§0: Einleitung

Hauptziel dieser Vorlesung ist die Vermittlung der wichtigsten statistischen Methoden sowie der Voraussetzungen und Grenzen ihrer Anwendbarkeit mit dem Ziel, daß die Hörer in der Lage sind,

1. elementare statistische Auswertungen durchzuführen *und*
2. statistische Auswertungen in der Fachliteratur zu verstehen und kritisch zu hinterfragen.

Die Grundaufgaben der Statistik sind

- Beschreiben
- Schätzen
- Entscheiden

1. Beschreiben

Beschreibende Statistik, der ältester Teil der Statistik: Im 18. Jahrhundert war Statistik der „Lehre von der Zustandsbeschreibung der Staaten". („Statistik" von *frz.* «statistique» von *neulat.* „statisticus" *(17. Jhd.)* = den Staat betreffend.

Beispiel: Beschreibe die Altersstruktur der Mannheimer Bevölkerung!

Maximale Information gibt die Liste aller Geburtstage – zu unübersichtlich. Die beschreibende Statistik stellt Methoden bereit, um mit deutlich weniger Zahlen einen möglichst großen Teil der Information darzustellen durch Ordnen und Strukturieren des Datenmaterials. *Hierbei muß bereits eine Entscheidung über Informationsverlust fallen!*

Beschreibung durch Tabellen, Diagramme, Indexzahlen, statistische Maßzahlen wie Mittelwerte, Streuungsmaße *usw.*

In der Geographie besonders wichtig: Beschreibung der *räumlichen Verteilung* von Daten und ihrer Veränderungen (z.B. Wanderungsbewegungen).

2. Schätzen

Analytische Statistik, entstanden aus der „Politischen Arithmetik", die aufgrund der Daten aus Tauf-, Heirats- und Sterberegistern Aussagen über Altersaufbau, Fruchtbarkeit, Sterberaten *usw.* ableitete. Allgemein: Schließe anhand von Beobachtungsdaten auf allgemeine Gesetzmäßigkeiten.

Beispiel: Was halten die Studenten der Universität Mannheim vom neuen MVG-Fahrplan?

Ausgangspunkt für die Beantwortung dieser Frage ist eine *Stichprobe,* interessant ist aber nicht die Stichprobe, sondern die *Gesamtheit,* aus der die Stichprobe gewählt wurde.

Probleme:

- Auswahl einer geeigneten Stichprobe. **Zufallsstichprobe?**

- Wie wahrscheinlich ist es, daß die aus der Stichprobe gewonnene Beschreibung auch eine ungefähre (?) Beschreibung der Grundgesamtheit gibt?

Grundlage für die Beantwortung solcher Fragen ist die *Wahrscheinlichkeitsrechnung,* insbesondere geht es dabei um *statistische Verteilungen.*

3. Entscheiden

Dies ist Aufgabe der *schließenden Statistik;* die Übergänge zur analytischen Statistik sind fließend.

Beispiel: Gab es in den letzten zehn Jahren eine allgemeine Klimaänderung, die zu einer langfristigen Erwärmung der Atmosphäre führt?

Zur Beantwortung solcher Fragen stehen *Signifikanztests* bereit: Man geht aus von einer *Nullhypothese* und macht Aussagen über die Wahrscheinlichkeit dafür, daß die vorliegenden Daten aus der durch diese Hypothese geforderten Verteilung stammen können. Falls diese Wahrscheinlichkeit zu klein ist, muß die Nullhypothese abgelehnt werden.

Ein spezielles, bislang noch nicht erwähntes Problem, bei dem alle drei genannten Aspekte auftreten, ist die

4. Regression und Korrelation

Hier geht es um die Abhängigkeit zweier (oder mehrerer) Größen voneinander.

Beispiel: Wie beeinflußt die berufliche Stellung eines Wählers seine Wahlentscheidung?

Ausgangspunkt sind parameterabhängige mathematische *Modelle,* deren Parameter so geschätzt werden, daß die beobachteten Daten unter dem Modell mit diesen Parametern möglichst wahrscheinlich sind *(maximum likelihood).* Hinzu kommen Aussagen über die Wahrscheinlichkeit, daß das Modell korrekt sein könnte.

In Abhängigkeit von ihren Zielen unterscheidet man wiederum drei Arten von Statistik:

- Die mathematische Statistik
- Die statistische Methodenlehre
- Die Wirtschafts- und Sozialstatistik

5. Die mathematische Statistik

Die *mathematische Statistik* stellt die mathematischen, insbesondere auch wahrscheinlichkeitstheoretischen Grundlagen bereit und entwickelt auf deren Basis Arbeitsmethoden: Sie sucht nach mathematischen Beschreibungen für statistische Verteilungen, untersucht deren Abhängigkeit von Parametern und entwickelt Berechnungsverfahren.

All dies ist üblicherweise nicht die Aufgabe eines Geographen; die mathematische Statistik wird in dieser Vorlesung daher keine größere Rolle spielen.

6. Die statistische Methodenlehre

Hier geht es um die Auswahl und den Einsatz der von der mathematischen Statistik bereitgestellten Methoden unter Beachtung der Tragweite und der Grenzen der einzelnen Methoden. *Dies ist der wesentliche Stoff dieser Vorlesung.* Die einzelnen Schritte sind

- Datenerhebung
- Datenaufbereitung (d.h. Verdichtung der Daten zu Tabellen, Diagrammen *usw.*
- Berechnung statistischer Maße wie Mittelwert, Streuung *usw.*
- Schätzen gesuchter Parameter, Prüfen von Hypothesen.

7. Die Wirtschafts- und Sozialstatistik

Hier geht es um die Erfassung und Sammlung von Daten nach gleichen Kriterien über längere Zeiträume hinweg; vor allem zu nennen sind hier die amtlichen Statistiken zu den verschiedensten Lebensbereichen, gesammelte Klimadaten des deutschen Wetterdienstes *usw.* Diese Statistiken werden hier in der Vorlesung nur gelegentlich in Beispielen auftreten.

8. Die Rolle der statistischen Methodenlehre in der empirischen Arbeit

Der Einsatz statistischer Methoden ist immer nur ein *Teil* eines Arbeitsprojekts: Am Anfang steht die Problemstellung und am Ende die Interpretation der Ergebnisse, was beides auf der *inhaltlichen* Ebene stattfindet: Hier hilft nicht die statistische Methodenlehre, sondern hier sind Geographiekenntnisse gefragt und die hier auftretenden Fehler sind Fehler bei der geographischen Arbeit, die im allgemeinen von keiner statistischen Auswertungsprozedur gefunden werden können. Umgekehrt führen Fehler auf der technisch-statistischen Ebene zu Fehlinterpretationen, die dann auf der geographischen Ebene schwer zu erkennen

sein können – falls die Ergebnisse nicht gerade komplett unsinnig sind.

Im einzelnen ergibt sich folgendes Zusammenspiel der beiden Ebenen:

— inhaltliche Ebene —

- Ausgangspunkt ist eine *Fragestellung,* die mit den vorliegenden Informationen (Stand der Forschung, vorhandene Daten) nicht beantwortet werden kann.
- Aufgrund der vorhandenen Information wird ein *Modell* entwickelt und werden eventuell bereits erste *Hypothesen* formuliert.
- Das Modell wird formalisiert: Exakte Definition der relevanten Größen durch präzise Meßvorschriften; Auswahl geeigneter Verfahren zur Schätzung der unbekannten Parameter

— empirisch-technische Ebene —

- Entwurf der Erhebungsbögen, Auswahl der Meßinstrumente *etc.* Festlegung der Stichprobe (Auswahlkriterien)
- *ggf.* Voruntersuchungen zur Kontrolle der Praktikabilität und der Konsistenz des Untersuchungsplans
- Datenerhebung
- Datenaufbereitung
- Statistische Datenanalyse

— inhaltliche Ebene —

- Interpretation der Daten

§1: Erfassung und Aufbereitung von Daten

Die Grundaufgabe beim Arbeiten mit statistischen Methoden besteht immer darin, daß man Informationen über eine gewisse *Grundgesamtheit* haben möchte. Diese Grundgesamtheit kann man sich etwa vorstellen als die Einwohnerschaft der Stadt Mannheim, als die sämtlichen Bäume des Schwarzwalds, als die Fläche des Landes Brasilien, als die Industrieunternehmen eines bestimmten Wirtschaftsraums, kurzum als so ziemlich alles, was einen Geographen interessiert und was aus mehreren wohlunterschiedenen Elementen besteht: Den einzelnen Mannheimern, den einzelnen Bäumen, den einzelnen Punkten (realistischerweise wohl eher irgendwie festgelegten Parzellen) des Landes Brasilien, oder den einzelnen Industrieunternehmen.

Diese einzelnen Elemente oder Individuen haben verschiedene *Merkmale,* die untersucht werden sollen: Einwohner von Mannheim haben etwa als solche Merkmale etwa ihren Geburtsort, ihren Arbeitgeber, ihr Einkommen, ihre politischen Präferenzen bei Gemeinderatswahlen *usw.;* die Bäume des Schwarzwalds haben als Merkmale unter anderem ihre biologische Art, ihr Alter, ihre Größe, ihre Waldschadensstufe *usw.* Bei der Fläche Brasiliens könnten die Vegetation und die Nutzung interessieren, oder die Bodenart, oder der Schadstoffgehalt des Bodens *usw.*

In diesem Kapitel soll es darum gehen, wie solche Daten mit vertretbarem Aufwand erhoben werden können und was man mit welchen Arten von Daten anfangen kann.

Beginnen wir mit den möglichen Datenarten.

1. Skalen

Für die Zwecke der Statistik ist eine Skala ein Wertebereich für Daten. Je nach Art der Daten unterscheidet man folgende Skalen:

a) **Nominalskalen:** Diese schwächste aller Skalen wird für qualitative Merkmale angewandt, die man zwar in wohlunterschiedene Klassen einteilen kann, für die man aber keine Rangfolge zwischen den einzelnen Klassen festlegen kann.

Beispiele hierfür sind Berufe, Nationalitäten, Institutionen wie Schulen oder Arbeitgeber, Aussagen über Flächennutzung *(landwirtschaftliche Fläche, Verkehrsfläche, Industriegebiet, . . .)*, Automarken *usw.*

Gemeinsam ist allen Beispielen (und allen Nominalskalen), daß man für jeden Skalenwert zählen kann, wie viele Elemente der Grundgesamtheit diesen Merkmalswert haben („120 Mannheimer sind von Beruf Malermeister“), daß man aber nicht von größeren oder kleineren oder gar mittleren Merkmalswerten sprechen kann („Der mittlere Beruf eines Mannheimer Bürgers ist . . . ?“).

Spezielle Nominalskalen sind die *Alternativskalen,* bei denen es nur zwei Merkmalsausprägungen gibt, beispielsweise „Ja“ und „Nein“. Ebenfalls unter den Nominalskalen kann man auch *geographische Skalen* einordnen, etwa eine Skala für Geburtsorte, allerdings werden wir noch sehen, daß es für solche Skalen doch mehr Möglichkeiten gibt als für beliebige Nominalskalen; beispielsweise kann man Begriffe wie *Bevölkerungsschwerpunkte* erklären, es gibt also so etwas wie geographische Mittelwerte.

Für Daten, die in Nominalskalen gemessen werden, ist eine Zusammenfassung mehrere Klassen nur auf der inhaltlichen Ebene möglich, wo man etwa Drechsler und Zimmerleute zur holzverarbeitenden Industrie zusammenfassen kann; auf der statistischen Ebene gibt es keine entsprechende Möglichkeit.

b) **Ordinalskalen:** Dies sind Skalen für Merkmale, bei denen man von einer Rangfolge zwischen den einzelnen Merkmalsausprägungen reden kann.

Ordinalskalen treten etwa auf für die Charakterisierung von Gemeindegrößen durch Begriffe wie „Großstadt“ oder „Dorf“, wobei man zwar klar unterscheiden kann, welcher Begriff eine größere Gemeinde beschreibt, wo man aber nicht sinnvoll die *Unterschiede* zwischen aufeinanderfolgende Klassen vergleichen kann: Der Einwohnerunterschied zwischen einer Millionenstadt und einer Großstadt ist sicherlich größer als der zwischen einer Kleinstadt und einem Dorf.

Weitere Beispiele von Ordinalskalen sind reine Rangskalen, wie sie etwa im Sport auftreten, die MOHsche Härteskala in der Gesteinskunde, Handelsklassen für Waren, Waldschadensstufen *usw.*

Daten, die in Ordinalskalen gemessen werden, lassen sich zusammenfassen, indem man aufeinanderfolgende Rangklassen zu einer Klasse zusammenfaßt und damit etwa die

Daten für die Klassen 1 bis 5 und 5 bis 10 *etc* jeweils als ganzes darstellt. Es ist im allgemeinen nicht sinnvoll, mittlere Ränge auszurechnen, da der Abstand zwischen zwei aufeinanderfolgenden Rängen sehr verschieden sein kann.

c) **Intervallskalen:** Dies sind strikt metrische Skalen, aller Merkmalsausprägungen sind also Zahlenwerte, und es ist auch sinnvoll, von *Differenzen* zwischen diesen Werten zu reden.

Typische Beispiele sind etwa Kalenderdaten oder Temperaturen in °C. In beiden Fällen gibt es keinen *natürlichen* Nullpunkt, auf den man sich beziehen kann, es ist also beispielsweise nicht sinnvoll, von der doppelten Temperatur oder dem doppelten Datum zu reden, aber man kann in sinnvollerweise von Mittelwerten wie etwa der mittleren Temperatur reden.

Daten, die mit Intervallskalen gemessen werden, lassen sich so zusammenfassen, daß Klassen gleicher Breite entstehen.

d) **Rationalskalen:** Auch dies sind strikt metrische Skalen, sie haben aber im Gegensatz zu Intervallskalen auch noch einen natürlichen Nullpunkt, so daß man auch von *Verhältnissen* zwischen den Daten sprechen kann.

Beispiele sind etwa Wirtschaftsdaten wie Einkommen, Sozialprodukte, Geldmengen *usw.*, außerdem Einwohnerzahlen, Hektarerträge und vieles mehr.

Mit solchen Daten läßt sich natürlich alles machen, was auch mit solchen auf einer Intervallskala möglich ist; hinzu kommen einige zusätzliche Möglichkeiten wie die Berechnung von Steigerungsraten, die Darstellung mit einer logarithmischen Skala, die Berechnung geometrischer Mittelwerte *usw.;* mit Einzelheiten werden wir uns später noch befassen.

2. Stichproben

Nachdem wir nun wissen, welche Arten von Daten in der Statistik unterschieden werden, ist es langsam an der Zeit, daß wir uns um die Beschaffung dieser Daten kümmern.

Das beste Verfahren besteht darin, von jedem Element der Grundgesamtheit für jedes interessierende Merkmal den Wert zu ermitteln, um so eine vollständige Übersicht zu haben. Leider ist eine solche *Totalerhebung* oftmals entweder schlichtweg unmöglich oder aber nur mit unvertretbar hohem Aufwand möglich. Man muß sich deshalb damit begnügen, nur einen *Teil* der Elemente aus der Grundgesamtheit zu untersuchen; diesen Teil bezeichnet man als eine *Stichprobe.*

Eine solche Stichprobe kann auf verschiedene Weisen aus der Grundgesamtheit ausgewählt werden; deren Vor- und Nachteile wollen wir uns an einem einfachen Beispiel klarmachen:

Auf dem Plan zum ersten Übungsblatt sind zweihundert Einfamilienhäuser einer englischen Ortschaft eingezeichnet. Durch Befragen von fünfzig Haushaltsvorständen soll geschätzt werden,

a) wie viele der Haushaltsvorstände Immigranten sind
b) wie viele der Haushalte (mindestens) ein Fernsehgerät besitzen
c) wie viele der Haushalte (mindestens) ein Auto besitzen

d) wie viele der Haushaltsvorstände bei der nächsten Wahl für die NDP stimmen werden. *(Die britische NDP ist ein Wahlbündnis aus Liberalen und von der Labourpartei abgespalteten Sozialdemokraten.)*

Die ebenfalls beigefügte Tabelle gibt für jeden der zweihundert Haushalte folgende Daten:

- In Spalte „R" eine willkürlich zugeordnete Ordnungszahl zwischen 1 und 200
- In Spalte „S" die Hausnummer
- In Spalte „I" eine 1, falls der Haushaltsvorstand Immigrant ist, sonst eine 0
- In Spalte „T" eine 1, falls der Haushalt ein Fernsehgerät besitzt, sonst eine 0
- In Spalte „C" eine 1, falls der Haushalt ein Auto besitzt, sonst eine 0
- In Spalte „N" eine 1, falls der Haushaltsvorstand angibt, bei der nächsten Wahl für die NDP zu stimmen, sonst eine 0.

Die Spalten „I" bis „N" dieser Tabelle sind hierbei natürlich zunächst noch nicht bekannt; sie wären für die Elemente der Stichprobe durch Befragen des Haushaltsvorstands zu ermitteln.

Im folgenden sollen einige Verfahren zur Auswahl dieser Stichprobe besprochen werden.

a) **Willkürliche Auswahl:** Dies ist die einfachste und schnellste Methode: Man wählt willkürlich irgendwie eine Stichprobe aus der Grundgesamtheit aus.

Im Beispiel würde das bedeuten, daß der Befrager durch die Stadt geht und dabei irgendwelche fünfzig Häuser besucht aus (oder gar auf der Straße fünfzig willkürlich ausgewählte Personen befragt).

Vorteil dieser Methode ist die einfache Durchführbarkeit, Nachteil ist die Gefahr, daß die Stichprobe zu homogen wird, da sie beispielsweise eines der Wohngebiete übermäßig berücksichtigt, was etwa im Falle des Gebiets bei der Fabrik oder des Gebiets „The Squerries" zu jeweils völlig anderen Ergebnissen führt.

Wesentliche Anforderungen an eine Stichprobe sollten also sein:

1.) Jedes Element aus der Grundgesamtheit muß die gleiche Chance haben, in der Stichprobe aufzutauchen und
2.) Die Auswahl eines Elements darf die Chancen der übrigen nicht beeinflussen.

Eine oft recht gute Annäherung an diese Forderungen liefert eine

b) **Systematische Auswahl:** Hier werden die Elemente der Grundgesamtheit in irgendeiner Weise geordnet, und dann wird für einen geeigneten Wert r jedes r-te Element in die Stichprobe aufgenommen.

Auch diese Methode läßt sich oft sehr einfach durchführen; falls es um Personenbefragungen geht, muß man für diese Art von Auswahl beispielsweise keine Adressenlisten haben, was die Kosten beträchtlich reduzieren kann. Zu Fehlern kann es kommen, wenn die Regularität der Auswahl auf eine Regularität der Grundgesamtheit trifft: Wird beispielsweise in einer Straße jede zehnte Hausnummer ausgewählt, so werden nur Häuser auf *einer* der beiden Straßenseiten (gerade *bzw.* ungerade) berücksichtigt, was bei Straßen an

der Grenze zwischen zwei Wohngebieten zu beträchtlichen Verfälschungen führen kann. Bei sehr regulären Stadtanlagen könnte es auch passieren, daß man möglicherweise alle Eckhäuser oder aber gar keine Eckhäuser erfaßt; da Geschäfte häufig in Eckhäusern zu finden sind, würde auch das zu einer Verfälschung des Ergebnisses führen.

Im hier betrachteten Beispiel sollen fünfzig von zweihundert Haushalten ausgewählt werden; im Falle einer systematischen Auswahl nimmt man also jeden vierten. Beginnt man mit dem Haushalt mit $R = 1$ und geht dann jeweils um vier weiter, also zu $R = 5, 9, 13, \ldots$, so findet man

- 3 eingewanderte Haushaltsvorstände
- 43 Haushalte mit Fernsehgeräten
- 23 Haushalte mit Auto
- 17 Haushaltsvorstände, die NDP wählen.

Die entsprechenden Zahlen für die Grundgesamtheit sollten etwa viermal so groß sein, wir schätzen also, daß es

- 12 eingewanderte Haushaltsvorstände gibt,
- 172 Haushalte mit Fernsehgeräten,
- 92 Haushalte mit Auto, und
- 68 Haushaltsvorstände, die NDP wählen.

Die korrekten Zahlen, die man hier ja aus der Tabelle ablesen kann, sind:

- 15 eingewanderte Haushaltsvorstände
- 160 Haushalte mit Fernsehgeräten
- 100 Haushalte mit Auto
- 70 Haushaltsvorstände, die NDP wählen,

vor allem bei den Immigranten ist die Schätzung also ziemlich ungenau: Sie ist um 20% niedriger als der korrekte

Wert. Hätte man alle Haushalt mit durch vier teilbarem R gefragt, hätte man fünf Immigranten gefunden und somit auf eine Gesamtzahl von 20 geschätzt, was sogar ein Fehler von einem Drittel gewesen wäre. Der Grund für dieses schlechte Ergebnis liegt sicherlich teilweise darin, daß häufig zwei Immigranten in benachbarten Häusern wohnen, von denen die hier getroffene Auswahl nie mehr als eines erfassen kann.

Dieser Nachteil wird vermieden durch die

c) **Reine Zufallsauswahl:** Hier werden die Elemente der Grundgesamtheit in irgendeiner Weise durchnummeriert. Dann wählt man zufällig eine Zahl zwischen 1 und der Anzahl der Elemente und nimmt das Element mit der entsprechenden Nummer in die Stichprobe auf. Dieses Verfahren wird wiederholt, bis eine Stichprobe des gewünschten Umfangs erreicht ist.

Im betrachteten Beispiel sind die Haushalte bereits durchnummeriert, wir müssen also zufällig fünfzig verschiedene Zahlen zwischen 1 und 200 wählen, und die Stichprobe besteht aus allen Haushalten, deren R-Wert unter diesen Zahlen vorkommt.

Die „zufällig erzeugten“ Zahlen kann man etwa von einem Computer berechnen lassen oder aus einer Tafel von Zufallszahlen entnehmen. Computer stellen im allgemeinen Zufallsgeneratoren zur Verfügung, die direkt Zufallszahlen aus einem gewünschten Intervall produzieren können, im vorliegenden Beispiel also Zufallszahlen zwischen 1 und 200, während Tabellen meist in Fünferblöcken von Ziffern organisiert sind und somit Zahlen zwischen 0 und 99999 liefern. Diese Ziffern lassen sich können aber auch (zumindest in

Gedanken) anders zusammenfassen, so daß man die Tabelle auch als eine Tabelle von Dreierblöcken oder Blöcken jeder anderen Länge auffassen kann. Hat man etwa die Folge

58251 40173 93179 14717 25291 98093

von Fünferblöcken, so kann man diese auch als Folge

582 514 017 393 179 147 172 529 198 093

auffassen, und daraus bekommt man eine Folge von Zufallszahlen zwischen 1 und 200, indem man etwa jede Zahl, die nicht in diesen Bereich fällt, wegläßt.

Dies kann allerdings zu einem ziemlich verschwenderischen Umgang mit Zufallszahlen führen, etwa falls man eine Stichprobe aus einer Grundgesamtheit vom Umfang 1001 ziehen möchte und dann von den vierstelligen Zufallszahlen zwischen 0 und 9999 fast 90% wegwerfen muß. Besser ist es, die Zufallszahlen auf einen Teiler von 10000 zu reduzieren, im vorliegenden Fall etwa 2000, indem man anstelle der Zahl selbst ihren *Rest* bei der Division durch 2000 nimmt. Damit erhält man Zufallszahlen zwischen 0 und 1999, von denen rund jede zweite verwendet werden kann.

Im vorliegenden Beispiel, bietet sich an, stattdessen die Reste bei der Division durch 200 zu nehmen; dann kann man sogar *jede* Dreiergruppe verwenden: Die Null würde natürlich in diesem Fall als 200 interpretiert, da wir zwar eine laufende Nummer 200, aber keine laufende Nummer 0 haben. Mit den obigen Zufallszahlen führt dieses Verfahren auf die Zahlen

182 114 017 193 179 147 172 129 198 093 ,

womit die ersten zehn Mitglieder der Stichprobe gefunden wären.

Dieses Verfahren der reinen Zufallsauswahl erfüllt die genannten Kriterien an eine Zufallsauswahl offensichtlich in idealer Weise – zumindest, wenn die verwendeten Zufallszahlen wirklich zufällig sind. Das sind sie in der *Praxis* allerdings leider fast nie: Ein Computer berechnet Zufallszahlen selbstverständlich nicht durch Würfeln, sondern nach einem strikt deterministischen Algorithmus, der zumindest die zweite der obigen Regeln in eklatanter Weise verletzt; die erste kann in den meisten Fällen dadurch (mehr oder weniger) gerettet werden, daß man einen hinreichend zufälligen Startwert wählt, wie etwa die Millisekunden der Systemzeit beim Start der Berechnung.

Trotzdem erzeugt ein solches Programm natürliche nie *echte* Zufallszahlen; zumindest in der theoretischeren Literatur spricht man daher korrekter von *Pseudozufallszahlen.* Die Stichproben *müssen* durch die Verwendung solcher Pseudozufallsgeneratoren nicht schlechter werden, aber leider sind die meisten Pseudozufallsgeneratoren von Computern oder gar Taschenrechnern deutlich schlechter als sie sein müßten und vor allem für umfangreichere Untersuchungen sehr mit Vorsicht zu genießen. Vor allem für extrem umfangreiche Simulationen, wie sie derzeit nur von Supercomputern bewältigt werden können, gibt es im Augenblick kein zuverlässiges Verfahren zur Erzeugung von Pseudozufallszahlen; für normale Stichproben, wie sie im geographischen Alltagsgeschäft auftreten, gibt es entsprechende Algorithmen, aber man sollte sich nicht darauf verlassen, daß die *random*-Funktion des verwenden Computers ein solches Verfahren benutzt.

Idealerweise sollte man eine Berechnung, die auf Zufallszahlen beruht, mindestens zweimal mit völlig verschiedenen Zufallsgeneratoren durchführen, in der Praxis des typischen Anwenders scheitert dies aber leider daran, daß er im allgemeinen nichts davon weiß, auf welchen Prinzipien sein Zufallsgenerator beruht.

Zuverlässiger sind im allgemeinen die veröffentlichten Tabellenwerke von Zufallszahlen, die Seite für Seite statistischen Test unterworfen werden und deren Ergebnisse auch angeben; sie sind aber natürlich für umfangreichere Erhebungen sehr viel umständlicher in der Benutzung.

Auch bei besten (Pseudo)Zufallsgeneratoren kann man gelegentlich noch ein besseres Verfahren zur Auswahl einer Stichprobe finden als die reine Zufallsauswahl, nämlich die

d) **Stratifizierte Zufallsauswahl:** Hier unterteilt man die Grundgesamtheit G in mehrere Teile $T_1, \ldots, T_m$, die zusammen G ergeben. Nun wählt man aus jeder der Teilgesamtheiten T_i eine Stichprobe und schätzt die gesuchte Größe einzeln für diese Teilgesamtheit. Diese Schätzwerte werden dann hochgerechnet auf die Grundgesamtheit.

Die Idee hinter diesem Verfahren ist folgende: In vielen geographischen Erhebungen hat man es mit räumlichen Verteilungen zu tun, die zwar *insgesamt* sehr heterogen sind, in Teilgebieten dagegen äußerst homogen. Beispiele sind etwa soziogeographische Untersuchungen in Städten, wo es sowohl „Bonzenviertel“ als auch Slums gibt, oder pflanzengeographische Untersuchungen in Gegenden mit vielen Monokulturen. Nimmt man Stichproben aus den homogenen Teilgesamtheiten, so kann man im allgemeinen davon ausgehen, daß man für diese nur sehr kleine Schätzfehler macht, so daß dann auch der Schätzwert für die ganze Grundgesamtheit genauer wird.

Die Art des Hochrechnens auf die Grundgesamtheit hängt etwas ab von der Art der betrachteten Daten: Falls es, wie hier im Beispiel, einfach um Anzahlen geht, wobei etwa a_i die geschätzte Anzahl für T_i sei, so ist die Schätzung für ganz G natürlich die Summe

$$a_1 + \cdots + a_n \, .$$

Falls es dagegen um Durchschnittswerte geht (dazu gehören auch Prozentzahlen), so muß man einen *gewichteten Mittelwert* bilden; sind s_i die entsprechenden Schätzwerte, so ist dieser definiert als

$$s = \frac{\#T_1}{\#G} \cdot s_1 + \cdots + \frac{\#T_m}{\#G} \cdot s_m \, ,$$

wobei $\#T_i$ und $\#G$ jeweils die Elementanzahlen der entsprechenden Gesamtheiten sind.

Diese letztere Formel kann auch verwendet werden, wenn die Teilgesamtheiten T_i zusammen nicht die ganze Grundgesamtheit bilden; in diesem Fall spricht man, falls auch die T_i zufällig gewählt sind, von einer **mehrstufigen Zufallsauswahl**. Beispielsweise könnte man für eine Stichprobe aus der deutschen Bevölkerung zunächst zufällig einige Regierungsbezirke auswählen und aus diesen wiederum entweder direkt oder durch Zwischenschalten weiterer Schichten (Kreise, Gemeinden, Ortsteile, …) zufällig eine Stichprobe der Bevölkerung ziehen. Ein Vorteil dieses Verfahrens ist, daß man erst im letzten Schritt der Auswahl, wenn es direkt um Personen geht, Adressenlisten benötigt, nicht aber für die gesamte Bundesrepublik, wie dies bei einer reinen Zufallsauswahl der Fall wäre.

Mit diesen kurzen Bemerkungen ist das Thema Stichprobenauswahl selbstverständlich bei weitem nicht erschöpfend dargestellt; für weitere Einzelheiten muß auf Spezialvorlesungen über Stichprobentheorie verwiesen werden oder auf die entsprechende Literatur. Dort wird auch genauer auf den Zusammenhang zwischen Kosten und Genauigkeit einer Stichprobe sowie auf entsprechende Optimierungen eingegangen.

Ein Klassiker, auf den auch neuere Werke immer wieder verweisen, ist

WILLIAM G. COCHRAN: Stichprobenverfahren, *de Gruyter, Berlin, New York,* 1972

3. Aufbereitung der Daten

Eine Liste von Rohdaten wie die Tabelle zu Aufgabe 1 des ersten Übungsblatts enthält zwar im Prinzip alle Informationen zu den vier dort gestellten Fragen, jedoch sind diese Informationen noch nicht so erschlossen, daß man ohne größere Arbeit etwas daraus ersehen könnte. Bevor dies möglich ist, muß man die Daten zunächst in geeigneter Weise aufbereiten und organisieren. Im vorliegenden Beispiel wäre dies etwa dadurch möglich, daß man für jedes der Wohnviertel die Prozentzahlen zu den einzelnen Fragen ermittelt und diese tabelliert; aus einer solchen Tabelle (oder einer entsprechenden graphischen Darstellung) sieht man „auf einen Blick“ den Charakter des jeweiligen Wohngebiets in Bezug auf die vier Fragen.

Etwas komplexer wird die Situation bei Daten, die nicht nur aus Ja/Nein-Entscheidungen bestehen oder sonst einer kleinen Nominalskala, sondern die einen umfangreichen Wertebereich etwa auf einer Ordinal- oder Intervallskala haben.

Am suggestivsten ist sicherlich eine graphische Darstellung der Daten. Umfragen unter den Adressaten statistischer Daten zeigen, daß die überwiegende Mehrzahl lieber eine Graphik als eine Tabelle benutzt. Allerdings kann eine Graphik oft auch zu falschen Schlußfolgerungen verleiten und kann, selbst wenn die dargestellten Daten völlig korrekt sind, viel leichter manipuliert werden als eine Tabelle. In der Tat ist es gelegentlich sogar unvermeidlich, den einen oder anderen Aspekt der vorliegenden Informationen zu verfälschen oder gar zu unterschlagen.

