Informationen ist ein Spiel mit zwei Urnen (Abb. 6.8). Man wählt eine Urne (ohne deren Inhalt zu sehen) zufällig aus und schätzt die Gewissheit ein, Urne 1 (U_1) oder Urne 2 (U_2) gewählt zu haben. Die Wahrscheinlichkeiten $P(S|U_1)$, $P(W|U_1)$, $P(S|U_2)$ und $P(W|U_2)$ sind durch die Wahrscheinlichkeitsverteilungen beider

[4] Die zwei HIV-Tests sind auch nicht als unabhängig modelliert, sondern es wird die Sensitivität und Spezifität der beiden Tests zusammen in einem Wert als frequentistische Wahrscheinlichkeit geschätzt.

Urnen gegeben. Legt man nach jedem Zug die Kugel der einmal ausgewählten Urne zurück, so sind die einzelnen Züge sinnvoll als stochastisch unabhängig zu modellieren.



Urne 1: Kugel	W	S
Wahrscheinlichkeit	0,8	0,2
Urne 2: Kugel	W	S
Wahrscheinlichkeit	0,2	0,8

Abbildung 6.8: Spiel mit zwei Urnen

Wir betrachten einmal die Verteilung der a-priori-Wahrscheinlichkeiten $P(U_1) = P(U_2) = 0,5$ und für eine zweite Serie die Verteilung der a-priori-Wahrscheinlichkeiten $\tilde{P}(U_1) = 0,05$ und $\tilde{P}(U_2) = 0,95$. Es wird eine schwarze Kugel (S) im ersten Versuch gezogen, was für die Urne 2 spricht. Je nach der Verteilung der a-priori-Wahrscheinlichkeiten neigt sich also die Entscheidung mehr oder weniger zu Urne 2:

$$\begin{aligned} P(U_1|S) &= \frac{P(S|U_1)\cdot P(U_1)}{P(S|U_1)\cdot P(U_1)+P(S|U_2)\cdot P(U_2)} \\ &= \frac{0,2\cdot 0,5}{0,2\cdot 0,5+0,8\cdot 0,5} = 0,2 \text{ bzw.} \\ \tilde{P}(U_1|S) &= \frac{0,2\cdot 0,05}{0,2\cdot 0,05+0,8\cdot 0,95} = 0,01 \end{aligned}$$

Die entsprechende Wahrscheinlichkeit für Urne 2 ist jeweils die Gegenwahrscheinlichkeit, also 0,8 bzw. 0,99. Da die Einzelzüge als stochastisch unabhängig modelliert werden können, ist etwa $P(S \cap S|U_1) = P(S|U_1)\cdot P(S|U_1) = 0,2\cdot 0,2 = 0,04$. So lässt sich eine Serie von 12 Informationen, wie z.B. S,S,S,W,W,W,S,W,W,W,S,W, wie folgt verarbeiten:

$$\begin{aligned} &P(U_1|S\cap S\cap S\cap W\cap W\cap W\cap S\cap W\cap W\cap W\cap S\cap W) = \\ &\frac{P(S\cap S\cap S\cap W\cap W\cap W\cap S\cap W\cap W\cap W\cap S\cap W|U_1)}{P(S\cap S\cap S\cap W\cap W\cap W\cap S\cap W\cap W\cap W\cap S\cap W|U_1)+P(S\cap S\cap S\cap W\cap W\cap W\cap S\cap W\cap W\cap W\cap S\cap W|U_2)} = \\ &\frac{0,8^7\cdot 0,2^5\cdot 0,5}{0,8^7\cdot 0,2^5\cdot 0,5+0,2^7\cdot 0,8^5\cdot 0,5} \approx 0,94 \end{aligned}$$

Die Entwicklung der Entscheidungsfindung bei 20 Ziehungen ist in der Abbildung 6.9 zu sehen. Hinsichtlich der eingehenden a-priori-Wahrscheinlichkeiten ist zu erkennen, dass auch die Verteilung $\tilde{P}(U_1) = 0,05;\ \tilde{P}(U_2) = 0,95$ ab einer bestimmten Versuchszahl von den Informationen überlagert wird und sich die Gewissheit der Entscheidung für eine der beiden Hypothesen stabilisiert.

Unbeeinflusst davon, ob die einzelnen Ereignisse (Informationen) stochastisch abhängig oder unabhängig sind, kann der subjektivistische Wahrscheinlichkeitsbegriff als mathematisches **Lernen aus Erfahrung** bezeichnet werden. So wird eine a-priori-Wahrscheinlichkeit (zu krank/nicht krank oder Urne 1/Urne 2) durch jede weitere Information revidiert und ermöglicht, wenn genug Informationen verarbeitet werden können, eine gute, datenbasierte Entscheidung.

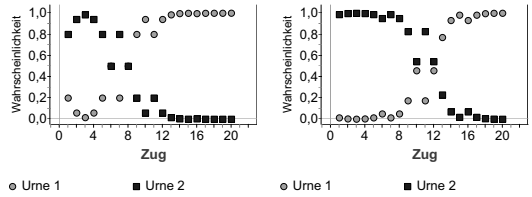


Abbildung 6.9: Serie $s,s,s,w,w,w,s,w,w,w,s,w,\ldots$, links mit $P(U_1) = P(U_2) = 0,5$ und rechts mit $\tilde{P}(U_1) = 0,05$ und $\tilde{P}(U_2) = 0,95$

6.4 Vom Baumdiagramm zu kombinatorischen Zählfiguren

„Stellen Sie den Baum zum 6-fachen Wurf eines Würfels oder zum Lotto 6 aus 49 dar." Diese Aufgabe ist zwar machbar, aber nicht sinnvoll. Überschlagen wir die Anzahl der notwendigen Äste, so haben wir beim einfachen Würfelwurf 6 Äste, beim zweifachen $6 \cdot 6 = 36$, beim dreifachen $6 \cdot 6 \cdot 6 = 216$ und schließlich beim 6-fachen Wurf $6^6 = 46656$. Hier hilft der Baum offensichtlich nicht mehr weiter. Je nach Fragestellung kann es möglich sein, den Baum selbst zu reduzieren. Wird etwa beim Würfeln nur nach den Ereignissen unterschieden, dass eine 6 fällt oder keine 6, so muss der Baum nicht mehr vollständig gezeichnet werden, sondern nur zwei Äste (das Ereignis $6 := \{6\}$ und $\overline{6} := \{1,2,3,4,5\}$. Aber auch hier wächst die Anzahl der Pfade exponentiell und bei 6 Würfen müsste man zumindest $2^6 = 64$ Äste in der letzten Ebene des Baumdiagramms zeichnen.

Beim Lotto wären es beim Zug der ersten Kugel 49 Äste, beim zweiten (da nicht zurückgelegt wird) $49 \cdot 48 = 2352$, beim dritten $49 \cdot 48 \cdot 47 = 110544$ und schließlich beim sechsten Zug rund 10 Milliarden Äste. Auch das ist nur noch theoretisch darstellbar. Kurz: Bei komplexeren zufälligen Vorgängen müssen Hilfsmittel eingesetzt werden, um ohne Zeichnung eines Baumdiagramms die Anzahl aller Pfade sowie die Anzahl der zu einem bestimmten Ereignis gehörenden Pfade zu zählen. Dieses Hilfsmittel stellt die Kombinatorik bereit, die wir in aller Kürze anhand des Ziehens von Kugeln aus einer Urne mit sechs von 1 bis 6 nummerierten Kugeln diskutieren (Abb. 6.10).

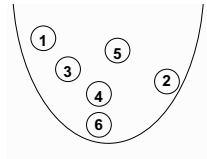


Abbildung 6.10: Ziehen aus einer Urne mit sechs von 1 bis 6 nummerierten Kugeln

Zählfigur mit Zurücklegen und mit Beachtung der Reihenfolge: Nach jedem Zug wird damit die Urne wieder in den Ausgangszustand mit den hier 6, allgemeiner n Kugeln versetzt. Zieht man 3, allgemeiner k Mal, so hat man im ersten Zug 6 (bzw. n) Möglichkeiten, im zweiten wieder 6 (bzw. n), also insgesamt $6 \cdot 6 = 6^2$ (bzw. $n \cdot n = n^2$) und schließlich im dritten Zug (k-ten Zug) 6^3 (bzw. n^k) verschiedene Möglichkeiten. Zusammengefasst ist die Anzahl von Möglichkeiten beim Ziehen mit Zurücklegen und mit Beachtung der Reihenfolge von k Elementen aus einer Menge mit n Elementen $A(n,k) = n^k$.

Zählfigur ohne Zurücklegen und mit Beachtung der Reihenfolge: Nach jedem Zug ist in der Urne mit den hier 6, allgemeiner n Kugeln eine Kugel weniger als vor dem Zug. Zieht man 3 Mal, allgemeiner k Mal, so hat man im ersten Zug 6 (bzw. n) Möglichkeiten, im zweiten nur noch 5 (bzw. $n-1$), also insgesamt $6\cdot 5$ (bzw. $n\cdot(n-1)$) und schließlich im dritten Zug (k-ten Zug) $6\cdot 5\cdot 4 = 6\cdot(6-1)\cdot(6-2) = 6\cdot(6-1)\cdot((6-3)+1)$, allgemein $n\cdot(n-1)\cdot\ldots\cdot((n-k)+1)$ verschiedene Möglichkeiten. Zusammengefasst ist die Anzahl von Möglichkeiten beim Ziehen ohne Zurücklegen und mit Beachtung der Reihenfolge von k Elementen aus einer Menge mit n Elementen $A(n,k) = n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$. Fasst man das Produkt als Bruch mit Nenner 1 auf und erweitert man diesen Bruch durch $(n-k)\cdot(n-k-1)\cdot\ldots\cdot 2\cdot 1$, so erhält man als Anzahl

$$\begin{aligned} A(n,k) &= n\cdot(n-1)\cdot\ldots\cdot(n-k+1) = \\ &= \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)\cdot((n-k)\cdot(n-k-1)\cdot\ldots\cdot 2\cdot 1)}{(n-k)\cdot(n-k-1)\cdot\ldots\cdot 2\cdot 1} \\ &= \frac{n!}{(n-k)!} \end{aligned}$$

wobei $n! = n\cdot(n-1)\cdot\ldots\cdot 1$ ist.[5]

Zählfigur ohne Zurücklegen und ohne Beachtung der Reihenfolge: Nach jedem Zug ist damit in der Urne mit den hier 6, allgemeiner n Kugeln eine Kugel weniger als vor dem Zug. Würde man die Reihenfolge beachten, so hätte man (siehe oben) $\frac{6!}{3!}$ bzw. $\frac{n!}{(n-k)!}$ Möglichkeiten. Beachtet man diese Reihenfolge nicht mehr, so müsste man z.B. die Zugfolge der Kugeln 1-2-3 mit der Zugfolge 3-2-1 *identifizieren*, da in beiden Zugfolgen die gleichen Kugeln, nur in einer anderen Reihenfolge, gezogen wurden. Man kann sich nun fragen, wie viele verschiedene zu identifizierende Zugfolgen es mit 3 (bzw. k) verschiedenen Kugeln es gibt. Das haben wir mit der letzten Zählfigur ebenfalls beantwortet: Es gibt $\frac{3!}{(3-3)!} = 3!$ (bzw. $\frac{k!}{(k-k)!} = k!$) Möglichkeiten aus einer Menge mit 3 (bzw. k) Elementen 3 (bzw. k) Mal ohne Wiederholung mit Beachtung der Reihenfolge zu ziehen. Die Anzahl der $\frac{6!}{3!}$ bzw. $\frac{n!}{(n-k)!}$ muss also noch durch die 3! bzw. $k!$ zu identifizierenden Zugfolgen dividiert werden, wenn die Zugfolge keine Rolle spielen soll. Zusammengefasst ist damit die Anzahl von Möglichkeiten beim Zug ohne Zurücklegen und ohne Beachtung der Reihenfolge von k Elementen aus einer Menge mit n Elementen:

$$A(n,k) = \frac{\frac{n!}{(n-k)!}}{k!} = \frac{n!}{(n-k)!\cdot k!} =: \binom{n}{k}$$

$\binom{n}{k}$ wird als **Binomialkoeffizient** bezeichnet. Der Binomialkoeffizient gibt also die Anzahl der k-elementigen Teilmengen einer n-elementigen Menge an. Der Binomialkoeffizient wird bei der im folgenden Kapitel erfolgenden systematischen Zusammenfassung von mehrstufigen zufälligen Vorgängen zu bestimmten Wahrscheinlichkeitsverteilungen bedeutsam.

Zählfigur mit Zurücklegen und ohne Beachtung der Reihenfolge: Nur einer gewissen Vollständigkeit halber entwickeln wir auch die folgende vierte Zählfigur, die für alle weiteren Überlegungen unerheblich ist und deswegen als Exkurs betrachtet werden kann.

Mit dem Zurücklegen wird auch hier wie bei der ersten Zählfigur die Urne wieder in den Ausgangszustand zurückversetzt. Im Unterschied zur ersten Zählfigur wird allerdings nicht die

[5] Die Fakultät ist eigentlich für $n \in \mathbb{N}$ rekursiv definiert durch $n! = n\cdot(n-1)!$ und $0! = 1! = 1$.

Reihenfolge der gezogenen Kugeln berücksichtigt. Wie viele Möglichkeiten gibt es, ohne Beachtung der Reihenfolge 3 Mal (k Mal) aus der Urne mit 6 (allgemein n) unterscheidbaren Kugeln zu ziehen, wenn die Kugeln nach jedem Zug zurückgelegt werden?

Die Frage nach einer Notationsmöglichkeit führt auf die Berechnung: Wird z.B. im ersten Zug 4, dann 5, schließlich noch 1 gezogen, lässt sich dies notieren als: $|123|4|56$, wobei „|“ für „gezogen“ vor der entsprechenden Kugelnummer steht. Die Notierung $12|34|5|6$ steht also dafür, dass die 3, die 5 und die 6 gezogen wurden. Die Reihenfolge ist aus dieser Notation nicht rekonstruierbar, was aber in dieser Zählfigur auch nicht verlangt ist. Bei Wiederholungen, wie z. B. der Zugfolge $3-3-4$, ergibt sich eine Mehrfachbelegung bezüglich der Trennstriche: $12||3|456$ oder bei $2-2-2$ entsprechend $1|||23456$. Für einen der Trennstriche gibt es $6-1+3=8$ Positionen: Die eigentlich 9. Position, also hinter der 6, kann nicht durch einen Trennstrich eingenommen werden, da die Notation so definiert ist, dass die Stellung *vor* einer Zahl das Ziehen dieser Zahl repräsentiert. Mit dieser Notation ergibt sich die Frage: Wie viele Möglichkeiten gibt es, 3 (bzw. allgemein k) „|“ (die Reihenfolge wird nicht berücksichtigt) auf insgesamt $6+3-1=8$ (bzw. allgemein auf insgesamt $n+k-1$) Positionen zu verteilen? Das entspricht der Auswahl von 3 Elementen (bzw. k Elementen) aus einer Menge mit 8 Elementen (bzw. $(n+k-1)$ Elementen). Die entsprechende Anzahl der Möglichkeiten ist durch den Binomialkoeffizienten aus der vorausgehenden Zählfigur mit $\binom{6+3-1}{3} = \frac{(6+3-1)!}{(6-1)!\cdot 3!}$ gegeben.

Zusammengefasst ist die Anzahl von Möglichkeiten beim Ziehen mit Zurücklegen und ohne Beachtung der Reihenfolge von k Elementen aus einer Menge mit n Elementen:

$$A(n,k) = \binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!\cdot k!}$$

6.5 Eigenschaften von Studierenden

Erhebung einer Eigenschaft in einer Stichprobe Wir nehmen die Einstiegsfrage zu diesem Kapitel auf und betrachten zunächst im Modell eine kleine Stichprobe zu den Ereignissen M: „männlich“ und W: „weiblich“ mit den auf der Basis der empirischen Resultate festgelegten Wahrscheinlichkeiten (für die PH Freiburg, Erhebung 2010) von $P(W) = 0,71$ und $P(M) = 0,29$. Betrachtet man mit diesem Modell von Wahrscheinlichkeiten eine Stichprobe von drei Studierenden, so sind die Ereignisse der einzelnen Auswahlen eines Studierenden stochastisch abhängig: Betrifft die erste Auswahl eine Studentin, so begünstigt das, wenn angenommen wird, dass genau diese Studentin nicht noch einmal zu ihrem Geschlecht erhoben wird, die Auswahl eines Studenten im folgenden Zug. Wenn man vereinfacht von 4000 Studierenden an der PH Freiburg ausgeht und damit von $0,29 \cdot 4000 = 1160$ Studenten und $0,71 \cdot 4000 = 2840$ Studentinnen, so würde sich das Szenario ergeben, das in Abbildung 6.11 dargestellt ist (wieder wird durch einen Index die Nummer des zufälligen Teilvorgangs angegeben).

Betrachtet man zunächst allein die ersten beiden Ebenen des Baums, so würde sich also im Modell beispielsweise ergeben $P(W_2|W_1) \approx 0,7099 \approx P(W_1)$, die Wahrscheinlichkeit für das Ereignis W_2 ändert sich also nur unwesentlich gegenüber $P(W_1)$, wenn man die erste Studentin aus der stochastischen Urne entfernt hat. Wir verfolgen die Wahrscheinlichkeiten, eine Studentin zu erheben (W) unter der Bedingung, davor ebenfalls nur Studentinnen erhoben (gezogen) zu haben:

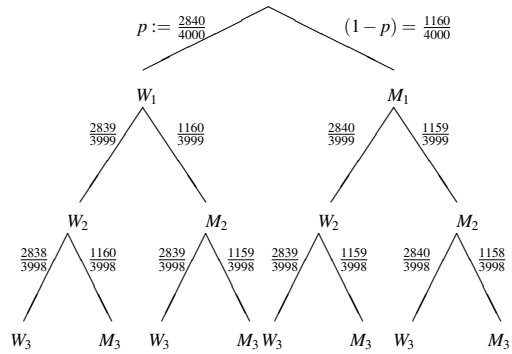


Abbildung 6.11: Baumdiagramm für drei Ebenen mit den Ereignissen M und W

Nummer des „Zugs“	Wahrscheinlichkeit
2	0,7099
10	0,7093
50	0,7063
100	0,7026
218	**0,6934**
500	0,6686
1000	0,6133

Zwar bildet das Urnen-Modell *ohne Zurücklegen* den Vorgang der Erhebung unter den Studierenden passender ab, offenbar lässt sich aber auch ohne erheblichen Unterschied das Modell *mit Zurücklegen* verwenden, das wegen der sich nicht ändernden Wahrscheinlichkeiten einfacher ist als das zuerst genannte. Dabei würde statt der sich ändernden Wahrscheinlichkeiten an den Ästen jeweils allein $P(W) = p$ und $P(M) = 1 - p$ gesetzt werden.

Im Vergleich ergibt sich, wenn man die Anzahl der Studierenden mit der Merkmalsausprägung „weiblich“ für die oben genannte Mini-Stichprobe von drei Studierenden analysiert:

Urnen-Modell mit Zurücklegen (stochastisch Unabhängigkeit)

Anzahl der Studentinnen	Wahrscheinlichkeit
0	$\left(\frac{1160}{4000}\right)^3 = \binom{3}{0} \cdot \left(\frac{1160}{4000}\right)^3 \cdot \left(\frac{2840}{4000}\right)^0 \approx 0,02$
1	$3 \cdot \left(\frac{1160}{4000}\right)^2 \cdot \left(\frac{2840}{4000}\right) = \binom{3}{1} \cdot \left(\frac{1160}{4000}\right)^2 \cdot \left(\frac{2840}{4000}\right)^1 \approx 0,18$
2	$3 \cdot \left(\frac{1160}{4000}\right) \cdot \left(\frac{2840}{4000}\right)^2 = \binom{3}{2} \cdot \left(\frac{1160}{4000}\right)^1 \cdot \left(\frac{2840}{4000}\right)^2 \approx 0,44$
3	$\left(\frac{2840}{4000}\right)^3 = \binom{3}{3} \cdot \left(\frac{1160}{4000}\right)^0 \cdot \left(\frac{2840}{4000}\right)^3 \approx 0,36$

Urnen-Modell ohne Zurücklegen (stochastische Abhängigkeit)

Anzahl der Studentinnen	Wahrscheinlichkeit
0	$\binom{3}{0} \cdot \left(\frac{1160 \cdot 1159 \cdot 1148}{4000 \cdot 3999 \cdot 3998}\right) \approx 0,02$
1	$\binom{3}{1} \cdot \left(\frac{1160 \cdot 1159 \cdot 2840}{4000 \cdot 3999 \cdot 3998}\right) \approx 0,18$
2	$\binom{3}{2} \cdot \left(\frac{1160 \cdot 2840 \cdot 2839}{4000 \cdot 3999 \cdot 3998}\right) \approx 0,44$
3	$\binom{3}{3} \cdot \left(\frac{2840 \cdot 2839 \cdot 2938}{4000 \cdot 3999 \cdot 3998}\right) \approx 0,36$

Gerundet auf zwei Stellen sind die Wahrscheinlichkeiten für beide Modelle identisch, wodurch die Modelle austauschbar werden. Das zweite Modell (ohne Zurücklegen) ist in seiner Berechnung etwas komplizierter als das erste Modell mit Zurücklegen (insbesondere, wenn die Größe der Stichprobe erhöht wird). In beiden Fällen ergibt sich die in Abbildung 6.12 zu sehende Wahrscheinlichkeitsverteilung.

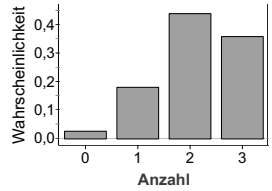


Abbildung 6.12: Wahrscheinlichkeitsverteilung für die Anzahl der weiblichen Studierenden in einer Stichprobe vom Umfang $n = 3$

Rückschluss mit der Formel von Bayes Wir nehmen das im Hauptteil behandelte Beispiel noch einmal auf. Tatsächlich haben wir in zwei Veranstaltungen insgesamt 40 Studentinnen und 40 Studenten zusätzlich zu den schon bekannten Eigenschaften zu folgenden Punkten gefragt:

- Sind Sie größer als 1,80 m (ja/nein; $G, \overline{G}$)?
- Rauchen Sie (ja/nein; $Z, \overline{Z}$)?
- Interessieren Sie sich für Fußball (ja/nein; $F, \overline{F}$)?
- Trinken Sie gerne Bier (ja/nein; $B, \overline{B}$)?
- Reparieren Sie in Ihrer Freizeit Dinge Ihres Haushalts selbst, z.B. Fahrrad, elektrische Geräte, Auto etc. (ja/nein; $R, \overline{R}$)?
- Machen Sie gerne einen Einkaufsbummel in der Innenstadt (ja/nein; $E, \overline{E}$)?

Wir betrachten hier nicht das Problem der Repräsentativität der Daten, sondern arbeiten mit den Ergebnissen der Umfrage so, als seien sie repräsentativ, um den Rückschlussgedanken der Formel von Bayes und das Problem der Verkettung von Informationen (Lernen aus Erfahrung) zu illustrieren. Das Geschlecht wurde zusätzlich auf dem Fragebogen angegeben.[6] Die einzelnen Merkmale führen (bis auf die Vorliebe für Bier) in der vorliegenden Stichprobe zu Festlegungen von durch das Geschlecht bedingten Wahrscheinlichkeiten:

[6] Der vollständige Datensatz ist in den Zusatzmaterialien unter den im Vorwort genannten Bezugsquellen zu erhalten.

Eigenschaft zu	Studentinnen (W)	Studenten (M)
Größe über 180 (G)	$P(G\|W) = 0,05$	$P(G\|M) = 0,70$
Raucher (Z)	$P(Z\|W) = 0,20$	$P(Z\|M) = 0,40$
Fußball (F)	$P(F\|W) = 0,18$	$P(F\|M) = 0,90$
Reparatur (R)	$P(R\|W) = 0,13$	$P(R\|M) = 0,40$
Einkaufsbummel (E)	$P(E\|W) = 0,85$	$P(E\|M) = 0,23$

Nimmt man ohne Wissen des Geschlechts eine einzelne dieser Informationen und bestimmt die subjektivistische Wahrscheinlichkeit dafür, einen Studenten bzw. eine Studentin vorliegen zu haben, die hinter dieser Information steht, so ergeben sich die folgenden Neueinschätzungen der a-priori-Wahrscheinlichkeiten $P(W) = P(M) = 0,5$:

Körpergröße $P(W|G) = \dfrac{0,05 \cdot 0,5}{0,05 \cdot 0,5 + 0,70 \cdot 0,5} \approx 0,07$ $P(M|G) = \dfrac{0,70 \cdot 0,5}{0,05 \cdot 0,5 + 0,70 \cdot 0,5} \approx 0,93$

Raucher $P(W|Z) = \dfrac{0,20 \cdot 0,5}{0,20 \cdot 0,5 + 0,40 \cdot 0,5} \approx 0,33$ $P(M|Z) = \dfrac{0,40 \cdot 0,5}{0,20 \cdot 0,5 + 0,40 \cdot 0,5} \approx 0,67$

Fußball $P(W|F) = \dfrac{0,18 \cdot 0,5}{0,18 \cdot 0,5 + 0,90 \cdot 0,5} \approx 0,16$ $P(M|F) = \dfrac{0,90 \cdot 0,5}{0,18 \cdot 0,5 + 0,90 \cdot 0,5} \approx 0,84$

Reparatur $P(W|R) = \dfrac{0,13 \cdot 0,5}{0,13 \cdot 0,5 + 0,40 \cdot 0,5} \approx 0,24$ $P(M|R) = \dfrac{0,40 \cdot 0,5}{0,13 \cdot 0,5 + 0,40 \cdot 0,5} \approx 0,76$

Einkauf $P(W|E) = \dfrac{0,85 \cdot 0,5}{0,85 \cdot 0,5 + 0,23 \cdot 0,5} \approx 0,79$ $P(M|E) = \dfrac{0,23 \cdot 0,5}{0,85 \cdot 0,5 + 0,23 \cdot 0,5} \approx 0,21$

Nimmt man etwa das Merkmal „Präferenz für Fußball“ heraus, so würde die Information *Fußballfan* (F) die a-priori-Wahrscheinlichkeit für eine männlichen Studenten (M) von $P(M) = 0,5$ auf die a-posteriori-Wahrscheinlichkeit $P(M|F) \approx 0,84$ anheben (begünstigen).

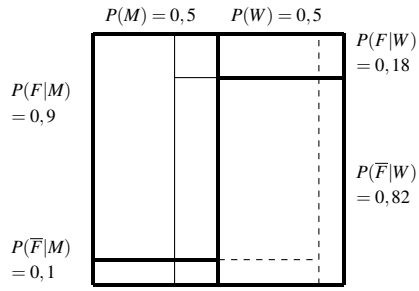


Abbildung 6.13: Auswirkung der Information F auf die Wahrscheinlichkeiten $P(M|F)$ und $P(W|F)$ im Einheitsquadrat. Die dünne Linie links gilt für die Übertragung auf die gesamte PH Freiburg, die gestrichelte Linie rechts für den Übertrag auf eine fiktive Gruppe von Elektrotechnik-Studierenden

Könnte man die Wahrscheinlichkeit auf den Datensatz der gesamten PH Freiburg mit der geschätzten a-priori-Wahrscheinlichkeit $P(M) = 0,30$ übertragen, so würde diese Wahrscheinlichkeit durch die Information F zur a-posteriori-Wahrscheinlichkeit $P(M|F) \approx 0,68$ angehoben werden, da der Anteil der männlichen Studierenden insgesamt geringer ist. Würde man dagegen die Wahrscheinlichkeit $P(F|M)$ auf eine Studierendengruppe aus einem Fachbereich Elektrotechnik mit $P(M) = 0,90$ beziehen, so würde sich $P(M|F) \approx 0,98$ ergeben (vgl. Abb. 6.13). Im Fall der geringen Basisrate erhöht die Information zwar relativ die a-posteriori-Wahrscheinlichkeit am deutlichsten, es ergibt sich aber insgesamt die unsicherste Aussage.

