Statistik für Verfahrenstechniker und Chemie-Ingenieure

Jürgen Raasch

unter Mitarbeit von Wulf Alex

2010

Karlsruhe

Ausgabedatum: 9. März 2012

Jürgen Raasch
Am Kirchberg 43
76229 Karlsruhe

juergen.k.raasch@t-online.de

Wulf Alex
Rieslingweg 14
76356 Weingarten (Baden)

alex-weingarten@t-online.de

Das Skriptum ist unvollständig und kann Fehler enthalten. Für Hinweise sind wir dankbar. Skriptum, Korrekturen und Ergänzungen finden sich unter:

`http://www.alex-weingarten.de/skripten/statistik/`

mit den Hyperlinks

- `http://www.alex-weingarten.de/skripten/statistik/buch.pdf` (Ganzes Skriptum)
- `http://www.alex-weingarten.de/skripten/statistik/vorwort.pdf` (Vorwort)
- `http://www.alex-weingarten.de/skripten/statistik/inhalt.pdf` (Inhalt)
- `http://www.alex-weingarten.de/skripten/statistik/probe.pdf` (Probeabschnitt)
- `http://www.alex-weingarten.de/skripten/statistik/errata.pdf` (Errata)

Vorwort

Stochastische Vorgänge – das heißt Vorgänge, bei denen der Zufall einen mehr oder weniger großen Einfluss hat – spielen in Wirtschaft, Wissenschaft und Technik eine erhebliche Rolle. Deshalb wäre es wünschenswert, wenn Kenntnisse der mathematischen Statistik, also der mathematische Beschreibung stochastischer Vorgänge, weit verbreitet wären. Das Gegenteil ist der Fall. In vielen Ausbildungsgängen wird die mathematische Statistik eher als ein lästiges Randgebiet angesehen mit der Folge, dass Ingenieure und Vertreter ähnlicher Berufe sich zwar gut mit Differentialgleichungen auskennen, aber unsicher werden, sobald es um Fragen der mathematischen Statistik geht.

Wir beide – der eine Maschinenbauer, der andere Elektrotechniker – haben unser Berufsleben in einem Hochschulinstitut für Verfahrenstechnik verbracht und dabei in Forschung und Lehre viel mit Statistik zu tun gehabt. Als Ingenieure sehen wir die Statistik unter praktischen Gesichtspunkten, sie ist für uns ein Werkzeug. Die theoretischen Hinter-, Unter- und Abgründe überlassen wir gern den Mathematikern.

Die mathematischen Schwierigkeiten und damit auch die Voraussetzungen zum Verständnis unseres Textes halten sich in Grenzen. Es kommen ein paar Differentialquotienten und Integrale vor, aber ein großer Teil der Rechnungen geht nicht über die Schulmathematik hinaus. Die eigentlichen Hürden sind die statistischen Begriffe und Vorstellungen, die wir deshalb ausführlich und mit vielen Beispielen erläutern, auch vor dem Hintergrund, dass einige der Begriffe im Alltag ungenau und missverständlich benutzt werden.

Die Statistik gehört zu den Grundlagen vieler spezieller Wissensgebiete wie Fehler- und Ausgleichsrechnung, Versuchsplanung und Probenahme, Qualitätskontrolle, Mischungsanalyse, Spieltheorie, Risikoanalyse, Epidemiologie, Ökonometrie, Kryptologie, Meinungsforschung, Signaltheorie, nichtlineare Optimierung, Chaostheorie und weiterer. An einigen Stellen gehen wir auf derartige Themen ein. Unser Text will und kann Monografien zu diesen Fragen jedoch nicht annähernd ersetzen.

Wir beginnen mit einigen Beispielen, die das Verständnis und das Interesse für Fragestellungen der Statistik wecken sollen, sowie der Klärung einiger allgemeiner Begriffe. Es folgt ein ausführliches Kapitel zur Darstellung von Stichprobenergebnissen als Häufigkeitsverteilungen. Erst danach wird der schwierigere Begriff der Wahrscheinlichkeit eingeführt, der benötigt wird, um zum einen Grundgesamtheiten zu beschreiben und zum anderen Zusammenhänge zwischen dem Stichprobenergebnis und der zugehörigen Grundgesamtheit zu formulieren.

In den anschließenden Kapiteln werden einige wichtige diskrete und stetige Verteilungsfunktionen vorgestellt. Deren Anzahl ließe sich beliebig vergrößern. Unsere Auswahl beschränkt sich auf solche Verteilungen, die für die im weiteren Verlauf erläuterten praktischen Anwendungen in der Verfahrenstechnik gebraucht werden.

Bei der Darstellung von Messergebnissen wie auch bei der Veröffentlichung von Umfrageergebnissen sollte die Angabe von Konfidenzintervallen selbstverständlich sein, ist es aber immer noch nicht. Dieses Thema wird deshalb in unserem Text eingehend behandelt, insbesondere was die jeweiligen Voraussetzungen betrifft. Statistische Prüfverfahren (Tests) gehören zu den wichtigsten Anwendungsgebieten der mathematischen Statistik. Es gibt unzählige spezielle Prüfverfahren. Wir beschränken uns auf die Darstellung der prinzipiellen Vorgehensweise an einem leicht verständlichen Beispiel. Das Thema Varianzanalyse haben wir beiseite gelassen, auch aus dem einfachen Grund, dass uns praktische Anwendungen sehr selten beschäftigt haben. Regression und Korrelation werden dagegen in einem eigenen Kapitel behandelt.

Im vorletzten Kapitel haben wir einige wichtigere Anwendungen der mathematischen Statistik in der Mechanischen Verfahrenstechnik, unserem ehemaligen Arbeitsgebiet, zusammengestellt. Den Abschluss bildet ein Kapitel zum Einsatz des Computers bei statistischen Rechnungen, das zum Zeitpunkt der ersten Veröffentlichung des Skriptums im Netz mehr ein Platzhalter für künftige Erweiterungen als eine Informationsquelle ist. Auch die vorangehenden Kapitel dürften in den ersten Jahren ihres Daseins im World Wide Web (WWW) manche Änderung oder Ergänzung erfahren.

Der Text geht auf ein Skriptum zu einer Vorlesung *Statistische Methoden in der Verfahrenstechnik* zurück, die der Erstautor von 1979 bis 2006 in der Universität Karlsruhe (TH) gehalten hat, und vor ihm Karl Sommer, jetzt Weihenstephan. In der Terminologie und der Wahl der Formelzeichen passen wir uns der deutschsprachigen Wikipedia an, um das Nachschlagen zu erleichtern. Wo es angebracht erscheint, nennen wir auch die englischen Fachausdrücke. Hierbei war uns das Glossar des *International Statistical Institute* (`http://isi.cbs.nl/`) in Den Haag eine Hilfe. Erstmals im Internet (Web) wurde das Skriptum am 17. Februar 2010 in Form einer pdf-Datei veröffentlicht. Die vorliegende Fassung des Skriptums wurde auf einem PC unter Debian GNU/Linux mit Hilfe der Programme `vi`, `gnuplot`, `xfig` und `pdflatex` hergestellt.

Es liegt uns am Herzen, dass Sie unsere Ausführungen verstehen. Scheuen Sie sich nicht, uns per Email zu fragen, wenn wir uns nicht klar genug ausdrücken oder wenn Sie meinen, dass ein wichtiges Thema fehlt. Es ist auch nicht auszuschließen, dass wir uns gelegentlich irren, aber das sollte ein seltenes Ereignis sein.

Karlsruhe, Anfang 2010

Jürgen Raasch Wulf Alex

Übersicht

Inhalt

Abbildungen

Tabellen

Formelzeichen

$\mathrm{A}, \mathrm{B}, \mathrm{C}$	Ereignisse
$\bar{\mathrm{A}}$	Ereignis nicht-A
a, b, c	reelle Zahlen
E	sicheres Ereignis
$E(x)$	Erwartungswert der Zufallsvariablen x
e	Eulersche Zahl 2,71828...
f	Formfaktor disperser Elemente
$F(x)$	Summenfunktion der rel. Häufigkeit bei einem stetigen Merkmal (2.25)
$f(x)$	Dichtefunktion der rel. Häufigkeit bei einem stetigen Merkmal (2.24)
$g(x)$	beliebige Funktion von x
H	Summenfunktion der rel. Häufigkeit bei einem diskreten M. (2.17)
h	relative Häufigkeit (2.2)
I_{50}	Interquartilbereich einer Stichprobe
I_{80}	Interdezilbereich einer Stichprobe
i, k, l	Index, laufende Nummer, natürliche Zahl
l	Schätzwert für einen Parameter λ der Verteilung einer Gg.
M	Moment, Messfehler
m	Medianwert einer Stichprobe, Masse
$\mathcal{N}(0, 1)$	normierte Normalverteilung (5.7)
n	natürliche Zahl, Anzahl der Klassen oder Freiheitsgrade
P, p, q	Wahrscheinlichkeit
$Q_0(x)$	Summenfunktion der Anzahlverteilung in der Partikelmesstechnik
$q_0(x)$	Dichtefunktion der Anzahlverteilung in der Partikelmesstechnik
$Q_3(x)$	Summenfunktion der Volumen- oder Massenverteilung in der P.
$q_3(x)$	Dichtefunktion der Volumen- oder Massenverteilung in der P.
R	Spannweite (range) einer Stichprobe
r	Korrelationskoeffizient einer Stichprobe (8.40)
S	Oberfläche (surface)
s	Standardabweichung einer Stichprobe
s^2	Varianz einer Stichprobe (2.19)
$\tilde{s}^2$	Varianz einer Stichprobe (2.20)
T	Lageparameter der Weibull-Verteilung (5.79)
t	Faktor, reelle Zahl, Variable, Zeit
u	Hilfsvariable
V	Volumen
$V(x)$	Varianz der Zufallsvariablen x

v	Zufallsvariable
x, y	Merkmal, Zufallsvariable, Partikelgröße
$\bar{x}$	Mittelwert einer Stichprobe (2.18)
z	Anzahl, Besetzungszahl, Inhalt einer Stichprobe
α, β	Wahrscheinlichkeiten bei Tests
$\Gamma(\alpha)$	Gammafunktion (5.62)
γ	Schiefe einer Verteilung, Konfidenzzahl
$\gamma(a, t)$	unvollständige Gammafunktion (9.123)
Δx	Klassenbreite
δ	Abweichung
ϵ	beliebig kleine reelle Zahl größer null, Porosität
κ, λ, ν	Index, laufende Nummer, Konstante, natürliche Zahl
λ	Parameter der Verteilung einer Grundgesamtheit
μ	Mittelwert einer Grundgesamtheit oder Verteilung
π	Kreiszahl 3,14159...
ξ, η	Merkmal, Zufallsvariable (wie x, y)
ρ	Korrelationskoeffizient einer Grundgesamtheit (8.42), Massendichte
σ^2	Varianz einer Grundgesamtheit oder Verteilung
σ_{xy}	Kovarianz (3.117)
$\Phi(x)$	Summenfunktion (3.31 und 3.34)
$\varphi(x)$	Dichtefunktion (3.35)
χ^2	Variable der Chi-Quadrat-Verteilung (5.59)

Kapitel 1

Grundbegriffe

Wir sehen uns einige Beispiele und allgemeine Begriffe an, um das Verständnis der nachfolgenden Kapitel zu erleichtern.

1.1 Kausalität und Zufall

Es gibt in der Natur Vorgänge, die vollkommen determiniert ablaufen und dem **Prinzip von Ursache und Wirkung** (Kausalität) uneingeschränkt gehorchen. Bei Kenntnis aller Ausgangsbedingungen lässt sich ihr Ablauf zuverlässig vorhersagen. Hierzu gehören die Bewegungen der Planeten im Sonnensystem, die Schwingungen eines einfachen Pendels, die Mischungstemperatur zweier Wasservolumina oder die Erwärmung eines von einem elektrischen Strom durchflossenen Drahtes.

Daneben gibt es aber auch Vorgänge, bei denen der **Zufall** eine Rolle spielt mit der Folge, dass sich ein eindeutiger Zusammenhang zwischen Ursachen und Wirkungen nicht herstellen lässt. Bei Wiederholung solcher zufälliger Vorgänge werden trotz gleicher Ausgangsbedingungen unterschiedliche Endzustände erreicht. Man spricht auch von stochastischen Vorgängen. Das aus dem Griechischen stammende Wort **Stochastik** bedeutet wörtlich die *Lehre von den Vermutungen*. Die heutige Stochastik ist ein Teilgebiet der Mathematik, das neben Statistik und Wahrscheinlichkeitstheorie auch die Kombinatorik umfasst.

Sehen wir uns einige **Beispiele für stochastische Vorgänge** an:
1. In der Medizin lässt sich die Frage, ob ein **Medikament** eine bestimmte Wirkung hat, nicht durch einen einzigen Versuch mit einer einzigen Testperson beantworten, weil Alter, Gesundheitszustand, Lebensgewohnheiten und Erbanlagen – aus unserer Sicht zufällige Einflüsse – das Ergebnis erheblich mitbestimmen. Wohlgemerkt: aus unserer Sicht zufällig. Ein Anderer, der mehr von den Zusammenhängen weiß oder einen größeren Versuchsaufwand treibt, wird vielleicht die Macht des Zufalls einengen und beispielsweise das Alter als mitbestimmende Ursache einbeziehen.
2. Jedes **Messverfahren** hat eine begrenzte Genauigkeit. Bei Wiederholungsmessungen erhält man deshalb notwendigerweise unterschiedliche Werte. Die durch

die Wiederholungen gewonnenen Informationen lassen sich zur Erhöhung der Genauigkeit und zur Beurteilung der Zuverlässigkeit der Messung heranziehen.
3. **Industriegüter** werden in großen Mengen hergestellt. Auch wenn ein Produktionsprozess weitgehend automatisiert ist, stimmen die Produkte niemals in ihren Eigenschaften vollkommen überein. Unter anderem stellt sich die Frage, wie groß die Abweichungen werden dürfen, ohne die Brauchbarkeit des Produktes zu gefährden.
4. Die **Laufzeit** von Datenpaketen in einem großen Computernetz wie dem Internet unterliegt zahlreichen Einflüssen, die nicht vorhersehbar sind. Deshalb schwankt die einfach zu ermittelnde Laufzeit der Datenpakete von einem Rechner A zu einem Rechner B und zurück (round-trip time, RTT) erheblich. Bei manchen Diensten stören nicht die Laufzeiten an sich, sondern ihre Schwankungen.
5. Beim **Würfelspiel** – mit unverfälschen (fairen) Würfeln – ist es auch einem geschickten Spieler nicht möglich, eine bestimmte Zahl zu würfeln. Deshalb ist ein Würfel ein beliebtes Gerät zum Erzeugen von Zufallszahlen. Der Computer als deterministischer Automat kann grundsätzlich keine Zufallszahlen erzeugen. Er kann jedoch Zahlenfolgen berechnen, die für viele Zwecke ausreichend zufällig wirken, und das rasend schnell.

Bei zufälligen Erscheinungen lassen sich der einzelne Vorgang und sein Ergebnis nicht festlegen. Man kann nicht voraussagen, ob ein gerade den Produktionsprozess durchlaufendes Werkstück den Qualitätsanforderungen genügen wird oder nicht. Einzelschicksale liegen außerhalb des Geltungsbereichs der Statistik. Für oftmals wiederholte Vorgänge lassen sich jedoch sehr wohl Gesetzmäßigkeiten angeben. Diese Gesetzmäßigkeiten aufzufinden und zu formulieren, ist Aufgabe der mathematischen Statistik. Die Gesetze der mathematischen Statistik erlauben es, aufgrund vergleichsweise weniger Beobachtungen oder Messungen an Stichproben Aussagen mit quantitativ bestimmbarer Zuverlässigkeit über die zugehörige Grundgesamtheit zu machen.

Hiermit werden zwei grundlegende Begriffe eingeführt. Eine **Stichprobe** (E: sample) ist eine Teilmenge der als **Grundgesamtheit** (Gesamtheit, E: population) bezeichneten Gesamtmenge, über die wir etwas wissen oder aussagen wollen. Zu Beginn einer statistischen Untersuchung muss feststehen beziehungsweise geklärt werden, woraus die Grundgesamtheit besteht. Der Inhalt oder Umfang der Teilmenge oder Stichprobe ist in der Regel viel kleiner als der Inhalt der Grundgesamtheit. Jede Stichprobe besteht ihrerseits – wie die Grundgesamtheit – aus einer mehr oder weniger großen Anzahl von **Elementen** (Beobachtungseinheit, Merkmalsträger, E: element). Eine Stichprobe, die nur aus einem Element bestünde, erlaubte keine statistischen Aussagen. Eine Stichprobe, die identisch mit ihrer Grundgesamtheit wäre, ließe sich zwar statistisch beschreiben, alle Überlegungen zum Zusammenhang zwischen Stichprobe und Grundgesamtheit entfielen jedoch.

Jedes Element der Stichprobe wird – wie im Folgenden gezeigt wird – durch einen bestimmten Wert eines **Merkmals** (E: attribute, property) gekennzeichnet. Im Falle nichttechnischer Anwendungen ist häufig nur zwischen zwei Merkmalswerten zu unterscheiden (gut–schlecht, schwarz–weiß, männlich–weiblich).

Grundgesamtheit und Stichprobe bestehen aus nur zwei Sorten unter sich gleicher Elemente. Zur Beschreibung kommt dann nur die Anzahl beziehungsweise der Anzahlanteil einer der beiden Komponenten in Betracht. Im Falle technischer oder naturwissenschaftlicher Anwendungen kann grundsätzlich jedes Element einer Stichprobe durch einen eigenen Merkmalswert gekennzeichnet werden. Als Merkmale kommen alle messbaren physikalischen Größen in Frage wie Masse, Länge, Fläche, Volumen, Geschwindigkeit, elektrische Ladung, Lichtstärke, pH-Wert usw. Die Merkmalswerte dürfen nur unter bestimmten Voraussetzungen als Ausprägungen einer **Zufallsvariablen** (E: random variable) aufgefasst werden; hierzu siehe Abschnitt 3.3.1 *Verteilungen einer Zufallsvariablen* auf Seite 32.

Um von einer Stichprobe auf die jeweilige Grundgesamtheit schließen zu können, muss das Beobachtungsmaterial eine Zufallsauswahl darstellen, das heißt, die Stichprobennahme ist so vorzunehmen, dass jedes Element der Grundgesamtheit grundsätzlich die gleiche Chance hat, in die Stichprobe zu gelangen. Wir verstehen unter *Stichprobe* stillschweigend immer eine zufällige Stichprobe. Andernfalls reden wir allgemein von Probe oder Teilmenge.

Das Wort *Stichprobe* rührt daher, dass früher mit einem rohrförmigen Probenstecher Proben aus Kaffeesäcken oder Käselaiben entnommen wurden. Man kann den Begriff *Stichprobe* im Sinne realer Materialmengen verstehen. Im mathematischen Sprachgebrauch bedeutet er jedoch die Menge der jeweils an den einzelnen Elementen der realen Materialmenge gemessenen Merkmalswerte.

In der mathematischen Statistik bezeichnet man einen stochastischen Vorgang gleich welcher Art als **Zufallsexperiment** (E: random experiment), auch wenn der Vorgang keineswegs experimentellen Charakter hat, falls

- der Vorgang so geartet ist, dass das Ergebnis nicht im Voraus zu bestimmen ist, und
- der Vorgang grundsätzlich beliebig oft wiederholbar ist und
- das Ergebnis einer Durchführung des Experimentes unabhängig von den Ergebnissen anderer Durchführungen des Experimentes ist.

Im Folgenden wird der Begriff *Stichprobe* nur dann verwendet, wenn die Stichprobenahme alle Kriterien eines Zufallsexperimentes erfüllt. Mit der Definition des Zufallsexperimentes wird zugleich klar und eindeutig festgelegt, wie der Begriff *Zufall* (E: random) zu verstehen ist, nämlich nicht etwa im Sinne von Beliebigkeit.

1.2 Merkmale und Messverfahren

Jedes Element der Stichprobe soll durch einen bestimmten Wert eines Merkmals gekennzeichnet werden. Hierzu muss für jede technische Aufgabe ein geeignetes Messverfahren bereit stehen. Dieses soll einerseits möglichst genau sein und andererseits die Verarbeitung einer großen Anzahl von Messwerten in kurzer Zeit ermöglichen.

Häufig ist der mögliche Wertebereich des Merkmals größer als der Messbereich des verwendeten Messgerätes (Beispiel: Partikelgrößenverteilungen). In solchen Fällen kann man sich damit helfen, dass man die durch eine Ausgleichskurve (Approximationsfunktion) vermittelten Messpunkte über den gemessenen Bereich hinaus extrapoliert. Damit können jedoch erhebliche Fehler verbunden sein. Sicherer ist es, in solchen Fällen den Gesamtbereich zu unterteilen und mehrere Messgeräte mit unterschiedlichem Messbereich einzusetzen, was den Aufwand leider in die Höhe treibt. Auch kann der Anschluss der jeweiligen Ergebnisse aneinander Schwierigkeiten bereiten, insbesondere dann, wenn Messgeräte verwendet werden, die nach unterschiedlichen physikalischen Prinzipien arbeiten.

Oftmals lässt man bei Rechnungen oder Darstellungen der Einfachheit halber den Wertebereich bei null beginnen oder gibt keine endliche obere Schranke an. Die Begriffe *null* und *unendlich* sind jedoch mit Vorsicht zu genießen. Wenn wir beispielsweise über einen Bereich des Merkmals x von null bis unendlich integrieren, ist damit gemeint, über alle vorkommenden Werte. Ist das Merkmal eine Partikellänge, so gibt es den Wert null nicht, denn selbst das kleinste Molekül hat eine Länge größer null. In manchen Zusammenhängen führt das Einbeziehen des Wertes null zu unsinnigen Ergebnissen. Ähnlich ist auch der Wert unendlich zu verstehen. In vielen Zusammenhängen bedeutet er nur größer als jeder vorkommende Wert.

1.3 Mengen

Zur Beschreibung einer Stichprobe sind außer den Merkmalswerten für die Stichprobenelemente auch Angaben zum **Mengeninhalt** oder **-umfang** (E: cardinality) erforderlich, in diesem Fall also zum Stichprobenumfang. Werden die Elemente gezählt – was sich nur bei endlichen Mengen durchführen lässt – so ist die **Anzahl** die gegebene Größe und unmittelbar zur Kennzeichnung des Mengeninhalts geeignet. Viele Messverfahren arbeiten auf diese Weise. In den üblichen Darstellungen der mathematischen Statistik wird davon ausgegangen, dass dies immer so ist. Im weiteren Sinn spricht man von der **Mächtigkeit** oder Kardinalität einer Menge, die auch für unendliche Mengen definiert ist.

Es gibt jedoch Messverfahren – beispielsweise in der Partikelgrößenanalyse – bei denen zur Kennzeichnung des Mengeninhalts andere Größen wie Masse oder Volumen verwendet werden. Auch ungemahlener Kaffee wird nicht nach der Anzahl der Bohnen gehandelt, sondern nach deren Masse. Gekochte Nordseegarnelen (Granat) werden in den Häfen literweise – nach Schüttvolumen – verkauft, nicht stückweise, ebenso Walnüsse auf dem Nussmarkt in Vianden. Auf die besonderen Fragestellungen, die sich aus der Verwendung verschiedener Größen zur Kennzeichnung des Mengeninhalts ergeben, wird in Abschnitt 9.8 *Umrechnung verschiedener Mengenarten* auf Seite 173 eingegangen.

Eine allgemeine Anforderung an Größen zur Kennzeichnung des Mengeninhalts ist ihre Additivität. Wenn wir zwei disjunkte Mengen vereinigen (Mengen, deren Schnittmenge leer ist; elementfremde Mengen), dann soll der Inhalt der ver-

einigten Menge gleich der Summe der Inhalte der beiden Teilmengen sein. Mengen können einen endlichen Inhalt haben – was bei Stichproben immer gegeben ist – oder einen unbegrenzten, unendlichen Inhalt, zumindest bei gedachten Objekten wie der Menge *aller* Würfe mit einem Würfel.

1.4 Modelle

Das Zurückführen einer gegebenen Situation oder Aufgabe auf das richtige statistische **Modell** ist ein Schritt, bei dem leicht Fehler gemacht werden. Hat man erst einmal das richtige Modell gefunden, kann der Rest zwar mühsam sein, aber die Schritte liegen fest.

Der **Münzwurf** ist das einfachste Zufallsexperiment. Sehen wir von der Möglichkeit ab, dass die Münze hochkant stehen bleibt, hat das Experiment zwei Ausgänge: Kopf (Zahl) oder Wappen. Deren Wahrscheinlichkeit ist nach allem, was wir wissen – und nicht etwa mathematisch herleiten können – gleich groß, das heißt je 0,5. Anders ausgedrückt ist der Ausgang eine diskrete gleichverteilte Zufallsvariable mit den sich gegenseitig ausschließenden Ausprägungen Kopf und Wappen. Die Anzahl der Münzwürfe in einer Versuchsreihe ist nur durch unsere Geduld begrenzt.

Beim Werfen eines idealen oder fairen **Würfels** – auch Laplace-Würfel genannt – hat jede der 6 Seiten die gleiche Chance, nach dem Ausrollen des Würfels oben zu liegen. Das ist unabhängig von der Zuordnung bestimmter Zahlen zu den Seiten des Würfels, weshalb es keine Rolle spielt, in welcher Reihenfolge die Seiten nummeriert sind. Üblich, aber für unsere Zwecke nicht notwendig ist, dass sich die Augenzahlen gegenüberliegender Seiten eines n-seitigen Würfels zu $n+1$ ergänzen. Beim gebräuchlichen Würfel lässt das immer noch zwei spiegelbildliche Anordnungen zu. Ein Münze kann als ein zweiseitiger Würfel aufgefasst werden.

Ein einzelner Wurf beziehungsweise sein Ergebnis stellt ein **Elementarereignis** dar, das nicht weiter unterteilt werden kann. Je nach Definition des Zufallsexperimentes lassen sich jedoch die Ergebnisse mehrerer Würfe zu einem zusammengesetzten Ereignis verbinden. Setzen wir voraus, dass die Reihenfolge der Einzelergebnisse keine Rolle spielt, ist es unerheblich, ob wir zehnmal mit einem Würfel, fünfmal mit zwei Würfeln oder einmal mit zehn Würfeln werfen. Von dieser Überlegung machen wir auf Seite 109 Gebrauch.

Simuliert man das Würfeln auf dem Computer[1], lassen sich große Anzahlen von Würfen verwirklichen und auswerten, ebenso auch Würfel mit beliebigen Seitenzahlen. Dabei wird von Pseudo-Zufallszahlen Gebrauch gemacht. Hierzu siehe Abschnitt 10.5 *Zufallszahlen* auf Seite 205.

Das **Urnenmodell** beschreibt ein Gedankenexperiment, bei dem in ein Gefäß (Urne) eine bestimmte Anzahl von Kugeln gegeben werden, die sich in einer

[1] Auf Rechnern unter einem Debian-System steht das Programm `rolldice` zur Verfügung. Es stellt eine nahezu beliebige Anzahl von Würfeln mit einer nahezu beliebigen Anzahl von Seiten bereit. Einen weiteren Würfel bietet `http://www.random.org/dice/` an.

Eigenschaft (Farbe, laufende Nummer oder dergleichen) unterscheiden. Aus der Urne werden nacheinander einige Kugeln entnommen, welche die Stichprobe bilden. Jede Kugel in der Urne muss dieselbe Chance haben, beim nächsten Zug entnommen zu werden. Deshalb müssen die Kugeln in Eigenschaften, welche die Entnahme beeinflussen (Durchmesser, Masse und dergleichen), übereinstimmen. Wir haben dann zu unterscheiden zwischen

- Ziehen mit oder ohne Zurücklegen,
- Ziehen mit oder ohne Beachtung der Reihenfolge,

was vier Fälle ergibt. Eine Lottoziehung beispielsweise entspräche einem Ziehen ohne Zurücklegen und ohne Beachtung der Reihenfolge. Will man eine zufällige Zeichenkette – ein Passwort etwa – erzeugen, beschriftet man die Kugeln mit den Zeichen des vorgesehenen Zeichensatzes, zieht, legt zurück, mischt erneut und beachtet die Reihenfolge.

Ein **Kartenspiel** besteht aus 52 Spielkarten, eingeteilt in die Farben Kreuz oder Treff (schwarz), Pik (schwarz oder grün), Herz (rot) und Karo (rot oder gelb). Daneben werden die Karten nach ihrem Wert von 1 (Ass) bis 10, Bube, Dame und König unterschieden. Joker sind in der Statistik nicht vorgesehen; ebensowenig berücksichtigen wir die zahlreichen Varianten. Ähnlich wie beim Urnenmodell lässt sich aus den 52 gut gemischten Karten eine Stichprobe ziehen, bei der jede Karte die gleiche Chance hat, in die Stichprobe zu gelangen. Anders als beim Urnenmodell werden die Karten nach dem Ziehen üblicherweise nicht zurückgelegt.

Letzten Endes stellt jede Beschreibung eines Zustands oder Vorgangs aus dem realen Leben mit mathematischen Mitteln eine Modellierung dar, die von der Wahrheit auf Grund von Vereinfachungen und/oder Verallgemeinerungen abweicht, manchmal auch auf Grund von falschen Beobachtungen oder Annahmen. Die oben genannten Modelle sind nur besonders gut bekannt und leicht nachvollziehbar.

Kapitel 2

Häufigkeit (Beschreibung von Stichproben)

Mit Messen, Zählen und Sortieren beginnt die Statistik. Das sind einfache Tätigkeiten, die jeder kennt. Wir schauen genauer hin.

2.1 Urliste, absolute Häufigkeit

Gegeben sei eine Stichprobe, die aus z Elementen besteht. Die Anzahl z wird **Inhalt** oder Umfang der Stichprobe genannt. An jedem Element werde der Wert x_i des interessierenden Merkmals x (beispielsweise Länge oder Masse oder Farbe) gemessen. Die nacheinander gemessenen Werte x_i bilden die unsortierte **Urliste** (Rohdaten, Primärdaten, E: raw data).

Im weiteren Verlauf beschäftigen uns nur noch die Werte, nicht mehr die Elemente, aber wir müssen die Art der Elemente und auch die Messmethode kennen, um die Zahlen später richtig zu deuten.

Die unsortierte Urliste enthält alle Informationen aus der Beobachtung oder Messung. Die nächsten Schritte dienen dazu, die für unsere Fragestellung wesentlichen Informationen herauszuschälen und überzeugend darzustellen. Neue Informationen kommen nicht hinzu, es sei denn, wir verfügten über zusätzliche Kenntnisse oder träfen Annahmen, die wir in die Auswertung einfließen lassen. Das Gebiet der Statistik, das sich mit der Beschreibung vollständig bekannter Mengen (= Stichproben) befasst, wird deskriptive oder beschreibende Statistik genannt. Von Wahrscheinlichkeiten ist noch nicht die Rede.

Im ersten Schritt sortieren wir die Urliste nach der Größe der Merkmalswerte x_i und verlieren damit die Information über die Reihenfolge der Ereignisse. Aus der sortierten Urliste lassen sich ohne viel Rechenarbeit einige statistische Größen ablesen. Der **Medianwert** m (E: median) oder Zentralwert einer Stichprobe ist bei ungerader Anzahl z der Werte der mittlere der nach ihrer Größe sortierten Werte, bei gerader Anzahl das arithmetische Mittel der beiden der Mitte benachbarten Werte. Der Medianwert kennzeichnet die Lage der Werte auf der Merkmalsskala und gehört zu den **Lageparametern** (E: positional parameter, location, central tendency).

Die **Spannweite** R (E: range) ist die Differenz von größtem und kleinstem Merkmalswert in der Stichprobe

$$R = x_{max} - x_{min} \tag{2.1}$$

Die Spannweite hat den Nachteil, dass sie sich auf zwei Werte stützt, die nur durch wenige Messungen belegt und daher unsicher sind. Ausreißer an den Enden der Verteilung gehen voll in die Spannweite ein. Die Spannweite kennzeichnet die Streuung der Werte und zählt zu den **Streuungsparametern** (E: variational parameter, dispersion, spread). Lage- und Streuungsparameter gehören zu den statistischen **Kennzahlen** (E: statististics), die weitere Größen wie die Schiefe oder die Kovarianz umfassen.

Unter einem **Quartil** (E: quartile) versteht man jeweils ein Viertel der sortierten Urliste. Der **Interquartilbereich** I_{50} (E: interquartile range) schließt die beiden mittleren Quartile und damit die Hälfte der Werte um den Medianwert ein, lässt also die äußeren Werte samt etwaigen Ausreißern unberücksichtigt. Mit Medianwert und Spannweite oder besser Interquartilbereich können wir bereits eine Stichprobe nach Lage und Breite grob beschreiben. In gleicher Weise wie Quartile werden auch **Dezile** und der **Interdezilbereich** I_{80} definiert. Der Oberbegriff für derartige Ausschnitte aus der gesamten Verteilung lautet **Quantil** (E: quantile).

Verallgemeinern wir unsere Betrachtungsweise und schließen beispielsweise auch das Werfen eines Würfels oder das Befragen von Personen ein – was man nicht als Messen bezeichnen wird – so ist jedes einzelne Messergebnis, jede Augenzahl oder jede Antwort ein statistisches **Ereignis** (E: event). Gleiche Ereignisse werden im Verlauf der Auswertung zusammengefasst, sie verlieren ihre Individualität. Wir können also bei einer Stichprobe sagen, dass in $z(\mathrm{A})$ Fällen das Ereignis A eingetreten ist:

- beim Messen der Länge gleichartiger Schrauben ist $z(\mathrm{A})$-mal das Ereignis A (Länge zwischen 50,0 und 50,1 mm) eingetreten,
- beim wiederholten Werfen eines Würfels ist $z(\mathrm{A})$-mal das Ereignis A (Augenzahl gleich 4) eingetreten,
- beim Befragen von Personen nach ihrem Befinden ist $z(\mathrm{A})$-mal das Ereignis A (Antwort: gut) eingetreten.

Umgangssprachlich würde man eher den Vorgang des Messens, Würfelns oder Befragens als Ereignis bezeichnen, der Statistiker sieht nur das Ergebnis.

2.2 Relative Häufigkeit bei einer Eigenschaft

Gegeben sei eine Stichprobe, die aus z Elementen besteht. Die Anzahl z wird – wie erwähnt – **Inhalt** oder Umfang der Stichprobe genannt. Die Untersuchung der Stichprobe möge ergeben, dass in $z(\mathrm{A})$ Fällen ein Ereignis A eingetreten ist. Dann bezeichnet man mit

$$h(\mathrm{A}) = \frac{z(\mathrm{A})}{z} \tag{2.2}$$

die relative Häufigkeit (E: relative frequency) des Ereignisses A in der Stichprobe. Die relative Häufigkeit ist stets dimensionslos, unabhängig davon, welche Dimension z möglicherweise hat.

Beispiel 2.1 : Ein Medikament sei an 88 Patienten getestet worden. In $z(\mathrm{A}) = 66$ Fällen sei eine Besserung des Befindens (Ereignis A) eingetreten. Dann ist die relative Häufigkeit des Ereignisses A

$$h(\mathrm{A}) = \frac{66}{88} = \frac{3}{4} = 0,75$$

Ende Beispiel

Im Folgenden werden Ereignisse mit großen, senkrechten Buchstaben wie $\mathrm{A}, \mathrm{B}, \mathrm{C}$ oder A_1, A_2, A_3 usw. bezeichnet. Für die relative Häufigkeit gilt stets

$$0 \leq h(\mathrm{A}) \leq 1 \tag{2.3}$$

das heißt, die relative Häufigkeit ist eine reelle nicht-negative Zahl, die höchstens gleich eins ist. Bezeichnet man mit $\bar{\mathrm{A}}$ (lies: nicht-A) das Ereignis, das genau dann eintrifft, wenn A nicht eintrifft (keine dritte Möglichkeit zugelassen), so folgt aus

$$z(\bar{\mathrm{A}}) = z - z(\mathrm{A})$$

durch Dividieren mit z

$$h(\bar{\mathrm{A}}) = 1 - h(\mathrm{A}) \tag{2.4}$$

Eine andere gängige Schreibweise für nicht-A ist $\neg\mathrm{A}$

Dass wir neben A und $\bar{\mathrm{A}}$ keine dritte Möglichkeit zulassen, ist einleuchtend und ein Grundprinzip der klassischen Logik (Satz vom ausgeschlossenen Dritten).

2.3 Relative Häufigkeit bei zwei Eigenschaften

Gegeben sei eine Stichprobe aus z Elementen. Die Untersuchung zeige, dass in $z(\mathrm{A})$ Fällen das Ereignis A und in $z(\mathrm{B})$ Fällen das Ereignis B eingetreten ist. Die Ereignisse A und B mögen sich nicht wechselseitig ausschließen.

Beispiel 2.2 : Die Stichprobe bestehe aus einer Gruppe von z Personen. Das Ereignis A sei definiert als *Alter über 50 Jahre*, das Ereignis B als *Körpergröße über 170 cm*. Die Ereignisse A und B betreffen zwei Eigenschaften, die sich offenbar nicht wechselseitig ausschließen.

Ende Beispiel

Hinter dem Zahlenmaterial obiger Stichprobe könnten selbstverständlich auch Stichproben aus anderen Gebieten stehen:

- Die Karten eines Kartenspiels, die einerseits in Farben (Kreuz, Pik, Herz, Karo), andererseits und unabhängig von den Farben in Kartenwerte (2 ... Ass) eingeteilt sind.

- Das Wetter gestern und heute, wenn wir nur die Merkmalswerte sonnig und regnerisch zulassen.
- Die Mitglieder eines Sportvereins, der aus zwei Abteilungen (Fußball, Leichtathletik) besteht und zwei Mitgliederkategorien (Jugendliche, Erwachsene) unterscheidet.
- Eine Gruppe von Erwachsenen bestehe aus Männern und Frauen, die Mitglied einer politischen Partei sind oder nicht.
- Die Flut der elektronischen Mails, die bei einem Empfänger eintreffen, bestehe aus deutsch- oder englischsprachigen Texten (andere Möglichkeiten bleiben außer Betracht). Ferner lasse sie sich in erwünschte und unerwünschte Mails (spam) einteilen.
- Ein Sortiment von Jeans, gekennzeichnet durch Länge und Bundweite.

Weiterhin wird mit A + B (A oder B; A vereinigt mit B) das Ereignis bezeichnet, das genau dann eintrifft, wenn A oder B oder beide zugleich eintreffen (inklusives Oder, nicht-ausschließende Disjunktion, Vereinigungsmenge). Das Ereignis A + B wird auch die **Summe** (E: union) der Ereignisse A und B genannt.

Mit AB (A und B; A geschnitten mit B) wird ferner das Ereignis bezeichnet, das genau dann eintrifft, wenn A und B zugleich eintreffen (Und, Konjunktion, Schnittmenge, Durchschnittsmenge). Das Ereignis AB heißt auch **Produkt** (E: intersection) der Ereignisse A und B.

Dann lassen sich die z Elemente der Stichprobe in vier Gruppen einteilen:

1. A und B zugleich eingetroffen, Anzahl der Elemente $z(\mathrm{AB})$,
2. A eingetroffen, B nicht, Anzahl der Elemente $z(\mathrm{A}\bar{\mathrm{B}})$,
3. A nicht eingetroffen, B eingetroffen, Anzahl der Elemente $z(\bar{\mathrm{A}}\mathrm{B})$,
4. weder A noch B eingetroffen, Anzahl der Elemente $z(\bar{\mathrm{A}}\bar{\mathrm{B}})$.

Ein Element der Stichprobe kann jeweils nur einer der vier Gruppen angehören, das heißt, die Ereignisse AB, $\mathrm{A}\bar{\mathrm{B}}$, $\bar{\mathrm{A}}\mathrm{B}$ und $\bar{\mathrm{A}}\bar{\mathrm{B}}$ schließen sich wechselseitig aus. Für die Gesamtzahl z der Elemente in der Stichprobe gilt daher

$$z(\mathrm{AB}) + z(\mathrm{A}\bar{\mathrm{B}}) + z(\bar{\mathrm{A}}\mathrm{B}) + z(\bar{\mathrm{A}}\bar{\mathrm{B}}) = z$$

Für die relativen Häufigkeiten folgt

$$h(\mathrm{AB}) + h(\mathrm{A}\bar{\mathrm{B}}) + h(\bar{\mathrm{A}}\mathrm{B}) + h(\bar{\mathrm{A}}\bar{\mathrm{B}}) = 1 \tag{2.5}$$

Das Ereignis A tritt nur in den beiden Fällen AB und $\mathrm{A}\bar{\mathrm{B}}$ ein; es gilt

$$z(\mathrm{A}) = z(\mathrm{AB}) + z(\mathrm{A}\bar{\mathrm{B}})$$

woraus folgt

$$h(\mathrm{A}) = h(\mathrm{AB}) + h(\mathrm{A}\bar{\mathrm{B}}) \tag{2.6}$$

und entsprechend

$$h(\mathrm{B}) = h(\mathrm{AB}) + h(\bar{\mathrm{A}}\mathrm{B}) \tag{2.7}$$

Für die Summe der Ereignisse A und B erhält man gemäß der Definition von A + B

$$z(\mathrm{A}+\mathrm{B}) = z(\mathrm{A}\bar{\mathrm{B}}) + z(\bar{\mathrm{A}}\mathrm{B}) + z(\mathrm{AB})$$

und nach Division mit z für die relative Häufigkeit

$$h(\mathrm{A}+\mathrm{B}) = h(\mathrm{A}\bar{\mathrm{B}}) + h(\bar{\mathrm{A}}\mathrm{B}) + h(\mathrm{AB})$$

Mit 2.6 und 2.7 folgt daraus der **Additionssatz für relative Häufigkeiten**

$$h(\mathrm{A}+\mathrm{B}) = h(\mathrm{A}) + h(\mathrm{B}) - h(\mathrm{AB}) \tag{2.8}$$

Das ist auch anschaulich klar: Ohne den Subtrahenden $h(\mathrm{AB})$ hätten wir das Ereignis AB (A und B zugleich eingetroffen) doppelt berücksichtigt, nämlich einmal bei der Gruppe A und zum zweiten Mal bei der Gruppe B. Im Sonderfall $h(\mathrm{AB}) = 0$, wenn sich also die Ereignisse A und B wechselseitig ausschließen, vereinfacht sich 2.8 zu

$$h(\mathrm{A}+\mathrm{B}) = h(\mathrm{A}) + h(\mathrm{B}) \tag{2.9}$$

Alle bisher behandelten relativen Häufigkeiten beziehen sich auf die Gesamtzahl z der Elemente in der Stichprobe. Bezieht man die Häufigkeiten nicht auf die Gesamtzahl, sondern auf eine Teilmenge der Stichprobe, so erhält man **bedingte relative Häufigkeiten** (E: conditional relative frequency). Mit $h(\mathrm{A}|\mathrm{B})$ wird die relative Häufigkeit von A in der Teilmenge der Stichprobe bezeichnet, auf deren Elemente B zutrifft. Man greift also aus der Stichprobe zuerst die Elemente heraus, auf die B zutrifft, und untersucht in einem zweiten Schritt das Auftreten von A in dieser Teilmenge, kurz gesprochen *A vorausgesetzt B*. Dann gilt

$$h(\mathrm{A}|\mathrm{B}) = \frac{z(\mathrm{AB})}{z(\mathrm{B})}$$

Dividiert man Zähler und Nenner mit z, so folgt

$$h(\mathrm{A}|\mathrm{B}) = \frac{h(\mathrm{AB})}{h(\mathrm{B})} \tag{2.10}$$

Entsprechend gilt für die relative Häufigkeit des Ereignisses B unter den Elementen der Stichprobe, auf die A zutrifft

$$h(\mathrm{B}|\mathrm{A}) = \frac{h(\mathrm{AB})}{h(\mathrm{A})} \tag{2.11}$$

Aus 2.10 und 2.11 folgt der **Produktsatz für relative Häufigkeiten**

$$h(\mathrm{AB}) = h(\mathrm{A}) \cdot h(\mathrm{B}|\mathrm{A}) = h(\mathrm{B}) \cdot h(\mathrm{A}|\mathrm{B}) \tag{2.12}$$

Der Additionssatz und der Produktsatz für relative Häufigkeiten lassen sich von zwei Ereignissen A und B auf beliebig viele Ereignisse erweitern.

Beispiel 2.3 : Gegeben seien folgende Beobachtungsergebnisse über die Augenfarben von Vätern und Söhnen:

	Vater hell (A)	Vater dunkel ($\bar{\text{A}}$)	Summe
Sohn hell (B)	471	148	619
Sohn dunkel ($\bar{\text{B}}$)	151	230	381
Summe:	622	378	1000

Die Elemente der Stichprobe sind Vater-Sohn-Paare, genauer ihre Augenfarben. Wie groß sind die relativen Häufigkeiten $h(\text{A})$, $h(\text{B})$, $h(\text{AB})$, $h(\text{A+B})$, $h(\text{B|A})$ und $h(\text{A|B})$, wenn wir die Ereignisse A und B gemäß obiger Tabelle definieren? In der Tabelle ist die Anzahl der Fälle verzeichnet, in denen die Ereignisse AB, $\text{A}\bar{\text{B}}$, $\bar{\text{A}}\text{B}$ und $\bar{\text{A}}\bar{\text{B}}$ aufgetreten sind. Wir lesen daraus $z(\text{AB}) = 471$ usw. ab. Die Gesamtzahl der Beobachtungen ist:

$$z = z(\text{AB}) + z(\text{A}\bar{\text{B}}) + z(\bar{\text{A}}\text{B}) + z(\bar{\text{A}}\bar{\text{B}}) = 1000$$

Ferner gilt

$$z(\text{A}) = z(\text{AB}) + z(\text{A}\bar{\text{B}}) = 471 + 151 = 622$$

$$z(\text{B}) = z(\text{AB}) + z(\bar{\text{A}}\text{B}) = 471 + 148 = 619$$

$$z(\text{A} + \text{B}) = z(\text{AB}) + z(\text{A}\bar{\text{B}}) + z(\bar{\text{A}}\text{B}) = 471 + 151 + 148 = 770$$

Damit lassen sich die gesuchten Häufigkeiten berechnen

$$h(\text{A}) = \frac{z(\text{A})}{z} = \frac{622}{1000} = 0,622$$

$$h(\text{B}) = \frac{z(\text{B})}{z} = \frac{619}{1000} = 0,619$$

$$h(\text{AB}) = \frac{z(\text{AB})}{z} = \frac{471}{1000} = 0,471$$

$$h(\text{A} + \text{B}) = \frac{z(\text{A} + \text{B})}{z} = \frac{770}{1000} = 0,770$$

$$h(\text{A|B}) = \frac{z(\text{AB})}{z(\text{B})} = \frac{471}{619} = 0,761$$

$$h(\text{B|A}) = \frac{z(\text{AB})}{z(\text{A})} = \frac{471}{622} = 0,757$$

Wir machen die Probe

$$h(\mathrm{AB}) = h(\mathrm{A}) \cdot h(\mathrm{B}|\mathrm{A}) = 0,622 \cdot 0,757 = 0,471$$

$$h(\mathrm{AB}) = h(\mathrm{B}) \cdot h(\mathrm{A}|\mathrm{B}) = 0,619 \cdot 0,761 = 0,471$$

$$h(\mathrm{A} + \mathrm{B}) = h(\mathrm{A}) + h(\mathrm{B}) - h(\mathrm{AB}) = 0,622 + 0,619 - 0,471 = 0,770$$

Das Beispiel ist eine typische Anwendung für eine Tabellenkalkulation wie MS Excel oder GNOME Gnumeric, siehe Abschnitt 10.2 *Tabellenkalkulation* auf Seite 194.

Ende Beispiel

2.4 Darstellung relativer Häufigkeiten

2.4.1 Häufigkeitsverteilung eines diskreten Merkmals

Gegeben sei eine Stichprobe, die aus z Elementen besteht. Die Untersuchung der Stichprobe zeige, dass die sich wechselseitig ausschließenden Ereignisse A_1, A_2 $\ldots \mathrm{A}_n$ mit den relativen Häufigkeiten $h(\mathrm{A}_1)$, $h(\mathrm{A}_1)$ $\ldots h(\mathrm{A}_n)$ eingetreten sind. Unter den z Elementen sei keines, das nicht als Realisation eines der Ereignisse A_1, A_2 $\ldots \mathrm{A}_n$ aufgefasst werden kann. Dann gilt

$$\sum_{\nu=1}^{n} h(\mathrm{A}_\nu) = h(\mathrm{A}_1) + h(\mathrm{A}_2) + \ldots + h(\mathrm{A}_n) = 1 \tag{2.13}$$

Dem Ereignis A_i ist oft ein bestimmter Zahlenwert x_i eines Merkmals x zugeordnet.

Beispiel 2.4 : Wir würfeln mit einem herkömmlichen Würfel. Das Merkmal ist die bei einem Wurf erzielte Augenzahl und kann sechs verschiedene diskrete Zahlenwerte annehmen. Man setzt

$$h(\mathrm{A}_i) \equiv h(x_i) \tag{2.14}$$

und kennzeichnet die Ereignisse A_i durch die Augenzahlen x_i.

Ende Beispiel

Die Bezeichnung der Ereignisse A_i, das heißt die Festlegung der Indizes i, soll stets so erfolgen, dass $x_{i+1} > x_i$ ist. Für die relativen Häufigkeiten gilt

$$h(x_i) = \frac{z(x_i)}{z} \tag{2.15}$$

worin $z(x_i)$ die Anzahl der Fälle bezeichnet, in denen das Merkmal x den Zahlenwert x_i angenommen hat. Aus 2.13 folgt mit 2.14:

$$\sum_{\nu=1}^{n} h(x_\nu) = 1 \tag{2.16}$$

Bei der Ermittlung und Darstellung von relativen Häufigkeiten geht man von der Urliste der Messwerte aus, in der die Beobachtungen in ihrer zeitlichen Reihenfolge (chronologisch, unsortiert) eingetragen worden sind.

Der am häufigsten vorkommende Wert des Merkmals wird als **Modalwert** oder Modus (E: mode) bezeichnet. Er stellt den dem Maximum der relativen Häufigkeit zugeordneten Merkmalswert dar und ist ein Lageparameter neben Median und Mittelwert. Hat die relative Häufigkeit nur ein einziges Maximum, spricht man von einem **unimodalen** oder eingipfligen Verlauf der Häufigkeit. Treten mehrere lokale Maxima auf, ist eines davon zugleich das globale Maximum; ein solcher Verlauf wird **multimodal** oder mehrgipflig genannt. Multimodale Häufigkeiten sind oft bei Mischungen zu finden. Bei Gleichverteilungen verliert der Begriff seinen Sinn.

Beispiel 2.5 : Bei 20 Würfen mit einem Würfel (eine Stichprobe aus der unendlichen Grundgesamtheit aller Würfe) sind folgende Augenzahlen eingetroffen (Urliste):

6	3	3	1	6
2	6	5	2	6
4	4	6	3	4
4	5	2	5	3

Der nächste Schritt ist ein Sortieren der Werte und Zusammenfassen gleicher Werte, gegebenenfalls mit Hilfe einer Strichliste. Da es sich bei der Augenzahl um ein diskretes Merkmal handelt, kommen genau gleiche Werte vor. Die Strichliste liefert zu jedem Zahlenwert x_i des Merkmals x (Augenzahl) die Anzahl $z(x_i)$ der Fälle, in denen der zugehörige Zahlenwert aufgetreten ist. Indem man die Anzahlen $z(x_i)$ auf den Stichprobeninhalt z bezieht, erhält man gemäß 2.15 die relativen Häufigkeiten $h(x_i)$. Aus obiger Urliste folgt die Tabelle:

x_i	1	2	3	4	5	6
$z(x_i)$	1	3	4	4	3	5
$h(x_i)$	0,05	0,15	0,20	0,20	0,15	0,25

Durch das Sortieren und Zusammenfassen der Werte verlieren wir Informationen über ihre zeitliche Reihenfolge. Beschränken wir uns auf die Angabe der relativen Häufigkeiten, geht auch die Information über den Probeninhalt verloren.

Das Ergebnis $h(x_i)$ (Häufigkeitsverteilung der Stichprobe) trägt man üblicherweise als **Stabdiagramm** (E: bar chart, bar diagram) auf, siehe Abbildung 2.1. Das Stabdiagramm enthält – von Zeichen- und Ableseungenauigkeiten abgesehen

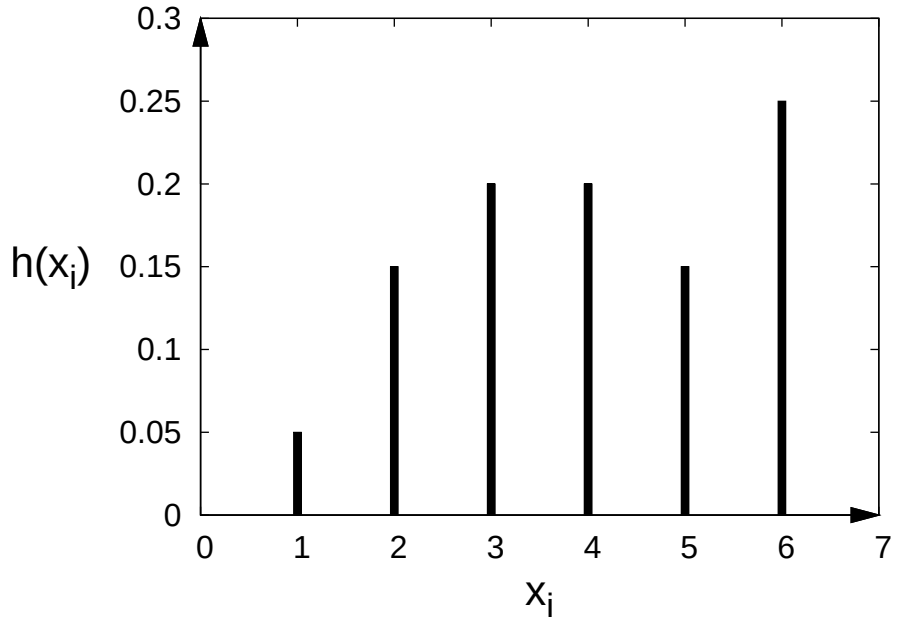


Abb. 2.1: Stabdiagramm einer Häufigkeitsverteilung der Augenzahlen bei Würfen mit einem Würfel

– dieselbe Information wie die Tabelle. Es kommt unserer Anschauung entgegen. Computer bevorzugen Zahlen.

Ende Beispiel

Man kann nun zu den relativen Häufigkeiten $h(x_i)$ einer Stichprobe die **relativen Häufigkeitssummen** (kumulative Häufigkeit, Summenfunktion der relativen Häufigkeit, E: cumulative frequency) bestimmen. Diese sind definiert durch

$$H(x_i) = h(x_1) + h(x_2) + \ldots + h(x_i) = \sum_{\nu=1}^{i} h(x_\nu) \tag{2.17}$$

Die Häufigkeitssummen sind gleich der Summe der relativen Häufigkeiten aller Stichprobenwerte, die kleiner oder gleich x_i sind. Trägt man $H(x_i)$ in einem Diagramm auf, so erhält man stets eine ansteigende Treppenkurve. Die zu obigem Stabdiagramm gehörige Summenfunktion ergibt gemäß 2.17 berechnet die Treppenkurve von Abbildung 2.2.

Man könnte auf den Gedanken verfallen, die Treppenkurve der Summenfunktion durch einen stetigen Kurvenzug auszugleichen. Das wäre aber eine Verfälschung, da das Merkmal x, die Augenzahl, nur diskrete Werte annimmt.

Zu jeder Stichprobe lassen sich zwei kennzeichnende Werte berechnen, ein **Mittelwert** (E: mean, mean value) und eine **Varianz**[1] (E: variance). Der Mittelwert einer Stichprobe ist definiert als das arithmetische Mittel aller Stichpro-

[1]Die Varianz wird in der Literatur teilweise als *Dispersion* bezeichnet. Da wir in der Verfahrenstechnik unter einer Dispersion ein Stoffsystem verstehen, vermeiden wir das Synonym.

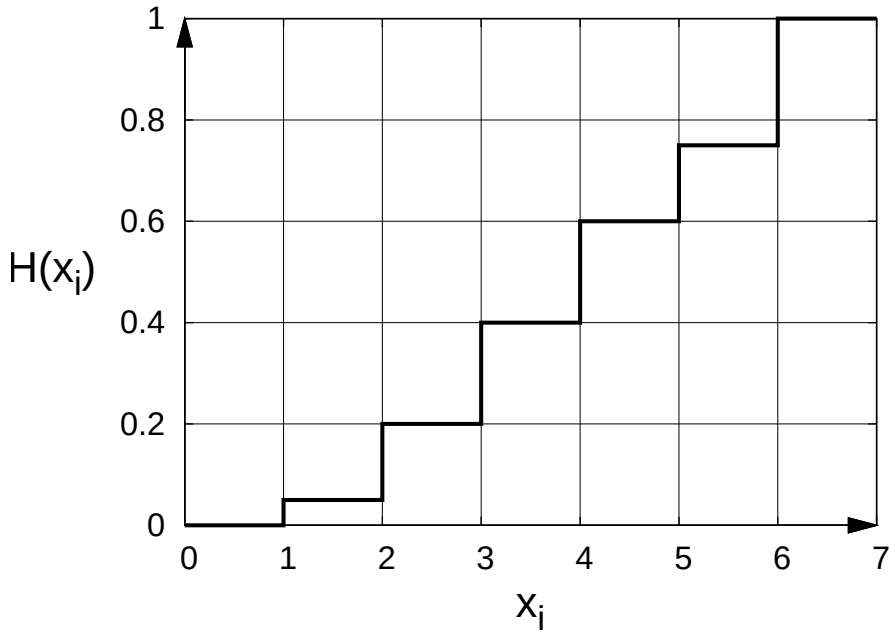


Abb. 2.2: Summenfunktion der relativen Häufigkeit zu obigem Stabdiagramm, als Treppe dargestellt

benwerte. Im Falle eines diskreten Merkmals x, das n verschiedene Werte x_i annehmen kann, berechnet sich der Mittelwert $\bar{x}$ zu

$$\bar{x} = x_1 h(x_1) + x_2 h(x_2) + \ldots + x_n h(x_n) = \sum_{\nu=1}^{n} x_\nu h(x_\nu) \tag{2.18}$$

Er kennzeichnet die Lage der Stichprobe beziehungsweise ihrer Werte auf der Merkmalsskala. Die Varianz s^2 einer Stichprobe ist ein Maß für die Streuung der Stichprobenwerte x_i um den Mittelwert $\bar{x}$. Sie wird üblicherweise durch die Gleichung

$$s^2 = \frac{z}{z-1} \sum_{\nu=1}^{n} (x_\nu - \bar{x})^2 \, h(x_\nu) \tag{2.19}$$

definiert, in der z die Anzahl der Elemente in der Stichprobe (Stichprobeninhalt oder -umfang) bezeichnet und $\bar{x}$ den gemäß 2.18 berechneten Mittelwert. Die positive Quadratwurzel aus der Varianz – also s – wird als **Standardabweichung** (E: standard deviation) einer Stichprobe bezeichnet.

Ist der Mittelwert μ_x der Grundgesamtheit bekannt, so kann man auch eine andere Definition der Varianz einer Stichprobe verwenden. Sie lautet

$$\tilde{s}^2 = \sum_{\nu=1}^{n} (x_\nu - \mu_x)^2 \, h(x_\nu) \tag{2.20}$$

Die beiden durch 2.19 und 2.20 definierten Stichprobenvarianzen unterscheiden sich. Das kann an obigem Beispiel gezeigt werden. Mit den Werten aus der Urliste auf Seite 14 erhält man aus 2.18 mit $z = 20$ und $n = 6$

$$\bar{x} = \sum_{\nu=1}^{n} x_\nu h(x_\nu) = 4,00$$

und aus 2.19

$$s^2 = \frac{z}{z-1} \sum_{\nu=1}^{n} (x_\nu - \bar{x})^2 \, h(x_\nu) = 2,5263$$

Wie später gezeigt wird, ist der Mittelwert μ_x der Grundgesamtheit beim unverfälschten Würfel $\mu_x = 3,5$. Hiermit berechnet man aus 2.20

$$\tilde{s}^2 = \sum_{\nu=1}^{n} (x_\nu - \mu_x)^2 \, h(x_\nu) = 2,6500$$

Die auf den Mittelwert bezogene Standardabweichung wird als **Variationskoeffizient** oder relative Standardabweichung bezeichnet. Er erlaubt besser zu beurteilen, ob eine Standardabweichung als groß oder klein anzusehen ist.

Die aus Stichprobenwerten berechneten Mittelwerte, Varianzen usw. werden auch als **empirische** Mittelwerte, Varianzen usw. bezeichnet, da sie sich auf empirisch, das heißt durch Messen, Beobachten, Befragen und dergleichen, gewonnene Werte stützen. Von den empirischen Größen zu unterscheiden sind die Größen einer Grundgesamtheit oder Verteilung, die sich entweder aus theoretischen Überlegungen ergeben oder nur mit gewissen Unsicherheiten abschätzen lassen.

Oft möchte man eine Verteilung durch möglichst wenige Parameter oder Maßzahlen beschreiben; das erleichtert Vergleiche. Allgemein werden Mittelwerte und ähnliche Größen als **Lageparameter** bezeichnet, Varianz, Standardabweichung und ähnliche Größen als **Streuungsparameter**. Eine Beschränkung auf Mittelwert und Varianz bedeutet einen weiteren Verlust an Information. Wir können nicht die vorhergehenden Werte von der Urliste bis zur relativen Häufigkeit aus Mittelwert und Varianz rekonstruieren.

2.4.2 Häufigkeitsverteilung eines stetigen Merkmals

Bisher haben wir Merkmale betrachtet, die nur bestimmte feste (= diskrete) Werte annehmen können wie die Augenzahl beim Würfeln. Kann ein Merkmal x auch beliebige Zwischenwerte annehmen, so spricht man von einem stetigen (kontinuierlichen) Merkmal.

In einer Stichprobe von z Elementen wird man in diesem Fall niemals zwei Elemente mit exakt gleichem Wert x_i des Merkmals finden. Den z Elementen der Stichprobe entsprechen folglich z verschiedene Ereignisse A_i, und die relative Häufigkeit dieser z Ereignisse A_i ist einheitlich

$$h(\mathrm{A}_i) = \frac{1}{z} \tag{2.21}$$

Für ihre Summe gilt

$$h(\mathrm{A}_1) + h(\mathrm{A}_2) + \ldots + h(\mathrm{A}_z) = \sum_{\nu=1}^{z} h(\mathrm{A}_\nu) = 1 \tag{2.22}$$

Ein Beispiel ist die Höhe von Bäumen. Das Merkmal *Baumhöhe* ist stetig. In einer Stichprobe von z Bäumen wird bei hinreichender Messgenauigkeit kein Baum genau so hoch sein wie der andere.

Da die Gleichungen 2.21 und 2.22 für jede Stichprobe das gleiche Ergebnis liefern, lassen sich unterschiedliche Merkmalsverteilungen hiermit offenbar nicht kennzeichnen. Es wird vielmehr deutlich, dass es keinen Sinn hat, den Begriff der relativen Häufigkeit von diskreten unverändert auf stetige Merkmale auszudehnen.

Um zu einer brauchbaren Darstellung der Häufigkeitsverteilung eines stetigen Merkmals zu gelangen, teilt man den Wertebereich des Merkmals in **Klassen** (E: class, bin) ein. Man geht dabei von dem Intervall des Merkmals aus, in dem alle Stichprobenwerte liegen, und unterteilt dieses in Teilintervalle der Breite Δx_i. Die Δx_i brauchen nicht gleich groß zu sein, doch erleichtert eine derartige Festlegung die weitere Rechnung. Alle Stichprobenwerte $z(x_i)$ in einem solchen Teilintervall bilden eine Klasse von Werten, die ihrer Intervall- oder Klassenmitte x_i zugeordnet wird. Die Definition von $z(x_i)$ lautet exakt

$$z(x_i) \equiv z\left((x_i - \frac{\Delta x_i}{2}) < x < (x_i + \frac{\Delta x_i}{2})\right) \tag{2.23}$$

Die Anzahlen $z(x_i)$ werden als **Besetzungszahlen** (E: absolute class frequency) bezeichnet: Die Klasse x_i ist mit $z(x_i)$ Werten besetzt. Die Klassenbildung kann als eine Diskretisierung angesehen werden, bei der das ursprünglich stetige Merkmal auf die diskreten Klassenmitten reduziert wird. Die Stichprobe wird damit verändert. Welche Auswirkungen dies hat, muss in jedem Fall bedacht werden.

Bezieht man $z(x_i)$ wie bisher auf den Stichprobeninhalt z und zusätzlich auf die Klassenbreite Δx_i, so erhält man mit

$$f(x_i) = \frac{z(x_i)}{z \cdot \Delta x_i} \tag{2.24}$$

die **relative Häufigkeitsdichte** oder **Dichtefunktion der relativen Häufigkeit**.

Die gemäß 2.24 berechnete relative Häufigkeitsdichte wird in einem **Histogramm** (E: histogram) oder **Säulendiagramm** dargestellt, siehe Abbildung 2.3. Das Säulendiagramm bei stetigem Merkmal entspricht dem Stabdiagramm bei diskretem Merkmal. Zu beachten ist jedoch, dass $h(x_i)$ und $f(x_i)$ infolge der Division durch die Klassenbreite in 2.24 unterschiedliche Dimensionen haben. Die relative Häufigkeitsdichte hat die reziproke Dimension des Merkmals.

Wie für ein diskretes Merkmal kann man auch für ein stetiges Merkmal eine **relative Häufigkeitssumme** $F(x_i)$ oder **Summenfunktion der relativen Häufigkeit** einführen

$$F(x_i) = f(x_1)\Delta x_1 + f(x_2)\Delta x_2 + \ldots + f(x_i)\Delta x_i = \sum_{\nu=1}^{i} f(x_\nu)\,\Delta x_\nu \tag{2.25}$$

Da vereinbart wurde, dass alle Stichprobenwerte, die in eine bestimmte Klasse fallen, der Klassenmitte x_i zugeordnet werden, erhält man eine Treppenkurve, die sich jeweils bei den Klassenmitten sprunghaft erhöht.

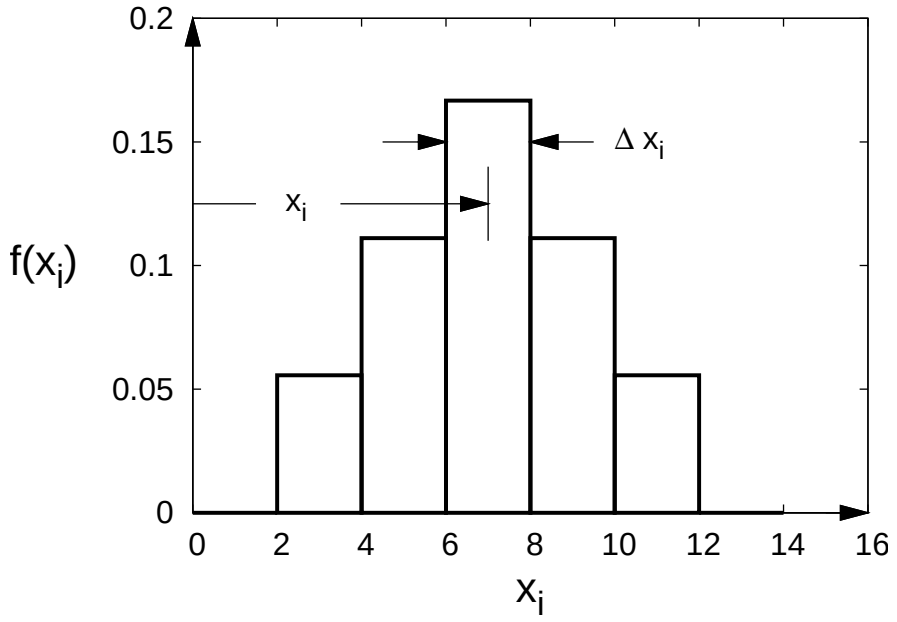


Abb. 2.3: Säulendiagramm (Histogramm) der Dichtefunktion der relativen Häufigkeit

Bei der Berechnung der Häufigkeitssummenwerte kommt man prinzipiell auch ohne Klassenbildung aus. Bei kleinem Stichprobeninhalt ist dies vorzuziehen und verbessert die Genauigkeit. Bei großem Stichprobeninhalt sind die Unterschiede in der Praxis vernachlässigbar. Auch gibt es viele Messgeräte, deren Auswerteelektronik bereits eine Klassenbildung vornimmt und die Messdaten nur klassenweise ausgibt.

Trägt man die Summenfunktion der relativen Häufigkeit $F(x_i)$ in einem Diagramm auf, so ergibt sich wie bei einem diskreten Merkmal eine monoton ansteigende Treppenkurve, siehe Abbildung 2.4.

Der einzige Unterschied ist der, dass es bei einem stetigen Merkmal gerechtfertigt und auch üblich ist, die zunächst erhaltene Treppenkurve durch einen stetigen Kurvenzug auszugleichen. Die Ausgleichskurve legt man nach Möglichkeit durch die Stufenmitten, da an der oberen Klassengrenze x_o, die der Stufenmitte entspricht, genau alle die Elemente berücksichtigt sind, deren Merkmalswert x kleiner oder gleich der oberen Klassengrenze x_o ist.

Bei der Berechnung von Mittelwert und Varianz einer Stichprobe sind zwei Möglichkeiten zu unterscheiden. Gehen wir von den Stichprobenwerten in der Urliste aus, so ist der Mittelwert der Stichprobe definiert als das arithmetische Mittel aller Stichprobenwerte. Da, wie in 2.21 festgehalten, jeder Stichprobenwert x_i nur einmal vorkommt, gilt

$$\bar{x} = \frac{1}{z}(x_1 + x_2 + \ldots + x_z) = \frac{1}{z}\sum_{\nu=1}^{z} x_\nu \tag{2.26}$$

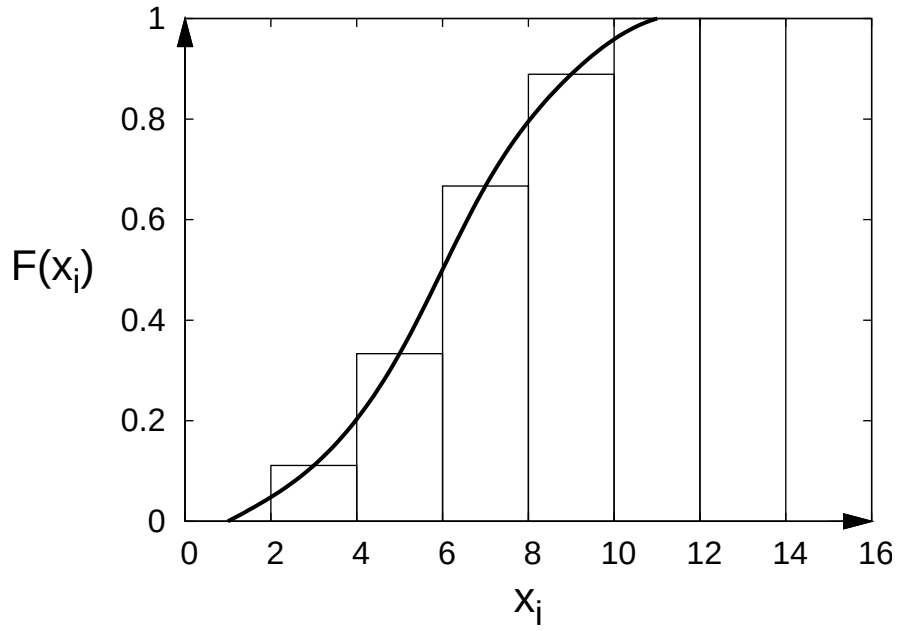


Abb. 2.4: Darstellung der Summenfunktion der relativen Häufigkeit

Für die Varianz einer Stichprobe erhält man unter den gleichen Voraussetzungen

$$s^2 = \frac{1}{z-1} \sum_{\nu=1}^{z} (x_\nu - \bar{x})^2 \tag{2.27}$$

oder, wenn man bei Kenntnis des Mittelwertes μ_x der Grundgesamtheit von der alternativen Definition 2.20 Gebrauch macht

$$\tilde{s}^2 = \frac{1}{z} \sum_{\nu=1}^{z} (x_\nu - \mu_x)^2 \tag{2.28}$$

Wie bisher bedeutet in vorstehenden Gleichungen z die Anzahl der Elemente in der Stichprobe (Stichprobeninhalt).

Die zweite Möglichkeit setzt eine vorhergegangene Klasseneinteilung voraus. Man tut dann so, als ob jeweils alle Werte, die zu einer Klasse zusammengefasst sind, in der jeweiligen Klassenmitte lägen. Für den Mittelwert einer Stichprobe gilt dann

$$\bar{x} = \sum_{\nu=1}^{n} x_\nu f(x_\nu) \Delta x_\nu \tag{2.29}$$

Entsprechend erhalten wir für die Varianz s^2 einer Stichprobe

$$s^2 = \frac{z}{z-1} \sum_{\nu=1}^{n} (x_\nu - \bar{x})^2 f(x_\nu) \Delta x_\nu \tag{2.30}$$

oder mit dem Mittelwert μ_x der Grundgesamtheit

$$\tilde{s}^2 = \frac{1}{z} \sum_{\nu=1}^{n} (x_\nu - \mu_x)^2 f(x_\nu) \Delta x_\nu \tag{2.31}$$

In den vorstehenden Gleichungen bezeichnet z wieder die Anzahl der Elemente in der Stichprobe und n die Anzahl der Klassen, in die der Wertebereich des Merkmals x aufgeteilt wurde. Bei der Berechnung von Mittelwert und Varianz einer Stichprobe macht man bei der zweiten Möglichkeit einen Fehler, der jedoch umso geringer ausfällt, je größer der Stichprobeninhalt ist. Dies ist deshalb wichtig zu wissen, weil im Fall großer Datenmengen – wie schon erwähnt – die Auswerteelektronik der Messgeräte häufig automatisch eine Klasseneinteilung vornimmt und die primären Messwerte (Urliste) gar nicht zur Verfügung stehen.

Im Fall kleiner Stichproben kennt man dagegen die primären Messwerte (Urliste) und kann damit von den Definitionsgleichungen 2.27 bis 2.29 Gebrauch machen.

Kapitel 3

Wahrscheinlichkeit (Beschreibung von Grundgesamtheiten)

Als wahrscheinlich werden im Alltag Ereignisse bezeichnet, die nicht sicher eintreten werden, deren Eintreffen aber der Erwartung entsprechen würde. Eine quantitative Bewertung dessen, was als wahrscheinlich in diesem Sinne zu gelten hat, ist daraus nicht abzuleiten.

