Mathematik in der Biologie

HD Dr. Günther Grün

Version 1.1

Rheinische Friedrich-Wilhelms-Universität Bonn
Wintersemester 2003/04 — Sommersemester 2004

Vorwort

Das vorliegende Skriptum ist aus einer zweisemestrigen Kursvorlesung entstanden, die ich im Wintersemester 2003/04 und im Sommersemester 2004 an der Rheinischen Friedrich-Wilhelms-Universität Bonn für Studierende der Biologie in den ersten beiden Fachsemestern gehalten habe.

Deskriptive Statistik, elementare Wahrscheinlichkeitstheorie und Methoden der beurteilenden Statistik waren die Inhalte der Vorlesung im Wintersemester. Im Sommersemester schloss sich eine Einführung in Matlab (und gegen Ende auch SPSS) an. Auf dieser Grundlage wurden rechnergestützt statistische Auswertungen vorgenommen und Zufallssimulationen durchgeführt.

Das Skriptum orientiert sich allerdings nicht an der Chronologie der Präsentation der Lerninhalte in den Vorlesungen. Vielmehr habe ich im wesentlichen die Gliederung der Wintersemestervorlesung übernommen. In diesen Rahmen sind ausgewählte Anwendungsbeispiele aus dem Sommersemester eingebettet, die die Theorie des Wintersemesters illustrieren und "mit Leben erfüllen" sollen. Insbesondere wird zu allen Beispielen der Quell-Code jener Matlab-Programme aufgeführt, mit denen sie gerechnet worden sind. Zu ihrem besseren Verständnis findet sich im Anhang eine Kurzreferenz zu Matlab.

Nicht nur für die Erstellung dieser Kurzreferenz gebührt Herrn Dipl.-Math. Thomas Roessler mein besonderer Dank. Er hat darüber hinaus mein Vorlesungsmanuskript in eine ansprechende Form übertragen und damit wesentlich zum Entstehen dieses Skriptums beigetragen.

Herr Dipl.-Math. Thomas Viehmann und Herr Dipl.-Chem. Christian Egler, die im Winter- bzw. Sommersemester die Vorlesung betreuten, haben mit unverwechselbaren Übungsaufgaben zum guten Gelingen dieser Lehrveranstaltungen beigetragen. Bei der inhaltlichen Gestaltung der Vorlesung habe ich profitiert von der einschlägigen Literatur [1]-[5] und den Vorlesungsskripten der vergangenen Jahre, die Herr Prof. Wolfgang Alt mir freundlicherweise zur Verfügung gestellt hat. Auch hierfür meinen herzlichen Dank.

Bonn, im Oktober 2004 Günther Grün

Inhaltsverzeichnis

Kapitel 1

Mathematik in der Biologie — ein erster Überblick

Die biologischen Wissenschaften werfen immer wieder Fragestellungen auf, deren Lösung eine Anwendung unterschiedlichster mathematischer Methoden verlangt. Dieser Abschnitt stellt einige einfache Beispiele vor, die im weiteren Verlauf der Vorlesung zum Teil wieder aufgegriffen werden. Ein erstes Beispiel ist die Ermittlung einfacher Gesetzmäßigkeiten zwischen unterschiedlichen anatomischen Messgrößen. Es geht hierbei um die Formulierung biologischer Skalengesetze und allometrischer Aussagen. Andere Anwendungen haben ihren Ursprung in Fragen der Prozessoptimierung und stehen damit in engem Zusammenhang mit der Modellierung dieser Prozesse. Abschnitt 1.2 behandelt ein elementares Optimierungsproblem; Abschnitt 1.3 zeigt auf, wie mittels Differentialgleichungsmodellen komplexe Vorgänge der Populationsdynamik erklärt werden können. Besonderen Raum werden in der Vorlesung statistische und stochastische Methoden einnehmen. Abschnitt 1.4 führt in diesen Problemkreis mit einigen beispielhaften Fragenstellungen ein.

1.1 Biologische Skalengesetze und Allometrie

Die *Allometrie* befasst sich mit der Messung von Unterschieden in den Wachstumsraten verschiedener Körperteile. Zum Beispiel wird qualitativ festgestellt, dass bei Hamstern das Gewicht schneller wächst als die Körperlänge. Mögliche Messergebnisse sind schematisch in Abbildung 1.1 dargestellt.

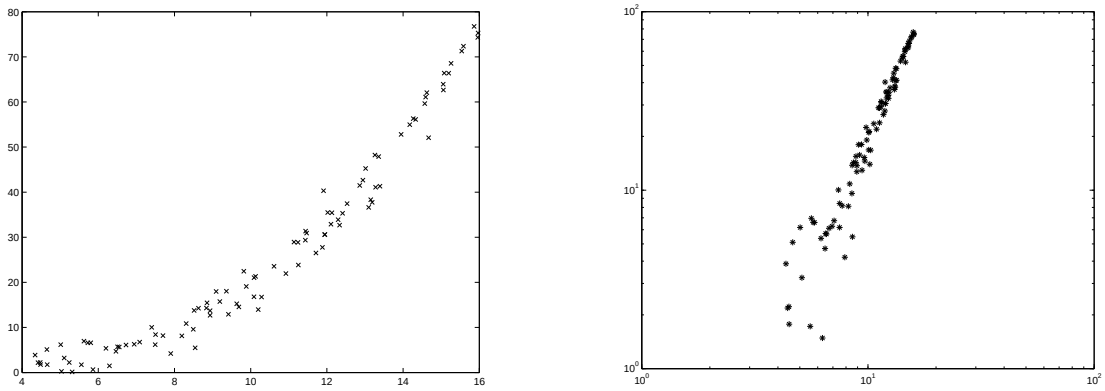


Abbildung 1.1: Allometrie: Gewicht vs. Körpergröße bei Hamstern in linearer und doppelt logarithmischer Darstellung

Wir vermuten, dass die Abhängigkeit zwischen Gewicht und Körpergröße durch ein Potenzgesetz dargestellt wird:

$$\text{Körpergewicht} = c \cdot \text{Körpergröße}^{\kappa} \tag{1.1}$$

c und κ sind zu bestimmen. Hierzu erweist sich eine *doppelt logarithmische Darstellung* als hilfreich. Und zwar gilt:

$$\log(\text{Körpergewicht}) = \kappa \log(\text{Körpergröße}) + \log c \tag{1.2}$$

Wenn ein Potenzgesetz gilt, sollte also in doppelt logarithmischer Darstellung ($\log$ Körpergewicht vs. $\log$ Körpergröße) eine lineare Abhängigkeit zu erkennen sein. Dies führt auf ein Problem der Approximationstheorie: Gesucht ist eine Gerade, die die Datenpunkte in einem noch zu präzisierenden Sinne optimal approximiert.

κ kann dann als Steigung, $\log c$ als Schnittpunkt der Geraden mit der y-Achse abgelesen werden.

1.2 Optimierungsaufgaben — ein einfaches Beispiel

Gegenstand einer Optimierungsaufgabe ist die Minimierung oder Maximierung von Größen. Als einführendes Beispiel betrachten wir das Problem der optimalen Beweidung

einer Weide durch Nutztiere. Sei dazu die Gesetzmäßigkeit zwischen Gesamttagesfraß y und der Anzahl der Tiere x auf der Weide bekannt und durch

$$y = \frac{\beta x}{1 + x^2} \tag{1.3}$$

gegeben. (Siehe auch Abbildung 1.2.)

Diesen Ausdruck interpretieren wir wie folgt: Der Tagesfraß auf einer Weide ist, naiv betrachtet, proportional zur Anzahl der Tiere ($y \approx \beta x$) — je mehr Tiere auf die Weide geschickt werden, desto mehr wird gefressen. Andererseits erhöht sich mit der Anzahl der Tiere auch der Stress, dem diese ausgesetzt sind. Die appetitzügelnde Wirkung dieser Anspannungen drücken wir durch den Korrekturfaktor $\frac{1}{1+x^2}$ aus.

Die Optimierungsfrage für den Landwirt lautet nun: *Bei welcher Anzahl von Tieren wird y maximiert?* Wie wird x_{opt} bestimmt?

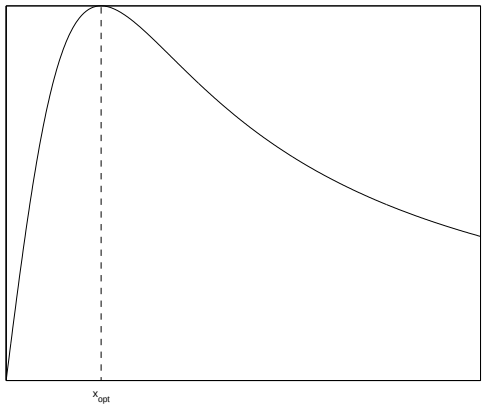


Abbildung 1.2: Gesamttagesfraß vs. Anzahl der Tiere

1.3 Prozessmodellierung — ein Beispiel aus der Populationsdynamik

Ziel der Prozessmodellierung ist das Verständnis von Prozessmechanismen. Dieses Verständnis eröffnet dann Möglichkeiten zur Prozesssteuerung.

Wir präsentieren ein Beispiel aus der Populationsdynamik. Eine Erhebung zum Anteil von Haien am Fischfang während der Jahre 1914–1923 in italienischen Adriahäfen hat

Jahr	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923
Haie / %	11,98	21,48	22,18	21,28	36,48	27,38	16,08	15,98	14,08	10,78

Tabelle 1.1: Fischfang in der Adria 1914–1923

(siehe Tabelle 1.1) zu der überraschenden Beobachtung geführt, dass während des ersten Weltkriegs anteilsmäßig mehr Haie gefangen wurden als in den übrigen Jahren.

Ein erster Erklärungsversuch führt diese Beobachtung auf die Verringerung des Fischfangs während des Krieges zurück. Nach dieser Erklärung hätte die Verringerung des Fischfangs während des Krieges zu mehr Nahrung für die Haie geführt, deren Population demzufolge gewachsen wäre.

Dieser Erklärungsversuch scheitert an der Erwartung, dass die Speisefischpopulationen ebenfalls steigen sollten — der *prozentuale* Anstieg des Haifischfangs wird hierdurch nicht erklärt.

Das Räuber-Beute-Modell von Lotka und Volterra liefert eine schlüssige Erklärung der beobachteten Fangmengen. Zunächst wird das folgende System von Differentialgleichungen für Räuber- und Beutepopulationen ohne Fischfang aufgestellt:

$$\left\{ \begin{array}{lcl} x' = \frac{dx}{dt} & = & ax - bxy \\ y' = \frac{dy}{dt} & = & -cy + dxy \end{array} \right\} \tag{1.4}$$

x' und y' geben dabei die Änderungsraten von Beute- und Räuberpopulationen an. ax ist das Wachstum der Beute unter der Annahme, dass keine Räuber auftreten. $-bxy$ modelliert den Verlust der Beutefische durch Haifraß. $-cy$ stellt die Sterberate der Räuber in einem beutefreien Biotop dar, während dxy die Wachstumsrate der Räuber in einem Biotop modelliert, in dem Beute verfügbar ist.

($a, \ldots, d$ werden jeweils als positiv angenommen.)

Um den Einfluss des Fischfangs zu modellieren, wird ein positiver Parameter ϵ eingeführt, der zu folgendem modifiziertem System führt:

$$\left\{ \begin{array}{lcl} x' & = & (a - \epsilon)x - bxy \\ y' & = & -(c + \epsilon)y + dxy \end{array} \right\} \tag{1.5}$$

Ein Vergleich von Lösungen der beiden Systeme (siehe Abbildung 1.3), liefert das *Volterra-Prinzip*: Durch eine Verringerung des Fischfangs wird im Durchschnitt die Räuberpopulation (hier: Haie) erhöht und die Beutepopulation vermindert. Wir beobachten also eine gegenseitige Abhängigkeit der beiden Tierarten.

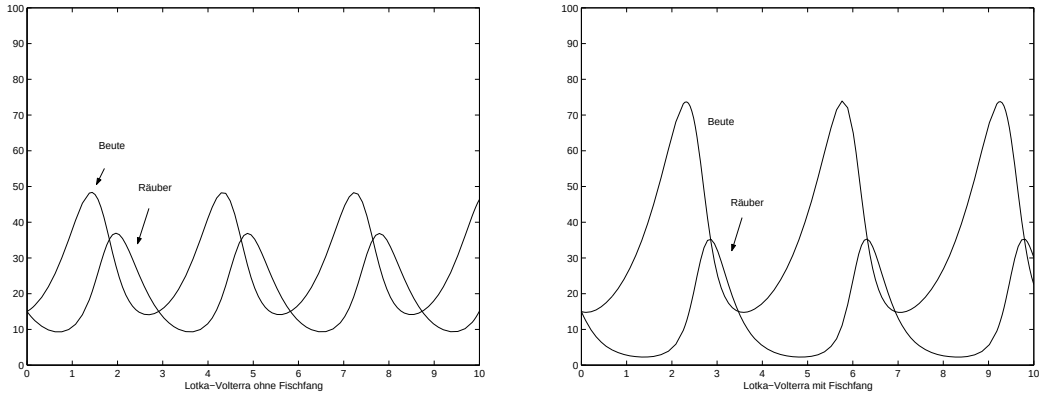


Abbildung 1.3: Lotka-Volterra: Räuber-Beute-Modell ohne und mit Fischfang

1.4 Datenerhebungen — beurteilende Statistik

Problem 1. Mit welcher Verlässlichkeit können auf der Basis einer Stichprobe Schlussfolgerungen auf eine Grundgesamtheit gezogen werden? Als Beispiel betrachte man eine Wahlvorhersage. Bei einer Umfrage werden 1000 Personen befragt; Partei A erhält 50 % der Stimmen. Wie sicher ist diese Aussage?

Ziel wäre eine Aussage der folgenden Form: *Mit einer Irrtumswahrscheinlichkeit von 5% wird die Partei A zwischen 48% und 52% der Stimmen erzielen.*

Problem 2. Test auf Inklusionsbeziehung. Welche Schlüsse lassen sich aus Datenerhebungen ziehen? In der Nähe einer Chemiefabrik häufen sich die Funde toter Tiere. Ein Biologe vermutet, dass eine bestimmte Chemikalie, die von der Fabrik produziert wird, für das Sterben verantwortlich sei. Messungen an 10 toten Hasen ergeben die folgenden Konzentrationen:

33 66 26 43 46 55 42 38 17 63

Die Vertreter der Chemiefabrik behaupten, diese Konzentrationen seien nicht höher als üblich, da die Chemikalie natürlich in der Umwelt auftrete.

Die Fragestellung für die beurteilende Statistik ist hier, ob die gemessenen Werte auf eine *signifikant höhere* Belastung hindeuten. Hierzu müssten etwa Vergleiche mit Erhebungen aus anderen Regionen herangezogen werden.

Problem 3. Korrelationen zwischen verschiedenen Merkmalen. Hier interessiert man sich dafür, ob es Zusammenhänge zwischen verschiedenen Merkmalen von Testperso-

nen oder Versuchseinheiten gibt — z.B. zwischen den Musik- und Lateinnoten der Schüler eines Jahrgangs.

Problem 4. Untersuchung des Einflusses verschiedener Faktoren auf ein Ergebnis Mögliche Anwendungen für diesen Typ von Fragestellung wären etwa Fragen nach der Wirksamkeit eines Medikamentes oder nach genetischen Komponenten bei Volkskrankheiten: Welche Gene beeinflussen die Erkrankungswahrscheinlichkeit?

Übungen

Aufgabe 1

In dieser Aufgabe sollen Sie sich ein wenig an den Logarithmus gewöhnen.

Erinnern Sie sich daran, dass für beliebige positive Zahlen a, b und x man den Logarithmus von x zur Basis a mit der Formel $\log_a x = \frac{\log_b x}{\log_b a}$ berechnen kann. Berechnen Sie mit dem Taschenrechner die folgenden Lorarithmen ($\ln x$ ist der natürliche Logarithmus, der Logarithmus zur Eulerschen Zahl $e = \exp(1)$ als Basis)

$$\log_{10} e, \quad \ln 10, \quad \log_2 12, \quad \ln 12, \quad \log_{10} 12, \quad \log_{12} 12.$$

Welches ist der Logarithmus von x zur Basis 10, falls der natürliche Logarithmus von x gerade 5 ist?

Schließen Sie aus $\exp(x + y) = \exp(x) \cdot \exp(y)$, dass für beliebige a,x,y und natürliche Zahlen m und n

$$\ln \frac{x}{y} = \ln x - \ln y$$

und

$$\ln(a^\kappa) = \kappa \ln a \quad \text{mit } \kappa = \frac{m}{n}$$

gilt.

Was ist $\log_b \frac{x}{y}$ und $\log_b a^\kappa$ für beliebiges positives b?

Kapitel 2

Beschreibende Statistik

Gegenstand der beschreibenden (*deskriptiven*) Statistik sind Methoden zur Darstellung und (An)ordnung empirischer Daten durch Tabellen und Graphiken. Empirische Daten sollen außerdem durch grundlegende Kenngrößen quantitativ beschrieben werden. In diesem Kapitel werden die wesentlichen Techniken vorgestellt. Zusätzlich präsentieren wir Programmbeispiele für die praktische Umsetzung dieser Techniken unter Matlab.

2.1 Grundbegriffe

Im Zuge einer *Datenerhebung* werden an ausgewählten *Versuchseinheiten* (oder *Merkmalsträgern*) (Englisch: *experimental units*) ein oder mehrere *Merkmale* (oder *Variablen*) festgestellt. Die Werte, die von einem Merkmal angenommen werden können, heißen Merkmalsausprägungen oder mögliche Variablenausprägungen.

Einige Beispiele:

Versuchseinheit	Merkmal	Merkmalsausprägung
Tiere einer Population	Gewicht	$\mathbb{R}^+$
	Geschlecht	M/W
	Cholesterinkonzentration	$\mathbb{R}^+$
	Rang in der Hierarchie	$\mathbb{N}$
Bäume eines Waldes	Schädlingsbefall	keiner, gering, mittel, stark
	Höhe, Gewicht	$\mathbb{R}^+$
Pflanzen	Blattlänge	$\mathbb{R}^+$
	Blütenzahl	$\mathbb{N}$
	Blütenfarbe	weiß/blau/gelb

2.2 Klassifikation von Variablen

<table>
<tr><td colspan="4">Merkmale</td></tr>
<tr><td colspan="2">quantitativ
(nur zahlenmäßig erfassbar)</td><td colspan="2">qualitativ
(artmäßig erfassbar)</td></tr>
<tr><td>diskret
Isolierte Zahlenwerte</td><td>stetig
Intervalle $[a, b]$, $-\infty \leq a < b \leq +\infty$</td><td>ordinal
Ausprägungen vergleichbar</td><td>nominal
Ausprägungen haben lediglich Bezeichnungscharakter — Vergleich nicht möglich</td></tr>
<tr><td colspan="2">auf metrischer Skala messbar, d.h., Differenzen sind interpretierbar.
1. Fall: “Intervallskala”, d.h., die Wahl des Nullpunkts ist willkürlich. Quotienten sind in diesem Fall nicht interpretierbar. Beispiel: Temperatur.
2. Fall: “Verhältnisskala”, d.h., der Nullpunkt ist eindeutig bestimmt, Quotientenbildung ist sinnvoll. Beispiel: Längen, Gewichte.</td><td colspan="2"></td></tr>
</table>

Falls es nur zwei Ausprägungen gibt, so heißt das Merkmal dichotomisch.

Warnung: Sind die Merkmalsausprägungen Zahlen, so folgt *nicht*, dass das Merkmal quantitativ ist. So werden etwa Befunde („positiv“, „negativ“) häufig durch die Zahlen 1 und 0 codiert.

2.3 Stichproben und Grundgesamtheit

Unter der *Grundgesamtheit* versteht man die Menge der Versuchseinheiten (Merkmalsträger), über die eine Aussage getroffen werden soll — etwa die Pflanzen eines Feldes, die Menschen einer Stadt oder die Regionen eines Landes. Wichtig ist eine genaue Definition der Grundgesamtheit bei einer jeden Datenerhebung.

In der Regel ist eine Untersuchung aller Elemente einer Grundgesamtheit — eine Totalerhebung — nicht möglich. Als Ausweg bedient man sich einer *repräsentativen Teilauswahl*.

"Repräsentativ" heißt hierbei, dass die Teilauswahl hinsichtlich aller relevanten Charakteristika im wesentlichen mit der Grundgesamtheit übereinstimmen soll. Die Statistik lehrt, dass eine Auswahl dann repräsentativ ist, wenn alle Elemente der Grundgesamtheit die gleiche Chance haben, ausgewählt zu werden. Man spricht von einer *Zufallsstichprobe*.

2.4 Darstellungsmethoden für empirische Daten

2.4.1 Urliste

Unter der *Urliste* versteht man die Auflistung aller erhobenen Daten. Zur übersichtlichen Darstellung verwendet man üblicherweise eine *Datenmatrix*. Daten über die Mitglieder einer Fußballmannschaft könnten etwa wie in Tabelle 2.1 dargestellt werden.

Merkmalsträger	Alter	Ruhepuls	geschossene Tore
1	33	53	0
2	25	57	1
3	27	56	1
4	26	55	0
5	27	53	2
6	25	54	1
7	28	55	4
8	24	54	1
9	25	55	5
10	26	56	3
11	26	55	0

Tabelle 2.1: Datenmatrix: Mitglieder einer Fußballmannschaft.

In diesem Fall werden $n = 11$ Merkmalsträger unterschieden; die Matrix hat demzufolge $n = 11$ Zeilen. Die i-te Zeile enthält den Laufindex i und die $m = 3$ erhobenen Daten für Merkmalsträger i. Die j-te Spalte enthält die $n = 11$ beobachteten Ausprägungen des j-ten Merkmals.

Sie wollen nun untersuchen, welchen Ruhepuls Fußballer haben. Aus der Datenmatrix in Tabelle 2.1 ermitteln Sie folgende Häufigkeitstabelle:

Ruhepuls	52	53	54	55	56	57	58
		||	||	||||	||	|	

Sie stellen fest: Die Merkmalsausprägung 55 kommt mit der *absoluten Häufigkeit* 4 vor. Häufig gibt man den Ausprägungen Namen, in unserem Beispiel etwa:

	53	54	55	56	57
Name	a_1	a_2	a_3	a_4	a_5

Die Ausprägungen werden dabei durch Indizes $(1, \ldots, 5)$ durchgezählt. Bei 80 möglichen Ausprägungen würde man entsprechend vorgehen und die Bezeichnungen $a_1, \ldots, a_{80}$ verwenden.

Sei allgemeiner $k \in \mathbb{N}$ die Zahl der Ausprägungen. Dann werden die einzelnen Ausprägungen mit $a_1, \ldots, a_k$ gekennzeichnet.

2.4.2 Visualisierung empirischer Daten

Sei eine Datenreihe $x_1, \ldots, x_n$ aus n Beobachtungen gegeben. Ein Merkmal liege in Ausprägungen $a_1, \ldots, a_k$ vor.

Im Fußballerbeispiel ist die Länge der Datenreihe $n = 11$; x_i, $i = 1, \ldots, n$, ist der Ruhepuls des i-ten Spielers. Die Anzahl der Merkmale ist $k = 5$, mit $\{a_1, \ldots, a_5\} = \{53, 54, 55, 56, 57\}$.

Dann sind die *absoluten Häufigkeiten* der Ausprägungen a_i gesucht: n_i ist die Anzahl der Indizes j, für die $x_j = a_i$. Die *relativen Häufigkeiten* h_i sind wie folgt definiert:

$$h_i = \frac{\text{absolute Häufigkeit von } a_i}{\text{Zahl der Beobachtungen}} = \frac{n_i}{n} \tag{2.1}$$

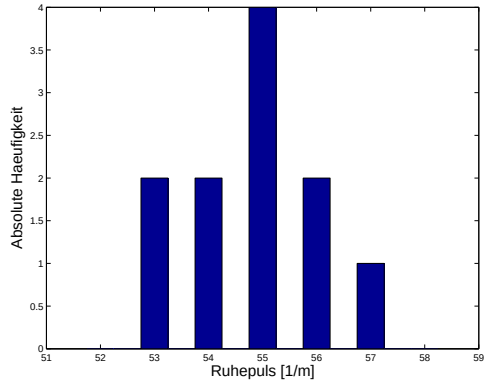
Beachte:

1. Die Summe der absoluten Häufigkeiten ist gleich der Zahl der Beobachtungen, d.h.,

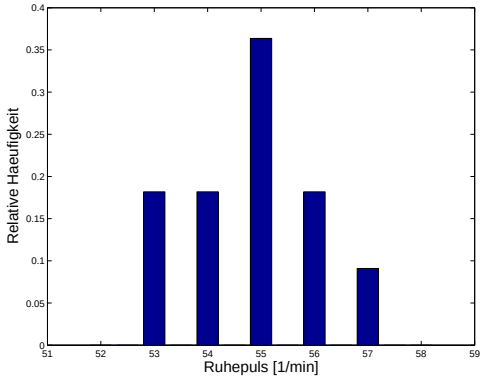
$$\sum_{i=1}^{k} n_i = n \quad (i = \text{Laufindex; } 1, k = \text{Summationsgrenzen}) \tag{2.2}$$

2. Die Summe der relativen Häufigkeiten ist 1,

$$\sum_{i=1}^{k} h_i = \frac{1}{n} \sum_{i=1}^{k} n_i = 1 \tag{2.3}$$




```
puls = [53 57 56 55 53 54 55 54 55 56 55];
n=hist (puls, [52:58]); % Elemente zählen
bar ([52:58], n); % Balkendiagramm zeichnen
```




```
puls = [53 57 56 55 53 54 55 54 55 56 55];
n=hist (puls, [52:58]);  % Elemente zählen
bar ([52:58], n/sum(n)); % Daten normalisieren und
                         % Balkendiagramm zeichnen.
```

Abbildung 2.1: Stabdiagramme der absoluten und relativen Häufigkeiten, einschließlich Matlab-Code

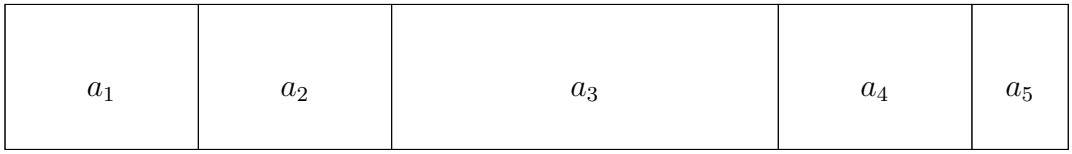


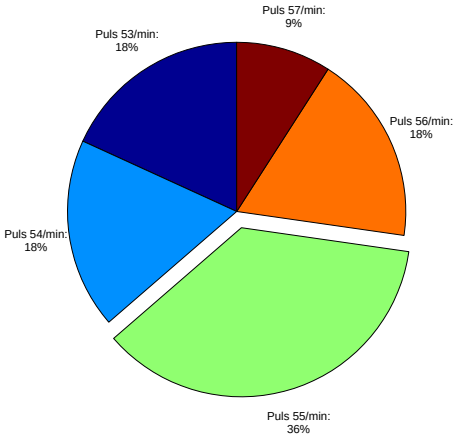
Abbildung 2.2: Balkendiagramm

Einfache Diagramme

Zur Visualisierung von absoluten bzw. relativen Häufigkeiten können verschiedene Diagrammtypen herangezogen werden. Allgemein werden bei ordinalen und quantitativen Daten die Ausprägungen entsprechend ihrer Ordnungsrelation von links nach rechts angeordnet. Bei quantitativen Daten ist es dabei sinnvoll, die Datenpunkte $x_1, \ldots, x_n$ zunächst auf der Zahlengeraden zu markieren (*Streudiagramm*); bei *zirkulären* Daten (Richtungen, Zeitangaben, Winkel) trägt man die Punkte auf einem Kreis auf.

Beim Stabdiagramm (Abbildung 2.1) wird der absoluten oder relativen Häufigkeit jeder Merkmalsausprägung a_i ein Stab der Länge h_i bzw. n_i zugeordnet; alle Stäbe haben die gleiche Breite. Das Balkendiagramm (Abbildung 2.2) ordnet der i-ten Ausprägung ein Balkensegment der Breite h_i zu. Beim Kreisdiagramm wird der i-ten Ausprägung ein Kreissektor mit dem Öffnungswinkel $\phi_i = h_i \cdot 360^o$ zugeordnet; es gilt $\sum_{i=1}^{k} \phi_i = 360^o$.

Jede dieser Visualisierungsmöglichkeiten folgt dem *Prinzip der Flächentreue*: Sollen Zahlen graphisch durch Flächenelemente visualisiert werden, so müssen die Flächen proportional zu den Zahlen gewählt werden.



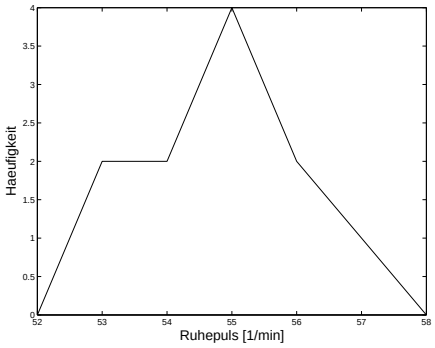

```
puls=[53 57 56 55 53 54 55 54 55 56 55];
n=hist(puls, [52:58]);  % Elemente zählen
nonnull=n(n~=0);        % Nur was vorkommt, zählt.
p=puls(nonnull);
n=n(nonnull);

explode=zeros(size(n)); % Was wird hervorgehoben?
explode(3)=1;

pie (nonnull, explode);
```

(Die Beschriftung, die im nebenstehenden Bild zu sehen ist, wurde per Hand hinzugefügt.)

Abbildung 2.3: Kuchendiagramm




```
puls = [53 57 56 55 53 54 55 54 55 56 55];
n = hist (puls, [52:58]);
plot ([52:58], n);
```

Abbildung 2.4: Häufigkeitspolygonzug

Bei stetigen Merkmalen kann die Darstellung von Häufigkeiten durch das *Häufigkeitspolygon* (Abbildung 2.4) nützlich sein: Die Punkte $(a_1, h_1), \ldots, (a_k, h_k)$ werden durch Geradenstücke verbunden. Bei stetigen Merkmalen erhält man so Hinweise auf die Häufigkeit von Zwischenwerten.

Summenhäufigkeiten und Summenhäufigkeitspolygone

Seien Ausprägungen $a_1 < \cdots < a_k$ gegeben, deren relative Häufigkeiten mit $h_1, \ldots, h_k$ und deren absolute Häufigkeiten mit $n_1, \ldots, n_k$ bezeichnet werden. Dann heißt

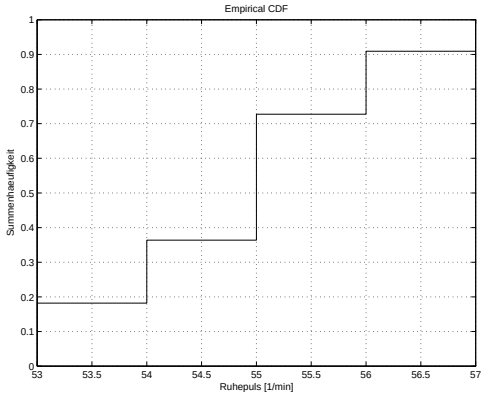
$$N_j := n_1 + n_2 + \cdots + n_j \tag{2.4}$$

die *absolute Summenhäufigkeit bis zur Ausprägung* a_j. Entsprechend wird

$$H_j := h_1 + h_2 + \cdots + h_j \tag{2.5}$$

als *relative Summenhäufigkeit bis zur Ausprägung* a_j bezeichnet.

Das *Summenhäufigkeitspolygon* bezüglich der relativen Häufigkeiten ist der Polygonzug durch die Punkte $(a_1, H_1), \ldots, (a_k, H_k)$. Dabei gilt $H_k = 1$; das Summenhäufigkeitspolygon ist monoton steigend.




```
puls = [53 57 56 55 53 54 55 54 55 56 55];
cdfplot (puls);
```

Abbildung 2.5: Summenhäufigkeitsdiagramm

Histogramme und Klasseneinteilungen

Wird eine große Zahl möglicher Ausprägungen eines Merkmals gemessen — etwa bei stetigen Merkmalen —, oder oszillieren die gemessenen Ausprägungen stark, so ist es zweckmäßig, den Wertebereich in ℓ Klassen zu unterteilen. Hierzu wird nach Abbildung 2.6 eine Treppenfunktion f mit

$$f(x) = \begin{cases} l_i = \frac{h_i}{g_{i+1} - g_i} & \text{falls } g_i < x \leq g_{i+1} \\ l_1 = \frac{h_1}{g_2 - g_1} & \text{falls } g_1 \leq x \leq g_2 \end{cases} \tag{2.6}$$

konstruiert, die *Häufigkeitsdichte* genannt wird. Die von Funktionsgraph und x-Achse eingeschlossene Fläche repräsentiert die relative Häufigkeit; die Höhe l_i repräsentiert die Dichte der Daten.

Histogramme von zirkulären Daten

Sollen zirkuläre Daten dargestellt werden, so bietet es sich an, die Daten auf einer Kreislinie einzutragen. Entsprechend kann unter Beachtung des Prinzips der Flächentreue ein zirkuläres Histogramm erzeugt werden.

Konstruktion eines Histogramms

1. Wähle $l+1$ Klassengrenzen aus:

   $$g_1 < g_2 < \cdots < g_{l+1}$$

2. Betrachte Klassen

   $$K_1 = [g_1, g_2]; \quad K_2 = (g_2, g_3]; \quad K_3 = (g_3, g_4]; \quad \ldots; \quad K_l = (g_l, g_{l+1}]$$

   Beachte: Bei der ersten Klasse zählen beide Randpunkte dazu, bei den übrigen nur der rechte.

3. Bestimme die Besetzungszahlen n_i der einzelnen Klassen:

   $$n_i = \text{Anzahl der Beobachtungen mit Werten in } K_i$$

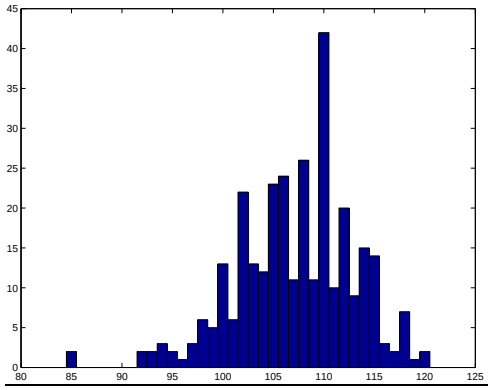
4. Ermittle relative Klassenhäufigkeiten

   $$h_i = \frac{n_i}{n}$$

5. Zeichne über der Klasse K_i ein Rechteck der *Fläche* h_i (Prinzip der Flächentreue!):

   $$\begin{aligned} h_i &= \text{Höhe} \cdot \text{Breite} = l_i \cdot (g_{i+1} - g_i) \\ \Rightarrow \; l_i &= \frac{h_i}{g_{i+1} - g_i} \end{aligned}$$

Abbildung 2.6: Konstruktion eines Histogramms



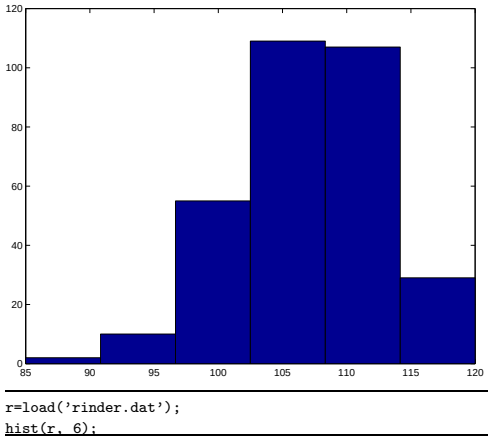
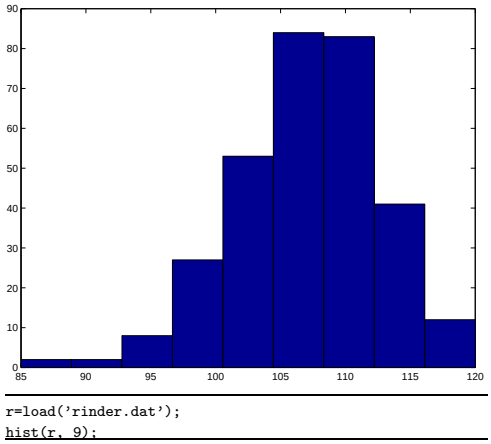


Abbildung 2.7: Häufigkeitsverteilung von Widerristhöhen in einer Herde von 312 einjährigen, weiblichen Frankenrindern. Links: Häufigkeitsverteilung aller Merkmalsausprägungen. Rechts: Histogramm mit 6 Klassen.



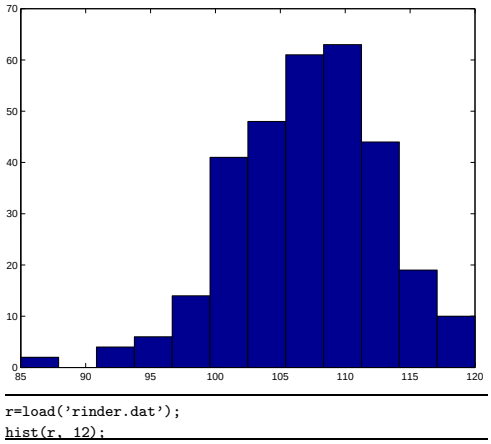
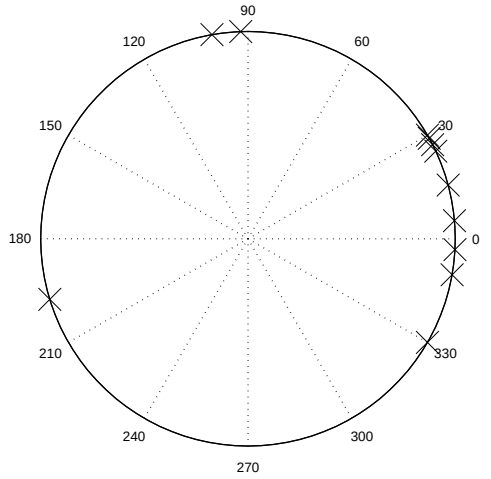
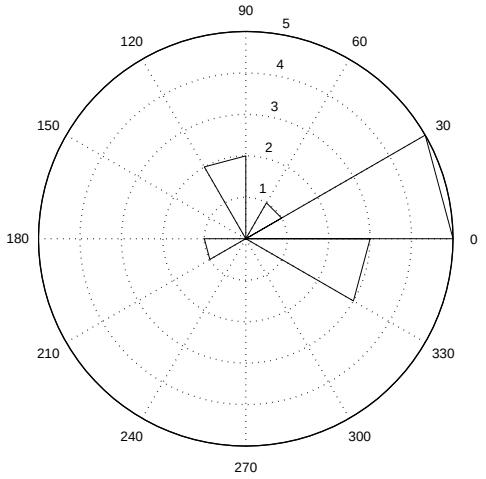


Abbildung 2.8: Widerristhöhen II: Histogramme mit 9 und 12 Klassen




```
voegel=[197 100 350 5 30 25 27 29 330 92 357 15];
phi=voegel/180 * pi; % Umrechnung in Bogenmaß.
polar (phi, ones (size (phi)), 'x');
```




```
voegel=[197 100 350 5 30 25 27 29 330 92 357 15];
phi=voegel/180 * pi; % Umrechnung in Bogenmaß.
rose (phi, 12);
```

Abbildung 2.9: Streudiagramm und Histogramm für zirkuläre Daten.

