SKRIPTUM

zur Lehrveranstaltung

STATISTIK FÜR LEHRAMT

von

Ferdinand Österreicher

Institut für Mathematik
der
Universität Salzburg

Salzburg

September 2001 / Juli 2010:
geringfügig überarbeitet und ergänzt

Inhaltsverzeichnis

Kapitel 1

BESCHREIBENDE STATISTIK

Zum Begriff "Stochastik"

"Wenn jemand von den Fertigkeiten und Künsten die Rechenkunst, die Messkunst und die Kunst des Wägens wegnimmt, so bleibt, um es offen zu sagen, nur etwas übrig, was fast minderwertig ist [...]. Es bleibt nichts anderes übrig, als ein Errarten, ein Schließen durch Vergleichen und ein Schärfen der Sinneswahrnehmung durch Erfahrung und durch eine gewisse Übung, wobei man die - von vielen als Künste titulierten - Fähigkeiten des geschickten Vermutens (στωχαστική sc. τέχνη) benützt, die durch stete Handhabung und mühevolle Arbeit herangebildet werden."

Platon, Philebos

Jabob Bernoullis 1713 posthum veröffentlichtes Buch hat übrigens den Titel *Ars conjectandi* (die Mutmaßungskunst). Dies ist die lateinische Entsprechung der Wortes Stochastik.

Zur Geschichte des Begriffes "Statistik"

Der Begriff "Statistik" stand in der Renaissance für eine an der Staatsräson orientierte Form der Politik, also kurz gesagt, für "Machiavellismus". Erst anfangs des 18. Jahrhunderts wurde der Begriff wertfrei und (1749) von *Gottfried Achenwall*, einem Vertreter der einflussreichen Göttinger Schule, im Sinn der beschreibenden Staatskunde verwendet. Letztere war ursprüglich verbal gehalten und beinhaltete auch feuilletonistische Schilderungen. Die

heute gebräuchliche "Tabellenstatistik" im politischen, sozialen und kulturellen Bereich ist das Ergebnis der Anwendung der astronomischen Methoden des Registrierens, Tabulierens und Berechnens.

sinngemäß aus *H. Rassem* und *J. Stagel (*1994*)*[1]

Im 18. Jahrhundert versteht man also unter "Statistik" die Zustandsbeschreibung des Staates. (Das neulateinische Wort "status" bedeutet übrigens Zustand, Staat.)

Nach einer stürmischen Mathematisierung, die insbesondere durch die englische statistische Schule des frühen 20. Jahrhunderts ausgelöst wurde, tritt erneut ein Bedeutungswandel bzw. eine Bedeutungserweiterung ein.

Statistik ist die Gesamtheit der Methoden zur Untersuchung von Massenerscheinungen.

Die Statistik hat ihre Grundlage im Gesetz der großen Zahlen (wenn die Zahl der untersuchten Erscheinungen genügend groß ist, werden die zufälligen Abweichungen aufgehoben und die typischen Zahlenverhältnisse kommen zum Vorschein) und in der Wahrscheinlichkeitsrechnung.

Bertelsmann Lexikon

"Statistik ist die Kunst und Wissenschaft, nützliche Information aus empirischen Daten zu ziehen."

F. Hampel, ETH Zürich

"Statistik ist eigentlich parasitär: sie lebt von den Arbeiten anderer. Dies ist keine Beleidigung, denn es ist inzwischen anerkannt, dass viele Wirte ohne die Parasiten, die sie unterhalten, sterben würden. Einige Tiere könnten ihre Nahrung nicht verdauen. So geht es auch mit vielen Bereichen des menschlichen Schaffens, sie würden zwar nicht sterben, wären aber auf jeden Fall wesentlich schwächer ohne Statistik."

L.J. Savage, US-amerikanischer Statistiker

[1] aus *H. Rassem* und *J. Stagel* (Hsg.): Geschichte der Staatsbeschreibung: Ausgewählte Quelltexte 1456 – 1813. Akademieverlag, Berlin 1994

1.1 ERHEBUNG VON DATEN

1.1.1 Datenbeschaffung (sekundärstatistische Datenerhebung)

Ehe man an eine eigene zeitaufwändige bzw. kostspielige Erhebung von Daten denkt, sollte man prüfen, ob nicht einige oder alle benötigten Daten bereits für andere Untersuchungen erhoben wurden und kostengünstig erhältlich sind. Als Quellen hiefür kommen in Frage: Jahrbücher und Berichte der amtlichen Statistik, Erhebungen anderer Institutionen wie von Markt- und Meinungsforschungsinstituten, meteorologische Stationen und anderen wissenschaftlichen Einrichtungen. Zu bedenken ist allerdings, dass Publikationen Daten zumeist bereits in "geraffter" Form enthalten.

1.1.2 Eigentliche Datenerhebung (primärstatistische Datenerhebung)

Das Medium eigener Erhebungen ist naturgemäß von der Fragestellung abhängig. **Erhebungsinstrumente** sind: Beobachtung, Messung, wissenschaftliches Experiment, schriftliche Befragung, Interview, usw.

Bei der **Beobachtung** wird der Untersuchungsgegenstand und dessen "Umgebung" prinzipiell nicht (de facto kaum) beeinflusst. Typisch dafür sind die Bevölkerungsstatistik (Demoskopie) und astronomische Beobachtungen.

Ein **Experiment** ist eine Beobachtung, die unter streng kontrollierten und prinzipiell wiederholbaren Bedingungen abläuft. Es dient zur Überprüfung von Vermutungen oder Theorien (sogenannter Hypothesen), die aufgrund von Beobachtungen aufgestellt wurden. Stimmen die Ergebnisse eines Experiments mit den von der Theorie vorhergesagten Werten gut überein, so "stützen" oder "erhärten" sie die Hypothese. Weichen die Ergebnisse jedoch beträchtlich von den vorhergesagten Werten ab, so legen sie nahe, "die Hypthese zu verwerfen". In diesem Sinn ist ein Experiment eher ein Falsifikations- als ein Verifikationsinstrument.

1.2 DATENTYPEN

1.2.1 Qualitative Daten

Nominale Daten: Die Werte einer Variablen drücken - gleichgültig, ob sie verbal beschrieben oder durch Zahlen kodiert werden - lediglich eine Verschiedenartigkeit aus.

Beispiele nominaler Variablen sind: Geschlecht, Blutgruppe, Familienstand, Kombinationsfach des Lehramtsstudiums, Oberflächenbeschaffenheit (Farbe?) von Materialien.

Ordinale (ordinal skalierte) Daten: Die Werte einer Variablen lassen sich, neben der Verschiedenartigkeit, auch in einer gewissen Weise ordnen.

Beispiele ordinaler Variablen sind: Schulnoten, Reihung in einem Wettkampfklassement (etwa Tabellen der Fußballliga), Güteklassen von Lebensmitteln, Tonhöhe in der üblichen Notenskala, Härtegrad von Mineralien.

Ein stets aktuelles Beispiel für ordinale Variable ist mit der schulischen Beurteilung und deren numerische Kodierung (Schulnoten) verknüpft. Mit der Notenskala ist stets eine Ordnungsrelation verbunden (hier durch $\succ$ für "besser als" gekennzeichnet. Die Skala selbst und deren numerische Kodierung ist in den verschiedenen Staaten sehr unterschiedlich. Dies ist ein deutlicher Hinweis darauf, dass das übliche Bilden des Notendurchschnitts eigentlich eine unzulässige Operation ist.

Die österreichische Notenskala ist bekanntlich

$$\begin{array}{lcllcllcllcl} \text{sehr gut} & \succ & \text{gut} & \succ & \text{befriedigend} & \succ & \text{genügend} & \succ & \text{nicht genügend}\,. \\ 1 & & 2 & & 3 & & 4 & & 5 \end{array}$$

Beim "Cambridge First Certificate", einem international anerkannten Zeugnis über die Kompetenz im Gebrauch der (nicht fachspezifischen) englischen Sprache, sind die "grades" wie folgt

$$\begin{array}{lcllcllcllcl} A & \succ & B & \succ & C & \succ & D & \succ & F\,, \\ & & \multicolumn{3}{c}{\text{bestanden}} & & \multicolumn{3}{c}{\text{nicht bestanden}} \end{array}$$

wobei F für "failure" steht. Auch in den U.S.A. wird die Notenskala durch Buchstaben beschrieben, was ein numerisches Rechnen zunächst verhindern würde. Die nachstehende numerische Kodierung

$$\begin{array}{lcllcllcllcl} A & \succ & B & \succ & C & \succ & D & \succ & F \\ 4 & & 3 & & 2 & & 1 & & 0 \end{array}$$

ermöglicht jedoch die Ermittlung eines sogenannten "gradepoint average", welcher über die Zulassung zu bestimmten Colleges mitentscheidet. In Italien war übrigens früher folgende Notenskala üblich

$$10 \succ 9 \succ 8 \succ 7 \succ \underbrace{6}_{\text{bestanden}} \succ \underbrace{5 \succ 4 \succ 3 \succ 2 \succ 1}_{\text{nicht bestanden}} .$$

1.2.2 Quantitative Daten

Quantitative Daten sind solche, die durch einen oder mehrere numerische Werte bestimmt sind. Demgemäß unterscheidet man ein- und mehrdimensionale quantitative Daten. In der Regel werden wir lediglich eindimensionale Daten betrachten, also Daten aus der Menge $\mathbb{R}$ der reellen Zahlen. Solche Daten nennt man in der beschreibenden Statistik üblicherweise *metrische Daten.*

Metrische (metrisch skalierte) Daten zeichnen sich dadurch aus, dass sich die Werte der Variablen nicht nur ordnen lassen, sondern dass sich auch Abstände zwischen den Werten angeben lassen. In der Regel sind Ordnung und Abstand durch die Ordnungsrelation ">" und die gewöhnliche Metrik auf $\mathbb{R}$ definiert.

Die metrischen Daten unterteilt man in *Intervalldaten* und *Verhältnisdaten.*

Verhältnisdaten (bezüglich einer Verhältnisskala skalierte Daten) sind metrische Daten, die auf einen "wahren" Nullpunkt bezogen sind, und zwar in dem in der nachstehenden Anmerkung präzisierten Sinn. Zumeist ist der Wertebereich einer bezüglich einer Verhältnisskala skalierten Variablen eine geeignete Teilmenge von $\mathbb{N}_0$ oder $[0, \infty)$.

Beispiele von metrischen Variablen sind: Anzahlen (Mengenangaben), Länge, Flächeninhalt, Korngröße von Steinen, Gewicht (Masse), Länge von Zeitintervallen, Tonhöhe in *Hertz*, Temperatur in Grad *Kelvin*, Höhe des Einkommens, Wahrscheinlichkeit eines Ereignisses. Der bei einem Glücksspiel zu erzielende Gewinn [2] ist ein Beispiel, bei dem der Wertebereich eine geeignete Teilmenge von $\mathbb{Z}$ ist.

Intervalldaten (bezüglich einer Intervallskala skalierte Daten) sind metrische Daten, die nicht auf einen wahren Nullpunkt bezogen sind.

[2] Dieser ist ein vom Ausfall des Zufallsexperiments abhängiges Vielfaches des Einsatzes. Dabei bedeutet - wie üblich - ein negativer Gewinn einen Verlust.

Beispiele von Intervalldaten sind folgende drei physikalische Größen, bei welchen der Nullpunkt in gewissem Sinn willkürlich festgelegt ist: potentielle Energie, Korngröße von Steinen, Zeitangabe in der Astronomie (modifiziertes julianisches Datum[3]) und Temperatur (in Grad *Celsius*, *Fahrenheit* oder *Reaumur*), und die beiden physiologischen Größen: Schallpegel (in Dezibel) und scheinbare Helligkeit von Sternen (Magnitudo). Man vergleiche dazu das Thema Logarithmische Skalen in Abschnitt 1.5.1.

Anmerkung: Die vier Skalentypen: nominale und ordinale Skalen, sowie Intervallskalen und Verhältnisskalen lassen sich durch die Familie von Transformationen (Abbildungen, Funktionen) charakterisieren, gegenüber welcher sie invariant sind. Nominale Skalen sind gegen umkehrbar eindeutige Abbildungen invariant. Ordinale Skalen sind gegenüber streng monotonen Funktionen invariant. Intervallskalen sind gegenüber linearen Funktionen mit positivem Anstieg und Verhältnisskalen gegenüber homogenen linearen Funktionen mit positivem Anstieg invariant.

Zirkuläre Daten

Zirkuläre Daten entsprechen Punkten auf dem Einheitskreis. Diese lassen sich bekanntlich durch die Zahlenpaare

$$(\cos\varphi, \sin\varphi)\ ,\ \varphi \in [0, 2\pi)$$

oder als Punkte

$$e^{i\varphi} = \cos\varphi + i\sin\varphi\ ,\ \varphi \in [0, 2\pi)$$

der Gaußschen Ebene beschreiben. In der Praxis wird der Winkel allerdings nicht durch das Bogenmaß sondern in Grad angegeben und bisweilen von Süden im Uhrzeigersinn gemessen ("Azimuth"). Die Addition von Winkeln ist also die Restklassenaddition bezüglich des Moduls 2π (im Bogenmaß) bzw. 360 (im Winkelmaß). So ist beispielsweise

$$(359 + 2) \bmod 360 = 1\,.$$

Zirkuläre Daten sind beispielsweise Daten der Orientierung, wie Windrichtung oder die Orientierung von magnetisierten Mineralen (Spuren von Paleomagnetismus) im Fels. Als zirkuläre Daten sind auch solche Daten zu betrachten, die mit Tages- und Jahreszyklus in Verbindung stehen, wie z.B.

[3] die Anzahl der Tage, die seit dem 17.11.1858, 0 Uhr, vergangen sind. Die Tageszeit wird durch die Nachkommastellen angegeben.

die Niederschlagsmenge bezogen auf die Stunden im Tag oder die Häufigkeit von Geburten bezogen auf die Tage im Jahr. Für diese Anwendungen wären als Module 24 bzw. 365 besser geeignet.

1.2.3 Unterscheidung hinsichtlich der Kardinalität des Wertebereichs

Schließlich unterscheidet man *diskrete* und *stetige* Daten.

Diskrete Daten: Eine diskrete Variable nimmt nur endlich oder abzählbar unendlich viele Werte an.

Stetige Daten: Eine stetige Variable nimmt überabzählbar viele Werte an. Dies ist einen bequeme mathematische Fiktion. Sie trägt dem Umstand Rechnung, dass die Messgenauigkeit der betrachteten Variablen prinzipiell nicht als beschränkt angesehen wird. Da jedoch in der Realität die Messgenauigkeit jedes noch so genauen Messgeräts beschränkt ist, kann davon ausgegangen werden, dass der Wertebereich jeder Variablen eine Gitterstruktur besitzt.

1.3 DARSTELLUNGSFORMEN EINDIMENSIONALER DATEN

1.3.1 Einführende Beispiele

Beispiel 1: Chuck-a-luck

In den Lehrbüchern der Wahrscheinlichkeitsrechnung aus dem angelsächsischen Raum findet sich häufig das folgende Spielchen, "Chuck-a-luck" genannt, das in geringfügig modifizierter Form auch im Casinospiel "Sic-Bo" integriert ist.