Wenn in Wissenschaft und Technik von Wahrscheinlichkeiten gesprochen wird, liegt dem dagegen eine präzise und eindeutige Definition zugrunde. Diese wird im folgenden Kapitel nebst einigen Folgerungen vorgestellt.

3.1 Axiome der Wahrscheinlichkeitsrechnung

Den Antrieb zur Entwicklung der mathematischen Wahrscheinlichkeitstheorie gab der Wunsch, die Gewinnaussichten bei Glücksspielen vorauszuberechnen. Es zeigte sich bald, dass ihre Ergebnisse auch auf andere Gebiete (Versicherungswesen, Physik, Medizin, Technik) angewendet werden können. Im Falle von Glücksspielen ist es möglich, Wahrscheinlichkeiten allein aufgrund theoretischer Überlegungen und ohne Rückgriff auf Beobachtungswerte zu bestimmen. Deshalb sind Glücksspiele nach wie vor der bevorzugte Gegenstand von Übungsbeispielen. Eine aus Vorwissen gewonnene Wahrscheinlichkeit wird als **A-priori-Wahrscheinlichkeit** oder Anfangswahrscheinlichkeit bezeichnet, eine aus Beobachtungswerten abgeleitete als **A-posteriori-Wahrscheinlichkeit** oder statistische Wahrscheinlichkeit.

Grundlage für die Berechnung von mathematischen Wahrscheinlichkeiten bei Glücksspielen ist der klassische Begriff der **Wahrscheinlichkeit** (E: probability), der von Pierre-Simon Laplace (1749–1827) im Jahr 1812 formuliert wurde:

> Gegeben seien n sich wechselseitig ausschließende gleichwahrscheinliche Fälle. Wenn davon $n(\mathrm{A})$ Fälle für den Eintritt eines Ereignisses A günstig sind, dann ist die Wahrscheinlichkeit von A bei einem Zufallsexperiment
>
> $$P(\mathrm{A}) = \frac{n(\mathrm{A})}{n}$$

Zur Veranschaulichung diene der ideale Würfel. Jede Augenzahl von 1 bis 6 ist gleichwahrscheinlich, andernfalls wäre es kein idealer Würfel. Die Wahrscheinlichkeit, eine 2 zu werfen (Ereignis A), ist $P(\mathrm{A}) = 1/6$, die Wahrscheinlichkeit, eine Zahl > 4 zu werfen (Ereignis B), ist $P(\mathrm{B}) = 2/6 = 1/3$. Ein solcher idealer Würfel wird auch als Laplace-Würfel bezeichnet.

Man erkennt, dass die klassische Definition der mathematischen Wahrscheinlichkeit darin besteht, dass der Begriff der Wahrscheinlichkeit auf den Begriff der **Gleichwahrscheinlichkeit** zurückgeführt wird. Letzterer gilt als grundlegend und bedarf keiner weiteren Definition.[1] Der klassische Wahrscheinlichkeitsbegriff kann deshalb nur auf solche Zufallsexperimente angewendet werden, bei denen eine Einteilung aller möglichen Ereignisse in endlich viele gleichwahrscheinliche Fälle möglich ist. Solche Zufallsexperimente werden als **Laplace-Experimente** bezeichnet.

Häufig fehlen hierfür jedoch alle Voraussetzungen. Bei den meisten praktisch wichtigen Problemen in Technik, Medizin, Naturwissenschaften und Wirtschaft lassen sich gleichwahrscheinliche Fälle nicht konstruieren. Man erkennt, dass die klassische Definition der mathematischen Wahrscheinlichkeit zwar nicht falsch oder in sich widersprüchlich ist, für viele Zwecke aber nicht ausreicht.

Zu einem allgemeineren Wahrscheinlichkeitsbegriff gelangt man, wenn man von der Erfahrung ausgeht, dass sich die relative Häufigkeit eines Ereignisses bei den meisten Zufallsexperimenten in großen Versuchsreihen als nahezu konstant erweist (Stabilität der relativen Häufigkeiten). Man nimmt an, dass sich in solchen Fällen zu jedem Ereignis A eine feste Zahl $P(\mathrm{A})$ angeben lässt, die man die Wahrscheinlichkeit des Ereignisses A bei dem betreffenden Zufallsexperiment nennt, und dass bei oftmaliger Ausführung des Experimentes es praktisch gewiss ist, dass die relative Häufigkeit $h(\mathrm{A})$ ungefähr gleich $P(\mathrm{A})$ ist. Im mathematischen Sinn lässt sich dies jedoch nicht beweisen. Diese Auffassung der Wahrscheinlichkeit als Grenzwert der Häufigkeit für unendlichfache Wiederholung des Zufallsexperiments geht auf Richard von Mises (1883–1953) zurück, der sie 1919 publiziert hat.

Es zeigte sich jedoch, dass man die mathematische Wahrscheinlichkeitstheorie – wie andere Gebiete der Mathematik auch – auf gewissen Grundannahmen (Axiomen) aufbauen muss. Die Axiome sind so zu wählen, dass die daraus abgeleiteten Sätze mit der Erfahrung in Einklang stehen. Die folgende Formulierung der Axiome der mathematischen Wahrscheinlichkeitstheorie geht auf Andrei Nikolajewitsch Kolmogorow (1903–1987, Axiome veröffentlicht 1933) zurück:

- **1. Axiom** Sind A und B Ereignisse, so sind auch $\bar{\mathrm{A}}$, AB und $\mathrm{A} + \mathrm{B}$ Ereignisse.
- **2. Axiom** Jedem Ereignis A ist eine reelle Zahl $P(\mathrm{A}) \geq 0$ zugeordnet.

[1] Gelegentlich liest man, dass Ereignisse dann als gleichwahrscheinlich anzusehen sind, wenn kein Grund für eine andere Annahme zu erkennen ist (Indifferenzprinzip oder Prinzip vom unzureichenden Grund).

- **3. Axiom** Für das sichere Ereignis E gilt $P(\mathrm{E}) = 1$.
- **4. Axiom** Wenn A und B sich wechselseitig (paarweise) ausschließen, gilt

  $$P(\mathrm{A} + \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B})$$

- **5. Axiom** (Stetigkeitsaxiom der Wahrscheinlichkeitstheorie) Sind A_1, A_2, $\ldots\,\mathrm{A}_n$ Ereignisse, die niemals alle zugleich als Ergebnis eines Zufallsexperimentes eintreffen können, so ist

  $$\lim_{n\to\infty} P(\mathrm{A}_1\mathrm{A}_2\ldots\mathrm{A}_n) = 0$$

Die Bedeutung der Ereignisse $\bar{\mathrm{A}}$, AB und $\mathrm{A}+\mathrm{B}$ wurde bereits im Zusammenhang mit der Definition entsprechender relativer Häufigkeiten auf Seite 9 erklärt. Neu ist die Einführung des **sicheren Ereignisses** E. Man kann E auch als die Menge aller Elementarereignisse auffassen, die als Ergebnisse eines Zufallsexperimentes (Ausprägungen) in Betracht zu ziehen sind.

Eine Erläuterung zum 5. Axiom: Die Ereignisse A_i können beliebiger Art sein. Es wird nicht angenommen, dass sie sich wechselseitig ausschließen, sondern lediglich gefordert, dass sie *niemals alle zugleich* eintreffen können.

Hinsichtlich der Formulierung der Axiome besteht kein einheitlicher Sprachgebrauch. Man findet formal unterschiedliche Darstellungen, die jedoch letztlich den gleichen Inhalt haben. Häufig wird das 1. Axiom durch die Voraussetzung ersetzt, dass A und B einen Borelschen Mengenkörper (eine σ-Algebra) bilden. Das 5. Axiom dehnt die Gültigkeit des 4. Axioms auf unendlich viele Ereignisse aus und wird oft weggelassen.

In größerem Zusammenhang lässt sich die Wahrscheinlichkeitsrechnung als die Theorie der normierten Booleschen Algebren (nach George Boole, 1815–1864) darstellen. Wir verfolgen diesen Ansatz jedoch nicht weiter und geben dem interessierten Leser nur das Stichwort an die Hand. Wer sich für die Entwicklung des Wahrscheinlichkeitsbegriffs näher interessiert, schlage in der deutschen Wikipedia unter *Wahrscheinlichkeit* und *Geschichte der Wahrscheinlichkeitsrechnung* nach.

Axiome – Aussagen, die sich weder beweisen noch widerlegen lassen – gibt es auch außerhalb der Mathematik. In der Philosophie werden sie als *Evidenzen* bezeichnet, von Immanuel Kant (1724–1804) als *anschauliche Gewissheiten*, von Edmund Husserl (1859–1938) als *Selbstgegebenheiten*. Mathematische Theorien bemühen sich, mit wenigen Axiomen auszukommen, aber ohne geht es nicht. Wieviele und welche das sind, ist im konkreten Fall Gegenstand der wissenschaftlichen Diskussion. **Definitionen** – mehr oder weniger willkürliche Festlegungen – lassen sich ebenfalls weder beweisen noch widerlegen. Sie können zweckmäßig oder unzweckmäßig sein, aber nicht richtig oder falsch, wenn wir von Fehlern beim Definieren absehen (Zirkel und andere).

3.2 Folgerungen, Additionssatz, Multiplikationssatz

Die Ereignisse A und $\bar{\text{A}}$ schließen sich wechselseitig aus. Gemäß dem 4. Axiom gilt daher

$$P(\text{A} + \bar{\text{A}}) = P(\text{A}) + P(\bar{\text{A}}) \tag{3.1}$$

Zugleich ist $\text{A} + \bar{\text{A}}$ ein sicheres Ereignis; aus dem 3. Axiom folgt

$$P(\text{A} + \bar{\text{A}}) = P(\text{E}) = 1 \tag{3.2}$$

Aus 3.1 und 3.2 ergibt sich

$$P(\text{A}) = 1 - P(\bar{\text{A}}) \tag{3.3}$$

Man wendet diese einfache Formel an, wenn $P(\bar{\text{A}})$ leichter zu bestimmen ist als die gesuchte Wahrscheinlichkeit $P(\text{A})$.

Da aufgrund des 1. und des 2. Axioms $P(\bar{\text{A}}) \geq 0$ ist, folgt aus 3.3 unmittelbar

$$0 \leq P(\text{A}) \leq 1 \tag{3.4}$$

In anderen Darstellungen ist dies in die Formulierung des entsprechenden Axioms mit hineingenommen.

Ist A ein *unmögliches* Ereignis, so ist

$$P(\text{A}) = 0 \tag{3.5}$$

Dies folgt aus der Überlegung, dass $\bar{\text{A}}$ offenbar ein sicheres Ereignis ist, und dass daher

$$P(\bar{\text{A}}) = 1 \tag{3.6}$$

ist. Aus 3.3 folgt damit sofort 3.5. Man darf aus 3.5 nicht den Schluss ziehen, dass ein Ereignis, dessen Wahrscheinlichkeit null ist, unmöglich ist. Der Umkehrschluss ist im Fall der Gleichung 3.5 unzulässig, wie später an einem einfachen Beispiel gezeigt wird. Ebenso darf man aus $P(A) = 1$ nicht folgern, A sei ein sicheres Ereignis. Bei weitem nicht jede Umkehrung einer gültigen Aussage ist wieder eine gültige Aussage.

Aus dem 4. Axiom folgt durch Induktion für n Ereignisse A_1, A_2 $\ldots \text{A}_n$, die sich bei einem Zufallsexperiment wechselseitig ausschließen

$$P(\text{A}_1 + \text{A}_2 + \ldots + \text{A}_n) = P(\text{A}_1) + P(\text{A}_2) + \ldots + P(\text{A}_n) \tag{3.7}$$

Man bezeichnet vorstehenden Satz als den **Additionssatz der Wahrscheinlichkeit**.

Sind A_1, A_2 $\ldots$ abzählbar unendlich viele Ereignisse, die sich bei einem Zufallsexperiment wechselseitig ausschließen, so gilt

$$P(\text{A}_1 + \text{A}_2 + \ldots) = P(\text{A}_1) + P(\text{A}_2) + \ldots \tag{3.8}$$

Bei der Herleitung von 3.8 wird vom 5. Axiom, dem Stetigkeitsaxiom, Gebrauch gemacht.

Haben zwei beliebige Ereignisse A und B bei einem Zufallsexperiment die Wahrscheinlichkeiten $P(\mathrm{A})$ und $P(\mathrm{B})$, so ist die Wahrscheinlichkeit des Ereignisses A + B, das heißt dafür, dass entweder A oder B oder beide gemeinsam eintreffen

$$P(\mathrm{A}+\mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{AB}) \tag{3.9}$$

Dies ist der **Additionssatz für beliebige Ereignisse**. Bei seiner Herleitung geht man davon aus, dass das Ereignis A + B in den drei sich wechselseitig ausschließenden Formen AB, $\mathrm{A\bar{B}}$ und $\mathrm{\bar{A}B}$ realisierbar ist. Das heißt, es gilt

$$P(\mathrm{A}+\mathrm{B}) = P(\mathrm{AB} + \mathrm{A\bar{B}} + \mathrm{\bar{A}B}) \tag{3.10}$$

Mit 3.7, dem verallgemeinerten 4. Axiom, folgt daraus

$$P(\mathrm{A}+\mathrm{B}) = P(\mathrm{AB}) + P(\mathrm{A\bar{B}}) + P(\mathrm{\bar{A}B}) \tag{3.11}$$

Weiterhin lässt sich das Ereignis A in den beiden sich wechselseitig ausschließenden Formen AB und $\mathrm{A\bar{B}}$ realisieren

$$P(\mathrm{A}) = P(\mathrm{AB} + \mathrm{A\bar{B}}) \tag{3.12}$$

Mit dem 4. Axiom folgt

$$P(\mathrm{A}) = P(\mathrm{AB}) + P(\mathrm{A\bar{B}}) \tag{3.13}$$

Entsprechend gilt für das Ereignis B

$$P(\mathrm{B}) = P(\mathrm{AB}) + P(\mathrm{\bar{A}B}) \tag{3.14}$$

Aus 3.11, 3.13 und 3.14 folgt 3.9.

Im Folgenden sollen analog zu den bedingten relativen Häufigkeiten **bedingte Wahrscheinlichkeiten** (E: conditional probability) eingeführt werden. Hierzu betrachtet man ein Zufallsexperiment, bei dem sich n gleichwahrscheinliche Fälle unterscheiden lassen. Davon seien $n(\mathrm{A})$ Fälle für den Eintritt eines Ereignisses A, $n(\mathrm{B})$ Fälle für den Eintritt eines Ereignisses B und $n(\mathrm{AB})$ Fälle für den gemeinsamen Eintritt der beiden Ereignisse A und B günstig. Da Gleichwahrscheinlichkeit vorausgesetzt wurde, kann die klassische Definition der mathematischen Wahrscheinlichkeit angewendet werden. Man erhält

$$P(\mathrm{A}) = \frac{n(\mathrm{A})}{n} \qquad P(\mathrm{B}) = \frac{n(\mathrm{B})}{n} \qquad P(\mathrm{AB}) = \frac{n(\mathrm{AB})}{n}$$

Bezeichnet man mit $P(\mathrm{A|B})$ die Wahrscheinlichkeit von A unter der zusätzlichen Bedingung, dass nur solche Fälle betrachtet werden, in denen B eintrifft (A vorausgesetzt B), so gilt offenbar

$$P(\mathrm{A|B}) = \frac{n(\mathrm{AB})}{n(\mathrm{B})}$$

Dividiert man Zähler und Nenner durch n, so folgt

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{AB})}{P(\mathrm{B})} \tag{3.15}$$

Man bezeichnet $P(\mathrm{A}|\mathrm{B})$ als die bedingte Wahrscheinlichkeit von A unter der Hypothese B. Die Definition 3.15 ist nur sinnvoll, wenn $P(\mathrm{B}) \neq 0$ ist. Entsprechend gilt für die bedingte Wahrscheinlichkeit von B unter der Hypothese A (B vorausgesetzt A)

$$P(\mathrm{B}|\mathrm{A}) = \frac{P(\mathrm{AB})}{P(\mathrm{A})} \tag{3.16}$$

wobei $P(\mathrm{A}) \neq 0$ sein muss.

Die bedingten Wahrscheinlichkeiten $P(\mathrm{A}|\mathrm{B})$ und $P(\mathrm{B}|\mathrm{A})$ wurden hier für solche Zufallsexperimente definiert, bei denen Gleichwahrscheinlichkeit gegeben ist und deshalb die klassische Wahrscheinlichkeitsdefinition angewendet werden kann. Die Definitionen 3.15 und 3.16 lassen sich jedoch auch auf beliebige andere Zufallsexperimente ausdehnen. Schließen sich die Ereignisse A und B wechselseitig aus, so ist $P(\mathrm{AB}) = 0$ und damit auch $P(\mathrm{A}|\mathrm{B}) = 0$ und $P(\mathrm{B}|\mathrm{A}) = 0$. Aus 3.15 und 3.16 folgt unmittelbar der Satz:

> Haben zwei Ereignisse A und B bei einem Zufallsexperiment die Wahrscheinlichkeiten $P(\mathrm{A})$ und $P(\mathrm{B})$, so ist die Wahrscheinleichkeit des gemeinsamen Eintreffens von A und B
>
> $$P(\mathrm{AB}) = P(\mathrm{A})P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})P(\mathrm{A}|\mathrm{B})$$

Obenstehender Satz ist der **Multiplikationssatz für Wahrscheinlichkeiten**. Durch Umstellen erhalten wir das **Bayes-Theorem** (nach Thomas Bayes, 1702–1761)

$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)} \tag{3.17}$$

das einige interessante Anwendungen hat.

Sind zwei Ereignisse A und B stochastisch unabhängig, so gilt

$$P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A}) \qquad P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})$$

und der Multiplikationssatz nimmt folgende einfache Form an

$$P(\mathrm{AB}) = P(\mathrm{A})P(\mathrm{B}) \tag{3.18}$$

Beispiel 3.1 : Es werden ohne Zurücklegen zwei Karten gezogen. Wie groß ist die Wahrscheinlichkeit, als erste Karte ein As und als zweite Karte einen König zu ziehen? Wir definieren folgende Ereignisse

- A: Erste Karte ist ein As.

- B: Zweite Karte ist ein König.
- AB: Erste Karte ist ein As und zweite Karte ist ein König.

Die Wahrscheinlichkeiten errechnen sich dann wie folgt

$$P(\mathrm{A}) = \frac{4}{52} \qquad P(\mathrm{B}|\mathrm{A}) = \frac{4}{51}$$

$$P(\mathrm{AB}) = P(\mathrm{A})\ P(\mathrm{B}|\mathrm{A}) = \frac{4}{52} \cdot \frac{4}{51} = 0,00603$$

Wenn die erste Karte zurückgesteckt wird, bevor die zweite Karte gezogen wird, besteht stochastische Unabhängigkeit, und es gilt

$$P(\mathrm{A}) = \frac{4}{52} \qquad P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B}) = \frac{4}{52}$$

$$P(\mathrm{AB}) = P(\mathrm{A})\ P(\mathrm{B}) = \frac{4}{52} \cdot \frac{4}{52} = 0,00592$$

Ende Beispiel

Der Additionssatz für beliebige Ereignisse und der Multiplikationssatz lassen sich von zwei Ereignissen auf beliebig viele Ereignisse erweitern. Für drei Ereignisse A, B und C mit den Wahrscheinlichkeiten $P(\mathrm{A})$, $P(\mathrm{B})$ und $P(\mathrm{C})$ werden die Erweiterungen im Folgenden abgeleitet

$$P(\mathrm{ABC}) = P(\mathrm{AB})\ P(\mathrm{C}|\mathrm{AB}) \tag{3.19}$$

Hier wurde AB, wie im 1. Axiom formuliert, als ein Ereignis aufgefasst. Andererseits gilt der Multiplikationssatz für Wahrscheinlichkeiten

$$P(\mathrm{AB}) = P(\mathrm{A})\ P(\mathrm{B}|\mathrm{A}) \tag{3.20}$$

Aus 3.19 und 3.20 folgt

$$P(\mathrm{ABC}) = P(\mathrm{A})\ P(\mathrm{B}|\mathrm{A})\ P(\mathrm{C}|\mathrm{AB}) \tag{3.21}$$

Das ist der **Multiplikationssatz für drei Ereignisse**.

Bei der Erweiterung des Additionssatzes wird zunächst $\mathrm{B} + \mathrm{C}$ als ein Ereignis aufgefasst, sodass gilt

$$P(\mathrm{A} + \mathrm{B} + \mathrm{C}) = P(\mathrm{A}) + P(\mathrm{B} + \mathrm{C}) - P(\mathrm{A}|(\mathrm{B} + \mathrm{C})) \tag{3.22}$$

Aus dem Multiplikationssatz folgt

$$P(\mathrm{A}(\mathrm{B} + \mathrm{C})) = P(\mathrm{A})P((\mathrm{B} + \mathrm{C})|\mathrm{A}) \tag{3.23}$$

Mit dem Additionssatz für zwei beliebige Ereignisse erhält man

$$P(\mathrm{B}+\mathrm{C}) = P(\mathrm{B}) + P(\mathrm{C}) - P(\mathrm{BC}) \tag{3.24}$$

Dies gilt ebenso für bedingte Wahrscheinlichkeiten. Es ist

$$P((\mathrm{B}+\mathrm{C})|\mathrm{A}) = P(\mathrm{B}|\mathrm{A}) + P(\mathrm{C}|\mathrm{A}) - P(\mathrm{BC}|\mathrm{A}) \tag{3.25}$$

Setzt man 3.23, 3.24 und 3.25 in 3.22 ein, so folgt

$$\begin{aligned} P(\mathrm{A}+\mathrm{B}+\mathrm{C}) = \\ & P(\mathrm{A}) + P(\mathrm{B}) + P(\mathrm{C}) - P(\mathrm{BC}) - P(\mathrm{A})P(\mathrm{B}|\mathrm{A}) - P(\mathrm{A})P(\mathrm{C}|\mathrm{A}) \\ & + P(\mathrm{A})P(\mathrm{BC}|\mathrm{A}) \end{aligned}$$

Nach mehrmaliger Anwendung des Multiplikationssatzes folgt schließlich

$$\begin{aligned} P(\mathrm{A}+\mathrm{B}+\mathrm{C}) = \\ & P(\mathrm{A}) + P(\mathrm{B}) + P(\mathrm{C}) - P(\mathrm{AB}) - P(\mathrm{BC}) - P(\mathrm{AC}) + P(\mathrm{ABC}) \end{aligned} \tag{3.26}$$

Das ist der **Additionssatz für drei beliebige Ereignisse**. Anwendungen dieser Erweiterungen zeigt das folgende Beispiel.

Beispiel 3.2 : Aus einem gut gemischten, verdeckten Kartenspiel werden nacheinander ohne Zurücklegen drei Karten gezogen. Wie groß ist die Wahrscheinlichkeit, als erste, zweite oder dritte Karte einen König zu ziehen? Wir definieren folgende Ereignisse

- A: die erste Karte ist ein König,
- B: die zweite Karte ist ein König,
- C: die dritte Karte ist ein König.

Dann gilt

$$P(\mathrm{A}) = \frac{4}{52} \qquad P(\bar{\mathrm{A}}) = \frac{48}{52}$$

$$P(\mathrm{B}) = P(\mathrm{AB} + \bar{\mathrm{A}}\mathrm{B}) = P(\mathrm{AB}) + P(\bar{\mathrm{A}}\mathrm{B})$$

worin

$$P(\mathrm{AB}) = P(\mathrm{A})\ P(\mathrm{B}|\mathrm{A}) \quad \text{mit} \quad P(\mathrm{B}|\mathrm{A}) = \frac{3}{51}$$

und

$$P(\bar{\mathrm{A}}\mathrm{B}) = P(\bar{\mathrm{A}})\ P(\mathrm{B}|\bar{\mathrm{A}}) \quad \text{mit} \quad P(\mathrm{B}|\bar{\mathrm{A}}) = \frac{4}{51}$$

mit Zahlenwerten

$$P(\mathrm{B}) = \frac{4}{52}\frac{3}{51} + \frac{48}{52}\frac{4}{51} = \frac{4}{52}$$

Weiterhin gilt

$$\begin{aligned} P(\mathrm{C}) &= P(\mathrm{ABC} + \bar{\mathrm{A}}\mathrm{BC} + \mathrm{A}\bar{\mathrm{B}}\mathrm{C} + \bar{\mathrm{A}}\bar{\mathrm{B}}\mathrm{C}) \\ &= P(\mathrm{ABC}) + P(\bar{\mathrm{A}}\mathrm{BC}) + P(\mathrm{A}\bar{\mathrm{B}}\mathrm{C}) + P(\bar{\mathrm{A}}\bar{\mathrm{B}}\mathrm{C}) \end{aligned}$$

worin

$$P(\mathrm{ABC}) = P(\mathrm{AB})P(\mathrm{C}|\mathrm{AB})$$

$$P(\bar{\mathrm{A}}\mathrm{BC}) = P(\bar{\mathrm{A}}\mathrm{B})P(\mathrm{C}|\bar{\mathrm{A}}\mathrm{B})$$

$$P(\mathrm{A}\bar{\mathrm{B}}\mathrm{C}) = P(\mathrm{A}\bar{\mathrm{B}}P(\mathrm{C}|\mathrm{A}\bar{\mathrm{B}})$$

$$P(\bar{\mathrm{A}}\bar{\mathrm{B}}\mathrm{C}) = P(\bar{\mathrm{A}}\bar{\mathrm{B}})P(\mathrm{C}|\bar{\mathrm{A}}\bar{\mathrm{B}})$$

$$P(\mathrm{A}\bar{\mathrm{B}}) = P(\mathrm{A})P(\bar{\mathrm{B}}|\mathrm{A}) \quad \text{mit} \quad P(\bar{\mathrm{B}}|\mathrm{A}) = \frac{48}{51}$$

$$P(\bar{\mathrm{A}}\bar{\mathrm{B}}) = P(\bar{\mathrm{A}})P(\bar{\mathrm{B}}|\bar{\mathrm{A}}) \quad \text{mit} \quad P(\bar{\mathrm{B}}|\bar{\mathrm{A}}) = \frac{47}{51}$$

$$P(\mathrm{C}|\mathrm{AB}) = \frac{2}{50}$$

$$P(\mathrm{C}|\bar{\mathrm{A}}\mathrm{B}) = \frac{3}{50}$$

$$P(\mathrm{C}|\mathrm{A}\bar{B}) = \frac{3}{50}$$

$$P(\mathrm{C}|\bar{\mathrm{A}}\bar{\mathrm{B}}) = \frac{4}{50}$$

und als Ergebnis

$$P(\mathrm{C}) = \frac{4}{52}\frac{3}{51}\frac{2}{50} + \frac{48}{52}\frac{4}{51}\frac{3}{50} + \frac{4}{52}\frac{48}{51}\frac{3}{50} + \frac{48}{52}\frac{47}{51}\frac{4}{50} = \frac{4}{52}$$

Für die drei Ereignisse A, B und C erhalten wir die gleiche Wahrscheinlichkeit. Bei einem Kartenspiel ist es daher gleichgültig, in welcher Reihenfolge die Karten gezogen beziehungsweise an die Mitspieler ausgeteilt werden.

Ende Beispiel

Zwischen den Sätzen, die zuvor für relative Häufigkeiten abgeleitet wurden, und den Sätzen für die entsprechenden Wahrscheinlichkeiten herrscht formal völlige Übereinstimmung, wie der Vergleich zeigt. Diese Übereinstimmung folgt daraus, dass zwischen relativer Häufigkeit und Wahrscheinlichkeit eine enge Beziehung besteht. Sie findet Ausdruck in dem **Gesetz der großen Zahlen** von Jakob Bernoulli (1655–1705):

> Es sei A ein Ereignis, das bei einem Zufallsexperiment die Wahrscheinlichkeit p besitzt. Mit $z(\mathrm{A})$ werde die Anzahl des Eintreffens von A bei z unabhängigen Ausführungen des Experiments bezeichnet. Dann gilt
>
> $$\lim_{z\to\infty} P\left\{\left|\frac{z(\mathrm{A})}{z} - p\right| > \epsilon\right\} = 0 \tag{3.27}$$
>
> worin ϵ eine beliebig kleine, aber von null verschiedene positive Zahl ist.

In Worten: Mit wachsendem z strebt die Wahrscheinlichkeit, dass die relative Häufigkeit des Ereignisses A um mehr als eine vorgegebene beliebig kleine positive Zahl ϵ von p abweicht, gegen null. Dieser Satz steht mit den zuvor aufgeführten Axiomen der Wahrscheinlichkeitstheorie in Einklang, das heißt, er lässt sich unter Verwendung der Axiome beweisen. Der Beweis kann allerdings an dieser Stelle noch nicht geführt werden. Wir kommen auf Seite 84 auf den Beweis zurück.

Das Gesetz der großen Zahlen bildet die Grundlage für die praktische Bestimmung von Wahrscheinlichkeiten, wenn sich bei Zufallsexperimenten keine endlich vielen gleichwahrscheinlichen Fälle unterscheiden lassen. Die unbekannten Werte der Wahrscheinlichkeiten werden aufgrund beobachteter relativer Häufigkeiten in langen Versuchsreihen festgelegt. Beispielsweise ermöglicht das Gesetz Versicherungsgesellschaften ungefähre Vorhersagen des zu erwartenden Schadensverlaufs. Je größer der Kreis der Versicherten ist, desto geringer ist der unvorhersagbare Einfluss des Zufalls.

3.3 Wahrscheinlichkeitsverteilungen

3.3.1 Verteilungen einer Zufallsvariablen

Im Folgenden werden nur solche Zufallsexperimente betrachtet, bei denen das Ergebnis einer einzelnen Ausführung (Realisation) jeweils durch eine einzige Zahl ausgedrückt wird. Bei Wiederholung eines derartigen Zufallsexperimentes werden einer Größe x unterschiedliche Zahlenwerte x_i zugeordnet. Man bezeichnet eine

solche Größe x als **Zufallsgröße**, **Zufallsvariable** oder **stochastische Variable** (E: random variable), wenn sie folgende Eigenschaften besitzt

- Die Werte x_i sind reelle Zahlen.
- Für jede Zahl a und für jedes Intervall I auf der Zahlengeraden ist die Wahrscheinlichkeit des Ereignisses x *hat den Wert* a beziehungsweise x *liegt in dem Zahlenintervall* I im Einklang mit den Axiomen der Wahrscheinlichkeitstheorie erklärt.

Diese Definition folgt der Darstellung von Erwin Kreyszig. Zu anderen Darstellungen siehe die deutsche und die englische Wikipedia.

Die meisten bei praktischen Problemen vorkommenden Zufallsvariablen lassen sich in zwei Gruppen einteilen, nämlich in diskrete und in stetige Variable. Eine Zufallsvariable x und ihre Verteilung heißen **diskret**, falls

- die Variable x nur endlich viele oder abzählbar unendlich viele Werte x_i mit von null verschiedener Wahrscheinlichkeit annehmen kann, und
- in jedem endlichen Intervall der reellen Zahlengeraden nur endlich viele Werte x_i liegen und für jedes Intervall $a < x \leq b$, das keinen solchen Wert enthält, die zugehörige Wahrscheinlichkeit $P(a < x \leq b)$ gleich null ist.

Da eine diskrete Zufallsvariable nur bestimmte Zahlenwerte x_i annimmt, wird ihre Wahrscheinlichkeitsverteilung durch

$$P(x) = \begin{cases} P(x_i) & \text{für} \quad x = x_i \\ 0 & \text{für} \quad \text{alle übrigen } x \end{cases} \tag{3.28}$$

beschrieben. Man kann sie als Stabdiagramm aufzeichnen, siehe Abbildung 3.1 auf Seite 34.

Beispiel 3.3 : Bei dem Wurf eines idealen Würfels gilt für die Augenzahl x

$$P(x) = \begin{cases} \dfrac{1}{6} & \text{für} \quad x = 1, 2, \ldots 6 \\ 0 & \text{für} \quad \text{alle übrigen } x \end{cases}$$

Ende Beispiel

Da sich die zu den Zahlenwerten x_i gehörigen Ereignisse ausschließen, gilt

$$P(x_1 \leq x_i \leq x_k) = P(x_1) + P(x_2) + \ldots + P(x_k) \tag{3.29}$$

Summiert man über alle n Werte, die die Zufallsvariable x annehmen kann, so folgt

$$\sum_{\nu=1}^{n} P(x_\nu) = 1 \tag{3.30}$$

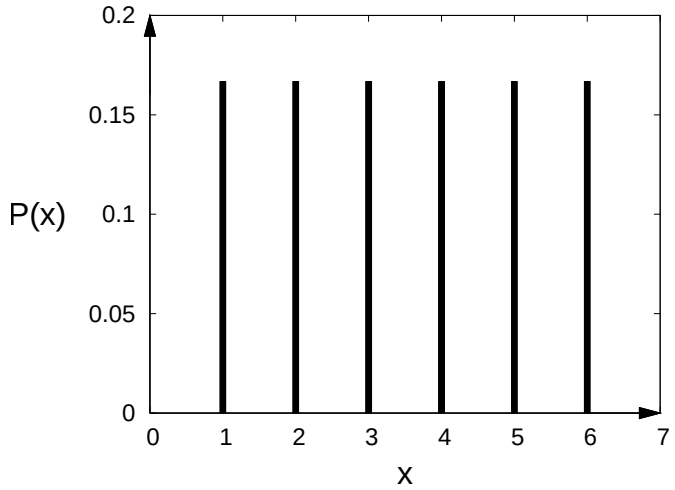


Abb. 3.1: Stabdiagramm der Wahrscheinlichkeit für die Ergebnisse beim Würfeln

Dies gilt auch dann, wenn die Zufallsvariable abzählbar unendlich viele Werte annehmen kann. In diesem Fall ist $n = \infty$ zu setzen.

Man kann zu der Wahrscheinlichkeitsverteilung 3.28 eine **Summenfunktion** einführen, indem man definiert

$$\Phi(x_i) = P(x \leq x_i) = \sum_{\nu=1}^{i} P(x_\nu) \tag{3.31}$$

Dabei ist vorausgesetzt, dass die Bezeichnung der Zahlenwerte x_i so vorgenommen wurde, dass stets $x_{i+1} > x_i$ ist. Trägt man $\Phi(x_i)$ in einem Diagramm auf, erhält man eine monoton steigende Treppenkurve.

Die Wahrscheinlichkeit dafür, dass eine diskrete Variable irgendeinen Wert innerhalb des beliebig vorgegebenen Intervalls $a < x \leq b$ annimmt, lässt sich mit Hilfe der Summenfunktion wie folgt berechnen

$$P(a < x \leq b) = \Phi(b) - \Phi(a) \tag{3.32}$$

Beweis: Die Ereignisse $x \leq a$ und $a < x \leq b$ schließen sich wechselseitig aus. Die Summe dieser Ereignisse ist $x \leq b$. Gemäß dem 4. Axiom ist daher

$$\begin{aligned} P(x \leq b) &= P(x \leq a) + P(a < x \leq b) \\ P(a < x \leq b) &= P(x \leq b) - P(x \leq a) \end{aligned} \tag{3.33}$$

Aus 3.31 folgt

$$P(x \leq a) = \Phi(a) \qquad P(x \leq b) = \Phi(b)$$

Damit erhält man aus 3.33

$$P(a < x \leq b) = \Phi(b) - \Phi(a)$$

was zu beweisen war.

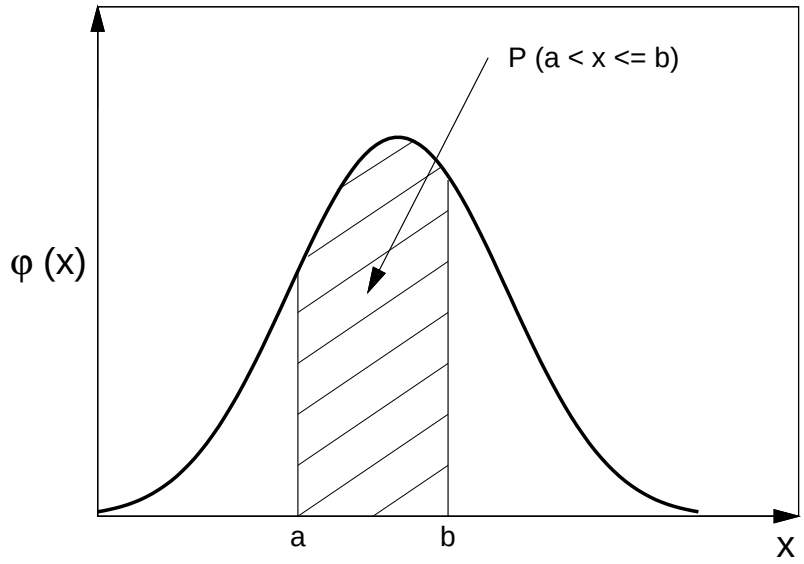


Abb. 3.2: Dichtefunktion der Wahrscheinlichkeit

Eine Zufallsvariable x und ihre Verteilung heißen **stetig**, wenn die Summenfunktion

$$\Phi(x_i) = P(x \leq x_i)$$

in Integralform dargestellt werden kann, wenn also

$$\Phi(x) = \int_{-\infty}^{x} \varphi(\xi)\mathrm{d}\xi \tag{3.34}$$

ist, wobei der Integrand eine nichtnegative und bis auf höchstens endlich viele Punkte stetige Funktion ist.

Der Integrand $\varphi(x)$ wird als **Dichtefunktion der Wahrscheinlichkeitsverteilung** bezeichnet. Aus 3.34 folgt für jedes x, bei dem $\varphi(x)$ stetig ist, durch Differenzieren

$$\frac{\mathrm{d}\Phi(x)}{\mathrm{d}x} = \varphi(x) \tag{3.35}$$

Die Dichtefunktion ist die Ableitung der Summenfunktion nach x.

Die Dichtefunktion der Wahrscheinlichkeit $\varphi(x)$ ist keine Wahrscheinlichkeit. Die Dimension von $\varphi(x)$ ist die reziproke Dimension der Zufallsvariablen x. Die Wahrscheinlichkeit dafür, dass die Zufallsvariable x irgendeinen Wert in dem differentiell schmalen Intervall zwischen x_i und $x_i + \mathrm{d}x$ annimmt, ist unter der Voraussetzung, dass $\varphi(x)$ in dem Intervall stetig ist

$$\mathrm{d}\Phi(x_i) = \varphi(x_i)\mathrm{d}x \tag{3.36}$$

Für ein beliebig vorgegebenes endliches Intervall $a < x \leq b$ folgt damit

$$P(a < x \leq b) = \int_{a}^{b} \varphi(\xi)\mathrm{d}\xi \tag{3.37}$$

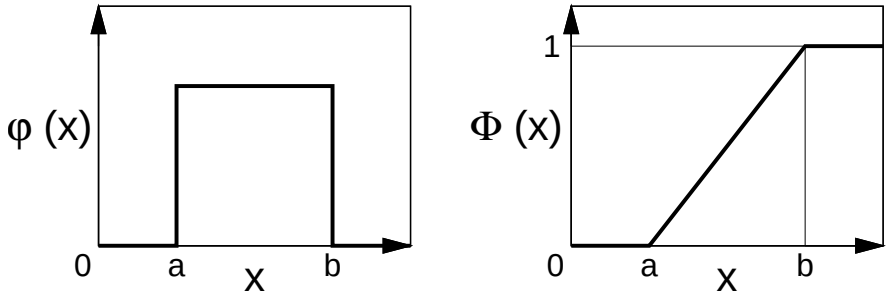


Abb. 3.3: Diagramme einer Rechteckverteilung, links die Verteilungsdichte, rechts die Verteilungssumme

Dieser Sachverhalt wird unmittelbar anschaulich, wenn man $\varphi(x)$ in einem Diagramm aufträgt. Die Wahrscheinlichkeit des Ereignisses $a < x \leq b$ ist bei linearer Auftragung gleich der Fläche unter der Kurve $\varphi(x)$ zwischen $x = a$ und $x = b$, siehe Abbildung 3.2 auf Seite 35.

Da $-\infty < x < +\infty$ ein sicheres Ereignis ist, gilt

$$P(-\infty < x < +\infty) = \int_{-\infty}^{+\infty} \varphi(\xi)\mathrm{d}\xi = 1 \tag{3.38}$$

das heißt, die ganze Fläche unter der Kurve $\varphi(x)$ hat den Wert eins.

Mit 3.34 lässt sich 3.37 umformen in

$$P(a < x \leq b) = \int_{a}^{b} \varphi(\xi)\mathrm{d}\xi = \Phi(b) - \Phi(a) \tag{3.39}$$

Dies stimmt formal mit 3.32 überein, das heißt, 3.32 gilt unabhängig davon, ob die Zufallsvariable x diskret oder stetig ist.

Aus 3.39 folgt wegen der immer gegebenen Stetigkeit der Summenfunktion $\Phi(x)$, dass die Wahrscheinlichkeit zu null wird, wenn man die Intervallbreite gegen null gehen lässt. Im Fall stetiger Verteilungen gilt daher für jeden Wert der Zufallsvariablen x

$$P(x = x_i) = 0 \tag{3.40}$$

Dies bedeutet nicht, dass $x = x_i$ ein unmögliches Ereignis ist, sonst wären ja alle Zahlenwerte x_i unmöglich. Wie in Abschnitt 3.2 auf Seite 26 erklärt wurde, darf man aus der Tatsache, dass ein unmögliches Ereignis die Wahrscheinlichkeit null hat, nicht den Umkehrschluss ziehen, ein Ereignis mit der Wahrscheinlichkeit null sei unmöglich.

Beispiel 3.4 : Die Verteilungsdichte einer Rechteckverteilung ist gegeben durch

$$\varphi(x) = \begin{cases} \dfrac{1}{b-a} & \text{für} \quad a < x < b \\ 0 & \text{für} \quad \text{alle übrigen } x \end{cases}$$

Für die Verteilungssumme folgt damit

$$\Phi(x) = \int_{-\infty}^{x} \varphi(\xi)\mathrm{d}\xi = \frac{1}{b-a}\int_{a}^{x} \mathrm{d}\xi = \frac{x-a}{b-a}$$

eine Gerade, wie Abbildung 3.3 veranschaulicht.

Ende Beispiel

3.3.2 Verteilungen mehrerer Zufallsvariablen

Die Überlegungen, die im Vorhergehenden für eine einzelne Zufallsvariable entwickelt wurden, lassen sich grundsätzlich auf beliebig viele Zufallsvariable erweitern. Mehrere Zufallsvariablen lassen sich unter gewissen Voraussetzungen zu einem **Zufallsvektor** zusammenfassen. Ihre Verteilung wird als **multivariat** bezeichnet.

Im Folgenden werden jedoch lediglich zweidimensionale (bivariate) Verteilungen behandelt. Diese Entscheidung beruht vor allem auf praktischen Erwägungen. Um eine eindimensionale Verteilung durch Messpunkte brauchbar anzunähern, sind etwa 10^2 Messwerte erforderlich. Für eine gleich gute Näherung einer zweidimensionalen Verteilung müssen etwa 10^4 gleichzeitig gemessene Wertepaare zur Verfügung stehen, im Falle einer dreidimensionalen Verteilung wären dies 10^6 simultan gemessene Wertetripel. Der Aufwand steigt rasant mit jeder zusätzlichen Variablen. Man wird deshalb kaum über zwei Dimensionen hinausgehen. Zweidimensionale Verteilungen sind jedoch immer dann unvermeidlich, wenn man es mit zwei Variablen zu tun hat, die *nicht* unabhängig voneinander sind, beziehungsweise wenn letztere Möglichkeit nicht auszuschließen ist.

Zu einem Zufallsexperiment, bei dem man eine einzelne Größe beobachtet, gehört eine einzelne Zufallsvariable, die mit x bezeichnet wird. Das Ergebnis einer jeden Ausführung des Zufallsexperimentes, wird durch eine einzelne Zahl x_i gekennzeichnet, den Wert, den x jeweils angenommen hat. Dieser Wert entspricht einem Punkt auf der x-Achse.

Kennt man für jedes Intervall $a < x \leq b$ die zugehörige Wahrscheinlichkeit $P(a < x \leq b)$, mit der x irgendeinen Wert in diesem Intervall annimmt, so ist die Wahrscheinlichkeitsverteilung der Zufallsvariablen x bekannt. Sie lässt sich durch die zugehörige Summenfunktion

$$\Phi(x_i) = P(x \leq x_i) \tag{3.41}$$

eindeutig beschreiben.

Zu einem Zufallsexperiment, bei dem man gleichzeitig zwei Größen beobachtet, gehören zwei Zufallsvariablen, die mit x und y bezeichnet werden sollen. Das Ergebnis einer jeden Ausführung des Zufallsexperimentes ist dann jeweils durch ein geordnetes Zahlenpaar x_i, y_k gekennzeichnet. Dabei ist x_i der Wert, den jeweils die Zufallsvariable x angenommen hat, und y_k der Wert, den jeweils die Zufallsvariable y angenommen hat. Diesem Zahlenpaar x_i, y_k entspricht ein Punkt in der x, y-Ebene, siehe Abbildung 3.4 auf Seite 38.

Kennt man für jedes Rechteck $a_1 < x \leq b_1$, $a_2 < y \leq b_2$ die zugehörige Wahrscheinlichkeit $P(a_1 < x \leq b_1, a_2 < y \leq b_2)$, mit der x, y irgendein Paar von Werten annimmt, das innerhalb der bezeichneten Rechteckfläche liegt, so ist

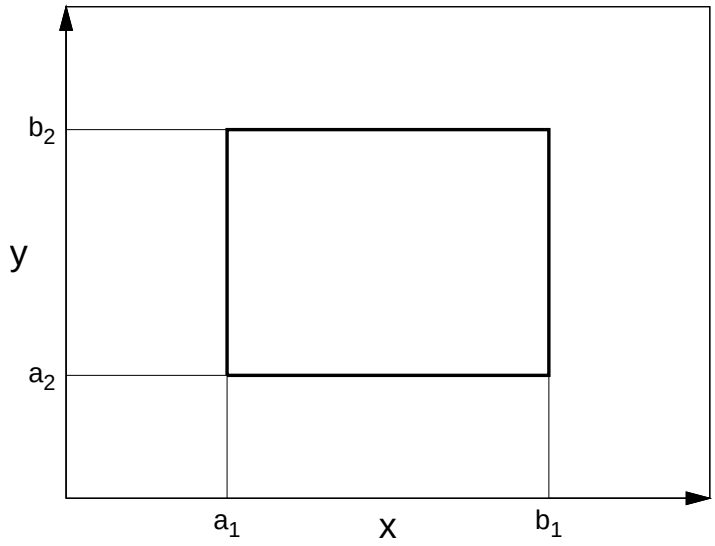


Abb. 3.4: Darstellung einer zweidimensionalen Zufallsvariablen in der x, y-Ebene

die **zweidimensionale Wahrscheinlichkeitsverteilung** der beiden Zufallsvariablen x und y oder der **zweidimensionalen Zufallsvariablen** (x, y) bekannt. Sie lässt sich durch die zugehörige Summenfunktion

$$\Phi(x_i, y_k) = P(x \leq x_i, y \leq y_k) \tag{3.42}$$

beschreiben.

Eine zweidimensionale Zufallsvariable (x, y) und deren Wahrscheinlichkeitsverteilung heißen **diskret**, wenn

- die Variable (x, y) nur endlich viele oder abzählbar unendlich viele Wertepaare x_i, y_k mit von null verschiedener Wahrscheinlichkeit annehmen kann und
- in jedem beschränktem Gebiet der x, y-Ebene nur endlich viele oder abzählbar unendlich viele solche Wertepaare x_i, y_k liegen und in jedem Gebiet, das keines dieser Wertepaare enthält, die zugehörige Wahrscheinlichkeit gleich null ist.

Die Wahrscheinlichkeitsverteilung wird ähnlich wie im eindimensionalen Fall durch

$$P(x, y) = \begin{cases} P(x_i, y_k) & \text{für} \quad x = x_i \quad y = y_k \\ 0 & \text{für} \quad \text{alle übrigen } (x, y) \end{cases} \tag{3.43}$$

beschrieben. Die Indices i und k durchlaufen unabhängig voneinander die Werte $1, 2, 3 \ldots$. Die Bezeichnungsweise ist so zu wählen, dass stets $x_{i+1} > x_i$ und $y_{k+1} > y_k$ ist.

Zwischen der Summenverteilung $\Phi(x_i, y_k)$ und der Wahrscheinlichkeitsverteilung $P(x, y)$ besteht der Zusammenhang

$$\Phi(x_i, y_k) = \sum_{\nu=1}^{i} \sum_{\lambda=1}^{k} P(x_\nu, y_\lambda) \tag{3.44}$$

Summiert man über alle Wertepaare, die die zweidimensionale Zufallsvariable (x, y) annehmen kann, so gilt

$$\sum_{\nu}\sum_{\lambda} P(x_\nu, y_\lambda) = 1 \tag{3.45}$$

Dies gilt auch dann, wenn die zweidimensionale Zufallsvariable (x, y) abzählbar unendlich viele Wertepaare x_i, y_k annehmen kann.

Die Wahrscheinlichkeit dafür, dass eine diskrete zweidimensionale Zufallsvariable irgendein Wertepaar annimmt, das innerhalb eines beliebig vorgegebenen rechteckigen Bereiches $a_1 < x \leq b_1, a_2 < y \leq b_2$ liegt, lässt sich mit Hilfe der Summenfunktion wie folgt berechnen

$$P(a_1 < x \leq b_1, a_2 < y \leq b_2) = \Phi(b_1, b_2) - \Phi(a_1, b_2) - \Phi(b_1, a_2) + \Phi(a_1, a_2) \tag{3.46}$$

Beweis: Die Ereignisse A_1, A_2, A_3 und A_4 schließen sich wechselseitig aus, siehe Abbildung 3.5 auf Seite 40. Daher gilt

$$P(\mathrm{A}_1 + \mathrm{A}_2 + \mathrm{A}_3 + \mathrm{A}_4) = P(\mathrm{A}_1) + P(\mathrm{A}_2) + P(\mathrm{A}_3) + P(\mathrm{A}_4)$$

oder auch

$$P(\mathrm{A}_1 + \mathrm{A}_2 + \mathrm{A}_3 + \mathrm{A}_4) = P(\mathrm{A}_1) + P(\mathrm{A}_2 + \mathrm{A}_3) + P(\mathrm{A}_3 + \mathrm{A}_4) - P(\mathrm{A}_3) \tag{3.47}$$

Nun ist aber

$$\begin{aligned}
P(\mathrm{A}_1 + \mathrm{A}_2 + \mathrm{A}_3 + \mathrm{A}_4) &= P(x \leq b_1, y \leq b_2) = \Phi(b_1, b_2) \\
P(\mathrm{A}_1) &= P(a_1 < x \leq b_1, a_2 < y \leq b_2) \\
P(\mathrm{A}_2 + \mathrm{A}_3) &= P(x \leq a_1, y \leq b_2) = \Phi(a_1, b_2) \\
P(\mathrm{A}_3 + \mathrm{A}_4) &= P(x \leq b_1, y \leq a_2) = \Phi(b_1, a_2) \\
P(\mathrm{A}_3) &= P(x \leq a_1, y \leq a_2) = \Phi(a_1, a_2)
\end{aligned}$$

Setzt man diese Beziehungen in 3.47 ein, so folgt 3.46.

Eine zweidimensionale Zufallsvariable (x, y) und ihre Verteilung heißen **stetig**, wenn sich die Summenfunktion

$$\Phi(x_i, y_k) = P(x \leq x_i, y \leq y_k)$$

durch ein Doppelintegral der Form

$$\Phi(x, y) = \int_{\eta=-\infty}^{y} \int_{\xi=-\infty}^{x} \varphi(\xi, \eta)\, \mathrm{d}\xi \mathrm{d}\eta \tag{3.48}$$

darstellen lässt, wobei der Integrand eine in der ganzen x, y-Ebene definierte, nichtnegative und bis auf höchstens endlich viele Kurven stetige Funktion ist.

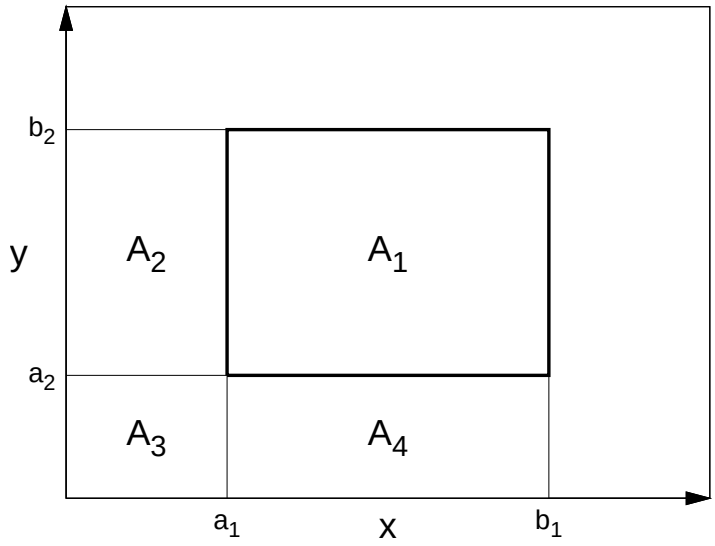


Abb. 3.5: Skizze zum Beweis

Der Integrand $\varphi(x, y)$ ist die Dichtefunktion der zweidimensionalen Wahrscheinlichkeitsverteilung. Aus 3.48 folgt für jedes Wertepaar (x, y), bei dem $\varphi(x, y)$ stetig ist, durch Differenzieren

$$\frac{\mathrm{d}^2\Phi(x,y)}{\mathrm{d}x\mathrm{d}y} = \varphi(x,y) \tag{3.49}$$

Die Wahrscheinlichkeit dafür, dass die zweidimensionale Zufallsvariable (x, y) irgendein Wertepaar in dem differentiell kleinen rechteckigen Bereich $x_i < x \leq x_i + \mathrm{d}x, y_k < y \leq y_k + \mathrm{d}y$ annimmt, ist unter der Voraussetzung, dass $\varphi(x, y)$ in diesem Bereich stetig ist

$$\mathrm{d}^2\Phi(x_i, y_k) = \varphi(x_i, y_k)\mathrm{d}x\mathrm{d}y \tag{3.50}$$

Daraus folgt für einen beliebig vorgegebenen rechteckigen Bereich $a_1 < x \leq b_1$, $a_2 < y \leq b_2$

$$P(a_1 < x \leq b_1, a_2 < y \leq b_2) = \int_{\eta=a_2}^{b_2}\int_{\xi=a_1}^{b_1} \varphi(\xi,\eta)\mathrm{d}\xi\mathrm{d}\eta \tag{3.51}$$

Integriert man 3.50 über die ganze x, y-Ebene, so folgt

$$\int_{\eta=-\infty}^{+\infty}\int_{\xi=-\infty}^{+\infty} \varphi(\xi,\eta)\ \mathrm{d}\xi\mathrm{d}\eta = 1 \tag{3.52}$$

das heißt, das Volumen unter der Fläche $\varphi(x, y)$ hat den Wert eins. Wie man zeigen kann, folgt aus 3.52 wieder 3.46

$$P(a_1 < x \leq b_1, a_2 < y \leq b_2)$$

$$= \int_{a_2}^{b_2}\left\{\int_{a_1}^{b_1} \varphi(\xi,\eta)\mathrm{d}\xi\right\}\mathrm{d}\eta$$

$$= \int_{-\infty}^{b_2} \left\{ \int_{-\infty}^{b_1} \varphi(\xi,\eta)\mathrm{d}\xi - \int_{-\infty}^{a_1} \varphi(\xi,\eta)\mathrm{d}\xi \right\} \mathrm{d}\eta$$

$$- \int_{-\infty}^{a_2} \left\{ \int_{-\infty}^{b_1} \varphi(\xi,\eta)\mathrm{d}\xi - \int_{-\infty}^{a_1} \varphi(\xi,\eta)\mathrm{d}\xi \right\} \mathrm{d}\eta$$

$$= \Phi(b_1,b_2) - \Phi(a_1,b_2) - \Phi(b_1,a_2) + \Phi(a_1,a_2)$$

Gleichung 3.46 gilt also unabhängig davon, ob die zweidimensionale Zufallsvariable (x, y) diskret oder stetig ist.

Die grafische Darstellung zweidimensionaler Wahrscheinlichkeitsverteilungen beziehungsweise ihrer Summenfunktionen in einer Ebene (Blatt Papier, Bildschirm) ist schwierig, aber grundsätzlich noch möglich. Wie in Abschnitt 10.1 *Grafische Darstellungen mittels gnuplot* auf Seite 187 gezeigt wird, bietet das Werkzeug `gnuplot` Möglichkeiten dazu. Wahrscheinlichkeitsverteilungen höherer Dimensionen lassen sich dagegen nicht mehr grafisch darstellen.

Jeder zweidimensionalen Verteilung kann man zwei eindimensionale Verteilungen zuordnen, die **Randverteilungen** (E: marginal distribution) genannt werden. Gegeben sei eine diskrete Verteilung einer zweidimensionalen Zufallsvariablen (x, y) durch 3.43

$$P(x,y) = \begin{cases} P(x_i,y_k) & \text{für} \quad x = x_i \quad y = y_k \\ 0 & \text{für} \quad \text{alle übrigen } (x,y) \end{cases} \tag{3.53}$$

Fragt man nach der Wahrscheinlichkeit, mit der x einen bestimmten Wert x_i annimmt, wobei es gleichgültig ist, welchen Wert y annimmt, so erhält man eine eindimensionale Wahrscheinlichkeitsverteilung

$$P_1(x) = \begin{cases} P_1(x_i) & \text{für} \quad x = x_i \\ 0 & \text{für} \quad \text{alle übrigen } x \end{cases} \tag{3.54}$$

worin $P_1(x_i)$ definiert ist durch

$$P_1(x_i) = P(x = x_i, y \text{ beliebig}) = \sum_{\lambda} P(x_i, y_\lambda) \tag{3.55}$$

Man bezeichnet $P_1(x)$ als die **Randverteilung der Variablen x** bezüglich der gegebenen zweidimensionalen Verteilung. Die zu 3.54 gehörige Summenfunktion ist

$$\Phi_1(x_i) = \sum_{\nu=1}^{i} P_1(x_\nu) = \sum_{\nu=1}^{i} \sum_{\lambda} P(x_\nu, y_\lambda) \tag{3.56}$$

Entsprechend bestimmt die eindimensionale Wahrscheinlichkeitsverteilung

$$P_2(y) = \begin{cases} P_2(y_k) & \text{für} \quad y = y_k \\ 0 & \text{für} \quad \text{alle übrigen } y \end{cases} \tag{3.57}$$

worin $P_2(y_k)$ definiert ist durch

$$P_2(y_k) = P(x \text{ beliebig}, y = y_k) = \sum_{\nu} P(x_\nu, y_k) \tag{3.58}$$

die **Randverteilung der Variablen y** bezüglich der gegebenen zweidimensionalen Verteilung. Die zugehörige Summenfunktion ist

$$\Phi_2(y_k) = \sum_{\lambda=1}^{k} P_2(y_\lambda) = \sum_{\nu} \sum_{\lambda=1}^{k} P(x_\nu, y_\lambda) \tag{3.59}$$

Wie man sieht, sind die Randverteilungen einer diskreten zweidimensionalen Verteilung ebenfalls diskret.

Ist die gegebene zweidimensionale Wahrscheinlichkeitsverteilung stetig, so besteht zwischen Dichte- und Summenverteilung die Beziehung 3.48

$$\Phi(x, y) = \int_{\eta=-\infty}^{y} \int_{\xi=-\infty}^{x} \varphi(\xi, \eta)\, \mathrm{d}\xi \mathrm{d}\eta \tag{3.60}$$

Fragt man nach der Wahrscheinlichkeit, mit der x einen Wert $\leq x_i$ annimmt, wobei es gleichgültig ist, welchen Wert y annimmt, so hat dieses Ereignis offenbar die Wahrscheinlichkeit

$$P(x \leq x_i, y \text{ beliebig}) = \int_{\xi=-\infty}^{x_i} \left(\int_{\eta=-\infty}^{+\infty} \varphi(\xi, \eta)\, \mathrm{d}\eta \right) \mathrm{d}\xi \tag{3.61}$$

Definiert man mit

$$\varphi_1(x) = \int_{y=-\infty}^{+\infty} \varphi(x, y)\, \mathrm{d}y \tag{3.62}$$

die **Dichtefunktion der Randverteilung von x** bezüglich der gegebenen stetigen zweidimensionalen Verteilung, so folgt für die Summenfunktion dieser Randverteilung aus 3.61

$$\Phi_1(x) = \int_{\xi=-\infty}^{x} \varphi_1(\xi)\, \mathrm{d}\xi \tag{3.63}$$

Entsprechend ist

$$\varphi_2(y) = \int_{x=-\infty}^{+\infty} \varphi(x, y)\, \mathrm{d}x \tag{3.64}$$

die **Dichtefunktion der Randverteilung von y** bezüglich der gegebenen stetigen zweidimensionalen Verteilung und

$$\Phi_2(y) = \int_{\eta=-\infty}^{y} \varphi_2(\eta)\, \mathrm{d}\eta \tag{3.65}$$

die zugehörige Summenfunktion. Wie man sieht, sind die Randverteilungen einer stetigen zweidimensionalen Verteilung ebenfalls stetig.

Die beiden Zufallsvariablen x und y einer zweidimensionalen Wahrscheinlichkeitsverteilung heißen **unabhängig**, wenn für alle (x, y) gilt

$$\Phi(x, y) = \Phi_1(x)\Phi_2(y) \tag{3.66}$$

Hierin ist $\Phi(x, y)$ die Summenverteilung der zweidimensionalen Wahrscheinlichkeitsverteilung, die entweder diskret nach 3.44 oder stetig nach 3.48 sein kann. $\Phi_1(x)$ und $\Phi_2(y)$ bezeichnen die Summenverteilungen der beiden Randverteilungen. Notwendig und hinreichend für die Unabhängigkeit ist dann entweder im Fall einer diskreten Verteilung

$$P(x_i, y_k) = P_1(x_i)P_2(y_k) \tag{3.67}$$

oder einer stetigen Verteilung

$$\varphi(x, y) = \varphi_1(x)\varphi_2(y) \tag{3.68}$$

Den Beweis führt man durch Einsetzen von 3.67 in 3.44 beziehungsweise von 3.68 in 3.48.

Die Wahrscheinlichkeit dafür, dass eine zweidimensionale Zufallsvariable (x, y) irgendein Wertepaar annimmt, das innerhalb eines beliebig vorgegebenen Bereiches $a_1 < x \leq b_1$, $a_2 < y \leq b_2$ liegt, lässt sich bei Unabhängigkeit der Zufallsvariablen wie folgt berechnen

$$P(a_1 < x \leq b_1,\, a_2 < y \leq b_2) = P_1(a_1 < x \leq b_1)P_2(a_2 < y \leq b_2) \tag{3.69}$$

Gemäß 3.32 gilt

$$P_1(a_1 < x \leq b_1) = \Phi(b_1) - \Phi(a_1)$$

$$P_2(a_2 < y \leq b_2) = \Phi(b_2) - \Phi(a_2)$$

Damit folgt aus 3.69

$$P(a_1 < x \leq b_1,\, a_2 < y \leq b_2) = \{\Phi_1(b_1) - \Phi_1(a_1)\}\{\Phi_2(b_2) - \Phi_2(a_2)\} \tag{3.70}$$

Bei Unabhängigkeit der Zufallsvariablen tritt 3.70 anstelle von 3.46. In diesem Fall wird die mathematische Behandlung, wie man sieht, viel einfacher. Unabhängigkeit der Variablen ist jedoch keineswegs die Regel. Sie muss in jedem Fall durch sorgfältige Prüfung nachgewiesen werden, andernfalls drohen schwerwiegende Fehler.

Neben den beiden Randverteilungen kann man jeder zweidimensionalen Verteilung **bedingte Verteilungen** zuordnen. Dies sind Verteilungen der einen Variablen, wenn der Wert der anderen Variablen festgehalten wird. Allgemein gilt

$$\varphi(x, y_0) = \varphi_2(y_0)\ \varphi_1^*(x|y_0) \tag{3.71}$$

Hierin ist $\varphi_1^*(x|y_0)$ die bedingte Verteilung der Variablen x, wenn die Variable y den festen Wert $y = y_0$ annimmt (x vorausgesetzt y). Entsprechend gilt für y vorausgesetzt x

$$\varphi(x_0, y) = \varphi_1(x_0)\ \varphi_2^*(y|x_0) \tag{3.72}$$

Randverteilungen und bedingte Verteilungen sind eindimensionale Verteilungen. Sie unterscheiden sich jedoch in einem Punkt grundsätzlich. Während im Fall der Randverteilungen nur nach dem Wert der einen Variablen gefragt wird und die andere Variable jeden beliebigen Wert annehmen darf, wird im Fall der bedingten Verteilungen der Wert der anderen Variablen festgehalten.

3.4 Erwartungswerte und Varianzen

3.4.1 Eindimensionale Wahrscheinlichkeitsverteilungen

Gegeben sei die Wahrscheinlichkeitsverteilung einer diskreten Zufallsvariablen (wie 3.28 auf Seite 33)

$$P(x) = \begin{cases} P(x_i) & \text{für} \quad x = x_i \\ 0 & \text{für} \quad \text{alle übrigen } x \end{cases} \tag{3.73}$$

Ferner sei $g(x)$ eine für alle n möglichen Werte von x definierte reellwertige Funktion. Dann bezeichnet man den Ausdruck

$$E(g(x)) = \sum_{\nu=1}^{n} g(x_\nu)P(x_\nu) \tag{3.74}$$

als den **mathematischen Erwartungswert** (E: mathematical expectation, expected value) der Funktion $g(x)$. Entsprechend gilt für eine stetige Wahrscheinlichkeitsverteilung mit der Dichtefunktion $\varphi(x)$

$$E(g(x)) = \int_{x=-\infty}^{+\infty} g(x)\varphi(x)\mathrm{d}x \tag{3.75}$$

Wie man aus 3.74 und 3.75 abliest, ist der Erwartungswert einer Funktion $g(x)$ ein Festwert und nicht etwa noch eine Funktion der Zufallsvariablen x. Setzt man speziell

$$g(x) = x^k \qquad \text{mit} \qquad k = 1, 2, 3, \ldots \tag{3.76}$$

so erhält man damit aus 3.74 und 3.75

$$E(x^k) = \sum_{\nu=1}^{n} x_\nu^k P(x_\nu) = M_k \tag{3.77}$$

$$E(x^k) = \int_{x=-\infty}^{+\infty} x^k\varphi(x)\mathrm{d}x = M_k \tag{3.78}$$

Man bezeichnet $E(x^k)$ als das **k-te Moment** M_k (E: k-th moment) der betreffenden Verteilung oder der betreffenden Zufallsvariablen. Allgemein ist ein Moment die Summe oder das Integral über dem Produkt einer Funktion und einer Potenz der Variablen. Der Begriff ist mit dem Moment der Mechanik verwandt. Umfasst die Summation beziehungsweise Integration den gesamten Bereich von x, spricht man von vollständigen Momenten im Gegensatz zu unvollständigen Momenten, die nur einen Teilbereich von x umfassen, dessen Grenzen anzugeben sind.

Das erste Moment (Erwartungswert von x)

$$E(x) = \sum_{\nu=1}^{n} x_\nu P(x_\nu) \tag{3.79}$$

$$E(x) = \int_{x=-\infty}^{+\infty} x\varphi(x)\mathrm{d}x \tag{3.80}$$

hat eine besondere Bedeutung. Man bezeichnet es als **Mittelwert der Verteilung** beziehungsweise der Zufallsvariablen x, wofür auch die Schreibweise

$$\mu_x = E(x) \tag{3.81}$$

gebräuchlich ist.

Beispiel 3.5 : Für die Augenzahl beim Werfen eines idealen Würfels x gilt

$$P(x) = \begin{cases} 1/6 & \text{für} \quad x = 1, 2, 3, \ldots 6 \\ 0 & \text{für} \quad \text{alle übrigen } x \end{cases}$$

Der Erwartungswert von x errechnet sich mit 3.79 zu

$$E(x) = \sum_{\nu=1}^{n} x_\nu \, P(x_\nu) = \sum_{\nu=1}^{6} \frac{\nu}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3,5$$

Der Erwartungswert $E(x)$ ist nicht notwendig ein Wert, den die Variable x auch tatsächlich annehmen kann, wie das Beispiel verdeutlicht.

Ende Beispiel

Setzt man anstelle 3.76

$$g(x) = (x - \mu_x)^k \qquad \text{mit} \qquad k = 1, 2, 3, \ldots \tag{3.82}$$

so erhält man aus 3.74 und 3.75

$$E\left((x - \mu_x)^k\right) = \sum_{\nu=1}^{n} (x_\nu - \mu_x)^k \, P(x_\nu) \tag{3.83}$$

$$E\left((x - \mu_x)^k\right) = \int_{x=-\infty}^{+\infty} (x - \mu_x)^k \, \varphi x \mathrm{d}x \tag{3.84}$$

Man bezeichnet $E\left((x - \mu_x)^k\right)$ als das **k-te zentrale Moment** der betreffenden Verteilung oder Zufallsvariablen.

Das erste zentrale Moment ist stets gleich null. Aus

$$E(x - \mu_x) = \sum_{\nu=1}^{n} (x_\nu - \mu_x) \, P(x_\nu)$$

folgt nämlich

$$E(x - \mu_x) = \sum_{\nu=1}^{n} x_\nu \, P(x_\nu) - \mu_x \sum_{\nu=1}^{n} P(x_\nu)$$

Andererseits gilt

$$\sum_{\nu=1}^{n} x_\nu P(x_\nu) = \mu_x \qquad \sum_{\nu=1}^{n} P(x_\nu) = 1$$

Damit ergibt sich

$$E(x - \mu_x) = 0 \tag{3.85}$$

Das zweite zentrale Moment

$$E\left((x-\mu_x)^2\right) = \sum_{\nu=1}^{n} (x_\nu - \mu_x)^2 \, P(x_\nu) \tag{3.86}$$

$$E\left((x-\mu_x)^2\right) = \int_{x=-\infty}^{+\infty} (x-\mu_x)^2 \, \varphi(x) \mathrm{d}x \tag{3.87}$$

hat wiederum eine besondere Bedeutung. Man verwendet es zur Kennzeichnung der Breite einer Verteilung und bezeichnet es als die **Varianz einer Verteilung** beziehungsweise der Variablen x. Es gilt die Schreibweise

$$\sigma_x^2 = E\left((x-\mu_x)^2\right) = \mathrm{Var}(x) \tag{3.88}$$

Das dritte zentrale Moment wird bisweilen ebenfalls zur Kennzeichnung einer Verteilung herangezogen. Man bildet eine Größe

$$\gamma = \frac{1}{\sigma_x^3} E\left((x-\mu_x)^3\right) \tag{3.89}$$

die man als die **Schiefe der Verteilung** (E: skewness) bezeichnet. Sie ist ein Maß für die Asymmetrie der Verteilung. Für symmetrische Verteilungen verschwindet die Schiefe, wie man im Fall einer stetigen Zufallsvariablen x leicht zeigen kann. Aus 3.84 erhält man für $k = 3$

$$E\left((x-\mu_x)^3\right) = \int_{x=-\infty}^{+\infty} (x-\mu_x)^3 \, \varphi(x) \mathrm{d}x$$

Wir substitutieren

$$\begin{array}{llll} u = x - \mu_x & \mathrm{d}u = \mathrm{d}x & \text{für} & x > \mu_x \\ v = \mu_x - x & \mathrm{d}v = -\mathrm{d}x & \text{für} & x < \mu_x \end{array} \tag{3.90}$$

Damit folgt

$$\begin{aligned} E\left((x-\mu_x)^3\right) &= \int_{v=+\infty}^{0} v^3 \, \varphi(\mu_x - v) \, \mathrm{d}v + \int_{u=0}^{\infty} u^3 \, \varphi(\mu_x + u) \, \mathrm{d}u \\ &= \int_{u=0}^{\infty} u^3 \, \varphi(\mu_x + u) \, \mathrm{d}u - \int_{v=0}^{\infty} v^3 \, \varphi(\mu_x - v) \, \mathrm{d}v \end{aligned} \tag{3.91}$$

Nach Voraussetzung ist $\varphi(x)$ symmetrisch bezüglich des Mittelwertes μ_x. Daher gilt

$$\varphi(\mu_x + u) = \varphi(\mu_x - u)$$

Durch Einsetzen folgt

$$E\left((x-\mu_x)^3\right) = 0 \tag{3.92}$$

Aus dem vierten zentralen Moment ergibt sich eine Größe, die als **Wölbung** (E: kurtosis) bezeichnet wird. Die Normalverteilung hat die Wölbung null. Verteilungen mit positiver Wölbung sind steilgipflig oder supergaußförmig, spitzer als die Normalverteilung. Schiefe und Wölbung werden jedoch nur selten verwendet.