Bei dieser Aufstellung der Informationen und ihrer Wirkung ist unmittelbar einleuchtend, dass eine zweite Information stochastisch abhängig von der ersten ist. Beispielsweise sind die Ereignisse G (groß) und F (Fußballfan) beide durch das Ereignis M (männlich) stark begünstigt. Es ist weiter anzunehmen, dass auch das Ereignis G (groß) ebenso das Ereignis F (Fußballfan) begünstigt, und damit die Ereignisse stochastisch abhängig sind. Dadurch wäre $P(G \cap F|M) \neq P(G|M) \cdot P(F|M)$. Wir untersuchen das anhand der Stichprobe. Dort ergibt sich für die Zusammenfassungen der Informationen $G \cap F$ sowie dem Gegenteil $\overline{G \cap F}$:

	M	W	Summe
$G \cap F$	18	1	19
$\overline{G \cap F}$	22	39	61
Summe	40	40	80

Setzt man wiederum die Wahrscheinlichkeit $P(G \cap F|M) = 0,450$ unmittelbar aufgrund der Schätzung auf der Basis der empirischen Ergebnisse, so ergibt sich zunächst:

$$P(G \cap F|M) = 0,450 \neq 0,630 = 0,70 \cdot 0,90 = P(G|M) \cdot P(F|M)$$

Schränkt man also die Ereignisse auf das bedingende Ereignis M (männlich) ein, so sind G und F stochastisch abhängig. Wir können damit bei der Zusammenfassung der Informationen diese nicht wie beim Urnenspiel in Kapitel 6.3 als Einzelinformationen verarbeiten, sondern von vornherein zusammenfassen und erhalten:

$$P(M|G \cap F) = \frac{P(G \cap F|M) \cdot P(M)}{P(G \cap F|M) \cdot P(M) + P(G \cap F|W) \cdot P(W)} = \frac{0,450 \cdot 0,5}{0,450 \cdot 0,5 + 0,025 \cdot 0,5} \approx 0,947$$

Durch die stochastische Abhängigkeit von $G|M$ und $F|M$ ist damit der Informationsgehalt von $G \cap F$ wegen $P(G \cap F|M) = 0,450 < P(G|M) \cdot P(F|M) = 0,630$ geringer, als er bei der stochastischen Unabhängigkeit beider Ereignisse wäre.

Dennoch erhöht die zusammengesetzte Information die Gewissheit, einen Studenten vorliegen zu haben. Wenn man von einem geringeren Anteil an Studenten ausgehen würde ($P(M) = 0,3$), wäre die Verbindung beider Ereignisse die Gewissheit, dass aufgrund des Zutreffens von G und F ein Student vorliegen muss, $P(M|G \cap F) \approx 0,86$ schon recht hoch. Unabhängig von der a-priori-Verteilung verschaffen also auch stochastisch abhängige Informationen einen Zuwachs an Gewissheit, der allerdings nicht mehr einfach zu bestimmen ist.[7]

[7] In gleicher Weise müssten auch verschiedene Diagnosemethoden etwa bei Tests auf HIV-Infizierung oder auch auf Doping betrachtet werden.

6.6 Ergänzungen

Wir wollen die Betrachtung von mehrstufigen zufälligen Vorgängen sowie der stochastischen Unabhängigkeit bzw. Abhängigkeit auf Zufallsgrößen übertragen. Dazu betrachten wir zunächst das Beispiel des doppelten Wurfs eines Würfels.

Unterscheidet man durch einen Index die beiden Würfe, so ist $\Omega_1 = \Omega_2 = \{1,2,3,4,5,6\}$ und $\Omega = \Omega_1 \times \Omega_2 = \{(1,1),(1,2),...,(1,6),(2,1),(2,2),...,(6,6)\}$ die Ergebnismenge des zweifachen Wurfs.

	Y	y_1	y_2	y_3	y_4	y_5	y_6
X		1	2	3	4	5	6
x_1	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
x_2	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
x_3	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
x_4	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
x_5	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
x_6	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Definieren wir X als Augenzahl des ersten Wurfs mit den Werten $x_1 = 1, x_2 = 2..., x_6 = 6$ und Y als Augenzahl des zweiten Wurfs mit den Werten $y_1 = 1, y_2 = 2..., y_6 = 6$, so sind in der Tabelle oben gemeinsam die Ergebnisse ω des zweifachen Wurfs sowie die Paare (x_i, x_j) ($i = 1,...,6$ und $j = 1,...,6$) der zweidimensionalen Zufallsgröße (X,Y) eingetragen.

Für ein bestimmtes Ergebnis ω und die Wahrscheinlichkeit eines Elementarereignisses $\{\omega\}$ verwenden wir bezogen auf die Zufallsgröße (X,Y) folgende Notation:

$$\begin{aligned} (X,Y)(\omega) &:= (X(\omega), Y(\omega)) \\ P(X(\omega), Y(\omega)) &:= P(\{(X(\omega), Y(\omega))\}) \end{aligned}$$

Wir betrachten weiterhin die Notation der Wahrscheinlichkeit dafür, dass $X(\omega)$ und $Y(\omega)$ bestimmte Werte x_i und y_j annehmen ($i = 1,...,6$ und $j = 1,...,6$):

$$\begin{aligned} P(\{X = x_i\} \cap \{Y = y_i\}) &= P(\{\omega \in \Omega | X(\omega) = x_i \text{ und } Y(\omega) = y_i\}) \\ &=: P(X = x_i, Y = y_i) \end{aligned}$$

Diese Schreibweise verändert nicht das betrachtete Ereignis selbst, sondern verkürzt lediglich die Notation. Die Schreibweise ist eine Entsprechung zu der im Hauptteil des Kapitels verwendeten Schreibweise für die Wahrscheinlichkeit von Schnittereignissen (etwa $P(A \cap B)$). Es ist also beispielsweise $P(X = 1, Y = 2) = \frac{1}{36}$ oder $P(X < 4, Y = 2) = \frac{3}{36}$, wenn man den zweifachen Wurf als Laplace-Experiment modelliert.

Mit der Entsprechung von $P(X = x, Y = y)$ und $P(A \cap B)$ erhält man unmittelbar auch die Definition der stochastischen Unabhängigkeit zweier Zufallsgrößen.

Definition 27
Zwei Zufallsgrößen X und Y heißen **stochastisch unabhängig**, wenn für alle Werte $x \in X$ und $y \in Y$ gilt:

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

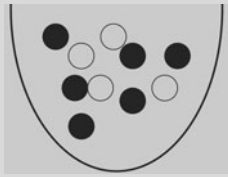
Wir hatten in Kapitel 6.1 einerseits die Unabhängigkeit zweier zufälliger Ereignisse A und B, andererseits die stochastische Unabhängigkeit zweier zufälliger Vorgänge über die paarweise stochastische Unabhängigkeit der die beiden Vorgänge festlegenden Ereignisse A_i und B_j definiert. Die Definition der stochastischen Unabhängigkeit zweier Zufallsgrößen ist, da die Forderung für *alle* x und y gelten soll, das Pendant zur stochastischen Unabhängigkeit zweier zufälliger Vorgänge. In dem Beispiel des zweifachen Wurf des Würfels können die beiden Zufallsgrößen X und Y als stochastisch unabhängig modelliert werden, da angenommen werden kann, dass sich die beiden Würfe nicht gegenseitig beeinflussen. Es gilt also (siehe auch schon oben):

$$P(X=1,Y=2) = P(X=1)\cdot P(Y=2) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36}$$

Können die beiden Zufallsgrößen X und Y nicht als stochastisch unabhängig modelliert werden, so gilt allgemein als Pendant der Definition der bedingten Wahrscheinlichkeit:

$$P(X=x,Y=y) = P(X=x)\cdot P(Y=y|X=x) \Leftrightarrow P(Y=y|X=x) = \frac{P(X=x,Y=y)}{P(X=x)}$$

Beispiel:
Gegeben ist folgende Urne (Abb. 6.14), aus der zwei Mal *ohne Zurücklegen* gezogen wird, wobei beide Züge jeweils als Laplace-Experiment modelliert werden (Gleichwahrscheinlichkeit der Elementarereignisse). Wir setzen X: Anzahl der schwarzen Kugeln im ersten Zug mit $x_1 = 0$ und $x_2 = 1$ sowie Y: Anzahl der schwarzen Kugeln im zweiten Zug mit $y_1 = 0$ und $y_2 = 1$.



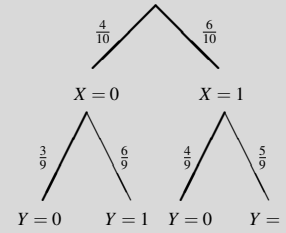


Abbildung 6.14: Zug aus einer Urne

Es gilt etwa:

$$P(Y=1|X=0) = \frac{P(X=0,Y=1)}{P(X=0)} = \frac{\frac{4}{10}\cdot\frac{6}{9}}{\frac{4}{10}} = \frac{6}{9}$$

und ebenso im Rückschluss (Umdrehen der chronologischen Abfolge) mit der Formel von Bayes:

$$\begin{aligned} P(X=0|Y=1) &= \frac{P(X=0,Y=1)}{P(Y=1)} \\ &= \frac{P(Y=1|X=0)\cdot P(X=0)}{P(Y=1|X=0)\cdot P(X=0)+P(Y=1|X=1)\cdot P(X=1)} \\ &= \frac{\frac{6}{9}\cdot\frac{4}{10}}{\frac{6}{9}\cdot\frac{4}{10}+\frac{5}{9}\cdot\frac{6}{10}} = \frac{24}{54} = \frac{4}{9} \end{aligned}$$

Wir führen schließlich noch eine Konvention ein. In Kapitel 5.1.3 hatten wir angemerkt, dass wir verschiedene Zufallsgrößen durch $X, Y, \ldots$ (wie oben) oder durch $X_1, X_2, \ldots$ bezeichnen.

Wir werden im Folgenden dann unterschiedliche Buchstaben verwenden, wenn sich die damit repräsentierten zufälligen Teilvorgänge im Kontext inhaltlich unterscheiden. Das ist etwa dann der Fall, wenn X die Anzahl der weiblichen Studierenden und Y die Anzahl der Raucher in einer Stichprobe beschreiben. Werden dagegen Teilvorgänge eines inhaltlich kohärenten zufälligen Vorgangs durch Zufallsgrößen beschrieben, so verwenden wir die Bezeichnungen $X_1, X_2, \ldots$, um die Teilvorgänge zu repräsentieren. Etwa bezeichnen wir mit X_1 die Augenzahl im ersten Wurf (die Anzahl der schwarzen Kugeln im ersten Zug, die Anzahl der weiblichen Studierenden bei der ersten Befragung) und X_2 die Augenzahl im zweiten Wurf (die Anzahl der schwarzen Kugeln im zweiten Zug, die Anzahl der weiblichen Studierenden bei der zweiten Befragung).

Beispiel:
Wir betrachten den zweifachen Wurf des Würfels, setzen also X_1: Augenzahl im ersten Wurf. Da die Zufallsgröße als Werte natürliche Zahlen annimmt, vereinfachen wir die Notation der Realisierungen von $x_1, \ldots, x_6$ auf $k_1 = 1, \ldots, 6$. Weiter ist X_2: Augenzahl im zweiten Wurf mit den Werten $k_2 = 1, \ldots, 6$. An den oben formulierten Aussagen ändert sich durch diese Notationsvariante nichts. Diese ermöglicht aber auf formal einfachere Weise die Summenbildung von Zufallsgrößen, die wir im folgenden Kapitel eingehend betrachten werden. So können wir im genannten Beispiel die Zufallsgröße X: Summe der beiden Augenzahlen auch durch $X = X_1 + X_2$ mit $k = k_1 + k_2 = 2, \ldots, 12$ definieren und erhalten:

X		X_2					
		1	2	3	4	5	6
X_1	1	1+1=2	1+2=3	1+3=4	1+4=5	1+5=6	1+6=7
	2	2+1=3	2+2=4	2+3=5	2+4=6	2+5=7	2+6=8
	3	3+1=4	3+2=5	3+3=6	3+4=7	3+5=8	3+6=9
	4	4+1=5	4+2=6	4+3=7	4+4=8	4+5=9	4+6=10
	5	5+1=6	5+2=7	5+3=8	5+4=9	5+5=10	5+6=11
	6	6+1=7	6+2=8	6+3=9	6+4=10	6+5=11	6+6=12

und daraus die Wahrscheinlichkeitsverteilung für den zweifachen Wurf (die etwa für das Spiel „Siedler von Catan“ wichtig ist):

k	2	3	4	5	6	7	8	9	10	11	12
$P(X=k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

6.7 Aufgaben

Aufgabe 6.1: Begründen Sie, warum die stochastische Unabhängigkeit zweier Ereignisse bzw. zufälliger Vorgänge nur ein Modell der Realität darstellen kann.

Aufgabe 6.2: Zeigen Sie, dass die drei Axiome der Wahrscheinlichkeitsrechnung für bedingte Wahrscheinlichkeiten gelten.

Aufgabe 6.3: Untersuchen Sie anhand des Datensatzes zu den Studierenden der Hochschulen in Münster und in Freiburg, welche Merkmale ein verstecktes Indiz für ein anderes sein können. Kann z.B. das Beförderungsmittel ein Indiz für die Parteipräferenz oder den Erhalt von BAföG sein?

Aufgabe 6.4: Eine Münze wird geworfen. X sei die Anzahl der Wappen (insgesamt), X_i die Anzahl der Wappen im i-ten Versuch. Begründen Sie, warum man die Teilvorgänge als stochastisch unabhängig modellieren kann. Bestimmen Sie die Wahrscheinlichkeitsverteilungen von X für den einfachen, den zweifachen, den dreifachen und den vierfachen Münzwurf. Zeichen Sie einen Baum zu diesem zufälligen Vorgang.

Aufgabe 6.5: Drei Urnen haben folgenden Inhalt. In der ersten Urne sind zwei weiße Kugeln, in der zweiten je eine schwarze und eine weiße, in der dritten zwei schwarze. Es wird eine Urne zufällig ausgewählt.

a) Eine weiße Kugel wird gezogen (ohne diese zurückzulegen). Bestimmen Sie die Wahrscheinlichkeit dafür, dass die zweite Kugel schwarz sein wird.

b) Im zweiten Zug ist eine schwarze Kugel gezogen worden. Bestimmen Sie die Wahrscheinlichkeit dafür, dass die erste Kugel weiß war.

c) Bestimmen Sie die beiden Wahrscheinlichkeiten aus a) und b), wenn die Kugeln nach dem Ziehen zurückgelegt werden.

Stellen Sie das Problem mit Hilfe eines Baums und dem Einheitsquadrat dar.

Aufgabe 6.6: In einer Urne sind 9 rote, 4 weiße und 7 blaue Kugeln. Es wird zufällig aus dieser Urne 3 Mal gezogen.

a) Dabei werden die gezogenen Kugeln wieder zurückgelegt. Bestimmen Sie die Wahrscheinlichkeit für die Zugfolge: Zuerst eine weiße, dann eine blaue, dann eine rote Kugel. Zeichnen Sie einen Baum zu diesem Zufallsexperiment.

b) Dabei werden die Kugeln nicht wieder zurückgelegt. Bestimmen Sie die Wahrscheinlichkeit für die Zugfolge in a).

c) Bestimmen Sie für den Vorgang mit und ohne Zurücklegen die Wahrscheinlichkeit für das Ereignis A: Es werden zwei blaue und eine weiße Kugel gezogen. Zeichnen Sie einen Baum zu dieser Fragestellung.

Aufgabe 6.7: Gegeben sei der erste HIV-Test mit: K: krank (bzw. infiziert), $\overline{K}$: nicht krank (nicht infiziert), D: Test positiv und $\overline{D}$: Test negativ. Weiterhin sei gegeben: $P(K) = 0,001$, $P(D|K) = 0,998$ (Sensitivität) und $P(D|\overline{K}) = 0,002$ (Spezifität).

a) Bestimmen Sie Wahrscheinlichkeit für $P(K|D)$.

b) Was bedeutet im Sachkontext $P(K|\overline{D})$?

c) Da das Ergebnis eines Tests (siehe a)) offensichtlich nicht ausreichend ist, um die Diagnose an einen Patienten zu geben, besteht der HIV-Test aus mehreren (unterschiedlichen) Tests. Bestimmen Sie die notwendige Spezifität des Tests insgesamt so, dass bei einer Sensitivität von $P(D|K) = 0,999$ die Wahrscheinlichkeit für $P(K|D)$ größer als 0,995 ist.

d) $P(K)$ ist die offizielle Basisrate für Deutschland. In anderen Ländern, insbesondere in Teilen Afrikas, ist die Basisrate erheblich höher. Welche Auswirkung hat das auf den HIV-Test? Bestimmen Sie die Wahrscheinlichkeit $P(K|D)$ in Abhängigkeit von $P(K)$.

e) Warum ist die Annahme $P(K) = 0,001$, also die Wahrscheinlichkeit, ohne es zu wissen und ohne Symptome mit HIV infiziert zu sein, im Allgemeinen keine sinnvolle Modellierung, wenn es um die Frage $P(K|D)$ für eine bestimmte Person geht?

7 Wahrscheinlichkeitsverteilungen

Einstiegsbeispiel

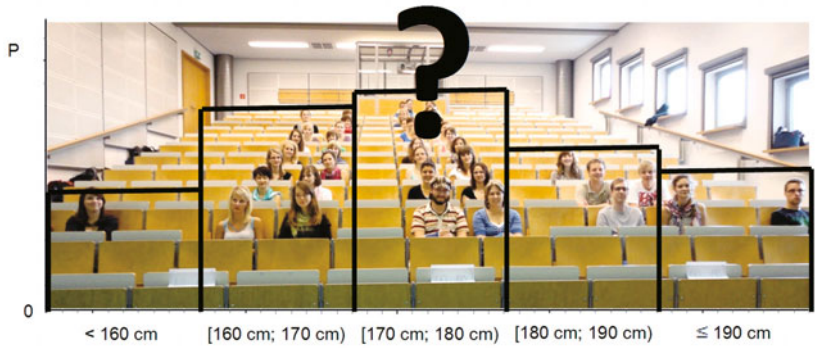


Abbildung 7.1: Muster in Prognosen untersuchen

Aufgabe 1: Ermitteln Sie die Verteilungen von Wahrscheinlichkeiten der Anzahlen von Studierenden mit einer bestimmten Eigenschaft. Bestimmen Sie ebenso die zu erwartende Anzahl und die möglichen Abweichungen von dieser erwarteten Anzahl. Ziehen Sie als Grundlage für Ihre Modellrechnungen die in Kapitel 5 geschätzten Wahrscheinlichkeiten für Eigenschaften der Studierenden heran.

Worum es geht

Besteht die Möglichkeit, ein Modell für die mehrfache Ausführung eines zufälligen Vorgangs aufzustellen (mit anderen Worten: den Entstehungsprozess zukünftiger relativer Häufigkeiten eines zufälligen Vorgangs zu modellieren), so lässt sich die Wahrscheinlichkeitsverteilung einer Zufallsgröße mit ihren charakteristischen Kennzahlen für zukünftige Daten in ähnlicher Weise bestimmen wie die Häufigkeitsverteilung und ihre charakteristischen Kennzahlen für bereits erhobene Daten. Wir wollen in diesem Kapitel wenige, mit elementar gehaltenen Methoden untersuchbare Verteilungen und ihre charakteristischen Kennzahlen diskutieren – zusammen mit der sich daraus ergebenden Möglichkeit, zukünftige zufällige Vorgänge abzuschätzen.

Wahrscheinlichkeitsverteilung Diesen Begriff haben wir bereits in Kapitel 5.2 eingeführt. In diesem Kapitel wollen wir den Begriff erweitern und bestimmte Verteilungstypen analysieren, die für die Schätzung zukünftiger Datenerhebungen wichtig sind.

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_7,

Charakteristische Kennzahlen Im Bereich der beschreibenden Datenanalyse hatten wir Mittelwerte, die Streuung um Mittelwerte wie auch den Begriff der Schiefe von Häufigkeitsverteilungen betrachtet. In gleicher Weise werden wir in diesem Kapitel die Mittelwerte, die Streuung und die Schiefe zukünftiger Verteilungen ermitteln.

Simulation Es muss stets beachtet werden, dass eine Abschätzung der Zukunft einerseits mit Unsicherheit behaftet ist, andererseits nur auf der Basis bestimmter Modellannahmen geschehen kann. Mit solchen Modellannahmen werden wir zukünftige Häufigkeitsverteilungen und deren charakteristische Kennzahlen simulieren, um einen Eindruck von ihnen zu erhalten.

7.1 Wahrscheinlichkeitsverteilungen

7.1.1 Die Gleichverteilung

Eine Klasse von Wahrscheinlichkeitsverteilungen mit dem Namen **Gleichverteilung** haben wir bereits implizit kennengelernt. Ein Prototyp für diese Klasse ist die Verteilung der Zufallsgröße X: Augenzahl beim einfachen Wurf des Würfels (X={1,2,3,4,5,6}):

k	1	2	3	4	5	6
$P(X=k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Abbildung 7.2: Gleichverteilung beim einfachen Wurf eines Würfels

Diese Gleichverteilung ergibt sich durch die Beschaffenheit des Würfels. Auch in anderen Glücksspielen werden Zufallsgeneratoren konstruiert, die eine Gleichverteilung erzeugen sollen. Die Lottokugeln werden durch einen komplexen Mischvorgang durcheinandergewirbelt und gezogen, Karten werden gemischt und verteilt, eine Münze wird geworfen usw. Dadurch soll gewährleistet werden, dass die den zufälligen Vorgängen zugrunde liegenden Elementarereignisse gleichwahrscheinlich sind.[1]

Bei nicht konstruierten zufälligen Vorgängen tritt im Allgemeinen keine Gleichverteilung auf.[2] Es wird aber versucht, etwa bei Erhebungen durch Befragung, zumindest annähernd eine Gleichverteilung zu erzeugen: So soll gewährleistet werden, dass jede statistische Einheit zumindest annähernd die gleiche Wahrscheinlichkeit hat, in die Erhebung aufgenommen zu werden.

Die Gleichverteilung ist ein Modell für eine Klasse zufälliger Vorgänge, die die Verteilung zukünftiger relativer Häufigkeiten „auf lange Sicht“ einschätzt. Dass diese Häufigkeitsverteilung bei einem Wurf der theoretischen Wahrscheinlichkeitsverteilung nicht ähnelt, sondern erst bei wachsender Versuchsanzahl, ist in den Simulationsergebnissen in Abbildung 7.3 zu sehen. Diese

[1] Auch das Glücksrad erzeugt eine solche Gleichverteilung, wenn die theoretisch unendlich vielen wählbaren Punkte des Kreises in Klassen von gleich großen Kreissektoren unterteilt werden.

[2] Es gibt allerdings auch z. B. physikalische Modelle, wie etwa die Bewegungsrichtungen von Gasmolekülen, bei denen von einer Gleichverteilung, allerdings einer stetigen Zufallsgröße mit überabzählbar unendlich vielen Werten, ausgegangen wird.

Simulation basiert auf der Gleichverteilung der Zahlen 1 bis 6 und ist für verschiedene Anzahlen von Wiederholungen realisiert.

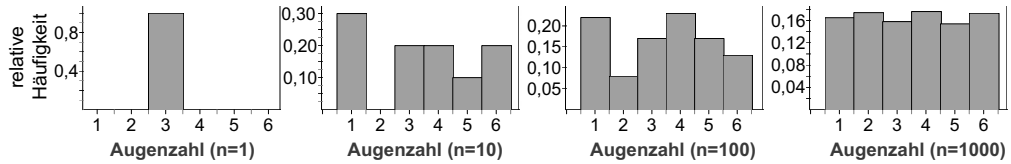


Abbildung 7.3: Simulation eine Würfels bei $n = 1, 10, 100$ und 1000 Würfen

7.1.2 Die Binomialverteilung

Ein Zufallsgenerator, der eine Klasse von spezifischen zufälligen Vorgängen, **Bernoulli-Experimente**, erzeugt, die wiederum auf eine Klasse von Wahrscheinlichkeitsverteilungen, die **Binomialverteilung**, führt, ist das Galton-Brett (vgl. Abb. 7.4). Beim einmaligen Experiment wird eine Kugel oben in das Galton-Brett geworfen. Sie wird vom ersten Hindernis nach links oder rechts abgelenkt, vom zweiten nach links oder rechts, vom dritten nach links oder rechts etc. Schließlich landet die Kugel in einem der sich unten befindenden Fächer. Welches Fach getroffen wird, weiß man nicht. Erhöht man die Anzahl der Kugeln auf z.B. $n = 1000$, so scheint sich eine eingipflige symmetrische Verteilung der Kugeln mit dem Maximum in der Mitte herauszukristallisieren. Wiederholt man den Versuch mit gleichem n, so ist zwar die Verteilung nicht gleich der ersten, die genannten Eigenschaften wiederholen sich aber.

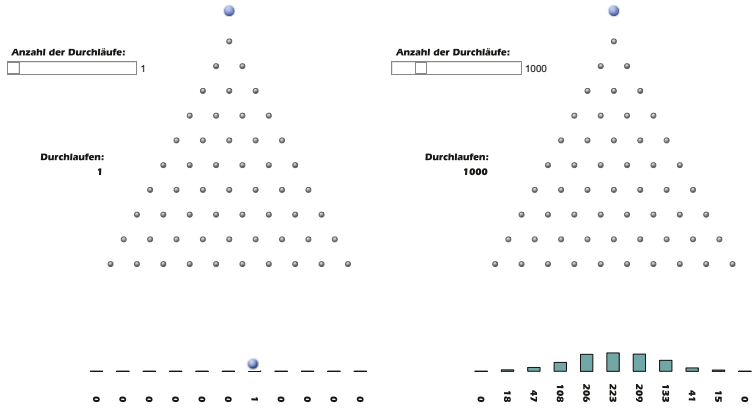


Abbildung 7.4: Galton-Brett beim Durchlauf von einer Kugel und von 1000 Kugeln

Betrachtet man eine Kugel auf dem Weg durch das Galton-Brett, so ist die Modellierung, dass man bei jedem Hindernis im Galton-Brett nur folgende zwei mögliche Ereignisse betrachtet, sofort plausibel:

A: *Kugel fällt nach rechts* und $\overline{A}$: *Kugel fällt nach links*

Eine andere Möglichkeit gibt es nicht (ein Steckenbleiben wird demnach ausgeschlossen). Allgemein führt ein zufälliger Vorgang mit äquivalenten Eigenschaften zu folgender Definition:

Definition 28
Ein zufälliger Vorgang mit der Ergebnismenge Ω und zwei sich gegenseitig ausschließenden Ereignissen A und $\overline{A} = \Omega \setminus A$ heißt **Bernoulli-Experiment**.

Bernoulli-Experimente stellen offensichtlich einen sehr einfachen Fall von zufälligen Vorgängen dar. Diese sind aber in dem Sinne universal, dass sich jeder zufällige Vorgang als Bernoulli-Experiment modellieren lässt:

Experiment	Ereignis A	Gegenereignis $\overline{A}$
Galton-Brett	Kugel rechts	Kugel links
Münze	Wappen	Zahl
Würfel	Sechs	keine Sechs
Jedes Experiment	$A \subset \Omega$	$\overline{A} = \Omega \setminus A$

Bei Bernoulli-Experimenten hat sich folgende Sprach- und Notationskonvention ergeben:

- Das Ereignis A wird häufig als „Erfolg“ und das Ereignis $\overline{A}$ als „Misserfolg“ bezeichnet.
- Man bezeichnet weiterhin $P(A) =: p$ und $P(\overline{A}) = 1 - p =: q$.

Für jedes Bernoulli-Experiment kann man eine Zufallsgröße X folgendermaßen definieren:

> X: Anzahl der Erfolge. Bei einem Einzelversuch nimmt die Zufallsgröße die Werte 0 (Misserfolg) oder 1 (Erfolg) an.[3]

Betrachtet man nun nicht nur die Kugelabweichung an *einem* Hindernis, sondern wie im abgebildeten Galton-Brett an mehreren Hindernissen hintereinander, dann erweisen sich folgende Modellannahmen als tragfähig:

- Man kann davon ausgehen, dass sich die Wiederholungen nicht gegenseitig beeinflussen, d. h., in welche Richtung die Kugel beim ersten (allgemein beim i-ten) Hindernis fällt, hat keinen Einfluss darauf, wohin die Kugel beim zweiten (allgemein beim $(i+1)$-ten) Hindernis fällt. Fasst man die Abweichungen der Kugel an den einzelnen Hindernissen als zufällige Teilvorgänge des Durchlaufs der Kugel durch das Galton-Brett auf, so können diese Teilvorgänge als stochastisch unabhängig modelliert werden. Die Möglichkeit, dass die Kugel vielleicht einen Drall bekommt und damit potentiell die stochastische Unabhängigkeit der Teilvorgänge aufgehoben wird, wird in dem *Modell* ausgeblendet.
- Geht man von der stochastischen Unabhängigkeit der zufälligen Teilvorgänge aus, so ist die Wahrscheinlichkeit für die Abweichung nach links (Erfolg A, $P(A) = p$) bzw. rechts (Misserfolg $\overline{A}$, $P(\overline{A}) = 1 - p$) an jedem Hindernis gleich, da sich die Teilvorgänge in diesem Modell unter gleichen Bedingungen wiederholen. Beim Experiment mit einem korrekt aufgestellten, voll funktionsfähigen Galton-Brett kann man zusätzlich $p = 1 - p = 0{,}5$ annehmen (Prinzip des unzureichenden Grundes).