Man setzt einen Einsatz e auf eine Zahl $z \in \{1, ..., 6\}$. Dann werden drei Würfel geworfen. Kommt z unter den drei Augenzahlen nicht vor, ist der Einsatz verloren. Kommt z genau ein-, zwei- oder dreimal vor, gewinnt man den ein-, zwei- bzw. dreifachen Einsatz.

Man spiele 36 Spiele und trage den Gewinn jedes Spieles gegen dessen Nummer auf.

Der Wertebereich des Gewinns X ist, bezogen auf den Einsatz $e = 1$, gleich $W_X = \{-1, 1, 2, 3\}$.

Beispiel 2: Augensumme beim Wurf mit drei Würfeln

Man werfe $6^3/2 = 108$ mal drei symmetrische Würfel, stelle die jeweilige Augensumme X fest und fertige eine Strichliste an. Danach erstelle man je ein Histogramm für die Häufigkeit der Ausfälle von X

a) für die einzelnen Werten $j \in \{3, ..., 18\}$ des Wertebereichs W_X

b) für die Klassen $K_1 = \{3, 4\}$, $K_2 = \{5, 6\}$, $K_3 = \{7, 8\}$, $K_4 = \{9, 10\}$, $K_5 = \{11, 12\}$, $K_6 = \{13, 14\}$, $K_7 = \{15, 16\}$, $K_8 = \{17, 18\}$

c) für die Klassen $K_1 = \{3, 4, 5, 6\}$, $K_2 = \{7, 8, 9, 10\}$, $K_3 = \{11, 12, 13, 14\}$, $K_4 = \{15, 16, 17, 18\}$

Der Wertebereich der Augensumme X ist die Menge $W_X = \{3, ..., 18\}$.

Beispiel 3: Warten bis zum ersten Erfolg

Ein Versuch bestehe darin, eine Münze so oft zu werfen, bis zum ersten Mal das Ereignis "Kopf" eintritt. X bezeichne die Anzahl der nötigen Münzwürfe.

Man führe $2^6 = 64$ Versuche durch und erstelle ein Histogramm für die Häufigkeiten der beobachteten "Wartezeiten" X_i, $i \in \{1, ..., 64\}$, und zwar für die Klassen $K_1 = \{1\}, ..., K_6 = \{6\}$ und $K_7 = \{7, 8, ...\}$.

Der Wertebereich einer "Wartezeit" X ist die abzählbar unendliche Menge $W_X = \mathbb{N} \cup \{\infty\}$.

1.3.2 Tabellarische Darstellung

Wir gehen auch weiterhin davon aus, dass der Wertebereich W_X der zugrundeliegenden Variablen X eine endliche oder abzählbar unendliche Teilmenge der reellen Zahlen ist, dass also gilt $W_X = \{\omega_1, \omega_2, ...\} \subset \mathbb{R}$. Wenn möglich - und das ist vielfach der Fall - nehmen wir an, dass gilt $\omega_1 < \omega_2 < ...$.

Ausgehend von einer Urliste, das ist eine Folge $x_1, x_2, ..., x_n \in W_X$ von Beobachtungswerten der Variablen X erstellen wir eine Häufigkeitstabelle.

Dann sind

$$\begin{aligned} H_j &= |\{i \in \{1, ..., n\} : x_i = \omega_j\}| \\ &= \text{die Anzahl der Versuche mit dem Ausfall } \omega_j \end{aligned}$$

bzw.

$$h_j = \frac{H_j}{n} = \frac{\text{Anzahl der Versuche mit dem Ausfall } \omega_j}{\text{Gesamtanzahl der Versuche}}$$

die *absolute Häufigkeit* bzw. die *relative Häufigkeit* des Ausfalls ω_j , $j \geq 1$.

1.3.3 Graphische Darstellung

a) **Die Strichliste**

Ist der Wertebereich der Variablen im Vergleich mit dem gewählten Stichprobenumfang klein und ist man von vornherein nicht an der Reihenfolge interessiert, in welcher die Daten erhoben werden, so ist zu erwägen, eine Strichliste anzufertigen [4]. Dabei sind die einzelnen Beobachtungswerte unmittelbar bei deren Erhebung durch je einen waagrechten (oder senkrechten)

[4] Es gibt Anwendungen, bei welchen man sich speziell für die Reihenfolge von Messwerten interessiert. Typisch dafür sind Folgen von Tagen mit und ohne Niederschlag. Hinsichtlich der Untersuchung sogenannter *Läufe von binären Zufallsfolgen* sei auf die Diplomarbeit von Frau *Radauer* [36] verwiesen.

Strich über (bzw. neben) dem zugehörigen Wert der Variablen in einer Strichliste einzutragen. Zur Wahrung der Übersicht - und um das spätere Abzählen zu erleichtern - formt man in der hergebrachten Weise Blöcke zu je 5 Strichen. Auf diese Weise kann bereits während der Datenerhebung eine graphische Darstellung der Verteilung der Beobachtungswerte entstehen.[5]

b) **Das Histogramm**

Bei einem Histogramm für die relativen Häufigkeiten der Werte einer metrischen Variablen trägt man über jeder Klasse die relative Häufigkeit pro Einheit der Variablen auf, also

$$\text{Höhe} = \frac{\text{relative Häufigkeit}}{\text{Klassenbreite}}$$

Ein Histogramm beruht auf einer geeigneten Klasseneinteilung. Eine solche ist die Partition eines Intervalls I von $\mathbb{R}$, welches den Wertebereich der Variablen umfasst, in endlich viele Intervalle (Klassen).

Für Variable mit abzählbar unendlichem oder überabzählbarem Wertebereich ist eine Klassenzusammenfassung unumgänglich. (Beispiele: Wartezeit bis zum ersten Erfolg, Anzahl der radioaktiven Zerfälle in einem Zeitintervall vorgegebener Länge. Messergebnisse mit prinzipiell beliebig genauer Messgenauigkeit.)

Bezeichnung: Ist der Wertebereich W_X einer Variablen X eine Teilmenge von $\{\mu + c \cdot z : z \in \mathbb{Z}\}$ mit $c > 0$ und $\mu \in \mathbb{R}$, so sagt man, W_X habe eine *Gitterstruktur mit der Gitterkonstante*[6] c. Seien $h_{\mu+c\cdot z}$ die relativen Häufigkeiten der Ausfälle $\mu + c \cdot z$ eines Versuches. Dann heißt die Funktion

$$f(x) = \sum_{z\in\mathbb{Z}} \frac{h_{\mu+c\cdot z}}{c} \cdot 1_{[\mu+c(z-1/2),\mu+c(z+1/2))}(x)\,,\ x \in \mathbb{R}$$

das zugehörige *Histogramm*. (Man stellt das Histogramm zumeist so dar, dass man zu jedem Wert $\mu + c \cdot z$ ein Rechteck mit der Basis $[\mu + c(z - 1/2), \mu + c(z + 1/2))$ und dem Flächeninhalt $h_{\mu+c\cdot z}$ - und somit der Höhe $h_{\mu+c\cdot z}/c$ - zeichnet.)

[5] Manche Beobachter verzichten bewusst darauf, während des Beobachtungsverlaufs Strichlisten anzufertigen, um jegliche subjektive Erwartungshaltung auszuschließen.

[6] Da man stets davon ausgehen kann, dass die Messgenauigkeit jedes noch so guten Messinstruments beschränkt ist, ist zumeist auch die Annahme einer Gitterstruktur angebracht.

Beispiel 4: Radioaktiver Zerfall

Ernest Rutherford $(1871-1937)$ und *Hans W. Geiger* $(1882-1945)$ verwendeten bei ihrem klassischen Experiment im Jahre 1910 eine Polonium-Quelle und registrierten für 2608 disjunkte Zeitintervalle von je 7.5 Sekunden Dauer die Anzahl der Szintillationen. Im Folgenden sind die beobachteten absoluten Häufigkeiten H_j der Intervalle mit j, $j \in \{0, ..., 14\}$ Szintillationen dargestellt (vgl. beispielsweise [20], S. 36).

j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	≥ 15
H_j	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1	0

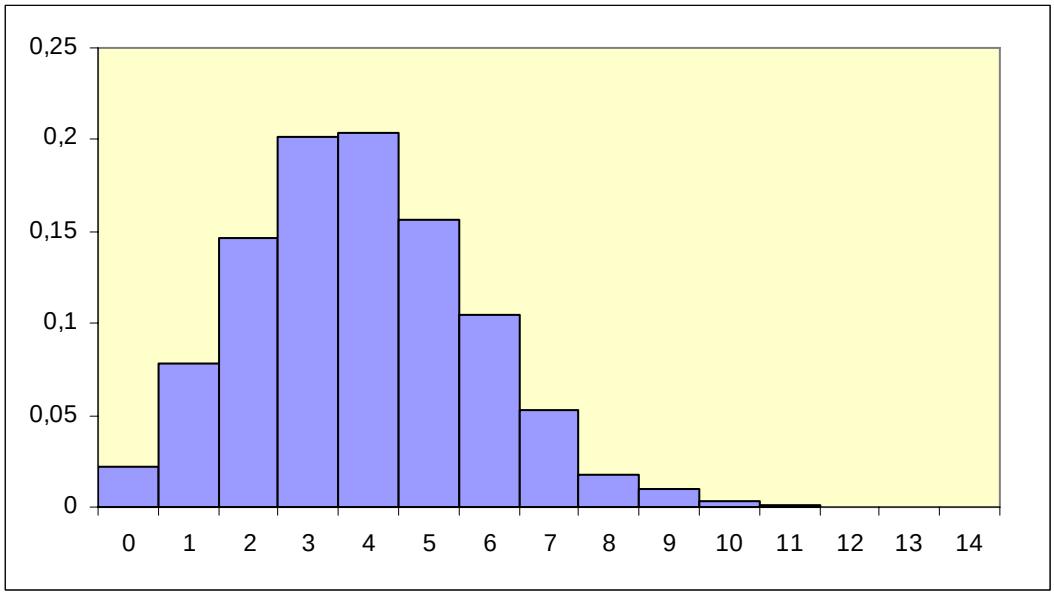


Abbildung: Zugehöriges Histogramm

Allgemeine Hinweise zur Klassenbildung

Generell lässt sich sagen, dass die Wahl der Klassen einige Erfahrung voraussetzt. Verwenden Sie nicht das erstbeste Histogramm. Experimentieren Sie, um eine möglichst charakteristische Form zu erzielen!

- Klassenbreite

  Ist der Wertebereich W_X diskret und besitzt er eine Gitterstruktur mit Gitterkonstante $c(W_X)$, dann ist diese eine untere Schranke für die Klassenbreite. Eine weitere Orientierungshilfe für die Klassenbreite

b_n ist die folgende, vom Stichprobenumfang n und der Standardabweichung s_n der Stichprobenwerte abhängige Größe $3.49 \cdot s_n \cdot n^{-1/3}$, welche sich an normalverteilten Grundgesamtheiten orientiert. [7]

Da unser Wahrnehmungsapparat eher gewohnt ist, Längen (und nicht Flächen) zu vergleichen, sollte man - insbesondere beim "Bulk" der Daten - möglichst gleiche Klassenbreiten wählen. Verschiedene Klassenbreiten, sollte man - wenn nötig - nur an den Rändern vorsehen.

- Anzahl der Klassen

  Diese sollte man nicht zu klein und nicht zu groß wählen. Zu viele Klassen erzeugen ein unerwünscht erratives Muster der Häufigkeiten, zu wenige ein unangebracht grobes Muster.

  Besitzt der Wertebereich W_X der Variablen X eine endliche Spannweite $s(W_X) = \max\{\omega \in W_X\} - \min\{\omega \in W_X\}$ und zudem eine Gitterstruktur mit Gitterkonstante $c(W_X)$, dann kann der Quotient $s(W_X)/c(W_X)$ als obere Schranke für die Anzahl der Klassen gelten.

  Als weitere Orientierungshilfe kann im Fall unimodaler (eingipfliger) Verteilungen der Funktionswert $1 + \lfloor ld(n) \rfloor$ des Stichprobenumfangs dienen. [8]

- Klassengrenzen

  sollte man möglichst so wählen, dass die Werte einer Variablen innerhalb der Klassen möglichst gleichmäßig verteilt sind. Eine Orientierung für die Wahl der Klassenmitten ist insbesondere für den Fall, dass die Gitterkonstante sehr klein ist, die Verteilung eher symmetrisch ist und keine Ausreißer [9] besitzt $\bar{x}_n \pm \frac{1}{2} b_n \cdot n$, $n \in \mathbb{N}_0$. Dabei bezeichnen $\bar{x}_n$ das Stichprobenmittel und b_n die Klassenbreite.

Verwendungszweck von Histogrammen

Histogramme werden verwendet,

[7] Die Stichprobenvarianz ist $s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$, wobei $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$ das Stichprobenmittel ist. Die merkwürdige Zahl 3.49 ist der numerische Wert von $2 \cdot 3^{1/3} \cdot \pi^{1/6}$.

[8] Für die Beispiele 2 und 3 ergibt sich für die Anzahl der Klassen mit Hilfe dieser Faustregel $1 + \lfloor ld(108) \rfloor = 1 + \lfloor ld(64) \rfloor = 7$.

[9] Ein Beobachtungswert wird dann als Ausreißer betrachtet, wenn sein Abstand von der überwiegenden Mehrheit der Daten unerwartet groß ist.

(i) um eine Datenmenge zwecks guter visueller Wahrnehmbarkeit von allgemeinen Charakteristika der Verteilung der Beobachtungswerte, wie repräsentativem Wert, Streuungsverhalten und charakteristischer Gestalt zu verdichten,

(ii) um Hinweise zur Wahl eines geeigneten Wahrscheinlichkeitsmodells und nötigenfalls einer geeigneten Variablentransformation zu erhalten - welcher später eine genauere statistische Analyse folgen kann - und

(iii) um ein unerwartetes Verhalten der zugrundeliegenden Variablen festzustellen und/oder ungewöhnliche Beobachtungswerte zu entdecken.

c) **Das Stängel-Blatt-Diagramm** [10]

Ist der Wertebereich einer Variablen vergleichsweise groß und ist man nicht an der Reihenfolge interessiert, in welcher die Daten erhoben werden, so kann die Anfertigung eines sogenannten *Stängel-Blatt-Diagramms* - wie die einer Strichliste - bereits während der Datenerhebung erfolgen. Im Wesentlichen entspricht dies nämlich dem Erstellen einer Strichliste mit zusätzlicher Klassenbildung. Im Folgenden sei diese, vom amerikanischen Statistiker *John W. Tukey* (1977), dem Begründer der sogenannten Explorativen Datenanalyse, vorgeschlagene visuelle Darstellung von Daten anhand eines Beispiels erläutert.

Anmerkungen zur Anfertigung eines Stängel-Blatt-Diagramms

für eine Liste von Rohdaten mit Werten beispielsweise aus $\{70, ..., 115\}$

Zunächst werden alle möglichen Zehnerzahlen (Zehnerziffer bzw. Paar, bestehend aus Hunderter- und Zehnerziffer), also $7, ..., 11$ im "Stängel" - d.h. vertikal - eingetragen,

anschließend wird die Liste der Rohdaten der Reihe nach durchgegangen und für jeden Wert der Liste wird hinter der zugehörigen Zehnerzahl im Stängel die Einerziffer als Blatt - d.h. horizontal - eingetragen.