Die zentralen Momente lassen sich durch die Momente $E(x)$, $E(x^2)$, $E(x^3)$ usw. ausdrücken. Aus 3.88 folgt unter Verwendung von 3.86 oder 3.87

$$\begin{aligned}\sigma_x^2 &= E\left((x-\mu_x)^2\right)\\ &= E\left(x^2-2\mu_x x+\mu_x^2\right)\\ &= E\left(x^2\right)-2\mu_x E\left(x\right)+\mu_x^2 \qquad (3.93)\end{aligned}$$

oder

$$\sigma_x^2 = E\left(x^2\right)-\left(E\left(x\right)\right)^2 \qquad (3.94)$$

Entsprechend beweist man

$$E\left((x-\mu_x)^3\right) = E\left(x^3\right)-3\,E\left(x\right)E\left(x^2\right)+2\left(E\left(x\right)\right)^3 \qquad (3.95)$$

Beispiel 3.6 : Wir wollen die Wahrscheinlichkeitsverteilung und den Erwartungswert einer Zufallsvariablen x bestimmen, die wie folgt definiert ist: x sei die Anzahl der Würfe mit einem idealen Würfel bis zum Erreichen der ersten 6. Zunächst definieren wir folgende Ereignisse

- A = Eintreffen der 6 bei einmaligem Würfeln
- A_1 = Eintreffen der 6 beim ersten Wurf
- A_2 = Eintreffen der 6 beim zweiten Wurf
- ...
- A_n = Eintreffen der 6 beim n-ten Wurf

Die Wahrscheinlichkeit des Ereignisses A ist bekannt

$$P(\mathrm{A}) = \frac{1}{6} \qquad P(\bar{\mathrm{A}}) = \frac{5}{6}$$

Die Zahl n, das heißt die Anzahl der Würfe bis zum Erreichen der ersten 6, ist offenbar nicht beschränkt. Die Variable x ist zwar diskret, nimmt aber abzählbar unendlich viele Werte an. Die Wahrscheinlichkeitsverteilung berechnet man wie folgt

$$\text{Für } x_1 = 1 \text{ gilt} \quad P(\mathrm{A}_1) = P(\mathrm{A}) = \frac{1}{6} = P(x_1)$$

$$\text{für } x_2 = 2 \text{ gilt} \quad P(\mathrm{A}_2) = P(\bar{\mathrm{A}}\mathrm{A}) = P(\bar{\mathrm{A}})\,P(\mathrm{A}) = \frac{5}{6}\,\frac{1}{6} = P(x_2)$$

$$\text{für } x_3 = 3 \text{ gilt} \quad P(\mathrm{A}_3) = P(\bar{\mathrm{A}}\bar{\mathrm{A}}\mathrm{A}) = P(\bar{\mathrm{A}})\,P(\bar{\mathrm{A}})\,P(\mathrm{A}) = \left(\frac{5}{6}\right)^2 \frac{1}{6} = P(x_3)$$

$$\ldots$$

$$\text{für } x_n = n \text{ gilt} \quad P(\mathrm{A}_n) = [P(\bar{\mathrm{A}})]^{n-1}\,P(\mathrm{A}) = \left(\frac{5}{6}\right)^{n-1} \frac{1}{6} = P(x_n)$$

Die Summe der Wahrscheinlichkeiten aller möglichen Werte, welche die Zufallsvariable x annehmen kann, muss nach 3.30 auf Seite 33 den Wert eins ergeben

$$\sum_{\nu=1}^{n} P(x_\nu) = 1$$

Dass diese Forderung erfüllt ist, lässt sich wie folgt zeigen: Setzt man die berechneten Wahrscheinlichkeiten in 3.30 ein, so erhält man

$$\sum_{\nu=1}^{n} P(x_\nu) = \sum_{n=1}^{\infty} \frac{1}{6} \left(\frac{5}{6}\right)^{n-1} = \frac{1}{6} + \frac{1}{6}\left(\frac{5}{6}\right) + \frac{1}{6}\left(\frac{5}{6}\right)^2 + \ldots$$

Das ist offenbar eine unendliche geometrische Reihe. Allgemein gilt für die Summe einer unendlichen geometrischen Reihe

$$a + aq + aq^2 + \ldots = \frac{a}{1-q} \qquad \text{falls} \qquad q < 1$$

Im vorliegenden Fall ist

$$a = \frac{1}{6} \qquad q = \frac{5}{6}$$

Daraus folgt

$$\sum_{\nu=1}^{\infty} P(x_\nu) = \frac{a}{1-q} = \frac{\frac{1}{6}}{1 - \frac{5}{6}} = 1$$

Für den Erwartungswert von x gilt allgemein (siehe 3.79 auf Seite 44)

$$E(x) = \sum_{\nu=1}^{n} x_\nu\, P(x_\nu)$$

Mit den berechneten Wahrscheinlichkeiten ergibt sich

$$E(x) = \sum_{n=1}^{\infty} n \frac{1}{6} \left(\frac{5}{6}\right)^{n-1}$$

Allgemein gilt für eine geometrische Reihe, dass man sie gliedweise differenzieren darf. Aus

$$\sum_{n=0}^{\infty} a\, q^n = \frac{a}{1-q} \qquad \text{falls} \qquad q < 1$$

folgt deshalb

$$\frac{\partial}{\partial q} \sum_{n=0}^{\infty} a\, q^n = \sum_{n=1}^{\infty} n\, a\, q^{n-1} = \frac{a}{(1-q)^2}$$

Damit ergibt sich der gesuchte Erwartungswert zu

$$E(x) = \frac{\frac{1}{6}}{(1-\frac{5}{6})^2} = 6$$

Das bedeutet, dass wir im Mittel sechsmal würfeln müssen, bis wir zum ersten Mal eine 6 erhalten.

Ende Beispiel

3.4.2 Mehrdimensionale Wahrscheinlichkeitsverteilungen

Gegeben sei die Wahrscheinlichkeitsverteilung einer diskreten zweidimensionalen Zufallsvariablen

$$P(x,y) = \begin{cases} P(x_i, y_k) & \text{für} \quad x = x_i,\ y = y_k \\ 0 & \text{für} \quad \text{alle übrigen } (x,y) \end{cases} \tag{3.96}$$

Ist $g(x,y)$ eine für alle möglichen Werte von x und y definierte reellwertige Funktion, so bezeichnet man

$$E(g(x,y)) = \sum_{\nu} \sum_{\lambda} g(x_\nu, y_\lambda)\, P(x_\nu, y_\lambda) \tag{3.97}$$

als den mathematischen Erwartungswert der Funktion $g(x,y)$. Entsprechend gilt für eine stetige Wahrscheinlichkeitsverteilung mit der Dichtefunktion $\varphi(x,y)$

$$E(g(x,y)) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} g(x,y)\, \varphi(x,y)\, \mathrm{d}x\mathrm{d}y \tag{3.98}$$

Hängt die Funktion nur von einer Variablen ab, zum Beispiel von x, so gilt für eine diskrete Verteilung

$$E(g(x)) = \sum_{\nu} g(x_\nu)\, P_1(x_\nu) \tag{3.99}$$

und für eine stetige Verteilung

$$E(g(x)) = \int_{x=-\infty}^{+\infty} g(x)\, \varphi_1(x)\, \mathrm{d}x \tag{3.100}$$

wobei $P_1(x)$ die durch 3.54 und 3.55 definierte Randverteilung von x bezüglich einer gegebenen diskreten zweidimensionalen Wahrscheinlichkeitsverteilung $P(x,y)$ ist und $\varphi_1(x)$ die durch 3.62 definierte Dichtefunktion der Randverteilung von x bezüglich einer gegebenen stetigen zweidimensionalen Dichteverteilung $\varphi(x,y)$.

Man erkennt, dass die Definition des Erwartungswertes gemäß 3.97 und 3.98 mit der Definition im eindimensionalen Fall in Einklang steht.

Setzt man für die Funktion $g(x, y)$ speziell

$$g(x, y) = x^k\, y^l \quad k = 0, 1, 2, 3, \ldots \quad l = 0, 1, 2, 3, \ldots \tag{3.101}$$

so folgt aus 3.97 und 3.98

$$E(x^k y^l) = \sum_\nu \sum_\lambda x_\nu^k y_\lambda^l\, P(x_\nu, y_\lambda) \tag{3.102}$$

$$E(x^k y^l) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} x^k y^l\, \varphi(x, y)\, \mathrm{d}x\mathrm{d}y \tag{3.103}$$

Man bezeichnet die $E(x^k y^l)$ wiederum als Momente der betreffenden Verteilung.

Die Momente erster Ordnung erhält man, wenn man entweder $k = 1, l = 0$ oder $k = 0, l = 1$ setzt. Im ersten Fall ergeben sich mit 3.99 und 3.100

$$E(x) = \sum_\nu \sum_\lambda x_\nu\, P(x_\nu, y_\lambda) = \sum_\nu x_\nu\, P_1(x_\nu) \tag{3.104}$$

$$E(x) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} x\, \varphi(x, y)\, \mathrm{d}x\mathrm{d}y = \int_{x=-\infty}^{+\infty} x\, \varphi_1(x)\, \mathrm{d}x \tag{3.105}$$

Im zweiten Fall folgt

$$E(y) = \sum_\nu \sum_\lambda y_\lambda\, P(x_\nu, y_\lambda) = \sum_\lambda y_\lambda\, P_2(y_\lambda) \tag{3.106}$$

$$E(y) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} y\, \varphi(x, y)\, \mathrm{d}x\mathrm{d}y = \int_{y=-\infty}^{+\infty} y\, \varphi_2(y)\, \mathrm{d}y \tag{3.107}$$

Man bezeichnet $E(x)$ und $E(y)$ auch als die **Mittelwerte** der beiden Variablen x und y. Es gilt die Schreibweise

$$\mu_x = E(x) \qquad \mu_y = E(y) \tag{3.108}$$

Setzt man für die Funktion $g(x, y)$

$$g(x, y) = (x - \mu_x)^k\, (y - \mu_y)^l \tag{3.109}$$

so erhält man die zentralen Momente der betreffenden Verteilung.

Die zentralen Momente 1. Ordnung verschwinden wie im eindimensionalen Fall. Es ist

$$E((x - \mu_x)) = 0 \qquad E((y - \mu_y)) = 0 \tag{3.110}$$

Die zentralen Momente 2. Ordnung haben wie im eindimensionalen Fall eine besondere Bedeutung. Man erhält sie, wenn man entweder $k = 2, l = 0$ oder $k = 0, l = 2$ oder $k = 1, l = 1$ setzt. Im ersten Fall ergibt sich mit 3.99 und 3.100

$$E((x - \mu_x)^2) = \sum_\nu (x_\nu - \mu_x)^2\, P_1(x_\nu) \tag{3.111}$$

$$E((x-\mu_x)^2) = \int_{x=-\infty}^{+\infty} (x-\mu_x)^2\,\varphi_1(x)\mathrm{d}x \tag{3.112}$$

Im zweiten Fall folgt

$$E((y-\mu_y)^2) = \sum_{\lambda} (y_\lambda-\mu_y)^2\,P_2(y_\lambda) \tag{3.113}$$

$$E((y-\mu_y)^2) = \int_{y=-\infty}^{+\infty} (y-\mu_y)^2\,\varphi_2(y)\mathrm{d}y \tag{3.114}$$

Man bezeichnet $E((x-\mu_x)^2)$ und $E((y-\mu_y)^2)$ als die **Varianzen** der beiden Variablen x und y. Es gilt die Schreibweise

$$\sigma_x^2 = E((x-\mu_x)^2) \qquad \sigma_y^2 = E((y-\mu_y)^2) \tag{3.115}$$

Im dritten Fall erhält man ein zentrales Moment der Form

$$E((x-\mu_x)(y-\mu_y)) = \sum_{\nu}\sum_{\lambda} (x_\nu-\mu_x)(y_\lambda-\mu_y)\,P(x_\nu,y_\lambda) \tag{3.116}$$

$$E((x-\mu_x)(y-\mu_y)) = \int_{y=-\infty}^{+\infty}\int_{x=-\infty}^{+\infty} (x-\mu_x)(y-\mu_y)\,\varphi(x,y)\mathrm{d}x\mathrm{d}y \tag{3.117}$$

Man bezeichnet $E((x-\mu_x)(y-\mu_y))$ als die **Kovarianz** (E: covariance) der Zufallsvariablen x und y. Die abgekürzte Schreibweise lautet

$$\sigma_{xy} = E((x-\mu_x)(y-\mu_y)) = \mathrm{Cov}(x,y) \tag{3.118}$$

Im Gegensatz zur Varianz kann die Kovarianz auch negative Werte annehmen. Eine positive Kovarianz bedeutet einen gleichsinnig linearen Zusammenhang zwischen x und y, eine negative einen gegensinnig linearen Zusammenhang. Sind x und y unabhängig voneinander, so ist σ_{xy} gleich null, wie man leicht zeigen kann. Der Zahlenwert der Kovarianz ist von den gewählten Maßeinheiten abhängig und besagt wenig. Erst der als **Korrelationskoeffizient** (E: correlation coefficient) der Grundgesamtheit bezeichnete Quotient

$$\rho = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y} = \mathrm{Kor}(x,y) \tag{3.119}$$

erlaubt Zusammenhänge zu vergleichen. Der Koeffizient geht auf Karl Pearson (1857–1936) zurück. Wie in Abschnitt 8.3.1 *Korrelationskoeffizient* auf Seite 142 gezeigt wird, gilt stets

$$-1 \le \rho \le 1 \tag{3.120}$$

Sind die Variablen x und y unabhängig voneinander, ist $\rho = 0$. Im Fall $|\rho| = 1$ besteht zwischen x und y eine gleich- beziehungsweise gegensinnig lineare Abhängigkeit.

3.4.3 Rechenregeln für Erwartungswerte und Varianzen

Eindimensionale Zufallsgrößen

Durch eine eindeutige, reellwertige, für alle möglichen Werte von x definierte Funktion

$$v = g(x) \tag{3.121}$$

wird jedem Wert x_i der Zufallsvariablen x ein Wert v_i einer neuen Zufallsvariablen v zugeordnet. Zu einer gegebenen Wahrscheinlichkeitsverteilung der Variablen x lässt sich mit 3.121 die Wahrscheinlichkeitsverteilung der Variablen v berechnen. Für Mittelwert und Varianz der neuen Variablen v gilt

$$\mu_v = E(v) = E(g(x)) \tag{3.122}$$

$$\sigma_v^2 = E((v - \mu_v)^2) = E((g(x) - \mu_v)^2) \tag{3.123}$$

Ist die Funktion $g(x)$ linear, so ist

$$v = g(x) = a + bx \tag{3.124}$$

Aus den Definitionsgleichungen 3.74 und 3.75 für $E(g(x))$ auf Seite 44 folgt direkt, dass stets – unabhängig davon, ob x diskret oder stetig ist – gilt

$$E(a + bx) = a + b\,E(x) \qquad \mu_v = a + b\,\mu_x \tag{3.125}$$

Für die Varianz von v erhält man aus den gleichen Gründen den Zusammenhang

$$\begin{aligned} E((g(x) - \mu_v)^2) &= E((a + bx - a - b\mu_x)^2) \\ &= b^2\,E((x - \mu_x)^2) \end{aligned} \tag{3.126}$$

oder kürzer

$$\sigma_v^2 = b\,\sigma_x^2 \tag{3.127}$$

Ist die Funktion $g(x)$ speziell von der Form

$$v = g(x) = \frac{x - \mu_x}{\sigma_x} \tag{3.128}$$

so ergeben sich Mittelwert und Varianz zu

$$\mu_v = 0 \qquad \sigma_v^2 = 1 \tag{3.129}$$

Man bezeichnet 3.128 auch als die Normal- oder Standardform der Variablen x, ebenso die sich aus der Transformation ergebende Verteilung.

Ist die Funktion $g(x)$ von beliebiger Form, so erhält man Erwartungswert und Varianz nur durch Ausführung der durch 3.122 und 3.123 festgelegten Summationen beziehungsweise Integrationen.

Beispiel 3.7 : Die Projektionsfläche A eines Kreiszylinders vom Radius R und der Höhe h, wie in Abbildung 3.6 dargestellt, ist eine Funktion des Winkels ϑ. Es gilt

$$A = 2R\,h\,cos\vartheta + R^2\,\pi\,sin\vartheta \tag{3.130}$$

Gesucht sei der Mittelwert von A unter der Voraussetzung, dass die Orientierung der Zylinderachse rein zufällig ist, das heißt, dass alle Raumrichtungen gleich wahrscheinlich sind.

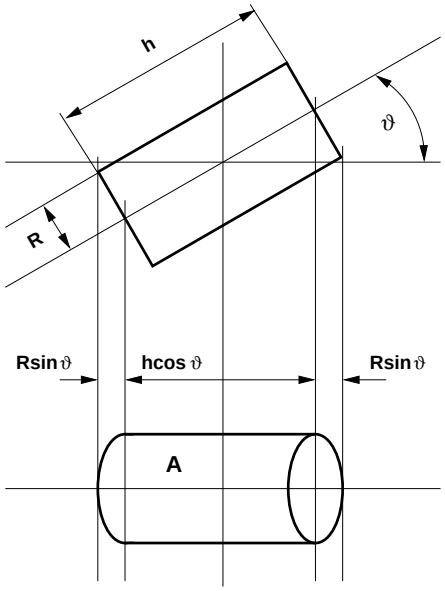


Abb. 3.6: Projektionsfläche eines Kreiszylinders

Die Wahrscheinlichkeitsverteilung des Winkels ϑ ergibt sich mit Abbildung 3.7 wie folgt. Wenn die Orientierung der Zylinderachse rein zufällig ist, sind ihre Durchstoßpunkte durch eine Halbkugel auf deren Oberfläche gleichverteilt. Die Oberfläche einer Halbkugel ist

$$F = 2R_0^2\pi \tag{3.131}$$

Das Flächendifferential auf der Halbkugel ergibt sich zu

$$\begin{aligned} \mathrm{d}F &= R_0^2\,\cos\vartheta\,\mathrm{d}\vartheta\mathrm{d}\psi \\ &= R_0\,\cos\vartheta\,\mathrm{d}\psi\,R_0\,\mathrm{d}\vartheta \end{aligned} \tag{3.132}$$

Die Wahrscheinlichkeit dafür, dass der Durchstoßpunkt der Zylinderachse auf dem Teil der Kugeloberfläche liegt, für den $\vartheta \leq \vartheta_0$ gilt, beträgt

$$\begin{aligned} \Phi(\vartheta_0) &= P(\vartheta \leq \vartheta_0) = \int_0^{2\pi}\int_0^{\vartheta_0}\frac{\mathrm{d}F}{F} \\ &= \int_0^{2\pi}\left(\frac{1}{2\pi}\int_0^{\vartheta_0}\cos\vartheta\,\mathrm{d}\vartheta\right)\mathrm{d}\psi \\ &= \int_0^{\vartheta_0}\cos\vartheta\,\mathrm{d}\vartheta = \sin\vartheta_0 \end{aligned} \tag{3.133}$$

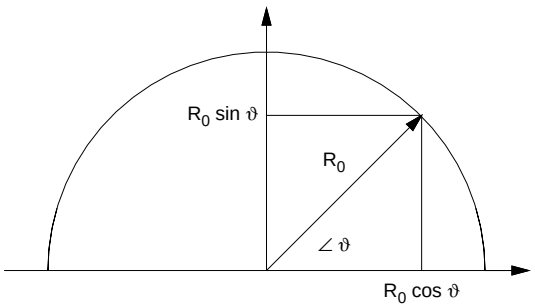


Abb. 3.7: Zur Berechnung der Wahrscheinlichkeitsverteilung des Winkels ϑ

Für die Dichtefunktion von ϑ folgt daraus

$$\varphi(\vartheta) = \frac{\mathrm{d}\Phi(\vartheta)}{\mathrm{d}\vartheta} = \cos\vartheta \tag{3.134}$$

Den gesuchten Mittelwert von A erhält man damit zu

$$\begin{aligned} E(A) &= E(2Rh\,\cos\vartheta + R^2\pi\,\sin\vartheta) \\ &= \int_0^{\pi/2} (2Rh\,\cos\vartheta + R^2\pi\sin\vartheta)\,\varphi(\vartheta)\,\mathrm{d}\vartheta \\ &= 2Rh\int_0^{\pi/2}\cos^2\vartheta\,\mathrm{d}\vartheta + R^2\pi\int_0^{\pi/2}\sin\vartheta\,\cos\vartheta\,\mathrm{d}\vartheta \\ &= 2Rh\,\frac{\pi}{4} + R^2\pi\,\frac{1}{2} \\ &= \frac{1}{4}\,(2R\pi h + 2R^2\pi) = \mu_A \end{aligned} \tag{3.135}$$

Der alternative Weg, zunächst Berechnung der Verteilungsfunktion von A und daraus dann $E(A)$, wäre hier schwieriger und aufwendiger. Das Ergebnis lässt sich mit S als der Oberfläche des Zylinders auch so schreiben

$$E(A) = \frac{S}{4} \tag{3.136}$$

Dies folgt auch aus dem Theorem von Augustin Louis Cauchy (1789–1857), das allgemein für konvexe Körper gilt.[2]

Außer dem Mittelwert $E(A)$ kann auch die Varianz σ_A^2 exakt berechnet werden. Es gilt

$$\sigma_A^2 = E((A - E(A))^2) = E(A^2) - (E(A))^2 \tag{3.137}$$

[2]siehe die deutsche Wikipedia unter *Cauchy-Theorem* oder die englische Wikipedia unter *Cauchy's theorem (geometry)*. Cauchys Oberflächenformel gehört in die *Konvexgeometrie* oder allgemeiner in die *Theorie konvexer Mengen*.

Für $E(A^2)$ erhält man

$$\begin{aligned} E(A^2) &= \int_0^{\pi/2} (2Rh\,\cos\vartheta + R^2\pi\,\sin\vartheta)^2\,\cos\vartheta\,\mathrm{d}\vartheta \\ &= 4R^2h^2\int_0^{\pi/2}\cos^2\vartheta\,\mathrm{d}\vartheta + 4R^3h\pi\int_0^{\pi\ 2}\cos^2\vartheta\,\sin\vartheta\,\mathrm{d}\vartheta \\ &\quad + R^4\pi^2\int_0^{\pi/2}\sin^2\vartheta\,\cos\vartheta\,\mathrm{d}\vartheta \\ &= 4R^2h^2\,\frac{2}{3} + 4R^3h\pi\,\frac{1}{3} + R^4\pi^2\,\frac{1}{3} \end{aligned} \tag{3.138}$$

Damit folgt für σ_A^2

$$\begin{aligned} \sigma_A^2 &= \frac{8}{3}\,R^2\,h^2 + \frac{4}{3}\,R^3\,h\,\pi + \frac{1}{3}\,R^4\,\pi^2 - (\frac{1}{2}\,Rh\pi + \frac{1}{2}\,R^2\,\pi)^2 \\ &= (\frac{8}{3} - \frac{\pi^2}{4})\,R^2h^2 + (\frac{4}{3} - \frac{\pi}{2})\,R^3h\pi + \frac{1}{12}\,R^4\,\pi^2 \end{aligned} \tag{3.139}$$

Ende Beispiel

Im Fall der Projektionsfläche eines Kreiszylinders ist es möglich, die in den Definitionsgleichungen 3.122 und 3.123 enthaltenen Integrale exakt zu lösen. Dies ist jedoch keineswegs immer so.

Näherungslösungen erhält man, indem man die Funktion $g(x)$ in eine Taylorreihe (benannt nach Brook Taylor, 1685–1731) um den Erwartungswert $\mu_x = E(x)$ entwickelt. Es ist

$$g(x) = g(\mu_x) + g'(\mu_x)(x-\mu_x) + \frac{1}{2!}\,g''(\mu_x)(x-\mu_x)^2 + \frac{1}{3!}\,g'''(\mu_x)(x-\mu_x)^3 + \ldots \tag{3.140}$$

Weiterhin wird angenommen, dass die Reihe beständig konvergent ist, das heißt für alle endlichen Werte von x konvergiert. Bildet man von 3.140 den Erwartungswert, so folgt

$$\begin{aligned} E(g(x)) &= g(\mu_x) + g'(\mu_x)\,E((x-\mu_x)) + \frac{1}{2!}\,g''(\mu_x)\,E((x-\mu_x)^2) \\ &\quad + \frac{1}{3!}\,g'''(\mu_x)\,E((x-\mu_x)^3) + \ldots \end{aligned} \tag{3.141}$$

Ist die Reihe 3.140 beständig konvergent, so konvergiert auch die Reihe 3.141. Bricht man sie nach dem 3. Glied ab, so gilt

$$\mu_v = E(g(x)) \approx g(\mu_x) + \frac{1}{2!}\,g''(\mu_x)\,\sigma_x^2 \tag{3.142}$$

Aus 3.142 folgt, dass man den Mittelwert von v nicht exakt, sondern nur näherungsweise erhält, wenn man in 3.121 die Variable x durch μ_x ersetzt. Den Fehler, den man dabei macht, kann man mit 3.142 abschätzen.

Beispiel 3.8 : Eine Näherungslösung für den Erwartungswert der Projektionsfläche eines Kreiszylinders ist nicht erforderlich, weil mit 3.135

$$E(A) = \frac{1}{4}\left(2R\pi h + 2R^2\pi\right)$$

die exakte Lösung leicht zu finden war. Interessant ist jedoch die Frage, wie genau die Näherung unter Verwendung von 3.142 in diesem Fall sein würde. Es gilt

$$E(A) \approx A(\mu_\vartheta) + \frac{1}{2}\left(\frac{\partial^2 A}{\partial \vartheta^2}\right)_{\mu\vartheta} \sigma_\vartheta^2 \tag{3.143}$$

Die Dichtefunktion von ϑ ist

$$\varphi(\vartheta) = \cos\vartheta \tag{3.144}$$

Ihren Mittelwert und ihre Varianz erhalten wir wie folgt

$$\begin{aligned} E(\vartheta) &= \int_0^{\pi/2} \vartheta\, \cos\vartheta\, \mathrm{d}\vartheta = \Big|\cos\vartheta + \vartheta\, \sin\vartheta\Big|_0^{\pi/2} \\ &= \frac{\pi}{2} - 1 = \mu_\vartheta \end{aligned} \tag{3.145}$$

$$\begin{aligned} E(\vartheta^2) &= \int_0^{\pi/2} \vartheta^2\, \cos\vartheta\, \mathrm{d}\vartheta = \Big|2\vartheta\cos\vartheta + (\vartheta^2 - 2)\, \sin\vartheta\Big|_0^{\pi/2} \\ &= \frac{\pi^2}{4} - 2 \end{aligned} \tag{3.146}$$

$$\sigma_\vartheta^2 = E(\vartheta^2) - (E(\vartheta))^2 = \frac{\pi^2}{4} - 2 - \frac{\pi^2}{4} + \pi - 1 = \pi - 3 \tag{3.147}$$

Die Projektionsfläche ist

$$A = 2\,R\,h\,\cos\vartheta + R^2\,\pi\,\sin\vartheta \tag{3.148}$$

Daraus folgt

$$\frac{\partial A}{\partial \vartheta} = -2\,R\,h\,\sin\vartheta + R^2\,\pi\,\cos\vartheta \tag{3.149}$$

$$\frac{\partial^2 A}{\partial \vartheta^2} = -2\,R\,h\,\cos\vartheta - R^2\,\pi\,\sin\vartheta \tag{3.150}$$

$$\begin{aligned} E(A) &\approx 2Rh\cos\mu_\vartheta + R^2\pi\cos\mu_\vartheta - \frac{1}{2}(2Rh\cos\mu_\vartheta + R^2\pi\sin\mu_\vartheta)\sigma_\vartheta^2 \\ &\approx 2Rh\cos\mu_\vartheta(1 - \frac{1}{2}\sigma_\vartheta^2) + R^2\pi\sin\mu_\vartheta(1 - \frac{1}{2}\sigma_\vartheta^2) \\ &\approx \frac{1}{4}\left(2R\pi h\left(\cos\mu_\vartheta\frac{4 - 2\sigma_\vartheta^2}{\pi}\right) + 2R^2\pi\left(\sin\mu_\vartheta\left(2 - \sigma_\vartheta^2\right)\right)\right) \\ &\approx \frac{1}{4}\left(2R\pi h \cdot 0,9955 + 2R^2\pi \cdot 1,0041\right) \end{aligned} \tag{3.151}$$

Wie man im Vergleich zu 3.135 sieht, ist die Näherung in diesem Fall außerordentlich gut. Verkürzt man die Näherung auf

$$E(A) \approx A(\mu_\vartheta)$$

so erhält man

$$E(A) \approx \frac{1}{4}\left(2R\pi h \cdot 1,0714 + 2R^2\pi \cdot 1,0806\right)$$

Ende Beispiel

Zur näherungsweisen Berechnung der Varianz von $g(x)$ wird zunächst die Differenz von 3.140 und 3.141 auf Seite 55 gebildet. Es ist

$$g(x) - \mu_v = g'(\mu_x)(x-\mu_x) + \frac{1}{2!}g''(\mu_x)\left((x-\mu_x)^2 - \sigma_x^2\right) + \ldots \tag{3.152}$$

Quadriert man 3.152, so folgt

$$\begin{aligned}(g(x)-\mu_v)^2 &= (g'(\mu_x))^2 \cdot (x-\mu_x)^2 + g'(\mu_x)\cdot g''(\mu_x)\cdot(x-\mu_x)^3 \\ &\quad -g'(\mu_x)\cdot g''(\mu_x)\cdot\sigma_x^2\cdot(x-\mu_x) + \frac{1}{4}(g''(\mu_x))^2\cdot(x-\mu_x)^4 \\ &\quad -\frac{1}{2}(g''(\mu_x))^2\cdot\sigma_x^2\cdot(x-\mu_x)^2 + \frac{1}{4}(g''(\mu_x))^2\cdot\sigma_x^4 + \ldots \end{aligned} \tag{3.153}$$

Bildet man von 3.153 den Erwartungswert, so erhält man die Varianz von v. In erster Näherung ist

$$\sigma_v^2 = E\left((g(x)-\mu_v)^2\right) \approx (g'(\mu_x))^2\cdot\sigma_x^2 \tag{3.154}$$

Beispiel 3.9 : Gegeben sei eine Zweikomponenten-Mischung von Partikeln der Massendichte ρ_1 und Partikeln der Massendichte ρ_2. Zwischen dem Volumenanteil x der ersten Komponente in einer Teilmenge (Probe) und ihrem zugehörigen Massenanteil $x^\star$ besteht folgender Zusammenhang

$$x^\star = \frac{x\rho_1}{x\rho_1 + (1-x)\rho_2} \tag{3.155}$$

Der Volumenanteil x soll zufällig schwanken. Der Erwartungswert $E(x) = p$ und die Varianz σ_x^2 seien bekannt. Gesucht sei die Varianz des Massenanteils $x^\star$. Eine exakte Lösung unter Verwendung von 3.123 auf Seite 52 ist in diesem Fall nicht zu haben. Stattdessen soll eine Näherung mit 3.154 versucht werden. Aus 3.155 folgt durch Differenzieren

$$\begin{aligned}\frac{\mathrm{d}x^\star}{\mathrm{d}x} &= \frac{\rho_1\left(x\rho_1 + (1-x)\rho_2\right) - x\rho_1(\rho_1-\rho_2)}{\left(x\rho_1+(1-x)\rho_2\right)^2} \\ &= \frac{\rho_1\rho_2}{\left(x\rho_1+(1-x)\rho_2\right)^2}\end{aligned} \tag{3.156}$$

und damit aus 3.154

$$\sigma_{x^\star}^2 \approx \frac{\rho_1^2 \rho_2^2}{(p\rho_1 + (1-p)\rho_2)^4}\, \sigma_x^2 \tag{3.157}$$

Ende Beispiel

Mehrdimensionale Zufallsgrößen

Analog zum eindimensionalen Fall wird durch

$$v = g(x, y) \tag{3.158}$$

eine eindeutige, reellwertige, für alle möglichen Werte von x und y definierte Funktion eingeführt, die jedem Wertepaar x_i, y_k der zweidimensionalen Zufallsvariablen (x, y) einen Wert v_j der neuen Zufallsvariablen v zuordnet. Für Mittelwert und Varianz der neuen Variablen v gilt

$$\mu_v = E(v) = E\left(g(x, y)\right) \tag{3.159}$$

$$\sigma_v^2 = E\left((v - \mu_v)^2\right) = E\left((g(x, y) - \mu_v)^2\right) \tag{3.160}$$

Ist die Funktion $g(x, y)$ von der Form

$$v = g(x, y) = x + y \tag{3.161}$$

so folgt mit den Definitionsgleichungen 3.97 und 3.98 für $E(g(x, y))$ auf Seite 49 und mit den Gleichungen 3.104, 3.105, 3.106 und 3.107 auf Seite 50 für $E(x)$ und $E(y)$, dass unabhängig davon, ob die zweidimensionale Zufallsvariable (x, y) diskret oder stetig ist, stets gilt

$$\mu_v = E(x + y) = E(x) + E(y) \tag{3.162}$$

Indem man y durch die Summe zweier Zufallsvariable ersetzt, erhält man aus 3.162 die entsprechende Gleichung für drei Zufallsvariable. Durch wiederholte Anwendung dieses Prozesses gelangt man zum **Additionssatz für Mittelwerte**

$$E(x_1 + x_2 + \ldots + x_n) = E(x_1) + E(x_2) + \ldots + E(x_n) \tag{3.163}$$

Der Mittelwert einer Summe von n diskreten oder stetigen Zufallsvariablen $x_1, x_2, \ldots x_n$ ist gleich der Summe ihrer Mittelwerte.

Für die Varianz von 3.161 folgt zunächst aus 3.160

$$\begin{aligned} \sigma_v^2 &= E\left((x + y - \mu_x - \mu_y)^2\right) \\ &= E\left(\left((x - \mu_x) + (y - \mu_y)\right)^2\right) \\ &= E\left((x - \mu_x)^2 + 2(x - \mu_x)(y - \mu_y) + (y - \mu_y)^2\right) \end{aligned} \tag{3.164}$$

Hieraus erhält man unabhängig davon, ob die zweidimensionale Zufallsvariable (x, y) diskret oder stetig ist, den Zusammenhang

$$\sigma_v^2 = E\Big((x-\mu_x)^2\Big) + 2E\Big((x-\mu_x)(y-\mu_y)\Big) + E\Big((y-\mu_y)^2\Big) \tag{3.165}$$

Mit den Definitionsgleichungen 3.115 und 3.118 auf Seite 51 folgt daraus schließlich

$$\sigma_v^2 = \sigma_x^2 + 2\sigma_{xy} + \sigma_y^2 \tag{3.166}$$

Hierin bezeichnet σ_{xy} die Kovarianz der Zufallsvariablen x und y. Sind x und y unabhängig voneinander, so ist die Kovarianz gleich null, und es gilt

$$\sigma_v^2 = \sigma_x^2 + \sigma_y^2 \tag{3.167}$$

Dieses Ergebnis lässt sich wie im Fall der Mittelwerte von zwei auf beliebig viele Zufallsvariable erweitern. Man erhält dann den **Additionssatz für Varianzen**

$$\sigma_v^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2 \tag{3.168}$$

Die Varianz einer Summe voneinander unabhängiger Zufallsvariablen ist gleich der Summe ihrer Varianzen.

Ist die Funktion $g(x, y)$ von der Form

$$v = g(x, y) = x \cdot y \tag{3.169}$$

so verwendet man zur Berechnung des Erwartungswertes den einfachen Zusammenhang

$$x \cdot y = (x-\mu_x)(y-\mu_y) + \mu_x y + \mu_y x - \mu_x \mu_y \tag{3.170}$$

Bildet man hiervon den Erwartungswert, so folgt

$$\mu_v = E(x \cdot y) = E\Big((x-\mu_x)(y-\mu_y)\Big) + \mu_x E(y) + \mu_y E(x) - \mu_x \mu_y \tag{3.171}$$

Mit den Definitionsgleichungen 3.108 auf Seite 50 und 3.118 auf Seite 51 erhält man

$$\mu_v = E(x \cdot y) = \sigma_{xy} + E(x) \cdot E(y) \tag{3.172}$$

Sind x und y unabhängig voneinandner, so ist die Kovarianz σ_{xy} gleich null, und es gilt

$$E(x \cdot y) = E(x) \cdot E(y) \tag{3.173}$$

Dieses Ergebnis lässt sich auf beliebig viele Zufallsvariable erweitern. Man erhält den **Multiplikationssatz für Mittelwerte**

$$E(x_1 \cdot x_2 \cdot \ldots \cdot x_n) = E(x_1) \cdot E(x_2) \cdot \ldots \cdot E(x_n) \tag{3.174}$$

Der Mittelwert des Produktes von n diskreten oder stetigen, voneinander unabhängigen Zufallsvariablen $x_1, x_2 \ldots x_n$ ist gleich dem Produkt ihrer Mittelwerte.

Ist die Funktion $g(x,y)$ von beliebiger Form, so erhält man wie im eindimensionalen Fall Erwartungswert und Varianz nur durch Ausführen der entsprechenden Summationen beziehungsweise Integrationen. Man kann sich jedoch ähnlich wie im eindimensionalen Fall Näherungslösungen verschaffen, indem man $g(x,y)$ in eine zweidimensionale Taylorreihe um die Erwartungswerte $\mu_x = E(x)$ und $\mu_y = E(y)$ entwickelt. Es gilt

$$\begin{aligned} g(x,y) &= g(\mu_x,\mu_y) + g_x(\mu_x,\mu_y)\cdot(x-\mu_x) + g_y(\mu_x,\mu_y)\cdot(y-\mu_y) \\ &\quad + \frac{1}{2!}\Big(g_{xx}(\mu_x,\mu_y)\cdot(x-\mu_x)^2 + g_{yy}(\mu_x,\mu_y)\cdot(y-\mu_y)^2 \\ &\quad + 2g_{xy}(\mu_x,\mu_y)\cdot(x-\mu_x)(y-\mu_y)\Big) + \ldots \end{aligned} \tag{3.175}$$

Wiederum wird vorausgesetzt, dass die Reihe beständig konvergiert. Bildet man hiervon den Erwartungswert, so ist

$$\begin{aligned} E\Big(g(x,y)\Big) &= g(\mu_x,\mu_y) + \frac{1}{2!}\Big(g_{xx}(\mu_x,\mu_y)\cdot E((x-\mu_x)^2) \\ &\quad + g_{yy}(\mu_x,\mu_y)\cdot E((y-\mu_y)^2) \\ &\quad + 2g_{xy}(\mu_x,\mu_y)\cdot E((x-\mu_x)(y-\mu_y))\Big) + \ldots \end{aligned} \tag{3.176}$$

Ist die Reihe 3.175 beständig konvergent, so konvergiert auch 3.176. Dann gilt näherungsweise

$$\begin{aligned} \mu_v = E(g(x,y)) &\approx g(\mu_x,\mu_y) + \frac{1}{2}\Big(g_{xx}(\mu_x,\mu_y)\cdot\sigma_x^2 \\ &\quad + g_{yy}(\mu_x,\mu_y)\cdot\sigma_y^2 + 2g_{xy}(\mu_x,\mu_y)\cdot\sigma_{xy}\Big) \end{aligned}$$

Zur Berechnung der Varianz von $g(x,y)$ wird die Differenz von 3.175 und 3.176 gebildet

$$\begin{aligned} g(x,y) - \mu_v &= g_x(\mu_x,\mu_y)\cdot(x-\mu_x) + g_y(\mu_x,\mu_y)\cdot(y-\mu_y) \\ &\quad + \frac{1}{2!}g_{xx}(\mu_x,\mu_y)\cdot\Big((x-\mu_x)^2 - \sigma_x^2\Big) \\ &\quad + \frac{1}{2!}g_{yy}(\mu_x,\mu_y)\cdot\Big((y-\mu_y)^2 - \sigma_y^2\Big) \\ &\quad + \frac{2}{2!}g_{xy}(\mu_x,\mu_y)\cdot\Big((x-\mu_x)(y-\mu_y) - \sigma_{xy}\Big) + \ldots \end{aligned}$$

Quadriert man diesen Ausdruck und bildet davon den Erwartungswert, so erhält man die Varianz σ_v^2. Als erste Näherung ergibt sich

$$\sigma_v^2 = E\Big((g(x,y)-\mu_v)^2\Big)$$

$$\approx \left(g_x(\mu_x, \mu_y)\right)^2 \cdot \sigma_x^2 + 2g_x(\mu_x, \mu_y) \cdot g_y(\mu_x, \mu y) \cdot \sigma_{xy} + \left(g_y(\mu_x, \mu_y)\right)^2 \cdot \sigma_y^2 \tag{3.177}$$

Sind x und y unabhängig voneinander, so ist die Kovarianz σ_{xy} gleich null, und es folgt aus 3.177

$$\sigma_v^2 \approx \left(g_x(\mu_x, \mu_y)\right)^2 \cdot \sigma_x^2 + \left(g_y(\mu_x, \mu_y)\right)^2 \cdot \sigma_y^2 \tag{3.178}$$

Dies ist das **Fehlerfortpflanzungsgesetz** von Carl Friedrich Gauss (1777–1855). Es kann von zwei Zufallsvariablen x, y auf beliebig viele Zufallsvariable erweitert werden.

Kapitel 4

Diskrete Wahrscheinlichkeitsverteilungen

Das Kapitel erläutert einige Wahrscheinlichkeitsverteilungen eines diskreten Merkmals, die in mathematischer oder technischer Hinsicht interessant sind.

4.1 Diskrete Gleichverteilung

Die diskrete **Gleichverteilung** wurde bereits in Abschnitt 3.3.1 *Verteilungen einer Zufallsvariablen* auf Seite 32 behandelt. Ein Beispiel ist die Verteilung der Augenzahlen beim Würfeln. Die Augenzahl nimmt jeden der diskreten Werte von 1 bis 6 mit gleicher Wahrscheinlichkeit an.

4.2 Binomialverteilung

Man erhält die **Binomialverteilung**, wenn man danach fragt, wie oft ein bestimmtes Ereignis A bei wiederholter Ausführung eines Zufallsexperimentes eintrifft, bei dem A bei jeder Ausführung dieselbe Wahrscheinlichkeit besitzt und die Ergebnisse der verschiedenen Ausführungen sich gegenseitig nicht beeinflussen. Ein Beispiel ist das Ziehen aus einer Urne mit Zurücklegen.

Die Wahrscheinlichkeit des Ereignisses A sei bei einmaliger Ausführung des Zufallsexperimentes

$$P(A) = p \tag{4.1}$$

Dann ist die Wahrscheinlichkeit, dass A nicht eintrifft

$$P(\bar{A}) = q \qquad \text{mit} \quad q = 1 - p \tag{4.2}$$

Wir betrachten nun die Zufallsvariable x, die gleich der Anzahl des Eintreffens von A bei n Ausführungen des Zufallsexperimentes ist. Ist $n = 1$, das heißt, wird das Zufallsexperiment nur einmal ausgeführt, so kann x nur die Werte 0 oder 1 annehmen. Die zugehörigen Wahrscheinlichkeiten sind

$$P(0) = P(\bar{A}) = q \qquad P(1) = P(A) = p$$

Ist $n = 2$, so kann x die Werte 0, 1, 2 annehmen. Die zugehörigen Wahrscheinlichkeiten ergeben sich zu

$$P(0) = P(\bar{\mathrm{A}}\bar{\mathrm{A}}) = P(\bar{\mathrm{A}})\ P(\bar{\mathrm{A}}) = q^2$$

$$P(1) = P(\mathrm{A}\bar{\mathrm{A}}) + P(\bar{\mathrm{A}}\mathrm{A}) = 2P(\mathrm{A})\ P(\bar{\mathrm{A}}) = 2pq$$

$$P(2) = P(\mathrm{AA}) = P(\mathrm{A})\ P(\mathrm{A}) = p^2$$

Wichtig ist die Feststellung, dass das Ergebnis $x = 1$ auf zweierlei Weise realisiert werden kann und dass die Ereignisse $\mathrm{A}\bar{\mathrm{A}}$ und $\bar{\mathrm{A}}\mathrm{A}$ sich nur durch die Reihenfolge, aber nicht durch die zugehörigen Wahrscheinlichkeiten unterscheiden. Entsprechende Überlegungen sind anzustellen, wenn das Zufallsexperiment beliebig oft, das heißt n-mal ausgeführt wird.

Fragt man nach der Wahrscheinlichkeit, dass zuerst x-mal das Ereignis A eintrifft und dann $(n - x)$-mal das Ereignis $\bar{\mathrm{A}}$, so ergibt sich

$$P(\underbrace{\mathrm{A}\cdot\mathrm{A}\cdot\mathrm{A}\ldots}_{x\text{-mal}}\underbrace{\bar{\mathrm{A}}\cdot\bar{\mathrm{A}}\cdot\bar{\mathrm{A}}\ldots}_{(n-x)\text{-mal}}) = \Big(P(\mathrm{A})\Big)^x \cdot \Big(P(\bar{\mathrm{A}})\Big)^{n-x} = p^x q^{n-x} \tag{4.3}$$

Das Ereignis *x-maliges Eintreffen des Ereignisses A bei n Realisationen des Zufallsexperimentes* lässt sich nicht nur in dieser, sondern auch in verschiedenen anderen Formen verwirklichen, die sich nur durch die Reihenfolge unterscheiden, in der die Ereignisse A und $\bar{\mathrm{A}}$ eintreten. Die zugehörigen Wahrscheinlichkeiten sind alle durch 4.3 gegeben.

Da sich diese Formen wechselseitig ausschließen, ist die Wahrscheinlichkeit $P(x)$ gleich der Summe der Wahrscheinlichkeiten aller möglichen Formen, also gleich dem Produkt von $p^x \cdot q^{n-x}$ und der Anzahl z dieser Formen. Zu fragen ist deshalb, wieviele verschiedene Formen, das heißt wieviele verschiedene Reihenfolgen der Ereignisse A und $\bar{\mathrm{A}}$ bei festgehaltenen Anzahlen x und n möglich sind. Wir zitieren zwei bekannte Sätze der Kombinatorik:

1. Satz: n verschiedene Elemente können auf

$$z = n! = 1 \cdot 2 \cdot 3 \ldots n$$

verschiedene Weisen angeordnet werden oder – anders ausgedrückt – die Anzahl der Permutationen von n verschiedenen Elementen ist z.

2. Satz: Wenn unter n Elementen mehrere Gruppen von n_i unter sich gleichen Elementen existieren, so ist die Anzahl der Permutationen

$$z = \frac{n!}{n_1! \cdot n_2! \cdot n_3! \ldots n_k!}$$

wobei gilt

$$\sum_{i=1}^{k} n_i = n \quad \text{und} \quad 0! = 1$$

Die Anzahl der möglichen unterschiedlichen Reihenfolgen der x Ereignisse A und der $(n-x)$ Ereignisse $\bar{A}$ ergibt sich daraus zu

$$z = \frac{n!}{x!(n-x)!} = \binom{n}{x} \tag{4.4}$$

Dieser Ausdruck wird **Binomialkoeffizient**[1] genannt, weil er bei der Berechnung der Koeffizienten der Potenzen eines Binoms vorkommt (Binomischer Satz)

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}\, y^k$$

Gelesen wird er *n über x* (E: n choose k). Wegen der Fakultäten stößt seine Berechnung schon bei mäßigen Werten von n und x auf praktische Schwierigkeiten. Die Zahl 20! hat bereits 19 Stellen in dezimaler Darstellung. Algorithmen und Rechenprogramme versuchen daher, die Multiplikationen und Divisionen abwechselnd auszuführen, um das Anwachsen der Zwischenergebnisse zu bremsen.

Die gesuchte Wahrscheinlichkeit des Ereignisses *x-maliges Eintreffen des Ereignisses* A *bei n Realisationen des Zufallsexperimentes* folgt mit 4.4 zu

$$\begin{aligned} P(x) &= \binom{n}{x} p^x q^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x q^{n-x} \quad \text{mit} \quad x = 0, 1, 2 \ldots n \end{aligned} \tag{4.5}$$

Dies ist die **Binomialverteilung**, die wohl wichtigste diskrete Verteilung.

Die Wahrscheinlichkeitssumme über alle x von 0 bis n ergibt sich direkt aus dem Binomischen Satz. Es gilt

$$(p+q)^n = \sum_{x=0}^{n} \binom{n}{k} p^x q^{n-x}$$

Wegen 4.2 folgt daraus

$$\sum_{x=0}^{n} P(x) = 1 \tag{4.6}$$

Um den Erwartungswert der Binomialverteilung zu berechnen, wird von einem Kunstgriff Gebrauch gemacht. Zunächst schreibt man den Binomischen Satz in der abgewandelten Form

$$(pt+q)^n = \sum_{x=0}^{n} \binom{n}{x} p^x t^x q^{n-x} \tag{4.7}$$

Diese Gleichung wird nun nach t differenziert

$$np\,(pt+q)^{n-1} = \sum_{x=0}^{n} x \binom{n}{x} p^x t^{x-1} q^{n-x} \tag{4.8}$$

[1] siehe die deutsche Wikipedia oder – ausführlicher – die englische

Setzt man danach $t = 1$, so folgt

$$np = \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x} \tag{4.9}$$

Der Ausdruck auf der rechten Seite von 4.9 entspricht vollständig der Definition des Erwartungswertes einer diskreten Variablen. Es ist daher

$$E(x) = \mu_x = np \tag{4.10}$$

Zur Berechnung der Varianz der Binomialverteilung wird 4.7 ein zweites Mal nach t differenziert. Wir erhalten

$$n(n-1)p^2(pt+q)^{n-2} = \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x t^{x-2} q^{n-x} \tag{4.11}$$

Wird nun $t = 1$ gesetzt, so folgt aus 4.11

$$n(n-1)p^2 = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x q^{n-x} - \sum_{x=0}^{n} x \binom{n}{x} p^x q^{n-x}$$

Hierfür können wir schreiben

$$n(n-1)p^2 = \mathrm{E}(x^2) - \mathrm{E}(x)$$

oder nach Umstellen

$$\mathrm{E}(x^2) = n(n-1)p^2 + \mathrm{E}(x) \tag{4.12}$$

Für die Varianz gilt gemäß Abschnitt 3.4.1 *Eindimensionale Wahrscheinlichkeitsverteilungen* auf Seite 44 allgemein

$$\sigma_x^2 = \mathrm{E}(x^2) - \Big(\mathrm{E}(x)\Big)^2$$

Mit 4.10 und 4.12 folgt daraus für die Binomialverteilung

$$\sigma_x^2 = n(n-1)p^2 + np - (np)^2$$

oder

$$\sigma_x^2 = np(1-p) = np\,q \tag{4.13}$$

Beispiel 4.1 : Gegeben seien N Dinge, beispielsweise Schrauben, darunter M unbrauchbare. Die Wahrscheinlichkeit, beim zufälligen Herausgreifen einer Schraube eine unbrauchbare zu erhalten (Ereignis A), ist dann

$$P = \frac{M}{N}$$

Greift man insgesamt n-mal eine einzelne Schraube heraus, nachdem man jeweils die zuvor herausgegriffene zurückgelegt hat, so ist die Wahrscheinlichkeit, dabei genau x unbrauchbare Schrauben zu erhalten, gemäß 4.5

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

Praktisch wichtiger ist das Ziehen ohne Zurücklegen. Wir werden damit auf die *Hypergeometrische Verteilung* geführt. Solange jedoch die Grundgesamtheit sehr viel größer ist als die gezogene Stichprobe

$$n \ll N$$

sind die Unterschiede zwischen Binomialverteilung und Hypergeometrischer Verteilung gering, und man darf auch das *Ziehen ohne Zurücklegen* mit der Binomialverteilung beschreiben.

Ende Beispiel

Beispiel 4.2 : Wie groß ist in einer Familie mit vier Kindern die Wahrscheinlichkeit

a) zwei Knaben und zwei Mädchen,

b) drei Knaben und ein Mädchen,

c) lauter Knaben

zu finden, wenn wir annehmen, dass Knaben- und Mädchengeburten gleichwahrscheinlich sind und nur durch den Zufall, nicht durch gezielte Eingriffe bestimmt werden? Mit den Bezeichnungen

- Ereignis A = Knabe
- Ereignis $\bar{\text{A}}$ = Mädchen
- Variable x = Anzahl der Knaben unter n Kindern

ergibt sich die gesuchte Wahrscheinlichkeit $P(x)$ aus 4.5, indem wir einsetzen

$$p = q = \frac{1}{2} \text{ und } n = 4$$

Wir finden

$$P(x=2) = \binom{4}{2}(\frac{1}{2})^4 = \frac{3}{8}$$

$$P(x=3) = \binom{4}{3}(\frac{1}{2})^4 = \frac{1}{4}$$

$$P(x=4) = \binom{4}{4}(\frac{1}{2})^4 = \frac{1}{16}$$

Wegen $p = q$ ist die Verteilung symmetrisch. Daher gilt

$$P(x=0) = P(x=4) \text{ und } P(x=1) = P(x=3)$$

Die Wahrscheinlichkeitssumme über alle x ergibt sich zu

$$\sum_{x=0}^{4} P(x) = \frac{1}{16} + \frac{1}{4} + \frac{3}{8} + \frac{1}{4} + \frac{1}{16} = 1$$

wie es 4.6 verlangt.

Ende Beispiel

Die **Null-Eins-Verteilung** oder **Bernoulli-Verteilung** ist ein Trivialfall der Binomialverteilung nach 4.5 mit $n = 1$

$$P(x) = \frac{1}{x!(1-x)!}\, p^x q^{1-x} \quad \text{mit} \quad x = 0, 1 \tag{4.14}$$

Aus 4.10 ergibt sich ihr Erwartungswert zu

$$E(x) = p \tag{4.15}$$

und aus 4.13 ihre Varianz zu

$$\sigma_x^2 = pq = p(1-p) \tag{4.16}$$

Beispiele sind das Werfen einer Münze (Zahl–Wappen mit jeweils gleicher Wahrscheinlichkeit), das Werfen eines Würfels, falls wir nur gerade und ungerade Augenzahl unterscheiden oder nur eins und nicht-eins, die Qualitätskontrolle (in Ordnung – nicht in Ordnung, mit hoffentlich ungleichen Wahrscheinlichkeiten) oder Aussagen über Personengruppen (Raucher–Nichtraucher).

4.3 Poisson-Verteilung

Die **Poisson-Verteilung** lässt sich als Näherung der Binomialverteilung für den Fall ableiten, dass die Wahrscheinlichkeit p für das Eintreten des Ereignisses A sehr klein, A also ein seltenes Ereignis ist, während die Anzahl n der Ausführungen des Zufallsexperimentes sehr groß ist. Sie wurde erstmals von dem französischen Physiker und Mathematiker Siméon-Denis Poisson (1781–1840) beschrieben, war jedoch schon vorher bekannt.

Wir gehen aus von der Binomialverteilung 4.5

$$P(x) = \frac{n!}{x!(n-x)!}\, p^x q^{n-x} \tag{4.17}$$

und verwenden die Formel von James Stirling (1692–1770) für Näherungswerte der Fakultät großer Zahlen

$$n! \approx \sqrt{2\pi n}\, (\frac{n}{\mathrm{e}})^n \tag{4.18}$$

mit e als der Basis der natürlichen Logarithmen (Eulersche Zahl, e = 2,718...). Die Näherung ist umso besser, je größer die Zahl n ist. Die Abweichungen vom

genauen Wert sind kleiner als 1 %, falls $n \geq 10$ ist. Eine verbesserte Näherung – brauchbar ab $n = 2$ – liefert die modifizierte Formel[2]

$$n! \approx \sqrt{2\pi n}\,(\frac{n}{\mathrm{e}})^n\,\left(1 + \frac{1}{12\,n}\right)$$

Tabelle 4.1 zeigt die exakten Werte, die Werte der einfachen und der verbesserten Näherung von J. STIRLING sowie in der letzten Spalte den relativen Fehler der einfachen Näherung.

Tab. 4.1: Fakultäten; exakte Werte, einfache und verbesserte Näherung nach J. STIRLING sowie relativer Fehler der einfachen Näherung

n	**Fakultät**	**Stirling**	**Stirling verb.**	**Fehler**
1	1	0.92	1.00	-0.077863
2	2	1.92	2.00	-0.040498
3	6	5.84	6.00	-0.027299
4	24	23.51	24.00	-0.020576
5	120	118.02	119.99	-0.016507
6	720	710.08	719.94	-0.013781
7	5040	4980.39	5039.68	-0.011827
8	40320	39902.38	40318.03	-0.010358
9	362880	359536.69	362865.73	-0.009213
10	3628800	3598693.55	3628682.66	-0.008297

Wenn, wie weiterhin vorausgesetzt, $p \ll 1$, $x \ll n$ und $n \gg 1$ gilt, lässt sich die Formel von STIRLING dazu benutzen, die beiden Ausdrücke $n!$ und $(n-x)!$ in 4.17 zu eliminieren. Wir erhalten

$$P(x) \approx \frac{(np)^x \mathrm{e}^{-x}}{x!}\;\frac{(1-p)^{n-x}}{(1-\frac{x}{n})^{n-x}}\;\sqrt{\frac{n}{n-x}} \tag{4.19}$$

Der zweite und dritte Faktor lassen sich unter den genannten Voraussetzungen erheblich vereinfachen. Es gilt

$$\begin{aligned}
\ln(1-p)^{n-x} &= (n-x)\ln(1-p) \\
&= (n-x)(-p - \frac{p^2}{2} - \frac{p^3}{3} - \ldots) \\
&\approx -np
\end{aligned}$$

woraus folgt

$$(1-p)^{n-x} \approx \exp(-np) \tag{4.20}$$

[2]siehe `http://algolist.manual.ru/count_fast/gamma_function.php` oder D. E. KNUTH, Band I

Ebenso ergibt sich

$$(1 - \frac{x}{n})^{n-x} \approx \exp(-x) \tag{4.21}$$

und

$$\sqrt{\frac{n}{n-x}} \approx 1 \tag{4.22}$$

Setzt man 4.20, 4.21 und 4.22 in 4.19 ein, so folgt

$$P(x) \approx \frac{(np)^x}{x!} \exp(-np) \quad \text{mit} \quad x = 0, 1, 2, \ldots n \tag{4.23}$$

Wir betrachten nun den Grenzfall, dass p gegen null geht und n gegen unendlich, wobei jedoch das Produkt

$$\mu_x = np \tag{4.24}$$

einen endlichen Wert behält. Für diesen Grenzfall folgt aus 4.23

$$P(x) = \frac{\mu_x^x}{x!} \exp(-\mu_x) \quad \text{mit} \quad x = 0, 1, 2, \ldots \tag{4.25}$$

Dies ist die **Poisson-Verteilung**. Sie ist eine eindimensionale diskrete Verteilung, die sich dadurch von der Binomialverteilung unterscheidet, dass die Variable x abzählbar unendlich viele Werte annimmt.

Die Wahrscheinlichkeitssumme über alle x erhält man, wenn man für die Exponentialfunktion folgende Reihenentwicklung verwendet

$$\exp z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{31} + \ldots$$

Hieraus folgt

$$\sum_{x=0}^{\infty} \frac{\mu_x^x}{x!} = \exp(\mu_x)$$

und damit für die Wahrscheinlichkeitssumme

$$\sum_{x=0}^{\infty} P(x) = \sum_{x=0}^{\infty} \frac{\mu_x^x}{x!} \exp(-\mu_x) = 1 \tag{4.26}$$

Für den Mittelwert finden wir

$$\begin{aligned} E(x) &= \sum_{x=0}^{\infty} x \, \frac{\mu_x^x}{x!} \, \exp(-\mu_x) \\ &= \mu_x \, \exp(-\mu_x) \sum_{x=1}^{\infty} \frac{\mu_x^{x-1}}{(x-1)!} \\ &= \mu_x \end{aligned} \tag{4.27}$$

Ebenso erhalten wir

$$\begin{aligned} E(x^2) &= \sum_{x=0}^{\infty} x^2 \, \frac{\mu_x^x}{x!} \, \exp(-\mu_x) \\ &= \mu_x \, \exp(-\mu_x) \, \sum_{x=1}^{\infty} x \, \frac{\mu_x^{x-1}}{(x-1)!} \\ &= \mu_x \, \exp(-\mu_x) \, \sum_{x=1}^{\infty} ((x-1)+1) \, \frac{\mu_x^{x-1}}{(x-1)!} \\ &= \mu_x^2 \, \exp(-\mu_x) \, \sum_{x=2}^{\infty} \frac{\mu_x^{x-2}}{(x-2)!} + \mu_x \, \exp(-\mu_x) \, \sum_{x=1}^{\infty} \frac{\mu_x^{x-1}}{(x-1)!} \\ &= \mu_x^2 + \mu_x \end{aligned} \tag{4.28}$$

Mit der allgemeinen Beziehung für die Varianz (3.94 auf Seite 47)

$$\sigma_x^2 = E(x^2) - (E(x))^2$$

ergibt sich schließlich

$$\sigma_x^2 = \mu_x^2 + \mu_x - \mu_x^2 = \mu_x \tag{4.29}$$

Mittelwert und Varianz der Poisson-Verteilung sind gleich. Man bezeichnet deshalb die Poisson-Verteilung als eine einparametrige Verteilung mit dem einzigen Parameter $\mu_x = \sigma_x^2$.

Es ist vorteilhaft, die Binomialverteilung unter den oben genannten Voraussetzungen durch die Poisson-Verteilung anzunähern, weil das Rechnen mit der Binomialverteilung für große Anzahlen n mühsam wird.

Beispiel 4.3 : Gefragt sei nach der Wahrscheinlichkeit, dass in einem Dorf mit 500 Einwohnern wenigstens einer am 24. Dezember Geburtstag hat. Man geht aus von der Annahme, dass die Geburtstage der Dorfbewohner zufällig über alle Tage des Jahres verteilt sind, und definiert

- Ereignis A = Ein einzeln herausgegriffener Dorfbewohner hat am 24. Dezember Geburtstag,
- Variable x = Anzahl derjenigen unter den $n = 500$ Dorfbewohnern, die am 24. Dezember Geburtstag haben.

Dann ist, wenn man Schaltjahre außer Acht lässt

$$P(\mathrm{A}) = p = \frac{1}{365} \qquad P(\bar{\mathrm{A}}) = q = \frac{364}{365} \qquad n = 500$$

Die exakte Lösung erhält man mit der Binomialverteilung nach 4.5. Die gesuchte Wahrscheinlichkeit ist

$$\begin{aligned} P(x \geq 1) &= 1 - P(x = 0) \\ &= 1 - \binom{500}{0} p^0 q^{500} \\ &= 1 - (\frac{364}{365})^{500} = 1 - 0,2537 = 0,7463 \end{aligned}$$

Die Näherungslösung bei Verwendung der Poisson-Verteilung 4.25 liefert

$$P(x \geq 1) = 1 - P(x = 0) = 1 - \frac{\mu_x^0}{0!} \exp(-\mu_x)$$

Mit dem Mittelwert

$$\mu_x = np = \frac{500}{365}$$

folgt

$$P(x \geq 1) = 1 - 0,2541 = 0,7459$$

Hier ist der Vorteil noch nicht erkennbar, den die Verwendung der Poisson-Verteilung anstelle der Binomialverteilung bietet. Dieser wird er jedoch deutlich, wenn etwa nach der Wahrscheinlichkeit $P(5 \geq x \geq 1)$ gefragt ist.

Ende Beispiel

Die Poisson-Verteilung lässt sich auch ohne Bezug auf die Binomialverteilung herleiten, wenn man von folgender Modellvorstellung ausgeht. Es werden Ereignisse beobachtet, die in zufälliger zeitlicher Folge eintreten. Dabei soll gelten

- Die Wahrscheinlichkeit für das Eintreten eines Ereignisses im Zeitintervall von t bis $t + \Delta t$ ist unabhängig davon, was im Zeitraum t geschehen ist.
- Die Wahrscheinlichkeit für das Eintreten genau eines Ereignisses im Zeitintervall Δt ist der Länge des Zeitintervalls proportional.
- Das Zeitintervall Δt wird so klein gewählt, dass immer höchstens ein Ereignis in das Zeitintervall fällt.

Die Anzahl x der Ereignisse, die im Zeitraum t beobachtet werden, ist eine Zufallsvariable, deren Wahrscheinlichkeit mit $P(x; t)$ bezeichnet werden soll.

Entsprechend dieser Festlegung ist $P(1; \Delta t)$ die Wahrscheinlichkeit, dass im Zeitintervall Δt genau ein Ereignis eintritt. Es gilt

$$P(1; \Delta t) = \lambda \, \Delta t \tag{4.30}$$

worin λ die mittlere Anzahl der Ereignisse in der Zeiteinheit bedeutet. Ist Δt so klein gewählt worden, dass in Δt höchstens ein Ereignis eintreten kann, so ist

$$P(0; \Delta t) + P(1; \Delta t) = 1$$

Mit 4.30 folgt

$$P(0;\Delta t) = 1 - \lambda\ \Delta t \tag{4.31}$$

Sind im Zeitraum $t + \Delta t$ genau x Ereignisse eingetreten, so kann dies nur auf zweierlei Weise geschehen sein:

- im Zeitraum t genau x Ereignisse, im Zeitintervall Δt kein Ereignis,
- im Zeitraum t genau $x - 1$ Ereignisse, im Zeitintervall Δt ein Ereignis.

Für die zugehörigen Wahrscheinlichkeiten gilt, da Unabhängigkeit vorausgesetzt wurde und die beiden genannten Möglichkeiten als sich wechselseitig ausschließende Ereignisse anzusehen sind

$$P(x;t+\Delta t) = P(x;t)\ P(0;\Delta t) + P(x-1;t)\ P(1;\Delta t)$$

Mit 4.30 und 4.31 folgt daraus

$$P(x;t+\Delta t) = P(x;t)\ (1 - \lambda\ \Delta t) + P(x-1;t)\ \lambda\ \Delta t$$

oder

$$\frac{P(x;t+\Delta t) - P(x;t)}{\Delta t} = \lambda(P(x-1;t) - P(x;t))$$

Macht man das Zeitintervall Δt sehr klein im Vergleich zu t, so darf man den Differenzenquotienten durch den Differentialquotienten ersetzen und erhält

$$\frac{\mathrm{d}P(x;t)}{\mathrm{d}t} + \lambda\ P(x;t) = \lambda\ P(x-1;t) \tag{4.32}$$

Die Differentialgleichung 4.32 ist rekursiv zu lösen. Man beginnt mit dem unmöglichen Ereignis $x = -1$ und setzt

$$P(-1;t) = 0$$

Dann folgt aus 4.32 für $x = 0$:

$$\frac{\mathrm{d}P(0;t)}{\mathrm{d}t} + \lambda P(0;t) = 0$$

und weiter

$$P(0;t) = c_0\ \exp(-\lambda t)$$

Für $t \to 0$ muss diese Wahrscheinlichkeit gegen 1 gehen (sicheres Ereignis). Daher ist $c_0 = 1$ und

$$P(0;1) = \exp(-\lambda t)$$

Für $t = 1$ ergibt sich hiermit

$$\frac{\mathrm{d}P(1;t)}{\mathrm{d}t} + \lambda P(1;t) = \lambda\ \exp(-\lambda t)$$

Die Lösung dieser einfachen Differentialgleichung ist

$$P(1;t) = \lambda t \exp(-\lambda t)$$

Für $x = 2$ erhält man damit die Differentialgleichung

$$\frac{\mathrm{d}P(2;t)}{\mathrm{d}t} + \lambda P(2;t) = \lambda(\lambda t)\ \exp(-\lambda t)$$

mit der Lösung

$$P(2;t) = \frac{(\lambda t)^2}{2}\ \exp(-\lambda t)$$

Die Lösung für beliebige x lautet

$$P(x;t) = \frac{(\lambda t)^x}{x!} \exp(-\lambda t) \quad \text{mit} \quad x = 0, 1, 2, \ldots \tag{4.33}$$

Den Beweis führt man durch vollständige Induktion.

Ist t ein vorgegebener konstanter Zeitraum, so ist 4.33, wie der Vergleich mit 4.25 zeigt, mit der Poisson-Verteilung identisch, und x hat den Mittelwert

$$\mu_x = \lambda t \tag{4.34}$$

Ein Prozess, der sehr genau dem Modell entspricht, das der Herleitung von 4.33 zugrunde liegt, ist der radioaktive Zerfall eines instabilen Isotops. Die Anzahl N der innerhalb eines Zeitraumes t ausgesandten Gammaquanten beispielsweise einer Cs-137-Quelle folgt deshalb einer Poisson-Verteilung, und es gilt entsprechend 4.29 auf Seite 71

$$\mu_N = \sigma_N^2$$

Außer dem radioaktiven Zerfall gibt es in Wissenschaft und Technik noch viele andere Vorgänge, die dadurch gekennzeichnet sind, dass irgendwelche Ereignisse in zufälliger zeitlicher Folge eintreten. In allen diesen Fällen liefert die Poisson-Verteilung eine exakte Beschreibung.

Wenn man anstelle des Zeitraums t eine Strecke s, eine Fläche F oder ein Volumen V betrachtet, über die bestimmte Dinge (Ereignisse) zufällig verteilt sind, so kann man dasselbe Modell auch auf diese Fälle anwenden. Die Konstante λ in 4.33 bedeutet dann die mittlere Anzahl der Dinge (Ereignisse) in der Strecken-, Flächen- oder Volumeneinheit.

Kapitel 5

Stetige Wahrscheinlichkeitsverteilungen

Das Kapitel erläutert einige Wahrscheinlichkeitsverteilungen eines stetigen Merkmals, die in mathematischer oder technischer Hinsicht interessant sind.

5.1 Stetige Gleichverteilung (Rechteckverteilung)

Die stetige **Gleichverteilung** oder **Rechteckverteilung** wurde bereits in Abschnitt 3.3.1 *Verteilungen einer Zufallsvariablen* auf Seite 32 behandelt. Ein Beispiel ist die Frequenzverteilung in Weißem Rauschen. Jede Frequenz ist gleich stark vertreten. Die spektrale Leistungsdichte ist über einen größeren Frequenzbereich konstant.

5.2 Eindimensionale Normalverteilung

5.2.1 Gewöhnliche Normalverteilung

Die eindimensionale (univariate) **Normal-** oder **Gauß-Verteilung** ist eine stetige Verteilung mit der Dichtefunktion

$$\varphi(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\ \exp\Big(-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2\Big) \tag{5.1}$$

Die Zufallsvariable x kann jeden Wert von $-\infty$ bis $+\infty$ annehmen. Die Bezeichnung *Normalverteilung* rührt von ihrer besonderen Bedeutung in Wissenschaft und Technik her. C. F. Gauss hat sie erstmals ausführlich beschrieben und bei astronomischen Berechnungen angewandt. Vor ihm hat sich jedoch schon Abraham de Moivre (1667–1754) mit der Normalverteilung befasst.

Die zugehörige Summenfunktion ist

$$\Phi(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\ \int_{-\infty}^{x} \exp\Big(-\frac{1}{2}(\frac{\xi-\mu_x}{\sigma_x})^2\Big)\ \mathrm{d}\xi \tag{5.2}$$

Gehorcht eine Zufallsvariable x dieser Verteilung, so ist die Wahrscheinlichkeit, dass x irgendeinen Wert in dem Intervall $a < x \leq b$ annimmt, gegeben durch

$$\begin{aligned} P(a < x \leq b) &= \frac{1}{\sigma_x\sqrt{2\pi}} \int_a^b \exp\Big(-\frac{1}{2}(\frac{\xi-\mu_x}{\sigma_x})^2\Big)\,\mathrm{d}x \\ &= \Phi(b) - \Phi(a) \end{aligned} \tag{5.3}$$

Das Integral in 5.2 und 5.3 ist nicht analytisch auswertbar. Es lässt sich jedoch durch die Substitution

$$t = \frac{x-\mu_x}{\sigma_x} \tag{5.4}$$

auf das spezielle Integral

$$H(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} \exp\Big(-\frac{1}{2}u^2\Big)\,\mathrm{d}u \tag{5.5}$$

zurückführen. Aus 5.2 folgt damit

$$\Phi(x) = H\Big(\frac{x-\mu_x}{\sigma_x}\Big) \tag{5.6}$$

und aus 5.3

$$P(a < x \leq b) = H\Big(\frac{b-\mu_x}{\sigma_x}\Big) - H\Big(\frac{a-\mu_x}{\sigma_x}\Big) \tag{5.7}$$

Man bezeichnet die durch 5.5 definierte Funktion $H(t)$ als die Summenfunktion der standardisierten, normalisierten oder **normierten Normalverteilung** $\mathcal{N}(0,1)$. Die Zufallsvariable t ist, wie man aus 5.4 abliest, normalverteilt mit den Parametern

$$\mu_t = 0 \qquad \sigma_t^2 = 1 \tag{5.8}$$

sofern die Zufallsvariable x normalverteilt ist mit dem Mittelwert μ_x und der Varianz σ_x^2.

Trotz überragenden Bedeutung der Verteilung hat sich eine einheitliche Definition der normierten Normalverteilung nicht durchgesetzt. Neben der hier mit 5.5 eingeführten Summenfunktion $H(t)$ findet man in der Literatur und manchen Computerprogrammen die so genannte **Gaußsche Fehlerfunktion** oder **Error-Funktion** $\mathrm{erf}\,(x)$

$$\mathrm{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\,(-\epsilon^2)\,\mathrm{d}\epsilon$$

Zwischen $H(t)$ und $\mathrm{erf}\,(x)$ besteht folgender Zusammenhang

$$H(t) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\,(\frac{t}{\sqrt{2}})$$

oder umgekehrt

$$\mathrm{erf}\,(x) = 2H(x\sqrt{2}) - 1$$

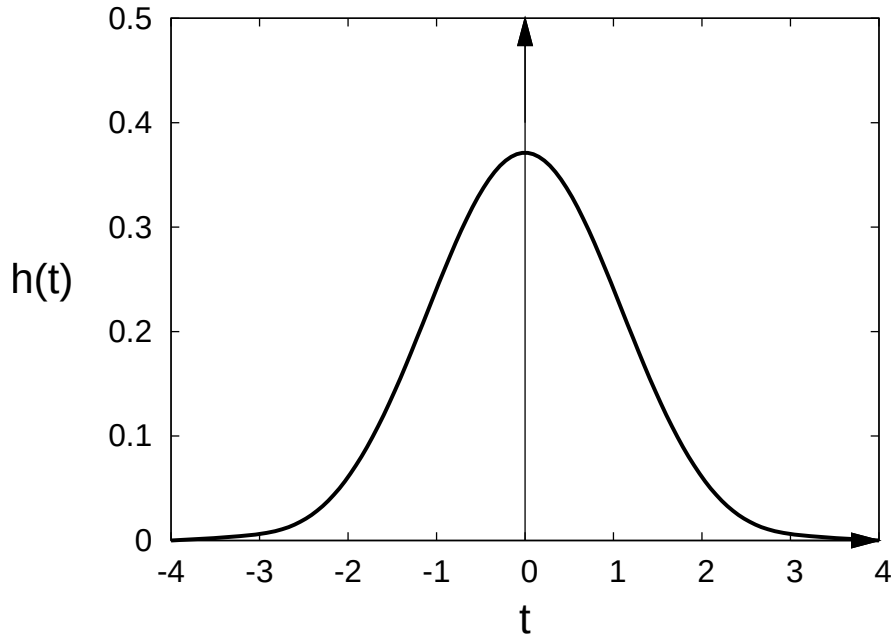


Abb. 5.1: Dichtefunktion der normierten Normalverteilung (Glockenkurve)

Die Summenfunktion $H(t)$ der normierten Normalverteilung lässt sich mittels einer Reihenentwicklung des Integranden numerisch für beliebige t berechnen. Die Ergebnisse solcher Berechnungen sind in Tabellenform in vielen Statistikbüchern zu finden, gebrauchsfertige Algorithmen in vielen Rechenprogrammen. Das gilt auch für andere Verteilungen wie die Chi-Quadrat-Verteilung oder die Student-Verteilung. Wegen 5.6 und 5.7 kann man mit den Tabellenwerten der normierten Normalverteilung auch alle Funktionswerte einer beliebigen Normalverteilung berechnen. Für die normierte Normalverteilung gibt es mit

$$H(t) \approx \frac{1}{\pi}\left(\frac{\pi + t}{2} + \sum_{n=1}^{3} \frac{1}{n} \exp\left(-\frac{n^2}{2}\right) \sin nt\right) \tag{5.9}$$

eine leicht handhabbare Näherung, deren maximale Abweichung vom genauen Wert kleiner als 10^{-4} ist.