[3] Eine Zufallsgröße X, die die Ergebnismenge eines zufälligen Vorgangs allein auf die Zahlen 0 und 1 abbildet, heißt auch *Indikatorfunktion*.

Zufällige Vorgänge, die sich in äquivalenter Weise modellieren lassen, genügen folgender Definition:

Definition 29
Die n-malige, paarweise stochastisch unabhängige Wiederholung eines Bernoulli-Experiments heißt Bernoulli-Kette der Länge n.

Wir wollen nun die Wahrscheinlichkeiten von Ereignissen in einer Bernoulli-Kette mit der Länge n betrachten. Für diese gilt:

- Sie sind alle Teilmengen von $\Omega = \Omega_1 \times \Omega_2 \times ... \times \Omega_n$ mit $\Omega_i = A \cup \overline{A}, \quad i = 1, ..., n$ und $P(A) = p, P(\overline{A}) = 1 - p$.
- Ein Ereignis in der Bernoulli-Kette ist ein Schnittereignis von n Ereignissen der n zufälligen und stochastisch unabhängigen Bernoulli-Experimente, von denen k dem Ereignis A und $n-k$ dem Ereignis $\overline{A}$ entsprechen, wobei $k = 0, 1, ..., n$ ist.
- Die Ereignisse jeder Wiederholung des Bernoulli-Experiments sind stochastisch unabhängig. Folglich besteht die Wahrscheinlichkeit eines solchen Ereignisses aus dem Produkt von p^k und $(1-p)^{n-k}$.

Beispiel:
Wir betrachten den 3-maligen Wurf einer Münze (vgl. Abb. 7.5). Dabei sei A: *Wappen* (Erfolg) und $\overline{A}$: *Zahl* (Misserfolg) mit $P(A) = p$ und $P(\overline{A}) = 1 - p$. Ein Ereignis in dieser Bernoulli-Kette der Länge 3 besteht, wenn man die beiden Ereignisse mit einem Index für die Nummer des Wurfs unterscheidet, etwa in dem Schnittereignis $A_1 \cap A_2 \cap \overline{A_3}$. Die Wahrscheinlichkeit dieses Ereignisses ist:

$$P(A_1 \cap A_2 \cap \overline{A_3}) = P(A_1) \cdot P(A_2) \cdot P(\overline{A_3}) = p \cdot p \cdot (1-p) = p^2 \cdot (1-p)^1$$

In gleicher Weise gilt auch $P(A_1 \cap \overline{A_2} \cap A_3) = P(\overline{A_1} \cap A_2 \cap A_3) = p^2 \cdot (1-p)^1$.

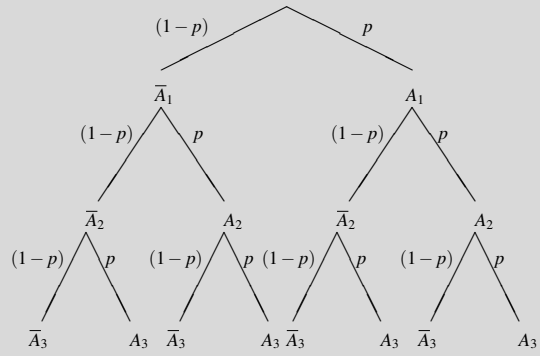


Abbildung 7.5: Baumdiagramm für drei Ebenen mit den Ereignissen Erfolg A und Misserfolg $\overline{A}$

Mit den Betrachtungen des vorangegangenen Beispiels ergibt sich einerseits die Frage nach einem Ereignis, das durch *einen Pfad* des zur Bernoulli-Kette gehörenden Baums repräsentiert ist, andererseits nach einem Ereignis, das durch eine bestimmte Anzahl von Erfolgen, also durch *mehrere Pfade* des Baums, repräsentiert ist.

Für diese Anzahl der Erfolge verwenden wir die Zufallsgröße X: Anzahl der Erfolge in einer Bernoulli-Kette der Länge n. Mit dem vorausgehenden Beispiel können wir die Anzahl der Pfade mit $X = k$ Erfolgen zunächst für Bernoulli-Ketten der Längen $n = 1,2,3$ entwickeln, indem wir fortwährend mehr Ebenen des Baumes in die Betrachtung einbeziehen. Wir erhalten dabei folgendes Ergebnis:

n	k	0	1	2	3	Summe
1	Pfade	1	1	–	–	2
2	Pfade	1	2	1	–	4
3	Pfade	1	3	3	1	8

Hier sind bereits zwei Muster erkennbar, die sich mit den in Kapitel 6.4 diskutierten kombinatorischen Zählfiguren analysieren lassen. So gibt es auf der ersten Ebene des Baumes 2, auf der zweiten 4, auf der dritten 8 und allgemein auf der n-ten Ebene 2^n Äste (erste kombinatorische Zählfigur).

Die Anzahl der Pfade mit einer bestimmten Anzahl von Erfolgen lässt sich mit Hilfe der dritten kombinatorischen Zählfigur bestimmen: Die n Äste eines Pfades haben hinsichtlich ihrer Zugehörigkeit zu einer Ebene des Baumes Positionen von 1 bis n. Diese Positionen sind die Elemente in einer Urne. Die Positionen, an denen die k Erfolge liegen sollen, werden durch das Ziehen von k Positionen aus der Urne ausgewählt. Für die Auswahl von k Elementen (Positionen) aus einer Urne mit n Elementen (Positionen) gibt es $\binom{n}{k}$ Möglichkeiten, da die Reihenfolge der Ziehung für die Positionierung der Erfolge im Baum unerheblich ist. Im Beispiel gibt es daher $\binom{3}{2} = 3$ Möglichkeiten, die zwei Erfolge auf die drei Positionen des Pfades mit drei Ästen zu legen.

Fassen wir die Überlegung zusammen, so gibt es $\binom{n}{k}$ Äste mit k Erfolgen in einem Baum, der eine Bernoulli-Kette der Länge n repräsentiert. Die Wahrscheinlichkeit für jeden dieser Pfade ist $p^k \cdot (1-p)^{n-k}$. Insgesamt ergibt sich also für die Zufallsgröße X: Anzahl der Erfolge in einer Bernoulli-Kette der Länge n:

$$P(X = k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}, \quad k = 0,1,...,n$$

Diese Wahrscheinlichkeitsverteilung erhält einen Eigennamen:

Definition 30
Hat die Wahrscheinlichkeitsverteilung einer Zufallsgröße X die unten angegebene Gestalt, so heißt diese **Binomialverteilung**:

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

Eine Zufallsgröße mit dieser Wahrscheinlichkeitsverteilung heißt binomialverteilt.

Diese Definition der Binomialverteilung geht über die Wahrscheinlichkeitsverteilung einer Bernoulli-Kette hinaus. Mit allen vorangegangenen Überlegungen erhalten wir aber unmittelbar:

Satz 13
Sei X die Anzahl der Erfolge in einer Bernoulli-Kette der Länge n, bestehend aus n stochastisch unabhängigen Wiederholungen eines Bernoulli-Experiments, so ist X binomialverteilt.

Dieser Satz bedeutet, dass immer, wenn die Voraussetzungen der stochastisch unabhängigen Wiederholung eines Bernoulli-Experiments gegeben sind, die entsprechend definierte Zufallsgröße X binomialverteilt ist. Der Satz bedeutet aber nicht umgekehrt, dass jede Binomialverteilung auf der stochastisch unabhängigen Wiederholung eines Bernoulli-Experiments basiert! Ein Gegenbeispiel ist etwa der in Abbildung 7.6 dargestellte zufällige Vorgang mit zwei Bernoulli-Experimenten, die nicht stochastisch unabhängig sind.

k	$P(X=k)$
0	$\frac{1}{3}\cdot\frac{3}{4}=\frac{1}{4}=\binom{2}{0}p^0\cdot(1-p)^2$
1	$\frac{1}{3}\cdot\frac{1}{4}+\frac{2}{3}\cdot\frac{5}{8}=\frac{1}{2}=\binom{2}{1}p^1\cdot(1-p)^1$
2	$\frac{2}{3}\cdot\frac{3}{8}=\frac{1}{4}=\binom{2}{2}p^2\cdot(1-p)^0$

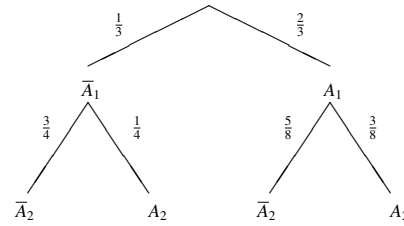


Abbildung 7.6: Gegenbeispiel zu einer binomialverteilten Zufallsgröße X mit $p=(1-p)=\frac{1}{2}$, die auf zwei stochastisch abhängigen zufälligen Vorgängen basiert

In diesem Beispiel gilt z. B. $P(A_1\cap A_2)\neq P(A_1)\cdot P(A_2)$, weil:

$$P(A_1\cap A_2)=\frac{2}{3}\cdot\frac{3}{8}=\frac{1}{4}\qquad\text{und}$$
$$P(A_1)\cdot P(A_2)=\frac{2}{3}\cdot\left(\frac{1}{3}\cdot\frac{1}{4}+\frac{2}{3}\cdot\frac{3}{8}\right)=\frac{2}{3}\cdot\left(\frac{1}{12}+\frac{3}{12}\right)=\frac{2}{9}$$

Dennoch ist X: Anzahl der Erfolge binomialverteilt mit $p=(1-p)=\frac{1}{2}$. Der vorangehende Satz hat dennoch erheblichen Nutzen, da in vielen Anwendungen die vorausgesetzten Annahmen als Modell ausreichend plausibel sind.

Beispiel:
Aus der Erhebung zu den Studierenden an der PH Freiburg lässt sich durch eine Schätzung die Wahrscheinlichkeit für das Ereignis A: Parteipräferenz für die Grünen durch $P(A)=p=0,59$ setzen ($n=119$ Studierende haben eine Angabe gemacht). Die Präferenz für eine andere Partei lässt sich daher durch $P(\overline{A})=1-P(A)=1-p=0,41$ setzen. Mit den Überlegungen aus Kapitel 6.5 modellieren wir eine Umfrage unter n Studierenden als Wiederholung stochastisch unabhängiger Einzelbefragungen. Damit ist in diesem Modell X: Anzahl der Studierenden mit Präferenz für die Grünen binomialverteilt mit den Parametern n und p.

Es ergibt sich für $n=10$, $n=50$ und $n=119$ die in Abbildung 7.7 dargestellte Wahrscheinlichkeitsverteilung für X:

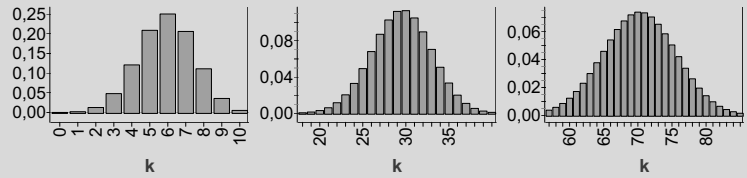
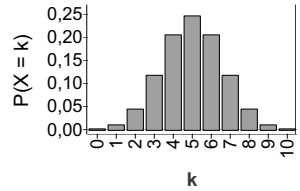


Abbildung 7.7: Binomialverteilungen mit $p = 0,59$ und $n = 10$, $n = 50$ und $n = 119$

Auf der Basis der Modellannahmen können also Vorhersagen zu relativen Häufigkeiten für die Anzahl von Studierenden mit einer Präferenz für die Partei Die Grünen getroffen werden: Bei einer Stichprobe von 50 Studierenden sind z. B. Anzahlen von 25 bis 35 Studierenden mit einer Präferenz für die Grünen recht wahrscheinlich, Anzahlen unter 25 oder über 35 dagegen weniger wahrscheinlich.

Abschließend betrachten wir wiederum das Verhältnis von Modell und Daten. Nachfolgend ist ein Modell für das Galton-Brett mit 10 Reihen von Nägeln bzw. Hindernissen gegeben.



z. B.:

$$
\begin{aligned}
P(X=k) &= \binom{10}{k} \cdot \left(\tfrac{1}{2}\right)^k \cdot \left(\tfrac{1}{2}\right)^{10-k} \\
P(X=2) &= \binom{10}{2} \cdot \left(\tfrac{1}{2}\right)^2 \cdot \left(\tfrac{1}{2}\right)^{10-2} \approx 0,044 \\
P(X<3) &= P(X=0)+P(X=1)+P(X=2) \\
&\approx 0,001+0,010+0,044 \approx 0,055
\end{aligned}
$$

Simuliert man dieses Modell über 10 gleichverteilte Zufallszahlen 0 und 1 (10 – Anzahl der Nägel, 0 – links, 1 – rechts) und wiederholt diese Simulation n Male ($n = 1, 10, 100$ und 1000, entspricht der Anzahl der Kugeln), so zeigen sich die Ergebnisse in Abbildung 7.8, die deutlich machen, dass sich die relativen Häufigkeiten erst auf lange Sicht dem Modell annähern.

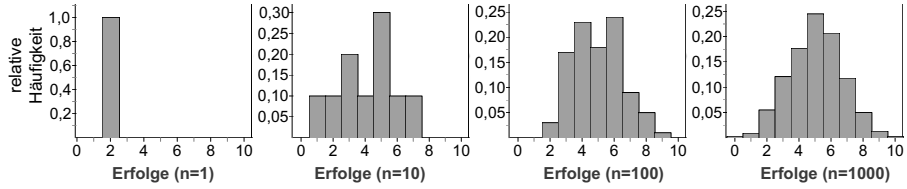


Abbildung 7.8: Simulation des Galton-Bretts bei $n = 1, 10, 100$ und 1000 Kugeln

7.1.3 Hypergeometrische Verteilung

Eine weitere Klasse von Wahrscheinlichkeitsverteilungen, die wir kursorisch betrachten, die **hypergeometrische Verteilung**, basiert ebenfalls auf Bernoulli-Experimenten, allerdings auf der Basis eines Modells, das sich vom Modell des vorangegangenen Abschnitts in einem wesentlichen Punkt unterscheidet. Das Galton-Brett aus dem vorausgehenden Abschnitt kann als wieder-

holtes Ziehen aus einer Urne mit einer weißen (Erfolg) und einer schwarzen Kugel (Misserfolg) *mit Zurücklegen* interpretiert werden. Die Reihenfolge der Erfolge/Misserfolge ist hier unerheblich. Allgemein können Bernoulli-Ketten analog als wiederholtes Ziehen aus einer Urne mit M weißen (Erfolg: E) und $N-M$ schwarzen Kugeln (Misserfolg) *mit Zurücklegen* aus einer Urne mit N Kugeln interpretiert werden, wobei für jeden Zug $P(E) = p = \frac{M}{N}$ gilt.

Bei der hypergeometrischen Verteilung betrachten wir dagegen das Ziehen aus einer solchen Urne *ohne Zurücklegen*. Als Beispiel nehmen wir das dreimalige Ziehen ($n = 3$) ohne Zurücklegen bei $N = 10$ Kugeln, von denen $M = 6$ weiß (Erfolg: A) und $N - M = 4$ schwarz (Misserfolg: $\overline{A}$) sind (vgl. Abb. 7.9).

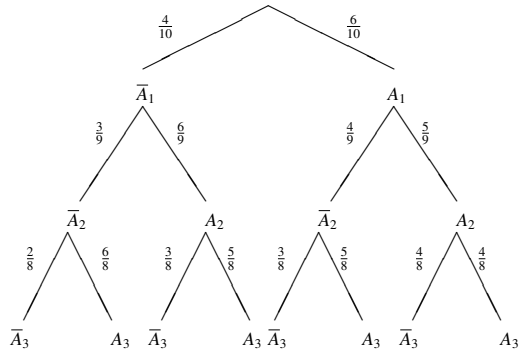


Abbildung 7.9: Baumdiagramm für drei Züge mit den Ereignissen Erfolg A und Misserfolg $\overline{A}$ ohne Zurücklegen

Man erhält z. B. (wobei die Anzahlen für die Erfolge in Klammern stehen):

$$P(A_1 \cap \overline{A}_2 \cap A_3) = \frac{6}{10} \cdot \frac{4}{9} \cdot \frac{5}{8} = \frac{(6 \cdot 5) \cdot 4}{10 \cdot 9 \cdot 8}$$

Dasselbe Ergebnis erhält man auch für alle anderen Zugfolgen mit zwei Erfolgen. Die Anzahl der Zugfolgen, bei denen 2 Erfolge auf drei Positionen verteilt werden, ist: $\binom{3}{2} = \frac{3!}{2! \cdot 1!}$ (zweite kombinatorische Zählfigur; vgl. Kap. 6.4). Damit erhalten wir, wenn die Zufallsgröße X die Anzahl der Erfolge misst, in unserem Beispiel durch Umformung:

$$\begin{aligned} P(X=k) &= \frac{3!}{2! \cdot 1!} \cdot \frac{(6 \cdot 5) \cdot 4}{10 \cdot 9 \cdot 8} = \frac{1}{2! \cdot 1!} \cdot \frac{1}{\frac{1}{3!}} \cdot \frac{\left(\frac{6 \cdot 5 \cdot \ldots \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1}\right) \cdot \frac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1}}{\frac{10 \cdot 9 \cdot \ldots \cdot 1}{7 \cdot 6 \cdot \ldots \cdot 1}} = \frac{\left(\frac{6!}{4!}\right) \cdot \frac{4!}{3!} \cdot \frac{1}{2! \cdot 1!}}{\frac{10!}{7!} \cdot \frac{1}{3!}} \\ &= \frac{\frac{6!}{4! \cdot 2!} \cdot \frac{4!}{3! \cdot 1!}}{\frac{10!}{7! \cdot 3!}} = \frac{\binom{6}{2} \cdot \binom{4}{1}}{\binom{10}{3}} \end{aligned}$$

Verallgemeinert man dieses Ergebnis, so erhält man für die Anzahl der Erfolge k beim n-maligen Ziehen aus einer Menge mit N Elementen, von denen K Erfolge sind:

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

Zufallsgrößen, die eine durch diese Formel bestimmte Wahrscheinlichkeitsverteilung haben, heißen **hypergeometrisch verteilt**. Die Verteilung heißt entsprechend **hypergeometrische Verteilung**. Ein populäres Beispiel aus der Glücksspielwelt ist die Anzahl der Richtigen im Lotto ($N = 49, K = 6, n = 6$ und $k = 0, 1, ..., 6$).

Bezogen auf die Umfrage unter den Studierenden scheint das Modell des Ziehens aus einer *Urne* ohne Zurücklegen passend für die Erhebung eines dichotomen Merkmals (etwa das Geschlecht) zu sein, wenn bei einer Befragung die Möglichkeit der Mehrfachbefragung einer Person ausgeschlossen sein soll. Betrachtet man aber solche Umfragen, bei denen die Stichprobe n gegenüber der Größe der Grundgesamtheit N klein ist, so ergibt sich bei der Wahl des Modells der Binomialverteilung bzw. hypergeometrischen Verteilung nur ein minimaler Unterschied. Geht man etwa im Modell von 71 Prozent weiblichen und 29 Prozent männlichen Studierenden an der PH Freiburg (Erhebung 2010) aus und soll eine Prognose für die relativen Häufigkeiten einer Stichprobe mit 100 der insgesamt (im Modell) 4000 Studierenden vorgenommen werden, so ergeben sich die Verteilungen, die in Abbildung 7.10 zu sehen sind.

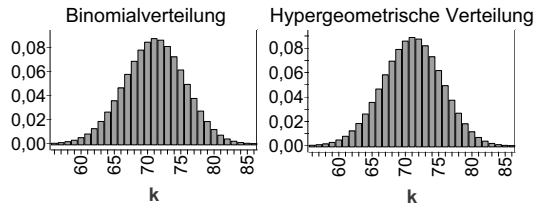


Abbildung 7.10: Binomialverteilung mit $n = 100$ und $p = 0,71$ (links), hypergeometrische Verteilung mit $N = 4000,\ n = 100$ und $K = 2840 = 0,71 \cdot 4000$

Die gute Näherung der hypergeometrischen Verteilung durch die Binomialverteilung, insbesondere im Zentrum der Verteilungen, zeigen die beiden Diagramme in Abbildung 7.11: Links sind die Abweichungen der Binomialverteilung zur hypergeometrischen Verteilung absolut aufgetragen und rechts entsprechend prozentual.

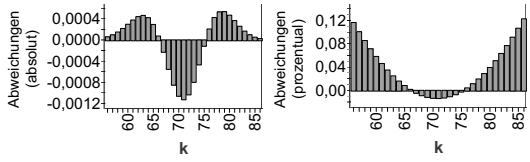


Abbildung 7.11: Binomialverteilung – hypergeometrische Verteilung, Abweichungen absolut und prozentual

Die Konvergenz der hypergeometrischen Verteilung gegen die Binomialverteilung (d. h. der jeweiligen Verteilungsfunktionen) ist ein Satz, den wir hier nicht beweisen werden (s. dazu z. B. Fisz, 1980). Durch diese Eigenschaft der hypergeometrischen Verteilung ist es häufig möglich, bei Anwendungen auf die einfacher zu handhabende Binomialverteilung zurückzugreifen.

7.2 Zentrum, Streuung und Form von Wahrscheinlichkeitsverteilungen

Im Zusammenhang mit dem empirischen Gesetz der großen Zahlen haben wir verdeutlicht, dass Wahrscheinlichkeiten ein Modell für die zu erwartenden zukünftigen relativen Häufigkeiten auf lange Sicht darstellen sollen. Bei theoretisch unendlich vielen Wiederholungen eines zufälligen Vorgangs, etwa eine Würfelwurfs, wird also *erwartet*, dass (im Modell) die relative Häufigkeit eines Ereignisses der Wahrscheinlichkeit entspricht. Dass das so ist, haben wir bisher allerdings entweder empirisch belegt oder theoretisch postuliert. Erst später in diesem Kapitel werden wir diesen Sachverhalt auch explizit theoretisch begründen. Wie bei relativen Häufigkeiten lassen sich auch zu erwartende Lage- oder Streumaße von zufälligen Vorgängen angeben.

7.2.1 Zentrum von Wahrscheinlichkeitsverteilungen: Der Erwartungswert

Wir betrachten das zunächst am Beispiel des einfachen Würfelwurfs mit der Zufallsgröße X: Augenzahl des Würfels. Betrachten wir die Wahrscheinlichkeiten $P(X=k)=\frac{1}{6}$ als die im modellhaften Wurf des Würfels zu erwartenden relativen Häufigkeiten, so würden wir bei vielen Würfen (n) ungefähr folgendes arithmetisches Mittel erwarten:

$$\begin{aligned}\bar{x} &= \frac{1}{n}\cdot(\underbrace{1+1+\ldots+1}_{\approx n/6 \text{ Summanden}} + \underbrace{2+2+\ldots+2}_{\approx n/6 \text{ Summanden}} + \underbrace{3+3+\ldots+3}_{\approx n/6 \text{ Summanden}} + \underbrace{4+4+\ldots+4}_{\approx n/6 \text{ Summanden}} \\ &\quad + \underbrace{5+5+\ldots+5}_{\approx n/6 \text{ Summanden}} + \underbrace{6+6+\ldots+6}_{\approx n/6 \text{ Summanden}}) \\ &\approx \frac{1}{n}\cdot\sum_{i=1}^{6} n\cdot\frac{i}{6} = \sum_{i=1}^{6} i\cdot\frac{1}{6} = \sum_{i=1}^{6} i\cdot P(X=i) = 21\cdot\frac{1}{6} = 3{,}5\end{aligned}$$

Das bedeutet zunächst, dass man *auf lange Sicht* (wenn nämlich jede Augenzahl etwa mit einem Anteil von $\frac{1}{6}$ erschienen sein wird) bei den zukünftigen Würfen ein arithmetisches Mittel von 3,5 beim einfachen Wurf des Würfels erwartet. Dieses arithmetische Mittel zukünftiger Würfe erhält per Definition den Namen **Erwartungswert.**

Definition 31
Sei X eine diskrete Zufallsgröße zu einem zufälligen Vorgang mit endlicher Ergebnismenge Ω, so heißt

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

Erwartungswert der Zufallsgröße X.

Im Einstiegsbeispiel gilt $X(\omega)=\omega$, die Zufallsgröße nimmt also die Werte 1 bis 6 an. Bezeichnet man die endlich vielen Werte der Zufallsgröße X als $x_1, x_2, \ldots, x_s$ und bestimmt die Wahrscheinlichkeiten für $P(X=x_i),\ i=1,\ldots,s$, so ist die Formulierung $E(X)=\sum_{i=1}^{s} x_i\cdot P(X=x_i)$ eine äquivalente Definition des Erwartungswerts der Zufallsgröße X. Bei Anwendungen werden

wir tendenziell eher die letztere Version der Definition und bei natürlichen Werten einer Zufallsgröße X wie bisher allein die Notation $k \in \mathbb{N}$ für einen spezifischen Wert der Zufallsgröße verwenden. Die Formulierung, die $X(\omega)$ verwendet, hilft allerdings, bei im Weiteren folgenden Beweisen, den formalen Aufwand gering zu halten. Daher werden wir diese Formulierung insbesondere bei theoretischen Betrachtungen verwenden. Die Bedeutung von *auf lange Sicht* diskutieren wir anhand des einfachen Würfelwurfs in zwei Varianten:

- Zunächst betrachten wir die Stabilisierung des arithmetischen Mittels von $n = 1, 2, ...$ (einfachen) Würfen des Würfels als Pendant zur Betrachtung der Stabilisierung der relativen Häufigkeiten (Abb. 7.12).

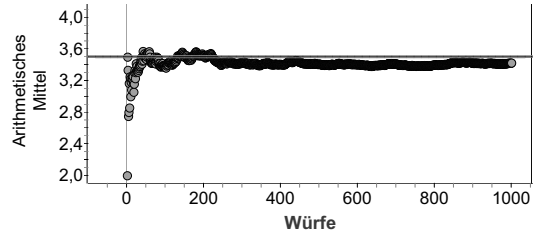


Abbildung 7.12: Stabilisierung des Ewartungswerts beim einfachen Würfelwurf

  Man sieht in Abbildung 7.12, dass sich die Mittelwerte bei höherer Versuchsanzahl in der Nähe des theoretischen Erwartungswerts von 3,5 zunehmend stabilisieren.

- In einem anderen Zugang werden die Mittelwerte der einfachen Würfe bei 1, 10 und 100 Versuchsdurchführungen betrachtet. Die jeweiligen Wurfserien werden 1000 Mal hintereinander ausgeführt (Abb. 7.13).

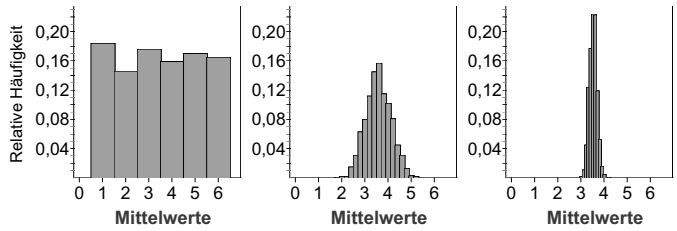


Abbildung 7.13: Mittelwerte beim 1-, 10- und 100-fachen Würfelwurf

  Man erkennt bei dem Vergleich der drei Verteilungen von Mittelwerten, dass das Prinzip des *auf lange Sicht* schon bei noch recht geringen Wurfwiederholungen deutlich wird: Die Streuung der empirischen Mittelwerte um den Erwartungswert 3,5 wird bei wachsender Versuchsanzahl immer geringer.

Zur Vertiefung betrachten wir das Beispiel des zweifachen Wurfs des Würfels. Es scheint mit dem Wissen um den Erwartungswert des einfachen Wurfs von 3,5 plausibel zu sein, dass in diesem Fall der Erwartungswert den doppelten Wert hat.

