Intelligente Datenanalyse

Vorlesungsskript

Christian Borgelt

20. Oktober 2004

Institut für Wissens- und Sprachverarbeitung
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, D-39106 Magdeburg

Inhaltsverzeichnis

He deals the cards to find the answer,
The sacred geometry of chance,
The hidden law of a probable outcome.
The numbers lead a dance.

Sting: “The Shape of My Heart”
aus dem Album “Ten Summoner’s Tales”

Kapitel 1

Einleitung

Durch die moderne Computertechnologie, die jedes Jahr leistungsfähigere Rechner hervorbringt, ist es heute möglich, mit sehr geringen Kosten enorme Datenmengen zu sammeln, zu übertragen, zusammenzuführen und zu speichern. Dadurch kann es sich eine immer größere Zahl von Industrieunternehmen, wissenschaftlichen Einrichtungen und staatlichen Institutionen leisten, riesige elektronische Archive von Tabellen, Dokumenten, Bildern und Tönen anzulegen. Es ist verlockend anzunehmen, daß, wenn man nur genug Daten hat, man jedes Problem lösen kann — wenigstens im Prinzip.

Eine nähere Betrachtung zeigt jedoch, daß Daten allein, wie umfangreich sie auch sein mögen, nicht ausreichen. Man kann sagen, daß man in großen Datenbanken den Wald vor lauter Bäumen nicht sieht. Obwohl Einzelinformationen abgerufen und einfache Aggregationen (z.B. der durchschnittliche monatliche Umsatz im Raum Frankfurt) leicht berechnet werden können, bleiben allgemeine Muster, Strukturen und Regelmäßigkeiten meistens unentdeckt. Oft sind jedoch gerade diese Muster besonders wertvoll, z.B., weil sie ausgenutzt werden können, um den Umsatz zu erhöhen. Wenn etwa ein Supermarkt herausfindet, daß bestimmte Produkte häufig zusammen gekauft werden, so kann manchmal der Absatz gesteigert werden, wenn man diese Produkte geschickt in den Regalen des Supermarktes anordnet (sie könnten z.B. nebeneinander plaziert oder als Paket angeboten werden, um noch mehr Kunden zu animieren, sie zusammen zu kaufen).

Diese Muster zu finden und damit einen größeren Teil der in den verfügbaren Daten enthaltenen Informationen auszunutzen, erweist sich jedoch als ziemlich schwierig. Im Gegensatz zu dem Überfluß an Daten gibt es einen Mangel an Werkzeugen, um diese Daten in nützliches Wissen zu verwandeln.

John Naisbett charakterisierte die Situation treffend so [Fayyad *et al.* 1996]:

> We are drowning in information, but starving for knowledge.
> (Wir ertrinken in Informationen, hungern aber nach Wissen.)

Als Folge hat sich eine neue Forschungsrichtung entwickelt, die unter den Namen **Wissensentdeckung in Datenbanken** (knowledge discovery in databases, KDD) und **Data Mining** bekannt geworden ist. Sie stellt sich der Herausforderung, Techniken zu entwickeln, die Menschen helfen können, nützliche Muster in ihren Daten zu entdecken.

In diesem einleitenden Kapitel geben wir einen kurzen Überblick über die Wissensentdeckung in Datenbanken und die intelligente Datenanalyse. Als ersten Schritt arbeiten wir den Unterschied zwischen „Daten“ und „Wissen“ heraus, um klare Begriffe zu erhalten, mit denen wir deutlich machen können, warum es nicht ausreicht, Daten zu sammeln, sondern wir versuchen müssen, aus ihnen Wissen zu gewinnen. Als Veranschaulichung betrachten wir ein Beispiel aus der Geschichte der Wissenschaft. Zweitens beschreiben wir den Prozeß der Wissensentdeckung in Datenbanken **(KDD-Prozeß)**, in dem Data Mining nur ein — wenn auch sehr wichtiger — Schritt ist. Wir charakterisieren einige Standardaufgaben der Datenanalyse und geben eine Liste einiger wichtiger Datenanalysemethoden an.

1.1 Daten und Wissen

In diesem Skript unterscheiden wir zwischen *Daten* und *Wissen*. Aussagen wie „Kolumbus entdeckte Amerika im Jahre 1492.“ oder „Herr Meier fährt einen VW Golf.“ sind **Daten**. Dabei sehen wir es als irrelevant an, ob wir diese Aussagen schon kennen, ob wir sie im Augenblick zu irgendeinem Zweck benötigen usw. Die wesentliche Eigenschaft dieser Aussagen ist, daß sie sich auf einzelne Ereignisse, Objekte, Personen, Zeitpunkte etc. beziehen, also allgemein auf Einzelfälle. Deshalb ist, selbst wenn sie wahr sind, ihr Geltungsbereich und damit auch ihr Nutzen sehr eingeschränkt.

Im Gegensatz dazu besteht **Wissen** in Aussagen wie „Alle Massen ziehen sich an.“ oder “Jeden Tag um 17 Uhr fährt ein Zug von Magdeburg nach Berlin.“. Wieder vernachlässigen wir die Relevanz der Aussagen für unsere augenblickliche Situation and ob wir sie bereits kennen. Wesentlich ist, daß sich diese Aussagen nicht auf Einzelfälle beziehen, sondern allgemeine Gesetze oder Regeln sind. Daher haben sie, wenn sie wahr sind, einen großen Geltungsbereich, vor allem aber erlauben sie uns, Voraussagen zu machen, und sind folglich sehr nützlich.

Zugegeben: Im Alltag nennen wir auch Aussagen wie „Kolumbus entdeckte Amerika im Jahre 1492“ Wissen. Hier vernachlässigen wir jedoch diese recht unscharfe Verwendung des Begriffs „Wissens“ und bedauern, daß sich offenbar keine mit der Alltagssprache völlig konsistente Terminologie finden läßt. Weder einzelne Aussagen über Einzelfälle noch Sammlungen solcher Aussagen sollen uns als Wissen gelten.

Zusammenfassend können Daten und Wissen so gekennzeichnet werden:

Daten

- beziehen sich auf Einzelfälle
  (einzelne Objekte, Personen, Ereignisse, Zeitpunkte usw.)
- beschreiben individuelle Eigenschaften
- sind oft in großen Mengen verfügbar
  (Datenbanken, Archive)
- sind meistens einfach zu sammeln oder zu beschaffen
  (z.B. Scannerkassen in Supermärkten, Internet)
- erlauben uns nicht, Voraussagen zu machen

Wissen

- beziehen sich auf *Klassen* von Fällen
  (*Mengen* von Objekten, Personen, Ereignissen, Zeitpunkten usw.)
- beschreiben allgemeine Muster, Strukturen, Gesetze, Prinzipien usw.
- bestehen aus so wenigen Aussagen wie möglich
  (dies ist eine Zielsetzung, siehe unten)
- ist meistens schwer zu finden oder zu erwerben
  (z.B. Naturgesetze, Ausbildung)
- erlaubt uns, Voraussagen zu machen

Aus diesen Charakterisierungen sieht man deutlich, daß im allgemeinen Wissen sehr viel wertvoller ist als (rohe) Daten. Es ist hauptsächlich die Allgemeinheit der Aussagen und die Möglichkeit, Voraussagen über das Verhalten und die Eigenschaften neuer Fälle machen zu können, die seine Überlegenheit ausmachen.

Allerdings ist nicht jede Art von Wissen genauso wertvoll wie jede andere. Nicht alle allgemeinen Aussagen sind gleich wichtig, gleich gehaltvoll, gleich nützlich. Wissen muß daher bewertet werden. Die folgende Liste, von der wir jedoch nicht behaupten wollen, daß sie vollständig sei, führt einige der wichtigsten Kriterien auf:

Kriterien zur Bewertung von Wissen

- Korrektheit (Wahrscheinlichkeit, Testerfolg)
- Allgemeinheit (Geltungsbereich, Geltungsbedingungen)
- Nützlichkeit (Relevanz, Vorhersagekraft)
- Verständlichkeit (Einfachheit, Klarheit, Sparsamkeit)
- Neuheit (vorher unbekannt, unerwartet)

In der (Natur-)Wissenschaft stehen Korrektheit, Allgemeinheit und Einfachheit (Sparsamkeit) im Vordergrund: Man kann (Natur-)Wissenschaft charakterisieren als die Suche nach einer korrekten Minimalbeschreibung der Welt. In Wirtschaft und Industrie dagegen wird mehr Wert auf Nützlichkeit, Verständlichkeit und Neuheit gelegt: Das Hauptziel besteht darin, einen Wettbewerbsvorteil zu erreichen und so höhere Gewinne zu erwirtschaften. Nichtsdestotrotz kann es sich keiner der beiden Bereiche leisten, die jeweils anderen Kriterien zu vernachlässigen.

Tycho Brahe und Johannes Kepler

Tycho Brahe (1546–1601) war ein dänischer Adeliger und Astronom, der mit der finanziellen Unterstützung von König Frederik II in den Jahren 1576 und 1584 zwei Sternwarten auf der Insel Hven, etwa 32 km nordöstlich von Kopenhagen, errichtete. Unter Verwendung der besten Instrumente seiner Zeit (Fernrohre waren noch unbekannt — sie wurden erst später von Galileo Galilei (1564–1642) und Johannes Kepler (siehe unten) zur Himmelsbeobachtung eingesetzt) bestimmte er die Positionen der Sonne, des Mondes und der Planeten mit einer Genauigkeit von weniger als einer Bogenminute, was alle bis dahin durchgeführten Messungen weit übertraf. Er erreichte in der Praxis die theoretische Genauigkeitsgrenze für Beobachtungen mit dem bloßen Auge. Sorgfältig zeichnete er die Bewegungen der Himmelskörper über mehrere Jahre hinweg auf [Greiner 1989, Zey 1997].

Tycho Brahe sammelte Daten über unser Planetensystem — große Mengen von Daten, jedenfalls vom Standpunkt des 16. Jahrhunderts aus. Aber er war nicht in der Lage, sie in einem einheitlichen Schema zusammenzufassen, z.T. deshalb, weil er am geozentrischen System festhielt. Er konnte genau sagen, wo z.B. der Mars an einem bestimmten Tag des Jahres 1582 gestanden hatte, aber er konnte die Positionen an verschiedenen Tagen nicht durch eine Theorie zueinander in Beziehung setzen. Alle Hypothesen, die er aufstellte, scheiterten an seinen hochgenauen Daten. Zwar entwickelte er das (im 17. Jahrhundert zeitweise populäre) tychonische Planetensystem,

in dem sich Sonne und Mond um die Erde, alle anderen Planeten aber (auf Kreisen) um die Sonne bewegen, doch bewährte sich dieses System nicht. Heute könnten wir sagen, daß Tycho Brahe ein „Data-Mining-“ oder „Wissensentdeckungsproblem“ hatte. Er verfügte über die notwendigen Daten, aber er konnte das enthaltene Wissen nicht herausziehen.

Johannes Kepler (1571–1630) war ein deutscher Astronom und Mathematiker und Assistent von Tycho Brahe. Er vertrat das kopernikanische Planetensystem und versuchte sein Leben lang, die Gesetzmäßigkeiten zu finden, die die Bewegungen der Planeten bestimmen. Er suchte nach einer mathematischen Beschreibung, was für seine Zeit ein geradezu radikaler Ansatz war. Sein Ausgangspunkt waren die von Tycho Brahe gesammelten Daten, die er selbst in späteren Jahren erweiterte. Nach vielen erfolglosen Versuchen und langen und mühsamen Berechnungen gelang es Kepler schließlich, die Daten Tycho Brahes in drei einfachen Gesetzen, den nach ihm benannten **Keplerschen Gesetzen**, zusammenzufassen. Nachdem er 1604 erkannt hatte, daß die Marsbahn eine Ellipse ist, veröffentlichte er die ersten beiden Gesetze in „Astronomia Nova“ im Jahre 1609, das dritte zehn Jahre später in seinem Hauptwerk „Harmonica Mundi“ [Greiner 1989, Zey 1997, Feynman *et al.* 1963].

1. Jeder Planet bewegt sich um die Sonne auf einer Ellipse, in deren einem Brennpunkt die Sonne steht.
2. Der Radiusvektor von der Sonne zum Planeten überstreicht in gleichen Zeitintervallen gleiche Flächen.
3. Die Quadrate der Umlaufzeiten zweier Planeten verhalten sich zueinander wie die Kuben der Hauptachsen ihrer Umlaufbahnen: $T \sim a^{\frac{3}{2}}$.

Tycho Brahe hatte eine große Menge astronomischer Daten gesammelt, Johannes Kepler fand die Gesetze, mit denen sie erklärt werden können. Er entdeckte das versteckte Wissen und wurde so zu einem der bekanntesten „Wissensentdecker“ der Geschichte.

Heute sind die Arbeiten von Tycho Brahe fast vergessen, seine Kataloge haben nur noch historischen Wert. Kein Lehrbuch der Astronomie enthält Auszüge seiner Messungen. Seine Beobachtungen und genauen Aufzeichnungen sind rohe Daten und haben daher einen entscheidenden Nachteil: Sie vermitteln uns keine Einsichten in die zugrundeliegenden Mechanismen und erlauben uns nicht, Voraussagen zu machen. Keplers Gesetze werden dagegen in allen Astronomie- und Physiklehrbüchern behandelt, denn sie geben die Prinzipien an, nach denen sich sowohl Planeten als auch Kometen bewegen. Sie fassen alle Beobachtungen Brahes in drei einfachen Aussagen

zusammen. Außerdem lassen sie Voraussagen zu: Kennt man die Position und die Geschwindigkeit eines Planeten zu einem bestimmten Zeitpunkt, so kann man mit Hilfe der Keplerschen Gesetze seine weitere Bahn berechnen.

1.2 Wissensentdeckung und Datenanalyse

Wie hat Johannes Kepler seine Gesetze gefunden? Wie ist es ihm gelungen, in Tycho Brahes langen Tabellen und umfangreichen Katalogen jene einfachen Gesetze zu entdecken, die die Astronomie revolutionierten? Wir wissen nur wenig darüber. Er wird eine große Zahl von Hypothesen ausprobiert haben, von denen die meisten scheiterten, und er muß lange und komplizierte Berechnungen angestellt haben. Wahrscheinlich haben außerordentliches mathematisches Talent, hartnäckige Arbeit und eine nicht geringe Portion Glück schließlich zum Erfolg geführt. Sicher ist auf jeden Fall, daß er nicht über eine universelle Methode zur Entdeckung physikalischer oder astronomischer Gesetze verfügte.

Auch heute kennen wir keine solche Methode. Es ist immer noch einfacher, Daten zusammenzutragen, von denen wir in der heutigen „Informationsgesellschaft“ (was auch immer das heißen mag) förmlich überschwemmt werden, als Wissen zu gewinnen. Wir brauchen nicht einmal mehr, wie Tycho Brahe es noch mußte, gewissenhaft und ausdauernd zu arbeiten, um Daten zu sammeln. Automatische Meßgeräte, Scanner, Digitalkameras und Computer haben uns diese Last abgenommen. Die moderne Datenbanktechnologie erlaubt uns außerdem, immer größere Mengen von Daten bequem und leicht abrufbar zu speichern. Es ist in der Tat, wie John Naisbett sagte: „We are drowning in information, but starving for knowledge.“

Wenn es einen solchen hervorragenden Forscher wie Johannes Kepler mehrere Jahre kostete, die von Tycho Brahe gesammelten Daten auszuwerten, deren Umfang — gemessen an heutigen Standards — vernachlässigbar klein erscheint und von denen er sogar nur die Daten über die Marsbahn benutzte, wie können wir dann hoffen, mit den enormen Datenmengen fertig zu werden, denen wir heute gegenüberstehen? „Manuelle“ Analysen sind schon lange nicht mehr durchführbar. Einfache Hilfsmittel, wie z.B. die Darstellung von Daten in Diagrammen und Schaubildern gelangen schnell an ihre Grenzen. Wenn wir nicht einfach vor der Datenflut kapitulieren wollen, sind wir gezwungen, intelligente, rechnergestützte Verfahren zu entwickeln, mit denen die Datenanalyse wenigstens teilweise automatisiert werden kann. Dies sind die Methoden, nach denen in den Forschungsgebieten *Wissensentdeckung in Datenbanken* (knowledge discovery in databases, KDD) und

Data Mining gesucht wird. Zwar können diese Methoden bei weitem nicht Menschen wie Johannes Kepler ersetzen, aber es ist nicht ganz unplausibel anzunehmen, daß er, unterstützt durch diese Methoden und Werkzeuge, sein Ziel etwas schneller erreicht hätte.

Oft werden die Begriffe „Wissensentdeckung“ und „Data Mining“ synonym gebraucht. Wir wollen sie hier jedoch unterscheiden. Mit dem Begriff **Wissensentdeckung in Datenbanken** bezeichnen wir einen Prozeß, der aus mehreren Schritten oder Stufen besteht, und der überlicherweise so charakterisiert wird [Fayyad *et al.* 1996]:

> Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.
>
> (Wissensentdeckung in Datenbanken ist der nicht-triviale Prozeß, gültige, neue, potentiell nützliche, und schließlich verständliche Muster in Daten zu finden.)

En Schritt dieses Prozesses, wenn auch sicherlich einer der wichtigstens, ist die eigentliche **Datenanalyse** oder **Data Mining**. In diesem Schritt werden Modellierungs-, Analyse- und Entdeckungstechniken angewandt.

1.2.1 Der Prozeß der Wissensentdeckung

In diesem Abschnitt unterteilen wir den Wissensentdeckungsprozeß (KDD-Prozeß) in zwei Vor- und fünf Hauptstufen. Die Struktur, die wir hier besprechen, ist jedoch nicht als bindend anzusehen. Bisher hat sich die Forschungsgemeinschaft noch nicht auf ein einheitliches, allgemein anerkanntes Schema einigen können. Die hier vorgestellte Struktur ist jedoch eng mit dem CRISP-DM-Modell (CRoss Industry Standard Process for Data Mining) [Chapman *et al.* 1999] verwandt, das eine genaue Erklärung des KDD-Prozesses liefert und einigen Einfluß hat, da es von großen Industrieunternehmen wie NCR und DaimlerChrysler unterstützt wird.

Vorstufen

- Abschätzung des erreichbaren Nutzens
- Zieldefinition, Machbarkeitsstudie

Hauptstufen

- Prüfen der Verfügbarkeit und Qualität der Daten, Datenauswahl, wenn nötig: Datensammlung

- Vorverarbeitung (60-80% des Gesamtaufwandes)
  - Vereinheitlichung und Transformation der Datenformate
  - Datenreinigung
    (Fehlerkorrektur, Ausreißererkennung, Auffüllen fehlender Werte)
  - Reduktion / Fokussierung
    (Ziehen von Stichproben, Attributauswahl, Prototyperzeugung)
- **Datenanalyse und Data Mining**
  (mit einer Vielzahl von Methoden)
- Visualisierung
  (auch parallel zu Vorverarbeitung, Datenanalyse und Interpretation)
- Interpretation, Bewertung und Test der Ergebnisse
- Anwendung und Dokumentation

Die Vorstufen dienen i.w. dazu, herauszufinden, ob die Haupstufen durchlaufen werden sollten. Denn nur wenn der erreichbare Nutzen groß genug ist und die Anforderungen durch Datenanalyse- und Data-Mining-Methoden erfüllt werden können, kann man überhaupt hoffen, daß die meist recht aufwendigen und teuren Hauptstufen Früchte tragen werden.

In den Hauptstufen werden die Daten, die nach verstecktem Wissen durchforstet werden sollen, zunächst gesammelt (wenn nötig), geeignete Teilmengen werden ausgewählt, und die Daten werden in ein einheitliches Format überführt, auf das sich Datenanalyse und Data-Mining-Methoden leicht anwenden lassen. Dann werden sie gereinigt und reduziert, um die Leistung, die Datenanalysealgorithmen erbringen können, durch Verbesserung der Voraussetzungen, insbesondere der Datenqualität, zu erhöhen. Diese Vorverarbeitungsschritte verschlingen i.a. den größten Teil der Gesamtkosten. Je nach den Zielen und Aufgaben (eine Liste wichtiger Aufgaben ist im nächsten Abschnitt zusammengestellt), die im Zieldefinitionsschritt (2. Vorstufe) identifiziert wurden, werden anschließend Datenanalyse- und Data-Mining-Methoden angewandt (eine Liste wichtiger Methoden ist im übernächsten Abschnitt zusammengestellt), deren Ergebnisse visualisiert werden können, um sie zu interpretieren und zu bewerten.

Da das gewünschte Ziel, z.B. die gewünschte Vorhersagegüte, selten im ersten Durchlauf erreicht wird, müssen mehrere Schritte der Vorverarbeitungsphase (z.B. die Attributauswahl) und die Anwendung von Datenanalyse- und Data-Mining-Methoden ggf. mehrfach wiederholt werden, um die Ergebnisse zu verbessern. Sollte es nicht schon vorher offensichtlich gewesen

sein, so wird dadurch deutlich, daß die Wissensentdeckung in Datenbanken kein völlig automatisierbarer, sondern ein interaktiver Prozeß ist. Ein Anwender muß die Ergebnisse bewerten, sie auf Plausibilität prüfen und sie gegen zurückbehaltene Daten testen. Wenn nötig, muß er den Gang des Prozesses ändern, so daß dieser seine Anforderungen erfüllt.

1.2.2 Datenanalyse: Aufgaben

Im Laufe der Zeit haben sich typische Aufgaben herauskristallisiert, die Datenanalyse- und Data-Mining-Aufgaben lösen können sollten (obwohl natürlich nicht jede Methode in der Lage sein muß, jede beliebige Aufgabe zu lösen — es ist gewöhnlich die Kombination von Methoden, die die Stärke eines Verfahrens ausmacht). Zu diesen Aufgaben gehören insbesondere die in der folgenden — sicherlich unvollständigen — Liste aufgeführten. Wir kennzeichnen sie nicht nur durch ihren Namen, sondern auch durch eine typische Fragestellung [Nakhaeizadeh 1998b].

- Klassifikation
  Ist dieser Kunde kreditwürdig?
- Segmentierung, Clustering
  Welche Kundengruppen habe ich?
- Konzeptbeschreibung
  Welche Eigenschaften kennzeichnen fehleranfällige Fahrzeuge?
- Vorhersage, Trendanalyse
  Wie wird sich der Wechselkurs des US Dollar entwickeln?
- Abhängigkeits-/Assoziationsanalyse
  Welche Produkte werden häufig zusammen gekauft?
- Abweichungsanalyse
  Gibt es saisonbedingte oder regionale Umsatzschwankungen?

Klassifikation und Vorhersage sind die häufigsten Aufgaben, da ihre Lösung direkte Auswirkungen z.B. auf den Umsatz und den Gewinn eines Unternehmens haben kann. Abhängigkeits- und Assoziationsanalyse stehen an zweiter Stelle, denn sie können z.B. benutzt werden, um eine Warenkorbanalyse durchzuführen (d.h., um Produkte zu identifizieren, die häufig zusammen gekauft werden) oder um die Fehlerdiagnose von Produkten zu unterstützen, und sind daher auch von hohem wirtschaftlichem Interesse.

1.2.3 Datenanalyse: Methoden

Die Forschung im Bereich der Datenanalyse und des Data Mining ist hochgradig interdisziplinär. Methoden, mit denen die im vorangehenden Abschnitt genannten Aufgaben bearbeitet werden können, wurden in einer Vielzahl von Forschungsbereichen entwickelt, unter anderem, um nur die wichtigsten zu nennen, Statistik, künstliche Intelligenz, maschinelles Lernen und Soft Computing. Es gibt daher ein ganzes Arsenal an Methoden, die auf einem weiten Spektrum von Ideen beruhen. Um einen Überblick zu geben, listen wir unten einige der bekanntesten Datenanalyse- und Data-Mining-Methoden auf. Jeder Listeneintrag verweist auf einige wichtige Veröffentlichungen zu der jeweiligen Methode und zeigt auf, für welche Aufgaben die Methode besonders geeignet ist. Natürlich ist diese Liste weit davon entfernt, vollständig zu sein. Außerdem sind die Verweise notwendigerweise unvollständig und u.U. nicht die bestmöglichen, da der Verfasser dieses Skriptes natürlich nicht Experte für alle diese Methoden sein, alle wesentlichen Veröffentlichungen kennen und natürlich auch nicht jeden nennen kann, der zur Entwicklung einer Methode beigetragen hat.

- Klassische Statistik
  (Parameterschätzung, Hypothesentest, Modelauswahl etc.)
  [Larsen and Marx 1986, Everitt 1998]
  Klassifikation, Vorhersage, Trendanalyse
- Entscheidungs- und Regressionsbäume
  [Breiman *et al.* 1984, Quinlan 1986, Quinlan 1993]
  Klassifikation, Vorhersage
- Bayes-Klassifikatoren
  [Good 1965, Duda and Hart 1973, Langley *et al.* 1992]
  Klassifikation, Vorhersage
- Probabilistische Netze (Bayes-Netze/Markow-Netze)
  [Pearl 1988, Lauritzen and Spiegelhalter 1988, Heckerman *et al.* 1995]
  Klassifikation, Abhängigkeitsanalyse
- (Künstliche) neuronale Netze
  [Anderson 1995a, Rojas 1996, Haykin 1994, Zell 1994]
  Klassifikation, Vorhersage, Clustering
- Neuro-Fuzzy-Regelgenerierung
  [Wang and Mendel 1992, Nauck and Kruse 1997, Nauck *et al.* 1997]
  Klassifikation, Vorhersage

- k-nächste-Nachbarn/fallbasiertes Schließen
  [Dasarathy 1990, Aha 1992, Kolodner 1993, Wettschereck 1994]
  Klassifikation, Vorhersage
- Induktive Logikprogrammierung
  [Muggleton 1992, de Raedt and Bruynooghe 1993]
  Klassifikation, Assoziationsanalyse, Konzeptbeschreibung
- Assoziationsregeln
  [Agrawal *et al.* 1993, Agrawal *et al.* 1996, Srikant and Agrawal 1996]
  Assoziationsanalyse
- Hierarchische und probabilistische Clusteranalyse
  [Bock 1974, Everitt 1981, Cheeseman *et al.* 1988, Mucha 1992]
  Segmentatierung, Clustering
- Fuzzy-Clusteranalyse
  [Bezdek and Pal 1992, Bezdek *et al.* 1999, Höppner *et al.* 1999]
  Segmentatierung, Clustering
- Conceptual Clustering
  [Michalski and Stepp 1983, Fisher 1987, Hanson and Bauer 1989]
  Segmentatierung, Konzeptbeschreibung
- und viele weitere

Obwohl es für jede Datenanalyseaufgabe mehrere verläßliche Methoden gibt, mit denen sie gelöst werden kann, gibt es, wie bereits oben angedeutet, nicht eine Methode, die alle Aufgaben lösen kann. Viele Verfahren sind auf eine bestimmte Aufgabe zugeschnitten und jede von ihnen hat unterschiedliche Stärken und Schwächen. Außerdem müssen gewöhnlich mehrere Methoden kombiniert werden, um gute Ergebnisse zu erzielen. Daher bieten kommerzielle Datenanalyseprogramme, wie z.B. Clementine (SPSS, Chicago, IL, USA), DataEngine (Management Intelligenter Technologien GmbH, Aachen, Deutschland), oder Kepler (Dialogis GmbH, Sankt Augustin, Deutschland) mehrere der oben aufgeführten Methoden unter einer einfach zu bedienenden graphischen Benutzeroberfläche an.[1] Bisher gibt es jedoch kein System, das alle oben aufgeführten Verfahren anbietet.

[1] Diese Datenanalyseprogramme wurden als Beispiele ausgewählt, weil der Verfasser dieses Skriptes mit ihnen recht gut vertraut ist. Für Clementine hat er das Assoziationsregel-Lernprogramm entwickelt, das dem „Apriori-Knoten“ zugrundeliegt, und das Plug-In „DecisionXpert“ für DataEngine basiert auf einem von ihm entwickelten Entscheidungsbaum-Lernprogramm [Borgelt 1998, Borgelt and Timm 2000]. Letzteres Programm wurde auch in Kepler eingebaut. Dies bedeutet jedoch nicht, daß die genannten Programme anderen auf dem Markt verfügbaren Programmen überlegen seien.

Einen Überblick über und eine Bewertung verschiedener Datenanalysewerkzeuge gibt [Gentsch 1999]. Ausführliche Listen verfügbarer Datenanalyseprogramme findet man z.B. auf den WWW-Seiten

`http://www.statserv.com/datamsoft.html` und
`http://www.xore.com/prodtable.html`.

Diese Seiten bieten Checklisten der von den Programmen angebotenen Methoden und Verweise auf die WWW-Seiten der einzelnen Programme und der Firmen, die sie anbieten.

Kapitel 2

Statistik

In [Sachs 1999] wird die Statistik so charakterisiert:

> Statistik ist die Kunst, Daten zu gewinnen, darzustellen, zu analysieren und zu interpretieren, um zu neuem Wissen zu gelangen.

Diese Charakterisierung legt bereits nahe, daß die Statistik für die Datenanalyse sehr wichtig ist. In der Tat gibt es eine ganze Reihe statistischer Verfahren, mit denen die Aufgabentypen bearbeitet werden können, durch die wir im vorangehenden Kapitel den Data-Mining-Schritt des KDD-Prozesses charakterisiert haben, oder die Hilfsfunktionen erfüllen. Einige dieser Verfahren stellen wir in diesem Kapitel vor, erheben jedoch keineswegs den Anspruch, einen vollständigen Überblick zu geben.

Die statistischen Verfahren, die wir betrachten werden, lassen sich grob in zwei Klassen einordnen, die den beiden Hauptgebieten entsprechen, in die sich die Statistik einteilen läßt, nämlich

- die *beschreibende (deskriptive) Statistik* (Abschnitt 2.2) und die
- die *beurteilende (schließende, induktive) Statistik* (Abschnitt 2.4).

In der beschreibenden Statistik versucht man, Daten durch graphische Darstellungen übersichtlicher zu machen und durch Berechnung von Kenngrößen zusammenzufassen. In der beurteilenden Statistik versucht man, Schlüsse über den datenerzeugenden Prozeß zu ziehen, also etwa Parameter dieses Prozesses zu schätzen oder ein Modell auszuwählen. Grundlage vieler Verfahren der beurteilenden Statistik ist die Wahrscheinlichkeitsrechnung (Abschnitt 2.3), Ziel ist gewöhnlich die Vorbereitung von Entscheidungen.

2.1 Begriffe und Notation

Ehe wir uns den Verfahren zuwenden können, müssen wir Begriffe einführen, mit denen wir über Daten reden können.

- **Objekt, Fall**
  Durch Daten werden Objekte, Fälle, Personen etc. beschrieben. Z.B. werden durch medizinische Daten gewöhnlich Patienten beschrieben, durch Lagerhaltungsdaten gewöhnlich Bauteile oder Produkte etc.
- **(Zufalls-)Stichprobe**
  Die Menge der Objekte bzw. Fälle, die durch einen Datensatz beschrieben wird, nennen wir *Stichprobe*, die Größe dieser Menge (Anzahl Elemente) den *Umfang* der Stichprobe. Sind die Objekte bzw. Fälle Ergebnisse eines Zufallsexperimentes (z.B. Ziehen von Lottozahlen), so sprechen wir von einer *Zufallsstichprobe*.
- **Merkmal, Attribut**
  Die Objekte bzw. Fälle der Stichprobe werden durch ihre *Merkmale* (Eigenschaften) beschrieben. Z.B. könnten Patienten durch die Merkmale Geschlecht, Alter, Gewicht, Blutgruppe etc., Bauteile durch ihre Abmessungen, ihre elektrischen Kenngrößen etc. beschrieben werden. Die Objekte bzw. Fälle einer Stichprobe nennt man daher oft auch allgemein *Merkmalsträger*.
- **Merkmalsausprägung, Merkmalswert, Attributwert**
  Die Merkmale, mit Hilfe derer die Objekte/Fälle beschrieben werden, können unterschiedliche *Ausprägungen* oder *Werte* annehmen. Z.B. kann das Geschlecht eines Patienten *weiblich* oder *männlich* sein, sein Alter eine positive ganze Zahl etc. Die Menge der Werte, die ein Merkmal annehmen kann, nennen wir seinen *Wertebereich*.
- **Stichprobenwert**
  Der Merkmalswert, den ein Merkmal für ein gegebenes Objekt bzw. einen gegebenen Fall der Stichprobe annimmt, heißt *Stichprobenwert*.

Nach der Art der möglichen Merkmalsausprägungen unterscheidet man verschiedene **Skalenarten** (Merkmalsarten, Attributtypen). Diese Unterscheidung ist wichtig, da sich z.B. bestimmte Kenngrößen nur für bestimmte Skalenarten berechnen lassen. Auch setzen einige statistische Verfahren Merkmale bestimmter Art voraus. In Tabelle 2.1 sind die wichtigsten Skalenarten — **nominal**, **ordinal** und **metrisch** — mit den auf ihnen möglichen Operationen und einigen Beispielen zusammengestellt.

Skalenart	mögliche Operationen	Beispiele
nominal (kategorial, qualitativ)	Gleichheit	Geschlecht Blutgruppe
ordinal (rangskaliert, komparativ)	Gleichheit größer/kleiner	Schulnote Windstärke
metrisch (intervallskaliert, quantitativ)	Gleichheit größer/kleiner Differenz evtl. Verhältnis	Länge Gewicht Zeit Temperatur

Tabelle 2.1: Die wichtigsten Skalenarten.

Bei Nominalskalen unterscheidet man oft noch nach der Anzahl der Ausprägungen *dichotome* bzw. *alternative* Merkmale (zwei Ausprägungen) und *polytome* Merkmale (mehr als zwei Ausprägungen). Bei metrischen Skalen unterscheidet man manchmal noch danach, ob nur Differenzen (Temperatur, Kalenderzeit) oder auch Verhältnisse (Länge, Gewicht, Zeitdauer) sinnvoll berechnet werden können. Im ersten Fall spricht man dann von einer *Intervallskala*, im zweiten von einer *Verhältnisskala*. Wir werden von diesen zusätzlichen Unterscheidungen hier jedoch absehen.

Wie aus den obigen Begriffserläuterungen bereits hervorgeht, besteht ein Datensatz aus der Angabe von Stichprobenwerten für die Merkmalsträger einer Stichprobe. Die Anzahl der Merkmale, die zur Beschreibung der Stichprobe verwendet werden, nennen wir die *Dimension* des Datensatzes.