Die Dichtefunktion der normierten Normalverteilung ist

$$h(t) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big) \tag{5.10}$$

Wenn man sie in einem Diagramm aufträgt, erhält man eine zu $t = 0$ symmetrische Kurve in Form einer Glocke (Glockenkurve), siehe Abbildung 5.1. Wegen der Symmetrie von $h(t)$ gilt

$$H(-t) = 1 - H(t) \qquad H(0) = \frac{1}{2} \tag{5.11}$$

Dass die Summenfunktion integriert über alle t den Wert eins ergibt, zeigt man wie folgt

$$\Big(\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\big(-\frac{1}{2}t^2\big)\,\mathrm{d}t\Big)^2$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\big(-\frac{1}{2}u^2\big)\,\mathrm{d}u\;\;\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp\big(-\frac{1}{2}v^2\big)\,\mathrm{d}v$$

$$= \frac{1}{2\pi}\int_{v=-\infty}^{v=+\infty}\int_{u=-\infty}^{u=+\infty}\exp\big(-\frac{1}{2}(u^2+v^2)\big)\,\mathrm{d}u\,\mathrm{d}v$$

$$= \frac{1}{2\pi}\int_{\varphi=0}^{\varphi=2\pi}\int_{r=0}^{r=\infty}\exp\big(-\frac{1}{2}r^2\big)\;r\;\mathrm{d}r\;\mathrm{d}\varphi$$

$$= \frac{1}{2\pi}\int_0^{2\pi}\mathrm{d}\varphi\;\int_0^{\infty}\exp\big(-\frac{1}{2}r^2\big)\,\mathrm{d}\big(\frac{1}{2}r^2\big)$$

$$= \int_0^{\infty}\exp(-s)\,\mathrm{d}s = 1$$

Den Erwartungswert von t erhält man auf ähnliche Weise. Es ist

$$E(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t\;\exp\big(-\frac{t^2}{2}\big)\,\mathrm{d}t$$

$$= \frac{1}{\sqrt{2\pi}}\int_{+\infty}^{+\infty}\exp\big(-\frac{t^2}{2}\big)\,\mathrm{d}\big(\frac{t^2}{2}\big) = 0$$

Damit ist gezeigt, dass $\mu_t = 0$ tatsächlich der Mittelwert der Variablen t ist und μ_x der Mittelwert der Variablen x. Für die Varianz von t gilt

$$\sigma_t^2 = E(t^2) = \frac{1}{\sqrt{2\pi}}\;\int_{-\infty}^{+\infty}t^2\;\exp(-\frac{t^2}{2})\,\mathrm{d}t$$

Man substituiert wie folgt

$$u = -t \qquad \mathrm{d}u = -\mathrm{d}t$$

$$v = \exp(-\frac{t^2}{2}) \qquad \mathrm{d}v = -t\;\exp(-\frac{t^2}{2})\,\mathrm{d}t$$

und erhält

$$\sigma_t^2 = \frac{1}{\sqrt{2\pi}}\;\int_{-\infty}^{+\infty}u\,\mathrm{d}v$$

Daraus durch partielle Integration

$$\sigma_t^2 = \frac{1}{\sqrt{2\pi}}\;\Big(\big|uv\big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty}v\,\mathrm{d}u\Big)$$

$$= \frac{1}{\sqrt{2\pi}}\;\Big(\big|-t\;\exp(-\frac{t^2}{2})\big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty}\exp(-\frac{t^2}{2})\,\mathrm{d}t\Big)$$

Nun ist

$$\Big| - t \ \exp\left(-\frac{t^2}{2}\right)\Big|_{-\infty}^{+\infty} = 0$$

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t = \sqrt{2\pi}$$

Daraus folgt

$$\sigma_t^2 = \frac{1}{\sqrt{2\pi}} \left(0 + \sqrt{2\pi}\right) = 1$$

Damit ist gezeigt, dass $\sigma_t^2 = 1$ tatsächlich die Varianz der Variablen t ist und σ_x^2 die Varianz der Variablen x.

Beispiel 5.1 : Auf einer Metall-Hobelmaschine werden Platten bearbeitet. Die Unvollkommenheiten von Maschine und Material bedingen zufällige Schwankungen der Plattenstärke x. Es sei angenommen, dass x normalverteilt ist und dass bei einer bestimmten Maschineneinstellung der Mittelwert $\mu_x = 10$ mm und die Varianz $\sigma_x^2 = 4 \cdot 10^{-4}$ mm^2 beträgt. Wieviel Ausschuss ist zu erwarten, wenn die brauchbaren Platten mindestens 9,97 mm und höchstens 10,05 mm stark sein sollen?

Die Wahrscheinlichkeit, dass irgendeine die Maschine durchlaufende Platte eine Stärke hat, die innerhalb der Toleranzgrenzen liegt, ist

$$P(a \leq x \leq b) = H(\beta) - H(\alpha)$$

mit

$$\alpha = \frac{a - \mu_x}{\sigma_x} = \frac{9,97 - 10,00}{0,02} = -1,5$$

$$\beta = \frac{b - \mu_x}{\sigma_x} = \frac{10,05 - 10,00}{0,02} = 2,5$$

Aus einer Tabelle der Normalverteilung lesen wir ab

$$H(2,5) = 0,9938 \qquad H(-1,5) = 1 - H(1,5) = 0,0668$$

und erhalten

$$P(a \leq x \leq b) = 0,9938 - 0,0668 = 0,9270$$

Damit ergibt sich die gesuchte Ausschusswahrscheinlichkeit zu

$$1 - P(a \leq x \leq b) = 0,0730 \quad \text{oder } 7{,}3\ \%$$

Ende Beispiel

Die Normalverteilung lässt sich als Näherung der Binomialverteilung für den Fall ableiten, dass die Anzahl n der Ausführungen eines Zufallsexperimentes sehr groß ist. Man geht dabei ähnlich vor wie bei der Ableitung der Poissonverteilung, das heißt, man geht aus von der Binomialverteilung in der Form 4.5 auf Seite 65 beziehungsweise 4.17 auf Seite 68

$$P(x) = \frac{n!}{x!(n-x)!}\, p^x q^{n-x} \tag{5.12}$$

und verwendet die Formel von Stirling, um die Fakultäten in 5.12 zu ersetzen. Dabei setzt man voraus, dass $n \gg 1$, $x \gg 1$ und $(n-x) \gg 1$ ist. Man erhält

$$\begin{aligned} P(x) &\approx \sqrt{\frac{n}{2\pi x(n-x)}}\, \Big(\frac{np}{x}\Big)^x \, \Big(\frac{nq}{n-x}\Big)^{n-x} \\ &\approx \frac{1}{\sqrt{2\pi npq}} \left(\Big(\frac{np}{x}\Big)^{x+0{,}5} \, \Big(\frac{nq}{n-x}\Big)^{n-x+0{,}5} \right) \end{aligned} \tag{5.13}$$

Unter der zusätzlichen Voraussetzung

$$\Big|\frac{x}{n} - p\Big| \ll 1 \tag{5.14}$$

lässt sich der Ausdruck in den großen Klammern – im Folgenden mit z bezeichnet – erheblich vereinfachen. Man führt durch

$$x = np + \delta \qquad (n-x) = n - np - \delta = nq - \delta \tag{5.15}$$

die Abweichung δ der Variablen x vom Mittelwert ein. Wegen 5.14 gilt dann auch

$$\Big|\frac{\delta}{n}\Big| \ll 1 \qquad \Big|\frac{\delta}{np}\Big| \ll 1 \qquad \Big|\frac{\delta}{nq}\Big| \ll 1 \tag{5.16}$$

Damit erhält man

$$\begin{aligned} z &= \Big(\frac{np}{x}\Big)^{x+0{,}5} \, \Big(\frac{nq}{n-z}\Big)^{n-x+0{,}5} \\ &= \Big(1 + \frac{\delta}{np}\Big)^{-(np+\delta+0{,}5)} \, \Big(1 - \frac{\delta}{nq}\Big)^{-(nq-\delta+0{,}5)} \end{aligned}$$

Durch Logarithmieren und mit der Reihenentwicklung

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n+1}\, \frac{x^n}{n} + \ldots \quad \text{für} \quad -1 < x \leq 1$$

folgt daraus

$$\begin{aligned} \ln z &= -(np + \delta + \frac{1}{2}) \left((\frac{\delta}{np}) - \frac{1}{2}(\frac{\delta}{np})^2 + \frac{1}{3}(\frac{\delta}{np})^3 - \ldots \right) \\ &\quad -(nq - \delta + \frac{1}{2}) \left(-(\frac{\delta}{nq}) - \frac{1}{2}(\frac{\delta}{nq})^2 - \frac{1}{3}(\frac{\delta}{nq})^3 - \ldots \right) \end{aligned}$$

$$
\begin{aligned}
&= -\delta - \frac{1}{2}\frac{\delta^2}{np} - \frac{1}{2}\frac{\delta}{np} + \frac{1}{6}\frac{\delta^3}{(np)^2} + \dots \\
&\quad +\delta + \frac{1}{2}\frac{\delta^2}{nq} + \frac{1}{2}\frac{\delta}{nq} - \frac{1}{6}\frac{\delta^3}{(nq)^2} + \dots \\
&= -\frac{1}{2}\frac{\delta^2}{npq} - \frac{1}{2}\frac{\delta}{np} + \frac{1}{2}\frac{\delta}{nq} + \frac{1}{6}\frac{\delta^3}{(np)^2} - \frac{1}{6}\frac{\delta^3}{(nq)^2} + \dots \\
&= -\frac{1}{2}\frac{\delta^2}{npq} - \frac{1}{\sqrt{n}}\left(\frac{1}{2}\sqrt{\frac{q}{p}}(\frac{\delta^2}{npq})^{\frac{1}{2}} - \frac{1}{2}\sqrt{\frac{p}{q}}(\frac{\delta^2}{npq})^{\frac{1}{2}}\right. \\
&\quad \left. +\frac{1}{6}p\sqrt{\frac{p}{q}}(\frac{\delta^2}{npq})^{\frac{3}{2}} - \frac{1}{6}q\sqrt{\frac{q}{p}}(\frac{\delta^2}{npq})^{\frac{3}{2}} + \dots\right)
\end{aligned}
$$

Geht man davon aus, dass

$$\frac{\delta^2}{npq} = (\frac{x-np}{\sqrt{npq}})^2 = t^2$$

beschränkt ist, während n sehr große Werte annimmt, so verschwindet in der obigen Gleichung der zweite Ausdruck, und es gilt

$$\ln z = -\frac{1}{2}\,(\frac{x-np}{\sqrt{npq}})^2$$

oder

$$z = \exp\Big(-\frac{1}{2}\,(\frac{x-np}{\sqrt{npq}})^2\Big)$$

Setzt man dies in 5.13 ein, so erhält man

$$P(x) \approx \frac{1}{\sqrt{2\pi npq}}\ \exp\Big(-\frac{1}{2}\frac{(x-np)^2}{npq}\Big) \tag{5.17}$$

Die Gleichung sagt aus, dass für solche Werte der Variablen x, für die die Bedingung 5.14 erfüllt ist, und für große n die Dichtefunktion der Normalverteilung eine gute Näherung für die Binomialverteilung darstellt. Beide Verteilungen sind asymptotisch gleich, das heißt, sie stimmen im Grenzfall $n \to \infty$ überein (**Lokaler Grenzwertsatz** von Abraham de Moivre und P.-S. Laplace).

Wie man zeigen kann, ist die Voraussetzung 5.14 im Allgemeinen erfüllt. Mit

$$t = \frac{x-np}{\sqrt{npq}}$$

folgt aus 5.14

$$\Big|\frac{t}{\sqrt{n}}\sqrt{pq}\Big| \ll 1$$

Aus Abbildung 5.1 auf Seite 77 ist abzulesen, dass der praktisch interessante Wertebereich von t sich auf $-3 \le t \le +3$ beschränkt. Damit folgt

$$3\,\sqrt{pq} \ll \sqrt{n}$$

Da $n \gg 1$ vorausgesetzt wurde, kann man davon ausgehen, dass diese Bedingung praktisch immer erfüllt ist. Das bedeutet, die Voraussetzung 5.14 schränkt den Anwendungsbereich der Näherung 5.17 praktisch nicht ein.

Mittelwert und Varianz der Binomialverteilung sind gemäß 4.10 und 4.13 auf Seite 66

$$\mu_x = np \qquad \sigma_x^2 = npq$$

Der Vergleich zwischen 5.1 und 5.17 zeigt, dass die mit 5.17 als Näherungsfunktion der Binomialverteilung eingeführte Normalverteilung den gleichen Mittelwert und die gleiche Varianz besitzt.

Fragt man nach der Wahrscheinlichkeit, dass die Variable x irgendeinen Wert in dem Intervall $a \le x \le b$ annimmt, so erhält man mit der Binomialverteilung den exakten Wert

$$P(a \le x \le b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \tag{5.18}$$

Nähert man in 5.18 die Binomialverteilung punktweise durch die Normalverteilung an, so folgt

$$P(a \le x \le b) \approx \frac{1}{\sqrt{2\pi npq}} \sum_{x=a}^{b} \exp\Big(-\frac{1}{2}\frac{(x-np)^2}{npq}\Big)\,\Delta x \tag{5.19}$$

Da $\Delta x = 1$ ist, bedeutet das Hinzufügen des Faktors Δx in 5.19 keine Veränderung. Der Mittelwertsatz der Integralrechnung lautet allgemein

$$\int_x^{x+\Delta x} f(u)\,\mathrm{d}u = f(\xi)\,\Delta x \tag{5.20}$$

mit $x \le \xi \le x + \Delta x$. Wendet man ihn auf 5.19 an, so ergibt sich

$$P(a \le x \le b) \approx \frac{1}{\sqrt{2\pi npq}} \int_a^b \exp\Big(-\frac{1}{2}\frac{(x-np)^2}{npq}\Big)\,\mathrm{d}x$$

Mit der Substitution

$$t = \frac{x-np}{\sqrt{npq}}$$

folgt daraus

$$P(a \le x \le b) \approx \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta \exp\Big(-\frac{t^2}{2}\Big)\,\mathrm{d}t = H(\beta) - H(\alpha) \tag{5.21}$$

Dies ist der **Integrale Grenzwertsatz** von A. DE MOIVRE und P.-S. LAPLACE. Für die Integrationsgrenzen in 5.21 ist einzusetzen

$$\alpha = \frac{a - np - 0,5}{\sqrt{npq}} \qquad \beta = \frac{b - np + 0,5}{\sqrt{npq}} \tag{5.22}$$

Die Addition beziehungsweise Subtraktion von $0,5$ in 5.22 verbessert die Näherung, da sie der endlichen Intervallbreite $\Delta x = 1$ Rechnung trägt.

Beispiel 5.2 : Eine Firma liefert Leuchtdioden in Kartons zu je 1000 Stück. Wie groß ist die Wahrscheinlichkeit, dass ein solcher Karton nicht mehr als 1 % = 10 Stück Ausschuss enthält, wenn man den Produktionsprozess als Zufallsexperiment mit $p = 0,01$ Wahrscheinlichkeit, eine unbrauchbare Leuchtdiode zu produzieren, ansehen kann?

Die exakte Antwort liefert die Binomialverteilung. Die gesuchte Wahrscheinlichkeit errechnet sich zu

$$P(0 \leq x \leq 10) = \sum_{x=0}^{10} \binom{n}{x} p^x q^{n-x}$$

mit $n = 1000$, $p = 0,01$ und $q = 0,99$. Die Rechnung ist äußerst mühsam. Man erhält schließlich den exakten Wert

$$P(0 \leq x \leq 10) = 0,58304$$

Verwendet man die Normalverteilung als Näherung, so ist

$$P(0 \leq x \leq 10) \approx H(\beta) - H(\alpha)$$

Für die Integrationsgrenzen gilt 5.22. Man erhält

$$\alpha = \frac{0 - 10 - 0,5}{\sqrt{9,9}} = -3,337 \qquad \beta = \frac{10 - 10 + 0,5}{\sqrt{9,9}} = +0,159$$

Aus einer Tabelle der normierten Normalverteilung liest man ab

$$H(0,16) = 0,5636 \qquad H(-3,34) = 0,0006$$

Damit erhält man für die gesuchte Wahrscheinlichkeit den Näherungswert

$$P(0 \leq x \leq 10) \approx 0,5636 - 0,0006 = 0,5630$$

Die Normalverteilung ist in diesem Fall keine gute Näherungsfunktion, da die Variable x nur relativ kleine Werte annimmt. Eine bessere Näherung ist von der Poisson-Verteilung zu erwarten. Tatsächlich ergibt sich

$$P(0 \leq x \leq 10) \approx \sum_{x=0}^{10} \frac{(np)^x}{x!} \exp(-np)$$

Mit Zahlenwerten:

$$\begin{aligned} P(0 \le x \le 10) &\approx 4,540 \cdot 10^{-5} \cdot (1 + 10 + 50 + 167 + 417 + 833 + 1389 \\ &\quad +1984 + 2480 + 2756 + 2756) \\ &\approx 4,540 \cdot 10^{-5} \cdot 12843 = 0,58307 \end{aligned}$$

Dieser Wert stimmt mit dem exakten Wert praktisch überein.

Ende Beispiel

Mit Hilfe des integralen Grenzwertsatzes 5.21 lässt sich das **Gesetz der großen Zahlen** 3.27 auf Seite 32 in der Form

$$\lim_{z\to\infty} P\left\{\left|\frac{z(A)}{z} - p\right| < \epsilon\right\} = 1 \tag{5.23}$$

beweisen. Hierin ist A ein Ereignis, das bei einem Zufallsexperiment die Wahrscheinlichkeit

$$P(\mathrm{A}) = p$$

besitzt. Mit $z(\mathrm{A})$ wird die Anzahl des Eintreffens von A bei z voneinander unabhängigen Ausführungen des Zufallsexperimentes bezeichnet; ϵ ist eine beliebig kleine, aber von Null verschiedene positive Zahl. Mit den geänderten Bezeichnungen

$$z(\mathrm{A}) \equiv x \qquad z \equiv n$$

nimmt das Gesetz der großen Zahlen 3.27 folgende Form an

$$\lim_{n\to\infty} P\left\{\left|\frac{x}{n} - p\right| > \epsilon\right\} = 0$$

wofür man auch schreiben kann

$$\lim_{n\to\infty} P\left\{\left|\frac{x}{n} - p\right| < \epsilon\right\} = 1 \tag{5.24}$$

Die Ungleichung in 5.24 lässt sich wie folgt umformen

$$\begin{aligned} \left|\frac{x}{n} - p\right| &< \epsilon \\ -\epsilon < (\frac{x}{n} - p) &< +\epsilon \\ (p-\epsilon)n < x &< (p+\epsilon)n \\ a < x &< b \end{aligned} \tag{5.25}$$

wobei für die Bereichsgrenzen gilt

$$a = (p-\epsilon)n \qquad b = (p+\epsilon)n \tag{5.26}$$

Der integrale Grenzwertsatz 5.21 von Seite 82 lautet

$$P(a \leq x \leq b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \approx H(\beta) - H(\alpha)$$

mit den Bereichsgrenzen

$$\alpha = \frac{a - np - \frac{1}{2}}{\sqrt{npq}} \qquad \beta = \frac{b - np + \frac{1}{2}}{\sqrt{npq}} \tag{5.27}$$

Setzt man 5.26 in 5.27 ein, so erhält man

$$\alpha = -\frac{\epsilon n - \frac{1}{2}}{\sqrt{npq}} \qquad \beta = +\frac{\epsilon n + \frac{1}{2}}{\sqrt{npq}} \tag{5.28}$$

Für hinreichend große Werte von n gilt $\epsilon n >> 1/2$ und $\alpha = -\beta$. Damit und mit 5.25 bis 5.27 erhält man schließlich

$$\begin{aligned} P(\left|\frac{x}{n} - p\right| < \epsilon) &= P(a \leq x \leq b) \\ &\approx H(\beta) - H(\alpha) \approx 2H(\beta) - 1 \end{aligned} \tag{5.29}$$

Hiermit kann man die Wahrscheinlichkeit berechnen, dass die Abweichung der relativen Häufigkeit von der Wahrscheinlichkeit innerhalb vorgegebener Grenzen liegt. Im Grenzfall $n \to \infty$ folgt 5.23.

Beispiel 5.3 : Vor wichtigen politischen Entscheidungen, beispielsweise Parlamentswahlen, werden regelmäßig Umfragen veranstaltet, um ein Meinungsbild der Bevölkerung zu gewinnen. Dabei wird aus Kostengründen immer nur ein Teil der Bevölkerung befragt. Von dieser Teilmenge (Stichprobe) schließt man auf die Gesamtheit. Solche Schlüsse, beispielsweise Prognosen des Wahlverhaltens, sind jedoch wie alle Aussagen der Statistik mit einer gewissen Wahrscheinlichkeit falsch. Hierzu wird bei einer Veröffentlichung der Umfrageergebnisse häufig leider nichts oder nur wenig gesagt.

Geht man davon aus, dass die Stichprobe wirklich repräsentativ ist (keine einfache Aufgabe), kann man sich mit Hilfe des integralen Grenzwertsatzes in der Gestalt von 5.29 die fehlenden Informationen leicht verschaffen. Für die relative Häufigkeit x/n ist hier das Umfrageergebnis (Zustimmungsrate zu einer bestimmten Frage) einzusetzen, für p die unbekannte Zustimmungsrate in der Gesamtbevölkerung. Gesucht ist die maximale Abweichung δ, die zwischen beobachteter relativer Häufigkeit x/n (Umfrageergebnis) und Wahrscheinlichkeit p (Gesamtbevölkerung) mit einer bestimmten Wahrscheinlichkeit γ zu erwarten ist

$$P(\left|\frac{x}{n} - p\right| < \delta) = \gamma$$

Aus 5.29 folgt für genügend große Stichprobenumfänge n

$$\gamma \approx 2H(\beta) - 1$$

und aus 5.28

$$\beta \approx \frac{\delta n}{\sqrt{npq}}$$

oder

$$\delta \approx \beta \sqrt{\frac{p(1-p)}{n}}$$

Eine Aussage mit einer Wahrscheinlichkeit von $\gamma = 0,95$ erhält man laut Tabelle der normierten Normalverteilung, wenn man $\beta = 1,960$ setzt. Die maximale Abweichung δ des Umfrageergebnisses x/n vom wahren Wert p ist dann nur noch eine Funktion von p und n. Hierzu folgende Tabelle.

Tab. 5.1: Stichprobenumfang n, Zustimmungsrate p und Abweichung δ bei Umfragen, Aussagewahrscheinlichkeit 0,95 (95 %)

Probenumfang n	**Rate** p	**Abw.** δ	**Rate** p	**Abw.** δ
100	0,5	0,098	0,1	0,059
200	0,5	0,069	0,1	0,042
500	0,5	0,044	0,1	0,026
1000	0,5	0,031	0,1	0,019
2000	0,5	0,022	0,1	0,013
5000	0,5	0,014	0,1	0,008
10000	0,5	0,010	0,1	0,006

Hieraus kann man beispielsweise ablesen, dass bei einer Zustimmungsrate von $p = 0,5$ und einem Stichprobenumfang von $n = 100$ die maximalen Abweichungen des Stichprobenergebnisses x/n vom wahren Wert in der Bevölkerung mit einer Wahrscheinlichkeit von 0,05 oder 5 % größer als $\delta = 0,098$ sind, das heißt p liegt mit einer Wahrscheinlichkeit von 0,95 zwischen 0,402 (40 %) und 0,598 (60 %), mit einer Wahrscheinlichkeit von 0,05 jedoch außerhalb dieser Grenzen. Selbst bei einem Stichprobenumfang von 1000 lässt sich bei einer Zustimmungsrate von 0,5 (50 %) nur behaupten, dass der Wert in der Grundgesamtheit mit einer Wahrscheinlichkeit von 0,95 zwischen 0,456 (46 %, Koalition) und 0,544 (54 %, absolute Mehrheit) liegt. Immer vorausgesetzt, dass die Stichprobe eine zufällige Auswahl aus der Wählerschaft darstellt und die Befragten so wählen, wie sie geantwortet haben.

Für den relativen Fehler bei Umfrageergebnissen ergibt sich aus obiger Gleichung

$$\frac{\delta}{p} = \beta \sqrt{\frac{1-p}{pn}}$$

Ist p klein – beispielsweise bei kleinen Parteien – so wird der relative Fehler des Umfrageergebnisses groß.

Wie man sieht, müssen bei Umfragen in jedem Fall mindestens 5000 Personen befragt werden, um zu einem brauchbaren Ergebnis zu kommen, wobei immer noch eine endliche, wenn auch kleine Wahrscheinlichkeit besteht, dass die auf der Umfrage basierende Prognose falsch ist. Die meisten veröffentlichten Prognosen beruhen jedoch auf auf kleineren Stichproben, ohne dass dabei auf die deutlich geringere Zuverlässigkeit hingewiesen wird.

Bei solchen Umfragen oder Erhebungen wird jedoch nicht immer von Zufallsproben (Stichproben) ausgegangen, sondern oft von *repräsentativen Proben*. Das ist eine andere Vorgehensweise mit eigenen Problemen. Die Verfahren der deskriptiven Statistik – siehe Abschnitt 2 *Häufigkeit* ab Seite 7 sind anwendbar, die der induktiven Statistik nur sehr eingeschränkt. Weiteres entnehme man der deutschen Wikipedia unter dem Begriff *Repräsentatitivtät* und der englischen unter *Sampling (statistics)*.

Ende Beispiel

5.2.2 Logarithmische Normalverteilung

Als eine besondere Form der Normalverteilung kann man die **Logarithmische Normalverteilung** auffassen. Nicht die Zufallsgröße x ist normalverteilt, sondern ihr natürlicher Logarithmus. Es gilt

$$y = \ln x \tag{5.30}$$

und die Dichtefunktion lautet

$$\varphi(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2\Big) \tag{5.31}$$

Hierin sind μ_y und σ_y^2 Mittelwert und Varianz von $\ln x$. Fragt man nach Mittelwert und Varianz der Variablen x selbst, so ist dies gleichbedeutend mit der Aufgabe, Mittelwert und Varianz einer neuen Variablen

$$x = \exp y \tag{5.32}$$

zu bestimmen, wenn die Dichtefunktion der Variablen y gegeben ist. Für den Mittelwert ergibt sich mit 3.75 von Seite 44

$$\mu_x = E(\exp(y)) = \int_{-\infty}^{+\infty} \exp(y)\ \varphi(y)\ \mathrm{d}y = \frac{1}{\sigma_y\sqrt{2\pi}}\ \int_{-\infty}^{+\infty} \exp\Big(y - \frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2\Big)\ \mathrm{d}y$$

Mit den Substitutionen

$$t = \frac{y-\mu_y}{\sigma_y} \qquad u = t - \sigma_y$$

folgt daraus

$$\begin{aligned}
\mu_x &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\Big(\mu_y + \sigma_y t - \frac{t^2}{2}\Big)\, \mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}(t-\sigma_y)^2 + \frac{1}{2}\sigma_y^2 + \mu_y\Big)\, \mathrm{d}t \\
&= \exp\Big(\mu_y + \frac{\sigma_y^2}{2}\Big) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\Big(-\frac{u^2}{2}\Big)\, \mathrm{d}u \\
&= \exp\Big(\mu_y + \frac{\sigma_y^2}{2}\Big) \qquad (5.33)
\end{aligned}$$

Für die Varianz ergibt sich in ähnlicher Rechnung

$$\sigma_x^2 = \mu_x^2(\exp(\sigma_y^2) - 1) \qquad (5.34)$$

Ist $\sigma_y^2 \ll 1$, dann gilt näherungsweise

$$\exp(\sigma_y^2) \approx 1 + \sigma_y^2 \qquad \exp\Big(\mu_y + \frac{\sigma_y^2}{2}\Big) \approx \exp(\mu_y)\,\Big(1 + \frac{\sigma_y^2}{2}\Big)$$

Damit folgt

$$\mu_x \approx \Big(1 + \frac{\sigma_y^2}{2}\Big)\, \exp(\mu_y) \qquad \sigma_x^2 \approx \sigma_y^2\, \exp(2\mu_y)$$

Zum gleichen Ergebnis kommt man, wenn man die Gleichungen 3.141 (Seite 55) und 3.154 (Seite 57) zur näherungsweisen Berechnung von Mittelwert und Varianz einer neuen Zufallsvariablen v aus Abschnitt 3.4.3 *Rechenregeln für Erwartungswerte und Varianzen* benutzt. Für den Mittelwert μ_x erhält man aus

$$\mu_x \approx g(\mu_y) + \frac{1}{2} g''(\mu_y)\sigma_y^2$$

mit

$$g(y) = \exp y \qquad g'(y) = \exp y \qquad g''(y) = \exp y$$

als Ergebnis

$$\mu_x \approx \Big(1 + \frac{\sigma_y^2}{2}\Big)\, \exp \mu_y$$

Für die Varianz σ_x^2 folgt aus

$$\sigma_x^2 \approx (g'(\mu_y))^2\, \sigma_y^2$$

das Ergebnis

$$\sigma_x^2 \approx \sigma_y^2\, \exp(2\mu_y)$$

Die Logarithmische Normalverteilung kommt für die Darstellung von Grundgesamtheiten in solchen Fällen in Betracht, in denen die Zufallsvariable keine negativen Werte annehmen kann, beispielsweise bei der Darstellung von Partikelgrößenverteilungen. Bei diesen ist folgende Bezeichnungsweise üblich

$$\varphi(\ln x) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\Big(-\frac{1}{2}\Big(\frac{\ln x - \ln x_{50}}{\sigma}\Big)^2\Big) \quad \text{mit} \quad \mu_y = \ln x_{50} \quad \sigma_y = \sigma$$

Die Darstellung Logarithmischer Normalverteilungen ist für die Dispersitätsanalyse in DIN 66144 genormt. Das in der Norm wiedergegebene Netz verfügt über Randskalen, die einige Rechnungen erleichtern. Eine Besonderheit der Logarithmischen Normalverteilung ist, dass sich die Umrechnung von Mengenarten (siehe Abschnitt 9.8 *Umrechnung verschiedener Mengenarten* auf Seite 173) als Parallelverschiebung der Verteilungsgeraden darstellt.

5.2.3 Approximation der Normalverteilung durch eine Fourierreihe

Gegeben sei eine Funktion $f(x)$ im Bereich von $-\pi$ bis $+\pi$. Ist $f(x)$ symmetrisch, das heißt gilt $f(-x) = f(x)$, so kann man folgende, nach Joseph Fourier (1768–1830) benannte Reihenentwicklung ansetzen

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{+\infty} a_n\ \cos nx \tag{5.35}$$

Die Konstanten a_n erhält man aus

$$a_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(x)\ \cos nx\ \mathrm{d}x \qquad n = 0, 1, 2, \ldots \tag{5.36}$$

Im speziellen Fall der Normalverteilung gilt

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\ \exp\,(-\frac{x^2}{2}) \tag{5.37}$$

Eine Approximation wird nur im Bereich von $-\pi$ bis $+\pi$ gesucht. Außerhalb des Bereiches ist $\varphi(x)$ sehr klein und daher praktisch bedeutungslos. Für $n = 0$ ergibt sich

$$a_0 = \frac{1}{\pi}\ \frac{1}{\sqrt{2\pi}}\ \int_{-\pi}^{+\pi} \exp\,(-\frac{x^2}{2})\ \mathrm{d}x = \frac{0,9984}{\pi} \approx \frac{1}{\pi} \tag{5.38}$$

Für die Werte $n = 1, 2, \ldots$ gilt

$$a_n = \frac{1}{\pi}\ \frac{1}{\sqrt{2\pi}}\ \int_{-\pi}^{+\pi} \exp\,(-\frac{x^2}{2})\ \cos nx\ \mathrm{d}x \tag{5.39}$$

Aus Formelsammlungen (beispielsweise *Handbook of Chemistry and Physics* oder *Taschenbuch der Mathematik* von I. N. Bronstein und K. A. Semendjajew) entnehmen wir

$$\int_0^{+\infty} \exp\,(-\frac{x^2}{2})\ \cos nx\ \mathrm{d}x = \sqrt{\frac{\pi}{2}}\ \exp\,(-\frac{n^2}{2}) \tag{5.40}$$

Damit folgt aus 5.36

$$a_n \approx \frac{1}{\pi}\ \exp\,(-\frac{n^2}{2}) \tag{5.41}$$

und weiter für $\varphi(x)$

$$\varphi(x) \approx \frac{1}{\pi}\ \Big(\frac{1}{2} + \sum_{n=1}^{\infty} \exp\,(-\frac{n^2}{2})\ \cos nx\Big) \tag{5.42}$$

Ist eine unendliche Reihe konvergent, darf sie gliedweise integriert werden. Daher gilt

$$\begin{aligned}
\Phi(x_0) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_0} \exp\,(-\frac{x^2}{2})\ \mathrm{d}x \\
&= \int_{-\infty}^{x_0} \varphi(x)\ \mathrm{d}x \\
&\approx \frac{1}{\pi} \int_{-\pi}^{x_0} \Big(\frac{1}{2} + \sum_{n=1}^{+\infty} \exp\,(-\frac{n^2}{2})\ \cos nx\Big)\ \mathrm{d}x \\
&\approx \frac{1}{\pi} \Big|\frac{x}{2} + \sum_{n=1}^{+\infty} \frac{1}{n} \exp\,(-\frac{n^2}{2})\ \sin nx\Big|_{-\pi}^{x_0} \\
&\approx \frac{1}{\pi} \Big(\frac{x_0+\pi}{2} + \sum_{n=1}^{+\infty} \frac{1}{n}\ \exp\,(-\frac{n^2}{2})\ \sin nx_0\Big)
\end{aligned} \tag{5.43}$$

An den Bereichsgrenzen ist die Approximation nicht ganz so gut wie im übrigen Bereich.

Nimmt man geringe Abweichungen in Kauf, kann man sich mit einer verkürzten Formel begnügen

$$\Phi(x) \approx \frac{1}{\pi}\Big(\frac{x+\pi}{2} + \sum_{n=1}^{3} \frac{1}{n} \exp\,(-\frac{n^2}{2})\ \sin nx\Big) \tag{5.44}$$

Das ist die in 5.9 auf Seite 77 angegebene Näherung für die Summenfunktion der normierten Normalverteilung, deren Genauigkeit für die meisten Anwendungen ausreicht.

5.3 Zweidimensionale Normalverteilung

Die zweidimensionale (bivariate) Normalverteilung ist eine stetige zweidimensionale Verteilung mit der Dichtefunktion

$$\varphi(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\ \exp\Big(-\frac{1}{2}f(x,y)\Big) \tag{5.45}$$

worin $f(x,y)$ eine Abkürzung ist für den Ausdruck

$$f(x,y) = \frac{1}{1-\rho^2}\left(\Big(\frac{x-\mu_x}{\sigma_x}\Big)^2 - 2\rho\Big(\frac{x-\mu_x}{\sigma_x}\Big)\Big(\frac{y-\mu_y}{\sigma_y}\Big) + \Big(\frac{y-\mu_y}{\sigma_y}\Big)^2\right) \tag{5.46}$$

Die zugehörige Summenfunktion ist

$$\Phi(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{\eta=-\infty}^{\eta=y}\int_{\xi=-\infty}^{\xi=x} \exp\Big(-\frac{1}{2}f(\xi,\eta)\Big)\,\mathrm{d}\xi\mathrm{d}\eta \tag{5.47}$$

Durch die Substitutionen

$$t_1 = \frac{x-\mu_x}{\sigma} \qquad t_2 = \frac{y-\mu_y}{\sigma_y} \tag{5.48}$$

wird das Integral in 5.47 auf das spezielle Integral

$$\Phi^\star(t_1,t_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{v=-\infty}^{v=t_2}\int_{u=-\infty}^{u=t_1} \exp\Big(-\frac{u^2-2\rho uv+v^2}{2(1-\rho^2)}\Big)\,\mathrm{d}u\mathrm{d}v \tag{5.49}$$

zurückgeführt, das nur den Parameter ρ enthält.

Die Dichtefunktionen der beiden Randverteilungen bezüglich der gegebenen Verteilung $\varphi(x,y)$ sind

$$\varphi_1(x) = \frac{1}{\sigma_x\sqrt{2\pi}}\ \exp\Big(-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2\Big) \tag{5.50}$$

$$\varphi_2(y) = \frac{1}{\sigma_y\sqrt{2\pi}}\ \exp\Big(-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2\Big) \tag{5.51}$$

wie man durch Einsetzen von 5.45 und 5.46 in die Definitionsgleichungen

$$\varphi_1(x) = \int_{-\infty}^{+\infty} \varphi(x,y)\,\mathrm{d}y \qquad \varphi_2(y) = \int_{-\infty}^{+\infty} \varphi(x,y)\,\mathrm{d}x$$

zeigt. Für die Mittelwerte μ_x und μ_y gilt

$$\mu_x = E(x) = \int_{-\infty}^{+\infty} x\ \varphi_1(x)\,\mathrm{d}x \qquad \mu_y = E(y) = \int_{\infty}^{+\infty} y\ \varphi_2(y)\,\mathrm{d}y \tag{5.52}$$

und für die Varianzen σ_x^2 und σ_y^2

$$\begin{aligned}\sigma_x^2 &= E((x-\mu_x)^2) = \int_{-\infty}^{+\infty} (x-\mu_x)^2\ \varphi_1(x)\,\mathrm{d}x \\ \sigma_y^2 &= E((y-\mu_y)^2) = \int_{-\infty}^{+\infty} (y-\mu_y)^2\ \varphi_2(y)\,\mathrm{d}y\end{aligned} \tag{5.53}$$

Für die Kovarianz gilt allgemein

$$\sigma_{xy} = E((x-\mu_x)(y-\mu_y))$$

Das Doppelintegral ist in diesem Fall nicht ganz einfach zu lösen. Nach umfangreichen Rechnungen findet man

$$\sigma_{xy} = \int_{y=-\infty}^{y=+\infty} \int_{x=-\infty}^{x=+\infty} (x-\mu_x)(y-\mu_y)\ \varphi(x,y)\ \mathrm{d}x\mathrm{d}y = \rho\ \sigma_x\sigma_y \tag{5.54}$$

Es besteht somit der Zusammenhang

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \tag{5.55}$$

Man bezeichnet ρ als den **Korrelationskoeffizienten der Grundgesamtheit**.

Sind zwei Zufallsvariable x und y unabhängig voneinander, so ist ihre Kovarianz σ_{xy} gleich null; sie sind unkorreliert. Im Allgemeinen gilt die Umkehrung dieser Aussage nicht. Im Fall der zweidimensionalen Normalverteilung gilt jedoch: Zwei normalverteilte Zufallsvariablen sind dann und nur dann unabhängig voneinander, wenn der zugehörige Korrelationskoeffizient gleich null ist. Dies folgt unmittelbar, wenn man in 5.45 und 5.46 den Korrelationskoeffizienten ρ gleich null setzt. Man erhält mit 5.50 und 5.51

$$\varphi(x,y) = \varphi_1(x)\ \varphi_2(y)$$

womit die Unabhängigkeit bewiesen ist.

Neben den beiden Randverteilungen kann man zu jeder zweidimensionalen Verteilung bedingte Verteilungen berechnen. Dies sind Verteilungen der einen Variablen, wenn der Wert der anderen Variablen festgehalten wird, wie schon auf Seite 43 erklärt. Allgemein gilt

$$\varphi(x,y_0) = \varphi_2(y_0)\ \varphi_1^\star(x|y_0) \tag{5.56}$$

Hierin ist $\varphi_1^\star(x|y_0)$ die bedingte Verteilung der Variablen x, wenn die Variable y den festen Wert y_0 annimmt (x vorausgesetzt y_0). Im Fall der zweidimensionalen Normalverteilung folgt mit 5.48

$$\begin{aligned}
\varphi(t_1,t_{2,0}) &= h(t_{2,0})\ \varphi_1^\star(t_1|t_{2,0}) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}}\ \exp\Big(-\frac{t_1^2-2\rho t_1 t_{2,0}+t_{2,0}^2}{2(1-\rho^2)}\Big) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}}\ \exp\Big(-\frac{t_{2,0}^2(1-\rho^2)+(t_1-\rho t_{2,0})^2}{2(1-\rho^2)}\Big) \\
&= \frac{1}{\sqrt{2\pi}}\ \exp\Big(-\frac{t_{2,0}^2}{2}\Big)\ \frac{1}{\sqrt{1-\rho^2}\sqrt{2\pi}}\ \exp\Big(-\frac{1}{2}\Big(\frac{t_1-\rho t_{2,0}}{\sqrt{1-\rho^2}}\Big)^2\Big)
\end{aligned}$$

und daraus

$$\varphi_1^\star(t_1|t_{2,0}) = \frac{1}{\sqrt{1-\rho^2}\sqrt{2\pi}}\ \exp\Big(-\frac{1}{2}\Big(\frac{t_1-\rho t_{2,0}}{\sqrt{1-\rho^2}}\Big)^2\Big)$$

Mittelwert und Varianz dieser Verteilung sind

$$E(t_1|t_{2,0}) = \rho t_{2,0} \qquad V(t_1|t_{2,0}) = 1 - \rho^2$$

Nach Rücktransformation erhält man

$$\varphi_1^\star(x|y_0) = \frac{1}{\sigma_x\sqrt{1-\rho^2}\sqrt{2\pi}} \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu_x-\rho\sigma_x t_{2,0}}{\sigma_x\sqrt{1-\rho^2}}\Big)^2\Big) \tag{5.57}$$

und für Mittelwert und Varianz

$$E(x|y_0) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y_0-\mu_x) \qquad V(x|y_0) = \sigma_x^2(1-\rho^2) \tag{5.58}$$

5.4 Chi-Quadrat-Verteilung

Gegeben seien n voneinander unabhängige Zufallsvariablen $x_1, x_2, \ldots x_n$, deren jede eine Normalverteilung mit dem Mittelwert 0 und der Varianz 1 besitzt. Die Summe der Quadrate dieser Variablen ist eine neue Zufallsvariable

$$\chi^2 = x_1^2 + x_2^2 + \ldots + x_n^2 \tag{5.59}$$

deren Verteilung von Friedrich Robert Helmert (1843–1917) im Rahmen geodätischer Aufgaben untersucht worden ist. Sie wird als **Chi-Quadrat-Verteilung** oder auch als Helmert-Pearson-Verteilung (nach Karl Pearson) bezeichnet und hat die Dichtefunktion

$$\varphi(\chi^2) = K_n\,(\chi^2)^{(n-2)/2}\,\exp(-\chi^2/2) \tag{5.60}$$

Wegen 5.59 kann die Variable χ^2 keine negativen Werte annehmen. In 5.60 ist n eine positive ganze Zahl, die man als die **Anzahl der Freiheitsgrade** der Verteilung bezeichnet.

In Abbildung 5.2 ist $\varphi(\chi^2)$ für verschiedene Freiheitsgrade n aufgetragen. Für $n = 1$ und $n = 2$ fallen die Kurven monoton. Für $n > 2$ erreichen sie ein Maximum bei $\chi^2 = n - 2$.

Die Summenfunktion zu 5.60 lautet

$$\Phi(\chi^2) = K_n \int_0^{\chi^2} u^{(n-2)/2}\,\exp(-u/2)\,\mathrm{d}u \tag{5.61}$$

Da das Integral in 5.61 nicht analytisch auswertbar ist, ist die Summenfunktion $\Phi(\chi^2)$ tabelliert worden. Der Wert der Konstanten K_n in 5.60 und 5.61 errechnet sich aus der Bedingung

$$\int_0^{+\infty} \varphi(\chi^2)\,\mathrm{d}\chi^2 = 1$$

Mit 5.60 folgt daraus

$$K_n \int_0^{+\infty} (\chi^2)^{(n-2)/2}\,\exp\Big(\frac{\chi^2}{2}\Big)\,\mathrm{d}\chi^2 = 1$$

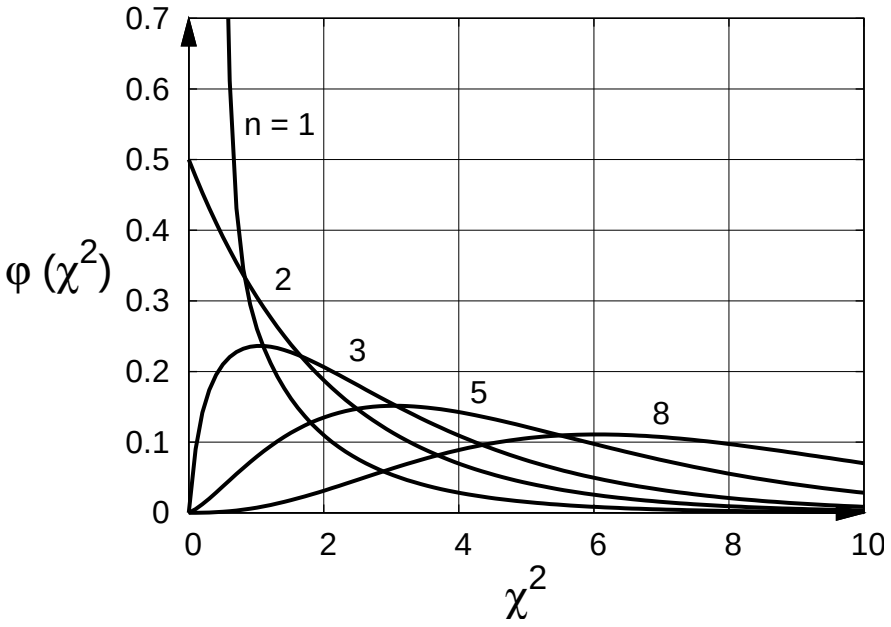


Abb. 5.2: Dichte der Chi-Quadrat-Verteilung für verschiedene Freiheitsgrade n

Wir substituieren

$$\chi^2 = 2t$$

und erhalten

$$K_n \int_0^{+\infty} 2^{(n-2)/2}\, t^{(n-2)/2}\, \exp(-t)\, 2\, \mathrm{d}t = K_n\, 2^{n/2} \int_0^{+\infty} \exp(-t)\, t^{(n-2)/2}\, \mathrm{d}t = 1$$

Das bestimmte Integral ist nicht analytisch lösbar. Es ist unter dem Namen (Eulersche) **Gammafunktion** bekannt, und zwar in der Form

$$\Gamma(\alpha) = \int_0^{+\infty} \exp(-t)\, t^{\alpha-1}\, \mathrm{d}t \qquad \text{für} \quad \alpha \in \mathbb{R} \quad \text{mit} \quad \alpha > 0 \tag{5.62}$$

Es gilt die Rekursionsformel

$$\Gamma(\alpha + 1) = \alpha\, \Gamma(\alpha) \tag{5.63}$$

Die Gammafunktion lässt sich als eine Erweiterung der Fakultätsfunktion verstehen. Zahlenwerte entnimmt man Tabellen, beispielsweise im *Taschenbuch der Mathematik* von I. N. Bronstein und K. A. Semendjajew.[1] Eine Näherung erhält man mit Hilfe der Stirlingschen Formel (siehe 4.18 auf Seite 68)

$$\Gamma(\alpha) \approx \sqrt{2\pi}\, \alpha^{\alpha-1/2}\, \exp(-\alpha)$$

Mit 5.63 ergibt sich schließlich folgender Ausdruck für die Konstante K_n

$$K_n = \frac{1}{2^{n/2}\, \Gamma(n/2)} \tag{5.64}$$

[1] zumindest in älteren Auflagen

Bei der Berechnung des Erwartungswertes von χ^2 geht man ähnlich vor. Aus

$$E(\chi^2) = \mu_{\chi^2} = \int_0^{+\infty} \chi^2 \, \varphi(\chi^2) \, \mathrm{d}\chi^2$$

folgt unter Verwendung von 5.62 und 5.63

$$E(\chi^2) = \mu_{\chi^2} = n \tag{5.65}$$

Auf gleiche Weise findet man

$$\sigma^2_{\chi^2} = 2n \tag{5.66}$$

Wie man zeigen kann, lässt sich die Chi-Quadrat-Verteilung für große Werte von n durch die normierte Normalverteilung brauchbar annähern. Es gilt

$$\Phi(\chi^2) \approx H\Big(\frac{\chi^2 - n}{\sqrt{2n}}\Big) \tag{5.67}$$

das heißt, die Zufallsvariable χ^2 ist asymptotisch normalverteilt mit dem Mittelwert $\chi^2 = n$ und der Varianz $\sigma^2_{\chi^2} = 2n$.

5.5 Student-Verteilung

Die **Student-Verteilung** oder **t-Verteilung**, auch Studentsche t-Verteilung genannt, ist von dem englischen Statistiker William Sealey Gosset (1876–1937) eingeführt worden, der unter dem Pseudonym *Student* veröffentlichte. Die Student-Verteilung ist die Verteilung der Zufallsvariablen

$$t = \frac{x}{\sqrt{y/n}} \tag{5.68}$$

worin x und y zwei voneinander unabhängige Zufallsvariablen sind. Vorausgesetzt wird, dass die Variable x normalverteilt ist mit dem Mittelwert 0 und der Varianz 1, und dass die Variable y eine χ^2-Verteilung mit n Freiheitsgraden besitzt.

Die Student-Verteilung hat die Dichtefunktion

$$\varphi(t) = C_n \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} \tag{5.69}$$

Die Konstante C_n ergibt sich mit der Gammafunktion zu

$$C_n = \frac{\Gamma\big(\frac{n+1}{2}\big)}{\sqrt{n\pi}\,\Gamma\big(\frac{n}{2}\big)}$$

Die Zufallsvariable t kann jeden Wert von $-\infty$ bis $+\infty$ annehmen.

In Diagramm 5.3 ist $\varphi(t)$ für einige Werte von n aufgetragen. Man erhält Kurven, die der normierten Normalverteilung ähnlich sind, nur flacher und breiter verlaufen als diese. Mit wachsendem n strebt die Student-Verteilung gegen die

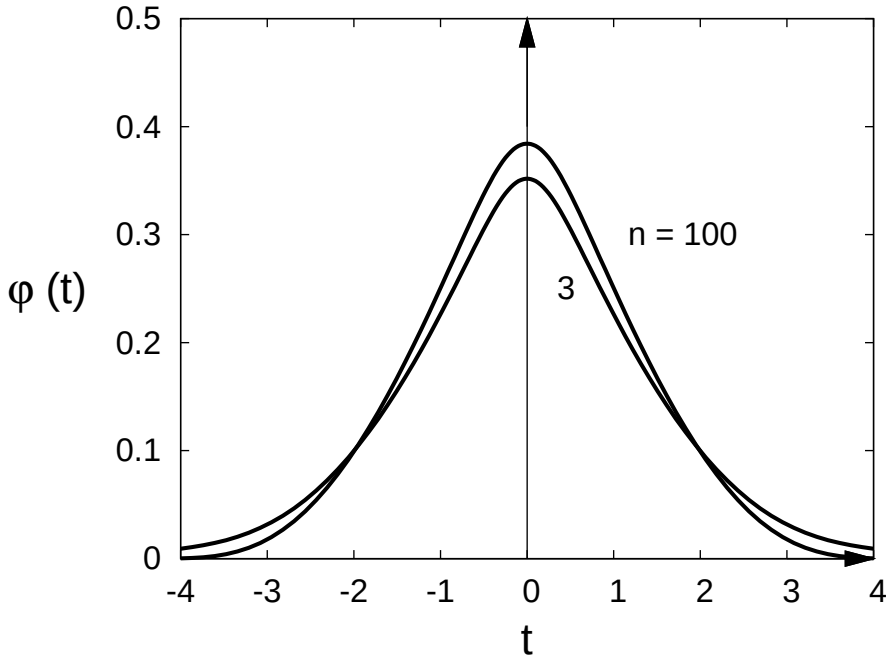


Abb. 5.3: Dichtefunktion der Student-Verteilung für einige Werte von n

normierte Normalverteilung. Die Glockenkurve der normierten Normalverteilung ist Grenzkurve einer Kurvenschar mit dem Parameter n.

Die Summenfunktion der Student-Verteilung ist

$$\Phi(t) = C_n \int_{-\infty}^{t} \Big(1 + \frac{u^2}{n}\Big)^{-(n+1)/2} \, \mathrm{d}u \tag{5.70}$$

Wegen der Symmetrie von $\varphi(t)$ hat die Student-Verteilung den Mittelwert

$$\mu_t = 0 \quad \text{für} \quad n > 1 \tag{5.71}$$

Für $n = 1$ und $n = 2$ hat die Student-Verteilung keine Varianz. Für $n > 2$ erhält man

$$\sigma_t^2 = \frac{n}{n-2} \quad \text{für} \quad n > 2 \tag{5.72}$$

Da das Integral in 5.70 nicht analytisch auswertbar ist, ist die Summenfunktion $\Phi(t)$ tabelliert worden.

Auch die Student-Verteilung kann für große Werte von n durch die normierte Normalverteilung brauchbar angenähert werden, indem man setzt

$$\Phi(t) \approx H\Big(\frac{t}{\sqrt{\frac{n}{n-2}}}\Big)$$

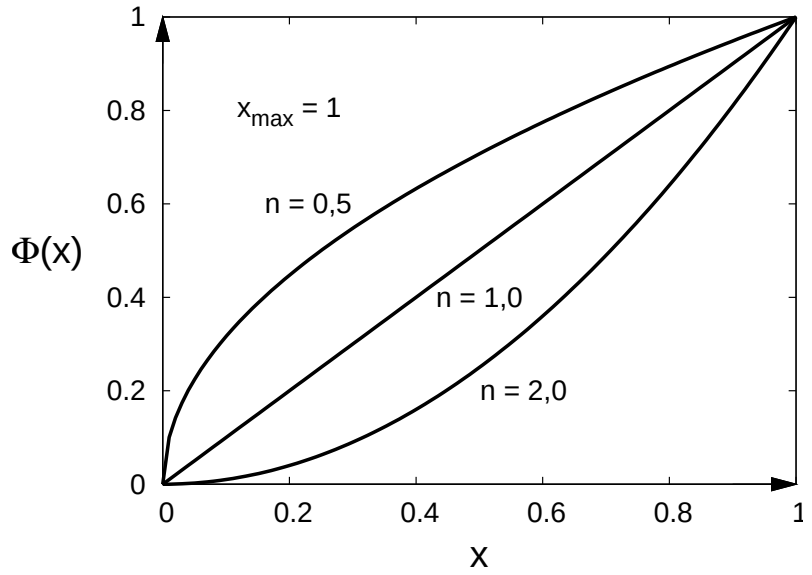


Abb. 5.4: Summenfunktion der Potenzverteilung mit $x_{max} = 1,0$ und verschiedenen Werten von n

5.6 Potenzverteilung

Die auf Antoine Marc Gaudin (1900–1974) und Reinhardt Schuhmann Jr. (1915–1996) zurückgehende **Potenzverteilung** ist eine stetige Verteilung mit der Dichtefunktion

$$\varphi(x) = \begin{cases} \frac{n}{x_{max}} \, (\frac{x}{x_{max}})^{n-1} & \text{für} \quad x \leq x_{max} \qquad n > 0 \\ 0 & \text{für} \quad x > x_{max} \end{cases} \tag{5.73}$$

und der Summenfunktion

$$\Phi(x) = \begin{cases} (\frac{x}{x_{max}})^{n} & \text{für} \quad x \leq x_{max} \qquad n > 0 \\ 1 & \text{für} \quad x > x_{max} \end{cases} \tag{5.74}$$

Gelegentlich findet sich auch die Bezeichnung Gates-Gaudin-Schuhmann-Verteilung (GGS) oder Andreasen-Verteilung (nach Alfred Herman Munch Andreasen, 1896–1978). Die Größe x_{max} ist ein Lageparameter, der Exponent n ein Streuungsparameter. Wegen ihrer einfachen mathematischen Form wird sie gern für Modellrechnungen herangezogen. Die Darstellung ist in DIN 66143 für die Volumen- oder Massenverteilungen aus Dispersitätsanalysen genormt.

Der Erwartungswert von x ergibt sich wie folgt

$$\begin{aligned} E(x) &= \mu(x) = \int_0^{x_{max}} x \, \varphi(x) \, \mathrm{d}x \\ &= \int_0^{x_{max}} n \, (\frac{x}{x_{max}})^{n} \, \mathrm{d}x \\ &= \left| \frac{n}{n+1} \, \frac{x^{n+1}}{x_{max}^{n}} \right|_0^{x_{max}} \end{aligned}$$

$$= \frac{n}{n+1}\, x_{max} \tag{5.75}$$

Für den Erwartungswert von x^2 erhält man entsprechend

$$E(x^2) = \frac{n}{n+2}\, x_{max}^2 \tag{5.76}$$

und damit für die Varianz

$$\sigma_x^2 = E(x^2) - (E(x))^2 = (\frac{n}{n+2} - \frac{n^2}{(n+1)^2})\, x_{max}^2$$

$$= \frac{n}{(n+1)^2(n+2)}\, x_{max}^2 \tag{5.77}$$

Die Summenfunktion $\Phi(x)$ erscheint bei Auftragung im doppeltlogarithmischen Netz (beide Achsen logarithmisch geteilt) als Gerade, siehe DIN 66143.

An ihrem feinen Ende $x/x_{max} << 1$ geht die RRSB-Verteilung (Weibull-Verteilung, siehe 5.81) in eine Potenzverteilung über. Der Unterschied beider Verteilungen liegt also hauptsächlich am groben Ende, wo die Potenzverteilung mit einem unschönen Knick in $\Phi(x > x_{max}) = 1$ einmündet.

5.7 Exponentielle Verteilungen

Die **Weibull-Verteilung**, benannt nach dem schwedischen Mathematiker und Ingenieur Waloddi Weibull (1887–1979), ist eine stetige Wahrscheinlichkeitsverteilung, die sich unter anderem für die Beschreibung der Materialermüdung spröder Werkstoffe und der Lebensdauer technischer Systeme verwenden lässt.

Ihre Dichteverteilung $\varphi(x)$ ist gegeben durch

$$\varphi(x) = abx^{b-1}\, \exp(-ax^b) \tag{5.78}$$

oder anders geschrieben

$$\varphi(x) = \frac{b}{T}\, (\frac{x}{T})^{b-1}\, \exp(-(\frac{x}{T})^b) \qquad \text{mit} \qquad a = \frac{1}{T^b} \tag{5.79}$$

Die Variable x kann nur Werte > 0 annehmen, ebenso die Konstanten a, b und T. Die Summenfunktion $\Phi(x)$ ergibt sich zu

$$\Phi(x) = \int_0^x \varphi(\xi)\, \mathrm{d}\xi = 1 - \exp(-ax^b) \tag{5.80}$$

beziehungsweise

$$\Phi(x) = 1 - \exp(-(\frac{x}{T})^b) \tag{5.81}$$

Für $x = T$ ergibt sich

$$\Phi(T) = 1 - \exp(-1) = 0,632 \tag{5.82}$$

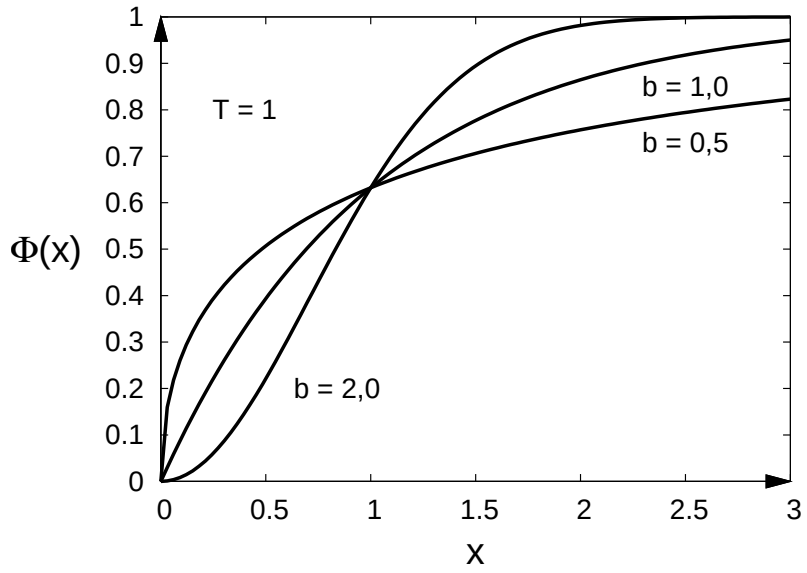


Abb. 5.5: Summenfunktion der Weibull-Verteilung mit $T = 1,0$ und verschiedenen Werten von b

Der zu $\Phi(x) = 0,632$ gehörende Wert von T ist daher ein Lageparameter der Weibull-Verteilung.

Den Erwartungwert von x erhält man aus

$$E(x) = \mu_x = \int_0^{+\infty} x\ \varphi(x)\ \mathrm{d}x \tag{5.83}$$

Mit der Substitution

$$t = (\frac{x}{T})^b \qquad \mathrm{d}t = \frac{b}{T}\ (\frac{x}{T})^{b-1}\ \mathrm{d}x$$

folgt aus 5.83

$$E(x) = \mu_x = T\ \int_0^{+\infty} t^{1/b}\ \exp{(-t)}\ \mathrm{d}t = T\ \Gamma(1 + \frac{1}{b}) \tag{5.84}$$

mit der auf Seite 94 eingeführten Gamma-Funktion. Entsprechend ergibt sich für den Erwartungswert von x^2

$$E(x^2) = T^2\ \Gamma(1 + \frac{2}{b}) \tag{5.85}$$

und damit für die Varianz

$$\sigma_x^2 = T^2\ \Gamma(1 + \frac{2}{b}) - T^2(\Gamma(1 + \frac{1}{b}))^2 = T^2\ (\Gamma(1 + \frac{2}{b}) - (\Gamma(1 + \frac{1}{b}))^2) \tag{5.86}$$

Die Größe b ist ein Streuungsparameter der Weibull-Verteilung und wird in manchen Zusammenhängen als Weibull-Modul bezeichnet.

Für $b = 1$ folgt aus 5.79 und 5.81 als Sonderfall die **Exponentialverteilung**

$$\begin{aligned} \varphi(x) &= \frac{1}{T}\ \exp\,(-\frac{x}{T}) \\ \Phi(x) &= 1 - \exp\,(-\frac{x}{T}) \end{aligned} \tag{5.87}$$

In der Dispersitätsanalyse kennt man die von Paul Rosin (1890–1967), Erich Rammler (1901–1986) und K. Sperling entwickelte RRS-Verteilung, meist in der von John Godolphin Bennett (1897–1974) eingeführten Schreibweise (**RRSB-Verteilung**). Sie ist nichts anderes als eine Weibull-Verteilung

$$q_3(x) = \frac{n}{x'}\ \Big(\frac{x}{x'}\Big)^{n-1}\ \exp\Big(-(\frac{x}{x'})^n\Big) \tag{5.88}$$

$$Q_3(x) = D(x) = 1 - \exp\Big(-(\frac{x}{x'})^n\Big) \tag{5.89}$$

mit

D	Durchgang (Siebdurchgang, Massenverteilungssumme)
x	Partikelgröße
x'	Lageparameter, $D(x') = 0,632$, in 5.81 als T bezeichnet
n	Streuungsparameter, in 5.81 als b bezeichnet

In dem für die Darstellung von Partikelgrößenverteilungen entwickelten RRSB-Netzpapier nach DIN 66145, verwendet vor allem in der Kohleaufbereitung, wird die Summenfunktion einer RRSB-Verteilung als Gerade wiedergegeben; ihre Parameter können in dem Netzpapier abgelesen werden. Ein Vorgänger des RRSB-Netzes nach DIN 66145 war das *Doppeltlogarithmische Körnungsnetz* nach DIN 4190. Das Netz lässt sich nicht nur für Partikelgrößenverteilungen verwenden, sondern auch für beliebige andere Anwendungen der Weibull-Verteilung.

Es ist ein verbreiteter Irrtum zu meinen, *jede* Partikelgrößenverteilung müsse sich im RRSB-Netz oder Körnungsnetz als Gerade darstellen. Die RRSB-Verteilung ist kein Naturgesetz. Es gibt auch keine verfahrenstechnische Maschine, die – warum auch immer – RRSB-Verteilungen erzeugt.

Netzpapiere gibt es auch für die Gauß-Verteilung und für die logarithmische Normalverteilung, die in dem jeweiligen Netz als Gerade abgebildet werden. Der Vorteil solcher Netzpapiere besteht darin, dass man durch Eintragen der Daten einer gemessenen Summenverteilung (Stichprobe) schnell entscheiden kann, ob diese durch die jeweils zugehörige Wahrscheinlichkeitsverteilung brauchbar angenähert werden kann. Die Bedeutung dieser speziellen Netze hat mit dem Aufkommen der Computer und numerischer Verfahren abgenommen.

Kapitel 6

Konfidenzintervalle für Verteilungsparameter

Hier geht es um die Zuverlässigkeit statistischer Aussagen. Wir sehen, dass es keine sicheren statistischen Aussagen gibt, sondern nur solche, die mit einer bestimmten Wahrscheinlichkeit zutreffen.

6.1 Grundbegriffe

Häufig besteht der Wunsch, mit Hilfe einer Stichprobe Näherungswerte für unbekannte Konstanten zu gewinnen, die als Parameter in der jeweiligen Verteilungsfunktion der Grundgesamtheit enthalten sind. Wir wollen Parameter der Grundgesamtheit auf Grund von Stichprobenergebnissen *schätzen*. Solche Parameter sind im Falle der Normalverteilung einer Variablen x deren Mittelwert μ_x und deren Varianz σ_x^2. Man kann sich jedoch auch andere Verteilungsfunktionen vorstellen, in denen Mittelwert μ_x und Varianz σ_x^2 als Parameter auftreten. Schließlich gibt es Verteilungsfunktionen mit anderen Parametern als Mittelwert und Varianz. Es brauchen auch nicht immer zwei Parameter zu sein, siehe Poisson-Verteilung. Die Verbindung von einer Stichprobe zur Grundgesamtheit herzustellen, ist Aufgabe der schließenden oder **induktiven Statistik**.

Es liegt nahe, den Mittelwert $\bar{x}$ einer Stichprobe als Näherung für den Mittelwert μ_x der Verteilung der zugehörigen Grundgesamtheit anzusehen. Man erhält damit die **Schätzung** (Schätzwert, Schätzer, E: estimator)

$$\mu_x \approx \bar{x} \qquad \text{mit} \qquad \bar{x} = \frac{1}{z} \sum_{i=1}^{z} x_i \tag{6.1}$$

Entsprechend kann man die Varianz s_x^2 einer Stichprobe als Näherung für die Varianz σ_x^2 der zugehörigen Grundgesamtheit auffassen. Dann ist

$$\sigma_x^2 \approx s_x^2 \qquad \text{mit} \qquad s_x^2 = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \tag{6.2}$$

Ist der Mittelwert μ_x der Verteilung der Grundgesamtheit bekannt, so verwendet man statt 6.2 die Schätzung

$$\sigma_x^2 \approx \tilde{s}_x^2 \qquad \text{mit} \qquad \tilde{s}_x^2 = \frac{1}{z} \sum_{i=1}^{z} (x_i - \mu_x)^2 \tag{6.3}$$

Im Fall der Poisson-Verteilung ist $\mu_x = \sigma_x^2$. Man kann dann entweder 6.1 oder 6.2 als Schätzung für den Parameter dieser Verteilung benutzen. Ganz allgemein hat man stets die Wahl zwischen mehreren Schätzungen.

Im Fall der Binomialverteilung ist der einzige Parameter die Wahrscheinlichkeit p. Diese ist durch $\mu_x = np$ mit dem Mittelwert verknüpft. Damit erhält man für p die Schätzung

$$p \approx \frac{\bar{x}}{n} \tag{6.4}$$

Man kann diese Überlegungen verallgemeinern und sich vorstellen, dass irgendein unbekannter Parameter λ_x durch eine **Schätzfunktion**

$$l_x = g(x_1, x_2, \ldots x_z) \tag{6.5}$$

angenähert wird. Das heißt, wir setzen

$$\lambda_x \approx l_x \qquad \text{mit} \qquad l_x = g(x_1, x_2 \ldots x_z) \tag{6.6}$$

Das wichtigste, wenn auch nicht das einzige Verfahren zur Gewinnung von Schätzfunktionen ist die **Maximum-Likelihood-Methode**, zu der Ronald Aylmer Fisher (1890–1962) wesentlich beigetragen hat. Man kann zeigen, dass man mit dieser Methode im Fall der Binomialverteilung die Schätzung

$$p \approx \frac{\bar{x}}{n}$$

und im Fall der Normalverteilung die Schätzung

$$\mu_x \approx \bar{x} \qquad \text{mit} \qquad \bar{x} = \frac{1}{z} \sum_{i=1}^{z} x_i$$

erhält. Eine mittels der Maximum-Likelihood-Methode gewonnene Schätzung wird als Maximum-Likelihood-Schätzung (E: maximum likelihood estimate, MLE) bezeichnet. Neben der Maximum-Likelihood-Methode ist die Momentenmethode von Karl Pearson verbreitet. Wir gehen jedoch nicht weiter darauf ein.

Der Grundgedanke, der zur Maximum-Likelihood-Methode führt, lässt sich am einfachsten verstehen, wenn man annimmt, die Grundgesamtheit werde durch eine diskrete Wahrscheinlichkeitsverteilung $P(x)$ gemäß 3.28 auf Seite 33 beschrieben. Die Wahrscheinlichkeiten für die einzelnen Merkmalswerte x_i der Stichprobe sollen von einem einzigen Parameter λ_x abhängen, für den ein Schätzwert gesucht wird.

Die in der Stichprobe enthaltenen z Merkmalswerte x_i sind mit der Wahrscheinlichkeit $P(x_1, x_2, \ldots x_z)$ eingetroffen. Da die Stichprobenwerte x_i voneinander unabhängig sind, gilt der Produktsatz in der Form

$$P(x_1, x_2, \ldots x_z) = P(x_1)\ P(x_2) \ldots P(x_z)$$

Dieses Wahrscheinlichkeitsprodukt hängt nur noch von dem unbekannten Parameter λ_x ab. Die Maximum-Likelihood-Methode besteht nun darin, dass man als Näherung für λ_x einen Wert sucht, bei dem das Wahrscheinlichkeitsprodukt einen möglichst großen Wert besitzt. Ist $P(x)$ eine differenzierbare Funktion des Parameters λ_x, erhält man folglich

$$\frac{\partial}{\partial \lambda_x} P(x_1, x_2 \ldots x_z) = 0$$

als Bestimmungsgleichung für λ_x. Diese Vorgehensweise lässt sich auf Wahrscheinlichkeitsverteilungen erweitern, die mehr als einen unbekannten Parameter enthalten, und auf stetige Verteilungen übertragen, indem man die Wahrscheinlichkeiten $P(x_i)$ durch die Wahrscheinlichkeitsdichten $\varphi(x_i)$ ersetzt.

Beispiel 6.1 : Eine einparametrige diskrete Wahrscheinlichkeitsverteilung, die die oben angegebenen Voraussetzungen erfüllt, ist die Poisson-Verteilung nach 4.25 auf Seite 70 in der Form

$$P(x) = \frac{\mu_x^x}{x!} \exp(-\mu_x)$$

Damit folgt aus dem Produktsatz

$$\begin{aligned} P(x_1, x_2, \ldots x_z) &= \frac{\mu_x^{x_1}}{x_1!} \exp(-\mu_x)\ \frac{\mu_x^{x_2}}{x_2!} \exp(-\mu_x) \ldots \frac{\mu_x^{x_z}}{x_z!} \exp(-\mu_x) \\ &= \frac{1}{x_1! x_2! \ldots x_z!}\ \mu_x^{x_1+x_2+\ldots+x_z}\ \exp(-z\mu_x) \\ &= \frac{1}{x_1! x_2! \ldots x_z!}\ \mu_x^{z\bar{x}}\ \exp(-z\mu_x) \end{aligned}$$

Hieraus erhält man durch Differenzieren und Nullsetzen die Bestimmungsgleichung

$$\frac{\partial}{\partial \mu_x} P(x_1, x_2 \ldots x_z) = \frac{1}{x_1! x_2! \ldots x_z!}\ \mu_x^{z\bar{x}}\ \left(\frac{z\bar{x}}{\mu_x} - z\right)\ \exp(-z\mu_x) = 0$$

woraus folgt

$$\mu_x \approx \bar{x} \qquad \text{mit} \qquad \bar{x} = \frac{1}{z} \sum_i x_i$$

das heißt wieder das Stichprobenmittel $\bar{x}$.

Ende Beispiel

Weiterhin wird häufig von einer Schlussweise Gebrauch gemacht, die zunächst ungewohnt ist, jedoch für das Verständnis des Weiteren unerlässlich. Die Stichprobenwerte $x_1, x_2 \ldots x_z$ sind z beobachtete Werte einer einzelnen Zufallsvariablen x. Man kann diese z Werte aber ebenso gut als einzelne Beobachtungswerte von z Zufallsvariablen $x_1, x_2 \ldots x_z$ auffassen, die alle dieselbe Verteilungsfunktion haben und unabhängig voneinander sind, da die Stichprobenwerte sich gemäß Voraussetzung aus z voneinander unabhängigen Ausführungen eines Zufallsexperimentes ergeben. Dann ist jeweils $g(x_1, x_2 \ldots x_z)$ als ein einzelner Wert der neuen Zufallsvariablen l_x anzusehen.

Beispiel 6.2 : Es sei x_1 die Augenzahl beim ersten Wurf, x_2 die Augenzahl beim zweiten Wurf usw. Für den Mittelwert einer Stichprobe

$$\bar{x} = \frac{1}{z}\,(x_1 + x_2 + \ldots + x_z)$$

kann auf Grund dieser Überlegung sehr einfach der Erwartungswert berechnet werden. Mit dem Additionssatz für Mittelwerte 3.163 auf Seite 58 ergibt sich

$$E(\bar{x}) = \frac{1}{z}\,\Big(E(x_1) + E(x_2) + \ldots + E(x_z)\Big)$$

Da alle x_i den gleichen Mittelwert $E(x_i) = \mu_x$ haben, erhält man

$$E(\bar{x}) = \mu_x \tag{6.7}$$

Ende Beispiel

An brauchbare Schätzfunktionen werden üblicherweise vier Forderungen gestellt. Sie sollen

- erwartungstreu (unverfälscht, unverzerrt, E: unbiased),
- wirksam (effizient, E: efficient),
- konsistent (passend, E: consistent) und
- suffizient (erschöpfend, E: sufficient)

sein. Was diese Gütekriterien bedeuten, sehen wir uns jetzt an.