Als triviales und extrem konstruiertes Beispiel betrachten wir folgende Umsatzdaten eines fiktiven Unternehmens:

1981	8,00	*Mio. DM*
1982	8,01	*Mio. DM*
1983	8,02	*Mio. DM*
1984	8,03	*Mio. DM*
1985	8,02	*Mio. DM*
1986	8,035	*Mio. DM*
1987	8,04	*Mio. DM*
1988	8,06	*Mio. DM*
1989	8,07	*Mio. DM*

Diese Daten können problemlos durch eine Kurve dargestellt werden, wobei auf der einen Achse die Jahreszahlen und auf der anderen die Umsatzzahlen aufgetragen werden. Bei den Jahreszahlen ist die Art der Darstellung völlig klar:

Da es sich um eine Intervallskala ohne natürlichen Nullpunkt handelt, schreibt man die Jahreszahlen mit gleichen Abständen zwischen den einzelnen Jahren an die Achse.

Die Umsatzzahlen sind allerdings Zahlen auf einer Rationalskala; Informationen, wie „Umsatzsteigerung von 10%“ lassen sich nur dann visuell sofort korrekt erfassen, wenn auch der Nullpunkt der Skala sichtbar ist, also müssen wohl die Umsatzzahlen von 0 bis etwas über acht Millionen DM aufgetragen werden. Das Ergebnis ist eine praktisch flache Gerade, denn die Schwankungen im Wertebereich sind natürlich viel zu klein, als daß sie auf einer solchen Skala sichtbar sein könnten. Damit ist also nicht mehr sichtbar, daß der Umsatz meist (sehr leicht) angestiegen, im Jahr 1985 aber auch zurückgegangen ist – eine Information, die für praktische Zwecke natürlich auch nicht sonderlich relevant ist, denn zumindest inflationsbereinigt ist der Umsatz ohnehin kontinuierlich gesunken.

Trotzdem möchte man gelegentlich gerade diese kleineren Schwankungen genauer untersuchen und sie dazu auch visuell sichtbar machen. Dies wird in solchen Fällen nur möglich, wenn man auf der Achse mit den Umsatzzahlen nicht bei Null beginnen läßt, sondern bei einem Wert knapp unter acht Millionen; nun ist das Auf und Ab der Umsätze sehr gut sichtbar, aber die Schwankungen werden natürlich – zumindest visuell – weit übertrieben: Trotz praktisch konstanter Umsätze entsteht beim nicht zu genauen Hinschauen der Eindruck von rasanten Steigerungen. (Siehe dazu die auf der nächsten Seite reproduzierten Diagramme aus

J.R. GESSLER: Statistische Graphik, *Birkhäuser, Basel, Boston, Berlin,* 1993, S. 35/36.

Es ist daher nicht einfach zu entscheiden, welche der beiden Darstellungen vorzuziehen ist: Grundsätzlich sollte bei der graphischen Darstellung von Daten, die auf einer Rationalskala gemessen werden, der Nullpunkt sichtbar sein; denn sonst werden alle Veränderungsraten optisch übertrieben. Es gibt allerdings Fälle, in der es vor allem auf die Darstellung von Veränderungen ankommt, und hier kann es gelegentlich auch gerechtfertigt sein, nur den Bereich darzustellen, in dem tatsächlich Daten liegen. Man sollte dies dann aber auch optisch so klar wie möglich machen, etwa durch eine auffällige Lücke zwischen dem niedrigsten Skalenwert und der etwas weiter unten angeschriebenen Null.

In extremen Fällen kann es sogar sinnvoll sein, im gleichen Diagramm Daten mit verschiedenen Skalen darzustellen und dabei eine der Skalen bei Null beginnen zu lassen, die andere aber nicht. Ein Beispiel dafür ist die Entwicklung der Beschäftigtenzahlen im Bauhauptgewerbe, wie sie auf einer der nächsten Seiten dargestellt ist. (Auch diese Diagramme wurden aus dem zitierten Buch von GESSLER übernommen.) Das obere Diagramm zeigt die Daten so, wie man sie üblicherweise darstellen sollte: Alle Daten auf derselben Skala, deren Nullpunkt sichtbar ist. Für einen Überblick über die Beschäftigtenzahlen des Bauhauptgewerbes ist dies die korrekte Art der Darstellung.

Nun ist aber eine der Charakteristiken dieser Beschäftigtenzahlen, daß die Anzahl der Bau*arbeiter* starken saisonalen Schwankungen unterworfen ist; nach dem ersten Diagramm könnte man glaube, daß sich die Zahl der Angestellten nur leicht und ohne Sprünge verändert. Die vergrößerte Darstellung im zweiten Diagramm (wobei nun für die Angestellten der Skalennullpunkt *nicht* sichtbar ist) zeigt, daß dies nicht der Fall ist: Die Schwankungen sind zwar kleiner und zu anderen Zeitpunkten als bei den Arbeitern, aber trotzdem gibt es auch bei den Angestellten ein Auf und Ab der Beschäftigtenzahlen. Will man dies untersuchen (und beispielsweise die Hypothese testen, daß man gerade für die Entlassung von Arbeitern zusätzliche Angestellte benötigt), ist das zweite Diagramm die geeignete Darstellungsweise, auch wenn es die Schwankungen der Angestelltenzahlen natürlich stark übertreibt.

Der für diese Vorlesung wichtigste Fall für die graphische Darstellung von Daten ist der, daß es für *eine* Größe verschiedene Werte gibt, deren Verteilung untersucht werden soll. Beispiele für solche Größen sind etwa Parteipräferenzen, wobei die einzelnen Werte die Antworten ausgewählter Bürger auf eine entsprechende Frage sind, oder Anzahlen radioaktiver Zerfälle pro Zeiteinheit bei einer radioaktiven Datierung, oder Beschäftigtenzahlen von Betrieben *usw.*

Betrachten wir ein konkretes Beispiel: In den 50 Jahren zwischen 1911 und 1960 wurden in München im Oktober folgende mittlere Temperaturen (in °C) gemessen (nach

C.-D. Schönwiese: Praktische Statistik für Meteorologen und Geowissenschaftler, *Berlin, Stuttgart,* ^{2}1992, S. 17)

8,5	6,3	9,8	7,3	5,2	8,1	7,0	6,4	5,3	6,6
11,0	5,4	10,8	8,5	8,1	8,5	7,2	8,3	9,3	8,0
6,4	8,4	8,3	8,3	8,9	5,5	8,2	8,5	7,0	8,0
7,7	11,9	9,6	8,1	8,4	6,1	7,2	8,7	10,2	7,8
6,5	7,5	9,9	8,2	7,4	7,6	8,0	7,9	8,1	8,8

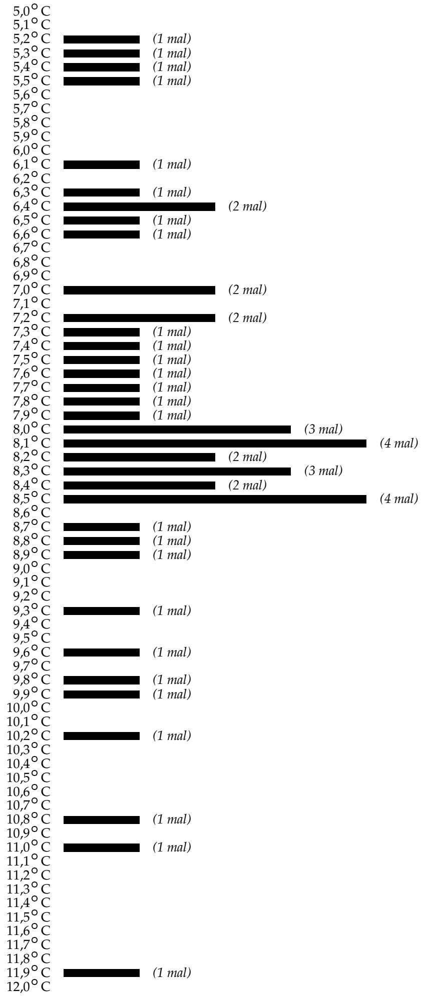
Was läßt sich anhand dieser Daten über die Verteilung der

Oktobertemperaturen in München sagen? (Der *zeitliche Verlauf* soll dabei vorerst noch völlig außer acht bleiben.) Die Tabelle aus fünfzig Werten ist offensichtlich nicht besonders übersichtlich, obwohl sie *im Prinzip* natürlich alle notwendigen Informationen enthält.

Die erste Stufe zur übersichtlicheren Darstellung dieser Informationen besteht darin, die Meßwerte der Größe nach zu ordnen und zu zählen, wie oft jeder von ihnen vorkommt. Dazu legt man am einfachsten eine Strichliste an, in der man zählt, wie oft die einzelnen Merkmalsausprägungen auftreten. Etwas aufwendiger, aber auch übersichtlicher, ist die Darstellung der Werte in einem Balkendiagramm, wobei man die Werte auf einer senkrechten oder waagrechten Achse aufträgt und neben (oder über) jeden Wert einen Balken gleicher Breite zeichnet, dessen Länge (oder Höhe) proportional zur Häufigkeit des entsprechenden Wertes ist; auf der folgenden Seite ist diese Darstellung für die hier betrachteten Daten zu sehen.

Mit diesem Balkendiagramm oder *Histogramm* sind die Daten schon deutlich übersichtlicher geworden, allerdings enthält die Darstellung auch eine ganze Reihe überflüssiger Details: Es ist sicherlich keine wesentliche Eigenschaft des Münchner Oktoberwetters, daß der Monatsmittelwert 8,3°C dreimal vorkommt, während die beiden Nachbarwerte 8,2°C und 8,4°C nur je zweimal auftreten: Durch einen einzigen Ablese- oder auch Rundungsfehler könnte die Situation genauso gut auch umgekehrt sein, und auch bei der Verschiebung des Zeitintervalls etwa um zwei Jahre auf den Zeitraum von 1913 bis 1962 hätte sich die Situation möglicherweise geändert. Das Diagramm dokumentiert also noch eine ganze Reihe von Zufallsschwankungen, die mit

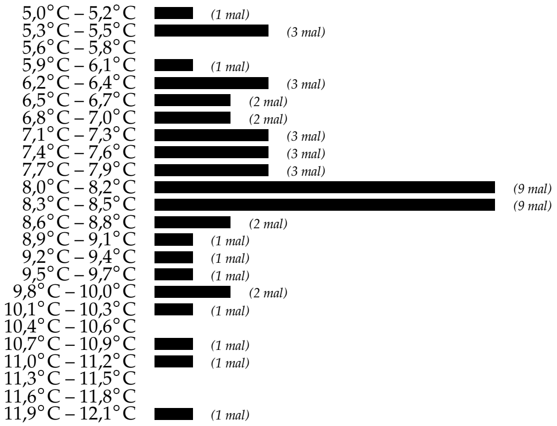
Mittlere Oktobertemperaturen in München



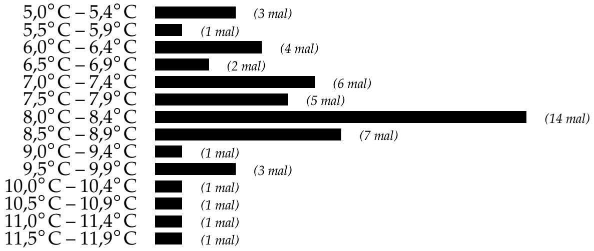
der eigentlich interessierenden allgemeinen Verteilung der mittleren Oktobertemperaturen von München wenig zu tun haben.

Es empfiehlt sich daher, aus Gründen der Übersichtlichkeit lieber auf einen Teil der vorliegenden Information zu verzichten und die Daten in Gruppen zusammenzufassen. Wie man die Daten im einzelnen gruppiert, ist zumindest teilweise Geschmacksache und hängt natürlich auch von den Daten ab: Im allgemeinen wird die Verteilung umso „glatter“ aussehen, je größer die dargestellten Gruppen sind. Als Beispiel dafür sind in den folgenden Histogrammen die mittleren Münchner Oktobertemperaturen mit Klassenbreiten von 0,3°C sowie einem halben, einem ganzen, zwei und drei Grad dargestellt.

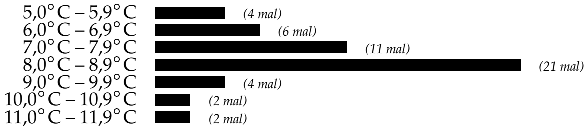
Mittlere Oktobertemperaturen in München
(in Intervallen von 0,3°C)



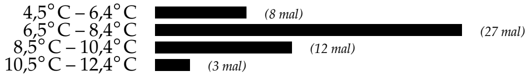
Mittlere Oktobertemperaturen in München
(in Intervallen von 0,5°C)



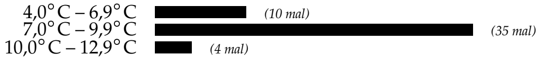
Mittlere Oktobertemperaturen in München
(in Intervallen von einem °C)



Mittlere Oktobertemperaturen in München
(in Intervallen von zwei °C)



Mittlere Oktobertemperaturen in München
(in Intervallen von drei °C)



Wie man sieht, werden die Darstellungen mit zunehmender

Klassenbreite immer übersichtlicher, enthalten aber auch immer weniger Information. Als Kompromiß bietet sich hier etwa die Darstellung mit den 1°C-Intervallen an, eventuell auch die mit einem halben Grad.

Durch die Literatur geistern einige starre Regeln, die für jeden Fall die Anzahl der darzustellenden Klassen aus der Anzahl der vorhandenen Beobachtungswerte berechnen wollen. Oft zitiert (und in einigen Statistiklehrbüchern für Geographen auch angewandt) wird insbesondere die Regel aus

H.A. STURGES: The choice of a class intervall, *J. Amer. Stat. Ass.* **26** (1926), 65–66,

wonach man für die Darstellung von N Datenwerten diese in

$$1 + 3{,}322 \log_{10} N$$

Klassen aufteilen soll. Da $3{,}322 \approx 1 / \log_{10} 2$ ist, kann man dies auch (besser) als $1+\log_2 N$ schreiben. In der Tat: Liest man die (nur eine Seite lange) Arbeit von STURGES, so sieht man, daß diese Regel ihren Hintergrund im von STURGES postulierten Dogma hat, wonach $N = 2^m$ Daten stets in $m + 1$ Klassen zerfallen sollten, wobei die Elementanzahl der i-ten Klasse (mit $i = 1, \ldots, m$) gleich dem Binomialkoeffizient $\binom{i}{m}$ sein sollte. (Was ein Binomialkoeffizient ist, werden wir im übernächsten Kapitel sehen.) Da viele in der Praxis auftretenden Verteilungen nicht einmal ungefähr so aussehen, wie von STURGES postuliert, ist die Regel im allgemeinen offensichtlich unbegründet.

Allerdings sind viele empirische Daten normalverteilt (auch für diesen Begriff muß auf das übernächste Kapitel verwiesen werden), und da die Normalverteilung gut durch Binomialkoeffizienten approximiert werden kann, hat man bei solchen Daten bei Beachtung der Regel von STURGES zumindest eine gewisse Chance, daß das Histogramm einer Stichprobe visuell ungefähr der Verteilung der Grundgesamtheit entspricht; vielleicht war das der (in seiner Arbeit nicht erwähnte) Gedanke hinter der Regel von STURGES. Trotzdem sollte man sie nicht zu ernst nehmen.

Im vorliegenden Beispiel würde die Regel von STURGES auf Grund der fünfzig Beobachtungswerte

$$1 + 3{,}322 \log_{10} 50 \approx 6{,}644 \approx 7$$

Klassen vorschlagen, was der (vernünftigen) Klassenbreite von einem Grad entspricht.

Weitgehend anerkannt werden folgende Regeln:

a) Der besseren Übersichtlichkeit wegen sollten im allgemeinen nicht mehr als etwa zwanzig Klassen dargestellt werden.

b) Die Klassenbreiten müssen gleich sein, da sonst die Höhe der Balken einen völlig falschen Eindruck vermitteln. Allenfalls kann man eventuell verschiedene Klassenbreiten durch verschiedene Balkenbreiten in der Weise kompensieren, daß zumindest die *Fläche* des Balkens wieder den richtigen Eindruck vermittelt.

c) Die Klassengrenzen sollten keine zu „krummen“ Zahlen sein, sondern zu einer übersichtlichen Aufteilung führen.

Es wäre beispielsweise recht künstlich erschienen, wenn bei den Klassenbreiten von einem Grad die erste Klasse von 5,2°C bis 6,2°C gegangen wäre; hier waren ganz- oder zumindest halbzahlige Klassengrenzen angebracht.

§3: Kennzahlen empirischer Verteilungen

Während graphische Darstellungen optisch eine gute Übersicht über einen Datensatz geben können, liefern sie keine objektiven quantifizierbaren Aussagen über Ähnlichkeiten und Nichtähnlichkeiten zweier Datensätze. Für solche Aussagen muß man dem Datensatz Zahlenwerte zuordnen, die ihn mit möglichst geringem Aufwand möglichst gut beschreiben. Die wichtigsten dieser Zahlen sollen in diesem Kapitel erklärt werden.

Wir gehen stets aus von einer Grundgesamtheit $x_1, \ldots, x_N$ aus N nicht notwendigerweise verschiedenen Werten x_i.

1. Lageparameter

Die bekanntesten Beschreibungszahlen sind die verschiedenen Mittelwerte; zusammen mit den noch zu erklärenden Quantilen gehören sie zu den sogenannten *Lageparametern;* man nennt sie auch gelegentlich *Parameter der Zentraltendenz.*

a) Der Modalwert

Unter dem *Modalwert* einer Grundgesamtheit versteht man den am häufigsten angenommenen Wert; im Falle der Daten $2, 3, 1, 3, 4, 5, 3, 2, 3, 4, 7, 2, 3, 1$ wäre das also der fünfmal vorkommende Wert 3.

Der Modalwert hat den Vorteil, daß man ihn für Daten auf beliebigen Skalen bilden kann; auch für Werte, die auf einer Nominalskala gemessen werden, kann man von Modalwerten reden: So ist beispielsweise *Betriebswirtschaftslehre* der Modalwert unter den Fachrichtungen der Mannheimer Studenten.

Für kontinuierliche Werte ist es im allgemeinen von der Meßgenauigkeit, also letztlich von einer mehr oder minder willkürlichen Klasseneinteilung abhängig, welcher Wert zum Modalwert wird, so daß der Modalwert hier weniger aussagekräftig ist. Wir werden später allerdings sehen, daß man zumindest bei theoretischen Verteilungen doch meist in sinnvoller Weise von Modalwerten reden kann.

Gelegentlich bezeichnet man auch Nebenmaxima der Häufigkeitsverteilung als Modalwerte (oder als Nebenmodi); eine Verteilung mit zwei oder mehr Modalwerten heißt dann *bimodal* oder *multimodal.*

Bimodale Verteilungen können ein Hinweis darauf sein, daß es sich bei der betrachteten Verteilung tatsächlich um eine Überlagerungen zweier verschiedener Verteilungen handelt, die man besser getrennt untersucht.

Ein Beispiel dafür ist das weiter unten reproduzierte Diagramm der Größen landwirtschaftlicher Nutzflächen im Burgund aus

G. CHADULE: Initiation aux pratiques statistiques en géographie, *Masson, Paris, Barcelona, Milano,* ^{3}1994, S. 45.

Modalwert ist die Klasse von 0–5 ha, es gibt aber nach dem Minimum bei 20–25 ha auch einen klar erkennbaren Nebenmodus bei 40–50 ha. Der Grund liegt im Nebeneinander zweier verschiedener Agrarsysteme: den meist recht kleinen Weingütern und den deutlich größeren Flächen für den Getreideanbau. Falls man sich genauer für die Landwirtschaft im Burgund interessiert, wird man wohl zweckmäßigerweise beide Gruppen getrennt untersuchen.

b) Der Medianwert

Der Median ist definiert als jener Wert, für den es genauso viele größere wie kleinere Werte gibt. Er ist somit nur für Daten erklärt, die mindestens auf einer Ordinalskala gemessen werden. Allerdings muß es für reine Ordinalskalen bei einer geraden Zahl von Meßwerten nicht unbedingt einen

wohldefinierten Medianwert geben: Ist etwa

$$\begin{aligned} x_1 &= \textit{Kleinstadt} \\ x_2 &= \textit{Millionenstadt} \\ x_3 &= \textit{Dorf} \\ x_4 &= \textit{Großstadt,} \end{aligned}$$

so liegt der Median zwischen *Kleinstadt* und *Großstadt,* jedoch gibt es auf der zugrundeliegenden Skala möglicherweise keinen Wert, der zwischen diesen beiden liegt, oder aber es gibt mehrere solche Werte, etwa *Kreisstadt* und *große Kreisstadt.* Bei einer Grundgesamtheit aus einer *ungeraden* Anzahl von Meßwerten kann dieses Problem offensichtlich nicht auftreten: Hier gibt es immer ein x_i, so daß genausoviele der übrigen Werte kleiner *bzw.* größer als x_i sind.

Beim (in der Praxis häufigsten Fall) einer metrischen Skalen läßt sich auch im geraden Fall das Problem leicht umgehen: Hier gibt es stets zwei Werte x_i und x_j mit $x_i \leq x_j$, so daß genauso viele vor x_i wie nach x_j liegen, und man *definiert* den Median einfach als

$$\frac{x_i + x_j}{2},$$

d.h. als Mittelwert dieser beiden Werte.

Falls man im obigen Beispiel anstelle einer Ordinalskala die metrische Skala der Einwohnerzahlen benutzt hätte, etwa mit

$$x_1 = 16\,555, \quad x_2 = 3\,141\,592, \quad x_3 = 827, \quad x_4 = 319\,439,$$

so wäre $x_3 < x_1 < x_4 < x_2$, in obiger Notation wäre also $x_i = x_1$ und $x_j = x_4$, denn von den vier Werten ist einer, nämlich x_3, kleiner als x_1, und genauso viele, nämlich der eine Wert x_2, größer als x_3. Der Median wäre demnach

$$\frac{x_1 + x_4}{2} = \frac{16\,555 + 319\,438}{2} = 167\,997{,}5\,.$$

Falls es noch einen fünften Wert $x_5 = 23\,456$ gäbe, so wäre dieser der Median, denn wegen

$$x_3 < x_1 < x_5 < x_4 < x_2$$

wären nun jeweils zwei Werte kleiner *bzw.* größer als x_5.

Formal definiert man daher den Median einer Grundgesamtheit aus den metrischen Daten $x_1, \ldots, x_N$ wie folgt: Man sortiere die Daten der Größe nach. Für ungerades $N = 2k + 1$ ist der Median gleich dem $(k + 1)$-ten Wert aus der sortierten Liste, für gerades $N = 2\ell$ ist er gleich dem (arithmetischen) Mittelwert aus dem ℓ-ten und dem $(\ell + 1)$-ten.

Dazu noch ein Beispiel: Wir betrachten wieder die fünfzig mittleren Oktobertemperaturen aus München. Hier ist $N = 50 = 2\ell$ mit $\ell = 25$, der Median ist also das arithmetische Mittel aus dem 25. und dem 26. Temperaturwert, sofern man der Größe nach ordnet. Durch Abzählen in den den Histogrammen findet man, daß beide Werte gleich 8,1°C sind. (Es gibt genau 24 Werte, die kleiner als 8,1°C sind.) Daher ist der Median als arithmetisches Mittel aus 8,1°C und 8,1°C selbst gleich 8,1°C.

Als Randbemerkung sei noch darauf hingewiesen, daß die Daten nicht unbedingt vollständig sortiert werden müssen (obwohl auch das selbst bei sehr großen Datensätzen für heutige Computer kein Problem ist): Um den k-ten Wert

der sortierten Liste aus den unsortierten Daten zu berechnen, wählt man sich irgendeinen Wert x_i, von dem man annimmt, daß er in der Nähe dieses Wertes liegen könnte, und man teilt die Grundgesamtheit auf in die Werte x_j mit $x_j \le x_i$ und die Werte x_ℓ mit $x_\ell > x_i$. Anhand der Elementanzahlen beider Teillisten sieht man, in welcher der beiden der k-te Wert liegt. Diese Teilliste unterteilt man weiter mit einem neuen Schätzwert x_m aus dieser Liste (im Zweifelsfall ist x_m ein zufällig gewähltes Listenelement) *usw.*, bis man schließlich bei einer Teilliste angelangt ist, deren Elemente alle den gleichen Wert haben: Dieser Wert ist der k-te Wert der sortierten Liste. (Das beschriebene Verfahren ist eine Variante des bekannten Sortieralgorithmus *quicksort* und geht wie dieser auf C.A.R. HOARE zurück.)

Mit Blick auf spätere Verallgemeinerungen auf geographische Mediane sei noch auf eine weitere Eigenschaft des Medians hingewiesen: Betrachtet man für ein beliebiges x die Summe der Abstände von x und x_i, wobei diese Abstände natürlich immer positiv zu nehmen sind, so wird dieser Ausdruck minimal, falls man für x den Median einsetzt. Dies ist allerdings keine ganz eindeutige Charakterisierung: Falls bei einer geraden Anzahl von Daten der Median zwischen zwei Werten x_k und x_ℓ liegt, überlegt man sich leicht, daß jedes x zwischen x_k und x_ℓ diese Eigenschaft hat, daß die Summe aber größer wird, sobald x das Intervall zwischen x_k und x_ℓ verläßt.

c) Das arithmetische Mittel

Dieser bekannteste Mittelwert ist nur für metrische Daten erklärt; für eine Grundgesamtheit $x_1, \ldots, x_N$ ist er gleich der durch N dividierten Summe der x_i.

Da wir es im folgenden häufig mit solchen und ähnlichen Summen zu tun haben, empfiehlt es sich, eine abkürzende Bezeichnung einzuführen: Wir schreiben die Summe der x_i als

$$x_1 + \cdots + x_N = \sum_{i=1}^{n} x_i \ .$$

Das Summenzeichen $\sum$, ein stilisiertes großes griechisches Sigma (Σ), soll dabei an den Anfangsbuchstaben von **S**umme erinnern.

Wir wollen diese Schreibweise nicht nur für die Summe der x_i verwenden, sondern auch beliebige Summen in analoger Weise schreiben. Beispielsweise ist also

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 \ ,$$

aber auch

$$\sum_{i=1}^{5} x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2$$

oder

$$\sum_{i=1}^{5} i^3 = 1^3 + 2^3 + 3^3 + 4^3 + 5^3 \ .$$

Das arithmetische Mittel von x_1 bis x_N ist demnach

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \ ,$$

im obigen Beispiel der $N = 5$ Einwohnerzahlen also

$$\begin{aligned}\frac{1}{5} \sum_{i=1}^{5} x_i &= \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} \\ &= \frac{16\,555 + 3\,141\,592 + 827 + 319\,439 + 23\,456}{5} \\ &= 700\,373{,}8 \ .\end{aligned}$$

d) Vergleich der drei Mittelwerte

Bei einer metrischen Skala lassen sich alle drei bislang betrachteten Mittelwerte definieren; natürlich sind sie im allgemeinen verschieden. Einige ihrer wesentlichen Stärken und Schwächen sind:

- Der *Modalwert* gibt nur den häufigsten Wert an, sagt aber nichts aus über die Verteilung der Werte. Er ist, vor allem bei einem stark ausgeprägten Häufigkeitsmaximum, nützlich, um typische Elemente zu charakterisieren. Er kann für jeden Skalentyp erklärt werden.
- Der *Medianwert* gibt die Lage der Verteilung am genauesten an, da genauso viele Werte links *bzw.* rechts von ihm liegen. Er ist unempfindlich gegen Ausreißer, da es bei extremen Werten nur darauf ankommt, ob sie rechts oder links liegen, nicht aber darauf *wie weit* rechts oder links sie liegen. Er ist im allgemeinen der beste Mittelwertparameter für eine empirische Verteilung, ist allerdings nur erklärt, wenn die Daten zumindest auf einer Ordinalskala gemessen werden.
- Das *arithmetische Mittel* hat den Vorteil der leichten Berechenbarkeit und der guten mathematischen Handhabbarkeit. Insbesondere ist das arithmetische Mittel eine guter Ausgangspunkt für die Definition weiterer

statistischer Kennzahlen. Sein wesentlicher Nachteil ist die starke Abhängigkeit von extremen Datenwerten: Bereits ein einziger Ausreißer kann das arithmetische Mittel wesentlich verändern. Es ist im übrigen nur definiert für Daten, die auf einer metrischen Skala gemessen werden können.

Als Beispiel für das gegenseitige Verhältnis der drei Werte sei auf das Diagramm aus

D. HUFF: How to Lie with Statistics, *W.W. Norton, New York,* 1954

auf der nächsten Seite verwiesen, das alle drei Werte für die Einkommensverteilung in einem fiktiven Unternehmen berechnet. (Dem Alter des Buchs entsprechend handelt es sich um Jahresgehälter in Dollar.)

e) Gewichtete Mittelwerte

Angenommen, wir möchten die durchschnittliche Studienzeit des Absolventenjahrgans 1995 an der Universität Mannheim berechnen. Bekannt seien die entsprechenden durchschnittlichen Studienzeiten der einzelnen Studiengänge.

Wegen der unterschiedlichen Studentenzahlen in den verschiedenen Fachrichtungen ist es sicher nicht sinnvoll, das arithmetische Mittel oder auch den Median oder Modalwert der Einzeldaten zu nehmen; das korrekte arithmetische Mittel ist

$$d = \frac{\text{Summe der Studienzeiten aller Absolventen}}{\text{Anzahl aller Absolventen}} .$$

Entsprechend ist das arithmetische Mittel d_i für den i-ten Studiengang

$$d_i = \frac{\text{Summe der Studienzeiten aller Absolventen des } i\text{-ten Studiengangs}}{\text{Anzahl aller Absolventen des } i\text{-ten Studiengangs}} .$$

Bezeichnen wir den Zähler dieses Bruchs mit z_i und den Nenner mit a_i, so ist die Summe der z_i gleich der Summe aller Studienzeiten von Absolventen des Jahrgangs 1995; entsprechend ist die Summe der a_i gleich der Gesamtzahl der Absolventen. Daher ist

$$d = \frac{\sum z_i}{\sum a_i} = \frac{\sum d_i \cdot a_i}{\sum a_i} ,$$

wobei sich die Summe jeweils über die sämtlichen Studiengänge der Universität Mannheim erstreckt.

Mit anderen Worten: Wir haben nicht den gewöhnlichen Mittelwert der d_i gebildet, sondern jedes d_i mit dem *Gewicht* a_i multipliziert, um so den „großen“ Studiengängen ein größeres Gewicht zu geben. Dividiert wurde dann am Ende natürlich nicht durch die Anzahl der Studiengänge, sondern durch die Anzahl der Absolventen, das heißt, durch die Summe der Gewichte.

Diese Konstruktion läßt sich natürlich verallgemeinern: Gegeben seien eine Grundgesamtheit bestehend aus den Daten $x_1, \ldots, x_N$ sowie *Gewichte* $g_1, \ldots, g_N$. Das *gewichtete Mittel* der x_i mit den Gewichten g_i ist der Wert

$$\frac{\sum_{i=1}^{N} g_i \cdot x_i}{\sum_{i=1}^{N} g_i} .$$

Wie im Beispiel kann ein gewichtetes Mittel dazu dienen, aus bekannten Durchschnittswerten von Teilgesamtheiten auf den Durchschnitt der Grundgesamtheit zu schließen; die Gewichte können aber auch willkürlich verteilt werden: Falls man etwa die vorhandenen Daten x_i einer Stichprobe für teils mehr, teils weniger zuverlässig hält, so wird man den zuverlässigeren Daten ein höheres Gewicht geben und anstelle des gewöhnlichen das gewichtete arithmetische Mittel bilden.

f) Quantilen

Hier handelt es sich um eine Verallgemeinerung des Medians: Der Median war so definiert, daß er, wenn man der Größe nach ordnet, der in der Mitte stehende Wert ist; im wesentlichen hat er also die Eigenschaft, daß jeweils 50% der Daten über und unter ihm liegen. (Das ist natürlich nicht ganz exakt, da man auch noch die Werte berücksichtigen muß, die gleich dem Median sind; außerdem ist der Median allein durch diese Vorschrift, wie wir gesehen haben, noch nicht eindeutig bestimmt.)