Eindimensionale Datensätze bezeichnen wir durch kleine Buchstaben vom Ende des Alphabetes, also z.B. x, y, z. Diese Buchstaben stehen für das Merkmal, das zur Beschreibung verwendet wird. Die Elemente des Datensatzes (die Stichprobenwerte) bezeichnen wir mit dem gleichen Buchstaben und geben ihre Position im Datensatz durch einen Index an, allgemein also $x = (x_1, x_2, \ldots, x_n)$ für eine Stichprobe vom Umfang n. (Ein Datensatz wird als Vektor und nicht als Menge geschrieben, da verschiedene Merkmalsträger den gleichen Stichprobenwert haben können.) Mehrdimensionale Datensätze schreiben wir als Vektoren von kleinen Buchstaben vom Ende des Alphabetes. Die Elemente des Datensatzes sind dann ihrerseits Vektoren von Stichprobenwerten. Einen zweidimensionalen Datensatz beispielsweise schreiben wir $(x, y) = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, wobei x und y für die beiden Merkmale stehen, durch die die Stichprobe beschrieben wird.

a_k	h_k	r_k	$\sum_{i=1}^{k} h_i$	$\sum_{i=1}^{k} r_i$
1	2	$\frac{2}{25} = 0.08$	2	$\frac{2}{25} = 0.08$
2	6	$\frac{6}{25} = 0.24$	8	$\frac{8}{25} = 0.32$
3	9	$\frac{9}{25} = 0.36$	17	$\frac{17}{25} = 0.68$
4	5	$\frac{5}{25} = 0.20$	22	$\frac{22}{25} = 0.88$
5	3	$\frac{3}{25} = 0.12$	25	$\frac{25}{25} = 1.00$

Tabelle 2.2: Eine einfache Häufigkeitstabelle, die die absoluten Häufigkeiten h_k, die relativen Häufigkeiten r_k, sowie die absoluten Summenhäufigkeiten $\sum_{i=1}^{k} h_i$ und die relativen Summenhäufigkeiten $\sum_{i=1}^{k} r_i$ zeigt.

2.2 Beschreibende Statistik

Die beschreibende Statistik hat die Aufgabe, Zustände und Vorgänge anhand von Beobachtungsdaten zu beschreiben. Hierzu dienen Tabellen und graphische Darstellungen sowie die Berechnung von Kenngrößen.

2.2.1 Tabellarische Darstellungen

Tabellen werden benutzt, um die Beobachtungsdaten selbst übersichtlich darzustellen, aber auch, um berechnete Kenngrößen zusammenzustellen.

Die einfachste tabellarische Darstellung eines (eindimensionalen) Datensatzes ist die **Häufigkeitstabelle**, die die Grundlage für die meisten graphischen Darstellungen ist. In eine Häufigkeitstabelle wird zu jedem Merkmalswert seine (absolute oder relative) Häufigkeit in der Stichprobe eingetragen, wobei die **absolute Häufigkeit** h_k einfach die Häufigkeit des Auftretens eines Merkmalswertes a_k in der Stichprobe ist und die **relative Häufigkeit** r_k definiert ist als $r_k = \frac{h_k}{n}$ mit dem Stichprobenumfang n. Außerdem können Spalten für die kumulierten (absoluten oder relativen) Häufigkeiten (auch *Summenhäufigkeiten* genannt, vorhanden sein. Wir betrachten als Beispiel den Datensatz

$$x = (3, 4, 3, 2, 5, 3, 1, 2, 4, 3, 3, 4, 4, 1, 5, 2, 2, 3, 5, 3, 2, 4, 3, 2, 3),$$

der z.B. die in einer Klausur verteilten Noten beschreiben könnte. Eine diesen Datensatz darstellende Häufigkeitstabelle ist in Tabelle 2.2 gezeigt. Offenbar vermittelt diese Tabelle einen wesentlich besseren Eindruck als der oben gezeigte Datensatz, der nur die Stichprobenwerte auflistet.

	a_1	a_2	a_3	a_4	$\sum$
b_1	8	3	5	2	18
b_2	2	6	1	3	12
b_3	4	1	2	7	14
$\sum$	14	10	8	12	44

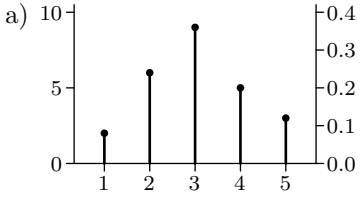
Tabelle 2.3: Eine Kontingenztafel für zwei Attribute A und B.

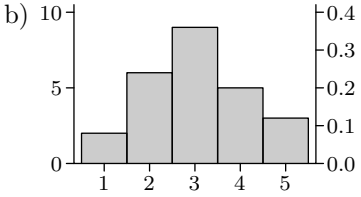
Eine zwei- oder mehrdimensionale Häufigkeitstabelle, in die für jede Merkmals*kombination* ihre (absolute oder relative) Häufigkeit eingetragen wird — nennt man auch **Kontingenztafel** (seltener: *Kontingenztabelle*). Ein Beispiel einer solchen Kontingenztafel für zwei Attribute A und B (mit absoluten Häufigkeiten), die auch die Zeilen- und Spaltensummen, also auch die Häufigkeiten der einzelnen Attribute enthält, zeigt Tabelle 2.3.

2.2.2 Graphische Darstellungen

Graphische Darstellungen dienen dazu, tabellarische Daten anschaulicher zu machen. Dazu nutzt man im wesentlichen aus, daß Zahlenwerte und Häufigkeiten durch geometrische Größen — z.B. Längen, Flächen oder Winkel — dargestellt werden können, die ein Mensch leichter erfassen kann als abstrakte Zahlenwerte. Die wichtigsten Typen graphischer Darstellungen sind:

- **Stab-/Balken-/Säulendiagramm**
  Zahlenwerte, die z.B. Häufigkeiten des Auftretens verschiedener Merkmalswerte in einer Stichprobe sein können, werden durch die Längen von Stäben, Balken oder Säulen dargestellt. Man erhält so einen guten Eindruck der Zahlenverhältnisse (siehe Abbildung 2.1a und b, in der die Häufigkeiten aus Tabelle 2.2 dargestellt sind).
- **Flächen- und Volumendiagramm**
  Eng verwandt mit Stab- und Balkendiagrammen sind Flächen- und Volumendiagramme, bei denen die Zahlenwerte — wie die Namen ja schon sagen — durch Flächen und Volumen statt durch Längen dargestellt werden (siehe Abbildung 2.2, in der wieder die Häufigkeiten aus Tabelle 2.2 dargestellt sind). Flächen- und Volumendiagramme sind meist weniger anschaulich (es sei denn, die darzustellenden Größen sind Flächen oder Volumen), da Menschen Schwierigkeiten haben, Flächeninhalte und Volumen zu vergleichen und sich dabei oft verschätzen. Dies sieht man schon in Abbildung 2.2: Kaum jemand





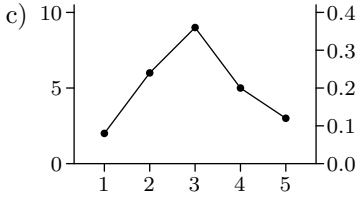


Abbildung 2.1: Stab- (a) und Balkendiagramm (b) sowie Häufigkeitspolygon (c) für die Daten aus Tabelle 2.2.

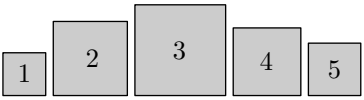


Abbildung 2.2: Flächendiagramm für die Daten aus Tabelle 2.2.

schätzt den Flächeninhalt des Quadrates für den Wert 3 (Häufigkeit 9) dreimal so groß wie den des Quadrates für den Wert 5 (Häufigkeit 3).

- **Häufigkeitspolygon und Liniendiagramm**
  Ein Häufigkeitspolygon entsteht, wenn man die Stabenden eines Stabdiagrams durch ein Polygon verbindet. Es kann vorteilhaft sein, wenn die Merkmalswerte eine Ordnung aufweisen und man die Entwicklung der Häufigkeit über dieser Ordnung darstellen will (siehe Abbildung 2.1c). Es kann auch verwendet werden, wenn z.B. Zahlenwerte in Abhängigkeit von der Zeit dargestellt werden sollen. In diesem Fall spricht man von einem Liniendiagramm.

- **Torten- und Streifendiagramm**
  Torten- und Streifendiagramme eignen sich besonders gut, wenn Anteile an einer Gesamtmenge, also z.B. relative Häufigkeiten, deutlich gemacht werden sollen. In einem Tortendiagramm werden die Anteile durch Winkel, in einem Streifendiagramm durch Längen dargestellt (siehe Abbildung 2.3).

- **Mosaikdiagramm**
  Mit einem Mosaikdiagramm können Kontingenztafeln (d.h. zwei- oder mehrdimensionale Häufigkeitstabellen) dargestellt werden. Für das er-

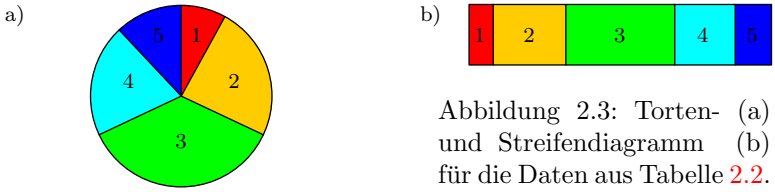


Abbildung 2.3: Torten- (a) und Streifendiagramm (b) für die Daten aus Tabelle 2.2.

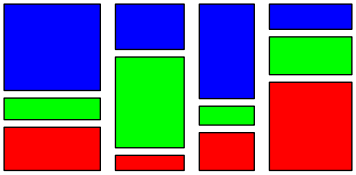


Abbildung 2.4: Ein Mosaikdiagramm für die Kontingenztafel aus Tabelle 2.3.

ste Attribut wird die horizontale Richtung wie in einem Streifendiagramm unterteilt. Jeder Abschnitt wird dann gemäß der Anteile des zweiten Attributes vertikal aufgeteilt, und zwar wieder wie in einem Streifendiagramm (siehe Abbildung 2.4). Im Prinzip können beliebig viele Attribute verwendet werden, indem man die entstehenden Mosaiksteine weiter abwechselnd in horizontaler und vertikaler Richtung teilt. Allerdings kann die Übersichtlichkeit schnell verlorengehen, wenn nicht die Fugenbreiten und Einfärbungen der Mosaiksteine zur Unterstützung der Zuordnung zu den Attributen verwendet werden.

- **Histogramm**
  Ein Histogramm sieht im Prinzip genauso aus wie ein Balkendiagramm, jedoch ist der Wertebereich des zugrundeliegenden Attributes metrisch. Dadurch kann man i.a. nicht einfach die Häufigkeiten der verschiedenen Attributwerte auszählen, sondern muß Zählabschnitte (engl.: bins, buckets) bilden. Die Breite (oder bei festem Wertebereich äquivalent: die Anzahl) dieser Zählabschnitte sind vom Anwender zu wählen. Alle Zählabschnitte sollten die gleiche Breite haben, da Histogramme mit unterschiedlich breiten Zählabschnitten schwer zu lesen sind, und zwar aus den gleichen Gründen, aus denen Flächendiagramme (siehe oben) schwerer zu interpertieren sind als Balkendiagramme. Außerdem kann es von der Abschnittsbreite und der Lage der Abschnittgrenzen abhängen, ob ein Histogramm einen guten Eindruck der Daten vermittelt.

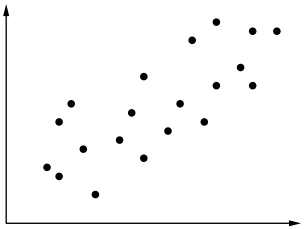


Abbildung 2.5: Ein einfaches Streudiagramm.

- **Streudiagramm (scatter plot)**
  In einem Streudiagramm wird ein zweidimensionaler Datensatz metrischer Größen dargestellt, indem die Stichprobenwerte als Punktkoordinaten verwendet werden (siehe Abbildung 2.5). Ein Streudiagramm ist sehr geeignet, wenn man herausfinden möchte, ob zwischen den beiden dargestellten Größen eine Abhängigkeit besteht (vergleiche auch Abschnitt 2.2.4 und Kapitel 3).

Beispiele, wie graphische Darstellungen täuschen können (was in der Presse und in Werbeprospekten von Unternehmen z.T. bewußt ausgenutzt wird), findet man in den sehr lesenswerten Büchern [Huff 1954, Krämer 1997].

2.2.3 Kenngrößen für eindimensionale Daten

Das Ziel der Berechnung von Kenngrößen ist, einen Datensatz durch wenige, möglichst charakteristische Maßzahlen zu beschreiben. Man unterscheidet im wesentlichen drei Arten von Kenngrößen:

- **Lagemaße (Lokalisationsmaße)**
  Lagemaße geben die Lage der Daten im Wertebereich eines Merkmals durch einen einzelnen Zahlenwert an.
- **Streuungsmaße (Dispersionsmaße)**
  Streuungsmaße geben an, wie stark die Daten um den Lageparameter streuen (wie stark sie von ihm abweichen) und zeigen so, wie gut der Lageparameter die Daten beschreibt.
- **Formmaße**
  Formmaße beschreiben die Form der Verteilung der Daten relativ zu einer Referenzform. Als Referenzform wird üblicherweise eine Gaußsche Glockenkurve gewählt.

Diese Kenngrößen, die sich später in der beurteilenden Statistik als sehr nützlich erweisen, werden wir im folgenden genauer besprechen.

2.2.3.1 Lagemaße (Lokalisationsmaße)

Wie bereits gesagt, wird durch ein *Lage-* oder *Lokalisationsmaß* die Lage der Daten im Wertebereich eines Merkmals durch einen einzelnen Merkmalswert beschrieben. Dieser Wert soll die Daten möglichst gut wiedergeben. Z.B. sollte die Summe der Abweichungen der Daten von diesem Wert möglichst gering sein. Die wichtigsten Lagemaße sind der *Modalwert*, der *Median* (oder *Zentralwert*) und seine Verallgemeinerung, die *Quantile*, sowie der *Mittelwert*, der am häufigsten verwendet wird.

Modalwert Als (empirischen) *Modalwert* x^* (auch *Modus* oder *Mode*) bezeichnet man einen Merkmalswert, der in dem zu beschreibenden Datensatz am häufigsten auftritt. Er muß offenbar nicht eindeutig bestimmt sein, da es mehrere Werte geben kann, die gleich häufig auftreten. Modalwerte können für alle Skalentypen bestimmt werden, da ihre Bestimmung nur einen Test auf Gleichheit erfordert. Der Modalwert ist daher das allgemeinste Lagemaß. Für metrische Daten ist er jedoch wegen der großen Zahl möglicher Merkmalswerte meist schlechter geeignet als andere Maße. Man kann sich im Falle metrischer Daten jedoch u.U. behelfen, indem man den Mittelwert eines höchsten Histogrammstreifens als Modalwert definiert.

Median (Zentralwert) Man kann den (empirischen) *Median* oder *Zentralwert* $\tilde{x}$ einführen als einen Wert, der die Summe der absoluten Abweichungen minimiert, für den also gilt

$$\sum_{i=1}^{n} |x_i - \tilde{x}| = \min.$$

Um einen Wert für $\tilde{x}$ zu bestimmen, leitet man die linke Seite des obigen Ausdrucks nach $\tilde{x}$ ab und setzt diese Ableitung gleich 0. Man erhält:

$$\sum_{i=1}^{n} \operatorname{sgn}(x_i - \tilde{x}) = 0,$$

wobei sgn die Vorzeichenfunktion ist.[1] Folglich ist der Median ein Wert, der „in der Mitte der Daten“ liegt. D.h., es gibt in den Daten genauso viele Werte, die größer, wie Werte, die kleiner sind (daher auch *Zentralwert*).

[1] Man beachte, daß bei der üblichen Definition der Vorzeichenfunktion diese Gleichung nicht immer erfüllbar ist. Man begnügt sich dann mit einer besten Annäherung an 0.

Mit der obigen Charakterisierung ist der Median jedoch nur bei einer ungeraden Anzahl von Merkmalsträgern immer eindeutig bestimmt. Bei einer geraden Anzahl muß dies dagegen nicht der Fall sein, wie das folgende Beispiel zeigt: Gegeben sei der Datensatz $(1, 2, 3, 4)$. Offenbar minimiert jeder Wert aus dem Intervall $[2, 3]$ die Summe der absoluten Abweichungen. Um einen eindeutigen Wert für den Median zu erhalten, definiert man bei einer ungeraden Anzahl von Merkmalsträgern den Median als das arithmetische Mittel der beiden Werte, die in der Mitte des Datensatzes liegen, in dem obigen Beispiel also $\tilde{x} = \frac{2+3}{2} = \frac{5}{2}$.

Formal wird der Median so definiert: Sei $x = (x_{(1)}, \ldots, x_{(n)})$ ein sortierter Datensatz, d.h., es gelte $\forall i, j : (j > i) \rightarrow (x_{(j)} \geq x_{(i)})$. Dann heißt

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

der *(empirische) Median* des Datensatzes x.

Der Median kann für ordinale und metrische Merkmale bestimmt werden, da lediglich ein Test auf größer oder kleiner benötigt wird, nämlich für die Sortierung der Merkmalsausprägungen. Die Berechnung des Mittelwertes der beiden mittleren Stichprobenwerte bei geradem Stichprobenumfang wird bei rangskalierten Merkmalen durch Auswahl eines der beiden mittleren Werte ersetzt, so daß keine Rechnung notwendig ist. Die obige Charakterisierung des Medians über die absoluten Abweichungen, für die eine Differenzbildung nötig ist, dient im wesentlichen dazu, die Analogie zum weiter unten behandelten Mittelwert aufzuzeigen.

Der Median metrischer Merkmale ist unempfindlich gegen lineare Transformationen der Daten, d.h., es gilt

$$\widetilde{ax + b} = a\tilde{x} + b.$$

Quantile Wie wir gerade gesehen haben, ist der Median ein Wert, der so bestimmt ist, daß die Hälfte der Stichprobenwerte des Datensatzes kleiner sind. Diese Idee kann man verallgemeinern, indem man Werte so bestimmt, daß ein Anteil p, $0 < p < 1$, der Stichprobenwerte des Datensatzes keiner ist. Diese Werte nennt man (empirische) p-Quantile. Der Median ist speziell das (empirische) $\frac{1}{2}$-Quantil eines Datensatzes.

Weitere wichtige Quantile sind das 1., 2. und 3. Quartil, für die $\frac{1}{4}$, $\frac{2}{4}$ bzw. $\frac{3}{4}$ des Datensatzes kleiner sind. (Der Median ist also auch identisch mit dem 2. Quartil.), sowie die Dezile (k Zehntel des Datensatzes kleiner) und die Perzentile (k Hundertstel des Datensatzes kleiner).

Man beachte, daß bei metrischen Merkmalen zur genauen Berechnung des p-Quantils je nach Umfang der Stichprobe Anpassungen notwendig sein können, die der Bildung des arithmetischen Mittels der beiden mittleren Stichprobenwerte im Falle des Medians entsprechen.

Mittelwert Während der Median die *absoluten Abweichungen* der Datenpunkte minimiert, ist der (empirische) Mittelwert $\bar{x}$ ist der Wert, der die Summe der *Abweichungsquadrate* von den Stichprobenwerten minimiert, für den also gilt

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

Um einen Wert für $\bar{x}$ zu bestimmen, leitet man die linke Seite des obigen Ausdrucks nach $\bar{x}$ ab und setzt diese Ableitung gleich 0. Man erhält:

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

also

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Der Mittelwert ist folglich das arithmetische Mittel der Stichprobenwerte.

Der Mittelwert ist wie der Median unempfindlich gegen lineare Transformationen der Daten, denn es gilt

$$\overline{ax + b} = a\bar{x} + b.$$

Zwar ist der Mittelwert das am häufigsten verwendete Lagemaß, jedoch ist ihm bei

- rangskalierten Merkmalen,
- wenigen Meßwerten,
- asymmetrischen Verteilungen und
- Verdacht auf Ausreißer

der Median vorzuziehen, da der Median in diesen Fällen robuster ist und einen besseren Eindruck der Daten vermittelt. Um den Mittelwert robuster gegen Ausreißer zu machen, läßt man bei seiner Berechnung manchmal den größten und den kleinsten Stichprobenwert, zuweilen auch mehrere Extremwerte, weg.

2.2.3.2 Streuungsmaße (Dispersionsmaße)

Wie bereits oben erwähnt, geben Streuungsmaße an, wie stark die Daten um einen Lageparameter streuen und damit, wie gut der Lageparameter die Daten beschreibt. Denn ein Lageparameter allein, der ja nichts über die Größe der Abweichungen aussagt, kann über die wahre Situation täuschen, wie der folgende Statistikerwitz verdeutlicht: "A man with his head in the freezer and feet in the oven is *on the average* quite comfortable." (Ein Mann mit seinem Kopf im Kühlschrank und Füßen im Ofen hat es *im Durchschnitt* recht angenehm.)

Spannweite Als Spannweite bezeichnet man die Differenz zwischen dem größten und dem kleinsten in der Stichprobe auftretenden Merkmalswert.

$$R = x_{\max} - x_{\min} = \max{}_{i=1}^{n} x_i - \min{}_{i=1}^{n} x_i$$

Die Spannweite ist zwar ein sehr intuitives Streuungsmaß, jedoch leider sehr anfällig für Ausreißer.

Interquantilbereich Als p-Interquantilbereich, $0 < p < \frac{1}{2}$, bezeichnet man die Differenz zwischen dem (empirischen) $(1-p)$- und dem (empirischen) p-Quantil des Datensatzes. Häufig verwendete Interquantilbereiche sind der Interquartilbereich ($p = \frac{1}{4}$, Differenz zwischen dem 3. und 1. Quartil), der Interdezilbereich ($p = \frac{1}{10}$, Differenz zwischen dem 9. und 1. Dezil) und der Interperzentilbereich ($p = \frac{1}{100}$, Differenz zwischen dem 99. und 1. Perzentil). Für kleines p wird mit dem p-Interquantilbereich die Idee, den Mittelwert durch Weglassen von Extremwerten robuster zu machen, auf die Spannweite übertragen.

Mittlere absolute Abweichung Die mittlere absolute Abweichung ist das arithmetische Mittel der absoluten Abweichungen vom Median oder vom (empirischen) Mittelwert. Es ist

$$d_{\tilde{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

die mittlere absolute Abweichung bzgl. des Medians und

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

die mittlere absolute Abweichung bzgl. des Mittelwertes. Es gilt $d_{\tilde{x}} \leq d_{\bar{x}}$, da der Median die Summe der absoluten Abweichungen und damit natürlich auch die mittlere absolute Abweichung minimiert.

Varianz und Standardabweichung In Analogie zur absoluten Abweichung könnte man auch die mittlere quadratische Abweichung berechnen. (Man erinnere sich, daß der Mittelwert die Summe der quadratischen Abweichungen minimiert.) Doch statt

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

verwendet man üblicherweise

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

als Streuungsmaß und bezeichnet es als die (empirische) Varianz der Stichprobe. Den Grund für den Nenner $n-1$ liefert die beurteilende Statistik, in der die Kenngrößen der beschreibenden Statistik eine wichtige Rolle spielen. Eine genauere Erklärung wird im Abschnitt 2.4.2 über Parameterschätzung nachgeliefert (Erwartungstreue des Schätzers für die Varianz).

Die positive Quadratwurzel aus der Varianz, also

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

nennt man die (empirische) Standardabweichung oder (empirische) Streuung der Stichprobe.

Man beachte, daß die (empirische) Varianz oft bequemer durch die sich aus der folgenden Umformung ergebende Formel berechnet werden kann:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)
\end{aligned}
$$

Der Vorteil dieser Formel ist, daß sie erlaubt, die (empirische) Varianz mit nur einem Durchlauf durch die Daten zu berechnen, indem man in diesem Durchlauf die Summe der Stichprobenwerte und die Summe ihrer Quadrate bestimmt. Mit der ursprünglichen Formel werden dagegen zwei Durchläufe benötigt: Im ersten Durchlauf wird der Mittelwert berechnet, im zweiten aus der Summe der quadratischen Abweichungen der Daten von Mittelwert die Varianz.

2.2.3.3 Formmaße

Zeichnet man ein Histogramm metrischer Beobachtungsdaten, so zeigt sich sehr oft ein etwa glockenförmiger Verlauf. Dieser weicht aber gewöhnlich mehr oder weniger stark von einer idealen Gaußschen Glockenkurve (Normalverteilung) ab. Die Verteilung ist z.B. unsymmetrisch oder anders gewölbt. Mit Formmaßen versucht man diese Abweichungen zu messen.

Schiefe (skewness) Die Schiefe (skewness) α_3 gibt an, ob, und wenn ja, wie stark eine Verteilung von einer symmetrischen Verteilung abweicht.[2] Sie wird berechnet als

$$\alpha_3 = \frac{1}{n \cdot s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n}\sum_{i=1}^{n} z_i^3 \qquad \text{mit} \qquad z_i = \frac{x_i - \bar{x}}{s}.$$

Für eine symmetrische Verteilung ist $\alpha_3 = 0$. Bei positiver Schiefe ist die Verteilung linkssteil, bei negativer Schiefe rechtssteil (siehe Abbildung 2.6).

Wölbung (Steilheit, Exzeß, Kurtosis) Die Wölbung α_4 gibt an, wie stark eine glöckenförmige Verteilung gewölbt ist[3] (verglichen mit der Gauß-

[2] Der Index 3 deutet an, daß es sich um das 3. Potenzmoment der Stichprobe handelt.
[3] Der Index 4 deutet an, daß es sich um das 4. Potenzmoment der Stichprobe handelt.

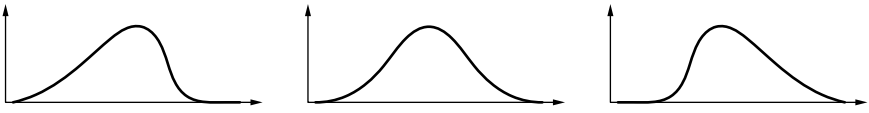


Abbildung 2.6: Illustration des Formmaßes „Schiefe“.

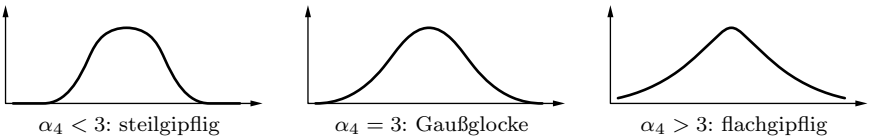


Abbildung 2.7: Illustration des Formmaßes „Wölbung“.

schen Glockenkurve). Sie wird berechnet als

$$\alpha_4 = \frac{1}{n \cdot s^4} \sum_{i=1}^{n} (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^{n} z_i^4 \qquad \text{mit} \qquad z_i = \frac{x_i - \bar{x}}{s}.$$

Eine ideale Gaußsche Glockenkurve hat eine Wölbung von 3. Bei einem kleineren Wert als 3 ist die Verteilung steilgipfliger (leptokurtic), bei einem größeren flachgipfliger (platikurtic) als eine Gaußsche Glockenkurve (siehe Abbildung 2.7). Manchmal wird auch $\alpha_4 - 3$ als Wölbung definiert, so daß die Gaußsche Glockenkurve eine Wölbung von 0 hat und sich so am Vorzeichen ablesen läßt, ob die Verteilung steil- oder flachgipflig ist.

2.2.3.4 Kastendiagramm (box plot)

Einige charakteristische Maße, nämlich der Median, der Mittelwert, die Spannweite und der Interquartilbereich werden gern in einem sogenannten **Kastendiagramm** (box plot) dargestellt (siehe Abbildung 2.8): Die äußeren Linien zeigen die Spannweite und der Kasten in der Mitte, der dem Diagramm seinen Namen gibt, den Interquartilbereich, in den der Median als durchgezogene, der Mittelwert als gestrichelte Linie eingezeichnet werden. Offenbar erhält man durch diese sehr einfache Graphik einen recht guten Eindruck der Verteilung der Daten. Manchmal wird der Kasten, der den Interquartilbereich darstellt, zum Mittelwert hin tailliert gezeichnet, um die Lage des Mittelwertes hervorzuheben.

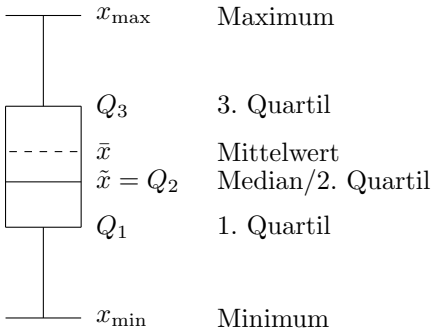


Abbildung 2.8: Ein Kastendiagramm (box plot) zur Darstellung einiger wichtiger statistischer Kenngrößen.

2.2.4 Kenngrößen für mehrdimensionale Daten

Die im vorangehenden Abschnitt betrachteten Lage- und Streuungsmaße lassen sich leicht auf mehrdimensionale Daten übertragen, indem man die Rechnungen mit Vektoren statt mit Skalaren ausführt. Wir betrachten hier beispielhaft die Übertragung des Mittelwertes und der Varianz, was uns zur Kovarianz führt. Über diese kommt man durch Normierung zum wichtigen Maß des Korrelationskoeffizienten.

Mittelwert Der Mittelwert wird für mehrdimensionale Daten zum Vektormittel der Datenpunkte, also etwa für zweidimensionale Daten:

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

Man beachte, daß man das gleiche Ergebnis erhält, indem man den Vektor der Mittelwerte der einzelnen Merkmale bildet.

Kovarianz und Korrelation Auch das Streuungsmaß der Varianz läßt sich leicht auf mehrdimensionale Daten übertragen, indem man die Rechnung mit Vektoren statt mit Skalaren ausführt. Problematisch ist höchstens das Quadrieren der Differenzen zwischen den Stichprobenelementen und dem Mittelwertsvektor, da diese Differenzen nun Vektoren sind. Man verwendet dazu das sogenannte **äußere Produkt** oder **Matrixprodukt** des Differenzvektors mit sich selbst, das definiert ist als $\vec{v}\vec{v}^{\top}$ (wobei $\top$ die Transponierung des Vektors bedeutet) und eine quadratische Matrix ergibt. Diese Matrizen werden summiert und wie bei der Varianz durch den um 1 verringerten Stichprobenumfang geteilt. Das Ergebnis ist eine quadratische,

symmetrische, positiv definite[4] Matrix, die sogenannte **Kovarianzmatrix**. Für zweidimensionale Daten ist die Kovarianzmatrix definiert als

$$\mathbf{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\begin{pmatrix}x_i\\y_i\end{pmatrix}-\begin{pmatrix}\bar{x}\\\bar{y}\end{pmatrix}\right)\left(\begin{pmatrix}x_i\\y_i\end{pmatrix}-\begin{pmatrix}\bar{x}\\\bar{y}\end{pmatrix}\right)^{\top} = \begin{pmatrix}s_x^2 & s_{xy}\\ s_{xy} & s_y^2\end{pmatrix}$$

mit

$$\begin{aligned}
s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right) && \text{(Varianz von } x\text{)}\\
s_y^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n}y_i^2 - n\bar{y}^2\right) && \text{(Varianz von } y\text{)}\\
s_{xy} &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}\right) && \text{(Kovarianz von } x \text{ und } y\text{)}
\end{aligned}$$

Neben den Varianzen der Einzeldimensionen tritt hier eine weitere Größe auf, die sogenannte **Kovarianz**, die Aussagen über die Stärke der (linearen) Abhängigkeit der beiden Merkmale ermöglicht. Da ihr Wert jedoch auch von den Varianzen der beiden Einzeldimensionen abhängt, wird sie durch Division durch die Standardabweichungen der beiden Einzeldimensionen normiert, womit man den sogenannten **Korrelationskoeffizienten** (genauer: Pearsonscher Produktmoment-Korrelationskoeffizient) erhält:

$$r = \frac{s_{xy}}{s_x s_y}.$$

Man beachte, daß der Korrelationskoeffizient der Kovarianz der in beiden Merkmalen auf Mittelwert 0 und Standardabweichung 1 normierten Daten entspricht. Der Korrelationskoeffizient hat einen Wert zwischen -1 und $+1$ und beschreibt die Stärke der *linearen* Abhängigkeit zweier Größen: Liegen alle Datenpunkte genau auf einer steigenden Gerade, so hat er den Wert $+1$, liegen sie genau auf einer fallenden Gerade, so hat er den Wert -1. Die anschauliche Bedeutung von Zwischenwerten zeigt Abbildung 2.9.

Man beachte, daß ein Korrelationskoeffizient von 0 *nicht* bedeutet, daß die betrachteten Größen (stochastisch) unabhängig sind. Liegen die Datenpunkte z.B. symmetrisch auf einer Parabel, so ist der Korrelationskoeffizient $r = 0$. Dennoch besteht eine funktionale Abhängigkeit der beiden Größen; diese Abhängigkeit ist nur nicht linear.

[4] Eine Matrix $\mathbf{M}$ heißt *positiv definit*, wenn für alle Vektoren $\vec{v} \neq \vec{0}$ gilt: $\vec{v}^{\top}\mathbf{M}\vec{v} > 0$.

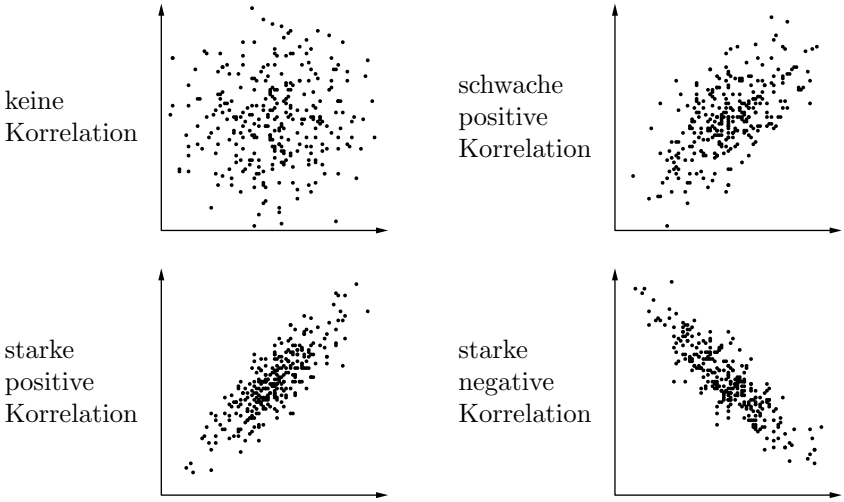


Abbildung 2.9: Illustration der Bedeutung des Korrelationskoeffizienten.

Da die Kovarianz/Korrelation die lineare Abhängigkeit zweier Größen beschreibt, ist es nicht verwunderlich, daß man mit ihrer Hilfe auch eine **Ausgleichsgerade** oder **Regressionsgerade** der Datenpunkte bestimmen kann. Sie ist definiert als

$$(y - \bar{y}) = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \qquad \text{bzw.} \qquad y = \frac{s_{xy}}{s_x^2}(x - \bar{x}) + \bar{y}.$$

Die Regressionsgerade kann als eine Art Mittelwertsfunktion aufgefaßt werden, mit der jedem x-Wert ein Mittelwert der y-Werte zugeordnet wird. Diese Deutung wird auch dadurch gestützt, daß die Regressionsgerade die Summe der Abweichungsquadrate der Datenpunkte (in y-Richtung) minimiert, wie es ja auch der Mittelwert tut. Genaueres zum Verfahren der Regression und der Methode der kleinsten Quadrate, auch Verallgemeinerungen auf größere Funktionenklassen, findet man in Kapitel 3.

2.2.5 Hauptkomponentenanalyse

Korrelationen zwischen den Merkmalen eines Datensatzes lassen sich ausnutzen, um die Dimension des Datensatzes zu reduzieren. Denn wenn ein Merkmal stark mit einem anderen korreliert ist, so ist ja das eine Merkmal

i.w. eine lineare Funktion des anderen. Es reicht dann oft aus, nur eines der beiden Merkmale zu betrachten, da ja das andere zur Not (über die Regressionsgerade) näherungsweise rekonstruiert werden kann. Dieser Ansatz hat jedoch den Nachteil, daß es, besonders bei mehreren korrelierten Merkmalen, nicht ganz leicht ist, die wegzulassenden Merkmale auszuwählen.