Eine Schätzfunktion $g(x_1, x_2 \ldots x_z)$ für einen Parameter λ_x heißt **erwartungstreu**, wenn

$$E\Big(g(x_1, x_2 \ldots x_z)\Big) = \lambda_x$$

ist. Man liest aus 6.7 ab, dass in diesem Sinn das Stichprobenmittel $\bar{x}$ ein erwartungstreuer Schätzwert für den Mittelwert μ_x der Grundgesamtheit ist.

In ähnlicher Rechnung kann man zeigen, dass auch die Stichprobenvarianz s_x^2 ein erwartungstreuer Schätzwert für die Varianz σ_x^2 der Grundgesamtheit ist. Es gilt

$$s_x^2 = \frac{1}{z-1}\,\sum_{i=1}^{z}(x_i - \bar{x})^2 = \frac{1}{z-1}\,\sum_{i=1}^{z}\Big((x_i - \mu_x) - (\bar{x} - \mu_x)\Big)^2$$

woraus durch Umformen folgt

$$(z-1)s_x^2 = \sum_{i=1}^{z}(x_i-\mu_x)^2 - 2\sum_{i=1}^{z}(x_i-\mu_x)(\bar{x}-\mu_x) + \sum_{i=1}^{z}(\bar{x}-\mu_x)^2$$

Für die zweite und dritte Summe ergeben sich Vereinfachungen

$$\begin{aligned}\sum_{i=1}^{z}(x_i-\mu_x)(\bar{x}-\mu_x) &= (\bar{x}-\mu_x)\sum_{i=1}^{z}(x_i-\mu_x)\\ &= (\bar{x}-\mu_x)\ z\ (\bar{x}-\mu_x)\\ &= z\ (\bar{x}-\mu_x)^2\end{aligned}$$

und

$$\sum_{i=1}^{z}(\bar{x}-\mu_x)^2 = z\ (\bar{x}-\mu_x)^2$$

Damit folgt weiter

$$(z-1)s_x^2 = \sum_{i=1}^{z}(x_i-\mu_x)^2 - z(\bar{x}-\mu_x)^2$$

Bildet man hiervon den Erwartungswert, so ist

$$(z-1)\ E(s_x^2) = \sum_{i-1}^{z} E\Big((x_i-\mu_x)^2\Big) - z\ E\Big((\bar{x}-\mu_x)^2\Big)$$

Für die Varianz von $\bar{x}$ findet man

$$\begin{aligned}E\Big((\bar{x}-\mu_x)^2\Big) &= \frac{1}{z^2}\,E\Big(\big((x_1-\mu_x)+(x_2-\mu_x)+\ldots+(x_z-\mu_x)\big)^2\Big)\\ &= \frac{1}{z^2}\sum_{i=1}^{z}E\Big((x_i-\mu_x)^2\Big) + \frac{1}{z^2}\sum_{i\neq j}E\Big((x_i-\mu_x)(x_j-\mu_x)\Big)\end{aligned}$$

Da Unabhängigkeit der Variablen x_i vorausgesetzt wurde, gilt der Multiplikationssatz für Mittelwerte 3.174 von Seite 59, und es folgt

$$E\Big((x_i-\mu_x)(x_j-\mu_x)\Big) = E(x_i-\mu_x)\ E(x_j-\mu_x) = 0$$

Damit ergibt sich

$$E\Big((\bar{x}-\mu_x)^2\Big) = \frac{1}{z^2}\sum_{i=1}^{z}E\Big((x_i-\mu_x)^2\Big)$$

Nun ist definitionsgemäß

$$E\Big((x_i-\mu_x)^2\Big) = \sigma_x^2$$

für alle Variablen x_i. Damit folgt für die Varianz von $\bar{x}$

$$E\Big((\bar{x} - \mu_x)^2\Big) = \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{z}$$

und weiter für den Erwartungswert von s_x^2

$$(z-1)\ E(s_x^2) = (z-1)\ \sigma_x^2$$

oder schließlich

$$E(s_x^2) = \sigma_x^2 \tag{6.8}$$

Ebenso wie s_x^2 ist auch $\tilde{s}_x^2$ ein erwartungstreuer Schätzwert für die Varianz der Grundgesamtheit. Es gilt genauso allgemein

$$E(\tilde{s}_x^2) = \sigma_x^2 \tag{6.9}$$

wie man leicht nachrechnen kann. Zu $\sigma_{\bar{x}}^2$ siehe 7.2 auf Seite 124. Die Erwartungstreue ist, wie man an den unterschiedlich definierten Schätzwerten s_x^2 und $\tilde{s}_x^2$ erkennt, allein noch kein Maß für die Güte einer Schätzfunktion.

Man bezeichnet eine erwartungstreue Schätzfunktion

$$l_x = g(x_1, x_2 \ldots x_z)$$

für den Parameter λ_x dann als **wirksam**, wenn die Varianz von l_x

$$\sigma_{l_x}^2 = E\Big((l_x - \lambda_x)^2\Big)$$

einen endlichen Wert hat und es keine andere Schätzfunktion gibt, die eine kleinere Varianz $\sigma_{l_x}^2$ besitzt

Eine Schätzfunktion muss nicht unbedingt erwartungstreu und wirksam sein, um für praktische Zwecke brauchbar zu sein. Sie muss aber in jedem Fall **konsistent** sein; das heißt, es muss gelten

$$E\Big((l_x - \mu_x)^2\Big) \to 0 \quad \text{für} \quad z \to \infty$$

Eine Vergrößerung des Stichprobenumfangs soll also dazu führen, dass der Schätzwert näher an dem Wert des zu schätzenden Parameters liegt. Hierzu folgende Beispiele: Die Schätzfunktion für den Mittelwert μ_x ist

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

Sie ist erwartungstreu, weil

$$E(\bar{x}) = \mu_x$$

Für die Varianz von $\bar{x}$ ergab sich

$$E\Big((\bar{x} - \mu_x)^2\Big) = \frac{\sigma_x^2}{z}$$

Daraus folgt, dass $\bar{x}$ konsistent ist. Die Schätzfunktionen für die Varianz σ_x^2 sind

$$s_x^2 = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \qquad \text{und} \qquad \tilde{s}_x^2 = \frac{1}{z} \sum_{i=1}^{z} (x_i - \mu_x)^2$$

Beide sind erwartungstreu, da

$$E(s_x^2) = \sigma_x^2 \qquad \text{und} \qquad E(\tilde{s}_x^2) = \sigma_x^2$$

Für die Varianzen finden wir

$$E\Big((s_x^2 - \sigma_x^2)^2\Big) = \frac{2\sigma_x^4}{z-1} \qquad \text{und} \qquad E\Big((\tilde{s}_x^2 - \sigma_x^2)^2\Big) = \frac{2\sigma_x^4}{z}$$

Beide Schätzwerte sind konsistent, der Schätzwert $\tilde{s}_x^2$ ist auch wirksam.

Eine Schätzung wird als **suffizient** bezeichnet, falls die geschätzten Parameter die Verteilung der Grundgesamtheit eindeutig oder erschöpfend kennzeichnen. Im Falle der Binomial- oder der Poisson-Verteilung reicht dazu ein einziger Parameter, im Falle der Normalverteilung brauchen wir zwei. Das Kriterium geht auf R. A. Fisher zurück. Es beschäftigt uns nicht weiter.

Die Angabe von Schätzwerten für die Verteilungsparameter ist für sich allein noch nicht befriedigend, Wir wollen in jedem Fall Angaben darüber haben, wie genau die angegebenen Näherungswerte sind. Wo immer man in der Mathematik Näherungswerte benutzt, ist man bemüht, die maximal möglichen Abweichungen der Näherungswerte von den genauen Werten abzuschätzen und feste Grenzen anzugeben, die den unbekannten genauen Wert der näherungsweise berechneten Größe mit Sicherheit einschließen.

Sichere Schlüsse von einer Stichprobe auf die Grundgesamtheit gibt es aber nicht. Man kann deshalb auch für die Verteilungsparameter keine sicheren Schranken angeben. Was man jedoch in jedem Fall angeben kann, sind Schranken, die den unbekannten genauen Wert des Verteilungsparameters mit einer bestimmten Wahrscheinlichkeit einschließen.

Bezeichnet man mit λ_{x_o} die obere und mit λ_{x_u} die untere Schranke für den unbekannten genauen Wert des Parameters λ_x, so ist die Wahrscheinlichkeit

$$P(\lambda_{x_u} \leq \lambda_x \leq \lambda_{x_o}) = \gamma \tag{6.10}$$

bei gegebener Verteilung der Grundgesamtheit und bei gegebenem Stichprobenumfang eine Funktion der Intervallgrenzen λ_{x_u} und λ_{x_o} und zugleich vom Ausfall der Stichprobe abhängig. Wird umgekehrt die Wahrscheinlichkeit γ vorgegeben, so sind die Intervallgrenzen λ_{x_u} und λ_{x_o} jeweils aus den Stichprobenwerten x_i zu berechnen. Wichtig ist die Feststellung, dass die so, das heißt mit gegebener Wahrscheinlichkeit γ, festgelegten Intervallgrenzen keine Konstanten sind, sondern selbst Zufallsvariablen.

Man bezeichnet das Intervall zwischen λ_{x_u} und λ_{x_o} als ein **Konfidenzintervall** (E: confidence interval, CI) oder einen **Vertrauensbereich** des Parameters λ_x. Die zugehörige Wahrscheinlichkeit γ wird **Konfidenzzahl** (E: confidence level) genannt.

Wählen wir beispielsweise $\gamma = 0,95$, so dürfen wir erwarten, dass die hierzu bestimmten Konfidenzintervalle den wahren Wert des Parameters λ_x in etwa 19 von 20 Fällen enthalten, in einem von 20 Fällen jedoch nicht. In etwa einem von 20 Fällen werden wir mithin aufgrund des Stichprobenergebnisses eine falsche Aussage über die Grundgesamtheit treffen, wenn wir behaupten, der Parameter λ_x läge innerhalb der Grenzen des Konfidenzintervalles. Wählen wir $\gamma = 0,99$, so haben wir mit einer falschen Aussage nur in einem von 100 Fällen zu rechnen. Das zugehörige Konfidenzintervall wird jedoch größer. In der Praxis wählt man meist $\gamma = 0,95$ oder $\gamma = 0,99$, seltener $\gamma = 0,999$.

Die Festlegung der Konfidenzzahl γ ist keine mathematische Frage, sondern wird durch die Anwendung bestimmt. Man muss sich jedesmal überlegen, welches Risiko mit einer falschen Aussage über die Grundgesamtheit verbunden ist.

6.2 Konfidenzintervalle für den Mittelwert einer Normalverteilung mit bekannter Varianz

Gegeben sei eine Stichprobe aus einer normalverteilten Grundgesamtheit mit bekannter Varianz σ_x^2. Der Mittelwert μ_x der Grundgesamtheit sei unbekannt. Als Schätzung für μ_x verwendet man den Mittelwert $\bar{x}$ der Stichprobe

$$\bar{x} = \frac{1}{z} \sum_{i-1}^{z} x_i \tag{6.11}$$

Um ein Konfidenzintervall für μ_x berechnen zu können, müssen wir die Verteilung der neuen Variablen $\bar{x}$ kennen. Deren Erwartungswert wurde bereits berechnet, siehe 6.7auf Seite 104

$$E(\bar{x}) = \mu_x \tag{6.12}$$

Die Varianz von $\bar{x}$ ergab sich bei der Ableitung von 6.8 auf Seite 106. Dort finden wir

$$E\Big((\bar{x} - \mu_x)^2\Big) = \frac{1}{z}\, E\Big((x - \mu_x)^2\Big) = \frac{\sigma_x^2}{z} \tag{6.13}$$

Nur der Typ der Verteilung von $\bar{x}$ ist noch unbekannt. Um diesen zu gewinnen, wird zunächst die Verteilung der Zufallsvariablen

$$v = x_1 + x_2 \tag{6.14}$$

unter der Voraussetzung untersucht, dass x_1 und x_2 voneinander unabhängige, normalverteilte Zufallsvariablen sind mit den Mittelwerten μ_{x_1}, μ_{x_2} und den Varianzen $\sigma_{x_1}^2$, $\sigma_{x_2}^2$. Für den Mittelwert der neuen Variablen v ergibt sich mit dem Additionssatz für Mittelwerte 3.163 von Seite 58

$$\mu_v = \mu_{x_1} + \mu_{x_2} \tag{6.15}$$

und entsprechend für die Varianz mit dem Additionssatz für Varianzen 3.168 von Seite 59

$$\sigma_v^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 \tag{6.16}$$

Die beiden eindimensionalen Zufallsvariablen x_1 und x_2 können wir uns auch als eine zweidimensionale Zufallsvariable (x_1, x_2) vorstellen. Dann sind die Dichtefunktionen der beiden Variablen x_1 und x_2 zugleich die Randverteilungen der zweidimensionalen Variablen (x_1, x_2), und es gilt

$$\begin{aligned} \varphi_1(x_1) &= \frac{1}{\sigma_{x_1}\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{x_1-\mu_{x_1}}{\sigma_{x_1}})^2\Big) \\ \varphi_2(x_2) &= \frac{1}{\sigma_{x_2}\sqrt{2\pi}} \exp\Big(-\frac{1}{2}(\frac{x_2-\mu_{x_2}}{\sigma_{x_2}})^2\Big) \end{aligned} \tag{6.17}$$

Da x_1 und x_2 voneinander unabhängige Variable sind, ist die zweidimensionale Dichtefunktion das Produkt ihrer Randverteilungen

$$\varphi(x_1, x_2) = \varphi_1(x_1)\ \varphi_2(x_2) \tag{6.18}$$

Man kann sich die zweidimensionale Dichtefunktion als gekrümmte Fläche über der x_1x_2-Ebene aufgespannt denken.

Die Summenfunktion der neuen Variablen v ist definiert durch

$$\Phi(v) = P\Big((x_1 + x_2) \le v\Big) \tag{6.19}$$

In Abbildung 6.1 auf Seite 110 wird die x_1x_2-Ebene durch die Gerade $x_1 + x_2 = v$ in zwei Bereiche unterteilt. Um die Summenfunktion $\Phi(v)$ zu bestimmen, ist die Dichtefunktion $\varphi(x_1, x_2)$ über den gesamten Bereich links von der Geraden $x_1 + x_2 = v$ zu integrieren. Eine Lösung ist

$$\Phi(v) = \int_{x_1=-\infty}^{x_1=+\infty} \Big(\varphi_1(x_1) \int_{x_2=-\infty}^{x_2=v-x_1} \varphi_2(x_2)\ \mathrm{d}x_2\Big)\ \mathrm{d}x_1 \tag{6.20}$$

Eine gleichwertige Lösung ist

$$\Phi(v) = \int_{x_2=-\infty}^{x_2=+\infty} \Big(\varphi_2(x_2) \int_{x_1=-\infty}^{x_1=v-x_2} \varphi_1 x_1\ \mathrm{d}x_1\Big)\ \mathrm{d}x_2 \tag{6.21}$$

Differenziert man 6.20 nach v, so erhält man die Dichtefunktion der neuen Variablen v zu

$$\varphi(v) = \frac{\mathrm{d}\Phi(v)}{\mathrm{d}v} = \int_{-\infty}^{+\infty} \varphi_1(x_1)\ \varphi_2(v - x_1)\ \mathrm{d}x_1 \tag{6.22}$$

In gleicher Weise folgt aus 6.21

$$\varphi(v) = \frac{\mathrm{d}\Phi(v)}{\mathrm{d}v} = \int_{-\infty}^{+\infty} \varphi_2(x_2)\ \varphi_1(v - x_2)\ \mathrm{d}x_2 \tag{6.23}$$

Integrale dieses Typs werden **Faltung** (E: convolution) genannt, ein Begriff aus der Funktionalanalysis, der in der Physik zahlreiche Anwendungen besitzt.

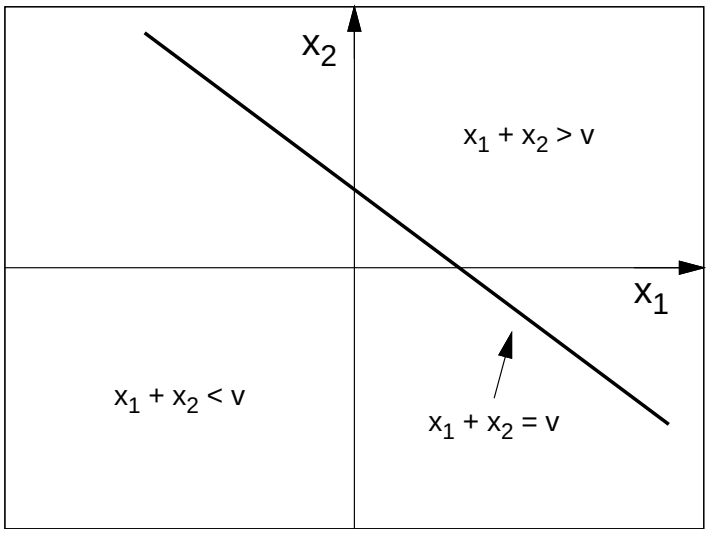


Abb. 6.1: Darstellung der x_1x_2-Ebene mit der Geraden $x_1 + x_2 = v$

Setzen wir 6.17 in 6.22 ein, so folgt

$$\varphi(v) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}} \int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}})^2 - \frac{1}{2}(\frac{v - x_1 - \mu_{x_2}}{\sigma_{x_2}})^2\Big)\, \mathrm{d}x_1 \qquad (6.24)$$

Wie man durch Nachrechnen bestätigen kann, gilt für das Argument der Exponentialfunktion

$$-\frac{1}{2}(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}})^2 - \frac{1}{2}(\frac{v - x_1 - \mu_{x_2}}{\sigma_{x_2}})^2$$

$$= -\frac{1}{2}(\frac{v - \mu_v}{\sigma_v})^2 - \frac{1}{2}\frac{\sigma_v^2}{\sigma_{x_1}^2\sigma_{x_2}^2}\Big(x_1 - \frac{\sigma_{x_2}^2\mu_{x_1} + \sigma_{x_1}^2(v - \mu_{x_2})}{\sigma_v^2}\Big)^2$$

Damit und mit der Substitution

$$y = \frac{\sigma_v}{\sigma_{x_1}\sigma_{x_2}}\Big(x_1 - \frac{\sigma_{x_2}^2\mu_{x_1} + \sigma_{x_1}^2(v - \mu_{x_2})}{\sigma_v^2}\Big) \qquad \text{daraus} \qquad \mathrm{d}y = \frac{\sigma_v}{\sigma_{x_1}\sigma_{x_2}}\, \mathrm{d}x_1$$

folgt weiter

$$\varphi(v) = \frac{1}{2\pi\sigma_v}\, \exp\Big(-\frac{1}{2}\,(\frac{v - \mu_x}{\sigma_v})^2\Big)\, \int_{-\infty}^{+\infty} \exp\Big(-\frac{1}{2}y^2\Big)\, \mathrm{d}y$$

Das verbleibende Integral hat den Wert $\sqrt{2\pi}$. Damit erhalten wir schließlich

$$\varphi(v) = \frac{1}{\sigma_v\,\sqrt{2\pi}}\, \exp\Big(-\frac{1}{2}\,(\frac{v - \mu_v}{\sigma_v})^2\Big) \qquad (6.25)$$

Die neue Variable v ist normalverteilt mit dem Mittelwert μ_v gemäß 6.15 und der Varianz σ_v^2 gemäß 6.16.

Stellen wir uns vor, die Variable x_1 sei die Summe zweier normalverteilter Variabler, so gelten alle Überlegungen unverändert für die Summe von drei Zufallsvariablen. Durch Wiederholen dieses Schrittes gelangen wir zu folgendem Satz:

Sind $x_1, x_2 \ldots x_n$ voneinander unabhängige, normalverteilte Zufallsvariable mit den Mittelwerten $\mu_{x_1}, \mu_{x_2} \ldots \mu_{x_n}$ und den Varianzen $\sigma^2_{x_1}, \sigma^2_{x_2} \ldots \sigma {x_n}^2$, dann ist auch die Zufallsvariable

$$v = x_1 + x_2 + \ldots + x_n$$

normalverteilt und hat den Mittelwert

$$\mu_v = \mu_{x_1} + \mu_{x_2} + \ldots + \mu_{x_n}$$

sowie die Varianz

$$\sigma^2_v = \sigma^2_{x_1} + \sigma^2_{x_2} + \ldots + \sigma^2_{x_n}$$

Zu untersuchen ist schließlich noch die Verteilung einer weiteren neuen Variablen

$$v = c_1 x + c_2 \tag{6.26}$$

unter der Voraussetzung, dass x normalverteilt ist mit dem Mittelwert μ_x und der Varianz σ^2_x. Für die Dichtefunktion muss gelten

$$\varphi^\star(v)\, \mathrm{d}v = \varphi(x)\, \mathrm{d}x \tag{6.27}$$

Aus 6.26 folgt $\mathrm{d}v = c_1 \mathrm{d}x$. Setzen wir dies und die Dichtefunktion der Normalverteilung von x in 6.27 ein, so erhalten wir

$$\varphi^\star(v) = \frac{1}{c_1}\, \varphi(x) = \frac{1}{c_1 \sigma_x \sqrt{2\pi}}\, \exp\Big(-\frac{1}{2} (\frac{x - \mu_x}{\sigma_x})^2 \Big)$$

Mit 6.26 folgt daraus

$$\varphi^\star(v) = \frac{1}{c_1 \sigma_x \sqrt{2\pi}}\, \exp\Big(-\frac{1}{2} (\frac{v - c_1\mu_x - c_2}{c_1 \sigma_x})^2 \Big)$$

Setzt man

$$\mu_v = c_1 \mu_x + c_2 \quad \text{und} \quad \sigma^2_v = c_1^2 \sigma^2_x \tag{6.28}$$

so folgt

$$\varphi^\star(v) = \frac{1}{\sigma_v \sqrt{2\pi}}\, \exp\Big(-\frac{1}{2} (\frac{v - \mu_x}{\sigma_v})^2 \Big) \tag{6.29}$$

Das Ergebnis kann man zusammenfassen in dem Satz:

Ist x normalverteilt mit dem Mittelwert μ_x und der Varianz σ_x^2, so ist

$$v = c_1 x + c_2$$

normalverteilt mit dem Mittelwert

$$\mu_v = c_1 \mu_x + c_2$$

und der Varianz

$$\sigma_v^2 = c_1^2 \sigma_x^2$$

Aus den beiden oben abgeleiteten Sätzen folgt direkt, dass das Stichprobenmittel normalverteilt ist mit

$$\mu_{\bar{x}} = \mu_x \quad \text{und} \quad \sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{z} \tag{6.30}$$

übereinstimmend mit 6.12 und 6.13 auf Seite 108. Durch die Substitution

$$t = \sqrt{z}\,\frac{\bar{x} - \mu_x}{\sigma_x} \tag{6.31}$$

wird diese Normalverteilung zurückgeführt auf die normalisierte Normalverteilung mit dem Mittelwert 0 und der Varianz 1.

Da die Normalverteilung symmetrisch zum Mittelwert ist, liegt es nahe, auch das Konfidenzintervall für den Mittelwert symmetrisch zu wählen. Wir setzen

$$P\Big(-c \le \sqrt{z}\,(\frac{\bar{x} - \mu_x}{\sigma_x}) \le c\Big) = H(c) - H(-c) = \gamma$$

Das gesuchte Konfidenzintervall für den Mittelwert μ_x der Grundgesamtheit ist dann

$$\bar{x} - c\frac{\sigma_x}{\sqrt{z}} \le \mu_x \le \bar{x} + c\frac{\sigma_x}{\sqrt{z}} \tag{6.32}$$

Hierin ist c eine eindeutige und umkehrbare Funktion von γ, siehe Tabelle 6.1.

Tab. 6.1: Konfidenzintervall einer Stichprobe, Werte der Konstanten c für gegebenes Intervall γ

γ	0,500	0.900	0,950	0,990	0,999
c	0,675	1,645	1,960	2,576	3,291

6.3 Konfidenzintervalle für die Varianz einer Normalverteilung mit bekanntem Mittelwert

Gegeben sei eine Stichprobe aus einer normalverteilten Grundgesamtheit mit bekanntem Mittelwert μ_x. Unbekannt sei die Varianz σ_x^2 der Grundgesamtheit. Als Schätzwert für σ_x^2 verwendet man die Stichprobenvarianz, vergleiche 6.3 auf Seite 102

$$\tilde{s}_x^2 = \frac{1}{z} \sum_{i=1}^{z} (x_i - \mu_x)^2 \tag{6.33}$$

Dividiert man beide Seiten von 6.33 durch σ_x^2 und multipliziert mit z, so folgt

$$z \, \frac{\tilde{s}_x^2}{\sigma_x^2} = \sum_{i=1}^{z} \Big(\frac{x_i - \mu_x}{\sigma_x} \Big)^2 \tag{6.34}$$

Die rechte Seite von 6.34 kann man auffassen als die Summe der Quadrate von z unabhängigen Zufallsvariablen, deren jede eine Normalverteilung mit dem Mittelwert 0 und der Varianz 1 besitzt. Daher ist

$$v = z \, \frac{\tilde{s}_x^2}{\sigma_x^2} \tag{6.35}$$

χ^2-verteilt mit der Anzahl der Freiheitsgrade $n = z$.

Wir können nun mit Hilfe der χ^2-Verteilung die Wahrscheinlichkeit berechnen, dass die Größe $z\tilde{s}_x^2/\sigma_x^2$ zwischen einer unteren Grenze χ_u^2 und einer oberen Grenze χ_o^2 liegt. Es ist dann

$$P\Big(\chi_u^2 \le z \, \frac{\tilde{s}_x^2}{\sigma_x^2} \le \chi_o^2 \Big) = \Phi(\chi_o^2) - \Phi(\chi_u^2) = \gamma$$

Das gesuchte Konfidenzintervall für die Varianz σ_x^2 ergibt sich daraus zu

$$z \, \frac{\tilde{s}_x^2}{\chi_o^2} \le \sigma_x^2 \le z \, \frac{\tilde{s}_x^2}{\chi_u^2} \tag{6.36}$$

Da die χ^2-Verteilung nicht symmetrisch ist, lassen sich auch keine symmetrischen Grenzen vorgeben. Meist wählt man die Grenzen so, dass gilt

$$\Phi(\chi_u^2) = 1 - \Phi(\chi_o^2) \tag{6.37}$$

Die Zahlenwerte entnimmt man aus Tabellensammlungen, beispielsweise aus Graf/Henning/Stange/Wilrich, *Formeln und Tabellen der angewandten mathematischen Statistik*, bei Springer 1997.

6.4 Konfidenzintervalle für die Varianz einer Normalverteilung mit unbekanntem Mittelwert

Sind Mittelwert und Varianz der Grundgesamtheit unbekannt, so verwendet man als Schätzwert für σ_x^2 die Stichprobenvarianz, siehe 6.2 auf Seite 101

$$s_x^2 = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \tag{6.38}$$

Die z Quadrate auf der rechten Seite von 6.38 lassen sich wegen

$$\sum_{i=1}^{z} (x_i - \bar{x}) = 0 \tag{6.39}$$

auf $z-1$ Quadrate zurückführen. Wenn man dann beide Seiten durch σ_x^2 dividiert und mit $(z-1)$ multipliziert, erhält man auf der rechten Seite die Summe der Quadrate von $z-1$ voneinander unabhängigen Zufallsvariablen, deren jede eine Normalverteilung mit dem Mittelwert 0 und der Varianz 1 besitzt. Daher ist die Größe

$$v = (z-1)\,\frac{s_x^2}{\sigma_x^2} \tag{6.40}$$

χ^2-verteilt mit der Anzahl der Freiheitsgrade $n = z-1$. Das gesuchte Konfidenzinterval für die Varianz σ_x^2 ist dann

$$(z-1)\,\frac{s_x^2}{\chi_o^2} \leq \sigma_x^2 \leq (z-1)\,\frac{s_x^2}{\chi_u^2} \tag{6.41}$$

6.5 Konfidenzintervalle für den Mittelwert einer Normalverteilung mit unbekannter Varianz

Sind Mittelwert und Varianz der Grundgesamtheit unbekannt, so verwendet man als Schätzwert für μ_x das Stichprobenmittel $\bar{x}$ nach 6.1 auf Seite 101

$$\bar{x} = \frac{1}{z} \sum_{i=1}^{z} x_i \tag{6.42}$$

von dem man weiß, dass es eine Normalverteilung mit dem Mittelwert μ_x und der Varianz σ_x^2/z besitzt. Wollte man diese Normalverteilung durch die Substitution 6.31 von Seite 112

$$w = \sqrt{z}\,\frac{\bar{x} - \mu_x}{\sigma_x} \tag{6.43}$$

auf die normierte Normalverteilung $\mathcal{N}(0,1)$ zurückführen, um daraus das Konfidenzintervall zu berechnen, so müsste σ_x^2 bekannt sein.

Wir ersetzen deshalb die unbekannte Varianz σ_x^2 in 6.43 durch die Stichprobenvarianz

$$s_x^2 = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \tag{6.44}$$

Im Unterschied zu 6.31 oder 6.43 ist

$$t = \sqrt{z}\, \frac{\bar{x} - \mu_x}{s_x} \tag{6.45}$$

jedoch nicht mehr normalverteilt. Aus 6.45 erhalten wir mittels Dividieren durch σ_x in Zähler und Nenner

$$t = \frac{\sqrt{z}\, \dfrac{\bar{x} - \mu_x}{\sigma_x}}{\sqrt{\dfrac{s_x^2}{\sigma_x^2}}} \tag{6.46}$$

Der Zähler

$$u = \sqrt{z}\, \frac{\bar{x} - \mu_x}{\sigma_x} \tag{6.47}$$

ist normalverteilt mit dem Mittelwert 0 und der Varianz 1. Die Variable im Nenner, leicht umgeformt

$$v = (z-1)\, \frac{s_x^2}{\sigma_x^2} \tag{6.48}$$

ist χ^2-verteilt mit dem Freiheitsgrad $n = z - 1$. Die Variable t ist, setzen wir 6.47 und 6.48 in 6.46 ein, von der Form

$$t = \frac{u}{\sqrt{\dfrac{v}{z-1}}} \tag{6.49}$$

Die Größe t ist daher student-verteilt mit dem Freiheitsgrad $n = z - 1$, vergleiche 5.68 auf Seite 95.

Da die Student-Verteilung symmetrisch ist, wird auch das Konfidenzintervall für den Mittelwert symmetrisch gewählt. Mit Hilfe der Summenfunktion der Student-Verteilung berechnen wir die Wahrscheinlichkeit

$$P\Big(-c \le \sqrt{z}\, \frac{\bar{x} - \mu_x}{s_x} \le +c\Big) = \Phi(+c) - \Phi(-c) = \gamma$$

Das gesuchte Konfidenzintervall für den Mittelwert μ_x ist dann

$$\bar{x} - c\, \frac{s_x}{\sqrt{z}} \le \mu_x \le \bar{x} + c\, \frac{s_x}{\sqrt{z}} \tag{6.50}$$

Hier ist c nicht nur von der gewählten Wahrscheinlichkeit γ abhängig, sondern auch vom Freiheitsgrad $n = z - 1$.

6.6 Konfidenzintervalle für die Parameter beliebiger Verteilungen

Ist die Art der Verteilung nicht bekannt, das heißt kann man nicht voraussetzen, dass die Grundgesamtheit normalverteilt ist, so kann man dennoch mit 6.41 und 6.50 Näherungswerte für die Konfidenzintervalle für Mittelwert μ_x und Varianz σ_x^2 angeben. Die Grundlage hierfür liefert der **Zentrale Grenzwertsatz**:

> Gegeben seien z voneinander unabhängige Zufallsvariablen $x_1, x_2, \ldots x_n$, die alle dieselbe Verteilungsfunktion, denselben Mittelwert μ_x und dieselbe Varianz σ_x^2 besitzen. Dann ist die Zufallsvariable
>
> $$t = \sqrt{z}\,\frac{\bar{x} - \mu_x}{\sigma_x}$$
>
> asymptotisch normalverteilt mit dem Mittelwert 0 und der Varianz 1.

Es gibt andere Formulierungen des Zentralen Grenzwertsatzes (E: central limit theorem). Analog gilt ferner:

> Die Zufallsvariable
>
> $$v = (z-1)\,\frac{s_x^2}{\sigma_x^2}$$
>
> ist asymptotisch χ^2-verteilt.

Die Güte der Näherung wächst mit dem Stichprobenumfang. Hierzu findet man beispielsweise bei E. Kreyszig folgende Angabe: Lässt die Stichprobe vermuten, dass die zugehörige Verteilung nicht allzu unsymmetrisch ist, so erhält man für den Mittelwert im Allgemeinen befriedigende Ergebnisse, sobald z wenigstens etwa gleich 30 ist, und für die Varianz, wenn z wenigstens etwa gleich 100 ist.

6.7 Beispiele für Konfidenzintervalle

Im Folgenden werden zwei Beispiele behandelt, denen gemeinsam ist, dass Konfidenzintervalle für Verteilungsparameter aus Messwerten (Ergebnissen von Stichproben) bestimmt werden. Zwischen den beiden Beispielen besteht jedoch ein bemerkenswerter Unterschied.

Im ersten Fall sind Wiederholungsmessungen ausgeführt worden, um eine Größe möglichst genau zu bestimmen, die aus einsichtigen physikalischen Gründen nur einen bestimmten Wert annehmen kann wie der Dampfdruck einer organischen Flüssigkeit bei einer vorgegebenen Temperatur. Die Abweichungen der Messwerte vom wahren Wert der Größe sind hier allein auf Unvollkommenheiten

der Messung zurückzuführen. Die Messwerte bilden – wenn systematische Fehler ausgeschlossen werden können – bei unendlichfacher Wiederholung (Grundgesamtheit) eine Verteilung, deren Mittelwert mit dem wahren Wert der Größe übereinstimmt.

Im zweiten Fall wurde eine Größe gemessen, von der man annehmen darf, dass sie annähernd normalverteilt ist. Ziel der Messungen war es, diese Verteilung, das heißt ihre Parameter, möglichst genau zu bestimmen. Die Unterschiede zwischen den Messwerten sind in diesem Fall nicht auf Messungenauigkeiten zurückzuführen, sondern vor allem darauf, dass der wahre Wert der gemessenen Größe selbst zufällig schwankt.

Beispiel 6.3 : Für eine Größe x liegen $z = 6$ Messwerte vor, siehe die erste Spalte der Tabelle 6.2. Wir berechnen zunächst Mittelwert und Varianz dieser Stichprobe. Der berechnete Mittelwert $\bar{x}$ ist ein Schätzwert für den wahren Wert μ_x der Größe x. Unter der Voraussetzung, dass die Messwerte normalverteilt sind, berechnen wir für μ_x ein Konfidenzintervall. Da es hier nicht auf die physikalische Bedeutung der Größe ankommt – statt Dampfdruck könnte es auch eine andere Größe sein – sehen wir von Maßeinheiten ab; es kommt nur auf die Zahlen an.

Tab. 6.2: Messergebnisse und Konfidenzintervall einer Stichprobe

x_i	$10^3 \cdot \xi_i$	$10^3 \cdot \xi_i z(\xi_i)$	$10^6 \cdot \xi_i^2 z(xi_i)$
0,842	2	2	4
0,846	6	6	36
0,835	-5	-5	25
0,839	-1	-1	1
0,843	3	3	9
0,838	-2	-2	4

Zur Vereinfachung der Zahlenrechnung setzen wir

$$x_i = a + \xi_i$$

wobei a so gewählt wird, dass es ungefähr in der Mitte des Streubereiches der Messwerte liegt. Für das Stichprobenmittel folgt hiermit

$$\bar{x} = \frac{1}{z} \sum_{i=1}^{z} x_i \; z(x_i) = a + \bar{\xi}$$

Darin ist $\bar{\xi}$ definiert durch

$$\bar{\xi} = \frac{1}{z} \sum_{i=1}^{z} \xi_i z(\xi_i)$$

Für die Stichprobenvarianz erhalten wir

$$s_x^2 \;\; = \;\; \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \; z(x_i)$$

$$= \frac{1}{z-1} \sum_{i=1}^{z} (\xi_i - \bar{\xi})^2 \, z(\xi_i)$$

$$= \frac{1}{z-1} \Big(\sum_{i=1}^{z} \xi_i^2 z(\xi_i) - z\bar{\xi}^2 \Big)$$

Bei den Zahlenwerten der ersten Spalte der Tabelle 6.2 wird $a = 0,840$ gewählt. Daraus ergeben sich die Werte ξ_i, $\xi_i z(\xi_i)$ und $\xi_i^2 z(\xi_i)$ der zweiten bis vierten Spalte. Durch Aufsummieren erhalten wir

$$\sum_{i=1}^{z} \xi_i \, z(\xi_i) = 3 \cdot 10^{-3} \quad \text{und} \quad \sum_{i=1}^{z} \xi_i^2 \, z(\xi_i) = 79 \cdot 10^{-6}$$

Damit folgt weiter

$$\bar{\xi} = \frac{1}{z} \sum_{i=1}^{z} \xi_i \, z(\xi_i) = 5 \cdot 10^{-4} \qquad \text{und daraus} \qquad \bar{\xi}^2 = 25 \cdot 10^{-8}$$

Stichprobenmittel und Stichprobenvarianz errechnen sich zu

$$\bar{x} = a + \bar{\xi} = 0,840 + 0,0005 = 0,8405$$

$$s_x^2 = \frac{1}{z-1} \Big(\sum_{i=1}^{z} \xi_i^2 z(\xi_i) - z\bar{\xi}^2 \Big) = \frac{1}{5} \cdot (79 - 1,5) \cdot 10^{-6} = 15,5 \cdot 10^{-6}$$

Das Stichprobenmittel $\bar{x}$ ist ein Schätzwert für den wahren Wert μ_x der Größe x. Die Messwerte sind nach Voraussetzung normalverteilt. Mittelwert μ_x und Varianz σ_x^2 dieser Normalverteilung sind unbekannt.

Wie in Abschnitt 6.5 auf Seite 114 gezeigt wurde, kann man für μ_x unter Verwendung der Student-Verteilung ein Konfidenzintervall berechnen. Es gilt 6.50 von Seite 115

$$\bar{x} - c \, \frac{s_x}{\sqrt{z}} \le \mu_x \le \bar{x} + c \, \frac{s_x}{\sqrt{z}}$$

Hierin ist c von der Konfidenzzahl γ und der Anzahl der Freiheitsgrade $n = z - 1$ abhängig. Zu $\gamma = 0,95$ und $n = 5$ entnehmen wir aus einer Tabelle der Student-Verteilung den Wert $c = 2,57$. Damit folgt

$$\mu_x = 0,8405 \pm 0,0041 \quad \text{oder} \quad 0,8364 \le \mu_x \le 0,8446$$

Ende Beispiel

Beispiel 6.4 : An $z = 300$ Glühlampen wurde die Lebensdauer unter kontrollierten Bedingungen gemessen. Die Ergebnisse schwankten zwischen 950 und 2150 Stunden. Der gesamte Bereich der Lebensdauer wurde in 12 Klassen gleicher Klassenbreite $\Delta x = 100$ h eingeteilt. Tabelle 6.3 gibt in ihren ersten beiden Spalten die Anzahl $z(x_i)$ der Fälle an, in denen die Lebensdauer x in das Intervall

Tab. 6.3: Gemessene Lebensdauer von Glühlampen

x_i	$z(x_i)$	$10^{-2} \cdot \xi_i$	$10^{-2} \cdot \xi_i \; z(\xi_i)$	$10^{-4} \cdot \xi_i^2 \; z(\xi_i)$
1000	4	-5	-20	100
1100	9	-4	-36	144
1200	19	-3	-57	171
1300	36	-2	-72	144
1400	51	-1	-51	51
1500	58	0	0	0
1600	53	1	53	53
1700	37	2	74	148
1800	20	3	60	180
1900	9	4	36	144
2000	3	5	15	75
2100	1	6	6	36

$x_i \pm 50$ h fiel. Wir wollen den Mittelwert $\bar{x}$ und die Varianz s_x^2 der Stichprobe berechnen und Konfidenzintervalle für Mittelwert μ_x und Varianz σ_x^2 der Grundgesamtheit angeben.

Zur Vereinfachung der Zahlenrechnungen setzen wir wie im ersten Beispiel

$$x_i = a + \xi_i$$

und wählen $a = 1500$ h. Damit ergeben sich für ξ_i, $\xi_i \; z(\xi_i)$ und $\xi_i^2 \; (\xi_i)$ die Zahlenwerte der Spalten 3 bis 5 der Tabelle 6.3. Durch Aufsummieren findet man

$$\sum_i \xi_i \; z(\xi_i) = 800 \text{ h} \qquad \text{und} \qquad \sum_i \xi_i^2 \; z(\xi_i) = 1246 \cdot 10^4 \text{ h}^2$$

Damit folgt

$$\bar{\xi} = \frac{1}{z} \sum_i \xi_i \; z(\xi_i) = 2,67 \text{ h} \qquad \bar{\xi_i}^2 = 7,11 \text{ h}^2$$

Stichprobenmittel und Stichprobenvarianz errechnen sich daraus zu

$$\bar{x} = a + \bar{\xi} = (1500 + 2,67) \text{ h} = 1502,7 \text{ h}$$

$$s_x^2 = \frac{1}{z-1} \left(\sum_i \xi_i^2 \; z(\xi_i) - z\bar{\xi}^2 \right) = \frac{1}{299} \left(1246 \cdot 10^4 - 300 \cdot 7,11\right) \text{h}^2 = 41665 \text{ h}^2$$

Das Stichprobenmittel $\bar{x}$ ist ein Schätzwert für den Mittelwert μ_x der Grundgesamtheit. Nimmt man an, dass die Grundgesamtheit normalverteilt ist, so kann man für μ_x unter Verwendung der Student-Verteilung beziehungsweise der normalisierten Normalverteilung – beide stimmen für $z = 300$ praktisch überein – ein Konfidenzintervall für μ_x berechnen. Es gilt 6.50 von Seite 115

$$\bar{x} - c \, \frac{s_x}{\sqrt{z}} \leq \mu_x \leq \bar{x} + c \, \frac{s_x}{\sqrt{z}}$$

Zu $\gamma = 0,95$ liest man aus einer Tabelle der Normalverteilung den Wert $c = 1,960$ ab.Damit folgt

$$\mu_x = (1502,67 \pm 23,10)\ \text{h} \quad \text{oder} \quad 1479,6\ \text{h} \leq \mu_x \leq 1525,8\ \text{h}$$

Die Stichprobenvarianz s_x^2 ist ein Schätzwert für die Varianz σ_x^2 der Grundgesamtheit. Nimmt man wiederum an, dass die Grundgesamtheit normalverteilt ist, so kann man für σ_x^2 mit Hilfe der χ^2-Verteilung ein Konfidenzintervall berechnen. Es gilt 6.41

$$(z-1)\,\frac{s_x^2}{\chi_o^2} \leq \sigma_x^2 \leq (z-1)\frac{s_x^2}{\chi_u^2}$$

Zu $\gamma = 0,95$ und $n = 299$ liest man aus Tabellen der χ^2-Verteilung ab

$$\chi_u^2 = 252,53 \qquad \chi_o^2 = 358,31$$

womit sich ergibt

$$35767\ \text{h}^2 \leq \sigma_x^2 \leq 49332\ \text{h}^2 \qquad \text{oder} \qquad 189,1\ \text{h} \leq \sigma_x \leq 222,1\ \text{h}$$

Bei der Berechnung der Konfidenzintervalle für μ_x und σ_x^2 wurde vorausgesetzt, dass die Grundgesamtheit normalverteilt ist. Diese Annahme ist durch den Aufgabentext nicht gedeckt. Der in Abschnitt 6.6 *Konfidenzintervalle für die Parameter beliebiger Verteilungen* auf Seite 116 zitierte Zentrale Grenzwertsatz erlaubt jedoch eine Berechnung von Konfidenzintervallen auch für die Parameter beliebiger Verteilungen, sofern der Stichprobenumfang ausreichend groß ist.

Welche für die Praxis bedeutenden Aussagen lassen sich aus den Ergebnissen ableiten? Für den Käufer einer Glühlampe ist nicht so wichtig, wo Mittelwert μ_x und Varianz σ_x^2 der Grundgesamtheit liegen, sondern vielmehr die Frage, welche Brenndauer eine einzelne Glühlampe mit einer Wahrscheinlichkeit von 0,95 mindestens erreicht. Diese Brenndauer erhält man wie folgt (in Stunden):

$$t_{95} = \frac{x_{95} - \mu_x}{\sigma_x}$$

$$H(t_{95}) = 0,05 \quad \text{woraus folgt} \quad t_{95} = -1,645$$

$$\begin{aligned} x_{95} &= \mu_x + t_{95}\sigma_x \\ &= 1502,67 - 1,645 \cdot 204 = 1167,1 \end{aligned}$$

Bei den allmählich verschwindenden Allgebrauchs-Glühlampen betrug die Nennlebensdauer 1000 h, bei modernen Lampen (Halogen, Leuchtstoff, LED) beträgt sie ein Mehrfaches davon.

Ende Beispiel

Die beiden vorstehenden Beispiele betreffen Sonderfälle. Zu einer Verallgemeinerung der Aufgabenstellung gelangt man durch eine Überlagerung dieser beiden Sonderfälle. Wird eine Größe x gemessen, von der man weiß, dass sie keinen bestimmten Wert annimmt, sondern eine Verteilung besitzt, und ist die Messung selbst ungenau, so ist der Messwert y die Summe vom jeweiligen wahren Wert x und dem Messfehler M

$$y = x + M$$

Für den Erwartungswert von y gilt:

$$E(y) = E(x)$$

da $E(M) = 0$ vorausgesetzt werden darf, wenn kein systematischer Messfehler vorliegt, sondern nur ein zufälliger. Wird Unabhängigkeit zwischen x und M vorausgesetzt, so gilt für die Varianzen:

$$\sigma_y^2 = \sigma_x^2 + \sigma_M^2$$

Will man die Varianz von x bestimmen, so muss man das Messverfahren derart wählen, dass:

$$\sigma_M^2 \ll \sigma_x^2$$

ist. Das Problem spielt eine erhebliche Rolle bei der Analyse von Mischungen, wo die Varianz σ_x^2 als Maß für die Mischgüte verwendet wird. Wir gehen in Abschnitt 9.6 *Mischgütemaß für Feststoffmischungen* auf Seite 166 darauf ein.

Kapitel 7

Statistische Prüfverfahren als Grundlage für Entscheidungen

Wir stehen vor dem Dilemma, dass einerseits keine sicheren Schlüsse von einer Stichprobe auf die jeweilige Grundgesamtheit möglich sind, andererseits aber auf Grund des Stichprobenergebnisses Entscheidungen in Bezug auf die Grundgesamtheit getroffen werden müssen. Statistische Tests liefern quantitative Entscheidungshilfen.

7.1 Grundbegriffe

In der Praxis werden statistische Untersuchungen nicht um ihrer selbst willen durchgeführt, sondern weil man sich eine Basis für unumgängliche Entscheidungen schaffen will. Diese Entscheidungen stehen im Zusammenhang mit bestimmten Eigenschaften einer Grundgesamtheit, über die man sich nur durch die Entnahme und Untersuchung von Stichproben Aufschluss verschaffen kann. Da aber sichere Schlüsse von einer Stichprobe auf die Grundgesamtheit nicht möglich sind, ist jede Entscheidung, die auf solchen Schlüssen beruht, mit einer bestimmten Wahrscheinlichkeit falsch, das heißt mit einem Risiko behaftet. Die Quantifizierung der Risiken von Entscheidungen ist der wesentliche Inhalt des vorliegenden Kapitels. Die unternehmerische Bewertung eines Risikos ist ein weiterer Schritt, den wir jedoch nicht vollziehen. Wie so oft empfiehlt sich ein Blick in die Wikipedia (*Risiko* beziehungsweise *Risk*).

Statistische Prüfverfahren oder Tests werden in den üblichen Darstellungen auf einer **Hypothese** (Annahme in Bezug auf die Grundgesamtheit, E: hypothesis) aufgebaut. Die zu treffende Entscheidung wird damit auf eine Entscheidung für oder gegen die Hypothese zurückgeführt. Zu einer Hypothese gehört ihre Alternativhypothese oder **Alternative** (E: alternative hypothesis). Hypothese und Alternative müssen disjunkt sein; beide können niemals gemeinsam zutreffen. Die gelegentlich anzutreffende Bezeichnung *Nullhypothese* ist nichts weiter als die Hypothese; wir vermeiden sie.

7.2 Durchführung eines Prüfverfahrens

Die Vorgehensweise ist immer die, dass man eine Stichprobe zieht und untersucht, ob deren Ergebnis unter der jeweiligen Hypothese eine kleine oder eine große Wahrscheinlichkeit besitzt. Ist die Wahrscheinlichkeit kleiner als ein vorgegebener Wert α, so verwirft man die Hypothese, man lehnt sie ab. Dabei nimmt man in Kauf, dass die Hypothese mit der Wahrscheinlichkeit α zu Unrecht verworfen wird. Das Verwerfen einer in Wirklichkeit richtigen Hypothese bezeichnet man als **Fehler erster Art**, englisch als *false positive*.

Wenn das Ergebnis der Stichprobe unter der jeweiligen Hypothese eine Wahrscheinlichkeit größer als α hat, nimmt man die Hypothese an. Man sagt vorsichtigerweise, dass das Ergebnis der Stichprobe die Hypothese nicht widerlege. Man sagt nicht, dass das Ergebnis der Stichprobe die Hypothese bestätige. Mit einer Wahrscheinlichkeit $(1-\beta)$ wird nämlich die Hypothese auch dann angenommen, wenn sie falsch ist, das heißt wenn die Annahme in Bezug auf die Grundgesamtheit nicht zutrifft. Das Annehmen einer in Wirklichkeit falschen Hypothese bezeichnet man als **Fehler zweiter Art**, englisch als *false negative*. Die Bezeichnungen samt den zugehörigen Überlegungen gehen auf den polnisch-amerikanischen Mathematiker Jerzy Neyman (1894–1981) und den englischen Statistiker Egon Sharpe Pearson (1895–1980) zurück.

Diese allgemeine Konzept werde im Folgenden durch ein Beispiel erläutert. Es sei die Aufgabe gestellt zu prüfen, ob eine gegebene Grundgesamtheit, bestehend aus Werten der Zufallsvariablen x, den Mittelwert $\mu_x = \mu_1$ besitzt. In diesem Fall geht man von der der Hypothese aus, die Grundgesamtheit habe eben diesen Mittelwert, die Hypothese lautet also $\mu_x = \mu_1$. Prüfgröße für μ_x ist das Stichprobenmittel $\bar{x}$, dessen Verteilungsfunktion für die weitere Rechnung benötigt wird. Setzt man voraus, dass x normalverteilt ist, so ist auch $\bar{x}$ normalverteilt. Es gilt

$$\varphi(\bar{x}) = \frac{1}{\sigma_{\bar{x}}\,\sqrt{2\pi}}\,\exp\Big(-\frac{1}{2}\,\Big(\frac{\bar{x}-\mu_x}{\sigma_{\bar{x}}}\Big)^2\Big) \tag{7.1}$$

mit der Varianz

$$\sigma_{\bar{x}}^2 = \frac{\sigma_x^2}{z} \tag{7.2}$$

Ist x nicht normalverteilt, so gilt, wie man aus dem Zentralen Grenzwertsatz folgert, dass $\bar{x}$ dennoch in guter Näherung normalverteilt ist, wenn der Stichprobenumfang z hinreichend groß ist.

In Abbildung 7.1 ist $\varphi(\bar{x})$ unter der Voraussetzung, dass die Grundgesamtheit den Mittelwert $\mu_x = \mu_1$ besitzt, als Kurve 1 über $\bar{x}$ aufgetragen. Dabei ist angenommen, dass die Varianz σ_x^2 bekannt ist und der Stichprobenumfang z festliegt.

Um die Grenzen des **Annahmebereichs** (E: acceptance region) für das Stichprobenmittel $\bar{x}$ festlegen zu können, muss man sich für eine bestimmte Wahrscheinlichkeit α entscheiden, beispielsweise für $\alpha = 0,05$ (5 %). Wir setzen

$$P(\mu_1 - \Delta x \le \bar{x} \le \mu_1 + \Delta x) = 1 - \alpha \tag{7.3}$$

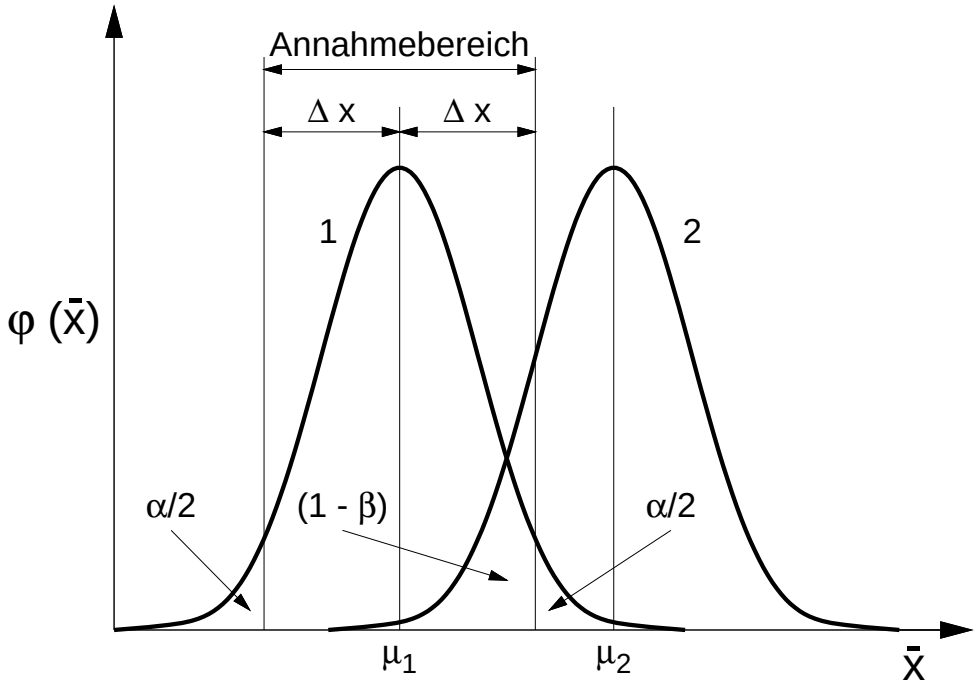


Abb. 7.1: Annahmebereich einer Hypothese. Wir empfehlen, die jeweiligen Bereiche ($\alpha/2$, $(1-\beta)$) farbig auszumalen.

womit ein zu μ_1 symmetrischer Annahmebereich der Breite $2\Delta x$ festgelegt wird. Durch die Substitution

$$t = \frac{\bar{x} - \mu_1}{\sigma_x/\sqrt{z}} \tag{7.4}$$

wird die Verteilung $\varphi(\bar{x})$ auf die normierte Normalverteilung $h(t)$ zurückgeführt. In der neuen Variablen t sind $-c$ und $+c$ die Grenzen des Annahmebereiches. Es gilt

$$P\Big(-c \leq \frac{\bar{x} - \mu_1}{\sigma_x/\sqrt{z}} \leq +c\Big) = H(c) - H(-c) = 1 - \alpha \tag{7.5}$$

Aus 7.3 und 7.5 folgt

$$\Delta x = c\,\frac{\sigma_x}{\sqrt{z}} \tag{7.6}$$

Der Annahmebereich für das Stichprobenmittel $\bar{x}$ ist dann

$$\mu_1 - c\,\frac{\sigma_x}{\sqrt{z}} \leq \bar{x} \leq \mu_1 + c\,\frac{\sigma_x}{\sqrt{z}} \tag{7.7}$$

Den Wert c entnimmt man Tabellen der normierten Normalverteilung zu einer vorgegebenen Wahrscheinlichkeit α. Aus 7.5 folgt

$$H(c) = 1 - \frac{\alpha}{2} \tag{7.8}$$

Liegt $\bar{x}$ innerhalb des Annahmebereiches, so nimmt man die Hypothese an, andernfalls verwirft man sie zu Gunsten der Alternative $\mu_x \neq \mu_1$.

Da $\bar{x}$ mit der Wahrscheinlichkeit α außerhalb des Annahmebereichs liegt, obwohl die Hypothese $\mu_x = \mu_1$ richtig ist, wird die Hypothese mit der Wahrscheinlichkeit α zu Unrecht abgelehnt (Fehler erster Art). In Abbildung 7.1 ist die Wahrscheinlichkeit α gleich der Größe der Flächen unter der Kurve 1 außerhalb des Annahmebereichs.

Wichtig für das Verständnis des Weiteren ist die Erkenntnis, dass $(1-\alpha)$ nicht die Wahrscheinlichkeit für die Gültigkeit der Hypothese ist, sondern lediglich die Wahrscheinlichkeit dafür, dass die Hypothese dann, wenn sie richtig ist, auch angenommen wird. Das ist kein kleiner, sondern ein fundamentaler Unterschied. Dies wird deutlich, wenn man den Annahmebereich beliebig vergrößern würde. Stark vom hypothetischen Mittelwert $\mu_x = \mu_1$ abweichende Stichprobenwerte $\bar{x}$ fielen dann immer noch in den Annahmebereich, während der Fehler erster Art α gegen null tendierte. Man käme zu dem unsinnigen Ergebnis, dass die Hypothese praktisch in jedem Fall anzunehmen wäre.

Besitzt die Grundgesamtheit einen Mittelwert $\mu_x = \mu_2$, der von dem hypothetischen Mittelwert μ_1 abweicht, so wird das Stichprobenmittel $\bar{x}$ dennoch mit einer bestimmten Wahrscheinlichkeit $(1-\beta)$ in den Annahmebereich fallen (Fehler zweiter Art). Dies wird durch Abbildung 7.1 veranschaulicht, in der die Kurve 2 die Verteilung des Stichprobenmittels $\bar{x}$ für den Fall darstellt, dass $\mu_x = \mu_2$ der Mittelwert der Grundgesamtheit ist. Der Teil der Fläche unter der Kurve 2, der in den Annahmebereich fällt, ist gleich der Wahrscheinlichkeit, dass die Hypothese $\mu_x = \mu_1$ angenommen wird, obwohl die Grundgesamtheit den Mittelwert $\mu_x = \mu_2$ besitzt.

Der Fehler zweiter Art lässt sich wie folgt quantifizieren. Die Ungleichung 7.7 wird durch Addition eines konstanten Wertes umgeformt zu

$$-c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}} \leq \frac{\bar{x} - \mu_2}{\sigma_x/\sqrt{z}} \leq c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}} \tag{7.9}$$

Durch die Substitution

$$t = \frac{\bar{x} - \mu_2}{\sigma_x/\sqrt{z}} \tag{7.10}$$

wird wiederum die Verteilung $\varphi(\bar{x})$ auf die normierte Normalverteilung $h(t)$ zurückgeführt. Die Wahrscheinlichkeit, dass das Stichprobenmittel $\bar{x}$ im Falle $\mu_x = \mu_2$ in den Annahmebereich fällt, ergibt sich dann mit 7.9 zu

$$\begin{aligned} P\Big(-c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}} \leq \frac{\bar{x} - \mu_2}{\sigma_x/\sqrt{z}} \;&\leq\; c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}}\Big) \\ &= H\Big(c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}}\Big) - H\Big(-c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}}\Big) \\ &= 1 - \beta \end{aligned}$$

Hieraus folgt weiter

$$1 - \beta = H\Big(c + \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}}\Big) + H\Big(c - \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}}\Big) - 1 \tag{7.11}$$

Wie man aus 7.11 erkennt, ist der Fehler zweiter Art nur vom Betrag der Abweichung vom hypothetischen Mittelwert μ_1 abhängig und nicht von ihrer Richtung. Setzt man

$$\Delta c = \frac{\mu_2 - \mu_1}{\sigma_x/\sqrt{z}} \tag{7.12}$$

so erhält man schließlich

$$1 - \beta = H(c + \Delta c) + H(c - \Delta c) - 1 \tag{7.13}$$

Aus 7.13 folgt, dass der Fehler zweiter Art umso kleiner wird, je stärker der wirkliche Wert μ_2 des Mittelwertes μ_x vom angenommenen Wert μ_1 abweicht. Dieses Ergebnis, das man auch ohne Rechnung aus Abbildung 7.1 entnehmen kann, bedeutet umgekehrt, dass eine falsche Hypothese, die sehr nahe bei der Wirklichkeit liegt, mit großer Wahrscheinlichkeit angenommen wird. Es wird deutlich, dass beliebig kleine Abweichungen Δc mit dem Prüfverfahren nicht nachgewiesen werden können. Wie groß die Abweichung Δc maximal sein darf beziehungsweise wie der Wert μ_2 im Einzelfall festzulegen ist, lässt sich nicht allgemein angeben, sondern ergibt sich aus der jeweiligen Anwendung.

Abbildung 7.1 verdeutlicht ferner, dass beide Fehler grundsätzlich unvermeidbar sind. Durch Verschieben der Grenzen des Annahmebereiches kann man die Größe der Fehler beeinflussen, jedoch nur in dem Sinn, dass man den einen Fehler auf Kosten des anderen verringert. Will man beide Fehler zugleich verringern, so lässt sich dies nur dadurch erreichen, dass man den Stichprobenumfang z vergrößert. Da die Varianz des Stichprobenmittels $\bar{x}$ wegen 7.2 mit wachsendem Stichprobenumfang kleiner wird, werden die Kurven 1 und 2 in Abbildung 7.1 entsprechend steiler und enger. Bei festgehaltenen Grenzen des Annahmebereiches vermindern sich damit beide Fehler.

Mit dem Stichprobenumfang z wächst allerdings der Aufwand für das statistische Prüfverfahren. Einer Vergrößerung des Stichprobenumfangs sind daher in der Praxis aus wirtschaftlichen Gründen Grenzen gesetzt. Wirtschaftliche Überlegungen sind auch ausschlaggebend für die Festlegung der Größe der Fehler α und $(1 - \beta)$. Hierfür sind die Kosten maßgeblich, die anfallen, wenn auf Grund des Prüfergebnisses falsche Entscheidungen getroffen werden.

Eine andere Situation liegt vor, wenn zwei unterschiedliche Hypothesen $\mu_x = \mu_1$ und $\mu_x = \mu_2$ gegeneinander stehen und mit dem Prüfverfahren die Entscheidung für die wahrscheinlichere Hypothese gefällt werden soll. Es gibt dann zwei verschiedene Annahmebereiche, die sich überschneiden können. Liegt das Stichprobenergebnis $\bar{x}$ näher bei μ_1 als bei μ_2, wird man sich für die erste Hypothese entscheiden. Auch diese Entscheidung ist jedoch mit einer gewissen Wahrscheinlichkeit falsch.

Ist der Annahmebereich beidseitig begrenzt wie in vorstehendem Fall, spricht man von einem **zweiseitigen Test** (E: two-sided test). Die Hypothese hat die Form $\mu_x = \mu_1$, die Alternative lautet $\mu_x \neq \mu_1$. Es gibt jedoch zahllose Fälle, in denen es sinnvoller ist, den Annahmebereich nur einseitig zu begrenzen. Ein Beispiel ist die Festigkeitsprüfung von Werkstoffen. Man wird ein Material nicht zurückweisen, wenn eine Stichprobe eine höhere als die verlangte Festigkeit erwarten lässt, sondern nur im umgekehrten Fall. Ebenso wird man Glühlampen mit einer längeren als der verlangten Lebensdauer nicht beanstanden. Diese Situation führt auf einen **einseitigen Test** (E: one-sided test). Die Hypothese formuliert man in der Form $\mu_x \geq \mu_1$ mit der Alternative $\mu_x < \mu_1$.

Kann man zwar voraussetzen, dass das Stichprobenmittel $\bar{x}$ normalverteilt ist, kennt aber nicht die Varianz σ_x^2 der Grundgesamtheit, so ergeben sich gegenüber dem vorstehend beschriebenen Test einige Änderungen. An Stelle von 7.4 auf Seite 125 tritt die Gleichung

$$t = \frac{\bar{x} - \mu_1}{s_x/\sqrt{z}} \tag{7.14}$$

in der die Varianz der Grundgesamtheit durch die Stichprobenvarianz

$$s_x^2 = \frac{1}{z-1} \sum_{i-1}^{z} (x_i - \bar{x})^2 \tag{7.15}$$

ersetzt ist. Die Verteilung der Größe t ist dann die Student-Verteilung mit dem Freiheitsgrad $n = z - 1$. Statt 7.5 auf Seite 125 gilt

$$P\Big(-c \leq \frac{\bar{x} - \mu_1}{\sigma_x/\sqrt{z}} \leq +c \Big) = \Phi(c) - \Phi(-c) = 1 - \alpha \tag{7.16}$$

Hierin bezeichnet $\Phi(c)$ den Wert der Summenfunktion der Studentverteilung für $t = c$ und $n = z - 1$. Ein Annahmebereich für das Stichprobenmittel lässt sich nicht mehr angeben, sondern nur noch ein Annahmebereich für t in der Form

$$-c \leq \frac{\bar{x} - \mu_1}{s_x/\sqrt{z}} \leq +c \tag{7.17}$$

Entsprechende Überlegungen sind anzustellen, wenn nicht das Stichprobenmittel, sondern die Stichprobenvarianz Prüfgröße ist. Dies trifft beispielsweise auf die Untersuchung von Feststoffmischungen zu. Als Mischgütemaß verwendet man dort die Varianz σ_x^2 der Grundgesamtheit.

Will man wissen, ob die Mischung einen bestimmten Homogenitätsgrad, gekennzeichnet durch einen bestimmten Wert σ_1^2 des Mischgütemaßes σ_x^2, erreicht hat, so zieht man eine Stichprobe im Umfang von z Einzelproben und bestimmt den Gehalt x der ersten Komponente der Mischung in den einzelnen Proben. Hieraus berechnet man mit 7.15 die Stichprobenvarianz s_x^2. Getestet wird die Hypothese $\sigma_x^2 \leq \sigma_1^2$ gegen die Alternative $\sigma_x^2 > \sigma_1^2$.

Zur Festlegung eines Annahmebereiches für die Stichprobenvarianz benötigt man deren Verteilungsfunktion. Wie in Abschnitt 6.4 *Konfidenzintervalle für die*

Varianz einer Normalverteilung mit unbekanntem Mittelwert auf Seite 114 gezeigt, wird die Verteilung von s_x^2 durch die Substitution

$$\chi^2 = (z-1)\,\frac{s_x^2}{\sigma_x^2} \tag{7.18}$$

auf die χ^2-Verteilung mit dem Freiheitsgrad $n = z - 1$ zurückgeführt.

Die Wahrscheinlichkeit, dass die Stichprobenvarianz in den einseitig begrenzten Annahmebereich

$$s_x^2 \leq s_{x_o}^2 \tag{7.19}$$

fällt, ist dann

$$P(s_x^2 \leq s_{x_o}^2) = P(\chi^2 \leq \chi_{01}^2) = \Phi(\chi_{01}^2) = 1 - \alpha \tag{7.20}$$

Hierin ist

$$\chi_{01}^2 = (z-1)\,\frac{s_{x_o}^2}{\sigma_1^2} \tag{7.21}$$

die Grenze des Annahmebereiches in der Variablen χ^2, wenn die Grundgesamtheit die Varianz $\sigma_x^2 = \sigma_1^2$ besitzt.

Auch dann, wenn die Hypothese nicht erfüllt ist, das heißt wenn die Varianz σ_x^2 einen Wert $\sigma_2^2 > \sigma_1^2$ angenommen hat, wird die Stichprobenvarianz dennoch mit einer bestimmten Wahrscheinlichkeit in den Annahmebereich fallen. Mit χ_{02}^2 wird die Grenze des Annahmebereiches in der Variablen χ^2 für den Fall bezeichnet, dass die Varianz σ_x^2 den Wert σ_2^2 besitzt. Analog zu 7.21 gilt

$$\chi_{02}^2 = (z-1)\,\frac{s_{x_0}^2}{\sigma_2^2} \tag{7.22}$$

Für den Fehler zweiter Art ergibt sich damit

$$1 - \beta = \Phi(\chi_{02}^2) \tag{7.23}$$

Wieder gilt, dass beide Fehler grundsätzlich unvermeidlich sind und dass bei dem Prüfverfahren immer beide Fehler zu berücksichtigen sind. Allgemein lässt sich sagen, dass man für jeden Verteilungsparameter, für den man ein Konfidenzintervall angeben kann, das heißt für den man eine Schätzfunktion samt deren Verteilung kennt, auch den entsprechenden Test formulieren kann.

Darüber hinaus gibt es jedoch zahlreiche spezielle Tests, zum Beispiel solche, die es festzustellen erlauben, ob die Mittelwerte oder Varianzen zweier Normalverteilungen übereinstimmen.

Außer den Tests für spezielle Verteilungsparameter (Mittelwert, Varianz) stehen auch **Tests für Verteilungen** zur Verfügung. Nimmt man beispielsweise an, eine Grundgesamtheit sei normalverteilt, so kann man diese Annahme (Hypothese) testen, indem man eine größere Stichprobe entnimmt und deren Verteilung ermittelt. Der **χ^2-Test** von K. Pearson und der **Kolmogorow-Smirnow-Test**

von A. N. Kolmogorow und Wladimir Iwanowitsch Smirnow (1887–1974) beantworten die Frage, ob die Verteilung der Stichprobe nicht gegen die Hypothese spricht. Beide Tests sind unabhängig von der speziellen, angenommenen Verteilungsfunktion. Der χ^2-Test eignet sich sowohl für stetige als auch für diskrete Verteilungen, der Kolmogorow-Smirnow-Test nur für stetige Verteilungen.

Kapitel 8

Regression und Korrelation

Mittels Regressions- und Korrelationsrechnung versuchen wir, Beziehungen zwischen zwei gleichzeitig beobachteten Variablen zu ermitteln. Solche empirisch gefundenen Beziehungen sind jedoch kein Beweis für einen Kausalzusammenhang.

8.1 Lineare Regression

8.1.1 Berechnung der Regressionsgeraden zu einer Stichprobe

Wird ein determinierter Vorgang durch einen eindeutigen Zusammenhang zwischen zwei Variablen x und y beschrieben, so kann man das Ergebnis stets in folgende Form bringen

$$y = f(x) \tag{8.1}$$

das heißt, zu jedem Wert der Variablen x lässt sich genau ein Wert der Variablen y angeben. Trägt man y über x in einem Diagramm auf, so erhält man eine Kurve wie in Abbildung 8.1.

In manchen Fällen lässt sich die Kurve ganz oder teilweise durch eine Gerade annähern. Falls dies bei linearer Auftragung (linear geteilte Achsen) nicht möglich ist, kann man eine einfach- oder doppeltlogarithmische Auftragung versuchen. Auch kompliziertere Achsenteilungen können verwendet werden. Die Erfahrung zeigt, dass sich bei geeignet gewählter Abszissen- und Ordinatenteilung viele funktionelle Zusammenhänge in guter Näherung durch Geraden darstellen lassen.

Sind zwei Variablen x und y nicht durch einen determinierten Vorgang, sondern durch einen stochastischen Vorgang miteinander verknüpft, so bedeutet dies, dass zu einem bestimmten Wert der Variablen x nicht ein bestimmter Wert der Variablen y gehört, sondern eine Verteilung der Werte von y. Trägt man experimentell ermittelte zusammengehörige Wertepaare x_i, y_i – das heißt die Ergebnisse aus einer Stichprobe – wie zuvor in einem **Streudiagramm** (E: scatterplot) auf, so liegen die Punkte niemals genau auf einer Kurve. Sie bilden eine mehr oder weniger ausgedehnte Punktwolke, die in manchen Fällen eine Struktur erkennen

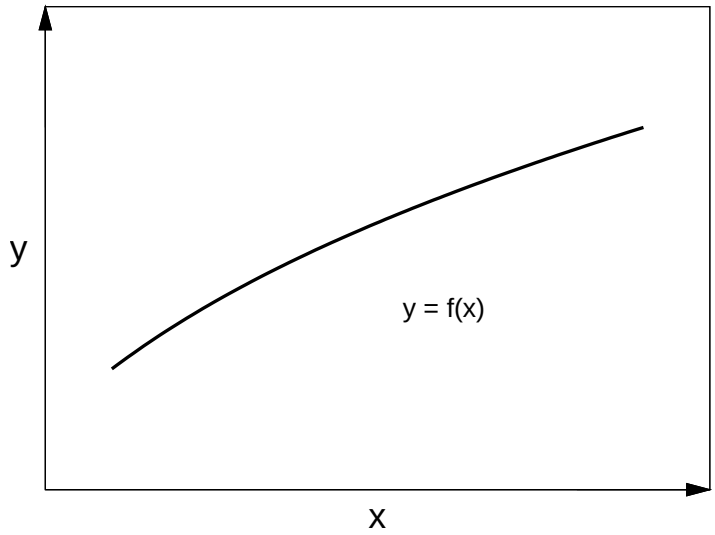


Abb. 8.1: Diagramm $y = f(x)$

lässt. Man kann jedoch versuchen, sie durch eine **Ausgleichskurve** (E: curve of best fit) zu vermitteln. Die Tätigkeit wird als **Regressionsanalyse** (E: regression analysis) bezeichnet.

Die einfachste Form einer Ausgleichskurve ist die **Ausgleichsgerade** (E: line of best fit). Man wird deshalb immer zuerst versuchen, die Messpunkte im linearen, einfach- oder doppeltlogarithmischen Netz durch eine Gerade zu vermitteln. Erscheint dieses Vorgehen möglich und streuen die Messpunkte nur wenig, so kann man die Ausgleichsgerade nach Augenmaß zeichnen. Streuen sie stärker, benötigen wir eine objektive Methode, um die Lage der Ausgleichsgeraden festzulegen. Üblich ist die Anwendung der auf C. F. Gauss zurückgehenden **Methode der kleinsten Quadrate** (E: Least Squares Method), aber es gibt auch andere Wege, beispielsweise die Hauptkomponentenanalyse (E: Pricipal Components Analysis, PCA) nach K. Pearson.

Die Ausgleichsgerade wird hierbei so gelegt, dass die Summe der Quadrate aller vertikalen Abstände zwischen den Messpunkten und der Geraden möglichst klein wird. Die x-Werte werden also als gegeben und genau betrachtet, die y-Werte als vom Zufall beeinflusst. Dies ist eine nicht selbstverständliche Vereinbarung. Berechnet wird hiermit, wie man sagt, eine **Regressionsgerade von y bezüglich** x. Wir kommen damit zur *einfachen linearen Regression*, auch bivariate lineare Regression genannt.

Die unabhängige Variable x wird als erklärende Variable, exogene Variable, Kovariable oder Regressor bezeichnet, die abhängige Variable y als interessierende Variable, endogene Variable, Zielvariable, Response oder Regressand.