Ein Biologe möge ein Vogelnest beobachten; er interessiere sich dafür, in welche Richtung der Vogel bevorzugt davonfliegt. Eine Messreihe ergibt folgende Werte:

$$197^\circ\ 100^\circ\ 350^\circ\ 5^\circ\ 30^\circ\ 25^\circ\ 27^\circ\ 29^\circ\ 330^\circ\ 92^\circ\ 357^\circ\ 15^\circ$$

Abbildung 2.9 zeigt ein Streudiagramm und ein zirkuläres Histogramm für diese Daten.

2.5 Quantifizierung der Gestalt empirischer Verteilungen

Ziel dieses Abschnitts ist, wesentliche Charakteristika von Messreihen herauszuarbeiten, zum Beispiel:

- Lage von Messwerten — verschiedene Begriffe von „Mittelwert“,
- Streuung von Messwerten — Variabilität,
- Gestalt der Verteilung — Symmetrie oder Schiefe?

2.5.1 Lagemaße

Lagemaße sollen das Zentrum repräsentieren, um das die Daten streuen. Mathematisch gesprochen: Einem Datenvektor $(x_1, \ldots, x_n)$ soll eine Zahl $L(x_1, \ldots, x_n) \in \mathbb{R}$ zugeordnet werden, die als "Zentrum" interpretiert werden kann.

Das arithmetische Mittel

Sei eine Datenreihe $x_1, \ldots, x_n$ gegeben. Dann heißt die Zahl

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + \cdots + x_n) \tag{2.7}$$

das *arithmetische Mittel*.

Man beachte, dass die Berechnung des arithmetischen Mittels alle Beobachtungen mit dem gleichen Gewicht $\frac{1}{n}$ einbezieht. Das arithmetische Mittel minimiert außerdem die Summe der Abstandsquadrate:

$$\sum_{i=1}^{k} (x_i - \bar{x})^2 \leq \sum_{i=1}^{n} (x_i - y)^2 \qquad \text{für alle } y \in \mathbb{R}.$$

Physikalisch kann man sich das arithmetische Mittel als Schwerpunkt vorstellen: Liegen Kugeln gleicher Masse an den Stellen $x_1, \ldots, x_n$ auf einem Lineal, das von $\min\{x_i\}$ bis $\max\{x_i\}$ reicht, so ist $\bar{x}$ genau die Stelle, an der man einen Stift ansetzen muss, damit das Lineal im Gleichgewicht ist.

Legt man Merkmalsausprägungen $a_1, \ldots, a_k$ und Messwerte $x_1, \ldots, x_n$ zugrunde, die zu absoluten Häufigkeiten $n_1, \ldots, n_k$ und relativen Häufigkeiten $h_1, \ldots, h_k$ führen, so gilt:

$$\bar{x} = \sum_{i=1}^{k} h_i a_i = \frac{1}{n} \sum_{i=1}^{k} n_i a_i = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Für die Datenerhebung zum Ruhepuls der Fußballer ergibt sich somit:

$$\bar{x} = \frac{1}{11} \sum_{i=1}^{5} n_i a_i = \frac{1}{11} (2 \cdot 53 + 2 \cdot 54 + 4 \cdot 55 + 2 \cdot 56 + 57) = \frac{603}{11} \approx 54, \overline{81}$$

Verhalten des arithmetischen Mittels unter linearen Transformationen

Werden alle Daten mit einem Faktor a multipliziert — etwa bei einer Umrechnung zwischen verschiedenen Maßeinheiten, z.B. von m zu cm —, so ändert sich auch das arithmetische Mittel um denselben Faktor a: Sei $z_j = a \cdot x_j$, $j = 1, \ldots, n$. Dann ist

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i = \frac{1}{n} \sum_{i=1}^{n} a x_i = a \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) = a\bar{x}.$$

Addiert man zu allen Daten in der Liste eine feste Zahl d, so muß d auch zum Mittelwert addiert werden: Sei $z_j = d + x_j$, $j = 1, \ldots, n$. Dann ist

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} (x_i + d) = \frac{1}{n} \left\{ \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} d \right\} = \bar{x} + d$$

Man sagt: *Das arithmetische Mittel ist invariant unter linearen Transformationen.*

Robustheit bei Ausreißern

Betrachten wir als Beispiel eine Hasenpopulation in der Nähe einer Chemiefabrik; die Konzentration eines Schadstoffs werde in den Nieren von erlegten Hasen gemessen. Das Messergebnis lautet:

$$3 \quad 5 \quad 8 \quad 6 \quad 38$$

Als arithmetisches Mittel erhält man

$$\bar{x} = \frac{1}{5} (3 + 5 + 8 + 6 + 38) = 12.$$

Dieser Wert ist aber nicht für die gemessenen Schadstoffkonzentrationen charakteristisch; er entspricht nicht einer umgangssprachlichen „mittleren Schadstoffkonzentration“: Nur ein Messwert ist größer als $\bar{x}$, alle übrigen sind kleiner.

Fazit: *Das arithmetische Mittel ist nicht robust gegenüber Ausreißern.*

Der Median

Gesucht ist ein Wert M mit der Eigenschaft, dass die gleiche Anzahl von Messwerten mindestens so groß bzw. höchstens so groß wie M ist. Im Hasenbeispiel aus dem vorigen Abschnitt — mit Messwerten 3 5 6 8 38 — wäre $M = 6$.

Zu geordneten Daten $x_1 \leq x_2 \leq \cdots \leq x_n$ (beachte: Einzelne Werte können mehrmals auftreten!) definieren wir den *Median*:

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

Hierbei ist zu beachten, dass wir die aufgetretenen Messwerte und nicht etwa die verschiedenen möglichen Merkmalsausprägungen betrachten. Die gleiche Merkmalsausprägung kann also durchaus mehrfach auftreten.

Der Median ist robust unter Veränderungen der Minimal- und Maximalwerte einer Datenreihe.

Werden die Daten einer affinen, monotonen Transformation Ψ unterworfen, so gilt

$$\text{med}\{\Psi(x_1), \ldots, \Psi(x_n)\} = \Psi(\text{med}\{x_1, \ldots, x_n\}).$$

Ist Ψ monoton, jedoch nicht affin, (z.B. $\log(x)$, e^x, $\sqrt{x}$), so gilt die Aussage nur, wenn die Anzahl der Datenpunkte ungerade ist.

Der Median hat die Minimaleigenschaft

$$Q(\tilde{x}_{\text{med}}) = \sum_{i=1}^{n} |x_i - \tilde{x}_{med}| \leq Q(m) \quad \forall m \in \mathbb{R}. \tag{2.8}$$

2.5.2 Streuung

Empirische Meßwerte auf einer metrischen Skala stimmen im Allgemeinen nicht mit einem Lagemaß wie Median oder Mittelwert überein. Man sagt: *Die Meßwerte streuen um das Lagemaß.* Ziel dieses Abschnitts ist die Quantifizierung der Abweichungen vom Lagemaß durch eine Kennzahl.

Empirische Varianz und Standardabweichung

Im ersten Ansatz wählen wir das arithmetische Mittel $\bar{x}$ als Lagemaß und betrachten quadratische Abstandsmaße

$$(x_1 - \bar{x})^2, \ldots, (x_n - \bar{x})^2 .$$

Wir erhalten n Kandidaten zur Streuungsmessung, die jeweils gleich gewichtet werden sollen. Als Kennzahl ziehen wir jedoch nicht das arithmetische Mittel dieser Kandidaten heran, sondern die *empirische Varianz*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 . \tag{2.9}$$

Man beachte, dass durch $n-1$, nicht aber durch n geteilt wird; die Gründe hierfür werden in Kapitel 4.3 klar werden.

Zur Berechnung der empirischen Varianz ist der *Verschiebungssatz* hilfreich:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} x_i^2 - \frac{n}{n-1} \bar{x}^2 \tag{2.10}$$

Der Beweis erfolgt durch Nachrechnen:

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

s^2 hat einen Schönheitsfehler: Diese Größe hat nicht die Dimension der Messungen, sondern die des zugehörigen Quadrates. Wir definieren daher die *empirische Standardabweichung*:

$$s = \sqrt{s^2}. \tag{2.11}$$

Mittlere absolute Abweichung

Wird der Median als Lagemaß verwendet, so liegt es nahe, die n Abstände

$$|x_1 - \tilde{x}_{\text{med}}|, \ldots, |x_n - \tilde{x}_{\text{med}}|$$

als Kandidaten für ein Streumaß zu wählen. Mittelwertbildung führt auf die *mittlere absolute Abweichung* (*Mean Average Deviation*):

$$\text{MAD} := \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}_{\text{med}}| \tag{2.12}$$

Warnung: Im Gegensatz zum Median selbst ist MAD *nicht* robust gegenüber Ausreißerabständen $x_i - \tilde{x}_{\text{med}}$. Daher wird stattdessen auch der Median der Abweichungen als Streumaß betrachtet:

$$\text{med}\left(|x_1 - \tilde{x}_{\text{med}}|, \ldots, |x_n - \tilde{x}_{\text{med}}|\right)$$

2.5.3 Quantile

Zur Beurteilung von Laborwerten — etwa Blutdruckmessungen — wird ein Normalbereich gesucht, der einen großen Teil der Population umfaßt. Gesucht ist also ein Intervall $[a, b]$, sodass für $p \in (0, 1)$ Werte außerhalb des Intervalls (also kleiner als a oder größer als b) mit einer Wahrscheinlichkeit von $p \cdot 100\%$ auftreten.

Definition: a heißt p-Quantil, b heißt $(1-p)$-*Quantil*.

Zu einer geordneten Stichprobe $x_1 \leq \cdots \leq x_n$ und einer Zahl $p \in (0, 1)$ können wir das empirische Gegenstück, das *empirische p-Quantil* $\tilde{x}_p$ definieren:

$$\tilde{x}_p := \begin{cases} \frac{1}{2}\left(x_{np+1} + x_{np}\right), & \text{falls } np \text{ ganzzahlig} \\ x_{[np]+1}, & \text{falls } np \text{ nicht ganzzahlig} \end{cases} \tag{2.13}$$

Dabei ist die *Gauss-Klammer* $[z]$ definiert als der ganzzahlige Anteil einer reellen Zahl z.

Das 0.5-Quantil haben wir bereits als Median kennengelernt. Das 0.25-Quantil heißt *unteres Qua***r***til*, das 0.75-Quantil wird als *oberes Qua***r***til* bezeichnet. Die Differenz zwischen unterem und oberem Quartil ist als *Quartilsabstand* bekannt; er umfaßt mindestens 50% der Daten.

Beispiel: Seien 10 Beobachtungen $x_1, \ldots, x_{10}$ gegeben,

2 4 7 11 16 22 29 37 46 56

Die Quantile ergeben sich dann wie folgt:

$$\begin{aligned} p = 0.25, np = 2.5 &\Rightarrow \tilde{x}_p = x_{[np]+1} = x_3 = 7 \\ p = 0.5, np = 5 &\Rightarrow \tilde{x}_p = \frac{1}{2}(x_6 + x_5) = 19 \\ p = 0.75, np = 7.5 &\Rightarrow \tilde{x}_p = x_{[np]+1} = x_8 = 37 \end{aligned}$$

2.5.4 Box-Plots

Box-Plots ermöglichen die geschlossene graphische Darstellung wichtiger Kenngrößen, um unterschiedliche Datenerhebungen vergleichen zu können. Hierzu wird die Fünf-Punkte-Zusammenfassung herangezogen: $x_{\min}$, unteres Quartil, Median, oberes Quartil, $x_{\max}$. Die Daten aus dem vorigen Abschnitt sind in Abbildung 2.11 dargestellt.

Konstruktion eines Boxplots

1. Zeichne ein Rechteck von $x_{0.25}$ bis $x_{0.75}$.
2. Trage den Median ein.
3. Markiere die Extremalwerte $x_{\min}$ und $x_{\max}$.

Abbildung 2.10: Konstruktion eines Boxplots

2.5.5 Symmetrieeigenschaften empirischer Verteilungen

Hat das Häufigkeitspolygon einer Verteilung nur eine Spitze, so spricht man von einer *eingipfligen Verteilung*. Andernfalls nennt man die Verteilung *mehrgipflig*. (Siehe Abbildung 2.13.)

Geht das Häufigkeitspolynom nach einer Spiegelung an einer zur y-Achse parallelen Geraden wieder in sich selbst über, so heißt die Verteilung *symmetrisch*, sonst: *schief*. (Siehe Abbildung 2.14.)

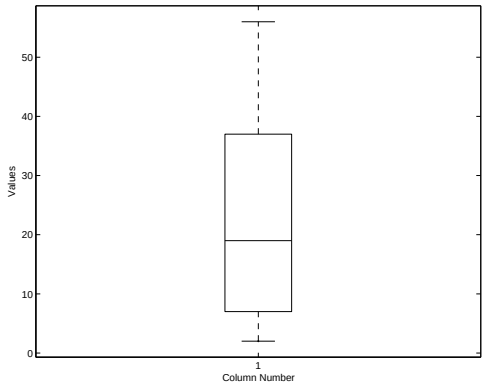
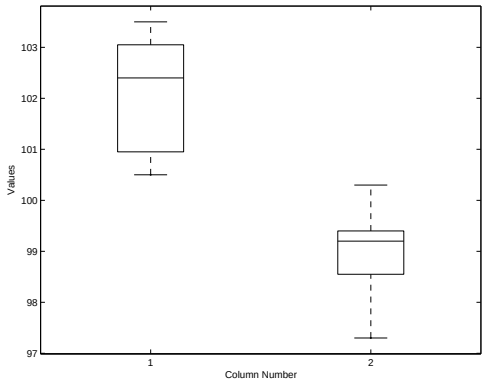


Abbildung 2.11: Boxplot-Beispiel

Körpertemperaturen in Grad Fahrenheit:

- Bei SARS-Patienten gemessen: 103.1 100.5 102.9 100.7 103.0 102.5 100.6 101.6 101.2 102.3 103.5 103.2
- Bei Gesunden gemessen: 100.3 98.3 99.3 99.7 97.8 97.3 99.3 99.2 98.8 98.9 99.2 99.5




```
t=[103.1 100.5 102.9 100.7 103.0 102.5 100.6 101.6 101.2 102.3 103.5 103.2;
   100.3 98.3 99.3 99.7 97.8 97.3 99.3 99.2 98.8 98.9 99.2 99.5]';
boxplot (t);
```

Abbildung 2.12: Boxplots zum Vergleich von Messreihen: Körpertemperaturen

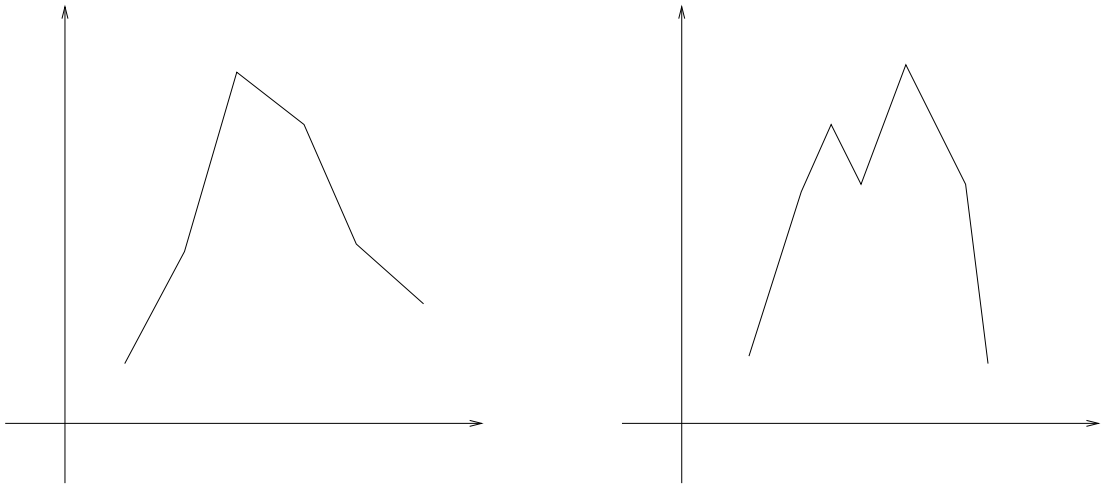


Abbildung 2.13: Ein- und mehrgipflige Verteilungen.

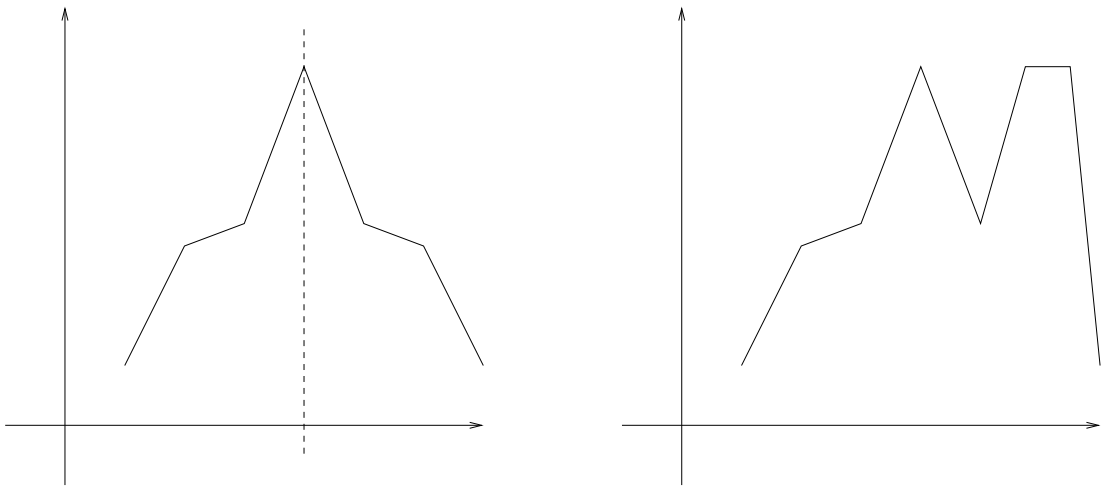


Abbildung 2.14: Symmetrische und schiefe Verteilungen.

Übungen

Aufgabe 1

In der folgenden Geschichte sind viele Zahlen enthalten. Bestimmen Sie die Merkmalsart und den Skalentyp (Nominal-, Ordinal-, Intervall- oder Verhältnisskala).

Bergtouristin X hatte in den Nachrichten um **20** Uhr gehört, dass nach **4** Tagen Regenwetter der Luftdruck nun endlich wieder über **1020**mbar steigen und die **Null**-Grad-Grenze auf **4000**m klettern werde. Als sie zum **5.** Mal in der Nacht aufwachte und nach draußen schaute, waren alle Wolken verschwunden. Sie schlug Tabelle **970** im Kursbuch auf und beschloss, den Zug um **8** Uhr **30** zu nehmen und die Gamsspitze (**3093,6**m über N.N.) zu besteigen. Auf dem Gipfel angekommen, staunte sie über das Panorama, griff zum Kompass und entdeckte in der Richtung **337**° NNW einen markanten Berg, der, wie ein Vergleich mit der Landeskarte L**9306** zeigte, nur das Weißhorn sein konnte.

Aufgabe 2

Schweinezüchter S. beobachtet folgende Anzahlen lebender Ferkel bei den Würfen seiner Sauen.

11, 11, 13, 14, 11, 8, 12, 10, 14, 14, 8, 10, 16, 13, 10, 8, 11, 12, 15, 9, 11, 7, 14, 7, 7, 11

1. Welches ist die Art des Merkmals und der Skalentyp?
2. Ermitteln Sie die absoluten und relativen Häufigkeiten der einzelnen Merkmalsausprägungen.
3. Zeichnen Sie ein Stabdiagramm und ein Kreisdiagramm der Verteilung.
4. Zeichnen Sie das Summenhäufigkeitspolygon und bestimmen Sie graphisch den Median.

Aufgabe 3

Der Weinmost von Winzerin W. erreicht in den Jahren 1990-2003 folgende Mostgewichte in Öchslegraden.

1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
95,5	59,0	82,8	101,1	77,8	59,1	72,7	105,0	76,5	90,1
2000	2001	2002	2003						
79,1	89,1	91,2	109,5						

1. Beurteilen Sie den Informationsgehalt der absoluten bzw. relativen Häufigkeit der einzelnen Merkmalsausprägungen.

2. Zeichnen Sie Histogramme mit der Klassenbreiten von 2, 5 und 10 Öchslegraden. (Ihre erste Klasse sollte bei 60 Enden.) Vergleichen Sie die Aussagekraft dieser Darstellungen.

Aufgabe 4

Auf dem Kahlen Asten wird neben zahlreichen anderen Wetterdaten auch die Windrichtung erhoben. An aufeinanderfolgenden Tagen werden dabei jeweils um 8 Uhr folgende Windrichtungen (auf ein Grad genau) gemessen. Folgende Werte traten auf (in Grad, 0° entspricht Nord).

7	59	97	5	27	190	308	352	143	78	87	344	340	8
302	85	255	340	8	160	42	26	60	305	24	172	23	357

1. Tragen Sie die Richtungen auf einem Kreis auf.

2. Zeichnen Sie ein zirkuläres Histogramm mit vier Klassen (den vier Himmelsrichtungen entsprechend).

3. Gibt es eine vorherrschende Windrichtung? Wie würden Sie diese ermitteln?

Aufgabe 5

Obstbauer O. gibt acht Testpersonen je einen Apfel seiner jüngsten Boskoop-Ernte, damit diese den Geschmack in Schulnoten (1 bis 6, in Schritten von $0,5$) bewerten. Die Ergebnisse lauten:

1 1,5 1,5 2 3 3 3,5 4.

Ein neunter Tester erscheint erst zu spät, daher liegt sein Urteil noch nicht vor.

1. Welche Möglichkeit gibt es für den Median der Noten, wenn das neunte Ergebnis hinzukommt?

2. Bestimmen Sie die beste und die schlechteste mögliche Durchschnittsnote.

Aufgabe 6

Bei einer Erhebung werden die n Werte $x_1, \ldots, x_n$ beobachtet. Dabei ergeben sich $k < n$ Merkmalsausprägungen $a_1, \ldots, a_k$ mit den relativen Häufigkeiten h_i und absoluten Häufigkeiten n_i, $i = 1, \ldots, k$. $\bar{x}$ sei ihr Mittelwert.

1. Zeigen Sie, dass für die empirische Varianz gilt

$$\frac{1}{n-1} \cdot \left(\sum_{i=1}^{k} n_i \cdot (a_i - \bar{x})^2 \right) = \frac{n}{n-1} \cdot \left(\left(\sum_{i=1}^{k} h_i \cdot a_i^2 \right) - \left(\sum_{i=1}^{k} h_i \cdot a_i \right)^2 \right)$$

2. Es werden k Paare (x_i, y_i), $i = 1, \ldots, k$ von Daten erhoben.
   Zeigen Sie: Der Mittelwert der Summen $z_i = x_i + y_i$ ist gleich der Summe $\bar{x} + \bar{y}$ der Mittelwerte.

3. Sei $\Psi : \mathbb{R} \to \mathbb{R}$ eine monoton steigende Funktion, d.h. aus $x \leq y$ folgt $\Psi(x) \leq \Psi(y)$. Ferner sei k ungerade.
   Zeigen Sie: Wird für Werte x_i, $i = 1, \ldots, k$ und für die Werte $y_i = \Psi(x_i)$ jeweils der Median $\tilde{x}_{\text{med}}$ bzw. $\tilde{y}_{\text{med}}$ bestimmt, so gilt $\tilde{y}_{\text{med}} = \Psi(\tilde{x}_{\text{med}})$.
   Gilt dies auch für monoton fallende Funktionen?

Aufgabe 7

Der kanadische SARS-Spezialist S. hat bei seinen Patienten mit Verdacht auf SARS am Einlieferungstag die Körpertemperatur gemessen. Später hat er die Messwerte danach sortiert, ob bei den Patienten SARS diagnostiziert wurde oder nicht.

Es ergaben sich folgende Messwerte in Grad Fahrenheit.

SARS	103.1	100.5	102.9	100.7	103.0	102.5	100.6	101.6	101.2
	102.3	103.5	103.2						
kein SARS	100.3	98.3	99.3	99.7	97.8	97.3	99.3	99.2	98.8

1. Welcher Anteil (in Prozent) der Patienten hatte SARS?

2. Bestimmen Sie oberes und unteres Quartil, Mittelwert, Median, empirische Varianz und Standardabweichung.

3. Nun möchte S. seine Daten europäischen Kollegen mitteilen. Berechnen Sie dazu die Größen aus Aufgabenteil b) in Grad Celsius. Wie können Sie dies besonders effizient machen und warum?

Hinweis: Um f° Fahrenheit in c° Celsius umzurechnen, benutzen Sie die Formel $c = 5/9(f - 32)$.

Aufgabe 8

Biologiestudentin B. wird von ihren zwei kleinen Geschwistern gefragt, wie lang Wattwürmer (*Arenicola marina*) werden. Jeder der drei stellt eine eigene Messreihe auf, insgesamt erheben sie die folgenden Daten.

Länge [cm]	14	15	16	17	18	19	20
abs. Häufigkeiten							
1. Messreihe	6	12	10	8	5	7	7
2. Messreihe	4	5	7	8	10	8	6
3. Messreihe	5	6	6	7	6	6	5

1. Welche Verteilungstypen (symmetrisch, schief, ein- oder mehrgipflig) repräsentieren die Häufigkeitsverteilungen der Messreihen.

2. So wie die empirische Varianz als Maß für die Streuung dienen kann, wird oft das standardisierte dritte Moment zur Messung der Schiefe benutzt, es ist für Verteilungen mit Merkmalsausprägungen $a_1, \ldots, a_k$ und relativen Häufigkeiten $h_1, ..., h_k$ als $m_3^* = \sum_{i=1}^{k} h_i \left(\frac{a_i - \bar{x}}{s}\right)^3$ definiert.
   Berechnen Sie die standardisierten dritten Momente der obigen Häufigkeitsverteilungen.

Aufgabe 9

Im „schwarzen Dreieck“, der gemeinsamen Grenzregion der Tschechischen Republik, Polens und Deutschlands werden folgende SO_2-Emissionen (in Kilotonnen) gemessen.

Jahr	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Tsch. R.	852	741	624	590	509	533	430	250	157	105	110	94
Polen	194	178	181	188	208	133	114	123	109	43	48	41
Sachsen	636	598	501	487	457	409	254	219	18	6	8	9

(Quelle: Gemeinsamer Bericht zur Luftqualität im Schwarzen Dreieck 2001, Umweltbundesamt)

Zeichnen Sie Boxplots der drei Verteilungen. Legen Sie zur besseren Vergleichbarkeit vorher eine Tabelle mit den benötigten Kenngrößen an.

Kapitel 3

Elementare Wahrscheinlichkeitstheorie

3.1 Einführung

Ziel dieses Kapitels ist die Modellierung von zufälligen Ereignissen oder von Experimenten mit zufälligem Ausgang.

Unter einem *zufälligen Ereignis* verstehen wir dabei ein Ereignis, das als Folge eines Vorgangs eintritt, dessen Ergebnis nicht vorhersagbar ist, etwa auf Grund der Komplexität des Vorgangs. Beispiele für zufällige Ereignisse sind etwa die Merkmalsvererbung, Mutationen als Fehler beim Kopieren von Erbinformationen und die Ausbreitungsrichtung von Tieren.

Wir unterscheiden *Zufallsexperimente* von *deterministischen Experimenten*: Das Ergebnis eines Zufallsexperimets ist nicht vorhersagbar; Aussagen sind lediglich über die Wahrscheinlichkeiten von Ereignissen möglich. Im Gegensatz dazu ist das Ergebnis eines deterministischen Experiments vorhersagbar.

Wir vereinbaren, dass Wahrscheinlichkeiten als Zahlen zwischen 0 und 1 oder in Prozent angegeben werden. (Siehe Abbildung 3.1.)

Beispiele zur Bestimmung von Wahrscheinlichkeiten

- Münzwurf: Es gibt zwei mögliche Ausgänge – „Kopf“ oder „Zahl“ —, die beide mit der gleichen Wahrscheinlichkeit auftreten. Dann gilt: $p = \frac{1}{2} (= 50\%)$.

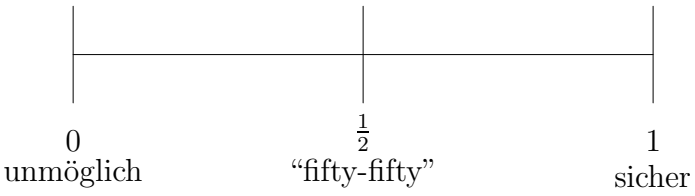


Abbildung 3.1: Wahrscheinlichkeiten als Zahlen.

- Gleichzeitiges Würfeln mit zwei perfekten Würfeln: Es gibt 36 mögliche und gleich wahrscheinliche Ausgänge, da jeder Würfel unabhängig vom anderen sechs Augenzahlen annehmen kann. Wie groß ist die Wahrscheinlichkeit, dass die Summe der Augenzahlen 3 ergibt? Es gibt zwei für dieses Ereignis günstige Ausgänge: (2, 1) und (1, 2). Die gesuchte Wahrscheinlichkeit ist also:

$$\frac{\text{Anzahl günstiger Fälle}}{\text{Anzahl möglicher Fälle}} = \frac{2}{36} = \frac{1}{18}$$

- Geschlecht von Neugeborenen: Mit der Wahrscheinlichkeit $p = 0.513$ ist das nächste in Zürich geborene Kind ein Junge. Grundlage: Auszählen aller Lebendgeburten in Zürich von 1931–1985 zeigt, dass 51,3% Jungen sind. Die Wahrscheinlichkeit ist also eine idealisierte relative Häufigkeit.

- Alltagsaussagen: „Mit 95% Wahrscheinlichkeit werde ich den 16:36 Uhr-Zug schaffen.“ Solche Aussagen sind ein rhetorisches Mittel; sie haben normalerweise keinerlei wissenschaftliche Aussagekraft.

3.2 Grundbegriffe

Wir führen nun einige elementare Grundbegriffe der Wahrscheinlichkeitstheorie ein.

Die Menge aller möglichen Ergebnisse eines Zufallsexperimentes heißt *Ergebnismenge* oder *Stichprobenraum* Ω. Elemente $\omega_i \in \Omega$ heißen *Elementarereignisse*. Teilmengen $E \subset \Omega$ heißen *Ereignisse*. Man sagt, *ein Ereignis E sei eingetreten*, wenn ein Zufallsexperiment ein Ergebnis $\omega \in E$ liefert.

Bei einem Würfel wären die Elementarereignisse $\omega_1 = 1, \ldots, \omega_6 = 6$. Als Ereignisse könnten etwa „Augenzahl gerade“ und „Augenzahl ungerade“ definiert werden, mit

$$\begin{aligned} \text{Augenzahl gerade} &= \{2,4,6\} \qquad (\subset \Omega = \{1,\ldots,6\}), \\ \text{Augenzahl ungerade} &= \{1,3,5\}. \end{aligned}$$

Das Ereignis $E = \Omega$ bezeichnen wir als das *sichere Ereignis*; $E = \emptyset$ heißt das *unmögliche Ereignis*: Alle möglichen Ergebnisse ω sind in Ω erhalten, das Ereignis Ω tritt also sicher ein. Andererseits ist kein Ergebnis ω in der leeren Menge enthalten, das Ereignis $\emptyset$ kann also niemals eintreten.

Das Ereignis $\bar{E} = \Omega \setminus E = \{\omega \in \Omega \quad : \quad \omega \notin E\}$ heißt das *zu E komplementäre Ereignis*. Ist im Würfelbeispiel A das Ereignis „Augenzahl gerade“, so ist $\bar{A}$ das Ereignis „Augenzahl ungerade“. Es gilt $A \cup \bar{A} = \Omega$.

$\bar{E}$ tritt genau dann ein, wenn E nicht eintritt.

Denn wird $\omega \in \bar{E}$ beobachtet, so ist $\omega \notin E$; ist umgekehrt $\omega \in E$, so kann nicht zugleich $\omega \notin E$ und somit $\omega \in \bar{E}$ gelten.

3.2.1 Rechnen mit Ereignissen — Bedeutung mengentheoretischer Operationen

Wie läßt sich in Mengenschreibweise formulieren,

- dass zwei Ereignisse gleichzeitig eintreten?
- dass mindestens eines von zwei Ereignissen eintritt?

Sei A das Ereignis „Augenzahl gerade“, B das Ereignis „Augenzahl ≤ 3“. Dann ist $A = \{2, 4, 6\}$, $B = \{1, 2, 3\}$.

„A und B treten gleichzeitig ein“ bedeutet also, dass das Ergebnis des Zufallsexperiments eine gerade Zahl ≤ 3 ist. Das gesuchte Ereignis ist also die Menge $\{2\} = A \cap B$.

Andererseits bedeutet „es tritt mindestens eines der Ereignisse A und B ein“, dass das Ergebnis des Zufallsexperiments gerade oder ≤ 3 ist. Das gesuchte Ereignis ist die Menge $\{1, 2, 3, 4, 6\} = A \cup B$.

Allgemeiner gilt: Seien $E \subset \Omega$ und $F \subset \Omega$ zwei Ereignisse zum Stichprobenraum Ω. Dann gilt:

- $E \cap F$ tritt genau dann ein, wenn E und F gleichzeitig eintreten.
- $E \cup F$ tritt genau dann ein, wenn E oder F gleichzeitig eintritt.

Was bedeutet $E \cap F = \emptyset$? Sei zum Beispiel $E = \{2, 4, 6\}$ das Ereignis „Augenzahl gerade“, $F = \{5\}$ das Ereignis „Augenzahl fünf“. Dann ist $E \cap F = \emptyset$; offenbar können die beiden Ereignisse nicht gleichzeitig eintreten.

Wir sagen, zwei Ereignisse E und F sind *disjunkt*, wenn sie nicht gleichzeitig eintreten können. E und F sind genau dann disjunkt, wenn $E \cap F = \emptyset$.

3.3 Axiome von Kolmogorov

Wie lassen sich nun Ereignissen Wahrscheinlichkeiten so zuordnen, dass zugleich praktikable Rechenmethoden zur Hand sind und reale Zufallsexperimente angemessen beschrieben werden können? Die Axiome von Kolmogorov liefern einen Lösungsansatz für dieses Problem.

Ein Wahrscheinlichkeitsmaß P ordnet jedem Ereignis $A \subset \Omega$ eine Zahl $P(A)$ zu, genannt *Wahrscheinlichkeit von* A, sodass folgende Eigenschaften erfüllt sind:

K1 Für alle Ereignisse $A \subset \Omega$ gilt: $0 \leq P(A) \leq 1$

K2 $P(\Omega) = 1$

K3 Sind A und B disjunkt (d.h., $A \cap B = \emptyset$), können sie also nicht gemeinsam eintreten, so gilt: $P(A \cup B) = P(A) + P(B)$.

Als Beispiel betrachten wir wieder das Würfelexperiment. Wir können den Elementarereignissen sechs Ereignisse $E_i = \{\omega_i\}$ zuordnen, konkret:

$$E_1 = \{1\} \qquad E_2 = \{2\} \qquad E_3 = \{3\} \qquad E_4 = \{4\} \qquad E_5 = \{5\} \qquad E_6 = \{6\}$$

Die Ereignisse E_i sind paarweise disjunkt, d.h.

$$E_i \cap E_j = \begin{cases} E_i, & \text{falls } i = j, \\ \emptyset & \text{sonst.} \end{cases}$$

Die Ereignisse E_i sind gleich wahrscheinlich: $P(E_i) = P(E_j)$ für alle $i, j \in \{1, \ldots, 6\}$. Wir benutzen die Kolomgorov-Axiome, um die einzelnen Wahrscheinlichkeiten auszurechnen:

$$\begin{aligned} 1 &= P(\Omega) && \text{K2} \\ &= P(E_1 \cup \cdots \cup E_6) \\ &= P(E_1) + P(E_2 \cup \cdots \cup E_6) && \text{K3: Die } E_i \text{ sind paarweise disjunkt.} \\ &= P(E_1) + \cdots + P(E_6) && \text{Wiederhole den vorigen Schritt.} \\ &= 6 \cdot P(E_i) && \text{für alle } i, \text{ da } P(E_i) = P(E_j) \text{ angenommen wurde} \end{aligned}$$

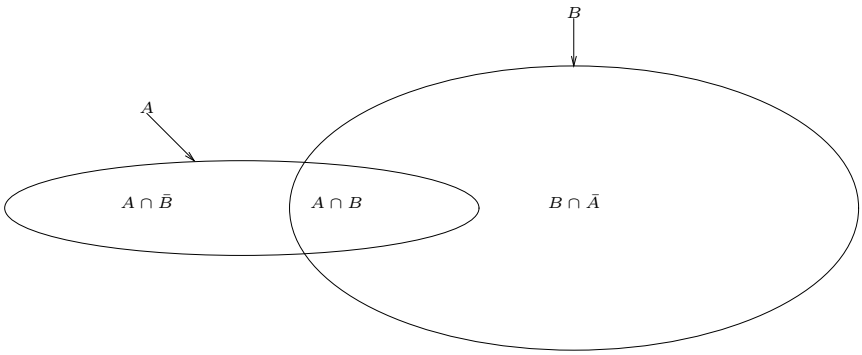


Abbildung 3.2: Venn-Diagramm: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Es gilt also $P(E_i) = \frac{1}{6}$.

Sei nun A das Ereignis „Augenzahl gerade“, also $A = \{\omega_2, \omega_4, \omega_6\} = E_2 \cup E_4 \cup E_6$. Da die E_i paarweise disjunkt sind, erhalten wir:

$$P(A) = P(E_2) + P(E_4) + P(E_6) = 3 \cdot P(E_i) = \frac{1}{2}$$

3.4 Elementare Rechenregeln für Wahrscheinlichkeiten

3.4.1 $P(\bar{A}) = 1 - P(A)$

A und $\bar{A}$ sind disjunkt, und es gilt $A \cup \bar{A} = \Omega$. Aus K2 und K3 folgt dann:

$$1 = P(\Omega) = P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

3.4.2 $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Wir zerlegen nach Abbildung 3.2 in disjunkte Teilmengen:

$$\begin{aligned} A \cup B &= (A \cap \bar{B}) \cup (A \cap B) \cup (B \cap \bar{A}) \\ A &= (A \cap \bar{B}) \cup (A \cap B) \\ B &= (B \cap \bar{A}) \cup (B \cap A) \end{aligned}$$

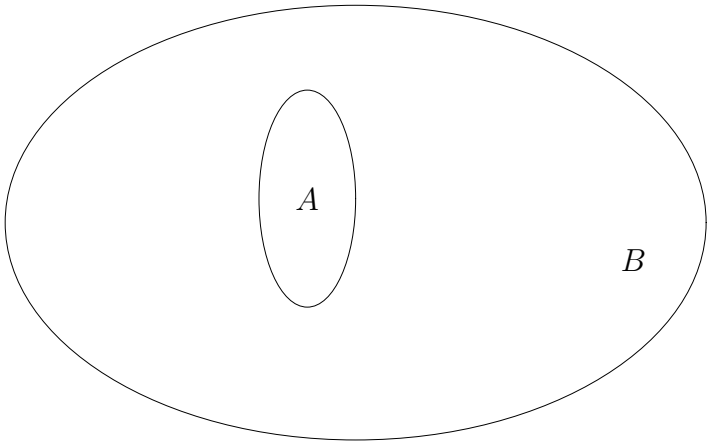


Abbildung 3.3: Venn-Diagramm zu $A \subset B \Rightarrow P(A) \leq P(B)$

Mit K3 ergibt sich hieraus:

$$\begin{aligned} P(A \cap \bar{B}) &= P(A) - P(A \cap B) \\ P(B \cap \bar{A}) &= P(B) - P(A \cap B) \end{aligned}$$

Wir haben dann:

$$\begin{aligned} P(A \cup B) &= P(A \cap \bar{B}) + P(A \cap B) + P(B \cap \bar{A}) \\ &= P(A) + P(B) - 2P(A \cap B) + P(A \cap B) = P(A) + P(B) - P(A \cap B). \end{aligned}$$

Das war zu zeigen.