Eine bessere Möglichkeit zur Dimensionsreduktion ist die sogenannte **Hauptkomponentenanalyse**. Die Idee dieses Verfahrens ist, nicht einen Teil der Merkmale auszuwählen, sondern eine geringere Anzahl neuer Merkmale als Linearkombinationen der alten zu konstruieren. Diese neuen Merkmale sollen den Großteil der Informationen des Datensatzes enthalten, wobei der Informationsgehalt über die (geeignet normierte) Varianz gemessen wird: Je größer die (normierte) Varianz, um so wichtiger das Merkmal.

Um die neuen Merkmale zu finden, normiert man zunächst die Daten auf Mittelwert 0 und Standardabweichung 1 in allen Merkmalen, damit die Skalierung der Daten (also z.B. die Größeneinheiten, in denen die Merkmale angegeben sind) keinen Einfluß hat. Anschließend sucht man eine neue Basis des Datenraums, also senkrecht zueinander stehende Richtungen, und zwar so, daß die Varianz der (normierten) Daten in der ersten Richtung am größten unter allen möglichen Richtungen im Datenraum ist, die Varianz in der zweiten am größten unter allen Richtungen, die senkrecht zur ersten stehen, die Varianz in der dritten am größten unter allen Richtungen, die senkrecht zur ersten und zur zweiten stehen usf. Schließlich transformiert man die Daten auf die neue Basis des Datenraums und läßt einige der neuen Merkmale weg, nämlich jene, bezüglich der die transformierten Daten die geringste Varianz aufweisen, wobei man anhand der Varianzsumme entscheidet, wie viele Merkmale weggelassen werden.

Formal lassen sich die oben beschriebenen Richtungen finden, indem man eine **Hauptachsentransformation** der **Korrelationsmatrix** (d.i. die Kovarianzmatrix der auf Mittelwert 0 und Standardabweichung 1 normierten Daten) durchführt. D.h., man sucht nach einer Rotation des Koordinatensystems, so daß die Korrelationsmatrix zu einer Diagonalmatrix wird. Die Elemente dieser Diagonalmatrix geben dann die Varianzen bezüglich der neuen Basis des Datenraums an, während alle Kovarianzen verschwinden. Wie aus der linearen Algebra (siehe z.B. [Jänich 1983]) bekannt, besteht eine Hauptachsentransformation in der Berechnung der **Eigenwerte** und **Eigenvektoren** einer Matrix. Die Eigenwerte sind die Diagonalelemente der sich ergebenden Diagonalmatrix, die Eigenvektoren sind die gesuchten Richtungen im Originalraum. Nach der Größe der Eigenwerte werden dann diejenigen Richtungen ausgewählt, bezüglich derer die Daten nun beschrieben werden und die Daten werden auf diese Richtungen projiziert.

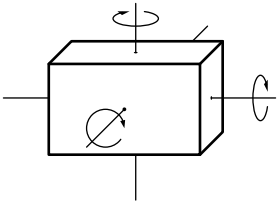


Abbildung 2.10: Die drei Hauptträgheitsachsen eines Quaders.

Die folgende physikalische Analogie macht die Idee vielleicht noch etwas klarer. Wie ein Körper auf eine Rotation um eine Achse reagiert, läßt sich i.w. durch den sogenannten **Trägheitstensor** beschreiben [Greiner 1989]. Der Trägheitstensor ist eine symmetrische 3×3-Matrix,

$$\Theta = \begin{pmatrix} \Theta_{xx} & \Theta_{xy} & \Theta_{xz} \\ \Theta_{xy} & \Theta_{yy} & \Theta_{yz} \\ \Theta_{xz} & \Theta_{yz} & \Theta_{zz} \end{pmatrix},$$

deren Diagonalelemente die **Trägheitsmomente** des Körpers bezüglich der Achsen sind, die durch seinen Schwerpunkt verlaufen[5] und parallel zu den Achsen des zur Beschreibung benutzten Koordinatensystems sind. Die übrigen Matrixelemente heißen **Deviationsmomente** und beschreiben bei der Rotation senkrecht zur Rotationsachse wirkende Kräfte, die sich daraus ergeben, daß i.a. der Drehimpulsvektor nicht parallel zum Winkelgeschwindigkeitsvektor steht. Es gibt jedoch für jeden Körper drei Achsen, bzgl. der die Deviationsmomente verschwinden, die sogenannten **Hauptträgheitsachsen** (siehe Abbildung 2.10, die die Hauptträgheitsachen eines Quaders zeigt).[6] Die Hauptträgheitsachsen stehen immer senkrecht aufeinander. In dem durch die Hauptträgheitsachsen gebildeten Koordinatensystem ist der Trägheitstensor eine Diagonalmatrix.

Formal findet man die Hauptträgheitsachsen eines Körpers, indem man eine Hauptachsentransformation eines Trägheitstensors (bezüglich eines beliebigen Koordinatensystems) durchführt: Die Eigenvektoren des Trägheitstensors geben die Richtung der Hauptträgheitsachsen an.

[5] Das Trägheitsverhalten bezüglich der Rotation um Achsen, die nicht durch den Schwerpunkt des Körpers verlaufen, läßt sich mit Hilfe des **Steinerschen Satzes** beschreiben, auf den wir hier jedoch nicht eingehen wollen. Interessenten seien auf die Standardlehrbücher der Mechanik, z.B. [Greiner 1989], verwiesen.

[6] Ein Körper kann mehr als drei Rotationsachsen besitzen, bezüglich der die Deviationsmomente verschwinden. Bei einer Kugel mit homogener Massenverteilung ist z.B. jede Achse durch den Mittelpunkt der Kugel eine solche Achse. Jeder Körper hat aber mindestens drei solcher Achsen.

Die Deviationsmomente bewirken in der Praxis durch die entstehenden Querkräfte ein Rütteln in den Lagern der Achse. Da ein solches Rütteln natürlich zu einem schnellen Verschleiß der Lager führt, versucht man, die Deviationsmomente zum Verschwinden zu bringen. Ein Automechaniker, der ein Rad auswuchtet, führt also in gewisser Weise eine Hauptachsentransformation durch, denn er bringt die Achse des Wagens mit einer Hauptträgheitsachse des Rades zur Deckung. Allerdings tut er dies nicht, indem er, wie bei der gewöhnlichen Hauptachsentransformation, die Lage der Rotationsachse ändert, denn die Richtung der Achse im Rad ist ja festgelegt. Stattdessen ändert er, durch Anbringen kleiner Gewichte, die Massenverteilung des Rades derart, daß die Deviationsmomente verschwinden.

Mit dieser Analogie kann man sagen: Ein Statistiker sucht im ersten Schritt der Hauptkomponentenanalyse nach den Achsen, um die sich eine Massenverteilung mit Einheitspunktgewichten an den Orten der Datenpunkte ohne „Rütteln in den Lagern“ drehen läßt. Dann wählt er aus den so gefundenen Hauptachsen einige aus, wobei die Achsen weggelassen werden, um die die Rotation „am schwersten geht“, für die also die Trägheitsmomente am höchsten sind (in Richtung dieser Achsen ist die Varianz am geringsten, senkrecht zu ihnen am höchsten).

Die Achsen werden formal über den **Prozentsatz der erklärten Varianz** ausgewählt. Man kann zeigen, daß die Summe der Eigenwerte einer Korrelationsmatrix gleich der Dimension m des Datensatzes, also gleich der Anzahl der Merkmale ist (z.B. [Kowalski 1979]). Dann ist es plausibel, den Anteil, den die j-te Hauptachse an der Gesamtvarianz hat, zu definieren als

$$p_j = \frac{\lambda_j}{m} \cdot 100\%,$$

wobei λ_j der zur j-ten Hauptachse gehörende Eigenwert ist.

Sei $p_{(1)}, \ldots, p_{(m)}$ eine absteigend sortierte Folge dieser Prozentsätze. Man sucht nun für diese sortierte Folge den kleinsten Wert k, für den

$$\sum_{j=1}^{k} p_{(j)} \geq \alpha \cdot 100\%$$

mit einem vom Anwender gewählten Anteil α gilt (z.B. $\alpha = 0.9$), wählt die zugehörigen k Hauptachsen als neue Merkmale und projiziert die Daten auf die zugehörigen Richtungen. Alternativ kann man festlegen, auf wieviele Dimensionen man den Datensatz reduzieren möchte, wählt entsprechend viele Achsen nach absteigendem Prozentsatz und erhält dann mit der obigen Summe eine Aussage über den Informationsverlust.

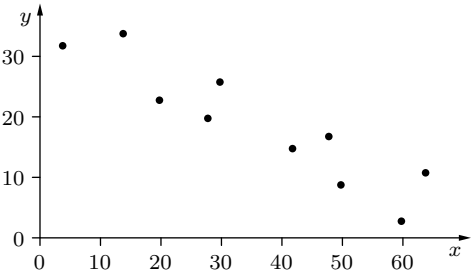


Abbildung 2.11: Daten des Beispiels zur Hauptkomponentenanalyse als Streudiagramm.

Beispiel zur Hauptkomponentenanalyse

Als Beispiel zur Hauptkomponentenanalyse betrachten wir die Reduktion eines Datensatzes zweier korrelierter Größen auf einen eindimensionalen Datensatz. Gegeben sei der folgende zweidimensionale Datensatz:

x	5	15	21	29	31	43	49	51	61	65
y	33	35	24	21	27	16	18	10	4	12

Schon aus dieser Tabelle — und noch besser in Abbildung 2.11 — ist zu sehen, daß die beiden Größen stark negativ korreliert sind, eine Reduktion auf eine Dimension also ohne großen Informationsverlust möglich ist. Als ersten Schritt berechnen wir die Korrelationsmatrix. Dazu normieren also die Daten auf Mittelwert 0 und Standardabweichung 1 und berechnen die Kovarianzmatrix der so normierten Daten. Es ist

$$\begin{aligned}
\bar{x} &= \frac{1}{10}\sum_{i=1}^{10} x_i &&= \frac{370}{10} &&= 37, \\
\bar{y} &= \frac{1}{10}\sum_{i=1}^{10} y_i &&= \frac{200}{10} &&= 20, \\
s_x^2 &= \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) &&= \frac{17290 - 13690}{9} = 400 \quad \Rightarrow \quad s_x = 20, \\
s_y^2 &= \frac{1}{9}\left(\sum_{i=1}^{10} y_i^2 - 10\bar{y}^2\right) &&= \frac{4900 - \ 4000}{9} = 100 \quad \Rightarrow \quad s_y = 10.
\end{aligned}$$

Mit diesem Ergebnis transformieren wir die Daten gemäß

$$x' = \frac{x - \bar{x}}{s_x} \qquad \text{und} \qquad y' = \frac{y - \bar{y}}{s_y}$$

und erhalten als normierten Datensatz:

x'	-1.6	-1.1	-0.8	-0.4	-0.3	0.3	0.6	0.7	1.2	1.4
y'	1.3	1.5	0.4	0.1	0.7	-0.4	-0.2	-1.0	-1.6	-0.8

Die Kovarianz dieser normierten Größen ist der Korrelationskoeffizient

$$r = s_{x'y'} = \frac{1}{9}\sum_{i=1}^{10} x_i y_i = \frac{-8.28}{9} = -\frac{23}{25} = -0.92$$

und die Korrelationsmatrix lautet folglich

$$\mathbf{\Sigma} = \frac{1}{9}\begin{pmatrix} 9 & -8.28 \\ -8.28 & 9 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{23}{25} \\ -\frac{23}{25} & 1 \end{pmatrix}.$$

(Man beachte, daß die Diagonalelemente die Korrelationskoeffizienten der beiden Variablen mit sich selbst und daher stets 1 sind.)

Für diese Korrelationsmatrix müssen wir eine Hauptachsentransformation durchführen. Dazu bestimmen wir die Eigenwerte und Eigenvektoren dieser Matrix, also diejenigen Werte λ_i und Vektoren $\vec{v}_i$, $i = 1, 2$, für die gilt

$$\mathbf{\Sigma}\vec{v}_i = \lambda_i \vec{v}_i \qquad \text{bzw.} \qquad (\mathbf{\Sigma} - \lambda_i \mathbf{E})\vec{v}_i = \vec{0}$$

mit der Einheitsmatrix $\mathbf{E}$. Zur Berechnung der Eigenwerte λ_i wählen wir hier den Weg über das charakteristische Polynom[7]

$$c(\lambda) = |\mathbf{\Sigma} - \lambda\mathbf{E}| = (1-\lambda)^2 - \frac{529}{625},$$

dessen Nullstellen die Eigenwerte

$$\lambda_{1/2} = 1 \pm \sqrt{\frac{529}{625}} = 1 \pm \frac{23}{25}, \qquad \text{also} \qquad \lambda_1 = \frac{48}{25} \quad \text{und} \quad \lambda_2 = \frac{2}{25}$$

sind. Zu diesen Eigenwerten gehören die (normierten) Eigenvektoren

$$\vec{v}_1 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) \qquad \text{und} \qquad \vec{v}_2 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right),$$

die man durch Einsetzen der Eigenwerte in

$$(\mathbf{\Sigma} - \lambda_i \mathbf{E})\vec{v}_i = \vec{0}$$

[7] Für größere Matrizen ist dieser Weg allerdings numerisch instabil und sollte daher durch ein anderes Verfahren ersetzt werden; siehe z.B. [Press *et al.* 1992].

und Lösen des zugehörigen unterbestimmten Gleichungssystems erhält.[8] Die Hauptachsentransformation ist folglich die aus den Eigenvektoren als Spalten zusammengesetzte orthogonale Matrix

$$\mathbf{T} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix},$$

für die

$$\mathbf{T}^\top \boldsymbol{\Sigma} \mathbf{T} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

gilt. Um einfache Zahlen zu erhalten, transformieren wir den (normierten) Datensatz jedoch nicht mit $\mathbf{T}^\top$ sondern mit $\sqrt{2}\mathbf{T}^\top$, also

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \sqrt{2} \cdot \mathbf{T}^\top \cdot \begin{pmatrix} x' \\ y' \end{pmatrix}.$$

Mit Hilfe dieser Transformation werden die Daten auf die Hauptachsen projiziert. Anschaulich werden bei der Multiplikation mit $\mathbf{T}^\top$ von jedem Datenpunkt Lote auf die beiden Hauptachsen gefällt und die Abstände der Fußpunkte vom Koordinatenursprung als neue Koordinaten angesetzt. Man erhält als neuen Datensatz:

x''	−2.9	−2.6	−1.2	−0.5	−1.0	0.7	0.8	1.7	2.8	2.2
y''	−0.3	0.4	−0.4	−0.3	0.4	−0.1	0.4	−0.3	−0.4	0.6

Dieser Datensatz beschreibt die Datenpunkte in dem durch die Hauptachsen gebildeten Koordinatensystem. Da die y''-Werte verglichen mit den x''-Werten nur wenig schwanken (ohne den Faktor $\sqrt{2}$ gäben die Eigenwerte $\lambda_1 = \frac{23}{25}$ und $\lambda_2 = \frac{2}{25}$ die Varianzen in diesen beiden Dimensionen an, mit ihm sind sie doppelt so groß), kann man sich auf die x''-Werte beschränken und hat so die Daten auf eine Dimension reduziert.

Man beachte, daß man diesen Datensatz direkt aus den Ausgangsdaten durch die Transformation

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \sqrt{2} \cdot \mathbf{T}^\top \cdot \begin{pmatrix} s_x^{-1} & 0 \\ 0 & s_y^{-1} \end{pmatrix} \cdot \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right)$$

erhält, die die Normierung der Daten und die Projektion auf die Hauptachsen zusammenfaßt.

[8] Man beachte, daß man bei zwei Variablen wegen der besonderen Form des charakteristischen Polynoms unabhängig von den Ausgangsdaten stets diese beiden Eigenvektoren für die (auf Mittelwert 0 und Standardabweichung 1) normierten Daten erhält.

2.3 Wahrscheinlichkeitsrechnung

Die im nächsten Abschnitt 2.4 besprochene beurteilende Statistik stützt sich stark auf die Wahrscheinlichkeitsrechnung. Sie benutzt wahrscheinlichkeitstheoretische Begriffe und Methoden, um Aussagen über den datenerzeugenden Prozeß zu machen. Dieser Abschnitt rekapituliert daher einige wesentliche Begriffe und Sätze der Wahrscheinlichkeitsrechnung.

2.3.1 Der Wahrscheinlichkeitsbegriff

Die Wahrscheinlichkeitsrechnung beschäftigt sich mit **zufälligen Ereignissen**. Bei diesen ist zwar bekannt, welche Ereignisse überhaupt eintreten können, nicht jedoch, welches der möglichen Ereignisse tatsächlich eintreten wird. Beispiele sind das Werfen einer Münze oder eines Würfels. In der Wahrscheinlichkeitsrechnung wird einem zufälligen Ereignis eine Größe — genannt *Wahrscheinlichkeit* — zugeschrieben. Diese Größe soll die Chance oder die Tendenz des Eintretens des Ereignisses beschreiben.

2.3.1.1 Intuitive Wahrscheinlichkeitsbegriffe

Um eine intuitive Vorstellung zu bekommen, betrachten wir zunächst die klassische Definition der Wahrscheinlichkeit. Diese Definition ist eigentlich eine Berechnungsvorschrift für Wahrscheinlichkeiten und ging aus Überlegungen über die möglichen Ausgänge in Glücksspielen hervor:

Als **Wahrscheinlichkeit** eines Ereignisses A bezeichnet man das Verhältnis der Anzahl der günstigen Ereignisse N_A zur Gesamtzahl der unvereinbaren gleichmöglichen Ereignisse N:

$$P(A) = \frac{N_A}{N}.$$

Daß die Voraussetzung der Gleichmöglichkeit erfüllt ist, wird gewöhnlich durch Symmetrieüberlegungen hergeleitet. Die unvereinbaren gleichmöglichen Ereignisse werden **Elementarereignisse** genannt. Sie bilden den **Ereignisraum**. Andere zufällige Ereignisse lassen sich als Kombination von Elementarereignissen darstellen. Wesentliches Hilfsmittel für die Bestimmung von Wahrscheinlichkeiten auf der Grundlage der obigen Definition ist die **Kombinatorik**, die Methoden zum Auszählen aller und der günstigen Fälle bereitstellt (siehe Abschnitt 2.3.2.1).

Als Beispiel betrachten wir einen symmetrischen und aus homogenem Material gefertigten Würfel. Der Ereignisraum besteht aus den Elementarereignissen „Die geworfene Augenzahl ist x.“, $x \in \{1, 2, 3, 4, 5, 6\}$. Das

Ereignis „Die geworfene Augenzahl ist gerade.“ ist aus den Elementarereignissen „Die geworfene Augenzahl ist x.“, $x \in \{2, 4, 6\}$, zusammengesetzt und tritt daher in drei der sechs gleichmöglichen Fälle ein. Seine Wahrscheinlichkeit ist folglich $\frac{3}{6} = \frac{1}{2}$.

Eine Verallgemeinerung des klassischen Wahrscheinlichkeitsbegriffs wird dadurch erreicht, daß man die Forderung nach Gleichmöglichkeit der Elementarereignisse aufgibt. Stattdessen wird ein Ereignisraum definiert und jedem Elementarereignis eine **Elementarwahrscheinlichkeit** zugeordnet. Der klassische Wahrscheinlichkeitsbegriff ergibt sich dann als der Spezialfall gleicher Elementarwahrscheinlichkeiten.

Eine andere intuitive Vorstellung der Wahrscheinlichkeit von Ereignissen ist die relative Häufigkeit. Ein Experiment mit zufälligem Ausgang werde unter denselben Bedingungen n-mal durchgeführt. Ist A ein zufälliges Ereignis, so tritt bei jeder Versuchsdurchführung A entweder ein oder nicht ein. Die Zahl der Fälle $h_n(A)$, in denen A eintritt, heißt **absolute Häufigkeit**, der Quotient $r_n(A) = \frac{h_n(A)}{n}$ **relative Häufigkeit** von A (vgl. Seite 16).

[von Mises 1928] hat versucht, die **Wahrscheinlichkeit** $P(A)$ eines Ereignisses zu definieren als den Grenzwert, dem sich die relativen Häufigkeiten für große n beliebig nähern, d.h.

$$P(A) = \lim_{n \to \infty} r_n(A).$$

Doch das gelingt nicht. Es kann nicht ausgeschlossen werden, daß Versuchsreihen auftreten, in denen

$$\forall n \geq n_0(\epsilon): \quad P(A) - \epsilon \leq r_n(A) \leq P(A) + \epsilon$$

nicht gilt, wie groß auch immer n_0 gewählt werden mag. (Es ist nicht unmöglich, wenn auch sehr unwahrscheinlich, daß beim Werfen eines Würfels nur Einsen geworfen werden.) Dennoch ist eine Vorstellung der Wahrscheinlichkeit als eine relative Häufigkeit oft hilfreich, insbesondere, wenn man sie als unsere Schätzung der relativen Häufigkeit eines Ereignisses (bei zukünftigen Durchführungen des entsprechenden Versuchs) interpretiert.

Als Beispiel betrachten wir das Geschlecht eines neugeborenen Kindes. Es werden etwa gleich viele Mädchen wie Jungen geboren. Die relative Häufigkeit einer Mädchengeburt ist folglich etwa gleich der relativen Häufigkeit einer Jungengeburt. Wir sagen daher, daß die Wahrscheinlichkeit, daß ein Mädchen geboren wird (genauso wie die Wahrscheinlichkeit, daß ein Junge geboren wird) $\frac{1}{2}$ beträgt. Man beachte, daß sich hier der Wert $\frac{1}{2}$ für die Wahrscheinlichkeit nicht — wie etwa bei einem Münzwurf — aus Symmetrieüberlegungen ableiten läßt.

2.3.1.2 Die formale Definition der Wahrscheinlichkeit

Der klassische Wahrscheinlichkeitsbegriff und die Interpretation als relative Häufigkeit sind stark in unserer Intuition verwurzelt. Die moderne Mathematik hat sich jedoch die axiomatische Methode zu eigen gemacht, bei der von der Bedeutung der Objekte, über die man spricht, abstrahiert wird. Sie nimmt Objekte als gegeben an, die keine anderen Eigenschaften haben als ihre Identität (d.h., sie sind voneinander unterscheidbar), und untersucht lediglich die Struktur der Relationen zwischen diesen Objekten, die sich aus vorausgesetzten Axiomen ergibt.

Auch die Wahrscheinlichkeitstheorie wird daher heute axiomatisch aufgebaut, und zwar über die Kolmogorow-Axiome [Kolmogorow 1933]. Unter einem Ereignis wird in diesen Axiomen einfach eine Menge von Elementareignissen verstanden, die unterscheidbar sind, d.h. eine Identität haben (s.o.). Eine Wahrscheinlichkeit ist dann eine Zahl, die einem Ereignis zugeordnet wird, so daß das System dieser Zahlen bestimmten Bedingungen genügt, die in den Axiomen festgelegt sind. Zunächst definieren wir jedoch die grundlegenden Begriffe „Ereignisalgebra“ und „σ-Algebra“.

Definition 2.1 *Sei $\mathcal{S}$ ein System von Ereignissen (Mengen) über einer Grundmenge Ω von Elementarereignissen (dem* **Ereignisraum***). $\mathcal{S}$ heißt* **Ereignisalgebra** *genau dann, wenn gilt*

1. *$\mathcal{S}$ enthält das* **sichere Ereignis** *Ω und das* **unmögliche Ereignis** *$\emptyset$.*
2. *Gehört A zu $\mathcal{S}$, so gehört auch das Ereignis $\overline{A} = \Omega - A$ zu S.*
3. *Gehören A und B zu $\mathcal{S}$, so gehören auch die Ereignisse $A \cap B$ und $A \cup B$ zu $\mathcal{S}$.*

Es kann außerdem die folgende Bedingung erfüllt sein:

3'. *Gehört für alle $i \in \mathbb{N}$ das Ereignis A_i zu $\mathcal{S}$, dann gehören auch die Ereignisse $\bigcup_{i=1}^{\infty} A_i$ und $\bigcap_{i=1}^{\infty} A_i$ zu $\mathcal{S}$.*

In diesem Fall nennt man $\mathcal{S}$ eine **σ-Algebra**.

Die Wahrscheinlichkeit wird nun über die folgenden Axiome definiert:

Definition 2.2 (Kolmogorow-Axiome)
Gegeben sei eine Ereignisalgebra $\mathcal{S}$ über einem endlichen Ereignisraum Ω.

1. *Die* **Wahrscheinlichkeit** *$P(A)$ eines Ereignisses $A \in \mathcal{S}$ ist eine eindeutig bestimmte, nicht-negative Zahl, die höchstens gleich Eins sein kann, d.h., es gilt $0 \leq P(A) \leq 1$.*

2. *Das sichere Ereignis Ω besitzt die Wahrscheinlichkeit Eins:* $P(\Omega) = 1$.
3. **Additionsaxiom**: *Für zwei unvereinbare (unverträgliche) Ereignisse A und B (also mit $A \cap B = \emptyset$) gilt* $P(A \cup B) = P(A) + P(B)$.

Bei Ereignisräumen Ω, die unendlich viele Elemente enthalten, muß $\mathcal{S}$ eine σ-Algebra sein und Axiom 3 ersetzt werden durch:

3'. **erweitertes Additionsaxiom**: *Sind $A_1, A_2, \ldots$ abzählbar unendlich viele, paarweise unvereinbare Ereignisse, so gilt*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} A_i.$$

Die Wahrscheinlichkeit $P(A)$ läßt sich also als eine auf einer Ereignisalgebra oder σ-Algebra definierte Funktion mit bestimmten Eigenschaften sehen. Aus Definition 2.2 ergeben sich unmittelbar die Folgerungen:

1. Für jedes Ereignis A gilt $P(\overline{A}) = 1 - P(A)$.
2. Das unmögliche Ereignis besitzt die Wahrscheinlichkeit 0: $P(\emptyset) = 0$.
3. Aus $A \subseteq B$ folgt $P(A) \leq P(B)$.
4. Für beliebige Ereignisse A und B gilt $P(A - B) = P(A \cap \overline{B}) = P(A) - P(A \cap B)$.
5. Für beliebige Ereignisse A und B gilt $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. Da $P(A \cap B)$ nicht-negativ ist, folgt unmittelbar $P(A \cup B) \leq P(A) + P(B)$.
7. Durch einfache Induktion erhält man aus dem Additionsaxiom: Sind $A_1, \ldots, A_m$ paarweise unvereinbare Ereignisse, so gilt $P(A_1 \cap \ldots \cap A_m) = \sum_{i=1}^{m} P(A_i)$.

Das Kolmogorowsche Axiomensystem ist widerspruchsfrei, da es Systeme gibt, die allen diesen Axiomen genügen. Die Axiomatik von Kolmogorow gestattet es, die Wahrscheinlichkeitstheorie als Teil der **Maßtheorie** aufzubauen und die Wahrscheinlichkeit als nichtnegative normierte additive Mengenfunktion, d.h. als **Maß** zu interpretieren.

Definition 2.3 *Sei Ω eine Menge von Elementarereignissen, $\mathcal{S}$ eine σ-Algebra über Ω und P eine Wahrscheinlichkeit, die auf $\mathcal{S}$ definiert ist. Dann heißt das Tripel* $(\Omega, \mathcal{S}, P)$ **Wahrscheinlichkeitsraum**.

2.3.1.3 Deutungen der Wahrscheinlichkeit

Die Anwendung der abstrakten axiomatischen Theorie der Wahrscheinlichkeit — wie jeder mathematischen Theorie — erfordert die Deutung der in ihr auftretenden Begriffe, d.h. ihre Abbildung auf die Welt. In diesem Abschnitt wenden wir uns kurz den damit verbundenen Problemen zu.

Eine Wahrscheinlichkeit wird gewöhnlich einem Subjekt A zugesprochen. Logisch erscheint sie also als ein Prädikat:

$$\begin{aligned} P(A) &\;\hat{=}\; A \text{ hat die Wahrscheinlichkeit } P \\ P(A \mid B) &\;\hat{=}\; \text{Wenn } B, \text{ so hat } A \text{ die Wahrscheinlichkeit } P \end{aligned}$$

Wir können wenigstens vier allgemeine Fragen über den Sinn solcher Aussagen stellen [von Weizsäcker 1992]:

1. Was ist die Natur der möglichen logischen Subjekte A und B?
2. Was bedeutet dann das Prädikat P?
3. Welche möglichen Werte haben die Aussagen $P(A)$ und $P(A \mid B)$?
4. Welche Gründe rechtfertigen uns, solche Aussagen zu machen?

Wir klassifizieren die üblichen Antworten auf diese Fragen:

1. Das Subjekt A oder B kann sein
   a) eine Aussage,
   b) ein Ereignis (eine Ereignisklasse),
   c) eine menschliche Verhaltensweise.
2. Das Prädikat P kann sein
   a) eine logische Relation,
   b) eine relative Häufigkeit,
   c) eine Verhaltensregel.
3. Der Wert von $P(A)$ oder $P(A \mid B)$ kann sein
   a) Wahrheit oder Falschheit,
   b) eine Wahrscheinlichkeit (im Sinne einer physikalischen Größe),
   c) ein Nutzen.
4. Der Grund, $P(A)$ oder $P(A \mid B)$ zu behaupten, kann sein
   a) logisch (sprachlich),
   b) physikalisch,
   c) psychologisch.

Die Antworten entsprechen den drei am weitesten verbreiteten Deutungen des Wahrscheinlichkeitsbegriffes [Savage 1954, von Weizsäcker 1992]:

a) Die **logische Deutung** faßt die Wahrscheinlichkeit einer Aussage oder eines Zustandes als Funktion eines möglichen Wissens auf. Gegeben ein Wissen und eine Aussage, so bestimmen die logischen Regeln der Sprache die logische Wahrscheinlichkeit dieser Aussage bezüglich dieses Wissens.

b) Die **empirische (frequentistische) Deutung** faßt Wahrscheinlichkeitsaussagen als Darstellung statistischer Wahrheiten über die Welt auf (etwa im Sinne physikalischer Eigenschaften), die sich in den Häufigkeiten des Auftretens von Ereignissen niederschlagen.

c) Die **subjektive (personalistische) Deutung** faßt die Wahrscheinlichkeit auf als Ausdruck einer Regel, der eine Person in einer gegebenen Situation „in ihrer nützlichen Gewohnheit“ folgt. Ein Zahlenwert kann ihr in dieser Deutung über die Interpretation als Wettbereitschaft zugeordnet werden.

Leider sind mit allen drei Deutungen Probleme verbunden:

a) Die (Zahlen-)Werte der bedingten Wahrscheinlichkeiten sind nicht logisch begründbar.

b) Die A-priori-Wahrscheinlichkeiten (speziell im Bayesschen Verfahren) sind willkürlich.

c) Der Erfolg oder Mißerfolg eines Wettsystems ist ein historisches Faktum, insofern also nicht subjektiv.

Den Erfolg der Anwendung der Wahrscheinlichkeitstheorie kann daher keine der drei Deutungen vollständig erklären. Die Lösung dieses (philosophischen) Problems steht noch aus.

2.3.2 Grundlegende Methoden und Sätze

Nachdem wir die Wahrscheinlichkeit formal definiert haben, wenden wir uns einigen grundlegenden Methoden und Sätzen der Wahrscheinlichkeitsrechnung zu, die wir an einfachen Beispielen erläutern. Hierzu gehören die Berechnung von Wahrscheinlichkeiten auf kombinatorischem und geometrischem Wege, die Begriffe der bedingten Wahrscheinlichkeit und unabhängiger Ereignisse, der Produktsatz, der Satz über die vollständige Wahrscheinlichkeit und schließlich der sehr wichtige Bayessche Satz.

2.3.2.1 Kombinatorische Methoden zur Bestimmung von Wahrscheinlichkeiten

Wie schon erwähnt, ist auf der Grundlage der klassischen Definition der Wahrscheinlichkeit die Kombinatorik ein wichtiges Hilfsmittel zur Bestimmung von Wahrscheinlichkeiten. Wir beschränken uns auf ein einfaches Beispiel, nämlich das berühmte **Geburtstagsproblem**:

m Personen werden zufällig ausgewählt. Wie groß ist die Wahrscheinlichkeit des Ereignisses A_m, daß mindestens zwei dieser Personen am gleichen Tag (Monat und Tag im Jahr, nicht jedoch Jahr) Geburtstag haben?

Vereinfachend vernachlässigen wir Schaltjahre. Außerdem nehmen wir an, daß jeder Tag des Jahres als Geburtstag gleichmöglich ist. Es ist offensichtlich, daß für $m \geq 366$ zwei Personen am gleichen Tag Geburtstag haben müssen. Damit gilt

$$\forall m \geq 366: \quad P(A_m) = 1.$$

Für $m \leq 365$ betrachten wir das komplementäre Ereignis $\overline{A_m}$, das eintritt, wenn alle m Personen an verschiedenen Tagen Geburtstag haben. Dieser Übergang ist eine oft angewandte Technik bei der Bestimmung von Wahrscheinlichkeiten. Numerieren wir die Personen, so kommen für die erste Person 365, für die zweite 364, für die m-te $365 - m + 1$ Tage in Frage. Da es insgesamt 365^m mögliche Fälle gibt, folgt

$$P(\overline{A_m}) = \frac{365 \cdot 364 \cdots (365 - m + 1)}{365^m},$$

also

$$P(A_m) = 1 - \frac{365 \cdot 364 \cdots (365 - m + 1)}{365^m}.$$

Schon für $m = 23$ erhält man überraschenderweise $P(A_{23}) \approx 0.507$.

2.3.2.2 Geometrische Wahrscheinlichkeiten

Geometrische Wahrscheinlichkeiten sind eine Verallgemeinerung der klassischen Definition der Wahrscheinlichkeit. Es wird nicht mehr das Verhältnis der Anzahl der günstigen Fälle zur Gesamtzahl der möglichen Fälle bestimmt, sondern die Anzahlen werden durch geometrische Größen, etwa Längen oder Flächen, ersetzt.