Grundsätzlich sind die Variablen x und y gleichwertig. Man kann auch eine Regressionsgerade von x bezüglich y berechnen, indem man die Summe aller horizontalen Abstände zwischen den Messpunkten und der Geraden minimiert. Die beiden Geraden fallen im Allgemeinen nicht zusammen. Dies ist kein Widerspruch, da sie unterschiedlichen Fragestellungen entsprechen.

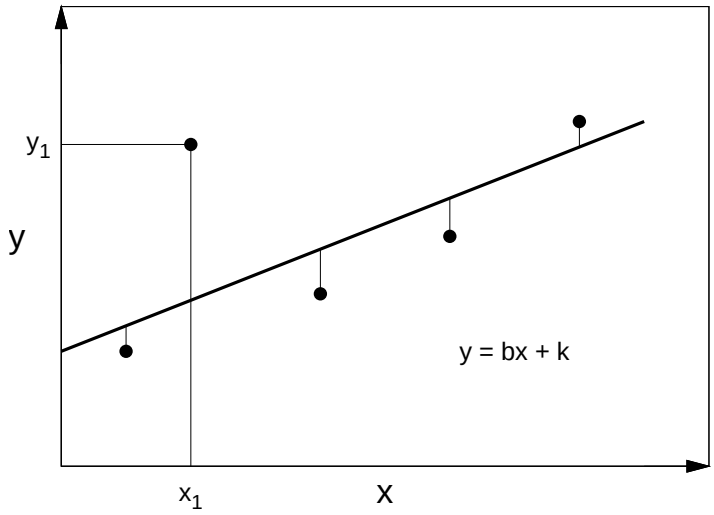


Abb. 8.2: Regressionsgerade

Die Regressionsgerade von y bezüglich x erlaubt, wie noch gezeigt wird, Aussagen über den Mittelwert der zu bestimmten Werten von x gehörigen Werte von y. Wie in Abbildung 8.2 veranschaulicht, setzen wir für die Geradengleichung an

$$y = bx + k \tag{8.2}$$

Der vertikale Abstand des Messpunktes x_i, y_i von der Geraden ist

$$\Delta y_i = y_i - bx_i - k$$

Damit wird die Summe a der Abstandsquadrate

$$a = \sum_{i=1}^{z} (y_i - bx_i - k)^2 \tag{8.3}$$

Dabei wird $z > 2$ vorausgesetzt, weil für $z = 2$ die Gerade durch die beiden Messpunkte läuft und $a = 0$ wird. Die Schreibweise von 8.3 lässt sich vereinfachen zu

$$a = \sum_{i} (y_i - bx_i - k)^2$$

Das i unter dem Summenzeichen bedeutet, dass über alle Indexwerte von 1 bis z summiert werden soll.

Die beiden unbekannten Parameter b und k der Geradengleichung erhalten wir aus den Bedingungen

$$\frac{\partial a}{\partial b} = -2 \sum_{i=1}^{z} x_i(y_i - bx_i - k) = 0 \tag{8.4}$$

$$\frac{\partial a}{\partial k} = -2 \sum_{i=1}^{z} (y_i - bx_i - k) = 0 \tag{8.5}$$

Aus 8.4 folgt weiter

$$\sum_{i=1}^{z} x_i y_i - b \sum_{i=1}^{z} x_i^2 - zk\bar{x} = 0 \tag{8.6}$$

und aus 8.5

$$\bar{y} - b\bar{x} - k = 0 \tag{8.7}$$

$$b = \frac{\sum x_i y_i - z\bar{x}\bar{y}}{\sum x_i^2 - z\bar{x}^2} \tag{8.8}$$

Mit 8.7 und 8.8 werden die Parameter b und k aus den Stichprobenwerten berechnet. Wir können 8.8 weiter umformen. Aus der Definition der Stichprobenvarianz 2.27 auf Seite 20 folgt

$$\begin{aligned} s_x^2 &= \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \\ &= \frac{1}{z-1} \left(\sum x_i^2 - 2\bar{x} \sum x_i + z\bar{x}^2 \right) \\ &= \frac{1}{z-1} \left(\sum x_i^2 - z\bar{x}^2 \right) \end{aligned} \tag{8.9}$$

So wie s_x^2 ein erwartungstreuer Schätzwert für σ_x^2 ist, gilt dies auch für die Kovarianz s_{xy} der Stichprobe gemäß der Definition 3.118 auf Seite 51

$$s_{xy} = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})(y_i - \bar{y}) \tag{8.10}$$

das heißt, man kann zeigen, dass

$$E(s_{xy}) = \sigma_{xy} \tag{8.11}$$

gilt, worin σ_{xy} die Kovarianz der Grundgesamtheit bedeutet. Aus 8.10 folgt

$$\begin{aligned} s_{xy} &= \frac{1}{z-1} \left(\sum x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i + z\bar{x}\bar{y} \right) \\ &= \frac{1}{z-1} \left(\sum x_i y_i - z\bar{x}\bar{y} \right) \end{aligned} \tag{8.12}$$

Mit 8.9 und 8.12 erhalten wir aus 8.8

$$b = \frac{s_{xy}}{s_x^2} \tag{8.13}$$

Man bezeichnet b als den **Regressionskoeffizienten der Stichprobe**. Die Geradengleichung nimmt mit 8.8 folgende Form an

$$y - \bar{y} = b\,(x - \bar{x}) \tag{8.14}$$

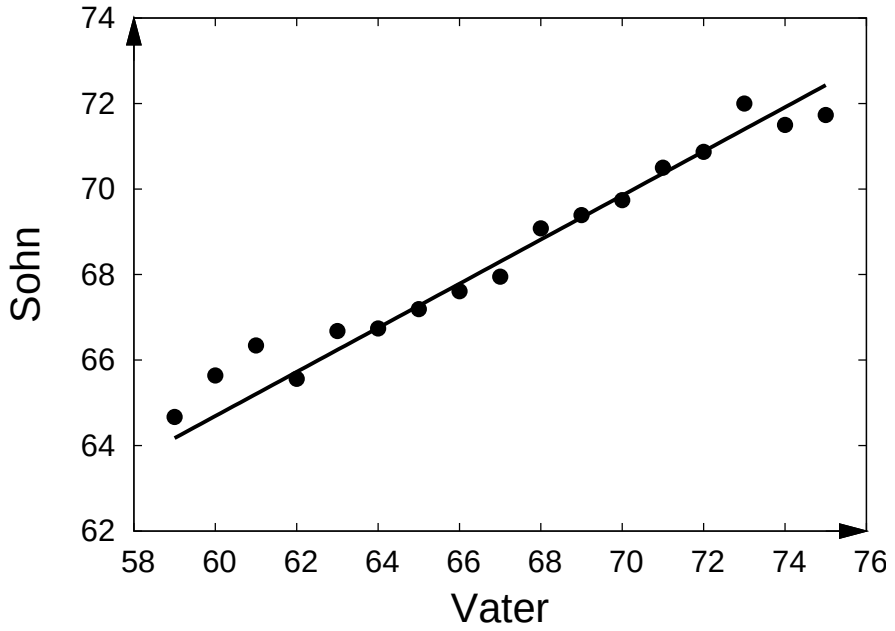


Abb. 8.3: Regressionsgerade, Körpergröße von Vätern (x) und Söhnen (y) in inch, nach K. Pearson und A. Lee

Die Ausgleichsgerade läuft durch den Punkt $\bar{x}, \bar{y}$ und hat die Steigung b. Aus 8.3 auf Seite 133 folgt mit 8.8

$$\begin{aligned} a &= \sum_{i=1}^{z} \Big((y_i - \bar{y}) - b\,(x_i - \bar{x}) \Big)^2 \\ &= \sum_{i=1}^{z} (y_i - \bar{y})^2 - 2b \sum_{i=1}^{z} (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum_{i=1}^{z} (x_i - \bar{x})^2 \\ &= (z-1)\,(s_y^2 - 2b s_{xy} + b^2 s_x^2) \end{aligned} \tag{8.15}$$

Mit 8.13 folgt daraus weiter

$$\begin{aligned} a &= (z-1)\,(s_y^2 - b s_{xy}) \\ &= (z-1)\,(s_y^2 - \frac{s_{xy}^2}{s_x^2}) \end{aligned} \tag{8.16}$$

Wir erkennen, dass a dann verschwindet, das heißt alle Messpunkte auf der Ausgleichsgeraden liegen, wenn

$$s_{xy}^2 = s_x^2\, s_y^2 \tag{8.17}$$

Dies ist ein Sonderfall. Allgemein gilt

$$s_{xy}^2 \leq s_x^2\, s_y^2 \tag{8.18}$$

Die Bezeichnung *Regression* für die Beschreibung des Zusammenhanges zwischen zwei Zufallsvariablen x und y geht auf die englischen Naturforscher Francis Galton (1822–1911) und K. Pearson zurück. Sie untersuchten die Frage, inwieweit sich Körpermerkmale von Menschen vererben. Im Diagramm 8.3 sind die Ergebnisse einer Untersuchung des Zusammenhanges zwischen der Körpergröße von Vätern (Variable x) und Söhnen (Variable y) dargestellt.[1] Die Punkte entsprechen Mittelwerten aus 1078 Beobachtungen. Wir erkennen, dass die Punkte sehr gut durch eine Gerade vermittelt werden können, laut Pearson und Lee mit der Steigung $0,516$ und dem Ordinatenabschnitt $33,73$.

Die Steigung der Geraden ist kleiner als 45 Grad, knapp 30 Grad. Dies bedeutet, dass oberhalb von etwa 70 inch die Körpergröße der Söhne im Mittel nicht die der Väter erreicht und dass unterhalb von 70 inch die Söhne im Mittel größer als ihre Väter werden. Es findet mithin eine Rückkehr (= lat. regressus) zu einer mittleren Körpergröße hin statt. Der Befund entspricht der Erwartung. Wäre es anders, würde sich die Menschheit fortlaufend auseinanderentwickeln in Riesen auf der einen und Zwerge auf der anderen Seite.

8.1.2 Konfidenzintervalle für den Regressionskoeffizienten

Der Wert des Regressionskoeffizienten b hängt vom jeweiligen Stichprobenergebnis ab und ist deshalb selbst eine Zufallsvariable. Gesucht ist der Erwartungswert β von b

$$\beta = E(b) \tag{8.19}$$

für den der mit 8.8 beziehungsweise 8.13 berechnete Wert lediglich einen Schätzwert darstellt. Der exakte Wert von β lässt sich nicht festlegen. Um für β Konfidenzintervalle angeben zu können, benötigen wir die Verteilungsfunktion von b. Aus 8.13 erhalten wir mit 8.10

$$b = \frac{1}{(z-1)\ s_x^2} \sum_{i=1}^{z} (x_i - \bar{x})(y_i - \bar{y}) \tag{8.20}$$

Hierin sind die x_i als feste, vorgegebene Werte anzusehen, während die zugehörigen y_i zufällig schwanken. Aus 8.20 folgt weiter

$$\begin{aligned} b &= \frac{1}{(z-1)\ s_x^2} \Big(\sum_{i=1}^{z} (x_i - \bar{x}) y_i - \bar{y} \sum_{i=1}^{z} (x_i - \bar{x}) \Big) \\ &= \frac{1}{(z-1)\ s_x^2} \sum_{i=1}^{z} (x_i - \bar{x}) y_i \end{aligned} \tag{8.21}$$

Um die Verteilungsfunktion von b berechnen zu können, müssen wir die Verteilungsfunktion von y für feste Werte x_i kennen, was im Allgemeinen nicht der Fall

[1] Nach K. Pearson und A. Lee *On The Laws Of Inheritance In Man*, Biometrika 2 (1903) 4, 357–462, Diagramm I und Tabelle XXII

ist. Die Situation ist ähnlich wie die bei der Berechnung von Konfidenzintervallen für die Verteilungsparameter μ_x und σ_x^2 einer eindimensionalen Verteilung, siehe Abschnitt 6 *Konfidenzintervalle für Verteilungsparameter* ab Seite 101. Dort wird vereinfachend angenommen, x sei normalverteilt, auch wenn dies keineswegs immer zutrifft. Die Annahme hat zur Folge, dass die berechneten Konfidenzintervalle nur näherungsweise gültig sind, wobei die Güte der Näherung mit dem Stichprobenumfang wächst (Zentraler Grenzwertsatz). Entsprechend gehen wir bei der Berechnung von Konfidenzintervallen für den Regressionskoeffizienten β vor. Wir nehmen an, dass

- die Variable y für feste Werte x_i normalverteilt ist und
- der Erwartungswert $E(y|x = x_i)$ durch die Beziehung

$$E(y|x = x_i) = \beta x_i + \kappa \tag{8.22}$$

  gegeben ist und
- die Varianz von y nicht von x_i abhängt.

Man kann zeigen, dass diese Annahmen immer dann exakt erfüllt sind, wenn die Verteilung $\varphi(x, y)$ durch eine zweidimensionale Normalverteilung dargestellt werden kann. Aus den in Abschnitt 6.2 *Konfidenzintervalle für den Mittelwert einer Normalverteilung mit bekannter Varianz* auf Seite 111 und 112 hergeleiteten Sätzen über die Verteilung einer Variablen v, die durch eine lineare Transformation $v = c_1x+c_2$ mit einer normalverteilten Variablen x verknüpft ist, beziehungsweise einer Variablen v, die die Summe aus einer Anzahl normalverteilter Variablen x_i ist, folgt nämlich, dass $\varphi_2^*(y|x_i)$ normalverteilt ist mit

$$E(y|x = x_i) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x_i - \mu_x) \qquad V(y|x = x_i) = \sigma_y^2(1 - \rho^2)$$

Dies entspricht vollkommen den Annahmen 8.22.

Ob die gegebene Grundgesamtheit durch eine zweidimensionale Normalverteilung approximiert werden kann, lässt sich nachprüfen, indem man die beiden Randverteilungen aufzeichnet. Das setzt allerdings voraus, dass hinreichend viele Messwertpaare vorliegen.

Mit den beiden oben genannten Sätzen folgt aus 8.21, dass die Variable b ebenfalls normalverteilt ist. Für den Erwartungswert dieser Variablen folgt aus 8.21 zunächst

$$E(b) = \frac{1}{(z-1)s_x^2} \sum_{i=1}^{z} (x_i - \bar{x})\ E(y|x = x_i) \tag{8.23}$$

Mit 8.22 folgt daraus weiter

$$E(b) \quad = \quad \frac{1}{(z-1)s_x^2} \sum_{i=1}^{z} (x_i - \bar{x})(\beta x_i + \kappa)$$

$$= \frac{1}{(z-1)s_x^2} \left(\beta \sum_{i=1}^{z} (x_i - \bar{x})\, x_i + \kappa \sum_{i=1}^{z} (x_i - \bar{x}) \right)$$

$$= \frac{\beta}{(z-1)s_x^2} \sum_{i=1}^{z} (x_i - \bar{x})\, x_i \tag{8.24}$$

Andererseits gilt

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

$$= \frac{1}{z-1} \left(\sum_{i=1}^{z} (x_i - \bar{x})\, x_i - \bar{x} \sum_{i=1}^{z} (x_i - \bar{x}) \right)$$

$$= \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})\, x_i \tag{8.25}$$

Hiermit ergibt sich aus 8.24

$$E(b) = \beta \tag{8.26}$$

Für die Varianz σ_b^2 von b liefert 8.21

$$\sigma_b^2 = \frac{1}{(z-1)^2 s_x^4} \sum_{i=1}^{z} (x_i - \bar{x})^2\, V(y|x = x_i)$$

$$= \frac{V(y|x = x_i)}{(z-1)s_x^2} \tag{8.27}$$

mit $V(y|x = x_i)$ als der Varianz von y unter der Bedingung $x = x_i$. Die transformierte Variable u

$$u = \frac{b - \beta}{\sigma_b} = (b - \beta)\, \frac{s_x \sqrt{z-1}}{\sqrt{V(y|x = x_i)}} \tag{8.28}$$

ist dann normalverteilt mit dem Mittelwert 0 und der Varianz 1. Ferner kann man zeigen, dass die Variable v

$$v = \frac{a}{V(y|x = x_i)} \tag{8.29}$$

eine χ^2-Verteilung mit $(z-2)$ Freiheitsgraden besitzt. Mit a ist in 8.29 die Summe der Abstandsquadrate gemäß 8.3 bezeichnet. Daher ist die Verteilung der Variablen t

$$t = \frac{u}{\sqrt{v/(z-2)}} = \frac{s_x \sqrt{(z-1)(z-2)}}{\sqrt{a}}\, (b - \beta) \tag{8.30}$$

eine Student-Verteilung mit $(z-2)$ Freiheitsgraden. Zu einer vorgegebenen Wahrscheinlichkeit γ liest man aus einer Tabelle der Student-Verteilung den Wert c ab und erhält damit aus 8.30 für β das folgende Konfidenzintervall

$$b - \frac{c\sqrt{a}}{s_x\sqrt{(z-1)(z-2)}} \leq \beta \leq b + \frac{c\sqrt{a}}{s_x\sqrt{(z-1)(z-2)}} \tag{8.31}$$

Umgekehrt kann man hierauf einen Test aufbauen, mit dem man prüft, ob der aus der Stichprobe berechnete Regressionskoeffizient b mit einer Hypothese bezüglich β – beispielsweise in der Form $\beta = \beta_0$ – verträglich ist.

Ferner kann man einen Test auf Linearität der Regression durchführen. Diesem Test liegt die Hypothese zu Grunde, dass die Grundgesamtheit im untersuchten Bereich durch eine Gerade nach 8.22 beschreibbar ist. Für diesen Test gilt jedoch wie für alle statistischen Tests, dass er nur bei deutlichen Abweichungen der Hypothese von der Wirklichkeit zur Ablehung der Hypothese führt. Im gegebenen Fall bedeutet dies, dass auch bei Annahme des Tests keineswegs gesichert ist, dass eine Gerade die bestmögliche Ausgleichskurve ist.

8.1.3 Konfidenzintervalle für den Mittelwert

Man benutzt die Regressionsgerade von y bezüglich x, um zu bestimmten Werten von x die zugehörigen Werte von y abzuschätzen beziehungsweise Angaben über $E(y|x = x_i)$ zu machen. Ein Schätzwert für $E(y|x = x_i)$ ist

$$\hat{y} = bx_i + k \tag{8.32}$$

Mit 8.7 von Seite 134 folgt daraus

$$\hat{y} = \bar{y} + b\,(x_i - \bar{x}) \tag{8.33}$$

Die Verteilungsfunktion von $\hat{y}$ ergibt sich unter den drei Annahmen auf Seite 137 aus den folgenden Überlegungen. Erstens ist die Variable $\bar{y}$ normalverteilt mit dem Erwartungswert

$$\begin{aligned} E(\bar{y}) &= E\Big(\frac{1}{z}\sum_{i=1}^{z} y_i\Big) \\ &= \frac{1}{z}\sum_{i=1}^{z} E(y|x = x_i) \\ &= \frac{1}{z}\sum_{i=1}^{z} (\beta x_i + \kappa) = \beta\bar{x} + \kappa \end{aligned}$$

und der Varianz

$$V(\bar{y}) = \frac{1}{z}\,V(y|x = x_i)$$

Zweitens ist die Variable $b(x_i - \bar{x})$ normalverteilt mit dem Erwartungswert

$$E\Big(b(x_i - \bar{x})\Big) = \beta(x_i - \bar{x})$$

und der Varianz

$$V\Big(b(x_i-\bar{x})\Big) = \frac{(x_i-\bar{x})^2}{(z-1)s_x^2}\, V(y|x=x_i)$$

Drittens ist dann die Variable $\hat{y}$ ebenfalls normalverteilt mit dem Erwartungswert

$$\begin{aligned} E(\hat{y}) &= E(\bar{y}) + E\Big(b(x_i-\bar{x})\Big) \\ &= \beta\bar{x} + \kappa + \beta(x_i-\bar{x}) \\ &= \beta x_i + \kappa = E(y|x=x_i) \end{aligned} \tag{8.34}$$

und der Varianz

$$\begin{aligned} V(\hat{y}) &= V(\bar{y}) + V\Big(b(x_i-\bar{x})\Big) \\ &= \frac{1}{z}\, V(y|x=x_i) + \frac{(x_i-\bar{x})^2}{(z-1)s_x^2}\, V(y|x=x_i) \\ &= \Big(\frac{1}{z} + \frac{(x_i-\bar{x})^2}{(z-1)s_x^2}\Big)\, V(y|x=x_i) \end{aligned} \tag{8.35}$$

Wir kürzen ab

$$h^2 = \frac{1}{z} + \frac{(x_i-\bar{x})^2}{(z-1)s_x^2}$$

Die transformierte Variable u

$$u = \frac{\hat{y}-E(\hat{y})}{\sqrt{V(\hat{y})}} = \frac{\bar{y}+b(x_i-\bar{x})-E(y|x=x_i)}{h\,\sqrt{V(y|x=x_i)}}$$

ist dann normalverteilt mit dem Mittelwert 0 und der Varianz 1. Mit 8.29 von Seite 138 wird hieraus wie im vorigen Abschnitt eine Variable t gebildet, die eine Student-Verteilung mit $(z-2)$ Freiheitsgraden besitzt. Es gilt

$$t = \frac{u}{\sqrt{v/(z-2)}} = \frac{\bar{y}+b(x_i-\bar{x})-E(y|x=x_i)}{h\,\sqrt{a/(z-2)}}$$

Mit der Abkürzung

$$l = c\,\frac{h\,\sqrt{a}}{\sqrt{z-2}}$$

erhalten wir für $E(y|x=x_i)$ das Konfidenzintervall

$$\bar{y}+b(x_i-\bar{x})-l \le E(y|x=x_i) \le \bar{y}+b(x_i-\bar{x})+l \tag{8.36}$$

Diese Rechnung können wir für alle Messpunkte x_i ausführen und erhalten so einen Konfidenzbereich für die Erwartungswerte $E(y|x=x_i)$.

Wichtig scheint an dieser Stelle der Hinweis, dass alle in diesem Abschnitt abgeleiteten Gleichungen nur unter der strengen Voraussetzung gelten, dass die Wertepaare x_i, y_k rein zufällig herausgegriffen sind. Dies gilt nicht für Zufallsexperimente, bei denen der Wert einer der beiden Variablen eingestellt oder vorgegeben werden kann. Dieser Wert wäre dann nicht mehr als Wert einer Zufallsvariablen, sondern als ein äußerer Versuchsparameter anzusehen. Die Regressionsrechnung ist auf solche Fälle nicht anwendbar. Die Lösung solcher Probleme ist jedoch bedeutend einfacher, wie an folgendem Beispiel gezeigt werden kann.

Gesucht sei der Mittelwert der Variablen y für feste, einstellbare Werte von x_i. Dazu werden z Wiederholungsmessungen für jedes x_i durchgeführt und wie in Abschnitt 6 *Konfidenzintervalle für Verteilungsparameter* ab Seite 101 beschrieben ausgewertet. Man erhält zu jedem x_i ein Konfidenzintervall für den Mittelwert $E(y|x_i)$ in der Form

$$\bar{y} - c\frac{s_y}{\sqrt{z}} \leq E(y|x = x_i) \leq \bar{y} + c\frac{s_y}{\sqrt{z}}$$

Dazu brauchen keine zusätzlichen Annahmen wie die auf Seite 137 formulierten erfüllt zu sein.

8.2 Nichtlineare Regression

Lassen sich die Messpunkte weder im linearen noch im einfach- oder doppeltlogarithmischen Netz durch eine Gerade vermitteln, so können wir versuchen, die Regression der Variablen y bezüglich der Variablen x durch ein Polynom höherer Ordnung zu beschreiben. Auch andere Funktionen als Polynome kommen in Frage, beispielsweise trigonometrische Funktionen. Die mathematischen Schwierigkeiten wachsen jedoch beträchtlich. Allgemein setzen wir für ein Polynom an

$$y = b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n \tag{8.37}$$

Zur Berechnung der Koeffizienten b_i in 8.37 wird wiederum von der Methode der kleinsten Quadrate (Seite 132) Gebrauch gemacht. Aus

$$a = \sum_{i=1}^{z} (y_i - b_0 - b_1 x_i - b_2 x_i^2 - \ldots - b_n x_i^n)^2 \tag{8.38}$$

erhalten wir mit

$$\frac{\partial a}{\partial b_0} = 0 \qquad \frac{\partial a}{\partial b_1} = 0 \qquad \frac{\partial a}{\partial b_2} = 0 \qquad \ldots \qquad \frac{\partial a}{\partial b_n} = 0 \tag{8.39}$$

genau $(n+1)$ Bestimmungsgleichungen für die $(n+1)$ unbekannten Koeffizienten $b_0, b_1, b_2 \ldots b_n$.

8.3 Korrelation

8.3.1 Korrelationskoeffizient

Gegeben sei eine Stichprobe aus z Wertepaaren x_i, y_i aus einer zweidimensionalen Grundgesamtheit. Hierzu lassen sich berechnen die Mittelwerte

$$\bar{x} = \frac{1}{z} \sum_{i=1}^{z} x_i \qquad \bar{y} = \frac{1}{z} \sum_{i=1}^{z} y_i$$

die Varianzen

$$s_x^2 = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{z-1} \sum_{i=1}^{z} (y_i - \bar{y})^2$$

und die Kovarianz

$$s_{xy} = \frac{1}{z-1} \sum_{i=1}^{z} (x_i - \bar{x})(y_i - \bar{y})$$

Der Quotient

$$r = \frac{s_{xy}}{s_x \; s_y} \tag{8.40}$$

wird als **Korrelationskoeffizient der Stichprobe** bezeichnet, sein Quadrat unter bestimmten Umständen als **Bestimmtheitsmaß** R^2 (E: coefficient of determination). Mit der Ungleichung 8.18 von Seite 135

$$s_{xy}^2 \leq s_x^2 \; s_y^2$$

folgt aus 8.40

$$-1 \leq r \leq +1 \tag{8.41}$$

Den Mittelwerten und Varianzen der Stichprobe entsprechen die Mittelwerte und Varianzen der Grundgesamtheit. Es gilt

$$\mu_x = E(x) \qquad \mu_y = E(y)$$

$$\sigma_x^2 = E\Big((x - \mu_x)^2\Big) \qquad \sigma_y^2 = E\Big((y - \mu_y)^2\Big)$$

$$\sigma_{xy} = E\Big((x - \mu_x)(y - \mu_y)\Big)$$

Der Quotient

$$\rho = \frac{\sigma_{xy}}{\sigma_x \; \sigma_y} \tag{8.42}$$

wird als **Korrelationskoeffizient der Grundgesamtheit** bezeichnet. Man kann zeigen, dass auch für den Korrelationskoeffizienten der Grundgesamtheit gilt

$$-1 \leq \rho \leq +1 \tag{8.43}$$

Hierzu wird eine neue Variable v eingeführt, die mit den Variablen x und y durch die Beziehung

$$v = t\,\frac{x-\mu_x}{\sigma_x} + \frac{y-\mu_y}{\sigma_y} \tag{8.44}$$

verbunden ist, wobei t eine feste reelle Zahl sein soll. Für den Erwartungswert von v folgt aus 8.44

$$\mu_v = E(v) = 0 \tag{8.45}$$

Für die Varianz von v erhalten wir

$$\begin{aligned}
\sigma_v^2 &= E\Big((v-\mu_v)^2\Big) = E(v^2) \\
&= t^2\,E\Big((\frac{x-\mu_x}{\sigma_x})^2\Big) + 2t\,E\Big(\frac{x-\mu_x}{\sigma_x}\;\frac{y-\mu_y}{\sigma_y}\Big) + E\Big((\frac{y-\mu_y}{\sigma_y})^2\Big) \\
&= t^2 + 2t\,\rho + 1 = (t+\rho)^2 + (1-\rho^2)
\end{aligned} \tag{8.46}$$

Die Varianz von v muss immer eine Zahl ≥ 0 sein, gleich welchen Wert die reelle Zahl t hat. Aus

$$(t+\rho)^2 + (1-\rho)^2 \geq 0 \tag{8.47}$$

folgt

$$(1-\rho)^2 \geq 0 \tag{8.48}$$

da andernfalls für $t = -\rho$ die Bedingung 8.47 nicht erfüllt wäre. Dieses Ergebnis ist identisch mit 8.43.

Eine wichtige Frage ist die, was der Zahlenwert des Korrelationskoeffizienten über den Zusammenhang zwischen den Variablen x und y aussagt. Ist $\rho = 0$, so dürfen wir im Fall der zweidimensionalen Normalverteilung folgern, dass die Variablen x und y voneinander unabhängig sind. Im Fall anderer zweidimensionaler Verteilungen ist dieser Schluss jedoch nicht erlaubt, wie in Abschnitt 5.3 *Zweidimensionale Normalverteilung* auf Seite 90 nachzulesen ist.

Für den anderen Extremfall $\rho = \pm 1$ gilt, dass zwischen zwei Zufallsvariablen x und y genau dann eine lineare Beziehung

$$y = \beta x + \gamma \tag{8.49}$$

besteht, wenn der Korrelationskoeffizient ρ den Wert $+1$ oder -1 hat.

Zum Beweis gehen wir davon aus, dass eine lineare Beziehung der Form 8.49 besteht. Daraus folgt

$$\mu_y = \beta\mu_x + \gamma \tag{8.50}$$

$$\sigma_y^2 = \beta^2\ \sigma_x^2 \quad \text{daraus:} \quad \sigma_y = \pm\beta\sigma_x \tag{8.51}$$

Subtrahiert man 8.50 von 8.49, so erhält man

$$y - \mu_y = \beta\ (x - \mu_x) \tag{8.52}$$

Hiermit erhalten wir für die Kovarianz

$$\sigma_{xy} = E\Big((x-\mu_x)(y-\mu_y)\Big) = \beta\ E\Big((x-\mu_x)^2\Big) = \beta\ \sigma_x^2 \tag{8.53}$$

Aus 8.42 folgt mit 8.51 und 8.52

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} = \pm 1$$

Gehen wir umgekehrt davon aus, dass $\rho = \pm 1$ ist, und setzen wir $t = -\rho$, so folgt aus 8.46

$$\sigma_v^2 = 0$$

Dies bedeutet, dass die Variable v in 8.44 nur einen einzigen Wert annehmen kann, nämlich

$$v = \mu_v = 0$$

Damit folgt aus 8.44

$$\frac{y-\mu_y}{\sigma_y} = \pm\frac{x-\mu_x}{\sigma_x} \tag{8.54}$$

Dies ist die gleiche lineare Beziehung wie 8.52.

Im Extremfall $\rho = \pm 1$ besteht zwischen den Variablen x und y die lineare Beziehung 8.49. Daraus können wir schließen, dass allgemein der Zusammenhang zwischen x und y umso genauer durch eine lineare Beziehung darstellbar ist, je näher der Wert von ρ bei $+1$ oder -1 liegt. Der Korrelationskoeffizient ρ ist also kein Maß für die Abhängigkeit schlechthin, wohl aber ein Maß für die lineare Abhängigkeit beziehungsweise für die Güte der Beschreibung des Zusammenhanges zwischen x und y durch einen linearen Ansatz, das heißt durch eine Regressionsgerade.

8.3.2 Konfidenzintervalle für den Korrelationskoeffizienten

Zu dem Korrelationskoeffizienten r der Stichprobe wird eine Hilfsgröße $z^\star$ berechnet

$$z^\star = \frac{1}{2}\ \ln\Big(\frac{1+r}{1-r}\Big) \tag{8.55}$$

Wie R. A. FISHER bewiesen hat, ist die Größe $z^\star$ normalverteilt mit dem Mittelwert

$$\mu_{z^\star} = \frac{1}{2}\ \ln\Big(\frac{1+\rho}{1-\rho}\Big) \tag{8.56}$$

und der Varianz

$$\sigma^2_{z^\star} = \frac{1}{z-3} \tag{8.57}$$

Damit lässt sich die Wahrscheinlichkeit

$$P\Big(-c \le \frac{z^\star - \mu_{z^\star}}{\sigma_{z^\star}} \le +c\Big) = \gamma \tag{8.58}$$

vorgeben und aus einer Tabelle der normalisierten Normalverteilung der zugehörige Wert c ablesen. Für $\mu_{z^\star}$ erhalten wir auf diese Weise das Konfidenzintervall

$$z^\star - \frac{c}{\sqrt{z-3}} \le \mu_{z^\star} \le z^\star + \frac{c}{\sqrt{z-3}} \tag{8.59}$$

Aus 8.56 folgt für den Korrelationskoeffizienten ρ der Grundgesamtheit

$$\rho = \tanh \mu_{z^\star} = \frac{\exp \mu_{z^\star} - \exp\left(-\mu_{z^\star}\right)}{\exp \mu_{z^\star} + \exp -\mu_{z^\star})} \tag{8.60}$$

Das Konfidenzintervall für ρ erhalten wir dann zu

$$\tanh\Big(z^\star - \frac{c}{\sqrt{z-3}}\Big) \le \rho \le \tanh\Big(z^\star + \frac{c}{\sqrt{z-3}}\Big) \tag{8.61}$$

Darin ist tanh() der Tangens hyperbolicus, eine den Kreisfunktionen verwandte Hyperbelfunktion.

Kapitel 9

Anwendungen

Hier folgen einige Anwendungen der Werkzeuge aus den vorhergehenden Abschnitten auf Fragen aus der Verfahrenstechnik.

9.1 Zweidimensionale Verteilungen von Partikelgröße und -geschwindigkeit disperser Phasen in Gas- oder Flüssigkeitsströmungen[1]

9.1.1 Aufgabenstellung

Zur Messung der Partikelgrößenverteilung suspendierter Feststoffpartikeln, Gasblasen oder Flüssigkeitstropfen in strömenden Kontinua wird man vorzugsweise Messverfahren anwenden, die die Strömung selbst nicht stören. Dies sind vor allem optische Verfahren der Dispersitätsanalyse.

Dabei können zwei unterschiedliche Messvorschriften angewendet werden, die zu unterschiedlichen Ergebnissen führen. Im ersten Fall werden zu einem bestimmten Zeitpunkt alle die Partikeln erfasst, die sich in einem abgegrenzten Volumen ΔV befinden. Ein entsprechendes Messverfahren ist die Auswertung eines Impulshologramms. Im zweiten Fall werden alle die Partikeln gemessen und registriert, die in einer Zeitspanne Δt durch ein kleines Messvolumen ΔV hindurchströmen. Hierzu ist die Streulicht-Partikelanalyse ein geeignetes Messverfahren.

Weshalb nicht in beiden Fällen die gleiche Partikelgrößenverteilung gemessen wird, liegt daran, dass die Partikeln sich im Allgemeinen mit unterschiedlichen Geschwindigkeiten durch das Messvolumen und relativ zum Kontinuum bewegen. Es zeigt sich, dass man neben der Partikelgröße x die Partikelgeschwindigkeit w als zweite Variable einführen muss, um die Unterschiede zwischen den nach den beiden Messvorschriften ermittelten Partikelgrößenverteilungen beschreiben zu können.

[1] J. Raasch und H. Umhauer, Grundsätzliche Überlegungen zur Messung der Verteilungen von Partikelgröße und Partikelgeschwindigkeit disperser Phasen in Strömungen, Chemie-Ingenieur-Technik 49 (1977) 12, S. 931–941

9.1.2 Mathematischer Ansatz

Für die Herleitung der nachfolgenden mathematischen Beziehungen wird vorausgesetzt, dass die gegebene Partikelströmung in dem Sinne stationär ist, dass während der Messdauer Δt

- die mittleren Geschwindigkeiten der beiden Phasen konstant sind und
- Menge (Konzentration) und Partikelgrößenverteilung der dispersen Phase sich nicht ändern.

Ferner sei hinsichtlich der Größe des Messvolumens vorausgesetzt, dass sich die zu messenden Größen (Partikelkonzentration, -größe, -geschwindigkeit) innerhalb des Messvolumens ΔV nicht merklich ändern.

Man denkt sich ein quaderförmiges Messvolumen innerhalb des Strömungsraumes abgegrenzt und so orientiert, dass die Hauptströmungsrichtung senkrecht auf einer der Seitenflächen steht. Weiterhin wird vorausgesetzt, dass die beiden anderen Komponenten der Partikelgeschwindigkeit gegenüber der Komponente w in der Hauptströmungsrichtung vernachlässigt werden können.

Zunächst soll untersucht werden, welche Aussage man erhält, wenn man mittels eines geeigneten Messverfahrens die Größe x und die Geschwindigkeit w aller der Partikeln bestimmt, die sich zu einem Zeitpunkt innerhalb des Messvolumens befinden (Messvorschrift 1). Werden x und w gleichzeitig an jeder einzelnen Partikel gemessen, so wird durch das Messverfahren jeder Partikel ein Wertepaar x_i, w_k zugeordnet, das heißt, es wird eine zweidimensionale Häufigkeitsverteilung aufgenommen, wie wir sie in Abschnitt 3.3.2 *Verteilungen mehrerer Zufallsvariablen* auf Seite 37 kennen gelernt haben. Bezieht man das Ergebnis auf die Einheit des Messvolumens, so bezeichnet

$$\mathrm{d}^2 n = f(x, w)\ \mathrm{d}x\mathrm{d}w \tag{9.1}$$

die Anzahl der Partikeln in der Volumeneinheit, deren Größe im Bereich von x bis $x + \mathrm{d}x$ und deren Geschwindigkeit im Bereich von w bis $w + \mathrm{d}w$ liegt. Integriert man über den ganzen Wertebereich beider Variablen, so erhält man mit

$$n = \int_0^\infty \int_0^\infty f(x, w)\ \mathrm{d}x\mathrm{d}w \tag{9.2}$$

die Gesamtanzahl der Partikeln in der Volumeneinheit an der untersuchten Stelle im Strömungsraum.

Integriert man nur über jeweils eine der beiden Variablen und dividiert durch n, so erhält man mit

$$q_0^{(1)}(x) = \frac{\int f(x, w)\ \mathrm{d}w}{\int\int f(x, w)\ \mathrm{d}x\mathrm{d}w} \tag{9.3}$$

$$h_0^{(1)}(w) = \frac{\int f(x, w)\ \mathrm{d}x}{\int\int f(x, w)\ \mathrm{d}x\mathrm{d}w} \tag{9.4}$$

die normierten Anzahldichteverteilungen von Partikelgröße x und Partikelgeschwindigkeit w als Randverteilungen einer zweidimensionalen Verteilung.

Nun soll der andere Fall untersucht werden, das heißt, es soll davon ausgegangen werden, dass mit einem geeigneten Messverfahren die Größe x und die Geschwindigkeit w aller der Partikeln bestimmt werden, die sich innerhalb der Messdauer Δt durch das Messvolumen ΔV hindurch bewegen (Messvorschrift 2). Dabei wird vereinfachend angenommen, dass nur solche Partikeln erfasst werden, die durch die der Hauptströmung zugewandte Seitenfläche des Messvolumens in dieses eindringen.

Werden wiederum x und w gleichzeitig an jeder einzelnen Partikel gemessen, so wird auch in diesem Fall eine zweidimensionale Häufigkeitsverteilung bestimmt. Bezieht man das Ergebnis auf die Zeit- und Flächeneinheit, so ist

$$\mathrm{d}^2\gamma = g(x,w)\,\mathrm{d}x\mathrm{d}w \tag{9.5}$$

die Anzahl der Partikeln, deren Größe im Bereich von x bis $x + \mathrm{d}x$ und deren Geschwindigkeit im Bereich von w bis $w + \mathrm{d}w$ liegt und die pro Zeit- und Flächeneinheit durch das Messvolumen hindurch strömen. Integriert man über den gesamten Wertebereich beider Variablen, so erhält man mit

$$\gamma = \int_0^\infty \int_0^\infty g(x,w)\,\mathrm{d}x\mathrm{d}w \tag{9.6}$$

die Partikelstromdichte an der untersuchten Stelle im Strömungsraum. Integriert man nur über jeweils eine der beiden Variablen und dividiert durch die Partikelstromdichte γ, so erhält man mit

$$q_0^{(2)}(x) = \frac{\int g(x,w)\,\mathrm{d}w}{\int\int g(x,w)\,\mathrm{d}x\mathrm{d}w} \tag{9.7}$$

$$h_0^{(2)}(w) = \frac{\int g(x,w)\,\mathrm{d}x}{\int\int g(x,w)\,\mathrm{d}x\mathrm{d}w} \tag{9.8}$$

die normierten Anzahldichteverteilungen von Partikelgröße x und Partikelgeschwindigkeit w wiederum als Randverteilungen einer zweidimensionalen Verteilung.

9.1.3 Folgerungen

Die gemäß Messvorschrift 1 bestimmten Verteilungen werden durch eine hochgestellte, geklammerte 1 gekennzeichnet, die gemäß Messvorschrift 2 durch eine entsprechende 2. Damit wird deutlich gemacht, dass die nach den beiden unterschiedlichen Messvorschriften bestimmten Verteilungen nicht identisch sind. Man könnte hieraus fälschlicherweise den Schluss ziehen, nach einer der beiden Messvorschriften erhielte man die *richtige* Verteilung für x oder w und nach der anderen Messvorschrift eine *falsche*. Tatsächlich erhält man jedoch in jedem Fall eine physikalisch sinnvolle und eindeutige Aussage. Zu prüfen ist lediglich, ob die

mit dem einen oder anderen Messverfahren bestimmte Verteilung der jeweiligen Fragestellung entspricht.

Wie man leicht zeigen kann, lassen sich die nach den beiden unterschiedlichen Messvorschriften bestimmten Verteilungen der Partikelgeschwindigkeit w problemlos ineinander umrechnen. Mit der einfachen Beziehung

$$g(x, w) = w\ f(x, w) \tag{9.9}$$

erhält man die Umrechnungsgleichungen

$$h_0^{(2)}(w) = \frac{w\ h^{(1)}(w)}{\int w h^{(1)}(w)\ \mathrm{d}w} \tag{9.10}$$

$$h_0^{(1)}(w) = \frac{w^{-1}\ h^{(2)}(w)}{\int w^{-1}\ h^{(2)}(w)\ \mathrm{d}w} \tag{9.11}$$

Entprechende einfache Beziehungen zwischen den nach den beiden Messvorschriften bestimmten Partikelgrößenverteilungen gibt es nicht. Es zeigt sich, dass eine Umrechnung nur bei Kenntnis der zugehörigen zweidimensionalen Verteilung unter Verwendung von 9.9 möglich ist.

Die Bestimmung zweidimensionaler Verteilungen ist messtechnisch sehr aufwendig. Man wird sie deshalb nach Möglichkeit vermeiden. Wenn eine Partikelgrößenverteilung gemessen werden soll, wird man immer zuerst nach einem Messverfahren suchen, das der Fragestellung entspricht.

Nicht zu vermeiden ist die Bestimmung einer zweidimensionalen Verteilung jedoch dann, wenn zusätzliche Fragen zu beantworten sind, wenn zum Beispiel die Geschwindigkeitsverteilung einer bestimmten Größenklasse x angegeben werden soll (bedingte Verteilung, siehe Abschnitt 3.3.2 *Verteilungen mehrerer Zufallsvariablen* auf Seite 37) oder wenn untersucht werden soll, ob eine funktionelle Abhängigkeit zwischen den Variablen x und w besteht beziehungsweise wenn der Grad der Korrelation zwischen den Variablen ermittelt werden soll.

9.2 Koinzidenzfehler bei der Streulicht-Partikelgrößen-Zählanalyse[2]

9.2.1 Aufgabenstellung

Mit einem Streulichtzählverfahren aus der Partikelgrößenanalyse (Dispersitätsanalyse) wird die Größenverteilung von Partikeln bestimmt, die sich mit einer Gas- oder Flüssigkeitsströmung mitbewegen. Der besondere Vorteil dieses Messverfahrens liegt darin, dass durch den Messvorgang die Strömung selbst nicht gestört wird. Verwirklicht wird das Prinzip dadurch, dass ein eng gebündelter

[2] J. Raasch und H. Umhauer, Der Koinzidenzfehler bei der Streulicht-Partikelgrößen-Zählanalyse, Fortschritt-Berichte VDI-Z. Reihe 3 Nr. 95, Düsseldorf: VDI-Verlag 1984

Lichtstrahl in den Strömungsraum gerichtet wird und dass senkrecht zur Richtung des Strahls ein Lichtempfänger das Streulicht solcher Partikeln registriert, die sich gerade in einem kleinen Teilvolumen ΔV, dem Messvolumen, aufhalten. Die Abgrenzung des Messvolumens wird mit rein optischen Mitteln (Linsen, Blenden) in beiden Strahlengängen erreicht. Partikel, die sich außerhalb des Messvolumens befinden, werden entweder nicht beleuchtet oder nicht beobachtet oder beides nicht.

Die Größe des Messvolumens muss den Gegebenheiten angepasst werden, das heißt so eingestellt werden, dass sich nach Möglichkeit stets nur eine Partikel im Messvolumen befindet. Andernfalls addiert der Lichtempfänger die Streulichtimpulse, und es werden größere Partikeln vorgetäuscht (Koinzidenzfehler). Die Größe des Messvolumens kann andererseits auch nicht beliebig klein gewählt werden, weil damit die Wahrscheinlichkeit wächst, dass sich während des Messvorgangs Partikeln teils innerhalb, teils außerhalb des Messvolumens befinden und damit kleinere Partikeln vorgetäuscht werden (Randzonenfehler). Hier soll allein die Frage untersucht werden, wie der Koinzidenzfehler von der Größe des Messvolumens abhängt.

9.2.2 Mathematischer Ansatz

Wir gehen davon aus, dass die Partikeln zufällig in der Strömung verteilt sind mit einer mittleren Partikelanzahl λ pro Volumeneinheit als Konzentrationsmaß. Ferner nehmen wir an, dass sich alle Partikeln mit gleicher Geschwindigkeit, das heißt mit der Geschwindigkeit des Trägermediums, durch das Messvolumen bewegen. Als Variable x wird die Anzahl von Partikeln bezeichnet, die sich zu einem bestimmten Zeitpunkt gleichzeitig im Messvolumen ΔV befinden. Dann gilt

$$P(x;\Delta V) = \frac{(\lambda\Delta V)^x}{x!}\exp\left(-\lambda\Delta V\right) \tag{9.12}$$

Die Wahrscheinlichkeit dafür, dass sich keine Partikeln im Messvolumen aufhalten, ergibt sich aus 9.12 zu

$$P(0;\Delta V) = \exp\left(-\lambda\Delta V\right) \tag{9.13}$$

Das Messgerät arbeitet mit kontinuierlicher Beleuchtung des Messvolumens. Eine Partikel sendet so lange Streulicht aus, wie sie sich im Messvolumen befindet. Die Wahrscheinlichkeit, dass sich das Streulichtsignal einer ersten Partikel nicht mit dem einer zweiten Partikel überschneidet, ist daher gleich der Wahrscheinlichkeit, dass auf die erste Partikel ein Gas- oder Flüssigkeitsvolumen ΔV folgt, in dem sich keine Partikeln befinden. Die Wahrscheinlichkeit, ein Streulichtsignal zu messen, das von einer einzigen Partikel stammt, ist folglich

$$P(1) = P(0;\Delta V) = \exp\left(-\lambda\Delta V\right) \tag{9.14}$$

Der Koinzidenzfehler f ist gleich der Wahrscheinlichkeit, dass sich das Streulichtsignal einer ersten Partikel mindestens mit dem einer zweiten Partikel überschneidet. Mit 9.14 folgt

$$f = 1 - P(1) = 1 - \exp\left(-\lambda\Delta V\right) \tag{9.15}$$

Mit Zahlenwerten: Das Messvolumen habe die Größe $\Delta V = 5 \cdot 10^{-5}$ cm^3, die Partikelkonzentration betrage $\lambda = 2000$ cm^{-3}. Dann gilt

$$\begin{aligned} \lambda\Delta V &= 2000 \cdot 5 \cdot 10^{-5} = 0,1 \\ P(1) &= P(0;\Delta V) = \exp(-\lambda\Delta V) = 0,90484 \\ f &= 1 - P(1) = 0,09516 \end{aligned}$$

Einen maximalen Fehler dieser Größe lässt man in der Praxis zu, indem man für Streulichtmessgeräte vorschreibt, dass zwischen Messvolumengröße ΔV und Partikelkonzentration λ der Zusammenhang

$$\lambda\Delta V \leq 0,1$$

bestehen soll.

Die meisten Koinzidenzen werden Zweier-Koinzidenzen sein, also Überlagerungen der Streulichtsignale von zwei Partikeln, die sich gleichzeitig im Messvolumen befinden. Ihre Wahrscheinlichkeit ist gleich dem Produkt der Wahrscheinlichkeiten zweier Ereignisse: Erstens muss die Entfernung zwischen erster und zweiter Partikel kleiner als und zweitens die zwischen zweiter und dritter Partikel größer als die Länge des Messvolumens sein. Daher gilt

$$\begin{aligned} P(2) &= \Big(1 - P(0,\Delta V)\Big)\, P(0,\Delta V) \\ &= \Big(1 - \exp(-\lambda\Delta V)\Big)\, \exp(-\lambda\Delta V) \end{aligned} \tag{9.16}$$

Dies lässt sich beliebig erweitern. Die Wahrscheinlichkeit für ein Streulichtsignal, das durch die Überlagerung der Signale von z Partikeln zustande kommt, ist

$$P(z) = \Big(1 - \exp(-\lambda\Delta V)\Big)^{z-1}\, \exp(-\lambda\Delta V) \tag{9.17}$$

Die Summe der geometrischen Reihe

$$\sum_{z=1}^{\infty} P(z) = \exp(-\lambda\Delta V) \sum_{z=1}^{\infty} \Big(1 - \exp(-\lambda\Delta V)\Big)^{z-1} \tag{9.18}$$

hat, wie man leicht nachrechnen kann, den Wert 1 (siehe Abschnitt 3.4.1 *Eindimensionale Wahrscheinlichkeitsverteilungen* auf Seite 44). Dies muss so sein, denn eine erste in das Messvolumen eintretende Partikel liefert immer ein Streulichtsignal, entweder allein oder zusammen mit einer zweiten oder zusammen mit beliebig vielen folgenden Partikeln.

Wie die Wahrscheinlichkeit von Koinzidenzen von $\lambda\Delta V$, dem Produkt aus Partikelkonzentration und Größe des Messvolumens, abhängt, zeigt Tabelle 9.1.

Tab. 9.1: Wahrscheinlichkeit von Koinzidenzen in Abhängigkeit von $\lambda\Delta V$

$\lambda\Delta V$	$P(1)$	$P(2)$	$P(3)$	$P(4)$
0,1	0,90484	0,08611	0,00819	0,00078
0,2	0,81873	0,14841	0,02690	0,00488
0,3	0,74082	0,19201	0,04976	0,01290
0,4	0,67032	0,22099	0,07286	0,02402
0,5	0,60653	0,23865	0,09390	0,03695

Die Streulicht-Partikelgrößen-Zählanalyse dient in erster Linie dazu, Partikelgrößenverteilungen zu bestimmen. Sie kann unter bestimmten Voraussetzungen jedoch auch zur Ermittlung der Partikelkonzentration verwendet werden. Wenn kein Totzeitfehler berücksichtigt werden muss, gilt

$$N_0 = N \sum_{z=1}^{\infty} zP(z) \tag{9.19}$$

Hierin bezeichnet N_0 die Anzahl der Partikeln, die während der Messdauer das Messvolumen durchquert haben, und N die Anzahl der in derselben Zeit registrierten Streulichtsignale. Mit 9.17 folgt aus 9.19

$$N_0 = N \ \exp\left(-\lambda\Delta V\right) \sum_{z=1}^{\infty} z\Big(1 - \exp\left(-\lambda\Delta V\right)\Big)^{z-1} \tag{9.20}$$

Den Wert dieser unendlichen Reihe erhält man durch gliedweises Differenzieren der geometrischen Reihe (siehe Abschnitt 3.4.1 *Eindimensionale Wahrscheinlichkeitsverteilungen* auf Seite 44). Dann ergibt sich

$$N_0 = N \ \exp\left(\lambda\Delta V\right) \tag{9.21}$$

Andererseits gilt auch

$$N_0 = \lambda V \tag{9.22}$$

wenn man mit V das Volumen bezeichnet, in dem die N_0 Partikel suspendiert sind und mit dem sie durch das Messvolumen hindurchgeströmt sind. Aus 9.21 und 9.22 folgt

$$N = \frac{V}{\Delta V}\Big(\lambda\Delta V \exp\left(-\lambda\Delta V\right)\Big) \tag{9.23}$$

Das Produkt $\lambda\Delta V \exp\left(-\lambda\Delta V\right)$ hat bei $\lambda\Delta V = 1$ ein Maximum und verschwindet sowohl für $\lambda\Delta V \to 0$ als auch für $\lambda\Delta V \to \infty$. Der Zusammenhang zwischen N und $\lambda\Delta V$ ist nicht umkehrbar eindeutig (eineindeutig). Eine kleine Anzahl von Streulichtsignalen kann sowohl mit einer sehr niedrigen als auch mit einer sehr hohen Partikelkonzentration erklärt werden. Im letzteren Fall bedeutet dies, dass die Signalüberlagerungen derart häufig werden, dass das Messsignal nur noch selten abreißt und nur noch selten Partikel registriert werden. Um zwischen den beiden Lösungen von 9.23 entscheiden zu können, bedarf es zusätzlicher Informationen, das heißt, man muss wissen, ob man sich im Bereich $\lambda\Delta V < 1$ oder im Bereich $\lambda\Delta V > 1$ bewegt.

9.2.3 Folgerungen

Der durch Koinzidenzen verursachte Fehler bei der Bestimmung der Partikelkonzentration kann, wie gezeigt wurde, vollständig korrigiert werden. Anders verhält es sich bei den gemessenen Partikelgrößenverteilungen. Hier sind Korrekturen nur mit Einschränkungen möglich.

Durch die Überlagerung der Streulichtimpulse mehrerer Partikeln kommen Messsignale zustande, die denen größerer Partikeln entsprechen. Deshalb ist die mittels eines Streulicht-Partikelgrößen-Analysators gemessene Partikelgrößenverteilung in jedem Fall gegenüber der wahren Verteilung zum Groben hin verschoben. Prinzipiell lässt sich diese Verschiebung beliebig genau vorausberechnen, wenn die wahre Partikelgrößenverteilung bekannt ist. Solche Rechnungen werden allerdings sehr kompliziert, wenn man versucht, dabei alle Formen der Signalüberlagerung einzubeziehen. Werden nur Einer- und Zweiersignale berücksichtigt, ist man auf den Bereich $\lambda \Delta V < 0,3$ beschränkt.

In der Praxis der Partikelmesstechnik ist die Situation jedoch anders als hier geschildert. Was vorliegt, ist die gemessene, durch Koinzidenzen verfälschte Verteilung, und es ist die Aufgabe gestellt, daraus die unbekannte wahre Verteilung so genau wie möglich zu errechnen, das heißt, den Koinzidenzfehler nach Möglichkeit zu korrigieren. Man kann zeigen, dass dies auf die Lösung einer nichtlinearen Integralgleichung hinausläuft.

9.3 Statistische Fehler bei der Partikelgrößenanalyse[3]

9.3.1 Aufgabenstellung

Um die Partikelgrößenverteilung eines Haufwerks (dispersen Feststoffs) zu bestimmen, zieht man eine Stichprobe und untersucht diese mit einem geeigneten Messverfahren. Das Ergebnis ist in jedem Fall eine Häufigkeitsverteilung, die von der eigentlich gesuchten Verteilung der Grundgesamtheit wegen des Probenahmefehlers mehr oder weniger stark abweicht. Um hierzu quantitative Aussagen treffen zu können, wird von folgenden Voraussetzungen ausgegangen:

- Das Messverfahren ist ein Zählverfahren, das Ergebnis der Analyse ist damit primär eine Anzahlverteilung.[4]
- Der Wertebereich der Partikelgröße x wird für die Analyse in n Klassen der einheitlichen Breite Δx eingeteilt.
- Die Partikelgröße x wird einzeln für jede Partikel der Stichprobe bestimmt.

[3] J. Raasch Seminarvortrag im Institut für Mechanische Verfahrenstechnik und Mechanik der Universität Karlsruhe (TH) am 14. Jan. 2008

[4] Für nicht-zählende Messverfahren lassen sich ähnliche Überlegungen anstellen, nur wächst der Rechenaufwand.

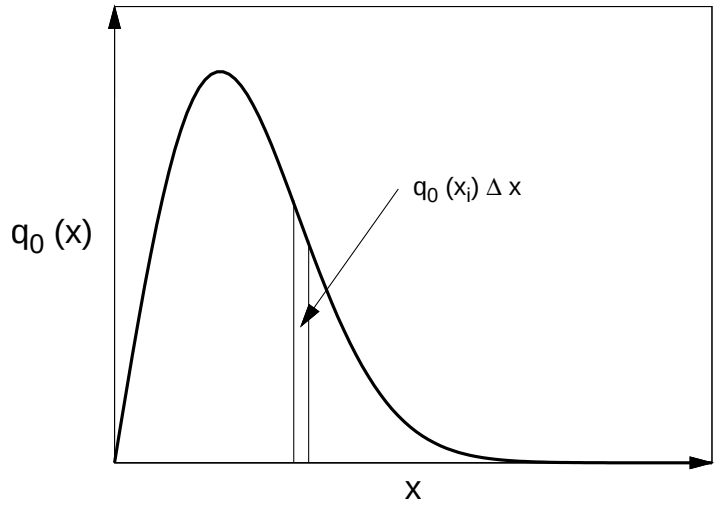


Abb. 9.1: Partikelgrößenverteilung, dargestellt als Dichtefunktion der Anzahlverteilung $q_0(x)$

- Für jede Größenklasse wird die Anzahl $z(x_i)$ aller der Partikeln registriert, deren Größe in diese Klasse fällt (Besetzungszahlen), wobei die exakte Definition für $z(x_i)$ folgendermaßen lautet

$$z(x_i) = z\Big((x_i - \frac{\Delta x}{2}) < x < (x_i + \frac{\Delta x}{2})\Big) \tag{9.24}$$

  Alle Werte x einer Klasse werden also der Klassenmitte x_i zugeordnet.

- Mit den Partikelzahlen $z(x_i)$ (Besetzungszahlen) wird für jede Klasse die Häufigkeitsdichte $f(x_i)$ berechnet, wobei gilt (2.24 auf Seite 18)

$$f(x_i) = \frac{z(x_i)}{z\ \Delta x} \tag{9.25}$$

  Die Häufigkeitsdichte $f(x)$ ist eine Schätzfunktion für die unbekannte Anzahldichte $q_0(x)$ der Grundgesamtheit. Eine solche Anzahlverteilung ist in Abbildung 9.1 dargestellt. Eingezeichnet ist nur eine der n Größenklassen.

Wiederholt man die Stichprobenahme, so erhält man für die Partikelzahlen $z(x_i)$ andere Werte. Die Partikelzahl $z(x_i)$ ist daher in jeder Größenklasse eine Zufallsvariable. Dieser Probenahmefehler ist eine Folge des endlichen Umfangs der Stichprobe und hat nichts mit Messfehlern zu tun. Messfehler führen zu weiteren Abweichungen der gemessenen Häufigkeitsdichte $f(x_i)$ von der Verteilung der Grundgesamtheit, sollen jedoch im Folgenden außer Betracht bleiben.

9.3.2 Mathematischer Ansatz

Die Entnahme einer Stichprobe aus einer Grundgesamtheit ist ein Vorgang, der alle Kriterien eines Zufallsexperimentes erfüllen muss. Andernfall wäre das Ergebnis keine Stichprobe, sondern nur eine Teilmenge ohne Wert für die Statistik. Die

gleichzeitige Entnahme von z Partikeln ist der z-maligen Entnahme von jeweils einer Partikel ohne Zurücklegen gleichwertig. Man kann deshalb die Stichprobenentnahme als z-malige Ausführung eines Zufallsexperimentes ansehen, bei dem jeweils nur eine einzige Partikel gezogen wird. Die Wahrscheinlichkeit, dass diese eine Partikel in die i-te Merkmalsklasse fällt, wird weiterhin mit p bezeichnet. Es gilt, wie man sich an Hand der Abbildung 9.1 klar machen kann,

$$p = q_0(x_i)\Delta x \tag{9.26}$$

Umgekehrt wird die Partikel mit der Wahrscheinlichkeit

$$q = 1 - q_0(x_i)\Delta x \tag{9.27}$$

in irgendeine andere Merkmalsklasse fallen. Das Zufallsexperiment wird so lange wiederholt, bis alle Partikeln der Stichprobe analysiert und in die jeweiligen Merkmalsklassen einsortiert sind.

Die Gesamtzahl z der Partikeln in der Stichprobe ist gleich der Summe über die in den einzelnen Merkmalsklassen registrierten Partikelzahlen $z(x_i)$

$$z = \sum_i z(x_i) \tag{9.28}$$

Die Partikelzahl $z(x_i)$ in der i-ten Klasse ist, wie bereits erklärt, eine Zufallsvariable. Gesucht wird ihre Verteilungsfunktion.

Die weiteren Überlegungen beschränken sich auf die i-te Merkmalsklasse. Wie die Verteilung der Variablen x im Übrigen aussieht, ist dabei gleichgültig. Um dies zu verdeutlichen und um die Schreibweise zu vereinfachen, wird die Zufallsvariable $z(x_i)$ umbenannt in y

$$y \equiv z(x_i) \tag{9.29}$$

Eine Binomialverteilung (siehe Abschnitt 4.2 *Binomialverteilung* auf Seite 63) ergibt sich immer dann, wenn man danach fragt, wie oft ein bestimmtes Ereignis bei z-maliger Ausführung eines Zufallsexperimentes eintrifft, wobei das Ereignis bei jeder Ausführung dieselbe Wahrscheinlichkeit p besitzt und die Ergebnisse der verschiedenen Ausführungen sich nicht gegenseitig beeinflussen. Genau diese Situation ist bei dem beschriebenen Vorgang der Probenahme gegeben. Die gesuchte Verteilungsfunktion von y lautet daher

$$P(y) = \binom{z}{y} p^y q^{z-y} \tag{9.30}$$

mit dem Mittelwert $E(y)$ und der Varianz $V(y)$

$$E(y) = zp \qquad V(y) = zpq \tag{9.31}$$

Die Wahrscheinlichkeit, dass y einen Wert annimmt, der im Bereich zwischen $y = a$ und $y = b$ liegt, errechnet sich mit 9.30 zu

$$P(a \le y \le b) = \sum_a^b \binom{z}{y} p^y q^{z-y} \tag{9.32}$$

Setzt man $z(x_i) \gg 1$ und $z \gg 1$ voraus, kann man die Binomialverteilung mit hoher Genauigkeit durch die Normalverteilung annähern und erhält

$$P(a \leq y \leq b) \approx \frac{1}{\sqrt{2\pi zpq}} \int_a^b \exp\Big(-\frac{1}{2}(\frac{y-zp}{\sqrt{zpq}})^2\Big)\,\mathrm{d}y \tag{9.33}$$

Mit der Substitution

$$t = \frac{y-zp}{\sqrt{zpq}} \tag{9.34}$$

folgt daraus

$$P(a \leq y \leq b) \approx \frac{1}{\sqrt{2\pi}} \int_\alpha^\beta \exp\,(-\frac{1}{2}t^2)\,\mathrm{d}t = H(\beta) - H(\alpha) \tag{9.35}$$

worin $H(t)$ die Summenfunktion der normalisierten Normalverteilung bezeichnet, deren Werte tabelliert sind. Für die Bereichsgrenzen α und β ist einzusetzen

$$\alpha = \frac{a-zp}{\sqrt{zpq}} \qquad \beta = \frac{b-zp}{\sqrt{zpq}} \tag{9.36}$$

Symmetrische Bereichsgrenzen erhält man, wenn man setzt

$$a = (p - \delta\Delta x)z \qquad b = (p + \delta\Delta x)z \tag{9.37}$$

Für α und β folgt damit

$$\alpha = -\frac{\delta z\Delta x}{\sqrt{zpq}} \qquad \beta = +\frac{\delta z\Delta x}{\sqrt{zpq}} \tag{9.38}$$

Damit ergibt sich für den Wert der Summenfunktion $H(\alpha)$

$$H(\alpha) = H(-\beta) = 1 - H(\beta) \tag{9.39}$$

Setzt man dies in 9.35 ein, so erhält man

$$P(a \leq y \leq b) \approx 2H(\beta) - 1 \tag{9.40}$$

Die linke Seite dieser Näherungsgleichung kann man mit 9.37, 9.29, 9.25 und 9.26 umformen und erhält

$$P\Big(|f(x_i) - q_0(x_i)| \leq \delta\Big) \approx 2H(\beta) - 1 \tag{9.41}$$

Hierin bezeichnet δ die Abweichung der aus der Stichprobe ermittelten Häufigkeitsdichte $f(x_i)$ von der unbekannten Anzahldichtefunktion $q_0(x_i)$ der Grundgesamtheit. δ hat die gleiche Dimension wie $f(x_i)$ und $q_0(x_i)$, nämlich die der reziproken Partikelgröße.

Der Zusammenhang zwischen δ und β ist durch 9.38 gegeben. Wählt man die Klassenbreite Δx hinreichend klein, kann man $p \ll 1$ und $q \approx 1$ voraussetzen. Damit folgt aus 9.38

$$\delta = \beta\sqrt{\frac{q_0(x_i)}{z\Delta x}} \tag{9.42}$$

Solange die Abweichungen δ klein sind, kann man $q_0(x_i)$ in 9.42 durch $f(x_i)$ ersetzen und erhält als gute Näherung

$$\delta \approx \beta \sqrt{z(x_i)} \tag{9.43}$$

Hiermit lässt sich 9.41 schließlich in die folgende einfache Form bringen

$$P\Big(f(x_i)\Big(1 - \frac{\beta}{\sqrt{z(x_i)}}\Big) \le q_0(x_i) \le f(x_i)\Big(1 + \frac{\beta}{\sqrt{z(x_i)}}\Big)\Big) \approx 2H(\beta) - 1 \tag{9.44}$$

9.3.3 Folgerungen

Gleichung 9.44 sagt aus, dass mit der Wahrscheinlichkeit $P = 2H(\beta) - 1$ der unbekannte Wert von $q_0(x_i)$ zwischen festen Grenzen eingeschlossen werden kann, die vom jeweiligen Stichprobenergebnis abhängen. Die Aussagewahrscheinlichkeit kann unterschiedlich gewählt werden; sie hängt allein von β ab. Ein üblicher Wert ist $\beta = 1,960$. Hierzu liest man aus einer Tabelle der Standardnormalverteilung den Wert $H(\beta) = 0.9750$ ab. Dies führt zu der Aussagewahrscheinlichkeit $P = 0,95$ oder 95 %. Die zu einer bestimmten Aussagewahrscheinlichkeit gehörige maximale relative Abweichung des Schätzwertes $f(x_i)$ vom wahren Wert $q_0(x_i)$ der Anzahldichteverteilung der Grundgesamtheit ist

$$\Big|\frac{f(x_i) - q_0(z_i)}{q_0(x_i)}\Big| = \frac{\beta}{\sqrt{z(x_i)}} \tag{9.45}$$

Sie hängt von der Partikelzahl $z(x_i)$ in der jeweiligen Größenklasse ab. An den Rändern der Partikelgrößenverteilung, wo die Partikelzahlen sehr klein werden, sind die maximalen relativen Abweichungen deshalb am größten.

Ist man nur an der Anzahlverteilung $q_0(x)$ interessiert, kann man an den Rändern der Verteilung die großen relativen Abweichungen hinnehmen, da dort $q_0(x)$ gegen Null strebt und damit die absoluten Abweichungen sehr klein werden.

Anders verhält es sich, wenn aus der Anzahlverteilung $q_0(x)$ die zugehörige Volumen- oder Massenverteilung $q_3(x)$ berechnet werden soll. Zwischen beiden besteht die Beziehung

$$q_3(x_i) = \frac{x_i^3\, q_0(x_i)}{M_{3,0}} \quad \text{mit dem Moment} \quad M_{3,0} = \sum_i x_i^3 q_0(x_i) \tag{9.46}$$

Man erhält $q_3(x_i)$, indem man $q_0(x_i)$ mit dem Faktor $x_i^3/M_{3,0}$ multipliziert. Dieser wird am rechten Rand der Verteilung (im Groben) sehr groß, weil x zur dritten Potenz eingeht. Folglich werden dort auch die Werte von $q_3(x_i)$ sehr viel größer als die Werte von $q_0(x_i)$ sein, solange $q_0(x_i) > 0$. Die großen relativen Abweichungen von $q_0(x_i)$ am rechten Rand der Verteilung haben gleichgroße relative Abweichungen von $q_3(x_i)$ zur Folge, die jedoch bei der Volumen- oder Massenverteilung mit sehr viel größeren absoluten Abweichungen verbunden sind.

Der Versuch, die Volumen- oder Massendichteverteilung $q_3(x)$ aus der gemessenen Häufigkeitsdichteverteilung $f(x)$ zu berechnen, führt aus diesem Grunde

meist zu wenig befriedigenden Ergebnissen. Die Aussage gilt jedoch nicht für die zugehörigen Summenverteilungen.

Karl Sommer hat die statistischen Schwankungen der Summenverteilung $Q_0(x)$ untersucht.[5] Eine Klasseneinteilung war hierfür nicht nötig. Es wurden nur Partikeln mit $x < x_i$ und $x > x_i$ unterschieden. Für die Varianz der aus der Stichprobe berechneten Häufigkeitssumme $F(x_i)$ erhielt er den Ausdruck

$$V\Big(F(x_i)\Big) = \frac{Q_0(x_i)(1 - Q_0(x_i))}{z} \tag{9.47}$$

Analog zu 9.44 kann man hiermit folgende Aussage über die Wahrscheinlichkeit der Abweichungen zwischen $F(x_i)$ und $Q_0(x_i)$ formulieren

$$P\Big(F(x_i)(1 - \beta \sqrt{\frac{1 - F(x_i)}{zF(x_i)}}) \le Q_0(x_i) \le F(x_i)(1 + \beta \sqrt{\frac{1 - F(x_i)}{zF(x_i)}})\Big)$$

$$\approx 2H(\beta) - 1 \tag{9.48}$$

Wie man aus dem Vergleich mit 9.44 erkennt, werden bei der Summenfunktion die maximalen relativen Abweichungen sehr viel kleiner als bei der Dichtefunktion, da im ersten Fall z im Nenner des Wurzelausdrucks steht und im zweiten Fall $z(x_i)$.

Die geringeren statistischen Schwankungen der Summenfunktion $F(x)$ werden verständlich, wenn man sich vorstellt, dass bei der Berechnung von $F(x)$, das heißt bei der Summierung über eine große Zahl von Werten $f(x_i)\Delta x$, sich deren statistische Schwankungen weitgehend ausgleichen.

9.4 Porosität von Schüttungen

9.4.1 Aufgabenstellung

Die Porosität (Quotient aus Porenvolumen und Gesamtvolumen) von Schüttungen kann man mit einer radiometrischen Methode bestimmen, die auf der Anwendung des folgenden Absorptionsgesetzes beruht

$$\frac{N'}{N_0'} = \exp\Big(-\mu(1-\epsilon)L\Big) = \exp\Big(-a(1-\epsilon)\Big) \tag{9.49}$$

Darin bedeuten

N_0'	Anzahl der vor der Schüttung in der Zeit τ eingestrahlten Quanten
N'	Anzahl der Quanten, die in der Zeit τ die Schüttung ohne Wechselwirkung durchdringen
ϵ	Porosität der Schüttung
μ	linearer Absorptionskoeffizient des Feststoffs
L	Dicke der Schüttung
a	$a = \mu L$ (zwecks einfacherer Schreibweise)

[5] K. Sommer, Probenahme von Pulvern und körnigen Massengütern, Springer, Berlin, Heidelberg 1979, S. 192

Vorausgesetzt wird, dass der Feststoff chemisch homogen ist und dass die Strahlungsquelle (radioaktives Isotop) annähernd monoenergetische Gammaquanten aussendet. Ist der lineare Absorptionskoeffizient bekannt, so kann die Porosität der Schüttung über Impulszählungen bestimmt werden. Eine für diesen Zweck entwickelte Messanordnung liefert den Quantenanzahlen N_0' und N' proportionale Impulsanzahlen N_0 und N. Es gilt

$$N_0 = \eta \cdot N_0' \qquad N = \eta \cdot N' \tag{9.50}$$

worin η den Ausbeutefaktor der Messanordnung bezeichnet. Mit 9.50 folgt aus 9.49

$$\frac{N}{N_0} = \exp\Big(- a(1-\epsilon)\Big) \tag{9.51}$$

Löst man nach der gesuchten Porosität ϵ durch Logarithmieren auf, so findet man

$$\epsilon = 1 + \frac{1}{a} \ln\left(\frac{N}{N_0}\right) \tag{9.52}$$

Es sollen die Messzeit τ und damit die Impulsanzahlen N_0 und N so festgelegt werden, dass die mit 9.52 berechnete Porosität mit einer Wahrscheinlichkeit von 0,95 um nicht mehr als ein vorgegebenes $\Delta\epsilon$ nach oben oder unten vom wahren Wert μ_ϵ abweicht Die Schwankungen von ϵ ergeben sich daraus, dass N und N_0 Zufallsvariable sind und ϵ wegen 9.52 eine Funktion dieser beiden Zufallsvariablen ist. Da sie in getrennten Messungen nacheinander bestimmt werden, sind N und N_0 voneinander unabhängig.

9.4.2 Mathematischer Ansatz

Die Verteilungsfunktion von ϵ ist nicht bekannt. Wir nehmen an, dass sie durch eine Normalverteilung mit dem Mittelwert μ_ϵ und der Varianz σ_ϵ^2 hinreichend genau angenähert werden kann. Dann gilt (siehe Abschnitt 6.2 *Konfidenzintervalle für den Mittelwert einer Normalverteilung mit bekannter Varianz* auf Seite 108)

$$P(\mu_\epsilon - \Delta\epsilon \le \epsilon \le \mu_\epsilon + \Delta\epsilon) = P(-c \le \frac{\epsilon - \mu_\epsilon}{\sigma_\epsilon} \le +c) = \gamma \tag{9.53}$$

wobei für c einzusetzen ist

$$c = \frac{\Delta\epsilon}{\sigma_\epsilon} \tag{9.54}$$

Zu der Wahrscheinlichkeit $\gamma = 0,95$ liest man aus der Tabelle der normierten Normalverteilung den Wert $c = 1,960$ ab.

Die in der Aufgabenstellung formulierte Forderung wird erfüllt, wenn

$$\sigma_\epsilon \le \frac{\Delta\epsilon}{1,960} \tag{9.55}$$

Für das Weitere werden Mittelwert μ_ϵ und Varianz σ_ϵ^2 der Verteilung von ϵ benötigt.