Entsprechend soll die *q%-Quantile* oder *q-te Perzentile* im wesentlichen so definiert sein, daß q% der Werte darunter und $(100 - q)$% darüber liegen. Natürlich treten auch hier die gleichen Probleme wie beim Median auf; im Gegensatz zu dort gibt es aber keine allgemein akzeptierte exakte Definition der Quantilen: Die verschiedenen Lehrbücher wie auch die Statistiksysteme für Computer benutzen jeweils verschiedene Konventionen. In der Praxis sind die Unterschiede zwischen den nach diesen Konventionen berechneten Werten natürlich gering; wir wollen uns daher mit *einer* einfachen und relativ weit verbreiteten Definition begnügen:

Die Grundgesamtheit sei wieder $x_1, \ldots, x_N$; wir wollen der Einfachheit halber annehmen, daß sie bereits der Größe nach sortiert sei mit x_1 als dem kleinsten und x_N als dem größten Wert.

Im Spezialfall $N = 100$ ist dann alles ganz einfach: Die kleinsten $q\%$ der Werte sind $x_1, \ldots, x_q$, die $(100-q)\%$ größten sind $x_{q+1}, \ldots, x_{100}$, die $q\%$-Quantile liegt also zwischen x_q und x_{q+1}. Wir könnten also das arithmetische Mittel dieser beiden Werte als q-te Perzentile definieren; tatsächlich nehmen wir das *gewichtete Mittel*

$$\frac{q \cdot x_q + (100-q) \cdot x_{q+1}}{100},$$

wir teilen also auch noch das Intervall von x_q bis x_{q+1} im Verhältnis q zu $100-q$.

Entsprechend gehen wir auch im allgemeinen Fall vor: Wir rechnen das Intervall von 1 bis N um auf das von 1 bis 100 und betrachten den Wert $k = q \cdot N \,/\, 100$. Falls das eine ganze Zahl ist, falls die Division also aufgeht, können wir wie oben vorgehen und erklären die $q\%$-Quantile als

$$M_q = \frac{q \cdot x_k + (100-q) \cdot x_{k+1}}{100}.$$

Falls k keine ganze Zahl ist, betrachten wir die nächstgrößere ganze Zahl ℓ und setzen

$$M_q = x_\ell .$$

Besonders wichtige Perzentilen sind

- Der *Median* M_{50} (Man überlege sich, daß M_{50} bei der hier verwendeten Quantildefinition in der Tat *exakt* gleich dem oben definierten Median ist.)
- Die *Quartilen* $Q_1 = M_{25}$ und $Q_3 = M_{75}$
- Die *Dezilen* $D_1 = M_{10}$ und $D_9 = M_{90}$.

Berechnen wir als Beispiel diese Quantilen für die Münchner Oktobertemperaturen! Den Median $M_{50} = 8{,}1°\text{C}$ kennen wir bereits, beginnen wir also mit den Quartilen. Da es $N = 50$ Meßwerte gibt, müssen wir dazu die Quotienten

$$k_1 = \frac{25 \cdot 50}{100} = 12{,}5 \quad \text{und} \quad k_2 = \frac{75 \cdot 50}{100} = 37{,}5$$

betrachten; da beide nicht ganzzahlig sind, ist

$$Q_1 = M_{25} = x_{13} = 7{,}2°C \text{ und } Q_3 = M_{75} = x_{38} = 8{,}5°C ,$$

wobei die x_i wieder *geordnet* angenommen wurden und aus dem Histogramm im letzten Kapitel abgelesen werden können.

Für die Dezilen $D_1 = M_{10}$ und $D_9 = M_{90}$ sind

$$k_3 = \frac{10 \cdot 50}{100} = 5 \quad \text{und} \quad k_4 = \frac{90 \cdot 50}{100} = 45$$

beide ganzzahlig, also ist

$$D_1 = \frac{5 \cdot x_5 + 95 \cdot x_6}{100} = \frac{5 \cdot 5{,}1°\text{C} + 95 \cdot 5{,}3°\text{C}}{100} = 5{,}29°\text{C}$$

und

$$D_9 = \frac{95 \cdot x_{45} + 5 \cdot x_{46}}{100} = \frac{95 \cdot 9{,}8°\text{C} + 5 \cdot 9{,}9°\text{C}}{100} = 9{,}805°\text{C} .$$

Wir können daher sagen, daß die mittlere Münchner Oktobertemperatur in 50% aller Fälle zwischen 7,2°C und 8,5°C liegt und in 80% aller Fälle zwischen 5,3°C und 9,8°C.

2. Streuungsparameter

Die letzte Bemerkung bringt uns bereits weg von den Lageparametern: Um wirklich etwas über eine Verteilung zu wissen, müssen wir natürlich außer ihrem Mittelwert (egal nach welcher Definition) auch noch etwas über ihre Ausbreitung wissen: Eine Durchschnittstemperatur für eine klimatisch begünstigte Insel bedeutet etwas ganz anderes als dieselbe Durchschnittstemperatur in einem extremen Kontinentalklima.

Ein erstes Streuungsmaß wäre etwa die *Spannweite* der Verteilung, die als Differenz zwischen dem Maximum und dem Minimum der x_i definiert ist. Da sie somit ausschließlich von der „Ausreißern" der Meßreihe abhängt, spielt sie keine große Rolle.

Als bessere Beschreibungen für die Streuung der Verteilung betrachtet man Maße für die mittlere Abweichung vom arithmetischen Mittel $\overline{x}$ der Grundgesamtheit $x_1, \ldots, x_N$.

Man könnte versucht sein, als mittlere Abweichung den Mittelwert der Abweichungen $\overline{x} - x_i$ zu definieren, aber dieser Wert ist natürlich nach Definition des arithmetischen Mittels gleich Null:

$$\frac{1}{N}\sum_{i=1}^{N}(\overline{x}-x_i) = \frac{1}{N}\sum_{i=1}^{N}\overline{x} - \frac{1}{N}\sum_{i=1}^{N}x_i = \overline{x}-\overline{x} = 0\,.$$

Wenn schon müßte man also $\overline{x} - x_i$ durch den stets nichtnegativen Abstand $|\overline{x} - x_i|$ ersetzen, wobei

$$|x| = \begin{cases} x & \text{falls } x \geq 0 \\ -x & \text{falls } x < 0 \end{cases}$$

den *Betrag* von x bezeichnet. In der Tat wird das entsprechende Maß

$$\frac{1}{N}\sum_{i=1}^{N}|\overline{x}-x_i|$$

gelegentlich benutzt und als *mittlere Abweichung* oder *durchschnittliche Abweichung* der Verteilung bezeichnet. Trotz dieser anschaulichen Interpretation spielt sie in der Praxis nur eine geringe Rolle, da sie mathematisch nur schlecht handhabbar ist.

Von erheblich größerer Bedeutung in der Statistik ist allerdings die mittlere *quadratische* Abweichung oder *Varianz* der Verteilung, also der Wert

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\overline{x}-x_i)^2$$

sowie seine (nichtnegative) Quadratwurzel

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{x}-x_i)^2}\,,$$

die *Standardabweichung*.

Diese Standardabweichung ist, wie wir sehen werden, einer der wichtigsten Parameter einer Verteilung überhaupt; insbesondere kann sie beispielsweise benutzt für die Charakterisierung des Fehlers von Meßdaten. Als Übungsaufgabe für die aus der Schule bekannte Berechnung von Extremwerten kann man sich überlegen, daß die Funktion

$$S(x) = \frac{1}{N}\sum_{i=1}^{N}(x-x_i)^2$$

genau für $x = \overline{x}$ ihr Minimum hat; das arithmetische Mittel steht also in gleicher Beziehung zur Varianz wie der Median zur mittleren Abweichung; denn der Median minimisiert ja die Funktion

$$A(x) = \frac{1}{N}\sum_{i=1}^{N} |x - x_i| \ .$$

Um Daten miteinander vergleichbar zu machen und um statistische Tabellenwerke benutzen zu können, ist es oft nützlich, sie zu *normalisieren:* Sind $x_1, \ldots, x_N$ die gegebenen Daten mit arithmetischem Mittel $\overline{x}$ und Standardabweichung σ, so bezeichnet man die Werte $y_1, \ldots, y_N$ mit

$$y_i = \frac{x_i - \overline{x}}{\sigma}$$

als die entsprechenden normalisierten Daten. Sie haben den Mittelwert

$$\overline{y} = \frac{1}{N\sigma}\sum_{i=1}^{N}(x_i - \overline{x}) = 0$$

und die Varianz

$$\sigma_y^2 = \frac{1}{N\sigma^2}\sum_{i=1}^{N}(x_i - \overline{x})^2 = 1 \ ,$$

also auch die Standardabweichung 1.

3. Höhere Momente

Analog zur Varianz, der mittleren *quadratischen* Abweichung

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

kann man auch die mittlere *kubische* Abweichung

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^3$$

und die entsprechenden Ausdrücke zu höheren Potenzen betrachten. Es hat sich allerdings eingebürgert, dies nicht für die Daten selbst, sondern für ihre Normalisierungen zu tun; man betrachtet also die Werte

$$\mu_n = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{\sigma}\right)^n = \frac{1}{N\sigma^n}\sum_{i=1}^{N}(x_i - \overline{x})^n \ ,$$

die gelegentlich auch als n-te zentrale Momente bezeichnet werden. Ihre Werte hängen stark davon ab, ob n gerade ist oder ungerade: Für ungerades n heben sich $(x_i - \overline{x})^n$ für ein $x_i < \overline{x}$ und $(x_j - \overline{x})^n$ für ein $x_j > \overline{x}$ gegenseitig weg, falls sich $x_i - \overline{x}$ und $x_j - \overline{x}$ nur durch das Vorzeichen unterscheiden. Bei geradem n sind dagegen alle Summanden größer oder gleich Null, so daß sich nie etwas wegheben kann.

Zur Interpretation der μ_n betrachten wir zunächst die Potenzfunktion x^n: Für $x < 1$ ist x^n kleiner als x; beispielsweise ist $0{,}1^3 = 0{,}001$. Für $x > 1$ dagegen ist x^n *größer* als x; beispielsweise ist $5^3 = 125$. Die Summe der $\big((x_i - \overline{x})/\sigma\big)^n$ wird also vor allem von den großen Abweichungen vom Mittelwert bestimmt, und dies in um so stärkerem Maße, je größer n ist.

Im ungeraden Fall ist μ_n also *positiv*, wenn die großen Abweichungen vom Mittelwert vorzugsweise rechts davon liegen, während links eher kleine Abweichungen zu finden sind. Im umgekehrten Fall ist μ_n *negativ,* und wenn sich die Abweichungen nach links und nach rechts die Waage halten, ist $\mu_n = 0$.

Speziell für $n = 3$ bezeichnet man μ_3 daher als *Schiefe* der Verteilung. Eine Verteilung mit $\mu_3 = 0$ heißt *symmetrisch,* Verteilungen mit negativem μ_3 heißen *linksschief* und solche mit positivem μ_3 *rechtsschief.*

Die Grundgesamtheit mit

$$x_1 = 1,\ x_2 = 2,\ x_3 = 3,\ x_4 = 4 \quad \text{und} \quad x_5 = 5$$

wäre demnach also symmetrisch, denn die Summe der kubischen Abweichungen vom Mittelwert $\overline{x} = 3$ ist

$$(-2)^3 + (-1)^3 + 0^3 + 1^3 + 2^3 = 0\,,$$

und damit ist auch $\mu_3 = 0$.

Wäre dagegen

$$x_1 = 0,\ x_2 = 1,\ x_3 = x_4 = 3 \quad \text{und} \quad x_5 = 8\,,$$

so wäre wieder $\overline{x} = 3$, aber die Summe der kubischen Abweichungen wäre nun

$$(-3)^3 + (-2)^3 + 0^3 + 0^3 + 5^3 = 90\,,$$

durch die große Abweichung zwischen x_5 und dem Mittelwert dominiert also die rechte Seite. Zur Berechnung der Schiefe müssen wir 90 noch durch fünf und durch die dritte Potenz der Standardabweichung dividieren:

$$\sigma^2 = \frac{(-3)^2 + (-2)^2 + 5^2}{5} = \frac{38}{5} = 7{,}6\,,$$

also

$$\sigma = \sqrt{7{,}6} \approx 2{,}7568 \quad \text{und} \quad \mu_3 = \frac{90}{5\sigma^3} \approx 0{,}8591\,.$$

Bei geradem n kann keiner der Summanden $(x_i - \overline{x})^n$ negativ sein, das Moment μ_n ist also stets positiv (oder Null, falls alle x_i gleich sind). μ_n wird um so größer, je ausgeprägter die *großen* Abweichungen vom Mittelwert sind, je deutlicher also die Verteilung in einer oder beiden Richtungen einen ausgeprägten „Schwanz“ hat. Das Moment μ_4 wird als *Wölbung, Exzeß* oder *Kurtosis* der Verteilung bezeichnet.

Berechnen wir als etwas ausführlicheres Beispiel zum Schluß noch Mittelwert, Standardabweichung, Schiefe und Wölbung der 50 mittleren Oktobertemperaturen aus München: Auf der nächsten Seite sind, jeweils am Anfang der fünfzig Zeilen, nochmals alle Werte zu sehen; da ihre Summe (wie in der letzten Zeile eingetragen) gleich 398,7°C ist, berechnet sich das arithmetische Mittel zu

$$\overline{x} = \frac{398{,}7}{50} = 7{,}974°\text{C}\,.$$

Die Differenzen zwischen den einzelnen Werten und diesem Mittelwert stehen in der zweiten Spalte, daneben stehen die zweiten, dritten und vierten Potenzen davon.

Diese Abweichungen *bzw.* ihre Potenzen müssen aufsummiert werden; die Ergebnisse sind in der letzten Zeile zu sehen. (Da alle Werte in gerundeter Form angegeben sind, die Rechnung aber mit größerer Genauigkeit durchgeführt

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^3$	$(x_i - \overline{x})^4$
8,5	0,526	0,277	0,146	0,0765
6,3	-1,674	2,802	-4,691	7,8528
9,8	1,826	3,334	6,088	11,1174
7,3	-0,674	0,454	-0,306	0,2064
5,2	-2,774	7,695	-21,346	59,2143
8,1	0,126	0,016	0,002	0,0003
7,0	-0,974	0,949	-0,924	0,9000
6,4	-1,574	2,477	-3,900	6,1379
5,3	-2,674	7,150	-19,120	51,1265
6,6	-1,374	1,888	-2,594	3,5641
11,0	3,026	9,157	27,708	83,8447
5,4	-2,574	6,625	-17,054	43,8970
10,8	2,826	7,986	22,569	63,7806
8,5	0,526	0,277	0,146	0,0765
8,1	0,126	0,016	0,002	0,0003
8,5	0,526	0,277	0,146	0,0765
7,2	-0,774	0,599	-0,464	0,3589
8,3	0,326	0,106	0,035	0,0113
9,3	1,326	1,758	2,331	3,0915
8,0	0,026	0,000	0,000	0,0000
6,4	-1,574	2,477	-3,900	6,1379
8,4	0,426	0,181	0,077	0,0329
8,3	0,326	0,106	0,035	0,0113
8,3	0,326	0,106	0,035	0,0113
8,9	0,926	0,857	0,794	0,7353
5,5	-2,474	6,121	-15,143	37,4627
8,2	0,226	0,051	0,012	0,0026
8,5	0,526	0,277	0,146	0,0765
7,0	-0,974	0,949	-0,924	0,9000
8,0	0,026	0,000	0,000	0,0000
7,7	-0,274	0,075	-0,021	0,0056
11,9	3,926	15,413	60,513	237,5750
9,6	1,626	2,644	4,299	6,9901
8,1	0,126	0,016	0,002	0,0003
8,4	0,426	0,181	0,077	0,0329
6,1	-1,874	3,512	-6,581	12,3333
7,2	-0,774	0,599	-0,464	0,3589
8,7	0,726	0,527	0,383	0,2778
10,2	2,226	4,955	11,030	24,5528
7,8	-0,174	0,030	-0,005	0,0009
6,5	-1,474	2,173	-3,203	4,7205
7,5	-0,474	0,225	-0,106	0,0505
9,9	1,926	3,709	7,144	13,7602
8,2	0,226	0,051	0,012	0,0026
7,4	-0,574	0,329	-0,189	0,1086
7,6	-0,374	0,140	-0,052	0,0196
8,0	0,026	0,000	0,000	0,0000
7,9	-0,074	0,005	-0,000	0,0000
8,1	0,126	0,016	0,002	0,0003
398,7	0,000	99,574	42,746	681,4936

wurde, ergeben sich beim Nachrechnen kleinere Abweichungen von den angegebenen Ergebnissen.) Damit berechnet sich die Varianz zu

$$\sigma^2 = \frac{99{,}574}{50} = 1{,}9915$$

und die Standardabweichung zu

$$\sigma = \sqrt{1{,}991} = 1{,}411\,.$$

Die Schiefe ist demnach gleich

$$\mu_3 = \frac{42{,}746}{50\sigma^3} = 0{,}304$$

und die Wölbung

$$\mu_4 = \frac{681{,}4936}{50\sigma^4} = 3{,}437\,.$$

4. Schätzung von Kennzahlen einer Grundgesamtheit aufgrund einer Stichprobe

Angenommen, wir wollen das mittlere Monatseinkommen eines Bundesbürgers schätzen. Da es unrealistisch wäre, alle Einzeleinkommen zu bestimmen und daraus dann den Mittelwert zu berechnen, muß statt dessen eine Stichprobe genügen. Trotzdem interessieren im Endeffekt natürlich nicht die Kennzahlen der zufällig gewählten Stichprobe, sondern die der Gesamtbevölkerung, die aufgrund der Stichprobe geschätzt werden soll.

In mathematischer Formulierung haben wir also eine Grundgesamtheit bestehend aus den Werten $x_1, \ldots, x_N$

und daraus eine Stichprobe $u_1 = x_{\nu_1}, \ldots, u_M = x_{\nu_M}$, wobei die M Indizes $\nu_1, \ldots, \nu_M$ zufällig aus der Menge $1, \ldots, N$ aller Indizes gewählt werden.

a) **Schätzung des arithmetischen Mittels**

Das arithmetische Mittel

$$\widehat{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

der Grundgesamtheit wird offensichtlich am besten geschätzt durch das arithmetische Mittel

$$\overline{x} = \frac{1}{M} \sum_{j=1}^{M} x_{\nu_j} = \frac{1}{M} \sum_{j=1}^{M} u_j$$

der Stichprobe; gesucht ist ein Maß für die Zuverlässigkeit dieser Schätzung.

Als ein solches Maß bietet sich die *Standardabweichung* von $\overline{x}$ an. Was darunter zu verstehen ist und wie man sie berechnet wird vielleicht am klarsten, wenn wir zunächst nur den Fall einer Stichprobe vom Umfang *zwei* betrachten; der Kürze halber bezeichnen wir die beiden Werte mit $u = x_{\nu_1}$ und $v = x_{\nu_2}$.

Dann ist

$$\widehat{x} - \overline{x} = \widehat{x} - \frac{u+v}{2} = \frac{\widehat{x} - u}{2} + \frac{\widehat{x} - v}{2}$$

und

$$(\widehat{x} - \overline{x})^2 = \frac{(\widehat{x} - u)^2}{4} + \frac{(\widehat{x} - u)(\widehat{x} - v)}{2} + \frac{(\widehat{x} - v)^2}{4} .$$

Da u und v jeweils zufällig unter den x_i ausgewählt wurden, lassen sich die quadratischen Abweichungen $(\widehat{x} - u)^2$ und $(\widehat{x} - v)^2$ am besten abschätzen durch die *mittlere* quadratische Abweichung und damit durch die Varianz $\widehat{\sigma}^2$ der *Grundgesamtheit.* Das gemischte Produkt $(\widehat{x} - u)(\widehat{x} - v)$ sollte im Mittel verschwinden, denn wenn u und v unabhängig voneinander zufällig aus der Grundgesamtheit ausgewählt werden, sollten sich die Fälle, in denen u und v auf derselben *bzw.* auf verschiedenen Seiten des Mittelwerts liegen gegenseitig aufheben. Also wäre $(\widehat{x} - \overline{x})^2$ zu schätzen als

$$\frac{\sigma^2}{4} + \frac{0}{2} + \frac{\sigma^2}{4} = \frac{\sigma^2}{2}$$

und der Fehler des Mittelwerts entsprechend als $\sigma/\sqrt{2}$.

Genauso kann man nun auch bei einer Stichprobe

$$u_1, \ldots, u_M$$

vom Umfang M vorgehen: Hier ist

$$\widehat{x} - \overline{x} = \widehat{x} - \frac{1}{M} \sum_{i=1}^{M} u_i = \frac{1}{M} \sum_{i=1}^{M} (\widehat{x} - u_i) .$$

Wieder können wir beim Quadrieren die gemischten Terme $(\widehat{x} - u_i)(\widehat{x} - u_j)$ für $i \neq j$ als Null schätzen und die Quadrate $(\widehat{x} - u_i)^2$ als $\widehat{\sigma}$, die quadratische Abweichung $(\widehat{x} - \overline{x})^2$ also durch $\widehat{\sigma}^2/M$.

Der *Fehler* des Mittelwerts kann also geschätzt werden auf $\widehat{\sigma}/\sqrt{M}$. Um ihn auf die Hälfte zu reduzieren, braucht man somit eine viermal so große Stichprobe, um ihn auf ein Zehntel zu reduzieren sogar eine hundertmal so große.

b) Schätzung der Standardabweichung

Das obige Ergebnis hat den Nachteil, daß wir die Standardabweichung $\widehat{\sigma}$ der Grundgesamtheit nicht kennen. Als Schätzung mag sich auf den ersten Blick die Standardabweichung σ der Stichprobe anbieten, aber dies ist keine gute Wahl, denn während

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(\widehat{x} - x_i)^2$$

die mittlere quadratische Abweichung vom Mittelwert der Grundgesamtheit mißt, ist

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(\overline{x} - u_i)^2$$

die mittlere quadratische Abweichung vom Mittelwert der *Stichprobe,* von dem wir gerade gesehen haben, daß er im allgemeinen von $\widehat{x}$ abweicht.

Eine sinnvolle Schätzung von $\widehat{\sigma}^2$ aufgrund der Stichprobe wäre also

$$\overline{\sigma} = \frac{1}{M}\sum_{i=1}^{M}(\widehat{x} - u_i)^2 .$$

Dieser Ausdruck läßt sich trotz des unbekannten Mittelwert $\widehat{x}$ recht gut schätzen: Zunächst ist

$$\widehat{x} - u_i = (\widehat{x} - \overline{x}) + (\overline{x} - u_i)$$

und damit

$$(\widehat{x} - u_i)^2 = (\widehat{x} - \overline{x})^2 + 2(\widehat{x} - \overline{x})(\overline{x} - u_i) + (\overline{x} - u_i)^2 .$$

Hier können wir $(\widehat{x} - \overline{x})^2$, wie wir gerade gesehen haben, als $\widehat{\sigma}^2/M$ schätzen, $\overline{\sigma}$ also durch

$$\begin{aligned}
&\frac{1}{M}\Big(\sum_{i=1}^{M}\frac{\widehat{\sigma}^2}{M} + 2\cdot\sum_{i=1}^{M}(\widehat{x} - \overline{x})(\overline{x} - u_i) + \sum_{i=1}^{M}(\overline{x} - u_i)^2\Big) \\
&= \frac{\widehat{\sigma}^2}{M} + 2\cdot\left(\frac{\widehat{x} - \overline{x}}{M}\right)\sum_{i=1}^{M}(\overline{x} - u_i) + \frac{1}{N}\sum_{i=1}^{M}(\overline{x} - u_i)^2 .
\end{aligned}$$

Die letzte Summe ist gerade gleich der Varianz σ^2 der Stichprobe, denn deren arithmetisches Mittel war ja gerade $\overline{x}$, und genau deshalb verschwindet auch die Summe der $\overline{x}-u_i$. Damit vereinfacht sich obiger Ausdruck zu

$$\frac{\widehat{\sigma}^2}{M} + \sigma^2 ,$$

und da wir $\widehat{\sigma}$ als $\overline{\sigma}$ schätzen wollen, erhalten wir die Beziehung

$$\overline{\sigma}^2 = \frac{\overline{\sigma}^2}{M} + \sigma^2$$

oder

$$\overline{\sigma}^2 - \frac{\overline{\sigma}^2}{M} = \frac{M-1}{M}\overline{\sigma}^2 = \sigma^2 .$$

Somit ist

$$\begin{aligned}
\overline{\sigma}^2 &= \frac{M}{M-1}\sigma^2 = \frac{M}{M-1}\cdot\frac{1}{M}\sum_{i=1}^{M}(\overline{x} - u_i)^2 \\
&= \frac{1}{M-1}\sum_{i=1}^{M}(\overline{x} - u_i)^2
\end{aligned}$$

und

$$\overline{\sigma} = \sqrt{\sum_{i=1}^{M} \frac{(\overline{x} - u_i)^2}{M-1}}\,.$$

Um also die Standardabweichung einer Grundgesamtheit zu schätzen, darf man nicht die Standardabweichung

$$\sigma = \sqrt{\sum_{i=1}^{M} \frac{(\overline{x} - u_i)^2}{M}}$$

der *Stichprobe* nehmen, sondern muß den etwas größeren Wert

$$\overline{\sigma} = \sqrt{\sum_{i=1}^{M} \frac{(\overline{x} - u_i)^2}{M-1}}$$

berechnen. Bei großem Stichprobenumfang M fällt dieser Unterschied nicht sonderlich ins Gewicht, bei kleinem M kann er den Wert aber spürbar vergrößern.

5. Zweidimensionale Mittelwerte

Die wichtigsten zweidimensionalen Verteilungen für Geographen sind Verteilungen auf der Erdoberfläche. Wir wollen dabei annehmen, daß das betrachtete Gebiet hinreichend klein sei, so daß man die Erdkrümmung vernachlässigen kann und jeden Punkt mit hinreichender Genauigkeit durch seine x- und y-Koordinate auf einer geeigneten Karte beschreiben kann.

Besteht eine Verteilung aus den Datenpaaren

$$(x_1, y_1), \ldots, (x_N, y_N)\,,$$

so berechnen wir ihr arithmetisches Mittel *komponentenweise* als

$$\left(\frac{1}{N}\sum_{i=1}^{N} x_i,\ \frac{1}{N}\sum_{i=1}^{N} y_i\right).$$

Entsprechend ist das *gewichtete Mittel* mit Gewichten $g_1, \ldots, g_N$ das Paar

$$\left(\frac{\sum_{i=1}^{N} g_i \cdot x_i}{\sum_{i=1}^{N} g_i},\ \frac{\sum_{i=1}^{N} g_i \cdot y_i}{\sum_{i=1}^{N} g_i}\right).$$

Ein Beispiel für ein gewichtetes zweidimensionales Mittel eines Landes oder Gebiets ist der *Bevölkerungsschwerpunkt:* Von der Idee her ist das das arithmetische Mittel der Koordinaten aller Einwohner des Gebiets. In dieser Form ist es natürlich unrealistisch, diesen Punkt berechnen zu wollen, schon weil sich die Positionen der einzelnen Personen ständig ändern. Eine realistische Annäherung besteht darin, daß man alle Gemeinden des betreffenden Gebiets betrachtet und das mit den Einwohnerzahlen gewichtete Mittel ihrer Koordinaten berechnet. Falls selbst das noch zu kompliziert ist, kann man auch noch größere Einheiten betrachten, beispielsweise könnte man den Bevölkerungsschwerpunkt von Brasilien grob dadurch abschätzen, daß man das Land in fünf Regionen aufteilt, jeder Region die Koordinaten eines irgendwie gewählten Mittelpunkts zuordnet, und dann das mit den Einwohnerzahlen der Regionen gewichtete Mittel dieser fünf Punkte bildet.

Einwohnerzahlen und Mittelpunkte der fünf Regionen sind aus der folgenden Tabelle ersichtlich:

Region	Mittelpunkt	Bevölkerung 1960	Bevölkerung 1970
N	*(1250 km, 3000 km)*	2,6 Mio.	3,6 Mio.
NO	*(3400 km, 2630 km)*	22,4 Mio.	28,7 Mio.
MW	*(2250 km, 2000 km)*	3,0 Mio.	5,2 Mio.
SO	*(2750 km, 1400 km)*	31,1 Mio.	40,3 Mio.
S	*(2250 km, 700 km)*	11,9 Mio.	16,7 Mio.

Damit wäre also die in Kilometern ausgedrückte x-Koordinate des aufgrund dieser Angaben geschätzten Bevölkerungsschwerpunkts 1960

$$\frac{2{,}6 \cdot 1250 + 22,4 \cdot 3400 + 3{,}0 \cdot 2250 + 31,1 \cdot 2750 + 11,9 \cdot 2250}{2{,}6 + 22{,}4 + 3{,}0 + 31{,}1 + 11,9} = \frac{198460}{71} \approx 2795{,}2\,.$$

Entsprechend berechnet man für die y-Koordinate

$$\frac{2{,}6 \cdot 3000 + 22,4 \cdot 2630 + 3{,}0 \cdot 2000 + 31,1 \cdot 1400 + 11,9 \cdot 700}{2{,}6 + 22{,}4 + 3{,}0 + 31{,}1 + 11,9} = \frac{124582}{71} \approx 1754{,}7\,,$$

der Bevölkerungsschwerpunkt lag 1960 also etwa bei

$$(2795{,}2\,km,\ 1754{,}7\,km)\,.$$

(Eigentlich hätten wir natürlich alle Gewichte noch mit einer Million multiplizieren müssen, aber da das den Zähler und den Nenner in gleicher Weise betroffen hätte, bleibt das Ergebnis unbeeinflußt.)

Für 1970 muß mit den Einwohnerzahlen aus der letzten Spalte als Gewichten gerechnet werden; hier ist die Summe der Gewichte gleich 94,5 Millionen. Die entsprechenden Rechnungen wie oben ergeben einen Bevölkerungsschwerpunkt von

$$(2774{,}4\,km,\ 1743{,}8\,km)\,.$$

Er hat sich also in den zehn Jahren zwischen 1960 und 1970 in südwestlicher Richtung verlagert; mit Hilfe des Satzes von PYTHAGORAS rechnet man leicht nach, daß die Verschiebung ungefähr 23,5 km betrug.

Entsprechend kann man etwa die Besiedlung des amerikanischen Westens und die zunehmende Bevölkerungsverlagerung Richtung Westküste gut durch die Bevölkerungsschwerpunkte der USA in den letzten zweihundert Jahren demonstrieren: Lag der Bevölkerungsschwerpunkg 1754 noch in der Nähe von Philadelphia, lag er 1960 schon im Staat Illinois. Eine entsprechende Karte, beruhend auf der 1962 erschienenen Dissertation von D.S. NEFT, ist abgedruckt in

G. BAHRENBERG, E. GIESE, J. NIPPER: Statistische Methoden in der Geographie I, *Teubner, Stuttgart,* 1990, Seite 76.

Eine andere Anwendung des Bevölkerungsschwerpunkt ist die *Standortplanung:* Ein Kaufhaus oder Bahnhof oder eine Behörde mit viel Publikumsverkehr liegen dann (im Mittel) besonders günstig für die zu versorgende Bevölkerung, wenn sie nahe am Bevölkerungsschwerpunkt des intendierten Einzugsbereichs liegen.

Ebenfalls für die Standortplanung interessant sind zweidimensionale *Medianwerte:* Plant man etwa ein Zentrallager, von dem aus andere Lager regelmäßig beliefert werden sollen, so soll die Summe der Entfernungen zwischen Zentrallager einerseits und den verschiedenen sonstigen Lagern andererseits möglichst gering sein.

Für eindimensionale Verteilungen hat der Median die Eigenschaft, daß er solche Summen minimiert; diese Eigenschaft verwendet man zur Definition eines Medians im

Mehrdimensionalen und sagt, der Median eine Punkteverteilung $P_1, \ldots, P_n$ sei ein Punkt P, für den die Summe der Abstände zwischen P und den P_i minimal ist.