Modifikationsmöglichkeit: Sollte sich die Klassenbildung nach Zehnerzahlen als zu grob erweisen, ist zu erwägen, die Zehnerzahlen im Stängel jeweils

[10] Das *Stängel-Blatt-Diagramm* (*stem-and-leaf-display*) ist - wie das *Kastenbild* (*boxplot*) in Abschnitt 1.4.3 eine Darstellungsform der *Explorativen Datenanalyse* (*Exploratory Data Analysis*). Dieser Zweig der modernen beschreibenden Statistik ist eine Schöpfung des US-amerikanischen Statistikers *John W. Tukey* ($1915 - 2000$), vom dem übrigens auch die Kurzbezeichnung *bit* für *b*inary dig*it* stammt.

zweimal anzuschreiben und hinter der jeweils ersten Zehnerzahl die Einerziffern von 0 bis 4 und hinter der jeweils zweiten Zehnerzahl die Einerziffern von 5 bis 9 als Blatt anzufügen.

d) **Die empirische Verteilungsfunktion**

Seien $x_1, ..., x_n$ reelle Beobachtungswerte, sei x eine weitere reelle Zahl und sei

$$|\{i \in \{1, ..., n\} : x_i \leq x\}|$$

die absolute Häufigkeit der Beobachtungen, deren Bobachtungswerte kleiner oder gleich x sind. Dann heißt die Abbildung

$$x \longmapsto F_n(x) = \frac{|\{i \in \{1, ..., n\} : x_i \leq x\}|}{n}$$

die *empirische Verteilungsfunktion* der Beobachtungswerte $x_1, ..., x_n$.

Anmerkung: Im folgenden wird die empirische Verteilungsfunktion mit Hilfe des Wertebereichs ihrer Funktionswerte definiert. Zu diesem Zweck ordnet man die Beobachtungswerte ihrer Größe nach: Es bezeichne

$x_{1:n}$	den kleinsten von n Beobachtungswerten,
$x_{2:n}$	den zweit-kleinsten Beobachtungswert,
...	...
$x_{n:n}$	den größten Beobachtungswert.

Dann gelten $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$, $\{x_{1:n}, ..., x_{n:n}\} = \{x_1, ..., x_n\}$ und

$$F_n(x) = \begin{cases} 0 & \text{für} \quad x < x_{1:n} \\ \frac{i}{n} & \text{für} \quad x_{i:n} \leq x < x_{i+1:n}, \quad i \in \{1, ..., n-1\} \\ 1 & \text{für} \quad x_{n:n} \leq x \end{cases}$$

Anmerkung zur "Kurve von Quetelet" [11]

Der nachstehende Bericht des Mathematikers *B.L.* v*an der Waerden* gibt einerseits einen interessanten Einblick in die Ideengeschichte der Statistik

[11] Der belgische Statistiker *Lambert-Adolphe-Jacques Quetelet* (1796 – 1874) erschloss der Normalverteilung in der Anthropometrie ein gänzlich neues und unvermutetes Anwendungsgebiet und übte damit einen prägenden Einfluss auf *Francis Galton* (1822 – 1911) aus. Auf seinen Einfluss geht auch die Gründung vieler statistischer Behörden in Europa zurück.

und andererseits eine Anregung, eine empirische Verteilungsfunktion plastisch darzustellen.

"Lebhaft erinnere ich mich noch, wie mein Vater mich als Knaben eines Tages an den Rand der Stadt führte, wo am Ufer die Weiden standen, und mich hundert Weidenblätter willkürlich pflücken ließ. Nach Aussonderung der beschädigten Spitzen blieben noch 89 unversehrte Blätter übrig, die wir dann zu Hause, nach abnehmender Größe geordnet, wie Soldaten in Reih und Glied stellten. Dann zog mein Vater durch die Spitzen eine gebogene Linie und sagte: "Dies ist die Kurve von QUETELET. Aus ihr siehst du, wie die Mittelmäßigen immer die große Mehrheit bilden und nur wenige nach oben hervorragen oder nach unten zurückbleiben." "

Beispiel 5: Zur "Kurve von Quetelet"

W. *Rohm* (HTL Saalfelden) und *F. Österreicher* haben 1982 gemäß *B.L. van der Waerden*s Beschreibung einer Föhre 100 Nadeln aufs Geratewohl entnommen und deren Längen (in *mm*) gemessen. Die geordneten Messergebnisse sind im nachstehenden Stängel-Blatt-Diagramm dargestellt.

7	044778
8	000134455677777888889999
9	000111112222333444445555566666777778899
10	00112222223333344456667777
11	0115

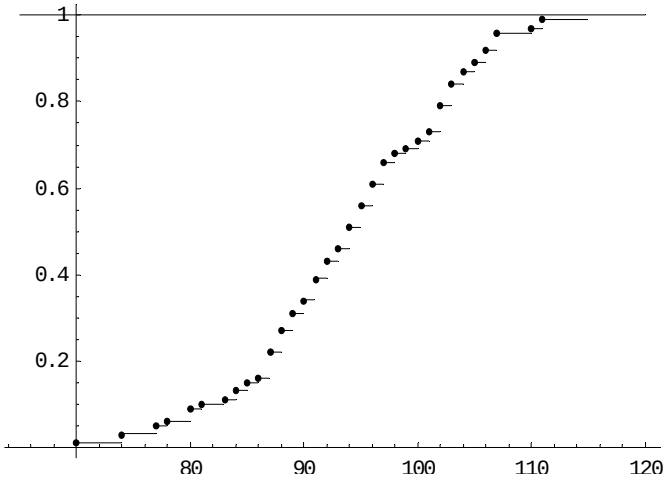


Abbildung: Zugehörige empirische Verteilungsfunktion

1.3.4 Weitere Beispiele für Beobachtungen bzw. Experimente

- Erzeugung gleichverteilter Zufallszahlen auf $\{0, ..., 999\}$
- In der Stadt Salzburg gemessene Tagesniederschlagsmengen (in Zehntel Millimeter), Quelle: Zentralanstalt für Meteorologie und Geodynamik in Salzburg (siehe [32])

1.4 KENNGRÖSSEN EINDIMENSIONALER DATEN: Zentral- und Streumaße

Wir gehen im folgenden - sofern nicht anders angenommen - von Daten $x_1, ..., x_n$ einer eindimensionalen Variablen (d.h. mit Werten aus $\mathbb{R}$), versehen mit der natürlichen Ordnung " $<$ " und dem üblichen Abstand, aus.

1.4.1 Stichprobenmittel und Standardabweichung

a) **Das Stichprobenmittel**

Das *arithmetische Mittel*

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

der Stichprobenwerte heißt *Stichprobenmittel* oder *Mittelwert*.

Anmerkung 1: Der britische Wissenschaftler *Thomas Simpson* (1710 – 1761) schlug für die Handhabung von mehrfachen, erfahrungsgemäß verschiedenen Messwerten einer Messgröße die Verwendung des Stichprobenmittels vor. Mit ein Motiv dafür war wohl, die Subjektivität der Beobachter beim Umgang mit solchen Daten hintanzuhalten und durch eine standardisierte Vorgangsweise den Austausch von Messergebnissen zu erleichtern. Er schreibt:

"Zusammenfassend scheint es, dass das Bestimmen des arithmetischen Mittels einer Anzahl von Messwerten die Chance kleiner Fehler beträchtlich verringert und nahezu jede Möglichkeit für große ausschließt. Diese Erwägung allein scheint ausreichend, um die Verwendung dieser Methode nicht nur Astronomen zu empfehlen, sondern allen, die Präzisionsmessungen durchführen. Je mehr Beobachtungen oder Experimente gemacht werden, desto weniger werden die Resultate fehleranfällig sein, vorausgesetzt, eine Wiederholung der Messungen ist unter gleichen Bedingungen möglich."

Anmerkung 2: Liegen bereits die relativen Häufigkeiten h_j der einzelnen Ausfälle ω_j , $j \geq 1$, der Variablen X vor, so ist es zweckmäßiger, das Stichprobenmittel gemäß

$$\bar{x}_n = \sum_{j \geq 1} \omega_j \cdot h_j$$

mit deren Hilfe auszudrücken. Falls möglich denken wir uns dabei die ω_j wieder der Größe nach geordnet: $\omega_1 < \omega_2 < \dots$.

Zu den beiden Darstellungen des Stichprobenmittels äußerte sich *Henri Lebesgue* [12], der Schöpfer des nach ihm benannten Lebesgue-Integrals, welches in der höheren Analysis das *Riemann*-Integral ersetzt, wie folgt.

"Man kann auch sagen, dass man sich bei der Verwendung der ersten Berechnungsart wie ein Kaufmann ohne System verhält, der Geldstücke und Banknoten in der zufälligen Reihenfolge zählt, wie er sie in die Hand bekommt. Während wir bei der zweiten Rechenart vorgehen wie ein umsichtiger Kaufmann, der sagt:

Ich habe	$H(E_1)$	Münzen zu einer Krone, macht	$1 \times H(E_1)$,
ich habe	$H(E_2)$	Münzen zu zwei Kronen, macht	$2 \times H(E_2)$,
ich habe	$H(E_5)$	Münzen zu fünf Kronen, macht	$5 \times H(E_5)$, usw.;

ich habe also insgesamt

$$S = 1 \times H(E_1) + 2 \times H(E_2) + 5 \times H(E_5) + \dots ,$$

weil er - wie reich er auch sein mag - nur eine endliche Anzahl von Banknoten zu zählen hat."

Die **Minimaleigenschaft des Stichprobenmittels** bezüglich der Summe der Abweichungsquadrate

Behauptung: Die Funktion

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

ist genau dann minimal, wenn $x = \bar{x}_n$ ist.

Beweisvariante 1: Die Hälfte der Ableitung von $f(x)$ ist

$$\frac{1}{2} f'(x) = -\frac{1}{n} \sum_{i=1}^{n} (x_i - x) = x - \bar{x}_n \begin{cases} < 0 & \text{für} \quad x < \bar{x}_n \\ = 0 & \text{für} \quad x = \bar{x}_n \\ > 0 & \text{für} \quad x > \bar{x}_n \, . \end{cases}$$

[12] *Henri Lebesgue* (1875-1941), französischer Mathematiker

Somit wird das Minimum im Punkt $x = \bar{x}_n$ angenommen.

Beweisvariante 2 bedient sich einer Aussage aus der Mechanik, welche trotz ihrer Einfachheit äußerst nützlich ist. Diese Aussage ist

Der Steinersche Verschiebungssatz[13]**:** Es gilt

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

Wegen $n(x - \bar{x}_n)^2 \geq 0$ ergibt sich daraus als **Folgerung** die gewünschte Aussage, nämlich

$$\sum_{i=1}^{n} (x_i - x)^2 \geq \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 ,$$

wobei Gleichheit offensichtlich genau dann gilt, wenn $x = \bar{x}_n$ ist.

Anmerkung 3: Interpretiert man die ω_j, $j \geq 1$, als Punkte der x-Achse, in denen die Massen $H(\omega_j)$ sitzen, und lässt man diese Massen um eine Achse rotieren, die durch den Punkt x der x-Achse geht und normal auf dieser steht, dann ist die Summe

$$\sum_{i=1}^{n} (x_i - x)^2 = \sum_{j \geq 1} (\omega_j - x)^2 H_j$$

das zugehörige Trägheitsmoment der Massenverteilung[14]. Die Folgerung des Steinerschen Verschiebungssatzes besagt also, dass das Trägheitsmoment genau dann kleinstmöglich ist, wenn die Drehachse durch den Massenmittelpunkt $\bar{x}_n$ der Massenverteilung geht.

Beweis des Satzes: Die Anwendung der unmittelbar einsichtigen (und auch geometrisch interpretierbaren) Beziehung

$$a^2 = b^2 + (a - b)^2 + 2(a - b)b$$

für $a = x_i - x$ und $b = x_i - \bar{x}_n$ ergibt

$$(x_i - x)^2 = (x_i - \bar{x}_n)^2 + (\bar{x}_n - x)^2 + 2(\bar{x}_n - x)\,(x_i - \bar{x}_n) .$$

[13] parallel–axis theorem, *Jakob Steiner* (1796-1863), Schweizer Geometer

[14] Der Begriff des Trägheitsmoments wurde - wie der des Erwartungswerts in der Wahrscheinlichkeitsrechnung - vom holländischen Wissenschaftler *Christiaan Huygens* (1629 – 1695) geprägt. Er ist überdies der Erfinder der Pendeluhr, der Entdecker der Saturnringe und der Urheber des *Huygensschen Prinzips* in der Optik.

Daraus ergibt sich die Aussage des Verschiebungssatzes durch Summation über $i \in \{1, ..., n\}$ und Berücksichtigen von $\sum_{i=1}^{n} (x_i - \bar{x}_n) = \sum_{i=1}^{n} x_i - n\,\bar{x}_n = 0$.

Anmerkung 4: Der Punkt $(\bar{x}_n, ..., \bar{x}_n) \in \mathbb{R}^n$ ist die Projektion von $(x_1, ..., x_n) \in \mathbb{R}^n$ auf die Diagonale

$$D = \{(x, ..., x) \in \mathbb{R}^n\,,\ x \in \mathbb{R}\}$$

des $\mathbb{R}^n$. Daher ist der Steinersche Verschiebungssatz ein Spezialfall des Pythagoräischen Lehrsatzes im $\mathbb{R}^n$.

b) **Stichprobenvarianz und Standardabweichung** [15]

Die Größe

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

heißt Stichprobenvarianz. Deren Quadratwurzel

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

heißt *Standardabweichung der Stichprobe*. Nicht die Stichprobenvarianz, sondern die Standardabweichung ist die eigentliche Maßzahl für die Abweichung der Daten vom Mittelwert. Dies kann man sich etwa mit Hilfe einer Dimensionsbetrachtung überlegen, wie sie in der Physik üblich ist. Sind beispielsweise die x_i die Ergebnisse der Messungen einer Länge, so sind auch die Differenzen $x_i - \bar{x}_n$ Längen. Deren Quadrate $(x_i - \bar{x}_n)^2$ sind demnach die Flächen der zugehörigen Quadrate, sodass auch s_n^2 ein Maß für eine Fläche ist. Erst deren Quadratwurzel, die Standardabweichung s_n, hat wieder die richtige "Dimension", nämlich die einer Länge.

Warum dividiert man durch $n-1$ und nicht durch n ?

Vordergründige Antwort: Wegen der Bedingung $\sum_{i=1}^{n} (x_i - \bar{x}_n) = 0$ ist eine der Abweichungen $x_i - \bar{x}_n$ durch die restlichen $n-1$ Abweichungen

[15] Die Bezeichnung "*standard deviation*" (Standabweichung) wurde vom englischen Statistiker *Karl Pearson* (1857−1936) im Jahre 1893 geprägt. Der Begriff selbst wurde jedoch bereits vom deutschen Mathematiker *Carl Friedrich Gauß* (1777 − 1855) im Rahmen der Fehlerrechung verwendet. Bei Gauß hieß diese Größe "*mittlerer zu befürchtender Fehler*".

festgelegt. Da also nur $n-1$ Summanden frei variieren können, dividiert man durch $n-1$. Einer physikalischen Tradition folgend nennt man diese Zahl auch die *Anzahl der Freiheitsgrade*.