Wie in Abschnitt 3.4.3 *Mehrdimensionale Zufallsgrößen* auf Seite 58 gezeigt wurde, gilt für eine Zufallsvariable v, die über eine allgemeine Beziehung

$$v = g(x, y) \tag{9.56}$$

von zwei voneinander unabhängigen Variablen x und y abhängt, dass Mittelwert und Varianz näherungsweise berechnet werden können zu

$$\mu_v \approx g(\mu_x,\, \mu_y) + \frac{1}{2}\left(g_{xx}(\mu_x,\, \mu_y) \cdot \sigma_x^2 + g_{yy}(\mu_x,\, \mu_y) \cdot \sigma_y^2\right) \tag{9.57}$$

$$\sigma_v^2 \approx \Big(g_x(\mu_x,\, \mu_y)\Big)^2 \cdot \sigma_x^2 + \Big(g_y(\mu_x,\, \mu_y)\Big)^2 \cdot \sigma_y^2 \tag{9.58}$$

Um diese allgemeinen Beziehungen auf den speziellen Fall von 9.52 anzuwenden, benötigt man zunächst einmal die partiellen Ableitungen von ϵ nach den beiden Variablen N und N_0. Aus 9.52 ergibt sich

$$\frac{\partial \epsilon}{\partial N} = \frac{1}{aN} \qquad \frac{\partial^2 \epsilon}{\partial N^2} = -\frac{1}{aN^2} \qquad \frac{\partial \epsilon}{\partial N_0} = -\frac{1}{aN_0} \qquad \frac{\partial^2 \epsilon}{\partial N_0^2} = \frac{1}{aN_0^2}$$

Damit folgt aus 9.57 für den Mittelwert μ_ϵ

$$\mu_\epsilon \approx 1 + \frac{1}{a} \ln\left(\frac{\mu_N}{\mu_{N_0}}\right) - \frac{1}{2a}\left(\frac{1}{\mu_N} - \frac{1}{\mu_{N_0}}\right)$$

Der zweite Term kann in allen praktischen Fällen gegenüber dem ersten Term vernachlässigt werden, weshalb im Weiteren mit

$$\mu_\epsilon \approx 1 + \frac{1}{a} \ln\left(\frac{\mu_N}{\mu_{N_0}}\right) \tag{9.59}$$

gerechnet wird. Aus 9.58 erhält man für die Varianz σ_ϵ^2

$$\sigma_\epsilon^2 \approx \frac{1}{a^2}\left(\frac{1}{\mu_N} + \frac{1}{\mu_{N_0}}\right) \tag{9.60}$$

Mit 9.59 folgt daraus weiter

$$\sigma_\epsilon^2 \approx \frac{1}{a^2}\frac{1}{\mu_{N_0}}\Big(1 + \exp\Big(a(1 - \mu_\epsilon)\Big)\Big) \tag{9.61}$$

Setzt man 9.61 in 9.55 ein, so ergibt sich

$$\mu_{N_0} \geq \left(\frac{1,96}{a\Delta\epsilon}\right)^2 \Big(1 + \exp\Big(a(1 - \mu_\epsilon)\Big)\Big) \tag{9.62}$$

9.4.3 Folgerungen

Mit 9.62 kann man den Mittelwert μ_{N_0} der gemessenen Impulszahlen N_0 berechnen, wenn μ_ϵ, μ und L bekannt sind und für das Konfidenzintervall von ϵ ein Wert $\Delta\epsilon$ vorgegeben wird.

Bei Untersuchungen an einer Schüttung von gemahlenem Kalkstein in einer Versuchsapparatur ergaben sich für μ_ϵ, μ und L folgende Werte:

- $\mu_\epsilon = 0,45$
- $\mu = 0,2085$ cm^{-1} für Kalkstein und eine ^{137}Cs-Quelle
- $L = 6$ cm, daraus $a = \mu L = 1,251$

Damit erhält man aus 9.62

für	$\Delta\epsilon$	=	0.0025	den Wert	$\mu_{N_0} = 1,174 \cdot 10^6$
für	$\Delta\epsilon$	=	0,0050	den Wert	$\mu_{N_0} = 2,936 \cdot 10^5$
für	$\Delta\epsilon$	=	0,0100	den Wert	$\mu_{N_0} = 7,339 \cdot 10^4$

Mit diesen Werten kann man bei bekannter Intensität μ_{N_0}/τ der Strahlenquelle die erforderliche Messzeit τ vorausberechnen beziehungsweise prüfen, ob mit der gewählten Messanordnung die Porosität ϵ bei vernünftigem Zeitaufwand hinreichend genau bestimmt werden kann.

Wie man aus 9.62 ablesen kann, hängt μ_{N_0} auf komplizierte Weise von a und damit von der Schichtdicke L ab. Sowohl für sehr große als auch für sehr kleine Schichtdicken nimmt μ_{N_0} sehr große Werte an. Dazwischen liegt ein Minimum, was bei der Dimensionierung einer Messeinrichtung beachtet werden sollte. Die in der Beispielrechnung angenommene Schichtdicke $L = 6$ cm führt zu sehr hohen Zahlenwerten für μ_{N_0}. Günstiger sind größere Schichtdicken im Bereich von $L = 20$ cm, wenn $\mu = 0.2085$ cm^{-1} für den linearen Absorptionskoeffizienten des Feststoffs vorausgesetzt wird.

9.5 Verteilung der Abstände zwischen optimal dispergierten Partikeln in einer Fluidströmung[6]

9.5.1 Aufgabenstellung

Eine genaue Kenntnis der räumlichen Verteilung von dispergierten Partikeln (Feststoffteilchen oder Flüssigkeitströpfchen) in strömenden Flüssigkeiten oder

[6] J. Raasch und H. Umhauer, Computation of the Frequency Distributions of Distances Between Particles Randomly Dispersed in a Fluid Flow, Part. Part. Syst. Charact. 6 (1989) 13–16

Gasen ist für zahlreiche verfahrenstechnische Prozesse (chemische Reaktionen, Verbrennung, Trennvorgänge) von erheblichem Interesse. Gefordert wird in solchen Fällen eine möglichst homogene räumliche Verteilung der Partikeln. Diese wird erreicht, wenn die Partikeln rein zufällig im strömenden Medium verteilt sind. Eine reguläre Anordnung wie in einem Kristallgitter wäre homogener, ist aber im Allgemeinen nicht erreichbar.

Bei zufälliger Anordnung variieren die Abstände zwischen benachbarten Partikeln in starkem Maße. Gesucht sei die Verteilungsfunktion des Abstands zweier Partikeln unter der einschränkenden Voraussetzung, dass die Volumenkonzentration der Partikeln insgesamt so gering ist, dass wechselseitige Verdrängungseffekte unberücksichtigt bleiben können. Die dispergierten Partikeln werden also wie mathematische Punkte behandelt, das heißt, es wird von ihrer räumlichen Ausdehnung abgesehen.

9.5.2 Mathematischer Ansatz

Die räumliche Verteilung der Punkte wird bei rein zufälliger Anordnung durch die Poisson-Verteilung beschrieben, siehe Abschnitt 4.3 *Poisson-Verteilung* auf Seite 68. Die Wahrscheinlichkeit, dass ein Volumen V genau x Partikeln enthält, ergibt sich damit zu

$$P(x;V) = \frac{(\lambda V)^x}{x!}\ \exp\left(-\lambda V\right) \tag{9.63}$$

worin λ für die Anzahlkonzentration der Partikeln steht. Weiterhin wird mit s_{12} der Abstand zwischen einer ersten Partikel und der nächsten benachbarten Partikel bezeichnet. In Abbildung 9.2 ist die Situation dargestellt. Im Mittelpunkt einer Kugel vom Radius s_{12} befinde sich Partikel 1. Das Volumen dieser Kugel ist

$$V = \frac{4\pi}{3} s_{12}^3 \tag{9.64}$$

Man denkt sich diese Kugel von einer zweiten Kugel vom Radius $s_{12} + \mathrm{d}s_{12}$ umschlossen. Das zwischen den beiden Kugelflächen eingeschlossene infinitesimal kleine Volumen $\mathrm{d}V$ ergibt sich zu

$$\mathrm{d}V = 4\pi s_{12}^2 \mathrm{d}s_{12} \tag{9.65}$$

Die Wahrscheinlichkeit, dass der Abstand s_{12} zwischen der Partikel 1 und der nächsten benachbarten Partikel 2 in dem Intervall zwischen s_{12} und $s_{12} + \mathrm{d}s_{12}$ liegt, ist offenbar gleich der Wahrscheinlichkeit, dass sich im Volumen V keine Partikel befindet und in dem Volumen $\mathrm{d}V$ genau eine Partikel

$$\varphi(s_{12})\mathrm{d}s_{12} = P(0;V)\ P(1,\mathrm{d}V) \tag{9.66}$$

Aus 9.63 erhält man

$$P(0;V) = \exp\left(-\lambda V\right) \tag{9.67}$$

$$P(1;\mathrm{d}V) = \lambda \mathrm{d}V \tag{9.68}$$

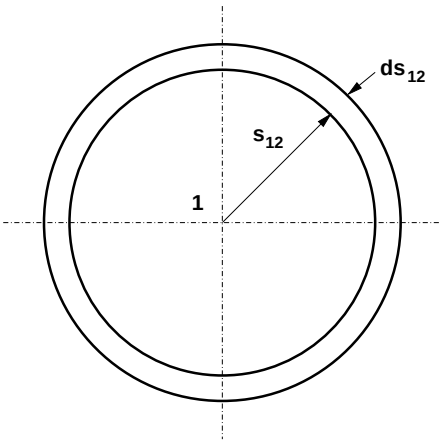


Abb. 9.2: Situation bei der Berechnung des Partikelabstands

Damit folgt aus 9.66

$$\varphi(s_{12}) = 4\pi\lambda s_{12}^2 \exp\left(-\frac{4\pi}{3}\lambda s_{12}^3\right) \tag{9.69}$$

Dies ist die Dichtefunktion der Verteilung der Abstände zur jeweils nächsten Partikel, eine Weibull-Verteilung gemäß 5.79 von Seite 98 mit

$$b = 3 \qquad \text{und} \qquad T^{\,b} = \frac{3}{4\pi\lambda}$$

Mittelwert und Varianz von s_{12} errechnen sich zu

$$E(s_{12}) = \int_0^\infty s_{12}\ \varphi(s_{12})\ \mathrm{d}s_{12} = 0,5540 \cdot \lambda^{-1/3} \tag{9.70}$$

$$V(s_{12}) = E(s_{12}^2) - (E(s_{12}))^2 = 0,0405 \cdot \lambda^{-2/3} \tag{9.71}$$

Da die Varianz von s_{12} mit zunehmender Anzahlkonzentration λ kleiner wird, wird die räumliche Partikelverteilung mit zunehmender Konzentration offenbar homogener.

Die Verteilung der Abstände zur zweitnächsten Partikel erhält man auf ähnliche Weise mit folgendem Ansatz

$$\varphi(s_{13})\mathrm{d}s_{13} = P(1;V)\ P(1;\mathrm{d}V) \tag{9.72}$$

oder allgemein für den Abstand zur n-nächsten Partikel

$$\varphi(s_{1n})\mathrm{d}s_{1n} = P(n-2;V)\ P(1;\mathrm{d}V) \tag{9.73}$$

wobei n beliebige Werte $n \geq 2$ annehmen kann. Im Allgemeinen wird man sich nur für die Abstände s_{12} zur jeweils nächsten benachbarten Partikel interessieren.

Bei der Auswertung von Impulshologrammen, die zur Untersuchung einer partikelbeladenen turbulenten Rohrströmung aufgenommen wurden, zeigte sich, dass mehr als die Hälfte aller Partikeln Paare bildeten in dem Sinne, dass jeweils zwei Partikeln wechselseitig füreinander die nächsten Nachbarn waren. Es stellte sich

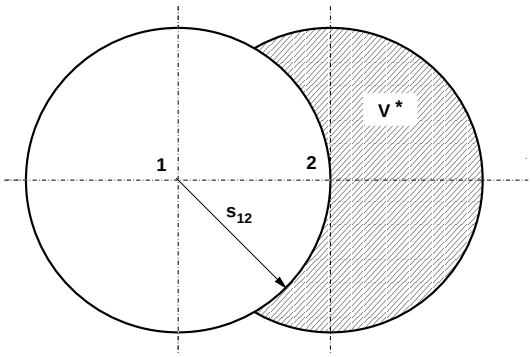


Abb. 9.3: Situation bei der Berechnung des Partikelabstands bei Partikelpaaren

die Frage, ob die beobachtete Paarbildung mit der Annahme einer rein zufälligen räumlichen Verteilung der Partikeln in Einklang stand oder ob sie auf zusätzliche Effekte – beispielsweise elektrostatische Aufladung der Partikeln – zurückzuführen war.

Auch diese Frage ließ sich mit einem ähnlichen mathematischen Ansatz beantworten. Abbildung 9.3 zeigt zwei benachbarte Partikeln im Abstand s_{12}. Damit die beiden Partikeln wechselseitig füreinander die nächsten Nachbarn sind, muss nicht nur das Volumen V um die Partikel 1 wie zuvor frei von weiteren Partikeln sein, sondern zusätzlich auch das Volumen $V^\star$ (in Abbildung 9.3 schaffiert). Dieses erhält man, wenn man vom Volumen V zweimal das Volumen einer Kugelkalotte abzieht, zu

$$V^\star = \frac{11\pi}{12} s_{12}^3 \tag{9.74}$$

Die gesuchte Wahrscheinlichkeit dafür, dass zwei Partikeln wechselseitig nächste Nachbarn füreinander sind (Paarbildung), errechnet sich aus

$$P(s_{12} = s_{21}) = \int_0^\infty P(0; V^\star)\, \varphi(s_{12})\, \mathrm{d}s_{12} \tag{9.75}$$

mit 9.74 und 9.69 zu

$$P(s_{12} = s_{21}) = \frac{16}{27} = 0,5926 \tag{9.76}$$

9.5.3 Folgerungen

Die an einer partikelbeladenen turbulenten Rohrströmung durch Auswertung von Impulshologrammen festgestellte Paarbildung bei mehr als der Hälfte aller Partikeln ist offenbar nicht auf irgendwelche Wechselwirkungskräfte zwischen den Partikeln zurückzuführen, sondern allein aus der rein zufälligen räumlichen Verteilung der Partikeln zu erklären.

Die aus dem mathematischen Modell der rein zufälligen räumlichen Verteilung der Partikeln hergeleiteten Aussagen über die Verteilungen der Partikelabstände und über die Wahrscheinlichkeit der Paarbildung konnten in einem speziellen Fall experimentell bestätigt werden. Sie gelten nur für den Zustand optimaler

Dispergierung der Partikeln, bieten sich jedoch zugleich als Vergleichsmaßstab für die Beschreibung realer Dispersionen an.

9.6 Mischgütemaß für Feststoffmischungen

9.6.1 Aufgabenstellung

Mischprozesse sind praktisch in allen Industriezweigen anzutrefffen. Mischungskomponenten können Gase, Flüssigkeiten oder feinkörnige Feststoffe (Pulver, Schüttgüter) sein. Hier sollen nur Feststoffmischungen in Betracht gezogen werden, wobei zunächst die grundsätzliche Frage zu beantworten ist, wie der jeweils erreichte Mischzustand beschrieben und beurteilt werden kann.

Die folgenden Überlegungen beschränken sich auf Mischungen aus zwei Komponenten. Dies ist keine schwerwiegende Einschränkung, weil sich viele Mehrkomponentenmischungen als Zweikomponentenmischungen behandeln lassen, nämlich als Mischung desjenigen Stoffes, auf den es aus irgendwelchen Gründen im konkreten Fall besonders ankommt (1. Komponente), und allen übrigen in der Mischung vorhandenen Stoffen (2. Komponente). Ein typisches Beispiel ist eine Mischung pharmazeutischer Substanzen in Pulverform als Ausgangsmaterial für Tabletten, die eine wirksame Substanz sowie einen Rest aus Füll-, Fließ-, Binde-, Sprengund Schmiermitteln enthält.

9.6.2 Mathematischer Ansatz

Im Falle einer Zweikomponentenmischung genügt eine einzige Zahlenangabe zur Kennzeichnung der Zusammensetzung der Mischung, nämlich die Volumenkonzentration p der 1. Komponente. Will man den Mischungszustand beschreiben, so muss man Proben (kleine Teilmengen) ziehen und darin die jeweilige Volumenkonzentration x der 1. Komponente mit einem geeigneten Analyseverfahren bestimmen. Die Konzentrationswerte x_i in den Proben werden im Allgemeinen von der Mischungskonzentration p abweichen. Je größer bei gleichem Probenumfang diese Abweichungen sind, desto schlechter ist die Mischung; je kleiner, desto besser. Je enger die Verteilung der Probenkonzentration x ist, desto kleiner ist ihre Varianz. Es bietet sich daher an, die Varianz σ_x^2 als Mischgütemaß zu verwenden.

Um mit einem bestimmten Wert der Varianz σ_x^2 ein Urteil über die Mischungsqualität verbinden zu können, sind jedoch zunächst zwei Fragen zu klären

- Welches ist der homogenste, theoretisch erreichbare Mischungszustand?
- Wie hängt die Varianz σ_x^2 vom Probenumfang ab?

Ein geordneter Mischungszustand würde dazu führen, dass die Konzentration x der 1. Komponente in jeder Probe exakt gleich p wäre und damit die Varianz σ_x^2 den Wert null hätte. Dies ist jedoch kein Mischungszustand, der mit technischen Mitteln erreicht werden kann. Der bestmögliche, unter idealen Bedingungen erreichbare Mischungszustand ist der einer **gleichmäßigen Zufallsmischung**.

Die Varianz der Konzentration von Proben aus einer Zufallsmischung lässt sich unter stark vereinfachenden Annahmen exakt berechnen. Hierzu wird im Folgenden vorausgesetzt, dass alle Partikeln gleich groß sind und bei der Probenahme jeweils genau n Partikeln aus der Mischung gezogen werden. Hierzu seien n_x Partikeln der 1. Komponente und n_y Partikeln der 2. Komponente zuzuordnen. Dann gilt in jedem Fall

$$n_z + n_y = n \tag{9.77}$$

oder wenn man mit

$$x = \frac{n_x}{n} \qquad y = \frac{n_y}{n} \tag{9.78}$$

die Anzahlkonzentrationen der beiden Komponenten einführt

$$x + y = 1 \tag{9.79}$$

Sind alle Partikeln dem Volumen nach gleich groß, so ist die Anzahlkonzentration der beiden Komponenten jeweils gleich ihrer Volumenkonzentration. Daher gilt

$$E(x) = p \qquad E(y) = 1 - p = q \tag{9.80}$$

Die gleichzeitige Entnahme von n Partikeln kann man sich durch die n-malige Entnahme jeweils einer Partikel ersetzt denken. Damit wird die Probenahme aus einer Zufallsmischung auf das statistische Modell des *Ziehens ohne Zurücklegen* zurückgeführt. Hierfür gilt die Binomialverteilung

$$P(n_x) = \binom{n}{n_x} p^{n_x} \, q^{n-n_x} \tag{9.81}$$

wenn vorausgesetzt werden darf, dass die Mischung durch die Probenahme nicht merklich verändert wird. Die Varianz dieser Verteilung ist

$$\sigma_{n_x}^2 = npq \tag{9.82}$$

oder in der Variablen x

$$\sigma_x^2 = \frac{pq}{n} \tag{9.83}$$

Dies ist der kleinste Wert, den die Varianz σ_x^2 im Fall einer Mischung gleichgroßer Partikeln annehmen kann.

Die Varianz der Probenkonzentration x ist, wie 9.83 zeigt, umgekehrt proportional zum Probenumfang. Dieses Ergebnis gilt zunächst nur für den speziellen Fall einer Mischung aus gleich großen Partikeln. Karl Sommer[7] hat in umfangreichen Modellrechnungen eine Annäherung an reale Aufgabenstellungen versucht, die dadurch gekennzeichnet sind, dass beide Komponenten aus Partikeln mit unterschiedlichen Größenverteilungen bestehen. Eine mathematisch exakte Berechnung der Varianz σ_x^2 ist dann nicht mehr möglich. Es bleibt aber der Befund, dass die Varianz σ_x^2 dem Probenumfang umgekehrt proportional ist. Dies gilt auch dann, wenn keine Zufallsmischung vorliegt.

[7] K. Sommer Probenahme von Pulvern und körnigen Massengütern, Springer-Verlag, Berlin, Heidelberg 1979, Seite 131 ff.

9.6.3 Folgerungen

Wenn man den Probenumfang (jeweils gezogene Teilmenge) sehr groß macht, kann man erreichen, dass die Varianz σ_x^2 sehr kleine Werte annimmt. Das ist jedoch kein Vorteil, wenn man bedenkt, dass zur Bestimmung der Probenkonzentration x ein Messverfahren eingesetzt werden muss, dessen Genauigkeit begrenzt ist. Das Analysenergebnis y setzt sich, wie in Abschnitt 6.7 *Beispiele für Konfidenzintervalle* auf Seite 116 gezeigt wurde, aus dem jeweiligen wahren Wert x der Probenkonzentration und einem Messfehler M zusammen

$$y = x + M \tag{9.84}$$

Für den Erwartungswert von y gilt

$$E(y) = E(x) \tag{9.85}$$

da $E(M) = 0$ vorausgesetzt werden darf, wenn kein systematischer Messfehler vorliegt. Wird Unabhängigkeit zwischen x und M vorausgesetzt, so gilt für die Varianz von y

$$\sigma_y^2 = \sigma_x^2 + \sigma_M^2 \tag{9.86}$$

Die gesuchte Varianz σ_x^2 darf man nur dann der Varianz des Analysenergebnisses σ_y^2 gleichsetzen, wenn dafür gesorgt ist, dass die Varianz des Messfehlers σ_M^2 sehr klein ist im Verhältnis zu σ_x^2

$$\sigma_M^2 \ll \sigma_x^2 \tag{9.87}$$

Diese Bedingung kann im Prinzip immer erfüllt werden, indem man den Probenumfang hinreichend klein wählt. Wird die Bedingung 9.87 verletzt, ist eine Beurteilung der Mischungsqualität nicht möglich.

9.7 Prüfverfahren in der Mischtechnik[8]

9.7.1 Aufgabenstellung

Die Wirksamkeit von Feststoffmischern und die nach einer bestimmten Mischdauer erreichte Homogenität von Feststoffmischungen lassen sich im Allgemeinen nur dadurch überprüfen, dass man Proben zieht und deren Zusammensetzung analysiert. Auf der Grundlage der Analysenergebnisse muss jeweils entschieden werden, ob die geforderte Homogenität erreicht wurde oder nicht. Dabei sind wie bei jedem statistischen Prüfverfahren Fehlentscheidungen unvermeidlich.

Gesucht sind quantitative Aussagen über die Höhe der Risiken, das heißt über die Größe der Fehler 1. Art (Wahrscheinlichkeit der Ablehnung hinreichend guter Mischungen) und 2. Art (Wahrscheinlichkeit der Annahme schlechter Mischungen).

[8] J. Raasch und K. Sommer, Anwendung von statistischen Prüfverfahren im Bereich der Mischtechnik, Chemie-Ingenieur-Technik 62 (1990) 1, S. 17–22

9.7.2 Mathematischer Ansatz

Die folgenden Überlegungen beschränken sich auf Feststoffmischungen aus zwei Komponenten und gelten nur, wenn das Analyseverfahren hinreichend genau ist, das heißt wenn die zufälligen Messfehler klein sind im Verhältnis zu den gemessenen Schwankungen der Probenzusammensetzung (siehe Abschnitt 9.6 *Mischgütemaß für Feststoffmischungen* auf Seite 166).

Als Mischgütemaß soll die Varianz σ_x^2 der Probenkonzentration x der 1. Komponente verwendet werden. Ein erwartungstreuer Schätzwert für σ_x^2 ist die Stichprobenvarianz $\tilde{s}_x^2$

$$\tilde{s}_x^2 = \frac{1}{z} \sum_{i=1}^{z} (x_i - p)^2 \tag{9.88}$$

Hierin bezeichnet p die Konzentration der 1. Komponente in der Mischung und z die Anzahl der Proben. Wenn die Mischungskonzentration p nicht bekannt ist, muss p durch das Stichprobenmittel $\bar{x}$ angenähert werden. Für die Stichprobenvarianz ist dann die alternative Definition

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

zu verwenden, wie in Abschnitt 6.1 *Grundbegriffe* auf Seite 101 dargestellt ist.

Weiterhin wird vorausgesetzt, dass p bekannt ist, was bei diskontinuierlichen Mischprozessen meist auch der Fall ist. Ob x und p als Anzahl-, Volumen- oder Massenkonzentrationen gemessen werden, ist gleichgültig.

Könnte man den Stichprobenumfang z unendlich groß machen, würden Stichprobenvarianz und Varianz der Grundgesamtheit übereinstimmen. Bei endlichem Stichprobenumfang weicht der Schätzwert von der Zielgröße σ_x^2 ab. Dadurch kommen statistische Überlegungen ins Spiel.

Schließlich wird davon ausgegangen, dass sich die Verteilung der Probenkonzentration x, gleichgültig wie weit der Mischprozess fortgeschritten ist, hinreichend genau durch eine Normalverteilung darstellen lässt. Abgesehen von einer Anfangsphase, in der die beiden Komponenten noch weitgehend voneinander getrennt vorliegen, ist diese Annahme bei den üblichen Probengrößen erfüllt.

Setzt man voraus, dass die Probenkonzentration x normalverteilt ist, folgt, wie in Abschnitt 6.3 *Konfidenzintervalle für die Varianz einer Normalverteilung mit bekanntem Mittelwert* auf Seite 113 nachzulesen ist, dass die Verteilung der Stichprobenvarianz $\tilde{s}_x^2$ durch die Substitution

$$\chi^2 = z \frac{\tilde{s}_x^2}{\sigma_x^2} \tag{9.89}$$

auf die χ^2-Verteilung zurückgeführt werden kann. Bezeichnet man mit $\Phi(\chi^2)$ die Summenfunktion der χ^2-Verteilung, so kann man mit

$$P(\chi_u^2 \le z \frac{\tilde{s}_x^2}{\sigma_x^2} \le \chi_o^2) = \Phi(\chi_o^2) - \Phi(\chi_u^2) \tag{9.90}$$

die Wahrscheinlichkeit berechnen, dass die Größe $z\tilde{s}_x^2/\sigma_x^2$ zwischen einer unteren Grenze χ_u^2 und einer oberen Grenze χ_o^2 liegt.

Statistische Prüfverfahren werden auf einer bestimmten Hypothese aufgebaut. Das Ergebnis einer Stichprobe soll dann darüber entscheiden, ob die Hypothese im konkreten Fall anzunehmen oder abzulehnen ist. Im Zusammenhang mit Mischprozessen interessiert die Frage, ob eine bestimmte Homogenität der Mischung, gekennzeichnet durch einen Wert σ_1^2 der Varianz σ_x^2, erreicht worden ist oder nicht. Die Hypothese wird man in der Form $\sigma_x^2 \leq \sigma_1^2$ formulieren, das heißt sie lautet: *Die Mischung hat eine Homogenität erreicht, die mindestens der gewünschten entspricht.* Verwirft man die Hypothese, so bedeutet das, dass man sich für die Alternative $\sigma_x^2 > \sigma_1^2$ entscheidet. Nur Abweichungen nach oben führen in diesem Fall zur Ablehnung der Hypothese (einseitiger Test).

Für die Stichprobenvarianz $\tilde{s}_x^2$ gibt man mit

$$\tilde{s}_x^2 \leq \tilde{s}_{xo}^2 \tag{9.91}$$

folgerichtig einen Annahmebereich vor, der nur nach oben begrenzt ist. Der oberen Grenze $\tilde{s}_{xo}^2$ entspricht eine obere Grenze χ_o^2 der Variablen χ^2. Setzt man voraus, dass die Varianz der Grundgesamtheit σ_x^2 exakt den Wert σ_1^2 erreicht hat, erhält man aus 9.89 für χ_o^2 speziell den Wert

$$\chi_{o1}^2 = z\frac{\tilde{s}_{xo}^2}{\sigma_1^2} \tag{9.92}$$

Hiermit kann man die Wahrscheinlichkeit berechnen, dass dann, wenn $\sigma_x^2 = \sigma_1^2$ ist, die Stichprobenvarianz $\tilde{s}_x^2$ in den Annahmebereich fällt

$$P(\tilde{s}_x^2 \leq \tilde{s}_{xo}^2) = \Phi(\chi_{o1}^2) \tag{9.93}$$

Umgekehrt fällt dann mit der Wahrscheinlichkeit α

$$\alpha = 1 - \Phi(\chi_{o1}^2) \tag{9.94}$$

die Stichprobenvarianz $\tilde{s}_x^2$ nicht in den Annahmebereich, obwohl σ_x^2 den Wert σ_1^2 erreicht hat. Dies bedeutet, dass im Fall $\sigma_x^2 = \sigma_1^2$ die Hypothese mit der Wahrscheinlichkeit α zu Unrecht abgelehnt wird (Fehler 1. Art).

In Abbildung 9.4 ist dieser Sachverhalt veranschaulicht. Die Kurve $\varphi_1(s_x^2/\sigma_1^2)$ beschreibt die Dichtefunktion der auf die Varianz σ_1^2 bezogenen Stichprobenvarianz s_x^2 unter der Voraussetzung, dass die Varianz der Grundgesamtheit σ_x^2 exakt den Wert σ_1^2 erreicht hat. Dabei wurde angenommen, dass die Stichprobe aus $z = 20$ einzelnen Proben besteht. Die Grenze des Annahmebereichs wurde mit $s_{xo}^2/\sigma_1^2 = 1,57$ festgelegt. Hierzu gehört ein Fehler 1. Art $\alpha = 0,05$. In Abbildung 9.4 entspricht dem die Fläche unter der Kurve φ_1 rechts von der Grenze des Annahmebereichs.

Wie man zeigen kann, wird der Fehler 1. Art dann kleiner als der mit 9.94 berechnete Wert α, wenn die Varianz der Grundgesamtheit Werte $\sigma_x^2 < \sigma_1^2$ annimmt. Für den Fehler 1. Art erhält man daher mit 9.94 eine obere Grenze für den Fall, dass die Hypothese $\sigma_x^2 \leq \sigma_1^2$ erfüllt ist.

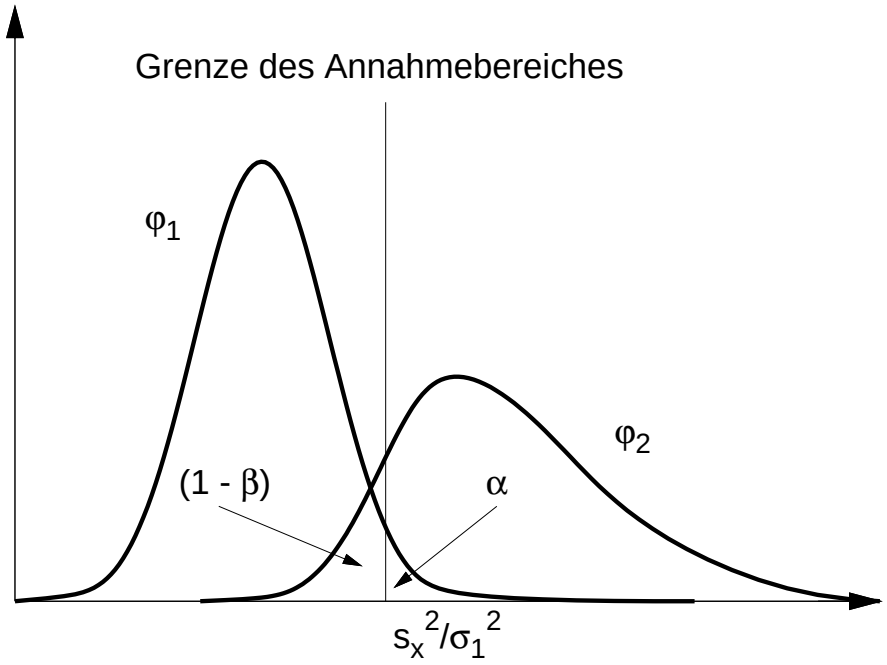


Abb. 9.4: Fehler 1. und 2. Art, dargestellt als Flächen unter den Kurven $\varphi_1(s_x^2/\sigma_1^2)$ beziehungsweise $\varphi_2(s_x^2/\sigma_1^2)$

Auch dann, wenn die Hypothese $\sigma_x^2 \leq \sigma_1^2$ nicht erfüllt ist, das heißt wenn die Varianz der Grundgesamtheit σ_x^2 einen Wert $\sigma_2^2 > \sigma_1^2$ angenommen hat, wird die Stichprobenvarianz s_x^2 dennoch mit einer bestimmten Wahrscheinlichkeit $(1-\beta)$ in den Annahmebereich fallen. Aus 9.89 erhält man mit

$$\chi_{o2}^2 = z\frac{s_{xo}^2}{\sigma_2^2} \tag{9.95}$$

die obere Grenze des Annahmebereichs in der Variablen χ^2 für den Fall, dass $\sigma_x^2 = \sigma_2^2$ ist. Aus

$$P(\chi_u^2 < \chi^2 < \chi_o^2) = \int_{\chi_u^2}^{\chi_o^2} \varphi(\chi^2)\,\mathrm{d}\chi^2 = \Phi(\chi_o^2) - \Phi(\chi_u^2) \tag{9.96}$$

folgt damit weiter

$$1 - \beta = \Phi(\chi_{o2}^2) \tag{9.97}$$

Dies ist die Wahrscheinlichkeit, dass die Hypothese $\sigma_x^2 \leq \sigma_1^2$ im Fall $\sigma_x^2 = \sigma_2^2$ zu Unrecht angenommen wird (Fehler 2. Art).

In Abbildung 9.4 ist zusätzlich zu $\varphi_1(s_x^2/\sigma_1^2)$ die Funktion $\varphi_2(s_x^2/\sigma_1^2)$ eingezeichnet, die sich für die auf die Varianz σ_1^2 bezogenen Stichprobenvarianz s_x^2 dann errechnet, wenn die Varianz der Grundgesamtheit σ_x^2 einen Wert $\sigma_2^2 = 2\,\sigma_1^2$ angenommen hat. Man erkennt, dass die zweite Kurve gegenüber der ersten deutlich nach rechts verschoben ist. Der Teil der Fläche unter der zweiten Kurve, der in den Annahmebereich fällt – also links von der Grenze liegt – ist gleich dem Fehler 2. Art. Im speziellen Fall ist $(1-\beta) = 0,265$.

Es ist leicht einzusehen, dass der Fehler 2. Art dann kleiner wird als der mit 9.97 berechnete Wert $(1-\beta)$, wenn die Varianz der Grundgesamtheit Werte $\sigma_x^2 > \sigma_2^2$ annimmt. Man erhält daher mit 9.97 eine obere Grenze für den Fehler 2. Art im Fall, dass $\sigma_x^2 > \sigma_2^2$ ist.

9.7.3 Folgerungen

Bei der Anwendung eines statistischen Prüfverfahrens sind Fehlentscheidungen unvermeidlich. Um die damit verbundenen Risiken zu begrenzen, ist es in jedem Fall notwendig, sowohl für den Fehler 1. Art als auch für den 2. Art Grenzen festzulegen. Im Zusammenhang mit der Untersuchung von Feststoffmischungen geschieht dies durch die beiden Forderungen

$$\sigma_x^2 \leq \sigma_1^2 \qquad \Phi(z\ s_{xo}^2/\sigma_x^2) \geq 1-\alpha \tag{9.98}$$

$$\sigma_x^2 \geq \sigma_2^2 \qquad \Phi(z\ s_{xo}^2/\sigma_x^2) \leq 1-\beta \tag{9.99}$$

Hiermit werden für σ_1^2, σ_s^2, α und β Zahlenwerte vorgegeben. Die beiden Unbekannten z und s_{xo}^2, das heißt Stichprobenumfang und obere Grenze des Annahmebereichs, erhält man dann durch iteratives Lösen von 9.94 und 9.97. Wenn man versucht, ein Prüfverfahren nur auf der Forderung 9.98 aufzubauen, kommt man zu dem überraschenden Ergebnis, dass der Stichprobenumfang scheinbar ins Belieben gestellt ist. Man könnte den Test dann folgerichtig auch mit beispielsweise nur $z = 3$ Proben durchführen, ohne die Forderung 9.98 zu verletzen. Dabei würde allerdings der Fehler 2. Art sehr groß, das heißt auch sehr schlechte Mischungen würden noch mit hoher Wahrscheinlichkeit angenommen. Im Grunde wird nämlich die Mischungsqualität erst durch die Forderung 9.99 definiert. Ein Test, der allein auf der Foderung 9.98 basiert, sagt über die Homogenität der Mischung überhaupt nichts aus.

Eher Sinn macht ein Test, der nur von der Forderung 9.99 ausgeht. Damit ist immerhin garantiert, dass schlechte Mischungen mit hoher Wahrscheinlichkeit abgelehnt werden. Wiederum ist der Stichprobenumfang scheinbar beliebig, das heißt der Test könnte auch mit einer sehr kleinen Probenanzahl durchgeführt werden. Dabei würde allerdings der Fehler 1. Art sehr groß werden. Auch gute Mischungen würden mit vergleichsweise hoher Wahrscheinlichkeit abgelehnt werden.

Für Planung und Auswertung von Mischungsuntersuchungen der hier geschilderten Art ist die Frage entscheidend, wie die Werte der Varianzen σ_1^2 und σ_2^2 festgelegt werden, die den Forderungen 9.98 und 9.99 zu Grunde liegen. Da die Qualität des jeweiligen Mischgutes von σ_2^2 abhängt, wird dieser Wert von der Anwendungstechnik bestimmt. Im Unterschied hierzu ist der Wert von σ_1^2 für die Anwendungstechnik ohne Bedeutung. Seine Festlegung ist allein unter dem Gesichtspunkt zu sehen, die Kosten des Mischprozesses zu begrenzen, das heißt sich gegen die Möglichkeit abzusichern, dass ausreichend gute Mischungen mit unannehmbar hoher Wahrscheinlichkeit abgelehnt werden. Äußerstenfalls kann σ_1^2 mit dem Wert σ_E^2 der Varianz σ_x^2 gleichgesetzt werden, der nach sehr langer Mischzeit

unter den gegebenen Bedingungen gerade noch erreichbar ist. Günstiger ist es, wenn für σ_1^2 ein höherer Wert gewählt werden kann.

Prinzipielle Schwierigkeiten für das Prüfverfahren ergeben sich dann, wenn die Werte der Varianzen σ_1^2 und σ_2^2 dicht beieinander liegen. Will man auch unter solchen Umständen sowohl den Fehler 1. Art als auch den Fehler 2. Art klein halten, so lässt sich dies, wenn überhaupt, nur mit einem großen Stichprobenumfang verwirklichen, das heißt mit hohem Kostenaufwand für das Prüfverfahren. Man wird in solchen Fällen auf der einen Seite versuchen müssen, die Mischtechnik zu verbessern mit dem Ziel, den Wert vo σ_E^2 beziehungsweise σ_1^2 niedriger ansetzen zu können, und auf der anderen Seite erwägen müssen, die Anforderungen der Anwendungstechnik zu reduzieren, um für σ_2^2 einen höheren Wert vorgeben zu können.

Ein wichtiger Grenzfall der Mischtechnik ist die **gleichmäßige Zufallsmischung**. Hiermit bezeichnet man einen Mischungszustand, der nur unter idealen Bedingungen (keinerlei Entmischungstendenzen, keine Toträume im Mischer) und nach langer Mischzeit zu erreichen ist und der durch ein Minimum σ_0^2 der Varianz σ_E^2 gekennzeichnet ist. Unter Zugrundelegen bestimmter Modellannahmen ist dieser Grenzwert σ_0^2 der Berechnung zugänglich.

Will man auf Zufallsmischung prüfen, das heißt setzt man sich zum Ziel, eine Mischung nur dann anzunehmen, wenn der Zustand der gleichmäßigen Zufallsmischung erreicht ist, so steht man vor dem Dilemma, dass dann $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$ gesetzt werden muss mit der Folge, dass auch $\alpha = \beta$ wird. Dies bedeutet, dass es dann prinzipiell unmöglich wird, die Fehler 1. und 2. Art gleichzeitig klein zu halten. Verlangt man beispielsweise, dass eine Mischung, die schlechter als eine Zufallsmischung ist, nur mit einer Wahrscheinlichkeit von 0,05 angenommen wird, so muss man hinnehmen, dass eine Zufallsmischung mit einer Wahrscheinlichkeit von 0,95 zu Unrecht abgelehnt wird. Würde man umgekehrt den Fehler 1. Art auf 0,05 begrenzen, so hätte dies zur Folge, dass schlechtere Mischungen mit einer Wahrscheinlichkeit von 0,95 angenommen würden.

9.8 Umrechnung verschiedener Mengen- und Merkmalsarten[9]

9.8.1 Aufgabenstellung

Bisher haben wir den Inhalt oder Umfang einer Menge durch Zählen ermittelt, also durch die Anzahl der in der Menge enthaltenen Elemente gekennzeichnet. Das ist die in fast allen Statistikbüchern ausschließlich verwendete **Mengenart**[10]. In Abschnitt 1.3 *Mengen* auf Seite 4 haben wir jedoch bereits angedeutet, dass in der Technik auch andere Größen wie Masse oder Volumen als Mengenart ver-

[9] K. Leschonski, W. Alex und B. Koglin, Teilchengrößenanalyse, Chemie-Ingenieur-Technik 46 (1974) 1, S. 23–26

[10] Richtiger wäre vielleicht *Mengenmaß*.

wendet werden. Bei einer Siebanalyse beispielsweise wird der Inhalt der einzelnen Siebfraktionen nicht durch die Anzahl der Partikel, sondern durch ihre Masse angegeben, das heißt die Siebfraktionen werden gewogen.

Hat man eine breite Partikelgrößenverteilung abschnittsweise durch Messverfahren bestimmt, die unterschiedliche Mengenarten verwenden, oder will man zwei Verteilungen miteinander vergleichen, die aus unterschiedlichen Messverfahren stammen, so sind die Mengenarten auf eine einheitliche Größe umzurechnen. Man darf nicht eine Anzahlverteilung mit einer Volumenverteilung vergleichen. Dass unter Umständen auch die Merkmale umzurechnen sind, ist eine andere Geschichte.[11]

Es kann verwirren, dass dieselbe physikalische Größe – beispielsweise die Masse – sowohl als Merkmal auf der Abszissenachse wie auch als Mengenart auf der Ordinatenachse auftritt. Als Merkmal handelt es sich jedoch um die Masse eines einzelnen dispersen Elementes, als Mengenart um die gesamte Masse der dispersen Elemente in einer Klasse, bezogen auf die Masse der Stichprobe.

Die Anzahl kann nie als Merkmal auftreten, da jedem Element unterschiedslos die Anzahl 1 zukommt. Umgekehrt könnte die Temperatur ein Merkmal sein, aber wegen der fehlenden Additivität nicht zur Kennzeichnung eines Mengeninhaltes dienen.

Gehen wir von einem **Partikelmerkmal** zu einem anderen über, das eine Funktion des ersten ist, so sind auch die zugehörigen Werte der Dichtefunktion der Verteilung umzurechnen. Dies ergibt sich aus der Forderung, dass die Fläche unter der Kurve der Dichtefunktion (das Integral über die Dichtefunktion von x_{min} bis x_{max}) gleich 1 sein soll.

9.8.2 Mathematischer Ansatz

Wir gehen aus von den Definitionen der Dichtefunktion (2.24 auf Seite 18) und der Summenfunktion (2.25) der relativen Häufigkeit

$$f(x_i) = \frac{z(x_i)}{z \cdot \Delta x_i} = \frac{1}{\sum\limits_i z(x_i)} \, \frac{z(x_i)}{\Delta x_i} \tag{9.100}$$

$$F(x_i) = f(x_1)\Delta x_1 + f(x_2)\Delta x_2 + \ldots + f(x_i)\Delta x_i = \sum_{\nu=1}^{i} f(x_\nu)\,\Delta x_\nu \tag{9.101}$$

und beschränken das Merkmal x auf eine kennzeichnende Länge der Elemente. Die Schreibweise $\sum\limits_i z(x_i)$ bedeutet *Summe über alle vorkommenden Indices i*.

Dann ist mit nicht näher bekannten Formfaktoren f, die wir als in der Probe konstant voraussetzen,

$$y = f \cdot x^n \qquad \text{mit} \qquad n = 0, 1, 2, 3 \tag{9.102}$$

[11] Siehe Veröffentlichungen über Formfaktoren in der Dispersitätsanalyse.

die Anzahl, die Länge, die Fläche (Oberfläche, Projektionsfläche) beziehungsweise das Volumen eines Elementes. Andere Exponenten haben keine praktische Bedeutung.

Die Anzahlen $z(x_i)$ in 9.100 multiplizieren wir mit y nach 9.102 und erhalten so eine verallgemeinerte Gleichung

$$\begin{aligned} f(x_i) &= \frac{1}{\sum_i \Big(z(x_i)\ y(x_i)\Big)}\ \frac{z(x_i)\ y(x_i)}{\Delta x_i} \\ &= \frac{1}{\sum_i \Big(z(x_i)\ x_i^n\Big)}\ \frac{z(x_i)\ x_i^n}{\Delta x_i} \end{aligned} \tag{9.103}$$

da sich der als konstant angenommene Formfaktor f herauskürzt. Die Gleichung besagt, dass beispielsweise die Dichtefunktion der Volumenverteilung sich aus dem Volumen aller Elemente in einer Klasse dividiert durch das Gesamtvolumen der Stichprobe und die Klassenbreite ergibt. Für $n = 0$ erhalten wir die Dichtefunktion der Anzahlverteilung nach 9.100, da für jedes endliche x gilt $x^0 = 1$.

In der Dispersitätsanalyse hat sich für die Dichtefunktion die Schreibweise $q_n(x)$ eingebürgert, um auf die Mengenart hinzuweisen, sofern sie die Potenz einer Länge ist. Bei konstanter Massendichte (spezifische Masse) ρ ist die Massenverteilung mit der Volumenverteilung identisch, darf also ebenso mit $q_3(x)$ bezeichnet werden. Andernfalls müsste man $q_m(x)$ schreiben. Für die Summenfunktion verwendet man entsprechend $Q_n(x)$. Aus 9.103 erhalten wir

$$\begin{aligned} q_0(x_i) &= \frac{1}{\sum_i z(x_i)}\ \frac{z(x_i)}{\Delta x_i} \\ q_1(x_i) &= \frac{1}{\sum_i \Big(z(x_i)\ x_i\Big)}\ \frac{z(x_i)\ x_i}{\Delta x_i} = \frac{\sum_i z(x_i)}{\sum_i \Big(z(x_i)\ x_i\Big)}\ q_0(x_i)\ x_i \\ q_2(x_i) &= \frac{1}{\sum_i \Big(z(x_i)\ x_i^2\Big)}\ \frac{z(x_i)\ x_i^2}{\Delta x_i} = \frac{\sum_i z(x_i)}{\sum_i \Big(z(x_i)\ x_i^2\Big)}\ q_0(x_i)\ x_i^2 \\ q_3(x_i) &= \frac{1}{\sum_i \Big(z(x_i)\ x_i^3\Big)}\ \frac{z(x_i)\ x_i^3}{\Delta x_i} = \frac{\sum_i z(x_i)}{\sum_i \Big(z(x_i)\ x_i^3\Big)}\ q_0(x_i)\ x_i^3 \end{aligned} \tag{9.104}$$

oder umgekehrt

$$q_0(x_i) = \frac{\sum_i \Big(z(x_i)\ x_i^3\Big)}{\sum_i z(x_i)}\ q_3(x)\ x^{-3} \tag{9.105}$$

und entsprechend für die anderen Kombinationen der Dichtefunktionen. Wegen

$$\frac{\sum\limits_i z(x_i)}{\sum\limits_i \Big(z(x_i)\ x_i^n\Big)} = \frac{1}{\sum\limits_i \Big(\dfrac{1}{\sum\limits_i z(x_i)}\ \dfrac{z(x_i)}{\Delta x_i}\ \Delta x_i\ x_i^n\Big)} = \frac{1}{\sum\limits_i \Big(q_0(x_i)\ \Delta x_i\ x_i^n\Big)}$$

(die Summen über alle i sind feste Werte) lässt sich 9.104 umschreiben zu

$$\begin{aligned}
q_0(x_i) &= \frac{1}{\sum\limits_i \Big(q_0(x_i)\ \Delta x_i\Big)}\ q_0(x_i) = \frac{1}{M_{0,0}}\ q_0(x_i) \qquad \text{mit} \qquad M_{0,0} = 1 \\
q_1(x_i) &= \frac{1}{\sum\limits_i \Big(x_i\ q_0(x_i)\ \Delta x_i\Big)}\ q_0(x_i)\ x_i = \frac{1}{M_{1,0}}\ q_0(x_i)\ x_i \\
q_2(x_i) &= \frac{1}{\sum\limits_i \Big(x_i^2\ q_0(x_i)\ \Delta x_i\Big)}\ q_0(x_i)\ x_i^2 = \frac{1}{M_{2,0}}\ q_0(x_i)\ x_i^2 \\
q_3(x_i) &= \frac{1}{\sum\limits_i \Big(x_i^3\ q_0(x_i)\ \Delta x_i\Big)}\ q_0(x_i)\ x_i^3 = \frac{1}{M_{3,0}}\ q_0(x_i)\ x_i^3 \qquad (9.106)
\end{aligned}$$

mit der Momentenschreibweise von 3.77 auf Seite 44. Umgekehrt gilt

$$q_0(x_i) = M_{3,0}\ q_3(x_i)\ x_i^{-3} \qquad \text{usw.} \qquad (9.107)$$

Das Moment $M_{1,0}$ ist das mittlere x der Anzahlverteilung; entsprechend ist $M_{3,0}$ das mittlere x^3 der Anzahlverteilung, bis auf einen Formfaktor also das mittlere Partikelvolumen. Für den Übergang $\Delta x_i \to 0$ werden aus den Summen bestimmte Integrale mit den Grenzen 0 und ∞ beziehungsweise x_{min} und x_{max}.

Für die Summenfunktion erhalten wir entsprechend aus 9.101 und 9.103

$$\begin{aligned}
F(x_i) = Q_n(x_i) &= \frac{1}{\sum\limits_\nu \Big(z(x_\nu)\ x_\nu^n\Big)}\ \sum_{\nu=1}^{i} \frac{z(x_\nu)\ x_\nu^n}{\Delta x_\nu}\ \Delta x_\nu \\
&= \frac{1}{\sum\limits_\nu \Big(z(x_\nu)\ x_\nu^n\Big)}\ \sum_{\nu=1}^{i} z(x_\nu)\ x_\nu^n \\
&= \frac{1}{\sum\limits_\nu \Big(q_0(x_\nu)\ x_\nu^n\ \Delta x_\nu\Big)}\ \sum_{\nu=1}^{i} \Big(q_0(x_\nu)\ x_\nu^n\ \Delta x_\nu\Big) \\
&= \frac{1}{M_{n,0}}\ \sum_{\nu=1}^{i} \Big(q_0(x_\nu)\ x_\nu^n\ \Delta x_\nu\Big) \qquad (9.108)
\end{aligned}$$

Für $i = i_{max}$ ergibt sich $Q_n(x_i) = 1$, wie es sein soll. Falls die Mengenart nicht die Potenz einer Länge oder das Merkmal nicht eine kennzeichnende Partikellänge ist, sondern beispielsweise eine Sinkgeschwindigkeit, gelten im Prinzip die gleichen Überlegungen, nur sehen die Formeln anders aus. Wir schenken uns – zumindest vorläufig – die Verallgemeinerungen. Ebenso werden die Formeln umfangreicher, wenn die Formfaktoren oder die Massendichte in der Probe nicht konstant sind.

Als Partikelmerkmal oder -größe kommen verschiedene physikalische Größen in Betracht, die Eigenschaften einer einzelnen Partikel sind und eine Ordnung der Partikel nach dem Zahlenwert der Größe erlauben. Gebräuchlich sind Längen verschiedener Art, aber auch die Sinkgeschwindigkeit in einem Gas oder einer Flüssigkeit oder die Masse werden verwendet, je nach Aufgabenstellung und Messmöglichkeiten.

Gegeben sei die Verteilung $\varphi_0(y)$ eines Partikelmerkmals y, das eine Funktion eines anderen Partikelmerkmals x ist. Gesucht ist die Verteilung $\varphi_1(x)$ des Partikelmerkmals x. Die Mengenbilanz erfordert, dass

$$\varphi_0(y)\,\mathrm{d}y = \varphi_1(x)\,\mathrm{d}x$$

oder umgestellt

$$\varphi_1(x) = \varphi_0(y)\,\frac{\mathrm{d}y}{\mathrm{d}x}$$

gilt.

Beispiel 9.1 : Es sei experimentell die Verteilung $\varphi_0(w)$ der Sinkgeschwindigkeit w der Partikeln in einer Flüssigkeit oder einem Gas bestimmt worden. Für den Zusammenhang zwischen der Sinkgeschwindigkeit w und dem Äquivalentdurchmesser x der Kugel gleicher Sinkgeschwindigkeit gilt im Bereich des Gesetzes von George Gabriel Stokes (1819–1903)

$$w = k\,x^2$$

Daraus folgt

$$\begin{aligned} \mathrm{d}w &= 2\,kx\,\mathrm{d}x \\ \varphi_0(w)\,\mathrm{d}w &= \varphi_1(x)\,\mathrm{d}x \\ \varphi_1(x) &= 2\,kx\,\varphi_0(w) = \frac{2\,w}{x}\,\varphi_0(w) \end{aligned}$$

Ende Beispiel

9.8.3 Folgerungen

Die in der Technik gebräuchlichen, unterschiedlichen Mengenarten beeinflussen die Form einer Verteilung. Das ist auch anschaulich klar: Wenige grobe Partikel

tragen zu einer Anzahlverteilung wenig bei, zu einer Volumen- oder Massenverteilung dagegen viel. Umgekehrt tragen Partikel am feinen Ende einer Verteilung viel zur Anzahl, aber wenig zur Masse bei.

Die vorstehend erläuterte Umrechnung der Mengenarten ist zwar mathematisch einwandfrei, berücksichtigt man aber die Streuung der Messwerte an den Enden einer Verteilung, so erkennt man, dass diese erheblich anwachsen kann, wie schon auf Seite 158 bemerkt. Die Folgerung ist, nach Möglichkeit schon bei der Messung die Mengenart zu wählen, die für die verfahrenstechnische Aufgabenstellung bedeutend ist. Leider lässt sich das nicht immer verwirklichen.

Eine gern übersehene Folgerung aus den verschiedenen Mengenarten ist, dass man nicht einfach von einer einzigen mittleren Partikellänge $\bar{x}$ reden darf, sondern dazu sagen muss, ob sie aus einer Anzahl- oder Volumen- beziehungsweise Massenverteilung berechnet worden ist, wenn wir einmal Länge und Fläche als Mengenart beiseite lassen. Es gibt zahlreiche Definitionen einer mittleren Partikellänge.

9.9 Spezifische Oberfläche[12]

9.9.1 Aufgabenstellung

Will man die Feinheit (Dispersität) eines dispersen Systems (Haufwerk oder ähnlich, Particulate Matter) durch eine einzige Zahlenangabe kennzeichnen, bieten sich verschiedene Mittelwerte an. Unter diesen spielt die spezifische Oberfläche eine herausragende Rolle, da viele Eigenschaften von ihr abhängen. Ein Stück Kohle oder Kandiszucker neigt kaum zur Explosion, dieselbe Stoffmenge mit durch Mahlen stark vergrößerter Oberfläche kann gefährlich werden. Auch beim vieldiskutierten Fein- und Feinststaub ist die große Oberfläche bei geringer Masse eine wesentliche Eigenschaft.

Als **spezifische Oberfläche** wird die auf das Volumen oder die Masse bezogene Oberfläche der dispersen Phase (Partikel, Körner, Tropfen) bezeichnet

$$S_V = \frac{S}{V} \qquad S_m = \frac{S}{m} \tag{9.109}$$

Die Dimension der volumenbezogenen Oberfläche ist 1/Länge; S_V kann daher mit aller Vorsicht als Kehrwert einer Partikellänge oder $1/S_V$ als eine mittlere Partikellänge gedeutet werden. Diese mittlere Partikellänge ist der Durchmesser einer Kugel, die ein Sechstel der gegebenen volumenbezogenen Oberfläche aufweist, eine ungebräuchliche Größe. Im Fall einer einheitlichen Massendichte ρ gilt die Umrechnung

$$S_V = \rho \, S_m \tag{9.110}$$

Die Oberfläche S der dispersen Phase ist eine sehr unscharf definierte Größe. Sieht man sich anscheinend glatte Partikel wie Sandkörner oder Zuckerkristalle

[12] H. Rumpf und K. F. Ebert, Darstellung von Kornverteilungen und Berechnung der spezifischen Oberfläche, Chemie-Ingenieur-Technik 36 (1964), S. 523–537

unter einem Mikroskop – insbesondere unter einem Rasterelektronenmikroskop – an, so entdeckt man zahlreiche Unebenheiten, die zur Oberfläche beitragen. Bei porösen Stoffen wie Koks oder Ziegelsteinen stellt sich schon dem unbewaffneten Auge die Frage, welche Rauhigkeiten und Poren zur Oberfläche gerechnet werden sollen oder nicht. Kurz und bündig: die Oberfläche disperser Systeme wird durch die Aufgabenstellung oder das Messverfahren definiert.

9.9.2 Mathematischer Ansatz

Ein Weg zur Bestimmung der spezifischen Oberfläche unter mehreren ist die Berechnung aus einer gemessenen Partikelgrößenverteilung. Es sei $q_0(x)$ die Dichtefunktion der gemessenen relativen Häufigkeit des Merkmals x, das eine kennzeichnende Partikellänge darstelle. Wir setzen voraus, dass die Partikel unabhängig von ihrer Größe dieselbe Form haben, das heißt geometrisch ähnlich sind. Dann ist die Oberfläche einer Partikel proportional x^2, ihr Volumen proportional x^3. In jeder Partikelgrößenklasse multiplizieren wir die Anzahl z_i der Partikel mit der Oberfläche pro Partikel, berechnet aus der Partikellänge x_i in der Klassenmitte

$$S_i = z_i \; f_S \; x_i^2 = z \; q_0(x_i) \; \Delta x_i \; f_S \; x_i^2 \qquad \text{mit} \qquad q_0(x_i) = \frac{z_i}{z \; \Delta x_i}$$

Darin ist f_S ein Proportionalitätsfaktor (Formfaktor). Dann summieren wir die Gesamtoberflächen jeder Klasse von x_u bis x_o auf und erhalten bis auf den unbekannten Proportionalitätsfaktor die Gesamtoberfläche S des dispersen Systems zwischen x_u und x_o

$$S(x_u, x_o) = z \; f_S \sum_{x_i = x_u}^{x_o} q_0(x_i) \; x_i^2 \; \Delta x_i$$

Die gleiche Rechnung führen wir für das Volumen $V(x_u, x_o)$ durch

$$V(x_u, x_o) = z \; f_V \sum_{x_i = x_u}^{x_o} q_0(x_i) \; x_i^3 \; \Delta x_i$$

Dann ergibt sich die volumenbezogene Oberfläche der Klassen von x_u bis x_o gemäß 9.109 zu

$$S_V(x_u, x_o) = 6 f_H \; \frac{\sum\limits_{x_i = x_u}^{x_o} x_i^2 \; q_0(x_i) \; \Delta x_i}{\sum\limits_{x_i = x_u}^{x_o} x_i^3 \; q_0(x_i) \; \Delta x_i} \qquad \text{mit} \qquad 6 f_H = \frac{f_S}{f_V} \tag{9.111}$$

Für $\Delta x_i \to 0$ werden aus den Summen Integrale

$$S_V(x_u, x_o) = 6 f_H \; \frac{\int\limits_{x_u}^{x_o} x^2 \; q_0(x) \; \mathrm{d}x}{\int\limits_{x_u}^{x_o} x^3 \; q_0(x) \; \mathrm{d}x} \tag{9.112}$$

Darin sind die unbekannten Proportionalitätsfaktoren von Oberfläche und Volumen in dem Faktor $6f_H$ zusammengefasst. Der nach HAROLD HEYWOOD (1925–1973) benannte Faktor f_H kennzeichnet die Partikelform im Vergleich zu einer Kugel; für diese wird der Heywoodfaktor zu eins. Für eine Kugel gilt

$$S_V = \frac{\pi x^2}{\frac{\pi}{6} x^3} = \frac{6}{x}$$

mit x als dem Durchmesser der Kugel. Dasselbe gilt für einen Würfel, wenn x die Kantenlänge bedeutet. Ein Kreiszylinder mit $d = h$ liefert $f_H = 4/3$. Für unregelmäßig geformte Partikel gilt $f_H > 1$, wenn als Partikelmerkmal der Äquivalentdurchmesser der Kugel gleichen Volumens gewählt wird. Außer dem Heywoodfaktor gibt es zahlreiche andere Formfaktoren. Kennen wir die Heywoodfaktoren zweier Proben nicht (was die Regel ist), dürfen wir ihre nach 9.112 berechneten spezifischen Oberflächen nur bei annähernd gleicher Partikelform miteinander vergleichen. Das ist beispielsweise bei zwei verschiedenen Chargen aus einem Produktionsgang oft gegeben.

Unter Verwendung der Momentenschreibweise nach 3.78 auf Seite 44 nimmt 9.112 folgende Form an

$$S_V(x_u, x_o) = 6f_H \, \frac{M_{2,0}(x_u, x_o)}{M_{3,0}(x_u, x_o)} \tag{9.113}$$

Der Term $M_{2,0}(x_u, x_o)$ wird gelesen als *zweites unvollständiges Moment der Anzahlverteilung*, unvollständig deshalb, weil sich der Bereich von x nicht über die gesamte Verteilung erstreckt.

Wir erweitern 9.113 mit $M_{3,0}$, dem dritten vollständigen Moment der Anzahlverteilung, einer Konstanten

$$\begin{aligned}
S_V(x_u, x_o) &= 6f_H \, \frac{M_{2,0}(x_u, x_o)}{M_{3,0}} \, \frac{M_{3,0}}{M_{3,0}(x_u, x_o)} \\
&= 6f_H \, \frac{\int_{x_u}^{x_o} x^2 \, q_0(x) \, \mathrm{d}x}{\int_{x_{min}}^{x_{max}} x^3 \, q_0(x) \, \mathrm{d}x} \, \frac{\int_{x_{min}}^{x_{max}} x^3 \, q_0(x) \, \mathrm{d}x}{\int_{x_u}^{x_o} x^3 \, q_0(x) \, \mathrm{d}x} \\
&= 6f_H \, \frac{\int_{x_u}^{x_o} x^2 \, q_0(x) \, \mathrm{d}x}{\int_{x_{min}}^{x_{max}} x^3 \, q_0(x) \, \mathrm{d}x} \, \frac{1}{Q_3(x_o) - Q_3(x_u)} \qquad (9.114) \\
&\text{mit} \qquad Q_3(x) = \frac{\int_{x_{min}}^{x} \xi^3 \, q_0(\xi) \, \mathrm{d}\xi}{\int_{x_{min}}^{x_{max}} \xi^3 \, q_0(\xi) \, \mathrm{d}\xi}
\end{aligned}$$

Gegenüber 9.112 hat sich nur die Bezugsgröße geändert. Für $x_u = x_{min}$ und $x_o = x_{max}$ vereinfacht sich obige Gleichung zu 9.112. Liegen x_u nahe bei x_{min}

und x_o nahe bei x_{max}, so ist $Q_3(x_o) - Q_3(x_u) \approx 1$, und der Fehler, den wir begehen, indem wir die Oberfläche eines Ausschnitts aus der Verteilung auf das Gesamtvolumen der Verteilung beziehen, ist vernachlässigbar, insbesondere in Anbetracht der physikalischen Unsicherheiten bei der Definition der Oberfläche.

In 9.112 kommt die Dichtefunktion der Anzahlverteilung $q_0(x)$ vor. Wir wissen jedoch aus Abschnitt 9.8 *Umrechnung verschiedener Mengenarten* auf Seite 173, dass es auch Volumen- oder Massenverteilungen $q_3(x)$ (und weitere) gibt. Müssen wir von einer solchen ausgehen, können wir entweder

- mit Hilfe von 9.107 auf Seite 176 die Formel 9.112 auf die Volumenverteilung umrechnen und die gegebene Volumenverteilung einsetzen oder
- die gegebene Volumenverteilung mit Hilfe von 9.107 in eine Anzahlverteilung umrechnen und dann in 9.112 einsetzen.

Beide Wege führen zu demselben Ergebnis. Mit 9.107 wird aus 9.112 und 9.113

$$S_V(x_u, x_o) = 6 f_H \frac{\int_{x_u}^{x_o} x^{-1}\, q_3(x)\, \mathrm{d}x}{\int_{x_u}^{x_o} q_3(x)\, \mathrm{d}x} = 6 f_H \frac{M_{-1,3}(x_u, x_o)}{M_{0,3}(x_u, x_o)} \tag{9.115}$$

Für $x_u = x_{min}$ und $x_o = x_{max}$ werden aus den unvollständigen Momenten vollständige Momente.

Bei einigen Verteilungstypen lässt sich die spezifische Oberfläche aus den Parametern der Verteilung berechnen. Aus der **Potenzverteilung** in der Form von 5.74 auf Seite 97

$$Q_3(x) = \begin{cases} (\frac{x}{x_{max}})^n & \text{für} \quad x \le x_{max} \quad n > 0 \\ 1 & \text{für} \quad x > x_{max} \end{cases}$$

mit $Q_3(x)$ als der Summenfunktion der Volumenverteilung erhalten wir durch Differenzieren nach x die Dichtefunktion der Volumenverteilung

$$q_3(x) = \frac{n}{x_{max}} (\frac{x}{x_{max}})^{n-1} \qquad \text{für} \qquad x \le x_{max} \quad n > 0$$

Mit 9.107 ergibt sich daraus die Dichtefunktion der Anzahlverteilung

$$\begin{aligned} q_0(x) &= M_{3,0}\, q_3(x)\, x^{-3} = M_{3,0} \frac{1}{x^3} \frac{n}{x_{max}} (\frac{x}{x_{max}})^{n-1} \\ &= M_{3,0} \frac{n}{x_{max}^n} x^{n-4} = \text{const} \cdot x^{n-4} \end{aligned}$$

eingesetzt in 9.112 folgt

$$\begin{aligned} S_V(x_u, x_o) &= 6 f_H \frac{\int_{x_u}^{x_o} x^2 x^{n-4}\, \mathrm{d}x}{\int_{x_u}^{x_o} x^3 x^{n-4}\, \mathrm{d}x} = 6 f_H \frac{\int_{x_u}^{x_o} x^{n-2}\, \mathrm{d}x}{\int_{x_u}^{x_o} x^{n-1}\, \mathrm{d}x} \\ &= 6 f_H \frac{\left| \frac{x^{n-1}}{n-1} \right|_{x_u}^{x_o}}{\left| \frac{x^n}{n} \right|_{x_u}^{x_o}} = 6 f_H \frac{n}{n-1} \frac{x_o^{n-1} - x_u^{n-1}}{x_o^n - x_u^n} \end{aligned} \tag{9.116}$$

Aus 9.115 folgt erwartungsgemäß dasselbe Ergebnis. Für $x_u = 0$ und $x_o = x_{max}$ wird aus 9.116

$$S_V(0, x_{max}) = 6 f_H \, \frac{n}{n-1} \, \frac{1}{x_{max}} \tag{9.117}$$

Gleichung 9.116 gilt nur für Potenzverteilungen mit $n \neq 1$. Der Fall $n = 1$ erfordert wegen der Integrationsregeln eine Sonderbehandlung. Mit dem Ansatz von 9.116 ergibt sich für $n = 1$

$$S_V(x_u, x_o) = 6 f_H \, \frac{\int_{x_u}^{x_o} x^{-1} \, \mathrm{d}x}{\int_{x_u}^{x_o} \mathrm{d}x} = 6 f_H \, \frac{\left| \ln x \right|_{x_u}^{x_o}}{\left| x \right|_{x_u}^{x_o}} = 6 f_H \, \frac{\ln(x_o) - \ln(x_u)}{x_o - x_u} \tag{9.118}$$

Hier dürfen wir zwar x_o bis x_{max} ausdehnen, aber wegen des Logarithmus nicht x_u bis null.

Die Einschränkungen bezüglich der Extrapolation bis $x_u = 0$ gehen noch weiter. Aus 9.107 folgt durch Integration und Einsetzen der Dichtefunktion $q_3(x)$ der Potenz-Volumenverteilung

$$\begin{aligned} Q_0(x_u, x_o) &= M_{3,0} \int_{x_u}^{x_o} \frac{1}{x^3} \, \frac{n}{x_{max}} \left(\frac{x}{x_{max}} \right)^{n-1} \mathrm{d}x = \text{const} \int_{x_u}^{x_o} x^{n-4} \, \mathrm{d}x \\ &= \text{const} \left| \frac{x^{n-3}}{n-3} \right|_{x_u}^{x_o} = \frac{\text{const}}{n-3} \left(x_o^{n-3} - x_u^{n-3} \right) \end{aligned}$$

Der Exponent wird für $n < 3$ negativ. Einsetzen von null für x_u ergäbe eine Division durch null. Für $n < 3$ dürfen wir daher nicht bis $x_u = 0$ extrapolieren. Der Fall $n = 3$ wirft gleichermaßen Fragen auf. Vereinfachend lässt sich sagen, dass man bei Verwendung einer Potenzverteilung stets eine realistische untere Partikelgröße x_u verwenden und nicht bis null extrapolieren sollte.

Beispiel 9.2 : Gegeben sei ein Schüttgut, dessen Summenfunktion der Volumenverteilung $Q_3(x)$ sich gut durch eine Potenzverteilung mit $x_o = x_{max} = 8$ mm und $n = 4,0$ wiedergeben lasse. Praktisch gesehen ist das eine ziemlich schmale Verteilung. Mit diesen Werten erhalten wir aus 9.117

$$S_V = 6 f_H \, \frac{4,0}{3,0} \, \frac{1}{8,00 \text{ mm}} = f_H \cdot 1,00 \text{ mm}^{-1}$$

Da $n > 3$ ist, dürfen wir am feinen Ende die Verteilung bis $x_u = 0$ extrapolieren. Mit einem Heywoodfaktor $f_H = 1,5$ ergibt sich eine volumenbezogene Oberfläche von $S_V = 1,50$ mm^{-1}, daraus mit einer Massendichte $\rho = 2,6$ g/cm^3 eine massenbezogene Oberfläche $S_m = 0,58$ cm^2/g.

Ende Beispiel

Wir untersuchen die volumenbezogene Oberfläche einer Potenzverteilung in Abhängigkeit von dem Exponenten n noch etwas und gehen dazu von 9.116 aus. Die Gleichung gilt wie die ganze Potenzverteilung nur für $n > 0$ und wegen der

Integrationsregeln nur für $n \neq 1$. Für die obere Grenze setzen wir $x_o = x_{max}$ und für die untere $x_u = a\,x_{max}$ mit $0 < a < 1$. Ein realistischer Wert für a ist $0,01$. Die Verteilung aus dem vorstehenden Beispiel würde sich damit von 0,08 bis 8,0 mm erstrecken und durch eine Siebanalyse zu ermitteln sein. Aus 9.116 erhalten wir mit unseren Annahmen

$$\frac{S_V(a\,x_{max}, x_{max})\;x_{max}}{6f_H} = \frac{n}{n-1}\;\frac{1-a^{n-1}}{1-a^n} \qquad \text{mit} \qquad 0 < a < 1 \tag{9.119}$$

In Abbildung 9.5 ist $y = S_V\;x_{max}/6f_H$ für drei Werte von a in Abhängigkeit von dem Exponenten n aufgetragen. Aus dem Diagramm ist zu ersehen, dass für $n > 2$ die durch a bestimmte untere Grenze x_u praktisch keine Rolle mehr spielt. Der Einfluss von a wird um so größer, je kleiner der Exponent n wird. Mit anderen Worten: Je breiter die Verteilung, desto kritischer ist die Wahl der unteren Grenze x_{min}. Der Einfluss der oberen Grenze x_{max} dagegen ist unabhängig von der Breite der Verteilung.

Im Fall der **Weibull-** oder **RRSB-Verteilung** gestaltet sich die Rechnung schwieriger. Wir gehen aus von 9.115, setzen 5.88 von Seite 100 ein und erhalten

$$S_V(x_u, x_o) = 6f_H\;\frac{\displaystyle\int_{x_u}^{x_o} \frac{1}{x}\,\frac{n}{x'}\,\Big(\frac{x}{x'}\Big)^{n-1}\;\exp\Big(-(\frac{x}{x'})^n\Big)\;\mathrm{d}x}{\displaystyle\int_{x_u}^{x_o} \frac{n}{x'}\,\Big(\frac{x}{x'}\Big)^{n-1}\;\exp\Big(-(\frac{x}{x'})^n\Big)\;\mathrm{d}x} \tag{9.120}$$

Der konstante Faktor n/x' kürzt sich heraus. Wir substituieren

$$t = \Big(\frac{x}{x'}\Big)^n \qquad \text{daraus} \qquad \frac{x}{x'} = t^{1/n} \qquad \text{und} \qquad dx = \frac{x'}{n}\,\Big(\frac{x}{x'}\Big)^{1-n}\,\mathrm{d}t$$

und kommen so auf

$$S_V(t_u, t_o) = \frac{6f_H}{x'}\;\frac{\displaystyle\int_{t_u}^{t_o} t^{-1/n}\;\exp(-t)\;\mathrm{d}t}{\displaystyle\int_{t_u}^{t_o} \exp(-t)\;\mathrm{d}t}$$

Wir setzen ferner (keine Variablensubstitution)

$$a = \frac{n-1}{n} \qquad \text{daraus} \qquad \frac{1}{n} = 1 - a$$

und erhalten

$$\begin{aligned} S_V(t_u, t_o) &= \frac{6f_H}{x'}\;\frac{\displaystyle\int_{t_u}^{t_o} t^{a-1}\;\exp(-t)\;\mathrm{d}t}{\displaystyle\int_{t_u}^{t_o} \exp(-t)\;\mathrm{d}t} \\ &= \frac{6f_H}{x'}\;\frac{\gamma(a, t_o) - \gamma(a, t_u)}{\exp(-t_u) - \exp(-t_o)} \end{aligned} \tag{9.121}$$

worin $\gamma(a,t)$ die *unvollständige Gammafunktion der oberen Grenze* t (E: lower incomplete gamma function) mit dem Parameter a ist ($t \geq 0$, $a > 0$, daraus

$n > 1$), bei M. Abramowitz und I. A. Stegun[13] die Formel Nr. 6.5.2. Wir haben bereits mit 5.62 auf Seite 94 die vollständige Gammafunktion $\Gamma(\alpha)$ kennen gelernt, bei der über einen Bereich von 0 bis $+\infty$ integriert wird. Üblicherweise wird die unvollständige Gammafunktion auf $\Gamma(a)$ normiert dargestellt, sodass sie für $t \to \infty$ den Wert 1 annimmt.