Analog folgt:

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

3.4.3 $A \subset B \Rightarrow P(A) \leq P(B)$

Es gilt: $B = A \cup (B \cap \bar{A})$. Mit K3 folgt dann $P(B) = P(A) + P(B \cap \bar{A})$, also $P(B) - P(A) = P(B \cap \bar{A})$. Wegen K1 muß $P(B \cap \bar{A}) \geq 0$ gelten. Das war zu zeigen.

3.4.4 A_j paarweise disjunkt $\Rightarrow P(\bigcup A_j) = \sum P(A_j)$

Seien $A_1, \ldots, A_n$ Ereignisse, so dass sämtliche Paare A_i, A_j verschiedener Ereignisse ($i \neq j$) disjunkt sind, also $A_i \cap A_j = \emptyset$, so gilt:

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$$

3.5 Laplace-Wahrscheinlichkeiten und elementare Kombinatorik

Definition: Falls die Elementarereignisse $\omega_1, \ldots, \omega_n$ eines Stichprobenraumes $\Omega = \{\omega_1, \ldots, \omega_n\}$ *gleichwahrscheinlich sind*, d.h., $P(\{\omega_1\}) = \cdots = P(\{\omega_n\}) = \frac{1}{n}$, so heißt (Ω, P) *Laplacescher Wahrscheinlichkeitsraum.*

Als Beispiel kann ein Würfelexperiment bei einem perfekten Würfel dienen. Aus der Definition folgt unmittelbar die folgende Aussage: Sei $A \subset \Omega$. Dann gilt $P(A) = \frac{|A|}{|\Omega|}$.

Ziel dieses Abschnitts ist nun, Methoden zur Bestimmung der Kardinalität von Ereignissen zu entwickeln, so dass Wahrscheinlichkeiten im Laplace-Wahrscheinlichkeitsraum einfach berechnet werden können.

Als Modell für die Urmenge dienen n numerierte, also unterscheidbare Kugeln. Wir diskutieren nun drei verschiedene Stichprobenexperimente[1].

3.5.1 Ziehen von k Kugeln ohne Zurücklegen und unter Beachtung der Reihenfolge

Beim ersten Zug einer Kugel gibt es n Möglichkeiten, beim zweiten Zug $(n-1)$ Möglichkeiten. Es gibt also $n(n-1)$ Möglichkeiten, zwei Kugeln zu ziehen. Durch Induktion folgt, dass es

$$n \cdot (n-1) \cdot (n-2) \cdots (n-k+1) = \frac{n!}{(n-k)!}$$

[1] Für eine Betrachtung des Urnenmodells *Ziehen mit Zurücklegen und unter Beachtung der Reihenfolge* verweisen wir auf [4].

Möglichkeiten gibt, k Kugeln zu ziehen. Dabei heisst $n! = n \cdot (n-1) \cdots 2 \cdot 1$ die *n-Fakultät*; wir vereinbaren $0! = 1$.

Als Beispiel betrachten wir die Weltmeisterschaft im Eisschnellauf. Es treten 20 Teilnehmer aus 12 Nationen an. Wie viele Möglichkeiten gibt es, die ersten zehn Plätze zu vergeben? Die Antwort: $\frac{20!}{10!}$.

3.5.2 Ziehen von k Kugeln ohne Zurücklegen und ohne Beachtung der Reihenfolge

Wir wissen: Es gibt $\frac{n!}{(n-k)!}$ Möglichkeiten, k Kugeln aus n Kugeln unter Beachtung der Reihenfolge zu ziehen. Es gibt außerdem $k!$ Möglichkeiten, k Kugeln anzuordnen. Die Zahl der Möglichkeiten unter Beachtung der Reihenfolge ist dann gerade

$$k! \cdot \text{Zahl der Möglichkeiten ohne Beachtung der Reihenfolge,}$$

d.h., die Zahl der Möglichkeiten *ohne* Beachtung der Reihenfolge ist $\frac{n!}{k!(n-k)!} =: \binom{n}{k} = \binom{n}{n-k}$. Wir nennen diese Zahl den *Binomialkoeffizienten* oder kurz „n über k“.

Für ein Anwendungsbeispiel begeben wir uns wieder auf dünnes Eis: Wie wahrscheinlich ist es, dass bei der Eisschnellauf-WM allein Sportler aus der Nation D., die fünf Teilnehmer stellt, auf dem Treppchen stehen? Die Antwort ist $\binom{5}{3}$: Wähle drei aus fünf ohne Beachtung der Reihenfolge.

3.5.3 Ziehen von k Kugeln mit Zurücklegen und unter Beachtung der Reihenfolge

Beim ersten Zug bestehen n Möglichkeiten, eine Kugel zu ziehen. Ebenso beim zweiten Zug; nach zwei Zügen gibt es also n^2 Möglichkeiten. Induktiv erhalten wir n^k Möglichkeiten in k Zügen.

Als Anwendungsbeispiel betrachten wir eine Familie mit k Kindern: Wie viele Möglichkeiten gibt es für das Geschlecht der Kinder unter Beachtung der Reihenfolge der Geburten? Antwort: 2^k.

3.6 Bedingte Wahrscheinlichkeiten

Wir beginnen mit einem Anwendungsbeispiel. Sei eine Population von 1000 Individuen gegeben. Eine Krankheit tritt auf; außerdem wird festgestellt, dass eine gewisse Zahl von Individuen untergewichtig ist. Konkret ergeben sich die Zahlen aus Tabelle 3.1.

	krank	gesund	Zeilensumme
untergewichtig	400	50	450
normalgewichtig	100	450	550
Spaltensumme	500	500	

Tabelle 3.1: Krankheiten und Untergewichtigkeit

Insgesamt sind 50% der Bevölkerung erkrankt, jedoch $8/9 = 88.\bar{8}\%$ der Untergewichtigen.

Abstrakt kann man die Frage stellen: Mit welcher Wahrscheinlichkeit tritt Ereignis A ein, wenn man bereits weiß, dass B eintritt?

Wir übersetzen zunächst die relativen Häufigkeiten des Krankheitsbeispiels in die Sprache der Wahrscheinlichkeitstheorie: Seien vier mögliche Ereignisse gegeben:

- K — das Individuum ist erkrankt.
- G — das Individuum ist gesund.
- U — das Individuum ist untergewichtig.
- N — das Individuum ist nicht untergewichtig.

Offensichtlich gilt mit $\Omega = \{1000 \text{ Individuen}\}$:

$$\Omega = K \cup G = U \cup N = (K \cap U) \cup (K \cap N) \cup (G \cap U) \cup (G \cap N)$$

Es gilt:

$$P(K \cap U) = \frac{|K \cap U|}{|\Omega|} = \frac{400}{1000} = 0.4,$$

jedoch

$$\text{„Anteil der kranken Individuen unter den Untergewichtigen“} = \frac{|K \cap U|}{|U|}$$
$$= \frac{400}{450} = 0.\bar{8} = P(K|U).$$

$P(K|U)$ bezeichnen wir dabei als die „bedingte Wahrscheinlichkeit von K unter der Bedingung U."

Allgemeiner: Seien $A, B \subset \Omega$ zwei Ereignisse zum Stichprobenraum Ω, wobei $P(B) > 0$. Dann heißt $P(A|B) = \frac{P(A \cap B)}{P(B)}$ die *bedingte Wahrscheinlichkeit von A unter der Bedingung B*.

$P(A|B)$ beschreibt die Wahrscheinlichkeit, dass A eintritt, falls B eintritt.

Als weiteres Beispiel betrachten wir die Mendelsche Vererbungslehre. Es mögen zwei Genorte berücksichtigt werden. Sei $E_1 = \{\text{aa}, \text{AA}\}$ das Ereignis „Erbse ist homozygot"; sei $E_2 = \{\text{aA}, \text{Aa}, \text{AA}\}$ das Ereignis „Erbse hat Phänotyp A". Wir betrachten nun das Ereignis „Erbse ist homozygot, wenn sie vom Phänotyp A ist", d.h., $E_1|E_2$.

Es gilt:

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{\#\{\text{AA}\}}{\#\{\text{aA}, \text{Aa}, \text{AA}\}} = \frac{1}{3}$$

Insbesondere gilt im allgemeinen nicht $P(E_1|E_2) = P(E_1 \cap E_2)$ — im vorliegenden Beispiel ist $P(E_1|E_2) = \frac{1}{3}$, während $P(E_1 \cap E_2) = \frac{1}{4}$.

Um das Konzept der bedingten Wahrscheinlichkeiten besser zu verstehen, betrachten wir das gerade diskutierte Beispiel in einem Baumdiagramm, Abbildung 3.4. In diesem Baum tragen wir an jedem Knoten ein mögliches Ereignis ein; die Wurzel ist das sichere Ereignis Ω. Die von jedem Ereignis ausgehenden Zweige beschreiben weitere Ereignisse, die unter der Bedingung dieses Ereignisses eintreten können. An den Zweigen werden die bedingten Wahrscheinlichkeiten dieser Ereignisse eingetragen. Man beachte, dass die Summe der Wahrscheinlichkeiten an den Zweigen, die von einem Knoten ausgehen, immer 1 betragen muß. Um die Wahrscheinlichkeit eines Ereignisses zu bestimmen, wird ein Weg vom sicheren Ereignis entlang der Zweige des Baums gesucht; die an den Zweigen notierten Wahrscheinlichkeiten werden multipliziert.

Im konkreten Beispiel kann — zusammen mit dem sicheren Ereignis — zunächst entweder E_2 oder das komplementäre Ereignis $\bar{E}_2$ auftreten, also „Phänotyp A" oder „nicht Phänotyp A". Wir tragen also beide Ereignisse in den Baum ein und notieren an den zugehörigen Ästen ihre Wahrscheinlichkeiten unter der Bedingung, dass das sichere Ereignis eingetreten ist. (Dies sind natürlich gerade die Wahrscheinlichkeiten der Ereignisse – offensichtlich gilt $P(A) = P(A|\Omega)$).

Im nächsten Schritt fragen wir für jeden der möglichen Phänotypen, ob die betrachteten Erbsen homozygot sind: Wir fragen nach den Ereignissen $E_1 \cap E_2$ (homozygot, Phänotyp

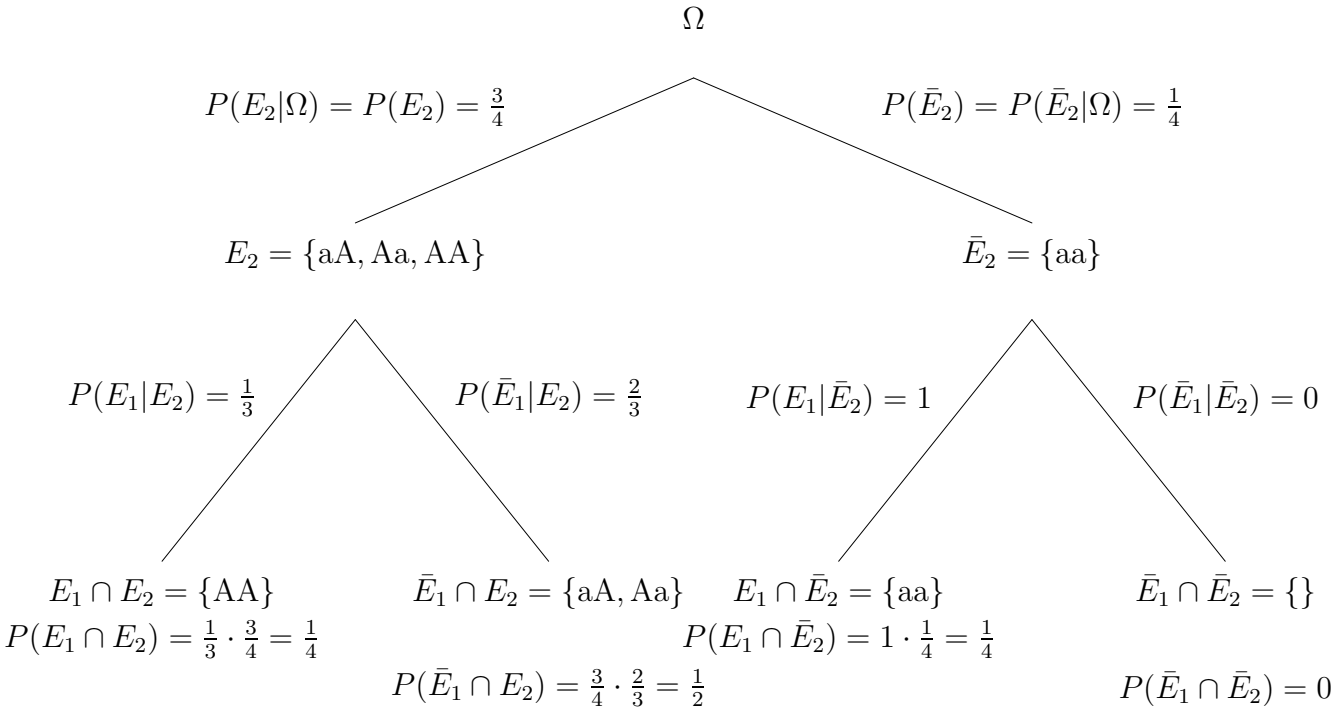


Abbildung 3.4: Bedingte Wahrscheinlichkeiten: Baumdiagramm.

A), $\bar{E}_1 \cap E_2$ (heterozygot, Phänotyp A), $E_1 \cap \bar{E}_2$ (homozygot, nicht Phänotyp A), $\bar{E}_1 \cap \bar{E}_2$ (heterozygot, nicht Phänotyp A). Alle vier Ereignisse werden ins Diagramm eingetragen und mit den Knoten für E_2 und $\bar{E}_2$ verbunden. An die Äste schreiben wir die *bedingten* Wahrscheinlichkeiten: Eine Erbse vom Phänotyp A ist mit einer Wahrscheinlichkeit von $\frac{1}{3}$ homozygot und mit einer Wahrscheinlichkeit von $\frac{2}{3}$ heterozygot. Eine Erbse, die nicht Phänotyp A ist, ist sicher homozygot.

Interessieren uns nun die Wahrscheinlichkeiten der vier kombinierten Ereignisse, so sind diese bedingten Wahrscheinlichkeiten mit den Wahrscheinlichkeiten zu multiplizieren. Diese Produkte finden sich ebenfalls in Abbildung 3.4.

Wir betrachten nun ein weiteres Beispiel für bedingte Wahrscheinlichkeiten: Farbenblindheit vs. Geschlecht; die zugrundeliegenden Informationen entnehme man Tabelle 3.2.

Wie groß ist nun die Wahrscheinlichkeit, dass ein männliches Individuum rot-grün-blind ist? Wir rechnen die bedingte Wahrscheinlichkeit aus:

$$P(B|M) = \frac{P(B \cap M)}{P(M)} = \frac{4.23}{52.71} = 0.0803 \approx 8\%$$

	Männlich (M)	Weiblich (W)	Σ
rot-grün-blind (B)	4.23%	0.65%	4.88%
normalsichtig (N)	48.48%	46.64%	95.12%
Σ	52.71%	47.29%	100.0%

Tabelle 3.2: Rot-Grün-Blindheit und Geschlecht

Wie wahrscheinlich ist es umgekehrt, dass ein rot-grün-blindes Individuum männlich ist?

$$P(M|B) = \frac{P(B \cap M)}{P(B)} = \frac{4.23}{4.88} = 0.8668 \approx 87\%$$

Wir stellen uns nun die Frage, wann $P(A|B) = P(A)$ gilt, wann also die Wahrscheinlichkeit für Ereignis A unabhängig davon ist, ob Ereignis B eintritt. Wir wissen, dass

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

Dann ist aber $P(A|B) = P(A)$ äquivalent zu $P(A \cap B) = P(A)P(B)$.

Wir sagen: Die Ereignisse $A, B \subset \Omega$ sind *unabhängig*, falls $P(A \cap B) = P(A)P(B)$. Zwei Ereignisse sind also genau dann unabhängig, wenn die Wahrscheinlichkeit, dass sie gemeinsam eintreten, durch das Produkt der Einzelwahrscheinlichkeiten gegeben wird. Entsprechend können wir eine endliche Anzahl von Ereignissen als unabhängig definieren, wenn die Produktregel für jede Teilauswahl dieser Ereignisse gilt.

Als Beispiel betrachten wir Farbenblindheit und Taubheit, siehe Tabelle 3.3.

	taub	nicht taub	Σ
farbenblind	0.04%	79.6%	8%
normalsichtig	0.46%	91.54%	92%
Σ	0.5%	99.5%	100%

Tabelle 3.3: Farbenblindheit und Taubheit

Es gilt: $P(\text{taub})P(\text{farbenblind}) = 0.5\% \cdot 8\% = 0.04\% = P(\text{taub} \cap \text{farbenblind})$. Farbenblindheit und Taubheit sind also unabhängig voneinander.

3.7 Satz von der totalen Wahrscheinlichkeit; Bayessche Formel

Wir beginnen mit einem Anwendungsbeispiel: Ein Testverfahren soll eine seltene Krankheit feststellen. Der Test ist nicht fehlerfrei, und folgende Tatsachen seien bekannt:

1. Die Krankheit tritt bei 0,1% der Population auf.
2. Bei kranken Individuen ergibt der Test mit Wahrscheinlichkeit 95% einen positiven Befund.
3. Bei gesunden Individuen ergibt der Test in 3% der Fälle einen positiven Befund.

Um zu bewerten, ob der Test auf die gesamte Population angewendet werden soll, ist folgende Frage zu beantworten: Wie groß ist die Wahrscheinlichkeit, dass ein Individuum mit positivem Testergebnis nicht krank ist?

Wir definieren die Ergebnisse K, Individuum erkrankt, und p, Testergebnis positiv. Wir wissen:

1. $P(K) = 0.001$
2. $P(p|K) = 0.95$
3. $P(p|\bar{K}) = 0.03$

Gesucht ist $P(\bar{K}|p)$.

Ein Hilfsmittel zur Beantwortung dieser Frage ist der *Satz von der totalen Wahrscheinlichkeit*: Sei $\Omega = B_1 \cup \cdots \cup B_n$, wobei die B_i paarweise disjunkt sind, d.h., $B_i \cap B_j = \emptyset$ für $i \neq j$. Ist $A \subset \Omega$ ein beliebiges Ereignis, so gilt:

$$P(A) = \sum_{j=1}^{n} P(A|B_j) \cdot P(B_j) \tag{3.1}$$

Zur Begründung bemerken wir, dass nach der Definition der bedingten Wahrscheinlichkeit gilt: $P(A|B_j)P(B_j) = P(A \cap B_j)$, wir haben also

$$\sum_{j=1}^{n} P(A|B_j)P(B_j) = \sum_{j=1}^{n} P(A \cap B_j) = P(A).$$

Die letzte Gleichheit folgt dabei aus der paarweisen Disjunktheit der B_j.

Dieser Satz stellt uns nun die für die Anfangsfragestellung wichtige *Bayessche Formel* zur Verfügung: Sei $\Omega = A_1 \cup \cdots \cup A_k$ eine disjunkte Zerlegung, d.h., $A_i \cap A_j = \emptyset$ für $i \neq j$. Dann gilt für jedes Ereignis B mit $P(B) > 0$:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} \tag{3.2}$$

Zur Herleitung notieren wir:

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

Nach Definition der bedingten Wahrscheinlichkeit ist $P(A_i \cap B) = P(B|A_i)P(A_i)$. Der Satz von der totalen Wahrscheinlichkeit impliziert $P(B) = \sum_{j=1}^{k} P(B|A_j)P(A_j)$. Setzen wir diese Beziehungen ein, so erhalten wir gerade

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)},$$

die Bayessche Formel.

Wir wenden diese Formel nun auf unser Beispiel an: $P(K) = 0.001$, $P(p|K) = 0.95$, $P(p|\bar{K}) = 0.03$. Gesucht ist $P(\bar{K}|p)$. Offenbar gilt $K \cup \bar{K} = \Omega$; $K \cap \bar{K} = \emptyset$. Die Bayessche Formel können wir also mit $A_1 = K$, $A_2 = \bar{K}$ und $B = p$ anwenden:

$$\begin{aligned} P(\bar{K}|p) &= \frac{P(p|\bar{K})P(\bar{K})}{P(p|K)P(K) + P(p|\bar{K})P(\bar{K})} = \frac{0.03 \cdot 0.999}{0.95 \cdot 0.001 + 0.03 \cdot 0.999} \\ &= \frac{0.02997}{0.00095 + 0.02997} = 0.96 \end{aligned}$$

96% der positiv getesteten wären also tatsächlich gar nicht erkrankt. Eine Anwendung des Tests auf die gesamte Population wäre nicht empfehlenswert.

3.8 Mehrstufige Bernoulli-Experimente

Definition. Zufallsexperimente, deren Stichprobenraum Ω nur zwei Elementarereignisse unfasst, heißen *Bernoulli-Experimente*.

Klassische Beispiele sind der Münzwurf (mit den Elementarereignissen „Kopf“ und „Zahl“) oder das Geschlecht von Nachkommen (mit den Elementarereignissen „männlich“ und „weiblich“).

Was passiert nun, wenn solche Experimente n-mal wiederholt werden? Falls die Experimente unabhängig sind, ergibt sich die in Abbildung 3.5 dargestellte Baumstruktur; die Elementarereignisse M und W mögen dabei die Wahrscheinlichkeiten $P(\mathrm{M}) = p$ und $P(\mathrm{W}) = 1 - p$ haben.

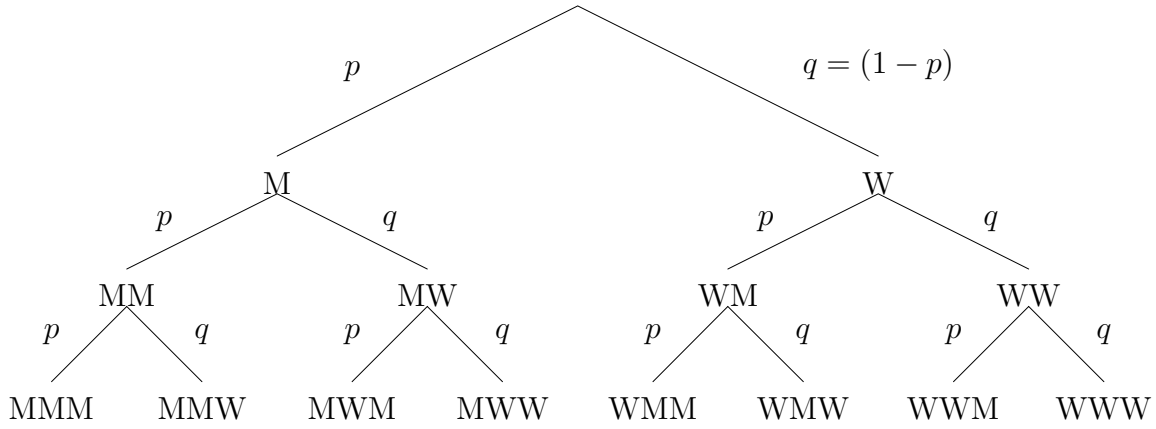


Abbildung 3.5: Entscheidungsbaum bei dreifach wiederholtem Bernoulli-Experiment.

Wie berechnet sich nun die Wahrscheinlichkeit von Elementarereignissen bei n Experimenten unter Berücksichtigung der Reihenfolge der Merkmale „M“ und „W“? Sei zum Beispiel $n = 3$. Dann haben wir:

$$\begin{aligned} P(\mathrm{MWW}) &= pq^2 \\ P(\mathrm{WWW}) &= q^3 \\ P(\mathrm{WMW}) &= pqp = pq^2 \end{aligned}$$

Sei allgemeiner

$$\omega = \left(\underbrace{\mathrm{M}, \ldots, \mathrm{M}}_{k \text{ mal}}, \underbrace{\mathrm{W}, \ldots, \mathrm{W}}_{(n-k) \text{ mal}} \right).$$

Dann gilt: $P(\omega) = p^k q^{n-k}$.

Für eine ökonomische Darstellung führen wir eine „Zählvariable“ $X : \Omega \to \mathbb{R}$ ein, die ω auf die Anzahl der Merkmale „M“ abbildet. Wir haben also:

$$X(\text{WWW}) = 0 \qquad X(\text{MWW}) = 1 \qquad X(\text{MWM}) = 2 \qquad \dots$$

Wir definieren: 1. Eine Abbildung $X : \Omega \to \mathbb{R}$, die jedem Elementarereignis $\omega \in \Omega$ eine reelle Zahl $X(\omega)$ zuordnet, heißt Zufallsvariable.

2. Mit $P(X = k)$ bezeichnet man die Wahrscheinlichkeit, dass die Zufallsvariable X den Wert k annimmt.

Die Notation $P(X = k)$ ist eine Abkürzung für $P(A_k)$, wobei das Ereignis A_k durch $A_k = \{\omega \in \Omega : X(\omega) = k\}$ definiert ist.

In unserem Beispiel mit $n = 3$ gilt:

$$\begin{aligned}
P(X = 0) &= P(\{\text{WWW}\}) = q^3 \\
P(X = 1) &= P(\{\text{MWW}\}) + P(\{\text{WMW}\}) + P(\{\text{WWM}\}) \\
&= 3pq^2 \\
P(X = 2) &= P(\{\text{MMW}\}) + P(\{\text{MWM}\}) + P(\{\text{WMM}\}) \\
&= 3p^2q \\
P(X = 3) &= P(\{\text{MMM}\}) = p^3
\end{aligned}$$

Sei nun n beliebig, $0 \le k \le n$. Dann ist $P(X = k)$ gesucht.

Zunächst sei ω ein Elementarereignis mit k mal „M“ und $(n - k)$ mal „W“. Dann ist die Wahrscheinlichkeit dieses Ereignisses

$$p^k q^{n-k}.$$

Wie viele solche Elementarereignisse gibt es nun? D.h., wie viele Möglichkeiten gibt es, k Kugeln aus n Kugeln ohne Berücksichtigung der Reihenfolge und ohne Zurücklegen zu ziehen? In Abschnitt 3.5.2 haben wir gesehen, dass die Antwort auf diese Frage $\binom{n}{k}$ lautet.

Wir erhalten damit:

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

Diese Formel gilt immer, wenn ein Bernoulli-Experiment n mal wiederholt wird.

Als Beispiel betrachten wir wieder den Mendelschen Versuch: Bei der Kreuzung homozygoter Erbsen ist die Wahrscheinlichkeit, einer runden Samenform in der F_2-Generation gleich $\frac{3}{4}$. Mendel erhielt in einem seiner Experimente 26 runde und 6 kantige Erbsen. Wie hoch war die Wahrscheinlichkeit für dieses Ereignis?

Es gilt: $n = 32$; gesucht ist die Wahrscheinlichkeit für 26 runde Erbsen.

$$P(X = 26) = \binom{32}{26} \left(\frac{3}{4}\right)^2 7 \left(\frac{1}{4}\right)^{32-26} = \frac{32 \cdot 31 \cdots 27}{2 \cdot 3 \cdots 6} \left(\frac{3}{4}\right)^{26} \left(\frac{1}{4}\right)^6 \approx 0.1249$$

3.9 Diskrete Zufallsvariable — Wahrscheinlichkeitsdichte

Definition. Eine Zufallsvariable X, die nur diskrete (d.h. isolierte) Werte annehmen kann, heißt *diskrete Zufallsvariable*.

Als Beispiel können wir die Anzahl der runden Samen uas dem obigen Beispiel betrachten — sie kann nur die diskreten Werte $x_1 = 0, \ldots, x_{33} = 32$ annehmen.

Ein weiteres Beispiel sind die Augensummen Y beim Würfeln mit zwei Würfeln. Y kann nur die Werte $y_1 = 2, y_2 = 3, \ldots, y_{11} = 12$ annehmen.

Definition. Eine Funktion f, die jedem Wert x_i, den die diskrete Zufallsvariable X annehmen kann, die Wahrscheinlichkeit $P(X = x_i)$ zuordnet, heißt *Wahrscheinlichkeitsfunktion* oder *diskrete Wahrscheinlichkeitsdichte*.

Die diskrete Zufallsvariable X möge die paarweise verschiedenen Werte x_i, $i = 1, \ldots, N$ annehmen. Da die Ereignisse $A_i = \{\omega : X(\omega) = x_i\}$ dann eine disjunkte Zerlegung von Ω bilden, gilt:

$$\sum_{i=1}^{N} f(x_i) = \sum_{i=1}^{N} P(X = x_i) = 1$$

Als Beispiel können wir ein Bernoulli-Experiment mit möglichen Ausgängen ω und $\bar{\omega}$ betrachten. Sei X die Zufallsvariable, die zählt, wie oft ω eintritt, wenn das Experiment n mal wiederholt wird.

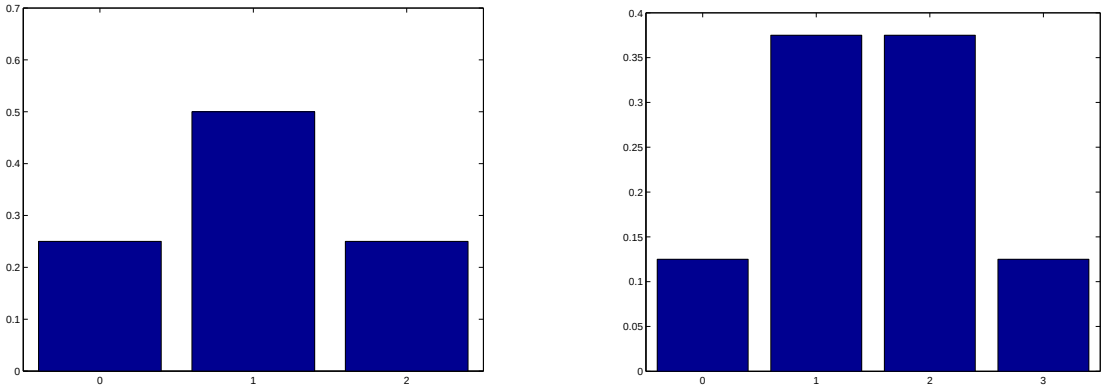


Abbildung 3.6: Binomialverteilungen zu $p = 0.5$, $n = 2, 3$.

X nimmt dann die Werte $x_0 = 0, \ldots, x_n = n$ an. Wir wissen: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

$$f(k) = B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

ist dann die diskrete Wahrscheinlichkeitsdichte von X. n gibt dabei die Anzahl der Wiederholungen an, p die Wahrscheinlichkeit für das Ereignis ω.

Definition. $B_{n,p}$ heißt *Binomialverteilung*. X heißt binomialverteilte Zufallsvariable.

Wir prüfen nach, dass $\sum_{i=0}^{n} f(i) = 1$:

$$\sum_{i=0}^{n} f(i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = (p+q)^n = 1^n = 1$$

Die entscheidende Gleichung

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

folgt dabei direkt aus dem binomischen Lehrsatz.

Als weitere Beispiele für Wahrscheinlichkeitsdichten betachten wir noch die Augensummen (bzw. die Augenzahl) beim Würfeln mit zwei (bzw. einem) Würfel(n).

Augensumme	2	3	4	5	6	7	8	9	10	11	12
Günstige Elementarereignisse	1	2	3	4	5	6	5	4	3	2	1
Wahrscheinlichkeitsdichte	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Augenzahl	1	2	3	4	5	6
Wahrscheinichkeitsfunktion	1/6	1/6	1/6	1/6	1/6	1/6

Definition. Kann eine Zufallsvariable die Werte $x_1, \ldots, x_n$ annehmen und gilt $P(X = x_i) = \frac{1}{n}$ für $i = 1, \ldots, n$, so heißt X *gleichverteilt.* Die zugehörige Wahrscheinlichkeitsdichte ist $f(x_i) = \frac{1}{n}$ für $i = 1, \ldots, n$.

3.10 Verteilung diskreter Zufallsvariablen

Welche Rückschlüsse lassen Kenngrößen einer Stichprobe auf die theoretische Verteilung der Grundgesamtheit zu?

Wir betrachten die analogen Größen bei Stichproben und theoretischen Wahrscheinlichkeitsmodellen:

Stichprobe	theoretisches Wahrscheinlichkeitsmodell
Verteilung der relativen Häufigkeiten	Wahrscheinlichkeitsdichte
Summenhäufigkeiten	Verteilungsfunktion

Definition. Sei X eine Zufallsvariable, die die Werte $x_1, \ldots, x_n$ annehmen kann. Es gelte $x_1 < x_2 < \cdots < x_n$, und f sei die zugehörige Wahrscheinlichkeitsdichte. Dann heißt

$$F(x_i) = \sum_{j=1}^{i} f(x_j) = \sum_{j=1}^{i} P(X = x_j)$$

die *Verteilungsfunktion der Zufallsvariablen* X.

Offenbar gilt $F(x_n) = 1$. Als Beispiel haben wir in Abbildung 3.7 Wahrscheinlichkeitsdichte und Verteilungsfunktion für einen Würfelwurf aufgetragen.

3.11 Erwartungswert und Varianz

Bei Stichproben kennen wir das arithemtische Mittel

$$\bar{x} = \sum_{i=i1}^{k} h_i a_i$$

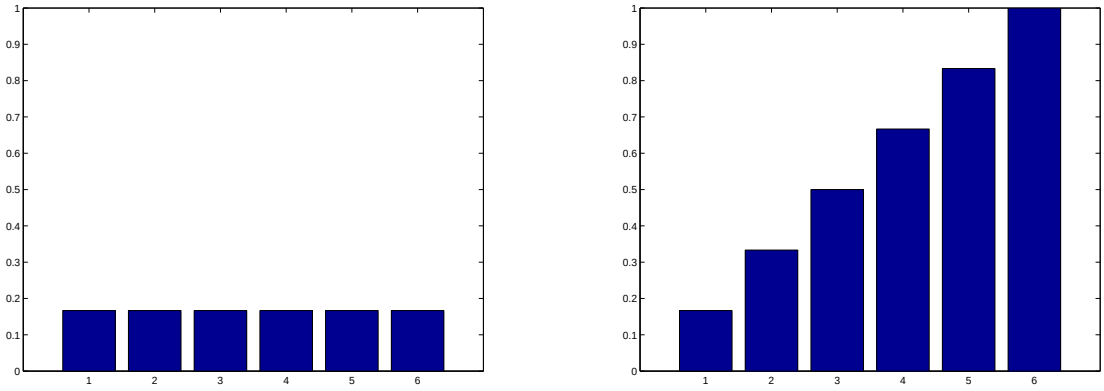


Abbildung 3.7: Wahrscheinlichkeitsdichte und Verteilungsfunktion für Würfelwurf

als Lagemaß einer Verteilung mit Ausprägungen $a_1, \ldots, a_k$ und relativen Häufigkeiten h_i.

Analog können wir einen „Durchschnittswert“ für Wahrscheinlichkeitsverteilungen definieren.

Definition. Sei X eine diskrete Zufallsvariable, die die Werte $x_1, \ldots, x_n$ annehmen kann. Dann heißt

$$E(X) := \sum_{i=1}^{k} P(X = x_i) x_i \tag{3.3}$$

Erwartungswert von X.

Im Beispiel des Würfelexperiments gilt:

$$E(X) = \sum_{i=1}^{6} P(X = i) i = \frac{1}{6}(1 + \cdots + 6) = 3.5$$

Als Maß für die Abweichung vom Mittelwert kennen wir bei Stichproben die empirische Varianz

$$s^2 = \frac{n}{n-1} \sum_{j=1}^{k} (a_j - \bar{x})^2 \, h_j.$$

Für eine Wahrscheinlichkeitsverteilung führen wir die Varianz ein.

Definition. Sei X eine diskrete Zufallsvariable, die die Werte $x_1, \ldots, x_k$ annehmen kann und den Erwartungswert $\mu = E(X)$ hat. Dann heißt

$$V(X) = \sum_{j=1}^{k} (x_j - \mu)^2 \, P(X = x_j) \tag{3.4}$$

die *Varianz von* X. Die Quadratwurzel von V heißt *Standardabweichung*.

Als Beispiel betrachten wir ein Bernoulli-Experiment. X nehme die Werte Null und Eins an, und es gelte $P(X = 1) = p$. Dann gilt:

$$\begin{aligned} \mu &= E(X) = p \cdot 1 + (1-p) \cdot 0 \\ &= p \\ V(X) &= (1-\mu)^2 P(X=1) + (0-\mu)^2 P(X=0) \\ &= (1-p)^2 p + (0-p)^2 (1-p) \\ &= p(1-p) \end{aligned}$$

3.12 Rechenregeln für Erwartungswert und Varianz

3.12.1 Summe und Produkt von Zufallsvariablen

Wir definieren zunächst Summen und Produkte von Zufallsvariablen.

Als Beispiel betrachten wir 100 Muttersauen: Seien $X_1, \ldots, X_{100}$ die Zufallsvariablen für die Anzahl der Nachkommen der i-ten Sau. Dann beschreibt die Zufallsvariable $X = \sum_{i=1}^{100}$ die Gesamtzahl der Ferkel.

Definition. Seien X und Y Zufallsvariablen, die die Werte $x_1, \ldots, x_n$ bzw. $y_1, \ldots, y_k$ annehmen können. Dann ist $Z = X + Y$ eine Zufallsvariable, die die Werte $x_i + y_j$, $i = 1, \ldots, n$, $j = 1, \ldots, k$ annehmen kann. $W = X \cdot Y$ ist die Zufallsvariable, die die Werte $x_i \cdot y_j$ annehmen kann.

Zwei Zufallsvariablen heißen *unabhängig*, falls

$$P(X = X_i \text{ und } Y = y_j) = P(X = x_i) \cdot P(Y = y_j)$$

für alle $i = 1, \ldots, n$, $j = 1, \ldots, k$ gilt.