Hintergründige Antwort: Dazu bedarf es eines stochastischen Modells. Sei X eine Zufallsvariable mit Erwartungswert μ und unbekannter Varianz σ^2. Wäre nun der Erwartungswert μ bekannt, so würde man σ^2 naturgemäß durch

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

schätzen. Ist jedoch der Erwartungswert unbekannt, so hat man diesen durch seinen Schätzwert $\bar{x}_n$ zu ersetzen. Gemäß der Minimaleigenschaft von $\bar{x}_n$ gilt

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

Damit schätzt man die Varianz "im Durchschnitt etwas zu kurz". Durch die Multiplikation mit dem Faktor $\frac{n}{n-1} > 1$ wird dieses Defizit behoben und man erhält den angegebenen Schätzwert

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

Wie wir im Folgenden zeigen werden, macht man damit den zugehörigen Schätzer S_n^2 *erwartungstreu* (unverfälscht, oder - englisch - *unbiased*) [16]. Dazu bedarf es freilich eines stochastischen Modells.

Behauptung: Seien $X_1,...,X_n$ unabhängige, identisch verteilte Zufallsvariable mit Erwartungswert $\mu = E(X_1)$ und Varianz $\sigma^2 = V(X_1)$ und ist $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ das Mittel dieser Zufallsvariablen. Dann ist die ebenfalls *Stichprobenvarianz* genannte Zufallsvariable

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

[16] Der englische Statistiker *William Searly Gosset* (1876 – 1937), der unter dem Pseudonym "*Student*" für die Brauerei Guinness arbeitete, verwendete aus diesem Grund anstelle des von *Karl Pearson* benützten Nenners n den Nenner $n-1$. Dies veranlasste Karl Pearson zu der Äußerung: "Only naughty brewers take n so small that the difference is not of the order of the probable error!"

ein erwartungstreuer Schätzer für σ^2 , d.h. dass - welchen Wert σ^2 auch immer besitzt - die Verteilung von S_n^2 den Erwartungswert σ^2 hat, oder - in Zeichen - dass gilt

$$E(S_n^2) = \sigma^2 .$$

Beweis: Die Anwendung des Erwartungswerts auf den Steinerschen Verschiebungssatz in der Form

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

ergibt wegen der Linearität des Erwartungswerts, wegen $E(\bar{X}_n) = \mu$, der Definition der Varianz und des Sachverhalts $V(\bar{X}_n) = \sigma^2/n$

$$\begin{aligned} E[\sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2] &= E[\sum_{i=1}^{n} (X_i - \mu)^2] - n \cdot E[\left(\bar{X}_n - \mu\right)^2] \\ &= \sum_{i=1}^{n} E[(X_i - \mu)^2] - n \cdot V(\bar{X}_n) \\ &= n \cdot \sigma^2 - \sigma^2 \\ &= (n-1) \cdot \sigma^2 . \end{aligned}$$

Dies ist - wiederum aufgrund der Linearität des Erwartungswerts - gleichbedeutend mit dem nachzuweisenden Sachverhalt.

Anmerkung 5: Die folgende Darstellung der Stichprobenvarianz, welche ebenfalls die Verwendung des Nenners $n-1$ nahelegt,

$$s_n^2 = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{(x_j - x_i)^2}{2}$$

bietet eine Möglichkeit, diese mit Hilfe der wechselseitigen Abweichungsquadrate der Beobachtungswerte, und damit ohne die Benützung des Stichprobenmittels zu definieren (vgl. etwa [26]).[17]

[17] Gemäß dieser Darstellung lässt sich s_n^2 übrigens als durchschnittliche potentielle Energie eines System von n Massepunkten der Masse 1 interpretieren, die paarweise durch virtuelle Hooksche Federkräfte verbunden sind.

Die obige Darstellung lässt sich folgendermaßen einsehen. Durch Anwendung der bereits beim Nachweis des Steinerschen Verschiebungssatzes benützten Beziehung

$$a^2 = b^2 + (a-b)^2 + 2b(a-b)$$

auf die Größen $a = x_j - x_i$ und $b = x_j - \bar{x}_n$ lassen sich die wechselseitigen Abstandsquadrate $(x_i - x_j)^2$ mit Hilfe des Stichprobenmittels wie folgt darstellen

$$(x_j - x_i)^2 = (x_j - \bar{x}_n)^2 + (\bar{x}_n - x_i)^2 + 2(x_j - \bar{x}_n)(\bar{x}_n - x_i)\,.$$

Die Summe der Hälften dieser Größen ist, da die Beiträge für $j = i$ verschwinden und weil bekanntlich $\sum_{j=1}^n (x_j - \bar{x}_n) = 0$ ist, gleich

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=i+1}^n \frac{(x_j - x_i)^2}{2} &= \frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n (x_j - x_i)^2 \\
&= \tfrac{1}{4}\left[\begin{array}{l} \sum_{i=1}^n \sum_{j=1}^n (x_j - \bar{x}_n)^2 + \sum_{i=1}^n \sum_{j=1}^n (\bar{x}_n - x_i)^2 + \\ +\, 2\sum_{j=1}^n (x_j - \bar{x}_n)\sum_{i=1}^n (\bar{x}_n - x_i) \end{array}\right] \\
&= \frac{1}{4} \times 2n \sum_{j=1}^n (x_j - \bar{x}_n)^2 \\
&= \frac{n}{2}\sum_{i=1}^n (x_i - \bar{x}_n)^2\,.
\end{aligned}$$

Da es $\binom{n}{2} = \frac{n(n-1)}{2}$ von 0 verschiedene wechselseitige Abstandsquadrate gibt, ist der betrachtete Durchschnittswert tatsächlich gleich der Stichprobenvarianz.

1.4.2 Stichprobenmedian und mittlere absolute Abweichung

a) **Der Stichprobenmedian** [18]

Zur Bestimmung des Stichprobenmedians ist es notwendig, die ursprünglichen Daten $x_1, ..., x_n$ der Größe nach zu ordnen. Seien also $x_{1:n}, ..., x_{n:n}$

[18] Vom lateinischen Wort *medius 3: der, die, das Mittlere* abgeleitet

die der Größe nach geordneten Daten (somit ist $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$ und es gilt $\{x_{1:n}, ..., x_{n:n}\} = \{x_1, ..., x_n\}$). Dann ist

$$\tilde{x}_n = \begin{cases} x_{(n+1)/2:n} & \text{für } n \text{ ungerade} \\ (x_{n/2:n} + x_{n/2+1:n})/2 & \text{für } n \text{ gerade} \end{cases}$$

der *Median* der Daten. Dabei sind

$x_{(n+1)/2:n}$	der Wert in der Mitte der geordneten Datenliste,
$(x_{n/2:n} + x_{n/2+1:n})/2$	das arithmetische Mittel der beiden Werte in der Mitte.

Anmerkung 1: Sofern n ungerade ist, verwendet man für die Definition des Stichprobenmedians nur die Ordnungsrelation "$<$" . Demnach ist die Definition des Stichprobenmedians in diesem Fall auch für ordinale Variable - wie z.B. für Schulnoten - möglich.

Anmerkung 2: Sei $\lceil x \rceil$ die kleinste ganze Zahl $\geq x$, d.h. sei $\lceil x \rceil = \min \{ i \in \mathbb{Z} : i \geq x \}$. Mit Hilfe dieser Bezeichnung kann der Stichprobenmedian ohne Fallunterscheidung definiert werden:

$$\tilde{x}_n = \left(x_{\lceil n/2 \rceil :n} + x_{\lceil (n+1)/2 \rceil :n} \right) /2 \, .$$

Anmerkung 3: Wird im Fall $n \geq 3$ ein Beobachtungswert $x_i = x_{n:n}$ durch einen extrem großen Wert ersetzt, so hat dies keinen Einfluss auf den Stichprobenmedian, hingegen einen beträchtlichen auf das Stichprobenmittel. Man sagt: Der Stichprobenmedian ist robust (resistent) gegen Ausreißer. Das Stichprobenmittel hingegen ist sensitiv gegen Ausreißer.

Die **Minimaleigenschaft des Stichprobenmedians** hinsichtlich der Summe der Absolutbeträge der Abweichungen

Behauptung: Die Funktion

$$g(x) = \frac{1}{n} \sum_{i=1}^{n} |x_i - x|$$

ist genau dann minimal, wenn $x \in \left[x_{\lceil n/2 \rceil :n}, x_{\lceil (n+1)/2 \rceil :n} \right]$ ist.

Beweis: Um die Funktion $g(x)$, welche sich äquivalent durch $g(x) = \frac{1}{n} \sum_{i=1}^{n} |x_{i:n} - x|$ ausdrücken lässt, zu diskutieren, unterscheiden wir alle möglichen Fälle.

Für $x < x_{1:n}$ ist $g(x) = \frac{1}{n}\sum_{i=1}^{n}(x_{i:n} - x) = \frac{1}{n}\sum_{i=1}^{n} x_i - x = \bar{x}_n - x$.

Für $x \geq x_{n:n}$ ist $g(x) = \frac{1}{n}\sum_{i=1}^{n}(x - x_{i:n}) = x - \frac{1}{n}\sum_{i=1}^{n} x_i = x - \bar{x}_n$.

Diese beiden Fälle lassen sich mit Hilfe von $x_{0:n} = -\infty$ und $x_{n+1:n} = +\infty$ als Spezialfälle des folgenden allgemeinen Falls betrachten:

Seien $i \in \{0, ..., n\}$ und $x \in [x_{i:n}, x_{i+1:n}]$ und bezeichne $b_i^{(n)} = [\sum_{j=i+1}^{n} x_{j:n} - \sum_{j=1}^{i} x_{j:n}]/n$. Dann ist

$$\begin{aligned} g(x) &= \frac{1}{n}[\sum_{j=1}^{i}(x - x_{j:n}) + \sum_{j=i+1}^{n}(x_{j:n} - x)] = x \cdot \frac{i - (n-i)}{n} + b_i^{(n)} \\ &= 2x(\frac{i}{n} - \frac{1}{2}) + b_i^{(n)}. \end{aligned}$$

Dies ist die Gleichung einer Geraden mit Anstieg $2\left(\frac{i}{n} - \frac{1}{2}\right)$. Dieser ist

$$\begin{cases} < 0 & \text{für} \quad i/n < 1/2 \\ = 0 & \text{für} \quad i/n = 1/2 \\ > 0 & \text{für} \quad i/n > 1/2. \end{cases}$$

Also nimmt $g(x)$ im Bereich $[x_{\lceil n/2 \rceil:n}, x_{\lceil (n+1)/2 \rceil:n}]$ sein Minimum an.

b) **Die mittlere absolute Abweichung** vom Stichprobenmedian ist

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

Diese ist ein mit der Standardabweichung vergleichbares Maß für die Abweichung der Beobachtungswerte vom zugehörigen Zentralmaß.

c) Das einfachste Streumaß ist die **Spannweite der Stichprobe**, nämlich die Differenz

$$x_{n:n} - x_{1:n}$$

zwischen Stichprobenmaximum und Stichprobenminimum.

Anmerkung 4: Für $n \geq 2$ und $x_{n:n} - x_{1:n} > 0$ gelten

$$\text{a)} \quad \tilde{s}_n = \frac{1}{n} \cdot \sum_{j=1}^{\lfloor n/2 \rfloor} (x_{n+1-j:n} - x_{j:n}) \leq \frac{1}{2} \cdot (x_{n:n} - x_{1:n}) ,$$

$$\text{b)} \quad \tilde{s}_n \leq \sqrt{\frac{n-1}{n}} \cdot s_n \,,$$

und

$$\text{c)} \quad s_n < \sqrt{\frac{1}{n-1} \cdot \sum_{j=1}^{\lfloor n/2 \rfloor} \left(x_{n+1-j:n} - x_{j:n}\right)^2} \leq \sqrt{\frac{\lfloor n/2 \rfloor}{n-1}} \cdot \left(x_{n:n} - x_{1:n}\right) \,,$$

wobei für $n = 2$ in allen drei Ungleichungen des Typs "$\leq$" Gleichheit gilt.

1.4.3 Quartile und Kasten-Bild

In der nachstehenden Definition betrachten wir eine Verallgemeinerung des Medians.

Definition: Sei $\alpha \in (0,1)$. Dann heißt

$$q_{\alpha,n} = \begin{cases} x_{\lceil \alpha \cdot n \rceil :n} & \text{für } \alpha \cdot n \notin \mathbb{N} \\ (1-\alpha) x_{\alpha \cdot n:n} + \alpha x_{\alpha \cdot n+1:n} & \text{für } \alpha \cdot n \in \mathbb{N} \end{cases}$$

das α-*Quantil* der Stichprobe.

Die Spezialfälle für $\alpha \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$, nämlich das *1. Quartil*, der Median und das *3. Quartil*, lassen sich folgendermaßen ohne Fallunterscheidung definieren

$$\begin{array}{lllll} \text{das 1. Quartil} & q_{1/4,n} & = & \left(3x_{\lceil n/4 \rceil :n} + x_{\lceil (n+1)/4 \rceil :n}\right)/4 \\ \text{der Median} & \tilde{x}_n = q_{1/2,n} & = & \left(x_{\lceil n/2 \rceil :n} + x_{\lceil (n+1)/2 \rceil :n}\right)/2 \\ \text{das 3. Quartil} & q_{3/4,n} & = & \left(x_{\lceil 3n/4 \rceil :n} + 3x_{\lceil (3n+1)/4 \rceil :n}\right)/4 \,. \end{array}$$

Zusammen mit dem Stichprobenminimum $x_{1:n}$ und dem Stichprobenmaximum $x_{n:n}$ benötigt man diese zum Anfertigen eines Kasten-Bilds[19].

Ein Streumaß, welches - im Unterschied zur Spannweite - robust gegen Ausreißer ist, ist der **Interquartilabstand**

$$q_{3/4,n} - q_{1/4,n} \,.$$

[19] Das *Kastenbild* (*box plot*) ist eine knappe visuelle Darstellung der Verteilung der Stichprobenwerte. Es stammt, wie das *Stängel-Blatt-Diagramm* aus Abschnitt 1.3.3, von *John W. Tukey*.

1.4.4 Ausblick: Stichprobenmittel und Abweichungsmaß für zirkuläre Daten

Bezeichnen $\varphi_1, ..., \varphi_n \in [0, 2\pi)$ die beobachteten Winkel und $\vec{r}_i = (\cos\varphi_i, \sin\varphi_i)$, $i \in \{1, ..., n\}$, die zugehörigen Richtungsvektoren.

Wir sind im folgenden am arithmetischen Mittel

$$\vec{R}_n = \frac{1}{n}\sum_{i=1}^{n} \vec{r}_i$$

dieser Vektoren, dem sogenannten *zirkulären Mittel,* interessiert.

Sofern $\vec{R}_n$ ungleich dem Nullvektor ist, ist dessen Richtung, *mittlere Richtung* genannt, eindeutig bestimmt. Da die x- und y-Koordinate von $\vec{R}_n$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} \cos\varphi_i \quad \text{bzw.} \quad \bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} \sin\varphi_i$$

sind, ist der zugehörige Winkel

$$\tilde{\varphi}_n = \begin{cases} \arccos(\bar{X}_n / \left|\vec{R}_n\right|) & \text{falls} \quad \bar{Y}_n \geq 0 \\ 2\pi - \arccos(\bar{X}_n / \left|\vec{R}_n\right|) & \text{falls} \quad \bar{Y}_n < 0\,, \end{cases}$$

wobei $\arccos(.)$ den Hauptwert des Arcus-Cosinus bezeichnet. Außer für den Stichprobenumfang $n = 2$ ist $\tilde{\varphi}_n$ typischerweise vom stets existierenden Stichprobenmittel $\bar{\varphi}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi_n$ der Winkel verschieden.