Als Beispiel betrachten wir das von [Buffon 1733] untersuchte Spiel **Franc-Carreau**: In diesem Spiel wird eine Münze auf eine mit gleichen Rechtecken bedeckte Fläche geworfen. Die Münze habe den Radius r, die Rechtecke die Kantenlängen a und b, wobei $2r \leq a$ und $2r \leq b$, so daß die

Münze völlig in das Innere des Rechtecks hineinpaßt. Gesucht ist die Wahrscheinlichkeit mit der die Münze eine oder zwei Rechteckkanten schneidet. Wenn wir in ein Rechteck ein kleineres mit den Kantenlängen $a - 2r$ und $b - 2r$ legen, wobei die Mittelpunkte der Rechtecke zusammenfallen und die Seiten parallel sind, so ist klar, daß die Münze die Seiten des Rechtecks genau dann nicht schneidet, wenn ihr Mittelpunkt in diesem inneren Rechteck liegt. Da die Fläche des inneren Rechtecks $(a - 2r)(b - 2r)$ ist, lautet die gesuchte Wahrscheinlichkeit folglich

$$P(A) = 1 - \frac{(a - 2r)(b - 2r)}{ab}.$$

2.3.2.3 Bedingte Wahrscheinlichkeiten und unabhängige Ereignisse

Oft ist die Wahrscheinlichkeit eines Ereignisses A zu bestimmen, wenn bereits bekannt ist, daß ein Ereignis B eingetreten ist. Solche Wahrscheinlichkeiten nennt man *bedingte* Wahrscheinlichkeiten und bezeichnet sie mit $P(A \mid B)$. Streng genommen, sind die „unbedingten" Wahrscheinlichkeiten, die wir bisher betrachtet haben, auch bedingte Wahrscheinlichkeiten, da wir sie stets unter bestimmten Bedingungen angegeben haben. So haben wir etwa angenommen, daß der Würfel, mit dem wir werfen, symmetrisch und aus homogenem Material gefertigt ist. Nur unter diesen, und gegebenfalls weiteren stillschweigend vorausgesetzten Bedingungen (z.B. keine elektromagnetische Beeinflußung des Würfels etc.) haben wir die Wahrscheinlichkeit jeder Augenzahl mit $\frac{1}{6}$ angegeben.

Definition 2.4 *Seien A und B zwei beliebige Ereignisse mit $P(B) > 0$. Dann heißt*

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

die **bedingte Wahrscheinlichkeit** *von A unter (der Bedingung) B.*

Dazu ein einfaches Beispiel: Es werde mit zwei Würfeln geworfen. Wie groß ist die Wahrscheinlichkeit, daß einer der beiden Würfel eine fünf zeigt, wenn bekannt ist, daß die Summe der Augenzahlen acht ist? Beim Wurf mit zwei Würfeln gibt es offenbar 36 Elementarereignisse, von denen fünf die Augensumme acht haben (4+4, 5+3 und 6+2, wobei die beiden letzten wegen der zwei möglichen Verteilungen der beiden Zahlen auf die beiden Würfel doppelt gezählt werden müssen), d.h., es ist $P(B) = \frac{5}{36}$. Das Ereignis „Augensumme gleich acht und ein Würfel zeigt eine fünf" kann durch genau zwei

Elementarereignisse erreicht werden: Entweder der erste Würfel zeigt eine fünf und der zweite eine drei oder umgekehrt. Folglich ist $P(A \cap B) = \frac{2}{36}$ und damit die gesuchte bedingte Wahrscheinlichkeit $P(A \mid B) = \frac{2}{5}$.

Satz 2.1 (Produktsatz/Multiplikationssatz)
Für beliebige Ereignisse A und B gilt

$$P(A \cap B) = P(A \mid B) \cdot P(B).$$

Dieser Satz folgt unmittelbar aus der Definition der bedingten Wahrscheinlichkeit unter Hinzunahme der offensichtlichen Beziehung $P(A \cap B) = 0$, wenn $P(B) = 0$.[9] Durch einfache Induktion über die Zahl der Ereignisse erhält man die Verallgemeinerung für m Ereignisse:

$$P\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} P\left(A_i \;\middle|\; \bigcap_{k=1}^{i-1} A_k\right).$$

Eine bedingte Wahrscheinlichkeit hat alle Eigenschaften einer Wahrscheinlichkeit, d.h., sie erfüllt die Kolmogorow-Axiome. Damit gilt:

Satz 2.2 *Für ein fest gewähltes Ereignis B mit $P(B) > 0$ stellt die durch $P_B(A) = P(A \mid B)$ definierte Funktion P_B eine Wahrscheinlichkeit dar, die die Bedingung $P_B(\overline{B}) = 0$ erfüllt.*

Mit Hilfe des Begriffs der bedingten Wahrscheinlichkeit wird nun der Begriff der (stochastischen) Unabhängigkeit von Ereignissen definiert. Dieser Begriff läßt sich so motivieren: Wenn z.B. Rauchen keinen Einfluß auf das Entstehen von Lungenkrebs hat, so müßte in der Gruppe der Raucher und der Nichtraucher der Anteil (die relative Häufigkeit) von Personen mit Lungenkrebs etwa gleich groß sein.

Definition 2.5 *Sei B ein Ereignis mit $0 < P(B) < 1$. Dann heißt das Ereignis A* **(stochastisch) unabhängig** *von B, wenn gilt*

$$P(A \mid B) = P(A \mid \overline{B}).$$

Zu dieser Beziehung sind die beiden folgenden, meist leichter handhabbaren Beziehungen äquivalent:

[9] Formal erscheint diese Argumentation jedoch nicht ganz sauber, da für $P(B) = 0$ die bedingte Wahrscheinlichkeit $P(A \mid B)$ nicht definiert ist (siehe Definition 2.4). Da die Beziehung jedoch in diesem speziellen Fall für eine beliebige Festlegung des Wertes von $P(A \mid B)$ gilt, sehen wir über diese Unsauberkeit hinweg.

Satz 2.3 *Das Ereignis A ist von dem Ereignis B mit $0 < P(B) < 1$ genau dann (stochastisch) unabhängig, wenn gilt*

$$P(A \mid B) = P(A)$$

oder äquivalent, wenn gilt

$$P(A \cap B) = P(A) \cdot P(B).$$

Man beachte, daß die Beziehung der (stochastischen) Unabhängigkeit symmetrisch ist, d.h., ist A (stochastisch) unabhängig von B, so ist auch B (stochastisch) unabhängig von A. Weiter läßt sich der Begriff der (stochastischen) Unabhängigkeit leicht auf mehr als zwei Ereignisse erweitern:

Definition 2.6 *m Ereignisse $A_1, \ldots, A_m$ heißen* **vollständig (stochastisch) unabhängig**, *wenn für jede Auswahl $A_{i_1}, \ldots, A_{i_t}$ von t Ereignissen mit $\forall r, s; 1 \leq r, s \leq t : i_r \neq i_s$ gilt*

$$P\left(\bigcap_{k=1}^{t} A_{i_k}\right) = \prod_{k=1}^{t} P(A_{i_k}).$$

Man beachte, daß für die vollständige (stochastische) Unabhängigkeit von mehr als zwei Ereignissen ihre paarweise Unabhängigkeit zwar notwendig, aber nicht hinreichend ist. Dazu betrachten wir ein einfaches Beispiel:

Es werde mit einem weißen und einem roten Würfel geworfen. Seien A das Ereignis „Die Augenzahl des weißen Würfels ist gerade“, B das Ereignis „Die Augenzahl des roten Würfels ist ungerade“ und C das Ereignis „Die Augensumme ist gerade“. Man rechnet leicht nach, das sowohl A und B, als auch B und C, als auch A und C paarweise (stochastisch) unabhängig sind. Wegen $P(A \cap B \cap C) = 0$ (die Summe einer geraden und einer ungeraden Augenzahl muß ungerade sein), sind sie jedoch nicht vollständig (stochastisch) unabhängig.

2.3.2.4 Vollständige Wahrscheinlichkeit und der Bayessche Satz

Oft hat man es mit Situationen zu tun, in denen die Wahrscheinlichkeit von disjunkten, zusammen den ganzen Ereignisraum abdeckenden Ereignissen A_i, sowie die bedingten Wahrscheinlichkeiten eines weiteren Ereignisses B unter den A_i bekannt sind. Gesucht ist die Gesamtwahrscheinlichkeit für das Ereignis B. Als Beispiel betrachte man etwa eine Fabrik, in der es eine bestimmte Anzahl von Maschinen zur Herstellung des gleichen Produktes gibt. Der Ausstoß der Maschinen sowie ihre jeweilige Ausschußrate seien

bekannt, die Gesamtausschußrate soll berechnet werden. Eine solche Berechnung ermöglicht der Satz über die vollständige Wahrscheinlichkeit. Vorher definieren wir aber noch den Begriff der vollständigen Ereignisdisjunktion.

Definition 2.7 *m Ereignisse $A_1, \ldots, A_m$ bilden eine* **vollständige Ereignisdisjunktion**, *wenn alle Paare A_i, A_k, $i \neq k$ unvereinbar sind (d.h., $A_i \cap A_k = \emptyset$ für $i \neq k$) und wenn gilt $A_1 \cup \cdots \cup A_m = \Omega$, sie also zusammen den ganzen Ereignisraum abdecken.*

Satz 2.4 (Satz über die vollständige Wahrscheinlichkeit)
Sei $A_1, \ldots, A_m$ eine vollständige Ereignisdisjunktion mit $\forall i; 1 \leq i \leq m : P(A_i) > 0$ (und, wie aus dem Additionsaxiom folgt, $\sum_{i=1}^{m} P(A_i) = 1$). Dann gilt für die Wahrscheinlichkeit eines beliebigen Ereignisses B

$$P(B) = \sum_{i=1}^{m} P(B \mid A_i) P(A_i).$$

Den Satz über die vollständige Wahrscheinlichkeit erhält man, indem man den Produktsatz (siehe Satz 2.1) auf die Beziehung

$$P(B) = P(B \cap \Omega) = P\left(B \cap \bigcup_{i=1}^{m} A_i\right) = P\left(\bigcup_{i=1}^{m} (B \cap A_i)\right) = \sum_{i=1}^{m} P(B \cap A_i)$$

anwendet, deren letzter Schritt aus dem Additionsaxiom folgt.

Mit Hilfe des Satzes über die vollständige Wahrscheinlichkeit läßt sich leicht der wichtige Bayessche Satz ableiten. Man muß sich nur klar machen, daß sich der Produktsatz für Wahrscheinlichkeiten auf das gleichzeitige Eintreten zweier Ereignisse A und B in zweierlei Weise anwenden läßt:

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A).$$

Dies führt nach Division durch die Wahrscheinlichkeit $P(B)$, die dazu als positiv vorausgesetzt werden muß, zum ersten Teil des Bayesschen Satzes. Die Anwendung des Satzes über die vollständige Wahrscheinlichkeit auf den Nenner liefert den zweiten Teil.

Satz 2.5 (Bayesscher Satz/Bayessche Formel)
Für eine vollständige Ereignisdisjunktion $A_1, \ldots, A_m$ mit $\forall i; 1 \leq i \leq m : P(A_i) > 0$ und jedes Ereignis B mit $P(B) > 0$ gilt

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{P(B)} = \frac{P(B \mid A_i) P(A_i)}{\sum_{k=1}^{m} P(B \mid A_k) P(A_k)}.$$

Diese Formel[10] heißt auch die Formel über die Wahrscheinlichkeit von Hypothesen, da man mit ihr die Wahrscheinlichkeit von Hypothesen, z.B. über das Vorliegen verschiedener Krankheiten bei einem Patienten, berechnen kann, wenn man weiß, mit welcher Wahrscheinlichkeit die verschiedenen Hypothesen zu den Ereignissen A_i (z.B. Krankheitssymptomen) führen.

Als Beispiel betrachten wir fünf Urnen folgenden Inhalts:
zwei Urnen vom Inhalt A_1 mit je zwei weißen und drei schwarzen Kugeln,
zwei Urnen vom Inhalt A_2 mit je einer weißen und vier schwarzen Kugeln,
eine Urne mit dem Inhalt A_3 mit vier weißen und einer schwarzen Kugel.

Aus einer willkürlich gewählten Urne werde eine Kugel entnommen. Sie sei weiß. (Dies sei das Ereignis B.) Wie groß ist die (A-posteriori-) Wahrscheinlichkeit dafür, daß die Kugel aus der Urne mit Inhalt A_3 stammt?

Nach Voraussetzung ist:

$$\begin{array}{llllll} P(A_1) & = \frac{2}{5}, & P(A_2) & = \frac{2}{5}, & P(A_3) & = \frac{1}{5}, \\ P(B \mid A_1) & = \frac{2}{5}, & P(B \mid A_2) & = \frac{1}{5}, & P(B \mid A_3) & = \frac{4}{5}. \end{array}$$

Zunächst wenden wir den Satz von der vollständigen Wahrscheinlichkeit an, um die Wahrscheinlichkeit $P(B)$ zu bestimmen:

$$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_2)P(A_3) \\ &= \frac{2}{5} \cdot \frac{2}{5} + \frac{1}{5} \cdot \frac{2}{5} + \frac{4}{5} \cdot \frac{1}{5} \quad = \quad \frac{10}{25}. \end{aligned}$$

Mit dem Bayesschen Satz erhalten wir anschließend

$$P(A_3 \mid B) = \frac{P(B \mid A_3)P(A_3)}{P(B)} = \frac{\frac{4}{5} \cdot \frac{1}{5}}{\frac{10}{25}} = \frac{2}{5}.$$

Genauso finden wir

$$P(A_1 \mid B) = \frac{2}{5} \quad \text{und} \quad P(A_2 \mid B) = \frac{1}{5}.$$

2.3.2.5 Das Bernoullische Gesetz der großen Zahlen

Schon in Abschnitt 2.3.1 haben wir die Beziehung zwischen der Wahrscheinlichkeit $P(A)$ und der relativen Häufigkeit $r_n(A) = \frac{h_n(A)}{n}$ eines Ereignisses A

[10] Man beachte, daß Thomas Bayes (1702–1761) diese Formel, obwohl sie seinen Namen trägt, nicht hergeleitet hat. Sie wurde erst später von Pierre-Simon Laplace (1749–1827) in dieser Form angegeben.

betrachtet, wobei $h_n(A)$ die absolute Häufigkeit dieses Ereignisses in n Versuchen ist. Wir haben dort gesehen, daß die Definition der Wahrscheinlichkeit als Grenzwert der relativen Häufigkeit nicht gelingt. Es gilt aber eine etwas schwächere Aussage, das berühmte *Gesetz der großen Zahlen.*

Satz 2.6 (Bernoullisches Gesetz der großen Zahlen)
Sei $h_n(A)$ die Anzahl des Eintretens eines Ereignisses A in n unabhängigen Versuchen, wobei in jedem dieser Versuche die Wahrscheinlichkeit des Eintretens dieses Ereignisses gleich $p = P(A)$, $0 \leq p \leq 1$, ist. Dann gilt für jedes $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{h_n(A)}{n} - p\right| < \epsilon\right) = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{z^2}{2}} \mathrm{d}z = 1.$$

Diese Eigenschaft der relativen Häufigkeit $r_n(A) = \frac{h_n(A)}{n}$ kann folgendermaßen interpretiert werden: Zwar ist nicht der Wert $p = P(A)$ der Wahrscheinlichkeit der Grenzwert der relativen Häufigkeit, wie [von Mises 1928] definieren wollte, aber es kann als sehr wahrscheinlich (praktisch sicher) angesehen werden, daß in einem Bernoulli-Experiment von großem Umfang n die relative Häufigkeit $r_n(A)$ von einer festen Zahl, der Wahrscheinlichkeit p, nur wenig abweicht. Damit haben wir die Beziehung zwischen der relativen Häufigkeit und der Wahrscheinlichkeit eines Ereignisses A gefunden.

2.3.3 Zufallsvariable

In vielen Fällen interessiert man sich nicht für die Elementarereignisse eines Ereignisraums und ihre Wahrscheinlichkeiten, sondern für die Wahrscheinlichkeiten von Ereignissen, die sich durch eine Partitionierung des Ereignisraums (Einteilung in disjunkte, den gesamten Raum abdeckende Teilmengen) ergeben. Die Wahrscheinlichkeiten dieser Ereignisse werden durch sogenannte *Zufallsvariablen* beschrieben, die man als Transformationen von einem Ereignisraum in einen anderen sehen kann [Larsen and Marx 1986].

Definition 2.8 *Eine auf dem Ereignisraum Ω definierte Funktion X mit Wertebereich* $\mathrm{dom}(X)$ *heißt* **Zufallsvariable**, *wenn das Urbild jeder Teilmenge ihres Wertebereichs eine Wahrscheinlichkeit besitzt. Das Urbild einer Teilmenge $U \subseteq \mathrm{dom}(X)$ ist dabei definiert als*

$$X^{-1}(U) = \{\omega \in \Omega \mid X(\omega) \in U\}.$$

Die Bezeichnung „Zufallsvariable“ ist etwas irreführend, eine bessere Bezeichnung wäre „Zufallsfunktion“. Wir halten uns hier jedoch an die konventionelle Bezeichnung.

2.3.3.1 Reellwertige Zufallsvariable

Die einfachste Zufallsvariable ist offenbar die, deren mögliche Werte die Elementarereignisse selbst sind. Im Prinzip kann der Wertebereich einer Zufallsvariablen eine beliebige Menge sein, meistens beschränkt man sich aber auf reellwertige Zufallsvariable.

Definition 2.9 *Eine auf einem Ereignisraum Ω definierte reellwertige Funktion X heißt* **(reellwertige) Zufallsvariable**, *wenn sie folgende Eigenschaften besitzt: Für jedes $x \in \mathbb{R}$ und jedes Intervall $(a, b]$, $a < b$ (wobei $a = -\infty$ zugelassen ist), besitzen die Ereignisse $A_x = \{\omega \in \Omega \mid X(\omega) = x\}$ und $A_{(a,b]} = \{\omega \in \Omega \mid a < X(\omega) \leq b\}$ Wahrscheinlichkeiten.*

Manchmal werden die geforderten Eigenschaften auch mit rechts offenem Intervall (a, b) angegeben. Dies führt zu keinen wesentlichen Unterschieden.

Definition 2.10 *Sei X eine reellwertige Zufallsvariable. Dann heißt die reellwertige Funktion*

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

Verteilungsfunktion *von X.*

2.3.3.2 Diskrete Zufallsvariable

Definition 2.11 *Eine Zufallsvariable X, deren Wertevorrat* $\mathrm{dom}(X)$ *nur endlich oder abzählbar unendlich ist, heißt* **diskret**. *Die Gesamtheit aller Zahlenpaare $(x_i, P(X = x_i))$, $x_i \in \mathrm{dom}(X)$, heißt* **(Wahrscheinlichkeits-) Verteilung** *der diskreten Zufallsvariablen X.*

Wenn sich die Wahrscheinlichkeiten $P(X = x)$ durch einen Funktionsausdruck angeben lassen, wird die Verteilung einer diskreten Zufallsvariablen oft auch als Funktion $v_X(x) = P(X = x)$ angegeben. Eventuelle Parameter werden in diesem Fall durch ein Semikolon getrennt nach dem Funktionsargument aufgeführt, z.B. bei der Binomialverteilung

$$b_X(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}$$

die Wahrscheinlichkeit p des Eintretens des betrachteten Ereignisses in einem Einzelversuch und der Umfang n der Versuchsreihe. (Die Binomialverteilung wird genauer in Abschnitt 2.3.5.1 besprochen, die sich daran anschließenden Abschnitte behandeln weitere wichtige Verteilungen.)

Die Funktionswerte der Verteilungsfunktion einer diskreten reellwertigen Zufallsvariablen lassen sich nach folgender Formel aus den Werten der (Wahrscheinlichkeits-)Verteilung berechnen

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i).$$

Jede diskrete reellwertige Zufallsvariable X besitzt als Verteilungsfunktion eine Treppenfunktion F, die nur an den Stellen x_i aus dem Wertebereich $\mathrm{dom}(X)$ Sprünge der Höhe $P(X = x_i)$ besitzt. Aus $x < y$ folgt $F(x) \leq F(y)$, d.h., F ist monoton nichtfallend. Die Funktionswerte $F(x)$ werden beliebig klein, wenn nur x klein genug gewählt wird, während sich die Funktionswerte $F(x)$ mit wachsendem x der Zahl 1 beliebig nähern. Es gilt also

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{und} \quad \lim_{x \to \infty} F(x) = 1.$$

Umgekehrt kann aus jeder Treppenfunktion F, welche die obigen Bedingungen erfüllt, die Verteilung $(x_i, P(X = x_i))$, $i \in \mathbb{N}$ einer diskreten reellwertigen Zufallsvariablen gewonnen werden.

Mit Hilfe der Verteilungsfunktion F läßt sich sehr einfach die Wahrscheinlichkeit dafür berechnen, daß X Werte aus einem Intervall annimmt.

Satz 2.7 *Ist F die Verteilungsfunktion einer diskreten reellwertigen Zufallsvariable X, so gelten folgende Gleichungen*

$$\begin{aligned} P(a < X \leq b) &= F(b) - F(a), \\ P(a \leq X \leq b) &= F(b) - F_L(a), \\ P(a < X) &= 1 - F(a), \end{aligned}$$

wobei $F_L(a)$ der linksseitige Grenzwert von $F(x)$ an der Stelle a ist. Dieser Grenzwert ist gleich $F(a)$, wenn a keine Sprungstelle ist, und sonst gleich dem Wert der Treppenstufe, die unmittelbar links von a liegt, oder formal

$$F_L(a) = \sup_{x < a} F(x).$$

2.3.3.3 Stetige Zufallsvariable

In Analogie zu diskreten Zufallsvariablen definiert man stetige (reellwertige) Zufallsvariable als solche, deren Wertebereich überabzählbar unendlich ist. Bei der Verteilungsfunktion muß man dann offenbar von einer Summe zu einem Integral übergehen.

Definition 2.12 *Eine reellwertige Zufallsvariable X heißt* **stetig**, *wenn eine nichtnegative, integrierbare Funktion f existiert, so daß für ihre Verteilungsfunktion $F(x) = P(X \leq x)$ die Integraldarstellung*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\mathrm{d}u$$

gilt. Die Funktion f heißt **(Wahrscheinlichkeits-)Dichtefunktion** — *oder kurz* **Dichte** — *der Zufallsvariablen X.*

Wegen $P(-\infty < X < \infty) = P(\{\omega \in \Omega \mid -\infty < X(\omega) < \infty\}) = P(\Omega)$ erfüllt eine Dichtefunktion f die Bedingung

$$\int_{-\infty}^{\infty} f(u)\mathrm{d}u = 1$$

Für stetige Zufallsvariable gelten ähnliche Beziehungen wie für diskrete reellwertige Zufallsvariable.

Satz 2.8 *Ist X eine stetige Zufallsvariable mit der Dichtefunktion f, so gilt für beliebige Zahlen $a, b, c \in \mathbb{R}$ mit $a < b$*

$$\begin{array}{lcll} P(a < X \leq b) & = & F(b) - F(a) & = & \int_a^b f(u)\mathrm{d}u \\ P(X > c) & = & 1 - F(c) & = & \int_c^\infty f(u)\mathrm{d}u \end{array}$$

2.3.3.4 Zufallsvektoren

Bisher haben wir nur einzelne Zufallsvariablen untersucht. Im folgenden erweitern wir unsere Betrachtungen auf mehrere Zufallsvariable und ihr Zusammenspiel, d.h. ihre gemeinsame Verteilung und ihre Abhängigkeit bzw. Unabhängigkeit. Dazu definieren wir den Begriff des Zufallsvektors bzw. der mehrdimensionalen Zufallsvariable.

Definition 2.13 *Auf dem gleichen Wahrscheinlichkeitsraum $(\Omega, \mathcal{S}, P)$, d.h. auf dem gleichen Ereignisraum Ω mit der gleichen σ-Algebra $\mathcal{S}$ und Wahrscheinlichkeit P, seien m Zufallsvariablen $X_1, \ldots, X_m$ definiert. In diesem Fall heißt der Vektor $(X_1, \ldots, X_m)$ ein* **Zufallsvektor** *oder eine* **m-dimensionale Zufallsvariable**.

Der Einfachheit halber betrachten wir im folgenden stets zweidimensionale Zufallsvariablen. Die Definitionen und Sätze lassen sich jedoch leicht auf mehrdimensionale Zufallsvariablen (Zufallsvektoren mit endlicher Stelligkeit m) verallgemeinern. Desweiteren beschränken wir uns, wie auch schon in den vorangehenden Abschnitten, auf reellwertige Zufallsvariablen.

Definition 2.14 *Es seien X und Y zwei reellwertige Zufallsvariablen. Die für alle Wertepaare $(x, y) \in \mathbb{R}^2$ durch*

$$F(x, y) = P(X \leq x, Y \leq y)$$

definierte Funktion F heißt **Verteilungsfunktion** *der zweidimensionalen Zufallsvariablen (X, Y). Die eindimensionalen Verteilungsfunktionen*

$$F_1(x) = P(X \leq x) \qquad \text{und} \qquad F_2(y) = P(Y \leq y)$$

heißen **Randverteilungsfunktionen**.

Analog läßt sich für diskrete Zufallsvariablen der Begriff der gemeinsamen Verteilung definieren.

Definition 2.15 *Es seien X und Y zwei diskrete Zufallsvariablen. Dann heißt die Gesamtheit $\forall i, j \in \mathbb{N} : ((x_i, y_i), P(X = x_i, Y = y_i))$ die* **gemeinsame Verteilung** *der Zufallsvariablen X und Y. Die eindimensionalen Verteilungen $\forall i \in \mathbb{N} : (x_i, \sum_j P(X = x_i, Y = y_j)$ und $\forall j \in \mathbb{N} : (y_j, \sum_i P(X = x_i, Y = y_j)$ heißen* **Randverteilungen**.

Stetige Zufallsvariablen werden ähnlich behandelt, nur wird die gemeinsame Verteilung durch die gemeinsame Dichte ersetzt.

Definition 2.16 *Die zweidimensionale Zufallsvariable (X, Y) heißt stetig, wenn eine nichtnegative Funktion $f(x, y)$ existiert, so daß für jedes $(x, y) \in \mathbb{R}^2$ gilt*

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) \mathrm{d}u \mathrm{d}v$$

Die Funktion $f(x, y)$ heißt **gemeinsame Dichte** *der Zufallsvariablen X und Y. Die eindimensionalen Dichtefunktionen*

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y) \mathrm{d}y \qquad \text{und} \qquad f_2(y) = \int_{-\infty}^{+\infty} f(x, y) \mathrm{d}x$$

heißen **Randdichten**.

In Erweiterung des Begriffes der Unabhängigkeit von Ereignissen kann man die Unabhängigkeit von Zufallsvariablen definieren.

Definition 2.17 *Zwei diskrete Zufallsvariablen X und Y mit der zweidimensionalen Verteilungsfunktion $F(x, y)$ und den Randverteilungsfunktionen $F_1(x)$ und $F_2(y)$ heißen* **(stochastisch) unabhängig**, *falls für alle Wertepaare $(x, y) \in \mathbb{R}^2$ gilt*

$$F(x, y) = P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) = F_1(x) \cdot F_2(y).$$

2.3.4 Kenngrößen von Zufallsvariablen

Wie in Abschnitt 2.2.3 für Datensätze durchgeführt, lassen sich auch Zufallsvariablen durch Kenngrößen charakterisieren. Es bestehen i.w. die gleichen Möglichkeiten, wobei man sich statt auf die Datenpunkte auf die Wahrscheinlichkeitsmasse bezieht, so daß man analoge Begriffe erhält.

2.3.4.1 Erwartungswert

Wendet man den Begriff der Zufallsvariable auf Glückspiele an, wobei z.B. der Gewinn (die Auszahlung) der Wert der Zufallsvariablen sein kann, so erscheint es sinnvoll, einen durchschnittlichen oder zu erwartenden Gewinn (bei einer genügend großen Anzahl von Spielen) zu betrachten. Dieses Vorgehen führt zum Begriff des Erwartungswertes.

Definition 2.18 *Ist X eine diskrete reellwertige Zufallsvariable mit der Verteilung $(x_i, P(X = x_i))$, $i \in \mathbb{N}$, und ist $\sum_i |x_i| P(X = x_i)$ endlich, so heißt (der dann auch existierende Grenzwert)*

$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \, P(X = x_i)$$

der **Erwartungswert** *der diskreten Zufallsvariablen X.*

Als Beispiel betrachten wir den Erwartungswert des Gewinns im Roulettespiel. Wir entscheiden uns *nicht* für eine sogenannte „einfache Chance“ (Rouge — Noir, Pair — Impair, Manque — Passe), um den Schwierigkeiten der mit diesen Chancen verbundenen Sonderregeln zu entgehen, sondern für das Setzen auf eine Kolonne (eine der Zahlenreihen 1–12, 13–24 und 14–36). Im Falle eines Gewinns gelangt der dreifache Einsatz zur Auszahlung. Da im Roulettespiel 37 Zahlen erzielt werden können (0–36), die, ideale Bedingungen angenommen, alle gleichwahrscheinlich sind, gewinnt man auf einer Kolonne mit der Wahrscheinlichkeit $\frac{12}{37}$ und verliert mit der Wahrscheinlichkeit $\frac{25}{37}$. Nehmen wir an, der Einsatz betrage m Jetons. Im Falle eines Gewinns erhält man den zweifachen Einsatz als Reingewinn (von dem dreifachen ausgezahlten Einsatz hat man ja den einfachen Einsatz, den man geleistet hat, abzuziehen), also $2m$ Jetons, im Falle eines Verlustes verliert man m Jetons. Als Erwartungswert ergibt sich folglich

$$E(X) = 2m \cdot \frac{12}{37} - m \cdot \frac{25}{37} = -\frac{1}{37} m \approx -0.027m.$$

Im Durchschnitt verliert man also pro Spiel 2.7% seines Einsatzes.

Um den Erwartungswert für stetige Zufallsvariablen zu definieren, braucht man nur von der Summe zum Integral überzugehen.

Definition 2.19 *Ist X eine stetige Zufallsvariable mit der Dichtefunktion f und existiert das Integral*

$$\int_{-\infty}^{\infty} |x| f(x) \mathrm{d}x = \lim_{\substack{a \to -\infty \\ b \to \infty}} \int_a^b |x| f(x) \mathrm{d}x,$$

so heißt das (dann auch existierende) Integral

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

der **Erwartungswert** *der stetigen Zufallsvariablen X.*

2.3.4.2 Eigenschaften des Erwartungswertes

In diesem Abschnitt sind einige Eigenschaften des Erwartungswertes einer Zufallsvariable zusammengestellt, die mitunter bei seiner Berechnung von Vorteil sein können.

Satz 2.9 *Sei X eine diskrete Zufallsvariable, die nur den Wert c annimmt. Dann ist ihr Erwartungswert gleich c: $\mu = E(X) = c$.*

Satz 2.10 (Linearität des Erwartungswertes)
Sei X eine (diskrete oder stetige) reellwertige Zufallsvariable mit dem Erwartungswert $E(X)$. Dann gilt für den Erwartungswert der Zufallsvariablen $Y = aX + b$, $a, b \in \mathbb{R}$

$$E(Y) = E(aX + b) = aE(X) + b.$$

Die Gültigkeit dieses Satzes läßt sich leicht durch das Einsetzen der Definition des Erwartungswertes, einmal für diskrete reellwertige und einmal für stetige Zufallsvariablen zeigen.

Als Beispiel betrachten wir die Verteilung der Zufallsvariablen X, die die Augensumme zweier Würfel beschreibt. Sie ist offenbar symmetrisch zum Wert 7, d.h. $\forall k \in \{0, 1, \ldots, 5\} : P(X = 7 + k) = P(X = 7 - k)$. Daraus folgt aber, daß die Erwartungswerte der Zufallsvariablen $Y_1 = X - 7$ und $Y_2 = -(X-7) = 7-X$ identisch sein müssen, da sie wegen dieser Symmetrie die gleiche Verteilung haben. Es gilt also

$$E(Y_1) = E(X - 7) = E(7 - X) = E(Y_2).$$

Nach dem obigen Satz erhält man so $E(X) - 7 = 7 - E(X)$ und folglich $E(X) = 7$. Allgemein können wir aus diesem Beispiel schließen, daß der Symmetriepunkt einer Verteilung, wenn sie einen besitzt, ihr Erwartungswert ist. Dies gilt auch für stetige Zufallsvariable.

Wir wenden uns nun dem Erwartungswert von Funktionen zweier Zufallsvariablen zu, und zwar ihrer Summe und ihrem Produkt.

Satz 2.11 (Erwartungswert einer Summe von Zufallsvariablen) *Der Erwartungswert einer Summe $Z = X + Y$ zweier beliebiger Zufallsvariablen X und Y, deren Erwartungswerte $E(X)$ und $E(Y)$ existieren, ist gleich der Summe ihrer Erwartungswerte,*

$$E(Z) = E(X + Y) = E(X) + E(Y).$$

Satz 2.12 (Erwartungswert eines Produktes von Zufallsvariablen) *Der Erwartungswert eines Produktes $Z = X \cdot Y$ zweier unabhängiger Zufallsvariablen X und Y, deren Erwartungswerte $E(X)$ und $E(Y)$ existieren, ist gleich dem Produkt ihrer Erwartungswerte,*

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

Wieder läßt sich die Gültigkeit der Sätze leicht durch das Einsetzen der Definition des Erwartungswertes, im Falle des Satzes über das Produkt von Zufallsvariablen unter Hinzunahme der Definition der Unabhängigkeit von Zufallsvariablen, ableiten. Beide Sätze lassen sich offenbar durch Induktion leicht auf Summen und Produkte endlich vieler (unabhängiger) Zufallsvariablen erweitern. Man beachte die Voraussetzung der Unabhängigkeit der Zufallsvariablen im zweiten Satz.

2.3.4.3 Varianz und Standardabweichung

Allein der Erwartungswert charakterisiert eine Zufallsvariable aber noch nicht hinreichend. Von Bedeutung ist ebenfalls, mit welcher Abweichung vom Erwartungswert man im Mittel rechnen muß (vgl. Abschnitt 2.2.3.2). Diese Abweichung beschreiben Varianz und Standardabweichung.

Definition 2.20 *Ist μ der Erwartungswert einer diskreten reellwertigen Zufallsvariable X, so heißt im Falle der Existenz der Zahlenwert*

$$\sigma^2 = D^2(X) = E([X - \mu]^2) = \sum_{i=1}^{\infty} (x_i - \mu)^2 \, P(X = x_i)$$

die **Varianz** *und die positive Quadratwurzel $\sigma = D(X) = +\sqrt{\sigma^2}$ die* **Standardabweichung** *oder* **Streuung** *der Zufallsvariablen X*

Betrachten wir auch hier als Beispiel wieder das Roulette. Beim Setzen von m Jetons auf eine Kolonne beträgt die Varianz

$$D^2(X) = \left(2m + \frac{1}{37}m\right)^2 \cdot \frac{12}{37} + \left(-m + \frac{1}{37}m\right)^2 \cdot \frac{25}{37} = \frac{99900}{50653}m^2 \approx 1.97m^2,$$

die Standardabweichung $D(X)$ also etwa $1.40m$. Im Vergleich dazu beträgt die Varianz beim Setzen auf „Plain“, d.h. auf eine einzelne Zahl, bei gleichem Erwartungswert

$$D^2(X) = \left(35m + \frac{1}{37}m\right)^2 \cdot \frac{1}{37} + \left(-m + \frac{1}{37}m\right)^2 \cdot \frac{36}{37} = \frac{1726272}{50653}m^2 \approx 34.1m^2,$$

die Standardabweichung $D(X)$ also etwa $5.84m$. Bei gleichem Erwartungswert ist die durchschnittliche Abweichung vom Erwartungswert beim Setzen auf „Plain“ folglich mehr als 4-mal so groß.

Um die Varianz für stetige Zufallsvariablen zu definieren, braucht man — wie beim Erwartungswert — nur von der Summe zum Integral überzugehen.