Beispiel:
Es sei die Zufallsgröße X: Augensumme beider Würfe. Wir erhalten für $k = 2,3,...,12$ folgende Wahrscheinlichkeitsverteilung mitsamt den Produkten $k \cdot P(X = k)$:

k	2	3	4	5	6	7	8	9	10	11	12	Summe
$P(X=k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1
$k \cdot P(X=k)$	$\frac{2}{36}$	$\frac{6}{36}$	$\frac{12}{36}$	$\frac{20}{36}$	$\frac{30}{36}$	$\frac{42}{36}$	$\frac{40}{36}$	$\frac{36}{36}$	$\frac{30}{36}$	$\frac{22}{36}$	$\frac{12}{36}$	$\frac{252}{36} = 7$

Es gilt also $P(X) = 7$, Das heißt, *auf lange Sicht* wird das arithmetische Mittel der Augensummen von Doppelwürfen des Würfels 7 betragen.

Was hinsichtlich der Augensumme des zweifachen Würfelwurfs plausibel war, lässt sich als Satz verallgemeinern und einfach beweisen.

Satz 14
Seien X und Y Zufallsgrößen hinsichtlich der endlichen Ergebnismenge Ω eines zufälligen Vorgangs, so gilt:

$$E(X+Y) = E\left(X(\omega)+Y(\omega)\right) = E(X(\omega)) + E(Y(\omega)) = E(X) + E(Y).$$

Beweis:

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

Dieser Satz lässt sich durch vollständige Induktion erweitern auf:

$$E\left(X_1 + X_2 + ... + X_n\right) = E\left(\sum_{j=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = E(X_1) + E(X_2) + ... + E(X_n)$$

Diese Erweiterung wird ein wesentliches Hilfsmittel in den folgenden Überlegungen haben.

Beispiel:
Wir wenden den vorherigen Satz auf den n-fachen Wurf des Würfels an und erhalten, da für X_j: Augenzahl in einem beliebigen Wurf j gilt $E(X_j) = 3,5$:

Anzahl der Würfe n	Erwartungswert der Augenzahl $X = X_1 + X_2 + ... + X_n$
$n = 1$	$E(X) = E(X_1) = 3,5$
$n = 2$	$E(X) = E(X_1 + X_2) = E(X_1) + E(X_2) = 3,5 + 3,5 = 2 \cdot 3,5 = 7$
$n = 3$	$E(X) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = 3,5 + 3,5 + 3,5 = 3 \cdot 3,5 = 10,5$
...	...
$n = m$	$E(X) = E(X_1 + X_2 + ... + X_m) = m \cdot 3,5$

Wir wenden die Überlegungen auf Bernoulli-Ketten der Länge n an, für die die Zufallsgröße X: Anzahl der Erfolge in der Bernoulli-Kette vom Umfang n sei. Für jedes der gleichartigen

Bernoulli-Experimente in dieser Kette ist X_i die Anzahl der Erfolge in der i-ten Wiederholung des Bernoulli-Experiments mit folgender Wahrscheinlichkeitsverteilung:

$X_i = k$	0	1
$P(X_i = k)$	$1-p$	p

Als Erwartungswert ergibt sich für das i-te Bernoulli-Experiment durch die spaltenweise Multiplikation in der Tabelle zur Wahrscheinlichkeitsverteilung und die Aufsummierung der Produkte: $E(X_i) = 0 \cdot (1-p) + 1 \cdot p = p$. Für den Erwartungswert $E(X)$ der Bernoulli-Kette der Länge n ergibt sich damit:

$$E(X) = E(\underbrace{X_1 + X_2 + ... + X_n}_{n \text{ Summanden}}) = \underbrace{E(X_1) + E(X_2) + ... + E(X_n)}_{n \text{ Summanden}} = n \cdot p$$

Der Beweis des folgenden Satzes ist etwas komplexer.[4] Wir werden diesen Beweis in den ergänzenden Bemerkungen dieses Kapitels aufnehmen (vgl. Kap. 7.4).

Satz 15
Ist a) X eine binomialverteilte Zufallsgröße mit den Parametern n und p bzw. ist b) X eine hypergeometrische Verteilung mit den Parametern N, K und n, so gilt:

a) $E(X) = n \cdot p$
b) $E(X) = n \cdot \frac{K}{N}$

Abschließend wollen wir noch eine nützliche Eigenschaften von Erwartungswerten einer Zufallsgröße betrachten:

Satz 16
Sei X eine Zufallsgröße mit den Werten $x_1, x_2, ..., x_s$ und $a, b \in \mathbb{R}$, so gilt:

$$E(aX + b) = aE(X) + b \text{ (Linearität des Erwartungswerts)}$$

Diese Linearität ergibt sich durch:

$$\begin{aligned} E(aX+b) &= \sum_{i=1}^{s} (ax_i + b) \cdot P(X = x_i) = a \sum_{i=1}^{s} x_i \cdot P(X = x_i) + b \underbrace{\sum_{i=1}^{s} P(X = x_i)}_{=1} \\ &= aE(X) + b \end{aligned}$$

Für andere Beschreibungen des Zentrums einer (theoretischen) Wahrscheinlichkeitsverteilung als erweiterte Betrachtung der Lageparameter einer (empirischen) Häufigkeitsverteilung ergeben sich kaum Veränderungen. Wir werden im Bereich der Wahrscheinlichkeitsanalyse auch nur am Rande auf diese Lageparameter eingehen:

[4] Binomialverteilte Zufallsgrößen müssen, wie wir bereits festgestellt haben, nicht notwendig auf Bernoulli-Ketten der Länge n basieren (Kap. 7.1.2), daher muss der Erwartungswert für binomialverteilte wie auch hypergeometrisch verteilte Zufallsgrößen direkt aus der Definition von Erwartungswerten bestimmt werden.

- Das p-Quantil einer Zufallsgröße X kann durch die beiden Ungleichungen $P(X \leq x_p) \geq p$ und $P(X \geq x_p) \geq (1-p)$ analog zur beschreibenden Datenanalyse (vgl. Kap. 2.3) bestimmt werden.
- Minimum und Maximum einer Zufallsgröße X ergeben sich (falls diese Werte existieren) durch den kleinsten bzw. größten Wert der Zufallsgröße mit einer Wahrscheinlichkeit größer 0.
- Der Modalwert einer Zufallsgröße X ist derjenige Wert der Zufallsgröße mit der größten Wahrscheinlichkeit.

Auf spezielle Quantile werden wir später kommen. Da wir im Weiteren stets als Zielrichtung annähernd symmetrische Verteilungen von Zufallsgrößen betrachten, ist die Unterscheidung von Median, Erwartungswert und Modalwert einer Zufallsgröße zumindest für die folgenden Betrachtungen sekundär.

7.2.2 Streuung, Varianz, Standardabweichung

Entsprechend zu der Analyse empirischer Daten kann man durch die Betrachtung der **Streuung** von Zufallsgrößen Prognosen für die Streuung in zukünftigen Erhebungen statistischer Daten abschätzen. In der Datenanalyse haben wir hauptsächlich Streuparameter wie den Quartilsabstand und die Spannweite analysiert, deren Berechnung und Interpretation sich mit Beachtung der Angaben am Ende des vorangegangenen Abschnitts nicht ändern. Bei der Wahrscheinlichkeitsanalyse wollen wir die (theoretische) Varianz und die (theoretische) Standardabweichung aufgrund ihrer hohen Bedeutung im Aufbau der Theorie stärker in den Mittelpunkt der Betrachtung rücken.

Wenn man Überlegungen zur Prognose zukünftiger Merkmalsausprägungen in einer Stichprobe anstellt, ist es im Zusammenhang mit der Prognose des zukünftigen Mittelwerts durch den Erwartungswert entscheidend, ob Abweichungen von dem in einer Stichprobe *erwarteten* Wert kleiner oder größer ausfallen können. Betrachtet man etwa exemplarisch die beiden Binomialverteilungen in Abbildung 7.14 ($n_1 = 30; p_1 = \frac{1}{6}; n_2 = 10, p_2 = 0,5$), so erkennt man bereits die unterschiedliche Streuung um den Erwartungswert (die hier aber insgesamt noch recht ähnlich ist).

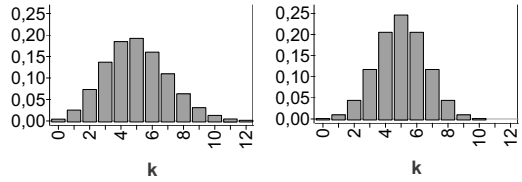


Abbildung 7.14: Zwei Verteilungen von Zufallsgrößen mit gleichem Erwartungswert

Um Unterschiede in der Streuung innerhalb der Wahrscheinlichkeitsanalyse messen zu können, verwenden wir die theoretische Varianz, die (vgl. auch Kap. 2.4.3) der Erwartungswert (arithmetisches Mittel) der quadratischen Abweichungen der Werte der Zufallsgröße zum Erwartungswert derselben Zufallsgröße ist:

Definition 32
Sei X eine Zufallsgröße, so heißt

$$V(X) = E\left((X - E(X))^2\right) = \sum_{\omega \in \Omega} (X(\omega) - E(X))^2 \cdot P(X(\omega))$$

Varianz und $\sigma(X) = \sqrt{V(X)}$ Standardabweichung der Zufallsgröße X.

Nummeriert man wiederum (wie in Kap. 7.2.1 beim Erwartungswert) die Werte der Zufallsgröße durch einen Index $(x_1, x_2, ..., x_s)$, so ergibt sich eine äquivalente Definition der Varianz mit:

$$V(X) = \sum_{i=1}^{n} (x_i - E(X))^2 \cdot P(X = x_i)$$

Sind die Werte der Zufallsgrößen natürliche Zahlen $k \in \mathbb{N}$, so können statt der x_i die für k zulässigen Werte verwendet werden.

Beispiel:
Wir betrachten wiederum den einfachen Würfelwurf:

k	1	2	3	4	5	6
$P(X=k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Es ergibt sich, da $E(X) = 3,5$ ist:

$$\begin{aligned} V(X) &= (1-3,5)^2 \cdot \frac{1}{6} + (2-3,5)^2 \cdot \frac{1}{6} + (3-3,5)^2 \cdot \frac{1}{6} + (4-3,5)^2 \cdot \frac{1}{6} + (5-3,5)^2 \cdot \frac{1}{6} + (6-3,5)^2 \cdot \frac{1}{6} \\ &= 17,5 \cdot \frac{1}{6} \approx 2,92 \end{aligned}$$

Es ist also $\sigma(X) \approx 1,7$, wobei die Standardabweichung geometrisch als Abstand vom Erwartungswert interpretiert werden kann.

Folgende beiden Eigenschaften der Varianz einer Zufallsgröße lassen sich bei der Berechnung spezieller Varianzen gut verwenden:

Satz 17
Sei X eine Zufallsgröße, so gilt

1. $V(X) = E(X^2) - E(X)^2$. Diese Eigenschaft kann bei der Berechnung von Varianzen nützlich sein.
2. $V(aX + b) = a^2 V(X)$

Die erste Eigenschaft ergibt sich aus:

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

Die zweite Eigenschaft ergibt sich aus:

$$\begin{aligned} V(aX+b) &= E\left(((aX+b)-E(aX+b))^2\right) = \underbrace{E\left((aX+b-aE(X)-b)^2\right)}_{\text{Linearität des Erwartungswerts}} \\ &= E\left((a(X-E(X)))^2\right) = E\left(a^2(X-E(X))^2\right) = a^2V(X) \end{aligned}$$

Wie beim Erwartungswert umfassen die theoretische Varianz und Standardabweichung eine Prognose für die empirische Varianz und Standardabweichung in einer Stichprobe *auf lange Sicht*. Wir machen das wiederum anhand der Stabilisierung von Varianz und Standardabweichung beim einfachen Würfelwurf deutlich. Dazu simulieren wir eine Serie von 1, 2, ..., 1000 Würfen des Würfels und berechnen nach jedem Wurf die insgesamt erzeugte Varianz.

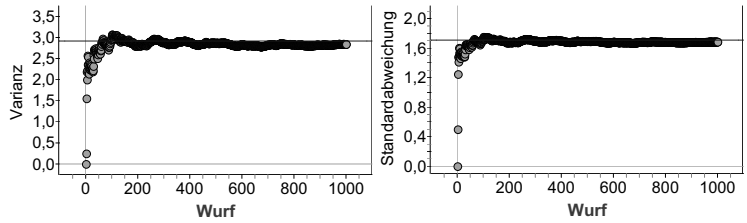


Abbildung 7.15: Varianz und Standardabweichungen bei 1, 2, ..., 1000 Würfen

Fast analog zum Erwartungswert lässt sich auch die Varianz einer Zufallsgröße $X+Y$ als Summe der Varianzen der einzelnen Zufallsgrößen betrachten. „Fast“ bedeutet hier, dass die Voraussetzung der stochastischen Unabhängigkeit bestehen muss.

Satz 18
Seien X und Y stochastisch unabhängige Zufallsgrößen. Dann gilt:

$$V(X+Y) = V(X)+V(Y)$$

Dies lässt sich zeigen durch:

$$\begin{aligned} V(X+Y) &= E\left((X+Y)^2\right) - E(X+Y)^2 \\ &= E(X^2)+E(2\cdot X\cdot Y)+E(Y^2)-(E(X)+E(Y))^2 \\ &= E(X^2)+E(2\cdot X\cdot Y)+E(Y^2)-E(X)^2-2\cdot E(X)E(Y)-E(Y)^2 \\ &= \left[E(X^2)-E(X)^2\right]+\left[E(Y^2)-E(Y)^2\right]+E(2\cdot X\cdot Y)-2\cdot E(X)E(Y) \\ &= V(X)+V(Y)+2\cdot E(X\cdot Y)-2\cdot E(X)E(Y) \end{aligned}$$

Betrachtet man die letzten beiden Terme, so wird bereits klar, dass der Satz bewiesen ist, wenn man – die stochastische Unabhängigkeit vorausgesetzt – beweisen kann, dass $E(X\cdot Y) = E(X)\cdot E(Y)$ gilt. Da dieser Nachweis formal etwas aufwändiger ist, betrachten wir zunächst die bei der Multiplikation zweier Zufallsgrößen X und Y mit s bzw. t Werten entstehenden Produkte:

X \ Y	y_1	y_2	y_3	...	y_t
x_1	$x_1 \cdot y_1$	$x_1 \cdot y_2$	$x_1 \cdot y_3$	...	$x_1 \cdot y_t$
x_2	$x_2 \cdot y_1$	$x_2 \cdot y_2$	$x_2 \cdot y_3$	...	$x_2 \cdot y_t$
x_3	$x_3 \cdot y_1$	$x_3 \cdot y_2$	$x_3 \cdot y_3$	...	$x_3 \cdot y_t$
...	...	...	...	...	...
x_s	$x_s \cdot y_1$	$x_s \cdot y_2$	$x_s \cdot y_3$	...	$x_s \cdot y_t$

Den Erwartungswert der Zufallsgröße $X \cdot Y$ erhält man, indem man alle im Inneren der Tafel stehenden Produkte gemäß der Definition mit der zugehörigen Wahrscheinlichkeit multipliziert und anschließend aufsummiert. Das Aufsummieren gehen wir so an, dass zunächst zeilenweise, beginnend mit der ersten (und damit mit einem festen x_1), vorgegangen wird und anschließend die folgenden Zeilen mit $x_2, x_3, ...$ einbezogen werden. Das lässt sich formal in einer Doppelsumme ausdrücken:

$$
\begin{aligned}
E(X \cdot Y) &= \sum_{i=1}^{s} \left(\sum_{j=1}^{t} (x_i \cdot y_j) \cdot P(X = x_i, Y = y_j) \right) \\
&\quad \big(\text{wegen der stochastischen Unabhängigkeit gilt } P(X = x_i, Y = y_j) = P(X = x_i) \cdot P(Y = y_i)\big) \\
&= \sum_{i=1}^{s} \left(\sum_{j=1}^{t} (x_i \cdot P(X = x_i)) \cdot (y_j \cdot P(Y = y_j)) \right) \\
&= \sum_{i=1}^{s} \left(\underbrace{(x_i \cdot P(X = x_i))}_{\text{Konstante hinsichtlich der inneren Summe}} \cdot \sum_{j=1}^{t} (y_j \cdot P(Y = y_j)) \right) \\
&= \sum_{i=1}^{s} ((x_i \cdot P(X = x_i)) \cdot E(Y)) \\
&= E(Y) \cdot \sum_{i=1}^{s} x_i \cdot P(X = x_i) = E(X) \cdot E(Y)
\end{aligned}
$$

Damit ist der Satz insgesamt bewiesen. Dieser Satz lässt sich mit Hilfe der vollständigen Induktion folgendermaßen erweitern:

Satz 19
Seien $X_1, X_2, ..., X_n$ paarweise stochastisch unabhängige Zufallsgrößen, so gilt:

$$V(X_1 + X_2 + ... + X_n) = V(X_1) + V(X_2) + ... + V(X_n)$$

Dieser Satz lässt sich wie im vergangenen Abschnitt für die einfache Berechnung der Varianz einer Zufallsgröße X: Anzahl der Erfolge in einer Bernoulli-Kette der Länge n verwenden. Wiederum bezeichnen wir mit X_i, $(i = 1, 2, ..., n)$ die n einzelnen Bernoulli-Experimente in der Kette. Wir berechnen anhand der Wahrscheinlichkeitsverteilung eines beliebigen Bernoulli-Experiments die Varianz von X_i:

k	0	1
$P(X_i = k)$	$1-p$	p

Den Erwartungswert einer solchen Zufallsgröße hatten wir bereits im vorangegangenen Abschnitt durch $E(X_i) = p$ bestimmt. Damit ergibt sich die Varianz durch

$$\begin{aligned} V(X_i) &= (0-E(X_i))^2 \cdot (1-p) + (1-E(X_i))^2 \cdot p \\ &= p^2 \cdot (1-p) + (1-p)^2 \cdot p \\ &= p \cdot (1-p) \cdot \underbrace{[p+(1-p)]}_{=1} \\ &= p \cdot (1-p) \end{aligned}$$

Für die Varianz von X: Anzahl der Erfolge in einer Bernoulli-Kette der Länge n ergibt sich dadurch insgesamt:

$$V(X_1 + X_2 + ... + X_n) = V(X_1) + V(X_2) + ... + V(X_n) = n \cdot p \cdot (1-p)$$

Damit lässt sich unmittelbar die Standardabweichung angeben:

$$\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$$

Ohne Beweis (siehe dazu Kap. 7.4) gilt hinsichtlich der bisher betrachteten Wahrscheinlichkeitsverteilungen:

Satz 20
Sei a) X eine binomialverteilte Zufallsgröße mit den Parametern n und p und b) X eine hypergeometrisch verteilte Zufallsgröße mit den Parametern N, K und n, so gilt:

a) $V(X) = n \cdot p \cdot (1-p) \quad \sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$

b) $V(X) = n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N-n}{N-1} \quad \sigma(X) = \sqrt{n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N}\right) \cdot \frac{N-n}{N-1}}$

Als Ergänzung der Betrachtung zu der Varianz der Summe unabhängiger Zufallsgrößen X und Y wollen wir abschließend den Fall betrachten, dass X und Y *nicht* stochastisch unabhängig sind. Hier würde sich ergeben (s.o.):

$$V(X+Y) = V(X) + V(Y) + 2 \cdot E(X \cdot Y) - 2 \cdot E(X)E(Y)$$

Formt man den Term $E(XY) - E(X)E(Y)$ um, so erhält man

$$\begin{aligned} E(XY) - E(X)E(Y) &= E(XY) - E(E(X)E(Y)) - E(E(X)E(Y)) + E(E(X)E(Y)) \\ &= E(XY) - E(XE(Y)) - E(E(X)Y) + E(E(X)E(Y)) \\ &= E[XY - XE(Y) + E(X)Y + E(X)E(Y)] \\ &= E[(X - E(X)) \cdot (Y - E(Y))] =: COV(X,Y) \end{aligned}$$

also die theoretische Entsprechung der empirischen Kovarianz s_{XY} (vgl. Kap. 3.3). Bei stochastisch unabhängigen Zufallsgrößen X und Y treten demnach Kovarianzen auf.

Damit lässt über Normierung dieser theoretischen Kovarianz – analog zur Normierung der empirischen Kovarianz – der theoretische Korrelationskoeffizient $r(X,Y)$ (als Prognose des empirischen) wie auch die Prognose für die Koeffizienten einer Regressionsgeraden $a(X,Y)$ und $b(X,Y)$ folgendermaßen konstruieren:

$$r(X,Y) = \frac{COV(X,Y)}{\sigma(X)\sigma(Y)}; \quad a(X,Y) = \frac{COV(X,Y)}{\sigma(X)}; \quad b(X,Y) = E(Y) - a(X,Y) \cdot E(X)$$

Wir werden hier allerdings nicht weiter auf die Prognose von Korrelationskoeffizienten bzw. Parametern einer Anpassungsgeraden eingehen.

7.2.3 Schiefe

Die Form der Verteilung ist einerseits für die Verwendung eines Mittelwerts wichtig (Kap. 2.5), aber auch, um die Streuung in Form der als Abstand deutbaren Standardabweichung einer Zufallsgröße beurteilen zu können. In Abbildung 7.16 sind als Beispiele eine rechtsschiefe (linkssteile) Binomialverteilung mit $n = 10$ und $p = 0,1$ sowie eine symmetrische Binomialverteilung mit $n = 10$ und $p = 0,5$ abgebildet. Die Standardabweichung ist hinsichtlich der eingezeichneten Senkrechten als Abstand vom Erwartungswert ($E(X) - \sigma(X)$ und $E(X) + \sigma(X)$) repräsentiert. Man erkennt, dass bei einer symmetrischen Verteilung die durch $E(X) \pm \sigma(X)$ gebildeten Intervalle die identische Wahrscheinlichkeitsmasse einschließen, während dies bei schiefen Verteilungen nicht oder nur annähernd der Fall sein kann. Gleiches gilt, wenn man statt der Standardabweichung etwa den Quartilsabstand als Streumaß verwendet.

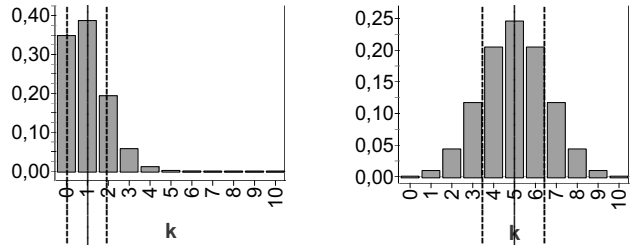


Abbildung 7.16: Erwartungswert und Standardabweichung als Abstand vom Erwartungswert

An der formalen Darstellung von Schiefemaßen ändert sich gegenüber Kapitel 2.5.1 nur, dass anstelle der empirischen Varianz die theoretische Varianz verwendet wird, und, dass an die Stelle des arithmetischen Mittels der Erwartungswert tritt. Entsprechend lässt sich die Schiefe der Wahrscheinlichkeitsverteilung einer Zufallsgröße X mit den Werten $x_1, ..., x_n$ angegeben durch:

$$g = \frac{\sum_{i=1}^{n}((X - E(X))^3) \cdot P(X = x_n)}{\sigma(X)^3}$$

Ohne auf die formale Berechnung dieses Schiefemaßes weiter einzugehen, betrachten wir hier allein die Schiefe der Binomialverteilung und der hypergeometrischen Verteilung in Abhängigkeit ihrer Parameter anhand weniger Beispiele.

Für die Binomialverteilung lassen sich folgende Beziehungen angeben, die mit Blick auf Abbildung 7.17 exemplarisch nachvollzogen werden können:

p	n	Form
klein	klein	rechtsschief (linkssteil)
$\approx 0,5$	klein	(annähernd) symmetrisch
groß	klein	linksschief (rechtssteil)
beliebig	(sehr) groß	annähernd symmetrisch

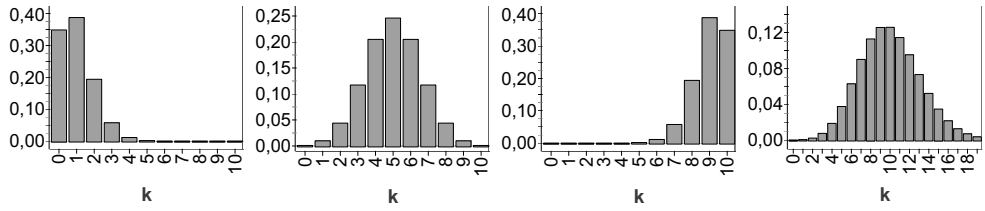


Abbildung 7.17: Binomialverteilung mit (von links nach rechts) mit $p = 0,1;\ n = 10$, $p = 0,5;\ n = 10$, $p = 0,9;\ n = 10$ und $p = 0,01;\ n = 1000$

Berechnet man die Schiefemaße der oben dargestellten Binomialverteilungen, so ergeben sich die nachfolgend tabellierten Werte. Ein Blick darauf zeigt: Während die erste und die dritte Verteilung schief sind, ist die zweite symmetrisch. Die vierte Verteilung ist anscheinend annähernd symmetrisch, obwohl das Schiefemaß noch einen Wert deutlich von 0 verschieden hat. Betrachtet man die Parameter $p = 0,05$ und $p = 0,1$ sowie $n = 1000$, so erhält man die Schiefemaße $g \approx 0,13$ und $g = 0,08$. Sie ergeben eine der grafischen Darstellung der Wahrscheinlichkeitsverteilung entsprechende annähernde Symmetrie.

Parameter	Schiefe
$p = 0,1;\ n = 10$	$g \approx 0,84$
$p = 0,5;\ n = 10$	$g \approx 0$
$p = 0,9;\ n = 10$	$g \approx -0,84$
$p = 0,01;\ n = 1000$	$g \approx 0,31$

Allgemein besteht die Faustformel, dass eine Binomialverteilung dann annähernd symmetrisch ist, wenn ihre Varianz größer 9 ist. Bei Parametern von $p \approx 0,5$ ist die Symmetrie aber auch für kleinere Varianzen annähernd gegeben.

Für die hypergeometrische Verteilung bestehen folgende Beziehungen:

K/N	n	Form
klein	klein	rechtsschief (linkssteil)
$\approx 0,5$	klein	(annähernd) symmetrisch
groß	klein	linksschief (rechtssteil)
beliebig	(sehr) groß	annähernd symmetrisch

7.2.4 Abschätzungen

Im Zusammenhang mit dem empirischen Gesetz der großen Zahlen (vgl. Kap. 5.2.2) haben wir auf *empirischer* Basis Belege gesammelt, dass eine möglichst große Stichprobe zu adäquaten Abschätzungen führt, etwa einer Wahrscheinlichkeit für ein Ereignis oder auch einem Erwartungswert als theoretischem Mittelwert in einer Grundgesamtheit. Hier wollen wir das empirische Phänomen der Stabilisierung noch einmal aufgreifen, dann aber den Zusammenhang von Stichprobengröße und Prognosegüte von der theoretischen Seite her betrachten.

Beispiel:
Gehen wir davon aus, dass der normale Würfel in Serien von 12, 120 und 1200 Würfen gewürfelt wird. Dabei sei X die Anzahl der Sechsen. Wir simulieren die Anzahlen der Sechsen in diesen Serien je 1000 Mal (vgl. Abb. 7.18).

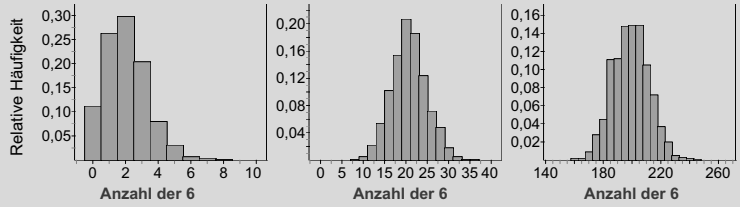


Abbildung 7.18: Simulation von je 1000 Wurfserien eines Würfels mit $n = 12, 120, 1200$ Würfen

Zunächst wird deutlich, dass absolut betrachtet die Streuung zunimmt. Dies entspricht einem empirischen Phänomen, nach dem bei größerem n mehr „Platz“ für Variation vorhanden ist. Es ergeben sich mit $\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$ der jeweils binomialverteilten Zufallsgrößen, die die Anzahl der Sechsen messen, die in der nachfolgenden Tabelle stehenden theoretischen Standardabweichungen. Diese entsprechen ungefähr den in der Simulation ermittelten (empirischen) Standardabweichungen (s). Wir betrachten in dieser Tabelle auch die mit der Anzahl der Würfe n normierten Standardabweichungen, also den Standardabweichungen *relativ* zur Stichprobengröße n.

n	σ	s	σ/n
12	1,291	1,327	0,108
120	4,082	4,112	0,034
1200	12,910	12,562	0,011

Für die normierten Standardabweichungen gilt: $\frac{\sigma_n}{n} = \frac{\sqrt{n \cdot p \cdot (1-p)}}{n} = \frac{\sqrt{p \cdot (1-p)}}{\sqrt{n}}$. Das bedeutet, dass bei der Erhöhung der Versuchsanzahl um einem Faktor m, im Beispiel ist dies $m = 10$, die normierte Standardabweichung der binomialverteilten Zufallsgrößen um den Faktor $\frac{1}{\sqrt{m}}$ verringert ist. Bezogen auf die Standardabweichung selbst bedeutet die Erhöhung der Versuchsanzahl um einen Faktor m eine Erhöhung der Standardabweichung um den Faktor $\sqrt{m}$.