3.12.2 Rechenregeln

Seien nun $c \in \mathbb{R}$, X, Y Zufallsvariablen. Dann gilt:

$$\begin{aligned} E(cX) &= cE(X) && (3.5)\\ E(X+Y) &= E(X)+E(Y) && (3.6)\\ V(cX) &= c^2V(X) && (3.7) \end{aligned}$$

Sind X und Y unabhängig, so folgt:

$$\begin{aligned} E(XY) &= E(X)E(Y) && (3.8)\\ V(X+Y) &= V(X)+V(Y) && (3.9) \end{aligned}$$

Als Beispiel berechnen wir Erwartungswert und Varianz der Binomialverteilung. Dazu bemerken wir, dass die Binomialverteilung zu den Parametern n und p bei Bernoulli-Experimenten interpretiert werden kann als die Wahrscheinlichkeitsverteilung einer Summe von n identisch verteilten, unabhängigen Zufallsvariablen X_i, $i = 1, \ldots, n$.

Mit $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$ folgt:

$$E(X_i) = p \qquad V(X_i) = p(1-p)$$

Für $X = \sum_{i=1}^{n} X_i$ folgt dann:

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} E(X_i) = np\\ V(X) &= \sum_{i=1}^{n} V(X_i) = np(1-p) \end{aligned}$$

3.12.3 Verschiebungssatz für die Varianz

Sei X eine diskrete Zufallsvariable, die die Werte $x_1, \ldots, x_n$ annehmen kann. Dann ist $V(X) = E(X^2) - (E(X))^2$.

Zum Beweis sei $\mu = E(X)$. Mit der zweiten binomischen Formel können wir schreiben:

$$\begin{aligned} V(X) &= \sum_{k=1}^{n}(x_k-\mu)^2 P(X=x_k) \\ &= \sum_{k=1}^{n}(x_k^2 - 2\mu x_k + \mu^2) P(X=x_k) \\ &= \underbrace{\sum_{k=1}^{n} x_k^2 P(X=x_k)}_{E(X^2)} - 2\mu \underbrace{\sum_{k=1}^{n} x_k P(X=x_k)}_{=E(X)} + \mu^2 \underbrace{\sum_{k=1}^{n} P(X=x_k)}_{=1} \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

Das war zu zeigen.

3.13 Die geometrische Verteilung

Sei eine Folge $X_1, X_2, \ldots$ von Bernoulli Experimenten gegeben; hierbei könnte es sich etwa um eine Serie von Münzwürfen handeln. In jedem Versuch wird ein Erfolg (z.B.: „Kopf“) oder Mißerfolg (z.B.: „Zahl“) beobachtet. Wie lange muß man warten, bis der erste Erfolg eintritt? Genauer: Wie viele Münzwürfe muß man im Durchschnitt durchführen, bis das erste Mal „Kopf“ erscheint?

Den Index eines Experiments können wir uns auch Zeitpunkt vorstellen. Dann fragen wir nach der durchschnittlichen Wartezeit ausgehend vom Zeitpunkt $t = 0$. Wird ein Erfolg als Elementarereignis „1“ kodiert, ein Mißerfolg als Elementarereignis „0“, so können wir auch schreiben:

$$\begin{aligned} T &= \text{Zeitpunkt des ersten Erfolges} = \min\{n \geq 1 : X_n = 1\} \\ W &= T - 1 = \text{Wartezeit auf ersten Erfolg} \end{aligned}$$

Eine Wartezeit von k Zeiteinheiten entspricht dann dem Ereignis

$$(\underbrace{0, \ldots, 0}_{k \text{ mal}}, 1),$$

also k Mißerfolgen, auf die ein Erfolg folgt.

Bezeichnen wir die für alle i identische Erfolgswahrscheinlichkeit $P(X_i = 1)$ mit p, so gilt:

$$P(X_1 = 0, \ldots, X_k = 0, X_{k+1} = 1) = \underbrace{(1-p)\cdots(1-p)}_{k \text{ mal}} \cdot p = (1-p)^k p$$

Wir haben also $P(T = k) = (1-p)^{k-1}p$ für $k \in \mathbb{N}$. Wir bestimmen nun den Erwartungswert:

$$\mu_T = E(T) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1}$$

Um diesen Ausdruck zu berechnen, greifen wir auf einen Trick aus der Differentialrechnung zurück. Und zwar ist mit $f_k(x) = -(1-x)^k$

$$f_k'(x) = k(1-x)^{k-1},$$

wir können also schreiben:

$$\mu_T = g(p) = \sum_{k=1}^{\infty} k(1-p)^{k-1} = \sum_{k=1}^{\infty} f_k'(p) = \frac{d}{dp}\left\{\sum_{k=1}^{\infty} f_k(p)\right\} = G'(p),$$

wobei $G(p) = -\sum_{k=1}^{\infty}(1-p)^k$ aus dem Grenzwert der geometrischen Reihe zu berechnen ist. Wir haben die folgende Summenformel:

$$G(q) = -\sum_{k=1}^{\infty} q^k = -\frac{q}{1-q},$$

es gilt also $G(p) = -\sum_{k=1}^{\infty}(1-p)^k = -\frac{(1-p)}{1-(1-p)} = \frac{p-1}{p}$. Wir müssen noch nach p ableiten und erhalten:

$$\mu_T = G'(p) = \left(\frac{p-1}{p}\right)' = \frac{1}{p}.$$

Für die Varianz ergibt sich:

$$V(T) = \frac{1-p}{p^2}$$

Wir tragen noch den Beweis der Summenformel für die geometrische Reihe nach: Sei $q \in (0,1)$. Dann ist $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$. Zum Beweis betrachten wir zunächst die Summenformel für Partialsummen. Und zwar sei $S_n = \sum_{k=0}^{n} q^k$. Dann ist $qS_n = \sum_{k=1}^{n+1} q^k$. Man rechnet leicht nach, dass dann $(1-q)S_n = 1 - q^{n+1}$ ist, also $S_n = \frac{1-q^{n+1}}{1-q}$. Wir führen den Grenzübergang für $n \to \infty$ durch: q^{n+1} ist wegen $q < 1$ Nullfolge, wir haben also:

$$\lim_{n\to\infty} S_n = \frac{1}{1-q} - \frac{1}{1-q} \lim_{n\to\infty} q^{k+1} = \frac{1}{1-q}$$

3.14 Die Multinomialverteilung

Wir wiederholen ein Zufallsexperiment n mal, das durch den Stichprobenraum $\Omega = \{\omega_1, \ldots, \omega_k\}$ gekennzeichnet ist. N_i, $i = 1, \ldots, k$, gebe die Anzahl der Beobachtungen des Elementarereignisses ω_i an. Wie lautet nun die Wahrscheinlichkeitsdichte für den Vektor $N = (N_1, \ldots, N_k)$, also $P(N_1 = x_1, \ldots, N_k = x_k)$? (Sei dabei $x_i \in \mathbb{R}$.)

Als Beispiel betrachten wir einen Genort mit zwei Allelen A und a, so dass sich vier Genotypen AA, Aa, aA, aa beobachten lassen. Gesucht ist dann die Wahrscheinlichkeit, dass unter n Individuen die Genotypen $\omega_1, \ldots, \omega_4$ $x_1, \ldots, x_4$ mal auftauchen, wobei $x_1, \ldots, x_4 = n$ und $x_i \in \mathbb{N} \cup \{0\}$ sein soll.

1. Wir betrachten die Wahrscheinlichkeitsdichte für das Einzelexperiment, $p_i := P(\{\omega_i\})$. Offenbar gilt $\sum_{i=1}^{k} p_i = 1$.

2. Wir bemerken weiter, dass die Wahrscheinlichkeit einer Stichprobe, bei der x_i mal die Ausprägung ω_i beobachtet wird,

$$p_1^{x_1} \cdots p_k^{x_k}$$

beträgt.

3. Wie viele Stichproben gibt es, so dass sich die Auszählung $(x_1, \ldots, x_k)$ ergibt?

- Es gibt $\binom{n}{x_1}$ Möglichkeiten, ω_1 zu beobachten.
- Es gibt $\binom{n-x_1}{x_2}$ Möglichkeiten, ω_2 zu beobachten.
- Es gibt $\binom{n-x_1-x_2}{x_3}$ Möglichkeiten, ω_3 zu beobachten.

Wir erhalten also als Anzahl der Möglichkeiten:

$$\binom{n}{x_1}\binom{n-x_1}{x_2}\cdots\binom{n-x_1-\cdots-x_{k-1}}{x_k} = \frac{n!}{x_1!x_2!\cdots x_k!}$$

Diesen Ausdruck bezeichnen wir auch als Multinomialkoeffizienten.

Für die Wahrscheinlichkeitsdichte erhalten wir entsprechend:

$$P(N = (x_1, \ldots, x_k)) = \frac{n!}{x_1!\cdots x_k!} p_1^{x_1}\cdots p_k^{x_k}$$

3.15 Poisson-Verteilung

Wir betrachten ein Beispiel aus der Ökologie: Gesucht ist das Verteilungsmuster einer bestimmten Pflanze über ein bestimmtes Gebiet (etwa ein Feld oder einen Wald). Dazu wird das Gebiet in eine große Zahl von Parzellen, etwa Quadrate oder Rechtecke gleicher Fläche, unterteilt. Man zählt die Anzahl der Pflanzen pro Parzelle.

Für eine zufällige Pflanzenverteilung könnte sich etwa die Situation aus Abbildung 3.8 ergeben.

Pfl. pro Parz.	Anz. Parz.	Pflanzen
0	1	0
1	7	7
2	7	14
3	3	9
4	2	8
Σ	20	38

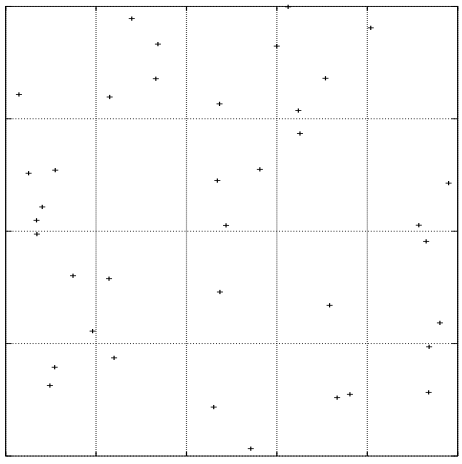


Abbildung 3.8: Poissonverteilung: Blumen.

Als wahrscheinlichkeitstheoretisches Modell können wir ansetzen, dass jede Pflanze einem Zufallsexperiment entspricht, in dem eine beliebige Parzelle — jede Parzelle entspricht einem Elementarereignis — mit Wahrscheinlichkeit p (hier: $p = \frac{1}{20}$) getroffen werden

kann. Dieses Experiment wird n mal wiederholt, und wir betrachten eine Zufallsvariable X, die angibt, wie oft eine bestimmte Parzelle getroffen wird. X kann offenbar die Werte $0, \ldots, n$ annehmen.

Wir wissen bereits, dass X $B_{n,p}$-verteilt ist, es gilt also

$$\begin{aligned} P(X = k) &= \binom{n}{k} p^k (1-p)^{n-k}, \\ E(X) &= \mu = np. \end{aligned}$$

Soll eine große Zahl von Individuen und Parzellen betrachtet werden — n sehr groß, p sehr klein — so ist $P(X = k)$ nur ungenau bestimmbar; die Berechnung ist außerdem mühselig.

Wir fragen daher, welche Grenzverteilung wir erhalten, wenn $m = np$ konstant bleibt und $n \to \infty, p = m/n \to 0$ gilt.

Wir rechnen, mit $q = 1 - p$:

$$\begin{aligned} \binom{n}{k} p^k q^{n-k} &= \binom{n}{k} \frac{p^k q^n}{q^k} = \binom{n}{k} \frac{m^k q^n}{n^k q^k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} \frac{1}{n^k q^k} m^k q^n \end{aligned}$$

Im Grenzübergang $n \to \infty$ sind $k!$ und m^k konstant; wir haben

$$q^n = (1-p)^n = \left(1 - \frac{m}{n}\right)^n.$$

Wir wissen, dass $\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = \exp(a)$, also

$$\lim_{n\to\infty} q^n = \exp(-m).$$

Wir betrachten noch den Ausdruck

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k \left(1 - \frac{m}{n}\right)^k} = \frac{n(n-1)\cdots(n-k+1)}{(n-m)^k}.$$

Zähler und Nenner sind Polynome k-ten Grades in n; multiplizieren wir aus und kürzen durch n^k, so erhalten wir in Zähler und Nenner Ausdrücke der Gestalt

$$1 + \frac{\cdots}{n} + \frac{\cdots}{n^2} + \cdots + \frac{\cdots}{n^k}$$

Diese Ausdrücke konvergieren jeweils gegen 1, wir haben also insgesamt:

$$\lim_{n\to\infty} \binom{n}{k} p^k q^{n-k} = \frac{m^k \exp(-m)}{k!}$$

Definition. Sei X eine Zufallsvariable mit der Ergebnismenge $\mathbb{N} \cup \{0\}$. Dann heißt X *Poisson-verteilt zum Parameter* m, wenn

$$P(X = k) = \frac{m^k \exp(-m)}{k!}.$$

Die Poissonverteilung zum Parameter $m = 1.9$ ist in Abbildung 3.9 dargestellt.

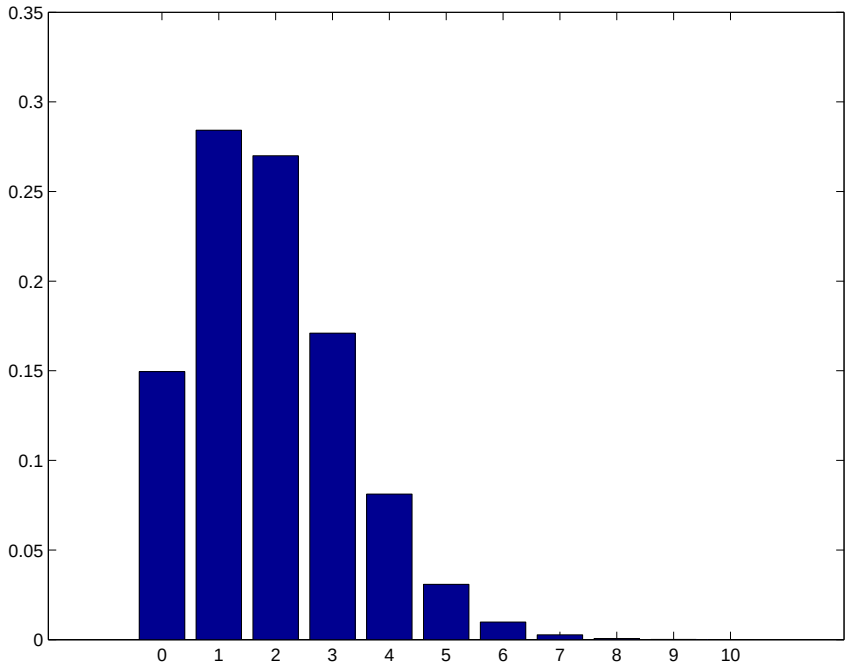


Abbildung 3.9: Poisson-Verteilung zum Parameter 1.9.

Wir zeigen noch, dass für Poisson-verteiltes X gilt:

$$\begin{aligned} E(X) &= m \\ V(X) &= m \end{aligned}$$

Zum Erwartungswert bemerken wir nur, dass m bei der Herleitung der Poisson-Verteilung bereits als Erwartungswert der einzelnen $B_{n,p}$ fixiert wurde.

```
mu=3; % Ereignisrate pro Zeiteinheit
dt=0.01; % Zeitinkrement
Tfin=50; % Ende des Beobachtungszeitraums
p=mu * dt; % Wahrscheinlichkeit eines Pulses
% pro Zeitinkrement.
Zeiten=[]; % Liste für Zeitpunkte, zu denen Pulse
% auftreten.
for t=dt/2:dt:Tfin
  if (rand < p)
    Zeiten = [Zeiten t];
  end
end
```

Abbildung 3.10: Erzeugen einer Zeitreihe

Die Varianz der Poisson-Verteilung ergibt sich aus derjenigen zu einem binomialverteilten $X_{B_{n,p}}$:

$$V(X_{B_{n,p}}) = npq = n\frac{m}{n}\left(1 - \frac{m}{n}\right) = m\left(1 - \frac{m}{n}\right) \to m$$

für $n \to \infty$.

Anwendungen findet die Poisson-Verteilung zum Beispiel bei der Beschreibung radioaktiven Zerfalls: Ein Geigerzähler zählt Pulse pro Zeitintervall; jedes Zeitintervall hat eine vorgegebene Länge.

Historisch wurde die Poisson-Verteilung entwickelt, um Todesfälle durch Hufschlag in den Kavallerie-Regimentern Napoleon Bonapartes zu modellieren.

3.15.1 Simulation und Analyse von Zeitreihen

Wir möchten eine Zeitreihe simulieren, etwa einen radioaktiven Zerfall. Dazu nutzen wir aus, dass sich die Poissonverteilung als Grenzfall von Binomialverteilungen $B_{n,p} \to P_\mu$ ergibt, $\mu = np = \text{const}$, $n \to \infty$. Genauer unterteilen wir die Zeiteinheit $[0, 1]$ in n Teilintervalle $I_1, \ldots, I_n$ der Länge dt. Sei p die Wahrscheinlichkeit eines Pulses pro Teilintervall; bei n-maliger Durchführung des Experiments erwarten wir np Pulse. Es gilt also $\mu = np$, $p = \mu \cdot |I_1| = \mu \cdot dt$. Wir nehmen an, dass dt klein gewählt ist, $dt \ll 1$.

Dann können wir eine Zeitreihe wie in Abbildung 3.10 dargestellt simulieren.

Die Vorgehensweise ist die folgende: Wir erzeugen in jedem Zeitintervall eine normalverteilte Zufallszahl r zwischen 0 und 1 und prüfen, ob $r < p$. Ist dies der Fall, so gehen wir davon aus, dass im laufenden Zeitintervall ein Puls beobachtet wurde; ist $r \geq p$, so nehmen wir an, dass kein Puls aufgetreten sei.

Es soll nun zur Klassenbreite 1 ein Histogramm gezeichnet werden; die Gesamtzahl der Pulse soll im Diagramm vermerkt werden. Der Programmcode für dieses Teilproblem ist in Abbildung 3.11 wiedergegegeben.

```
figure(1);
Klassenbreite=1;
Klassenmitten=[0.5*Klassenbreite:Klassenbreite:Tfin];
Pulse=hist(Zeiten,Klassenmitten);
bar(Klassenmitten, Pulse);
MaxP=max(Pulse);
PulseTotal=sum(Pulse);
txt=['Anzahl der Punkte =' num2str(PulseTotal)];
text (0.75 * TFin, 0.75 * MaxP, txt);
```

Abbildung 3.11: Zeitreihen: Erzeugen eines Diagramms.

Wir setzen zunächst die Klassenbreite und die Mittelpunkte der einzelnen Klassen fest. Dann wird ein Histogramm mittels `hist` erzeugt und zunächst in der Variablen `Pulse` zwischengespeichert. `bar` zeichnet das Histogramm. Um einen geeigneten Ort zum Eintragen der Gesamtzahl der Pulse zu bestimmen, finden wir durch `max(Pulse)` zunächst heraus, wie viele Pulse maximal in einem Zeitintervall beobachtet wurden; dieser Wert wird in `MaxP` abgespeichert. `sum(Pulse)` bestimmt die Gesamtzahl der Pulse; diese wird zur Konstruktion des Textes benutzt, der dann mit Hilfe des `text`-Befehls in die Graphik eingetragen wird.

Wir möchten nun noch überprüfen, ob die Zufallsvariable $\frac{\text{Anzahl Pulse}}{\text{Zeiteinheit}}$ tatsächlich näherungsweise Poisson-verteilt ist. Hierzu benutzen wir die schon eben erzeugte Liste der Pulszahlen pro Klasse (`Pulse`): Wir berechnen Mittelwert und Varianz dieser Daten. Diese Werte tragen wir in ein Histogramm von `Pulse` ein und fügen außerdem eine entsprechend skalierte Poisson-Verteilung hinzu. Der hierzu benötigte Programmcode findet sich in Abbildung 3.12.

```
figure(2);
Mittelwert=mean(Pulse);
Varianz=var(Pulse);
PKlassen=[0:MaxP];
PHist = hist (Pulse, PKlassen);
bar(PKlassen, PHist);
axis ([-1 MaxP+1 0 max(PHist)]);
hold on;
Poiss=poisspdf(PKlassen, Mittelwert);
Breite=0.5; % Breite der Bars
bar(PKlassen, PulseTotal*Poiss, Breite, 'g');

text(0.75 * MaxP, 0.875 * max(PHist), ['Mittelwert = ' num2str(Mittelwert)]);
text(0.75 * MaxP, 0.75  * max(PHist), ['Varianz    = ' num2str(Varianz)]);
```

Abbildung 3.12: Zeitreihen: Vergleich mit einer Poissonverteilung

3.16 Stetige Verteilungen

Als Beispiel betrachten wir die Verteilung des α-Globulin-Gehaltes im Blutplasma einer großen Anzahl von Personen. Der Anteil werde in $\frac{\text{g}}{100\text{ml}}$ gemessen.

In einem ersten Schritt werden die Messergebnisse durch immer feinere Histogramme zusammengefaßt. Zunächst betrachten wir sechs Klassen, dann elf.

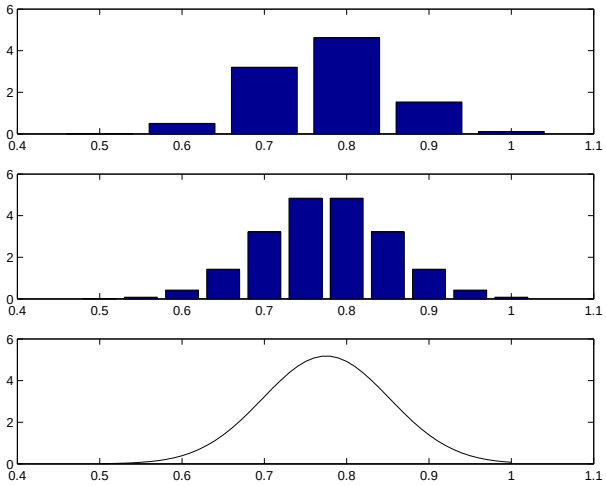


Abbildung 3.13: α-Globulin-Gehalt

Im zweiten Schritt führen wir einen Grenzübergang durch, bei dem beliebig viele, beliebig feine Klassen betrachten werden. Geht die Klassenbreite gegen Null, so erhält man im Grenzfall eine stetige Funktion. Offenbar beschreibt diese Funktion Wahrscheinlichkeiten; sie heißt deshalb *stetige Wahrscheinlichkeitsdichte*. Formal betrachten wir also eine Abbildung $f : \mathbb{R} \to \mathbb{R}_0^+$ mit $f \geq 0$, $\int_{\mathbb{R}} f = 1$.

3.16.1 Zusammenhang zwischen stetiger Wahrscheinlichkeitsdichte und tatsächlichen Wahrscheinlichkeiten

Sei X eine Zufallsvariable mit stetiger Wahrscheinlichkeitsdichte f. Dann gilt:

$$P(a \leq X \leq b) = \int_a^b f(t)\ dt$$

Geometrisch interpretieren wir also — wie schon beim Histogramm — die Wahrscheinlichkeit als Fläche zwischen x-Achse und dem Graph der Dichtefunktion.

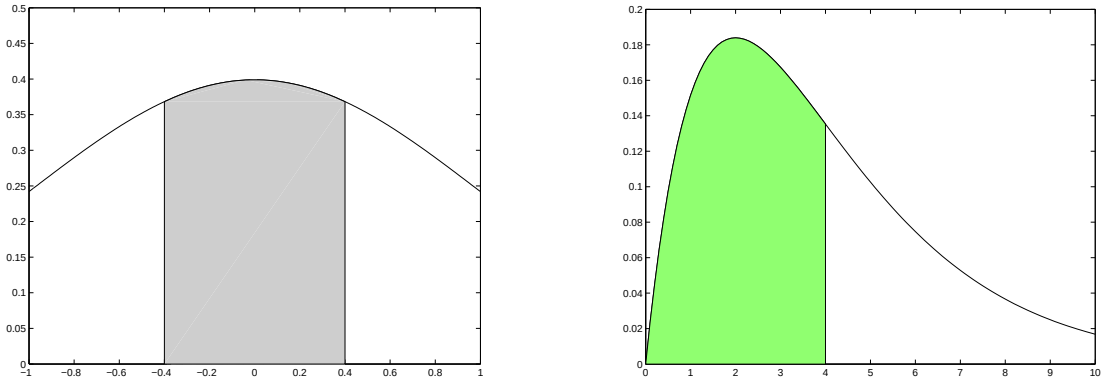


Abbildung 3.14: Geometrische Interpretation von Wahrscheinlichkeiten bei einer stetigen Wahrscheinlichkeitsdichtefunktion; geometrische Bedeutung der Verteilungsfunktion am Beispiel einer χ^2-Verteilung.

3.16.2 Bemerkungen

1. Die Wahrscheinlichkeit des Elementarereignisses $X = a$, $a \in \mathbb{R}$ ist Null,

$$P(X = a) = \int_a^a f(t)\ dt = 0$$

Diese Beobachtung ist nicht überraschend, da es auch keinen Laplace-Raum mit unendlich vielen Elementarereignissen geben kann. Die Wahrscheinlichkeit für jedes einzelne Elementarereignis wäre Null.

2. Stets muss gelten:

$$\int_{-\infty}^{\infty} f(t)\ dt = 1,$$

die Wahrscheinlichkeit des sicheren Ereignisses muss Eins sein.

3. Für stetige Zufallsvariablen lässt sich ebenso wie für diskrete Zufallsvariablen eine Verteilungsfunktion definieren. Zur Erinnerung zunächst der diskrete Fall: Nehme X die Werte $x_1 < x_2 < \cdots < x_n$ an. Dann ist $F(x_k) := \sum_{i=1}^{k} P(X = x_i) = P(X \leq x_k)$ die Verteilungsfunktion. Im stetigen Fall definieren wir analog:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\ dt$$

Die geometrische Bedeutung der Verteilungsfunktion wird in Abbildung 3.14 deutlich.

Der Hauptsatz der Differential- und Integralrechnung zeigt $F' = f$.

4. Berechnung von Wahrscheinlichkeiten mittels Verteilungsfunktion:

$$P(a \leq X \leq b) = F(b) - F(a)$$

In der Tat gilt

$$P(a \leq X \leq b) = P(X \leq b) - P(X < a) - P(X = a) = F(b) - F(a).$$

Dabei haben wir $P(X = a) = 0$ ausgenutzt.

3.17 Erwartungswert und Varianz stetiger Verteilungen

Analog zum diskreten Fall definiert man Erwartungswert und Varianz für stetige Verteilungen; f sei dabei die Wahrscheinlichkeitsdichte zur Zufallsvariablen X:

$$E(X) = \int_{-\infty}^{\infty} t f(t)\, dt \tag{3.10}$$

$$V(X) = \int_{-\infty}^{\infty} (t - E(X))^2 f(t) dt \tag{3.11}$$

Die in 3.12 vorgestellten Rechenregeln für Erwartungswert und Varianz gelten auch im stetigen Fall, insbesondere also der Verschiebungssatz:

$$V(X) = E(X^2) - (E(X))^2$$

3.18 Beispiele stetiger Verteilungen

3.18.1 Gleichverteilung

Als Experiment möge ein Ornithologe eine Reihe von Vögeln freilassen und ihre zufälligen Flugrichtungen festhalten. Der Winkel α werde relativ zur Nordrichtung beschrieben. α nehme alle Werte in $[0, 360^0)$ mit gleicher Wahrscheinlichkeit an.

Die Wahrscheinlichkeitsdichte $f : [0, 360^0) \to \mathbb{R}$ ist dann konstant. Wegen

$$\int_0^{360^0} f(t) = 1$$

gilt dann $f(\alpha) = \frac{1}{360}$ für alle α sowie $F(\alpha) = \int_0^\alpha f(s)ds = \alpha/360$.

Für eine Gleichverteilung auf $[a, b)$ erhalten wir analog:

$$f(t) = \begin{cases} \frac{1}{b-a} & t \in [a, b) \\ 0 & \text{sonst} \end{cases}$$

3.18.2 Normalverteilung

Häufig sind stetige Verteilungen glockenförmig und symmetrisch bezüglich des Erwartungswerts. Von besonderer Bedeutung ist die Normalverteilung, die wir auch als Grenzwert von Binomialverteilungen bei geeigneter Parameterwahl auffassen können.

Die generische Verteilung ist gegeben durch die Wahrscheinlichkeitsdichte

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \tag{3.12}$$

die „Gaußsche Glockenkurve."

Einige Eigenschaften der Glockenkurve:

- Der Exponent ist negativ, es gilt also $\phi(x) < 1$ für alle $x \in \mathbb{R}$.
- $\phi(x) = \phi(-x)$
- $\int_{-\infty}^{\infty} \phi(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-t^2}{2}\right) dt = 1$
- Sei ϕ Wahrscheinlichkeitsdichte zu einer Zufallsvariablen X. Dann gilt $E(X) = 0$, $V(X) = 1$.

Wir betrachten nun die allgemeine Normalverteilung: Durch Verschieben der Glockenkurve längs der x-Achse erhalten wir einen Erwartungswert $\mu \neq 0$; die zugehörige Wahrscheinlichkeitsdichte ist dann

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

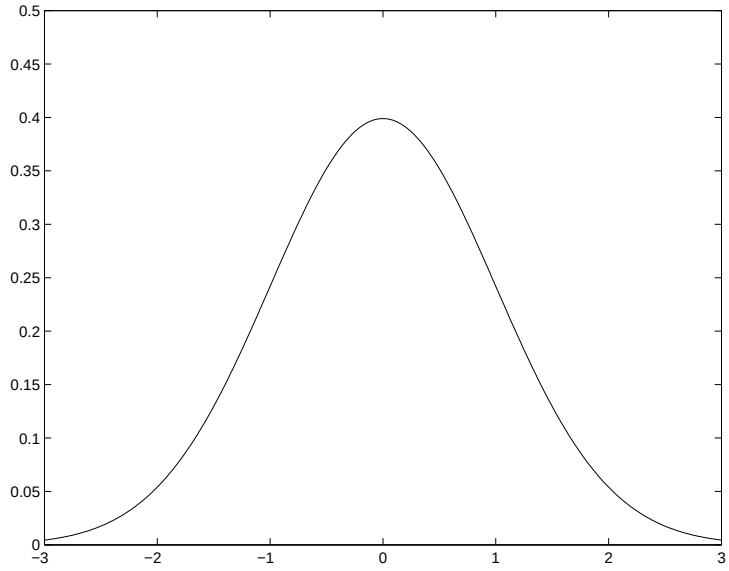


Abbildung 3.15: Gaußsche Glockenkurve zu Erwartungswert 0 und Varianz 1.

Stauchung oder Streckung der Glockenkurve mit einem Parameter $\sigma \neq 1$ ergibt die Wahrscheinlichkeitsdichte

$$\phi_{0,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Wir definieren also die Normalverteilung mit Erwartungswert μ und Parameter σ:

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

Zu $\phi_{\mu,\sigma}$ geben wir noch die Verteilungsfunktion an:

$$\Phi_{\mu,\sigma}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{y} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

Es existiert keine geschlossene Formel für Φ. Überschreitungswahrscheinlichkeiten $P(X > x) = 1 - \Phi_{0,1}(x)$ sind daher Tabellen zu entnehmen oder durch Statistikprogramme zu ermitteln. Wir diskutieren weiter unten, wie sich hieraus Werte für $\Phi_{\mu,\sigma}$ ergeben.

Welche statistische Bedeutung hat der Parameter σ in $\phi_{\mu,\sigma}$? Wächst σ, so wird die Glocke breiter; für kleines σ erhalten wir eine schmale Glocke. σ gibt also ein Maß für die Streuung

der Wahrscheinlichkeitsverteilung $\phi_{\mu,\sigma}$ an. Tatsächlich gilt für eine $\phi_{\mu,\sigma}$-verteilte Zufallsvariable:

$$E(X) = \mu \qquad V(X) = \sigma^2 \tag{3.13}$$

Der Beweis erfolgt durch Nachrechnen:

$$\begin{aligned} E(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx + \frac{\mu}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \end{aligned}$$

Aufgrund der Symmetrie des Integranden verschwindet das erste Integral. Das zweite Integral ergibt gerade $\sqrt{2\pi\sigma^2}$, wir erhalten also $E(X) = \mu$.

Die Varianz berechnen wir mit Hilfe der Substitutionsregel:

$$\begin{aligned} V(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma y^2 \exp\left(-\frac{y^2}{2}\right) dy \\ &= \sigma^2 \end{aligned}$$

Im letzten Schritt haben wir benutzt, dass die Varianz für standard-normalverteilte Zufallsverteilungen gerade 1 ist. Das war zu zeigen.

Es sei nun bekannt, dass die Zufallsvariable X μ, σ^2-normalverteilt ist (kurz: $N(\mu, \sigma^2)$); beschreibt X die Größe der Einwohner einer Stadt, so könnte etwa $\mu = 178$cm sein, $\sigma = 8$cm. Wie wird nun eine Wahrscheinlichkeit $P(a \leq X \leq b)$ aus den tabellierten Werten für $1 - \Phi_{0,1}$ berechnet?

Wir betrachten zunächst die standardisierte Zufallsvariable $Y = \frac{X-\mu}{\sigma}$. Y ist $N(0,1)$-verteilt. Es gilt dann.

$$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Als Beispiel fragen wir, wieviel Prozent der Bevölkerung in der oben angesprochenen $N(178\text{cm}, 8\text{cm})$-verteilten Stadt zwischen 186 und 190 cm groß sind. Wir haben also auszurechnen:

$$\begin{aligned} P\left(186\text{cm} < X < 190\text{cm}\right) &= P\left(\frac{186-178}{8} \leq Y \leq \frac{190-178}{8}\right) \\ &= \Phi(1.5) - \Phi(1) = 0.9332 - 0.8413 = 0.0919 \approx 9.2\% \end{aligned}$$

Seien die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig und identisch $N(\mu, \sigma^2)$-verteilt. Dann ist das arithmetische Mittel $\bar{X}$ $N(\mu, \sigma^2/n)$-verteilt. Mit Hilfe der üblichen Rechenregeln für unabhängige Zufallsvariablen X, Y und reelle Parameter a, nämlich

$$\begin{aligned} E(X+Y) = E(X) + E(Y) \qquad & V(X+Y) = V(X) + V(Y) \\ E(aX) = aE(X) \qquad & V(aX) = a^2 V(X), \end{aligned}$$

rechnen wir aus:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum X_n\right) = \frac{1}{n}\sum E(X_n) = \frac{n\mu}{n} = \mu \\ V(\bar{X}) &= V\left(\frac{1}{n}\sum X_n\right) = \frac{1}{n^2} V\left(\sum X_n\right) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Beachte: Wir haben nur Erwartungswert und Varianz berechnet. Wir haben *nicht* gezeigt, dass $\bar{X}$ tatsächlich normalverteilt ist.

Die getroffene Aussage können wir wie folgt interpretieren: Das arithmetische Mittel von n unabhängig und normalverteilten Zufallsvariablen hat den gleichen Erwartungswert wie jede einzelne dieser Zufallsvariablen, jedoch geht die Standardabweichung für $n \to \infty$ gegen Null. Je größer n wird, desto wahrscheinlicher liegt das Mittel der Beobachtungen nahe dem Erwartungswert.

Gilt ein ähnliches Resultat für beliebige Verteilungen? Diese Frage werden wir in Abschnitt 3.19 näher betrachten.

3.18.3 Exponentialverteilung: Wartezeiten

Wir betrachten den radioaktiven Zerfallsprozess aus Abschnitt 3.15, der mit Parameter $\mu = 3$ Poisson-verteilt sei. Mit welcher Wahrscheinlichkeit wird im Zeitintervall $[0, t]$,

$t < 1$, ein Puls beobachtet? Oder, äquivalent, wie groß ist die Wahrscheinlichkeit, dass die Wartezeit auf das erste Ereignis kleiner als t ist?

Die Zufallsvariable Y beschreibe also die Anzahl der Ereignisse während des Zeitintervalls $[0, 1]$, d.h., es gelte

$$P(Y = k) = \frac{\mu^k e^{-\mu}}{k!}.$$

Sei $X_t : [0, 1] \ni t \to \{0, 1\}$ eine Zufallsvariable, die für jeden Zeitpunkt $t \in [0, 1]$ markiert, ob ein Puls gemessen wird ($X_t = 1$) oder nicht ($X_t = 0$).

Die Wartezeit T auf das erste Ereignis ist dann gegeben durch

$$T = \min\{t : X_t = 1\}$$

Wie ist T verteilt? Sei dazu N_t die Anzahl der Ereignisse bis zur Zeit t. Dann ist N_t Poisson-verteilt zum Parameter μt. N_t verschwindet nun genau dann, wenn $T > t$ gilt. Es folgt, dass

$$P(T > t) = P(N_t = 0) = \frac{(\mu t)^0 e^{-\mu t}}{0!} = e^{-\mu t}$$

Mithin ergibt sich als Verteilungsfunktion von T

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - e^{-\mu t}$$

und als Wahrscheinlichkeitsdichte

$$f(t) = F'(t) = \mu e^{-\mu t}.$$

Wir sagen, T sei *exponentialverteilt mit Parameter* μ.

Es gilt:

$$E(T) = \frac{1}{\mu} \qquad V(T) = \frac{1}{\mu^2}$$

Wir möchten überprüfen, wie gut die Wartezeiten der in Abschnitt 3.15.1 erstellten Poisson-Zeitreihe exponentialverteilt sind. Dazu entwickeln wir ein Matlab-Programm,

```
Wartezeiten=diff(Zeiten);
WZMittel = mean (Wartezeiten);
WZMax = max (Wartezeiten);

figure (3)
subplot (2,1,1);
hold off;
[nWZ, tWZ] = hist (Wartezeiten, 20);
bar (tWZ, nWZ);
AnzWZ=sum(nWZ);


hold on;
tdots = [0:dt:WZMax];
Klassenbreite=tWZ(2) - tWZ(1);
Flaeche=AnzWZ*Klassenbreite;
plot (tdots, Flaeche * exppdf(tdots, WZMittel), '+g');

ymax=1.1*max([nWZ Flaeche*WZMittel])
axis ([0 WZMax 0 ymax]);

subplot (2, 1, 2);
cdfplot (Wartezeiten);
axis ([0 WZMax 0 1.1]);
hold on;
plot (tdots, expcdf(tdots, WZMittel), '+g');

hold off;
```

Abbildung 3.16: Wartezeiten: Vergleich mit einer Exponentialverteilung

das das Histogramm der Wartezeiten der Exponentialverteilung gegenüberstellt und zusätzlich die Verteilungsfunktionen vergleicht. Den Programmcode entnehme man Abbildung 3.16.

Die Wartezeiten werden aus den Differenzen zwischen den verschiedenen simulierten Pulsen erzeugt (`diff (Zeiten)`). Dann wird mit `[nWZ, tWZ] = hist (Wartezeiten, 20);` ein Histogramm mit 20 Klassen erzeugt; die Klassenmitten werden dabei in `tWz`, die Besetzungszahlen in `tWZ` abgespeichert. Histogramm und geeignet skalierte Exponentialverteilung werden dann in das mit `subplot` angelegte erste Koordinatensystem eingezeichnet.