Mit dieser Definition ist es leider nicht ganz einfach, einen solchen Median auch wirklich zu berechnen. Zwar kennt die Mathematik natürlich Verfahren, um solche Funktionen wie die Summe aller Abstände von einem variablen Punkt zu minimieren, aber deren Behandlung würde im Rahmen einer solchen Einführungsveranstaltung zu weit führen.

Als ein praktikabler Ausweg bietet sich an, die Karte mit einem Quadratnetz zu überziehen, die Abstandssummen für einen Punkt aus einem der kleinen Quadrate abzuschätzen und dies mit den Werten von Nachbarquadraten sowie anderen Quadraten zu vergleichen, von denen man meint, daß ungefähr dort der Median liegen könnte. Zur Erhöhung der Genauigkeit kann man anschließend das Quadrat mit dem besten Wert gegebenenfalls noch weiter unterteilen und das ganze mit den kleineren Quadraten wiederholen *usw.*

§4: Grundzüge der Wahrscheinlichkeitstheorie

Die nützlichsten Informationen über eine empirische Verteilung erhält man durch Vergleich mit einer *theoretischen* Verteilung. Zum Verständnis theoretischer Verteilungen müssen wir uns zunächst mit den grundlegendsten Begriffen der Wahrscheinlichkeitstheorie auseinandersetzen.

1. Der Begriff der Wahrscheinlichkeit

Wenn wir von „Wahrscheinlichkeit" reden, ist das Wort meist im Sinne einer der drei folgenden Bedeutungen gemeint:

a) Kombinatorische Definition der Wahrscheinlichkeit

Dies ist der einfachste und anschaulichste Wahrscheinlichkeitsbegriff; er ist allerdings nur anwendbar für festumrissene endliche Grundgesamtheiten. Falls von den Individuen in der Grundgesamtheit einige eine gewisse Eigenschaft E haben, andere dagegen nicht, so definieren wir die Wahrscheinlichkeit $p(E)$ für das Auftreten von E in dieser Grundgesamtheit als

$$p(E) = \frac{\text{Anzahl der Individuen mit Eigenschaft } E}{\text{Gesamtzahl aller Individuen}} .$$

Insbesondere ist also stets

$$0 \le p(E) \le 1 ;$$

in der Praxis redet man auch häufig von einer Wahrscheinlichkeit von $100 \cdot p(E)$ Prozent.

Besteht etwa die Grundgesamtheit aus allen Studenten der Universität Mannheim und ist E die Eigenschaft *studiert Geographie,* so ist die Wahrscheinlichkeit dafür, daß ein Student der Universität Mannheim Geographie studiert

$$p(E) = \frac{\text{Anzahl der Mannheimer Geographiestudenten}}{\text{Gesamtzahl der Mannheimer Studenten}} .$$

Im Wintersemester 1993/94 waren an der Universität Mannheim 13 057 Studenten eingeschrieben, von denen 385 Geographie studierten; die Wahrscheinlichkeit, daß ein Mannheimer Student Geographie studiert, war damals also

$$p(E) = \frac{385}{13\,057} \approx 0{,}0295 = 2{,}95\% .$$

b) Häufigkeitsdefinition der Wahrscheinlichkeit

Die obige Definition ist nicht anwendbar, wenn wir etwa die Wahrscheinlichkeit bestimmen wollen, mit der eine gegebene Münzen nach dem Hochwerfen mit Kopf (oder Zahl) nach oben auf den Boden fällt, und sie ist auch nicht anwendbar um festzustellen, mit welcher Wahrscheinlichkeit ein Fuchs in Deutschland an Tollwut leidet. In beiden Fällen gibt es keine genau umrissene Grundgesamtheit, da ich die Münze immer noch einmal werfen kann, und da auch immer neue Füchse nachwachsen und alte sterben. (Im Beispiel mit den Füchsen könnte man sich natürlich dadurch aus der Affäre ziehen, daß man als Grundgesamtheit alle Füchse betrachtet, die zu einem genau bestimmten Zeitpunkt in Deutschland leben; in der Praxis wäre es allerdings auch hier unrealistisch, die genau Zahl sowohl aller Füchse wie auch der mit Tollwut infizierten Füchse festzustellen.)

In solchen Fällen muß man sich mit einer *Stichprobe* begnügen, mit dem Nachteil, daß nun die Wahrscheinlichkeit noch von der Auswahl der Stichprobe abhängt. Bei einer hinreichend großen Stichprobe ist diese Abhängigkeit aber fast zu vernachlässigen: Falls etwa eine Münze mehrmals je viermal geworfen wird, wird man sicherlich alle Kombinationen von Kopf und Zahl bekommen; wird sie dagegen mehrfach je 4 000 mal geworfen, so erhält man zwar sicherlich wieder jeweils verschiedene Ergebnisse, sie werden aber alle zumindest in der Nähe von jeweils 2 000 mal Kopf *bzw.* Zahl liegen. Wir definieren daher in solchen Fällen

$$p(E) \approx \frac{\text{Anzahl der Stichprobenelemente mit } E}{\text{Gesamtumfang der Stichprobe}}$$

für eine *hinreichend große* Stichprobe.

Um also die Wahrscheinlichkeit für das Ereignis *Kopf* beim Werfen einer Münze zu bestimmen, würde man die Münze ziemlich oft werfen und dann den Anteil der mit Kopf nach oben liegenden Münzen berechnen.

Im Fall einer endlichen Grundgesamtheit ist diese Definition kompatibel mit der kombinatorischen: Zumindest für die *größtmögliche* Stichprobe, die Grundgesamtheit selbst, stimmen beide Definitionen überein.

c) Subjektive Wahrscheinlichkeiten

Dieser dritte Wahrscheinlichkeitsbegriff soll nur ganz kurz erwähnt werden, da er für die Statistik ohne Bedeutung ist. Hier geht es um Aussagen wie „Mit einer Wahrscheinlichkeit von 60% gewinnt der 1. FC X sein nächstes Spiel“ oder „Mit einer Wahrscheinlichkeit von 55% heißt der nächste Bundeskanzler Y“. Da es jeweils um ein Einzelereignis geht, das entweder eintritt oder nicht, sind solche Aussagen ohne mathematische Bedeutung. Man könnte sie höchstens interpretieren als Aussagen über die *Erwartungen* einer Vielzahl von Individuen, wie es beispielsweise Wettbüros bei der Festlegung ihrer Quoten praktizieren, aber solche Erwartungen müssen nichts mit dem realen Ausgang einer Wahl oder eines Fußballspiels zu tun haben. In dieser Vorlesung werden subjektive Wahrscheinlichkeiten keine Rolle spielen.

3. Zufallsvariablen

Wie werden in diesem Kapitel hauptsächlich mit der zweiten, der Häufigkeitsdefinition der Wahrscheinlichkeit arbeiten. Um sie bequem anwenden zu können, definieren wir den Begriff der *Zufallsvariablen:*

Eine Zufallsvariable *ist eine Vorschrift, die jeder Durchführung eines Experiments einen zufallsabhängigen Wert zuordnet. Die Zufallsvariable heißt* diskret, *wenn ihr Wertebereich aus einer diskreten Menge besteht; sie heißt* kontinuiertlich, *wenn sie beliebige reelle Werte aus einem gewissen Bereich annehmen kann.*

Der Begriff „Experiment“ ist hier im weitesten Sinne zu verstehen: Bei einer statistischen Erhebung etwa wäre ein „Experiment“ die Befragung einer zufällig ausgewählten

Person, und jede Frage auf dem Erhebungsbogen würde eine Zufallsvariable definieren, die diesem „Experiment“ die Antwort der befragten Person zuordnet. Insbesondere muß der Wertebereich einer Zufallsvariablen also nicht aus Zahlen bestehen; es gibt auch Zufallsvariablen etwa mit den Werten *Ja, Nein, Weiß nicht* und *Keine Antwort*. Während eine solche Zufallsvariable diskret ist, könnte man etwa die Körpergröße eines Probanden als kontinuierliche Zufallsvariable auffassen, da hier natürlich beliebige Zwischenwerte möglich sind.

Wir reden bei einer Zufallsvariablen von der Wahrscheinlichkeit des Ereignisses

$$X \textit{ hat die Eigenschaft } E$$

in dem Sinne, daß für hinreichend viele Experimente der Anteil aller Experimente bestimmt wird, bei denen das durch X ausgedrückte Ergebnis die Eigenschaft E hat. Ist insbesondere X eine zahlenwertige Zufallsvariable, so reden wir im gleichen Sinne auch kurz von den Wahrscheinlichkeiten der Ereignisse

$$X = a, \quad X < a \quad \text{oder} \quad X > a \textit{ usw.}$$

Beispielsweise kann beim Roulette die jeweils gefallene Zahl zwischen 0 und 36 als Wert einer Zufallsvariablen X interpretiert werden, und bei nichtmanipulierter Schüssel sollte

$$p(X = 5) = \tfrac{1}{37} \quad \text{und} \quad p(X > 30) = \tfrac{6}{37}$$

sein.

3. Rechenregeln für Wahrscheinlichkeiten

Statistische Tests beruhen im allgemeinen darauf, daß man die Wahrscheinlichkeiten relativ komplex zusammengesetzter Ereignisse abschätzt. Zur Vorbereitung darauf wollen wir hier die einfachsten Regeln für das Berechnen der Wahrscheinlichkeiten zusammengesetzter Ereignisse betrachten.

Beginnen wir mit einem einfachen Beispiel, dem Werfen eines Würfels, wobei uns die Zufallsvariable X die geworfene Augenzahl angebe. X ist also eine diskrete Zufallsvariable, deren Wertebereich genau die sechs Zahlen $1, 2, 3, 4, 5$ und 6 enthält. Bei einem *perfekten* Würfel wird man erwarten, daß er keine dieser Zahlen bevorzugt, so daß alle sechs gleichwahrscheinlich sind; die Wahrscheinlichkeit des Ereignisses $X = i$ ist also für jede der möglichen Augenzahlen $1/6$.

Was ist dann die Wahrscheinlichkeit, daß eins *oder* drei gewürfelt wird? Da in einem von sechs Fällen eine Eins fällt und auch in einem von sechs Fällen eine Drei, fällt in zwei von sechs Fällen eine Eins oder eine Drei, d.h.

$$p(X = 1 \text{ oder } X = 3) = p(X = 1) + p(X = 3) = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3}\,.$$

Es stimmt allerdings nicht, daß nun allgemein für zwei Ereignisse E_1 und E_2 gälte

$$p(E_1 \text{ oder } E_2) = p(E_1) + p(E_2) \quad \textbf{(falsch)}.$$

Betrachten wir dazu einen zweiten Würfel, dessen Augenzahl durch eine weitere Zufallsvariable Y beschrieben werde; zur besseren Unterscheidung sei etwa der erste Würfel rot, der zweite grün. Wir betrachten das Ereignis

$$X = 1 \quad \text{oder} \quad Y = 3\,,$$

d.h. der rote Würfel soll eine Eins, der grüne eine Drei zeigen. Beim Werfen zweier Würfel gibt es $6 \cdot 6 = 36$ Möglichkeiten für das Ergebnis, die bei perfekten Würfeln alle gleichwahrscheinlich sind, d.h. sie haben jeweils die Wahrscheinlichkeit $\frac{1}{36}$. Bei *sechs* dieser Möglichkeiten zeigt der rote Würfel eine Eins; der grüne zeigt dabei irgendeine der sechs Zahlen. Entsprechend zeigt der grüne Würfel bei sechs der möglichen Ausgänge eine drei, wobei jetzt die Augenzahl des roten Würfels beliebig ist, aber trotzdem sind es nicht *zwölf,* sondern nur *elf* Fälle, in denen der rote Würfel eine Eins und der grüne eine Drei zeigt, denn der Fall, daß der rote auf eins und *gleichzeitig* der grüne auf drei fällt, wurde zweimal betrachtet. Also ist

$$p(X = 1 \text{ oder } Y = 3) = \tfrac{6}{36} + \tfrac{6}{36} - \tfrac{1}{36} = \tfrac{11}{36} .$$

Genauso überlegt man sich, daß allgemein gilt

$$p(E_1 \text{ oder } E_2) = p(E_1) + p(E_2) - p(E_1 \text{ und } E_2) .$$

Nur falls sich die beiden Ereignisse gegenseitig ausschließen, vereinfacht sich das zur obigen Formel.

Ein spezielles Ereignis, das E_1 auf jeden Fall ausschließt, ist das Ereignis *nicht* E_1; da eines der Ereignisse E_1 oder *nicht* E_1 eintreten *muß,* ist hier

$$1 = p(E_1 \text{ oder } \textit{nicht } E_1) = p(E_1) + p(\textit{nicht } E_1)$$

oder

$$p(\textit{nicht } E_1) = 1 - p(E_1) .$$

Als nächstes müssen wir uns überlegen, wie man die Wahrscheinlichkeit für das Eintreten eines Ereignisses E_1 *und* E_2

berechnen kann. Offensichtlich reicht es dazu nicht aus, nur die Wahrscheinlichkeiten $p(E_1)$ und $p(E_2)$ zu kennen, denn mit den beiden oben definierten Zufallsvariablen X und Y ist zwar

$$p(X = 1 \text{ und } Y = 3) = p(X = 1) \cdot p(Y = 3) = \tfrac{1}{36} ,$$

aber natürlich ist

$$p(X = 1 \text{ und } X = 3) = 0 \neq p(X = 1) \cdot p(X = 3) ,$$

denn man kann nicht gleichzeitig mit demselben Würfel sowohl eine Eins als auch eine Drei würfeln.

Der Unterschied zwischen den beiden Fällen ist offensichtlich, daß die beiden Ereignisse $X = 1$ und $Y = 3$ voneinander unabhängig sind: Die Augenzahl des roten Würfels beeinflußt die des grünen nicht und umgekehrt. Die Ereignisse $X = 1$ und $X = 3$ allerdings sind alles andere als unabhängig voneinander: Sie schließen sich gegenseitig aus. Daher gilt

$$p(E_1 \text{ und } E_2) = p(E_1) \cdot p(E_2) ,$$

falls die Ereignisse E_1 und E_2 voneinander unabhängig sind. Da es bei komplizierteren Anwendungen meist kaum möglich ist, *a priori* zu entscheiden, ob zwei Ereignisse voneinander unabhängig sind oder nicht, liest man den letzten Satz in der Statistik rückwärts:

Zwei Ereignisse E_1 und E_2 heißen unabhängig *voneinander, wenn*

$$p(E_1 \text{ und } E_2) = p(E_1) \cdot p(E_2)$$

ist.

In späteren Teilen der Vorlesung werden wir Techniken kennenlernen, wie man aufgrund von Beobachtungsdaten die gegenseitige Abhängigkeit *bzw.* Unabhängigkeit von Ereignissen untersuchen kann ohne die beiden Seiten dieser Gleichung explizit zu kennen.

4. Verteilungen diskreter Zufallsvariabler

Bei der Berechnung von Kennzahlen einer Zufallsvariablen müssen für diskrete und für kontinuierliche Zufallsvariablen unterschiedliche Techniken angewandt werden. Wichtiger ist dabei der kontinuierliche Fall; da der diskrete aber leichter verständlich ist, soll zunächst dieser behandelt werden. Er liefert auch eine gute Grundlage für das Verständnis des kontinuierlichen Falls. Wir nehmen in diesem Paragraphen an, daß alle Daten numerisch sind.

Gegeben sei also eine Zufallsvariable X, die nur die endlich vielen Zahlen $x_1, \dots, x_n$ annehmen kann; die Wahrscheinlichkeit des Wertes x_i sei p_i, d.h.

$$p(X = x_i) = p_i \,.$$

(Streng genommen gibt es natürlich auch diskrete Zufallsvariablen, die unendlich viele Werte annehmen können, etwa eine Zufallsvariable, deren Wertebereich alle natürlichen Zahlen sind. Solche Zufallsvariablen spielen aber in der Praxis kaum eine Rolle, wir können sie hier also getrost vernachlässigen.)

Für diese Zufallsvariable X definieren wir zunächst die *Wahrscheinlichkeitsfunktion* von X als jene Funktion, die jedem Wert x_i seine Wahrscheinlichkeit

$$f(x_i) = p_i$$

zuordnet; falls die Zahl x nicht unter den x_i vorkommt, setzen wir $f(x) = 0$.

Ist also beispielsweise X die Zufallsvariable, die die Augenzahl beim Werfen eines Würfels angibt, und nehmen wir der Einfachheit halber an, der Würfel sei so perfekt, daß jede Zahl mit gleicher Wahrscheinlichkeit fällt, so sind die möglichen Werte von x die Zahlen $x_i = i$ von eins bis sechs, und alle p_i sind gleich einem Sechstel. Also ist

$$f(1) = f(2) = f(3) = f(4) = f(5) = f(6) = \tfrac{1}{6}$$

und beispielsweise

$$f(0) = f(8) = f(4{,}5) = f(\pi) = 0\,.$$

Analog definieren wir die *Verteilungsfunktion*

$$F(x) = p(X \le x)$$

durch die Wahrscheinlichkeit dafür, daß $X \le x$ ist. Hier wäre für den Würfel also

$$F(1) = \tfrac{1}{6},\; F(2) = \tfrac{2}{6},\; F(3) = \tfrac{3}{6},\; F(4) = \tfrac{4}{6},\; F(5) = \tfrac{5}{6}$$

und $F(6) = 1$; entsprechend auch

$$F(0) = 0,\; F(8) = 1,\; F(4{,}5) = \tfrac{4}{6} \quad \text{und} \quad F(\pi) = \tfrac{3}{6}\,.$$

Da Wahrscheinlichkeiten immer zwischen null und eins liegen, ist offenbar stets

$$0 \le f(x) \le 1 \quad \text{und} \quad 0 \le F(x) \le 1\,.$$

Außerdem ist $F(x)$ eine mit zunehmendem x ansteigende Funktion, denn falls $y > x$ ist, umfaßt natürlich das Ereignis $X \leq y$ insbesondere auch das Ereignis $X \leq x$ und kann somit keine kleinere Wahrscheinlichkeit haben als dieses. Also ist $F(y) \geq F(x)$. Wie das Beispiel des Würfels zeigt, muß $F(y)$ nicht *größer* als $F(x)$ sein: Für $x = 3$ und $y = \pi$ etwa ist

$$F(x) = F(y) = \tfrac{1}{2} ,$$

denn ein Würfel kann offensichtlich nur dadurch das Ereignis „Augenzahl kleiner oder gleich π" produzieren, wenn die Augenzahl tatsächlich höchstens gleich drei ist.

Als nächstes wollen wir uns die Frage stellen, welches *arithmetische Mittel* wir erwarten, wenn wir das durch die Zufallsvariable X beschriebene Experiment hinreichend oft ausführen.

Betrachten wir dazu wieder das Beispiel des Würfels: Bei hinreichend vielen Versuchen erwarten wir, daß jede Augenzahl etwa gleich oft fallen wird, der Mittelwert der Augenzahlen sollte also ungefähr

$$\frac{1+2+3+4+5+6}{6} = \frac{21}{6} = \frac{7}{2} = 3{,}5$$

sein.

Ganz entsprechend sollte auch im allgemeinen Fall bei einer großen Zahl N von Versuchen der Wert x_i mit Wahrscheinlichkeit p_i, also in etwa $N \cdot p_i$ der N Fälle auftreten; als Mittelwert erwarten wir somit etwa den Wert

$$\frac{1}{N} \sum_{i=1}^{n} N \cdot p_i \cdot x_i = \sum_{i=1}^{n} p_i \cdot x_i ,$$

da die Gesamtzahl N der Fälle offensichtlich herausgekürzt werden kann. (Man beachte den Unterschied zwischen N und n: Es gibt N Versuche, bei denen jeweils einer der n Werte $x_1, \ldots, x_n$ als Ergebnis auftreten kann.)

Den gerade berechneten Wert bezeichnet wir allgemein als den *Erwartungswert* der Zufallsvariablen X; in Formeln schreiben wir

$$E(X) = \sum_{i=1}^{n} p_i \cdot x_i = \sum_{i=1}^{n} f(x_i) \cdot x_i .$$

Als nächstes wollen wir auch für Zufallsvariablen eine *Varianz* erklären. Für Verteilungen war die Varianz definiert als die mittlere quadratische Abweichung vom arithmetischen Mittel. Für eine Zufallsvariable haben wir anstelle des arithmetischen Mittels den Erwartungswert, wir sollten also die mittlere quadratische Abweichung von $\overline{x} = E(X)$ betrachten. Wieder ist diese Mittelbildung nicht direkt möglich, da das durch X beschriebene Zufallsexperiment beliebig oft wiederholt werden kann; wir nehmen also wieder anstelle des Mittelwerts den Erwartungswert und definieren die Varianz von X als

$$\sigma^2 = E\big((X - \overline{x})^2\big) .$$

Dabei soll der Erwartungswert eines beliebigen Ausdrucks $A(X)$ in der Zufallsvariablen X völlig analog zu oben definiert sein als

$$E\big(A(x)\big) = \sum_{i=1}^{n} p_i \cdot A(x_i) = \sum_{i=1}^{n} f(x_i) \cdot A(x_i) ,$$

also ist beispielsweise für $A(X) = (X - \overline{x})^2$

$$E\big((X-\overline{x})^2\big) = \sigma^2 = \sum_{i=1}^{n} p_i \cdot (x_i - \overline{x})^2 \,.$$

Für den Würfel ist also

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^{6} \tfrac{1}{6} \cdot (i-3{,}5)^2 = \tfrac{1}{6} \cdot \sum_{i=1}^{6} (i-3{,}5)^2 \\ &= \tfrac{1}{6} \cdot \big((1-3{,}5)^2 + (2-3{,}5)^2 + (3-3{,}5)^2 + (4-3{,}5)^2 \\ &\qquad + (5-3{,}5)^2 + (6-3{,}5)^2\big) \\ &= \tfrac{1}{6} \cdot (2{,}5^2 + 1{,}5^2 + 0{,}5^2 + 0{,}5^2 + 1{,}5^2 + 2{,}5^2) \\ &= \tfrac{1}{6} \cdot 17{,}5 = 2{,}91\overline{6}\,.\end{aligned}$$

Die (nichtnegative) Quadratwurzel aus σ^2 bezeichnen wir auch hier wieder als *Standardabweichung* der Zufallsvariablen; im Beispiel des Würfels ist also

$$\sigma = \sqrt{2{,}91\overline{6}} \approx 1{,}7078\,.$$

Im allgemeinen wird man den Erwartungswert und die Standardabweichung einer Zufallsvariablen anhand einer hinreichend großen Stichprobe schätzen; da die Wahrscheinlichkeiten von Ereignissen ohnehin durch ihre Häufigkeiten in Stichproben bestimmt sind, ist die beste Schätzung für den Erwartungswert der Zufallsvariablen natürlich das arithmetische Mittel der Stichprobe.

Im Falle der Varianz und der Standardabweichung müssen wir wieder anders vorgehen; da Varianz und Standardabweichung der Stichprobe aus den quadratischen Abweichungen vom arithmetischen Mittel der Stichprobe berechnet

werden, muß noch zusätzlich noch *dessen* Fehler berücksichtigt werden, und die gleiche Rechnung wie im letzten Paragraphen zeigt, daß bei einer Stichprobe vom Umfang N mit arithmetischem Mittel $\overline{x}$ die Varianz am besten durch

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}$$

geschätzt werden kann; die Standardabweichung entsprechend durch die Wurzel daraus.

Schließlich können wir zum Abschluß dieses Abschnitts auch noch *Schiefe* und *Wölbung* einer Zufallsvariablen X mit Erwartungswert $\overline{x} = E(X)$ und Standardabweichung σ definieren: In vollständiger Analogie zur Varianz definieren wir die *Schiefe* von X als

$$\frac{E\big((X-\overline{x})^3\big)}{\sigma^3}$$

und die *Wölbung* als

$$\frac{E\big((X-\overline{x})^4\big)}{\sigma^4}\,.$$

Im Beispiel des Würfels ist die *Schiefe* natürlich null, denn zu jeder positiven Abweichung vom Mittelwert 3,5 gibt es eine genauso große negative Abweichung.

Für die *Wölbung* berechnen wir zunächst den Erwartungswert der biquadratischen Abweichung

$$\begin{aligned}E\big((X-\overline{x})^4\big) &= \tfrac{1}{6}(2{,}5^4 + 1{,}5^4 + 0{,}5^4 + 0{,}5^4 + 1{,}5^4 + 2{,}5^4) \\ &= \tfrac{1}{6} \cdot 88{,}375 = 14{,}7291\overline{6}\,.\end{aligned}$$

Also ist die Wölbung gleich

$$\frac{14{,}7291\overline{6}}{2{,}91\overline{6}^2} \approx 1{,}7314\,.$$

5. Verteilungen kontinuierlicher Zufallsvariablen

Bei einer kontinuierlichen Zufallsvariablen X ist es im allgemeinen witzlos, von der Wahrscheinlichkeit eines Ereignisses $X = x$ zu sprechen: Obwohl etwa $\pi/2 \approx 1{,}57$ durchaus als Körpergröße in Frage kommt, ist die Aussage „Seine Körpergröße ist $\pi/2$ Meter“ sinnlos, da die Messung spätestens im Bereich der hundertstel Millimeter problematisch wird.

Selbst wenn man von diesen Schwierigkeiten bei der Messung absieht und annimmt, es gäbe einen exakten Wert, bleiben Probleme: Betrachten wir etwa eine Zufallsvariable, die für das Werfen von Dartspfeilen angibt, um wieviel der Pfeil die Mitte der Scheibe verfehlt hat. In diesem Fall wird wohl die Wahrscheinlichkeit dafür, daß diese Entfernung 3 cm ist, etwa genauso groß wie die, daß sie 3,00001 cm ist, und natürlich ist auch jeder der unendlich vielen Zwischenwerte wieder genauso wahrscheinlich; da die Summe aller dieser Wahrscheinlichkeiten aber höchstens eins sein kann, muß also jede einzelne Wahrscheinlichkeit p(X = x) null sein.

Eine positive Wahrscheinlichkeit kann es also erst dann geben, wenn wir danach fragen, wann die Entfernung in einem gewissen *Intervall* liegt, wie groß also etwa die Wahrscheinlichkeit dafür ist, daß der Pfeil eine Entfernung zwischen 2,8 und 3,3 cm vom Mittelpunkt der Scheibe hat.

Somit spielt die *Häufigkeitsfunktion* für kontinuierliche Zufallsvariablen keine Rolle, während die *Verteilungsfunktion*

$$F(x) = p(X \le x)$$

genau wie im diskreten Fall erklärt werden kann. Die Wahrscheinlichkeit, daß der Wert von X zwischen zwei Zahlen a und b liegt mit $a \le b$ ist dann

$$p(a \le x \le b) = F(b) - F(a) = p(X \le b) - p(X \le a)\,.$$

Trotz dieser nützlichen Eigenschaft, mit der die Verteilungsfunktion alle sinnvollen Wahrscheinlichkeiten für die Verteilung der Werte von X gibt, hat sie doch zumindest optisch einen Nachteil: Man sieht ihr nicht gleich an, in welcher Gegend die Werte besonders häufig angenommen werden, ob also etwa ein Dartspieler häufiger Abstände um 3 cm oder Abstände um 30 cm wirft.

Ein Maß dafür, daß X in der Gegend von x liegt, wäre etwa die Wahrscheinlichkeit dafür, daß X zwischen x und $x + h$ liegt, wobei h eine kleine Zahl ist. Da diese Wahrscheinlichkeit

$$F(x+h) - F(x)$$

im allgemeinen mit sinkendem h ungefähr proportional absinkt, ist ein besseres Maß der Quotient

$$\frac{F(x+h) - F(x)}{h}$$

für hinreichend kleines h; als kanonische Wahl bietet sich der Grenzwert für h gegen Null an, d.h. die Ableitung

$$f(x) = F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h}\,.$$

Diese Funktion, so sie existiert, heißt *Wahrscheinlichkeitsdichte* von X. (Sie muß nicht existieren, da nicht jede Funktion eine Ableitung hat; für die Zwecke der Statistik können wir aber im allgemeinen davon ausgehen, daß wir es bei kontinuierlichen Verteilungen mit differenzierbaren Verteilungsfunktionen zu tun haben.)

Nach dem Hauptsatz der Differential- und Integralrechnung ist F die *Stammfunktion* seiner Ableitung f; für $a \le b$ ist also

$$F(b) - F(a) = \int_a^b f(x)\,dx\,.$$

Wir können die Wahrscheinlichkeit dafür, daß der Wert einer Zufallsvariable in einem Intervall liegt, also auch durch die über das Intervall integrierte Wahrscheinlichkeitsdichte ausdrücken.

Diese Formel sollte man vergleichen mit dem diskreten Fall: Dort ist

$$F(x) = \sum_{x_i \le x} f(x_i)\,,$$

wobei die Summe hier bedeuten soll, daß über alle $x_i \le x$ summiert werden soll. Damit ist dann (mit der entsprechenden Interpretation der Summe)

$$F(b) - F(a) = \sum_{a < x_i \le b} f(x_i)\,,$$

die Wahrscheinlichkeitsdichte im kontinuierlichen Fall entspricht also auch formal der Wahrscheinlichkeitsfunktion des diskreten Falls, wenn man nur das Integral durch eine Summe ersetzt. In der Tat ist ein Integral seiner Idee nach nichts anderes als eine Summe (man erinnere sich an die Integrale über Treppenfunktionen, die in der Schule wahrscheinlich irgendwann einmal auftauchten) und auch das Integralzeichen $\int$ ist nichts anderes als ein stilisiertes „S“ für Summe.

Trotzdem muß man klar unterscheiden zwischen der *Wahrscheinlichkeitsfunktion* $f(x)$ im diskreten Fall, die ja die Wahrscheinlichkeit für das Ereignis $X = x$ angibt, und der *Wahrscheinlichkeitsdichte* $f(x)$ im kontinuierlichen Fall: Als Ableitung einer Funktion, die eine Wahrscheinlichkeit beschreibt, ist dieses $f(x)$ nicht selbst wieder eine Wahrscheinlichkeit, insbesondere kann $f(x)$ auch größer als eins werden: Für eine Zufallsvariable, die bevorzugt Werte in der Umgebung von drei annimmt, könnte etwa

$$f(x) = \begin{cases} 0 & \text{für } x \le 2{,}5 \\ 4(x-2{,}5) & \text{für } 2{,}5 \le x \le 3 \\ 2-4(x-3) & \text{für } 3 \le x \le 3{,}5 \\ 0 & \text{für } x \ge 3,5 \end{cases}$$

sein, d.h. der Graph von $f(x)$ wäre ein Zacken, der vom Punkt (2,5, 0) aus hochgeht zum Punkt (3, 2) und dann wieder herunter zum Punkt (3,5, 0); die Verteilungsfunktion wäre dann

$$F(x) = \begin{cases} 0 & \text{für } x \le 2{,}5 \\ 2x^2 - 10x + 12{,}5 & \text{für } 2{,}5 \le x \le 3 \\ -2x^2 + 14x - 23{,}5 & \text{für } 3 \le x \le 3{,}5 \\ 1 & \text{für } x \ge 3{,}5 \end{cases}.$$

Der Leser mag sich fragen, *wozu* die Wahrscheinlichkeitsdichte und die darauf beruhende Integraldarstellung der

Wahrscheinlichkeit gut sind, wo doch die Verteilungsfunktion F in sehr viel bequemerer Weise alle Informationen enthält, die man für die Berechnung dieser Wahrscheinlichkeit benötigt. Die traurige Wahrheit ist, daß man bei den für die Praxis wichtigsten kontinuierlichen Verteilungen zwar die Wahrscheinlichkeitsdichte $f(x)$ durch einen mehr oder weniger einfachen analytischen Ausdruck angeben kann, daß aber die Verteilungsfunktion $F(x)$ *nicht* in einer solchen geschlossenen Form angeben kann. Falls man Wahrscheinlichkeiten wirklich berechnen will, *muß* man also mit der Integraldarstellung arbeiten. Die gute Nachricht ist allerdings, daß die Funktionen $F(x)$ natürlich für alle interessanten Verteilungen (in geeigneter Normalisierung) tabelliert sind und daß auch die Statistiksysteme für Computer Berechnungsroutinen für diese Funktionen enthalten; in der Praxis kann man also weithin davon ausgehen, daß $F(x)$ bekannt ist.