Wir betrachten diese Erkenntnis, die auch als $\sqrt{n}$-Gesetz bezeichnet wird, aus einer anderen Perspektive bezogen auf den Erwartungswert einer binomialverteilten Zufallsgröße X. Erhöht man die Größe der Stichprobe n, so ist die Wahrscheinlichkeit einer festgelegten prozentualen Abweichung $\tilde{\varepsilon} > 0$ vom Erwartungswert geringer als bei kleineren Stichproben. Da für binomialverteilte Zufallsgrößen X gilt: $E(X) = n \cdot p$, können wir die prozentuale Abweichung auch durch $\varepsilon = \tilde{\varepsilon} \cdot p$ und damit in Abhängigkeit von n formulieren, also eine Abweichung von $\varepsilon \cdot n$ vom Erwartungswert betrachten.

Beispiel:
Gegeben seien wieder die Würfe eines Würfels im Beispiel oben. Unterscheiden wir die Wurfanzahl (die Größe der Stichprobe) durch einen Index, so ergibt sich in dem Beispiel $E_{12}(X) = 2;\ E_{120}(X) = 20$ und $E_{1200}(X) = 200$.
Wir betrachten nun das Ereignis, dass die Abweichung vom Erwartungswert größer als 50% des Erwartungswerts selbst beträgt ($\tilde{\varepsilon} = 0{,}5$), also die Zufallsgröße einen Wert außerhalb des Intervalls $[E(X) - 0{,}5 \cdot E(X); E(X) + 0{,}5 \cdot E(X)]$ annimmt. Bezogen auf den Umfang der Stichprobe n, wird also eine Abweichung größer als $\varepsilon = 0{,}5 \cdot p = 0{,}5 \cdot \frac{1}{6} = 1/12 \approx 0{,}083 = 8{,}3\%$ des Stichprobenumfangs vom Erwartungswert bzw. ein Wert außerhalb des Intervalls $[E(X) - \varepsilon \cdot n; E(X) + \varepsilon \cdot n]$ betrachtet. Es ergibt sich für die Stichprobenumfänge des Würfelbeispiels:

n	Intervall	Wahrscheinlichkeit	Gegenwahrscheinlichkeit
12	$[1;3]$	$P(1 \geq X \geq 3) \approx 0{,}493$	$\approx 0{,}507$
120	$[10;30]$	$P(10 \geq X \geq 30) \approx 0{,}986$	$\approx 0{,}014$
1200	$[100;300]$	$P(100 \geq X \geq 300) \approx 1$	≈ 0

Während also bei einer geringen Stichprobengröße von $n = 12$ eine Abweichung von mehr als $\frac{1}{12} \approx 8{,}3\%$ des Stichprobenumfangs (mehr als 50% des Erwartungswerts) noch recht wahrscheinlich ist, ist diese bei einer Stichprobengröße von $n = 1200$ nahezu 0.

Erweitert man diese Erkenntnis auf beliebige prozentuale Abweichungen $\varepsilon \cdot n$ vom Erwartungswert einer binomialverteilten Zufallsgröße, so ergibt sich, dass die Wahrscheinlichkeit dieser prozentualen Abweichung vom Erwartungswert gegen 0 geht, wenn man den Stichprobenumfang nur hinreichend (und theoretisch unbeschränkt) erhöht.

Beispiel:
Wir betrachten wiederum die binomialverteilte Zufallsgröße X: Anzahl der Sechsen bei n Wiederholungen des Würfelwurfs und geben diesmal mit $\varepsilon = 0{,}001$ eine prozentuale Abweichung von n vor und betrachten wiederum das symmetrische Intervall um den Erwartungswert $[E(X) - \varepsilon \cdot n; E(X) + \varepsilon \cdot n]$. Nun erhöhen wir den Stichprobenumfang n immer um den Faktor 10 und berechnen die Wahrscheinlichkeiten dafür, dass ein Wert der Zufallsgröße X *außerhalb* des genannten Intervalls angenommen wird $1 - P(E(X) - \varepsilon \cdot n \leq X \leq E(X) + \varepsilon \cdot n)$. Wir erhalten etwa beginnend bei $n = 600$:

n	Intervall	Wahrscheinlichkeit für Wert außerhalb des Intervalls
600	$[99;101]$	$\approx 0,913$
6000	$[994;1006]$	$\approx 0,835$
60000	$[9940;10060]$	$\approx 0,511$
600000	$[99400;100600]$	$\approx 0,038$
6000000	$[994000;1006000]$	≈ 0

Diesen bisher nur empirisch gesicherten Sachverhalt formulieren wir in einem Satz, der eine Form des **Bernoullischen Gesetzes der großen Zahlen** darstellt.

Satz 21
Sei X_n eine Zufallsgröße, die die Anzahl der Erfolge in einer Bernoulli-Kette vom Umfang n misst, und $\varepsilon \in \mathbb{R},\ \varepsilon > 0$, so gilt:

$$\lim_{n\to\infty} P(|X_n - E(X_n)| \geq \varepsilon \cdot n) = 0$$

Wir formulieren zunächst diesen Satz in einer alternativen Form, bei der wir statt X_n nun $\frac{X_n}{n}$, statt $E(X_n)$ nun $\frac{E(X_n)}{n} = p$ und statt $\varepsilon \cdot n$ nun ε verwenden:

Satz 22
Sei X_n eine Zufallsgröße, die die Anzahl der Erfolge in einer Bernoulli-Kette vom Umfang n misst, und $\varepsilon \in \mathbb{R},\ \varepsilon > 0$, so gilt:

$$\lim_{n\to\infty} P(|\frac{X_n}{n} - p| \geq \varepsilon) = 0$$

Die Bedeutung beider Satzvarianten ist groß. Sie schließen die Lücke zwischen Empirie und Theorie. So war bisher mit dem *empirischen Gesetz der großen Zahlen* die Stabilisierung der relativen Häufigkeiten bzw. der Mittelwerte gesichert. Anhand von speziellen Versuchen, bei denen die Wahrscheinlichkeit für ein Ereignis (p) bzw. der Erwartungswert ($E(X) = n \cdot p$) einer Bernoulli-Kette vom Umfang n zumindest im Modell bekannt war, konnte sogar eine Stabilisierung in der Nähe von p bzw. von $E(X)$ plausibel vermutet werden. Diese Beobachtung ist auch in die Interpretation der Wahrscheinlichkeit und des Erwartungswerts als Schätzung zukünftiger relativer Häufigkeiten bzw. zukünftiger Mittelwerte eingegangen.

Die beiden Satzvarianten sichern nun aber aus der theoretischen Perspektive die oben genannte empirische Erkenntnis: Wir führen n stochastisch unabhängige Bernoulli-Experimente aus und beobachten eine bestimmte Anzahl von Erfolgen. Ist nun n hinreichend groß, so ist die Wahrscheinlichkeit, dass die relative Häufigkeit der Erfolge sich wenig von der Erfolgs-Wahrscheinlichkeit unterscheidet, nahe 1. Gleiches gilt für eine vorab festgelegte Abweichung vom Erwartungswert.

Wir wollen den zweiten Satz (und dabei implizit auch den ersten) in drei Schritten beweisen. In einem ersten Schritt zeigen wir zunächst, dass der folgende Hilfssatz gilt:

Hilfssatz:
Sei Y eine Zufallsgröße, die nur nichtnegative Werte annimmt und sei $k \in \mathbb{R}^+$. Dann gilt:

$$P(Y \geq k) \leq \frac{E(Y)}{k}$$

Diesen Satz beweisen wir anhand der Definition des Erwartungswerts und anhand von zwei Abschätzungen:

$$\begin{aligned}
E(Y) &= \sum_{\omega \in \Omega} Y(\omega) \cdot P(\{\omega\}) = \sum_{\omega \in \Omega} Y(\omega) \cdot P(Y(\omega)) \\
&= \sum_{\substack{\omega \in \Omega \\ Y(\omega) < k}} Y(\omega) \cdot P(Y(\omega)) + \sum_{\substack{\omega \in \Omega \\ Y(\omega) \geq k}} Y(\omega) \cdot P(Y(\omega)) \\
&\geq \sum_{\substack{\omega \in \Omega \\ Y(\omega) \geq k}} Y(\omega) \cdot P(Y(\omega)) \\
&\quad \text{(Aufteilen der Summe durch den Wert } k \text{ und Betrachten nur der zweite Summe)} \\
&\geq k \cdot \sum_{\substack{\omega \in \Omega \\ Y(\omega) \geq k}} P(Y(\omega)) = k \cdot P(Y \geq k)
\end{aligned}$$

Im zweiten Schritt setzen wir $Y = (X - E(X))^2$, wobei X eine weitere beliebige Zufallsgröße sei und $k = (\varepsilon \cdot n)^2$, $\varepsilon \in \mathbb{R}^+$, $n \in \mathbb{N}$. Es gilt nun mit Ausnutzen der Definition der Varianz sowie des Hilfssatzes:

$$P(Y \geq (\varepsilon \cdot n)^2) = P\left((X - E(X))^2 \geq (\varepsilon \cdot n)^2\right) \leq \frac{E\left((X - E(X))^2\right)}{(\varepsilon \cdot n)^2} = \frac{V(X)}{(\varepsilon \cdot n)^2}$$

Da $(X - E(X))^2 \geq (\varepsilon \cdot n)^2$ äquivalent zu $|X - E(X)| \geq (\varepsilon \cdot n)$ und ebenso $|\frac{X}{n} - \frac{E(X)}{n}| = |\frac{X}{n} - p| \geq (\varepsilon)$ ist (die Ungleichungen haben die gleiche Lösungsmenge), gilt schließlich folgende Ungleichung, die **Tschebycheff-Ungleichung**:

$$P(|X - E(X)| \geq \varepsilon \cdot n) = P(|\frac{X}{n} - p| \geq \varepsilon) \leq \frac{V(X)}{(\varepsilon \cdot n)^2}$$

Die Tschebycheff-Ungleichung ermöglicht eine hier nicht weiter betrachtete Abschätzung der Abweichungen für Werte einer Zufallsgröße vom Erwartungswert dieser Zufallsgröße, wenn es sich um eine *beliebige* Zufallsgröße X handelt. Daher wollen wir in einem dritten Schritt allein *binomialverteilte* Zufallsgrößen betrachten und erhalten:

$$P(|X - E(X)| \geq \varepsilon \cdot n) = P(|\frac{X}{n} - p| \geq \varepsilon) \leq \frac{n \cdot p \cdot (1-p)}{(\varepsilon \cdot n)^2} = \frac{p \cdot (1-p)}{\varepsilon^2 \cdot n}$$

Für $n \to \infty$ geht der Term auf der rechten Seite der Ungleichung gegen 0, womit der Satz bewiesen ist.

Für den Abstand eines empirisch erhaltenen Werts einer Zufallsgröße X lassen sich durch die oben erzielten Ergebnisse Abschätzungen machen, die allerdings sehr grob sind. Dazu verwenden wir anstatt k im Hilfssatz das Produkt der Standardabweichung und einer reellen Zahl c, also $k = c \cdot \sigma(X)$. Es ergibt sich demnach:

$$P(|X - E(X)| \geq c \cdot \sigma) \leq \frac{n \cdot p \cdot (1-p)}{(c \cdot \sigma)^2} = \frac{1}{c^2}$$

bzw., wenn wir die Wahrscheinlichkeit dafür bestimmen, dass der Wert von X innerhalb des angegebenen Intervalls liegt:

$$P(|X - E(X)| \leq c \cdot \sigma) \geq 1 - \frac{n \cdot p \cdot (1-p)}{(c \cdot \sigma)^2} = 1 - \frac{1}{c^2}$$

Betrachtet man die sogenannten **Sigma-Umgebungen** für $c = 1,2$ und $c = 3$, so ist die Wahrscheinlichkeit, dass ein empirisch ermittelter Wert der Zufallsgröße

- innerhalb des 1-Sigma-Intervalls liegt, größergleich 0,
- innerhalb des 2-Sigma-Intervalls liegt, größergleich 0,75 und
- innerhalb des 3-Sigma-Intervalls liegt, größergleich 0,89

Bei binomialverteilten Zufallsgrößen sind, insbesondere, wenn diese annähernd symmetrisch sind, die Abschätzungen wesentlich genauer möglich.[5] Wir werden aber im Folgenden entweder eine Abschätzung per Simulation oder eine direkte Berechnung favorisieren.

7.3 Eigenschaften von Studierenden: Verteilungen

Wir betrachten im Folgenden drei Eigenschaften von Studierenden in unterschiedlichen Kontexten, konstruieren jeweils ein Modell und schätzen anhand dieses Modells eine zukünftige Stichprobe von Studierenden zu diesen Merkmalen ab.

Ein Berechnungsbeispiel: Aus der Stichprobe der Studierenden der PH Freiburg (Erhebung 2010) lässt sich folgende relative Häufigkeit ermitteln:

- A: *Altstudent* (mit einem Alter über 25 Jahre); $h_{218}(A) \approx 0,220$

Wir bezeichnen nun A als Erfolg, alle anderen möglichen Merkmalsausprägungen als Misserfolg und setzen schließlich $p = P(A) := 0,220$ in einem Bernoulli-Experiment.

Weiter kann man bei einer Stichprobe davon ausgehen, dass Studierende nur einmal befragt werden, ein passendes Modell also insbesondere das Ziehen ohne Zurücklegen aus einer Urne und entsprechend das Modell der hypergeometrischen Verteilung sinnvoll anzunehmen ist. Ist aber die Größe der Stichprobe klein in Bezug auf die Grundgesamtheit von etwa 4000 Studierenden, so ist auch das Modell der Binomialverteilung möglich. Um eine Prognose für eine zukünftige Stichprobe von $n = 218$ Studierenden zu treffen, verwenden wir beide genannten

[5] Das sind die sogenannten Sigma-Regeln, die auf der Kenntnis der Normalverteilung sowie der Kenntnis, dass die Normalverteilung die Grenzverteilung der Binomialverteilung ist, beruhen.

Modelle und simulieren zusätzlich das zufällige Ziehen von 218 Studierenden aus der Grundgesamtheit von 4000 Studierenden 1000 Mal. Dieses führen wir hinsichtlich des zuletzt genannten Merkmals des Alters der Studierenden durch.

Es ergeben sich die Verteilungen, die in Abbildung 7.19 zu sehen sind.

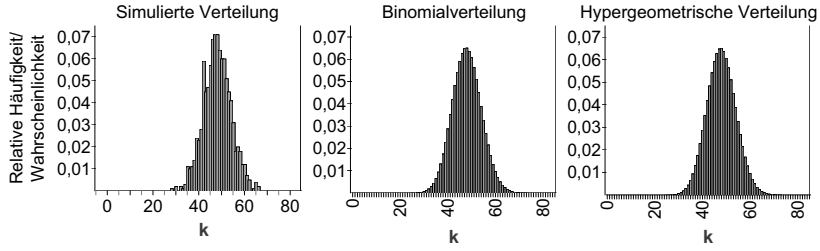


Abbildung 7.19: Simulierte Verteilung der Ziehung von 218 aus 4000 Studierenden, von denen 22% „alt“ sind, Binomialverteilung mit $n = 218$ und $p = 0,22$, hypergeometrische Verteilung mit $N = 4000, n = 218$ und $K = 880 = 0,22 \cdot 4000$

Während die simulierte Verteilung noch erkennbare Unterschiede zu den beiden theoretischen Verteilungen aufweist, ist zwischen den Letzteren kein Unterschied mehr erkennbar (aber bei exakter Berechnung durchaus noch in geringem Maße vorhanden). Wir betrachten Prognosen basierend auf allen drei Verteilungen. Zunächst kann analysiert werden, welche Anzahl von Altstudierenden erwartet werden kann (bei der simulierten Verteilung wird dazu das arithmetische *Mittel* als Schätzwert verwendet):

Modell	Erwartungswert
Simulation	$\bar{x} \approx 48$
Binomialverteilung	$E(X) = 0,22 \cdot 218 \approx 48$
hypergeometrische Verteilung	$E(X) = 218 \cdot \frac{880}{4000} \approx 48$

Um diesen bis auf Dezimalen identischen Erwartungswert konstruieren wir ein symmetrisches Intervall (Erwartungswert ± 10) und ermitteln die Wahrscheinlichkeit dafür, dass in einer zukünftigen Stichprobe eine Anzahl von Altstudierenden in diesem Intervall erzeugt wird (bei der Simulation bestimmen wir die Häufigkeit der Simulationen in diesem Intervall):

Modell	Intervall-Wahrscheinlichkeit
Simulation	$h_{218}(38 \leq X \leq 58) = \sum_{k=38}^{54} h_{218}(k) \approx 0,916$
Binomialverteilung	$P(38 \leq X \leq 58) = \sum_{k=38}^{54} \binom{218}{k} \cdot 0,22^k \cdot 0,78^{218-k} \approx 0,915$
hypergeometrische Verteilung	$P(38 \leq X \leq 58) = \sum_{k=38}^{54} \frac{\binom{880}{k}\binom{3120}{218-k}}{\binom{4000}{218}} \approx 0,923$

Hier sind Abweichungen zwischen den einzelnen Modellen erkennbar, die aber, beachtet man, dass alle drei Ansätze auf Modellen und damit nicht exakt auf der Realität basieren, als gering einzuschätzen sind. Auch wenn die Simulation des Vorgangs noch auf einer recht geringen Wiederholungsanzahl beruht, können offenbar alle drei Modelle nahezu als gleich geeignet für die Prognose einer zukünftigen Stichprobe angesehen werden. Das soll keinesfalls bedeuten, man solle nicht das bestmögliche Modell für eine Prognose wählen, sofern es mathematisch verfügbar ist und es bessere Einsichten ermöglicht als weniger ausgefeilte Modelle. Betrachten wir aber in diesem Beispiel die drei Modelle, so wird in allen die Prognose ausreichend ähnlich sein, dass nämlich auf lange Sicht in 92% der Stichproben eine Anzahl zwischen 38 und 58 Altstudierenden enthalten wird, sofern die Stichprobenerzeugung zufällig ist.

Ein Beispiel zur Beurteilung eines Modells: Wir betrachten als zweites Beispiel die Merkmalsausprägung der Raucher ($h_{218}(R) \approx 0,275$), die (zumindest auf mathematischer Ebene) als Erfolg bezeichnet wird. Wir legen das Modell der Binomialverteilung für X: Anzahl der Raucher mit $p = 0,275$ für eine zukünftige Stichprobe von 218 Studierenden zugrunde. Für diese Prognose betrachten wir die Binomialverteilung mit $p = 0,275$ und $n = 218$.

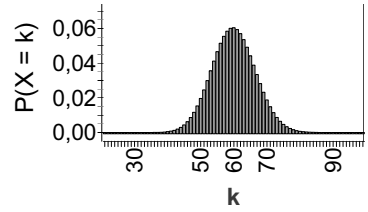


Abbildung 7.20: Binomialverteilung mit $n = 218$ und $p = 0,275$

Wann sollte man an dem Modell zweifeln? Angenommen, in der Stichprobe werden 65 Rauchende gezählt, obwohl nur $E(X) = 218 \cdot 0,275 \approx 60$ zu erwarten wären. Da $P(X \geq 65) \approx 0,243$ ist, kann auf der Basis dieses Modells solch ein Ereignis in fast einem Viertel der Fälle erwartet werden. Das ist sicher kein Grund der Ablehnung.

Angenommen, es werden 75 Raucher gezählt. Hier ist $P(X \geq 75) \approx 0,015$ schon recht klein. Nur in etwa 1,5% der Fälle könnte theoretisch mit solch einer Anzahl gerechnet werden. Was heißt das aber für das Modell hinsichtlich dieses *einen* Stichprobenergebnisses? Da gibt es zwei Möglichkeiten: Entweder man geht davon aus, in der einen Stichprobe ein seltenes, aber dennoch mögliches Ereignis beobachtet zu haben, oder man zweifelt nachhaltig an dem Modell und verwirft es (in diesem Fall würde man von einem höheren Anteil der Raucher ausgehen können). Verwirft man aber das Modell, so macht man auf lange Sicht bei 1,5% der so beurteilten Stichproben einen Fehler in diesem Modell, nämlich dann, wenn auf der Basis des Modells ein seltenes Ereignis mit 75 oder mehr Rauchern erzeugt wird.

Das ist die Idee des Testens von Hypothesen: Man bestimmt ein Intervall von Werten einer Zufallsgröße (im vorausgehenden Raucher-Beispiel war das $[0;74]$) und beurteilt ein Ereignis, in dem ein Wert der Zufallsgröße (wir hatten 75 Raucher angenommen) außerhalb des Intervalls liegt. Das Intervall kann statt *einseitig* auch *zweiseitig* und symmetrisch um den Erwartungswert

liegen, also z.B. im Beispiel $[46;74]$ sein. Bei der Beurteilung eines Ereignisses, in dem ein Wert der Zufallsgröße außerhalb des Intervalls liegt, haben sich in den Statistik verwendenden Wissenschaften (wie etwa der Psychologie) – ohne mathematisch tiefere Begründung – Konventionen durchgesetzt, Intervalle so zu konstruieren, dass die Wahrscheinlichkeit des Ereignisses „beobachteter Wert der Zufallsgröße liegt innerhalb des Intervalls“ 0,9 bzw. 0,95 oder 0,99 beträgt. Liegt dann der beobachtete Wert außerhalb des so gebildeten Intervalls, so beträgt die Wahrscheinlichkeit weniger als 0,1, was *tendenziell* an dem Modell zweifeln lässt, bzw. weniger als 0,05, was als **signifikant** bezeichnet wird, oder 0,01, was als **hochsignifikant** bezeichnet wird und wie ein signifikantes Ereignis zum Verwerfen des Modells führt. In dem vorausgehenden Raucher-Beispiel haben wir also ein signifikantes Ereignis, da basierend auf dem Modell die Wahrscheinlichkeit, dass der beobachtete Wert der Zufallsgröße ($X = 75$) außerhalb des konstruierten Intervalls $[0;74]$ liegt, kleiner als 0,05 ist. In diesem Fall hätte man in der Regel die Grenzen des Zweifels auf beiden Seiten des Erwartungswerts gezogen und wegen $P(49 \leq X \leq 71) \geq 0,95$ jede Anzahl außerhalb des Intervalls $[49;71]$ als signifikant bezeichnet und daraufhin das Modell verworfen.

Ein Beispiel zur Schätzung: In einem dritten Beispiel gehen wir von der unmittelbar einleuchtenden Vermutung aus, dass $h_{218}(W) \approx 0,711$ als relative Häufigkeit der weiblichen Studierenden an der PH Freiburg wie auch jede andere in der Stichprobe ermittelte Häufigkeit nicht exakt der tatsächlich in der Grundgesamtheit vorhandenen Häufigkeit entspricht. Der Anteil an weiblichen Studierenden an der PH Freiburg wird offiziell mit 0,76 angegeben. Nun ist aber die Frage, ob das zufällige Ergebnis in der Stichprobe (155 Studentinnen) zu dem als tatsächlich angenommenen Anteil von 0,76 passt. Wir stellen dazu die Binomialverteilung (dieses Modell legen wir wiederum zugrunde) für X: Anzahl der Studentinnen mit $p = 0,76$ und $n = 218$ dar.

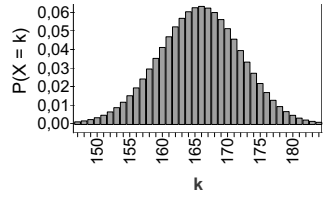


Abbildung 7.21: Binomialverteilung mit $n = 218$ und $p = 0,76$

Die in der Stichprobe erhaltene Anzahl von 155 weiblichen Studierenden (Anteil ca. 0,711) scheint noch zum offiziell angegebenen Anteil von 0,76 (vgl. Kap. 5.4) auf der Basis des gewählten Modells zu passen. So ist zwar die Wahrscheinlichkeit 155 oder weniger weibliche Studierende und damit auch weniger als die mit $p = 0,76$ erwarteten 166 Studentinnen mit rund 6% recht gering, das müsste aber noch nicht zur Ablehnung des Modells führen (s. o.).

Man kann sich nun fragen, zu welchen tatsächlichen Anteilen das empirische Ergebnis von 155 Studentinnen in einer Stichprobe von 218 Studierenden noch gepasst hätte. Unmittelbar einleuchtend ist, dass es sowohl zu einem Anteil größer 0,711 als auch zu einem Anteil kleiner 0,711 passen könnte. Wir könnten also systematisch symmetrische Intervalle um die in der Stichprobe ermittelte Häufigkeit bilden, begrenzt von zwei Modellparametern p_u und p_o ($[p_u < h_{218}(W) < p_o]$)

zweier Binomialverteilungen. Zusammen überdecken beide Binomialverteilungen die beobachtete Häufigkeit (hier 155) mit einer bestimmten Wahrscheinlichkeit α. Ein so gebildetes Intervall wird als **Konfidenzintervall** für den unbekannten Parameter p der Grundgesamtheit und α als **Konfidenzniveau** bezeichnet, wobei sich per Konvention wiederum $\alpha = 0,05$ und $\alpha = 0,01$ eingebürgert haben.

In unserem Beispiel suchen wir diese Grenzen für $\alpha = 0,05$ einerseits durch Simulation, andererseits durch Berechnung der Binomialverteilung mit p_u und p_o sowie $n = 218$. Per Simulation ergibt sich das Intervall $[0,650;0,772]$, so dass die obere ($p = 0,772$) und untere Binomialverteilung ($p = 0,650$) den Wert von 155 insgesamt mit einer Wahrscheinlichkeit von etwa 0,05 überlappen (vgl. Abb. 7.22).

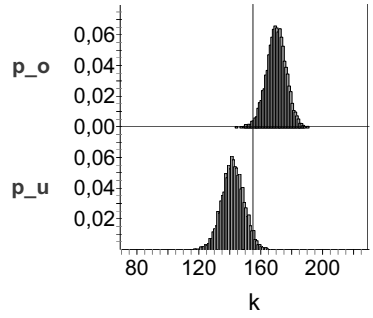


Abbildung 7.22: Binomialverteilungen mit $n = 218$ und $p_u = 0,650$ und $p_o = 0,772$ simuliert

Per Berechnung würde sich mit gerundeten Grenzen das Intervall $[0,651;0,772]$ ergeben, das nahezu identisch zu dem über Simulation ermittelten Intervall ist (vgl. Abb. 7.23).

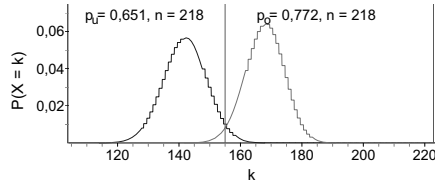


Abbildung 7.23: Binomialverteilungen mit $n = 218$ und $p_u = 0,651$ und $p_o = 0,772$ berechnet

Die Bedeutung eines solchen simulativ oder per Berechnung ermittelten Intervalls ist frequentistischer Natur: Da auf der Basis eines in aller Regel unbekannten Parameters p einer Zufallsgröße X in jeder Stichprobe zufällig eine Anzahl von Erfolgen k ergibt, ist das mit dieser Anzahl erzeugte Konfidenzintervall ebenso zufällig wie die beobachtete Anzahl k der Erfolge. Da man aber weiß, dass mit dem unbekannten p auf lange Sicht nur in 5% der Fälle ein Wert von k außerhalb eines Intervalls um den Erwartungswert $E(X)$ mit $P(E(X) - c \leq X \leq E(X) + c) \approx 0,95$ und $c \in \mathbb{R}$ liegt, werden auf lange Sicht auch 95% der zufällig erzeugten Konfidenzintervalle den Parameter p enthalten. Wir machen diesen letzten Aspekt am Beispiel der Studentinnen klar und gehen von dem hier ausnahmsweise (bis auf Rundungen) bekannten wahren Parameter von $p = 0,76$ für die weiblichen Studierenden an der PH aus. Mit diesem Parameter simulieren wir im

Modell der Binomialverteilung 100 Stichproben mit $n = 218$, also 100 verschiedene Umfragen unter 218 Studierenden. Jede Stichprobe ergibt eine neue relative Häufigkeit für das Ereignis *Studentin* und damit eine neues Konfidenzintervall für den Parameter p. Da p in diesem speziellen Fall bekannt ist, kann überprüft werden, wie viel Prozent der simulierten Intervalle p enthalten.