Definition 2.21 *Ist μ der Erwartungswert einer stetigen Zufallsvariable X, so heißt im Falle der Existenz der Zahlenwert*

$$\sigma^2 = D^2(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \mathrm{d}x$$

die **Varianz** *von X und $\sigma = D(X) = +\sqrt{\sigma^2}$ die* **Standardabweichung** *oder* **Streung** *der Zufallsvariablen X.*

2.3.4.4 Eigenschaften der Varianz

In diesem Abschnitt sind einige Eigenschaften der Varianz einer Zufallsvariable zusammengestellt.

Satz 2.13 *Sei X eine diskrete Zufallsvariable, die nur den Wert c annimmt. Dann ist ihre Varianz gleich 0: $\sigma^2 = D^2(X) = 0$.*

Satz 2.14 *Sei X eine (diskrete oder stetige) reellwertige Zufallsvariable mit der Varianz $D^2(X)$. Dann gilt für die Varianz der Zufallsvariablen $Y = aX + b$, $a, b \in \mathbb{R}$*

$$D^2(Y) = D^2(aX + b) = a^2 D^2(X)$$

und folglich für die Streuung

$$D(Y) = D(aX + b) = |a| D(X).$$

Die Gültigkeit dieses Satzes (wie auch des folgenden) läßt sich leicht durch das Einsetzen der Definition der Varianz, einmal für diskrete reellwertige und einmal für stetige Zufallsvariablen zeigen.

Satz 2.15 *Für die Varianz σ^2 einer (diskreten oder stetigen) reellwertigen Zufallsvariable gilt die Beziehung*

$$\sigma^2 = E(X^2) - \mu^2.$$

Satz 2.16 (Varianz einer Summe von Zufallsvariablen, Kovarianz) *Sind X und Y zwei (diskrete oder stetige) reellwertige Zufallsvariablen, deren Varianzen $D^2(X)$ und $D^2(Y)$ existieren, so gilt*

$$D^2(X+Y) = D^2(X) + D^2(Y) + 2[E(X \cdot Y) - E(X) \cdot E(Y)].$$

Den Ausdruck $E(X \cdot Y) - E(X) \cdot E(Y) = E[(X - E(X))(Y - E(Y))]$ nennt man die **Kovarianz** *der Zufallsvariablen X und Y. Aus der (stochastischen) Unabhängigkeit von X und Y folgt*

$$D^2(Z) = D^2(X+Y) = D^2(X) + D^2(Y),$$

d.h., die Kovarianz unabhängiger Zufallsvariablen verschwindet.

Wieder läßt sich die Gültigkeit des Satzes leicht durch das Einsetzen der Definition der Varianz zeigen. Durch Induktion läßt er sich problemlos auf endlich viele Zufallsvariablen erweitern.

2.3.4.5 Quantile

Quantile werden in direkter Analogie zu den Quantilen eines Datensatzes definiert, wobei — wie auch naheliegend — der Anteil des Datensatzes durch die Wahrscheinlichkeitsmasse ersetzt wird.

Definition 2.22 *Sei X eine reellwertige Zufallsvariable. Dann heißt jeder Wert x_α, $0 < \alpha < 1$, für den*

$$P(X \leq x_\alpha) \geq \alpha \qquad \textit{und} \qquad P(X \geq x_\alpha) \geq 1 - \alpha$$

gelten, ein **α-Quantil** *der Zufallsvariable X (bzw. ihrer Verteilung).*

Man beachte, daß u.U., etwa für diskrete Zufallsvariable, mehrere Werte beide Ungleichungen erfüllen können. Das Ungleichungspaar ist übrigens der Doppelungleichung

$$\alpha - P(X = x) \leq F_X(x) \leq \alpha$$

äquivalent, wobei $F_X(x)$ die Verteilungsfunktion der Zufallsvariable X ist. Für eine stetige Zufallsvariable X kann man meist einfacher definieren, daß das α-Quantil derjenige Wert ist, der $F_X(x) = \alpha$ erfüllt. In diesem Fall kann ein Quantil aus der Umkehrfunktion der Verteilungsfunktion F_X bestimmt werden (wenn diese existiert und angegeben werden kann).

2.3.5 Einige spezielle Verteilungen

In diesem Abschnitt betrachten wir einige spezielle Verteilungen, die oft in der Praxis benötigt werden (vgl. Abschnitt 2.4 über beurteilende Statistik).

2.3.5.1 Die Binomialverteilung

Definition 2.23 *Ein Zufallsexperiment, bei dem das Ereignis A eintreten kann, werde n-mal wiederholt. A_i sei das Ereignis, daß beim i-ten Versuch das Ereignis A eintritt. Dann heißt die Versuchsreihe vom Umfang n ein* **Bernoulli-Experiment**[11] *für das Ereignis A, wenn folgende Bedingungen erfüllt sind:*

1. $\forall 1 \leq i \leq n : P(A_i) = p$
2. *Die Ereignisse $A_1, \ldots, A_n$ sind vollständig unabhängig.*

Beschreibt die Zufallsvariable X die Anzahl der Versuche, bei denen in einem Bernoulli-Experiment vom Umfang n das Ereignis A mit $p = P(A)$ eintritt, so besitzt X die Verteilung $\forall x \in \mathbb{N} : (x; P(X = x))$ mit

$$P(X = x) = b_X(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}$$

und heißt **binomialverteilt** mit den Parametern p und n. Diese Formel ist auch als **Bernoullische Formel** bekannt. Der Ausdruck $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ heißt **Binomialkoeffizient**. Es gilt die Rekursionsformel

$$\forall x \in \mathbb{N}_0 : b_X(k+1; p, n) = \frac{(n-x)p}{(x+1)(1-p)}\, b_X(x; p, n)$$
$$\text{mit} \quad b_X(0; p, n) = (1-p)^n$$

Für den Erwartungswert und die Varianz gelten

$$\mu = E(X) = np; \qquad \sigma^2 = D^2(X) = np(1-p).$$

[11] Der Begriff „Bernoulli-Experiment“ wurde zu Ehren des schweizer Mathematikers Jakob Bernoulli (1654–1705) eingeführt.

2.3.5.2 Die Polynomialverteilung

Bernoulli-Experimente lassen sich leicht auf mehr als zwei unvereinbare Ereignisse verallgemeinern. Man gelangt so zur Polynomialverteilung, die eine mehrdimensionale Verteilung ist: Ein Zufallsexperiment werde n-mal unabhängig durchgeführt. $A_1, \ldots, A_k$ seien paarweise unabhängige Ereignisse, von denen bei jedem Versuch genau eines eintreten muß, d.h., $A_1, \ldots, A_k$ sei eine vollständige Ereignisdisjunktion. Bei jedem Versuch trete das Ereignis A_i mit konstanter Wahrscheinlichkeit $p_i = P(A_i)$, $1 \leq i \leq k$, ein. Dann ist die Wahrscheinlichkeit dafür, daß bei den n Versuchen x_i-mal das Ereignis A_i, $i = 1, \ldots, k$, $\sum_{i=1}^{k} x_i = n$, eintritt, gleich

$$\begin{aligned} P(X_1 = x_1, \ldots, X_k = x_k) &= \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k} \\ &= \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}. \end{aligned}$$

Die Gesamtheit der Wahrscheinlichkeiten aller Vektoren $(x_1, \ldots, x_k)$ mit $\sum_{i=1}^{k} x_i = n$ heißt (k-dimensionale) **Polynomialverteilung** mit den Parametern $p_1, \ldots, p_k$ und n. Die Binomialverteilung erhält man offenbar als Spezialfall der Polynomialverteilung, nämlich für $k = 2$. Der Ausdruck $\binom{n}{x_1 \ldots x_k} = \frac{n!}{x_1! \cdots x_k!}$ heißt **Polynomialkoeffizient**, in Analogie zum Binomialkoeffizient $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

2.3.5.3 Die geometrische Verteilung

Die Zufallsvariable X beschreibe die bis zum erstmaligen Eintreten des Ereignisses A mit $p = P(A) > 0$ notwendigen Versuche in einem Bernoulli-Experiment. Sie besitzt die Verteilung $\forall x \in \mathbb{N} : (x; P(X = x))$ mit

$$P(X = x) = g_X(x; p) = p(1-p)^{x-1}$$

und heißt **geometrisch verteilt** mit dem Parameter p. Für die Berechnung der Wahrscheinlichkeiten ist die Rekursionsformel

$$\forall x \in \mathbb{N} : P(X = x+1) = (1-p)P(X = x) \quad \text{mit} \quad P(X = 1) = p$$

hilfreich. Für den Erwartungswert und die Varianz gelten

$$\mu = E(X) = \frac{1}{p}; \qquad \sigma^2 = D^2(X) = \frac{1-p}{p^2}.$$

2.3.5.4 Die hypergeometrische Verteilung

Aus einer Urne, welche M schwarze und $N-M$ weiße, insgesamt also N Kugeln enthält, werden ohne zwischenzeitliches Zurücklegen n Kugeln gezogen. Ist X die Anzahl der gezogenen schwarzen Kugeln, so besitzt X die Verteilung $\forall x; \max(0, n-(N-M)) \leq x \leq \min(n, M) : (x; P(X=x))$ mit

$$P(X=x) = h_X(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

und heißt **hypergeometrisch verteilt** mit den Parametern n, M und N. Es gilt die Rekursionsformel

$$\begin{aligned}&\forall x; \max(0, n-(N-M)) \leq x \leq \min(n, M) :\\ &\quad h_X(x+1; n, M, N) = \frac{(M-x)(n-x)}{(x+1)(N-M-n+x+1)} h_X(x; n, M, N)\\ &\qquad \text{mit} \quad h_X(1; n, M, N) = \frac{M}{N}.\end{aligned}$$

Mit $p = \frac{M}{N}$ und $q = 1-p$ gelten für den Erwartungswert und die Varianz

$$\mu = E(X) = np; \qquad \sigma^2 = D^2(X) = npq\frac{N-n}{N-1}.$$

2.3.5.5 Die Poissonverteilung

Eine Zufallsvariable X mit der Verteilung $\forall x \in \mathbb{N} : (x; P(X=x))$ mit

$$P(X=x) = \Lambda_X(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

heißt **poissonverteilt**[12] mit dem Parameter λ. Es gilt die Rekursionsformel

$$\forall x \in \mathbb{N}_0 : \Lambda_X(x+1; \lambda) = \frac{\lambda}{x+1}\Lambda_X(x; \lambda) \quad \text{mit} \quad \Lambda_X(0; \lambda) = e^{-\lambda}.$$

Für den Erwartungswert und die Varianz gelten

$$\mu = E(X) = \lambda; \qquad \sigma^2 = D^2(X) = \lambda.$$

[12] Die Poissonverteilung erhielt ihren Namen zu Ehren des französischen Mathematikers Siméon-Denis Poisson (1781-1840).

Mit der Poissonverteilung kann man die Häufigkeit des Auftretens von Ereignissen in einem festen Zeitabschnitt beschreiben, etwa die Zahl der tödlichen Verkehrsunfälle pro Jahr.

Für seltene Ereignisse A und eine große Anzahl n von Versuchen kann die Binomialverteilung durch die Poissonverteilung angenähert werden, denn es gilt: Geht in der Binomialverteilung n gegen unendlich, und zwar so, daß $np = \lambda$ konstant bleibt, so gilt

$$\forall x \in \mathbb{N} : \lim_{\substack{n \to \infty \\ np=\lambda}} b_X(x; p, n) = \frac{\lambda^x}{x!} e^{-\lambda}$$

und somit für große n und kleine p die Näherungsformel

$$\forall x \in \mathbb{N} : b_X(x; p, n) \approx \frac{np^x}{x!} e^{-np}.$$

Für poissonverteilte Zufallsvariablen gilt folgendes Reproduktionsgesetz:

Satz 2.17 *Sind X und Y zwei (stochastisch) unabhängige poissonverteilte Zufallsvariablen mit den Parametern λ_X bzw. λ_Y, so ist die Summe $Z = X + Y$ poissonverteilt mit dem Parameter $\lambda_Z = \lambda_X + \lambda_Y$.*

2.3.5.6 Die Gammaverteilung

Während die Poissonverteilung die Häufigkeit des Auftretens von Ereignissen in einem festen Zeitabschnitt beschreibt, bezieht sich die **Gamma-Verteilung** auf die Zeit, die es dauert, bis eine bestimmte Anzahl r von Ereignissen eingetreten ist. Zur Ableitung der Wahrscheinlichkeitsdichte einer gammaverteilten Zufallsvariable X betrachten wir eine Zufallsvariable W, die die Häufigkeit des Auftretens des betrachteten Ereignisses im Zeitinterval $[0, x]$ beschreibt. Aus dem Reproduktionssatz (Satz 2.17) sieht man leicht, daß W poissonverteilt ist mit dem Parameter λx, wobei λ der Parameter der zugehörigen Poissonverteilung ist, die die Häufigkeit des Auftretens des Ereignisses im Einheitsinterval $[0, 1]$ beschreibt. Die Verteilungsfunktion von X ist folglich [Larsen and Marx 1986]

$$\begin{aligned} F_X(x; r) &= P(X \leq x) \\ &= 1 - P(X > x) \\ &= 1 - P(\text{weniger als } r \text{ Ereignisse in } [0, x]) \\ &= 1 - F_W(r - 1) \\ &= 1 - \sum_{k=0}^{r-1} e^{-\lambda x} \frac{(\lambda x)^k}{k!}, \end{aligned}$$

Daher ist die Wahrscheinlichkeitsdichte von X

$$\begin{aligned} f_X(x;r) &= -\frac{\mathrm{d}}{\mathrm{d}x}\sum_{k=0}^{r-1} e^{-\lambda x}\frac{(\lambda x)^k}{k!} \\ &= -\left[-\lambda e^{-\lambda x} + \sum_{k=1}^{r-1}\left((-\lambda)e^{-\lambda x}\frac{(\lambda x)^k}{k!} + e^{-\lambda x}\,\lambda\frac{(\lambda x)^{k-1}}{(k-1)!}\right)\right] \\ &= \sum_{k=0}^{r-1}\lambda e^{-\lambda x}\frac{(\lambda x)^k}{k!} - \sum_{k=0}^{r-2}\lambda e^{-\lambda x}\frac{(\lambda x)^k}{k!} \quad = \quad \frac{\lambda^r}{(r-1)!}x^{r-1}e^{-\lambda x}. \end{aligned}$$

Für den Erwartungswert und die Varianz gelten

$$\mu = E(X) = \frac{r}{\lambda}; \qquad \sigma^2 = D^2(X) = \frac{r}{\lambda^2}.$$

2.3.5.7 Die gleichmäßige Verteilung oder Gleichverteilung

Eine Zufallsvariable X mit der Dichtefunktion

$$f_X(x;a,b) = \begin{cases} \frac{1}{b-a}, & \text{für } x \in [a,b], \\ 0, & \text{sonst,} \end{cases}$$

mit $a, b \in \mathbb{R}$, $a < b$, heißt **gleichverteilt** in $[a,b]$. Ihre Verteilungsfunktion F_X lautet

$$F_X(x;a,b) = \begin{cases} 0, & \text{für } x \le a, \\ \frac{x-a}{b-a}, & \text{für } a \le x \le b, \\ 1, & \text{für } x \ge b. \end{cases}$$

Für den Erwartungswert und die Varianz gelten

$$\mu = E(X) = \frac{a+b}{2}; \qquad \sigma^2 = D^2(X) = \frac{(b-a)^2}{12}.$$

2.3.5.8 Die Normalverteilung

Eine Zufallsvariable X mit der Dichtefunktion

$$N_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

heißt **normalverteilt** mit den Parametern μ und σ^2.
Für den Erwartungswert und die Varianz gelten

$$E(X) = \mu; \qquad D^2(X) = \sigma^2.$$

Die Normalverteilung mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$ nennt man auch **Standardnormalverteilung**.

Die Dichtefunktion $f_X(x;\mu,\sigma^2)$ besitzt ein Maximum an der Stelle μ und Wendepunkte an den Stellen $x = \mu \pm \sigma$. Die Verteilungsfunktion von X ist nicht geschlossen darstellbar; sie wird daher tabelliert, und zwar gewöhnlich für die Standardnormalverteilung, aus der sich die Werte für beliebige Normalverteilungen leicht durch Skalierung gewinnen lassen.

In der Praxis steht man oft auch vor dem umgekehrten Problem, nämlich das Argument der Verteilungsfunktion der Standardnormalverteilung zu bestimmen, für daß sie eine gegebene Wahrscheinlichkeit annimmt (oder anders ausgedrückt: man möchte ein Quantil der Normalverteilung bestimmen). Zur Lösung dieses Problems kann man natürlich auch auf eine Tabelle zurückgreifen. Auf der WWW-Seite zur Vorlesung steht allerdings auch das Programm `ndqtl` zur Verfügung, das die Umkehrfunktion der Verteilungsfunktion der Standardnormalverteilung durch eine gebrochen rationale Funktion approximiert und recht genaue Werte liefert.

Die Normalverteilung ist wohl die wichtigste stetige Verteilung, da sehr viele Zufallsprozesse, speziell physikalische Messungen einer Größe, durch diese Verteilung sehr gut beschrieben werden können. Die theoretische Begründung für diese Tatsache liefert der zentrale Grenzwertsatz:

Satz 2.18 (zentraler Grenzwertsatz)
Für jedes m seien die reellwertigen Zufallsvariablen $X_1, \ldots, X_m$ (stochastisch) unabhängig. Weiter mögen sie die sogenannte **Lindeberg-Bedingung** *erfüllen, d.h., wenn $F_i(x)$ die Verteilungsfunktion der Zufallsvariable X_i, $i = 1, \ldots, m$, μ_i ihr Erwartungswert und σ_i^2 ihre Varianz sind, so möge für jedes $\epsilon > 0$*

$$\lim_{m\to\infty} \frac{1}{V_m^2} \sum_{i=1}^{m} \int_{|x_i-\mu_i|>\epsilon V_m^2} (x_i - \mu_i)^2 \mathrm{d}F_i(x) = 0$$

mit $V_m^2 = \sum_{i=1}^m \sigma_i^2$ gelten. Dann gilt für die standardisierten Summen

$$S_m = \frac{\sum_{i=1}^m (X_i - \mu_i)}{\sqrt{\sum_{i=1}^m \sigma_i^2}}$$

(d.h., standardisiert auf Erwartungswert 0 und Varianz 1), daß

$$\lim_{m\to\infty} P(S_m \le x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t$$

für jedes $x \in \mathbb{R}$, wobei $\Phi(x)$ die Verteilungsfunktion der Standardnormalverteilung ist.

Anschaulich besagt dieser Satz, daß die Summe einer großen Zahl fast beliebig verteilter Zufallsvariablen (die Lindeberg-Bedingung ist eine sehr schwache Einschränkung) annähernd normalverteilt ist. Da nun aber z.B. Messungen physikalischer Größen gewöhnlich einer großen Zahl von Zufallseinflüssen aus verschiedenen unabhängigen Quellen unterliegen, die sich alle addieren, ist das Ergebnis meist näherungsweise normalverteilt. Dies erklärt, warum man so häufig normalverteilte Größen vorfindet.

Die Normalverteilung kann wie die Poissonverteilung für große n auch als Näherung für die Binomialverteilung verwendet werden, nur fällt hier die Einschränkung kleiner Wahrscheinlichkeiten p weg.

Satz 2.19 (Grenzwertsatz von de Moivre-Laplace)[13]
Wenn die Wahrscheinlichkeit für das Auftreten eines Ereignisses A in n unabhängigen Versuchen konstant und gleich p, $0 < p < 1$, ist, so genügt die Wahrscheinlichkeit $P(X = x)$ dafür, daß in diesen Versuchen das Ereignis A genau x-mal eintritt, für $n \to \infty$ der Beziehung

$$\sqrt{np(1-p)}\; P(X=x)\; \frac{1}{\sqrt{2\pi}}\; e^{-\frac{1}{2}y^2} \to 1 \qquad \textit{mit} \qquad y = \frac{x-np}{\sqrt{np(1-p)}},$$

und zwar gleichmäßig für alle x, für die sich y in einem beliebigen endlichen Intervall (a, b) befindet.

Dieser Satz erlaubt es, für große n die Wahrscheinlichkeiten der Binomialverteilung anzunähern durch

$$\forall x; 0 \le x \le n:$$

$$\begin{aligned} P(X=x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &\approx \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(x-np)^2}{2np(1-p)}\right) \qquad \text{oder} \\ P(X=x) &\approx \Phi\left(\frac{x-np+\frac{1}{2}}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right) \qquad \text{und} \end{aligned}$$

$$\forall x_1, x_2; 0 \le x_1 \le x_2 \le n:$$

$$P(x_1 \le X \le x_2) \approx \Phi\left(\frac{x_2-np+\frac{1}{2}}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x_1-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right).$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Schon für $np(1-p) > 9$ sind diese Näherungen brauchbar.

[13] Benannt nach den französischen Mathematikern Abraham de Moivre (1667–1754) und Pierre-Simon Laplace (1749–1827).

2.3.5.9 Die χ^2-Verteilung

Summiert man m unabhängige, standardnormalverteilte Zufallsvariablen (Erwartungswert 0 und Varianz 1), so erhält man eine Zufallvariable X mit der Dichtefunktion

$$f_X(x;m) = \begin{cases} 0, & \text{für } x < 0, \\ \dfrac{1}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right)} \cdot x^{\frac{m}{2}-1} \cdot e^{-\frac{x}{2}}, & \text{für } x \geq 0, \end{cases}$$

wobei Γ die **Gammafunktion** (Verallgemeinerung der Fakultät)

$$\Gamma(x) = \int_0^\infty t^{x-1} e^t \mathrm{d}t$$

mit $x > 0$ ist. Diese Zufallsvariable heißt **χ^2-verteilt** mit m Freiheitsgraden. Für den Erwartungswert und die Varianz gelten

$$E(X) = m; \qquad D^2(X) = 2m.$$

Die χ^2-Verteilung spielt in der statistischen Theorie des Hypothesentests (siehe Abschnitt 2.4.3), z.B. bei Abhängigkeitstests, eine wichtige Rolle.

2.3.5.10 Die Exponentialverteilung

Eine Zufallsvariable X mit der Dichtefunktion

$$f_X(x;\alpha) = \begin{cases} 0, & \text{für } x \leq 0, \\ \alpha e^{-\alpha x}, & \text{für } x > 0, \end{cases}$$

mit $\alpha > 0$ heißt **exponentialverteilt** mit dem Parameter α. Ihre Verteilungsfunktion F_X lautet

$$F_X(x;\alpha) = \begin{cases} 0, & \text{für } x \leq 0, \\ 1 - e^{-\alpha x}, & \text{für } x > 0. \end{cases}$$

Für den Erwartungswert und die Varianz gelten

$$\mu = E(X) = \frac{1}{\alpha}; \qquad \sigma^2 = D^2(X) = \frac{1}{\alpha^2}.$$

Die Exponentialverteilung wird gern benutzt, um die Zeiten zwischen den Ankünften von Personen oder Aufträgen, die sich in eine Warteschlange einreihen, zu beschreiben.

2.4 Beurteilende Statistik

Mit der beurteilenden Statistik versucht man die Frage zu beantworten, ob beobachtete Erscheinungen typisch bzw. gesetzmäßig sind oder auch durch Zufallseinflüsse hervorgerufen worden sein könnten. Außerdem versucht man Wahrscheinlichkeitsverteilungen zu finden, die gute Modelle des die gegebenen Daten erzeugenden Prozesses sind, und ihre Parameter zu schätzen. Wichtige Aufgaben der beurteilenden Statistik sind daher:

- **Parameterschätzung**
  Unter Voraussetzung eines Modells des die Daten erzeugenden Prozesses, speziell einer Annahme über die Familie der Verteilungen der zugrundeliegenden Zufallsvariablen, werden die Parameter dieses Modells aus den Daten geschätzt.

- **Hypothesentest**
  Es werden Hypothesen über den datenerzeugenden Prozeß anhand der gegebenen Daten geprüft. Spezielle Hypothesentests sind:

  - Parametertest: Test, ob ein Parameter einer Verteilung einen bestimmten Wert haben kann bzw. ob die Parameter der zwei verschiedenen Datensätzen zugrundeliegenden Verteilungen verschieden sind.
  - Anpassungstest: Test, ob eine bestimmte Verteilungsannahme zu den Daten paßt, oder ob Abweichungen von den unter dieser Annahme zu erwartenden Datencharakteristiken nicht mehr durch Zufallseinflüsse erklärt werden können.
  - Abhängigkeitstest: Test, ob zwei Merkmale abhängig sind, oder ob eventuelle Abweichungen von einer unabhängigen Verteilung durch Zufallseinflüsse erklärbar sind.

- **Modellauswahl**
  Unter verschiedenen Modellen, mit denen man die Daten erklären könnte, wird das am besten passende ausgewählt, wobei auch die Komplexität der Modelle berücksichtigt wird.

Die Parameterschätzung kann als Spezialfall der Modellauswahl gesehen werden, bei der die Klasse der Modelle, aus der gewählt werden kann, stark eingeschränkt ist. Der Anpassungstest steht auf der Grenze zwischen Hypothesentest und Modellauswahl, da er dazu dient, zu prüfen, ob ein Modell einer bestimmten Klasse zur Erklärung der Daten geeignet ist.

2.4.1 Zufallsstichproben

Im Kapitel über beschreibende Statistik haben wir sogenannte *Stichproben* (Vektoren von beobachteten/gemessenen Merkmalswerten) betrachtet und sie in Tabellen und Schaubildern übersichtlich dargestellt. Auch in der beurteilenden Statistik untersucht man Stichproben. Hier wendet man das mathematische Instrumentarium der Wahrscheinlichkeitsrechnung an, um aus Stichproben (neues) Wissen zu gewinnen oder Hypothesen zu überprüfen. Damit dies möglich ist, müssen die Stichprobenwerte durch Zufallsexperimente gewonnen werden. Solche Stichproben heißen **Zufallsstichproben**.

Die Zufallsvariable, welche bei der Durchführung des entsprechenden Zufallsexperimentes den Stichprobenwert x_i liefert, nennen wir X_i. Der Wert x_i heißt **Realisierung** der Zufallsvariable X_i, $i = 1, \ldots, n$. Somit können wir eine Zufallsstichprobe $x = (x_1, \ldots, x_n)$ als *Realisierung des Zufallsvektors* $X = (X_1, \ldots, X_n)$ auffassen. Eine Zufallsstichprobe heißt **unabhängig**, wenn die Zufallsvariablen $X_1, \ldots, X_n$ (stochastisch) unabhängig sind, also wenn gilt

$$\forall c_1, \ldots, c_n \in \mathbb{R}: \quad P\Big(\bigwedge_{i=1}^{n} X_i \leq c_i\Big) = \prod_{i=1}^{n} P(X_i \leq c_i).$$

Eine unabhängige Zufallsstichprobe heißt **einfach**, wenn die Zufallsvariablen $X_1, \ldots, X_n$ alle dieselbe Verteilungsfunktion besitzen. Analog werden wir auch den zugehörigen Zufallsvektor *unabhängig* bzw. *einfach* nennen.

2.4.2 Parameterschätzung

Wie bereits gesagt, werden bei der Parameterschätzung unter Voraussetzung eines Modells des die Daten erzeugenden Prozesses, speziell einer Annahme über die Familie der Verteilungen der zugrundeliegenden Zufallsvariablen, die Parameter dieses Modells aus den Daten geschätzt.

Gegeben:
- Ein Datensatz und
- eine Familie von gleichartigen, parametrisierten Verteilungen $f_X(x; \theta_1, \ldots, \theta_k)$.

Solche Familien können z.B. sein:

Die Familie der Binomialverteilungen $b_X(x; p, n)$ mit den Parametern p, $0 \leq p \leq 1$, und $n \in \mathbb{N}$, wobei n allerdings bereits durch die Stichprobengröße gegeben sind.

Die Familie der Poissonverteilungen $\Lambda_X(x; \lambda, n)$ mit den Parametern $\lambda > 0$ und $n \in \mathbb{N}$, wobei n wieder durch die Stichprobengröße gegeben sind.

Die Familie der Normalverteilungen $N_X(x; \mu, \sigma^2)$ mit den Parametern μ (Erwartungswert) und σ^2 (Varianz).

Annahme: Der Prozeß, der die Daten erzeugt hat, wird durch ein Element der betrachteten Familie von Verteilungen angemessen beschrieben: *Verteilungsannahme.*

Gesucht: Dasjenige Element der betrachteten Familie von Verteilungen (bestimmt durch die Werte der Parameter), das das beste Modell für die gegebenen Daten ist.

Schätzer für Parameter sind **Statistiken**, d.h. Funktionen der Stichprobenwerte eines gegebenen Datensatzes. Sie sind daher Funktionen von (Realisierungen von) Zufallsvariablen und folglich selbst (Realisierungen von) Zufallsvariablen. Wir können daher das gesamte Instrumentarium, das die Wahrscheinlichkeitsrechnung zur Untersuchung von Zufallsvariablen bereitstellt, auf Parameterschätzer anwenden.

Man unterscheidet i.w. zwei Arten der Parameterschätzung:

- **Punktschätzer** bestimmen den besten Wert eines Parameters relativ zu den Daten und bezüglich bestimmter Qualitätskriterien.
- **Intervallschätzer** geben einen Bereich (Konfidenzintervall) an, in dem der wahre Wert des Parameters mit hoher Sicherheit liegt, wobei man den Grad der Sicherheit wählen kann.

2.4.2.1 Punktschätzung

Offenbar ist nicht jede Statistik, also nicht jede aus den Stichprobenwerten berechnete Funktion, ein brauchbarer Punktschätzer für einen gesuchten Parameter θ. Vielmehr muß eine Statistik bestimmte Eigenschaften aufweisen, um ein brauchbarer Schätzer zu sein. Wünschenswerte Eigenschaften sind:

- **Konsistenz (consistency)**
  Wächst die Menge der zur Verfügung stehenden Daten, so sollte der Schätzwert immer näher an dem wahren Wert des geschätzten Parameters liegen, zumindest mit immer größerer Wahrscheinlichkeit. Dies kann man formalisieren, indem man fordert, daß die Schätzfunktion für zunehmende Stichprobengröße in Wahrscheinlichkeit gegen

den wahren Wert des Parameters konvergiert. Ist z.B. T ein Schätzer für den Parameter θ, so sollte gelten:

$$\forall \varepsilon > 0 : \quad \lim_{n \to \infty} P(|T - \theta| < \varepsilon) = 1,$$

wobei n die Stichprobengröße ist. Diese Bedingung sollte jeder Punktschätzer mindestens erfüllen.

- **Erwartungstreue (unbiasedness)**
  Der Schätzer sollte nicht zu einer Über- oder Unterschätzung des Parameters neigen, sondern vielmehr im Mittel den richtigen Wert liefern. Formal ausgedrückt bedeutet das, daß der Erwartungswert des Schätzers der wahre Wert des Parameters sein sollte. Ist z.B. wieder T ein Schätzer für den Parameter θ, so sollte gelten

  $$E(T) = \theta$$

  und zwar unabhängig von der Größe der Stichprobe.

- **Wirksamkeit/Effizienz (efficiency)**
  Die Schätzung sollte so genau wie möglich sein, d.h. die Abweichung vom wahren Wert sollte möglichst klein sein. Dies läßt sich formalisieren, indem man fordert, daß die Varianz des Schätzers möglichst klein sein soll, denn die Varianz ist ein naheliegendes Maß für die Genauigkeit. Seien etwa T und U erwartungstreue Schätzer für den Parameter θ. Dann heißt T *wirksamer* oder *effizienter als* U, wenn

  $$D^2(T) < D^2(U).$$

  Ob ein Schätzer die größte Wirksamkeit erreicht, die möglich ist, läßt sich leider nicht in allen Fällen bestimmen.

- **Suffizienz/Informationsausschöpfung (sufficiency)**
  Die Schätzfunktion sollte die in den Daten enthaltene Information über den gesuchten Parameter optimal nutzen. Dies kann man dadurch präzisieren, daß man fordert, daß verschiedene Stichproben, die den gleichen Schätzwert liefern, auch (gegeben den geschätzen Wert des Parameters) gleich wahrscheinlich sein sollten. Denn sind sie nicht gleich wahrscheinlich, läßt sich offenbar aus den Daten weitere Information über den Parameter gewinnen. Formal definiert man, daß ein Schätzer T für einen Parameter θ *suffizient* oder *erschöpfend* ist,

wenn für alle Zufallsstichproben $x = (x_1, \ldots, x_n)$ mit $T(x) = t$ der Ausdruck

$$\frac{f_{X_1}(x_1;\theta)\cdots f_{X_n}(x_n;\theta)}{f_T(t;\theta)}$$

nicht von θ abhängt [Larsen and Marx 1986].

Man beachte, daß die in der Definition der Wirksamkeit bzw. Effizienz betrachteten Schätzer erwartungstreu sein müssen, da sonst beliebige Konstanten (Varianz $D^2 = 0$) effiziente Schätzer wären. Auf die Konsistenz als zusätzliche Bedingung kann man dagegen verzichten, da man allgemein zeigen kann, daß ein erwartungstreuer Schätzer T für einen Parameter θ, der zusätzlich die Bedingung

$$\lim_{n\to\infty} D^2(T) = 0$$

erfüllt, konsistent ist [Bosch 1994] (wie ja auch naheliegend).

2.4.2.2 Beispiele zur Punktschätzung

Gegeben: Eine Familie von Gleichverteilungen auf dem Intervall $[0, \theta]$, d.h.

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{falls } 0 \le x \le \theta, \\ 0, & \text{sonst.} \end{cases}$$

Gesucht: Schätzwert für den unbekannten Parameter θ.

a) Wähle als Schätzwert das Maximum der Stichprobenwerte, d.h. wähle die Schätzfunktion $T = \max\{X_1, \ldots, X_n\}$.