Machen wir die Ersetzungen rückgängig, gelangen wir von 9.121 zu

$$S_V(x_u, x_o) = \frac{6 f_H}{x'} \; \frac{\gamma\Big(\frac{n-1}{n}, (\frac{x_o}{x'})^n\Big) - \gamma\Big(\frac{n-1}{n}, (\frac{x_u}{x'})^n\Big)}{\exp\left(-(\frac{x_u}{x'})^n\right) - \exp\left(-(\frac{x_o}{x'})^n\right)} \tag{9.122}$$

In Einklang mit DIN 66145[14] werde vereinbart

$$x_u = x(Q_3 = 0,001) \qquad x_o = x(Q_3 = 0,999)$$

Dann ist mit 5.89 von Seite 100

$$t_u = \Big(\frac{x_u}{x'}\Big)^n \;=\; -\ln\, 0,999 = 0,0010005 \qquad \text{für} \qquad Q_3(x) = 0,001$$

$$t_o = \Big(\frac{x_o}{x'}\Big)^n \;=\; -\ln\, 0,001 = 6,9077553 \qquad \text{für} \qquad Q_3(x) = 0,999$$

Damit wird aus 9.122

$$S_V(x_u, x_o) = \frac{6 f_H}{x'} \; \frac{\gamma\Big(\frac{n-1}{n},\; 6,9078\Big) - \gamma\Big(\frac{n-1}{n},\; 0,0010\Big)}{0,999 - 0,001} \tag{9.123}$$

Wie kommen wir nun zu Zahlenwerten? Für die unvollständige Gammafunktion gibt es die Reihenentwicklung[15]

$$\gamma(a, x) = \int_0^x t^{a-1}\, \exp(-t)\, \mathrm{d}t = x^a\, \exp(-x)\, \sum_{n=0}^{\infty} \frac{\Gamma(a)}{\Gamma(a+1+n)}\, x^n \tag{9.124}$$

die sich mit Computerhilfe auswerten lässt. Wir dürfen hoffen, dass andere die Programmierarbeit bereits erledigt haben. Das Zeichenprogramm Gnuplot und viele Statistikprogramme kennen die unvollständige Gammafunktion.

Beispiel 9.3 :

Gegeben sei ein Schüttgut, dessen Summenfunktion der Volumenverteilung $Q_3(x)$ sich gut durch eine RRSB- beziehungsweise Weibull-Verteilung mit den Parametern $x' = 8$ mm und $n = 4,0$ wiedergeben lasse. Mit diesen Werten erhalten wir aus 9.123

$$S_V = \frac{6 f_H}{8\ \mathrm{mm}}\; \frac{1}{0,999 - 0,001}\; \Big(\gamma\Big(\frac{3}{4},\; 6,9078\Big) - \gamma\Big(\frac{3}{4},\; 0,0010\Big)\Big)$$

[13] siehe Anhang

[14] Die Norm DIN 4190 von 1966 berücksichtigte den Bereich von $Q_3 = 0,007$ bis $Q_3 = 0,999$.

[15] siehe *Numerical Recipes in C* oder die englische Wikipedia

Aus Abbildung 9.6 entnehmen wir, dass die Werte der normierten unvollständigen Gammafunktion für diese Zahlenwerte ausreichend genau zu $1,0$ beziehungsweise 0 angenommen werden dürfen und erhalten in guter Näherung

$$S_V \approx \frac{6 f_H}{8\ \mathrm{mm}}\,\Gamma(3/4) = f_H \cdot 0,92\ \mathrm{mm}^{-1}$$

Beim Vergleich mit dem Beispiel zur Potenzverteilung verwundert zunächst, dass der Streuungsparameter n nicht im Ergebnis erscheint. Er steckt aber in der unvollständigen Gammafunktion, siehe Abbildung 10.3 auf Seite 193. Im vorliegenden Fall (großes n) ist der Parameter x' zugleich der Durchmesser einer Kugel mit derselben volumenbezogenen Oberfläche wie die Verteilung.

Ende Beispiel

9.9.3 Folgerungen

Zur Oberfläche trägt der Anteil der Partikel am feinen Ende der Verteilung viel bei, zum Volumen oder zur Masse jedoch wenig. Infolgedessen hängt die volumen- oder massenbezogene Oberfläche stark vom Verlauf der Verteilung am feinen Ende ab. Muss man mangels Messpunkten dort extrapolieren, so ist das nur vertretbar, wenn die Verteilung schmal ist.

Lassen sich gemessene Partikelgrößenverteilungen durch eine der genannten Verteilungen nur stückweise befriedigend annähern, so kann man die vorstehenden Gleichungen entsprechend erweitern. Man hat dann die Oberflächen aller Abschnitte zu addieren, ebenso die Volumina, und daraus den Quotienten zu bilden. In solchen Fällen wird man jedoch heute auf parameterfreie numerische Verfahren ausweichen, die von der Definition 9.111 ausgehen. Auch die in DIN 66143 bis 66145 vorgeschlagenen grafischen Verfahren sind überholt.

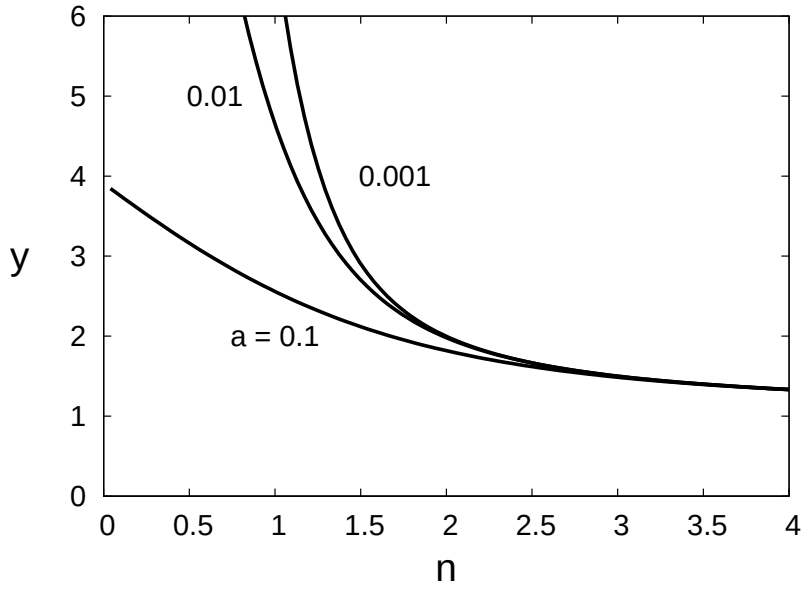


Abb. 9.5: Einflussfaktoren der volumenbezogenen Oberfläche der Potenzverteilung (Erklärung siehe Text)

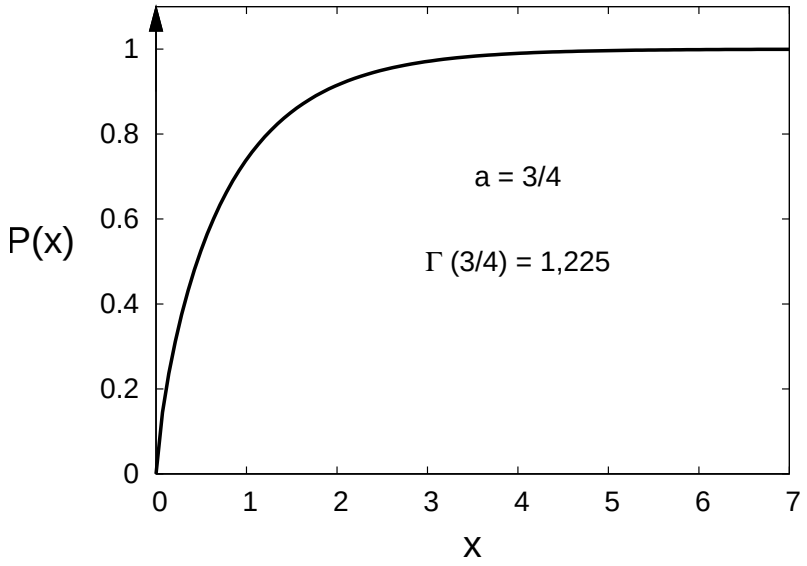


Abb. 9.6: Normierte unvollständige Gammafunktion der oberen Grenze

Kapitel 10

Statistisches Rechnen auf dem Computer

In der Statistik – wenn man sie nicht nur theoretisch betreibt – hat man ständig mit Zahlen und Diagrammen zu tun. Das Ringen mit der Zahlenflut ist heute dank der Computer nicht mehr so mühsam wie früher. Sogar auf einem PC oder Mac lassen sich viele Auswertungen in kurzer Zeit bewältigen. Wir sehen uns einige Vertreter von häufig gebrauchten Werkzeugen an, erfahren, wozu wir sie einsetzen können, und gehen ein paar Beispiele durch. Die folgenden Abschnitte ersetzen nicht die Dokumentation zu den Werkzeugen.

10.1 Grafische Darstellungen mittels Gnuplot

Das bewährte Programm **Gnuplot** dient dem Zeichnen von Diagrammen nach Wertetabellen oder Funktionsgleichungen. Es eignet sich nicht für künstlerische, geometrische oder technische Zeichnungen, für Schaltpläne oder allgemein zur Bildbearbeitung. Sein Heimathafen ist `http://www.gnuplot.info/`, wo außer dem Programm auch Beispiele und Dokumentation zu finden sind. Gnuplot ist im Quellcode erhältlich sowie compiliert für die Intel-386-Architektur und weitere. Es ist Bestandteil vieler Linux-Distributionen.

Gnuplot beherrscht cartesische und polare Koordinaten, zwei- und dreidimensionale Diagramme, linear oder logarithmisch geteilte Achsen sowie zahlreiche Ausgabeformate. In begrenztem Umfang kann Gnuplot auch rechnen. Es lässt sich im Dialog benutzen; gebräuchlicher ist jedoch die Steuerung durch ein Skript. Die Diagramme des vorliegenden Textes wurden mit Gnuplot erstellt. Das Werkzeug ist einfach zu handhaben, allerdings nicht so selbsterklärend wie ein Bleistift.

Um ein Diagramm wie Abbildung 8.1 auf Seite 132 zu zeichnen – ein bewusst einfaches Beispiel – brauchen wir zweierlei

- eine Datei mit den Messwerten (datafile) und
- eine Datei mit den Anweisungen an Gnuplot (plotscript).

Zeichnen wir eine Kurve nach einer Funktionsgleichung, entfällt die Datei mit den Messwerten. Das Ergebnis ist in den meisten Fällen eine Datei in einem gängi-

gen Grafikformat wie PostScript, die mit einem anderen Programm (Betrachter, Viewer) auf den Bildschirm oder einen Drucker ausgegeben wird.

Das **Datafile** trägt einen beliebigen Namen und muss bestimmten formalen Vorschriften genügen. Sehen wir uns das Datafile zu Abbildung 8.1 an:

```
# data_81, 2. und 3. Wurzel

1.00  1.000  1.000
2.00  1.414  1.260
3.00  1.732  1.442
4.00  2.000  1.587
5.00  2.236  1.710
6.00  2.449  1.817
7.00  2.646  1.913
8.00  2.828  2.000
9.00  3.000  2.080
```

Mit einem Doppelkreuz eingeleitete Zeilen sind Kommentar und werden von Gnuplot nicht beachtet. Leerzeilen spielen eine Rolle. Eine einzelne Leerzeile unterbricht einen Kurvenzug, doppelte Leerzeilen trennen Datensätze voneinander so, als ob sie in getrennten Dateien lägen. Jede Zeile enthält die zu einem Punkt gehörenden Zahlenwerte, wobei die Spalten im Plotskript ausgewählt werden. Die erste Spalte enthält üblicherweise die Abszissenwerte (x-Werte), die folgenden Spalten Ordinatenwerte (y-Werte). Die Spalten sind durch Leerzeichen oder Tabs getrennt. Dezimalzeichen ist der Punkt.

Das **Plotskript** ist Text im US-ASCII-Zeichensatz und mit jedem einfachen Texteditor abzufassen. Textprozessoren wie Microsoft Word fügen von sich aus gelegentlich Zeichen ein und sind daher weniger geeignet. Man hat gewisse Freiheiten in der Gestaltung; das folgende Plotskript zu Abbildung 8.1 ist nur als Vorschlag anzusehen:

```
# plot_81 fuer Statistikskriptum

reset

set terminal postscript enhanced 24
set output "abbildung.ps"

set size 1,1
set lmargin 6

set xrange [0:10]
set yrange [0:4]
unset tics

set style line 1 linetype 1 linecolor rgb "black" linewidth 1
```

```
set style line 2 linetype 1 linecolor rgb "black" linewidth 2
set style line 3 linetype 1 linecolor rgb "black" linewidth 6

set style arrow 1 head filled size screen 0.03,15 linestyle 2

# set grid linestyle 1
set xzeroaxis linestyle 2
set yzeroaxis linestyle 2

set arrow 1 from first 8.0,0 to 10.0,0 arrowstyle 1
set arrow 2 from first 0,3.0 to 0,4.0 arrowstyle 1

set xlabel "x" font ",32"
set ylabel "y" norotate font ",32"
set nokey

set label 1 "y = f(x)" at 6.0,1.5 center

plot "data_81" using 1:2 smooth bezier linestyle 3
```

Gehen wir das Plotskript Zeile für Zeile durch:

- Das Doppelkreuz leitet Kommentarzeilen beliebigen Inhalts ein. Damit lassen sich auch Kommandozeilen vorübergehend außer Betrieb nehmen, die man nicht löschen möchte. Leerzeilen im Plotskript dienen der Gliederung des Textes für den Leser und interessieren Gnuplot nicht.

- Als erstes Gnuplot-Kommando folgt ein `reset`, das sämtliche Einstellungen auf die Grundwerte zurücksetzt, schafft klare Verhältnisse.

- `set terminal` legt das Ausgabeformat fest und damit zugleich eine Reihe von Möglichkeiten, die uns zur Verfügung stehen oder nicht. Wir haben uns für `postscript enhanced` entschieden, weil dieser Treiber einen Symbolfont mit griechischen Buchstaben kennt und ein für unsere Zwecke geeignetes Ausgabeformat (PostScript) liefert. Der Zahlenwert `24` legt die Schriftgröße in Punkten fest. Davor könnten wir noch einen Fontnamen wie `"LucidaSansRegular"` angeben. Natürlich muss der Font verfügbar sein.

- `set output` benennt die Ausgabedatei. Sie wird überschrieben beziehungsweise im Arbeitsverzeichnis neu angelegt.

- `set size 1,1` legt das Seitenverhältnis fest, hier auf den Vorgabewert, daher überflüssig.

- `set lmargin 6` legt den linken Rand des Diagramms auf 6-mal Zeichenbreite fest. Nicht immer erforderlich, hier wegen der Breite der Beschriftung.

- `set xrange [0:10]` Falls wir keine automatische Skalierung wünschen, wird hiermit der x-Bereich bestimmt. Entsprechend `yrange`.
- `unset tics` unterbindet Skalenstriche auf den Achsen, so man will.
- `set style line 1 linetype 1 linecolor rgb "black" linewidth 1` definiert einen Linienstil nach Typ, Farbe und Breite. Der Stil bekommt von uns eine laufende Nummer, auf die wir uns hernach beziehen. Ebenso für Linienstil Nr. 2 und 3.
- `set style arrow 1 head filled size screen 0.03,15 linestyle 2` definiert einen Stil für Pfeile.
- `# set grid linestyle 1` ist auskommentiert, da wir kein Gitternetz haben wollen. Es würde im Linienstil Nr. 1 (dünne Linien) gezeichnet.
- `set xzeroaxis linestyle 2` bestimmt, dass die x-Achse im Linienstil Nr. 2 (mittelstarke Linien) gezeichnet werden soll. Entsprechend die y-Achse.
- `set arrow 1 from first 8.0,0 to 10.0,0 arrowstyle 1` besagt, dass Pfeil Nr. 1 von $x = 8.0, y = 0$ bis $x = 10.0, y = 0$ zu zeichnen ist, also ans Ende der x-Achse, und zwar im Pfeilstil Nr. 1 (einen anderen haben wir auch nicht definiert). Entsprechend Pfeil Nr. 2 am Ende der y-Achse. Die Koordinaten werden in dem Koordinatensystem des Diagramms angegeben.
- `set xlabel "x" font ",32"` bestimmt die Beschriftung der x-Achse mit dem Buchstaben x. Bei der Angabe des Fonts und der Schriftgröße muss das Komma zwischen beiden stehen, auch wenn wir keinen besonderen Font auswählen. Entsprechend für die y-Achse. `norotate` verhindert eine Drehung der Beschriftung, je nach Wunsch.
- `set nokey` verhindert das Einblenden einer Legende zum Diagramm.
- `set label 1 "y = f(x)" at 6.0,1.5 center` definiert eine beliebige Beschriftung im Diagramm. Die Beschriftung wird zu den angegebenen Koordinaten zentriert.
- Die letzte Zeile `plot "data_81" using 1:2 smooth bezier linestyle 3` zeichnet das Diagramm. Auf das `plot`-Kommando folgen das Datafile und die Angabe, welche Spalten auszuwählen sind. `smooth` verlangt einen durchgehenden Kurvenzug, und zwar in diesem Fall eine Bezier-Approximation. Die Kurve soll im Linienstil Nr. 3 (kräftige Linien) gezeichnet werden.

In vielen Zeilen könnte man andere oder zusätzliche Wünsche äußern. Wir hoffen, dass der wesentliche Aufbau klar geworden ist. Der Aufruf von Gnuplot lautet dann

```
$ gnuplot plot_81
```

Sekundenbruchteile später finden wir die Ausgabedatei `abb_81.ps` in unserem Arbeitsverzeichnis und schauen sie uns mit einem PostScript-Betrachter unserer Wahl an

```
$ gv abbildung.ps
```

siehe Abbildung 8.1 auf Seite 132.

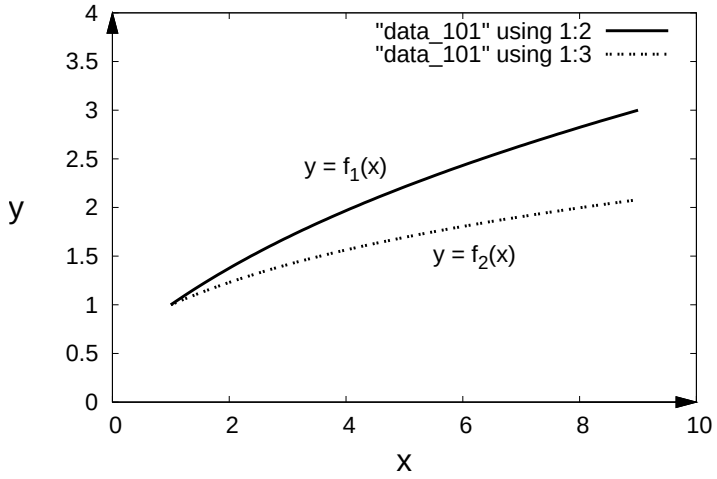


Abb. 10.1: Ausgabe von Gnuplot, 2. und 3. Wurzel

Um Abbildung 10.1 zu zeichnen, erweitern wir das Plotskript um folgende Änderungen:

- Wir kommentieren mit einem Doppelkreuz die Zeile `unset tics` aus, verlangen also Tics (Skalenstriche), ebenso die Zeile `set nokey`.
- Wir definieren einen vierten Linienstil ähnlich Linienstil 3, nur mit einem anderen Linientyp, beispielsweise `linetyp 7`. Was für ein Linientyp sich hinter einer Nummer verbirgt, ist eine Frage des Treibers.
- Wir verschieben das 1. Label und erzeugen ein 2. Label:

```
set label 1 " y = f_1(x)" at 4.0,2.4 center
set label 2 " y = f_2(x)" at 6.2,1.5 center
```

- Wir fügen an die `plot`-Zeile ein Komma und dann noch einmal alle Argumente an, jedoch mit der Anweisung `1:3`, um einen Kurvenzug aus der 1. und 3. Spalte des Datafiles zu erzeugen, und unserem neu geschaffenen Linienstil 4. Das Ganze bleibt eine einzige Zeile

```
plot "data_101" using 1:2 smooth bezier linestyle 3,
     "data_101" using 1:3 smooth bezier linestyle 4
```

Wollen wir eine formelmäßig vorliegende Funktion zeichen, kann Gnuplot einige Berechnungen selbst durchführen. Wir gehen wieder von obigem Plotskript zu Abbildung 8.1 aus und

- kommentieren die Zeilen für `xrange` und `yrange` aus, um Gnuplot das Skalieren zu überlassen, und infolgedessen auch die Zeilen mit `set arrow`,
- dito die Zeilen mit `unset tics` und `set label`,
- ersetzen die `plot`-Zeile durch

```
plot sin(x)/x linestyle 3
```

Das Ergebnis ist in Abbildung 10.2 zu bewundern. Gnuplot kennt auch die Gammafunktion und die normierte Normalverteilung.

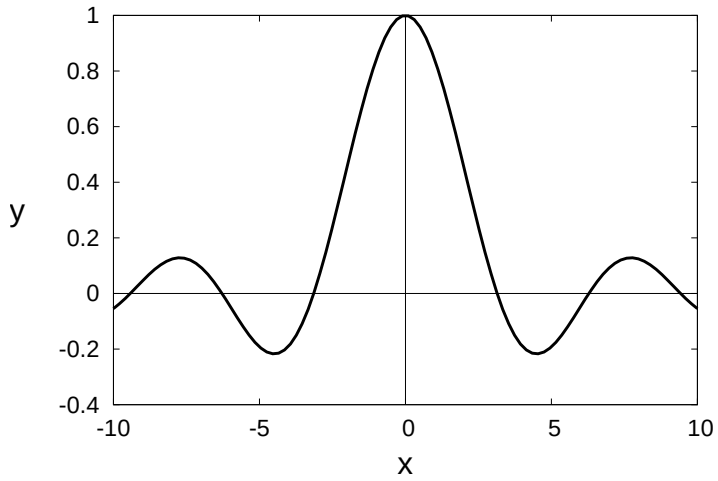


Abb. 10.2: Ausgabe von Gnuplot, $\sin(x)/x$

Um ein Diagramm der normierten unvollständigen Gammafunktion $\gamma(a, x)$ für mehrere Parameterwerte a zu zeichnen, gehen wir wieder vom Plotskript der Abbildung 8.1 aus und

- setzen den `xrange` auf $0 : 16$, den `yrange` auf $0 : 1.1$ und schieben die Pfeile an den Achsen an die richtigen Stellen,
- verlangen `tics`, vergrößern den linken Rand auf 9,
- setzen die Labels für die vier Parameterwerte, aber erst nachdem alles Andere fertig ist,
- plotten die Funktion für die Parameter $a = 0.5, 1.0, 3.0$ und 10.0 (einzeilig):

```
plot igamma(0.5,x) linestyle 3, igamma(1.0,x) linestyle 3,
    igamma(3.0,x) linestyle 3, igamma(10.0,x) linestyle 3
```

Abbildung 10.3 zeigt das Diagramm. Die Magie dahinter bleibt uns verborgen. Nach einem Aufruf von Gnuplot im Dialog hilft die Eingabe von `help igamma` etwas weiter. Ansonsten bleiben ein Blick in die Programmquellen und eine Suche im Netz nach näheren Auskünften.

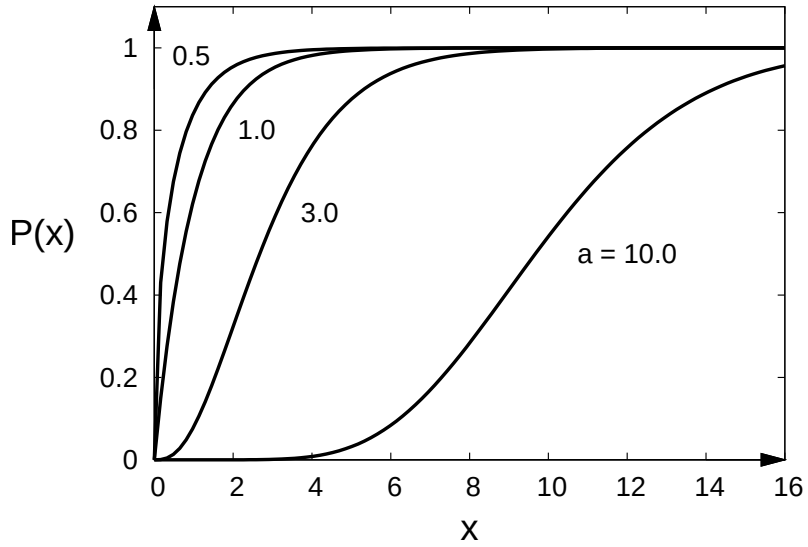


Abb. 10.3: Ausgabe von Gnuplot, normierte unvollständige Gammafunktion für verschiedene Parameter a

Zum Abschluss lassen wir Gnuplot eine lineare Regression rechnen und zeichnen, und zwar von den Daten von K. Pearson und A. Lee ausgehend. In Abbildung 8.3 auf Seite 135 ist das Ergebnis der beiden Wissenschaftler dargestellt; sie hatten für die Regressionsgerade die Werte $a = 0,516$ und $b = 33,73$ ermittelt.

Wir schenken uns hier die vorbereitenden Zeilen und wenden uns gleich den entscheidenden zu:

```
f(x) = a*x + b
fit f(x) "data_83P" using 1:2 via a,b
plot "data_83P" using 1:2 with points pointtype 7 pointsize 2,
                                                    f(x) linestyle 2
```

Die erste Zeile definiert eine Funktion $f(x)$. Die zweite Zeile passt die benutzerdefinierte Funktion $f(x)$ an eine Menge von Datenpunkten x, y an, deren Koordinaten in den Spalten 1 und 2 des Datafiles `data_83P` vorliegen. Dazu werden die beiden Parameter a und b berechnet, laut Manual nach dem Nonlinear-Least-Squares Algorithmus von Donald Marquardt und Kenneth Levenberg.[1] Die dritte Zeile veranlasst Gnuplot, die Datenpunkte und den Graphen der Funktion $f(x)$ in ein Diagramm zu zeichnen.

[1] siehe die deutsche oder englische Wikipedia unter *Levenberg-Marquardt algorithm* oder *Numerical Recipes in C* im Kapitel über nichtlineare Modelle.

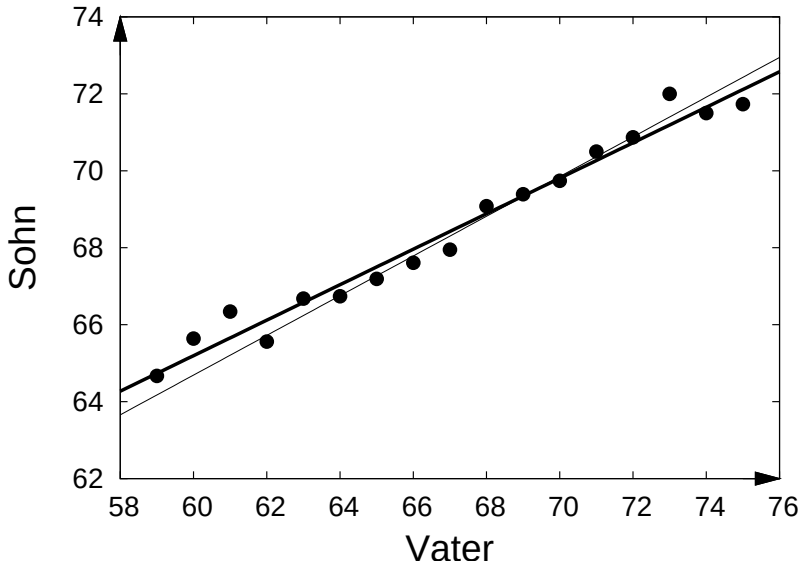


Abb. 10.4: Regressionsgerade der Ergebnisse von K. Pearson und A. Lee, berechnet und gezeichnet mittels Gnuplot (vgl. Abbildung 8.3)

Außer einer PostScript-Datei mit der Grafik gibt Gnuplot auf der Konsole (Textbildschirm oder -fenster, aus dem wir Gnuplot aufgerufen haben), eine Handvoll Werte aus, darunter $a = 0,461176$ und $b = 37,5241$ samt zugehörigen Standardabweichungen. Die Werte weichen von den originalen ab. Wir vertrauen sowohl K. Pearson wie Gnuplot und vermuten, dass der Unterschied von verschiedenen Algorithmen herrührt. Der Unterschied beleuchtet, dass es zu einem Satz von Messpunkten durchaus verschiedene Regressionsgeraden gibt, je nach verwendetem Algorithmus. Abbildung 10.4 zeigt das Diagramm. Die Regressionsgerade von K. Pearson und A. Lee ist dünn eingezeichnet.

Zum Umwandeln der PostScript-Datei in andere Grafikformate (eps, jpg, pdf, png, xfig ...) stehen Werkzeuge bereit. Alles Weitere entnehme man bitte der Dokumentation.

10.2 Tabellenkalkulation (Gnumeric)

Tabellenkalkulationen (E: spreadsheet) sind Werkzeuge für die interaktive Verarbeitung von numerischen und alphanumerischen Daten in Tabellenform. Seit 1979 mit VisiCalc für den Apple II die erste Tabellenkalkulation auf dem Markt erschien, haben sich die Werkzeuge zu einer Standardanwendung im Büro entwickelt und werden oft sogar als Ersatz für Datenbanken missbraucht. Als Vertreter einer Tabellenkalkulation sehen wir uns **Gnumeric** aus dem GNOME-Projekt an, das für eine Reihe von Rechnerarchitekturen verfügbar ist. Sein Heimathafen ist `http://projects.gnome.org/gnumeric/`.

Eine **Tabelle** besteht aus **Spalten** (E: column) und **Zeilen** (E: row). In den Kreuzungspunkten der Spalten und Zeilen befinden sich **Zellen** oder Felder (E: cell), die Werte, Text oder Formeln aufnehmen. Zellen kann ein Kommentar hin-

zugefügt werden. Eine Zelle kann auf andere Zellen verweisen. Formeln stammen aus einer Bibliothek oder sind vom Benutzer definiert. Beispielsweise ist es möglich, Funktionen der R-Statistik-Software aus einer Gnumeric-Tabelle aufzurufen. Wird der Wert in einer Zelle verändert, so können die Werte in abhängigen Zellen automatisch oder auf Anweisung hin nachgeführt werden. Auf diese Weise lassen sich einfach Fragestellungen der Art *Was wäre, wenn ...* durchspielen, eine Art numerischer Experimente.

Mit einer solchen Tabelle, auch **Arbeitsblatt** (E: sheet, worksheet) genannt, lassen sich zwei Variable darstellen. Ordnet man mehrere Arbeitsblätter in Art einer Kartei oder eines Buches (**Arbeitsbuch**, Arbeitsmappe, E: workbook) hintereinander an, verfügt man über drei Variable, gekennzeichnet durch die Nummern von Zeile, Spalte und Blatt. Mehrere Karteikästen lassen sich auf einen Regalboden in einem Regal stellen usw. Der Index einer Zelle wird so zu einem n-dimensionalen Vektor. Weil es jedoch immer schwieriger wird, sich die Zusammenhänge vorzustellen, belässt man es in der Regel bei einem zweidimensionalen Arbeitsblatt oder einem dreidimensionalen Arbeitsbuch. Als Vorgabe öffnet Gnumeric ein Arbeitsbuch mit drei Arbeitsblättern unter einem Dateinamen.

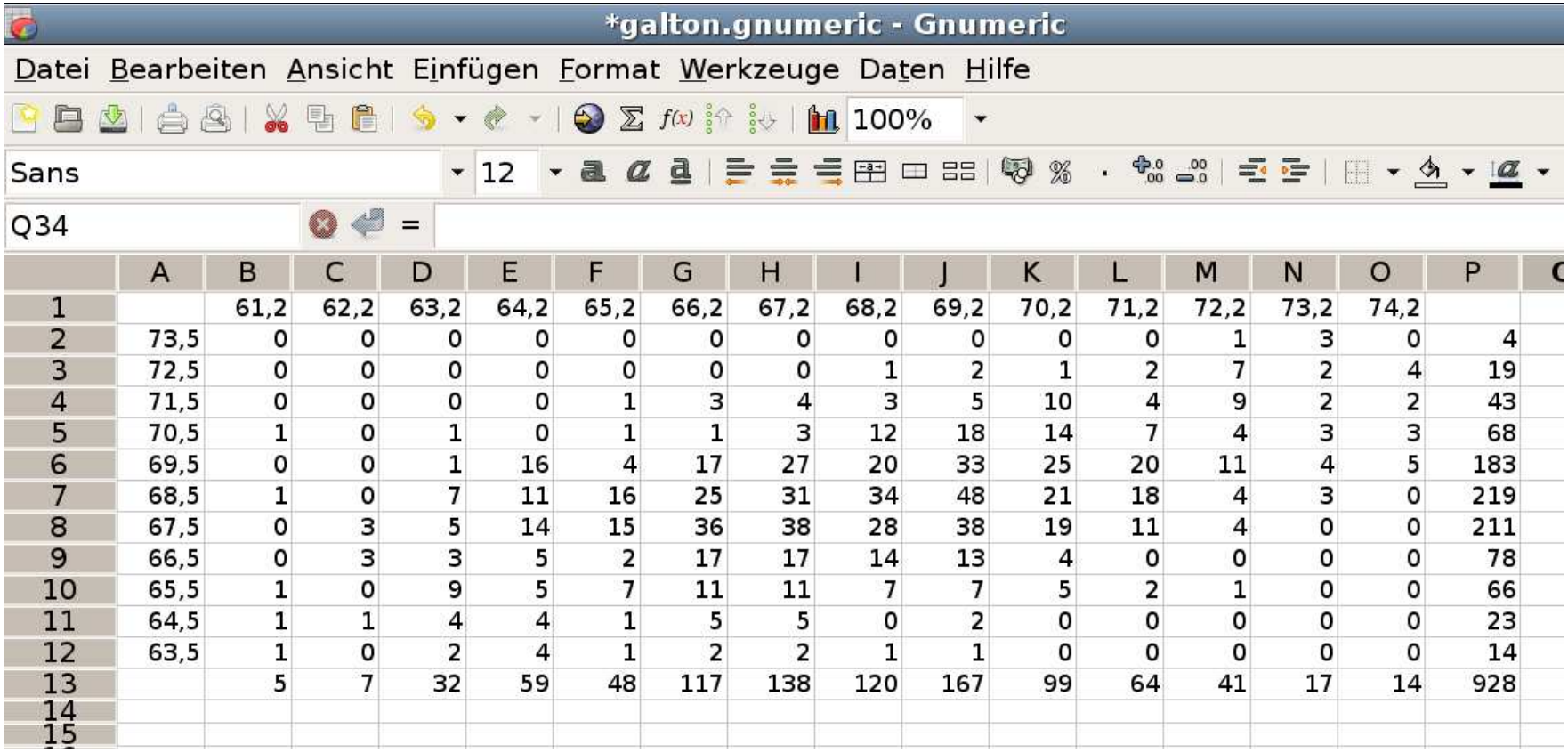
Die Anzahl der Zeilen und Spalten ist begrenzt. Gnumeric bringt als Vorgabe maximal 256 Spalten und 65000 Zeilen mit, die sich auf 16000 Spalten und 16 Mio. Zeilen erweitern lassen. Braucht man mehr, ist die Datei `gnumeric.h` zu editieren und das Programm erneut zu übersetzen. Irgendwann reichen dann der Arbeitsspeicher und die Geschwindigkeit von Festplatte und CPU nicht mehr aus. Der Umgang mit riesigen Datenmengen ist eher die Aufgabe von Datenbanken, deren mathematische Fähigkeiten dafür bescheidener sind als die von Tabellenkalkulationen. Häufig liefert auch eine Datenbank das vorverdichtete Rohmaterial für eine Tabellenkalkulation.

Sehen wir uns als Beispiel die Ergebnisse von Francis Galton an, die auch der Tabelle 10.1 auf Seite 204 zu Grunde liegen. Es geht um die Körpergröße von Kindern in Abhängigkeit von der Körpergröße ihrer Eltern.[2] Wir legen ein Verzeichnis für unsere Übung an, wechseln hinein und rufen Gnumeric auf. Es öffnet sich ein Fenster mit der linken oberen Ecke – etwa 12 Spalten und 30 Zeilen umfassend – des ersten Arbeitsblattes unseres Arbeitsbuches. Dann übertragen wir Galtons Tabelle in das Arbeitsblatt:

- Spalte A nimmt die Größen der Eltern auf, insgesamt 11 Werte.
- Zeile 1 nimmt die Größen der Kinder auf, insgesamt 14 Werte.
- In die Zellen kommen die Anzahlen der Beobachtungen (Besetzungszahlen), wobei für Galtons Pünktchen der Wert 0 einzusetzen ist. Zelle A1 bleibt leer.

Für die Bezeichnungen *Above* und *Below* extrapolieren wir die Skalen. Da die zugehörigen Zellen schwach besetzt sind, verursacht das keinen nennenswerten Fehler.

[2] F. Galton, *Natural Inheritance*, Macmillan, 1889, Table 11, Seite 208. Zu finden auf `http://galton.org/books/natural-inheritance/`



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1		61,2	62,2	63,2	64,2	65,2	66,2	67,2	68,2	69,2	70,2	71,2	72,2	73,2	74,2	
2	73,5	0	0	0	0	0	0	0	0	0	0	0	1	3	0	4
3	72,5	0	0	0	0	0	0	0	1	2	1	2	7	2	4	19
4	71,5	0	0	0	0	1	3	4	3	5	10	4	9	2	2	43
5	70,5	1	0	1	0	1	1	3	12	18	14	7	4	3	3	68
6	69,5	0	0	1	16	4	17	27	20	33	25	20	11	4	5	183
7	68,5	1	0	7	11	16	25	31	34	48	21	18	4	3	0	219
8	67,5	0	3	5	14	15	36	38	28	38	19	11	4	0	0	211
9	66,5	0	3	3	5	2	17	17	14	13	4	0	0	0	0	78
10	65,5	1	0	9	5	7	11	11	7	7	5	2	1	0	0	66
11	64,5	1	1	4	4	1	5	5	0	2	0	0	0	0	0	23
12	63,5	1	0	2	4	1	2	2	1	1	0	0	0	0	0	14
13		5	7	32	59	48	117	138	120	167	99	64	41	17	14	928
14																
15																

Abb. 10.5: Screenshot eines Arbeitsblattes der Tabellenkalkulation Gnumeric, obere Hälfte, mit einer Tabelle nach F. Galton, Körpergröße von Eltern und Kindern in inch, Spalte A = Eltern (untere Klassengrenzen), Zeile 1 = Kinder, Zellen = Besetzungszahlen, Spalte P = Zeilensummen, Zeile 13 = Spaltensummen, Dezimalzeichen ist das Komma (deutsche Umgebung)

Wir fahren den Cursor über die jeweilige Zelle, tippen den Wert ein, drücken die `<Return>`-Taste oder fahren gleich zur nächsten Zelle. Spaltenweises Eingeben geht am einfachsten. Bei Fehleingaben überschreiben. In Zelle H8 sollte beispielsweise der Wert 38 stehen. Haben wir die Zahlen im Kasten, klicken wir auf *Datei->Speichern*, um unsere Arbeit zu sichern.

Nun überlegen wir uns, was wir ausrechnen wollen. Als erstes bieten sich die Spalten- und Zeilensummen an, das sind die Anzahlen der jeweiligen Messwerte, von GALTON als *Totals* bezeichnet. Wir wählen Zelle P2 aus und geben in dem Feld oberhalb der Tabelle nach dem Gleichheitszeichen ein

```
=sum(B2:O2)
```

Das Gleichheitszeichen kennzeichnet den Zellinhalt als Formel. Der Doppelpunkt benennt einen zusammenhängenden Bereich von Zellen. Groß- und Kleinbuchstaben werden nicht unterschieden. Es gibt vereinfachte Wege für derartige Eingaben, siehe Manual. Nach Abschluss der Eingaben finden wir in der Zelle sofort den Wert 4. In dem Feld oberhalb lässt sich eine Formel jederzeit ändern (editieren). In Zelle P13 haben wir die Wahl zwischen der Zeilen- und der Spaltensumme, die Werte sind identisch gleich 928. Haben wir uns nicht vertippt, stimmen unsere Summen mit GALTONS *Totals* überein. Arbeit sichern.

Wir verlassen GALTONS Vorlage und gehen eigene Wege. Gnumeric bietet unter seinen rund 150 statistischen Funktionen `AVERAGE()` an, aber wir wollen nicht die durchschnittlichen Besetzungszahlen wissen, sondern die durchschnittlichen Körpergrößen in den Zeilen und Spalten. Dazu brauchen wir eine zweite Tabelle, die in den Zellen die Besetzungszahlen multipliziert mit den zugehörigen Größen der Kinder enthält. Die Tabelle könnten wir auf dem zweiten Arbeitsblatt anlegen, aber angesichts des geringen Umfangs der Tabelle ist es einfacher, die Zeilen 21 bis 33 des ersten Arbeitsblattes dafür zu verwenden. In Zelle H27 steht dann beispielsweise die Formel

```
=H1*H7
```

Wir bilden auch hier wie gehabt die Zeilen- und Spaltensummen. Die Gesamtsumme beträgt 63190,6. In Spalte Q bilden wir den Qotienten aus der Summe der Größen der Kinder (Spalte P der zweiten Tabelle) und der Summe der Besetzungszahlen beziehungsweise der Anzahl der Kinder (Spalte P der ersten Tabelle), das heißt die mittlere Größe der Kinder, die zu einer Elterngrößenklasse (Spalte A) gehören. In Zelle Q22 steht also

```
=P22/P2
```

In Tabelle 10.1 ist dieser Zusammenhang wiedergegeben, allerdings mit dem Unterschied, dass GALTON die Medianwerte verwendet, nicht die Mittelwerte.

Das Kopieren (copy) und Einfügen (paste) von Zeilen oder Spalten lässt sich über das Menü *Bearbeiten* bequem gestalten. Man selektiert mit der linken Maustaste den zu kopierenden Zellenbereich, klickt den Menüpunkt *Kopieren* an, selektiert den Zielbereich und klickt den Menüpunkt *Einfügen* an.

Ebenso lassen sich Formeln kopieren, wobei Gnumeric die Zellenindizes nachführen kann. Man klickt die Zelle mit der zu kopierenden Formel, verlässt bei gedrückter linker Maustaste die Zelle über die rechte, untere Ecke – wobei der Cursor sein Aussehen wechselt – und selektiert den Zielbereich. Nachprüfen, ob sich die Zellenindices in gewünschter Weise angepasst haben.

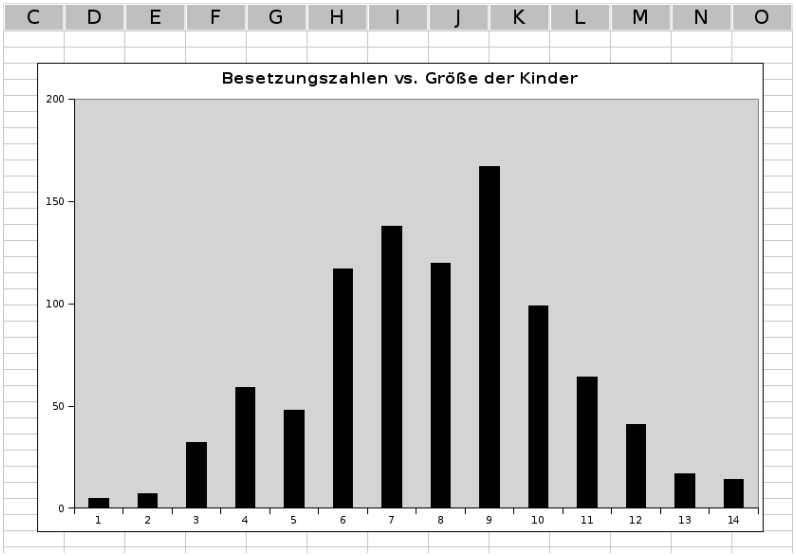


Abb. 10.6: Screenshot von Gnumeric mit Graph zur Tabelle 10.5

Nun soll Gnumeric ein Diagramm zeichnen. Das Werkzeug ist bei dieser Aufgabe nicht so vielseitig wie Gnuplot, aber für viele Zwecke reichen seine Fähigkeiten aus. Wir wollen die aufaddierten Besetzungszahlen über den Größenklassen der Kinder als Balkendiagramm auftragen. Dazu selektieren wir den Bereich B13:O13, klicken in der Mitte der Werkzeugleiste auf das Symbol, das wie ein kleines Histogramm (oder ein Bücherregal, graphing button) aussieht und gelangen in ein Fenster, in dem wir den Graphtyp auswählen. Im zweiten Schritt legen wir einige grafische Parameter fest oder bleiben bei den Vorgaben. Zum Abschluss wählen wir *OK* oder *Einfügen*, das Hilfsfenster verschwindet, und der Cursor auf dem Arbeitsblatt ändert sein Aussehen. Wir fahren das Fadenkreuz auf eine größere freie Fläche, das heißt unter unsere zweite Tabelle. Wenn wir die linke Maustaste loslassen, erscheint ein Balkendiagramm wie in Abbildung 10.6, das sich mit der Maus skalieren und positionieren lässt. Unter der Abszissenachse stehen die laufenden Nummern der zugehörigen Spalte, neben der Ordinatenachse die Besetzungszahlen.

Um das Diagramm auszugestalten, klicken wir mit der rechten Maustaste in seine Fläche und wählen *Eigenschaften* aus. Im weiteren Verlauf bestimmen wir Farben und Muster, fügen Texte hinzu und dergleichen mehr. Beim Abspeichern wird das Diagramm Bestandteil des Arbeitsblattes oder -buches und steht beim nächsten Aufruf wieder zur Verfügung.

Hätten wir die Rohdaten (Urliste) gleich in die Tabellenkalkulation eingetragen, wäre die Arbeit einfacher gewesen, aber wir kennen die Urliste nicht. Gal-

TONS Tabelle enthält bereits verdichtete Daten. Eine Tabellenkalkulation ist bei der Auswertung von Versuchen von Anfang an eine wertvolle Hilfe.

10.3 Octave, Euler (in Vorbereitung)

Octave aus dem GNU-Projekt ist eine Skriptsprache zur numerischen Lösung mathematischer Aufgaben, mit dem Heimathafen `http://www.octave.org/`. Es ist für die wichtigsten Rechnerarchitekturen verfügbar. **Euler** mit dem Heimathafen `http://euler.sourceforge.net/` ist ein ähnliches Werkzeug. Beide haben Anklänge an die kommerzielle Software MATLAB der Firma *The MathWorks*. Alle drei sind keine Computeralgebrasysteme, sondern beherrschen ausschließlich numerische Berechnungen und grafische Darstellungen.

10.4 Die GNU-R-Umgebung (in Vorbereitung)

GNU-R ist eine für mehrere Rechnerarchitekturen verfügbare Statistik-Software aus dem GNU-Projekt. Ihr Heimathafen ist `http://www.r-project.org/`, Ergänzungen liegen auf `http://cran.r-project.org/`. Die zentralen Themen sind numerische Rechnungen vorwiegend statistischer Art und grafische Darstellungen von Ergebnissen. R stellt eine freie Alternative zu der Statistik-Software S aus den Bell Laboratories dar, mit vielen Gemeinsamkeiten. Entsprechend ihrer Mächtigkeit erfordert sie mehr Einarbeitung als ein Taschenrechner.

Wie Gnuplot kann R interaktiv aus der Kommandozeile oder durch Skripte gesteuert werden, wobei Letzteres vorzuziehen ist, sobald die Rechnungen wiederholt werden sollen oder umfangreicher werden. Wie das Textsatzprogramm LaTeX kann R durch externe Pakete beliebig ergänzt werden. Da es seit 1997 öffentlich zugänglich ist, finden sich fertige Lösungen für viele Aufgaben. R verfügt über Schnittstellen zu Funktionen in C, Fortran und weiteren Programmiersprachen. Zu R gibt es eine grafische Benutzeroberfläche namens Rcmdr. Auch bei ihrer Verwendung muss man wissen, was man will und was die Ergebnisse bedeuten. Einzig die Syntax von R braucht man nicht genau zu kennen.

Beginnen wir – nach Einrichten der Grundsoftware – mit einem Beispiel aus *An Introduction to R*, leicht verändert. Da R alle Daten und die Kommandogeschichte im Arbeitsverzeichnis speichert, empfiehlt es sich, für jede Aufgabe ein eigenes Verzeichnis anzulegen. Wir tun das, wechseln hinein und starten die Software aus einer Kommandozeile

```
$ R
```

Das Dollarzeichen ist der Systemprompt, der vom Rechner geschrieben wird und anzeigt, dass er auf Eingaben wartet. Es erscheint ein Begrüßungstext und als letztes eine Eingabezeile mit dem R-Prompt. Wollen wir Schluss machen, tippen wir dort `q()` (quit) ein und schicken die Eingabe mit der Eingabetaste (CR, Return, Enter ...) ab. Aber so weit sind wir noch nicht. Wir geben ein

```
> help.start()
```

und bekommen unseren Web-Browser mit der Startseite der lokalen Dokumentation zu Gesicht, darunter die eben erwähnte Einführung. Wir schaffen das Fenster mit dem Browser aus dem Weg und machen in der R-Zeile weiter

```
> demo()
```

Das Angebot verfügbarer Demos verlassen wir durch Eingabe von `<q>` ohne Eingabetaste, da es von einem Hilfsprogramm, einem Pager wie `more` oder `less`, dargestellt wird. Nun wollen wir wissen, wie eine Demo gestartet wird, und geben ein

```
> help(demo)
```

Mit den daraus gezogenen Erkenntnissen geben wir ein

```
> demo(lm.glm)
```

starten die Demo und sehen außer einer Menge Text auch ein Grafikfenster mit einem Diagramm.

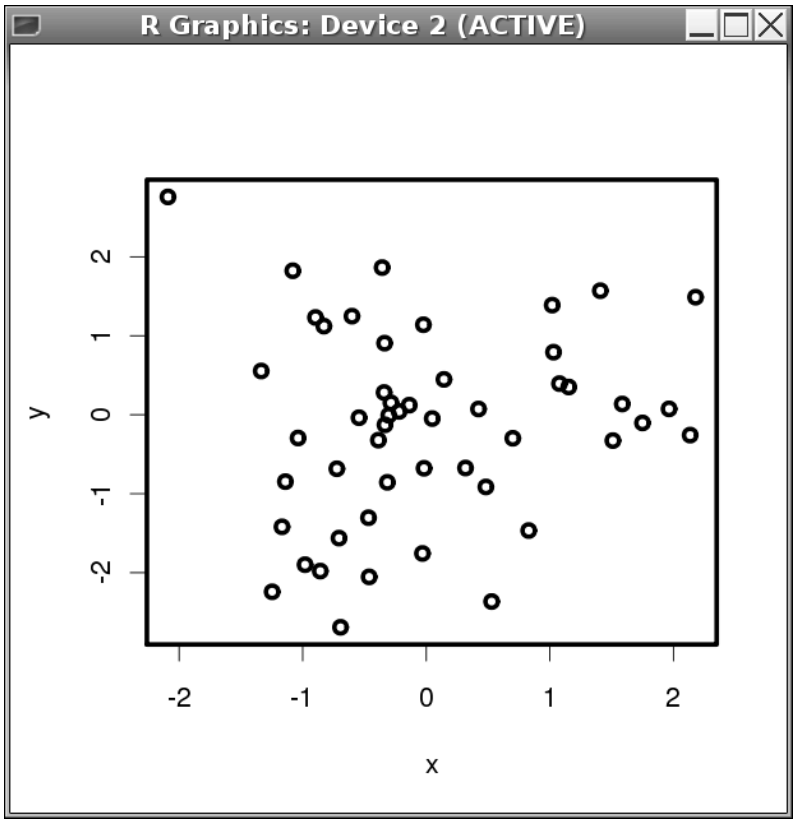


Abb. 10.7: Screenshot eines Beispiels zur Statistik-Software R: Punkte mit normalverteilten zufälligen Koordinaten (Scatterplot)

Nachdem wir die Ergebnisse bewundert haben, gehen wir an eine kleine Rechnung. In der R-Zeile geben wir ein

```
> x <- rnorm(50); y <- rnorm(x)
```

woraufhin wir nichts als eine neue R-Zeile bekommen. R ist nicht sehr gesprächig. Wir wollen erfahren, was die Funktion `rnorm()` bewirkt

```
> help(rnorm)
```

Die erste Zuweisung weist dem Vektor x 50 standardnormalverteilte Pseudo-Zufallszahlen zu, die zweite dem Vektor y so viele standardnormalverteilte Pseudo-Zufallszahlen, wie der Vektor x Komponenten hat, also ebenfalls 50. Zahlen werden in R grundsätzlich als Vektoren mit ganzzahligen, reellen oder komplexen Komponenten aufgefasst. Beide Zuweisungen werden durch ein Semikolon oder einen Zeilenwechsel getrennt. Wir sehen uns x an:

```
> show(x)
```

und zeichnen die Punkte in ein Diagramm

```
> par(lwd=4, cex=1.5); plot(x, y)
```

Die optionale Funktion `par(lwd=4, cex=1.5)` setzt einige grafische Parameter. Das Diagramm ist ein Quadrat mit einer leichten Häufung der Punkte um den Koordinatenursprung herum, siehe Abbildung 10.7.

Die Vektoren x und y sind allgemein gesprochen Objekte. R verarbeitet Objekte[3]. Das sind bestimmte Datenstrukturen wie Vektoren, Listen, Funktionen oder Ausdrücke. Objekte haben einen Modus oder Typ. Vektoren bestehen aus Komponenten (Werten) desselben Typs: numerisch, komplex, logisch, Zeichen oder raw (roh). Wir können den Modus eines Objekts abfragen

```
> mode(x)
```

Eine ähnliche Information liefert

```
> typeof(x)
```

Zu dem Unterschied zwischen Modus und Typ befrage man die *R Language Definition*. Ferner besitzen Objekte eine Länge, bei Vektoren die Anzahl der Komponenten

```
> length(x)
```

Modus und Länge gehören zu den Eigenschaften (property) eines Objekts. Wir listen unsere gegenwärtigen Objekte auf

```
> ls()
```

und finden erwartungsgemäß x und y nebst einigen Objekten aus der Demo. Einzelheiten zu einem Objekt erfahren wir mittels der Funktion `show()`, wie bereits ausprobiert. Wir löschen y, da nicht länger benötigt,

[3]Für C-Programmierer: R-Objekte sind Pointer auf Strukturen

```
> rm(y)
```

oder alle Objekte auf einen Schlag

```
> rm(list = ls())
```

Das Ergebnis von `ls()` bildet das Argument von `rm()`. Kontrolle mittels

```
> ls()
```

Falls wir uns vertippen, lässt sich die R-Zeile wie eine Kommandozeile korrigieren, bevor sie abgeschickt wird (Cursortasten, Backspace).

Wir setzen das Übungsbeispiel fort und weisen dem neu zu erzeugenden Vektor x ganzzahlige Werte von 1 bis 20 zu

```
> x <- 1:20
```

Dann erzeugen wir einen Vektor w

```
> w <- 1 + sqrt(x)/2
```

und ein Objekt *dummy*

```
> dummy <- data.frame(x = x, y = x + w * rnorm(x))
```

Was ist ein `data.frame`?

```
> help(data.frame)
```

Aha, eine Datenstruktur aus miteinander verbundenen Variablen, hier eine Tabelle von x, y-Paaren. Schaun wir mal

```
> dummy
```

Wir geben eine Datenanalyse von *dummy* in Auftrag

```
$ fm <- lm(y ~ x, data = dummy)
```

Der Tildenausdruck, das Datenmodell, besagt, dass hier y in Abhängigkeit von x betrachtet werden soll. Die höhere Funktion $lm()$ ruft ein lineares Modell mit niederen Funktionen auf und erspart uns deren Einzelaufruf. Das Objekt `fm` ist vom Modus `list` und hat die Länge 12. Um etwas von den Ergebnissen der Analyse zu sehen, rufen wir `summary()` auf

```
> summary(fm)
```

Nun sollten wir etwas von Statistik verstehen, um die Ausgabe zu deuten, und auch etwas Englisch können:

- **Residuals**
- **Coefficients**

- **Significance codes** bezieht sich auf die vorangegangene Zeile, in der das Signifikanzniveau durch Sternchen angegeben wird, hier drei Sternchen = 0.001.
- **Residual standard error**
- **Multiple R-squared**
- **Adjusted R-squared**
- **F-statistic**

Wir rechnen noch ein bisschen

```
> fm1 <- lm(y ~ x, data = dummy, weight = 1/w^2)
```

schauen uns wieder die Ergebnisse summarisch an:

```
> summary(fm1)
```

machen die Spalten des `data.frame` als Variable zugänglich, was sie nicht von vornherein sind

```
> attach(dummy)
```

definieren eine nichtparametrische lokale Regressionsfunktion:

```
> lrf <- lowess(x, y)
```

und plotten wieder:

```
> plot(x, y)
```

Dann zeichnen wir die Funktion lrf() in den vorhandenen Plot als Linie ein:

```
> lines(x, lrf$y)
```

hängen den `data.frame` wieder aus;

```
> detach()
```

und zeichnen ein etwas hübscheres Diagramm:

```
> plot(fitted(fm), resid(fm),
xlab = "Fitted values",
ylab = "Residuals",
main = "Residuals vs. Fitted")
```

Mit der mehrzeiligen Eingabe kommt R zurecht; der gewöhnliche R-Prompt kehrt erst wieder, wenn die Eingabe vollständig ist. Wir räumen auf:

```
> rm(dummy, fm, fm1, lrf, w, x)
```

und sagen Tschüss:

```
> q()
```

Für das nächste Mal aufzubewahren gibt es nichts. Das war eine kleine Kostprobe von R. Bis man mit R arbeiten kann, ist noch viel zu lesen und durchzuprobieren. Ausreichend Dokumentation steht zur Verfügung.

Wir wollen abschließend eine Regression rechnen und zeichnen, ähnlich wie in Abbildung 8.3 auf Seite 135. Ausgangspunkt ist eine Tabelle von FRANCIS GALTON, in der er die Ergebnisse einer Untersuchung der Körpergröße von Eltern und ihren Kinder (im Erwachsenenalter) darstellt.[4] Die Tabelle wird beschrieben in seinem Buch ab Seite 87, insbesondere auf Seite 91. Die Größe der Eltern (mid-parent) ist das Mittel aus der Größe des Vaters und dem 1,08-fachen der Größe der Mutter in Inch gemessen. Mit demselben Faktor wurde die Größe der Töchter multipliziert. Die Probe umfasste 205 Eltern und 928 Kinder. Die Größe der Eltern wurde in Klassen gleicher Breite (1 inch) eingeteilt und in jeder Klasse die Größe der Kinder verzeichnet, ebenfalls klassiert. Tabelle 10.1 zeigt die Klassenmitten der Elterngrößen und die zugehörigen Medianwerte der Größe der Kinder.

Tab. 10.1: Körpergröße von Eltern und ihren Kindern, gemessen in Inch, Medianwerte Kinder (k) vs. Klassenmitten Eltern (e), nach F. GALTON (1889)

e	k		
65,0	65,8	70,0	68,9
66,0	66,7	71,0	69,5
67,0	67,2	72,0	69,9
68,0	67,6	73,0	72,2
69,0	68,2		

Die Tabelle schreiben wir zweispaltig – mit Spaltenüberschrift und Dezimalpunkt statt -komma – in eine Datei namens `galton`. In R lesen wir die Datei in ein Objekt gleichen Namens ein

```
> galton <- read.table("galton", header = TRUE)
```

schauen uns den Erfolg an

```
> galton
```

machen die Spalten als Variable zugänglich

```
> attach(galton)
```

und zeichnen

```
> plot(e, k, xlab = "Eltern",
ylab = "Kinder",
main = "Kindergroesse vs. Elterngroesse")
```

[4] F. GALTON, *Natural Inheritance*, Macmillan, 1889, Table 11, Seite 208. Zu finden auf `http://galton.org/books/natural-inheritance/`

10.5 Zufallszahlen (in Vorbereitung)

10.5.1 Wofür Zufallszahlen?

Im Leben braucht man gelegentlich Zufallszahlen, genauer gesagt **zufällige Folgen von Zahlen**, denn von einer einzelnen Zahl lässt sich nicht behaupten, sie sei zufällig oder nicht. Sehen wir uns einige Anwendungen von Zufallszahlenfolgen an:

- Die Wikipedia bietet an, einen *Zufälligen Artikel* (random article, article au hazard) anzuzeigen.
- Eine Sammlung von Musikstücken soll in zufälliger Reihenfolge abgespielt werden.
- Ein Passwort soll als zufällige Zeichenfolge erzeugt werden, um ein Erraten zu erschweren.
- Für eine Simulation braucht man Zahlenwerte, die in einem bestimmten Bereich zufällig verteilt sind.
- Bei Online-Banküberweisungen braucht man eine Liste mit Zufallszahlen als Transaction Numbers (TAN).
- Aus einer endlichen Grundgesamtheit (Wählerverzeichnis, Seiten eines Buches, Packungen einer Charge eines Medikamentes) sollen Elemente für eine Stichprobe zufällig entnommen werden.
- Einige Verfahren zur Verschlüsselung von Daten benötigen Folgen von Zufallszahlen.
- Bei der Ziehung von Lottozahlen erwarten die Teilnehmer, dass die Zahlen rein zufällig gezogen werden.

Offensichtlich werden an die *Zufälligkeit* der Zahlenfolgen unterschiedlich strenge Anforderungen gestellt, wobei obige Auflistung grob nach wachsender Strenge sortiert ist.

Was beinhaltet der Begriff *Zufälligkeit*? Eingangs hatten wir von einem Zufallsexperiment verlangt, dass der Vorgang grundsätzlich beliebig oft wiederholbar sei, dass die Ergebnisse der Wiederholungen voneinander unabhängig und dass sie drittens nicht vorhersagbar seien. Mit der ersten Forderungen haben wir keine Schwierigkeiten. Die zweite Forderung lässt sich dadurch erfüllen, dass der Zahlenvorrat, aus dem der Zufallszahlengenerator seine Ergebnisse zieht (Urnenmodell), sich durch eine Ziehung nicht verändert (Ziehen mit Zurücklegen oder unbegrenzter Vorrat). Es ist also prinzipiell möglich, dass mehrmals nacheinander dieselbe Zufallszahl ausgegeben wird. Das ist bei manchen Anwendungen unerwünscht und wird ausgefiltert, was einen Eingriff in die Zufälligkeit darstellt. Dasselbe gilt für einfache Ziffernfolgen wie 12345 oder 77777.

Bleibt die dritte Forderung, die Unvorhersagbarkeit. Kein deterministischer Vorgang erzeugt zufällige Zahlenfolgen. Das wäre ein Widerspruch in sich. Ein Computer ist ein deterministischer Automat, der bei gleicher Ausgangslage nach einem Algorithmus stets dieselbe Zahlenfolge erzeugt. Es gibt allerdings Hardware zur Erzeugung zufälliger Zahlenfolgen, die auf Rauschen, radioaktivem Zerfall, Quanteneffekten oder ähnlichen Erscheinungen beruht.

Schraubt man die Anforderungen an die Zufälligkeit herab – verlangt man also keine absolute Unvorhersagbarkeit, sondern begnügt sich mit einer mehr oder weniger schwierigen Vorhersagbarkeit – bestehen Möglichkeiten, mit einem Computer fast zufällige Zahlenfolgen zu erzeugen. Deren Zahlen werden **Pseudo-Zufallszahlen** genannt, weil sie dem Betrachter als zufällig erscheinen, in Wirklichkeit jedoch berechenbar sind. Ein gelegentlich entscheidender Vorteil algorithmisch erzeugter Zahlenfolgen ist ihre Wiederholbarkeit. Man kann ein Experiment identisch wiederholen, was für seine Verifizierung und damit seine wissenschaftliche Anerkennung entscheidend ist.

Die Zufallszahlen stammen stets aus einem begrenzten Intervall wie $[0, 1]$, $[-1, +1]$ oder $[0, 2^{32} - 1]$. Sie haben ferner eine endliche Genauigkeit oder Auflösung, da wir sie mit endlichen Mitteln darstellen müssen. Man kann sich daher im Prinzip auf die ganzen Zahlen beschränken. In dem Intervall können sie gleichverteilt sein. Es finden sich aber auch Aufgabenstellungen, die eine Normal- oder Poisson-Verteilung verlangen. In jedem Fall muss man sich davon überzeugen, dass der gewählte Algorithmus (Generator) zur Aufgabe passt.

10.5.2 Erzeugung

10.5.3 Prüfung auf Zufälligkeit

Anhang A

Zum Weiterlesen

Die Auswahl ist subjektiv und enthält Werke, die wir noch lesen wollen, schon gelesen haben oder oft benutzen. Einige der Bücher können vergriffen oder inzwischen in neuerer Auflage erschienen sein.

1. Material aus dem World Wide Web (WWW)
   - Statistisches Bundesamt Deutschland
     `http://www.destatis.de/`
   - Statistik Austria
     `http://www.statistik.at/`
   - Bundesamt für Statistik (Schweiz)
     `http://www.bfs.admin.ch/`
   - Statistisches Landesamt Baden-Württemberg
     `http://statistik-bw.de/`
   - Statistisches Landesamt des Freistaates Sachsen
     `http://www.statistik.sachsen.de/`
   - ISI Multilingual Glossary of Statistical Terms
     `http://isi.cbs.nl/glossary/`
   - Engineering Statistics Handbook
     `http://www.itl.nist.gov/div898/handbook/`
   - GSL – GNU Scientific Library
     `http://www.gnu.org/software/gsl/`
   - Mathematik Online
     `http://mo.mathematik.uni-stuttgart.de/`
   - The R Project for Statistical Computing
     `http://www.r-project.org/`
   - Startistics (U. Wien)
     `http://learnserver.csd.univie.ac.at/statistics/`
   - Statistik am URZ Heidelberg
     `http://www.urz.uni-heidelberg.de/statistik/`

- StatWiki (HU Berlin)
  http://statwiki.hu-berlin.de/
- StatProb – The Encyclopedia sponsored by Statistics and Probability Societies
  http://statprob.com/
- Wikibook: Lehrbuch der Mathematik, Band 7: Statistik
  http://de.wikibooks.org/wiki/Mathematik:_Statistik
  auch http://en.wikibooks.org/wiki/Statistics
  oder http://fr.wikibooks/wiki/Statistiques

F. Böker Statistik III
Universität Göttingen, Vorlesungsskript 2004
http://www.statoek.wiso.uni-goettingen.de/
veranstaltungen/statistik3alt/daten/

P. L. Davies Wahrscheinlichkeitstheorie I
Universität Essen, Vorlesungsskript SS 2000
http://wwwstat.mathematik.uni-essen.de/~kovac/wt1/

H. Diefenbacher, A. Frank Einfach Lernen! Statistik
http://studentensupport.de/

R. Dutter Statistik und Wahrscheinlichkeitsrechnung für InformatikerInnen
Technische Universität Wien, Vorlesungsskript SS 2006
http://www.statistik.tuwien.ac.at/public/
dutt/vorles/inf_bak/

J. C. Pezzullo Webpages that Perform Statistical Calculations
http://www.statpages.org/

M. Scheutzow Stochastische Modelle
ftp://ftp.math.tu-berlin.de/pub/Lehre/StochMod/WS03-04/
main.ps

V. Schmidt Stochastik für Informatiker, Physiker, Chemiker und Wirtschaftswissenschaftler
Universität Ulm, Vorlesungsskript SS 2001
http://www.mathematik.uni-ulm.de/stochastik/
lehre/ss01/stochInfWi/vs1/vs1.html

V. Schmidt Wahrscheinlichkeitsrechnung
Universtiät Ulm, Vorlesungsskript WS 2003/4
http://www.mathematik.uni-ulm.de/stochastik/
lehre/ws03_04/wr/skript/

M. P. H. Wolff, P. Hauck, W. Küchlin Mathematik für Informatik und Bioinformatik
Universität Tübingen, 2004
http://mfb.informatik.uni-tuebingen.de/book/
WolffHauckKuechlin.html

K. Zwerenz, K. Kehrle Grundkurs Statistik
Virtuelle Hochschule Bayern, ohne Jahresangabe
http://www.lerne-statistik.de/

2. DIN-Normen

1319 Grundbegriffe der Messtechnik

55 302 Häufigkeitsverteilung, Mittelwert und Streuung

55 350 Begriffe der Qualitätssicherung und Statistik (mehrteilig)

66 141 Darstellung von Korngrößenverteilungen; Grundlagen

66 143 Potenznetz

66 144 Logarithmisches Normalverteilungsnetz

66 145 RRSB-Netz (Ersatz für DIN 4190 Teil 1)

3. Bücher zur Statistik

M. Abramowitz, I. A. Stegun Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables
Dover Publications, New York, 1965, 1046 S.

I. N. Bronstein u. a. Taschenbuch der Mathematik
Harri Deutsch, Frankfurt (M), 2008, 1216 S.

E. Cantoni, P. Huber, E. Ronchetti Maîtriser l'aléatoire
Springer, Berlin + Heidelberg, 2009, 300 S.

A. Caputo et al. Arbeitsbuch Statistik
Springer, Berlin + Heidelberg, 2009, 304 S.

A. J. Chorin, O. H. Hald Stochastic Tools in Mathematics and Science
Springer, Berlin + Heidelberg, 2009, 162 S.

L. Devroye Non-Uniform Random Variate Generation
Springer, Berlin + Heidelberg, 1986, 843 S.

L. Fahrmeir et al. Statistik – Der Weg zur Datenanalyse
Springer, Berlin + Heidelberg, 2010, 610 S.

L. Fahrmeir, T. Kneib, S. Lang Regression
Springer, Berlin + Heidelberg, 2009, 502 S.

M. Fisz Wahrscheinlichkeitsrechnung und Mathematische Statistik
VEB Deutscher Verlag der Wissenschaften, Berlin 1976, 777 S.

J. E. Gentle Computational Statistics
Springer, Berlin + Heidelberg, 2009, 752 S.

S. Givant, P. Halmos Introduction to Boolean Algebras
Springer, New York, 2009, 574 S.

U. Graf et al. Formeln und Tabellen der angewandten mathematischen Statistik
Springer, Berlin + Heidelberg, 1997, 529 S.

R. L. Graham, D. E. Knuth, O. Patashnik Concrete Mathematics
Addison-Wesley, Boston 1994, 657 S.

A. Gut An Intermediate Course in Probability
Springer, Berlin + Heidelberg, 2009, 302 S.

R, Hafner Nichtparametrische Verfahren der Statistik
Springer, Wien, 2001, 233 S.

R.-D. Hilgers, P. Bauer, V. Scheiber Einführung in die Medizinische Statistik
Springer, Berlin + Heidelberg, 2007, 330 S.

P. D. Hoff A First Course in Bayesian Statistical Methods
Springer, Berlin + Heidelberg, 2009, 268 S.

D. E. Knuth The Art of Computer Programming
Vol. 1: Fundamental Algorithms, Chapter 1: Basic Concepts
Vol. 2: Seminumerical Algorithms, Chapter 3: Random Numbers
Addison-Wesley, Boston, 1998, 762 S.

E. Kreyszig Statistische Methoden und ihre Anwendungen
Vandenhoeck + Ruprecht, Göttingen, 1998, 451 S.

M. Kühlmeyer Statistische Auswertungsmethoden für Ingenieure
Springer, Berlin + Heidelberg, 2001, 417 S.

H. Meschkowski Wahrscheinlichkeitsrechnung
Bibliographisches Institut, Mannheim, 1968, 233 S.

W. H. Press et al. Numerical Recipes in C
Cambridge University Press, Cambridge (UK) 1992, 994 S.

S. M. Ross Statistik für Ingenieure und Naturwissenschaftler
Spektrum (Elsevier), 2006, 561 S.

D. Ruelle Zufall und Chaos
Springer, Berlin + Heidelberg, 1994, 254 S.

L. Sachs Statistische Methoden – Ein Soforthelfer
Springer, Berlin + Heidelberg, 1970, 103 S.

L. Sachs, J. Hedderich Angewandte Statistik
Springer, Berlin + Heidelberg, 2006, 702 S.

L. Sachs, J. Hedderich Angewandte Statistik – Methodensammlung mit R
Springer, Berlin + Heidelberg, 2009, 813 S.

M. W. Smirnow, L. W. Dunin-Barkowski Mathematische Statistik in der Technik
VEB Deutscher Verlag der Wissenschaften, Berlin, 1963, 431 S.

A. Steland Basiswissen Statistik
Springer, Berlin + Heidelberg, 2010, 270 S.

D. Stoyan Stochastik für Naturwissenschaftler und Ingenieure
Wiley-VCH, Weinheim, 1998, 307 S.

G. und S. Teschl Mathematik für Informatiker
Bd. 2: Analysis und Statistik
Springer, Berlin + Heidelberg, 2007, 387 S.

B. L. van der Waerden Mathematische Statistik
Springer, Berlin + Heidelberg, 1971, 360 S.

E. Wagemann Narrenspiegel der Statistik
Hanseatische Verlagsanstalt, Hamburg 1935, 255 S.

E. Weber Grundriß der biologischen Statistik
Urban + Fischer, München 1991, 652 S.

A. F. Zuur, E. N. Ieno, E. H. W. G. Meesters A Beginner's Guide to R
Springer, Berlin + Heidelberg, 2009, 220 S.

4. Bücher zur Partikeltechnologie

T. Allen Powder Sampling and Particle Size Determination
Elsevier Science, Amsterdam 2003, 682 S.

W. Batel Einführung in die Korngrößenmeßtechnik
Springer, Berlin + Heidelberg 1971, 214 S.

C. Bernhardt Granulometrie
Deutscher Verlag für Grundstoffindustrie, Leipzig 1990, 400 S.

G. Herdan Small Particle Statistics
Butterworths, London 1960, 418 S.

K. Sommer Probenahme von Pulvern und körnigen Massengütern
Springer, Berlin + Heidelberg 1979, 305 S.

M. Stieß Mechanische Verfahrenstechnik – Partikeltechnologie I
Springer, Berlin + Heidelberg 2007, 498 S.

Personenverzeichnis

Sachverzeichnis