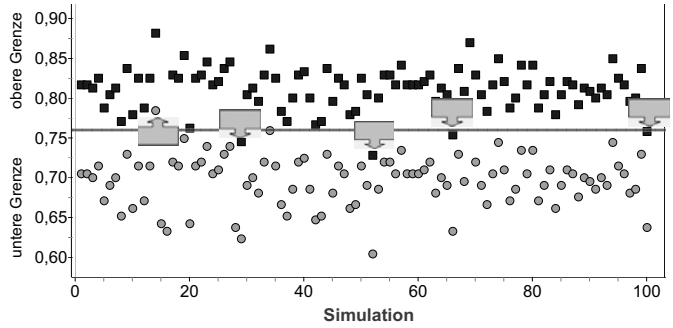


Abbildung 7.24: 100 Konfidenzintervalle zu Stichproben einer simulierten Binomialverteilung mit $n = 218$ und $p = 0,76$

In der Abbildung 7.24 markieren die Kreise die untere, die Quadrate die obere Grenze der Konfidenzintervalle. Tatsächlich hat sich in 5 der 100 simulierten Stichproben von 218 Studierenden eine Anzahl von Studentinnen ergeben, die zu einem Konfidenzintervall geführt haben, das den wahren Parameter nicht überdeckt (diese sind mit dem Pfeil markiert).[6]

7.4 Ergänzungen

7.4.1 Verteilungen

Im Hauptteil haben wir uns neben der Gleichverteilung auf die einfachen Klassen der Binomialverteilung und der hypergeometrischen Verteilung beschränkt. Einfach deshalb, da beide Verteilungen durch das Ziehen von Kugeln mit (Binomialverteilung) und ohne Zurücklegen (hypergeometrische Verteilung) aus einer Urne mit nur zwei Sorten von Kugeln (Erfolg und Misserfolg) erzeugt werden können.

Eine natürliche Erweiterung von Verteilungsmodellen basiert daher auf dem Ziehen mit und ohne Zurücklegen aus Urnen, in denen in bestimmten Anteilen $p_1, p_2, \ldots p_s$ nun s unterschiedliche Sorten von Kugeln liegen. Zieht man n Mal mit Zurücklegen, so ergibt sich, wenn man mit $X_1, X_2, \ldots, X_s$ die Zufallsgrößen für die Anzahl der k_i gezogenen Kugeln der Sorte i mit $i = 1, \ldots, s$ bezeichnet, folgende sogenannte **Multinomialverteilung**:

$$P(X_1 = k_1, X_2 = k_2, \ldots, X_s = k_s) = \frac{n!}{k_1! \cdot k_2! \cdot \ldots \cdot k_s!} \cdot p_1^{k_1} \cdot p_2^{k_2} \cdot \ldots \cdot p_s^{k_s}$$

[6] Dass bei 100 Simulationen *genau* 5 Konfidenzintervalle den wahren Parameter p nicht enthalten, ist Zufall! Ebenso könnte es weniger oder mehr solcher Intervalle geben. Erst auf lange Sicht, die mit 100 Simulationen noch nicht erreicht ist, ergibt sich tatsächlich der Anteil nicht treffender Konfidenzintervalle von etwa 5%.

Eine Herleitung dieser Formel könnte analog zur Binomialverteilung über die Betrachtung eines Baumes und der Identifizierung der maßgeblichen Äste und Pfade in diesem Baum erfolgen. Ebenso könnten Überlegungen zur verallgemeinerten hypergeometrischen Verteilung angestellt werden, wenn die Kugeln ohne Zurücklegen gezogen werden, die auf

$$P(X_1 = k_1, X_2 = k_2, ..., X_s = k_s) = \frac{\binom{K_1}{k_1}\binom{K_2}{k_2}...\binom{K_s}{k_s}}{\binom{N}{n}}$$

führt, wobei K_i die Anzahl der Kugeln einer Sorte, N die Gesamtanzahl der Kugeln und n die Anzahl der gezogenen Kugeln bezeichnet. Beide Verteilungen werden wir hier aber nicht weiter diskutieren (vgl. dazu z. B. Sachs, 1999).

7.4.2 Binomialkoeffizient

Der in Kapitel 7.2 hergeleitete Binomialkoeffizient ist der Ausdruck für die Zahlen im **Pascalschen Dreieck**. Dieses entsteht, beginnend bei der Spitze (in der 0-ten Zeile und 0-ten Spalte) mit dem Eintrag 1 so, dass die Einträge in den Zeilen darunter jeweils aus der Summe der Zahl in der darüber liegenden Zeile und vorhergehenden Spalte und der Zahl in der darüber liegenden Zeile und der identischen Spalte entstehen (siehe Tabelle). Alle nicht vorhandenen Einträge sind als Eintrag 0 anzusehen. Wir unterschieden mit dem Index $i = 0, 1, ..., r$ die Zeile und mit $j = 0, 1, ..., s$ die Spalte, so dass jede Zahl $a \in \mathbb{N}$ im Dreieck durch $a_{i,j}$ beschrieben werden kann.

Spalte / Zeile	0	1	2	3	4	5	6	7	8	9
0	1									
1	1	1								
2	1	2	1							
3	1	3	3	1						
4	1	4	6	4	1					
5	1	5	10	10	5	1				
6	1	6	15	20	15	6	1			
7	1	7	21	35	35	21	7	1		
8	1	8	28	56	70	56	28	1		
9	1	9	36	84	126	126	84	36	9	1

Formal entsteht also ein Element $a_{i,j}$ durch $a_{i,j} = a_{i-1,j-1} + a_{i-1,j}$, Beispielsweise gilt demnach für das Element in der 6. Zeile und 2. Spalte: $a_{6,2} = a_{5,1} + a_{5,2} = 5 + 10 = 15$.

Die Einträge im Pascalschen Dreieck entsprechen dem Binomialkoeffizienten $\binom{i}{j}$. So ist etwa $\binom{6}{2} = \frac{6!}{2!\cdot 4!} = 15$. Weiter folgt aus der Konstruktion des Pascalschen Dreiecks unmittelbar $\binom{i}{j} = \binom{i-1}{j-1} + \binom{i-1}{j}$, wodurch sich ein beliebiger Binomialkoeffizient rekursiv berechnen lässt, sofern $\binom{0}{0} := 1$ und $\binom{1}{0} := 1$ festgelegt werden.

7.4.3 Erwartungswerte, Varianzen

In Kapitel 7.2 sind wir die Beweise für die allgemeine Herleitung der Erwartungswerte und Varianzen der Binomialverteilung wie auch der hypergeometrischen Verteilung schuldig geblieben. Wir werden diese Beweise im Folgenden skizzieren.

Erwartungswert und Varianz einer binomialverteilten Zufallsgröße X Wir verwenden die Definition des Erwartungswerts einer binomialverteilten Zufallsgröße X sowie im letzten Schritt den binomischen Satz $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$, der wiederum mit vollständiger Induktion bewiesen werden kann.

$$\begin{aligned}
E(X) &= \sum_{r=0}^{n} r \cdot \binom{n}{r} p^r \cdot (1-p)^{n-r} \\
&= 0 + \sum_{r=1}^{n} r \cdot \frac{n!}{r! \cdot (n-r)!} p^r \cdot (1-p)^{n-r} \\
&= \sum_{r=1}^{n} r \cdot \frac{n \cdot (n-1)!}{r \cdot (r-1)! \cdot (n-r)!} p \cdot p^{r-1} \cdot (1-p)^{n-r} \\
&= np \cdot \sum_{r=1}^{n} \frac{(n-1)!}{(r-1)! \cdot (n-r)!} p^{r-1} \cdot (1-p)^{n-r} \\
&\quad \text{(Setze } k = r-1\text{.)} \\
&= np \cdot \sum_{k=0}^{n-1} \frac{(n-1)!}{k! \cdot (n-1-k)!} p^k \cdot (1-p)^{n-1-k}
\end{aligned}$$

Setzt man nun in der Summe $m = n-1$, so ergibt sich:

$$E(X) = np \cdot \underbrace{\left[\sum_{k=0}^{m} \frac{m!}{k! \cdot (m-k)!} p^k \cdot (1-p)^{m-k} \right]}_{\text{Binomischer Satz}} = np \cdot \underbrace{[p + (1-p)]^m}_{=1} = np$$

Für die Varianz gilt allgemein $V(X) = E(X^2) - E(X)^2$ (vgl. Kap. 7.2), wobei für eine binomialverteilte Zufallsgröße X unmittelbar $E(X)^2 = (np)^2$ folgt. Um den Term $E(X^2)$ zu berechnen, gehen wir wie im vorangehenden Beweis vor.

$$\begin{aligned}
E(X^2) &= \sum_{r=0}^{n} r^2 \cdot \binom{n}{r} p^r \cdot (1-p)^{n-r} \\
&= 0 + \sum_{r=1}^{n} r^2 \cdot \frac{n!}{r! \cdot (n-r)!} p^r \cdot (1-p)^{n-r} \\
&= \sum_{r=1}^{n} r^2 \cdot \frac{n \cdot (n-1)!}{r \cdot (r-1)! \cdot (n-r)!} p \cdot p^{r-1} \cdot (1-p)^{n-r} \\
&= np \cdot \sum_{r=1}^{n} r \cdot \frac{(n-1)!}{(r-1)! \cdot (n-r)!} p^{r-1} \cdot (1-p)^{n-r}
\end{aligned}$$

Setze wiederum $k = r-1$:

$$= \quad np \cdot \sum_{k=0}^{n-1} (k+1) \cdot \frac{(n-1)!}{k! \cdot (n-1-k)!} p^k \cdot (1-p)^{n-1-k}$$

Setze in der Summe $m = n-1$:

$$\begin{aligned}
&= \quad np \cdot \sum_{k=0}^{m} (k+1) \cdot \frac{m!}{k! \cdot (m-k)!} p^k \cdot (1-p)^{m-k} \\
&= \quad np \left[\underbrace{\sum_{k=0}^{m} k \cdot \frac{m!}{k! \cdot (m-k)!} p^k \cdot (1-p)^{m-k}}_{= \, mp \, = \, (n-1)p \text{ (wie oben)}} + \underbrace{\sum_{k=0}^{m} \frac{m!}{k! \cdot (m-k)!} p^k \cdot (1-p)^{m-k}}_{= \, 1 \text{ (wie oben)}} \right] \\
&= \quad np[(n-1)p+1] \\
&= \quad np(np-p+1) \\
&= \quad (np)^2 + np(1-p)
\end{aligned}$$

Damit ist $V(X) = E(X^2) - E(X)^2 = (np)^2 + np(1-p) - (np)^2 = np(1-p)$.

Erwartungswert und Varianz einer hypergeometrisch verteilten Zufallsgröße X Wir gehen beim Erwartungswert der hypergeometrisch verteilten Zufallsgröße X mit der gleichen Strategie wie oben vor und nutzen dabei folgende Beziehung:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n \cdot (n-1)!}{k \cdot (k-1)! \cdot (n-k)!} = \frac{n}{k} \cdot \binom{n-1}{k-1}$$

Wir erhalten damit:

$$\begin{aligned}
E(X) &= \sum_{r=0}^{n} r \cdot \frac{\binom{K}{r}\binom{N-K}{n-r}}{\binom{N}{n}} \\
&= 0 + \sum_{r=1}^{n} r \cdot \frac{\frac{K}{r}\binom{K-1}{r-1}\binom{N-K}{n-r}}{\frac{N}{n}\binom{N-1}{n-1}} \\
&= n \cdot \frac{K}{N} \sum_{r=1}^{n} \frac{\binom{K-1}{r-1}\binom{N-K}{n-r}}{\binom{N-1}{n-1}} \\
&= n \cdot \frac{K}{N} \sum_{r=1}^{n} \frac{\binom{K-1}{r-1}\binom{N-1+1-K}{n-1+1-r}}{\binom{N-1}{n-1}}
\end{aligned}$$

Setzt man $k = r-1$, erhält man:

$$E(X) = n \cdot \frac{K}{N} \underbrace{\sum_{k=0}^{n-1} \frac{\binom{K-1}{k}\binom{N-1-(K-1)}{n-1-k}}{\binom{N-1}{n-1}}}_{= \, 1} = n \cdot \frac{K}{N}$$

hypergeometrische Verteilung mit $(N-1)$, $(K-1)$ und $(n-1)$

Für die Varianz nutzen wir erneut die gleiche Strategie sowie $V(X) = E(X^2) - E(X)^2$. Wir berechnen zunächst $E(X^2)$:

$$E(X^2) = \sum_{r=0}^{n} r^2 \cdot \frac{\binom{K}{r}\binom{N-K}{n-r}}{\binom{N}{n}} = \ldots = n \cdot \frac{K}{N} \sum_{r=1}^{n} r \frac{\binom{K-1}{r-1}\binom{N-1-(K-1)}{n-1-(r-1)}}{\binom{N-1}{n-1}}$$

Setzt man wiederum $k = r-1$, so erhält man:

$$\begin{aligned} E(X^2) &= n \cdot \frac{K}{N} \sum_{k=0}^{n-1} (k+1) \frac{\binom{K-1}{k}\binom{N-1-(K-1)}{n-1-k}}{\binom{N-1}{n-1}} \\ &= n \cdot \frac{K}{N} \left[\underbrace{\sum_{k=0}^{n-1} k \frac{\binom{K-1}{k}\binom{N-1-(K-1)}{n-1-k}}{\binom{N-1}{n-1}}}_{= (n-1)\frac{K-1}{N-1}} + \underbrace{\sum_{k=0}^{n-1} \frac{\binom{K-1}{k}\binom{N-1-(K-1)}{n-1-k}}{\binom{N-1}{n-1}}}_{= 1} \right] \\ &= n \cdot \frac{K}{N} \cdot \left((n-1)\frac{K-1}{N-1} + 1 \right) \end{aligned}$$

Es ergibt sich also insgesamt:

$$\begin{aligned} V(X) &= E(X^2) - E(X)^2 = n \cdot \frac{K}{N} \cdot \left((n-1)\frac{K-1}{N-1} + 1 \right) - n^2 \frac{K^2}{N^2} \\ &= n \cdot \frac{K}{N} \cdot \left((n-1)\frac{K-1}{N-1} + 1 - n\frac{K}{N} \right) \\ &= n \cdot \frac{K}{N} \cdot \left(\frac{nK - K - n + 1 + N - 1}{N-1} - \frac{nK}{N} \right) \\ &= n \cdot \frac{K}{N} \cdot \left(\frac{NnK - NK - Nn + N^2 - NnK + nK}{N(N-1)} \right) \\ &= n \cdot \frac{K}{N} \cdot \left(\frac{N^2 - Nn}{N(N-1)} - \frac{K(N-n)}{N(N-1)} \right) \\ &= n \cdot \frac{K}{N} \cdot \left(\frac{N-n}{N-1} - \frac{K}{N} \cdot \frac{N-n}{N-1} \right) \\ &= n \cdot \frac{K}{N} \cdot \left(1 - \frac{K}{N} \right) \cdot \frac{N-n}{N-1} \end{aligned}$$

Die letzten Umformungen sind vom Ende her gedacht, um die endgültige Form zu erhalten, in der man mit Ersetzung von $\frac{K}{N}$ durch p die Struktur der Varianz der Binomialverteilung mit einem weiteren Faktor $\frac{N-n}{N-1}$ erkennt.

Da der Erwartungswert der hypergeometrischen Verteilung bei festem n identisch zu dem der Binomialverteilung ist mit $p = \frac{K}{N}$, erkennt man, dass für N sehr groß im Gegensatz zu n auch die Varianz beider Verteilungen fast identisch ist, da in diesem Fall $\frac{N-n}{N-1} \approx 1$ gilt.

7.5 Aufgaben

Aufgabe 7.1: Gegeben sei ein Würfel, der die Form eines Tetraeders hat und die Augenzahlen 1 bis 4 aufweist.

a) Bestimmen Sie die Wahrscheinlichkeitsverteilung X: *Augensumme* für den zweifachen Wurf des Tetraeders.

b) Bestimmen Sie Erwartungswert, Varianz und Standardabweichung von X.

Der Doppelwurf des Tetraeders wird nun 100 Mal durchgeführt. Es wird die Anzahl der Augensumme 6 betrachtet

c) Begründen Sie, dass X: *Anzahl der Augensumme 6* als binomialverteilte Zufallsgröße modelliert werden kann.

d) Bestimmen Sie Erwartungswert, Varianz und Standardabweichung für X_i: *Augensumme 6 im i-ten Doppelwurf.*

e) Bestimmen Sie Erwartungswert, Varianz und Standardabweichung für X: *Augensumme 6* für die 100 Doppelwürfe.

f) Zeichnen Sie die Wahrscheinlichkeitsverteilung von X.

g) Ab welcher Anzahl der Augensumme 6 würden Sie an dem Modell für p, das Sie oben aufgestellt haben, zweifeln?

h) Bestimmen Sie die Wahrscheinlichkeit für das Intervall $[E(X)-\sigma; E(X)+\sigma]$.

i) Bestimmen Sie ein symmetrisch zum Erwartungswert von X konstruiertes Intervall $I=[a,b]$ so, dass die Wahrscheinlichkeit für $P(a \leq X \leq b) \approx 0,95$ ist.

Aufgabe 7.2: Das Zahlenlotto 6 aus 49 soll analysiert werden. Bei diesem Zahlenlotto werden 6 Erfolgskugeln aus einer Urne mit 49 Kugeln mit den Zahlen von 1 bis 49 als Aufschrift zufällig und ohne Zurücklegen gezogen. Die Reihenfolge der Erfolgskugeln ist dabei unerheblich, da diese am Ende der Ziehung der Größe nach geordnet werden.

a) Ein Spieler tippt 6 Zahlen. Bestimmen Sie die Wahrscheinlichkeitsverteilung für X: *Anzahl der richtig getippten Zahlen*. Eine richtig getippte Zahl sei dann vorhanden, wenn sie mit der Aufschrift einer der Erfolgskugeln gleich ist.

b) Bestimmen Sie die Wahrscheinlichkeit, dass mindestens 3 richtige Zahlen bzw. höchstens 2 richtige Zahlen getippt werden.

Ein Spieler gibt jeden Samstag, also etwa 52 Mal im Jahr, jeweils einen Tipp von 6 Zahlen ab.

c) Begründen Sie, dass X: *Anzahl der Dreier (genau drei richtige Zahlen)* bei n Samstagen eine binomialverteilte Zufallsgröße ist.

d) Welche Anzahl von Dreiern kann der Spieler in einem, zwei, m Jahren erwarten?

Aufgabe 7.3: Gegeben ist eine Urne mit 2 weißen und 8 schwarzen Kugeln. Aus dieser wird 4 Mal ohne Zurücklegen gezogen.

a) Bestimmen Sie die Wahrscheinlichkeitsverteilung für X: *Anzahl der schwarzen Kugeln*.

b) Simulieren Sie diesen Versuch 10, 100 und 1000 Mal und vergleichen Sie die simulierte Häufigkeit der schwarzen Kugeln mit der Wahrscheinlichkeitsverteilung.

c) Bestimmen Sie die Wahrscheinlichkeitsverteilung von X unter der Voraussetzung, dass die Kugeln nach dem Zug wieder zurückgelegt wird. Vergleichen Sie die entstehende Verteilung mit der in a) ermittelten.

Aufgabe 7.4:

a) Teilen Sie beliebige Merkmale zu den Studierenden (siehe Zusatzmaterial) dichotom auf (mit zwei Merkmalsausprägungen) und bezeichnen diese als Erfolg/Misserfolg eines zufälligen Vorgangs. Setzen Sie anhand der relativen Häufigkeiten einer Merkmalsausprägung die Wahrscheinlichkeiten für Erfolg und Misserfolg fest.

b) Berechnen Sie die Wahrscheinlichkeitsverteilungen jeweils für X: *Anzahl der Erfolge* bei Stichprobenumfängen von $n = 10$ und $n = 100$. Begründen Sie jeweils die Wahl des von Ihnen verwendeten Modells einer Wahrscheinlichkeitsverteilung.

c) Bestimmen Sie die Wahrscheinlichkeiten für die Intervalle $I_1 = [E(X) - \sigma; E(X) + \sigma]$, $I_2 = [E(X) - 2\sigma; E(X) + 2\sigma]$, $I_3 = [E(X) - 3\sigma; E(X) + 3\sigma]$

Aufgabe 7.5: „On August 18, 1913, at the casino in Monte Carlo, black came up a record tewnty-six times in succession [in roulette]. ... [There] was a near-panicky rush to bet on red, beginning about the time black had come up a phenomenal fifteen times. In application of the maturity [of the chances] doctrine, players doubled and tripled their stakes, this doctrine leading them to believe after black came up the twentieth time that there was not a chance in a million of another repeat. In the end the unusual run enriched the Casino by some millions of francs. “ (Huff & Geis, 1959, pp. 28–29)
Der Gewinn des Casinos an diesem Tag beruhte auf der auch unter Spielern weitverbreiteten Fehlvorstellung *gambler's fallacy* (vgl. Kap. 6.1, S. 118 Fußnote).

a) Bestimmen Sie die Wahrscheinlichkeit für die Farbe rot im 21. Spiel und die Wahrscheinlichkeit, dass die Farbe schwarz 21 Mal in Serie auftritt. (Zur Erinnerung: Das Rouletterad hat 37 Felder, die von 0 bis 36 durchnummeriert sind. 18 Felder sind rot, 18 Felder sind schwarz gefärbt. Die 0 („Zero“) ist grün.)

c) Wie wahrscheinlich tritt bei 100 Spielrunden eine Serie von 5 mal schwarz hintereinander auf? Simulieren Sie diesen Versuch 10, 100 und 1000 Mal und leiten Sie daraus Aussagen über die Wahrscheinlichkeit P(Anzahl von 5 mal schwarz hintereinander) bei einem Stichprobenumfang von $n = 100$ Spielen ab.

d) Entwickeln Sie eine rekursive Formel, mit der die Wahrscheinlichkeit errechnet werden kann, dass bei 100 Spielrunden k mal schwarz hintereinander auftritt. Setzen Sie $k = 5$ und überprüfen Sie Ihre Formel mit den Simulationsergebnissen aus der vorangehenden Teilaufgabe c).

8 Daten beurteilen mit Simulationen

Einstiegsbeispiel

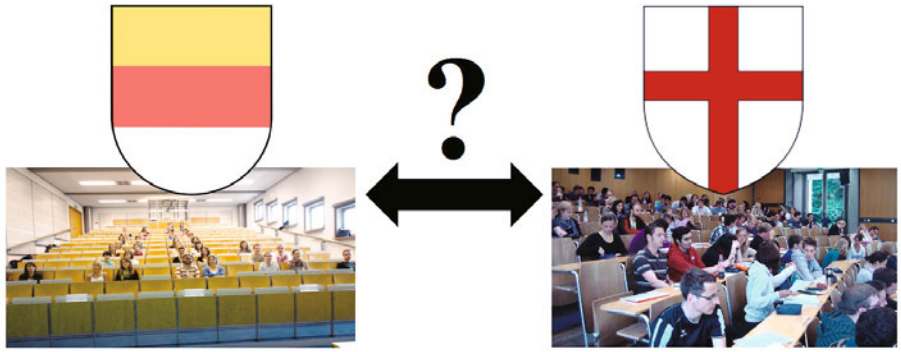


Abbildung 8.1: Münster und Freiburg, Uni und PH im Vergleich

Aufgabe 1: Identifizieren Sie Unterschiede zwischen den Studierenden in Münster und Freiburg und beurteilen Sie, ob diese zufällig oder systematisch sind.

Worum es geht

In den vorangegangenen Kapiteln hatten wir häufiger Unterschiede zwischen den Studierenden (beider Hochschulen, beider Geschlechter usw.) notiert und die Aussagekraft der deskriptiven Aussagen zu beurteilen versucht, konnten aber nicht entscheiden, ob diese zufällig waren oder auf systematischen Unterschieden basierten. So ergeben sich in aller Regel bei zwei Stichproben (auch aus der gleichen Grundgesamtheit, z.B. bei zwei Serien von Würfen des normalen Würfels) unterschiedliche Ergebnisse, etwa unterschiedliche Mittelwerte der Augenzahlen. Dadurch kann aber noch nicht auf einen systematischen Unterschied des Würfels in beiden Serien geschlossen werden, sondern die Unterschiede lassen sich (vermutlich) allein mit der Zufälligkeit der Ausprägungen verschiedener Stichproben begründen. Wir wollen in diesem Kapitel gerade diese wichtige Unterscheidung zufälliger und systematischer Abweichungen und – in Erweiterung der Aufgabe oben – insbesondere auch die eines empirischen Ergebnisses von einem theoretischen Modell untersuchen.

Mit Hilfe von Simulationen und Verteilungsmodellen haben wir bereits im vorangegangenen Kapitel 7 begonnen, empirische Phänomene auf der Basis eines vorgegebenen Modells zu be-

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_8,

urteilen. Diese Beurteilung mit Hilfe von Simulationen wollen wir in diesem Abschnitt propädeutisch ausbauen, da für die folgenden Fragestellungen in der Regel das mathematische Handwerkszeug in diesem Band nicht zur Verfügung gestellt wurde bzw. wird. Mit dieser propädeutischen, also fachlich verkürzten und auf Phänomene beschränkten Betrachtung geht es uns im Kern darum, die Idee der statistisch basierten Beurteilung empirischer Datenphänomene so zu diskutieren, dass in einem späteren Schritt die entsprechenden mathematischen Verfahren auf einer vorhandenen inhaltlichen Grundvorstellung aufgebaut werden können. Dadurch wird sich in diesem Kapitel auch die Abfolge der Teilkapitel verändern. So werden wir direkt mit der Erkundung der Datensätze zu den Studierenden einsteigen und erst danach in den Ergänzungen wenige fachliche Überlegungen über die Beurteilung von Simulationsergebnissen hinaus machen.

8.1 Eigenschaften von Studierenden: Beurteilungen von Modellen

8.1.1 Tests

Berechnung und Simulation: In einem ersten Beispiel werden wir der Simulation noch eine Berechnung zur Seite stellen und anhand eines Beispiels die Repräsentativität der Stichprobe zu den Studierenden aus Münster beurteilen.

Für die Universität Münster wird offiziell ein Anteil von 0,53 weiblicher Studierender angegeben. In der vorliegenden Stichprobe haben wir allerdings einen Anteil von $\frac{617}{1081}$, also etwas mehr als $0,57$, erhalten. Wir stellen zwei Fragen:

1. Wie wahrscheinlich ist eine Abweichung von mehr als 0,04, also 4% bei einer Stichprobe von 1081 Studierenden unter der Annahme, dass der Anteil von 0,53 (bis auf Rundungen) korrekt ist? Absolut bedeutet das, eine Anzahl von weniger als 530 oder mehr als 616 Studentinnen zu erhalten (zweiseitiger Signifikanztest).
2. Wie wahrscheinlich ist es, dass 617 oder mehr Studierende in einer Stichprobe von 1081 Studierenden weiblich sind, unter der Annahme, dass der Anteil von 0,53 (bis auf Rundungen) korrekt ist (einseitiger Signifikanztest)?

Welches Modell ist zu wählen? Da es rund 37 000 Studierende an der Universität Münster gibt, von denen 1081 erhoben wurden, ist $N = 37\,000$ groß gegenüber $n = 1081$. Neben den Erwartungswerten eines hypergeometrischen Modells einerseits und eines binomialen Modells andererseits sind auch die Varianzen in beiden Fällen nahezu identisch, da der Term $\frac{N-n}{N-1} \approx 0,97$, der die Varianz beider Modelle unterscheidet, fast 1 ist (vgl. Kap. 7.4). Wir greifen deshalb sowohl bei der Simulation als auch bei der Berechnung auf das Modell stochastisch unabhängiger Bernoulli-Experimente zurück.

Für eine Stichprobe vom Umfang $n = 1081$ bezogen auf eine binomialverteilte Zufallsgröße X: Anzahl weiblicher Studierende mit $p = 0,53$ ergeben die Berechnung und eine Simulation der genannten binomialverteilten Zufallsgröße das in Abbildung 8.2 dargestellte Ergebnis. Die Form der Verteilungen ist ähnlich, auch wenn die simulierte Verteilung noch vergleichsweise unregelmäßig erscheint. Wir betrachten nun die beiden Verteilungen hinsichtlich der oben gestellten Fragen.