`subplot (2, 1, 2);` wählt dann das zweite Koordinatensystem aus; dort werden die Verteilungsfunktion der Wartezeiten (mit `cdfplot`) und die Verteilungsfunktion der Exponentialverteilung (mit `expcdf` berechnet) eingezeichnet.

3.18.4 Exponentialverteilung und Überlebenszeiten

Wir betrachten das Wartezeitproblem nun aus einer anderen Perspektive: Gegeben sei eine Population, die unter Nahrungsentzug leidet — etwa Mückenlarven in einem ausgetrockneten Flußbett. Wir nehmen an, dass die Larven mit konstanter Rate α umkommen: Pro Zeiteinheit sterben αN von N Individuen. Jedes Individuum überlebt ein Zeiteinheitsintervall mit Wahrscheinlichkeit $(1 - \alpha)$.

Wir führen zunächst eine numerische Simulation durch: Betrachtet werden $N = 100$ Individuen über 30 Zeiteinheiten hinweg. Als Zeitschrittweite nehmen wir $dt = 0.1$ an; die Sterbewahrscheinlichkeit pro Zeiteinheit betrage $\alpha = 0.2$.

Wir möchten die Überlebensdauern aller Individuen auftragen. In einem separaten Plot wird die Überlebenskurve dargestellt: Die Daten aus dem ersten Bild sollen nach Überlebenszeit sortiert werden. In einer dritten Graphik schließlich wird die Populationskurve dargestellt, die Zahl der lebenden Individuen als Funktion der Zeit.

In Abbildung 3.18 ist der benutzte Programmcode dargestellt; ein mögliches Ergebnis findet sich in Abbildung 3.17.

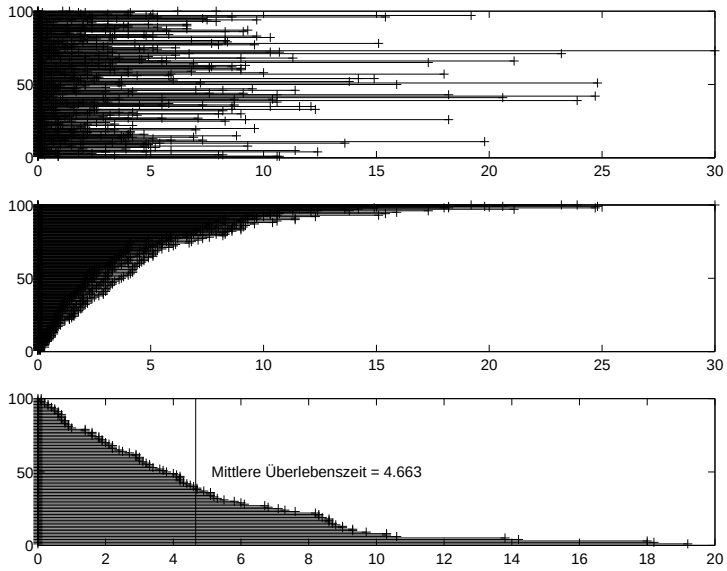


Abbildung 3.17: Simulation von Überlebenszeiten: Ausgabe

Zeichnet man in das Diagramm der Populationskurve zusätzlich den Graphen der Funktion

$$N(t) = N \exp\left(-t/t_0\right)$$

ein, wobei t_0 die mittlere Überlebenszeit (im Programmcode: `meansurv`) sei, so findet sich eine gute Übereinstimmung zwischen den beiden Kurven.

Aus der Graphik können wir ablesen, dass die mittlere Überlebenszeit mit der Zeit übereinstimmt, zu der die Populationsgröße auf den Wert N_0/e abgefallen ist. Wir be-

```
N=100;
Tfin=30 ;
dt=0.1;
alpha=0.2;
zeit=Tfin*ones(N,1);
% Ermittlung Ueberlebenszeit fuer jedes Einzelindividuum
t=0;
while (t<Tfin-dt)
    t=t+dt;
    for i=1:N
        if (zeit(i)>=Tfin) & (rand<alpha*dt)
            zeit(i)=t;
        end
    end
end
meansurv=mean(zeit);
% graphische Darstellung
figure(1)
subplot(3,1,1)
axis=([0 Tfin 0 N]);
for i=1:N
    plot([0 zeit(i)],[i i ], 'r-+')
    hold on
end

% Ueberlebenskurve
subplot(3,1,2)
ordertime=sort(zeit);
axis=([0 Tfin 0 N]);
for i=1:N
    plot([0 ordertime(i)],[i i ], 'r+-')
    hold on
end

% Populationskurve
subplot(3,1,3)
axis=([0 Tfin 0 N]);
for i=1:N
    plot([0 ordertime(N+1-i)],[i i ], 'r+-')
    hold on
end
plot ([meansurv meansurv],[0 N],'-k')
hold on
txt=['Mittlere Überlebenszeit = ' num2str(meansurv)];
text(1.1*meansurv, 0.5*N,txt)
hold off
```

Abbildung 3.18: Simulation von Überlebenszeiten

merken außerdem, dass die kontinuierliche Funktion N der folgenden gewöhnlichen Differentialgleichung genügt:

$$\frac{d}{dt}N(t) = -\frac{1}{t_0}N(t)$$

Nach Division durch N haben wir:

$$\frac{1}{N(t)}\frac{dN(t)}{dt} = -\frac{1}{t_0}.$$

$-\frac{1}{t_0}$ mißt also die mittlere Abnahme der Population pro Zeiteinheit. Wir erwarten, dass

$$\alpha \approx \frac{1}{t_0}$$

gilt. t_0^{-1} ist also ein empirischer Schätzer für die Sterberate α.

Wir betrachten nun den Zusammenhang zwischen Überlebens- und Sterbewahrscheinlichkeit:

$$P(t) = N(t)/N_0 = \exp(-\alpha t)$$

können wir als Wahrscheinlichkeit des Überlebens bis zum Zeitpunkt t interpretieren. Andererseits ist

$$Q(t) = 1 - P(t)$$

die Wahrscheinlichkeit, dass ein Individuum bis zur Zeit t bereits gestorben ist. $Q(t)$ ist also die Verteilungsfunktion der Sterbezeiten, und

$$q(t) = Q'(t) = \alpha \exp(-\alpha t)$$

die Wahrscheinlichkeitsdichte der Sterbezeiten. Die Sterbezeiten sind also exponentialverteilt mit Ereignisrate α.

3.19 Grenzwertsätze und ihre Anwendung

Die am Ende von Abschnitt 3.17 gestellte Frage wird durch den folgenden Satz beantwortet:

Gesetz der großen Zahlen. Seien X_i, $i = 1, \ldots,$ unabhängige und identisch verteilte Zufallsvariablen mit Erwartungswert $\mu = E(X_i)$ und Varianz $0 < \sigma^2 = V(X_i) < \infty$. Dann gilt mit $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ für jedes $\epsilon > 0$:

$$\lim_{n\to\infty} P\left(\mu - \epsilon \leq \bar{X}_n \leq \mu + \epsilon\right) = 1$$

Gleichgültig, wie klein ϵ wird, so ist für genügend großes n die Wahrscheinlichkeit, dass $\bar{X}_n$ Werte zwischen $\mu - \epsilon$ und $\mu + \epsilon$ annimmt, beliebig nahe an 1.

Zum Beweis benutzen wir die *Ungleichung von Tschebyscheff*: Sei X eine beliebig verteilte Zufallsvariable mit Erwartungswert μ. Dann gilt:

$$P(|X - \mu| > \epsilon) \leq \frac{V(X)}{\epsilon^2}$$

Wegen $V(\bar{X}_n) = \frac{\sigma^2}{n}$ gilt dann:

$$P(|\bar{X}_n - \mu| \leq \epsilon) = 1 - P(|X - \mu| > \epsilon) \geq 1 - \underbrace{\frac{\sigma^2}{n\epsilon^2}}_{\to 0}.$$

Das war zu zeigen.

Wir leiten noch die Tschebyscheffsche Ungleichung im diskreten Fall her. Wegen

$$V(X) = \sum_i (a_i - \mu)^2 P(X = a_i)$$

gilt folgende Abschätzung:

$$\begin{aligned} V(X) &\geq \sum_{a_i < \mu - \epsilon} (a_i - \mu)^2 P(X = a_i) + \sum_{a_i > \mu + \epsilon} (a_i - \mu)^2 P(X = a_i) \\ &\geq \epsilon^2 \sum_{a_i < \mu - \epsilon} P(X = a_i) + \epsilon^2 \sum_{a_i > \mu + \epsilon} P(X = a_i) \\ &= \epsilon^2 P(|X - \mu| > \epsilon). \end{aligned}$$

Das war zu zeigen.

Eine Anwendung der Tschebyscheffschen Ungleichung liegt in der „$k\sigma$-Prognose“: Sei ϵ ein Vielfaches der Standardabweichung, $\epsilon = k\sigma$. Dann gilt

$$P(|X - \mu| \leq k\sigma) = 1 - P(|X - \mu| > k\sigma) \geq 1 - \frac{1}{k^2}.$$

Wir können also folgern, dass der Ausgang eines Experiments mit einer Wahrscheinlichkeit von mindestens 75% in einer 2σ-Umgebung des Erwartungswerts liegt. Man beachte, dass diese Abschätzung unabhängig von der Verteilung der zugrundeliegenden Zufallsvariablen gilt.

Bei speziellen Verteilungen – etwa der Normalverteilung – können entsprechend stärkere Abschätzungen erfüllt sein.

Zentraler Grenzwertsatz. In der Praxis ist der *zentrale Grenzwertsatz* (Satz von de Moivre-Laplace) wichtig: Seien X_i, $i = 1, ...$ unabhängig und identisch verteilt mit Erwartungswert μ und Varianz $\sigma^2 \in (0, \infty)$. Dann ist das arithmetische Mittel $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$

näherungsweise normalverteilt mit Erwartungswert μ und Varianz $\frac{\sigma^2}{n}$. Insbesondere konvergiert die Verteilungsfunktion der standardisierten Zufallsvariablen

$$\bar{X}_n^* = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

gegen die Normalverteilung.

In anderen Worten: Wird eine große Anzahl von Zufallsexperimenten durchgeführt — d.h., ist die Stichprobe genügend groß —, so kann für das arithmetische Mittel Normalverteilung angenommen werden. Dies gilt insbesondere in dem Fall, dass das Resultat eines Experiments durch eine große Anzahl sich gegenseitig nicht beeinflussender Fehlerquellen additiv gestört wird.

Wir bemerken noch, dass die unbekannte Varianz σ^2 im Zentralen Grenzwertsatz durch die empirische Varianz S^2 ersetzt werden kann; es gilt also approximativ

$$\bar{X}_n^* \sim N(\mu, S^2/n),$$

d.h., X_n^* ist $\mu, S^2/n$-normalverteilt.

Wir betrachten ein Beispiel: Eine Labormaschine fülle Flüssigkeit in $n = 36$ Reagenzgläser ein. Mit einer Streuung $s = 0.12$ erreicht die Maschine einen mittleren Abfüllwert von $\mu = 1$g. Die 36 Proben werden nun in einen einzigen Erlemeyerkolben gefüllt. Mit welcher Wahrscheinlichkeit weicht die Endmenge um höchstens 1g vom Zielwert ab?

Mögen die Zufallsvariablen X_i, $i = 1, ..., 36$ die Füllmengen in den 36 Reagenzgläsern beschreiben. Dann beschreibt die Zufallsvariable $Y = X_1 + \cdots + X_n$ die Endmenge. Wir haben also $P(35 \leq Y \leq 37)$ zu berechnen.

Wegen $\mu = 1$ und $2 = 0.12$ gilt:

$$\begin{aligned} P(35 \leq Y \leq 37) &= P\left(\frac{35}{36} \leq \bar{X}_n \leq \frac{37}{36}\right) \\ &= P\left(\frac{\frac{35}{36} - 1}{0.12/6} \leq \frac{\bar{X}_n - \mu}{s/\sqrt{n}} \leq \frac{\frac{37}{36} - 1}{0.12/6}\right) \\ &\approx P(-1.389 \leq \bar{X}_n^* \leq 1.389) \approx 0.8354 \end{aligned}$$

$\bar{X}_n^*$ ist dabei $N(0, 1)$-verteilt.

3.20 Multivariate Zufallsvariablen

3.20.1 Diskreter Fall

Häufig werden mehrere Variablen $X_1, \ldots, X_p$ gleichzeitig an einer Versuchseinheit beobachtet. Zum Beispiel mögen gleichzeitig p Würfel geworfen werden, deren Augenzahlen dann durch die Zufallsvariablen $X_1, \ldots, X_p$ bezeichnet werden könnten. Die Resultate dieses Experiments lassen sich als *Zufallsvektor*

$$\begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_p \end{pmatrix}$$

darstellen. Wie sieht dann die gemeinsame Verteilung aus?

Zur Vereinfachung betrachten wir zunächst eine bivariate Zufallsvariable, sei also $p = 2$. Konkret seien Zufallsvariablen X und Y gegeben mit Wertebereichen $\chi_X = \{a_1, \ldots, a_k\}$, $\chi_Y = \{b_1, \ldots, b_\ell\}$. Dann wird die gemeinsame Verteilung durch die Wahrscheinlichkeiten $p_{ij} = P(X = a_i, Y = b_j)$ gegeben mit $i = 1, \ldots, k$ und $j = 1, \ldots, \ell$. Zur graphischen Darstellung kann man etwa Säulen der Höhe p_{ij} über den Ausprägungen (a_i, b_j) benutzen.

Die Verteilungen der einzelnen Komponenten heißen *Randverteilungen*. Die Randverteilung von X erhält man durch Aufsummieren über Y, d.h.,

$$P(X = a_i) = p_{i1} + \cdots + p_{i\ell}.$$

3.20.2 Stetiger Fall

Ein Zufallsvektor (X, Y) heißt stetig verteilt, wenn es eine Dichtefunktion $f(x, y)$ gibt mit der folgenden Eigenschaft: Die Wahrscheinlichkeit, dass (X, Y) Werte in einem Rechteck $[a, b] \times [c, d]$ der xy-Ebene annimmt, wird durch das Volumen gegeben, das durch $f(x, y)$ und das Rechteck beschrieben wird:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dx dy$$

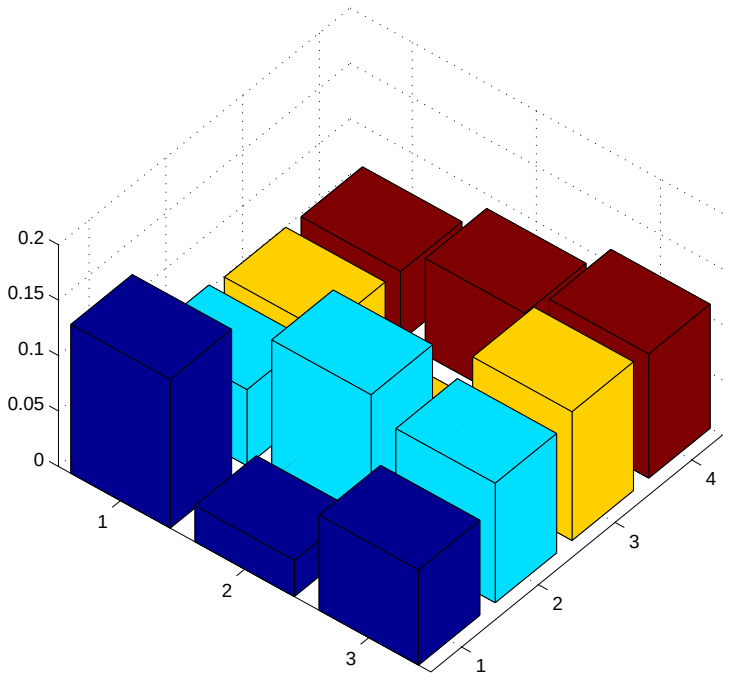


Abbildung 3.19: Darstellung einer bivariaten Zufallsvariablen durch ein Balkendiagramm

Die Randdichten sind dann durch

$$f_X(x) = \int f(x,y)dy \qquad \text{und} \qquad f_Y(y) = \int f(x,y)dx$$

gegeben.

X und Y sind genau dann unabhängig, wenn

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

gilt.

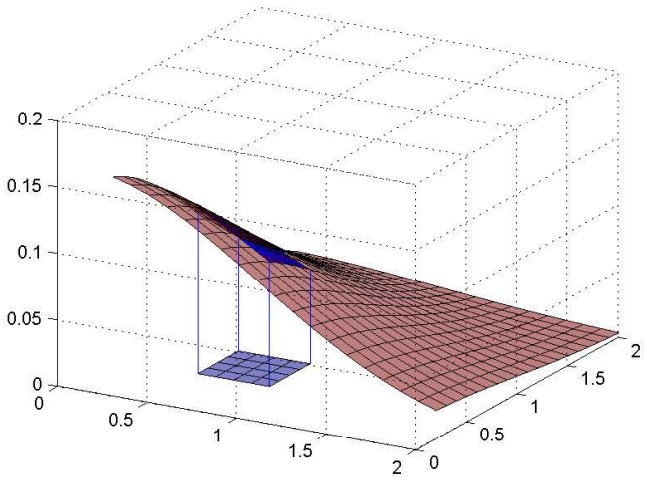


Abbildung 3.20: Stetige Wahrscheinlichkeitsverteilung einer bivariaten Zufallsvariablen.

Übungen

Aufgabe 1

Von 20 bekannten Biologen werden die folgenden Geburtstage auf je ein Los geschrieben.

16. Januar,	25. Januar,	3. Februar,	12. Februar,	16. Februar,
17. Februar,	14. März,	5. April,	6. April,	4. Mai,
27. Mai,	8. Juni,	20 Juli,	22. Juli,	10. September,
13. Oktober,	7. Dezember,	11. Dezember,	15. Dezember,	27. Dezember.

Es wird nun ein Los gezogen.

Betrachten Sie im Folgenden die Ereignisse

$A := \{$Geburtstage, die im Dezember liegen$\}$,

$B := \{$Geburtstage, die in der ersten Jahreshälfte liegen$\}$,

$C := \{$Geburtstage, die in einem Monat liegen, der den Buchstaben „**i**“ enthält$\}$.

1. Bestimmen Sie - elementweise und deskriptiv - die folgenden Ereignisse: $A \cap B$, $\bar{A} \cap C$, $B \cap \bar{C}$, $\overline{B \cap C}$, $A \cup \bar{B}$, $A \cup C$.

2. Bestimmen Sie $P(A), P(B), P(C)$ mittels der Formel

$$P(E) := \frac{\text{Anzahl der in } E \text{ enthaltenen Elementarereignisse}}{\text{Gesamtzahl der Elementarereignisse}}.$$

3. Bestimmen Sie $P(B \cup C)$ auf zwei verschiedene Arten (d.h. mengentheoretisch oder via Abzählen und Definition).

4. Benutzen Sie Teil c), um $P(\bar{B} \cap \bar{C})$ zu ermitteln.

Aufgabe 2

Mönch M. macht Beobachtungen zur Vererbung bei Erbsen (*pisum sativum*). Er stellt die Farbe der Erbsen fest, gelb (A) ist dabei dominant über grün (a). Außerdem beobachtet er die Farbe der Blüten. Hier sind farbige Blüten (B) dominant über weiße (b).

Er kreuzt eine homozygote (reinerbige) gelbe Erbse mit weißer Blüte (Phänotyp: Ab) wird mit einer homozygoten grünen Erbse mit farbiger Blüte (Phänotyp: aB). Damit er gezielt

verschiedene Varietäten kreuzen kann, entfernt er bei unreifen Blüten die Staubblätter, um so Autogamie (Selbstbestäubung) zu verhindern.

1. Geben Sie die Genotypen der Eltern (P) und die Geno- und Phänotypen der ersten Filialgeneration (F_1) an.

2. Erstellen Sie für die F_2-Generation ein Kombinationsquadrat.

3. Geben Sie für das Zufallsexperiment „Genotyp einer Pflanze der F_2-Generation“ den Stichprobenraum an.

4. Bestimmen Sie die folgenden Ereignisse E_i und berechnen Sie $P(E_i)$.

   E_1 : heterozygot an beiden Genorten
   E_2 : heterozygot an mindestens einem Genort
   E_3 : gelbe Erbse mit weißer Blüte (Phänotyp: aB)
   E_4 : homozygot an mindestens einem Genort

Aufgabe 3

In einem Versuch müssen Mäuse einen von vier Wegen wählen, um zu einem Stück Käse zu gelangen. Es nehmen vier (durch eine Markierung unterscheidbare) Mäuse an dem Experiment teil.

1. Wie viele verschiedene Ergebnisse bei den Wahlen der Wege sind insgesamt möglich?

2. Bestimmen Sie Zufallsexperiment, Stichprobenraum, Elementarereignisse und Ereignisse. (Letztere nicht explizit, dafür sind es viel zu viele. Wissen Sie, wie viele?)

3. Bestimmen Sie das Ereignis „jede Maus wählt einen anderen Weg“ als Menge und berechnen Sie die Wahrscheinlichkeit, wenn alle Wege für die Mäuse gleich attraktiv sind.

4. Den Studenten, die dieses Experiment durchführen, stehen acht Versuchstiere zur Verfügung. Wie viele Möglichkeiten gibt es, die vier benötigten Mäuse auszuwählen?

Aufgabe 4

Im Garten wurden 80 Äpfel und 20 Birnen geerntet. Aus 40 Äpfeln und 6 Birnen wird Gelée gekocht, die übrigen werden für Kuchen verwendet. Wir verfolgen die Verarbeitung einer zufällig ausgewählten Frucht.

1. Bestimmen Sie den Stichprobenraum und die Ereignisse.
2. Bestimmen Sie Ereignisse und Wahrscheinlichkeiten für die folgenden Aussagen:
   (a) Die Frucht ist ein Apfel und wird zu Gelée verarbeitet.
   (b) Die Frucht wird zu Gelée verarbeitet.
   (c) Ein Kuchen wird mit der Frucht belegt.
   (d) Die Frucht ist eine zu Gelée zu verarbeitende Birne oder ein Apfel, mit dem Kuchen gebacken wird.
   (e) Die Frucht ist ein Apfel oder wird nicht zu Gelée verarbeitet.

Aufgabe 5

In einem Wald unweit von Bonn sind 7 von 10 Pilzen Lamellenpilze und die übrigen Röhrenpilze. Der Anteil essbarer Pilze unter den Röhrenpilzen ist $\frac{4}{5}$, unter den Lamellenpilzen $\frac{1}{3}$.
Pilzkenner P. wird von seinem völlig unbedarften Bekannten B. begleitet. B. findet einen Pilz.

1. Bestimmen Sie den Stichprobenraum und die Wahrscheinlichkeiten der einelementigen Ereignisse.
2. Bestimmen Sie die Wahrscheinlichkeiten dafür, dass der von B. gefundene Pilz
   (a) ein Lamellenpilz ist,
   (b) essbar ist,
   (c) ein essbarer Röhrenpilz ist,
   (d) ein Röhrenpilz oder essbar ist,
   (e) essbar ist, falls bekannt ist, dass es ein Lamellenpilz ist,
   (f) ein Lamellenpilz ist, falls bekannt ist, dass er essbar ist,
   (g) ein Röhrenpilz ist, falls bekannt ist, dass er nicht essbar ist.

Aufgabe 6

In der Prüfung zu einer Vorlesung, bei der acht Themen behandelt wurden, werden von diesen zwei Themen zur Auswahl gestellt, nur eines der beiden muss bearbeitet werden. Die Prüfung ist so angelegt, dass genau solche Kandidaten, die sich auf eines der Prüfungsthemen vorbereitet haben, bestehen.

1. Kandidat K. ist sehr risikobereit. Er bereitet sich nur auf ein Thema vor. Mit welcher Wahrscheinlichkeit besteht er?
2. Prüfling P. gibt sich mit einer Erfolgswahrscheinlichkeit von nur 75% zufrieden. Wie viele Themen muss er vorbereiten?
3. Studentin S. verfährt nach dem Motto *Mut zur Lücke* und bereitet sich auf sechs Themengebiete vor. Welche Chancen hat sie, die Prüfung zu bestehen?

Aufgabe 7

Es wird mit zwei nicht unterscheidbaren Würfeln gewürfelt.

1. Beschreiben Sie Zufallsexperiment, Stichprobenraum, Elementarereignisse und Ereignisse. (Schreiben Sie dabei zu große Mengen nicht explizit auf.) Welches ist hier das unmögliche und welches das sichere Ereignis?
2. Bestimmen Sie die Wahrscheinlichkeit dafür, dass mindestens eine 6 gewürfelt wird.
3. Für zwei Ereignisse A und B gelte $A \cup B = \emptyset$. Was können Sie über A und B sagen?
4. Für die Ereignisse C und D gelte $C \cap D = \emptyset$. Welche Folgen hat das für die Ereignisse und ihre Wahrscheinlichkeiten? Was gilt für $P(C \cup D)$ und $P(C \cap D)$?

Aufgabe 8

Fertigen Sie Boxplots zu den Messreihen in Aufgabe 8 des 2. Kapitels an. Tragen Sie zuvor die benötigten Kenngrößen in eine Tabelle ein.

Aufgabe 9

1. Wir betrachten einen Stichprobenraum Ω. Es sei A ein Ereignis. Ferner seien die Ereignisse E_i, $i = 1, \ldots, k$ disjunkt, d.h. für $i \neq j$ gelte $E_i \cap E_j = \emptyset$. Es gelte $\bigcup_{i=1}^{k} E_i = \Omega$.
   Dann gilt $\sum_{i=1}^{k} P(E_i|A) = 1$.
   Wie kann man dies für $k = 2$ prägnant formulieren?
   Beweisen Sie obige Formel.
   Hinweis: Benutzen Sie den Satz von der totalen Wahrscheinlichkeit.

2. Es sei $k = 3$ und $P(E_1|A) = \frac{1}{3}$, $P(E_2|A) = \frac{1}{6}$, $P(\bar{A} \cap E_1) = \frac{1}{4}$, $P(\bar{A} \cap E_2) = \frac{1}{8}$ sowie $P(E_3|\bar{A}) = \frac{1}{3}$. Stellen Sie ein passendes Baumdiagramm auf und berechnen Sie die darin fehlenden Wahrscheinlichkeiten.

Aufgabe 10

Ein Feuermelder funktioniert im Brandfall mit einer Wahrscheinlichkeit von 99%, ein Fehlalarm tritt mit der Wahrscheinlichkeit 2% auf. Mit der Wahrscheinlichkeit $0,001$ brennt es. (Alle Angaben beziehen sich auf ein festes Beobachtungsintervall von einem Jahr.)

1. Zeichnen Sie ein Baumdiagramm.

2. Bestimmen Sie die Wahrscheinlichkeiten dafür, dass

   (a) es brennt, wenn Alarm ausgelöst wird,
   (b) ein Alarm, der ausgelöst wird, ein Fehlalarm ist,
   (c) es brennt, wenn kein Alarm ausgelöst ist,
   (d) es nicht brennt, wenn kein Alarm ausgelöst wird,
   (e) es brennt und Alarm ausgelöst wird,
   (f) Alarm ausgelöst wird, wenn es brennt,
   (g) es nicht brennt und Alarm ausgelöst wird.

Aufgabe 11

Ein neues Medikament wird auf Nebenwirkungen getestet. Bei 7 von 100 Patienten treten diese auf. Das Medikament wird an zwölf Probanden gestestet. Es wird die Anzahl der Patienten, bei denen Nebenwirkungen auftreten, ermittelt.

1. Welche Verteilung hat die betrachtete Zufallsvariable?
2. Berechnen Sie den Erwarungswert und die Varianz der Zufallsvariable.
3. Was ist die Wahrscheinlichkeit dafür, dass
   (a) bei keinem, genau einem bzw. höchstens zwei Patienten Nebenwirkungen auftreten,
   (b) mindestens neun Patienten nebenwirkungsfrei bleiben.

Aufgabe 12

Bei einem Spiel wird gewürfelt. Wirft die Spielerin im ersten Wurf mehr als 3 Augen, so darf sie noch einmal würfeln. Am Ende wird die Gesamtzahl der Augen ermittelt, dies ist die Zufallsvariable X.

Bestimmen Sie die diskrete Wahrscheinlichkeitsdichte, die Verteilungsfunktion, den Erwartungswert und die Varianz von X.

Aufgabe 13

In Bonn gibt es mit einer Wahrscheinlichkeit von $0,08$ eine weiße Weihnacht. Die Zufallsgröße X sei die Anzahl der Jahre (ab dem Jahr 2003), die wir auf eine weiße Weihnacht warten müssen.

1. Berechnen Sie Erwartungswert und Varianz.
2. Welches ist die Wahrscheinlichkeit dafür, dass von den nächsten zwanzig Jahren (d.h. 2003-2022) die Jahre mit weißer Weihnacht genau die Schaltjahre sind.

Aufgabe 14

In den Lieferungen der Schokoladenfabrik S. geht ein Schokoladenweihnachtsmann von 200 bei der Lieferung zu Bruch. Die Fabrik garantiert Ihren Abnehmern, dass in den Chargen von je 1000 Weihnachtsmännern höchstens 10 kaputt angeliefert werden. Bestimmen Sie näherungsweise mit der Poisson- und exakt mit der Binomialverteilung die Wahrscheinlichkeit dafür, dass eine gegebene Lieferung diese Anforderung nicht erfüllt.

Aufgabe 15

In einer Stadt bleibt der Weihnachtsmann in jedem Jahr im Durchschnitt in drei Schornsteinen stecken. Wie groß ist die Wahrscheinlichkeit, dass er in diesem Jahr ohne Probleme die Geschenke verteilen kann?

Aufgabe 16

Die Höhe (in Metern), die die Sylvesterrakete von Feuerwerkerin F. erreicht, ist eine normalverteilte Zufallsgröße X mit Erwartungswert 75 und Standardabweichung 8.

1. Bestimmen die Höhe, die mit Wahrscheinlichkeit 0.8 überschritten wird.
2. Bestimmen Sie eine Höhe, so dass die Wahrscheinlichkeit, dass die Rakete mindestens diese, aber höchstens 80 Meter erreicht, gerade $\frac{1}{2}$ beträgt.
3. Bestimmen Sie Wahrscheinlichkeit dafür, dass $67 \leq X \leq 83$.
4. Bestimmen Sie eine Zahl t, so dass die erreichte Höhe mit 90% Wahrscheinlichkeit zwischen $75 - t$ und $75 + t$ Metern liegt.

Aufgabe 17

Sei X eine gleichverteilte stetige Zufallsvariable auf dem Intervall $[-5, 10]$. Berechnen Sie Erwartungswert und Varianz.

Aufgabe 18

Eine normalverteilte Population hat die durchschnittliche Körpergröße $180cm$. 20% der Personen sind mindestens $188cm$ groß.

1. Welches sind die Verteilungsparameter μ und σ^2?
2. Wie groß ist die Wahrscheinlichkeit dafür, dass eine zufällig ausgewählte Person zwischen $172cm$ und $180cm$ misst?

Kapitel 4

Beurteilende Statistik

Zu den zentralen Aufgaben der Beurteilenden Statistik gehört es einerseits, das zugrundeliegende Verteilungsmodell aus einer Stichprobe zu bestimmen – etwa die Wahrscheinlichkeit für „Kopf“ aus endlich vielen Würfen einer nicht perfekten Münze. Andererseits stellt sich die Frage, wie verläßlich Schätzungen für Kenngrößen einer Verteilung sind. Es werden etwa n normalverteilte Meßwiederholungen vorgenommen. Wie verläßlich sind die abgeleiteten Schätzungen von μ und σ?

4.1 Schätzung unbekannter Wahrscheinlichkeiten

Gegeben sei ein Bernoulli-Experiment mit Ausgängen E und $\bar{E}$. Gesucht ist $p = P(E)$.

Das Experiment werde n-mal wiederholt, k-mal trete E ein. Was ist dann eine gute Schätzung für p?

Ein erster, einfacher Ansatz könnte den Erwartungswert $E(X) = np$ durch k abschätzen, dann wäre $p = k/n$.

Ein anderer Ansatz besteht im *Maximum-Likelihood-Prinzip*: Bei gegebenen Daten ist dasjenige Verteilungsmodell am plausibelsten, das die vorliegenden Daten mit der höchsten Wahrscheinlichkeit erzeugt.

Ein Alltagsbeispiel: Ein Restaurant habe zwei Köche, die abwechselnd arbeiten. Koch A versalzt die Suppe mit Wahrscheinlichkeit 0.3, Koch B mit Wahrscheinlichkeit 0.1. Ihre Suppe ist versalzen. Als Vermutung drängt sich dann auf, dass wohl Koch A wieder

einmal Dienst hat. Diese Vermutung lässt sich ebenso mittels des Maximum-Likelihood-Prinzips begründen. In der Tat vermuten wir Koch A am Werk, weil er mit höherer Wahrscheinlichkeit (0.3) die vorliegenden Daten ("Suppe versalzen") verursachen würde.

Im Fall des Bernoulliexperiments würde das Maximum-Likelihood-Prinzip wie folgt angewandt: Ist $P(E) = q$, so gilt $P(X = k) = \binom{n}{k} q^k (1-q)^{n-k}$. Das Maximum-Likelihood-Prinzip sagt dann aus, dass der Wert von q am plausibelsten ist, für den $f(q) = \binom{n}{k} q^k (1-q)^{n-k}$ maximal wird. Bestimmung des Maximums ergibt wiederum $q = k/n$.

4.2 Schätzung von Maßzahlen einer Grundgesamtheit

In diesem Kapitel beschäftigen wir uns mit der Frage, in welcher Weise sich Methoden der deskriptiven Statistik, z.B. das Ziehen von Stichproben, in einen wahrscheinlichkeitstheoretischen Zusammenhang einbetten lassen.

Allgemein gilt, dass eine Grundgesamtheit durch die Wahrscheinlichkeitsverteilung einer Zufallsvariablen X beschrieben wird. Bezeichne etwa Ω die Gesamtheit der Einwohner einer Stadt; die normalverteilte Zufallsvariable X könnte die Größe der Einwohner bezeichnen.

Eine Stichprobe der Länge n entspricht einer Realisierung von n unabhängigen Kopien $X_1, ..., X_n$ der Zufallsvariablen X, in unserem Beispiel also der Auswahl und Ermittlung der Körpergröße von n Einwohnern.

Als Maßzahlen von X haben wir den Erwartungswert $E(X)$ und die Varianz $V(X)$ kennengelernt. Diese sind a priori unbekannt. Wie können wir nun aus einer empirischen Stichprobe Aussagen über den Erwartungswert und die Varianz gewinnen?

Gegeben seien Werte $x_1, \ldots, x_n$ als Ergebnisse einer Stichprobe. Dann liegt es nahe, den Erwartungswert μ als Mittelwert zu interpretieren und das arithmetische Mittel $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ als Schätzwert zu benutzen.

Da die einzelnen Ergebnisse zufällig sind, können $\bar{x}_n$ und μ natürlich stark voneinander abweichen.

Definition. Eine Abbildung Θ_n, die jeder Stichprobe vom Umfang n aus einer Grundgesamtheit einen Schätzwert für eine bestimmte Maßzahl der Grundgesamtheit zuordnet, heißt *Schätzfunktion* für diese Maßzahl.

Im Beispiel hätten wir:

$$\Theta_n : (x_1, \ldots, x_n) \mapsto \frac{1}{n} \sum_{i=1}^{n} x_i$$

als Schätzfunktion für den Erwartungswert.

4.3 Erwartungstreue Schätzfunktionen

Schätzfunktionen konfrontieren uns mit dem folgenden grundlegenden Problem: Sei Θ_n Schätzfunktion; $x_1, \ldots, x_n$ seien Realisierungen der Zufallsvariablen $X_1, ..., X_n$. Dann ist $\Theta_n(X_1, \ldots, X_n)$ selbst eine Zufallsvariable.

Es ist dann eine plausibel zu verlangen, dass der Erwartungswert der Schätzfunktion wiederum die Maßzahl ergibt.

Definition. Eine Schätzfunktion Θ_n für eine unbekannte Maßzahl κ der Grundgesamtheit heißt *erwartungstreu bezüglich* κ, falls

$$E\left(\Theta_n(X_1, \ldots, X_n)\right) = \kappa.$$

Ein einfaches Beispiel ist das arithmetische Mittel, das als erwartungstreue Schätzfunktion für den Erwartungswert dienen kann.

In der Tat gilt:

$$E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{n}{n}\mu = \mu$$

Als nächstes drängt sich die Frage auf, ob Funktionen von erwartungstreuen Schätzern ihrerseits erwartungstreu sind, etwa: Ist $\bar{X}_n^2$ erwartungstreu für μ^2?

Seien $X_1, \ldots, X_n$ unabhängig. Dann haben wir $E((\bar{X})^2) = V(\bar{X}) + E(\bar{X})^2$ und $V(\bar{X}) = \sigma^2/n$. Es folgt: $E((\bar{X})^2) = \sigma^2/n + \mu^2$.

$(\bar{X}_n)^2$ ist also lediglich *asymptotisch erwartungstreu* für μ^2, da im allgemeinen $E((\bar{X}_n)^2) \neq \mu^2$, aber $\lim_{n\to\infty} E((\bar{X}_n)^2) = \mu^2$.

Definition. Die Größe $E((\bar{X}_n)^2) - \mu^2 =: \text{Bias}\left(\bar{X}^2; \mu^2\right)$ heißt *Verzerrung der Schätzfunktion.*

Wir zeigen noch, dass die empirische Varianz $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ eine erwartungstreue Schätzfunktion für σ^2 ist; seien dabei $X_1, \ldots, X_n$ unabhängig und identisch verteilt mit Erwartungswert μ und positiver Varianz $\sigma^2 = V(X_i)$.

Zum Beweis rechnen wir

$$E\left(\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right)$$

aus. Und zwar impliziert der Verschiebungssatz

$$\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \sum_{i=1}^{n} X_i^2 - n\left(\bar{X}_n\right)^2;$$

oben haben wir bereits

$$E\left(\left(\bar{X}_n\right)^2\right) = \frac{\sigma^2}{n} + \mu^2$$

ausgerechnet. Da $\sigma^2 = V(X_i) = E(X_i^2) - \mu^2$ gilt, folgt somit:

$$E\left(\sum_{i=1}^{n} X_i^2\right) = nE(X_i^2) = n(\sigma^2 + \mu^2)$$

Wir haben also:

$$\begin{aligned} E\left(\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2\right) &= E\left(\sum_{i=1}^{n} X_i^2\right) - nE\left(\left(\bar{X}_n\right)^2\right) = n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= (n-1)\sigma^2 \end{aligned}$$

Das war zu zeigen.

4.4 Konfidenzintervalle

Wir möchten die Güte einer Schätzung für μ abschätzen. Dazu fragen wir nach einer Konstanten $b > 0$, sodass der Erwartungswert μ mit einer Wahrscheinlichkeit von 95% im Intervall $[\bar{x}_n - b, \bar{x}_n + b]$ liegt.

Anders formuliert: Wir betrachten die Zufallsvariable $\bar{X}_n - \mu$ und versuchen, $P(|\bar{X}_n - \mu| \leq b)$ zu bestimmen.