Nun zur Länge von $\vec{R}_n$. Für die Länge des Summenvektors $\sum_{i=1}^{n} \vec{r}_i$ gilt aufgrund der Dreiecksungleichung und der Tatsache, dass alle Vektoren $\vec{r}_i$ die Länge 1 haben, $|\sum_{i=1}^{n} \vec{r}_i| \leq \sum_{i=1}^{n} |\vec{r}_i| = n$ bzw., gleichbedeutend damit,

$$\left|\vec{R}_n\right| \leq 1\,.$$

Dabei gilt Gleichheit offenbar genau dann, wenn alle Vektoren $\vec{r}_i$ gleich sind.

Das bedeutet, dass das zirkuläre Mittel durch einen Punkt im oder am Rand des Einheitkreises beschrieben wird, wobei letztes - wie gesagt - nur

dann zutreffen kann, wenn alle Vektoren $\vec{r}_i$ gleich sind. Dieser Sachverhalt motiviert es auch, die Größe

$$d_n = 1 - \left|\vec{R}_n\right| ,$$

nämlich den (minimalen) Abstand des Punktes $\vec{R}_n$ vom Rand des Einheitskreises, als *zirkuläres Abweichungsmaß* zu definieren.

Um eine Vorstellung davon zu erhalten, wie diese Größe durch die Winkel beschrieben werden kann, betrachten wir den Spezialfall zweier Vektoren.

Spezialfall $\mathbf{n = 2}$: Wir setzen, eigentlich ohne Einschränkung der Allgemeinheit, $\varphi_1 < \varphi_2$ und $\varphi_2 - \varphi_1 < \pi$ voraus. Dann ist die Länge des zirkulären Mittels $\frac{1}{2}(\vec{r}_1 + \vec{r}_2)$ gleich der durch den Scheitel $(0,0)$ gelegten Höhe des durch die Ecke $(0,0)$ und die Schenkel $\vec{r}_1$ und $\vec{r}_2$ bestimmten gleichschenkeligen Dreiecks und somit gleich

$$\left|\vec{R}_2\right| = \left|\cos(\frac{\varphi_2 - \varphi_1}{2})\right| .$$

Die nachstehende Behauptung ist eine Verallgemeinerung dieses Sachverhalts.

Behauptung: Die Größe $\vec{R}_n^2$ lässt sich mit Hilfe der Cosinus-Quadrate der halben wechselseitigen Abweichungen der Winkel in folgender Weise darstellen

$$\vec{R}_n^2 = \frac{2}{n}[1 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\cos^2(\frac{\varphi_j - \varphi_i}{2})] - 1 .$$

Diese Darstellung entspricht jener der Stichprobenvarianz in Anmerkung 4 aus Abschnitt 1.4.1. Dementsprechend gilt für das zirkuläre Abweichungsmaß

$$d_n = 1 - \sqrt{\frac{2}{n}[1 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\cos^2(\frac{\varphi_j - \varphi_i}{2})] - 1} .$$

Beweis: Offensichtlich gilt $\vec{R}_n^2 = \bar{X}_n^2 + \bar{Y}_n^2$. Ausquadrieren und Berücksichtigen des Summensatzes $\cos\alpha\cos\beta + \sin\alpha\sin\beta = \cos(\alpha - \beta)$ liefert die dritte Zeile der folgenden Umformungen. Addieren und gleichzeitiges Abziehen von 1 , Anwenden des Spezialfalls $\cos\alpha + 1 = 2\cos^2\left(\frac{\alpha}{2}\right)$ der

Beziehung $\cos\alpha + \cos\beta = 2\cos\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)$ für $\beta = 0$, sowie Berücksichtigen von $\cos 0 = 1$ für die Summanden $j = i$ und $\cos\alpha = \cos(-\alpha)$ für die einander entsprechenden Indexpaare (i,j) und (j,i) ergeben schließlich das gewünsche Resultat

$$\begin{aligned}
\vec{R}_n^2 &= \frac{1}{n^2}[(\sum_{i=1}^{n}\cos\varphi_i)^2 + (\sum_{i=1}^{n}\sin\varphi_i)^2] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\cos\varphi_j\cos\varphi_i + \sin\varphi_j\sin\varphi_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos(\varphi_j - \varphi_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\cos(\varphi_j - \varphi_i) + 1\right) - 1 \\
&= \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\cos^2(\frac{\varphi_j - \varphi_i}{2}) - 1 \\
&= \frac{2}{n}[1 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\cos^2(\frac{\varphi_j - \varphi_i}{2})] - 1\,.
\end{aligned}$$

1.5 WEITERE MITTELWERTE

1.5.1 Das geometrische Mittel

Beispiel 1: Man bestimme die Wachstumsrate für das im Schaubild in *Küttings* Artikel [24] dargestellte Wirtschaftswachstum. Offensichtlich wächst die Wirtschaft im angegebenen Zeitraum von 6 Jahren mit dem Faktor

$$(1+0.044)\cdot(1+0.004)\cdot(1-0.02)\cdot(1+0.057)\cdot(1+0.026)\cdot(1+0.034) = 1.1519\,.$$

Um dieses Resultat bei einer für jedes Jahr gleichen Wachstumsrate r zu erreichen, muss gelten

$$(1+r)\cdot(1+r)\cdot(1+r)\cdot(1+r)\cdot(1+r)\cdot(1+r) = (1+r)^6 = 1.1519$$

und somit

$$1+r = \sqrt[6]{1.1519} = 1+0.0238\,.$$

Anmerkung 1: Es gilt

$$1+\frac{0.044+0.004-0.02+0.057+0.026+0.034}{6} = 1+0.143 > 1+0.0238\,.$$

Anders ausgedrückt: Bei einem durchschnittlichen Prozentsatz von 14.3 % würde die Wirtschaft gemäß $1.143^6 = 2.2299$ mit dem Faktor 2.2299 wachsen!

Definition: Seien $x_1, ..., x_n \geq 0$. Dann nennt man

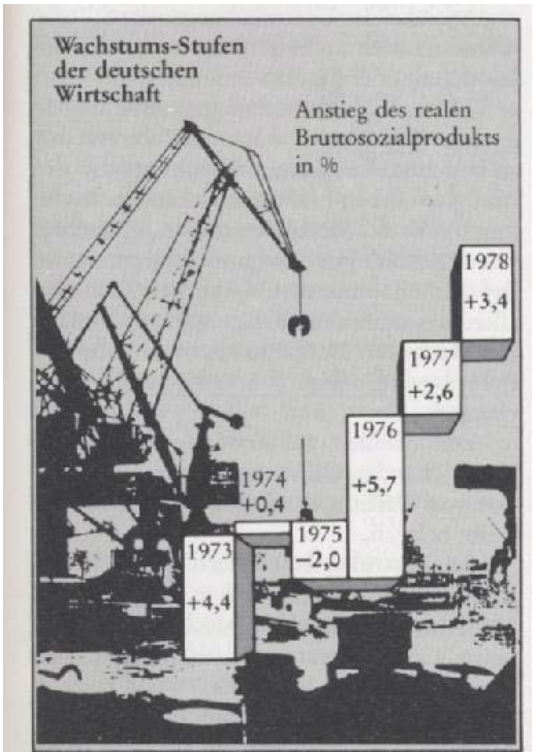
$$\hat{x}_n = \sqrt[n]{x_1 \cdot ... \cdot x_n}$$

das *geometrische Mittel* der Zahlen $x_1, ..., x_n$.

Bezeichnung: Für $x_i = 1 + r_i\,,\ i \in \{1, ..., n\}$ nennt man die Größe

$$\hat{x}_n - 1 = \sqrt[n]{(1+r_1)\cdot ... \cdot (1+r_n)} - 1$$

die *Wachstumsrate* (*growth rate*).



Anmerkung 2: Das geometrische Mittel eignet sich für Größen, die multiplikativ verknüpft werden. Typisch dafür ist der durchschnittliche *Aufzinsungsfaktor* in der Zinseszinsrechnung. Demgemäß wird es häufig als Durchschnittswert bei zeitlich aufeinanderfolgenden Veränderungsraten verwendet, also etwa bei jährlich erhobenen Kosten- und Preisindizes, den Quotienten der Börsenkurse von Ende zu Beginn eines Tages und den Wachstumsraten von Individuen und Populationen.

Behauptung 1: Es gilt

$$\sqrt[n]{\prod_{i=1}^{n} x_i} \leq \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$$

mit Gleichheit genau dann, wenn gilt $x_1 = ... = x_n$.

Beweis: Ist eines der x_i gleich 0, so ist die Gültigkeit der Ungleichung unmittelbar einsichtig. Im weiteren seien also die $x_1, ..., x_n >$

0 vorausgesetzt. Wir gehen vom folgenden **fundamentalen Sachverhalt** aus:

$\ln u \leq u - 1 \;\; \forall\, u \in (0, \infty)$ mit Gleichheit genau dann, wenn $u = 1$ ist.

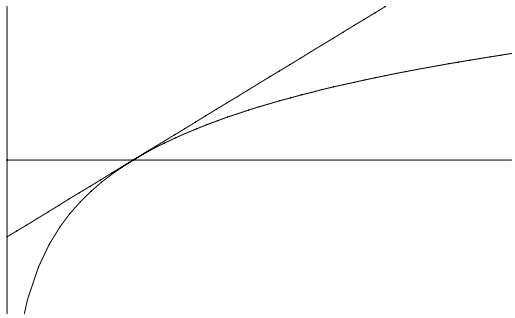


Abbildung der Funktion $u \mapsto \ln u$ und deren Tangente $u \mapsto u - 1$ im Punkt $(1, 0)$

Wendet man diesen Sachverhalt auf die Quotienten $\frac{x_i}{\bar{x}_n}$, $i \in \{1, ..., n\}$, an, summiert über alle i und dividiert schließlich durch n, so erhält man

$$\frac{1}{n}\sum_{i=1}^{n} \ln(\frac{x_i}{\bar{x}_n}) \leq \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i}{\bar{x}_n} - 1) = \frac{1}{\bar{x}_n} \cdot \frac{1}{n}\sum_{i=1}^{n} x_i - 1 = \frac{\bar{x}_n}{\bar{x}_n} - 1 = 0\,,$$

wobei Gleichheit genau dann gilt, wenn alle Quotienten $\frac{x_i}{\bar{x}_n}$ gleich 1 sind, oder gleichbedeutend, wenn $x_1 = ... = x_n = \bar{x}_n$ ist.

Wegen der Additivität des Logarithmus ist die obige Ungleichung gleichbedeutend mit

$$\frac{1}{n}\sum_{i=1}^{n} \ln x_i \leq \ln \bar{x}_n\,.$$

Wendet man auf diese Ungleichung die Exponentialfunktion an und berücksichtigt die Monotonie derselben, so erhält man die Behauptung. □

Der oben verwendete **fundamentale Sachverhalt** ist äquivalent mit folgender

Aussage: Die durch

$$g(u) = u - 1 - \ln u$$

auf dem Intervall $(0, \infty)$ definierte Funktion besitzt an der Stelle $u = 1$ ihr eindeutiges Minimum $g(1) = 0$. Diese Aussage ist aus der Betrachtung der Ableitung $g'(u) = 1 - \frac{1}{u}$ unmittelbar einsichtig. Denn letztere ist

$$1 - \frac{1}{u} = \begin{cases} < 0 & \text{für} \quad u < 1 \\ = 0 & \text{für} \quad u = 1 \\ > 0 & \text{für} \quad u > 1\,. \end{cases}$$

Wegen

$$\ln(\sqrt[n]{\prod_{i=1}^{n} x_i}) = \frac{1}{n} \sum_{i=1}^{n} \ln(x_i)$$

sind logarithmische Skalen mit dem geometrischen Mittel eng verknüpft.

Logarithmische Skalen sind in verschiedenen Wissenschaftsdiziplinen gebräuchlich, wie zum Beispiel die beiden physiologischen Größen Schallpegel (in Dezibel) und die scheinbare Helligkeit von Sternen (Magnitudo).[20]

Der *pH-Wert*

$$pH = -\log_{10}\left[H^+\right]$$

von wasserhaltigen Flüssigkeiten misst die Konzentration der H^+-Ionen und gibt an, wie sauer oder basisch eine Lösung ist. So bedeuten

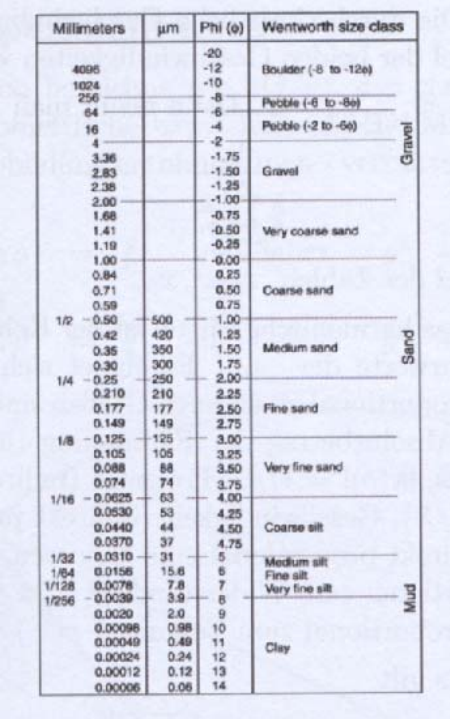
$$pH = 0: \quad \text{extrem sauer} \qquad pH = 7: \quad \text{neutral} \qquad pH = 14: \quad \text{extrem basisch.}$$

Das Symbol "p" steht übrigens für "negativer dekadischer Logarithmus von" und ist somit ein Synomym für die logarithmische Skalierung.

Die in der Petrographie verwendete *Udden-Wentworth-Skala* wird zur Klassifikation von Korngrößen herangezogen. Ist d der Korndurchmesser (in mm), dann ist der zugehörige *Udden-Wentworth-Skalenwert* $\Phi(d)$ gegeben durch

$$\Phi(d) = -\log_2(d)\,.$$

[20] Die physiologischen Größen Schallpegel und Helligkeit messen die durch die physikalischen Größen Schallintensität bzw. Strahlungsstrom beim Menschen hervorgerufene Empfindung. Dabei kommt das *Weber-Fechner*sche Gesetz der Psychophysik zur Anwendung, nach dem die Empfindungsstärke proportional zum Logaritmus der Reizstärke ist.