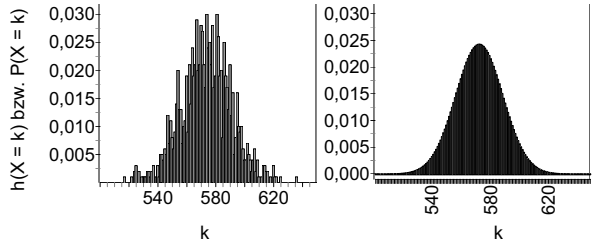


Abbildung 8.2: Simulierte und berechnete Binomialverteilung mit $p = 0,53$ und $n = 1081$

Eine Abweichung von mehr als 4% vom als wahr angenommenen Anteil von 0,53 würde bedeuten, dass in der Stichprobe weniger als $0,49 \cdot 1081 \approx 530$ bzw. mehr als $0,57 \cdot 1081 \approx 616$ Studentinnen vorhanden sind. Es ergibt sich bei der Simulation $h_{1081}(530 \leq X \leq 616) = 0,988$ sowie durch Berechnung $P(530 \leq X \leq 616) \approx 0,991$. Zwar sind auch diese beiden Ergebnisse noch leicht unterschiedlich, erzeugen aber die gleiche Aussage: Eine Abweichung von über 4% ist in der Stichprobe vom Umfang $n = 1081$ sehr unwahrscheinlich und kommt auf lange Sicht nur in etwa 1% der Stichproben vor (z. B. $1 - 0,988 = 0,012$). Der Zweifel daran, dass die Stichprobe bezogen auf den Anteil der Studentinnen dem Modell $P(W) = 0,53$ genügt, führt dazu, das Modell aufgrund des *signifikanten* Stichprobenresultats zu verwerfen. In etwa 1% der so erzeugten Stichproben macht man dabei allerdings den Fehler, das Modell fälschlicherweise zu verwerfen, da die Stichprobe ein seltenes, aber dennoch mögliches Ereignis erzeugt hat.

Für die Stichprobe bedeutet das Verwerfen des Modells, dass diese bezogen auf den Anteil der weiblichen Studierenden offenbar nicht repräsentativ war, da die weiblichen Studierenden *überrepräsentiert* sind. Das kann nicht zuletzt daran liegen, dass bei den Studierenden des Lehramts überproportional viele Lehramtsstudierende der Schulformen Grund- und Hauptschule befragt worden waren, bei denen der Anteil an Studentinnen hoch ist, so dass dadurch die Stichprobe zumindest hinsichtlich des Merkmals Geschlecht verzerrt wurde.

Die gleiche Aussage ergibt sich in noch verschärfter Form bei der zweiten Frage, da hier das Stichprobenergebnis nur in eine Richtung (mehr als 616 Studentinnen) überprüft wird. Hierbei ergibt sich $h_{1081}(X > 616) = 0,005$ und $P(X > 616) \approx 0,003$.

Für das hier verwendete Verfahren der Simulation ist es wichtig herauszustellen, dass das Simulationsergebnis zur identischen Interpretation führt wie das berechnete Ergebnis. Die Simulation kann sich zwar nicht auf ein formalisiertes algorithmisches Verfahren stützen und birgt damit das Risiko, Begründungszusammenhänge nicht entsprechend fundieren zu können. Sie erzeugt aber eine Entscheidungshilfe, die dem formalisierten Verfahren hier nicht nachsteht.

Test mit Simulation: Im zweiten Beispiel werden wir allein mit Simulationen arbeiten, da wir in diesem Buch auf die Darstellung der entsprechenden mathematischen Verfahrensweisen verzichtet haben, die eine Beurteilung auf algorithmisch-formalen Wege erlauben würde. In Kapitel 3.5 hatten wir den Zusammenhang zwischen dem Studienort (M: Münster und F: Freiburg) und dem Rauchverhalten (R: Raucher und $\overline{R}$: Nichtraucher) untersucht und auf dem deskriptiven Wege die in Abbildung 8.3 dargestellten Ergebnisse in der Vierfeldertafel und im Einheitsquadrat erhalten:

	M	F	Summe
R	270	124	394
$\overline{R}$	810	272	1082
Summe	1080	396	1476

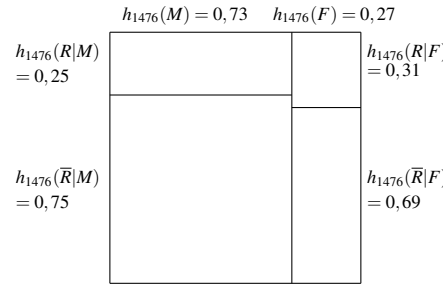


Abbildung 8.3: Zusammenhang zwischen Studienort und Rauchverhalten

Den beobachteten Unterschied hinsichtlich des Rauchverhaltens in den Stichproben beider Hochschulen hatten wir durch ein einfaches Assoziationsmaß A und die odds ratio ρ gemessen, nämlich durch:

$$A = h_{1476}(R|M) - h_{1476}(R|F) \approx 0,25 - 0,31 = -0,06; \quad \rho = \frac{270 \cdot 272}{810 \cdot 124} \approx 0,73$$

Beide Werte, A nahe 0 und ρ nahe 1, deuten auf die Unabhängigkeit der Merkmale.

Beurteilen lassen sich die Werte von A und ρ bezogen auf die Größe der Stichprobe nur mit wahrscheinlichkeitstheoretischen Mitteln oder aber zumindest annähernd mit einer Simulation. Wie bei einem formalen Vorgehen basiert auch bei der Simulation die Beurteilung des beschriebenen Unterschieds auf einem Modell:

- Es wird davon ausgegangen, dass es *keinen* Zusammenhang zwischen Rauchverhalten und Hochschulzugehörigkeit gibt, d.h. die *Hypothese* lautet: Es gibt keinen Zusammenhang zwischen Hochschulzugehörigkeit und Rauchverhalten.
- Als erstes Merkmal wird mit den zwei Merkmalsausprägungen Münster und Freiburg sowie insgesamt 1476 Merkmalsträgern (1080 aus Münster, 396 aus Freiburg) die Hochschulzugehörigkeit gesetzt. Die Zuordnung der 394 Raucher und 1082 Nichtraucher zu den Merkmalsträgern der verschiedenen Hochschulen wird *zufällig* vorgenommen und damit das Modell der Unabhängigkeit beider Merkmale simuliert.
- Diese Simulation wird 1000 Mal wiederholt und der Anteil derjenigen Zuordnungen bestimmt, für den gilt $|A| \geq 0,06$ bzw. $\rho \leq 0,73$. Hat das Ereignis von $|A| \geq 0,06$ bzw. $\rho \leq 0,73$ eine Häufigkeit von weniger als 0,05 (alternativ: Wenn das Ereignis $|A| < 0,06$ bzw. $\rho > 0,73$ in mehr als 95% der Fälle auftritt), so wird der Unterschied als signifikant bewertet. Das Modell der Unabhängigkeit, das das unwahrscheinliche Ereignis produziert hat, könnte verworfen und angenommen werden, dass ein systematischer Zusammenhang zwischen Rauchverhalten und Hochschulzugehörigkeit besteht.

Bildlich wird folgendermaßen vorgegangen. Die 1080 Studierenden aus Münster und die 396 Studierenden aus Freiburg haben bei der Erfassung ein Schild angeheftet bekommen, auf dem

den Tatsachen entsprechend ihr Rauchverhalten steht. Nun werden diese Schilder abgenommen, gut durchmischt und anschließend zufällig den Studierenden wieder ausgeteilt. Zuletzt wird wie oben die absolute Häufigkeit in der Vierfeldertafel sowie das Assoziationsmaß A und die odds ratio ρ bestimmt. In Abbildung 8.4 sind links die Vierfeldertafel sowie das Assoziationsmaß *einer* Simulation, rechts das Ergebnis für das Assoziationsmaß sowie für die odds ratio nach 1000 Simulationen dargestellt (beim odds ratio sind Werte ρ, die größer 1 sind, als $1/\rho$ dargestellt).

	M	F	Summe
R	284	110	394
$\overline{R}$	796	286	1082
Summe	1080	396	1476

$$A \approx 0,263 - 0,278 = -0,015$$

$$\rho \approx \frac{284 \cdot 286}{110 \cdot 796} \approx 0,927$$

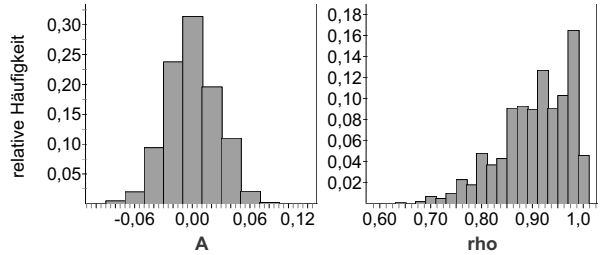


Abbildung 8.4: Verteilung des Assoziationsmaßes A und der odds ratio ρ nach 1 bzw. nach 1000 Simulationen

Tatsächlich ist der Anteil der Assoziationsmaße mit $|A| \geq 0,06$ wie auch von $\rho \leq 0,73$ kleiner als 0,05, genauer gilt $h(|A| \geq 0,06) \approx 0,016$ und $h(\rho \leq 0,73) \approx 0,015$. Nimmt man diese relativen Häufigkeiten als Schätzung der Wahrscheinlichkeiten $P(|A| \geq 0,06)$ und $P(\rho \leq 0,73)$, so kann man das Modell, dass kein Zusammenhang zwischen Hochschulzugehörigkeit und Rauchverhalten besteht, entgegen der auf dem deskriptiven Ergebnis in Kapitel 3.5 basierenden Aussage verwerfen. D. h., die Hypothese des nicht bestehenden Zusammenhangs wird verworfen und angenommen, dass in Freiburg systematisch mehr Raucher vorhanden sind als in Münster. Eine Erklärung für dieses Phänomen kann allerdings aus der Stichprobe nicht gewonnen werden.

Das verwendete Verfahren, der Permutationstest (mit Simulation), ist für viele Merkmale anwendbar, wenn der Unterschied zwischen den beiden Hochschulen (oder einem anderen gruppierenden Merkmal) beurteilt werden soll. Der Vorteil dieses Verfahrens ist, dass allein von dem empirischen Ergebnis in der Stichprobe ausgegangen wird und keine Zusatzannahmen, etwa das Modell der Binomialverteilung, wie im oben ausgeführten Beispiel notwendig sind. Dadurch lässt sich der Permutationstest mit dem angegebenen Verfahren stets gut simulieren.[1]

Beurteilen von Unterschieden zu Lage- und Streuparametern: In einem dritten Beispiel wollen wir die Unterschiede der Abiturleistung von Studentinnen und Studenten beurteilen. Die deskriptiv nicht beeindruckenden Unterschiede sind in der folgenden Tabelle dargestellt. Es gibt leichte Unterschiede bezogen auf die Mittelwerte (arithmetisches Mittel und Median). Zudem scheinen die Abiturleistungen der Studentinnen zumindest bezogen auf den Quartilsabstand homogener als die der Studenten zu sein.

[1] Zu dem formalisierten Verfahren sowie dem Einsatz des Permutationstests siehe etwa Sachs (1999). Ein Permutationstest ist solchen Tests, die eine Verteilungsannahme machen, in dem Sinne unterlegen, dass für den Nachweis der Signifikanz eines bestimmten Ereignisses eine größere Stichprobe genommen werden muss.

Messwert \ Geschlecht	weiblich	männlich	|Differenz|
$\bar{x}$	2,25	2,33	$\lvert d_{\bar{x}}\rvert = 0,08$
$x_{0,5}$	2,2	2,4	$\lvert d_{x_{0,5}}\rvert = 0,2$
$Q_{0,5}$	0,7	0,9	$\lvert d_{Q_{0,5}}\rvert = 0,2$

Auch das beurteilen wir per Simulation, indem wir wie oben die Zuordnung der Abiturleistungen zum Geschlecht aufheben, danach neu verteilen und die drei genannten Werte für die neuen Stichproben erneut berechnen. Wir erhalten bei 1000 solchen simulierten neuen Zuordnungen von links nach rechts die Verteilung der Unterschiede bezogen auf das arithmetische Mittel, der Mediane sowie der Streuung (Quartilsabstand) der Abiturleistungen, die Abbildung 8.5 zu sehen sind.

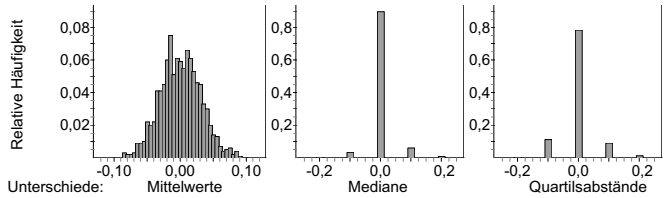


Abbildung 8.5: Simulierte Unterschiede bei Parametern der Häufigkeitsverteilung der Abiturnoten

Alle drei in der obigen Tabelle genannten Differenzen werden bezogen auf die Simulationen als Zufallsgrößen betrachtet. Für diese ergibt sich:

- $P(|d_{\bar{x}}| \geq 0,08) \approx (h_{1000}(|d_{\bar{x}}| \geq 0,08) = 0,009$
- $P|d_{x_{0,5}}| \geq 0,2) \approx (h_{1000}(|d_{x_{0,5}}| \geq 0,2) = 0,011$
- $P(|d_{Q_{0,5}}| \geq 0,2) \approx (h_{1000}(|d_{Q_{0,5}}| \geq 0,2) = 0,016$

Die Wahrscheinlichkeiten der in der Stichprobe erhaltenen Unterschiede hinsichtlich dieser drei Parameter sind also, wenn man von dem Modell der Unabhängigkeit von Geschlecht und Abiturnote ausgeht, so gering, dass das Modell der Unabhängigkeit verworfen werden kann. Allerdings ist hier zu beachten, dass die Unterschiede in den Stichproben zwar signifikant, absolut dennoch aber recht gering sind.

Wir weisen bei der Frage der Signifikanz hier in aller Kürze auf ein statistisches Phänomen hin: Die Möglichkeit, die Signifikanz von Unterschieden zwischen einem Modell und einem empirischen Phänomen zu erzeugen, gelingt umso besser, je größer der Stichprobenumfang ist. D. h., bei einer kleinen Stichprobe müssen die Unterschiede schon sehr auffällig sein, damit deren Signifikanz nachgewiesen werden kann, bei sehr großen Stichproben reicht dagegen ein eigentlich völlig unerheblicher Unterschied aus, um Signifikanz nachweisen zu können (zu letzteren Phänomenen zählen wir auch den oben untersuchten Unterschied). Wir machen diesen Gedanken am Beispiel des Würfels deutlich und dort für die als binomialverteilt modellierte Zufallsgröße X: Anzahl der Sechsen. Wir betrachten n Wiederholungen des Wurfs und bestimmen zwei symmetrisch um den Erwartungswert von X liegende Anzahlen der Sechs, k_u und k_o, so dass

$P(k_u \leq X \leq k_o) \geq 0,95$ gilt. Realisierungen der Zufallsgröße X außerhalb des Intervalls $[k_u;k_o]$ wären demnach *signifikante* Ereignisse, da die Wahrscheinlichkeit für Werte der Zufallsgröße außerhalb des Intervalls kleiner als 5% ist. Zu den Werten der Zufallsgröße, die gerade außerhalb des Intervalls liegen, nämlich $X = k_u - 1$ und $X = k_o + 1$, bestimmen wir die relativen Häufigkeiten der Sechsen, die an dem Modell $P(6) = p = \frac{1}{6}$ als signifikantes Ereignis Zweifel erzeugen:

n	$[k_u;k_o]$	$P(k_u \leq X \leq k_o)$	$h_n(k_u - 1)$	$h_n(k_o + 1)$
$n = 6$	$[0;3]$	0,991	—	0,667
$n = 60$	$[4;16]$	0,977	0,050	0,283
$n = 600$	$[82;118]$	0,957	0,135	0,198
$n = 6000$	$[943;1057]$	0,954	0,157	0,176
$n = 60000$	$[9821;10179]$	0,951	0,164	0,170

Während also bei sehr wenigen Versuchen nur erhebliche Abweichungen vom Modellparameter $p = \frac{1}{6}$ zu einem Verwerfen des Modells führen würden (beim 60maligen Würfeln etwa die relative Häufigkeit kleinergleich 0,050 oder größergleich 0,283), werden die entsprechenden Abweichungen bei hohen Versuchsanzahlen immer geringer.

Alle hier dargestellten Tests sind *Signifikanztests*. Das bedeutet, dass ein in einer Stichprobe erhaltenes Ereignis auf der Basis einer theoretischen Wahrscheinlichkeitsverteilung mit einem vorgegebenen Modellparameter *beurteilt* wird. Für andere Formen eines statistischen Tests sowie Gütekriterien an einen Test siehe z. B. Sachs (1999).

8.1.2 Schätzungen

In Kapitel 7.3 hatten wir per Simulation und Berechnung ein Intervall (Konfidenzintervall) für den Anteil der Studentinnen ausgehend von der Stichprobe geschätzt. Als Voraussetzung für diese Intervall-Schätzung war ein *Verteilungsmodell*, nämlich das der Binomialverteilung, notwendig. Diese Intervall-Schätzung einer relativen Häufigkeit in der Grundgesamtheit und damit dem Parameter p einer Binomialverteilung ist stets durch Berechnung möglich. Dagegen lassen sich andere Parameter einer Verteilung, wenn diese unbekannt ist, mit den hier beschriebenen Verfahren bezogen auf ein Konfidenzintervall nicht schätzen.

Es gibt dennoch intuitive und per Simulation mögliche Verfahren, Konfidenzintervalle bestehend auf einer Stichprobe zu schätzen. Von diesen skizzieren wir hier ein sogenanntes **Bootstrap-Verfahren**.[2] Dazu wird

- aus einer Stichprobe vom Umfang n eine neue Stichprobe, ebenfalls vom Umfang n und zwar *mit Zurücklegen* gezogen, und
- aus dieser Stichprobe heraus der interessierende Parameter einer unbekannten Verteilung geschätzt. Diese *Punktschätzung* werden wir stets so vornehmen, indem wir die empirische Entsprechung des theoretischen Parameters verwenden, also etwa die relative Häufigkeit in einer Stichprobe als Punktschätzung einer Wahrscheinlichkeit verwenden.[3]

[2] „Bootstrap“ heißt „Stiefelschlaufe“ und kann sinngemäß bedeuten, sich an den eigenen Haaren (Daten) aus dem Sumpf zu ziehen.

[3] Für andere Parameter gibt es zum Teil Punktschätzungen, die nicht in der direkten Übernahme des empirischen Parameters bestehen und sich dabei an Gütekriterien für Punktschätzungen orientieren, die wir hier nicht diskutieren werden (vgl. dazu etwa Sachs, 1999).

- Dieses Verfahren wird B Mal wiederholt und damit B Mal eine Punktschätzung zum unbekannten Parameter ausgeführt.
- Geht man wie oben von einem *Konfidenzniveau* von 95% aus, so wird das Konfidenzintervall durch die mittleren 95% der erfolgten Punktschätzungen erzeugt.[4]

Wir führen dieses Verfahren am Beispiel des Anteils von Studentinnen durch, zu dem wir in Kapitel 7.3 ein Konfidenzintervall $[0,651;0,772]$ zum Niveau 95% berechnet haben. Nun führen wir das oben beschriebene Verfahren $B = 1000$ Mal durch, d.h. wir schätzen in 1000 neu erzeugten Stichproben die Wahrscheinlichkeit p durch die in der Stichprobe erhaltene relative Häufigkeit der Studentinnen und erhalten die Verteilung der Punktschätzungen der Anteile von Studentinnen, die Abbildung 8.6 zu sehen ist.

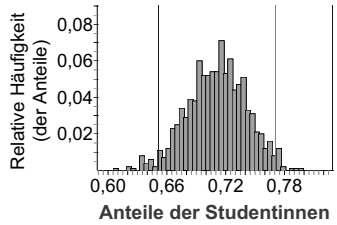
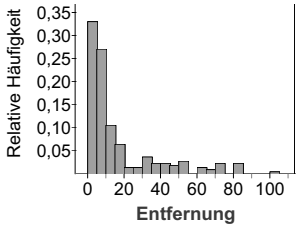


Abbildung 8.6: Simulierte Anteile von Studentinnen in 1000 Stichproben vom Umfang $n = 218$

Als Konfidenzintervall zum Konfidenzniveau 95% ergibt sich das Intervall, das die mittleren 95% der Punktschätzungen enthält, nämlich $[0,651;0,771]$. Dieses Konfidenzintervall entspricht nahezu dem in Kapitel 7.3 berechneten. Es basiert auf einer virtuellen Vergrößerung der Stichprobe (die entnommenen Elemente aus der Stichprobe werden zurückgelegt), aus der mit der Annahme der stochastischen Unabhängigkeit eine Stichprobe vom Umfang $n = 218$ entnommen wird. [5]

Als weiteres Beispiel wenden wir dieses Verfahren auf den Parameter der Schiefe der Verteilung der Entfernung der Studierenden der PH Freiburg zur Hochschule an, die auf der Häufigkeitsverteilung und Schiefeberechnung beruht, die in Abbildung 8.7 zu sehen ist.



$$QS_{0,25} = \frac{(x_{0,75} - x_{0,5}) - (x_{0,5} - x_{0,25})}{x_{0,75} - x_{0,25}} \approx 0,385$$

Abbildung 8.7: Häufigkeitsverteilung und Schiefeberechnung bzgl. der Entfernung der PH-Studierenden zur Hochschule

[4] Allgemeiner: Die mittleren $(1-\alpha)\cdot 100\%$ Schätzungen zum Konfidenzniveau $(1-\alpha)\cdot 100\%$.

[5] Vgl. für einen Überblick Sachs (1999) und zur Entwicklung der Verfahren Efron (1979) und Efron & Tibshirani (1993).

Es werden entsprechend der oben dargestellten Vorgehensweise 1000 Stichproben vom Umfang n aus der bestehenden Stichprobe erzeugt und jeweils die empirische Schiefe der Verteilung als Schätzung der theoretischen Schiefe berechnet. Man erhält die in Abbildung 8.8 zu sehende Verteilung der Schiefen.

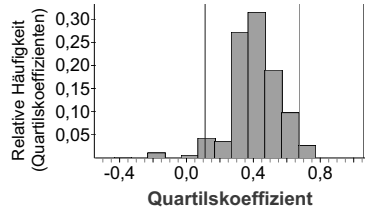


Abbildung 8.8: Simulierte Verteilung der Schiefen

Als Intervall zum Konfidenzniveau 95% ergibt sich $[0,11;0,67]$ und weist auf die vorhandene Rechtsschiefe der Verteilung der Entfernungen hin, wie sie auch in Kapitel 2.6 beschrieben worden ist. Ein Fülle teilweise deutlich komplexerer Beispiele ist in einer Ausgabe der Zeitschrift *Statistical Science* (2/2003) enthalten. Wir haben hier dieses Verfahren allein propädeutisch verwendet, ohne auf die Gültigkeit eines solchen rechnerbasierten Verfahrens einzugehen. Die Gültigkeit haben wir hier nur empirisch an einem Fall, der Intervallschätzung für einen unbekannten Parameter p, durch den Vergleich mit der dort möglichen Berechnung eines Konfidenzintervalls plausibel gemacht. Darüber hinaus gehende Hinweise zu dieser Gültigkeit findet man in der angegebenen Literatur oder exemplarisch bei Engel und Grübel (2008).

8.2 Ergänzungen

Wir wollen als Ergänzung einen kurzen Blick auf Standardverfahren des Testens und Schätzens werfen, die Verbindung der Intervallschätzung und des Testens eines Parameters einer Verteilung diskutieren und schließlich noch eine Überlegung zur sinnvollen Größe einer Stichprobe und damit auch einer durch Simulation erzeugten virtuellen Stichprobe (wir haben standardmäßig 1000 Simulationen verwendet) anstellen.

8.2.1 Testen und Schätzen

Die Verfahren, die wir im vorangegangenen Kapitel verwendet haben, sind frei von einem Verteilungsmodell. Die klassischen statistischen Verfahren des Testens und Schätzens, die auch in statistischer Software implementiert sind, setzen dagegen häufig ein Verteilungsmodell (meist das der Normalverteilung) voraus. Diese Verfahren sind damit prinzipiell in dem Zugang dieses Buches, das die Beschränkung auf diskrete Zufallsgrößen umfasst, nicht anwendbar. Wir machen daher nur wenige Anmerkungen zum grundsätzlichen Verfahren.

Testen: Beim Testen besteht das Verfahren darin, eine sogenannte **Testgröße** zu einer bestimmten Fragestellung zu entwickeln. In dem Beispiel des Vergleichs der Mittelwerte zweier Verteilungen (vgl. Kap. 8.1) geht man davon aus, dass, wenn die beiden Verteilungen (hier der Studierenden aus Münster und Freiburg) auf dem identischen, unbekannten Mittelwert basieren,

die Differenz der Mittelwerte zweier Stichproben theoretisch 0 sein müsste und empirisch bei einer großen Stichprobe nicht weit von 0 abweichen sollte. Mit diesem Modell sowie der weiteren Modellannahme der Normalverteilung kann die Verteilung der Differenzen theoretisch analysiert werden und in dieser theoretischen Verteilung wiederum ein Ablehnungsbereich bestimmt werden.

Bei diesem Test auf Mittelwertsunterschied wird die Testgröße durch

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

konstruiert, bei der die Differenz der Mittelwerte im Zähler erkennbar ist. Ist die Grundgesamtheit normalverteilt, so ist die Verteilung von t bekannt: Diese ist die sogenannte **student-t-Verteilung**[6], die ebenfalls stetig ist. Mit Hilfe dieser Verteilung kann (wie bei den bisher auf Berechnung beruhenden Beispielen) überprüft werden, ob mit dem empirisch ermittelten Mittelwertsunterschied ein bezogen auf das Verteilungsmodell seltenes Ereignis vorliegt. Im Beispiel der Zensurenverteilungen von Studentinnen und Studenten (vgl. Kap. 8.1) hat die Testgröße den Wert $t \approx 2,53$ sowie die Eigenschaft, dass ein Wert größergleich $t = 2,53$ in der zugehörigen student-t-Verteilung die Wahrscheinlichkeit von etwa 0,012 hat ($P(T \geq t) \approx 0,012$). Die zugehörige student-t-Verteilung hat die in Abbildung 8.9 dargestellte Gestalt. Auch nach diesem Test wäre also die Gleichheit der Abiturnoten der weiblichen und männlichen Studierenden abzulehnen. Hinsichtlich der Interpretation ergibt sich somit zum simulierten Permutationstest kein Unterschied.

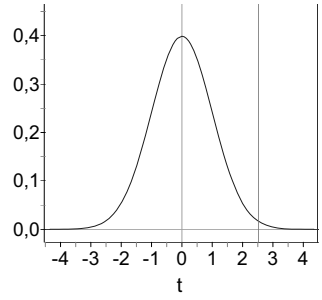


Abbildung 8.9: student-t-Verteilung im Beispiel der Zensurenverteilungen von Studentinnen und Studenten

Über den beschriebenen Permutationstest hinaus gibt es noch zahlreiche Testvarianten, die ebenfalls *parameterfrei*, d.h. ohne Voraussetzung eines Verteilungsmodells, konstruiert sind (vgl. Sachs, 1999).

Schätzen von Konfidenzintervallen: Am Beispiel der Schätzung des unbekannten Parameters p einer binomialverteilten Zufallsgröße erläutert, basiert die klassische Berechnung von Konfidenzintervallen auf folgenden Schritten:

[6] Diese Verteilung ist durch den Stichprobenumfang charakterisiert, durch den die sogenannten **Freiheitsgrade** festgelegt sind.