Um Aussagen über die Eigenschaften dieses Schätzers machen zu können, bestimmen wir zunächst seine Wahrscheinlichkeitsdichte[14]:

$$\begin{aligned} f_T(t;\theta) &= \frac{\mathrm{d}}{\mathrm{d}t}F_T(t;\theta) = \frac{\mathrm{d}}{\mathrm{d}t}P(T \le t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}P(\max\{X_1, \ldots, X_n\} \le t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}P\Big(\textstyle\bigwedge_{i=1}^{n} X_i \le t\Big) = \frac{\mathrm{d}}{\mathrm{d}t}\prod_{i=1}^{n} P(X_i \le t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}(F_X(t;\theta))^n = n\cdot(F_X(t;\theta))^{n-1} f_X(t,\theta) \end{aligned}$$

[14] Man erinnere sich daran, daß Schätzer Funktionen von Zufallsvariablen und damit selbst Zufallsvariablen sind. Folglich haben sie auch eine Wahrscheinlichkeitsdichte.

mit

$$F_X(x;\theta) \;=\; \int_{-\infty}^{x} f_X(x;\theta)\mathrm{d}x \;=\; \begin{cases} 0, & \text{falls } x \le 0, \\ \frac{x}{\theta}, & \text{falls } 0 \le x \le \theta, \\ 1, & \text{falls } x \ge \theta. \end{cases}$$

Damit folgt

$$f_T(t;\theta) = \frac{n \cdot t^{n-1}}{\theta^n} \quad \text{für} \quad 0 \le t \le \theta.$$

T ist ein konsistenter Schätzer für θ:

$$\begin{aligned} \lim_{n\to\infty} P(|T-\theta| < \epsilon) &= \lim_{n\to\infty} P(T > \theta - \epsilon) \\ &= \lim_{n\to\infty} \int_{\theta-\epsilon}^{\theta} \frac{n \cdot t^{n-1}}{\theta^n}\mathrm{d}t \;=\; \lim_{n\to\infty} \left[\frac{t^n}{\theta^n}\right]_{\theta-\epsilon}^{\theta} \\ &= \lim_{n\to\infty} \left(\frac{\theta^n}{\theta^n} - \frac{(\theta-\epsilon)^n}{\theta^n}\right) \\ &= \lim_{n\to\infty} \left(1 - \left(\frac{\theta-\epsilon}{\theta}\right)^n\right) \;=\; 1 \end{aligned}$$

T ist jedoch kein erwartungstreuer Schätzer für θ. Dies ist bereits anschaulich klar, da T den Wert von θ ja nur unterschätzen, niemals jedoch überschätzen kann. Der Schätzwert wird daher „fast immer“ zu klein sein. Formal erhält man:

$$\begin{aligned} E(T) &= \int_{-\infty}^{\infty} t \cdot f_T(t;\theta)\mathrm{d}t \;=\; \int_0^{\theta} t \cdot \frac{n \cdot t^{n-1}}{\theta^n}\mathrm{d}t \\ &= \left[\frac{n \cdot t^{n+1}}{(n+1)\theta^n}\right]_0^{\theta} \;=\; \frac{n}{n+1}\theta \;<\; \theta \quad \text{für } n < \infty. \end{aligned}$$

b) Wähle als Schätzfunktion $U = \frac{n+1}{n}T = \frac{n+1}{n}\max\{X_1, \ldots, X_n\}$.

Der Schätzer U für den Parameter θ der Gleichverteilung ist konsistent und erwartungstreu. Auf einen formalen Nachweis, der wie unter a) geführt wird, sei hier verzichtet. Die Wahrscheinlichkeitsdichtefunktion dieses Schätzers werden wir allerdings später noch einmal brauchen. Sie lautet (vergleiche die Ableitung der Wahrscheinlichkeitsdichte von T):

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n}\frac{u^{n-1}}{\theta^n}.$$

Gegeben: Eine Familie von Normalverteilungen $N(x;\mu,\sigma)$, d.h.

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gesucht: Schätzwerte für die unbekannten Parameter μ und σ^2.

a) Der (empirische) Median und der (empirische) Mittelwert sind beide konsistente und erwartungstreue Schätzer für den Parameter μ. Der Median ist weniger effizient als der Mittelwert. Da er nur Rangordnungsinformationen aus den Daten nutzt, ist er auch nicht suffizient bzw. erschöpfend. Obwohl der Median für kleine Stichproben wegen seiner geringeren Anfälligkeit für Ausreißer vorzuziehen ist, ist seine Varianz für hinreichenden Stichprobenumfang größer als die des (empirischen) Mittelwertes.

b) Die Funktion $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ist ein konsistenter, aber nicht erwartungstreuer Schätzer für den Parameter σ^2, denn diese Funktion neigt dazu, die Varianz zu unterschätzen. Die (empirische) Varianz $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ ist dagegen ein konsistenter und erwartungstreuer Schätzer für den Parameter σ^2. (Allerdings ist $\sqrt{S^2}$, wegen der Wurzel, *kein* erwartungstreuer Schätzer für die Standardabweichung.)

Gegeben: Eine Familie von Polynomialverteilungen, d.h.

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k;\theta_1,\ldots,\theta_k,n) = \frac{n!}{\prod_{i=1}^{k} x_i!}\prod_{i=1}^{k}\theta_i^{x_i},$$

wobei die θ_i die Wahrscheinlichkeiten des Auftretens von Merkmalswerten a_i sind und die Zufallsvariablen X_i beschreiben, wie oft die Merkmalswerte a_i in der Stichprobe auftreten.

Gesucht: Schätzwerte für die unbekannten Parameter $\theta_1,\ldots,\theta_k$.

Die relativen Häufigkeiten $R_i = \frac{X_i}{n}$ der Merkmalswerte in der Stichprobe sind konsistente, erwartungstreue, wirksamste und suffiziente Schätzer für die unbekannten Parameter θ_i, $i = 1,\ldots,k$ ($\rightarrow$ Übungsaufgabe).

2.4.2.3 Maximum-Likelihood-Schätzung

Bisher haben wir Schätzfunktionen für Parameter direkt angegeben. In der Tat sind für viele Problemstellungen geeignete (konsistente, erwartungstreue und effiziente) Schätzfunktionen bekannt, so daß man sie einfach aus

Statistikbücher entnehmen kann. Dennoch wollen wir uns hier kurz damit beschäftigen, wie man Schätzfunktionen finden kann.

Neben der Momentenmethode, auf die wir hier nicht eingehen wollen, ist die i.w. von R.A. Fisher entwickelte Maximum-Likelihood-Schätzung[15] eines der bekanntesten Verfahren zum Auffinden von Schätzfunktionen. Das Prinzip ist sehr einfach: Es wird der Wert des zu schätzenden Parameters (bzw. der Wertesatz der zu schätzenden Parameter) gewählt, der die vorliegende Zufallsstichprobe am wahrscheinlichsten macht. Man geht dazu wie folgt vor: Wären die Parameter der wahren, den Daten zugrundeliegenden Verteilung bekannt, so könnte man die Wahrscheinlichkeit berechnen, daß bei einem Zufallsexperiment die beobachtete Zufallsstichprobe erzeugt wird. Diese Wahrscheinlichkeit kann man jedoch auch mit den unbekannten Parametern aufschreiben (wenn auch nicht numerisch ausrechnen). Man erhält eine Funktion, die die Wahrscheinlichkeit der Zufallsstichprobe in Abhängigkeit von den unbekannten Parametern angibt. Diese Funktion heißt *Likelihood-Funktion*. Aus dieser Funktion berechnet man durch (partielle) Ableitung nach dem zu schätzenden Parameter und Nullsetzen eine Schätzfunktion.

2.4.2.4 Beispiel zur Maximum-Likelihood-Schätzung

Gegeben: Eine Familie von Normalverteilungen $N_X(x;\mu,\sigma^2)$, d.h.

$$N_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gesucht: Schätzwerte für die unbekannten Parameter μ und σ^2.

Die Likelihood-Funktion einer einfachen Zufallsstichprobe $x = (x_1,\ldots,x_n)$, die eine Realisierung eines Vektors $X = (X_1,\ldots,X_n)$ von mit den Parametern μ und σ^2 normalverteilten Zufallsvariablen ist, lautet

$$L(x_1,\ldots,x_n;\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right),$$

denn sie soll ja die Wahrscheinlichkeit der dem Datensatz zugrundeliegenden Stichprobe in Abhängigkeit von den Parametern μ und σ^2 beschreiben. Durch Ausnutzung der bekannten Rechenregeln für die Exponentialfunktion

[15] R.A. Fisher hat die Methode der Maximum-Likelihood-Schätzung allerdings nicht erfunden; schon C.F. Gauß und D. Bernoulli haben sie verwendet. Aber erst Fisher hat sie systematisch untersucht und in der Statistik etabliert [Larsen and Marx 1986].

können wir diesen Ausdruck umformen in

$$L(x_1,\ldots,x_n;\mu,\sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}\ \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Zur Bestimmung des Maximums dieser Funktion bezüglich der Parameter μ und σ^2 ist es günstig, den natürlichen Logarithmus zu bilden. Das Maximum ändert sich dadurch nicht, da der natürliche Logarithmus eine monotone Funktion ist. Wir erhalten so:

$$\ln L(x_1,\ldots,x_n;\mu,\sigma^2) = -n\ln\left(\sqrt{2\pi\sigma^2}\right)\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Um das Maximum zu bestimmen, setzen wir nun die partiellen Ableitungen nach μ und σ^2 gleich 0. Die partielle Ableitung nach μ ist

$$\frac{\partial}{\partial\mu}\ln L(x_1,\ldots,x_n;\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \stackrel{!}{=} 0,$$

woraus

$$\sum_{i=1}^{n}(x_i-\mu) = \left(\sum_{i=1}^{n}x_i\right) - n\mu \stackrel{!}{=} 0,$$

also der Schätzwert

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

für den Parameter μ folgt. Die partielle Ableitung der logarithmierten Likelihood-Funktion nach σ^2 liefert

$$\frac{\partial}{\partial\sigma^2}\ln L(x_1,\ldots,x_n;\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \stackrel{!}{=} 0.$$

Daraus folgt unter Einsetzen des oben bestimmten Schätzwertes $\hat{\mu}$ für den Parameter μ die Schätzfunktion

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}x_i\right)^2$$

für den Parameter σ^2. Man beachte, daß man eine nicht erwartungstreue Schätzfunktion erhält. (Wir haben bereits oben gesehen, daß die empirische

Varianz mit einem Vorfaktor von $\frac{1}{n}$ statt $\frac{1}{n-1}$ nicht erwartungstreu ist.) Dies zeigt, daß es für die Varianz einer Normalverteilung keine Schätzfunktion mit allen wünschbaren Eigenschaften gibt. Unter dem vom erwartungstreuen Schätzer gelieferten Wert sind die Daten nicht maximal wahrscheinlich und der von einem Maximum-Likelihood-Schätzer gelieferte Schätzwert ist nicht erwartungstreu.

2.4.2.5 Maximum-A-posteriori-Schätzung

Als Alternative zur Maximum-Likelihood-Schätzung gibt es die Maximum-A-posteriori-Schätzung, die auf dem Bayesschen Satz beruht. Bei dieser Schätzmethode setzt man eine A-priori-Verteilung über dem Wertebereich des bzw. der Parameter voraus und berechnet mit Hilfe dieser Verteilung und der gegebenen Daten über den Bayesschen Satz eine A-posteriori-Verteilung über dem Wertebereich des bzw. der Parameter. Der Parameterwert mit der größten A-posteriori-Wahrscheinlichkeit(sdichte) wird als Schätzwert gewählt. D.h., es wird der Wert θ gewählt, der

$$f(\theta \mid D) = \frac{f(D \mid \theta) f(\theta)}{f(D)} = \frac{f(D \mid \theta) f(\theta)}{\int_{-\infty}^{\infty} f(D \mid \theta) f(\theta) \mathrm{d}\theta}$$

maximiert, wobei D die gegebenen Daten und $f(\theta)$ die vorausgesetzte A-priori-Verteilung sind.[16] Zum Vergleich: Bei der oben besprochenen Maximum-Likelihood-Schätzung wird der Wert des Parameters θ gewählt, der $f(D \mid \theta)$, also die Wahrscheinlichkeit der Daten, maximiert.

2.4.2.6 Beispiel zur Maximum-A-posteriori-Schätzung

Um klar zu machen, warum es sinnvoll sein kann, eine A-priori-Verteilung über den möglichen Werten des Parameters θ anzugeben, betrachten wir folgende drei Situationen:

- Ein Betrunkener behauptet, voraussagen zu können, wie eine Münze fallen wird (Kopf oder Zahl). Bei zehn Würfen sagt er tatsächlich immer die richtige Seite vorher.
- Eine Teetrinkerin behauptet, herausschmecken zu können, ob zuerst Tee oder zuerst Milch in die Tasse gegossen wurde. Bei zehn Versuchen gibt sie tatsächlich stets die richtige Reihenfolge an.

[16] Man beachte, daß diese Sichtweise nur in der subjektiven (personalistischen) Deutung der Wahrscheinlichkeit möglich ist, da der Wert des Parameters in der frequentistischen (empirischen) Sichtweise fest und nicht Gegenstand eines Zufallsexperimentes ist, also auch keine Wahrscheinlichkeitsdichte besitzt.

- Ein Musikkenner behauptet, anhand einer einzelnen Notenseite erkennen zu können, ob die Musik von Mozart stammt oder nicht. Bei zehn zufällig ausgewählten Notenseiten liegt er tatsächlich stets richtig.

Sei nun θ der (unbekannte) Parameter, der die Wahrscheinlichkeit angibt, mit der die richtige Aussage gemacht wird. Die Datenlage ist in allen drei Fällen gleich: 10 richtige, 0 falsche Aussagen. Dennoch werden wir es kaum für angemessen halten, diese drei Fälle gleich zu behandeln, wie es die Maximum-Likelihood-Schätzung tut. Wir werden nicht glauben, daß der Betrunkene tatsächlich die Münzseite vorhersagen kann, sondern annehmen, daß er einfach „Glück gehabt“ hat. Auch der Teetrinkerin stehen wir skeptisch gegenüber, wenn auch unsere Ablehnung weniger strikt ist als bei dem Betrunkenen. Vielleicht gibt es ja chemische Veränderungen je nach der Tee-Milch-Reihenfolge, die sich in minimalen Geschmacksunterschieden zeigen, die eine passionierte Teetrinkerin herausschmecken kann. Aber wir halten diese Möglichkeit für sehr unwahrscheinlich. Dagegen sind wir leicht bereit, dem Musikkenner Glauben zu schenken. Schließlich gibt es Unterschiede zwischen den Stilen verschiedener Komponisten, die einem versierten Kenner wohl auch schon anhand einer Notenseite erlauben, zu erkennen, ob die Musik von Mozart komponiert wurde oder nicht.

Die beschriebenen Haltungen den drei Situationen gegenüber können wir in der A-priori-Verteilung über dem Wertebereich des Parameters θ zum Ausdruck bringen: Im Falle des Betrunkenen werden wir nur dem Wert 0.5 ein Wahrscheinlichkeitsdichte verschieden von 0 zuschreiben[17]. Im Falle der Teetrinkerin werden wir eine A-priori-Verteilung wählen, die den Werten nahe 0.5 eine hohe, zur 1 hin sehr schnell abnehmende Wahrscheinlichkeitsdichte zuordnet. Im Falle des Musikkenners werden wir dagegen auch höheren Werten (zur 1 hin) eine beträchtliche Wahrscheinlichkeitsdichte zuerkennen. Das führt dazu, daß wir im Falle des Betrunkenen unabhängig von den Daten stets 0.5 als Parameterwert schätzen. Im Falle der Teetrinkerin kann uns erst eine sehr deutliche Datenlage dazu bringen, höhere Werte des Parameters θ zu akzeptieren. Im Falle des Musikkenners dagegen reichen wenige positive Beispiele, um zu einem hohen Schätzwert für θ zu kommen.

Die A-priori-Verteilung enthält folglich Hintergrundwissen über den die Daten erzeugenden Prozeß und drückt aus, welche Parameterwerte wir erwarten bzw. wie leicht wir bereit sind, sie zu akzeptieren. Die Wahl der A-priori-Verteilung ist allerdings ein kritisches Problem, da sie subjektiv getroffen werden muß. Je nach Erfahrungsschatz werden Menschen unterschiedliche A-priori-Verteilungen wählen.

[17] Formal: Dirac-Puls bei $\theta = 0.5$.

2.4.2.7 Intervallschätzung

Ein aus einer Zufallstichprobe berechneter Schätzwert für einen Parameter wird i.a. von dem tatsächlichen Wert des Parameters abweichen. Es ist daher nützlich, wenn man Aussagen über diese (unbekannten) Abweichungen und ihre zu erwartende Größe machen kann. Die einfachste Möglichkeit ist sicherlich, neben einem (Punkt-)Schätzwert t die Standardabweichung $D(T)$ des Schätzers anzugeben, also

$$t \pm D(T) = t \pm \sqrt{D^2(T)}.$$

Eine bessere Möglichkeit besteht darin, Intervalle — sogenannte *Konfidenz-* oder *Vertrauensintervalle* — zu bestimmen, in denen der wahre Wert des Parameters mit hoher Wahrscheinlichkeit liegt.

Die Grenzen dieser Konfidenzintervalle werden durch bestimmte Berechnungvorschriften aus den Stichprobenwerten gewonnen. Sie sind also Statistiken, daher wie Punktschätzer (Realisierungen von) Zufallsvariablen und können folglich analog behandelt werden. Formal werden sie so definiert:

Sei $X = (X_1, \ldots, X_n)$ ein einfacher Zufallsvektor, dessen Zufallsvariablen die Verteilungsfunktion $F_{X_i}(x_i; \theta)$ mit dem (unbekannten) Parameter θ besitzen. Weiter seien $A = g_A(X_1, \ldots, X_n)$ und $B = g_B(X_1, \ldots, X_n)$ zwei auf X definierte Schätzfunktionen, so daß

$$P(A < \theta < B) = 1 - \alpha, \qquad P(\theta \leq A) = \frac{\alpha}{2}, \qquad P(\theta \geq B) = \frac{\alpha}{2}.$$

Dann heißt das zufällige Intervall $[A, B]$ (oder eine Realisierung $[a, b]$ dieses zufälligen Intervalls) $(1-\alpha)\cdot 100\%$ **Konfidenz-** oder **Vertrauensintervall** (confidence interval) für den (unbekannten) Parameter θ. Der Wert $1 - \alpha$ heißt **Konfidenzniveau** (confidence level).

Man beachte, daß sich der Ausdruck „Konfidenz“ in „Konfidenzintervall“ auf das *Verfahren* und *nicht* auf ein *Ergebnis* des Verfahrens (eine Realisierung des zufälligen Intervalls) bezieht. *Vor* dem Sammeln von Daten enthält ein $(1-\alpha)\cdot 100\%$ Konfidenzintervall den wahren Wert mit Wahrscheinlichkeit $1 - \alpha$. *Nach* dem Sammeln von Daten und dem Berechnen des Intervalls sind die Intervallgrenzen dagegen keine Zufallsvariablen mehr. Folglich enthält das Intervall den Parameter θ oder nicht (Wahrscheinlichkeit 1 oder 0 — allerdings ist unbekannt, welcher Wert der richtige ist).

Die obige Definition eines Konfidenzintervalls ist nicht spezifisch genug, um aus ihr eine Berechnungsvorschrift ableiten zu können. In der Tat sind die Schätzer A und B gar nicht eindeutig bestimmt: Die Mengen der Realisierungen des Zufallsvektors $X_1 \ldots, X_n$, für die $A \geq \theta$ und $B \leq \theta$ gilt,

müssen ja lediglich disjunkt sein und jeweils die Wahrscheinlichkeit $\frac{\alpha}{2}$ besitzen. Um eine Berechnung der Grenzen A und B eines Konfidenzintervalls zu ermöglichen, schränkt man diese Schätzer meist wie folgt ein: Sie werden nicht als allgemeine Funktionen des Zufallsvektors, sondern als Funktionen eines Punktschätzers T für den Parameter θ angesetzt, d.h.

$$A = h_A(T) \quad \text{und} \quad B = h_B(T).$$

Damit lassen sich Konfidenzintervalle allgemein bestimmen, indem man statt des Ereignisses $A < \theta < B$ das entsprechende Ereignis in bezug auf den Schätzer T betrachtet, also $A^* < T < B^*$. Natürlich ist dieser Weg nur gangbar, wenn sich aus den in bezug auf T anzusetzenden Umkehrfunktionen $A^* = h_A^{-1}(\theta)$ und $B^* = h_B^{-1}(\theta)$ die Funktionen $h_A(T)$ bzw. $h_B(T)$ bestimmen lassen.

$$\begin{array}{lrcl}
(\text{Idee:} & P(A^* < T < B^*) & = & 1-\alpha \\
\Rightarrow & P(h_A^{-1}(\theta) < T < h_B^{-1}(\theta)) & = & 1-\alpha \\
\Rightarrow & P(h_A(T) < \theta < h_B(T)) & = & 1-\alpha \\
\Rightarrow & P(A < \theta < B) & = & 1-\alpha.)
\end{array}$$

Dies ist jedoch nicht immer (hinreichend leicht) möglich.

2.4.2.8 Beispiele zur Intervallschätzung

Gegeben: Eine Familie von Gleichverteilungen auf dem Intervall $[0, \theta]$, d.h.

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta}, & \text{falls } 0 \le x \le \theta, \\ 0, & \text{sonst.} \end{cases}$$

Gesucht: Ein Konfidenzintervall für den unbekannten Parameter θ.

Hier gelangt man zum Erfolg, indem man mit dem erwartungstreuen Punktschätzer $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$ ansetzt:

$$\begin{aligned}
P(U \le B^*) &= \int_0^{B^*} f_U(u;\theta)\mathrm{d}u &= \frac{\alpha}{2} \\
P(U \ge A^*) &= \int_{A^*}^{\frac{n+1}{n}\theta} f_U(u;\theta)\mathrm{d}u &= \frac{\alpha}{2}
\end{aligned}$$

Wie wir aus dem Abschnitt über Punktschätzung wissen, ist

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n}.$$

Damit erhalten wir

$$B^* = \sqrt[n]{\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta \qquad \text{und} \qquad A^* = \sqrt[n]{1-\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta,$$

also

$$P\left(\sqrt[n]{\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta < U < \sqrt[n]{1-\frac{\alpha}{2}}\,\frac{n+1}{n}\,\theta\right) = 1-\alpha,$$

woraus man leicht

$$P\left(\frac{U}{\sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}} < \theta < \frac{U}{\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}}\right) = 1-\alpha$$

errechnet. Aus diesem Ausdruck lassen sich die Grenzen A und B des Konfidenzintervall direkt ablesen.

Gegeben: Eine Familie von Binomialverteilungen, d.h.

$$b_X(x;\theta,n) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Gesucht: Ein Konfidenzintervall für den unbekannten Parameter θ.

Für eine exakte Berechnung eines Konfidenzintervalls für den Parameter θ setzt man analog zu dem obigen Beispiel an:

$$\begin{aligned} P(X \geq A^*) &= \sum_{i=A^*}^{n} \binom{n}{i}\theta^i(1-\theta)^{n-i} \leq \frac{\alpha}{2}, \\ P(X \leq B^*) &= \sum_{i=0}^{B^*} \binom{n}{i}\theta^i(1-\theta)^{n-i} \leq \frac{\alpha}{2}. \end{aligned}$$

Da diese Ausdrücke jedoch nicht leicht auszuwerten sind, wählt man meist einen anderen Weg, nämlich die näherungsweise Berechnung eines Konfidenzintervalls unter Zuhilfenahme des Grenzwertsatzes von de Moivre-Laplace (siehe Satz 2.19 auf Seite 65). Dieser Satz erlaubt es, die Binomialverteilung durch eine Standardnormalverteilung anzunähern:

$$b_X(x;\theta,n) \approx N\left(\frac{x-n\theta}{\sqrt{n\theta(1-\theta)}};0,1\right) = \frac{1}{\sqrt{2\pi n\theta(1-\theta)}}\exp\left(-\frac{(x-n\theta)^2}{2n\theta(1-\theta)}\right).$$

Diese Näherung ist bereits für $n\theta(1-\theta) > 9$ sehr gut.

Ähnlich wie oben setzt man nun an:

$$\begin{aligned} P\left(\frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} \le B^*\right) &= \frac{\alpha}{2}, \\ P\left(\frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} \ge A^*\right) &= \frac{\alpha}{2}. \end{aligned}$$

Man beachte, daß hier nicht (wie oben) direkt der Schätzer (hier $T = \frac{X}{n}$) für den unbekannten Parameter θ auftritt, sondern (wegen der verwendeten Näherung) eine Funktion dieses Schätzers.

Durch die Symmetrie der Normalverteilung vereinfachen sich die Ableitungen. So wissen wir wegen dieser Symmetrie, daß $B^* = -A^*$. Wir können daher schreiben:

$$\begin{aligned} & P\left(-A^* < \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} < A^*\right) \\ &= \int_{-A^*}^{A^*} \frac{1}{\sqrt{2\pi n\theta(1-\theta)}} \exp\left(-\frac{(x-n\theta)^2}{2n\theta(1-\theta)}\right) \mathrm{d}x \\ &= \Phi(A^*) - \Phi(-A^*) \;=\; 2\Phi(A^*) - 1 \;=\; 1-\alpha, \end{aligned}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Diese Funktion läßt sich zwar nicht analytisch berechnen, ist aber tabelliert, so daß man zu einem gegebenen Wert $\Phi(x)$ den zugehörigen Wert x bestimmen kann. Damit brauchen wir nur noch aus dem obigen Ausdruck einen Ausdruck $P(A < \theta < B)$ ableiten. Dies geschieht wie folgt:

$$\begin{aligned} & -A^* < \frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} < A^* \\ \Rightarrow\quad & |X-n\theta| < A^*\sqrt{n\theta(1-\theta)} \\ \Rightarrow\quad & (X-n\theta)^2 < (A^*)^2 n\theta(1-\theta) \\ \Rightarrow\quad & \theta^2(n(A^*)^2+n^2) - \theta(2nX + (A^*)^2 n) + X^2 < 0. \end{aligned}$$

Aus der so erhaltenen quadratischen Gleichung lassen sich die Werte von A und B berechnen zu

$$A/B \;=\; \frac{1}{n+(A^*)^2}\left(X + \frac{(A^*)^2}{2} \mp A^*\sqrt{\frac{X(n-X)}{n} + \frac{(A^*)^2}{4}}\right),$$

mit $\Phi(A^*) = 1 - \frac{\alpha}{2}$, womit das gesuchte Konfidenzintervall gefunden ist.

2.4.3 Hypothesentest

Ein Hypothesentest ist ein statistisches Verfahren, mit dem eine Entscheidung zwischen zwei einander entgegengesetzten Hypothesen über den datenerzeugenden Prozeß getroffen wird. Die Hypothesen können sich auf den Wert eines Parameters *(Parametertest)*, auf eine Verteilungsannahme *(Anpassungstest)* oder auf die Abhängigkeit bzw. Unabhängigkeit zweier Größen beziehen *(Abhängigkeitstest)*. Eine der beiden Hypothesen wird bevorzugt, d.h., die Entscheidung fällt im Zweifelsfalle zu ihren Gunsten aus. Diese bevorzugte Hypothese nennt man die **Nullhypothese** H_0, die andere heißt die **Alternativhypothese** H_a. Nur bei ausreichend starker Evidenz gegen die Nullhypothese wird die Alternativhypothese angenommen (und damit die Nullhypothese verworfen). (Man sagt auch, die Nullhypothese erhalte den „Vorteil des Zweifels“ — engl.: “benefit of the doubt”).[18]

Die Testentscheidung wird auf der Grundlage einer **Teststatistik** getroffen, also einer Funktion der Stichprobenwerte des gegebenen Datensatzes. Die Nullhypothese wird abgelehnt, wenn der Wert der Teststatistik im sogenannten **kritischen Bereich** C (critical region) liegt. Die Entwicklung eines statistischen Tests besteht darin, zu einer Verteilungsannahme und einem Parameter eine geeignete Teststatistik zu wählen und dann zu einem vorzugebenden Signifikanzniveau (siehe nächster Abschnitt) den zugehörigen kritischen Bereich C zu bestimmen (siehe folgende Abschnitte).

2.4.3.1 Fehlerarten und Signifikanzniveau

Da die Daten, auf die sich die Testentscheidung stützt, Ergebnis eines Zufallsprozesses sind, können wir natürlich nicht sicher sein, daß die Entscheidung, die wir mit einem Hypothesentest treffen, richtig ist. Wir können auch eine Fehlentscheidung treffen, und zwar auf eine von zwei Arten:

Fehler 1. Art: Die Nullhypothese H_0 wird, obwohl richtig, verworfen.
Fehler 2. Art: Die Nullhypothese H_0 wird, obwohl falsch, angenommen.

Fehler 1. Art gelten als schwerwiegender, da ja die Nullhypothese den Vorteil des Zweifels besitzt, also nicht so leicht verworfen wird wie die Alternativhypothese. Wird die Nullhypothese trotzdem verworfen, obwohl sie richtig ist, haben wir einen schweren Fehler gemacht. Man versucht daher in einem

[18] Man könnte auch sagen: Es wird ein Gerichtsverfahren gegen die Nullhypothese geführt, wobei die Daten (Stichprobe) als Indizien fungieren. Im Zweifelsfalle wird für den Angeklagten (die Nullhypothese) entschieden. Nur wenn die Beweise belastend genug sind, wird der Angeklagte verurteilt (die Nullhypothese verworfen).

Hypothesentest die Wahrscheinlichkeit für einen Fehler 1. Art auf einen Maximalwert α zu begrenzen. Dieser Maximalwert α heißt das **Signifikanzniveau** des Hypothesentestes. Es ist vom Anwender zu wählen. Typische Werte für das Signifikanzniveau sind 10%, 5% oder 1%.

2.4.3.2 Parametertest

Bei einem Parametertest machen die gegenübergestellten Hypothesen Aussagen über den Wert eines oder mehrerer Parameter. So kann etwa die Nullhypothese die Aussage sein, daß der wahre Wert eines Parameters θ mindestens (oder höchstens) θ_0 beträgt:

$$H_0: \quad \theta \geq \theta_0, \qquad\qquad H_a: \quad \theta < \theta_0.$$

Man spricht in diesem Falle von einem **einseitigen Test**. Bei einem **zweiseitigen Test** besteht die Nullhypothese dagegen darin, daß der wahre Wert eines Parameters innerhalb eines bestimmten Intervalles liegt oder gleich einem bestimmten Wert ist. Weitere Formen von Parametertests vergleichen etwa die Parameter der Verteilungen, die zwei verschiedenen Stichproben zugrundeliegen. Wir betrachten hier nur den einseitigen Test als Beispiel.

Bei einem einseitigen Test wie dem oben beschriebenen wählt man gewöhnlich einen Punktschätzer T für den Parameter θ als Teststatistik. In diesem Fall werden wir die Hypothese H_0 nur dann ablehnen, wenn der Wert des Punktschätzers T einen Wert c, den **kritischen Wert** (critical value), nicht übersteigt. Die kritische Region ist folglich $C = (-\infty, c]$. Es ist klar, daß der Wert c links von θ_0 liegen muß, denn wenn schon der Wert des Punktschätzers T θ_0 übersteigt, werden wir H_0 kaum sinnvoll ablehnen können. Aber auch ein Wert, der nur wenig unterhalb von θ_0 liegt, wird nicht ausreichen, um die Wahrscheinlichkeit eines Fehler 1. Art (die Hypothese H_0 wird, obwohl richtig, verworfen) hinreichend klein zu machen. c wird daher ein Stück weit links von θ_0 liegen müssen. Formal wird der kritische Wert c so bestimmt: Wir betrachten

$$\beta(\theta) = P_\theta(H_0 \text{ wird verworfen}) = P_\theta(T \in C),$$

was sich für den hier betrachteten einseitigen Test noch zu $\beta(\theta) = P(T \leq c)$ vereinfachen läßt. $\beta(\theta)$ nennt man auch die **Stärke** (power) des Testes. Sie beschreibt die Wahrscheinlichkeit einer Ablehnung von H_0 in Abhängigkeit vom Wert des Parameters θ. Ihr Wert soll für alle Werte θ, die die Nullhypothese erfüllen, kleiner als das Signifikanzniveau α sein. Denn wenn die Nullhypothese wahr ist, wollen wir sie ja nur mit einer Wahrscheinlichkeit

von höchstens α ablehnen, um nur mit dieser Wahrscheinlichkeit einen Fehler 1. Art zu machen. Also muß gelten

$$\max_{\theta:\theta \text{ erfüllt } H_0} \beta(\theta) \leq \alpha.$$

Für den hier betrachteten Test sieht man leicht, daß die Stärke $\beta(\theta)$ des Testes ihr Maximum gerade für $\theta = \theta_0$ annimmt. Denn je größer der wahre Wert von θ ist, um so unwahrscheinlicher ist es, daß die Teststatistik (der Punktschätzer T) einen Wert von höchstens c liefert. Also müssen wir den kleinsten Wert θ wählen, der die Nullhypothese $H_0 : \theta \geq \theta_0$ erfüllt. Der Ausdruck reduziert sich dadurch zu

$$\beta(\theta_0) \leq \alpha.$$

Es bleibt nun nur noch $\beta(\theta_0)$ aus der Verteilungsannahme und dem Punktschätzer T zu berechnen, um den Test zu vervollständigen.

2.4.3.3 Beispiel zum Parametertest

Als Beispiel für einen Parametertest betrachten wir einen einseitigen Test des Erwartungswertes μ einer Normalverteilung $N(\mu, \sigma^2)$ mit bekannter Varianz σ^2 [Berthold and Hand 2002]. D.h., wir betrachten die Hypothesen

$$H_0 : \quad \mu \geq \mu_0, \qquad\qquad H_a : \quad \mu < \mu_0.$$

Als Teststatistik verwenden wir den üblichen Punktschätzer für den Erwartungswert einer Normalverteilung, also

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

d.h. das arithmetische Mittel der Stichprobenwerte. (n ist der Umfang der Stichprobe.) Dieser Schätzer hat, wie man leicht nachrechnen kann, die Wahrscheinlichkeitsdichte

$$f_{\bar{X}}(x) = N\left(x; \mu, \tfrac{\sigma^2}{n}\right).$$

Folglich ist

$$\alpha = \beta(\mu_0) = P_{\mu_0}(\bar{X} \leq c) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z \leq \frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$$

mit der standardnormalverteilten Zufallsvariablen Z. (Um eine Aussage über eine solche Zufallsvariable zu erhalten, wurde gerade der dritte Umformungsschritt ausgeführt.) Also ist

$$\alpha = \Phi\left(\frac{c-\mu_0}{\sigma/\sqrt{n}}\right),$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist, die man in vielen Wahrscheinlichkeitstheorie- und Statistiklehrbüchern tabelliert findet. Aus einer solchen Tabelle liest man denjenigen Wert z_α ab, für den $\Phi(z_\alpha) = \alpha$. Der kritische Wert ist dann

$$c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Man beachte, daß z_α wegen des kleinen Wertes von α negativ ist, und c folglich, wie auch oben plausibel gemacht, kleiner ist als μ_0.

Um ein Zahlenbeispiel zu erhalten, wählen wir [Berthold and Hand 2002] $\mu_0 = 130$ und $\alpha = 0.05$. Außerdem sei $\sigma = 5.4$, $n = 125$ und $\bar{x} = 128$. Aus einer Tabelle der Standardnormalverteilung lesen wir $z_{0.05} \approx -1.645$ ab[19] und erhalten

$$c_{0.05} \approx 130 - 1.645\frac{5.4}{\sqrt{25}} \approx 128.22.$$

Da $\bar{x} = 128 < 128.22 = c$ wird die Hypothese H_0 abgelehnt. Hätten wir dagegen $\alpha = 0.01$ gewählt, so hätte sich mit $z_{0.01} \approx -2.326$

$$c_{0.01} \approx 130 - 2.326\frac{5.4}{\sqrt{25}} \approx 127.49$$

ergeben und H_0 wäre nicht abgelehnt worden.

Alternativ kann man auch das Signifikanzniveau unbestimmt lassen und stattdessen den Wert α angeben, ab dem die Hypothese H_0 abgelehnt wird. Dieser Wert α wird als **p-Wert** (p-value) bezeichnet. Für das obige Beispiel hat er den Wert

$$p = \Phi\left(\frac{128-130}{5.4/\sqrt{25}}\right) \approx 0.032.$$

D.h., die Hypothese H_0 wird bei einem Signifikanzniveau über 0.032 abgelehnt, bei einem Signifikanzniveau unter 0.032 nicht abgelehnt.