Millimeters	µm	Phi (ϕ)	Wentworth size class	
		-20		
4096		-12	Boulder (-8 to -12ϕ)	
1024		-10		
256		-8	Pebble (-6 to -8ϕ)	
64		-6		
16		-4	Pebble (-2 to -6ϕ)	Gravel
4		-2		
3.36		-1.75		
2.83		-1.50	Gravel	
2.38		-1.25		
2.00		-1.00		
1.68		-0.75		
1.41		-0.50	Very coarse sand	
1.19		-0.25		
1.00		-0.00		
0.84		0.25		
0.71		0.50	Coarse sand	
0.59		0.75		
1/2 – 0.50	500	1.00		
0.42	420	1.25		
0.35	350	1.50	Medium sand	Sand
0.30	300	1.75		
1/4 – 0.25	250	2.00		
0.210	210	2.25		
0.177	177	2.50	Fine sand	
0.149	149	2.75		
1/8 – 0.125	125	3.00		
0.105	105	3.25		
0.088	88	3.50	Very fine sand	
0.074	74	3.75		
1/16 – 0.0625	63	4.00		
0.0530	53	4.25		
0.0440	44	4.50	Coarse silt	
0.0370	37	4.75		
1/32 – 0.0310	31	5	Medium silt	
1/64 0.0156	15.6	6	Fine silt	
1/128 0.0078	7.8	7	Very fine silt	
1/256 – 0.0039	3.9	8		Mud
0.0020	2.0	9		
0.00098	0.98	10		
0.00049	0.49	11	Clay	
0.00024	0.24	12		
0.00012	0.12	13		
0.00006	0.06	14		

Udden-Wentworth-Scale grain-size scale

1.5.2 Das harmonische Mittel

Beispiel 2: Die ersten 100 km legt ein Zug mit einer konstanten Geschwindigkeit von 70 km/h zurück, die zweiten 100 km mit einer konstanten Geschwindigkeit von 110 km/h. Mit welcher Durchschnittsgeschwindigkeit fährt der Zug?

Bezeichnen $s_1, s_2; v_1, v_2; t_1, t_2$ Weg, Geschwindigkeit und Zeit der jeweiligen Teilstrecken, und $\tilde{v}$ die gesuchte Durchschnittsgeschwindigkeit, dann gilt wegen der Beziehungen $s = \tilde{v} \cdot t$ (Weg = Geschwindigkeit×Zeit), $s = s_1 + s_2$ und $t = t_1 + t_2$

$$\begin{aligned} \tilde{v} &= \frac{s}{t} = \frac{s_1 + s_2}{t_1 + t_2} = \frac{s_1 + s_2}{\frac{s_1}{v_1} + \frac{s_2}{v_2}} \\ &= \frac{1}{\frac{s_1}{s_1+s_2} \cdot \frac{1}{v_1} + \frac{s_2}{s_1+s_2} \cdot \frac{1}{v_2}} . \end{aligned}$$

Die numerische Lösung ist daher $\tilde{v} = \frac{1}{\frac{1}{2} \cdot \frac{1}{70} + \frac{1}{2} \cdot \frac{1}{110}} = 85.56$ km/h.

Anmerkung 1: Die durchschnittliche Geschwindigkeit wäre 90 km/h, das geometrische Mittel der beiden Geschwindigkeiten wäre 87.75 km/h!

Definition: Seien $x_1, ..., x_n > 0$. Dann nennt man

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

das *harmonische Mittel* der Zahlen $x_1, ..., x_n$.

Anmerkung 2: Das harmonische Mittel ist der Kehrwert des arithmetischen Mittels der Kehrwerte der x_i. Es eignet sich als Mittelwert von Größen, die indirekt proportional zu anderen Größen sind. Anwendungsmöglichkeiten sind z.B.: Absolutbetrag der Krümmung (indirekt proportional zum Krümmungsradius, $|\kappa(\rho)| = 1/\rho$), Frequenz (indirekt proportional zur Wellenlänge, $\nu(\lambda) = 1/\lambda$), Geschwindigkeit (indirekt proportional zur Zeit, $v(t) = \frac{s}{t}$), Dichte (indirekt proportional zum Volumen, $\rho(V) = \frac{m}{V}$), Stromstärke (indirekt proportional zum Widerstand, $I(R) = \frac{U}{R}$), Druck eines idealen Gases (indirekt proportional zum Volumen, $p(V) = \frac{RT}{V}$).

Behauptung 2: Es gilt

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \leq \sqrt[n]{\prod_{i=1}^{n} x_i}$$

mit Gleichheit genau dann, wenn gilt $x_1 = ... = x_n$.

Beweis: Diese Ungleichung ist gleichbedeutend mit

$$\sqrt[n]{\prod_{i=1}^{n}\frac{1}{x_i}} \leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i},$$

was identisch mit der Ungleichung zwischen dem geometrischen und dem arithmetischen Mittel der Kehrwerte $y_i = \frac{1}{x_i}$, $i \in \{1, ..., n\}$ ist. □

Aus den Behauptungen 1 und 2 ergibt sich zusammenfassend

Behauptung 3: Für $x_i \in (0, \infty)$, $i \in \{1, ..., n\}$ gilt

$$\min(x_1, ..., x_n) \leq \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} \leq \hat{x}_n \leq \bar{x}_n \leq \max(x_1, ..., x_n)$$

mit Gleichheit jeweils genau dann, wenn gilt $x_1 = ... = x_n$.

1.5.3 Der Spezialfall n = 2

Für $0 < x < y < \infty$ bezeichne $a = a(x,y)$, $g = g(x,y)$ und $h = h(x,y)$ das arithmetische, geometrische bzw. harmonische Mittel von x und y. Der nachstehenden Abbildung entnimmt man - vermittels der Ähnlichkeit jeweils zweier Dreiecke -

$$\frac{x}{g} \ (= \tan\alpha) \ = \ \frac{g}{y} \qquad \text{und daher} \qquad g^2 \ = \ x \cdot y \ \ ... \ \ (1_\alpha)$$

und

$$\frac{h}{g} \ (= \cos\beta) \ = \ \frac{g}{a} \qquad \text{und daher} \qquad g^2 \ = \ h \cdot a \ \ ... \ \ (1_\beta)$$

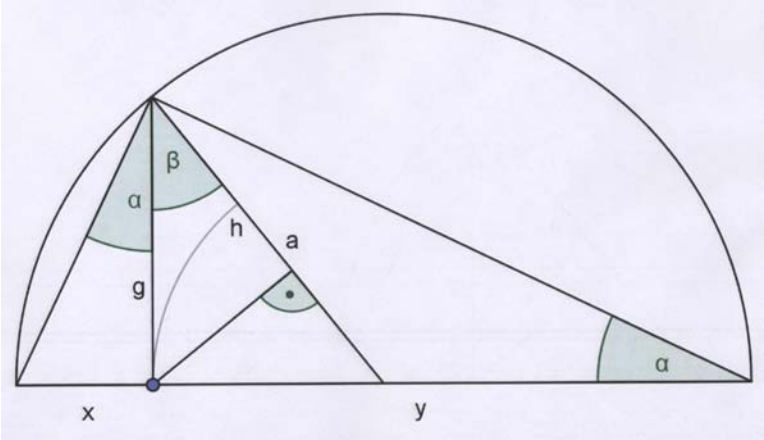


Abbildung: **a**rithmetisches, **g**eometrisches und **h**armonisches Mittel von x und y

Also gilt

$$h(x,y) \cdot a(x,y) = xy \tag{2}$$

und somit im Hinblick auf (1_α) $g(x,y) = \sqrt{xy}$ und wegen $a(x,y) = \frac{x+y}{2}$ im Hinblick auf (2)

$$h(x,y) = \frac{1}{\frac{1}{2}(\frac{1}{x} + \frac{1}{y})} \, .$$

Ferner entnimmt man der Abbildung die Gültigkeit der Ungleichungskette

$$x < h(x,y) < g(x,y) < a(x,y) < y \, . \tag{3}$$

Für den Fall $x = y$ gilt offensichtlich überall das Gleichheitszeichen.

Auf den beiden Beziehungen (2) und (3) beruht der sogenannte *babylonische Wurzelalgorithmus* zur numerischen Approximation der Quadratwurzel (siehe Aufgabe M 6).

Anmerkung 3: Analog zu (1_β) erhält man auch

$$(a-x)^2 = (a-h)a\,.$$

Daraus folgt zusammen mit (1_β) der Pythagoräische Lehrsatz

$$g^2 + (a-x)^2 = a^2\,.$$

1.5.4 Einfache Bewegungsaufgaben zum arithmetischen, geometrischen und harmonischen Mittel

Beispiel 3: Ein Zug fährt mit konstanter Geschwindigkeit v auf einer schnurgeraden Strecke. Ein Fahrgast geht von seinem Abteil am Ende des Zuges mit der relativ zum Zug konstanten - und vergleichsweise kleinen - Geschwindigkeit $k < v$ zum Speisewagen vor und von dort wieder mit derselben konstanten Geschwindigkeit k zu seinem Abteil zurück.

Wie stellt sich die durchschnittliche Geschwindigkeit $\bar{v}$ des Fahrgastes für einen ruhenden Beobachter dar?

Lösung: Da die Zeit t_0, die der Fahrgast bis zum Speisewagen benötigt, und die, welche er von dort wieder zurück zu seinem Platz benötigt, übereinstimmen, und die beiden zurückgelegten Wegstrecken für den ruhenden Beobachter demnach $(v+k)\,t_0$ und $(v-k)\,t_0$ sind, ist die Lösung das arithmetische Mittel von $v+k$ und $v-k$, nämlich

$$\begin{aligned} \bar{v} &= \frac{(v+k)\,t_0 + (v-k)\,t_0}{2t_0} \\ &= \frac{v+k+v-k}{2} \\ &= v\,. \end{aligned}$$

Die folgenden beiden Aufgaben könnten als vorbereitende Überlegungen zu einem fächerübergreifenden Thema dienen, welches mit dem Experiment von *Michelson* und *Morley* aus dem Jahr 1886 verknüpft ist. In diesem bahnbrechenden Experiment wurde nachgewiesen, dass die Lichtgeschwindigkeit c unabhängig davon ist, ob diese in der Richtung der Erdbewegung oder normal dazu gemessen wird, und zirka $300\,000\,km/s$ beträgt. Diese Tatsache

ist eine der tragenden Säulen von Einsteins spezieller Relativitätstheorie[21], [22].

Beispiel 4.I: Ein (gläserner) Zug fährt mit konstanter Geschwindigkeit v auf einer schnurgeraden Strecke. Ein Fahrgast wirft einen Gegenstand mit einer vergleichsweisen großen Geschwindigkeit $\hat{v} >> v$ vertikal in die Höhe. (Wir lassen dabei die Erdanziehung unberücksichtigt, sodass wir davon ausgehen können, daß der Gegenstand mit konstanter Geschwindigkeit zur Decke des Waggons fliegt.) Ein gegenüber dem Zug ruhender Beobachter, der sich in einem großen Abstand von der Bahntrasse befindet, sieht, dass sich der Gegenstand mit der Geschwindigkeit $c >> v$ schräg aufwärts bewegt.

a) Stellen Sie die vertikale Geschwindigkeitskomponente $\hat{v}$ - also die Geschwindigkeit, die der Fahrgast wahrnimmt - in Abhängigkeit von c und v dar und

b) überlegen Sie, dass sich $\hat{v}$ als geometrisches Mittel der Größen $c + v$ und $c - v$ interpretieren lässt.

Lösung: Nimmt man - ohne Beschränkung der Allgemeinheit - an, dass der Gegenstand die Zeiteinheit $t_0 = 1$ nach oben fliegt, dann sind Geschwindigkeit und zurückgelegter Weg identisch. Demnach gilt aufgrund des pythagoräischen Lehrsatzes und der Beziehung $a^2 - b^2 = (a - b)(a + b)$

$$\begin{aligned} \hat{v} &= \sqrt{c^2 - v^2} = (\, c \cdot \sqrt{1 - \frac{v^2}{c^2}}\,) \\ &= \sqrt{(c - v)(c + v)}\,. \end{aligned}$$

Beispiel 4.II: Angenommen, der Gegenstand wird mit der relativ zur Bahntrasse konstanten und sehr großen Geschwindigkeit $c >> v$ vom Ende des Zuges zu dessen Anfang befördert. Dort wendet der Gegenstand und wird mit derselben (relativ zur Bahntrasse) konstanten Geschwindigkeit c wieder zum Zugsende zurückbefördert.

Bestimmen Sie die durchschnittliche Geschwindigkeit des Gegenstandes, die ein sitzender Fahrgast im Zug wahrnimmt.

Lösung: Die beiden Geschwindigkeiten, die unser Fahrgast beobachtet, sind $c - v$ und $c + v$. Bezeichnet nun d die Länge des Zuges, dann

[21] *Albert Einstein*: Zur Elektrodynamik bewegter Körper. Annalen der Physik (1905), S. 891-921.

[22] Eine schülergerechte Darstellung dieses Themas findet man etwa in [17], S. 7 ff.

sind für die beiden Zuglängen benötigten Zeiten $\frac{d}{c-v}$ und $\frac{d}{c+v}$. Daher ist die durchschnittliche Geschwindigkeit gleich dem harmonischen Mittel

$$\begin{aligned}\tilde{v} &= \frac{2d}{\frac{d}{c-v}+\frac{d}{c+v}} \\ &= \frac{2}{\frac{1}{c-v}+\frac{1}{c+v}}\,.\end{aligned}$$

Indem wir die beiden Summanden im Nenner auf gemeinsamen Nenner bringen, ergibt sich dieses im konkreten Fall zu

$$\begin{aligned}\tilde{v} &= \frac{2}{\frac{c+v}{c^2-v^2}+\frac{c-v}{c^2-v^2}} = \frac{c^2-v^2}{c} \\ &= c\cdot(1-\frac{v^2}{c^2}) \\ &< \hat{v} = c\cdot\sqrt{1-\frac{v^2}{c^2}}\,.\end{aligned}$$

1.5.5 Ausblick 1: Zum Ursprung der Mittelwerte bei den Pythagoräern

Über das Leben von *Pythagoras* (Samos um 570 - Metapont um 496 v.Chr.) gibt es viele Legenden, jedoch kaum zuverlässige Informationen. Der Geschichtsschreiber *Jamblichos von Chalkis* (ca. 250-330 n.Chr.) berichtet, dass er von einem Aufenthalt in Mesopotamien (dem Gebiet des heutigen Irak) die Kenntnisse der drei "*musikalischen Proportionen*" mitgebracht habe, welche schon ca. 2000 v.Chr. von den Babyloniern für ihren Quadratwurzelalgorithmus benutzt worden seien. Jedenfalls gründete *Pythagoras* in Kroton, einer griechischen Kolonie in Süditalien, eine religiös-kultisch orientierte Lebensgemeinschaft, deren Mitglieder sich um die Erfüllung der asketischen, auf Reinhaltung der Seele (*katharsis*) abzielende Verhaltensregeln des Meisters bemühten.

Ob der Pythagoräische Lehrsatz tatsächlich *Pythagoras* zuzuschreiben ist, ist ungewiss. Seine Beschäftigung mit Musik gilt jedoch als erwiesen. Der spätantike Musiktheoretiker *Gaudentius* (4. Jh.n.Chr.) berichtet über diese:

"[Pythagoras] spannte eine Saite über einen Kanon [ein gerades Holz] und teilte ihn in zwölf Teile. Dann ließ er zunächst eine Saite ertönen, darauf die Hälfte, das heißt sechs Teile, und er fand, dass die ganze Saite zu ihrer Hälfte

konsonant sei, und zwar nach dem Zusammenklang der Oktave. Nachdem er darauf erst die ganze Saite, dann Dreiviertel von ihr hatte erklingen lassen, erkannte er die Konsonanz der Quarte und analog für die Quinte."