Wirklich benötigt wird die Wahrscheinlichkeitsdichte für die Definition von Erwartungswert, Varianz *usw.*; hier gehen wir aus von den entsprechenden Definitionen für diskrete Zufallsvariablen und ersetzen jeweils die Summe durch ein Integral und die Wahrscheinlichkeitsfunktion durch die Wahrscheinlichkeitsdichte: Bei einer kontinuierliche Zufallsvariablen X, deren mögliche Werte zwischen dem Minimum a und dem Maximum b liegen, ist der Erwartungswert

$$E(X) \underset{\text{def}}{=} \int_a^b x \cdot f(x)\, dx\ .$$

Für die Verteilungen, mit denen wir es in Zukunft hauptsächlich zu tun haben werden, ist diese Definition leider zu speziell: Bei den meisten theoretischen Verteilungen

sind *theoretisch* beliebig große Abweichungen vom Erwartungswert möglich, auch wenn die Wahrscheinlichkeit für eine große Abweichung mit zunehmender Größe der Abweichung rapide zurückgeht. Daher müssen wir den Erwartungswert einer Zufallsvariablen X für die später relevanten Fälle definieren als

$$E(X) \underset{\text{def}}{=} \int_{-\infty}^{\infty} x \cdot f(x)\, dx\ ,$$

wobei dieses Integral zu verstehen ist als Grenzwert für immer größer werdendes positives a und immer kleiner werdendes negatives b (wobei kleiner bei negativen Werten natürlich bedeutet, daß der *Betrag* immer größer wird: $-1000 < -10$) von

$$\int_a^b x \cdot f(x)\, dx\ .$$

Ein solcher Grenzwert muß leider nicht immer existieren, und er kann auch von der Art es Grenzübergangs abhängen; wir wollen diese Schwierigkeiten hier allerdings ignorieren, da das entsprechende Problem in der Praxis eines Geographen wahrscheinlich nur sehr selten auftritt. (Bei dem Wort *wahrscheinlich* im letzten Satz handelt es sich natürlich um eine *subjektive* Wahrscheinlicheit, d.h. um etwas mathematisch sinnloses. Mit etwas Pech kann es also auch in der Praxis eines Geographen Probleme mit der Existenz dieses Grenzwerts geben; falls ein solches Problem vermutet wird, sollte man einen Statistiker zu Rate ziehen.)

Entsprechen geht man vor bei der Definition der Varianz: Diese ist der Erwartungswert der quadratischen Abweichung vom Erwartungswert, d.h.

$$\sigma^2 = \int_{-\infty}^{\infty} \big(x - E(x)\big) \cdot f(x)\,dx\ ,$$

wobei man natürlich auch hier wieder voraussetzen muß, daß dieses Integral existiert, und wobei auch hier zu erwarten ist, daß es in der Praxis des Geographen hier *wahrscheinlich* keine Probleme gibt.

Die Definition von *Schiefe* und *Wölbung* ist nun nur noch eine reine Formalität: Die Schiefe ist

$$\frac{1}{\sigma^3} \int_{-\infty}^{\infty} \big(x - E(x)\big)^3 \cdot f(x)\,dx$$

und die Wölbung ist

$$\frac{1}{\sigma^4} \int_{-\infty}^{\infty} \big(x - E(x)\big)^4 \cdot f(x)\,dx\ .$$

Auch der *Modalwert* einer kontinuierlichen Verteilung kann leicht mit Hilfe der Wahrscheinlichkeitsdichte definiert werden: Er ist derjenige Wert, für den $f(x)$ maximal wird. Falls es kein einziges wohlbestimmtes Maximum gibt, reden wir wie bei den empirischen Verteilungen von einer *bimodalen* oder gar *multimodalen* Verteilung.

Bei der Definition von *Median* und *Quantilen* haben wir bei kontinuierlichen Verteilungen sehr viel weniger Probleme als im diskreten Fall: Wir wollen der Einfachheit halber annehmen, daß die Verteilungsfunktion $F(x)$ stetig ist und dort, wo sie echt zwischen null und eins liegt, monoton ansteigt, d.h. falls $F(a)$ und $F(b)$ nicht beide gleich null oder eins sind und $a < b$ ist, dann soll auch $F(a) < F(b)$ sein, nicht nur (was selbstverständlich wäre) $F(a) \le F(b)$. Beide Annahmen müssen nicht immer erfüllt sein, sie sind es aber bei den wichtigsten Verteilungen, die uns hier begegnen werdne.

Unter diesen Annahmen gibt es dann für jeden Prozentsatz q mit $0 < q < 100$ *genau* eine Zahl x mit $F(x) = q/100$; dieses x bezeichnen wir als q-te *Quantile.* Speziell für $q = 50$ ist dies also gerade jenes x mit $F(x) = \frac{1}{2}$, d.h. die Wahrscheinlichkeit, daß ein Wert von X unter x liegt, ist $\frac{1}{2}$, und damit natürlich auch die, daß er darüber liegt. Somit bezeichnen wir dieses x als *Median* von X.

6. Die Gleichverteilung

Das einfachste Beispiel einer kontinuierlichen Verteilung ist die *Gleichverteilung:* Hier kann die Zufallsvariable X nur Werte x annehmen, die zwischen zwei vorgegebenen Werten a und b liegen, und jeder dieser Werte ist gleich wahrscheinlich, d.h., exakt ausgedrückt, die *Wahrscheinlichkeitsdichte* dieser Verteilung ist gleich einer Konstanten γ im Intervall zwischen a und b und ist null außerhalb dieses Intervalls. Da

$$p(a \le x \le b) = \int_a^b f(x)\,dx = \int_a^b \gamma\,dx = \gamma \cdot (b - a) = 1$$

sein muß, ist also

$$f(x) = \gamma = \frac{1}{b-a} \quad \text{für } a \le x \le b\,.$$

Dementsprechend ist für $a \le x \le b$

$$F(x) = \int_a^x f(x)\,dx = \gamma \cdot (x-a)\,.$$

Für $x \le a$ ist natürlich $F(x) = f(x) = 0$, und für $x \ge b$ ist $F(x) = 1$ und $f(x) = 0$.

Der *Erwartungswert* einer gleichverteilten Zufallsvariablen ist

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_a^b x \cdot f(x)\,dx \\ &= \int_a^b x \cdot \gamma\,dx = \gamma \int_a^b x\,dx \\ &= \gamma \cdot \left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right) = \frac{\frac{1}{2}b^2 - \frac{1}{2}a^2}{b-a} \\ &= \frac{a+b}{2}\,; \end{aligned}$$

wie zu erwarten war, ist es also der Mittelpunkt des Intervalls.

Die Varianz errechnet sich demnach als

$$\sigma^2 = \int_a^b \gamma \cdot \left(x - \frac{a+b}{2}\right)^2 dx\,;$$

da $\frac{1}{3}(x-c)^3$ eine Stammfunktion von $(x-c)^2$ ist, folgt

$$\begin{aligned} \sigma^2 &= \frac{\gamma}{3} \cdot \left(\left(b - \frac{a+b}{2}\right)^3 - \left(a - \frac{a+b}{2}\right)^3\right) \\ &= \frac{\gamma}{3} \cdot \left(\left(\frac{b-a}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right) \\ &= \frac{\gamma}{3 \cdot 8} \cdot 2 \cdot (b-a)^3 = \frac{2 \cdot (b-a)^3}{3 \cdot 8 \cdot (b-a)} \\ &= \frac{a^2 + ab + b^2}{12}\,. \end{aligned}$$

Die *Schiefe* ist natürlich aus Symmetriegründen gleich null, und genauso offensichtlich ist der *Median* gleich dem Erwartungswert. Von einem *Modalwert* kann man natürlich nicht in sinnvoller Weise reden, denn jeder Wert aus dem Intervall $[a,b]$ hat ja genau dieselbe Wahrscheinlichkeitsdichte. Die Berechnung der *Wölbung* sei dem interessierten Leser als Übungsaufgabe in Integralrechnung überlassen; das Ergebnis ist $-6/5$.

7. Die Binomialverteilung

Als nächstes Beispiel einer konkreten Verteilung wollen wir eine Verteilung für *diskrete* Zufallsvariablen betrachten, die uns später auf *die* wichtigste kontinuierliche Verteilung, die *Normalverteilung* führen wird.

Ausgangspunkt ist das n-malige Werfen einer Münze; diese falle jeweils mit Wahrscheinlichkeit p so, daß *Kopf* oben liegt und dementsprechend mit Wahrscheinlichkeit $q = 1 - p$ so, daß die *Zahl* zu sehen ist. Wir wollen die Wahrscheinlichkeit dafür berechnen, daß beim n-maligen Werfen k-mal *Kopf* und somit $(n-k)$-mal *Zahl* erscheint.

Anstelle des Münzwurfs kann man sich natürlich genausogut jedes andere Ereignis vorstellen, das genau zwei mögliche Ausgänge hat; lediglich der Anschaulichkeit halber soll vorläufig vom Werfen einer Münze die Rede sei – auch wenn für ein stark von $\frac{1}{2}$ abweichendes p die Anschaulichkeit vielleicht nicht allzu groß erscheint.

Wir definieren eine Zufallsvariable X durch

$$X = \text{Anzahl der Würfe mit Ergebnis } \textit{Kopf}$$

und suchen deren Wahrscheinlichkeitsfunktion

$$f(k) = p(X = k)$$

für $k = 0, \dots, n$. Die Verteilung dieser Zufallsvariablen heißt *Binomialverteilung* oder BERNOULLI-*Verteilung* mit Parametern n und p. (Der Parameter $q = 1 - p$ muß nicht eigens erwähnt werden, da er sich einfach aus p berechnen läßt.)

Im Fall $n = 1$ ist alles klar: Die Wahrscheinlichkeit für *Kopf* ist p, die für *Zahl* entsprechend $q = 1 - p$, d.h.

$$f(0) = q \quad \text{und} \quad f(1) = p\,.$$

Für $n = 2$ gibt es drei mögliche Ausgänge des Zufallsexperiments: zweimal *Kopf,* zweimal *Zahl* oder je einmal *Kopf* und *Zahl.* Zweimal *Kopf* ist nur möglich, wenn beim ersten wie beim zweiten Wurf *Kopf* oben liegt. Beide Ereignisse haben, jeweils für sich betrachtet, die Wahrscheinlichkeit p; da sie

voneinander unabhängig sind, ist die Wahrscheinlichkeit ihres gemeinsamen Auftretens gleich dem Produkt der beiden Einzelwahrsheinlichkeiten, also $p \cdot p = p^2$. Entsprechend ist die Wahrscheinlichkeit für zweimal *Zahl* natürlich q^2.

Das Ereignis „einmal *Kopf* und einmal *Zahl*“ kann auf zweierlei Weise zustande kommen: Entweder fällt beim ersten Wurf *Kopf* und beim zweiten Wurf *Zahl,* oder umgekehrt. Die Wahrscheinlichkeit für die erste Möglichkeit berechnet sich analog zu oben als $p \cdot q$, die für die zweite als $q \cdot p$, was natürlich genau der gleiche Wert ist. Da beide Möglichkeiten offensichtlich nie gleichzeitig auftreten können, ist die Wahrscheinlichkeit für das Auftreten von einer der beiden gerade die Summe der beiden Einzelwahrscheinlichkeiten, also $2pq$. Somit ist hier

$$f(0) = q^2, \; f(1) = 2pq \quad \text{und} \quad f(2) = p^2\,.$$

Die Wahrscheinlichkeiten für die drei möglichen Ausgänge des Zufallsexperiments addieren sich, wie es sich gehört, zu *eins,* denn

$$p^2 + q^2 + 2pq = (p+q)^2 = 1\,,$$

da sich p und q zu eins ergänzen.

Bei *dreimaligem* Werfen der Münze gibt es *vier* Möglichkeiten: Dreimal *Kopf*, dreimal *Zahl* sowie zweimal *Kopf* und einmal *Zahl* oder einmal *Kopf* und zweimal *Zahl.*

Anstatt diese Wahrscheinlichkeiten wieder einzeln zu berechnen, können wir vom Fall der zwei Würfe ausgehen und für jeden möglichen Ausgang den Effekt des dritten Wurfs betrachten: Für das Ereignis „zweimal *Kopf* und einmal *Zahl*“ müssen dann nur zwei Fälle betrachtet werden: Entweder war der dritte Wurf *Zahl* und damit die ersten beiden Würfe „zweimal *Kopf*“, oder der dritte Wurf lieferte *Kopf,* und die ersten beiden hatten das Ergebnis „einmal *Kopf* und einmal *Zahl*“. Die Wahrscheinlichkeit ist also

$$q \cdot p^2 + p \cdot pq = 3p^2 q \, .$$

Entsprechend ist die Wahrscheinlichkeit des Ereignisses „zweimal *Zahl* und einmal *Kopf*“ gleich $3pq^2$, und die Wahrscheinlichkeiten von „dreimal *Kopf*“ beziehungsweise „dreimal *Zahl*“ sind natürlich p^3 beziehungsweise q^3. Somit ist

$$f(0) = q^3, \; f(1) = 3pq^2, \; f(2) = 3p^2 q \quad \text{und} \quad f(3) = p^3 \, ,$$

und wieder ist die Summe aller Wahrscheinlichkeiten eins, denn

$$1 = (p+q)^3 = p^3 + 3p^2 q + 3pq^2 + q^3 \, .$$

Entsprechend können wir argumentieren bei einer beliebigen Anzahl n von Würfen: X kann die $n+1$ Werte $0 \le k \le n$ annehmen. Wie oben setzt sich das Ereignis $X = k$ zusammen aus verschiedenen Einzelereignissen wie etwa „zuerst k mal *Kopf,* dann $(n-k)$ mal Zahl“ und entsprechend den Ereignissen, bei denen k andere Wurfnummern für *Kopf* verlangt werden. Jedes dieser Ereignisse hat die Wahrscheinlichkeit $p^k q^{n-k}$, so daß $f(k)$ ein ganzzahliges Vielfaches von $p^k q^{n-k}$ ist. Wir schreiben dieses Vielfache als

$$f(k) = \binom{n}{k} p^k q^{n-k}$$

und nennen $\binom{n}{k}$ den *Binomialkoeffizienten „n über k“.*

Zu seiner Berechnung können wir vorgehen wie im Fall $n = 3$, indem wir alles auf das Werfen von $n-1$ Münzen

zurückführen und somit versuchen, $\binom{n}{k}$ durch geeignete Binomialkoeffizienten der Form $\binom{n-1}{j}$ auszudrücken:

Das Ereignis $X = k$ kann zustandekommen

- *entweder* dadurch, daß bei den ersten $n-1$ Würfen schon k mal *Kopf* gefallen ist und daß dann beim n-ten Wurf *Zahl* fällt
- *oder* dadurch, daß bei den ersten $n-1$ Würfen nur $(k-1)$ mal *Kopf* gefallen ist und daß dann beim n-ten Wurf noch einmal *Kopf* fällt.

Definieren wir also eine neue Zufallsvariable Y dadurch, daß Y die Häufigkeit des Ereignisses *Kopf* bei den ersten $n-1$ Würfen zählt, und eine Zufallsvariable Z als 0 oder 1, je nachdem, ob im n-ten Wurf *Zahl* oder *Kopf* fällt, so ist

$$\begin{aligned} p(X=k) &= p(Y=k-1) \cdot p(Z=1) + p(Y=k) \cdot p(Z=0) \\ &= \binom{n-1}{k} p^k q^{n-1-k} \cdot q \\ &\quad + \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \cdot p \\ &= \binom{n-1}{k} p^k q^{n-k} + \binom{n-1}{k-1} p^k q^{n-k} \, . \end{aligned}$$

Andererseits ist

$$p(X=k) = \binom{n}{k} p^k q^{n-k} \, ,$$

also erhalten wir die Beziehung

$$\binom{n}{k} = \binom{n-k}{k-1} + \binom{n-1}{k} \, ,$$

aus der sich $\binom{n}{k}$ rekursiv berechnen läßt.

Dies geschieht am besten über das PASCALsche Zahlendreieck: Wenn wir die Binomialkoeffizienten in der Form

$$
\begin{array}{lccccccccccc}
n=0 & & & & & & \binom{0}{0} & & & & & \\
n=1 & & & & & \binom{1}{0} & & \binom{1}{1} & & & & \\
n=2 & & & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
n=3 & & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
n=4 & & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} & \\
n=5 & \binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5}
\end{array}
$$

anordnen, besagt die gerade hergeleitete Beziehung, daß jeder Binomialkoeffizient die *Summe* der beiden rechts und links über ihm stehenden ist. Außen muß offenbar immer 1 stehen, denn die Wahrscheinlichkeit für 0 mal *Kopf,* d.h. n mal *Zahl,* ist $1 \cdot q^n$, und die für n mal *Kopf* ist $1 \cdot p^n$. Damit läßt sich das Dreieck leicht hinschreiben:

$$
\begin{array}{lccccccccccc}
n=0 & & & & & & 0 & & & & & \\
n=1 & & & & & 1 & & 1 & & & & \\
n=2 & & & & 1 & & 2 & & 1 & & & \\
n=3 & & & 1 & & 3 & & 3 & & 1 & & \\
n=4 & & 1 & & 4 & & 6 & & 4 & & 1 & \\
n=5 & 1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}
$$

Für kleine Werte von n läßt sich $\binom{n}{k}$ auf diese Weise einfach bestimmen; für großes n ist dieses Verfahren allerdings zu umständlich. Hier verwendet man besser die explizite Formel

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} ,$$

in der $n!$ (gesprochen n *Fakultät*) für das Produkt

$$n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot (n-1) \cdot n$$

der Zahlen von 1 bis n steht und verabredungsgemäß das „leere“ Produkt $0! = 1$ sein soll. (Wer sich noch an seine Schulmathematik erinnert, kann diese Formel leicht mit elementarer Bruchrechnung durch Induktion nach n beweisen.) Für das praktische Rechnen ist es oft von Vorteil, durch $(n-k)!$ zu kürzen; dadurch wird obige Formel zu

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

Für große Werte von n wird die Binomialverteilung nur selten benutzt, da man sie dann meist durch handlichere Verteilungen annähern kann. Trotzdem sei der Vollständigkeit halber die sogenannte STIRLINGsche Formel angegeben, über die man $n!$ für große Werte von n näherungsweise berechnen kann:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} ,$$

wobei

$$e \approx 2{,}71828\,18284\,59045\,23536$$

die EULERsche Zahl ist und π die Kreiszahl. Der Schätzwert ist stets zu klein mit einem Fehler von höchstens $1/12n$.

Die Berechnung von Erwartungswert und Standardabweichung unmittelbar aus der Definition ist wegen der dazu auszuwertenden Summen mit vielen Binomialkoeffizienten eher unangenehm; glücklicherweise kann man darauf aber verzichten und alles auf den Fall $n = 1$ zurückführen: Für $n = 1$ ist der Erwartungswert

$$E(X) = p \cdot 1 + 1 \cdot 0 = p$$

und die Varianz

$$\begin{aligned} E\big((X-p)^2\big) &= p \cdot (1-p)^2 + q \cdot (0-p)^2 \\ &= p \cdot q^2 + q \cdot p^2 = pq(q+p) = pq\,, \end{aligned}$$

die *Standardabweichung* ist also $\sigma = \sqrt{pq}$.

Um daraus die entsprechenden Werte für beliebiges n zu berechnen, können wir uns zunutze machen, daß die verschiedenen Würfe der Münze voneinander *unabhängig* sind, so daß sich die Erwartungswerte einfach addieren, d.h.

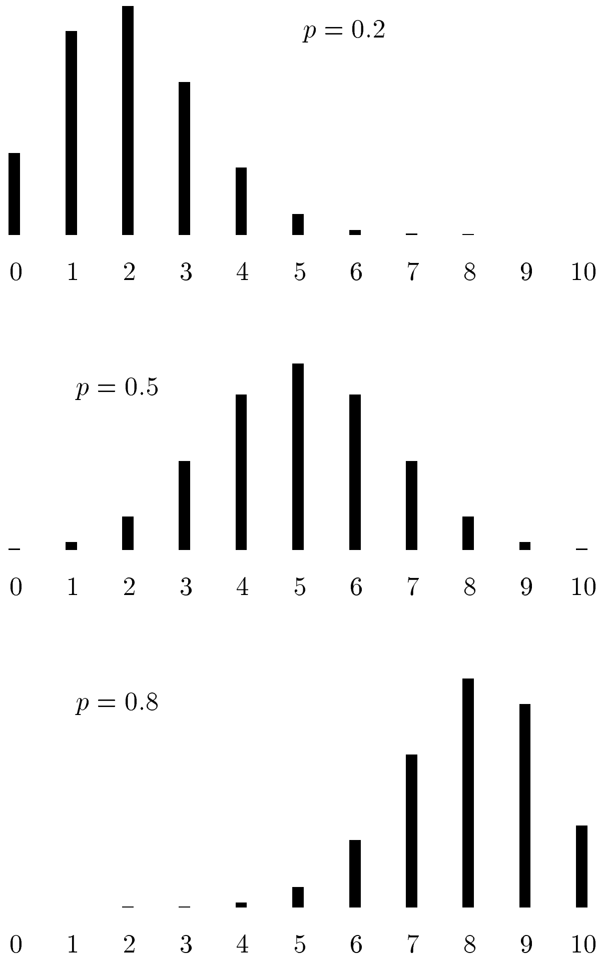
$$E(X) = n \cdot p \quad \text{bei } n \text{ Würfen.}$$

Nicht ganz so klar ist, daß sich auch die Varianzen addieren; hierzu müssen wir uns an die Rechnung erinnern, in der wir den Fehler des arithmetischen Mittels aus n Werten einer Zufallsvariablen berechnet haben, indem wir aus dem Verschwinden der Erwartungswerte der gemischten Produkte genau dies geschlossen haben. Somit ist die Varianz im Falle von n Würfen gleich npq und die Standardabweichung entsprechend $\sqrt{npq}$.

Auf die Berechnung von *Schiefe* und *Wölbung* sei hier verzichtet; lediglich das Ergebnis soll der Vollständigkeit halber angegeben werden:

$$\text{Schiefe} = \frac{(q-p)}{\sqrt{npq}} \quad \text{und} \quad \text{Wölbung} = 3 + \frac{1-6pq}{npq}\,.$$

Die Binomialverteilung ist also, wie zu erwarten war, genau dann symmetrisch, wenn $p = q = \frac{1}{2}$ ist, andernfalls ist sie schief, wird aber mit zunehmendem n immer symmetrischer.



Binomialverteilungen mit $n = 10$

Auf der vorigen Seite ist die Wahrscheinlichkeitsfunktion der Binomialverteilung für $n = 10$ und drei verschiedene p-Werte aufgezeichnet; wie man sieht, ist die Verteilung in der Tat nur für $p = 0{,}5$ symmetrisch. Die Verteilungen für die Parameter $p = 0{,}2$ und $p = 0{,}8$ sind erwartungsgemäß spiegelsymmetrisch zueinander, denn da $\binom{n}{i} = \binom{n}{n-i}$ ist, werden hier gerade die Werte von p und q miteinander vertauscht.

8. Die Poissonverteilung

Auch hier handelt es sich um eine diskrete Verteilung, nämlich den Grenzfall der Binomialverteilung für großes n und kleines p, wobei der Erwartungswert $\lambda = np$ konstant gehalten wird. Dieses Produkt λ ist demgemäß der einzige Parameter der POISSONverteilung.

In Alltagssprache übersetzt bedeuten großes n und kleines p, daß wir die Häufigkeit eines seltenen Ereignisses betrachten; insbesondere interessieren also nur die kleinen Werte von k, da die großen ohnehin extrem unwahrscheinlich sind.

Führt man den Grenzübergang explizit aus, so erhält man die Wahrscheinlichkeitsfunktion

$$f(k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} ;$$

wir bezeichnen eine Zufallsvariable demnach als POISSONverteilt, wenn sie diese Wahrscheinlichkeitsfunktion hat.

Für interessierte Leser sei wenigstens hier im Skriptum im Kleindruck kurz dargelegt, wie man die Wahrscheinlichkeitsfunktion der POISSONverteilung aus der der Binomialverteilung herleiten kann:

Für die Binomialverteilung mit Parametern n und p ist

$$\begin{aligned} f(k) = \binom{n}{k} p^k q^{n-k} &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k!} p^k q^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot (1-p)^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \cdots \cdot \frac{n-k}{n} \cdot \frac{\lambda^k}{k!} \cdot \frac{(1-\frac{\lambda}{n})^n}{\left(1-\frac{\lambda}{n}\right)^k} , \end{aligned}$$

da $q = 1 - p$ und $p = \lambda/n$ ist. Wie aus der Schule bekannt (sein sollte), ist

$$\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda} ,$$

und da k klein, n aber groß sein soll, machen wir keinen großen Fehler, wenn wir

$$\frac{n}{n} \approx \frac{n-1}{n} \approx \cdots \approx \frac{n-k}{n} \approx 1 \quad \text{und} \quad \left(1 - \frac{\lambda}{n}\right)^k \approx 1$$

setzen. Damit ist

$$f(k) \approx \frac{e^{-\lambda} \cdot \lambda^k}{k!} .$$

Die Parameter der POISSONverteilung bestimmt man am einfachsten über die der Binomialverteilung: Da die POISSONverteilung Grenzwert von Binomialverteilungen mit Erwartungswert $\lambda = np$ ist, hat sie erwartungsgemäß auch selbst den Erwartungswert λ. Die *Varianz* einer Binomialverteilung mit Parametern n und p ist

$$n \cdot p \cdot q = n \cdot p \cdot (1-p) = np - np^2 = \lambda - p \cdot \lambda .$$

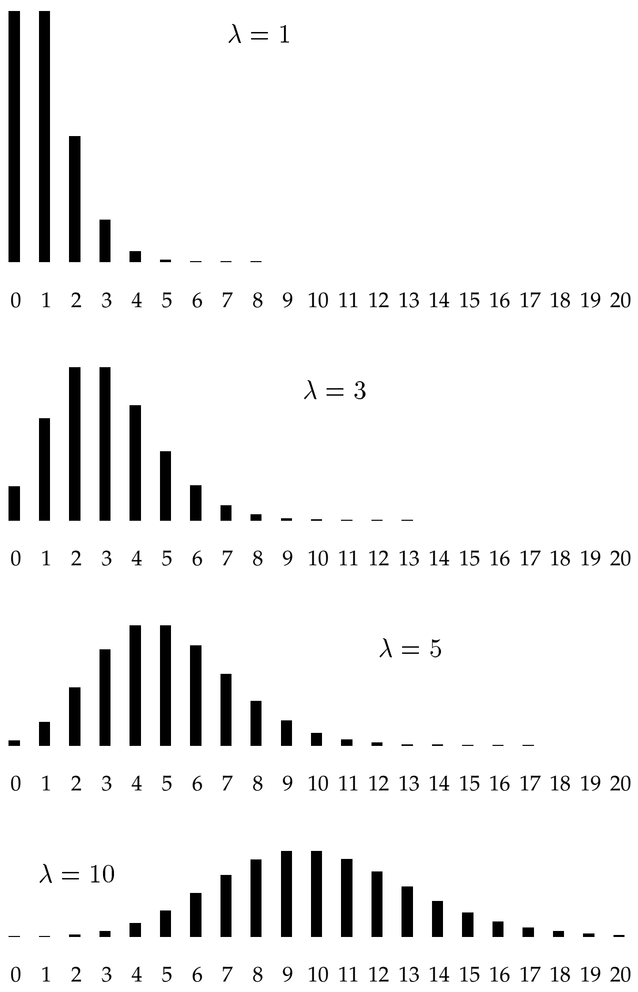
Da die POISSONverteilung der Grenzwert für $p \to 0$ ist, hat sie also die Varianz λ und damit die Standardabweichung $\sqrt{\lambda}$.

Auf die etwas trickreiche Berechnung der Schiefe möchte ich verzichten und nur das Ergebnis $1/\sqrt{\lambda}$ angeben. Die POISSONverteilung wird also umso symmetrischer, je größer der Parameter λ wird. Dies sieht man auch deutlich auf der nächsten Seite, wo die Wahrscheinlichkeitsverteilung der POISSONverteilung für $\lambda = 1, 3, 5$ und 10 aufgezeichet ist. Die Wölbung ist $3 + \frac{1}{\lambda}$, nähert sich also mit steigendem λ immer mehr dem Wert drei.

Praktisch kann man davon ausgehen, daß man für Werte von n ab etwa $n = 100$ und für Wahrscheinlichkeiten bis etwa 5%, d.h. $p \leq 0{,}05$, die Binomialverteilung durch die erheblich einfacher zu berechnende POISSONverteilung ersetzen kann, ohne einen nennenswerten Fehler zu machen. Auch für etwas kleinere Werte von n ist die Übereinstimmung im allgemeinen schon recht gut.

Beispiele für POISSONverteilte Zufallsvariablen gibt es viele: Häufigkeiten von Naturkatastrophen wie Erdbeben, Überschwemmungen *usw.*, die Anzahl radioaktiver Zerfälle pro Minute eines schwach radioaktiven Materials wie etwa C^{14} oder Co^{60} und viele mehr.

Da die POISSONverteilung nur einen Parameter hat, ist ihre Standardabweichung durch ihren Erwartungswert bestimmt; in der Tat ist sie einfach die Quadratwurzel davon. Daher steigt die Standardabweichung mit zunehmendem λ nur sehr viel langsamer als der Erwartungswert. Interpretiert man die Standardabweichung als Abweichung vom Mittelwert, so wird daher die *relative* Abweichung für große λ immer geringer und für sehr große λ praktisch vernachlässigbar. Dies ist einer der Gründe dafür, daß zahlreiche Vorgänge, die auf der mikroskopischen Ebene eigentlich



POISSONverteilungen für verschiedene Werte von λ

nur statistisch beschrieben werden können, auf der makroskopischen Ebene im Rahmen der erzielbaren Meßgenauigkeit ein deterministisches Verhalten zeigen.

Ein zwar etwas exotisches, wegen der ungewöhnlich guten Übereinstimmung von Theorie und Praxis aber häufig in Statistiklehrbüchern anzutreffendes Beispiel für POISSONverteilte Daten sind die Anzahlen der jährlich durch Hufschlag getöteten Offiziere preußischer Kavallerieregimenter: Eine Untersuchung für zehn Regimenter über zwanzig Jahre hinweg ergab insgesamt 122 Todesfälle, also im Mittel

$$\lambda = \frac{122}{10 \cdot 20} = 0{,}61$$

pro Jahr und Regiment.

Die Wahrscheinlichkeit dafür, daß es in einem Regiment genau k Todesfälle in einem Jahr gab, sollte damit unter der Annahme einer POISSONverteilung mit dem Parameter (= Erwartungswert) 0,61 ungefähr gleich

$$\frac{e^{-0{,}61} \cdot 0{,}61^k}{k!} \approx \frac{0{,}331444 \cdot 0{,}61^k}{k!}$$

sein; wir erwarten also, daß dies in etwa

$$200 \cdot \frac{0{,}331444 \cdot 0{,}61^k}{k!} = \frac{66{,}2888 \cdot 0{,}61^k}{k!}$$

der zweihundert Fälle vorkommt. Die folgende Tabelle zeigt die tatsächlichen und die (gerundeten) berechneten Werte:

k	*tatsächliche Fallzahl*	*berechnete Fallzahl*
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥ 5	0	0,08

Wie man sieht, ist die Übereinstimmung in der Tat erstaunlich gut.