[19] Oder berechnen ihn mit Hilfe des auf der WWW-Seite zur Vorlesung zur Verfügung gestellten Programms `ndqtl.c` zur Berechnung der Quantile einer Normalverteilung.

2.4.3.4 Anpassungstest

Mit einem Anpassungstest wird geprüft, ob zwei Verteilungen, und zwar entweder zwei empirische Verteilungen oder eine empirische und eine theoretische, übereinstimmen. Oft werden Anpassungstests verwendet, um eine Verteilungsannahme, wie sie die Parameterschätzung voraussetzt, zu testen. Wir betrachten als Beispiel den χ^2-Anpassungstest für eine Polynomialverteilung: Gegeben sei ein eindimensionaler Datensatz vom Umfang n über den k Attributwerten $a_1, \ldots, a_k$ sowie eine Annahme über die Wahrscheinlichkeiten p_i^*, $1 \leq i \leq k$, mit denen die Attributewerte a_i auftreten. Wir wollen prüfen, ob die Hypothese auf den Datensatz paßt, d.h., ob die tatsächlichen Wahrscheinlichkeiten p_i mit den hypothetischen p_i^*, $1 \leq i \leq k$, übereinstimmen oder nicht. Dazu stellen wir die Hypothesen

$$H_0 : \forall i, 1 \leq i \leq k : p_i = p_i^* \qquad \text{und} \qquad H_a : \exists i, 1 \leq i \leq k : p_i \neq p_i^*$$

einander gegenüber. Eine geeignete Teststatistik ergibt sich aus dem folgenden Satz über polynomialverteilte Zufallsvariablen, die die Häufigkeit des Auftreten der verschiedenen Werte a_i in einer Stichprobe beschreiben.

Satz 2.20 *Sei $(X_1, \ldots, X_k)$ eine k-dimensionale, mit den Parametern $p_1, \ldots, p_k$ und n polynomialverteilte Zufallsvariable. Dann ist die Zufallsvariable*

$$Y = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

näherungsweise χ^2-verteilt mit $k-1$ Freiheitsgraden. (Damit diese Näherung hinreichend gut ist, sollte $\forall i, 1 \leq i \leq k : np_i \geq 5$ gelten. Dies läßt sich durch Zusammenfassung von Attributwerten/Zufallsvariablen stets erreichen.)

In der Berechnungsvorschrift für die Zufallsvariable Y werden die Werte der Zufallsvariablen X_i mit ihren Erwartungswerten np_i verglichen, die Abweichungen quadriert (unter anderem, damit sich nicht positive und negative Abweichungen gegeneinander wegheben) und gewichtet aufsummiert, wobei eine Abweichung um so stärker gewichtet wird, je kleiner der Erwartungswert ist. Da Y χ^2-verteilt ist, sind große Werte unwahrscheinlich.

Die Zahl der Freiheitsgrade ergibt sich übrigens aus der Zahl der freien Parameter der Verteilung. n ist kein freier Parameter, da er sich aus dem Umfang des Zufallsexperimentes ergibt. Von den k Parametern $p_1, \ldots, p_k$ können nur $k-1$ frei gewählt werden, da ja $\sum_{i=1}^{k} p_i = 1$ gelten muß. Von den insgesamt $k+1$ Parametern der Polynomialverteilung bleiben daher nur $k-1$ übrig, die die Zahl der Freiheitsgrade bestimmen.

Indem man die tatsächlichen Wahrscheinlichkeiten p_i durch die hypothetischen p_i^* ersetzt und für die Zufallsvariablen X_i ihre Instanziierung in einer Stichprobe (absolute Häufigkeit des Auftretens von a_i in der Stichprobe) ersetzt, erhält man eine Teststatistik für den Anpassungstest, nämlich

$$y = \sum_{i=1}^{k} \frac{(x_i - np_i^*)^2}{np_i^*}$$

Trifft die Nullhypothese H_0 zu, stimmen also alle hypothetischen Wahrscheinlichkeiten mit den tatsächlichen überein, so ist es unwahrscheinlich, daß y einen großen Wert annimmt, da y dann eine Instanziierung der Zufallsvariable Y ist, die χ^2-verteilt ist. Man wird daher die Nullhypothese H_0 ablehnen, wenn der Wert von y einen vom gewählten Signifikanzniveau abhängenden kritischen Wert c überschreitet. Die kritische Region ist folglich $C = [c, \infty)$. Der kritische Wert c wird aus der χ^2-Verteilung mit $k - 1$ Freiheitsgraden bestimmt, und zwar als derjenige Wert, für den $P(Y > c) = \alpha$ bzw. äquivalent $P(Y \leq c) = F_Y(c) = 1 - \alpha$ gilt, wobei F_Y die Verteilungsfunktion der χ^2-Verteilung mit $k - 1$ Freiheitsgraden ist.

Man beachte, daß man in der Praxis ggf. die Zahl k der Attributewerte durch Zusammenfassen von Attributwerten verringern muß, um sichzustellen, daß $\forall i, 1 \leq i \leq k : np_i \geq 5$ gilt, da sonst die Näherung durch die χ^2-Verteilung nicht gut genug ist.

Für stetige Verteilungen kann man den Wertebereich in nicht überlappende Intervalle einteilen, die hypothetische Verteilung durch eine Treppenfunktion annähern (konstanter Funktionswert je Intervall) und dann den Anpassungstest für eine Polynomialverteilung durchführen, wobei jedes Intervall einen Experimentausgang darstellt. In diesem Fall kann man entweder mit einer vollständig bestimmten hypothetischen Verteilung arbeiten (alle Parameter vorgegeben, Test ob eine bestimmte Verteilung auf die Daten paßt) oder die Parameter ganz oder teilweise aus den Daten schätzen (Test, ob eine Verteilung eines bestimmten Typs auf die Daten paßt). Im letzteren Fall ist jedoch die Zahl der Freiheitsgrade der χ^2-Verteilung um 1 für jeden aus den Daten geschätzen Parameter zu verringern. Das ist auch plausibel, denn mit aus den Daten geschätzten Parametern sollte natürlich die Stärke des Testes sinken und genau diesen Effekt hat die Verringerung der Zahl der Freiheitsgrade.

Als Alternative zur Anwendung des χ^2-Anpassungstestes auf eine Intervalleinteilung steht für stetige Verteilungen der **Kolmogorow-Smirnow-Test** zur Verfügung, der ohne Intervalleinteilung auskommt, sondern direkt die empirische mit der hypothetischen Verteilungsfunktion vergleicht.

2.4.3.5 Beispiel zum Anpassungstest

Von einem Würfel werde vermutet, daß er nicht fair sei, daß also die verschiedenen Augenzahlen nicht mit gleicher Wahrscheinlichkeit auftreten. Um diese Hypothese zu prüfen, werfen wir den Würfel 30-mal und zählen, wie häufig die verschiedenen Augenzahlen auftreten. Das Ergebnis sei:

$$x_1 = 2,\ x_2 = 4,\ x_3 = 3,\ x_4 = 5,\ x_5 = 3,\ x_6 = 13,$$

d.h., die Augenzahl 1 wurde zweimal geworfen, die Augenzahl 2 viermal etc. Dann stellen wir die Hypothesen

$$H_0 : \forall i, 1 \leq i \leq 6 : p_i = \frac{1}{6} \qquad \text{und} \qquad H_a : \exists i, 1 \leq i \leq 6 : p_i \neq \frac{1}{6}$$

einander gegenüber. Da wegen $n = 30$ gilt, daß $\forall i : np_i = 30\frac{1}{6} = 5$, sind die Voraussetzungen des Satzes 2.20 erfüllt und daher ist die χ^2-Verteilung mit 5 Freiheitsgraden eine gute Näherung der Verteilung der Zufallsvariable Y. Wir berechnen die Teststatistik

$$y = \sum_{i=1}^{6} \frac{(x_i - 30 \cdot \frac{1}{6})^2}{30 \cdot \frac{1}{6}} = \frac{1}{5} \sum_{i=1}^{6} (x_i - 5)^2 = \frac{67}{5} = 13.4.$$

Für ein Signifikanzniveau von $\alpha_1 = 0.05$ (5% Wahrscheinlichkeit für einen Fehler 1. Art) beträgt der kritische Wert $c \approx 11.07$, da für eine χ^2-verteilte Zufallsvariable Y mit 5 Freiheitsgraden gilt, daß

$$P(Y \leq 11.07) = F_Y(11.07) = 0.95 = 1 - \alpha_1,$$

wie man z.B. aus Tabellen der χ^2-Verteilung ablesen kann. Da $13.4 > 11.07$, läßt sich die Nullhypothese, der Würfel sei fair, auf einem Signifikanzniveau von $\alpha_1 = 0.05$ ablehnen. Sie läßt sich jedoch nicht auf einem Signifikanzniveau von $\alpha_2 = 0.01$ ablehnen, da

$$P(Y \leq 15.09) = F_Y(15.09) = 0.99 = 1 - \alpha_2$$

und $13.4 < 15.09$. Der p-Wert beträgt

$$p = 1 - F_Y(13.4) \approx 1 - 0.9801 = 0.0199.$$

D.h., für ein Signifikanzniveau über 0.0199 wird die Nullhypothese H_0 abgelehnt, für ein Signifikanzniveau unter 0.0199 dagegen angenommen.

2.4.3.6 Abhängigkeitstest

Mit einem Abhängigkeitstest wird geprüft, ob zwei Größen abhängig sind. Im Prinzip läßt sich aus jedem Anpassungstest leicht ein Abhängigkeitstest ableiten: Man vergleicht die empirische gemeinsame Verteilung zweier Größen mit einer hypothetischen unabhängigen Verteilung, die aus den Randverteilungen berechnet wird. Dabei werden die Randverteilungen gewöhnlich aus den Daten geschätzt.

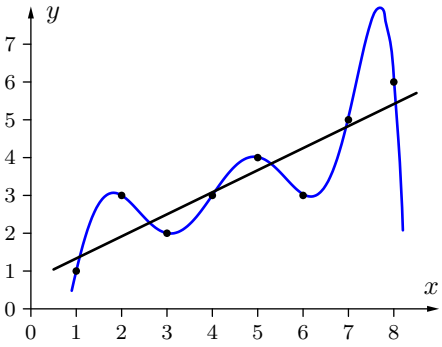
Wir betrachten als Beispiel den χ^2-Abhängigkeitstest für zwei nominale Größen, der sich aus dem χ^2-Anpassungstest ableitet. Seien X_{ij}, $1 \le i \le k_1$, $1 \le j \le k_2$, Zufallsvariablen, die die absolute Häufigkeit des gemeinsamen Auftretens der Werte a_i und b_j zweier Attribute A bzw. B beschreiben. Weiter seien $X_{i.} = \sum_{j=1}^{k_2} X_{ij}$ und $X_{.j} = \sum_{i=1}^{k_1} X_{ij}$ die Randhäufigkeiten (absolute Häufigkeiten der Attributwerte a_i und b_j). Dann berechnet man als Teststatistik aus den Instanziierungen x_{ij}, $x_{i.}$ und $x_{.j}$ dieser Zufallsvariablen, die man aus einer Stichprobe vom Umfang n auszählt, bzw. den aus ihnen geschätzten Verbundwahrscheinlichkeiten $p_{ij} = \frac{x_{ij}}{n}$ und Randwahrscheinlichkeiten $p_{i.} = \frac{x_{i.}}{n}$ und $p_{.j} = \frac{x_{.j}}{n}$, die Teststatistik

$$y = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(x_{ij} - \frac{1}{n} x_{i.} x_{.j})^2}{\frac{1}{n} x_{i.} x_{.j}} = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n \frac{(p_{ij} - p_{i.} p_{.j})^2}{p_{i.} p_{.j}}.$$

Der kritische Wert c wird mit Hilfe des Signifikanzniveaus aus einer χ^2-Verteilung mit $(k_1 - 1)(k_2 - 1)$ Freiheitsgraden berechnet, wobei die Zahl der Freiheitsgrade wie folgt zustande kommt: Für die $k_1 \cdot k_2$ Wahrscheinlichkeiten p_{ij}, $1 \le i \le k_1$, $1 \le j \le k_2$, für das Auftreten der verschiedenen Kombinationen von a_i und b_j muß $\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} p_{ij} = 1$ gelten. Bleiben noch $k_1 \cdot k_2 - 1$ freie Parameter. Aus den Daten werden die k_1 Wahrscheinlichkeiten $p_{i.}$ und die k_2 Wahrscheinlichkeitem $p_{.j}$ geschätzt. Für diese muß allerdings wieder $\sum_{j=1}^{k_2} p_{i.} = 1$ und $\sum_{i=1}^{k_1} p_{.j} = 1$ gelten, so daß sich die Zahl der Freiheitsgrade nur um $(k_1 - 1) + (k_2 - 1)$ verringert. Insgesamt haben wir $k_1 k_2 - 1 - (k_1 - 1) - (k_2 - 1) = (k_1 - 1)(k_2 - 1)$ Freiheitsgrade.

2.4.4 Modellauswahl

Eine wesentliche Aufgabe der beurteilenden Statistik ist, ein geeignetes Modell für die gegebenen Daten auszuwählen, gewöhnlich aus einer vorgegebenen Menge in Frage kommender Modelle. Bei dieser Wahl müssen Modellkomplexität und Passung des Modells auf die Daten gegeneinander abgewogen werden. Denn das komplexere Modell wird i.a. besser auf die Daten



schwarze Linie:
Ausgleichsgerade
(2 Parameter)

blaue Line:
Ausgleichspolynom 7. Grades
(8 Parameter)

Abbildung 2.12: Modelle verschiedener Komplexität für die gleichen Daten.

passen (geringere Abweichung der Datenpunkte, höhere Wahrscheinlichkeit der Stichprobe), aber es ist die Frage, ob die Passung so viel besser ist, daß sie die höhere Modellkomplexität rechtfertigt. Insbesondere ist zu berücksichtigen, daß ein bestimmtes Modell für eine Stichprobe zwar vielleicht die Daten dieser Stichprobe gut beschreibt, aber den datenerzeugenden Prozeß nur schlecht erfaßt, weil es zu sehr die zufälligen Besonderheiten der Stichprobe modelliert. In einem solchen Fall wird das Modell zur Vorhersage, wozu statistische Modelle ja oft benutzt werden, wenig taugen, denn die zufälligen Besonderheiten der Stichprobe werden in der Zukunft, für neue Daten, kaum wieder genau so auftreten.

Zur Veranschaulichung betrachten wir den folgenden, aus 8 Datenpunkten bestehenden zweidimensionalen Datensatz:

x	1	2	3	4	5	6	7	8
y	1	3	2	3	4	3	5	6

Mit Hilfe der Regression, die in Kapitel 3 genauer besprochen wird, können wir für diesen Datensatz eine Ausgleichsgerade oder ein Ausgleichspolynom bestimmen. Wenn wir den Grad des Polynoms zu 7 wählen, erhalten wir ein perfektes Modell, das genau die Datenpunkte trifft, während es natürlich Abweichungen von der Ausgleichsgerade gibt (siehe Abbildung 2.12). Trotzdem wird man die Regressiongerade als das bessere Modell ansehen, da es besser für Vorhersagen geeignet ist: So werden wir kaum annehmen, das das Ausgleichpolynom für x-Werte zwischen 7 und 8 oder gar für x-Werte größer als 8 eine sinnvolle Extrapolation erlaubt, während die Ausgleichsgerade hier sehr plausible Werte liefert (siehe Abbildung 2.12).

2.4.4.1 Informationskriterien

Eine beliebte Methode zur statistischen Modellauswahl sind sogenannte **Informationskriterien** (information criteria). Sie werden gewöhnlich definiert als der Logarithmus der Wahrscheinlichkeit der Daten gegeben ein betrachtetes Modell (auch als **Log-Likelihood** der Daten bezeichnet, vergleiche die Maximum-Likelihood-Schätzung) zuzüglich eines Terms, der von der Zahl der Parameter des Modells abhängt. Auf diese Weise berücksichtigt ein Informationskriterium sowohl die statistische Güte der Passung des Modells (durch die Wahrscheinlichkeit der Daten) als auch die Anzahl der Parameter, die geschätzt werden müssen, um diese Passungsgüte zu erzielen, indem ein Erhöhen der Zahl der Parameter mit einer Art „Strafe“ belegt wird [Everitt 1998]. Allgemein werden Informationskriterien so definiert:

$$\mathrm{IC}_\kappa(M \mid D) \;=\; -2 \ln P(D \mid M) + \kappa |\Theta(M)|,$$

wobei D der gegebene Datensatz, M das betrachtete Model, $\Theta(M)$ die Menge der Parameter des Modells M, und κ ein Parameter sind. $P(D \mid M)$ ist die Wahrscheinlichkeit der Daten, wie sie aus dem Modell berechnet werden kann (Likelihood der Daten). Ein Modell ist um so besser, je kleiner der Wert des Informationskriteriums ist.

Es sollte klar sein, daß eine Wahl von $\kappa = 0$ einem reinen Maximum-Likelihood-Ansatz entspricht, da ja dann die Komplexität des Modells keinen Einfluß hat. Ein solcher Ansatz ist jedoch nicht sehr sinnvoll, da meist das komplexeste Modell am besten auf die Daten paßt (vgl. obiges Beispiel).

Wichtige Spezialfälle der oben angegebenen allgemeinen Form sind das **Akaike-Informationskriterium** (AIC) [Akaike 1974] und das **Bayessche Informationskriterium** (BIC) [Schwarz 1978]. Ersteres ergibt sich für $\kappa = 2$. Es wird aus asymptotischen entscheidungstheoretischen Überlegungen hergeleitet. Für letzteres ist $\kappa = \ln n$, wobei n der Umfang des Datensatzes D ist, für den ein Modell gesucht wird. Es wird aus einem asymptotischen Bayesschen Argument abgeleitet [Heckerman 1998].

2.4.4.2 Minimale Beschreibungslänge

Um die Güte der Passung eines Modells und seine Komplexität kommensurabel, also mit dem gleichen Maßstab meß- und dadurch vergleichbar zu machen, kann man das Prinzip der **minimalen Beschreibungslänge** (minimum description length, MDL) oder **minimalen Nachrichtenlänge** (minimal message length, MML) verwenden: Man stellt sich vor, daß ein gegebener Datensatz von einem Sender zu einem Empfänger zu übertragen

ist. Da die Übertragung Kosten verursacht, soll die Länge der zu sendenden Nachricht so klein wie möglich sein. Nun ist klar, das man, wenn man ein gutes Modell der Daten hat, dieses ausnutzen kann, um die Daten kürzer zu kodieren, nämlich indem man sich auf das Modell bezieht und nur Abweichungen von diesem Modell überträgt oder indem man die durch das Modell spezifizierten Wahrscheinlichkeiten ausnutzt, um häufigen Attributewerten kurze, seltenen Attributwerten dagegen lange Kodes zuzuordnen (vergleiche auch Abschnitt **??** über den Informationsgewinn als Auswahlmaß für Entscheidungsbäume). Im Extremfall paßt das Modell exakt auf die Daten, so daß eigentlich gar keine Nachricht mehr gesendet werden muß.

Das Problem eines solchen Ansatzes besteht darin, daß der Empfänger das Modell, daß der Sender zur kürzeren Kodierung der Daten benutzt hat, nicht kennt und daher die erhaltene Nachricht nicht entschlüsseln kann. Daher muß das Modell, wenn eines verwendet wird, ebenfalls übertragen werden. Das verlängert allerdings die Nachricht, und zwar um so mehr, je komplexer das Modell ist, da komplexe Modelle natürlich aufwendiger zu beschreiben sind. Es kann dann günstiger sein, nicht das am besten passende Modell zur Kodierung der Daten zu verwenden, sondern ein einfacheres, weniger gut passendes Modell, das dafür kürzer beschrieben werden kann.

Allgemein wählt man beim Ansatz der minimalen Beschreibungslänge dasjenige Modell (aus einer Menge vorgegebener Modelle), das die Länge der zu sendenden Nachricht, also

Beschreibungslänge	=	**Länge der Modellbeschreibung**
	+	**Länge der Datenbeschreibung**

minimiert. Die Länge der Datenbeschreibung gibt an, wie gut das Modell auf die Daten paßt: Je besser das Modell paßt, um so kürzer lassen sich die Daten kodieren. Die Länge der Modellbeschreibung gibt an, wie komplex das Modell ist: Je komplexer das Modell, um so länger seine Kodierung. Da beide Größen Nachrichtenlängen (Anzahl Bits) sind, können sie miteinander verglichen werden und damit die Güte der Passung und die Modellkomplexität gegeneinander abgewogen werden.

2.4.4.3 Beispiel zur minimalen Beschreibungslänge

Als Beispiel zur Modellauswahl mit Hilfe der minimalen Beschreibunsglänge betrachten wir die Kodierung eines Datensatzes vom Umfang n über den k Attributwerten $a_1, \ldots, a_k$. Diesen Datensatz können wir erstens ohne ein Modell übertragen, wobei wir implizit annehmen, daß die Attributwerte alle

mit der gleichen Wahrscheinlichkeit auftreten. Dann brauchen wir

$$l_1 = n \log_2 k$$

Bits, um die Daten zu übertragen. Denn um einen einzelnen Wert zu übertragen, brauchen wir $\log_2 k$ Bits, weil man im Durchschnitt etwa $\log_2 k$ Ja-Nein-Fragen braucht, um mit einem der binären Suche analogen Verfahren einen Wert aus einer Menge von k Werten zu identifizieren: Die Menge wird in zwei etwa gleich große Teilmengen aufgeteilt und dann wird nach dem Enthaltensein in einer dieser Mengen gefragt. Unabhängig von der Antwort kann die Hälfte der Werte ausgeschlossen werden. Anschließend wird das Verfahren rekursiv auf die andere Teilmenge angewandt. Der Kode besteht einfach in der Folge der Antworten, die durch $ja \mathrel{\hat{=}} 1$ und $nein \mathrel{\hat{=}} 0$ kodiert werden (weitere Details zu Kodierung finden sich in Abschnitt **??**).

Alternativ können wir die Daten übertragen, indem wir als Modell eine Polynomialverteilung verwenden, deren Parameter $p_1, \ldots, p_k$ (bzw. äquivalent die absoluten Häufigkeiten $x_1, \ldots, x_k$ der Attributewerte im Datensatz, aus denen die $p_1, \ldots, p_k$ geschätzt werden) wir ausnutzen, um die Daten kürzer zu beschreiben. In diesem Fall brauchen wir

$$l_2 = \underbrace{\log_2 \frac{(n+k-1)!}{n!(k-1)!}}_{\text{Modellbeschreibung}} + \underbrace{\log_2 \frac{n!}{x_1! \ldots x_k!}}_{\text{Datenbeschreibung}}$$

Bits zur Übertragung. Der Term für die Modellbeschreibung kommt folgendermaßen zustande: Wir stellen uns ein Kodebuch vor, in dem alle möglichen Häufigkeitsverteilungen von n Beispielen (Kugeln) auf k Attributwerte (Kästen) verzeichnet sind, und zwar eine je Seite. Dieses Buch hat $s_1 = \frac{(n+k-1)!}{n!(k-1)!}$ Seiten, da wir die möglichen Anordnungen von $n + k - 1$ Objekten (n Kugeln und $k - 1$ Kastenwände) aufführen müssen, von denen n (die Kugeln) bzw. $k - 1$ (die Kastenwände) ununterscheidbar sind. Zur Übertragung des Modells übermitteln wir einfach die Seitennummer, die (auf die gleiche Weise wie oben) wir in $\log_2 s_1$ Bits kodieren können.

Auf ähnliche Weise werden die Daten beschrieben. Wir kennen bereits aus der Modellbeschreibung die Häufigkeiten x_i der verschiedenen Attributwerte a_i, $1 \leq i \leq k$. Wieder stellen wir uns ein Kodebuch vor, in dem alle möglichen Anordnungen von n Objekten verzeichnet sind, eine je Seite, wobei Gruppen von $x_1, \ldots, x_k$ Objekten ununterscheidbar sind. Dieses Buch hat, wie die Kombinatorik lehrt, $s_2 = \frac{n!}{x_1! \ldots x_k!}$ Seiten. Wie das Modell übertragen wir auch die Daten durch Übermittlung der Seitennummer, wofür wir $\log_2 s_2$ Bits benötigen.

Ist nun die Beschreibungslänge l_2 kleiner als die Beschreibungslänge l_1, was bei hinreichend großen Abweichungen von einer Gleichwahrscheinlichkeit aller Attributwerte und einem hinreichend großen Datensatz der Fall sein wird, so wird man das Polynomialverteilungsmodell der modellosen Annahme gleicher Wahrscheinlichkeiten vorziehen. Ist l_1 dagegen kleiner als l_2, so ist die Polynomialverteilung ein zu komplexes Modell, dessen Verwendung durch die Daten nicht gerechtfertigt wird.

Man beachte, daß man auf diese Weise auch einen Hypothesentest (ohne Signifikanzniveau) durchführen kann, nämlich, ob hinreichende Gründe vorliegen, um die Hypothese einer Gleichwahrscheinlichkeit aller Attributwerte abzulehnen. Dies ist der Fall, wenn $l_2 < l_1$.

Kapitel 3

Regression

In diesem Kapitel betrachten wir die in der Analysis und Statistik wohlbekannte **Methode der kleinsten Quadrate**, auch **Regression** genannt, zur Bestimmung von Ausgleichsgeraden (Regressionsgeraden) und allgemein Ausgleichpolynomen. Die Darstellung folgt im wesentlichen [Heuser 1988] (außer multilineare und nicht-polynomiale Regression und Lösen von Zwei-Klassen-Problemen mit Hilfe der Regression).

(Physikalische) Meßdaten zeigen selten exakt den gesetzmäßigen Zusammenhang der gemessenen Größen, da sie unweigerlich mit Fehlern behaftet sind. Will man den Zusammenhang der gemessenen Größen dennoch (wenigstens näherungsweise) bestimmen, so steht man vor der Aufgabe, eine Funktion zu finden, die sich den Meßdaten möglichst gut anpaßt, so daß die Meßfehler „ausgeglichen“ werden. Natürlich sollte dazu bereits eine Hypothese über die Art des Zusammenhangs vorliegen, um eine Funktionenklasse wählen und dadurch das Problem auf die Bestimmung der Parameter einer Funktion eines bestimmten Typs reduzieren zu können.

3.1 Lineare Regression

Erwartet man z.B. bei zwei Größen x und y einen linearen Zusammenhang (z.B. weil ein Diagramm der Meßpunkte einen solchen vermuten läßt oder weil man einen betragsmäßig großen Korrelationskoeffizienten berechnet hat), so muß man die Parameter a und b der Gerade $y = g(x) = a + bx$ bestimmen. Wegen der unvermeidlichen Meßfehler wird es jedoch i.a. nicht möglich sein, eine Gerade zu finden, so daß alle gegebenen n Meßpunkte

(x_i, y_i), $1 \leq i \leq n$, genau auf dieser Geraden liegen. Vielmehr wird man versuchen müssen, eine Gerade zu finden, von der die Meßpunkte möglichst wenig abweichen. Es ist daher plausibel, die Parameter a und b so zu bestimmen, daß die Abweichungsquadratsumme

$$F(a,b) \;=\; \sum_{i=1}^{n} (g(x_i) - y_i)^2 \;=\; \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

minimal wird. D.h., die aus der Geradengleichung berechneten y-Werte sollen (in der Summe) möglichst wenig von den gemessenen abweichen. Die Gründe für die Verwendung des Abweichungsquadrates sind i.w. die folgenden: Erstens ist die Fehlerfunktion durch die Verwendung des Quadrates überall (stetig) differenzierbar, während die Ableitung des Betrages, den man alternativ verwenden könnte, bei 0 nicht existiert/unstetig ist. Zweitens gewichtet das Quadrat große Abweichungen von der gewünschten Ausgabe stärker, so daß vereinzelte starke Abweichungen von den Meßdaten tendenziell vermieden werden.[1]

Aus der Analysis ist bekannt, daß eine notwendige Bedingung für ein Minimum der oben definierten Fehlerfunktion $F(a,b)$ ist, daß die partiellen Ableitungen dieser Funktion nach den Parametern a und b verschwinden, also

$$\begin{aligned} \frac{\partial F}{\partial a} &= \sum_{i=1}^{n} 2(a + bx_i - y_i) &&= 0 \qquad \text{und} \\ \frac{\partial F}{\partial b} &= \sum_{i=1}^{n} 2(a + bx_i - y_i)x_i &&= 0 \end{aligned}$$

gilt. Aus diesen beiden Gleichungen erhalten wir nach wenigen einfachen Umformungen die sogenannten **Normalgleichungen**

$$\begin{aligned} na \;+\; \left(\sum_{i=1}^{n} x_i\right) b \;&=\; \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a \;+\; \left(\sum_{i=1}^{n} x_i^2\right) b \;&=\; \sum_{i=1}^{n} x_i y_i, \end{aligned}$$

[1] Man beachte allerdings, daß dies auch ein Nachteil sein kann. Enthält der gegebene Datensatz „Ausreißer“ (das sind Meßwerte, die durch zufällig aufgetretene, unverhältnismäßig große Meßfehler sehr weit von dem tatsächlichen Wert abweichen), so wird die Lage der berechneten Ausgleichsgerade u.U. sehr stark von wenigen Meßpunkten (eben den Ausreißern) beeinflußt, was das Ergebnis unbrauchbar machen kann.

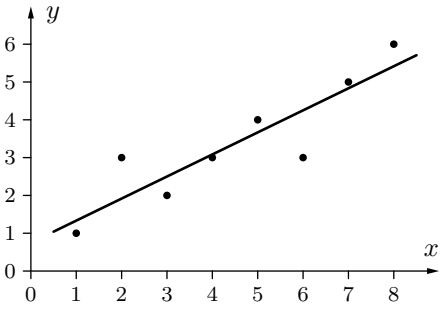


Abbildung 3.1: Beispieldaten und mit der Methode der kleinsten Quadrate berechnete Ausgleichsgerade.

also ein lineares Gleichungssystem mit zwei Gleichungen und zwei Unbekannten a und b. Man kann zeigen, daß dieses Gleichungssystem eine eindeutige Lösung besitzt, es sei denn, die x-Werte aller Meßpunkte sind identisch (d.h., es ist $x_1 = x_2 = \ldots = x_n$), und daß diese Lösung tatsächlich ein Minimum der Funktion F beschreibt [Heuser 1988]. Die auf diese Weise bestimmte Gerade $y = g(x) = a + bx$ nennt man die **Ausgleichsgerade** oder **Regressionsgerade** für den Datensatz $(x_1, y_1), \ldots, (x_n, y_n)$.

Man beachte, daß man die Bestimmung einer Regressionsgerade auch als **Maximum-Likelihood-Schätzung** (vergleiche Abschnitt 2.4.2.3) der Parameter des linearen Modells

$$Y = aX + b + \xi$$

sehen kann, wobei X eine beliebig verteilte und ξ eine normalverteilte Zufallsvariable mit Erwartungswert 0 und beliebiger Varianz ist: Diejenigen Parameter, für die die Summe der Abweichungsquadrate in y-Richtung von den Datenpunkten minimiert, maximieren die Wahrscheinlichkeit der Daten für diese Modellklasse.

Zur Veranschaulichung des Verfahrens betrachten wir ein einfaches Beispiel. Gegeben sei der aus acht Meßpunkten $(x_1, y_1), \ldots, (x_8, y_8)$ bestehende Datensatz, der in der folgenden Tabelle gezeigt ist [Heuser 1988]:

x	1	2	3	4	5	6	7	8
y	1	3	2	3	4	3	5	6

Um das System der Normalgleichungen aufzustellen, berechnen wir

$$\sum_{i=1}^{8} x_i = 36, \qquad \sum_{i=1}^{8} x_i^2 = 204, \qquad \sum_{i=1}^{8} y_i = 27, \qquad \sum_{i=1}^{8} x_i y_i = 146.$$

Damit erhalten wir das Gleichungssystem (Normalgleichungen)

$$\begin{aligned} 8a &+ 36b &= 27, \\ 36a &+ 204b &= 146, \end{aligned}$$

das die Lösung $a = \frac{3}{4}$ und $b = \frac{7}{12}$ besitzt. Die Ausgleichsgerade ist also

$$y = \frac{3}{4} + \frac{7}{12}x.$$

Diese Gerade ist zusammen mit den Datenpunkten, von denen wir ausgegangen sind, in Abbildung 3.1 dargestellt.

3.2 Polynomiale Regression

Das gerade betrachtete Verfahren ist natürlich nicht auf die Bestimmung von Ausgleichsgeraden beschränkt, sondern läßt sich mindestens auf Ausgleichspolynome erweitern. Man sucht dann nach einem Polynom

$$y = p(x) = a_0 + a_1 x + \ldots + a_m x^m$$

mit gegebenem, festem Grad m, das die n Meßpunkte $(x_1, y_1), \ldots, (x_n.y_n)$ möglichst gut annähert. In diesem Fall ist

$$F(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} (p(x_i) - y_i)^2 = \sum_{i=1}^{n} (a_0 + a_1 x_i + \ldots + a_m x_i^m - y_i)^2$$

zu minimieren. Notwendige Bedingung für ein Minimum ist wieder, daß die partiellen Ableitungen nach den Parametern a_0 bis a_m verschwinden, also

$$\frac{\partial F}{\partial a_0} = 0, \quad \frac{\partial F}{\partial a_1} = 0, \quad \ldots \quad , \frac{\partial F}{\partial a_m} = 0$$

gilt. So ergibt sich das System der Normalgleichungen [Heuser 1988]

$$\begin{aligned} n a_0 &+ \left(\sum_{i=1}^{n} x_i\right) a_1 &+ \ldots &+ \left(\sum_{i=1}^{n} x_i^m\right) a_m &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) a_0 &+ \left(\sum_{i=1}^{n} x_i^2\right) a_1 &+ \ldots &+ \left(\sum_{i=1}^{n} x_i^{m+1}\right) a_m &= \sum_{i=1}^{n} x_i y_i \\ &\vdots & & & \vdots \\ \left(\sum_{i=1}^{n} x_i^m\right) a_0 &+ \left(\sum_{i=1}^{n} x_i^{m+1}\right) a_1 &+ \ldots &+ \left(\sum_{i=1}^{n} x_i^{2m}\right) a_m &= \sum_{i=1}^{n} x_i^m y_i, \end{aligned}$$

aus dem sich die Parameter a_0 bis a_m mit den üblichen Methoden der linearen Algebra (z.B. Gaußsches Eliminationsverfahren, Cramersche Regel, Bildung der Inversen der Koeffizientenmatrix etc.) berechnen lassen. Das so bestimmte Polynom $p(x) = a_0 + a_1x + a_2x^2 + \ldots + a_mx^m$ heißt **Ausgleichspolynom** oder **Regressionspolynom** m-ter Ordnung für den Datensatz $(x_1, y_1), \ldots, (x_n, y_n)$.

Wie die lineare Regression kann auch die Bestimmung eines Ausgleichspolynoms als Maximum-Likelihood-Schätzung gedeutet werden, wenn man einen normalverteilten Fehlerterm ansetzt.