Unter den Pythagoräern versteht man Personen, die die von Pythagoras initiierten Gedanken weiterverfolgten. Von diesen sind uns in den sogenannten Fragmenten verbürgte Aussagen überliefert. So findet man in Fragment 6 des Pythagoäers *Philolaos von Kroton* (-), einem Zeitgenossen des *Sokrates*, folgende Aussage:

"... Der Harmonie (Oktave) Größe umfasst die Quarte und die Quinte. Die Quinte ist aber um einen Ganzton größer als Quarte. ... Die Quarte aber hat das Verhältnis $3:4$ *, die Quinte* $2:3$ *, die Oktave* $1:2$ *. So besteht die Oktave aus fünf Ganztönen und zwei Halbtönen, die Quinte aus drei Ganztönen und einem Halbton, die Quarte aus zwei Ganztönen und einem Halbton."*

Besondere Verehrung genoss bei den Pythagoräern die Zahl Zehn, die als Summe der ersten vier natürlichen Zahlen $(1+2+3+4=10)$ eine Vierergruppe (*tetraktys*) bildet und als Punktmenge in einem gleichseitigen Dreieck dargestellt wurde.

Mit der *tetraktys* ist die Vorstellung der beschriebenen Harmonie eng verknüpft: Das Verhältnis der Anzahlen der Steine (*calculi*) zweier aufeinanderfolgender Lagen beschreiben der Reihe nach Oktave $(1:2)$, Quinte $(2:3)$ und Quarte $(3:4)$.

Im Fragmente 2 der Harmonik des Pythagoräers *Archytas von Tarent* (428 - 365 v.Chr.), einem Schüler von *Philolaos* und Freund von *Platon*, werden die drei behandelten Mittel erstmals explizit beschrieben. Und zwar vermittels der für die Griechen typischen Proportionen, d.h. durch Verhältnisse natürlicher Zahlen.

"Es gibt aber drei Proportionen in der Musik: einmal die arithmetische, zweitens die geometrische, drittens die entgegengesetzte, sogenannte harmonische.

Die arithmetische, wenn drei Zahlbegriffe analog folgende Differenz aufweisen: um wieviel der erste den zweiten übertrifft, um soviel übertrifft der zweite den dritten. Und bei dieser Analogie trifft es sich, dass das Verhältnis der größeren Zahlbegriffe kleiner, das der kleineren größer ist."

"Die geometrische: wenn sich der erste Begriff zum zweiten, wie der

zweite zum dritten verhält. Die größeren von ihnen haben das gleiche Verhältnis wie die geringeren."

"Die entgegengesetzte, sogenannte harmonische Proportion, wenn sich die Begriffe so verhalten: um den wievielten Teil der eigenen Größe der erste Begriff den zweiten übertrifft, um diesen Teil des dritten übertrifft der Mittelbegriff den zweiten. Bei dieser Analogie ist das Verhältnis der größeren Begriffe größer, das der kleineren kleiner."

Anmerkung: Seien $0 < c < b_1, b_2, b_3 < a$. Dann lautet die Aussage über das arithmetische Mittel b_1 in moderner Notation

$$a - b_1 \;=\; b_1 - c \quad \Rightarrow \quad a : b_1 \;<\; b_1 : c \;\; .$$

Wie man sich mit Hilfe der Bruchrechnung - die den Griechen freilich nicht zur Verfügung stand - leicht überzeugen kann, ergibt sich das arithmetische Mittel aus der Beziehung $a - b_1 = b_1 - c$ (nach Addition von $b_1 + c$ und anschließender Division durch 2)

$$b_1 = \frac{a+c}{2} \, .$$

Multiplikation der Bruchdarstellung $\frac{a}{b_1} < \frac{b_1}{c}$ der Ungleichung $a : b_1 < b_1 : c$ mit $b_1 \cdot c$ und anschließendes Wurzelziehen ergibt

$$b_2 = \sqrt{a \cdot c} < b_1 \, ,$$

wobei sich die erste Identität aus der Definition

$$a : b_2 \;=\; b_2 : c$$

des geometrische Mittels b_2 ergibt. D**ie Definition des harmonischen Mittels** b_3 lautet in moderner Notation

$$(a - b_3) : a \;=\; (b_3 - c) : c \quad \Rightarrow \quad a : b_3 \;>\; b_3 : c \;\; .$$

Aus der entsprechenden Bruchdarstellung $\frac{a-b_3}{a} = \frac{b_3-c}{c}$ ergibt sich durch einfache Umformung

$$b_3 = \frac{2ac}{a+c} = (\frac{\frac{1}{a}+\frac{1}{c}}{2})^{-1} \, .$$

Die getroffene Aussage in Bruchdarstellung, nämlich $\frac{a}{b_3} > \frac{b_3}{c}$, ist schließlich gleichbedeutend mit

$$\sqrt{a \cdot c} > b_3 = \frac{2ac}{a+c} \, (= \frac{ac}{b_1}) \, .$$

Zusammenfassend werden das arithmetische, geometrische und harmonische Mittel durch die Proportionen

$$\begin{array}{lcl} a - b_1 = b_1 - c & \Leftrightarrow & b_1 = \frac{a+c}{2} \\ a : b_2 = b_2 : c & \Leftrightarrow & b_2 = \sqrt{a \cdot c} \\ (a - b_3) : (b_3 - c) = a : c & \Leftrightarrow & b_3 = \frac{1}{\frac{1}{2}\left(\frac{1}{a}+\frac{1}{c}\right)} \end{array}$$

definiert. Das arithmetische und das harmonische Mittel entsprechen einander - wie bereits festgestellt - im folgenden Sinn

$$a : b_1 = b_3 : c \, .$$

Anmerkung zur Harmonik: Teilt man - wie von *Gautentius* beschrieben - den Kanon in $a = 12$ Teile und sei $c = 6$. Dann sind das arithmetische Mittel dieser beiden Zahlen $b_1 = 9$ und das harmonische Mittel $b_3 = 8$.

Wird die Länge der Saite im Verhältnis $c : a = 1 : 2$ geteilt, so ergibt das die Oktave.

Wird die Länge der Saite im Verhältnis $b_1 : a = 3 : 4$ geteilt, so ergibt das die Quarte.

Wird die Länge der Saite im Verhältnis $b_3 : a = 2 : 3$ geteilt, so ergibt das die Quinte.

Die oben angeführte Entsprechung des arithmetischen und des harmonischen Mittels, nämlich $b_1 : a = 3 : 4 = c : b_3$, entspricht also dem Zusammenhang zwischen Quarte und Quinte.

Hinsichtlich detaillierterer Ausführungen zu den Pythagoräern und zur Musik sei auf die fächerübergreifende Diplomarbeit [37] von Frau *Fritz* verwiesen.

1.5.6 Ausblick 2: Die Klasse der Komogorow-Nagumo-Mittel

Eine allgemeine Klasse von Mitteln, die die drei behandelten Mittel, nämlich das arithmetische Mittel (für $\alpha = 1$), das geometrische Mittel (für $\alpha = 0$) und das harmonisches Mittel (für $\alpha = -1$) als Spezialfälle enthält, wurde von *Kolmogorow* und *Nagumo*[23] vorgeschlagen. Sie ist - in ihrer positiv ho-

[23] *Andrej Nikolajewitsch Kolmogorow* (1903 – 1987), russischer Mathematiker, Begründer der modernen Wahrscheinlichkeitstheorie

Mitio Nagumo (1905 – ....), japanischer Mathematiker

mogenen Form - gegeben durch

$$M_\alpha(x_1,...,x_n) = \begin{cases} (\frac{1}{n}\sum\limits_{i=1}^{n} x_i^\alpha)^{1/\alpha} & \text{für } \alpha \in \mathbb{R}\setminus\{0\} \\ \sqrt[n]{\prod\limits_{i=1}^{n} x_i} & \text{für } \alpha = 0\,, \end{cases}$$

wobei folgende einschränkenden Voraussetzungen hinsichtlich der Beobachtungswerte $x_1,...,x_n \in \mathbb{R}$ zu treffen sind

$$\begin{array}{ll} x_1,...,x_n \geq 0 & \text{für } \alpha \in [0,\infty)\setminus\mathbb{N} \\ x_1,...,x_n > 0 & \text{für } \alpha \in (-\infty,0)\,. \end{array}$$

Anmerkung 1: Analog zum Stichprobenmittel lässt sich auch die vorliegende Verallgemeinerung mit Hilfe der relativen Häufigkeiten h_j der einzelnen Ausfälle ω_j, $j \geq 1$, der Variablen X darstellen:

$$M_\alpha(x_1,...,x_n) = \begin{cases} (\sum\limits_{j\geq 1} \omega_j^\alpha \cdot h_j)^{1/\alpha} & \text{für } \alpha \in \mathbb{R}\setminus\{0\} \\ \prod\limits_{j\geq 1} \omega_j^{h_j} & \text{für } \alpha = 0\,. \end{cases}$$

Die so definierte Klasse von Mitteln hat folgende Eigenschaften

- $M_\alpha(cx_1,...,cx_n) = cM_\alpha(x_1,...,x_n) \quad \forall c > 0$

  Diese Eigenschaft der positiven Homogenität ist charakteristisch für diese Klasse.

- $M_0(x_1,...,x_n) = \lim\limits_{\alpha\to 0} M_\alpha(x_1,...,x_n)$

- $\lim\limits_{\alpha\downarrow-\infty} M_\alpha(x_1,...,x_n) = \min(x_1,...,x_n)$ und $\lim\limits_{\alpha\uparrow\infty} M_\alpha(x_1,...,x_n) = \max(x_1,...,x_n)$

- $M_{\alpha_1}(x_1,...,x_n) \leq M_{\alpha_2}(x_1,...,x_n) \quad \forall \alpha_1 < \alpha_2\,,$

  wobei Gleichheit jeweils genau dann gilt, wenn $x_1 = ... = x_n$.

Definition: Der Spezialfall für $\alpha = 2$, nämlich

$$M_2(x_1, ..., x_n) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2},$$

heißt *quadratisches Mittel*.

Anmerkung 2: Die Gültigkeit von

$$M_1(x_1, ..., x_n) \leq M_2(x_1, ..., x_n),$$

mit Gleichheit genau dann, wenn $x_1 = ... = x_n$, ist eine unmittelbare Folgerung der nachstehenden Spezialfall des *Steinerschen Verschiebungssatzes* für $x = 0$, nämlich

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

1.6 ANALYSE ZWEIDIMENSIONALER DATENMENGEN: Lineare Regression und Korrelation

1.6.1 Aufgabenstellung und einführendes Beispiel

Gegeben seinen Wertepaare $\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^2$, derart, dass weder alle x_i noch alle y_i gleich sind. Gesucht ist eine Ausgleichsgerade, d.h. die Gleichung

$$y = kx + d \quad bzw. \quad x = k'y + d'$$

einer Geraden, die die gegebene "Punktwolke" möglichst gut beschreibt.

Beispiel 1: Case Study 10.2.1 aus [11]

Präzisierung der Zielfunktion: Eine formale Behandlung der Aufgabenstellung setzt voraus, dass wir zunächst zwei Fragen klären.

Frage 1: Welche Abstände von Punkten und Geraden werden gemessen?

Variante 1: Vertikale Abstände $y_i - (k\,x_i + d)\,,\ \ i \in \{1, ..., n\}$. Durch diese Wahl zeichnet man die Variable X als *unabhängige Variable* und die Variable Y als *abhängige Variable* aus.

Variante 1': Horizontale Abstände $x_i - (k'\,y_i + d')$, $i \in \{1, ..., n\}$. Durch diese Wahl zeichnet man die Variable Y als *unabhängige Variable* und die Variable X als *abhängige Variable* aus.

Variante 2: Normalabstände

Frage 2: Wie werden die Abstände von Punkten und Geraden gemessen?

Variante 1: Durch die Abstandsquadrate, etwa $(y_i - (kx_i + d))^2$, $i \in \{1, ..., n\}$.

Variante 2: Durch die Absolutbeträge der Abweichungen.

Als Zugeständnis an die mathematische Bequemlichkeit schließen wir die beiden Varianten 2 aus. Darüber hinaus beschränken wir uns auf die Behandlung der Vertikalabstände, da die der Horizontalabstände lediglich auf eine Vertauschung der beiden Variablen X und Y hinausläuft.

Demnach haben wir die Absicht, der Liste von Paaren $\{(x_1, y_1), ...,$

$(x_n, y_n)\} \subseteq \mathbb{R}^2$ eine Gerade $y = kx+d$ durch geeignete Wahl der Parameter k und d so anzupassen, dass die Zielfunktion

$$\sum_{i=1}^{n} (y_i - (kx_i + d))^2$$

minimal wird. Die so bestimmte Gerade heißt **Regressionsgerade**. Die verwendete Methode nennt man *die Methode der kleinsten Quadrate* [24], [25].

1.6.2 Herleitung der Gleichung der homogenen Regressionsgeraden

Gegeben seien Wertepaare $\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^2$ derart, dass weder alle x_i noch alle y_i gleich 0 sind.

Gesucht ist die sogenannte *homogene Regressionsgerade*, das ist jene Gerade $y = kx$ durch den Koordinatenursprung, die die gegebene "Punktwolke" im obigen Sinn möglichst gut beschreibt. Demgemäß werden wir den Anstieg k der Geraden so wählen, dass die Summe der Quadrate der sogenannten *Residuen*

$$y_i - kx_i \ , \quad i \in \{1, ..., n\} \ ,$$

möglichst klein ist.

1. Lösungsvariante: Durch Differenzieren

[24] Die Methode der kleinsten Quadrate wurde vom französischen Mathematiker *Adrien Marie Legendre* (1752 – 1833) in seiner Arbeit "*Methode de la Moindre Quaree*" 1805 publiziert. Unabhängig von diesem wurde sie auch von *Carl Friedrich Gauß* entwickelt und in der Astronomie und der Geodäsie angewandt.

[25] Die beiden Begriffe "*Regression*" und "*Korrelation*" wurden vom englischen Statistiker *Sir Francis Galton* (1822-1911) im Rahmen seiner Studien zur Vererbung geprägt. Das Wort "*Regression*" wurde von ihm im Zusammenhang mit "regression towards mediocrity", also dem Rückschritt zum Mittelmaß verwendet. Das Wort "Korrelation" könnte man mit "wechselseitiger Beziehung" übersetzen.

Galton war äußerst vielseitig: Er ist Schöpfer des nach ihm benannten *Galtonbretts*, der Urheber des Einsatzes von Fingerabdrücken zur Identifikation von Personen in der Kriminologie und der Entdecker der Antizyklonen in der Meteorologie.

Die Hälfte der Ableitung unserer Zielfunktion

$$f(k) = \sum_{i=1}^{n} (kx_i - y_i)^2$$

ist

$$\frac{1}{2} f'(k) = \sum_{i=1}^{n} x_i \cdot (kx_i - y_i) = k \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot y_i = \sum_{i=1}^{n} x_i^2 \cdot \left(k - \hat{k}\right)$$

mit $\hat{k} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sum_{i=1}^{n} x_i^2}$. Wegen

$$f'(k) \begin{cases} < 0 & \text{für} \quad k < \hat{k} \\ = 0 & \text{für} \quad k = \hat{k} \\ > 0 & \text{für} \quad k > \hat{k} \end{cases}$$

nimmt $f(k)$ für $k = \hat{k}$ sein Minimum an.