- X sei eine Zufallsgröße, die einer Verteilung mit dem unbekannten Parameter u genügt. Die Verteilung könnte hier etwa eine Binomialverteilung mit dem Parameter $u = p$ sein.
- Eine Stichprobe vom Umfang n zeigt das Ergebnis $x_1, ..., x_n$. Diese Stichprobe interpretieren wir als die Realisierung von n identisch verteilten Zufallsgrößen $X_1, ..., X_n$, die alle dem unbekannten Parameter u genügen. Bezogen auf eine binomialverteilte Zufallsgröße erhalten wir die Realisierungen der n binomialverteilten Zufallsgrößen $X_1, ..., X_n$, die alle dem Parameter p genügen und die eine Reihe von Erfolgen (A) und Misserfolgen ($\overline{A}$) bzw. Einsen und Nullen repräsentieren.
- Mit Hilfe dieser Stichprobe wird *ein* Schätzwert $\hat{u}$ mittels einer konstruierten Schätzfunktion $\hat{U}$ geschätzt: $\hat{U}(x_1, ..., x_n) = \hat{u}$. Das Ergebnis dieser einen (Punkt-)Schätzung ist eine Realisierung einer Zufallsgröße $\hat{U}(X_1, ..., X_n)$. Ein Wert dieser Zufallsgröße ist dabei der unbekannte Parameter u. Im Falle der Schätzung von p einer binomialverteilten Zufallsgröße könnten wir mit $\hat{U}(x_1, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} = \hat{p}$ eine solche (Punkt-)Schätzung durchführen, wobei $\hat{p} = h_n(A)$ gilt, also der relativen Häufigkeit für Erfolg (A) entspricht. Diese eine (Punkt-)Schätzung ist die Realisierung der Zufallsgröße $\hat{U}(x_1, ..., x_n) = \overline{X}$. Ein Wert dieser Zufallsgröße ist der unbekannte Parameter p.
- Ist die Klasse der Verteilung von $\hat{U}(X_1, ..., X_n)$ bekannt, so lässt sich in Abhängigkeit des Wertes $\hat{u}$ ein Intervall $[a;b]$ angeben mit $P\left(a \leq \hat{U}(X_1, ..., X_n) \leq b\right) \geq 0,95$. Da der unbekannte Parameter u ein Wert der Zufallsgröße $\hat{U}(X_1, ..., X_n)$ ist, gilt $P(a \leq u \leq b) \geq 0,95$. Das Intervall $[a;b]$ stellt damit das Konfidenzintervall zum Konfidenzniveau 0,95 dar. Dieses ist, da es abhängig von dem in der empirischen Stichprobe ermittelten Wert $\hat{u}$ ist, zufällig und überdeckt in 95% der Fälle den gesuchten Parameter u. Bezogen auf den unbekannten Parameter p einer Binomialverteilung kann $\hat{U}$ durch

$$\hat{U}(X_1, ..., X_n) = \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

konstruiert werden[7], wobei diese Zufallsgröße annähernd standardnormalverteilt ist. Die Standardnormalverteilung ist bekannt, ihr Erwartungswert ist 0, ihre Standardabweichung ist 1 und für eine standardnormalverteilte Zufallsgröße X gilt $P(-1,96 \leq X \leq 1,96) \approx 0,95$. Unter der Annahme das $\hat{U}$ annähernd standardnormalverteilt ist erhalten wir:

$$\begin{aligned} P\left(-1,96 \leq \hat{U}(X_1, ..., X_n) \leq 1,96\right) &= P\left(-1,96 \leq \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1,96\right) \\ &= P\left(\overline{X} - 1,96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \overline{X} + 1,96\sqrt{\frac{p(1-p)}{n}}\right) \\ &\approx 0,95 \end{aligned}$$

Verwendet man zusätzlich die auf der empirischen Stichprobe basierende Abschätzung $\overline{X} = h_n(A) \approx P(A) = p$, so erhält man

[7] Zu der Schwierigkeit, diesen Parameter durch ein Intervall abzuschätzen, siehe auch die Originalarbeit Neyman (1935) hinsichtlich der Konstruktion von Konfidenzintervallen.

$$P\left(h_n(A)-1{,}96\sqrt{\frac{h_n(A)(1-h_n(A))}{n}} \le p \le h_n(A)+1{,}96\sqrt{\frac{h_n(A)(1-h_n(A))}{n}}\right) \approx 0{,}95$$

und insgesamt das Konfidenzintervall:

$$\left[h_n(A)-1{,}96\sqrt{\frac{h_n(A)(1-h_n(A))}{n}};h_n(A)+1{,}96\sqrt{\frac{h_n(A)(1-h_n(A))}{n}}\right]$$

Das Verfahren, die Zufallsgröße $\hat{U}$ so zu konstruieren, dass sie der Standardnormalverteilung genügt, ist ein häufig verwendetes, wenn auch nicht universal einsetzbares Verfahren. Im Falle des hier skizzierten Bootstrap-Verfahrens haben wir dagegen die Verteilung der Zufallsgröße $\hat{U}(X_1,...,X_n)$ simuliert und über das Ergebnis der Simulation das Konfidenzintervall erhalten.

Wir fassen exemplarisch zusammen, wie ein Konfidenzintervall für den unbekannten Parameter p einer binomialverteilten Zufallsgröße X bestimmt werden kann, und zwar am Beispiel der in einer Stichprobe von 218 Studierenden erhaltenen 155 Studentinnen.

1. Per Simulation von binomialverteilten Zufallsgrößen hatten wir Parameter p_u und p_o ermittelt, so dass $P_{p_u}(X \ge 155)+P_{p_o}(X \le 155) \ge 0{,}05$ war, die Binomialverteilungen also den Wert 155 mit einer Wahrscheinlichkeit von (gerade) 0,05 überlappen. Das Ergebnis war hier $[0{,}650;0{,}772]$.
2. Die Parameter p_u und p_o lassen sich in gleicher Weise berechnen mit dem Ergebnis des Intervalls $[0{,}651;0{,}772]$.
3. Wendet man die Methode des Bootstrap an, ergibt sich das Intervall $[0{,}651;0{,}771]$.
4. Verwendet man schließlich die Approximation mit der Normalverteilung, die wir oben skizziert haben, so ergibt sich schließlich das Intervall $[0{,}651;0{,}771]$.[8]

Alle verwendeten Verfahren[9] ergeben nahezu identische Intervalle. Insbesondere die ersten drei Varianten basieren einerseits auf der Verwendung diskreter Zufallsgrößen und des Rechners, andererseits auf einer Idee, mit der sich Hypothesentests und Konfidenzintervalle gemeinsam interpretieren lassen, was wir im Folgenden skizzieren.

8.2.2 Vergleich Hypothesentest – Konfidenzintervalle

Wir haben die ersten drei Konfidenzintervalle mit folgender Idee gesucht:

- Die in einer Stichprobe von $n = 218$ Studierenden erhaltene Anzahl von Studentinnen ist ein Wert einer binomialverteilten Zufallsgröße X: Anzahl der Studentinnen mit dem unbekannten Parameter p.

[8] Es ergibt sich, wenn mit A das Ereignis einer Studentin bezeichnet wird, $h_{218}(A) = \frac{155}{218} \approx 0{,}711$ und $h_n(A) - 1{,}96\sqrt{\frac{h_n(A)(1-h_n(A))}{n}} \approx 0{,}651$ sowie $h_n(A)+1{,}96\sqrt{\frac{h_n(A)(1-h_n(A))}{n}} \approx 0{,}771$.

[9] Dies sind nicht alle tatsächlich möglichen Verfahren. So ergibt sich auch im Zusammenhang mit dem Bernoullischen Gesetz der großen Zahlen eine Abschätzung eines Konfidenzintervalls, die allerdings aufgrund der recht groben Abschätzung größer als die genannten ist. Andere Möglichkeiten findet man z.B. in Sachs (1999).

- Für diesen Parameter p existiert ein (ebenfalls nicht bekanntes) Intervall $[a \leq X \leq b]$, in dem die Ergebnisse von 95% zukünftiger Stichproben liegen werden, wobei wir nicht wissen, ob p kleiner oder größer der erhaltenen Häufigkeit für Studentinnen von $\frac{155}{218} \approx 0,711$ ist.
- Gehen wir davon aus, dass dem Stichprobenergebnis von 155 Studentinnen der Fall vorliegt, dass 155 in dem Intervall $[a;b]$ liegt, was in 95% der Fälle zutrifft, so könnte im Extremfall $a = 155$ oder $b = 155$ gelten.
- Für beide Extremfälle lässt sich ein Parameter per Simulation oder Berechnung finden. Da wir vorausgehend die Wahrscheinlichkeit dafür, dass ein Wert außerhalb des Intervalls $[a;b]$ in einer Stichprobe angenommen wird, auf 5% gesetzt haben, verwenden wir dies auch für die beiden extremen Parameter p_u zum Intervall $[\tilde{a};155]$ und p_o zum Intervall $[155;\tilde{b}]$. Ist die Wahrscheinlichkeit für einen Wert rechts des ersten und links des zweiten Intervalls zusammen etwa 5%, so haben wir das Konfidenzintervall konstruiert.

Bezogen auf den Hypothesentest bedeutet diese nochmalig skizzierte Konstruktion: Setzt man $\tilde{p}_u = p_u - \varepsilon$ oder $\tilde{p}_o = p_o + \varepsilon$ mit einem positiven ε nahe 0, d. h., erniedrigt man p_u bzw. erhöht man p_o sehr gering, so erhält man zwei weitere Parameter einer binomialverteilten Zufallsgröße mit folgender Eigenschaft:

- Überprüft man mittels eines Hypothesentests die Anzahl von 155 Studentinnen in einer Stichprobe von 218 Studierenden, so wäre $X = 155$ sowohl hinsichtlich von $\tilde{p}_u$ als auch von $\tilde{p}_o$ ein signifikantes Ereignis.

Wir wählen $\tilde{p}_u = 0,645$ und $\tilde{p}_o = 0,775$ und erhalten folgende zwei Binomialverteilungen.

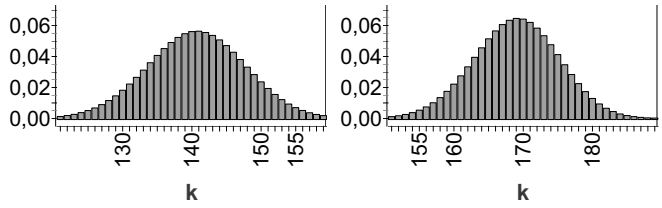


Abbildung 8.10: Binomialverteilungen mit $\tilde{p}_u$ links und $\tilde{p}_o$ rechts

Für $\tilde{p}_u = 0,645$ ergibt sich ein symmetrisches Intervall um den Erwartungswert von $[126;154]$ mit $P_{p_u}(126 \leq X \leq 154) > 0,95$ und für $\tilde{p}_o = 0,775$ ergibt sich ein symmetrisches Intervall um den Erwartungswert von $[157;181]$ mit $P_{p_o}(157 \leq X \leq 181) > 0,95$. Bezogen auf beide Parameter wäre demnach $X = 155$ ein signifikantes Ereignis.

Wir verbinden in einem weiteren Beispiel die Konstruktion eines Konfidenzintervalls sowie die Idee des Testens. In Kapitel 3.5 hatten wir den Zusammenhang zwischen Geschlecht und der Präferenz für die Parteien CDU bzw. Die Grünen betrachtet und vermutet, dass Studentinnen seltener eine Präferenz für die Partei CDU bzw. häufiger für die Partei Die Grünen besitzen als die Studenten. Wir bilden mit dem Bootstrap-Verfahren 1000 neue Stichproben aus der vorhandenen Stichprobe mit 405 Studierenden ohne Zurücklegen und bestimmen auf diese Weise 1000 (Punkt-)Schätzungen zum Assoziationsmaß $A = h_{405}(C|M) - h_{405}(C|W)$, wobei C für die Partei CDU

und *M* für männlich sowie *W* für weiblich steht. Dabei ergibt sich die in Abbildung 8.11 (links) dargestellte Verteilung dieser Schätzungen.

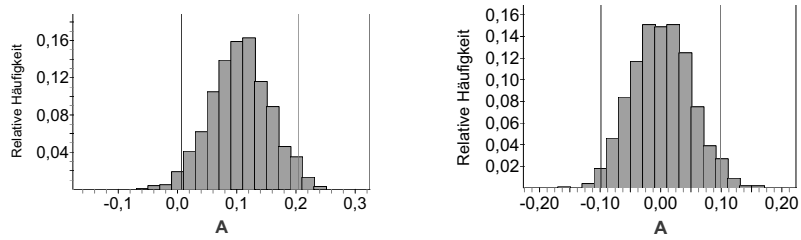


Abbildung 8.11: Verteilung des Assoziationsmaßes *A* in 1000 Bootstrap-Stichproben (links) und Verteilung des Assoziationsmaßes *A* in 1000 simulierten Permutationen der Merkmale Geschlecht und Partei

Durch die Bestimmung der Grenzen der mittleren 95% der so konstruierten Schätzungen (um das arithmetische Mittel der Schätzungen) erhalten wir das Konfidenzintervall (zum Konfidenzniveau 95%) für das Assoziazionsmaß *A* als $[0,01;0,20]$. In diesem Intervall ist der Wert $A = 0$, der die Unabhängigkeit beider Merkmale implizieren würde, nicht enthalten. Das bedeutet, dass das Modell der Unabhängigkeit beider Merkmale nicht zu der empirisch ermittelten Stichprobe und dem empirisch ermittelten Assoziationsmaß passt. Umgekehrt sollte also damit das empirisch ermittelte Assoziationsmaß bezogen auf das Modell der Unabhängigkeit beider Merkmale ein signifikantes Ereignis darstellen. Wir überprüfen dies wie in Kapitel 8.1 mit dem simulierten Permutationstest und erhalten bei 1000 Simulationen auf der Basis des Modells der Unabhängigkeit die in Abbildung 8.11 (rechts) dargestellten Assoziationsmaße.

Bei den 1000 Simulationen ist der Anteil der Assoziationsmaße mit $|A| \geq 0,11$ kleiner als 0,05 (liegt außerhalb der in der Abbildung eingezeichneten Grenzen). Das in der Stichprobe erhaltene Assoziationsmaß ist demnach bezogen auf das Modell der Unabhängigkeit signifikant, die Unabhängigkeit kann als Modell abgelehnt werden. Zusammengefasst gilt also, dass ein Konfidenzintervall für einen unbekannten Parameter der Verteilung in der Grundgesamtheit diejenigen Modell-Parameter außerhalb des Konfidenzintervalls angibt, für die das empirische Ereignis (in der Stichprobe) ein signifikantes Ereignis darstellen würde.[10]

8.2.3 Simulationsanzahl

Wir betrachten schließlich mit Hilfe von Konfidenzintervallen die potentiellen Abweichungen von einem wahren Parameter und zwar bezogen auf die Stichprobengrößen von $n = 218$ (Erhebung Freiburg 2010), $n = 1082$ (Erhebung Münster) und 1000 (virtuelle Erhebung per Simulation).

Während die Stichprobengrößen bei der Erhebung von Studierenden durch die Rahmenbedingungen vorgegeben waren, haben uns bei der Festlegung der virtuellen Stichprobengröße von 1000 folgende Gründe geleitet:

[10] Hier ist allerdings zu beachten, dass das Konfidenzintervall und der Signifikanztest mit unterschiedlichen Methoden ausgeführt wurden.

- Diese Simulationsanzahl kann von schulbezogener Software sinnvoll verarbeitet werden. Professionelle Software oder Programmierung erlauben dagegen deutlich höhere Anzahlen.
- Diese Simulationsanzahl erlaubt es noch deutlich, zwischen simulierter Verteilung und theoretischer Verteilung zu unterscheiden. Gerade dieser Unterschied zwischen erzeugten Daten (simuliert oder regulär erhoben) und einem Modell solcher Datenerzeugungen, das auf einer theoretischen Verteilung basiert, war uns wichtig.
- Diese Simulationsanzahl ermöglicht es dennoch, Parameter einer Verteilung relativ eng einzugrenzen.

Das zuletzt genannte Argument betrachten wir für Konfidenzintervalle zu den drei Stichprobengrößen und zu einer Auswahl von relativen Häufigkeiten der Erfolge (A) bei n Wiederholungen einer binomialverteilten Zufallsgröße X: Anzahl der Erfolge:

	$n = 218$	$n = 1082$	$n = 1000$
$h_n(A) = 0,01$	$[-0,003; 0,023]$	$[0,040; 0,016]$	$[0,004; 0,016]$
$h_n(A) = 0,1$	$[0,006; 0,140]$	$[0,082; 0,118]$	$[0,081; 0,119]$
$h_n(A) = 0,3$	$[0,239; 0,361]$	$[0,273; 0,327]$	$[0,272; 0,328]$
$h_n(A) = 0,5$	$[0,434; 0,566]$	$[0,470; 0,530]$	$[0,469; 0,531]$
$h_n(A) = 0,7$	$[0,639; 0,761]$	$[0,673; 0,727]$	$[0,672; 0,728]$
$h_n(A) = 0,9$	$[0,860; 0,940]$	$[0,882; 0,918]$	$[0,881; 0,919]$
$h_n(A) = 0,99$	$[0,977; 1,003]$	$[0,984; 0,996]$	$[0,984; 0,996]$

Die negativen „Wahrscheinlichkeiten“ wie auch die größer 1 sind hier aufgenommen worden, um die Länge der Intervalle in Abhängigkeit von der beobachteten relativen Häufigkeit zu analysieren, die in Abbildung 8.12 für $n = 218$ und $n = 1000$ dargestellt ist.

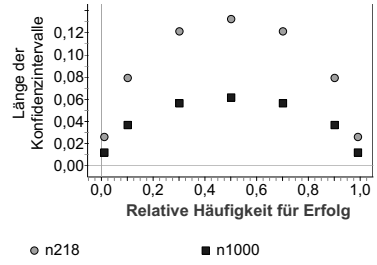


Abbildung 8.12: Längen von Konfidenzintervallen für $n = 218$ und $n = 1000$

Man erkennt, dass die Länge der Konfidenzintervalle, die bezogen auf eine beobachtete Häufigkeit von 0,5 symmetrisch ist, für $n = 218$ noch recht groß ist, für $n = 1000$ dagegen maximal etwa 0,06 beträgt. Letzteres war uns in diesem Buch ausreichend.

8.3 Aufgaben

Aufgabe 8.1: Gegeben ist der Datensatz zu den Studierenden der Hochschulen in Freiburg und Münster. Untersuchen Sie mit dem simulierten Permutationstest, ob die Studierenden beider Hochschulen sich

a) in der Eigenschaft, noch bei den Eltern zu wohnen, unterscheiden.

b) im Beziehungsverhalten unterscheiden.

Aufgabe 8.2: Gegeben ist der Datensatz zu den Studierenden der Hochschulen in Freiburg und Münster. Untersuchen Sie mit dem simulierten Permutationstest, ob die Studierenden beider Geschlechter sich

a) in der Eigenschaft, noch bei den Eltern zu wohnen, unterscheiden.

b) in der Präferenz für das Fortbewegungsmittel Fahrrad unterscheiden.

Aufgabe 8.3: Betrachten Sie eine der beiden Hochschulen, Freiburg oder Münster. Bestimmen Sie per Berechnung, Simulation und Bootstrap-Verfahren Konfidenzintervalle zu der in der Grundgesamtheit auftretenden Häufigkeit (p) für

- Single,
- Radfahrer,
- „schlechte“ Schüler mit einem Abiturschnitt über 3,0.

Aufgabe 8.4: In Kapitel 3.7 hatte wir folgende Unterschiede zwischen den Studierenden in Freiburg und Münster betrachtet:

„Gegeben sind unten die Datensätze zu Studierenden (Münster und Freiburg) zu der Parteipräferenz, dem bevorzugten Beförderungsmittel sowie zum Erhalt von BAföG.
Sind Grünen-Wähler umweltbewusster? Sind BAföG-Bezieher SPD-Wähler?“

Beurteilen Sie nun die deskriptiv ermittelten Unterschiede.

	Grüne	CDU/FDP	Summe
Auto	13	9	22
Fahrrad	116	132	248
Summe	129	141	270

	BAföG	kein BAföG	Summe
SPD	49	107	156
CDU/FDP	66	163	229
Summe	115	270	385

9 Daten- und Wahrscheinlichkeitsanalyse: Rückschau

In Kapitel 4 haben wir eine vorsichtige Umschreibung des Begriffs *Datenanalyse* vorgenommen. Gleiches wollen wir auch für den Begriff der *Wahrscheinlichkeitsanalyse* machen, der häufig als Wahrscheinlichkeitsrechnung oder Wahrscheinlichkeitstheorie bezeichnet wird. Der Begriff umfasst natürlich die Analyse von Gesetzmäßigkeiten zufälliger Vorgänge (vgl. Fisz, 1980) oder die „Mathematik des Zufalls“ (Henze, 2010). Um das Anliegen dieses Buches, die Stochastik durchweg aus der Perspektive von Daten zu betrachten, zu verdeutlichen, wählen wir aber folgende Umschreibung:

> Die Wahrscheinlichkeitsanalyse umfasst Methoden zur Beschreibung zufälliger Vorgänge und ermöglicht insbesondere das Aufstellen, die Analyse sowie die Beurteilung von Modellen zur Verteilung zukünftig erhobener Daten.

Mit dieser Umschreibung geht der Ansatz einher, für Methoden, die in der Datenanalyse zur Beschreibung statistischer Daten verwendet werden, jeweils Modelle zur Beschreibung zukünftiger Daten zu untersuchen. Auch wenn die Schwerpunkte in der Beschreibung von Methoden innerhalb der Daten- sowie der Wahrscheinlichkeitsanalyse unterschiedlich gesetzt sind und nicht *jede* Methode in beiden Teilen des Buches zur Sprache kommt, gibt es etwa folgende Querverbindungen:

Datenanalyse empirische Daten-Welt	Wahrscheinlichkeitsanalyse Modell-Welt
Daten ordnen, Merkmale, Merkmalsausprägungen	Festlegen von Ereignissen Definition einer Zufallsgröße
Häufigkeiten	Schätzen und Setzen von Wahrscheinlichkeiten
Häufigkeitsverteilung	Wahrscheinlichkeitsverteilung einer Zufallsgröße X
empirische Lage- und Streuparameter	theoretische Lage- und Steuparameter (insbesondere Erwartungswert, Varianz und Standardabweichung)
(empirische) Schiefe	(theoretische) Schiefe
bedingte Häufigkeiten	bedingte Wahrscheinlichkeiten, Abhängigkeit, Unabhängigkeit
empirische Korrelation	theoretische Korrelation

Trotz dieser Gemeinsamkeiten oder Querverbindungen gibt es einen grundsätzlichen Unterschied bei der Untersuchung von empirischen Häufigkeitsverteilungen und theoretischen Wahr-

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5_9,

scheinlichkeitsverteilungen. So kann innerhalb der Datenanalyse (fast) ohne weitere Überlegungen das *Produkt* einer Erhebung, also die Daten, untersucht werden. Innerhalb der Wahrscheinlichkeitsanalyse müssen dagegen erhebliche Überlegungen zum *Prozess* der Genese zukünftiger Daten investiert werden. Während so in der Datenwelt Aussagen für beliebige Häufigkeitsverteilungen gemacht werden können, kann der Prozess der Genese zukünftiger Daten nur hinsichtlich weniger Typen eines solchen Prozesses beschrieben werden, will man sich wie hier auf möglichst elementar gehaltene Methoden beschränken. Auch über dieses Buch hinaus gilt, dass die Analyse des Prozesses im Allgemeinen wesentlich komplexer als die Analyse des Produkts ist.

In den Fällen, in denen wir den Prozess der Datengenese in einem Modell beschreiben konnten, haben wir versucht, den Zusammenhang zwischen der Analyse empirischer und zukünftiger Daten im Sinne von Abbildung 9.1 deutlich zu machen:

- Ziel der Datenanalyse ist die Beschreibung von Mustern in den Daten (unter Beachtung von Abweichungen);
- Ziel der Wahrscheinlichkeitsanalyse ist es, solche Muster in theoretischen Modellen zu verarbeiten und damit die Möglichkeit von Prognosen zukünftiger Datenerhebungen zu schaffen;
- Im Rückschluss lassen sich diese Modelle anhand von erneut erhobenen Daten beurteilen.

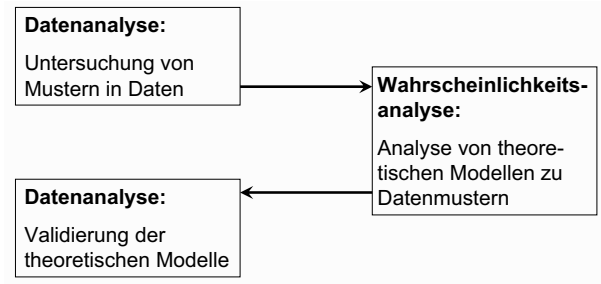


Abbildung 9.1: Datenanalyse – Wahrscheinlichkeitsanalyse – Rückschluss

Weiterhin war uns die Simulation als Mittlerin zwischen empirischen Daten und der theoretischen Beschreibung zukünftiger Daten durch Wahrscheinlichkeitsverteilungen wichtig. Simulationen können den Unterschied zwischen einem Modell (wie z.B. einer Wahrscheinlichkeitsverteilung) und empirischen Daten verdeutlichen und dabei der Formulierung „auf lange Sicht“ bezogen auf die Interpretation einer Wahrscheinlichkeit oder eines Erwartungswerts verdeutlichen. Mit Simulationen können schließlich auch Grundideen der beurteilenden Datenanalyse (beurteilenden Statistik) skizziert werden oder können dort sogar als eigenständige Methode fungieren, etwa in den Methoden des Bootstrap (Kap. 8).

Diesen in Abbildung 9.1 enthaltenen Dreischritt in der elementaren Analyse von Daten aufzuzeigen, war das Anliegen des Buches. Die Güte der Ergebnisse eines solchen Dreischritts basiert am Anfang wie auch am Ende auf der Güte der Datenerhebung. Damit sind wir am Ende dieses Buches auch wieder zum Anfang, der Datenerhebung, zurückgekehrt.

Literaturverzeichnis

Büchter, A. & Henn, H.-W. (2005). *Elementare Stochastik – Eine Einführung in die Mathematik der Daten und des Zufalls*. Berlin: Springer.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.

Eichler, A. & Vogel, M. (2009). *Leitidee Daten und Zufall – Von konkreten Beispielen zur Didaktik der Stochastik*. Wiesbaden: Vieweg+Teubner.

Engel, J. (2010). *Anwendungsorientierte Mathematik: Von Daten zur Funktion*. Berlin: Springer.

Engel, J. & Grübel, R. (2008). Bootstrap – oder die Kunst, sich selbst aus dem Sumpf zu ziehen. *Stochastik in der Schule*, *55* (2), 113–130.

Fisz, M. (1980). *Probability theory and mathematical statistics* (3. Aufl.). Malabar: Robert E. Krieger Publishing Company.

Hartung, J., Elpelt, B. & Klösener, K.-H. (2009). *Statistik: Lehr- und Handbuch der angewandten Statistik*. München: Oldenbourg.

Henze, N. (2010). *Stochastik für Einsteiger* (8. Aufl.). Wiesbaden: Vieweg+Teubner.

Hoffrage, U. (2003). Risikokommunikation bei Brustkrebsfrüherkennung und Hormonersatztherapie. *Zeitschrift für Gesundheitspsychologie*, *11* (3), 76–86.

Huff, D. & Geis, I. (1959). *How to take a chance*. New York: Norton.

Kolmogoroff, A. N. (1973). *Grundbegriffe der Wahrscheinlichkeitsrechnung (reprint)*. Heidelberg: Springer.

Kreyszig, E. (1998). *Statistische Methoden und ihre Anwendungen* (7. Aufl.). Göttingen: Vandenhoeck & Ruprecht.

Kütting, H. (1994). *Beschreibende Statistik im Schulunterricht*. Mannheim: B.I. Wissenschaftsverlag.

Neyman, J. (1935). On the problem of confidence intervalls. *The Annals of Mathematical Statistics*, *6* (3), 112–116.

Polasek, W. (1994). *EDA Explorative Datenanalyse – Einführung in die deskriptive Statistik* (2. Aufl.). Berlin: Springer.

Sachs, L. (1999). *Angewandte Statistik – Anwendung statistischer Methoden*. Berlin: Springer.

Schneider, I. (1988). *Die Entwicklung der Wahrscheinlichkeitstheorie von den Anfängen bis 1933*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Schupp, H. (1982). Zum Verhältnis statistischer und wahrscheinlichkeitstheoretischer Komponenten im Stochastik-Unterricht der Sekundarstufe I. *Journal für Mathematik-Didaktik*, *3* (3/4), 207–226.

Sill, H.-D. (1993). Zum Zufallsbegriff in der stochastischen Allgemeinbildung. *Zentralblatt der Mathematikdidaktik*.

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Berlin: Springer.

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5,

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review* (3), 223–248.

Zimbardo, P. G. & Gerrig, R. J. (2004). *Psychologie. Eine Einführung*. München: Pearson Studium.

Sachverzeichnis

A. Eichler, M. Vogel, *Leitfaden Stochastik*, DOI 10.1007/978-3-8348-9909-5,