9. Die Normalverteilung, das Laplacesche Fehlermodell und der zentrale Grenzwertsatz

Die *Normalverteilung* mit der charakterischen *Glockenkurve* als Wahrscheinlichkeitsdichte ist die bekannteste und wichtigste statistische Verteilung; zahlreiche weitere Verteilungen sind von ihr abgeleitet, und insbesondere bei statistischen Test spielt es oft eine große Rolle, ob man mit normalverteilten Daten arbeitet oder nicht: Für normalverteilte Daten gibt es spezialisierte Tests, die mit relativ wenigen Daten genauso sichere Schlußfolgerungen erlauben, wie man sie im allgemeinen Fall nur mit erheblich größerem Aufwand erhält. Aus diesem Grund ist es von fundamentaler Bedeutung, zu wissen, welche Daten typischerweise normalverteilt sind und welche nicht.

Beginnen wir also mit einer Situation, die für normalverteilte Daten typisch ist: Wir betrachten eine Meßgröße, die von verschiedenen Störungen beeinflußt wird. Als Beispiel kann man sich etwa ein Laborexperiment vorstellen, das „theoretisch“ auf einen wohlbestimmten Wert a führen sollte, bei dem aber der wirklich gemessene Wert natürlich noch von einer Vielzahl nicht vollständig kontrollierbarer Umgebungsbedingungen abhängt und somit bei jeder Durchführung des Experiments ein bißchen anders ist. Der Meßwert kann daher als eine Zufallsvariable X aufgefaßt werden

Der französische Mathematiker PIERRE SIMON DE LAPLACE betrachtete eine stark vereinfachte Version dieser Situation: Er ging aus von einer großen Zahl n möglicher Fehlerquellen, die alle voneinander unabhängig sein sollen, und er nahm an, daß jeder dieser Fehler das Ergebnis um den gleichen Betrag ε verändert; das Vorzeichen dieser Störung soll mit gleicher Wahrscheinlichkeit positiv oder negativ sein.

Im Falle von nur einer Fehlerquelle wäre der Fehler also mit gleicher Wahrscheinlichkeit entweder $+\varepsilon$ oder $-\varepsilon$, bei zwei Fehlerquellen wäre er in jeweils 25% aller Fälle $+2\varepsilon$ oder $-\varepsilon$, während sich in 50% der Fälle die beiden Fehler aufheben würden.

Allgemein sind wir damit in einer ähnlichen Situation wie bei einer Binomialverteilung mit $p = q = \frac{1}{2}$: Wir können so tun, als werfe jede der n Fehlerursachen eine perfekte Münze, um sich dann bei *Kopf* für den Fehler $-\varepsilon$ und bei *Zahl* für den Fehler $+\varepsilon$ zu entscheiden. Falls k der n Fehlerursachen *Kopf* werfen, ist der Gesamtfehler also

$$-\varepsilon \cdot k + \varepsilon \cdot (n-k) = (n-2k) \cdot \varepsilon \,.$$

Die Wahrscheinlichkeit für diesen Fehler ist daher gleich der Wahrscheinlichkeit für das Ereignis „k mal *Kopf*“ bei n Würfen einer fairen Münze und damit gleich

$$\binom{n}{k} \cdot 2^{-n} \,.$$

Definieren wir eine Zufallsvariable Y_n dadurch, daß ihr Wert gerade den Gesamtfehler angeben soll, so entspricht der Fehler $u = (n-2k)\cdot\varepsilon$ dem Wert $k = \frac{1}{2}(n - \frac{u}{\varepsilon})$, d.h.

$$p(Y_n = u) = \binom{n}{k} \cdot 2^{-n} \quad \text{mit } k = \tfrac{1}{2}(n - \tfrac{u}{\varepsilon}) \,.$$

Mittelwert und Standardabweichung dieser Zufallsvariablen lassen sich genauso wie im Fall der Binomialverteilung auf den Fall $n = 1$ zurückführen: In diesem Fall ist der Erwartungswert gleich null, da die Fehler $+\varepsilon$ und $-\varepsilon$ mit gleicher Wahrscheinlichkeit auftreten, und die Varianz ist $\frac{1}{2}(+\varepsilon)^2 + \frac{1}{2}(-\varepsilon)^2 = \varepsilon^2$. Im Falle von n Fehlerquellen ist also der Erwartungswert ebenfalls gleich null, die Varianz demnach $n\varepsilon^2$ und die Standardabweichung $\sqrt{n}\varepsilon$.

Die Idee des LAPLACEschen Fehlermodells ist nun ähnlich wie beim Übergang von der Binomialverteilung zur POISSONverteilung: Wir lassen n immer größer und ε immer kleiner werden; dabei soll der zu erwartende Gesamtfehler, also die Standardabweichung $\sigma = \sqrt{n}\varepsilon$, konstant bleiben. Bei gegebenem n müssen wir also

$$\varepsilon = \sigma/\sqrt{n}$$

setzen.

Außerdem müssen wir noch bedenken, daß wir hier (im Gegensatz zum letzten Abschnitt) eine *kontinuierliche* Verteilung suchen, denn ein wirklicher Meßfehler kann natürlich kontinuierliche Werte annehmen. Deshalb suchen wir nicht nach einem Grenzwert der *Wahrscheinlichkeit*

$$p(Y_n = u) = \binom{n}{k} 2^{-n} ,$$

sondern wir suchen nach einer *Wahrscheinlichkeitsdichte.* (In der Tat wäre der Grenzwert der *Wahrscheinlichkeit* für jedes u gleich null, wie man es von einer kontinuierlichen Zufallsvariablen auch nicht anders erwartet.)

Wenn wir mit einer Binomialverteilung mit Parameter n arbeiten, können nur Fehler der Form $n = (n-2k)\varepsilon$ vorkommen, zwei benachbarte Werte haben also den Abstand 2ε, und das Ereignis $Y_n = u$ repräsentiert tatsächlich das Ereignis

$$u - \varepsilon \leq Y < u + \varepsilon$$

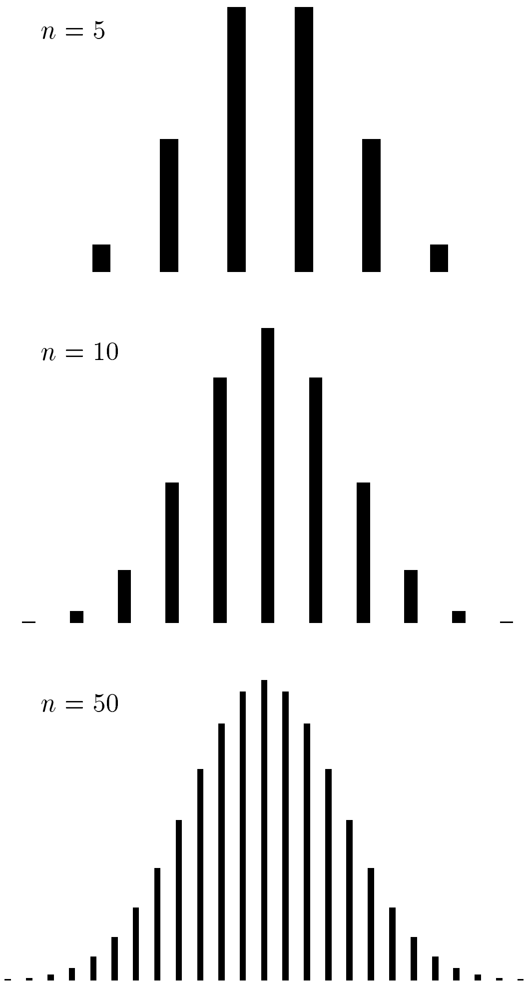
für die Zufallsvariable $Y = X - a$ des *kontinuierlich* gedachten Fehlers. Falls wir annehmen, daß die Wahrscheinlichkeitsdichte in diesem Intervall ungefähr konstant ist, können wir sie ausrechnen als

$$\varphi(u) \approx \frac{f(k)}{2\varepsilon} = \binom{n}{k} \frac{2^{-n}}{2\varepsilon} = \binom{n}{k} \frac{2^{-n-1}\sqrt{n}}{\sigma} .$$

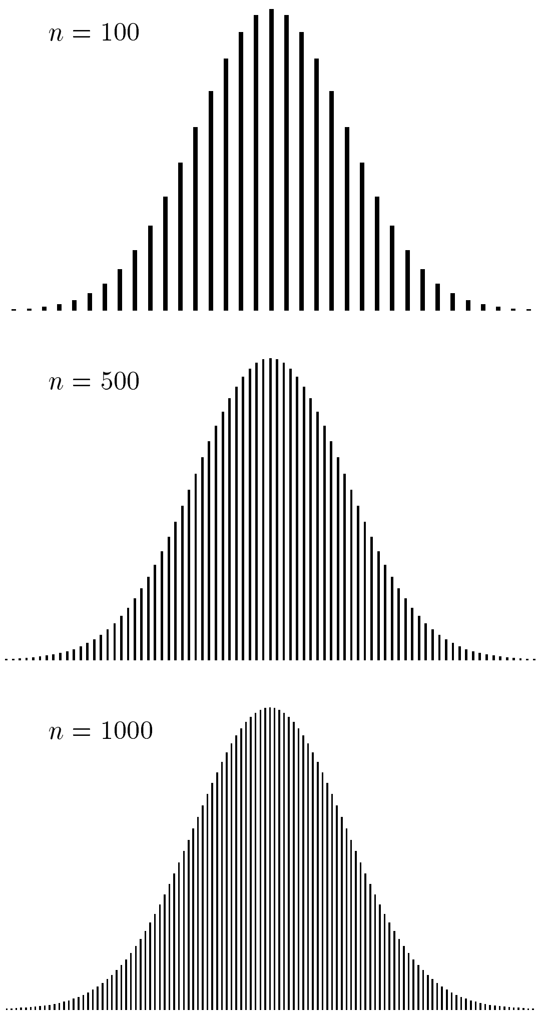
Diese Funktion ist auf den beiden folgenden Seiten für Fehler zwischen -3σ und $+3\sigma$ für $n = 5, 10, 50, 100, 500$ und 1000 aufgezeichnet; wie man sieht, nähert sie sich mit zunehmendem n in der Tat immer besser der bekannten Glockenkurve.

Eine genauere mathematische Analyse des Sachverhalts zeigt, daß diese Glockenkurve durch die Funktion

$$\varphi(u) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{u^2}{2\sigma^2}}$$



Binomialverteilungen mit $p = q = \frac{1}{2}$ für wachsendes n



Binomialverteilungen mit $p = q = \frac{1}{2}$ für wachsendes n

beschrieben wird, wobei u wie oben den *Fehler* bezeichnet. Betrachten wir stattdessen den *Meßwert* x, der idealerweise den Wert a haben sollte, so ist der Fehler gleich $x - a$, und die Wahrscheinlichkeitsdichte für den *Meßwert* ist gegeben durch

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} .$$

Auch hier soll wieder im Kleindruck kurz skizziert werden, wie man auf die angegebene Form der Wahrscheinlichkeitsdichte φ kommt: Gesucht ist der Grenzwert von

$$\binom{n}{k} 2^{-n-1} \frac{\sqrt{n}}{\sigma} = \frac{n!}{k!(n-k)!} 2^{-n-1} \frac{\sqrt{n}}{\sigma}$$

für $n \to \infty$ und konstantes

$$u = (n - 2k)\varepsilon = \frac{(n-2k)\sigma}{\sqrt{n}} .$$

Nach der STIRLINGschen Formel, die übrigens genau für diese Rechnung erstmalig aufgestellt wurde, ist

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{n^n e^{-n} \sqrt{2\pi n}}{k^k e^{-k} \sqrt{2\pi k}\,(n-k)^{n-k} e^{-(n-k)} \sqrt{2\pi (n-k)}}$$
$$= \frac{1}{\sqrt{2\pi}} \frac{n^n}{k^k\,(n-k)^{n-k}} \sqrt{\frac{n}{k \cdot (n-k)}} .$$

Dabei ist wegen obiger Beziehung

$$k = \frac{n}{2} - \frac{u\sqrt{n}}{2\sigma} .$$

Setzen wir zur Vereinfachung der Schreibweise

$$m = \frac{n}{2}, \quad v = \frac{u}{2\sigma} \quad \text{und} \quad \ell = \sqrt{2m} \cdot v ,$$

so ist also

$$k = m - \sqrt{2m} \cdot v = m - \ell \quad \text{und} \quad n - k = m + \sqrt{2m} \cdot v = m + \ell ,$$

und wir erhalten

$$
\begin{aligned}
&\binom{n}{k} 2^{-n-1} \frac{\sqrt{n}}{\sigma} \\
\approx\ & \frac{1}{\sqrt{2\pi}} \frac{n^n}{k^k\,(n-k)^{n-k}} \sqrt{\frac{n}{k\cdot(n-k)}}\, 2^{-n-1} \frac{\sqrt{n}}{\sigma} \\
=\ & \frac{1}{\sqrt{2\pi}\,\sigma} \frac{n^n \cdot 2^{-n}}{k^k\,(n-k)^{n-k}} \frac{n \cdot 2^{-1}}{\sqrt{k\cdot(n-k)}} \\
=\ & \frac{1}{\sqrt{2\pi}\,\sigma} \frac{(2m)^{2m} \cdot 2^{-2m}}{(m-\ell)^{m-\ell}(m+\ell)^{m+\ell}} \frac{2m \cdot 2^{-1}}{\sqrt{(m-\ell)(m+\ell)}} \\
=\ & \frac{1}{\sqrt{2\pi}\,\sigma} \frac{m^{2m}}{(m-\ell)^m (m+\ell)^m} \left(\frac{m-\ell}{m+\ell}\right)^{\ell} \frac{m}{\sqrt{m^2-\ell^2}} \\
=\ & \frac{1}{\sqrt{2\pi}\,\sigma} \frac{m^{2m}}{(m^2-\ell^2)^m} \left(\frac{m-\ell}{m+\ell}\right)^{\ell} \frac{m}{\sqrt{m^2-\ell^2}} \\
=\ & \frac{1}{\sqrt{2\pi}\,\sigma} \frac{1}{\left(1-\frac{\ell^2}{m^2}\right)^m} \left(\frac{1-\ell/m}{1+\ell/m}\right)^{\ell} \frac{1}{\sqrt{1-\ell^2/m^2}} \\
=\ & \frac{1}{\sqrt{2\pi}\,\sigma} \frac{1}{\left(1-\frac{2v^2}{m}\right)^m} \left(\frac{1-\sqrt{2/m}\cdot v}{1+\sqrt{2/m}\cdot v}\right)^{\sqrt{2m}\cdot v} \frac{1}{\sqrt{1-2v^2/m}}\ .
\end{aligned}
$$

Nun können wir langsam daran denken, n (und damit auch m) gegen unendlich gehen zu lassen; wir verwenden dazu wieder die aus der Schule bekannte Beziehung

$$e^x = \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n\ .$$

Danach ist insbesondere

$$\lim_{m\to\infty} \left(1-\frac{2v^2}{m}\right)^m = e^{-2v^2}$$

und

$$
\begin{aligned}
&\lim_{m\to\infty} \left(1 \pm \sqrt{\frac{2}{m}}\cdot v\right)^{\sqrt{2m}\cdot v} = \lim_{m\to\infty} \left(1 \pm \frac{\sqrt{2}\cdot v}{\sqrt{m}}\right)^{\sqrt{m}\cdot\sqrt{2}\cdot v} \\
&= \lim_{q\to\infty} \left(1 \pm \frac{\sqrt{2}\cdot v}{q}\right)^{q\cdot\sqrt{2}\cdot v} = \left(e^{\pm\sqrt{2}\cdot v}\right)^{\sqrt{2}\cdot v} = e^{\pm 2v^2}\ ,
\end{aligned}
$$

denn es bleibt sich natürlich gleich, ob m oder $q = \sqrt{m}$ gegen unendlich geht. Da der Term v^2/m gegen null geht, erhalten wir somit als Grenzwert des gesamten obigen Ausdrucks

$$\frac{1}{\sqrt{2\pi}\,\sigma} \cdot \frac{1}{e^{-2v^2}} \cdot \frac{e^{-2v^2}}{e^{+2v^2}} \cdot 1 = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-2v^2}\ .$$

Beachten wir nun noch, daß $v = u/2\sigma$ und $u = x - a$ war, erhalten wir als Endergebnis den oben angegebenen Ausdruck

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-u^2/2\sigma^2} = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-a)^2/2\sigma^2}\ .$$

Als *Normalverteilung mit Mittelwert a und Standardabweichung σ* bezeichnen wir daher die Verteilung mit Wahrscheinlichkeitsdichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}\ .$$

Diese Wahrscheinlichkeitsdichte hängt offensichtlich nur von der *normierten* Zufallsvariablen

$$Z = \frac{X-a}{\sigma}$$

ab; daher gibt es für die Normalverteilung nicht – wie etwa für Binomialverteilung und POISSONverteilung – je nach Parameterwerten verschiedene Tabellen, sondern man findet in allen Tabellenwerken nur die Normalverteilung mit Mittelwert null und Standardabweichung eins, man findet also die Wahrscheinlichkeitsdichte

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

und die Verteilungsfunktion

$$F(z) = \int_{-\infty}^{z} f(u)\,du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\,du\,.$$

Diese Verteilungsfunktion läßt sich nicht weiter vereinfachen, da sich die Stammfunktion von $e^{-u^2/2}$ nicht durch elementare Funktionen ausdrücken läßt. Für die Bestimmung von $F(x)$ ist man daher auf Tabellen oder Computerprogramme angewiesen.

Für Leser, die ihren Computer selbst programmieren und keine spezielle Statistiksoftware haben, sei hier eine Näherungsformel für $F(z)$ angegeben: Mit einem Fehler von höchstens $7{,}5 \cdot 10^{-8}$ ist

$$F(z) = 1 - \varphi(z) \cdot (a_1 t + a_2 t^2 + a_3 a^3 + a_4 t^4 + a_5 t^5)$$

mit $t = \dfrac{1}{1+pz}$ und $\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ sowie

$$\begin{array}{lll} a_1 = 0{,}319\,381\,530 & a_2 = -0{,}\,356\,563\,782 & a_3 = 1{,}781\,477\,973 \\ a_4 = -1{,}821\,255\,978 & a_5 = 1{,}330\,274\,429 & p = 0{,}231\,641\,9 \end{array}\quad.$$

Beim Rechnen mit dem Taschenrechner kann man sich auch mit einer vereinfachten Version begnügen, bei der $a_4 = a_5 = 0$ ist und

$$a_1 = 0{,}436\,1836 \quad a_2 = -0{,}120\,1676 \quad a_3 = 0{,}937\,2980 \quad p = 0{,}332\,67\,;$$

hier kann der Fehler bis zu 10^{-5} betragen.

Beispiele für normalverteilte Daten sind sehr häufig (daher auch der Name *Normal*verteilung); selbst ohne die sehr speziellen Voraussetzungen des LAPLACEschen Fehlermodells sind Summen (und damit auch Mittelwerte) von Zufallsvariablen meist normalverteilt; genauer besagt der sogenannte *zentrale Grenzwertsatz* der Statistik folgendes: Sind

$X_1, \ldots, X_n$ voneinander unabhängige Zufallsvariablen mit (fast) beliebiger Verteilung, so ist die Zufallsvariable

$$\frac{X_1 + \cdots + X_n}{n}$$

für große n annähernd normalverteilt. Als „groß“ kann man dabei in der Praxis je nach Genauigkeitsanforderungen bei Zufallsvariablen vergleichbarer Größenordnung ein n ab etwa zwanzig oder dreißig betrachten, und das eingeklammerte Wort „fast“ bedeutet, daß es einem Mathematiker zwar nicht schwer fällt, Verteilungen zu konstruieren, für die dieser Satz falsch wird, daß diese Verteilungen aber in der Praxis „praktisch“ nicht auftreten.

In die Umgangssprache übersetzt bedeutet der zentrale Grenzwertsatz, daß man immer dann von einer Normalverteilung ausgehen kann, wenn die Abweichung vom Erwartungswert als kumulativer Effekt einer Vielzahl voneinander unabhängiger Ereignisse interpretiert werden kann. Ein Beispiel ist etwa die *Körpergröße* eines Menschen, die als Produkt einer Vielzahl von genetischen Faktoren und Umwelteinflüssen interpretiert werden kann, und in der Tat zeigen die auf der nächsten Seite abgedruckten Daten von Scotland Yard, daß die Verteilung sehr gut mit der Normalverteilung mit gleichem Mittelwert und gleicher Standardabweichung übereinstimmt.

Ein geographisch interesanteres Beispiel sind die Steigungswinkel bei durch Erosion entstandenen Hängen; nach Untersuchungen von

A.N. STRAHLER: Equilibrium theory of erosional slopes approached by frequenct distribution analysis, *Am. J. Science* **248** (1950), *S.* 673–696 *und* 800–814

Verteilung der Körpergröße von 3000 Kriminellen
nach W.R. MACDONNEL, Biometrika **1** (1901), *S.* 177

Körpergröße (in inch)	*beobachtete Häufigkeit*	*erwartete Häufigkeit*
55	0	0
56	1	1
57	1	2
58	6	7
59	23	20
60	48	48
61	90	103
62	175	187
63	317	293
64	393	395
65	462	458
66	458	455
67	413	390
68	264	287
69	177	182
70	97	99
71	46	46
72	17	19
73	7	6
74	4	2
75	0	1
76	0	0
77	1	0

stimmt deren Verteilung an einem festen Hang gut mit einer Normalverteilung überein, wobei man als Grund auch hier wieder davon ausgehen kann, daß zahlreiche Einflüsse wie Variation des Bodens, der Pflanzenbewachsung, Bewässerungseffekte *usw.* eine Rolle spielen.

Typische Beispiele sind auch Meßfehler bei mehrfacher Bestimmung derselben Größe oder bereits gemittelte Größen wie etwa die mittlere Niederschlagsmenge im Monat März in Mannheim. Nicht normalverteilt wäre aber etwa die mittlere Niederschlagsmenge im Monat März bei einer Klimastation in einem Wüstenort, wo es im März typischerweise überhaupt nicht regnet: Hier wäre die Anzahl der Regentage POISSONverteilt mit kleinem λ, und dementsprechend wäre wohl die Standardabweichung der Niederschlagsmenge relativ groß im Vergleich zum Mittelwert; die Normalverteilung gäbe also eine relativ große Wahrscheinlichkeit für das offensichtlich sinnlose Ereignis „Niederschlagsmenge ist negativ“. Allgemein muß man bei Zufallsvariablen, deren Wertebereich nach einer Seite hin begrenzt ist, vorsichtig sein: Sie können höchstens dann einigermaßen normalverteilt sein, wenn der Grenzwert mindestens zwei bis drei Standardabweichungen vom Mittelwert entfernt ist.

Betrachten wir ein *Beispiel* für den Umgang mit normalverteilten Daten: Die Zufallsvariable X gebe den jährlichen Höchstwasserstand am Pegel Haltern des Flusses Stever an, von dem die bekannten Daten zeigen, daß er ungefähr normalverteilt ist mit Mittelwert $a = 4\,m$ und Standardabweichung $\sigma = 2\,m$.

Wir fragen uns, wie groß die Wahrscheinlicheit dafür ist, daß es dieses Jahr ein Hochwasser gibt, das über die $7\,m$-

Marke steigt (die etwa gleich der Deichhöhe in einer gewissen Region sein könnte). Der normalisierte Wert für $X = 7\,m$ ist

$$Z = \frac{X-a}{\sigma} = \frac{7\,m - 4\,m}{2\,m} = \frac{3\,m}{2\,m} = 1{,}5\,,$$

wir müssen also in einer Tabelle der Verteilungsfunktion der normalisierten Normalverteilung bei $z = 1{,}5$ nachschlagen. Dort finden wir

$$F(1{,}5) \approx 0{,}933\,193\,.$$

Nun ist aber

$$F(z) = p(Z \leq z)$$

während wir uns für die Wahrscheinlichkeit des komplementären Ereignisses „$Z > 1{,}5$“ *bzw.* „$X > 7\,m$“ interessieren. Die gesuchte Wahrscheinlichkeit ist also

$$1 - 0{,}933\,193 = 0{,}066\,807 \approx 6{,}7\%\,.$$

Genauso können wir uns fragen, mit welcher Wahrscheinlichkeit der Pegelhöchststand eines Jahres etwa zwischen $3\,m$ und $4\,m$ liegt. Da $4\,m$ der Mittelwert der betrachteten Verteilung ist, entspricht dies der normalisierten Variablen $Z = 0$, während $X = 3\,m$ der normalisierten Variablen

$$Z = \frac{3\,m - 4\,m}{2\,m} = -0{,}5\,m$$

entspricht. Wir interessieren uns also für

$$p(-0{,}5 \leq z \leq 0) = F(0) - F(-0{,}5)\,.$$

$F(0)$ ist leicht zu finden: Wie aufgrund der Symmetrie der normalisierten Normalverteilung nicht anders zu erwarten war, ist $F(0) = \frac{1}{2}$. Der Wert $F(-0{,}5)$ ist allerdings in kaum einer Tabelle zu finden: Diese beginnen im allgemeinen erst bei $z = 0$. Grund dafür ist wieder die Symmetrie der Wahrscheinlichkeitsdichte φ, aufgrund derer $\varphi(u) = \varphi(-u)$ und damit

$$\begin{aligned}
F(-z) &= \int_{-\infty}^{-z} \varphi(u)\,du = -\int_{\infty}^{z} \varphi(-u)\,du = \int_{z}^{\infty} \varphi(-u)\,du \\
&= \int_{z}^{\infty} \varphi(u)\,du = \int_{-\infty}^{\infty} \varphi(u)\,du - \int_{-\infty}^{z} \varphi(u)\,du \\
&= F(\infty) - F(z) = 1 - F(z)
\end{aligned}$$

ist. Also können wir bei $z = 0{,}5$ nachschlagen und erhalten

$$F(-0{,}5) = 1 - F(0{,}5) \approx 1 - 0{,}691\,462 = 0{,}308\,538\,.$$

Die gesuchte Wahrscheinlichkeit ist somit

$$F(0) - F(-0{,}5) \approx 0{,}5 - 0{,}308\,538 = 0{,}191\,462 \approx 19\%\,.$$

Als nächstes wollen wir ausrechnen, wie hoch ein Hochwasser sein muß, damit man es als *Jahrhunderthochwasser* bezeichnen kann. Da die Zufallsvariable X die jährlichen Höchststände angibt, haben wir genau dann ein Jahrhunderthochwasser, wenn $X \geq x_0$ ist, wobei

$$p(X \geq x_0) = \frac{1}{100} = 0{,}01\,.$$

Für dasselbe x_0 ist

$$p(X < x_0) = 1 - 0{,}01 = 0{,}99\,;$$

wir müssen also in der Tabelle nachschauen, für welchen Wert z_0 die Verteilungsfunktion der normierten Normalverteilung den Wert 0,99 annimmt; dieser Wert liegt etwa bei $z_0 = 2{,}326$ und entspricht dem Wert

$$x_0 = \sigma \cdot z_0 + a = 2{,}326 \cdot 2\,m + 4\,m = 8{,}652\,m\,.$$

Ab diesem Pegelstand können wir also von einem Jahrhunderthochwasser sprechen.

Noch ein weiterer Typ von Fragestellungen ist typisch bei normalverteilten Daten: Falls wir eine Zufallsvariable X betrachten, die einen fehlerbehafteten Meßwert beschreibt, interessiert natürlich die Wahrscheinlichkeit dafür, daß der Fehler eine gewisse Größe ε nicht überschreitet; falls a der Erwartungswert der Zufallsvariablen ist, der theoretisch „richtige" Meßwert also, interessiert demnach die Wahrscheinlichkeit

$$p(a - \varepsilon \le X \le a + \varepsilon)$$

oder, durch die normalisierte Zufallsvariable $(X - a)/\sigma$ ausgedrückt, die Wahrscheinlichkeit

$$p(-\varepsilon/\sigma \le Z \le \varepsilon/\sigma)\,.$$

Setzen wir kurz $u = \varepsilon/\sigma$, so ist

$$p(-u \le Z \le u) = F(u) - F(-u) = F(u) - \bigl(1 - F(u)\bigr) = 2F(u) - 1\,.$$

Besonders interessant ist dabei der Fall $u = 1$, der die Wahrscheinlichkeit dafür angibt, daß der Fehler höchstens gleich einer Standardabweichung ist; wegen

$$F(1) = 0{,}84134\,47461\,68543$$

ist diese Wahrscheinlichkeit gleich

$$2 \cdot F(1) - 1 = 0{,}68268\,94921\,37086$$

oder stark zwei Drittel. In etwa zwei von drei Fällen wird der Fehler also höchstens eine Standardabweichung betragen.

Da Fehlerbalken in einer Meßkurve im allgemeinen die Standardabweichung angeben, bedeutet dieses Ergebnis: Etwa ein Drittel aller Meßwerte liegt außerhalb der um die *theoretische* Kurve eingezeichneten Fehlerbalken. Falls dies bei einer längeren Meßreihe nicht gilt, wurden entweder die Fehler falsch berechnet oder die Daten manipuliert.

Auch die Wahrscheinlichkeiten dafür, daß ein Fehler höchstens zwei, drei, vier oder fünf Standardabweichungen beträgt, sind einer Betrachtung wert; die entsprechenden Werte

$$p(a - k\sigma \le X \le a + k\sigma) = p(-k \le Z \le k) = F(k) - 2$$

sind in der untenstehenden Tabelle zu finden:

k	$p =$	$1 - p =$
1	0,68268 94921	0,31731 05079
2	0,95449 97361	0,04550 02639
3	0,99730 02039	0,00269 97961
4	0,99993 66576	0,00006 33424
5	0,99999 94266	0,00000 05734

Damit ist also die Wahrscheinlichkeit eines Fehlers von mindestens 2σ geringer als 5%, ein Fehler von 3σ oder mehr kommt gerade noch in einem Viertel Prozent aller Fälle vor, und die Wahrscheinlichkeit für einen Fehler von mindestes 4σ oder gar 5σ ist geringer als $1 : 15\,000$ *bzw.* $1 : 1\,700\,000$. Für die Praxis geht man daher im allgemeinen davon aus, daß eine Abweichung von mehr als 3σ kein zufälliger Fehler mehr ist, sondern ein zu ignorierender *Ausreißer,* der durch einen groben Ablesefehler oder einen falsch notierten Zahlenwert oder etwas ähnliches entstanden ist.

Zum Schluß dieses Kapitels seien noch die Parameter der Normalverteilung genannt; auf ihre tatsächliche Berechnung muß hier verzichtet werden: Schon der exakte Nachweis, daß wirklich $F(\infty) = 1$ ist, erfordert nichttriviale Tricks bei der Bestimmung des Integrals.

Klar aufgrund der Symmetrie der Wahrscheinlichkeitsdichte ist allerdings, daß *Mittelwert, Median* und *Modus* wirklich gleich jenem Parameter a sind, den wir als Mittelwert bezeichnet haben, und daß die *Schiefe* gleich null ist. Als Standardabweichung stellt sich ebenfalls der zu erwartende Wert σ heraus, und als *Wölbung* erhält man den Wert drei.

Da an der Normalverteilung alles „normal“ sein soll, stört dieser Wert *drei* manche Leute, die deshalb anstelle der Wölbung eine neue Kenngröße

$$\text{Exzeß} = \text{Wölbung} - 3$$

einführen, die somit für die Normalverteilung gleich null ist. Leider wird auch diese neue Kenngröße gelegentlich als Wölbung bezeichnet, genau wie die hier definierte Wölbung gelegentlich auch als Exzeß bezeichnet wird; falls man Werte für die Wölbung aus der Literatur übernimmt, ist also Vorsicht geboten und man sollte sich klarmachen, was der Autor meint.

Schiefe und Wölbung der Normalverteilung geben einen ersten Hinweis darauf, welche empirische Daten normalverteilt sind und welche nicht: Wenn die Schiefe zu stark von null abweicht und die Wölbung zu stark von drei, können die Daten offensichtlich *nicht* normalverteilt sein.

Beispielsweise ergibt sich bei den Körpergrößen der 3 000 Kriminellen die Schiefe zu 0,064 und die Wölbung zu 3,170, was noch in guter Übereinstimmung mit den theoretischen Werten ist. Bei den Münchner Oktobertemperaturen dagegen hatten wir die Schiefe als 0,304 berechnet und die Wölbung als 3,437, und in der Tat zeigen die Histogramme in Kapitel 2 ein Bild, das schon recht deutlich von einer Normalverteilung abweicht.