3.3 Multivariate Regression

Weiter läßt sich die Methode der kleinsten Quadrate nicht nur verwenden, um, wie bisher betrachtet, Ausgleichspolynome zu bestimmen, sondern kann auch für Funktionen mit mehr als einem Argument eingesetzt werden. In diesem Fall spricht man von **multipler** oder **multivariater Regression**. Wir untersuchen hier beispielhaft nur den Spezialfall der **multilinearen Regression**, wobei wir uns außerdem zunächst auf eine Funktion mit zwei Argumenten beschränken. D.h., wir betrachten, wie man zu einem gegebenen Datensatz $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ eine Ausgleichsfunktion der Form

$$z = f(x, y) = a + bx + cy$$

so bestimmen kann, daß die Summe der Abweichungsquadrate minimal wird. Die Ableitung der Normalgleichungen für diesen Fall ist zu der Ableitung für Ausgleichspolynome völlig analog. Wir müssen

$$F(a, b, c) \;=\; \sum_{i=1}^{n} (f(x_i, y_i) - z_i)^2 \;=\; \sum_{i=1}^{n} (a + bx_i + cy_i - z_i)^2$$

minimieren. Notwendige Bedingungen für ein Minimum sind

$$\begin{aligned}
\frac{\partial F}{\partial a} &= \sum_{i=1}^{n} 2(a + bx_i + cy_i - z_i) &&= 0, \\
\frac{\partial F}{\partial b} &= \sum_{i=1}^{n} 2(a + bx_i + cy_i - z_i)x_i &&= 0, \\
\frac{\partial F}{\partial c} &= \sum_{i=1}^{n} 2(a + bx_i + cy_i - z_i)y_i &&= 0.
\end{aligned}$$

Also erhalten wir das System der Normalgleichungen

$$\begin{aligned}
na &+ \left(\sum_{i=1}^{n} x_i\right) b &+ \left(\sum_{i=1}^{n} y_i\right) c &= \sum_{i=1}^{n} z_i \\
\left(\sum_{i=1}^{n} x_i\right) a &+ \left(\sum_{i=1}^{n} x_i^2\right) b &+ \left(\sum_{i=1}^{n} x_i y_i\right) c &= \sum_{i=1}^{n} z_i x_i \\
\left(\sum_{i=1}^{n} y_i\right) a &+ \left(\sum_{i=1}^{n} x_i y_i\right) b &+ \left(\sum_{i=1}^{n} y_i^2\right) c &= \sum_{i=1}^{n} z_i y_i
\end{aligned}$$

aus dem sich a, b und c leicht berechnen lassen.

Im allgemeinen Fall der multilinearen Regression (Funktion mit m Argumenten) ist ein Datensatz $((x_{11}, \ldots, x_{1m}, y_1), \ldots, (x_{n1}, \ldots, x_{nm}, y_n))$ (oder auch dargestellt als $((\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n))$ mit einem Eingabevektor $\vec{x}_i$ und der zugehörigen Ausgabe y_i, $1 \leq i \leq n$) gegeben, für den eine lineare Ausgleichsfunktion

$$y = f(x_1, \ldots, x_m) = a_0 + \sum_{k=1}^{m} a_k x_k$$

gesucht ist. Zur Ableitung der Normalgleichungen stellt man in diesem Fall das zu minimierende Funktional bequemer in Matrixform dar, nämlich als

$$F(\vec{a}) = (\mathbf{X}\vec{a} - \vec{y})^\top (\mathbf{X}\vec{a} - \vec{y}),$$

wobei

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nm} \end{pmatrix} \qquad \text{und} \qquad \vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

den Datensatz wiedergeben und $\vec{a} = (a_0, a_1, \ldots, a_m)^\top$ der Vektor der zu bestimmenden Koeffizienten ist.[2] (Man beachte, daß die Einsen in der ersten Spalte der Matrix $\mathbf{X}$ dem Koeffizienten a_0 zugehören.) Wieder ist eine notwendige Bedingung für ein Minimum, daß die partiellen Ableitungen nach den Koeffizienten a_k, $0 \leq k \leq m$, verschwinden, was wir mit Hilfe des Differentialoperators ∇ (gesprochen: nabla) schreiben können als

$$\nabla_{\vec{a}} F(\vec{a}) = \frac{\mathrm{d}}{\mathrm{d}\vec{a}} F(\vec{a}) = \left(\frac{\partial}{\partial a_0} F(\vec{a}), \frac{\partial}{\partial a_1} F(\vec{a}), \ldots, \frac{\partial}{\partial a_m} F(\vec{a}) \right) = \vec{0}.$$

[2] $\top$ bedeutet die Transponierung eines Vektors oder einer Matrix, also die Vertauschung von Zeilen und Spalten.

Die Ableitung läßt sich am leichtesten berechnen, wenn man sich klar macht (wie man durch elementweises Ausschreiben leicht nachrechnet), daß sich der Differentialoperator

$$\nabla_{\vec{a}} = \left(\frac{\partial}{\partial a_0}, \frac{\partial}{\partial a_1}, \ldots, \frac{\partial}{\partial a_m} \right)$$

formal wie ein Vektor verhält, der von links an die Summe der Fehlerquadrate „heranmultipliziert“ wird. Alternativ kann man die Ableitung komponentenweise ausschreiben. Wir verwenden hier jedoch die erstere, wesentlich bequemere Methode und erhalten

$$\begin{aligned}
\vec{0} &= \nabla_{\vec{a}} \, (\mathbf{X}\vec{a} - \vec{y})^\top (\mathbf{X}\vec{a} - \vec{y}) \\
&= (\nabla_{\vec{a}} \, (\mathbf{X}\vec{a} - \vec{y}))^\top (\mathbf{X}\vec{a} - \vec{y}) + ((\mathbf{X}\vec{a} - \vec{y})^\top (\nabla_{\vec{a}} \, (\mathbf{X}\vec{a} - \vec{y})))^\top \\
&= (\nabla_{\vec{a}} \, (\mathbf{X}\vec{a} - \vec{y}))^\top (\mathbf{X}\vec{a} - \vec{y}) + (\nabla_{\vec{a}} \, (\mathbf{X}\vec{a} - \vec{y}))^\top (\mathbf{X}\vec{a} - \vec{y}) \\
&= 2\mathbf{X}^\top (\mathbf{X}\vec{a} - \vec{y}) \\
&= 2\mathbf{X}^\top \mathbf{X}\vec{a} - 2\mathbf{X}^\top \vec{y},
\end{aligned}$$

woraus sich unmittelbar das System

$$\mathbf{X}^\top \mathbf{X}\vec{a} = \mathbf{X}^\top \vec{y}$$

der Normalgleichungen ergibt. Dieses System ist offenbar lösbar, wenn $\mathbf{X}^\top \mathbf{X}$ invertierbar ist. Dann gilt

$$\vec{a} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \vec{y}.$$

Den Ausdruck $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ nennt man auch die (Moore-Penrose-) **Pseudoinverse** der Matrix $\mathbf{X}$ [Albert 1972]. Mit ihr kann man unmittelbar die Lösung der Regressionsaufgabe angeben.

Es dürfte klar sein, daß sich die Methode der kleinsten Quadrate auch auf Polynome in mehreren Variablen erweitern läßt. Am einfachsten geht man dabei ebenfalls von der oben verwendeten Matrixdarstellung aus, wobei man in die Matrix $\mathbf{X}$ als Eingangsgrößen auch die zu verwendenden Potenzprodukte der unabhängigen Variablen einträgt. Die Ableitung der Normalgleichungen kann dann einfach übernommen werden.

Ein Programm zur multipolynomialen Regression, das zur schnellen Berechnung der benötigten Potenzprodukte eine auf Ideen der dynamischen Programmierung beruhende Methode benutzt, findet man unter

`http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html#regress`

3.4 Logistische Regression

Die Bestimmung eines Ausgleichspolynoms läßt sich in einigen Spezialfällen auch zur Bestimmung anderer Ausgleichsfunktionen verwenden, nämlich dann, wenn es gelingt, eine geeignete Transformation zu finden, durch die das Problem auf das Problem der Bestimmung eines Ausgleichspolynoms zurückgeführt wird. So lassen sich z.B. auch Ausgleichsfunktionen der Form

$$y = ax^b$$

durch die Bestimmung einer Ausgleichgeraden finden. Denn logarithmiert man diese Gleichung, so ergibt sich

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

Diese Gleichung können wir durch die Bestimmung einer Ausgleichsgeraden behandeln. Wir müssen lediglich die Datenpunkte (x_i, y_i) logarithmieren und mit den so transformierten Werten rechnen. Man beachte allerdings, daß bei einem solchen Vorgehen zwar die Fehlerquadratsumme im transformierten Raum (Koordinaten $x' = \ln x$ und $y' = \ln y$), aber damit nicht notwendig die Fehlerquadratsumme im Originalraum (Koordinaten x und y) minimiert wird. Dennoch führt der Ansatz meist zu sehr guten Ergebnissen.

Für die Praxis ist wichtig, daß es auch für die sogenannte **logistische Funktion**

$$y = \frac{Y}{1 + e^{a+bx}},$$

wobei Y, a und b Konstanten sind, eine Transformation gibt, mit der wir das Problem der Bestimmung einer Ausgleichsfunktion dieser Form auf die Bestimmung einer Ausgleichsgerade zurückführen können (sogenannte **logistische Regression**). Die logistische Funktion ist in vielen Anwendungen wichtig, da sie Wachstumsprozesse mit Größenbeschränkung beschreibt, also z.B. dem Wachstum einer Tierpopulation bei begrenztem Lebensraum oder den Absatz eines neuen Produktes bei endlichem Markt. Außerdem wird sie gern in künstlichen neuronalen Netzen (speziell mehrschichtigen Perzeptren, siehe Abschnitt **??**) als Aktivierungsfunktion verwendet.

Um die logistische Funktion zu „linearisieren“, bestimmen wir zunächst den Reziprokwert der logistischen Gleichung:

$$\frac{1}{y} = \frac{1 + e^{a+bx}}{Y}.$$

Folglich ist

$$\frac{Y-y}{y} = e^{a+bx}.$$

Durch Logarithmieren dieser Gleichung erhalten wir

$$\ln\left(\frac{Y-y}{y}\right) = a + bx.$$

Diese Gleichung können wir durch Bestimmen einer Ausgleichsgerade behandeln, wenn wir die y-Werte der Datenpunkte entsprechend der linken Seite dieser Gleichung transformieren. (Man beachte, daß dazu der Wert von Y bekannt sein muß, der i.w. eine Skalierung bewirkt.) Diese Transformation ist unter dem Namen **Logit-Transformation** bekannt. Sie entspricht einer Umkehrung der logistischen Funktion. Indem wir für die entsprechend transformierten Datenpunkte eine Ausgleichsgerade bestimmen, erhalten wir eine logistische Ausgleichskurve für die Originaldaten.

Man beachte wieder, daß bei diesem Vorgehen zwar die Fehlerquadratsumme im transformierten Raum (Koordinaten x und $z = \ln\left(\frac{Y-y}{y}\right)$), aber damit nicht notwendig die Fehlerquadratsumme im Originalraum (Koordinaten x und y) minimiert wird. Sollte man an den Parameterwerten interessiert sein, die tatsächlich die Fehlerquadratsumme im Originalraum minimieren, so kann man die oben beschriebene Lösung als Startpunkt benutzen und einen Gradientenabstieg für die Fehlerquadratsumme im Originalraum durchführen. Dies führt auf ein iteratives Verfahren, das dem Gradientenabstieg zum Training eines neuronalen Netzes, speziell eines mehrschichtigen Perzeptrons, entspricht (siehe Abschnitt **??**).

Zur Veranschaulichung des Vorgehens betrachten wir ein einfaches Beispiel. Gegeben sei der aus den fünf Punkten $(x_1, y_1), \ldots, (x_5, y_5)$ bestehende Datensatz, der in der folgenden Tabelle gezeigt ist:

x	1	2	3	4	5
y	0.4	1.0	3.0	5.0	5.6

Wir transformieren diese Daten mit

$$z = \ln\left(\frac{Y-y}{y}\right), \qquad Y = 6.$$

Die transformierten Datenpunkte sind (näherungsweise):

x	1	2	3	4	5
z	2.64	1.61	0.00	-1.61	-2.64

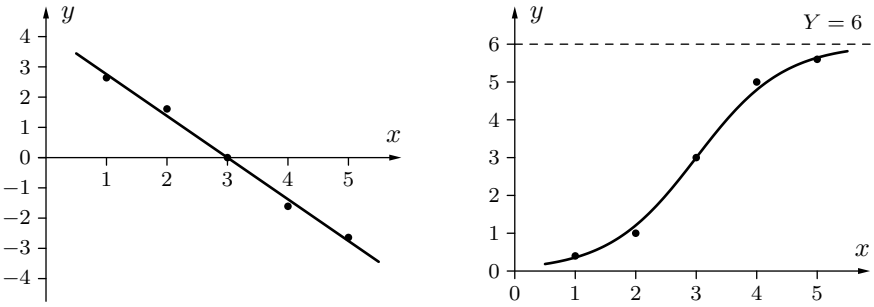


Abbildung 3.2: Transformierte Daten (links) und Originaldaten (rechts) sowie mit der Methode der kleinsten Quadrate berechnete Ausgleichsgerade (transformierte Daten) und zugehörige Ausgleichskurve (Originaldaten).

Um das System der Normalgleichungen aufzustellen, berechnen wir

$$\sum_{i=1}^{5} x_i = 15, \qquad \sum_{i=1}^{5} x_i^2 = 55, \qquad \sum_{i=1}^{5} z_i = 0, \qquad \sum_{i=1}^{5} x_i z_i \approx -13.775.$$

Damit erhalten wir das Gleichungssystem (Normalgleichungen)

$$\begin{aligned} 5a &+ 15b &&= 0, \\ 15a &+ 55b &&= -13.775, \end{aligned}$$

das die Lösung $a \approx 4.133$ und $b \approx -1.3775$ besitzt. Die Ausgleichsgerade für die transformierten Daten ist daher

$$z \approx 4.133 - 1.3775x$$

und die Ausgleichskurve für die Originaldaten folglich

$$y \approx \frac{6}{1 + e^{4.133 - 1.3775x}}.$$

Diese beiden Ausgleichsfunktionen sind zusammen mit den (transformierten bzw. Original-) Datenpunkten in Abbildung 3.2 dargestellt.

3.5 Zwei-Klassen-Probleme

Die logistische Regression, wie sie im vorangehenden Abschnitt beschrieben wurde, wird auch gern benutzt, um Zwei-Klassen-Probleme zu lösen. D.h.,

die Datenpunkte sind jeweils einer von zwei Klassen zugeordnet und es ist eine Funktion gesucht, die aus den Werten der übrigen Attribute die Klasse bestimmt — oder genauer: die Wahrscheinlichkeit für die Zugehörigkeit eines Datenpunktes zu einer der beiden Klassen. Dieser Ansatz ist besonders in Bankwesen zur Kreditwürdigkeitsprüfung beliebt.

Formal wird die zu lösende Aufgabe so beschrieben: Sei C ein Klassenattribut mit dem Wertebereich $\mathrm{dom}(C) = \{c_1, c_2\}$ und $\vec{X} = (X_1, \ldots, X_m)$ ein m-dimensionaler Zufallsvektor. Weiter sei

$$\begin{array}{lcll} P(C = c_1 \mid \vec{X} = \vec{x}) & = & p(\vec{x}) & \text{und folglich} \\ P(C = c_2 \mid \vec{X} = \vec{x}) & = & 1 - p(\vec{x}). & \end{array}$$

Schließlich sei ein Datensatz $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\}$ gegeben, dessen Elementen jeweils eine der beiden Klassen c_1 oder c_2 zugeordnet ist.

Gesucht ist eine einfache Beschreibung der Funktion $p(\vec{x})$, deren Parameter aus dem Datensatz $\mathcal{X}$ zu schätzen sind. Wir setzen die Funktion $p(\vec{x})$ als logistische Funktion an, also als

$$p(\vec{x}) \quad = \quad \frac{1}{1 + e^{a_0 + \vec{a}\vec{x}}} \quad = \quad \frac{1}{1 + \exp\left(a_0 + \sum_{i=1}^{r} a_i x_i\right)}.$$

Indem wir auf diese Funktion die im vorangehenden Abschnitt besprochene Logit-Transformation anwenden, erhalten wir

$$\ln\left(\frac{1 - p(\vec{x})}{p(\vec{x})}\right) \quad = \quad a_0 + \vec{a}\vec{x} \quad = \quad a_0 + \sum_{i=1}^{r} a_i x_i,$$

also eine multilineare Regressionsaufgabe, die wir auf die in Abschnitt 3.3 besprochene Weise lösen können.

Es bleibt damit nur noch zu klären, wie wir die Werte $p(\vec{x})$ bestimmen. Wenn der Datenraum klein genug ist, so daß für jeden möglichen Punkt (d.h. für jede mögliche Instanziierung der Zufallsvariablen $X_1, \ldots, X_m$) hinreichend viele Datenpunkte zur Verfügung stehen, dann können wir diese Klassenwahrscheinlichkeit einfach aus der relativen Häufigkeit der Klasse schätzen (zur Parameterschätzung siehe Abschnitt 2.4.2).

Ist dies nicht der Fall, so kann man eine sogenannte **Kernschätzung** verwenden, um die Wahrscheinlichkeiten an den Datenpunkten zu bestimmen. Die Idee einer solchen Schätzung besteht darin, eine Kernfunktion K zu definieren, die angibt, wie stark ein Datenpunkt den Schätzwert der

Wahrscheinlichkeit(sdichte) an einem benachbarten Punkt beeinflußt. Üblicherweise wird eine Gaußsche Funktion verwendet, also

$$K(\vec{x},\vec{y}) \quad = \quad \frac{1}{(2\pi\sigma^2)^{\frac{m}{2}}} \exp\left(-\frac{(\vec{x}-\vec{y})^\top(\vec{x}-\vec{y})}{2\sigma^2}\right),$$

wobei die Varianz σ^2 vom Anwender zu wählen ist. Mit Hilfe dieser Kernfunktion wird die Wahrscheinlichkeitsdichte an einem Punkt $\vec{x}$ aus einem Datensatz $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\}$ geschätzt als

$$\hat{f}(\vec{x}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} K(\vec{x},\vec{x}_i).$$

Bei einem Zwei-Klassen-Problem setzt man zur Schätzung der Klassenwahrscheinlichkeiten die Wahrscheinlichkeitsdichte für eine Klasse ins Verhältnis zur Gesamtwahrscheinlichkeitsdichte, benutzt also

$$\hat{p}(\vec{x}) \quad = \quad \frac{\sum_{i=1}^{n} c(\vec{x}_i)K(\vec{x},\vec{x}_i)}{\sum_{i=1}^{n} K(\vec{x},\vec{x}_i)},$$

wobei

$$c(\vec{x}_i) = \begin{cases} 1, & \text{falls } x_i \text{ zur Klasse } c_1 \text{ gehört,} \\ 0, & \text{falls } x_i \text{ zur Klasse } c_2 \text{ gehört.} \end{cases}$$

Als Ergebnis der Regression erhält man eine (mehrdimensionale) logistische Funktion, die die Wahrscheinlichkeit der einen Klasse beschreibt. Für diese Funktion ist nun noch ein Schwellenwert zu wählen, oberhalb dem die eine und unterhalb dem die andere Klasse vorhergesagt wird. Wird dieses Verfahren im Bankwesen zur Kreditwürdigkeitsprüfung eingesetzt, so bedeutet die eine Klasse, daß der Kredit gewährt, die andere dagegen, daß der Kreditantrag abgelehnt wird. Man wählt daher meist nicht nur einen Schwellenwert, sondern mehrere, denen unterschiedliche Konditionen (Zinssatz, Sicherheiten, Tilgungsrate etc.) zugeordnet sind.

Man beachte, daß mit dem Schwellenwert und der logistischen Funktion eine lineare Trennung des Eingaberaums beschrieben wird. Details hierzu findet man im Kapitel **??** über neuronale Netze, deren einfachste Form auf sehr ähnliche Weise Klassifikationsprobleme löst.

Literaturverzeichnis

[Agrawal *et al.* 1993] R. Agrawal, T. Imielienski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. *Proc. Conf. on Management of Data (Washington, DC, USA)*, 207–216. ACM Press, New York, NY, USA 1993

[Agrawal *et al.* 1996] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast Discovery of Association Rules. In: [Fayyad *et al.* 1996], 307–328

[Aha 1992] D.W. Aha. Tolerating Noisy, Irrelevant and Novel Attributes in Instance-based Learning Algorithms. *Int. Journal of Man-Machine Studies* 36(2):267–287. Academic Press, San Diego, CA, USA 1992

[Akaike 1974] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Trans. on Automatic Control* 19:716–723. IEEE Press, Piscataway, NJ, USA 1974

[Albert 1972] A. Albert. *Regression and the Moore-Penrose Pseudoinverse.* Academic Press, New York, NY, USA 1972

[Anderson 1995a] J.A. Anderson. *An Introduction to Neural Networks.* MIT Press, Cambridge, MA, USA 1995

[Anderson 1995b] J.R. Anderson. *Cognitive Psychologie and its Implications (4th edition).* Freeman, New York, NY, USA 1995

[Anderson and Rosenfeld 1988] J.A. Anderson and E. Rosenfeld. *Neurocomputing: Foundations of Research.* MIT Press, Cambridge, MA, USA 1988

[Azvine *et al.* 2000] B. Azvine and N. Azarmi and D. Nauck, eds. *Soft Computing and Intelligent Systems: Prospects, Tools and Applications.* LNAI 1804, Springer-Verlag, Berlin, Germany 2000

[Berthold and Hand 2002] M.R. Berthold and D. Hand, eds. *Intelligent Data Analysis: An Introduction* (2. Auflage). Springer, Berlin, Germany 2002

[Bezdek and Pal 1992] J.C. Bezdek and N. Pal. *Fuzzy Models for Pattern Recognition.* IEEE Press, New York, NY, USA 1992

[Bezdek *et al.* 1999] J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer, Dordrecht, Netherlands 1999

[Bock 1974] H.H. Bock. *Automatische Klassifikation (Cluster-Analyse).* Vandenhoek & Ruprecht, Göttingen, Germany 1974

[Boden 1990] M.A. Boden, ed. *The Philosophy of Artificial Intelligence.* Oxford University Press, Oxford, United Kingdom 1990

[Borgelt 1998] C. Borgelt. A Decision Tree Plug-In for DataEngine. *Proc. 2nd Int. Data Analysis Symposium.* MIT GmbH, Aachen, Germany 1998. Reprinted in: *Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT'98, Aachen, Germany)*, Vol. 2:1299–1303. Verlag Mainz, Aachen, Germany 1998

[Borgelt and Timm 2000] C. Borgelt and H. Timm. Advanced Fuzzy Clustering and Decision Tree Plug-Ins for DataEngine. In: [Azvine *et al.* 2000], 188–212.

[Bosch 1994] K. Bosch. *Elementare Einführung in die angewandte Statistik. (5. Auflage).* Vieweg, Braunschweig/Wiesbaden, Germany 1994

[Bosch 1987] K. Bosch. *Elementare Einführung in die Wahrscheinlichkeitsrechnung. (5. Auflage).* Vieweg, Braunschweig/Wiesbaden, Germany 1987

[Breiman *et al.* 1984] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, USA 1984

[Buffon 1733] G.-L.L. Buffon. *Mémoire sur le Jeu Franc-Carreau.* France 1733

[Chapman *et al.* 1999] P. Chapman, J. Clinton, T. Khabaza, T. Reinartz, and R. Wirth. *The CRISP-DM Process Model.* NCR, Denmark 1999. `http://www.ncr.dk/CRISP/`.

[Cheeseman *et al.* 1988] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. AutoClass: A Bayesian Classification System. *Proc. 5th Int. Workshop on Machine Learning*, 54–64. Morgan Kaufmann, San Mateo, CA, USA 1988

[Dasarathy 1990] B.V. Dasarathy. *Nearest Neighbor (NN) Norms: NN Pattern Classifcation Techniques.* IEEE Computer Science Press, Los Alamitos, CA, USA 1990

[Duda and Hart 1973] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* J. Wiley & Sons, New York, NY, USA 1973

[Everitt 1981] B.S. Everitt. *Cluster Analysis.* Heinemann, London, United Kingdom 1981

[Everitt 1998] B.S. Everitt. *The Cambridge Dictionary of Statistics.* Cambridge University Press, Cambridge, United Kingdom 1998

[Fayyad *et al.* 1996] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds. *Advances in Knowledge Discovery and Data Mining.* AAAI Press and MIT Press, Menlo Park and Cambridge, MA, USA 1996

[Feynman *et al.* 1963] R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics, Vol. 1: Mechanics, Radiation, and Heat.* Addison-Wesley, Reading, MA, USA 1963

[Fisher 1936] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179–188. Cambridge University Press, Cambridge, United Kingdom 1936

[Fisher 1987] D.H. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. *Machine Learning* 2:139-172. Kluwer, Dordrecht, Netherlands 1987

[Fredkin and Toffoli 1982] E. Fredkin and T. Toffoli. Conservative Logic. *Int. Journal of Theoretical Physics* 21(3/4):219–253. Plenum Press, New York, NY, USA 1982

[Gentsch 1999] P. Gentsch. *Data Mining Tools: Vergleich marktgängiger Tools.* WHU Koblenz, Germany 1999

[Good 1965] I.J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, MA, USA 1965

[Greiner 1989] W. Greiner. *Theoretische Physik: Mechanik I.* Verlag Harri Deutsch, Thun/Frankfurt a.M., Germany 1989

[Hanson and Bauer 1989] S.J. Hanson and M. Bauer. Conceptual Clustering, Categorization, and Polymorphy. *Machine Learning* 3:343-372. Kluwer, Dordrecht, Netherlands 1989

[Haykin 1994] S. Haykin. *Neural Networks — A Comprehensive Foundation.* Prentice Hall, Upper Saddle River, NJ, USA 1994

[Hebb 1949] D.O. Hebb. *The Organization of Behaviour.* J. Wiley & Sons, New York, NY, USA 1949
Chap. 4: "The First Stage of Perception: Growth of an Assembly" reprinted in [Anderson and Rosenfeld 1988], 45–56.

[Heckerman *et al.* 1995] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning* 20:197–243. Kluwer, Dordrecht, Netherlands 1995

[Heckerman 1998] D. Heckerman. *A Tutorial on Learning with Bayesian Networks.* In: [Jordan 1998], 301–354.

[Heuser 1988] H. Heuser. *Lehrbuch der Analysis, Teil 1+2.* Teubner, Stuttgart, Germany 1988

[Heuser 1989] H. Heuser. *Gewöhnliche Differentialgleichungen.* Teubner, Stuttgart, Germany 1989

[Hopfield 1982] J.J. Hopfield. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. of the National Academy of Sciences* 79:2554–2558. USA 1982

[Hopfield 1984] J.J. Hopfield. Neurons with Graded Response have Collective Computational Properties like those of Two-state Neurons. *Proc. of the National Academy of Sciences* 81:3088–3092. USA 1984

[Höppner *et al.* 1999] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis.* J. Wiley & Sons, Chichester, United Kingdom 1999

[Huff 1954] D. Huff. *How to Lie with Statistics.* W.W. Norton, New York, NY, USA 1954

[Ising 1925] E. Ising. Beitrag zur Theorie des Ferromagnetismus. Zeitschrift für Physik, 31(253), 1925

[Jänich 1983] K. Jänich. *Lineare Algebra* (3. Auflage). Springer-Verlag, Heidelberg, Germany 1983

[Jordan 1998] M.I. Jordan, ed. *Learning in Graphical Models.* MIT Press, Cambridge, MA, USA 1998

[Kolmogorow 1933] A.N. Kolmogorow. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Heidelberg, Germany 1933. English edition: *Foundations of the Theory of Probability.* Chelsea, New York, NY, USA 1956

[Kolodner 1993] J. Kolodner. *Case-Based Reasoning.* Morgan Kaufmann, San Mateo, CA, USA 1993

[Kowalski 1979] H.-J. Kowalski. *Lineare Algebra* (9. Auflage). de Gruyter, Berlin, Germany 1979

[Krämer 1997] W. Krämer. *So lügt man mit Statistik (7. Auflage)*. Campus-Verlag, Frankfurt, Germany 1997

[Langley *et al.* 1992] P. Langley, W. Iba, and K. Thompson. An Analysis of Bayesian Classifiers. *Proc. 10th Nat. Conf. on Artificial Intelligence (AAAI'92, San Jose, CA, USA)*, 223–228. AAAI Press/MIT Press, Menlo Park/Cambridge, CA, USA 1992

[Larsen and Marx 1986] R.J. Larsen and M.L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall, Englewood Cliffs, NJ, USA 1986

[Lauritzen and Spiegelhalter 1988] S.L. Lauritzen and D.J. Spiegelhalter. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, 2(50):157–224. Blackwell, Oxford, United Kingdom 1988

[McCulloch 1965] W.S. McCulloch. *Embodiments of Mind*. MIT Press, Cambridge, MA, USA 1965

[McCulloch und Pitts 1943] W.S. McCulloch and W.H. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5:115–133. USA 1943
Reprinted in [McCulloch 1965], 19–39, in [Anderson and Rosenfeld 1988], 18–28, and in [Boden 1990], 22–39.

[Michalski and Stepp 1983] R.S. Michalski and R.E. Stepp. Learning from Observation: Conceptual Clustering. In: [Michalski *et al.* 1983], 331–363

[Michalski *et al.* 1983] R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, ed. *Machine Learning: An Artificial Intelligence Approach*. Morgan Kaufmann, San Mateo, CA, USA 1983

[Minsky und Papert 1969] L.M. Minsky and S. Papert. *Perceptrons*. MIT Press, Cambridge, MA, USA 1969

[von Mises 1928] R. von Mises. *Wahrscheinlichkeit, Statistik und Wahrheit*. Berlin 1928

[Mucha 1992] H.-J. Mucha. *Clusteranalyse mit Mikrocomputern*. Akademie-Verlag, Berlin, Germany 1992

[Muggleton 1992] S. Muggleton, ed. *Inductive Logic Programming*. Academic Press, San Diego, CA, USA 1992

[Nakhaeizadeh 1998a] G. Nakhaeizadeh, ed. *Data Mining: Theoretische Aspekte und Anwendungen*. Physica-Verlag, Heidelberg, Germany 1998

[Nakhaeizadeh 1998b] G. Nakhaeizadeh. Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick. In: [Nakhaeizadeh 1998a], 1–33

[Nauck and Kruse 1997] D. Nauck and R. Kruse. A Neuro-Fuzzy Method to Learn Fuzzy Classification Rules from Data. *Fuzzy Sets and Systems* 89:277–288. North-Holland, Amsterdam, Netherlands 1997

[Nauck *et al.* 1997] D. Nauck, F. Klawonn, and R. Kruse. *Foundations of Neuro-Fuzzy Systems.* J. Wiley & Sons, Chichester, United Kingdom 1997

[Newell and Simon 1976] A. Newell and H.A. Simon. Computer Science as Empirical Enquiry: Symbols and Search. *Communications of the Association for Computing Machinery* 19. Association for Computing Machinery, New York, NY, USA 1976.
Reprinted in [Boden 1990], 105–132.

[Nilsson 1965] N.J. Nilsson. *Learning Machines: The Foundations of Trainable Pattern-Classifying Systems.* McGraw-Hill, New York, NY, 1965

[Nilsson 1998] N.J. Nilsson. *Artificial Intelligence: A New Synthesis.* Morgan Kaufmann, San Francisco, CA, USA 1998

[Pearl 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA, USA 1988 (2nd edition 1992)

[Press *et al.* 1992] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C — The Art of Scientific Computing (2nd edition).* Cambridge University Press, Cambridge, United Kingdom 1992

[Quinlan 1986] J.R. Quinlan. Induction of Decision Trees. *Machine Learning* 1:81–106. Kluwer, Dordrecht, Netherlands 1986

[Quinlan 1993] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, USA 1993

[de Raedt and Bruynooghe 1993] L. de Raedt and M. Bruynooghe. A Theory of Clausal Discovery. *Proc. 13th Int. J. Conf. on Artificial Intelligence.* Morgan Kaufmann, San Mateo, CA, USA 1993

[Rojas 1996] R. Rojas. *Theorie der neuronalen Netze — Eine systematische Einführung.* Springer, Berlin, Germany 1996

[Rosenblatt 1958] F. Rosenblatt. The Perceptron: A Probabilistic Modell for Information Storage and Organization in the Brain. *Psychological Review* 65:386–408. USA 1958

[Rosenblatt 1962] F. Rosenblatt. *Principles of Neurodynamics.* Spartan Books, New York, NY, USA 1962

[Rumelhart und McClelland 1986] D.E. Rumelhart and J.L. McClelland, eds. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. 1: Foundations.* MIT Press, Cambridge, MA, USA 1986

[Rumelhart *et al.* 1986a] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning Internal Representations by Error Propagation. In [Rumelhart und McClelland 1986], 318–362.

[Rumelhart *et al.* 1986b] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning Representations by Back-Propagating Errors. *Nature* 323:533–536. 1986

[Sachs 1999] L. Sachs. *Angewandte Statistik - Anwendung statistischer Methoden (9. Auflage).* Springer, Berlin/Heidelberg, Germany 1999

[Savage 1954] L.J. Savage. *The Foundations of Statistics.* J. Wiley & Sons, New York, NY, USA 1954. Reprinted by Dover Publications, New York, NY, USA 1972

[Schwarz 1978] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics* 6:461–464. Institute of Mathematical Statistics, Hayward, CA, USA 1978

[Srikant and Agrawal 1996] R. Srikant and R. Agrawal. Mining Quantitative Association Rules in Large Relational Tables. *Proc. Int. Conf. Management of Data (Montreal, Quebec, Canada)*, 1–12. ACM Press, New York, NY, USA 1996

[Wang and Mendel 1992] L.-X. Wang and J.M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Trans. on Systems, Man, & Cybernetics* 22:1414–1227. IEEE Press, Piscataway, NJ, USA 1992

[Wasserman 1989] P.D. Wasserman. *Neural Computing: Theory and Practice.* Van Nostrand Reinhold, 1989

[von Weizsäcker 1992] C.F. Weizsäcker. *Zeit und Wissen.* Hanser, München, Germany 1992

[Werbos 1974] P.J. Werbos. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences.* Ph.D. Dissertation, Havard University, Cambridge, MA, USA 1974

[Wettschereck 1994] D. Wettschereck. *A Study of Distance-Based Machine Learning Algorithms.* PhD Thesis, Oregon State University, OR, USA 1994

[Widner 1960] R.O. Widner. Single State Logic. *AIEE Fall General Meeting*, 1960. Reprinted in [Wasserman 1989].

[Widrow und Hoff 1960] B. Widrow and M.E. Hoff. Adaptive Switching Circuits. *IRE WESCON Convention Record*, 96–104. IRE, New York, NY, USA 1960

[Zell 1994] A. Zell. *Simulation Neuronaler Netze.* Addison-Wesley, Stuttgart, Germany 1996

[Zey 1997] R. Zey, ed. *Lexikon der Forscher und Erfinder.* Rowohlt, Reinbek/Hamburg, Germany 1997

Index