Definition. Das Intervall $[x_\alpha^-, x_\alpha^+]$ heißt für $\alpha \in (0,1)$ *α-Konfidenzintervall für X*, falls

$$P(x_\alpha^- \leq X \leq x_\alpha^+) = 1 - \alpha.$$

Wir bestimmen nun ein Konfidenzintervall zu $X \sim N(\mu, \sigma^2)$, das symmetrisch zu μ liegt.

Dazu setzen wir an:

$$x_\alpha^- = \mu - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \qquad\qquad x_\alpha^+ = \mu + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

Dann ist $z_{\alpha/2}$ so zu bestimmen, dass $\Phi(z_{\alpha/2}) = 1 - \alpha/2$.

Insbesondere sind die folgenden Aussagen äquivalent:

1. Mit einer Wahrscheinlichkeit von $1 - \alpha$ liegt $\bar{X}_n$ im Intervall
$$\left[\mu - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \mu + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right].$$
2. Mit einer Wahrscheinlichkeit von $1 - \alpha$ gilt
$$\left|\bar{X}_n - \mu\right| \leq \frac{\sigma}{\sqrt{n}} z_{\alpha/2}.$$
3. Mit einer Wahrscheinlichkeit von $1 - \alpha$ liegt μ im Intervall
$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right].$$

Eine Zusammenfassung der Konstruktion eines Konfidenzintervalls ist in Abbildung 4.1 dargestellt.

Als Beispiel betrachten wir eine Apfelernte: Das Gewicht der Äpfel sei normalverteilt mit Varianz σ^2, $\sigma = 10$g. Gesucht ist der Erwartungswert μ.

Konstruktion eines Konfidenzintervalls

Gegeben sei eine annähernd normalverteilte Zufallsvariable mit bekannter Varianz σ^2. Gesucht ist eine Schätzung des Erwartungswertes bei vorgegebener Irrtumswahrscheinlichkeit. Dazu werden folgende Schritte durchgeführt:

1. Bestimme den Mittelwert $\bar{x}_n$ der Stichprobe $(x_1, \ldots, x_n)$.
2. Wähle eine Irrtumswahrscheinlichkeit α. Typische Werte sind 0.01, 0.05 oder 0.1.
3. Bestimme $z_{\alpha/2}$, sodass $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$. Hierzu wird typischerweise eine Tabelle herangezogen.
4. Dann gilt mit einer Wahrscheinlichkeit von $1 - \alpha$:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

Abbildung 4.1: Konstruktion eines Konfidenzintervalls

1. 100 Äpfel werden gewogen. Wir erhalten $\bar{x} = 142$g.
2. Wir wählen die Irrtumswahrscheinlichkeit zu $\alpha = 0.1$.
3. $z_{\alpha/2}$ muss erfüllen:

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2 = 0.95 \quad \Rightarrow \quad z_{\alpha/2} \approx 1.645.$$

4. Mit einer Wahrscheinlichkeit von 90% gilt dann:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

   Einsetzen ergibt:

$$140.355 \leq \mu \leq 143.645.$$

4.5 Konfidenzintervalle für den Erwartungswert einer normalverteilten Zufallsvariablen X bei unbekannter Varianz

Sei wiederum eine $N(\mu, \sigma^2)$-verteilte Zufallsvariable gegeben. Sowohl μ als auch σ sei unbekannt. Die Stichprobendaten seien $x_1, \ldots, x_n$.

Gesucht ist wieder ein Schätzwert für μ. Unsere Strategie wird der aus dem vorigen Abschnitt ähneln, allerdings stehen wir vor dem zusätzlichen Problem, dass σ unbekannt ist.

Wir wählen daher die empirische Varianz

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

als Schätzwert für σ^2; dabei ist $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ Schätzwert für μ.

Die Zufallsvariable $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ ist $N\left(\mu, \frac{\sigma^2}{n}\right)$-verteilt. Nach Standardisierung ist

$$\bar{X}_n^* = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

$N(0,1)$-verteilt.

Da σ unbekannt ist, ersetzen wir es durch S und betrachten die Zufallsvariable

$$T = \frac{\bar{X}_n - \mu}{S/\sqrt{n}}.$$

T ist *nicht* normalverteilt. Die korrekte Verteilung wurde von W. S. Gosset (1876–1937) ermittelt und unter dem Pseudonym "Student" veröffentlicht. Die Verteilung von T ist daher unter dem Namen "Studentsche t-Verteilung" bekannt.

Wir fassen einige Eigenschaften dieser Verteilung zusammen:

- Die Wahrscheinlichkeitsdichtefunktion ist

  $$f_{n-1}(x) = C_{n-1} \left(1 + \frac{x^2}{n-1}\right)^{-n/2};$$

  dabei ist $C_{n-1} \in \mathbb{R}$ ein Normierungsfaktor, so dass $\int_{\infty}^{\infty} f_{n-1} = 1$. Eine Zufallsvariable mit Dichtefunktion f_{n-1} heißt *t-verteilt* ("Student"-verteilt) mit $n-1$ Freiheitsgraden.

- Die Student-Verteilung ist symmetrisch bezüglich Null. Insbesondere gilt:

  $$E(T) = 0 \qquad V(T) = \frac{n-1}{n-3} \quad \text{für } n > 3; \text{ sonst } V(T) = \infty.$$

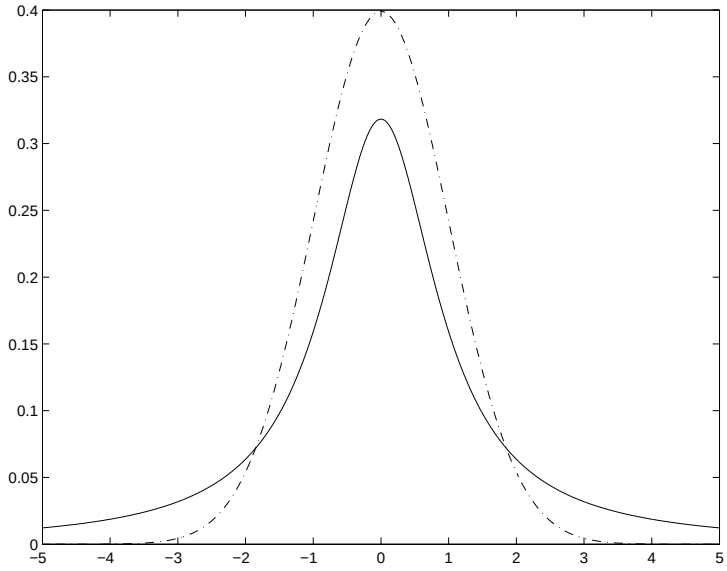


Abbildung 4.2: Student-t-Verteilung zu $n-1=1$ (durchgezogene Linie); zum Vergleich ist die Normalverteilung zu $\mu=0, \sigma=1$ eingezeichnet.

- Asymptotisch nähert sich T für große n der Normalverteilung an.

$T=\sqrt{n}\frac{\bar{X}_n-\mu}{S}$ ist t-verteilt mit $(n-1)$ Freiheitsgraden. In einem ersten Schritt bestimmen wir nun Konfidenzintervalle für T. Dazu haben wir $t_{n-1,\alpha/2}$ zu suchen, so dass

$$P(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}) = 1-\alpha.$$

Die Verteilungsfunktion der t-Verteilung ist tabelliert; diese Tabellen liefern die gewünschten Werte.

Im zweiten Schritt bestimmen wir zu den gegebenen Stichprobendaten $x_1, \ldots, x_n$ Mittelwert $\bar{x}$ und empirische Varianz S^2. Mit Wahrscheinlichkeit $1-\alpha$ gilt dann:

$$\left|\frac{\bar{x}-\mu}{S/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}.$$

Eine dazu äquivalente Aussage ist, dass mit Wahrscheinlichkeit $1-\alpha$ gilt:

$$\bar{x} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2} \leq \mu \leq \bar{x} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$$

In Abbildung 4.4 geben wir wieder ein kurzes Rezept zur Bestimmung des Konfidenzintervalls an. Abbildung 4.3 zeigt ein Matlab-Programm, das zu gegebenem Konfidenzniveau α die Meßwerte zusammen mit Mittelwerten und zugehörigen Konfidenzintervallen gegen die Reihenindizes aufträgt.

```
function [M,MCI] = Konfiplot(X,alpha)
M=[]; MCI=[];

J=size(X,2);    % Anzahl der Spalten von X
Xmin = 1.1*min(min(X));
Xmax= 1.1*max(max(X));
axis([0 J+1 Xmin Xmax])
hold on

for j=1:J

    Werte = X(:,j);
    Werte = Werte(~isnan(Werte));
    laenge=length(Werte);
    Mittel = mean(Werte);
    Varianz = var(Werte);

    plot(j*ones(laenge,1),Werte,'.')
    plot(j,Mittel,'or')

    X_unten = Mittel - Tzwei(laenge-1,alpha)*
                    *sqrt(Varianz/laenge);
    X_oben  = Mittel + Tzwei(laenge-1,alpha)*
                    *sqrt(Varianz/laenge);
    plot([j j], [X_unten X_oben], 'r+-')

    M= [M,Mittel];
    MCI=[MCI,[X_unten;X_oben]];
end
```

Abbildung 4.3: konfiplot.m: Datenplot mit Konfidenzintervallen.

4.6 Konfidenzintervalle für die Varianz bei normalverteilten Daten

Sei eine normalverteilte Zufallsvariable X gegeben; μ, σ seien unbekannt. $X_1, \ldots, X_n$ seien Kopien von X; $x_1, \ldots, x_n$ seien Stichprobendaten.

Gesucht ist eine Schätzung für σ^2.

Als Ansatz für einen Schätzer wählen wir die empirische Varianz

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 .$$

Die Verteilung der Zufallsvariablen $\hat{S}^2 := \frac{(n-1)S^2}{\sigma^2}$ wird nun als χ^2_{n-1}-Verteilung bezeichnet. Man sagt, $\hat{S}^2$ sei *Chi-Quadrat-verteilt mit* $(n-1)$ *Freiheitsgraden.* Man beachte, dass

$$\hat{S}^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

Konstruktion eines Konfidenzintervalls bei unbekannter Varianz

Sei eine (annähernd) normalverteilte Zufallsvariable X gegeben. Es mögen n empirische Messungen $x_1, \ldots, x_n$ vorliegen.

1. Bestimme Mittelwert $\bar{x}$ und Streuung
$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$
2. Wähle die Irrtumswahrscheinlichkeit α.
3. Bestimme mittels Tabelle $t_{n-1,\alpha/2}$.
4. Dann gilt:
$$P\left(\bar{x} - \frac{S}{\sqrt{n}}t_{n-1,\alpha/2} \leq \mu \leq \bar{x} + \frac{S}{\sqrt{n}}t_{n-1,\alpha/2}\right) \leq 1 - \alpha.$$

Die Voraussetzung, dass X zumindest annähernd normalverteilt ist, ist kritisch für diese Abschätzung.

Abbildung 4.4: Konstruktion eines Konfidenzintervalls bei unbekannter Varianz.

gilt; die einzelnen Summanden sind dabei $N(0,1)$-verteilt.

Allgemein gilt, dass die Summe von n unabhängig und identisch standardnormalverteilten Zufallsvariablen einer χ^2-Verteilung mit n Freiheitsgraden folgt.

Als Wahrscheinlichkeitsdichte ergibt sich:

$$f_n(t) = \begin{cases} 0 & t \leq 0 \\ C_n t^{n-2} 2e^{-t/2} & \text{sonst} \end{cases}$$

Dabei ist $C_n > 0$ eine geeignete Normierungskonstante.

Ferner gilt:

$$E(\chi_n^2) = n \qquad V(\chi_n^2) = 2n$$

Wir konstruieren Konfidenzintervalle für unseren Varianzschätzer:

Sei

$$F_n(t) = \int_0^t f_n(s)ds$$

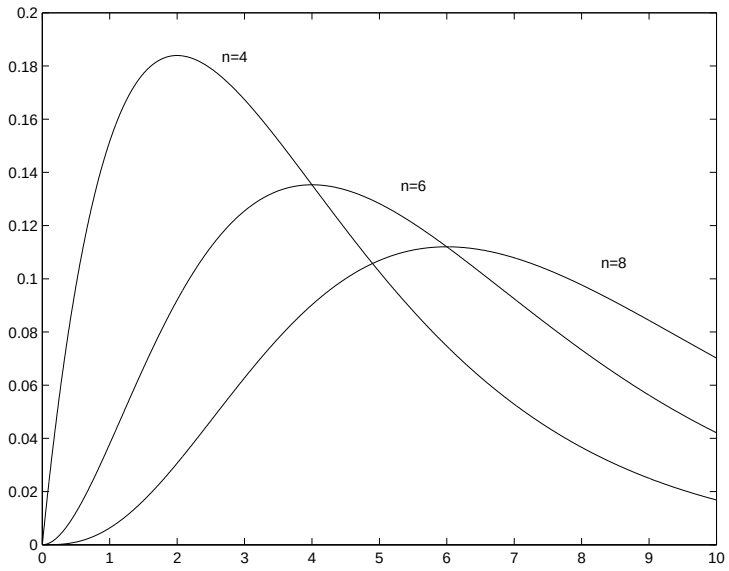


Abbildung 4.5: χ^2-Verteilung mit $n = 4, 6, 8$.

die Verteilungsfunktion der χ^2_n-Verteilung. Dann wählen wir zu vorgewählter Irrtumswahrscheinlichkeit α das α-Konfidenzintervall $\left[C_{n-1,\alpha/2}, C_{n-1,1-\alpha/2}\right]$ so, dass $F_{n-1}(C_{n-1,\alpha/2}) = \alpha/2$ sowie $F_{n-1}(C_{n-1,1-\alpha/2}) = 1 - \alpha/2$. Es folgt dann

$$P(C_{n-1,\alpha/2} \leq \chi^2_{n-1} \leq C_{n-1,1-\alpha/2}) = 1 - \alpha.$$

Wir wissen, dass $\hat{S}^2 = \frac{(n-1)S^2}{\sigma^2}$ χ^2_{n-1}-verteilt ist. Daher gilt mit Wahrscheinlichkeit $1 - \alpha$:

$$C_{n-1,\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq C_{n-1,1-\alpha/2}$$

oder äquivalent dazu

$$\frac{(n-1)S^2}{C_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{C_{n-1,\alpha/2}}.$$

Anders gesagt:

$$P\left(\frac{(n-1)S^2}{C_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{C_{n-1,\alpha/2}}\right) = 1 - \alpha.$$

Wir fassen das Verfahren in Abbildung 4.6 zusammen.

Konstruktion eines Konfidenzintervalls für die Varianz

Sei eine (annähernd) normalverteilte Zufallsvariable X gegeben. Es mögen n empirische Messungen $x_1, \ldots, x_n$ vorliegen. Gesucht ist eine Schätzung für die Varianz.

1. Bestimme Mittelwert $\bar{x}$ und empirische Varianz
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$
2. Wähle die Irrtumswahrscheinlichkeit α.
3. Bestimme mittels Tabelle $C_{n-1,\alpha/2}, C_{n-1,1-\alpha/2}$, so dass
$$F_{n-1}\left(C_{n-1,\alpha/2}\right) = \alpha/2, \qquad F_{n-1}\left(C_{n-1,1-\alpha/2}\right) = 1 - \alpha/2$$
4. Dann gilt mit Wahrscheinlichkeit $1 - \alpha$:
$$\frac{(n-1)S^2}{C_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{C_{n-1,\alpha/2}}.$$

Abbildung 4.6: Konstruktion eines Konfidenzintervalls für die Varianz.

4.7 Parametertests — Begriffsbildung

Wir diskutieren zunächst ein Beispiel: Eine Brauerei besitze eine Abfüllanlage, die in jeder Flasche genau 500ml Bier abfüllen soll. Kleinere Abweichungen sind unvermeidlich.

Es wird also behauptet, dass die Anlage im Mittel 500ml Bier in eine Flasche füllt. In der Sprache der Wahrscheinlichkeitstheorie sei X eine Zufallsvariable, die die Abfüllmenge beschreibt. Dann soll $E(X) = 500\text{ml}$ gelten. Diese Aussage soll mittels einer Stichprobe überprüft werden.

In der Statistik drückt man das folgendermaßen aus: Zu prüfen ist die *Null-Hypothese* H_0,

$$H_0 : E(X) = \mu_0 = 500\text{ml}.$$

Um diese Hypothese zu prüfen, bestimmen wir zunächst den Mittelwert der Stichprobe. Die Frage ist dann: Welche Abweichung des Mittelwerts vom Erwartungswert ist als so signifikant einzustufen, dass die Null-Hypothese abgelehnt wird?

Die Antwort auf diese Frage kann von der Interessenlage abhängen. Der Verband der

Biertrinker etwa wird die Hypothese nur dann ablehnen, wenn im Mittel zu wenig Bier abgefüllt wird. Die Gegenhypothese würde dann lauten:

$$H_1 : E(X) < \mu_0 = 500\text{ml}$$

Diese Art der Gegenhypothese nennt man *linksseitige Fragestellung*. Hier wird man H_0 nur dann ablehnen, wenn der Mittelwert $\bar{x}$ signifikant kleiner als μ_0 ist.

Der Bierproduzent andererseits könnte vor allem daran interessiert sein, nicht zu viel Bier abzufüllen. Seine Gegenhypothese wäre dann

$$H_1 : E(X) > \mu_0;$$

man spricht von einer *rechtsseitigen Fragestellung*.

Dem Produzenten der Abfüllanlage schließlich könnte vor allem an der Funktionstüchtigkeit der Anlage gelegen sein. Seine Gegenhypothese könnte also

$$H_1 : E(X) \neq \mu_0$$

lauten — eine *zweiseitige Fragestellung*.

Zusammenfassend sind Null-Hypothese und Gegenhypothese wichtigste Bestandteile eines statistischen Tests. Spricht unter Berücksichtigung der Gegenhypothese die Stichprobe signifikant gegen die Null-Hypothese, so wird diese abgelehnt. Gleichzeitig wird damit die Gegenhypothese angenommen.

4.8 Fehler und Risiken bei Statistik-basierten Entscheidungsverfahren

Im vorigen Abschnitt haben wir bereits angedeutet, dass in der Praxis Anreize bestehen können, Entscheidungsverfahren so auszuwählen, dass ein bestimmtes Ergebnis favorisiert wird — „traue keiner Statistik, die Du nicht selbst gefälscht hast.“

Ein naturwissenschaftliches Experiment soll Gesetzmäßigkeiten über Prozesse aufdecken; das Resultat wird häufig durch eine stetige Zufallsvariable T beschrieben. Im einfachsten Fall konkurrieren — wie schon im vorigen Abschnitt — zwei Theorien (oder Hypothesen)

zur Beschreibung des zugrundeliegenden Prozesses. Das Experiment wird dann häufig so angelegt, dass T unter der Theorie T_0 tendenziell kleine Werte annimmt, große Werte aber unwahrscheinlich sind. Unter der Theorie T_1 möge T tendenziell große Werte annehmen, kleine Werte seien unwahrscheinlich. Das heißt: Die Wahrscheinlichkeitsdichten f_i zu den Theorien T_i sind unterschiedlich.

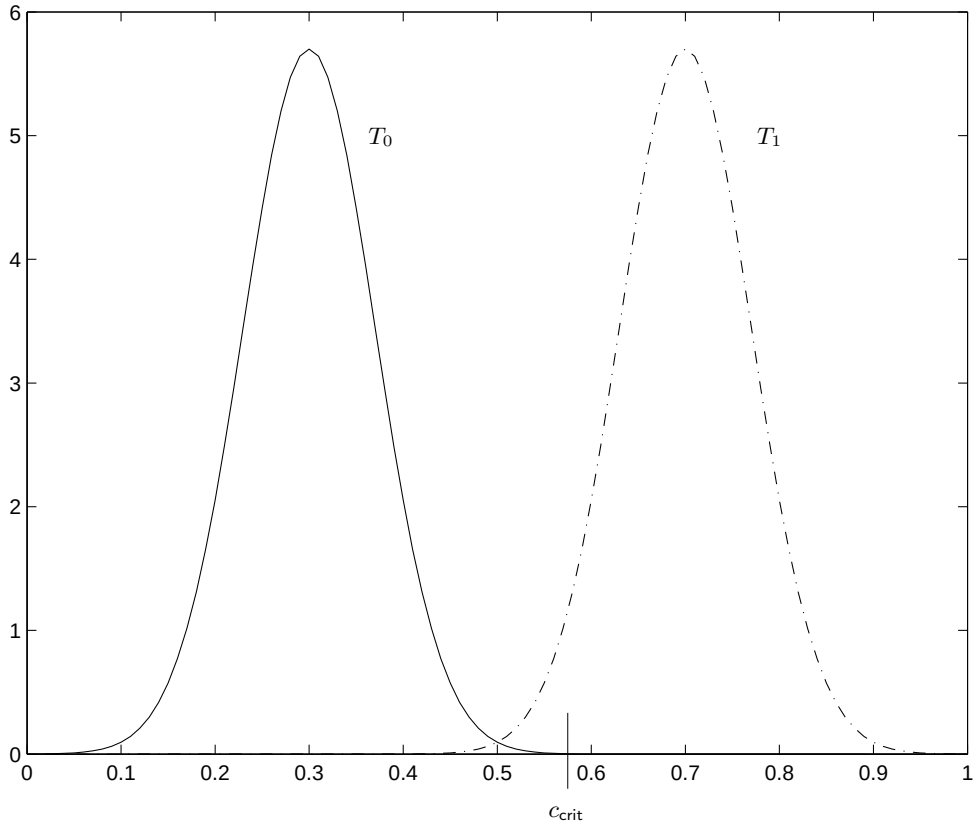


Abbildung 4.7: Hypothesentest

Bezeichne f die experimentell gemessene Häufigkeitsdichte. Die Nullhypothese H_0 besage, dass $f = f_0$. Die Gegenhypothese $H1$ besage, dass $f = f_1$ ist. Unsere Entscheidungsregel wird dann sein, dass H_1 angenommen wird (Schreibweise: „H_1“), wenn T einen kritischen Wert c_{krit} überschreitet. Es gilt:

$$[0, 1] = [0, c_{\text{krit}}] \cup [c_{\text{krit}}, 1] =: A \cup B,$$

der Wertebereich von T wird also in Annahmebereiche von H_0 (A) und H_1 (B) zerlegt.

Wir unterscheiden Fehler erster und zweiter Art:

- Einen *Fehler erster Art* nennen wir eine Entscheidung für H_1, obwohl H_0 richtig ist.
- Einen *Fehler zweiter Art* nennen wir eine Entscheidung für H_0, obwohl H_1 richtig ist.

Wie ist nun c_{krit} zu wählen? Verschieben wir c_{krit} nach rechts, so sinkt die Wahrscheinlichkeit für einen Fehler 1. Art, und die für einen Fehler 2. Art steigt. Verschieben wir c_{krit} umgekehrt nach links, so ändern sich die Fehlerwahrscheinlichkeiten entsprechend umgekehrt. Es ist also nicht möglich, beide Fehlertypen gleichzeitig völlig auszuschließen.

Um Fehler handzuhaben, gibt es zwei mögliche Verfahrensweisen.

4.8.1 Nachweisproblematik

Sei T_0 der aktuell etablierte Stand des Wissens, T_1 eine neue Theorie, die die herkömmliche ersetzen will. Dann sollte H_0 nur dann verworfen werden, wenn sich ein experimenteller Ausgang einstellt, der unter f_0 höchst unwahrscheinlich wäre. Kontrolliert wird in diesem Fall der Fehler erster Art, wir fordern also

$$P_{H_0}(\text{„}H_1\text{“}) \leq \alpha.$$

α wird als *Signifikanzniveau* bezeichnet; ist β die sich daraus ergebende Wahrscheinlichkeit für einen Fehler 2. Art ($P_{H_1}(H_0)$), so heißt $1 - \beta$ die „Schärfe“ des Tests.

4.8.2 Risikoüberlegung

Untersucht werden soll die Verseuchung von Milchprodukten durch radioaktiven Fallout nach der Katastrophe von Tschernobyl. T_0 sei die Theorie, dass die Milch verseucht wurde, T_1 die Theorie, dass sie nicht verseucht wurde.

Wir betrachten die Risiken der möglichen Fehler:

- Wird trinkbare Milch für verseucht erklärt, so führt dies zur unnötigen Vernichtung der Existenzen von Bauern.
- Wird verseuchte Milch für unbedenklich erklärt, so sind Missbildungen bei Kleinkindern zu erwarten.

Das Risiko von Missbildungen bei Kleinkindern ist schwerer zu gewichten; daher wird in diesem Fall die Wahrscheinlichkeit kontrolliert, dass verseuchte Milch für unbedenklich gehalten wird. Die Nullhypothese ist in diesem Fall, dass die Milch verseucht sei; gefordert wird ein statistischer Nachweis, dass die Milch trinkbar ist.

4.9 Test des Erwartungswertes μ_0 einer Normalverteilung (zweiseitiger t-Test)

Sei X normalverteilt mit unbekanntem Erwartungswert $\mu = E(X)$ und unbekannter Varianz $\sigma^2 = V(X)$. $X_1, \ldots, X_n$ seien unabhängige Kopien von X und $x_1, \ldots, x_n$ seien Stichprobendaten. Wir betrachten die folgenden Hypothesen:

Nullhypothese: $H_0 : E(X) = \mu_0$ Gegenhypothese: $H_1 : E(X) \neq \mu_0$

Wir wählen eine feste Irrtumswahrscheinlichkeit α. Als Prüfgröße für die Entscheidung über die Nullhypothese wählen wir:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

T ist Student-verteilt mit $n-1$ Freiheitsgraden. Wir wählen $t_{n-1,\alpha/2}$ so, dass

$$P(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}) = 1 - \alpha.$$

Ist also die Nullhypothese richtig, so produziert die Prüfgröße T mit Wahrscheinlichkeit $1-\alpha$ Werte im Intervall $\left[-t_{n-1,\alpha/2}, t_{n-1,\alpha/2}\right]$.

Sei nun $t^* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ der empirische Wert. Falls $t^* \notin \left[-t_{n-1,\alpha/2}, t_{n-1,\alpha/2}\right]$, so wird die Nullhypothese auf Signifikanzniveau α abgelehnt.

4.10 t-Test auf Lageunterschied bei nichtverbundenen Stichproben

Wir beginnen mit einem Anwendungsbeispiel: Gegeben seien zwei Gruppen von Meerschweinchen mit unterschiedlicher Fütterung. Wir betrachten Gewichtszunahmen in den verschiedenen Gruppen:

Zweiseitiger t-Test

Sei eine (annähernd) normalverteilte Zufallsvariable X gegeben. Es mögen n empirische Messungen $x_1, \ldots, x_n$ vorliegen.
Es werden zwei Hypothesen aufgestellt, die

Nullhypothese: $H_0 : E(X) = \mu_0$

und die

Gegenhypothese: $H_1 : E(X) \neq \mu_0$.

1. Bestimme Mittelwert $\bar{x}$ und empirische Varianz

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

2. Wähle die Irrtumswahrscheinlichkeit α.
3. Bestimme mittels Tabelle $t_{n-1,\alpha/2}$.
4. Ist

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq t_{n-1,\alpha/2},$$

so behalte H_0 bei, sonst lehne H_0 ab und akzeptiere H_1.

Abbildung 4.8: Zweiseitiger t-Test für den Erwartungswert.

Gruppe 1	134	146	104	...	123
Gruppe 2	120	94	146	...	133

Unterscheidet sich die mittlere Gewichtszunahme?

Wir formulieren die Fragestellung in der Sprache der Wahrscheinlichkeitstheorie: Seien X, Y normalverteilte Zufallsvariablen, und es gelte $\mu_1 = E(X)$, $\mu_2 = E(Y)$. Außerdem sei $V(X) = V(Y) = \sigma^2$; die beiden Verteilungen mögen also die gleiche Varianz haben.

Dann stellen wir die folgenden Hypothesen auf:

Nullhypothese:	$H_0 : \mu_1 = \mu_2$	(kein Lageunterschied)
Gegenhypothese:	$H_1 : \mu_1 \neq \mu_2$	(Lageunterschied)

Sollen tendenziell größere Beobachtungen in Gruppe 2 nachgewiesen werden, wird stattdessen die folgende Gegenhypothese aufgestellt:

$$H_1 : \mu_1 < \mu_2$$

Wir betrachten die Differenz der arithmetischen Mittel

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n} X_i \qquad \bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n} Y_i.$$

Da X und Y unabhängig sein sollen, ist die Varianz von $\bar{Y} - \bar{X}$ gegeben durch

$$\sigma_+^2 = V(\bar{Y} - \bar{X}) = V(\bar{Y}) + V(\bar{X}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

Außerdem sind $\bar{X}$ und $\bar{Y}$ normalverteilt; damit ist

$$\bar{Y} - \bar{X} \sim N\left(\mu_2 - \mu_1; \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Leider ist im allgemeinen Fall σ^2 nicht bekannt. Wir benutzen ohne Beweis das folgende Resultat: Ein erwartungstreuer Schätzer für σ^2 wird gegeben durch das gewichtete Mittel der Schätzer

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2, \qquad S_Y^2 = \frac{1}{n_2 - 1} \sum_{i=2}^{n_2} (y_i - \bar{y})^2.$$

Und zwar soll gelten:

$$S^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_X^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_Y^2$$

Wie vorhin folgt dann die Prüfgröße

$$T = \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2}} = \frac{\bar{Y} - \bar{X}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2}}$$

einer Student-Verteilung mit $(n_1 + n_2 - 2)$ Freiheitsgraden.

Als α-Konfidenzintervall ergibt sich dann

$$I_\alpha = \left[-t_{n_1+n_2-2,\alpha/2}, t_{n_1+n_2-2,\alpha/2}\right].$$

Liegt also der tatsächliche Wert

$$t^* = \frac{\bar{y} - \bar{x}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2}}$$

nicht in I_α, so kann H_0 mit Irrtumswahrscheinlichkeit α verworfen werden.

4.11 t-Test auf Lageunterschied bei verbundenen Stichproben

Es soll untersucht werden, ob ein neues Futtermittel die Milchleistung von Kühen erhöht. Dazu wird die Milchleistung bei fünf Kühen vor und nach Gabe des neuen Futtermittels erhoben. Folgende Werte ergeben sich:

vorher	5ℓ	3ℓ	4ℓ	6ℓ	5ℓ
nachher	4.5ℓ	6ℓ	7ℓ	4ℓ	5ℓ

t-Test auf Lageunterschied

Seien normalverteilte Zufallsvariablen X, Y mit derselben Varianz gegeben.

1. Berechne Mittelwerte $\bar{X}$, $\bar{Y}$ sowie empirische Varianzen S_X^2, S_Y^2.
2. Berechne
$$S^2 = \frac{n_X - 1}{n_X + n_Y - 2} S_X^2 + \frac{n_Y - 1}{n_X + n_Y - 2} S_Y^2$$
als empirische Varianz für $\bar{Y} - \bar{X}$.
3. Wähle die Irrtumswahrscheinlichkeit α.
4. Bestimme $t_{n_X+n_Y-2,\alpha/2}$.
5. Bestimme die empirische Prüfgröße
$$t^* = \frac{\bar{y} - \bar{x}}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_X}\right) S^2}}.$$
6. Lehne H_0 ab, falls $|t^*| > t_{n_X+n_Y-2,\alpha/2}$.

Abbildung 4.9: t-Test auf Lageunterschied.

Als Nullhypothese nehmen wir an, dass $\mu_1 = \mu_2$, also kein Effekt eingetreten ist. Gegenhypothese ist $\mu_1 \neq \mu_2$. (Ein Effekt wird gemessen.)

Um diese Hypothesen zu überprüfen, betrachten wir die Differenzen $D_i = Y_i - X_i$, $i = 1, \ldots, n$. Die D_i sind dann unabhängig normalverteilt mit unbekanntem μ und σ. Unsere Hypothesen lassen sich wie folgt umformulieren:

$$H_0 : \delta = E(\bar{D}) = 0 \qquad H_1 : \delta \neq 0$$

Man kann dann einen t-Test mit Daten $d_1, \ldots, d_n$ durchführen, $d_i = y_i - x_i$. Zu gegebener Irrtumswahrscheinlichkeit α wird also wieder $t_{n-1,\alpha/2}$ bestimmt, so dass für die t_{n-1}-verteilte Zufallsvariable T gilt:

$$P\left(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}\right) = 1 - \alpha$$

Mit

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2$$

gilt dann: Falls $\left|\frac{\bar{d}}{S_D/\sqrt{n}}\right| > t_{(n-1,\alpha/2)}$, so ist H_0 zu verwerfen.

Im konkreten Beispiel die Milchproduktion wählen wir ein Signifikanzniveau von 10%. Dann ist $t_{4,0.95} = 2.132$. Als Daten haben wir $\{d_i\} = \{-0.5, 3, 3, -2.0\}$. Es gilt also $\bar{d} = 0.7$. Wir erhalten außerdem $S_D^2 = 4.95$, $t^* = 0.7/\sqrt{4.95/5} \approx 0.7 < 2.132$. Die Hypothese H_0 wird also beibehalten.

4.12 Test auf Varianzgleichheit bei normalverteilten Zufallsvariablen

Haben Grizzly-Bären in Alaska und in Kanada im Mittel das gleiche Körpergewicht? Die übliche Vorgehensweise wäre eine Gewichtserhebung in Kanada und Alaska, die Stichprobenwerte $x_1, \ldots, x_{n_1}$ und $y_1, \ldots, y_{n_2}$ erzeugen würde. Wir würden dann einen t-Test bei unverbundenen Stichproben durchführen. Dabei wurde stillschweigend vorausgesetzt, dass die Varianz in beiden Populationen identisch ist.

Läßt sich diese Annahme anhand der Stichproben überprüfen?

Wir führen einen *Test auf Varianzhomogenität* durch. Nullhypothese H_0 ist Gleichheit der Varianzen, $\sigma_1^2 = \sigma_2^2$; Gegenhypothese H_1 sind unterschiedliche Varianzen, $\sigma_1^2 \neq \sigma_2^2$.

Wir betrachten wieder die empirischen Varianzschätzer

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} \left(X_j - \bar{X}\right)^2, \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left(Y_j - \bar{Y}\right)^2.$$

Da X, Y als normalverteilt angenommen wurden, gilt

$$U_i^2 := \frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2(n-1).$$

Sind $U_1 \sim \chi^2(df_1)$, $U_2 \sim \chi^2(df_2)$ unabhängige Zufallsvariablen, so genügt

$$\frac{\frac{1}{df_1}U_1}{\frac{1}{df2}U_2}$$

einer Fischer-Verteilung $F(df_1, df_2)$ mit df_1 Zähler- und df_2 Nennerfreiheitsgraden.

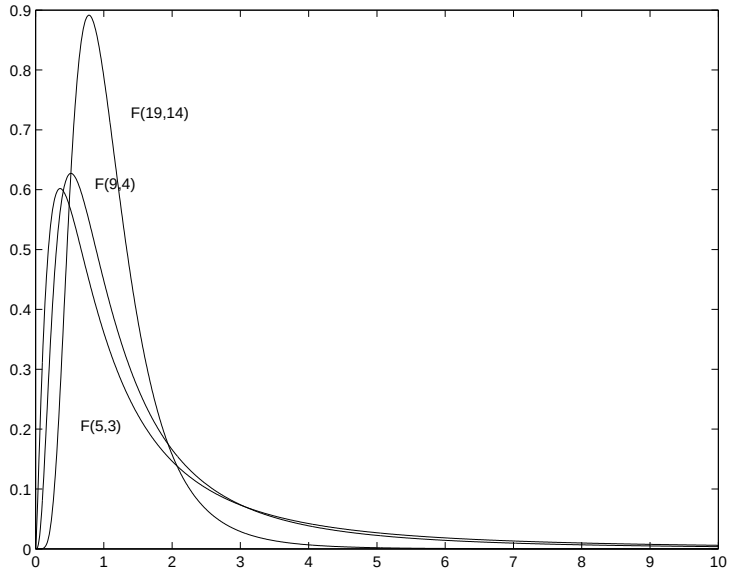


Abbildung 4.10: Fischer-Verteilungen mit (19,14), (9,4), (5,3) Zähler- und Nennerfreiheitsgraden.

In unserem Fall haben wir

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \chi^2(n_i - 1),$$

also

$$\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Ist die Nullhypothese erfüllt ($\sigma_1 = \sigma_2$), so ist $Q = S_1^2/S_2^2$ $F(n_1 - 1, n_2 - 1)$-verteilt.

Wir gehen wieder wie von den früheren Hypothesentests her gewohnt vor: Wir bestimmen Parameter $F(n_1 - 1, n_2 - 1)_{\alpha/2}$, $F(n_1 - 1, n_2 - 1)_{1-\alpha/2}$, so dass

$$P(F(n_1 - 1, n_2 - 1)_{\alpha/2} \leq Q \leq F(n_1 - 1, n_2 - 1)_{1-\alpha/2}) = 1 - \alpha.$$

Ist dann $Q < F(n_1 - 1, n_2 - 1)_{\alpha/2}$ oder $Q > F(n_1 - 1, n_2 - 1)_{1-\alpha/2}$, so wird H_0 verworfen.

Im Umgang mit der Fischer-Verteilung ist die folgende Inversionsrelation hilfreich:

$$F(df_1, df_2)_{\alpha/2} = \frac{1}{F(df_2, df_1)_{1-\alpha/2}}$$

4.13 Korrelation und Regression

Seien zwei Zufallsvariablen X und Y gegeben. Wir interessieren uns dafür, ob ein Zusammenhang oder eine Abhängigkeit zwischen den Werten der beiden Variablen bestehen könnte.

Wir nennen einige Beispiele:

1. X beschreibe die Körpergröße von Vätern, Y die ihrer Söhne.
2. X beschreibe das Alter von Personen, Y ihren Blutdruck.
3. X sei der Milchertrag der EU in den Jahren 1980-2000, Y die Weinproduktion im gleichen Zeitraum.

In den ersten beiden Fällen ist X offensichtlich eine unabhängige Variable; gefragt wird hier, wie Y von X *abhängt*: Gibt es eine Funktion, die diese Abhängigkeit beschreibt? Und wie lautet diese Funktion? Diese Fragen sind Thema der *Regressionsrechnung*.

Im letzten Fall ist fraglich, ob eine Abhängigkeit besteht. Hier ist zunächst zu fragen, ob ein Zusammenhang zwischen X und Y besteht: Treten etwa große X-Werte immer gemeinsam mit kleinen Y-Werten auf? Dies ist die Frage nach der *Korrelation* zwischen Zufallsvariablen.

4.13.1 Regressionsrechnung — Prinzip der kleinsten Quadrate

Ziel der (linearen) Regressionsrechnung ist die statistische Analyse linearer Zusammenhänge zwischen einer abhängigen Variable Y (Zielgröße, Regressor) und der unabhängigen Variablen X (Regressand).

Als Modell betrachten wir n unabhängige Paare von Meßwerten $(x_1, y_1), \ldots, (x_n, y_n)$. Gesucht wird dann eine Gerade

$$m(x) = b_1 + b_2 x,$$

die die Messwerte „optimal“ approximiert.