§5: Schätz- und Teststatistik

Wir haben uns schon in Kapitel 3 damit beschäftigt, wie man aufgrund einer *Stichprobe* Aussagen über eine größere *Grundgesamtheit* machen kann. In diesem Kapitel soll dieses Thema weiter ausgebaut werden, wobei insbesondere die *Zuverlässigkeit* dieser Aussagen eine Rolle wird.

1. Punktschätzungen

Unter diesem Oberbegriff faßt man die uns bereits aus Kapitel 3, §4 bekannten Schätzungen zusammen; zwar ging es dort nicht um Verteilungen von Zufallsvariablen, sondern um die Schätzung von Parametern einer Grundgesamtheit aus den Daten einer Stichprobe, aber aufgrund der Häufigkeitsdefinition der Wahrscheinlichkeit können wir die Ergebnisse auf Zufallsvariablen ausdehnen, indem wir uns vorstellen, daß die gegebene Stichprobe Teil einer sehr viel größerem Stichprobe sein, deren Parameter wir wie in §4 schätzen können und die nach der Häufigkeitsdefinition der Wahrscheinlichkeit die Verteilung der Zufallsvariablen wiederspiegeln sollte.

Ist X eine Zufallsvariable und

$$x_1, \ldots, x_N$$

eine *Stichprobe* von Werten dieser Zufallsvariablen, so kann man den *Erwartungswert* von X schätzen durch das arithmetisches Mittel

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

der Stichprobe und die *Standardabweichung* σ durch

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}\,.$$

Letzteren Wert, den man – vor allem für kleine N – von der *Standardabweichung*

$$\overline{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}}$$

der Stichprobe unterscheiden muß, wollen wir im folgenden kurz als die *Streuung* der Stichprobe bezeichnen: Die Standardabweichung der Zufallsvariablen X wird also am besten geschätzt durch die Streuung der Stichprobe.

2. Intervallschätzungen

Die Schätzungen aus dem letzten Paragraphen haben den Nachteil, daß wir noch keine Aussage über ihre Qualität machen können. Um zu solchen Aussagen zu kommen, müssen wir Mittelwert und Streuung der Stichprobe selbst als *Zufallsvariablen* betrachten und deren Verteilungen untersuchen. Dazu sei der Stichprobenumfang N für den Augenblick festgehalten, und $\overline{X}$ sei jene Zufallsvariable, die jeder zufällig gewählten Stichprobe vom Umfang N zur Zufallsvariablen X deren arithmetisches Mittel zuordnet.

Falls die ursprüngliche Zufallsvariable X normalverteilt war, gilt dasselbe auch für $\overline{X}$; falls N genügend groß ist (in der Praxis ab etwa $N = 30$ oder $N = 40$), gilt dies nach dem zentralen Grenzwertsatz sogar ohne die Annahme, daß X normalverteilt sei. Der Erwartungswert von $\overline{X}$ ist natürlich gleich dem von X, und in Kapitel 3, §4*a)* haben wir auch bereits die Standardabweichung von $\overline{X}$ berechnet als $\sigma/\sqrt{N}$, wobei σ die Standardabweichung von X ist.

Mit den Kenntnissen aus dem letzten Kapitel können wir uns nun genauer überlegen, was diese Zahl eigentlich aussagt: Mit einer Wahrscheinlichkeit von etwa $2:3$ gilt für den Mittelwert $\overline{x}$ einer Stichprobe vom Umfang N, daß

$$E(X) - \frac{\sigma}{\sqrt{N}} \leq \overline{x} \leq E(X) + \frac{\sigma}{\sqrt{N}}\,.$$

Entsprechende Schranken können wir auch für jede andere Wahrscheinlichkeit angeben:

Dazu betrachten wir die normalisierte Zufallsvariable zu $\overline{X}$, also die Variable

$$Z = \frac{\overline{X} - E(X)}{\sigma/\sqrt{N}}\,,$$

die einer normalisierten Normalverteilung genügt. Wie wir am Ende des vorigen Kapitels gesehen haben, ist die Wahrscheinlichkeit p dafür, daß ein Wert einer solchen Zufallsvariablen der Ungleichung

$$-z \leq Z \leq +z$$

genügt, gleich $2F(z) - 1$, wobei F die Verteilungsfunktion der normalisierten Normalverteilung ist.

Rückwärts gelesen heißt das: Ist für eine *vorgegebene* Wahrscheinlichkeit p zwischen null und eins z_p jener Wert, für den

$$2F(z_p) - 1 = p$$

ist, so gilt mit Wahrscheinlichkeit p, daß

$$E(X) - \frac{z_p \cdot \sigma}{\sqrt{N}} \leq \overline{x} \leq E(X) + \frac{z_p \cdot \sigma}{\sqrt{N}}\,.$$

Um von hier auf eine Abschätzung für die eigentlich interessierende Größe $E(X)$ zu kommen, müssen wir die Ungleichungen noch etwas umformen:

$$E(X) - \frac{z_p \cdot \sigma}{\sqrt{N}} \leq \overline{x} \iff E(X) \leq \overline{x} + \frac{z_p \cdot \sigma}{\sqrt{N}}$$

und

$$\overline{x} \leq E(X) + \frac{z_p \cdot \sigma}{\sqrt{N}} \iff \overline{x} - \frac{z_p \cdot \sigma}{\sqrt{N}} \leq E(X)\,,$$

die obige Ungleichungskette ist also äquivalent zu

$$\overline{x} - \frac{z_p \cdot \sigma}{\sqrt{N}} \leq E(X) \leq \overline{x} + \frac{z_p \cdot \sigma}{\sqrt{N}}\,,$$

so daß auch diese mit Wahrscheinlichkeit p gilt.

Damit haben wir eine erste Antwort auf das Zuverlässigkeitsproblem:

Ist $\overline{x}$ das arithmetische Mittel einer Stichprobe vom Umfang N und ist $2F(z_p) - 1 = p$, so liegt der Erwartungswert $E(X)$ der

Zufallsvariablen mit Wahrscheinlichkeit p zwischen den Schranken $\overline{x} - z_p \cdot \sigma/\sqrt{N}$ *und* $\overline{x} + z_p \cdot \sigma/\sqrt{N}$.

Ein solches Intervall, in dem der zu schätzende Wert mit einer vorgegebenen Wahrscheinlichkeit p liegt, bezeichnen wir als *Konfidenzintervall* zur Wahrscheinlichkeit p.

Das Problem mit den so berechneten Konfidenzintervallen ist, daß es wohl kaum Fälle gibt, in denen wir zwar den *Erwartungswert* einer Zufallsvariablen nicht kennen und somit schätzen müssen, in denen wir aber die *Standardabweichung* σ derselben Zufallsvariablen kennen, so daß wir die beiden Schranken des Konfidenzintervalls wirklich ausrechnen können.

Als Ausweg bietet sich an, die unbekannte Standardabweichung σ der Zufallsvariablen X durch die bekannte Streuung s der Stichprobe zu ersetzen. Damit betrachten wir anstelle der normalisierten Zufallsvariablen

$$Z = \frac{\overline{X} - E(X)}{\sigma/\sqrt{N}}$$

die neue Zufallsvariable

$$T = \frac{\overline{X} - E(X)}{s/\sqrt{N}} .$$

Dem Vorteil der Berechenbarkeit dieser neuen Variablen T steht der Nachteil entgegen, daß wir nichts wissen über die Verteilung von T: Da sich s und σ zumindest für kleine Werte von N teilweise beträchtlich unterscheiden können und s im Gegensatz zu σ auch noch von der Stichprobe abhängt, können wir sicherlich nicht davon ausgehen,

daß auch T wie Z einer normalisierten Normalverteilung genügt: Zwar ist zu erwarten, daß Z und T für große Werte von N recht gut übereinstimmen, für kleine N ist aber mit teilweise beträchtlichen Abweichungen zu rechnen.

In der Tat genügt T einer eigenen Verteilung, die für jedes N leicht verschieden ist und STUDENTsche t-Verteilung heißt. Ihre Wahrscheinlichkeitsdichte ist glockenförmig wie die der Normalverteilung, und sie nähert sich dieser mit zunehmendem N immer besser an.

„STUDENT“ war das Pseudonym, unter dem der irische Mathematiker WILLIAM S. GOSSET (1876–1936) seine Arbeiten veröffentlichte. GOSSET arbeitete für die Brauerei Guiness, die es ihren Angestellten nicht erlaubte, unter eigenem Namen zu publizieren. Bei seine Arbeiten für die Brauerei, in denen es unter anderem um die Unterscheidung verschiedener Behandlungsarten von Braugerste mit statistischen Mitteln ging, mußte er häufig mit Stichproben auskommen, die zu klein waren, als daß man einfach $\sigma \approx s$ setzen konnte; um auch bei solchen kleinen Stichproben zu statistisch signifikanten Aussagen zu kommen, führte er 1908 die nach ihm benannte t-Verteilung ein. Zu seinen vielen sonstigen Beiträgen zur Statistik gehört auch die Berechnung der POISSONverteilung als Grenzwert der Binomialverteilung, auf die er unabhängig von POISSON kam im Zusammenhang mit dem Zählen von Hefepilzen. „STUDENT“ gilt aufgrund dieser und anderer Arbeiten auch als einer der Gründerväter der *Biometrik*.

Auch für die t-Verteilung kann man nach dem Intervall fragen, in dem T mit vorgegebener Wahrscheinlichkeit liegt; da bei dieser Verteilung im allgemeinen *nur* diese Frage interessiert, findet man in Tabellenwerken nur ganz selten die *Verteilungsfunktion* von T, dafür aber immer diese Grenzwerte. Anstelle der Wahrscheinlichkeit p betrachtet man dort allerdings die komplementäre Wahrscheinlichkeit $\alpha = 1 - p$ dafür, daß der Erwartungswert *nicht* im entsprechenden Konfidenzintervall liegt, und anstelle des Stichprobenumfangs N benutzt man den Parameter $f = N - 1$, der als *Anzahl der Freiheitsgrade* bezeichnet wird. Schließlich unterscheiden die Tabellen noch zwischen *einseitigen* und *zweiseitigen* Test; aus Gründen, die im nächsten Abschnitt klar werden sollten, benötigt man hier die Werte für den zweiseitigen Test. Für diesen findet man für gegebenes $\alpha = 1-p$ und $f = N-1$ einen Grenzwert t_α, für den gilt:

Ist $\overline{x}$ das arithmetische Mittel einer Stichprobe vom Umfang N, so liegt der Erwartungswert $E(X)$ der Zufallsvariablen mit Wahrscheinlichkeit p zwischen $\overline{x} - t_\alpha \cdot s/\sqrt{N}$ und $\overline{x} + t_\alpha \cdot s/\sqrt{N}$.

Meist wird in den Tabellen auch noch das t_α für unendlich viele Freiheitsgrade angegeben; da sich die t-Verteilung für immer größer werdende Anzahl von Freiheitsgraden immer mehr der normalisierten Normalverteilung nähert, ist dies der entsprechende Grenzwert für die normalisierte Normalverteilung.

Betrachten wir ein konkretes Beispiel:

In der Gemeinde Telgte wurden 56 zufällig ausgewählte Haushalte nach ihren monatlichen Ausgaben für Lebensmittel gefragt. Der Mittelwert der Antworten ergibt sich zu 850 DM bei einer Streuung von 400 DM. Was wirklich interessiert, ist natürlich nicht der Mittelwert über diese Stichprobe, sondern der über die gesamte Grundgesamtheit, d.h. für alle Haushalte von Telgte. Allein anhand der Stichprobe läßt dieser sich nicht exakt bestimmen, gesucht ist also ein Intervall, in dem dieser Mittelwert μ mit beispielsweise 90%-iger Sicherheit liegt.

Für $p = 90\%$ müssen wir in der Tabelle des zweiseitigen t-Tests unter der komplementären Wahrscheinlichkeit

$\alpha = 1 - p = 10\%$ nachschauen bei $f = N - 1 = 55$ Freiheitsgraden; wir finden den Wert

$$t_\alpha = 1{,}673\,.$$

Da bei den vorliegenden Daten

$$t_\alpha \cdot \frac{s}{\sqrt{N}} = 1{,}673 \cdot \frac{400\,\mathrm{DM}}{\sqrt{56}} \approx 89{,}43\,\mathrm{DM}$$

ist, können wir also mit 90%-iger Sicherheit davon ausgehen, daß

$$760{,}57\,\mathrm{DM} \le \mu \le 939{,}43\,\mathrm{DM}\,.$$

Würden wir auf einer 95%-igen Sicherheit bestehen, müßten wir entsprechend bei $\alpha = 5\%$ nachsehen und erhielten $t_\alpha = 2{,}004$; mit 95%-iger Sicherheit ist demnach

$$740{,}96\,\mathrm{DM} \le \mu \le 959{,}04\,\mathrm{DM}\,.$$

3. Einführendes Beispiel zur Teststatistik

Angenommen, fünf zufällig ausgewählte Studenten einer Vorlesung *Quantitative Methoden in der Geographie* werden einem standardisierten Intelligenztest unterworfen; als Ergebnis ergebe sich ein mittlerer Intelligenzquotient von 111,11. Da der Intelligenzquotient, bezogen auf die Gesamtbevölkerung, ungefähr normalverteilt ist mit einem Erwartungswert von 100 und einer Standardabweichung von 16, liegt der gemessene Mittelwert also über dem Durchschnitt. Ist dieser Effekt rein zufällig, so daß Studenten der Vorlesung *Quantitative Methoden in der Geographie* im Mittel genauso intelligent sind wie die Gesamtbevölkerung, oder gibt es signifikante Unterschiede?

Falls der Effekt rein zufällig ist, bilden die fünf Studenten bezüglich ihrer Intelligenz (insoweit diese etwas mit dem Ergebnis eines Intelligenztests zu tun haben sollte) eine Zufallsstichprobe aus der Gesamtbevölkerung; die Zufallsvariable für den Mittelwert ihrer Testergebnisse ist also normalverteilt mit Erwartungswert 100 und Standardabweichung

$$\overline{\sigma} = \frac{16}{\sqrt{5}} \approx 7{,}1554728\,.$$

Der gemessene Mittelwert 111,11 unterscheidet sich um

$$11{,}11 \approx 1{,}55\overline{\sigma}$$

vom Erwartungswert; die Wahrscheinlichkeit für eine Abweichung höchstens dieser Größe ist

$$2F(1{,}55) - 1 = 2 \cdot 0{,}939429 - 1 = 0{,}878858\,.$$

Mit anderen Worten: Mit einer Wahrscheinlichkeit von etwa 88% unterscheidet sich der Mittelwert einer Zufallsstichprobe vom Umfang fünf aus der Gesamtbevölkerung um *weniger* als 11,11 vom Erwartungswert 100; die Wahrscheinlichkeit dafür, daß ein Unterschied von *mehr* als 11,11 auftritt, liegt also nur bei etwa 12%. Falls man bereit ist, eine Fehlerwahrscheinlichkeit von 12% zu tolerieren, könnte man also sagen, daß der Effekt *nicht* zufällig ist.

Gegen obige Überlegungen und Berechnungen kann man einwenden, daß man Abweichungen vom Mittelwert nach *unten* hier wohl praktisch sicher ausschließen kann: Um Student in einer Vorlesung *Quantitative Methoden in der Geographie* werden zu können, muß ein Mitglied der Gesamtbevölkerung schließlich zum Hochschulstudium zugelassen werden und im allgemeinen wohl auch die ersten

beiden Semester des Geographiestudiums überstehen; da auch heute noch deutlich weniger als 50% eines Jahrgangs an eine Hochschule kommen und der Hochschulzugang idealerweise etwas mit Intelligenz zu tun haben sollte, wird man wohl davon ausgehen können, daß Studenten einer Vorlesung *Quantitative Methoden in der Geographie* im Mittel kaum weniger intelligent sein können als der Durchschnitt der Gesamtbevölkerung. Die obige Wahrscheinlichkeit ist aber die Wahrscheinlichkeit dafür, daß entweder $\overline{x} \geq 111{,}11$ ist oder aber $\overline{x} \leq 100 - 11{,}11 = 88{,}89$; man spricht daher bei Betrachtung dieser Wahrscheinlichkeit von einem *zweiseitigen* Test. Wenn man die Möglichkeit $\overline{x} \leq 88{,}89$ ausschließt ist die Irrtumswahrscheinlichkeit gleich der Wahrscheinlichkeit des Ereignisses $\overline{x} \geq 111{,}11$, also

$$1 - F(1{,}55) \approx 1 - 0{,}939429 = 0{,}060571 \approx 6\%$$

und damit nur noch halb so groß; in diesem Fall spricht man von einem *einseitigen* Test.

Auf den ersten Blick ist dieses Ergebnis paradox: Wir können mit einer Irrtumswahrscheinlichkeit von nur 6% sagen, daß der mittlere Intelligenzquotient der Studenten *über* dem Gesamtbevölkerung liegt, aber wenn wir nur sagen, die beiden Mittelwerte seien *verschieden,* verdoppelt sich die Irrtumswahrscheinlichkeit auf 12% ?

Selbstverständlich sind die beiden Mittelwerte voneinander verschieden, wenn der eine größer ist als der andere, aber wir haben beim einseitigen und beim zweiseitigen Test nicht die gleiche Frage gestellt: Betrachten wir die Daten nur unter dem Gesichtspunkt der Statistik, so ist es durchaus möglich, daß der mittlere Intelligenzquotient der Studenten *unter* dem der Gesamtbevölkerung liegt, obwohl der Mittelwert der Stichprobe mit 111,11 deutlich darüber liegt: Da die Differenz 11,11 wie wir oben gesehen haben 1,55 Standardabweichungen des Mittelwerts entspricht, ist die Wahrscheinlichkeit dafür

$$F(-1{,}55) = 1 - F(1{,}55) \approx 1 - 0{,}939429 = 0{,}060571 \approx 6\% \,.$$

Allein aufgrund der statistischen Daten müssen wir also mit einer Wahrscheinlichkeit von 6% damit rechnen, daß der mittlere Intelligenzquotient der Studenten *unter* dem der Gesamtbevölkerung liegt. Beim einseitigen Test haben wir dies aufgrund *inhaltlicher* Überlegungen, die nichts mit der statistischen Verteilung der Daten zu tun haben, ausgeschlossen; die Irrtumswahrscheinlichkeit beim einseitigen Test ist also dadurch kleiner geworden, daß wir ein Ereignis, das aufgrund der Datenlage mit 6%-iger Wahrscheinlichkeit eintreten kann, für *unmöglich* erklären und ignorieren.

Allgemein geht man in der statistischen Testpraxis meistens aus von einer vorgegebenen Irrtumswahrscheinlichkeit α, die man gerade noch zu tolerieren bereit ist; typische Werte dafür sind 20%, 10%, 5%, 1%, 0,1%, 0,01% *usw.* Welchen Wert man in einzelnen wählt, hängt natürlich von der jeweiligen Anwendung ab: Geht es nur um eine Vorabschätzung, ob eine gewisse Hypothese einer näheren Untersuchung wert ist, wird man im allgemeinen auch noch mit einer Irrtumswahrscheinlichkeit von 10% oder gar 20% zufrieden sein; falls es aber darum geht, ob gewisse Maßnahmen ausreichen, um eine Katastrophe zu verhindern, wird man sicherlich erheblich höhere Ansprüche stellen und nur eine deutlich kleinere Irrtumswahrscheinlichkeit tolerieren.

Als nächstes wird eine *Nullhypothese* aufgestellt; im obigen Fall wäre dies für den zweiseitigen Test die Hypothese

H_0: *Der Unterschied zwischen dem mittleren Intelligenzquotienten der fünf Studenten der Vorlesung* Quantitive Methoden in der Geographie *und dem der Gesamtbevölkerung ist nur zufallsbedingt.*

Diese Nullhypothese ist genau dann richtig, wenn die folgende *Alternativhypothese* falsch ist:

$H_{!A}$: *Der in der Stichprobe ermittelte Wert ist signifikant verschieden von dem der Gesamtbevölkerung.*

Entsprechend haben wir beim zweiseitigen Test die Nullhypothese

H_0: *Es ist nur zufallsbedingt, daß der mittleren Intelligenzquotienten der fünf Studenten der Vorlesung* Quantitive Methoden in der Geographie *größer ist als der der Gesamtbevölkerung.*

und die Alternativhypothese

$H_{!A}$: *Der in der Stichprobe ermittelte Wert ist signifikant größer als der der Gesamtbevölkerung.*

Der Test hat praktisch immer das Ziel, die Nullhypothese zu wiederlegen, indem man zeigt, daß die Wahrscheinlichkeit für ihre Richtigkeit unterhalb der noch tolerierten Irrtumswahrscheinlichkeit liegt. Falls dies der Fall ist, betrachtet man die Alternativhypothese als bestätigt mit einer Irrtumswahrscheinlichkeit von höchstens α oder, wie man auch sagt, auf einem *Signifikanzniveau* von $1-\alpha$; andernfalls ist die Widerlegung der Nullhypothese nicht mit der geforderten Sicherheit gelungen und man kann *nichts* sagen.

Im obigen Beispiel wäre die Widerlegung der Nullhypothese beim *zweiseitigen* Test gelungen, falls die noch tolerierte Irrtumswahrscheinlichkeit über 12% liegt; im Falle des einseitigen Test hätten 6% genügt, so daß man beim einseitigen Test beispielsweise sagen kann, daß bei vorgegebener Irrtumswahrscheinlichkeit von höchstens 10% nachgewiesen wäre, daß der Intelligenzquotient der Studenten *signifikant* über dem der Gesamtbevölkerung liegt oder, anders ausgedrückt, daß auf einem Signifikanzniveau von 90% nachgewiesen wurde, daß der Intelligenzquotient der Studenten über dem der Gesamtbevölkerung liegt.

Bei einer vorgegebenen Irrtumswahrscheinlichkeit von höchstens 5% wäre die Widerlegung der Nullhypothese weder im einseitigen noch im zweiseitigen Fall gelungen. Das bedeutet aber selbstverständlich nicht, daß damit die Nullhypothese *bestätigt* wäre, im Gegenteil: Die Zahlen zeigen deutlich, daß die Alternativhypothese sehr viel wahrscheinlicher ist als die Nullhypothese; wir können nur nicht mit der geforderten 95%-igen Sicherheit ausschließen, daß es sich hierbei nicht um einen Zufallseffekt handelt. Falls wir dies feststellen wollen, müssen wir den Test mit einer größeren Stichprobe wiederholen und hoffen, daß sich dabei ein klareres Bild ergibt.

4. Der Studentsche t-test

Im Beispiel von §3 kannten wir die Standardabweichung $\sigma = 16$ der betrachteten Zufallsvariablen. In den Anwendungen ist so etwas nur sehr selten der Fall: Normalerweise muß man die Standardabweichung schätzen aufgrund der Streuung der Stichprobe. Damit sind wir wieder in derselben Situation wie in §2, d.h. anstelle der normalisierten Normalverteilung muß die STUDENTsche t-Verteilung benutzt werden. Den entsprechenden Test bezeichnet man entsprechend als STUDENTschen t-Test. Er kommt in zwei Varianten:

a) Einfacher t-Test auf den Erwartungswert

Hier sind wir – bis auf die unbekannte Standardabweichung – in der Situation des vorigen Paragraphen:

X sei eine normalverteilte Zufallsvariable, und $x_1, \ldots, x_n$ sei eine Stichprobe aus der zu X gehörenden Grundgesamtheit mit arithmetischem Mittel $\overline{x}$ und Streuung s. Wir möchten wissen, ob diese Stichprobe mit der Annahme kompatibel ist, daß X den (vorgegebenen) Erwartungswert μ hat.

(Im obigen Beispiel wäre X die Zufallsvariable, die den Intelligenzquotienten eines Studenten einer Vorlesung *Quantitative Methoden in der Geographie* beschreibt, und $\mu = 100$.)

Beim *zweiseitigen* Test haben wir also die Nullhypothese

$$H_0\colon E(X) = \mu$$

und die Alternativhypothese

$$H_{!A}\colon E(X) \neq \mu\,.$$

In Analogie zu oben berechnen wir die Testgröße

$$t = \frac{\overline{x} - \mu}{s/\sqrt{N}}\,,$$

wobei $\overline{x}$ den Mittelwert und s die Streuung der Stichprobe bezeichnet, und wir lehnen die Nullhypothese ab mit einer Irrtumswahrscheinlichkeit von höchstens α, wenn

$$|t| > t_\alpha$$

ist, wobei t_α der Grenzwert für α aus einer Tabelle für den zweiseitigen t-Test mit $f = N - 1$ Freiheitsgraden ist.

Beim *einseitigen* Test haben wir entsprechend die Nullhypothese

$$H_0\colon E(X) \leq \mu$$

und die Alternativhypothese

$$H_{!A}\colon E(X) > \mu\,.$$

Auch hier berechnen wir die Testgröße

$$t = \frac{\overline{x} - \mu}{s/\sqrt{N}}$$

und lehnen die Nullhypothese ab mit einer Irrtumswahrscheinlichkeit von höchstens α, wenn

$$t > t_\alpha$$

ist, wobei t_α der Grenzwert für α aus einer Tabelle für den einseitigen t-Test mit $f = N - 1$ Freiheitsgraden ist.

Gelegentlich werden nur Grenzwerte für entweder den einseitigen oder (seltener) den zweiseitigen t-Test angegeben. Dies ist kein Problem, denn der Grenzwert des zweiseitigen Tests für vorgegebenes α ist wie oben im Beispiel gerade gleich dem Grenzwert des einseitigen Tests für $\alpha/2$ und dieselbe Anzahl von Freiheitsgraden.

b) *t*-Test auf die Gleichheit zweier Erwartungswerte bei gleicher Standardabweichung

Hier gehen wir aus von zwei Zufallsvariablen X und Y mit gleicher (aber i.a. unbekannter) Standardabweichung,

einer Stichprobe $x_1, \ldots, x_M$ zu X und einer Stichprobe $y_1, \ldots, y_N$ zu Y. Wir wollen entscheiden, ob die Erwartungswerte der beiden Zufallsvariablen gleich sind, d.h. wir haben die Nullhypothese

$$H_0\colon E(X) = E(Y)$$

und die Alternativhypothese

$$H_{!A}\colon E(X) \neq E(Y)\,.$$

Eine mathematische Begründung des hier anzuwendenden Testverfahrens würde im Rahmen dieser Vorlesung zu weit führen, daher soll nur die Vorgehensweise angegeben werden:

Zunächst berechnet man die arithmetischen Mittel $\overline{x}$ und $\overline{y}$ sowie die Streuungen s_X und s_Y der beiden Stichproben; dann berechnet man daraus den Wert

$$t = \frac{\overline{y} - \overline{x}}{\sqrt{(M-1)s_X^2 + (N-1)s_Y^2}} \cdot \sqrt{\frac{MN(M+N-2)}{M+N}}\,;$$

die entsprechend definierte Zufallsvariable genügt einer t-Verteilung mit $f = N + M - 2$ Freiheitsgraden, so daß die Nullhypothese mit einer Irrtumswahrscheinlichkeit von höchstens α abgelehnt werden kann, wenn

$$|t| > t_\alpha$$

ist, wobei t_α der Grenzwert für α aus einer Tabelle für den zweiseitigen t-Test mit $f = M + N - 2$ Freiheitsgraden ist.

Analog kann auch ein einseitiger Test für die Nullhypothese $E(X) \leq E(Y)$ durchgeführt werden.

5. Der Welsh-Test

Dies ist die aufwendigere Variante des t-Tests für den Fall, daß X und Y zwar normalverteilt sind, daß aber nichts über die Gleichheit der Standardabweichungen bekannt ist. Mit Bezeichnungen wie oben berechnet man in diesem Fall

$$t = \frac{\overline{y} - \overline{x}}{\sqrt{\frac{s_X^2}{M} + \frac{s_Y^2}{N}}}$$

und berechnet die Anzahl f der Freiheitsgrade als größte ganze Zahl, die kleiner oder gleich dem Wert

$$\frac{(M-1)(N-1)\left(\frac{s_X^2}{M} + \frac{s_Y^2}{N}\right)^2}{(N-1)\left(\frac{s_X^2}{M}\right)^2 + (M-1)\left(\frac{s_X^2}{N}\right)^2}$$

ist. Mit diesen Werten verfährt man dann genauso wie beim gewöhnlichen t-Test.

6. Der *F*-Test

Der t-Test auf Gleichheit zweier Erwartungswerte funktioniert nur bei gleicher Standardabweichung der beiden Zufallsvariablen; ansonsten muß man den aufwendigeren WELSH-Test anwenden. Der FISHERsche F-Test überprüft, ob die Annahme gleicher Standardabweichung erfüllt ist. Er ist damit auch anwendbar auf die Untersuchung, ob verschiedene Instrumente gleiche Meßgenauigkeit haben oder ob sich Disparitäten von Verteilungen unterscheiden: Beispielsweise könnte man mit einem F-Test auch untersuchen, ob sich die regionalen Einkommensunterschiede in Deutschland in den letzten zehn Jahren vergrößert oder verkleinert haben.

SIR RONALD AYLMER FISHER (1890–1962), der wohl bedeutendste Statistiker dieses Jahrhunderts, untersuchte die Verteilung der Quotienten s_X^2/s_Y^2 für die Streuungen zweier Stichproben vom Umfang M *bzw.* N zu normalverteilten Zufallsvariablen X und Y gleicher Standardabweichung; die resultierende Verteilung, auf deren (komplizierte) mathematische Form wir zum Glück nicht eingehen müssen, hängt von den beiden Parametern M, N ab und heißt FISHERsche F-Verteilung mit $(M-1, N-1)$ Freiheitsgraden; auch für diese Verteilung sind wieder die Quantilen für gängige Irrtumswahrscheinlichkeiten tabelliert.

Zur Durchführung des F-Tests berechnet man daher den Quotienten

$$F = \frac{s_X^2}{s_Y^2} \quad \text{oder} \quad F = \frac{s_Y^2}{s_X^2}$$

der Streuungen der gegebenen Stichproben, und zwar nimmt man denjenigen der beiden Werte, bei dem die größere Zahl im *Zähler* steht, so daß immer $F \geq 1$ ist. Sodann vergleicht mit dem Grenzwert zur gewählten Irrtumswahrscheinlichkeit. Dabei ist allerdings zu beachten, daß sich die Quantilen natürlich auf einen *einseitigen* Test beziehen, während wir hier auf Gleichheit testen und somit einen *zweiseitigen* Test durchführen; ist α die vorgegebene Irrtumswahrscheinlichkeit, muß man in der Tabelle also bei $\alpha/2$ oder – falls dort statt der Irrtumswahrscheinlichkeit das Signifikanzniveau genannt ist – bei $1-\alpha/2$ nachschauen: Für eine 10%-ige Irrtumswahrscheinlichkeit müßte man also unter dem Signifikanznivau

$$100\% - \frac{1}{2} \cdot 5\% = 95\%$$

nachschauen. Die relevanten Freiheitsgrade sind entweder $(M-1, N-1)$ oder $(N-1, M-1)$, je nachdem ob s_X oder s_Y im Zähler steht. (Man beachte, daß sich die F-Verteilungen mit $(M-1, N-1)$ und mit $(N-1, M-1)$ Freiheitsgraden beträchtlich unterscheiden!)

Dort findet man dann den jeweiligen Grenzwert F_α, und wie üblich wird die Nullhypothese auf dem gegebenen Signifikanzniveau abgelehnt, wenn $F > F_\alpha$ ist.

7. Der Mann-Whitney-Test

Während alle bisher betrachteten Tests von *normalverteilten* Zufallsvariablen ausgingen, wollen wir hier einen sogenannten *nichtparametrischen* Test betrachten, d.h. einen Test, der auch ohne Normalverteilungsannahme funktioniert.

Wieder geht es um zwei Zufallsvariablen X und Y, und wieder haben wir Stichproben $x_1, \ldots, x_M$ und $y_1, \ldots, y_M$ zu den beiden, aber über die Verteilungen von X und Y werden keinerlei Annahmen gemacht. Die Nullhypothese ist

$$H_0\colon X \textit{ und } Y \textit{ haben dieselbe Verteilung}\,,$$

die Alternativhypothese entsprechend

$$H_{!A}\colon X \textit{ und } Y \textit{ haben verschiedene Verteilungen}\,.$$

Zur praktischen Durchführung des Tests werden die beiden Stichproben vereinigt und die entstehende Reihe wird der Größe nach sortiert. Wir sagen, ein Stichprobenelement habe die *Rangzahl* r, wenn es in dieser sortierten Liste an der r-ten Stelle steht, und setzen

$$R_X = \text{Summe der Rangzahlen der sämtlichen } x_i$$

und

$$R_Y = \text{Summe der Rangzahlen der sämtlichen } y_j$$

und berechnen daraus

$$U_X = NM + \frac{M(M+1)}{2} - R_X$$

sowie

$$U_Y = NM + \frac{N(N+1)}{2} - R_Y = NM - U_X\,.$$

Der eigentliche Testwert ist

$$U = \min(U_X, U_Y)\,,$$

das heißt also, der kleinere der beiden Werte. Falls U *kleiner* ist als der für das Paar (M, N) und die gewählte Irrtumswahrscheinlichkeit tabellierte Grenzwert, wird die Nullhypothese gleicher Verteilungen abgelehnt. Teilweise beziehen sich die Tabellen auch auf die *einseitige* Version des Tests, der die Hypothese $X \geq Y$ prüft; für die zweiseitige Version kann man dann mit denselben Zahlen rechnen, wobei diese sich jedoch auf die *doppelte* Irrtumswahrscheinlichkeit beziehen.

Anstelle von U findet man gelegentlich auch Tabellen für

$$T = \frac{U - \frac{MN}{2}}{\sqrt{\frac{MN(M+N+1)}{12}}}\,;$$

diese Zufallsvariable ist für große Werte von M, N annähernd normalverteilt mit Mittelwert null und Standardabweichung σ, so daß man Tabellen der Standardnormalverteilung betrachten kann. In der Praxis kann man ab etwa

$$N \geq 4,\ M \geq 4 \quad \text{und} \quad M + N \geq 20$$

gute Übereinstimmung mit einer Standardnormalverteilung erwarten.

Der Test wird auch gelegentlich als WILCOXON-Test bezeichnet, da WILCOXON denselben Test bereits früher als MANN und WHITNEY im Spezialfall $M = N$ betrachtet hatte.

8. Abschließende Bemerkungen zur Teststatistik

Die meisten Tests, die in diesem Kapitel behandelt wurden, sind sogenannte *parametrische* Tests, das heißt, sie wurden abgeleitet unter der Voraussetzung, daß die betrachtete Zufallsvariable einer vorgegebenen Verteilung (hier immer der Normalverteilung) genügt. Nur unter dieser Voraussetzung ist bewiesen, daß die jeweilige Testgröße durch eine *t- bzw.* F-Verteilung beschrieben werden kann und der Test somit zu einer sinnvollen Aussage führt.

Nun haben wir zwar im vorigen Kapitel einige Regeln darüber kennengelernt, wann man typischerweise von einer Normalverteilung ausgehen kann, und wir werden auch im nächsten Kapitel sehen, daß man die Normalverteilungshypothese mit einem χ^2-Test überprüfen kann, aber manchmal wird man doch nicht ganz sicher sein, ob alle Voraussetzungen für die Anwendbarkeit eines Tests vorliegen und man wird Zweifel über die Relevanz der Ergebnisse haben.

Mit dem Aufkommen von Computern konnte genauer untersucht werden, wie empfindlich die einzelnen Tests auf die Verletzung der zugrundegelegten Annahmen reagieren: In einer sogenannten *Monte-Carlo-Simulation* erzeugt man eine große Anzahl von Zufallsstichproben aus einer vorgegebenen Verteilung und bestimmt (über die Häufigkeitsdefinition der Wahrscheinlichkeit) direkt die Irrtumswahrscheinlichkeit eines

Tests. Die für uns wesentlichsten Ergebnisse solcher Simulationen lassen sich wie folgt zusammenfassen:

a) Beim t-Test auf Gleichheit der Mittelwerte ist die Annahme gleicher Standardabweichungen unkritisch bei *gleichen* Stichprobenumfängen; bei Stichproben verschiedener Größe reagiert der Test aber sehr empfindlich auf die Verletzung dieser Annahme, so daß man dann im Zweifelsfall immer lieber den WELSH-Test nehmen sollte.

b) Falls die Verteilung nicht normal aber immerhin noch symmetrisch ist, liefern sowohl t-Test als auch WELSH-Test immer noch recht brauchbare Ergebnisse. Schiefe Verteilungen dagegen können zu extremen Abweichungen zwischen Testaussage und Wirklichkeit führen, vor allem wenn die Stichprobenumfänge verschieden sind.

c) Je kleiner die Stichprobenumfänge sind, desto empfindlicher reagieren die Tests auf schiefe Verteilungen; die Irrtumswahrscheinlichkeit kann dann bei kleinen Stichproben erheblich über dem Wert liegen, den der Test angibt. Dies ist umso kritischer, als man bei kleinen Stichproben anhand der Daten keine zuverlässige Aussage über die Schiefe machen kann, so daß sich in solchen Fällen empfiehlt, grundsätzlich immer den WELSH-Test anstelle des t-Tests zu verwenden. Außerdem sollte man erstens wann immer möglich mit gleichen Stichprobenumfängen arbeiten und zweitens zur Vorsicht lieber den Grenzwert für eine *geringere* Irrtumswahrscheinlichkeit als die eigentlich interessierende zu benutzen.

d) Bei sehr schiefen Verteilungen mit verschiedenen Standardabweichung liefert der Mann-Whitney-Test keine zuverlässige Information über die Gleichheit von Mittelwerten; hier ist im Zweifelsfall der WELSH-Test zuverlässiger (wenn auch nicht unbedingt zuverlässig).

e) Der F-Test reagiert sehr sensibel auf Abweichungen von der Normalverteilungsannahme, auch wenn die Verteilungen noch symmetrisch sind. Bei kleinen Stichproben, über deren Verteilung man keine genaueren Kenntnisse hat, kann ein F-Test daher keine brauchbaren Ergebnisse liefern.

Genaueres zu den hier angesprochenen Problemen sowie noch zahlreiche hier nicht angesprochene Testverfahren findet man in speziellen Lehrbüchern der Teststatistik, sehr ausführlich beispielsweise in

J.M. DIEHL, R. ARBINGER: Einführung in die Inferenzstatistik, *Eschborn*, ^{2}1992

Ein weiteres Problem sollte noch wenigstens kurz angesprochen werden: Bei den hier vorgestellten Tests ist immer eine Irrtumswahrscheinlichkeit vorgegeben dafür, daß die Nullhypothese abgelehnt wird, obwohl sie tatsächlich richtig ist; diese fälschliche Ablehnung wird gelegentlich als *Fehler erster Art* bezeichnet. Ein *Fehler zweiter Art* liegt dann vor, wenn die Nullhypothese nicht abgelehnt werden kann, obwohl sie falsch ist.

Da sich die Nullhypothese im allgemeinen auf die Gleichheit zweier Zahlen bezieht, muß sich zunächst genauer überlegen, was man unter der Falschheit der Nullhypothese verstehen will: Eine relative Abweichung von einem Milliardstel Prozent wird man üblicherweise noch nicht als Verschiedenheit interpretieren. Man muß sich also Schranke vorgegeben und redet nur dann von einem Fehler zweiter Art, wenn der tatsächliche Unterschied der beiden Werte über der Schranke liegt und der Test trotzdem nicht die Ablehnung der Nullhypothese erlaubt.

Die Wahrscheinlichkeit für einen solchen Fehler läßt sich nur schwer abschätzen; meist arbeitet man dazu mit Monte-Carlo-Simulationen. Klar ist, daß Fehlers zweiter Art umso wahrscheinlicher sind, je geringer die Irrtumswahrscheinlichkeit für einen Fehler erster Art angesetzt wird, so daß man nicht beide Fehlerwahrscheinlichkeiten gleichzeitig gering halten kann.

Tabellen für die Wahrscheinlichkeiten von Fehlern zweiter Art findet man zum Beispiel in

J. COHEN: Statistical power analysis for the behavioral sciences, *Hillsdale*, ^{2}1988

§6: Anpassungstests und lineare Regression

Bei den Test im letzten Kapitel ging es um die Überprüfung einer fertig vorliegenden Hypothese. Häufig steht man aber auch vor der allgemeineren Situation, daß man seine Hypothese zwar *grundsätzlich* formuliert hat, daß diese aber noch von Parametern abhängt, die erst aufgrund der vorliegenden Daten zu schätzen sind. Tests, die sich auf solche Fragestellungen beziehen, haben daher *zwei* Aufgaben:

- Als erstes müssen die freien Parameter in der Hypothese so gewählt werden, daß die Hypothese (in der Grenzen ihrer grundsätzlich vorgegebenen Struktur) optimal an das vorliegende Datenmaterial angepaßt ist, und
- dann muß getestet werden, ob die so angepaßte Hypothese mit hinreichender Sicherheit die vorliegenden Daten erklärt.

Das Standardverfahren in einer solchen Situation ist der χ^2-Test, der beide Aufgaben sowohl einzeln als auch gemeinsam erfüllen kann.

1. Die χ^2-Verteilung

Ausgangspunkt ist die folgende Überlegung: Angenommen, wir haben zwei Größen, bei denen wir *wissen*, daß die eine durch die andere vollständig bestimmt ist wie beispielsweise der Siedepunkt des Wassers durch den Luftdruck oder die Leitfähigkeit vieler Lösungen durch die Konzentration des gelösten Stoffs. Dann gibt es einen funktionalen Zusammenhang

$$y = f(x)$$

zwischen den beiden Größen, der genau angibt, wie man die eine aus der anderen berechnen kann. Ist etwa $x = I$ die Stromstärke, die durch einen Widerstand R fließt, und ist $y = U$ die an R abfallende Spannung, so gibt des OHMsche Gesetz diesen Zusammenhang in der Form

$$U = R \cdot I \, .$$

Falls man nun allerdings die sowohl Stromstärke als auch Spannung mehrfach mißt und die jeweiligen Paare (U_i, I_i) von Meßwerten betrachtet, so wird aufgrund allfälliger Meßfehler im allgemeinen $U_i \neq R \cdot I_i$ sein; die Abweichung zwischen beiden sollte, wie bei Meßfehlern üblich, ungefähr normalverteilt sein mit Mittelwert 0 und einer Standardabweichung σ, die von der Meßgenauigkeit abhängt. Die Größen

$$\frac{U_i - R \cdot I_i}{\sigma}$$

genügen also einer Standardnormalverteilung.

Um ein Maß für den Gesamtfehler zu bekommen, bei dem sich die einzelnen Fehler nicht aufgrund verschiedener Vorzeichen kompensieren können, kann man beispielsweise die Größe

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{U_i - R \cdot I_i}{\sigma} \right)^2$$

zu betrachten, und genau diese Größe ist es, die dem χ^2-Test seinen Namen gab.

Die zugrundeliegende Verteilung, die χ^2-Verteilung läßt sich unmittelbar aus der Normalverteilung ableiten: Sind $X_1, \ldots, X_n$ Zufallsvariablen, die jeweils einer Standardnormalverteilung genügen, so bezeichnet man die Verteilung der Zufallsvariablen

$$X_1^2 + \cdots + X_n^2$$

als *χ^2-Verteilung mit n Freiheitsgraden.*

Wahrscheinlichkeitsdichte und Verteilungsfunktion der χ^2-Verteilung lassen sich daher mathematisch ableiten aus der Normalverteilung; da aber diese schon relativ unangenehm war und wir, wie stets in der Test- und Schätzstatistik nur die in Tabellen zu findenden Grenzwerte für vorgegebene Irrtumswahrscheinlichkeit brauchen, können wir im Rahmen dieser Vorlesung auf eine genauere mathematische Untersuchung der χ^2-Verteilung verzichten.

Die Wahrscheinlichkeitsdichte der χ^2-Verteilung kann – da es sich ja um die Verteilung einer Summe von Quadraten handelt – nur für positive Werte der Zufallsvariablen von Null verschieden sein; für diese ergibt sie sich zu

$$f(x) = K_n x^{(n-2)/2} e^{-x/2}$$

mit

$$K_n = \begin{cases} \frac{1}{2^{n/2}\left(\frac{n}{2}-1\right)!} & \text{für gerades } n \\ \frac{1}{2^{(n-1)/2}\sqrt{2\pi}\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{1}{2}\right)} & \text{für ungerades } n \, . \end{cases}$$

Die Verteilung hat den Erwartungswert n und die Standardabweichung $\sqrt{2n}$; für großes n kann sie näherungsweise durch eine Normalverteilung beschrieben werden.

Im obigen Beispiel sollte das dort definierte χ^2 also einer χ^2-Verteilung mit N Freiheitsgraden genügen; falls der beobachtete Wert von χ^2 stark von zu stark von deren Erwartungswert abweicht, wird man davon ausgehen, daß die Nullhypothese $U = R \cdot I$ im vorliegenden Fall falsch ist.

Meist wird man allerdings nicht mehrere Messungen von Spannung und Stromstärke durchführen nur um die Hypothese „Widerstand $= R$“ zu testen, sondern man wird messen, um die Größe des Widerstands zu bestimmen. Hier sind wir in der Situation eines Anpassungstest: Wir betrachten die Größe

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{U_i - R \cdot I_i}{\sigma} \right)^2$$

nun nicht mehr für festes R, sondern wir betrachten sie als Funktion des unbekannten Parameters R und wählen als wahrscheinlichsten Wert für R denjenigen Wert, der diesen Ausdruck minimal macht *(maximum likelihood Ansatz).* Im vorliegenden Fall wäre dies, wie man leicht durch Differenzieren nachrechnet, einfach das arithmetische Mittel der Quotienten U_i/I_i, also das, was man auch ohne diesen komplizierten Ansatz erwartet hätte, aber in allgemeineren Fällen kann die Bestimmung des optimalen Werts oder der optimalen Werte auch sehr viel komplizierter sein und auf ganz und gar nicht offensichtliche Ergebnisse führen.

2. Test auf das Vorliegen einer erwarteten Verteilung

Die für uns wichtigste Anwendung ist die Überprüfung von Annahmen über die Verteilung einer Grundgesamtheit. Hierzu teilt man die Stichprobe in mehrere Klassen ein und berechnet die theoretischen Klassenstärken aufgrund der angenommenen Verteilungsfunktion. Falls man genügend viele Klassen hat, kann man davon ausgehen, daß die tatsächlich beobachteten Elementanzahlen ungefähr POISSONverteilt sind, wobei der Erwartungswert (bei korrekter Annahme über die Verteilung) gleich dem theoretischen Wert sein sollte. Da die Standardabweichung einer POISSONverteilung gleich der Quadratwurzel aus dem Erwartungswert ist, bestimmt sich χ^2 hier als

$$\chi^2 = \sum_{i=1}^{r} \frac{(b_i - e_i)^2}{e_i} ,$$

wobei r die Anzahl der Klassen bezeichnet und b_i, e_i die beobachtete und die theoretisch erwartete Elementanzahl der i-ten Klasse bezeichnen. Diese Größe sollte also χ^2-verteilt sein; die Anzahl der Freiheitsgrade ist

$$f = r - a ,$$

wobei a die Anzahl der Parameter der theoretischen Verteilung ist.

Betrachten wir dazu ein einfaches Beispiel, in dem es im wesentlichen nur um die Überprüfung einer Hypothese geht: Angenommen, an 120 Tagen wurden an jeweils derselben Stelle folgende vorherrschende Windrichtungen beobachtet:

Richtung	N	NW	W	SW	S	SO	O	NO
an Tagen	17	19	18	16	13	12	12	13

Läßt sich aufgrund dieser Daten sagen, daß eine bestimmte Windrichtung häufiger auftritt als die anderen, oder sind die Richtungen bis auf Zufallsschwankungen gleichverteilt?

Falls die Richtungen exakt gleichverteilt sind, sollte man erwarten, daß jede der acht Richtungen an jeweils etwa

$$\frac{120}{8} = 15$$

Tagen beobachtet wird; hier ist also $r = 8$ und $e_i = 15$ für alle i. Somit ist

$$\begin{aligned}\chi^2 &= \frac{(17-15)^2}{15} + \frac{(19-15)^2}{15} + \frac{(18-15)^2}{15} + \frac{(16-15)^2}{15}\\ &\quad + \frac{(13-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(12-15)^2}{15} + \frac{(13-15)^2}{15}\\ &= \frac{2^2+4^2+3^2+1^2+(-2)^2+(-3)^2+(-3)^2+(-2)^2}{15}\\ &= \frac{56}{15} = 3{,}7\overline{3}\,.\end{aligned}$$

Da die Gleichverteilung genau einen Parameter hat, ist $a = 1$, wir müssen also eine χ^2-Verteilung mit $r - 1 = 7$ Freiheitsgraden betrachten. Geben wir uns eine Irrtumswahrscheinlichkeit von 5% vor, so zeigt die Tabelle, daß der entsprechende Grenzwert der χ^2-Verteilung mit sieben Freiheitsgraden gleich 14,1 ist; da der berechnete Wert deutlich unter dieser Schranke liegt, gibt es keinen Grund, die Nullhypothese abzulehnen.

Etwas knapper wird es, wenn wir die mittleren Oktobertemperaturen aus München an eine Normalverteilung anpassen: Nehmen wir als Klassen die Intervalle von jeweils

einem Grad, so ergeben sich aufgrund der Normalverteilung folgende Schätzungen:

Temperatur	*beobachtet*	*theoretisch*
5,0°C–5,9°C	4	3,1
6,0°C–6,9°C	6	8,0
7,0°C–7,9°C	11	12,9
8,0°C–8,9°C	21	12,9
9,0°C–9,9°C	4	8,1
10,0°C–10,9°C	2	3,2
11,0°C–11,9°C	2	0,8

Somit ist

$$\chi^2 = \frac{(4-3{,}1)^2}{3{,}1} + \cdots + \frac{(2-0{,}8)^2}{0{,}8} \approx 10{,}45\,.$$

Da die Normalverteilung *zwei* Parameter hat, die aufgrund der Daten geschätzt werden mußten (nämlich den Erwartungswert und die Standardabweichung), haben wir bei sieben Klassen fünf Freiheitsgrade; der Grenzwert für eine Irrtumswahrscheinlichkeit von 5% ist daher gleich 11,07. Somit kann die Nullhypothese nicht abgelehnt werden. Sie kann allerdings abgelehnt werden, wenn wir die Irrtumswahrscheinlichkeit auf 10% erhöhen, denn dann sinkt der Grenzwert auf 9,24. Die Annahme einer Normalverteilung ist also zwar mit mindestens 90%-iger Wahrscheinlichkeit falsch, nicht aber mit 95%-iger.

3. Vierfeldertafeln

Ein wichtiger Sonderfall des vorigen Paragraphen sind die sogenannten Vierfeldertafeln: Untersucht werden zwei

Stichproben, die sich in irgendeinem äußeren Merkmal unterscheiden: Beispielsweise könnte man die Sterblichkeit von Kranken untersuchen, von denen die eine Gruppe mit einem neuen Medikament behandelt wurde, die andere mit einem alten, oder man könnte bei Kunden zweier Kaufhäuser danach fragen, ob sie aus derselben Stadt oder von auswärts kommen. In jedem Fall hat man nach Ende der Untersuchung vier Werte a, b, c, d, wobei a die Anzahl der Probanden aus Gruppe I ist, bei denen Resultat I vorliegt, b die, bei denen Resultat II vorliegt, und entsprechend seien c, d die Zahlen für Gruppe II. Nullhypothese ist, daß sich die beiden Gruppen nicht unterscheiden.

Sind daher M und N die beiden Gruppenstärken, und haben insgesamt r Probanden das Merkmal I und die restlichen s das Merkmal II (d.h. $M + N = r + s$), so erwarten wir unter der Annahme der Nullhypothese, daß

$$E_a = \frac{r}{M+N} \cdot M$$

der M Probanden aus Gruppe I Merkmal I haben und

$$E_b = \frac{s}{M+N} \cdot M$$

das Merkmal II. Die entsprechenden Zahlen für Gruppe II sind entsprechend

$$E_c = \frac{r}{M+N} \cdot N$$

für Merkmal I und

$$E_d = \frac{s}{M+N} \cdot N$$

für Merkmal II. Die zu überprüfende Summe ist also

$$\chi^2 = \frac{(a-E_a)^2}{E_a} + \frac{(b-E_b)^2}{E_b} + \frac{(c-E_c)^2}{E_c} + \frac{(d-E_d)^2}{E_d} .$$

Da die Summen $M + N$ und $r + s$ vorgegebene Zahlen sind, die keinen statistischen Schwankungen unterliegen, und wir mit den Quotienten $\frac{r}{M+N}$ einen Parameter der Gleichverteilung geschätzt haben, genügt dieser Wert einer χ^2-Verteilung mit nur *einem* Freiheitsgrad.

Sind beispielsweise bei einer neuen Therapie $a = 4$ von $M = 81$ Probanden gestorben und mit der alten Therapie $c = 15$ von $N = 100$, so ist zunächst

$$b = M - a = 77 \quad \text{und} \quad d = N - c = 85 ,$$

also

$$r = a + c = 19 \quad \text{und} \quad s = b + d = 162 .$$

Damit ist

$$\frac{r}{M+N} = \frac{19}{181} \approx 0{,}10497 \quad \text{und} \quad E_a \approx 0{,}10497 \cdot M \approx 8{,}5 .$$

Entsprechend ergeben sich

$$E_b \approx 72{,}5, \quad E_c \approx 10{,}5 \quad \text{und} \quad E_d \approx 89{,}5 .$$

Dies führt zu

$$\chi^2 \approx 4{,}822 ;$$

falls man davon ausgeht, daß die neue Therapie sicher nicht schlechter ist als die alte und somit einen einseitigen Test durchführt, ergibt sich als Grenzwert für die Irrtumswahrscheinlichkeit von 5% der Wert 2,706; da das berechnete χ^2 größer ist, kann die Nullhypothese mit mindestens 95%-iger Sicherheit abgelehnt werden; die neue Therapie scheint also etwas zu nützen.

3. Lineare Regression

Hier geht es darum, zu gegebenen Datenpaaren $(x_i y_i)$, zwischen denen ein linearer Zusammenhang vermutet wird, die besten Koeffizienten a, b zu bestimmen, so daß

$$y_i \approx a x_i + b$$

ist. Im Sinne obiger Strategie muß also

$$\sum_{i=1}^{N} \frac{\big((y_i - (a x_i + b)\big)^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \big((y_i - (a x_i + b)\big)^2$$

minimiert werden, wobei es offensichtlich auf den (i.a. unbekannten) Faktor σ^2 nicht ankommt.

Wie aus der Schule bekannt ist, kann eine Funktion nur dort ein relatives Minimum haben, wo ihre Ableitung verschwindet. Hier hängt die Funktion von zwei Variablen a, b ab, also müssen die beiden Ableitungen nach a und b verschwinden. Wir betrachten also die Funktion

$$F(a, b) = \sum_{i=1}^{N} (y_i - a \cdot x_i - b)^2$$

Bevor wir sie nach den Parametern a und b ableiten, empfiehlt es sich, zunächst auszumultiplizieren und a und b auszuklammern:

$$F(a, b) = \sum_{i=1}^{N} y_i^2 + a^2 \cdot \sum_{i=1}^{N} x_i^2 + n \cdot b^2 - 2 \cdot a \cdot \sum_{i=1}^{N} y_i \cdot x_i$$
$$- 2 \cdot b \cdot \sum_{i=1}^{N} y_i + 2 \cdot a \cdot b \cdot \sum_{i=1}^{N} x_i .$$

Damit ist

$$F_a(a, b) = 2 \cdot a \cdot \sum_{i=1}^{N} x_i^2 - 2 \cdot \sum_{i=1}^{N} y_i \cdot x_i + 2 \cdot b \cdot \sum_{i=1}^{N} x_i$$

und

$$F_b(a, b) = 2 \cdot n \cdot b - 2 \cdot \sum_{i=1}^{N} y_i + 2 \cdot a \cdot \sum_{i=1}^{N} x_i .$$

Da beide partielle Ableitungen verschwinden sollen, müssen beide Ausdrücke Null sein; nach Division durch zwei erhalten wir also die beiden Gleichungen

$$a \cdot \sum_{i=1}^{N} x_i^2 + b \cdot \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i \cdot y_i$$

$$\text{und} \qquad a \cdot \sum_{i=1}^{N} x_i + b \cdot n = \sum_{i=1}^{N} y_i .$$

Um b zu eliminieren, multiplizieren wir die zweite Gleichung mit $\frac{1}{n} \sum_{i=1}^{N} x_i$

$$a \cdot \frac{1}{n} \left(\sum_{i=1}^{N} x_i \right)^2 + b \cdot \sum_{i=1}^{N} x_i = \frac{1}{n} \left(\sum_{i=1}^{N} x_i \right) \cdot \left(\sum_{i=1}^{N} y_i \right)$$

und subtrahieren dies von der ersten Gleichung:

$$a \cdot \left(\sum_{i=1}^{N} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{N} x_i \right)^2 \right) = \sum_{i=1}^{N} x_i \cdot y_i - \frac{1}{n} \left(\sum_{i=1}^{N} x_i \right) \cdot \left(\sum_{i=1}^{N} y_i \right) ,$$

d.h.

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

Entsprechend läßt sich a eliminieren, wenn wir die erste Gleichung mit $\sum x_i$ und die zweite mit $\sum x_i^2$ multiplizieren, und wir erhalten

$$b = \frac{\left(\sum_{i=1}^{N} x_i \right) \cdot \left(\sum_{i=1}^{N} x_i \cdot y_i \right) - \left(\sum_{i=1}^{N} x_i^2 \right) \cdot \left(\sum_{i=1}^{N} y_i \right)}{\left(\sum_{i=1}^{N} x_i \right)^2 - n \sum_{i=1}^{N} x_i^2} .$$

Als Ergebnis können wir also festhalten, daß

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

und

$$b = \frac{\left(\sum_{i=1}^{N} x_i\right)\cdot\left(\sum_{i=1}^{N} x_i\cdot y_i\right) - \left(\sum_{i=1}^{N} x_i^2\right)\cdot\left(\sum_{i=1}^{N} y_i\right)}{\left(\sum_{i=1}^{N} x_i\right)^2 - n\sum_{i=1}^{N} x_i^2}\,.$$

Als Beispiel dazu können wir die Voraussagen des Sachverständigenrats über den realen Anstieg des Bruttosozialprodukt (in den alten Bundesländern) mit den tatsächlichen Zahlen vergleichen:

Jahr:	1982	1983	1984	1985	1986	1987
vorausgesagt:	$+\frac{1}{2}$	$+1$	$+2\frac{1}{2}$	$+3$	$+3$	$+2$
tatsächlich:	$-1,1$	$+1,9$	$+3,1$	$+1,8$	$+2,2$	$+1,5$

Jahr:	1988	1989	1990	1991	1992
vorausgesagt:	$+1\frac{1}{2}$	$+2\frac{1}{2}$	$+3$	$+3$	$+2$
tatsächlich:	$+3,7$	$+4,0$	$+4,9$	$+3,6$	$+0,9$

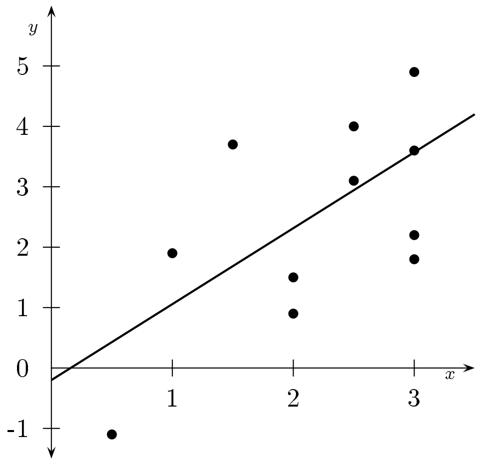
Mit x_i = *vorausgesagtes Wachstum* und y_i = *tatsächliches Wachstum* ergibt sich nach Einsetzen in die obigen Formeln $a \approx 1,2$ und $b = -0,2$; die wirklichen Zahlen liegen also im Durchschnitt etwa 20% höher als die vorausgesagten. Genauer können wir sagen, daß der Sachverständigenrat *Trends*, das heißt Veränderungen der Wachstumsrate, *unterschätzt*, denn $(y_i - y_j) = a\cdot(x_i - x_j)$ ist im Durchschnitt 20% größer als das vorhergesagte $(x_i - x_j)$, und daß er im übrigen einen gewissen Grundoptimismus hat, der durch die Subtraktion von $0,2$ ausgeglichen werden muß.

Graphisch zeigt dies die Verteilung der Wertepaare für vorhergesagtes und tatsächliches Wachstum ist in der folgenden Abbildung; dort ist auch die Gerate $u = a\cdot t + b$ eingezeichnet. Wie man sieht, geht die Gerade zwar mitten durch die Punktwolke, aber viele Punkte liegen doch recht weit entfernt von ihr.

4. Korrelation

Das gerade diskutierte Beispiel wirft die Frage auf, wie man die Qualität einer Ausgleichsgerade messen kann. Eine Antwort darauf liefert ein Argument, daß wir schon bei der Bestimmung des Fehlers des Mittelwerts in §4 benutzt haben: Seien im folgenden $\bar{x}$ der Mittelwert der x_i und $\bar{y}$ der Mittelwert der y_i. Falls die y_i und die x_i aus voneinander unabhängigen Prozessen stammen, ist der Erwartungswert von $\sum_{i=1}^{N}(x_i - \bar{x})\cdot(y_i - \bar{y})$ Null, da die Summen der $x_i - \bar{x}$ und der $y_i - \bar{y}$ jeweils Null sind. Falls aber im anderen Extrem $y_i = a\cdot x_i + b$ ist für alle i, so ist $\bar{y} = a\cdot\bar{x} + b$ und

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



eine (fast immer) positive Zahl, die eng mit der Standardabweichung der Verteilung der x_i zusammenhängt. Um sie weiter auszurechnen, beachten wir, daß

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

ist und damit

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

oder

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

wobei im Zähler immer die nichtnegative Quadratwurzel gewählt werden soll. Damit ist also der Quotient

$$r = \frac{\sum_{i=1}^{N}(x_i-\bar{x}) \cdot (y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{N}(y_i-\bar{y})}}$$

gleich ± 1, je nach Vorzeichen von a, wenn alle $y_i = a \cdot x_i + b$ sind, und $r = 0$, wenn die y_i unabhängig von den x_i sind. Mit elementaren Grundkenntnissen der Vektorrechnung läßt sich zeigen, daß r immer zwischen -1 und $+1$ liegt, und auch ohne Vektorrechnung ist das eine zwar etwas längliche, aber nicht allzu schwierige Rechnung. Dieses r nennen wir den *Korrelationskoeffizienten;* sein Betrag gibt an, wie gut die Beziehung zwischen den y_i und den x_i durch eine Gerade beschrieben werden kann, während sein Vorzeichen angibt, ob diese Gerade ansteigt oder sinkt. Im obigen Zahlenbeispiel ist $r \approx 0{,}62$, die Approximation ist also eher mittelprächtig.

Selbst bei sehr viel besserer Korrelation müßten wir uns allerdings immer noch fragen, ob wirklich ein *inhaltlicher* Zusammenhang besteht, oder ob es sich nur um eine *formale* Korrelation handelt. Ein beliebtes Beispiel, zu dem leider nirgends Zahlenwerte zu finden sind, ist die gute Korrelation zwischen dem Rückgang der Storchpopulationen und der Geburtenrückgang zur Zeit der Industrialisierung: Trotz guter Korrelationskoeffizienten glauben hier nur wenige Forscher an einen inhaltlichen Zusammenhang.