Wir betrachten abhängig von b_1, b_2 die Residuen

$$\epsilon_i = y_i - b_1 - b_2 x_i,$$

also die Abstände zwischen den y-Koordinaten von Meßwert und Schätzwert. „Optimal“ sind dann solche Paare b_1, b_2, für die die ϵ_i gleichmäßig klein werden. Dies führt auf die Methode der kleinsten Quadrate (C. F. Gauß): Wir minimieren den Ausdruck

$$Q(\beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2; \quad \beta_1, \beta_2 \in \mathbb{R}.$$

Satz. Das o.g. Minimierungsproblem hat die eindeutige Lösung

$$\begin{aligned} b_2 &= \frac{\sum_{i=1}^n y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} = \frac{\mathrm{Cov}(x,y)}{V(x)} \\ b_1 &= \bar{y} - b_2 \bar{x}. \end{aligned}$$

Dabei gilt:

$$\mathrm{Cov}(x,y) := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad V(x) := \frac{1}{n-1} \left\{ \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 \right\}$$

Wir bezeichnen diese Ausdrücke als empirische Varianz und empirische Covarianz[1].

Zum Beweis zeigen wir, dass $Q(b_1, b_2) \leq Q(\beta_1, \beta_2)$ für alle $\beta_1, \beta_2 \in \mathbb{R}$. Wir ergänzen zunächst:

$$\begin{aligned} Q(\beta_1, \beta_2) &= \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - b_1 - \beta_i x_i + (b_1 - \beta_1) + (b_2 - \beta_2 x_i)^2 \right) \end{aligned}$$

Dann rechnen wir mit der binomischen Formel aus:

$$\begin{aligned} Q(\beta_1, \beta_2) &= \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2 \\ &\quad + 2 \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)(b_1 - \beta_1 + (b_2 - \beta_2) x_i) \\ &\quad + \sum_{i=1}^{n} (b_1 - \beta_1 + (b_2 - \beta_2) x_i)^2 \\ &= Q(b_1, b_2) + \underbrace{2 \sum \cdots}_{I} + \underbrace{\sum (\cdots)^2}_{II} \end{aligned}$$

[1] Dabei haben wir die Abkürzungen $x := (x_1, \cdots, x_n)$ und $y := (y_1, \cdots, y_n)$ benutzt.

Term II ist als Summe von Quadraten positiv. Wir müssen also nur noch den gemischten Term I betrachten. $Q(\beta_1, \beta_2) \leq Q(b_1, b_2)$ ist sicher dann wahr, wenn I verschwindet. Wir wählen also b_1, b_2, so dass gilt:

$$\begin{aligned} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) &= 0 \\ \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)\, x_i &= 0 \end{aligned}$$

Die erste dieser Bedingungen führt unmittelbar zu

$$nb_1 = \sum_{i=1}^{n} y_i - b_2 \sum_{i=1}^{n} x_i,$$

also zum behaupteten Ausdruck $b_1 = \bar{y} - b_2\bar{x}$. Die zweite Bedingung ergibt:

$$\begin{aligned} \sum b_2 x_i^2 &= \sum_{i=1}^{n} (y_i - b_1)\, x_i \\ &= \sum_{i=1}^{n} y_i x_i - (\bar{y} - b_2\bar{x}) \sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} y_i x_i - \bar{y} n \bar{x} + n b_2 \bar{x}^2, \end{aligned}$$

also

$$b_2 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Das war zu zeigen.

4.13.2 Ein lineares Regressionsproblem unter Matlab

Wir interessieren uns für den funktionalen Zusammenhang zwischen Körpergröße und Ohrenlänge bei Maultieren. Dazu werden bei zehn Mulis Widerristhöhe und Ohrenlänge gemessen; diese Daten seien in einer 2×10-Matrix `mulis` abgelegt. Die erste Spalte enthalte dabei die Widerristhöhe, die zweite die Ohrenlänge.

Das zu erstellende Programm soll zunächst die Daten nach steigender Widerristhöhe sortieren und Widerristhöhen sowie Ohrenlängen in einem Diagramm gegen die Nummer des Maultiers auftragen.

Zusätzlich soll ein Streudiagramm der Ohrenlänge in Abhängigkeit von der Widerristöhe gezeichnet werden, in das zusätzlich die Regressionsgerade eingezeichnet wird.

Im letzten Schritt sollen die Residuen durch ein Stabdiagramm dargestellt werden.

Der Programmcode ist in Abbildung 4.11 dargestellt.

```
% Datenerzeugung
widerrist=[];
ohren=[];
for i=1:10
    widerrist=[widerrist,72+16*rand];
    ohren=[ohren,0.25*widerrist(i)-10+0.2*randn];
end
muli=[widerrist; ohren];

% Visualisierung Messreihe
muli=muli';
ordermuli=sortrows(muli,1);
figure(1)
subplot(2,2,1)
title(['Biometrische Daten von 10 Maultieren'])
hold on
plot(ordermuli,'d');

% Streudiagramm und Regressionsgerade
subplot(2,2,3)
title(['Ohrenlänge gegen Widerristhöhe [cm]'])
hold on
plot(widerrist,ohren,'+k')
X=[ones(size(widerrist))' widerrist'];
linregress=X\ohren'
mulimin=min(widerrist);
mulimax=max(widerrist);
xx=[mulimin-2:(mulimax+4-mulimin)/600:mulimax+2];
yy=linregress(1)+linregress(2)*xx;
plot(xx,yy,'-r')

% Darstellung Residuen
residuen=linregress(1)+linregress(2)*widerrist-ohren;
subplot(2,2,4)
title(['Residuen'])
hold on
index=[1:1:10];
plot([0 10],[0 0],'k')
hold on
plot(index,residuen,'*k')
for i=1:10
    plot ([i i],[0 residuen(i)],'-k')
end
```

Abbildung 4.11: Regressionsproblem unter Matlab

Kern des Programms ist der „Backslash-Operator“, der benutzt wird, um das eigentliche Regressionsproblem zu lösen. In seiner einfachsten Version dient dieser Operator dazu, lineare Gleichungssysteme zu lösen: Sei A eine reguläre Matrix. Dann ist $A \setminus y$ die eindeutige Lösung x des Gleichungssystems $Ax = y$.

Diese Lösung minimiert insbesondere den folgenden Ausdruck:

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} A_{ij} x_j - y_i \right)^2 \geq 0$$

Ist das Gleichungssystem überbestimmt, hat die Matrix A also mehr Zeilen als Spalten, d.h., gibt es mehr Gleichungen als Unbekannte, so existiert im allgemeinen keine exakte Lösung mehr. Das gerade genannte Minimierungsproblem hat aber dennoch eine eindeutige Lösung, die der Matlab-Ausdruck $A \setminus y$ berechnet.

Das oben besprochene lineare Regressionsproblem ist ein Spezialfall dieser Anwendung des Backslash-Operators. Dazu müssen wir das Ausgleichsproblem

$$\sum_{i=1}^{n} (\beta_2 x_i + \beta_1 - y_i)^2 \stackrel{!}{=} \min$$

in Matrixschreibweise übersetzen: A ist nun die Matrix

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \\ 1 & x_n \end{pmatrix},$$

und wir haben

$$\sum_{i=1}^{n} (\beta_2 x_i + \beta_1 - y_i)^2 = \left(A \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} - y_i \right)^2.$$

In Matlab können wir also den Vektor b_1, b_2 als $A \setminus y$ schreiben.

4.13.3 Allometrische Regressionsrechnung

Wir kehren zum Hamsterproblem aus der Einleitung zurück: Gesucht ist die Abhängigkeit zwischen Körpergewicht und Körperlänge bei Goldhamstern; eine Datenerhebung zeigt einen nichtlinearen Zusammenhang — die Meßwerte liegen nicht auf einer Geraden (Abb. 4.12).

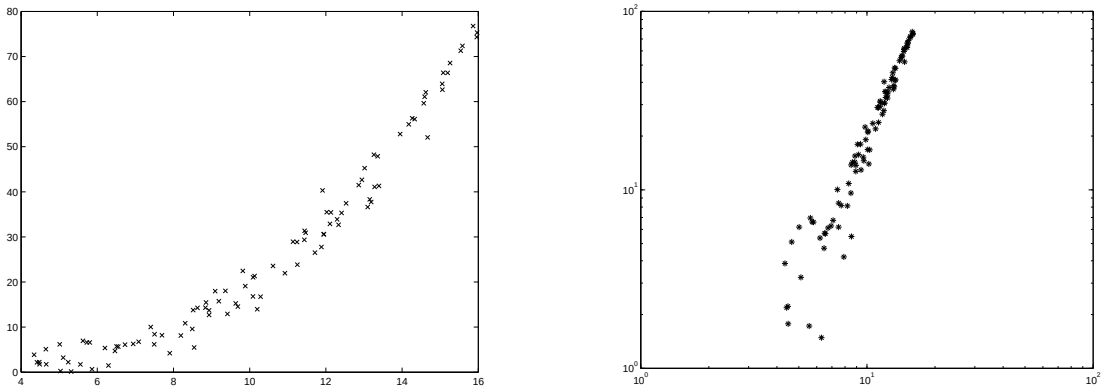


Abbildung 4.12: Allometrie: Gewicht vs. Körpergröße bei Hamstern und doppelt logarithmische Darstellung

Wir setzen

$$\text{Gewicht} = \alpha \cdot (\text{Länge})^{\kappa}$$

mit positivem α, κ an. Diese Werte sollen nun bestimmt werden.

Wie schon in der Einleitung angedeutet, geht unsere Lösungsidee von einer logarithmierten Version des Ansatzes an. Es gilt:

$$\log \text{Gewicht} = \alpha + \kappa \log \text{Länge}.$$

Wir haben das nichtlineare Ausgangsproblem also durch Logarithmieren auf ein lineares Regressionsproblem zurückgeführt.

Seien präziser Meßwerte $(x_1, y_1), \ldots, (x_n, y_n)$ gegeben, die Länge und Gewichte von Goldhamstern beschreiben. Dann möchten wir — wieder mit Hilfe von Matlab — κ, α näherungsweise bestimmen sowie die Kurve $y(x) = \alpha x^{\kappa}$ in das Streudiagramm einzeichnen.

```
% Programm  Hamster.m
% Datenerzeugung Hamstergewicht
laenge=[];
gewicht=[];
for i=1:20
    laenge=[laenge;4+12*rand];
    gewicht=[gewicht;0.019*laenge(i)^3+2*randn];
end
hamster=[laenge, gewicht];
% Visualisierung Messreihe
%orderhamster=sortrows(hamster,1);
figure(1)
%subplot(2,2,1)
%title(['Biometrische Daten von 10 Hamstern'])
hold on
%plot(hamster,'d');
% Streudiagramm und Regressionsgerade
%subplot(2,2,3)
title(['Gewicht[g] gegen Länge [cm]'])
hold on
plot(laenge,gewicht,'+k')
% Allometrische Regression
loglaenge=log(laenge);
loggewicht=log(gewicht);
X=[ones(size(loglaenge)) loglaenge];
linregress=X\loggewicht
hammin=min(laenge);
hammax=max(laenge);
xx=[hammin-2:(hammax+4-hammin)/600:hammax+2];
yy=exp(linregress(1))*exp(linregress(2)*log(xx));
plot(xx,yy,'-r')
% Kubische Regression
X3=[ones(size(laenge)) laenge.^3];
cubregress=X3\gewicht
zz=cubregress(1)+cubregress(2)*xx.^3;
plot(xx,zz,'-g')
```

Abbildung 4.13: Nichtlineares Regressionsproblem unter Matlab

Der erste Teil des Programms in Abbildung 4.13 leistet dies.

Ein alternativer Ansatz — die nichtlineare Regression — baut auf einem Skalierungsargument auf: Wir erwarten, dass das Gewicht der Hamster ungefähr proportional zum Körpervolumen ist; dieses skaliert mit der dritten Potenz der Länge. Wir schätzen also $\kappa = 3$.

Für das Hamstergewicht wird dann ein polynomialer Ansatz gemacht:

$$y = a_1 + a_2 x + a_3 x^2 + a_4 x^3;$$

in diesem Fall vereinfachen wir auf

$$y = b_1 + b_2 x^3.$$

Wie schon bei der linearen Regression haben wir es mit einem überbestimmten Glei-

chungssystem

$$\begin{aligned} y_1 &= b_1 + b_2 x_1^3 \\ &\vdots \\ y_n &= b_1 + b_2 x_n^3 \end{aligned}$$

zu tun, dessen Lösung wir wiederum mit Hilfe des Backslash-Operators bestimmen können. Dies wird im letzten Teil des Programms in Abbildung 4.13 durchgeführt.

4.13.4 Konfidenzintervalle für Regressionskoeffizienten

Seien Zufallsvariablen X, Y sowie unabhängige Paare von Messwerten $(x_1, y_1), \ldots, (x_n, y_n)$ gegeben. Dann erwarten wir einen linearen Zusammenhang zwischen X und Y, also

$$Y = b_1 + b_2 X.$$

Mittels linearer Regression bestimmen wir aus den Daten Schätzwerte

$$\hat{b}_1 = \bar{y} - \hat{b}_2 \bar{x}, \qquad \hat{b}_2 = \frac{\mathrm{Cov}(x, y)}{V(x)}.$$

Wie gut approximieren $\hat{b}_1, \hat{b}_2$ die tatsächlichen Werte? Wie lautet das Konfidenzintervall für die Schätzwerte $\hat{b}_1, \hat{b}_2$?

Wir betrachten die Residuen

$$\hat{\epsilon}_i = y_i - \hat{b}_1 - \hat{b}_2 x_i, \quad i = 1, \ldots, n,$$

und setzen

$$S_n^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

S_n^2 wird als erwartungstreuer Schätzer des Fehlers des linearen Modells bezeichnet. Sei nun

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

die empirische Varianz der X-Daten.

Dann kann man zeigen, dass folgendes gilt: Falls b_2 die tatsächliche Steigung ist, so ist

$$T = \frac{\hat{b}_2 - b_2}{\sqrt{\frac{S_n^2}{(n-1)S_x^2}}}$$

t-verteilt zum Parameter $n-2$. Damit erhalten wir das folgende Konfidenzintervall zum Signifikanzniveau α:

$$b_2 \in \left[\hat{b}_2 - t_{n-2,\alpha/2}\sqrt{\frac{S_n^2}{(n-1)\,S_x^2}}, \hat{b}_2 + t_{n-2,\alpha/2}\sqrt{\frac{S_n^2}{(n-1)\,S_x^2}}\right]$$

Analog können wir Hypothesen zu Regressionskoeffizienten testen: Sei $b_2 = \bar{b}_2$ die Nullhypothese H_0, $b_2 \neq \bar{b}_2$ die Gegenhypothese. Dann wird H_0 verworfen auf Signifikanzniveau α, falls $|T| > t_{n-2,\alpha/2}$.

Der Parameter $n-2$ erklärt sich dadurch, dass für $n \leq 2$ Eindeutigkeit des Regressionsproblems gegeben ist.

4.13.5 Korrelationsrechnung

In diesem Abschnitt entwickeln wir eine Methode, um den Abhängigkeitsgrad zwischen Zufallsvariablen zu bestimmen. Als Beispiel betrachten wir etwa den Zusammenhang zwischen Milch- und Weinproduktion in der Europäischen Union.

Gegeben seien also zwei Zufallsvariablen X und Y sowie eine Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ aus der X,Y-Grundgesamtheit. Wir wissen aus Abschnitt 4.13.1: Gibt es einen linearen Zusammenhang zwischen x_i und y_i, so gilt für $i = 1, \ldots, n$

$$y_i = \frac{\mathrm{Cov}(x,y)}{V(x)}x_i + \bar{y} - \frac{\mathrm{Cov}(x,y)}{V(x)}\bar{x}.$$

Dabei ist

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

die empirische Covarianz. Die obige Gleichung ist nun äquivalent zu

$$y_i - \bar{y} = \frac{\mathrm{Cov}(x,y)}{V(x)}\,(x_i - \bar{x})\,,$$

es gilt also

$$V(y) = \frac{\mathrm{Cov}^2(x,y)}{V^2(x)} V(x).$$

Daraus folgt:

$$\frac{\mathrm{Cov}(x,y)}{\sqrt{V(x)V(y)}} = \pm 1,$$

falls ein linearer Zusammenhang besteht.

Man kann zeigen:

$$-\sqrt{V(x)}\sqrt{V(y)} \leq \mathrm{Cov}(x,y) \leq \sqrt{V(x)}\sqrt{V(y)}.$$

Definition. Die Zahl

$$\rho(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{V(x)}\sqrt{V(y)}}$$

heißt *Pearsonscher empirischer Korrelationskoeffizient.*

Einige Bemerkungen:

- Eine andere Bezeichungsweise ist

  $$\rho = \mathrm{Cor}(x,y).$$

- Kovarianz und Korrelation lassen sich auch auf der Ebene von Zufallsvariablen definieren. So ist z.B.

  $$\mathrm{Cov}(X,Y) := E(X \cdot Y) - E(X) \cdot E(Y)$$

und entsprechend

$$\mathrm{Cor}(X,Y) := \frac{Cov(X,Y)}{\sqrt{V(x)}\sqrt{V(Y)}}.$$

Sind die Zufallsvariablen X und Y also unabhängig voneinander, so ist

$$\mathrm{Cov}(X,Y) = \mathrm{Cor}(X,Y) = 0.$$

Die Umkehrung gilt z.B., wenn X und Y normalverteilt sind.

- $\mathrm{Cov}(X,X) = V(X)$

Sind x_i, y_i Messdaten, so lassen sich umgangssprachlich folgende Zusammenhänge deuten:

1. $\mathrm{Cor}(X,Y) \sim +1$ bedeutet, dass $Y \sim b_1 + b_2 X$ mit $b_2 > 0$.
2. $\mathrm{Cor}(X,Y) \sim -1$ bedeutet, dass die gleiche Beziehung mit $b_2 < 0$ gilt.
3. $\mathrm{Cor}(X,Y) \sim 0$ bedeutet, dass kein linearer Zusammenhang erkennbar ist.

4.13.6 Realisierung von Kovarianzen und Korrelationen unter Matlab

Kovarianzen werden durch den Matlab-Befehl `cov` realisiert. Dieser erzeugt für Vektoren x, y die 2×2-Matrix

$$\begin{pmatrix} V(x) & \mathrm{Cov}(x,y) \\ \mathrm{Cov}(x,y) & V(y) \end{pmatrix}.$$

`corrcoef` erzeugt hingegen die Matrix

$$\begin{pmatrix} 1 & \rho(x,y) \\ \rho(x,y) & 1 \end{pmatrix}.$$

Unter Matlab läßt sich der Korrelationskoeffizient also durch

```
A = corrcoef (x, y);
corr = A(1, 2);
```

berechnen.

4.13.7 Bemerkungen zu Korrelation und Regression

Lokale Korrelationskoeffizienten implizieren nicht notwendigerweise einen linearen Zusammenhang.

Wir weisen insbesondere auf das Phänomen der Scheinkorrelation hin: Besteht die Grundgesamtheit zu X aus mehreren Subpopulationen, und unterscheiden sich die Y-Werte in Abhängigkeit der Zugehörigkeit zur Subpopulation wesentlich, so erhält man betragsmäßig hohe Korrelationskoeffizienten auch dann, wenn X und Y *innerhalb* der Subpopulationen völlig unkorreliert sind.

4.14 Test auf Lageunterschied bei nicht normalverteilten Daten

Wir sind bislang davon ausgegangen, dass die untersuchten Daten annähernd normalverteilt sind; zum Vergleich zweier Stichproben wurde dann der t-Test auf Lageunterschied herangezogen.

Wie geht man aber vor, wenn die Normalverteilungsannahme nicht gerechtfertigt ist, wie etwa bei schiefen Verteilungen, Ausreißern in den Daten oder sehr kleinen Stichproben?

Als Beispiel betrachten wir das Lernverhalten von Ratten: Können Ratten ein erlerntes Verhalten verallgemeinern, wenn sie in eine veränderte Situation mit anderen Motivationsstrukturen versetzt werden? In einem Experiment werden zwei Gruppen von Ratten gebildet. Ratten aus Gruppe 1 sind solche mit Training, Ratten aus Gruppe 2 solche ohne. X bezeichne die Anzahl der Tage, die in Gruppe 1 bis zum Erreichen des Lernziels benötigt werden; Y die Anzahl der Tage, die in Gruppe 2 benötigt werden. Es werden folgende Werte gemessen:

X	110	53	70	51	
Y	78	64	75	45	82

Für solche Fälle eignet sich etwa der Wilcoxon-Regressionstest; es handelt sich um ein nichtparametrisches Verfahren.

Seien dazu zwei unabhängige Stichproben $x_1, \ldots, x_{n_1}$, $y_1, \ldots, y_{n_2}$ gegeben, die gemäß F_1 bzw. F_2 verteilt seien. Besteht ein Lageunterschied Δ zwischen den X- und Y-Daten, d.h. ist

$$P(X_1 - \Delta \leq x) = P(Y_1 \leq x) \text{ für alle } x \in \mathbb{R},$$

oder äquivalent

$$F_1(x - \Delta) = F_2(x)?$$

Wird diese Gleichung mit $\Delta > 0$ erfüllt, so interpretieren wir dies als Aussage, dass die Daten in der ersten Stichprobe tendenziell um Δ größer sind als die der zweiten Stichprobe.

Wir formulieren ein Testproblem; zur Vereinfachung sei dabei $\Delta = 0$ (d.h., kein Lageunterschied). Unsere Hypothesen lauten:

$$\begin{aligned} H_0 &: \Delta = 0 \quad \Leftrightarrow \quad F_1 = F_2 \\ H_1 &: \Delta \neq 0 \quad \Leftrightarrow \quad F_1 \neq F_2 \end{aligned}$$

Die Gesamtstichprobe $x_1, \ldots, x_{n_1}, \ldots, y_1, \ldots, y_{n_2}$ wird zunächst in aufsteigender Reihenfolge angeordnet, und es wird jeweils die Zugehörigkeit zu X oder Y markiert. Jedem Ergebnis wird eine Rangzahl $R_{x_1}, \ldots, R_{x_{n_1}}, R_{y_1}, \ldots, R_{y_{n_2}}$ zugeordnet.

Ergebnis:	45	51	53	64	70	75	78	82	110
Gruppe:	Y	X	X	Y	X	Y	Y	Y	X
Rang:	1	2	3	4	5	6	7	8	9

Sei o.B.d.A. $n_1 \leq n_2$. Dann bilden wir die Summen der Rangzahlen:

$$T_{n_1} = \sum_{i=1}^{n_1} R_{x_i} \qquad T_{n_2} = \sum_{i=1}^{n_2} R_{y_i}$$

Dann muß gelten:

$$T_{n_1} + T_{n_2} = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

Wir erwarten, dass bei gültiger Nullhypothese H_0 die durchschnittlichen Rangsummen in etwa gleich sein sollten. Kombinatorische Betrachtungen liefern:

$$E(T_{n_1}) = \frac{n_1(n_1 + n_2 + 1)}{2} \quad V(T_{n_1}) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

Ist $\max\{n_1, n_2\} < 10$, so läßt sich die Verteilung von T_{n_1} kombinatorisch bestimmen. Bei größeren Stichproben dürfen wir annehmen, dass $T_{n_1} \sim N(E(T_{n_1}), V(T_{n_1}))$.

Um nun einen beidseitigen Test vorzunehmen, standardisieren wir T_{n_1} noch:

$$W = \frac{T_{n_1} - E(T_{n_1})}{\sqrt{V(T_{n_1})}}$$

Um unsere Nullhypothese zu testen, müssen wir nurmehr W auf $(0, 1)$-Normalverteilung testen.

4.15 Rangkorrelation nach Spearman

Zunächst ein Beispiel: Es soll der Zusammenhang zwischen berufsbedingtem Stress (X) und Schlaflosigkeit (Y) untersucht werden. Dazu markieren 6 Personen auf einer Skala von 0–8, die nur ordinal interpretiert werden soll, wie sehr sie unter Stress bzw. Schlaflosigkeit leiden:

i	1	2	3	4	5	6
x_i	3.4	6.1	11.2	5.2	3.3	7.8
y_i	2.8	5.4	1.7	3.9	3.5	7.2
R_{x_i}	3	5	1	4	2	6
R_{y_i}	2	5	1	4	3	6

Um diese Datensätze quantitativ miteinander in Beziehung zu setzen, wird der Korrelationskoeffizient nach Spearman benutzt. In einem ersten Schritt ordnen wir den Ausgangsmessungen x_i und y_i Rangzahlen zu; wir erhalten Paare von Rankings (R_{x_i}, R_{y_i}). Im zweiten Schritt ermitteln wir den Korrelationskoeffizienten nach Pearson für die Rangzahlpaare, wir fragen also nach $R_{Sp} = \rho(R_x, R_y)$. Hierfür gilt die folgende einfache Formel:

$$R_{Sp} = 1 - \frac{6 \sum D_i^2}{n(n+1)(n-1)}, \text{ wobei } D_i = R_{y_i} - R_{x_i}$$

die Rangdifferenz bezeichnet.

R_{Sp} mißt die Stärke des monotonen Zusammenhangs; ist $R_{Sp} = -1$, so sind die Daten monoton fallend; ist $R_{Sp} = +1$, so sind sie monoton wachsend.

Sollen Hypothesen getestet werden, so wird die Teststatistik

$$T = \frac{R_{Sp}\sqrt{n-2}}{\sqrt{1-R_{Sp}^2}}$$

mit den Quantilen der $t(n-2)$-Verteilung verglichen.

4.16 Kontingenztafeln und χ^2-Unabhängigkeitstests

Sie interessieren sich für die Abhängigkeit zwischen „guten“ Noten in den Fächern Mathematik und Biologie. Als „gut“ gelten dabei die Noten „gut und sehr gut“, wir betrachten also ein nominales Merkmal mit zwei möglichen Ausprägungen (auch: *binäres* Merkmal).

4.16.1 Φ-Kontingenzkoeffizient für 2×2-Tafeln

Die beobachtete Notenverteilung werde in einer Kontingenztafel eingetragen:

	X		
Y	0	1	
0	a	b	$a+b$
1	c	d	$c+d$
	$a+c$	$b+d$	n

	Mathematik		
Biologie	1–2	3–6	
1–2	20	30	50
3–6	10	40	50
	30	70	100

Für ein beliebiges binäres Merkmal kodieren wir die möglichen Ausprägungen des Merkmals mit 0 und 1; wir haben dann Beobachtungspaare (x_i, y_i) mit $i = 1, \ldots, n$. Wir berechnen den Pearsonschen Korrelationskoeffizienten.

Mit den Bezeichnungen aus der Kontingenztafel gilt:

$$\sum_{i=1}^{n} x_i y_i = d \quad \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 = c+d \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i^2 = b+d$$

Daraus erhalten wir:

$$\rho(x,y) = \frac{ad-bc}{\sqrt{(c+d)(a+b)(b+d)(a+c)}} \in [-1,1].$$

Wir definieren:

$$\Phi(x,y) = \rho^2(x,y) = \frac{(ad-bc)^2}{(c+d)(a+b)(b+d)(a+c)} = \frac{\text{Determinante}}{\text{Produkt der Randsummen}}$$

Φ bezeichnen wir als den Φ-*Kontingenzkoeffizienten*. Er nimmt Werte zwischen 0 und 1 an und ist ein Maß für den Zusammenhang zwischen den betrachteten Größen.

4.16.2 Vergleich diskreter Verteilungen

Die Wirksamkeit eines Medikaments soll untersucht werden. Für eine Kontrollgruppe und eine Testgruppe wird gemessen, ob der Zustand sich verbessert oder verschlechtert, oder ob keine Änderung meßbar ist. Wir erhalten die folgende Kontingenztafel:

	schlechter (1)	unverändert (2)	besser (3)	Zeilensumme
Kontrollgruppe (1)	$p_{11} = 8$	$p_{12} = 15$	$p_{13} = 7$	$Z_1 = 30$
Testgruppe (2)	$p_{21} = 5$	$p_{22} = 10$	$p_{23} = 15$	$Z_2 = 30$
Spaltensummen	$S_1 = 13$	$S_2 = 25$	$S_3 = 22$	$N = 60$

Wir stellen die Nullhypothese auf, dass das Medikament unwirksam ist, also $p_1 = p_{11} = p_{21}$, $p_2 = p_{12} = p_{22}$, $p_3 = p_{13} = p_{23}$. Für H_1 gilt dann entsprechend, dass $(p_{11}, p_{12}, p_{13}) \neq (p_{21}, p_{22}, p_{23})$.

Unter der Nullhypothese folgt aus einer Maximum-Likelihood-Schätzung für die Wahrscheinlichkeiten p_i die Schätzwerte $\hat{p}_i = S_i/N$. Der Erwartungswert für die Belegungszahl der (i,j)-ten Zelle ist dann $\hat{E}_{ij} = Z_i \hat{p}_j = \frac{Z_i S_j}{N}$.

Wir vergleichen nun den geschätzten Erwartungswert mit der tatsächlichen Anzahl N_{ij}; im allgemeinen Fall mit k Zeilen und ℓ Spalten haben wir dann folgende Teststatistik:

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(N_{ij} - Z_i S_j/N)^2}{Z_i S_j/N}$$

Unter H_0 ist Q näherungsweise χ^2-verteilt mit $df = (k-1)(\ell-1)$ Freiheitsgraden.

Im Beispiel haben wir $Q = 4.601$, $df = (3-1)(2-1) = 2$. Es gilt $\chi^2(2)_{0.95} = 5.991 > 4.601$, H_0 kann also auf einem Signifikanzniveau von 0.05 nicht verworfen werden.

4.16.3 Unabhängigkeitstest und Kreuzklassifikation

Gefragt wird, ob es eine statistische Abhängigkeit zwischen dem Kindstod in der Perinatalperiode (28. Schwangerschaftswoche bis 7. Lebenstag des Kindes) und Rauchen der Mutter während der Schwangerschaft gibt. Eine Erhebung an 19380 Schwangeren ergibt folgende Ergebnisse (Daten aus [4]):

Mortalität	Raucherin	Nichtraucherin	$\sum$
ja	246	264	$Z_1 = 510$
nein	8160	10710	$Z_2 = 18870$
$\sum$	$S_1 = 8406$	$S_2 = 10974$	$N = 19380$

Seien wieder p_{ij} die Zellenwahrscheinlichkeiten, p_{iZ} die Wahrscheinlichkeit des i-ten Zeilenkriteriums (Mortalität), p_{Sj} die Wahrscheinlichkeit des j-ten Spaltenkriteriums (Raucherin).

Wir wissen: Gilt $p_{ij} = p_{iZ}p_{Sj}$ für alle i, j, so sind die beiden Zufallsvariablen stochastisch unabhängig. Die Nullhypothese H_0 sei, dass $p_{ij} = p_{iZ}p_{Sj}$ für alle i, j erfüllt ist; H_1 besage, dass dieses Kriterium für mindestens ein i, j nicht erfüllt ist.

Wir nehmen wieder die Nullhypothese an; Maximum-Likelihood-Schätzung der Zellenwahrscheinlichkeiten ergibt Schätzwerte

$$\hat{p}_{iZ} = \frac{Z_i}{N} \qquad \hat{p}_{Sj} = \frac{S_j}{N}.$$

Die erwarteten Belegungszahlen unter H_0 sind dann

$$\hat{E}_{ij} = N\hat{p}_{iZ}\hat{p}_{Sj} = \frac{S_j Z_i}{N}.$$

Wir erhalten wieder eine χ^2-verteilte Teststatistik:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{\ell} \frac{(N_{ij} - Z_i S_j/N)^2}{Z_i S_j/N};$$

die Anzahl der Freiheitsgrade beträgt $df = (k-1)(\ell-1)$.

In unserem Beispiel ist $Q = 5.038 > \chi^2(1)_{0.95} = 3.842$, H_0 muss auf Signifikanzniveau $\alpha = 0.05$ verworfen werden.

4.16.4 Tests auf Trends

Bei der Einführung eines neuen Medikaments soll getestet werden, ob höhere Dosierungen zu höheren Genesungsraten führen. 43 kranke Versuchstiere erhalten die Substanz mit den Dosierungen $d_0 = 0$, $d_1 = 1$, $d_2 = 3$, $d_3 = 4$.

Folgendes Ergebnis wird beobachtet:

Dosierung	0	1	3	4	$\sum$
gesund	3	4	6	7	$Z_1 = 20$
krank	7	6	6	4	$Z_2 = 23$
S_i	10	10	12	11	43

Wir interpretieren die Situation wahrscheinlichkeitstheoretisch: Für jede Dosierung d_i liegt ein Binomialexperiment mit n_i Wiederholungen und Erfolgswahrscheinlichkeit p_i vor. Ein Trend liegt dann vor, wenn die p_i durch eine streng monoton wachsende Funktion F, $F(d_i) = p_i$ dargestellt werden können.

Wir stellen wiederum Hypothesen auf: Die Nullhypothese „kein Trend“ entspricht Gleichheit der $p_0 = p_1 = \cdots = p_k$; H_1 entspreche einen Trend $p_0 \leq p_1 \leq \cdots \leq p_i < p_{i+1} \leq \cdots p_k$, wobei mindestens eine der Ungleichungen strikt sein soll. Der Trend-Test von Cochran-Armitage betrachtet dann die folgende Testgröße:

$$Q = \frac{\left(\sum_{i=1}^{k} x_i d_i - \hat{p} \sum_{i=1}^{k} S_i d_i\right)^2}{\hat{p}\hat{q}\left(\sum_{i=1}^{k} S_i d_i^2 - \frac{1}{N}\left(\sum_{i=1}^{k} S_i d_i\right)^2\right)}$$

Dabei sei $\hat{p} = Z_1/N$, $\hat{q} = 1 - \hat{p} = Z_2/N$. Man beachte, dass

$$E_{H_0}\left(\sum_{i=1}^{k} x_i d_i\right) = \sum_{i=1}^{k} S_i \hat{p} d_i = \hat{p} \sum_{i=1}^{k} S_i d_i$$

ist. Der Zähler sollte also verschwinden, wenn H_0 erfüllt ist.

Man kann zeigen, dass Q unter H_0 in großen Stichproben annähernd $\chi^2(1)$-verteilt ist.

4.17 Anpassungstests

Gegeben sei eine Zufallsstichprobe, etwa die Körpergröße der Hörer einer Vorlesung im Biologie-Grundstudium. Ist diese Zufallsstichprobe mit einem Verteilungsmodell, z.B. einer Normalverteilung, verträglich?

4.17.1 Quantildiagramme

Ein erster Lösungsansatz besteht darin, ausgewählte empirische Quantile mit den theoretischen Quantilen einer Verteilung zu vergleichen. Dazu ordnen wir die Daten $x_1, \ldots, x_n$ zunächst in aufsteigender Reihenfolge an; wir nehmen also an, dass $x_1 \leq x_2 \leq \cdots \leq x_n$. Sei Φ die Verteilungsfunktion der zu prüfenden Verteilung. Dann tragen wir die Werte

$$\Phi^{-1}\left(\frac{i-0.5}{n}\right)$$

gegen die x_i auf: Liegen die Punktepaare näherungsweise auf der Winkelhalbierenden, so kann die Verteilungsfunktion Φ akzeptiert werden; dieses Diagramm bezeichnen wir als QQ-Plot.

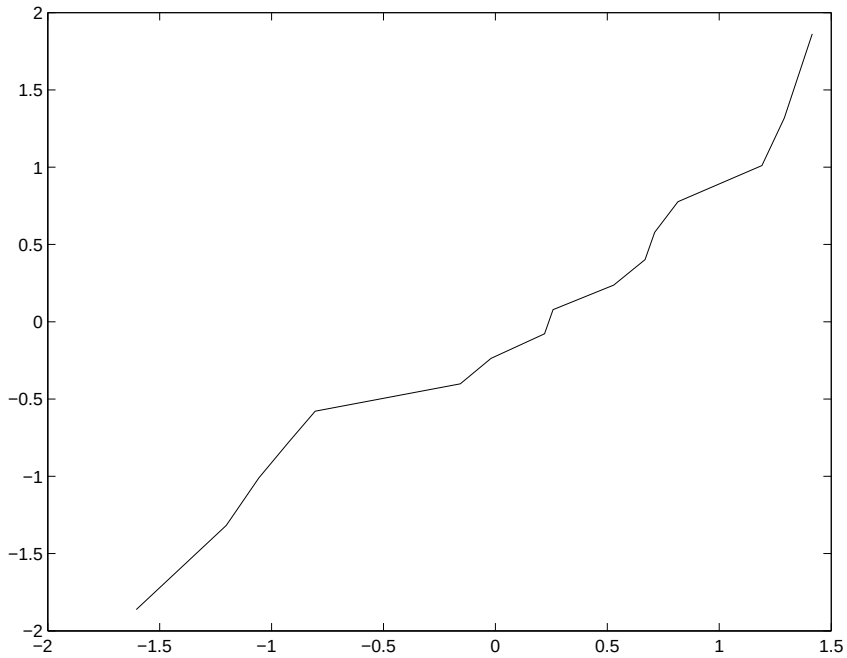


Abbildung 4.14: QQ-Plot für eine (0,1)-normalverteilte Stichprobe

In der Praxis ist häufig lediglich von Interesse, ob Daten normalverteilt sind; Φ soll dann die Verteilungsfunktion zur Normalverteilung sein.

Die q-Quantile der $N(\mu, \sigma)$-Verteilung lassen sich linear aus den q-Quantilen der $N(0, 1)$-Verteilung berechnen:

$$z_q(\mu, \sigma) = \mu + z_q(0, 1)\sigma$$

Für den QQ-Plot ziehen wir also die Quantile der $N(0, 1)$-Verteilung heran; Normalverteilung $N(\mu, \sigma)$ akzeptieren wir dann, wenn die Punktepaare auf einer Geraden mit Steigung σ und Achsenabschnitt μ liegen.

4.17.2 Korrelationstests

Die Anpassungsgüte des QQ-Plots soll durch Bestimmung des Korrelationskoeffizienten getestet werden. Zu den Datenpaaren

$$(x_i, q_i) = \left(x_i, \Phi^{-1} \left(\frac{i - 0.5}{n} \right) \right)$$

berechnen wir

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

Ist R hinreichend groß, so wird Φ als Verteilungsmodell akzeptiert.

Die kritischen Werte für eine Akzeptanz hängen dabei von Verteilungsmodell und Stichprobenumfang ab, z.B. gilt:

$$C_{0,05}(n) = \begin{cases} \frac{4.7+0.745n}{5.7+0.745n} & \text{Normalverteilung} \\ \frac{0.107+1.12n}{1.107+1.12n} & \text{Gleichverteilung auf } [0, 1]. \end{cases}$$

4.17.3 χ^2-Anpassungstest

Gegeben sei eine Stichprobe $x_1, \ldots, x_n$ von unabhängigen, identisch verteilten Daten. Gefragt ist wieder nach der Verträglichkeit mit der Verteilung Φ.

Wir unterteilen dann die x-Achse in k Klassen

$$A_1 = [g_1, g_2], A_2 = (g_2, g_3], \ldots, A_k = (g_k, g_{k+1}]$$

und zählen die Beobachtungen in den einzelnen Klassen aus; sei

$$g_\ell = \#\{x_i \in A_\ell\}.$$

Wir ermitteln nun die Wahrscheinlihckeit, dass eine Beobachtung in der ℓ-ten Klasse auftritt:

$$p_\ell = P(x_\ell < X \leq g_{\ell+1}) = \Phi(g_{\ell+1}) - \Phi(g_\ell)$$

Dann ist $e_\ell = np_\ell$ der Erwartungswert für die Anzahl an Beobachtungen in Klasse A_k.

Wir formulieren wie üblich unsere Hypothesen: Nullhypothese H_0 sei, dass die Beobachtungen nach Φ verteilt sind, also

$$P(A_1) = p_1, \ldots, P(A_k) = p_k$$

Gegenhypothese H_1 ist, dass $P(A_\ell) \neq p_\ell$ für ein ℓ.

Die Prüfgröße ist wie folgt gegeben:

$$Q = \sum_{\ell=1}^{k} \frac{(b_\ell - e_\ell)^2}{e_\ell}.$$

Q ist χ^2-verteilt mit $(k-1)$ Freiheitsgraden. H_0 ist auf Signifikanzniveau α zu verwerfen, falls

$$Q > \chi^2(k-1)_{1-\alpha}.$$

Übungen

Aufgabe 1

Ungefähr jedes achte Schaf ist schwarz. In einer Zufallspopulation von 10000 Schafen wird die Anzahl X der schwarzen Schafe ermittelt.

1. Wie müssen Sie die Parameter μ und σ^2 der Normalverteilung wählen, damit diese die Verteilung von X approximiert?

2. Berechnen Sie näherungsweise die Wahrscheinlichkeit dafür, dass die Anzahl der schwarzen Schafe

   (a) genau 1250 ist,

   (b) mindestens 1600 beträgt,

   (c) zwischen 1000 und 1500 liegt,

   (d) weniger als 10 ist.

Aufgabe 2

Das Gewicht von Mäusen (in g) aus einer Population sei eine normalverteilte Zufallsgrösse X mit unbekannten Parametern. Um den Mittelwert μ zu schätzen, bestimmen Sie die Gewichte x_i von n Mäusen.

Zeigen Sie, dass das arithmetische Mittel $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ der Maximum-Likelihood-Schätzer für μ ist.

Hinweis: Betrachten Sie für beliebiges $\bar{\mu} \in \mathbb{R}$ die Wahrscheinlichkeitsdichte

$$f(X_1, ..., X_n|\bar{\mu}) := \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{(X_1-\bar{\mu})^2}{2\sigma^2}\right) \cdots \exp\left(-\frac{(X_n-\bar{\mu})^2}{2\sigma^2}\right),$$

setzen Sie für X_i die Messwerte x_i ein und bestimmen Sie das Maximum bezüglich $\bar{\mu}$.

Aufgabe 3

Eine Stichprobe von 20 Amseleiern ergibt folgende Gewichte (in g):

8.4	7.0	7.0	7.2	7.8	7.9	7.6	8.3	7.5	9.0
8.7	7.9	7.0	7.7	7.1	6.7	7.8	6.6	7.6	7.3

1. Schätzen Sie (erwartungstreu) Erwartungswert und Varianz der zu Grunde liegenden Verteilung.

2. Angenommen, das Gewicht sei normalverteilt, wieviel Prozent aller Amseleier wiegen dann zwischen 7 und 8 Gramm?

3. Nehmen Sie an, das Gewicht sei normalverteilt. Geben Sie ein Konfidenzintervall für die Varianz zur Irrtumswahrscheinlichkeit $\alpha = 5\%$ an.

Aufgabe 4

Die Zufallsvariable X sei $N(10, 4)$–verteilt. Bestimmen Sie das Konfidenzintervall für den Erwartungswert von X zur Irrtumswahrscheinlichkeit $\alpha = 0.05$. Wie ändert sich das Konfidenzintervall, wenn Sie die Irrtumswahrscheinlichkeit auf $\alpha' = 2\%$ drücken wollen? (Argumentation genügt.)

Aufgabe 5

Die Imker der Region vergleichen Ihre Bienenpopulationen. Dabei werden folgende Größen der Bienenvölker erhoben.

22822	29168	24170	21445	28899	25671	26175	30283	21475	27292
23429	14765	25288	20524	16678	11891	26864	23751	23932	29339

Bestimmen Sie das jeweils Konfidenzintervall für die mittlere Bienenzahl pro Volk zu Irrtumswahrscheinlichkeiten von $\alpha_1 = 5\%$ und $\alpha_2 = \frac{1}{100}$.

Aufgabe 6

Bei einem in Tablettenform produzierten Medikament ist der Wirkstoff gleichmäßig in der Tablette verteilt, die Menge des Wirkstoffs also proportional zum Gewicht. Folgende Gewichte (in g) wurden ermittelt:

0.96	0.94	0.94	1.01	0.91	1.09	1.00	1.01	0.99	1.02
1.03	0.95	0.97	0.99	0.91	0.97	0.96	0.96	1.05	0.95

1. Testen Sie die Nullhypothese $H_0 : \mu = 1\text{g}$ gegen die Gegenhypothesen $H_1 : \mu \neq 1\text{g}$ und $H_2 : \mu < 1\text{g}$, jeweils auf einem Signifikanzniveau von 5%.
2. Testen Sie die Nullhypothese $H_0 : \mu = 0.95\text{g}$ gegen die Gegenhypothese $H_3 : \mu > 0.95\text{g}$ mit Signifikanzniveau 1%.

Aufgabe 7

An zehn Personen wird eine neue Diät ausprobiert. Es wird die folgende Messreihe (in kg) erhoben.

vorher	87.9	74.2	71.6	74.0	52.1	72.0	61.3	59.1	76.6	76.7
nacher	87.0	74.6	71.8	73.1	52.8	70.3	61.2	59.8	77.2	72.4

Bewirkt die Diät eine Gewichtsabnahme? Stellen Sie Null- und Gegenhypothese auf und testen Sie zu Signifikanzniveaux von $\alpha_1 = 5\%$ und $\alpha_2 = 1\%$.

Aufgabe 8

Die Hasenmaulfledermaus (*noctilio leporinus*) ernährt sich unter anderem von Fischen. Bei der Gewichtsverteilung soll zwischen zwei Populationen, die eine an einem See, die andere am Meer beheimatet, verglichen werden. Biologen erfassen folgende Gewichte (in g):

See	45.8	48.3	49.9	48.4	56.2	46.1
Meer	52.8	55.0	50.0	53.8		

Testen Sie die Hypothese, dass beide Populationen gleiches erwartetes Gewicht haben, zum Signifikanzniveau $\frac{5}{100}$. Geben Sie explizit Nullhypothese und Gegenhypothese an.

Literaturverzeichnis

(1) Batschelet, F.: Einführung in die Mathematik für Biologen. Berlin Heidelberg New York: Springer 1980.

(2) Bohl, E.: Mathematik in der Biologie. Berlin Heidelberg New York: Springer 2001.

(3) Murray, J. D.: Mathematical Biology I. An Introduction. Berlin Heidelberg New York: Springer 2002.

(4) Steland, A.: Mathematische Grundlagen der empirischen Forschung. Heidelberg: Springer 2004.

(5) Timischl, W.: Biostatistik. Eine Einführung für Biologen. Wien New York: Springer 1990.

Anhang A

Matlab: Kurzreferenz

A.1 Einführung

A.1.1 Eingabe von Befehlen

Matlab-Befehle können sowohl im Command Window, als auch im Programmeditor eingegeben werden. Ein Befehl endet grundsätzlich am Zeilenende; sollen mehrere Befehle in einer Zeile eingegeben werden, so können sie durch Semikolon getrennt werden.

Matlab gibt das Resultat des letzten Befehls in einer Zeile aus. Ist der letzte Befehl in einer Zeile leer, d.h., schließt die Zeile mit einem Semikolon ab, so wird keine Ausgabe erzeugt.

Werden Befehle im Command Window eingegeben, so wird jede Zeile unmittelbar nach Abschluß der Eingabe ausgeführt. Werden die Befehle im Editor eingegeben, so wird das Programm durch Aufruf des Menüpunkts „Run“ im Menü „Debug“ ausgeführt.

Beispiele in diesem Skript sind üblicherweise Eingaben im Command Window und die von Matlab gegebenen Antworten. Dabei sind die Zeilen, die mit `>>` beginnen, Eingaben des Nutzers. Zeilen, die mit `???` beginnen, sind Fehlermeldungen von Matlab. Mit `ans` beginnt Matlab die Anzeige eines Rückgabewertes.

```
>> 5*5

ans =

    25

>> 5*5;
>> 5*5; 2*2

ans =

    4

>>
```

A.1.2 Daten: Zahlen, Zeichenketten, Matrizen

Matlab kann insgesamt vierzehn fundamentale Datentypen bearbeiten. Wir führen einige der meistgebrauchten vor.

Zahlen können als Dezimalbrüche oder auch in wissenschaftlicher Schreibweise eingegeben werden:

```
>> 1.05
ans =
    1.0500

>> 1.05e3
ans =
    1050

>>
```

Zeichenketten werden in Apostrophen eingeschlossen:

```
>> Zeichenkette
??? Undefined function or variable 'Zeichenkette'.

>> 'Zeichenkette'

ans =

     Zeichenkette

>>
```

Die eigentliche Stärke von Matlab liegt im Verarbeiten von Vektoren und Matrizen. Diese werden in eckigen Klammern eingegeben; ein Semikolon beendet eine Zeile.

```
>> [1 2 3]

ans =

1     2     3

>> [1; 2; 3]

ans =

1
2
3

>> [1 2 ; 3 4 ]

ans =

1     2
        3     4

>> [1 2 ; 3 4] * [1; 2]

ans =

      5
      11
```

In diesem Beispiel wird zunächst ein Zeilen-, dann ein Spaltenvektor angegeben, dann eine 2×2-Matrix. In der letzten Eingabezeile wird das Produkt zwischen einer Matrix und einem Spaltenvektor ausgerechnet, das wiederum einen Spaltenvektor ergibt.

Zeilenvektoren können außerdem mit Hilfe des Doppelpunktoperators erzeugt werden: Der Ausdruck *Start*:*Schritt*:*Ende* erzeugt einen Zeilenvektor, der mit der gegebenen Schrittweite vom Start- zum Endwert hochzählt.

```
>> 0:0.1:1

ans =

  Columns 1 through 6

         0    0.1000    0.2000    0.3000    0.4000    0.5000

  Columns 7 through 11

         0.6000    0.7000    0.8000    0.9000    1.0000

>>
```

Die Schrittweite kann auch ausgelassen werden: *Start*:*Ende* zählt in Einer-Schritten hoch.

```
>> 0.1:10

ans =

  Columns 1 through 6

    0.1000    1.1000    2.1000    3.1000    4.1000    5.1000

  Columns 7 through 10

    6.1000    7.1000    8.1000    9.1000

>>
```

A.1.3 Variablen und Zuweisungen

Werte können zwecks Zwischenspeicherung und weiterer Verwendung Variablen zugewiesen werden. Variablennamen müssen mit einem Buchstaben beginnen; es können beliebige Kombinationen aus Buchstaben, Zahlen und Unterstrichen folgen. Groß- und Kleinschreibung wird unterschieden, `A` und `a` sind also unterschiedliche Variablennamen.

Eine Zuweisung wird mit einem einfachen Gleichheitszeichen notiert:

```
>> a=55

a =

    55

>> b=a*a

b =
    3025

>>
```

Man beachte, dass Matlab in der Antwort nun nicht mehr `ans` verwendet, sondern den Namen der verwendeten Variablen. Tatsächlich ist `ans` eine spezielle Variable, die das Ergebnis von Ausdrücken aufnimmt, die nicht explizit zugewiesen werden:

```
>> 5*5;
>> ans

ans =

    25
```

A.1.4 Kontrollstrukturen: `if` und `while`

Wir haben bereits gesehen, dass Matlab-Instruktionen in der Reihenfolge ausgeführt werden, in der sie im Kommando-Fenster eingegeben werden, oder in der sie in einem im Editor bearbeiteten „M-File“ stehen.

Soll Programmcode wiederholt oder nur bedingt ausgeführt werden, so werden hierzu Kontrollstrukturen verwendet. Wir führen zunächst die `if`-Anweisung und die `while`-Schleife ein.

`if` ermöglicht die bedingte Ausführung von Programmteilen: Abhängig von einer Bedingung werden unterschiedliche Teile des Programms ausgeführt.

```
if (Bedingung 1)
  Code 1
elseif (Bedingung 2)
  Code 2
else
  Code 3
end
```

Im Beispiel wird, falls *Bedingung 1* erfüllt ist, *Code 1* ausgeführt. Ist *Bedingung 1* nicht erfüllt, trifft aber *Bedingung 2* zu, so wird *Code 2* ausgeführt. Ist keine der Bedingungen erfüllt, so wird *Code 3* durchlaufen.

Allgemein können auf eine `if`-Abfrage beliebig viele `elseif`-Zweige sowie höchstens ein `else`-Zweig folgen. Der letzte dieser Zweige muß mit `end` abgeschlossen werden. Es ist natürlich ebenfalls legitim, überhaupt keine `else`- oder `elseif`-Zweige zu verwenden; in diesem einfachsten Fall ergibt sich die folgende einfache Struktur:

```
if (Bedingung)
  Instruktionen
end
```

`while` ermöglicht die bedingte wiederholte Ausführung eines Programmteils: Ein Block wird ausgeführt, *solange* eine Bedingung erfüllt ist.

```
while (Bedingung)
  Instruktionen
end
```

Die Bedingung wird dabei vor *jedem* (auch dem ersten!) Durchlauf durch die Schleife überprüft.

A.1.5 Logische Ausdrücke

Für `while`-Schleifen wie für `if`-Abfragen werden Ausdrücke benötigt, die die Wahrheitswerte „wahr“ oder „falsch“ annehmen können.

Hierzu können einerseits einfache Relationen herangezogen werden:

`<`	`<=`	`>`	`>=`	`==`	`~=`
$<$	$\leq$	$>$	$\geq$	$=$	$\neq$

Diese Relationen können auf Skalare oder auf Matrizen und Vektoren gleicher Dimension angewandt werden; dabei erfolgt der Vergleich elementweise. Es ist ein Fehler, die Relationen auf Matrizen oder Vektoren unterschiedlicher Dimension anzuwenden.

Verschiedene Bedingungen können mit Klammern gruppiert und mit den üblichen logischen Operatoren verknüpft werden:

`&`	`|`	`~`
und	oder	nicht

A.1.6 `for`-Schleife

Eine `for`-Schleife durchläuft der Reihe nach alle Spalten einer Matrix:

```
for Variable=Matrix
  Anweisungen
end
```

Dabei werden die gegebenen Anweisungen für jede Spalte der Matrix einmal ausgeführt, die angegebene Variable enthält dann jeweils den Spaltenvektor.

In der geläufigsten Form der `for`-Schleife ist die angegebene Matrix ein Zeilenvektor, der wie schon oben besprochen aus Startwert, Schrittweite und Endwert zusammengesetzt wird. Die Laufvariable ist dann ein Skalar:

```
for Variable=Start:Schritt:Ende
  Anweisungen
end
```

A.1.7 Arithmetische Ausdrücke und Operatoren

Wir listen kurz die arithmetischen Operatoren in Matlab auf:

`+`	Addition
`-`	Subtraktion
`.*`	Komponentenweise Multiplikation
`./`	Komponentenweise Division nach rechts
`.\`	Komponentenweise Division nach links
`:`	Doppelpunktoperator
`.^`	Exponentiation
`.'`	Transposition
`'`	Komplex konjugierte Transposition
`*`	Matrixmultiplikation
`/`	Matrixdivision nach rechts
`\`	Matrixdivision nach links

Einer zusätzlichen Erklärung bedarf vor allem die "Matrixdivision nach links," der "Backslash-Operator." Und zwar berechnet `A \ b` die Lösung c des Gleichungssystems

$$Ac = b,$$

wenn diese wohlbestimmt ist. Ist A eine singuläre quadratische Matrix, so wird der Backslash-Operator einen Fehler erzeugen; ist A nahezu singulär, so wird eine Warnung erzeugt.

Der Backslash-Operator kann jedoch auch mit überbestimmten Systemen umgehen: Hat A etwa mehr Zeilen als Spalten, so existiert im allgemeinen kein c, das das Gleichungssystem lösen würde. Stattdessen greift Matlab zur Ausgleichsrechnung: c wird bestimmt, so dass $|Ac - b|^2$ minimal ist.

A.1.8 Mehr über Matrizen: Zugriff auf Matrixelemente

Zum Abschluss des Einführungskapitels noch einige weitere Hinweise über Matrizen. Sei dazu `A` eine Variable, die eine Matrix beschreibt. Ein Eintrag dieser Matrix wird dann durch `A(i,j)` gegeben, wobei `i` die Zeile, `j` die Spalte angibt. Dem Element in der linken oberen Ecke der Matrix entspricht dabei `A(1,1)` – nicht etwa `A(0,0)`, wie es von manchen anderen Programmiersprachen her bekannt ist.

Der Doppelpunktoperator kann benutzt werden, um Teilbereiche aus einer Matrix auszulesen: `A(1,:)` ist der erste Zeilenvektor einer Matrix; `A(:,1)` der erste Spaltenvektor. Mit `A(2,3:5)` könnte man das dritte bis fünfte Element aus der zweiten Zeile der Matrix auslesen.

Allgemeiner können beliebige Zeilenvektoren als Indizes beim Zugriff auf Matrixelemente benutzt werden: So ist es etwa völlig legitim und auch gelegentlich nützlich, Teile einer Matrix etwa wie folgt auszulesen:

```
>> A = magic(10)
...
>> A ([1 3 5], [1 3 5])

ans =
92   1  15
 4  88  22
       86  25   9

>>
```

Der Zugriff auf einzelne Elemente eines Zeilenvektors erfolgt analog hierzu.

A.1.9 Funktionen

Benutzerdefinierte Funktionen werden in `m`-Dateien mit dem gleichen Namen abgelegt, die Funktion `Konfiplot` etwa in einer Datei `Konfiplot.m` (siehe Abbildung 4.3 für ein vollständiges Funktionsbeispiel). Diese Dateien beginnen mit einer Funktionsdeklaration, etwa der folgenden:

```
function [M,MCI] = Konfiplot(X,alpha)
```

Diese Deklaration teilt Matlab mit, dass die Funktion mit dem Namen `Konfiplot` zwei Variablen (namens `X` und `alpha`) als Eingabe erwartet und einen Vektor mit zwei Spalten zurückgibt. Die Elemente dieses Spaltenvektors werden `M` und `MCI` genannt. Sie können im auf die Deklaration folgenden Matlab-Code beschrieben werden; ihre Werte am Ende der Matlab-Instruktionen, die die Funktion ausmachen, werden an den aufrufenden Programmcode zurückgegeben.

Aufgerufen wird eine Funktion wie folgt:

```
[a,b] = Konfiplot (X, alpha);
```

A.2 Bibliotheks-Funktionen

Wir geben in diesem Abschnitt eine alphabetische Referenz einiger der wichtigsten Matlab-Funktionen, die im Rahmen der Vorlesung vorgestellt wurden.

Implementiert sind viele dieser Funktionen als `m`-Files; wo dies der Fall ist, kann diese Implementierung mit dem Befehl

```
type Funktion
```

ausgegeben werden. Befehle wie etwa `plot` sind direkt in Matlab implementiert; `type` wird in diesem Fall eine geeignete Fehlermeldung ausgeben.

A.2.1 `axis`

Der Befehl `axis` setzt die Eigenschaften der Achsen in einem mit Matlab erzeugten Funktionsplot. Übergeben wird ein Zeilenvektor, der nacheinander die minimalen und maximalen x- und y-Werte angibt. Soll also die x-Achse von 0 bis 10 reichen, die y-Achse aber von -10 bis $+10$, so würde man folgenden Befehl benutzen:

```
axis ([ 0 10 -10 10]);
```

A.2.2 `bar`

Der Befehl `bar` erzeugt eine Balkengraphik. Wird nur eine Matrix übergeben (`bar(X)`), so wird für jedes Element der Matrix ein Balken gezeichnet; dabei werden die Balken zeilenweise gruppiert.

Wird zusätzlich ein Vektor mit Positionen übergeben (`bar(x,Y)`), so werden die Balken an den angegebenen x-Positionen dargestellt. Elemente werden wiederum zeilenweise gruppiert.

A.2.3 `boxplot`

Der Befehl `boxplot` erzeugt den aus Abschnitt 2.5.4 bekannten Boxplot. Dabei werden für jede *Spalte* der übergebenen Matrix Median, unteres und oberes Quartil sowie Maximum und Minimum berechnet und eingezeichnet.

A.2.4 `cdfplot`

`cdfplot` zeichnet eine empirische kumulative Verteilungsfunktion zu einem übergebenen Zeilenvektor `X`. Dabei ist die gezeichnete Funktion $F(x)$ definiert als der Anteil von Werten in `X`, der kleiner oder gleich x ist.

A.2.5 `corrcoef`

`corrcoef` berechnet für zwei übergebene Spaltenvektoren die Matrix

$$\begin{pmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{pmatrix},$$

siehe Abschnitt 4.13.6.

A.2.6 `cov`

`cov` berechnet für zwei übergebene Spaltenvektoren die Matrix

$$\begin{pmatrix} V(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & V(Y) \end{pmatrix},$$

siehe Abschnitt 4.13.6.

A.2.7 `diff`

`diff` berechnet die Differenzen zwischen den benachbarten Einträgen des übergebenen Vektors.

A.2.8 `exp`

`exp` ist die elementweise Exponentialfunktion. Auf einen Skalar angewandt, wird der skalare Wert der Exponentialfunktion zurückgegeben. Auf Vektoren oder Matrizen angewendet, wird die Exponentialfunktion für die einzelnen Einträge berechnet.

Man beachte, dass dies *nicht* die aus der Mathematik bekannte *Matrix-Exponentialfunktion* ist.

A.2.9 `figure`

`figure (`N`)` erzeugt ein neues Fenster mit der Nummer N oder bringt das Fenster Nummer N in den Vordergrund. Darauf folgende Graphikbefehle beziehen sich dann auf das ausgewählte Fenster.

A.2.10 `hist`

`hist` dient der Erzeugung von Histogrammen. In der einfachsten Variante wird ein einziger Vektor übergeben. `hist` erzeugt und zeichnet dann ein Histogramm mit zehn Klassen.

```
hist(X)
```

In einer weiteren Variante wird `hist` mit zwei Vektoren aufgerufen. In diesem Fall werden die Klassen um die Elemente des zweiten Vektors zentriert. Sollen die Klassen etwa um die Punkte 0.5, 1, 2 zentriert sein, so würde der Aufruf wie folgt lauten:

```
hist(X, [0.5 1 2])
```

Will der Nutzer die Mittelpunkte der Klassen nicht selbst bestimmen, so kann er auch deren Anzahl angeben:

```
hist (X, 30)
```

wird ein Histogramm mit 30 verschiedenen Klassen erzeugen.

Wird `hist` mit Ausgabeargumenten aufgerufen, so kann das Histogramm auch zur weiteren Verwendung gespeichert werden:

```
[n, xout] = hist (X, 30)
bar (xout, n)
```

ist äquivalent zu

```
hist(X, 30)
```

A.2.11 `hold`

Mit dem Befehl `hold on` wird Matlab instruiert, weitere Graphiken zum aktuellen Graphikfenster hinzuzufügen; alte Graphikelemente werden nicht gelöscht. `hold off` stellt das Standard-Verhalten wieder her, d.h., Matlab entfernt alte Graphiken, bevor neue hinzugefügt werden.

A.2.12 `isnan`

Die Funktion `isnan` wird auf beliebige Matrizen angewandt. Zurückgegeben wird eine Matrix gleicher Größe, deren Einträge da 1 sind, wo die ursprüngliche Matrix „not a number“-Einträge enthält. Typische Anwendung:

```
Werte=Werte(~isnan(Werte));
```

Diese Zuweisung entfernt aus dem Vektor `Werte` sämtliche „not a number“-Einträge.

A.2.13 `length`

Dieser Befehl gibt die Länge eines Vektors zurück. Wird `length` auf eine Matrix angewandt, so wird die größte Dimension der Matrix zurückgegeben.

A.2.14 `load`

Dieser Befehl wird benutzt, um eine Datei in eine Variable einzulesen.

```
B = load ('blatt.dat')
```

A.2.15 `log`

Die Logarithmus-Funktion. Diese Funktion wird analog zu `exp` (A.2.8) benutzt.

A.2.16 `max`, `min`

Diese Funktionen geben das größte bzw. kleinste Element eines übergebenen Vektors zurück. Wird eine Matrix übergeben, so wird ein Zeilenvektor erzeugt, der zu jeder Spalte das größte bzw. kleinste Element enthält.

A.2.17 `mean`

Wird `mean` auf einen Vektor angewandt, so wird der Mittelwert aller Einträge zurückgegeben. Wird `mean` auf eine Matrix angewandt, so berechnet Matlab den Mittelwert einer jeden Spalte und gibt einen Zeilenvektor zurück.

A.2.18 `median`

Diese Funktion verhält sich analog zu `mean`; allerdings wird jeweils der Median der übergebenen Daten berechnet.

A.2.19 `normcdf`

Die Funktion `normcdf (`x`, `μ`, `σ`)` berechnet die Verteilungsfunktion der Normalverteilung ($F_{\mu,\sigma}(x)$) zum Erwartungswert μ und der Standardabweichung σ an der Stelle x. Ist x ein Vektor, so wird die Verteilungsfunktion für jede Komponente von x ausgerechnet; der Rückgabewert ist dann ein Vektor.

A.2.20 `normrnd`

`normrnd(`μ`, `σ`, `m`, `n`)` erzeugt eine $m \times n$-Matrix von μ, σ-normalverteilten Zufallszahlen.

Werden nur zwei Parameter übergeben, so erzeugt `normrnd (`μ`, `σ`)` eine einzelne μ, σ-normalverteilte Zufallszahl.

A.2.21 `num2str`

Diese Funktion wandelt eine Zahl in eine Zeichenkette um. Nützlich im Zusammenhang mit `text`.

A.2.22 ones

`ones (`m`, `n`)` erzeugt eine $m \times n$-Matrix, der Einträge sämtlich den Wert 1 haben.

A.2.23 pie

`pie` erzeugt ein Kuchendiagramm; jedes Element des übergebenen Vektors wird durch ein Kuchenstück proportionaler Fläche repräsentiert. Ist die Elementsumme des übergebenen Vektors kleiner 1, so wird ein teilweises Kuchendiagramm gezeichnet; ist die Elementsumme größer oder gleich 1, so werden die Elemente normalisiert.

Ein optional übergebener zweiter Vektor gibt an, ob einzelne Kuchenstücke herausgerückt werden:

```
pie(X, [0 0 0 1 0]);
```

wird etwa ein Kuchendiagramm zeichnen, bei dem das zu `X(4)` korrespondierende Kuchenstück abgerückt ist.

A.2.24 polar

`polar` stellt in Polarkoordinaten übergebene Daten dar. Übergeben wird ein Vektor von Winkeln und ein Vektor von zugehörigen Radien. Zusätzlcih kann eine Zeichenkette angegeben werden, die Details der graphischen Darstellung kontrolliert; siehe hierzu `plot` (A.2.25).

A.2.25 plot

`plot(`X`, `Y`)` zeichnet die Komponenten des Vektors Y gegen diejenigen des Vektors X. Wird X ausgelassen, so werden die Komponenten des Vektors Y gegen ihre Indizes aufgetragen.

Graphikdetails können durch die Angabe einer Zeichenkette als dritter Parameters festgelegt werden: Durch Angabe von `- : -. --` kann der Linienstil festgelegt werden; die Zeichen `. o x + * s d v ^ < > p h` legen fest, wie Punkte markiert werden. Die Buchstaben `bgrcmyk` schließlich geben die zu verwendende Farbe an.

`plot` kann auch als einfacher Funktionsplotter genutzt werden, wenn man davon Gebrauch macht, dass die meisten Funktionen und Operatoren komponentenweise auf Vektoren operieren.

Als Beispiel werde die Exponentialfunktion auf dem Intervall $[0, 1]$ dargestellt:

```
X=[0:0.05:1];
plot (X, exp(X), '-xk');
```

`X` ist in diesem Beispiel der Vektor, der die gewünschten x-Koordinaten enthält. `exp(X)` ist der Vektor der zugehörigen Funktionswerte. `-xk` gibt an, dass die Zeichnung mit schwarzen Linien erfolgt, wobei die Stützstellen durch Kreuze markiert werden.

A.2.26 poisspdf

`poisspdf` (x, λ) gibt den Wert der Poissonverteilung zum Parameter λ an der Stelle x zurück. x kann wie schon gewohnt ein Vektor sein; in diesem Fall wird die Funktion komponentenweise angewandt.

A.2.27 rand

Ohne Parameter benutzt, gibt `rand` eine gleichverteilte Zufallszahl zwischen 0 und 1 zurück. In der Form `rand(`N`)` erzeugt die Funktion eine $N \times N$-Matrix von gleichverteilten Zufallszahlen.

A.2.28 rose

`rose` zeichnet ein Histogramm zu Winkeldaten. In der einfachsten Version (`rose(theta)`) wird eine Vektor von Winkeln angegeben, zu dem dann ein Histogramm mit 20 Klassen gezeichnet wird. Alternativ können Klassenmitten (`rose(theta,x)`) oder auch die Anzahl von Klassen (`rose(theta,20)` übergeben werden.

A.2.29 save

`save` wird zum Abspeichern von Variablen in Dateien benutzt.

```
save 'Blatt.dat' X -ASCII
```

Der Parameter `-ASCII` stellt sicher, dass die erzeugte Datendatei eine einfach lesbare Textdarstellung der gespeicherten Daten enthält.

A.2.30 `size`

`size` gibt die Dimensionen einer Matrix als Vektor zurück, also etwa `[2, 3]` für eine 2×3-Matrix.

A.2.31 `sort`

`sort (`X`)` gibt eine aufsteigend sortierte Version des Vektors X zurück.

A.2.32 `sortrows`

Diese Funktion sortiert die Zeilen der übergebenen Matrix in aufsteigender Reihenfolge (`sortrows(A)`). Wird zusätzlich eine Spaltennummer angegeben, so wird entsprechend den Werten in der angegebenen Spalte sortiert (`sortrows(A, 3)`).

A.2.33 `sqrt`

Die elementweise Wurzelfunktion. Diese Funktion wird ähnlich `exp` (A.2.8) verwendet.

A.2.34 `std`

`std (`X`)` gibt die empirische Standardaweichung der Elemente des Vektors X zurück. Dabei wird der erwartungstreue Varianzschätzer

$$V = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

zugrundegelegt; $\bar{X}$ sei dabei der Mittelwert der Elemente von X. `std` gibt $\sqrt{V}$ zurück.

A.2.35 `subplot`

`subplot (`n`, `m`, `i`)` instruiert Matlab, ein Graphikfenster in ein $n \times m$-Raster von Koordinatensystemen aufzuteilen; das i-te Koordinatensytem wird für den nächsten `plot`-Befehl benutzt.

A.2.36 `sum`

`sum` gibt die Elementsumme eines Vektors zurück. Wird eine Matrix übergeben, so wird ein Zeilenvektor zurückgegeben, der die Spaltensummen dieser Matrix enthält.

A.2.37 `text`

`text (`x`, `y`, '`*Text*`')` fügt den angegebenen Text an der Position (x, y) in die aktuelle Graphik ein. Um Zahlen einzufügen, benutzt man `num2str`: Die Zahlen werden zunächst in Zeichenketten konvertiert und dann als Text an der entsprechenden Stelle in die Graphik eingefügt.

A.2.38 `title`

`title ('`*Text*`')` fügt den angegebenen Text als Überschrift zum aktuellen Koordinatensystem hinzu.

A.2.39 `var`

`var (`X`)` gibt eine erwartungstreue Varianzsschätzung der Elemenete des Vektors X zurück.

A.2.40 `zeros`

Siehe `ones` (A.2.22).

Anhang B

Tabellen

B.1 Tabelle der kumulativen Normalverteilung

$$\Phi(z) = P[Z \leq z], Z \sim N(0,1)$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499
3.3	0.999517	0.999534	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999651
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758

B.2 Quantile der χ^2-Verteilung

f	$\chi^2_{f,0.995}$	$\chi^2_{f,0.990}$	$\chi^2_{f,0.975}$	$\chi^2_{f,0.950}$	$\chi^2_{f,0.025}$	$\chi^2_{f,0.101}$	$\chi^2_{f,0.005}$
1	7.879439	6.634897	5.023886	3.841459	0.000982	0.016110	0.000039
2	10.596635	9.210340	7.377759	5.991465	0.050636	0.212944	0.010025
3	12.838156	11.344867	9.348404	7.814728	0.215795	0.588763	0.071722
4	14.860259	13.276704	11.143287	9.487729	0.484419	1.070015	0.206989
5	16.749602	15.086272	12.832502	11.070498	0.831212	1.618526	0.411742
6	18.547584	16.811894	14.449375	12.591587	1.237344	2.214025	0.675727
7	20.277740	18.475307	16.012764	14.067140	1.689869	2.844556	0.989256
8	21.954955	20.090235	17.534546	15.507313	2.179731	3.502443	1.344413
9	23.589351	21.665994	19.022768	16.918978	2.700389	4.182432	1.734933
10	25.188180	23.209251	20.483177	18.307038	3.246973	4.880754	2.155856
11	26.756849	24.724970	21.920049	19.675138	3.815748	5.594593	2.603222
12	28.299519	26.216967	23.336664	21.026070	4.403789	6.321787	3.073824
13	29.819471	27.688250	24.735605	22.362032	5.008751	7.060631	3.565035
14	31.319350	29.141238	26.118948	23.684791	5.628726	7.809753	4.074675
15	32.801321	30.577914	27.488393	24.995790	6.262138	8.568030	4.600916
16	34.267187	31.999927	28.845351	26.296228	6.907664	9.334531	5.142205
17	35.718466	33.408664	30.191009	27.587112	7.564186	10.108470	5.697217
18	37.156451	34.805306	31.526378	28.869299	8.230746	10.889181	6.264805
19	38.582257	36.190869	32.852327	30.143527	8.906516	11.676090	6.843971
20	39.996846	37.566235	34.169607	31.410433	9.590777	12.468699	7.433844
21	41.401065	38.932173	35.478876	32.670573	10.282898	13.266576	8.033653
22	42.795655	40.289360	36.780712	33.924438	10.982321	14.069338	8.642716
23	44.181275	41.638398	38.075627	35.172462	11.688552	14.876649	9.260425
24	45.558512	42.979820	39.364077	36.415029	12.401150	15.688206	9.886234
25	46.927890	44.314105	40.646469	37.652484	13.119720	16.503743	10.519652
26	48.289882	45.641683	41.923170	38.885139	13.843905	17.323016	11.160237
27	49.644915	46.962942	43.194511	40.113272	14.573383	18.145808	11.807587
28	50.993376	48.278236	44.460792	41.337138	15.307861	18.971921	12.461336
29	52.335618	49.587884	45.722286	42.556968	16.047072	19.801176	13.121149
30	53.671962	50.892181	46.979242	43.772972	16.790772	20.633407	13.786720
40	66.765962	63.690740	59.341707	55.758479	24.433039	29.091509	20.706535
50	79.489978	76.153891	71.420195	67.504807	32.357364	37.735637	27.990749
60	91.951698	88.379419	83.297675	79.081944	40.481748	46.511303	35.534491
70	104.214899	100.425184	95.023184	90.531225	48.757565	55.386343	43.275180
80	116.321057	112.328793	106.628568	101.879474	57.153173	64.339891	51.171932
90	128.298944	124.116319	118.135893	113.145270	65.646618	73.357497	59.196304
100	140.169489	135.806723	129.561197	124.342113	74.221927	82.428666	67.327563

B.3 Quantile der Student-t-Verteilung

m	$t_{m,0.9000}$	$t_{m,0.9500}$	$t_{m,0.9750}$	$t_{m,0.9900}$	$t_{m,0.9950}$	$t_{m,0.9995}$
1	3.077684	6.313752	12.706205	31.820516	63.656741	636.619249
2	1.885618	2.919986	4.302653	6.964557	9.924843	31.599055
3	1.637744	2.353363	3.182446	4.540703	5.840909	12.923979
4	1.533206	2.131847	2.776445	3.746947	4.604095	8.610302
5	1.475884	2.015048	2.570582	3.364930	4.032143	6.868827
6	1.439756	1.943180	2.446912	3.142668	3.707428	5.958816
7	1.414924	1.894579	2.364624	2.997952	3.499483	5.407883
8	1.396815	1.859548	2.306004	2.896459	3.355387	5.041305
9	1.383029	1.833113	2.262157	2.821438	3.249836	4.780913
10	1.372184	1.812461	2.228139	2.763769	3.169273	4.586894
11	1.363430	1.795885	2.200985	2.718079	3.105807	4.436979
12	1.356217	1.782288	2.178813	2.680998	3.054540	4.317791
13	1.350171	1.770933	2.160369	2.650309	3.012276	4.220832
14	1.345030	1.761310	2.144787	2.624494	2.976843	4.140454
15	1.340606	1.753050	2.131450	2.602480	2.946713	4.072765
16	1.336757	1.745884	2.119905	2.583487	2.920782	4.014996
17	1.333379	1.739607	2.109816	2.566934	2.898231	3.965126
18	1.330391	1.734064	2.100922	2.552380	2.878440	3.921646
19	1.327728	1.729133	2.093024	2.539483	2.860935	3.883406
20	1.325341	1.724718	2.085963	2.527977	2.845340	3.849516
21	1.323188	1.720743	2.079614	2.517648	2.831360	3.819277
22	1.321237	1.717144	2.073873	2.508325	2.818756	3.792131
23	1.319460	1.713872	2.068658	2.499867	2.807336	3.767627
24	1.317836	1.710882	2.063899	2.492159	2.796940	3.745399
25	1.316345	1.708141	2.059539	2.485107	2.787436	3.725144
26	1.314972	1.705618	2.055529	2.478630	2.778715	3.706612
27	1.313703	1.703288	2.051831	2.472660	2.770683	3.689592
28	1.312527	1.701131	2.048407	2.467140	2.763262	3.673906
29	1.311434	1.699127	2.045230	2.462021	2.756386	3.659405
30	1.310415	1.697261	2.042272	2.457262	2.749996	3.645959
40	1.303077	1.683851	2.021075	2.423257	2.704459	3.550966
50	1.298714	1.675905	2.008559	2.403272	2.677793	3.496013
60	1.295821	1.670649	2.000298	2.390119	2.660283	3.460200
70	1.293763	1.666914	1.994437	2.380807	2.647905	3.435015
80	1.292224	1.664125	1.990063	2.373868	2.638691	3.416337
90	1.291029	1.661961	1.986675	2.368497	2.631565	3.401935
100	1.290075	1.660234	1.983972	2.364217	2.625891	3.390491
110	1.289295	1.658824	1.981765	2.360726	2.621265	3.381179
120	1.288646	1.657651	1.979930	2.357825	2.617421	3.373454
130	1.288098	1.656659	1.978380	2.355375	2.614177	3.366942
140	1.287628	1.655811	1.977054	2.353278	2.611403	3.361378
150	1.287221	1.655076	1.975905	2.351465	2.609003	3.356569
200	1.285799	1.652508	1.971896	2.345137	2.600634	3.339835
2000	1.281975	1.645616	1.961151	2.328214	2.578290	3.295398