2. Lösungsvariante: Durch quadratisches Ergänzen

Ausquadrieren der Summanden der Funktion $f(k)$ und quadratisches Ergänzen in der folgenden Form

$$A \cdot k^2 + B \cdot k + C = A(k + \frac{B/2}{A})^2 + C(1 - \frac{(B/2)^2}{AC})$$

ergibt

$$\begin{aligned} f(k) &= \sum_{i=1}^{n} (k \cdot x_i - y_i)^2 \\ &= \sum_{i=1}^{n} \left(k^2 x_i^2 - 2k x_i y_i + y_i^2\right) \\ &= k^2 \sum_{i=1}^{n} x_i^2 - 2k \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \\ &= \sum_{i=1}^{n} x_i^2 (k - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2})^2 + \sum_{i=1}^{n} y_i^2 (1 - \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}) \,. \end{aligned}$$

Anmerkung 1: Sei nun $\hat{k}$ wie oben definiert und

$$\hat{r} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

der sogenannte *homogene Korrelationskoeffizient* (der Datenmenge $\{(x_1, y_1), ..., (x_n, y_n)\}$). Dann nimmt die Funktion $f(k)$ ihr Minimum

$$f(\hat{k}) = \sum_{i=1}^{n} (y_i - \hat{k} \cdot x_i)^2 = \sum_{i=1}^{n} y_i^2 \left(1 - \hat{r}^2\right)$$

offensichtlich genau dann an, wenn $k = \hat{k}$ ist. Die Größe $\hat{r}^2$ heißt *homogenes Bestimmtheitsmaß* und ist eine Maß für die Güte der Anpassung, denn $f(\hat{k})$ ist als Summe von Quadraten nichtnegativ und genau dann gleich Null, wenn $y_i = \hat{k} \cdot x_i \ \forall\, i \in \{1, ..., n\}$ gilt. Dies ist gleichbedeutend damit, dass

$$\hat{r}^2 \leq 1$$

ist, wobei $\hat{r}^2 = 1$ genau dann gilt, wenn alle Punkte der Punktwolke auf einer Geraden durch den Ursprung liegen.

Für den homogenen Korrelationskoeffizienten $\hat{r}$ gilt demnach

$$-1 \leq \hat{r} \leq 1\,,$$

wobei Gleichheit genau im eben angesprochenen Fall besteht. Sofern $\hat{r} > 0$ ist, ist der Anstieg $\hat{k}$ der Regressionsgeraden positiv; sofern $\hat{r} < 0$ ist, ist dieser negativ.

Anmerkung 2: Die Gültigkeit der Beziehung $\hat{r}^2 \leq 1$ ist, zusammen mit der zugehörigen Aussage über den Fall $\hat{r}^2 = 1$, gleichbedeutend mit der *Cauchy-Schwarz*schen Ungleichung

$$(\sum_{i=1}^{n} x_i y_i)^2 \leq \sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2\,,$$

für die bekanntlich Gleichheit genau dann zutrifft, wenn $y_i = \hat{k} \cdot x_i \ \forall\, i \in \{1, ..., n\}$ gilt.

1.6.3 Anwendungsbeispiele

Beispiel 2 [26]**: Zur Schätzung von** π

[26] Die Zahl π wird in den Anwendungen gelegentlich nach dem deutschen Mathematiker *Ludolf van Ceulen* $(1540 - 1610)$ Ludolfsche Zahl genannt.

Gegeben seien Messungen $(d_i, u_i),\ i \in \{1, ..., n\}$, von Durchmesser und Umfang zylindrischer Gefäße mit kreisförmiger Grundfläche. Seien $\frac{u_i}{d_i},\ i \in \{1, ..., n\}$, die Verhältnisse von Kreisumfang und Kreisdurchmesser. Dann ist - wegen des bekannten linearen Zusammenhangs $u = d \cdot \pi$ zwischen Kreisumfang und Kreisdurchmesser - das arithmetische Mittel

$$\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{u_i}{d_i}$$

ein naheliegender Schätzwert für π.

Wendet man hingegen die Methode der kleinsten Quadrate auf die Punktwolke $(d_i, u_i),\ i \in \{1, ..., n\}$, an, so erhält man nicht bloß den alternativen Schätzwert

$$\begin{aligned} \hat{\pi}_2 &= \frac{\sum_{i=1}^{n} d_i u_i}{\sum_{i=1}^{n} d_i^2} \\ &= \sum_{i=1}^{n} \frac{d_i^2}{\sum_{j=1}^{n} d_j^2} \frac{u_i}{d_i} \end{aligned}$$

für π, sondern zudem eine empirische Bestätigung des obigen linearen Zusammenhangs zwischen Kreisumfang und Kreisdurchmesser. Aus der zweiten Darstellung erkennt man, dass dieser Schätzwert ein mittels der Gewichte $\frac{d_i^2}{\sum_{j=1}^{n} d_j^2}$ gewichtetes Mittel der Quotienten $\frac{u_i}{d_i},\ i \in \{1, ..., n\}$, ist. Durch diese Gewichte werden Quotienten mit großem Durchmesser sehr viel stärker bewertet als solche mit kleinem. Dieser Sachverhalt ist verträglich mit der Vorstellung, dass Messergebnisse von größeren Gefäßen genauer sind als solche von kleinen.

Beispiel 3: Keplers drittes Gesetz [27]

Dieses lautet

[27] Keplers erstes Gesetz lautet bekanntlich: Die Planeten bewegen sich auf Ellipsen, in deren gemeinsamen Brennpunkt die Sonne steht.

In seinem 1604 veröffentlichten Werk *Paralipomena*, welches er auch einfach *Optica* nannte, beschäftigte sich *Johannes Kepler* $(1571 - 1630)$ auch mit Linsen, deren Krümmungskurven die Form von allgemeinen Kegelschnitten besitzen. In diesem Zusammmenhang prägte er den Begriff *Brennpunkt*.

Das Quadrat der Umlaufzeit eines Planeten ist proportional zum Kubus der großen Halbachse seiner Bahn. In Zeichen $U^2 = k \cdot a^3$

Die Daten der jeweiligen Planeten sind wie folgt

Planet	große Halbachse	siderische Umlaufzeit
Merkur	0.3870938	0.2408
Venus	0.7233276	0.6152
Erde	1.0009071	1
Mars	1.5237020	1.8808
Jupiter	5.2026255	11.8618
Saturn	9.5402393	29.4566
Uranus	19.2685300	84.0120
Neptun	30.2080026	164.7819
Pluto	39.8397880	247.6867

In moderner Darstellung ist die Beziehung zwischen großer Halbachse a und Umlaufszeit U

$$U = c \cdot a^{\frac{3}{2}} .$$

In dieser Form ist es jedoch für die Anwendung der homogenen linearen Regression nicht brauchbar. Um letztere anwenden zu können, ist es zweckmäßig, auf die ursprügliche Formulierung des Gesetzes zurückzugreifen und die Größen a^3 und U^2 gegeneinander aufzutragen.

Der Schätzwert für die Konstante k ist

$$\hat{k} = \frac{\sum_{i=1}^{9} a_i^3 u_i^2}{\sum_{i=1}^{9} a_i^6} = .97271 .$$

Das Bestimmtheitsmaß ist

$$\hat{r}^2 = \frac{(\sum_{i=1}^{9} a_i^3 u_i^2)^2}{\sum_{i=1}^{9} a_i^6 \cdot \sum_{i=1}^{9} u_i^4} = .99997 .$$

Damit ist die Gültigkeit des Gesetzes denkbar gut bestätigt.

1.6.4 Herleitung der Gleichung der allgemeinen Regressionsgeraden

Gegeben seien Wertepaare $\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^2$, derart, dass weder alle x_i noch alle y_i gleich sind.

Gesucht ist die Regressionsgerade, d.h. die Gleichung $y = kx + d$ jener Geraden, die diese "Punktwolke" im obigen Sinn möglichst gut beschreibt. Demgemäß haben wir die Parameter k und d so zu wählen, dass die Summe der Quadrate der Residuen

$$y_i - (kx_i + d) \ , \quad i \in \{1, ..., n\}$$

möglichst klein ist. Der erste Schritt, die Zielfunktion

$$f(d, k) = \sum_{i=1}^{n} (y_i - k\, x_i - d)^2$$

zu minimieren, besteht darin, den Spezialfall

$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} (z_i - \bar{z}_n)^2 + n \cdot \bar{z}_n^2$$

des Steinerschen Verschiebungssatzes auf die Größen

$$z_i = y_i - k\, x_i - d$$

anzuwenden und $\bar{z}_n = \bar{y}_n - k\bar{x}_n - d$ zu berücksichtigen. Auf diese Weise erhält man

$$\sum_{i=1}^{n} (y_i - k\, x_i - d)^2 = \sum_{i=1}^{n} (y_i - \bar{y}_n - k\,(\, x_i - \bar{x}_n))^2 + n\,(\bar{y}_n - k\bar{x}_n - d)^2 \,.$$

Für den - zunächst noch vom Anstieg k abhängigen - Ordinatenabschnitt $\tilde{d} = \tilde{d}(k) = \bar{y}_n - k\bar{x}_n$ gilt demnach

$$f(d, k) = f(\tilde{d}, k) + n(d - \tilde{d})^2 \geq f(\tilde{d}, k)$$

mit Gleichheit genau dann, wenn $d = \tilde{d}$ ist.

Die neue Zielfunktion $g(k) = f(\tilde{d}(k), k) = \sum_{i=1}^{n} (y_i - \bar{y}_n - k\,(\, x_i - \bar{x}_n))^2$ zu minimieren, bedeutet jedoch, der Punktwolke $\{(x_i', y_i') = (x_i -$

$\bar{x}_n, y_i - \bar{y}_n),\ i \in \{1, ..., n\}\}$ eine Gerade durch den Koordinatenursprung anzupassen. Somit können wir das Resultat aus dem vorangehenden Abschnitt auf die transformierte Punktwolke anwenden.

Mit Hilfe der Stichprobenvarianzen der x- und y-Werte und der Kovarianz der Stichprobe

$$\begin{aligned} s_x^2 &= \tfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \,, \\ s_y^2 &= \tfrac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 \quad \text{und} \\ s_{xy} &= \tfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n) \cdot (y_i - \bar{y}_n) \end{aligned}$$

lassen sich der entsprechende Anstieg der Regressionsgeraden und der entsprechende Korrelationskoeffizient folgendermaßen ausdrücken

$$\tilde{k} = \frac{s_{xy}}{s_x^2} \quad \text{und} \quad \tilde{r} = \frac{s_{xy}}{s_x \cdot s_y} \,.$$

Für diese Größen gilt

$$\begin{aligned} g(k) &= \sum_{i=1}^{n} (y_i - \bar{y}_n - k\,(\,x_i - \bar{x}_n))^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x}_n)^2\,(k - \tilde{k})^2 + \sum_{i=1}^{n} (y_i - \bar{y}_n)^2\,(1 - \tilde{r}^2) \\ &\geq \sum_{i=1}^{n} (y_i - \bar{y}_n)^2\,(1 - \tilde{r}^2) \end{aligned}$$

mit Gleichheit genau dann, wenn $k = \tilde{k}$ ist.

Die Gleichung der *Regressionsgeraden der Variablen Y bezüglich der unabhängigen Variablen X* ist daher

$$y = \bar{y}_n + \frac{s_{xy}}{s_x^2}(\,x - \,\bar{x}_n) \,.$$

Der zugehörige Ordinatenabschnitt ist somit $\tilde{d}(\tilde{k}) = \bar{y}_n - \frac{s_{xy}}{s_x^2}\bar{x}_n$.

Anmerkung: Die Größe $\tilde{r}^2$ lässt sich mit Hilfe der durch das lineare Modell für die gegebenen Werte x_i der unabhängigen Variablen X für die abhängige Variable Y vorhergesagten Werte

$$\tilde{y}_i = \bar{y}_n + \frac{s_{xy}}{s_x^2}(\,x_i - \bar{x}_n) \,,\ i \in \{1, ..., n\} \,,$$

gemäß

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\tilde{y}_i-\bar{y}_n\right)^2=\frac{s_{xy}^2}{\left(s_x^2\right)^2}\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2=\frac{s_{xy}^2}{s_x^2}$$

folgendermaßen ausdrücken

$$\tilde{r}^2=\frac{\sum_{i=1}^{n}\left(\tilde{y}_i-\bar{y}_n\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_n\right)^2}\,.$$

Dies ermöglicht die

Interpretation: $\tilde{r}^2$ ist der Anteil der Variation in der abhängigen Variablen, der durch das lineare Modell erklärt wird.

Variante 1': Indem man die Variablen X und Y vertauscht, erhält man die Gleichung der *Regressionsgeraden von* X *bezüglich der unabhängigen Variablen* Y, nämlich

$$x=\bar{x}_n+\frac{s_{yx}}{s_y^2}\left(y-\bar{y}_n\right)=\bar{x}_n+\frac{s_{xy}}{s_y^2}\left(y-\bar{y}_n\right)\,.$$

In ihrer üblichen Form

$$y_{X(Y)}(x)=\bar{y}_n+\frac{s_y^2}{s_{xy}}(x-\bar{x}_n)$$

besitzt diese den Anstieg $\tilde{k}_{X(Y)}=\frac{s_y^2}{s_{xy}}$, während die *Regressionsgerade von* Y *bezüglich der unabhängigen Variablen* X den Anstieg $\tilde{k}_{Y(X)}=\frac{s_{xy}}{s_x^2}$ besitzt.

Für das *Bestimmtheitsmaß* $\tilde{r}^2=\frac{s_{xy}^2}{s_x^2\cdot s_y^2}$ der Korrelation, welches - wie im homogenen Fall - ein Maß für die Stärke des linearen Zusammenhanges von X und Y ist, gilt im Hinblick auf die Anstiege $\tilde{k}_{Y(X)}$ und $\tilde{k}_{X(Y)}$ folgende Aussage

$$\tilde{r}^2\leq 1\Longleftrightarrow\begin{cases}\frac{s_{xy}}{s_x^2} & \leq & \frac{s_y^2}{s_{xy}} & \text{für} & s_{xy} & > & 0\\ y=\bar{y}_n & \text{und} & x=\bar{x}_n & \text{für} & s_{xy} & = & 0\\ \frac{s_{xy}}{s_x^2} & \geq & \frac{s_y^2}{s_{xy}} & \text{für} & s_{xy} & < & 0\,.\end{cases}$$

Der Fall $\tilde{r}^2=1$ ist mit jeder der beiden folgenden Aussagen gleichbedeutend.

- Alle Punkte (x_i, y_i) liegen auf einer Geraden.
- Die beiden Regressionsgeraden fallen zusammen.

Der *Pearsonsche Korrelations-Koeffizient* $\tilde{r} = \frac{s_{xy}}{s_x \cdot s_y}$ ist naturgemäß ein Wert aus $[-1, 1]$. Er bleibt selbstverständlich unverändert, wenn die Maßeinheiten in den beiden Variablen geändert werden. Positive Korrelation liegt vor, wenn zu großen Werten der einen Variablen auch große Werte der anderen Variablen gehören; negative Korrelation, wenn zu großen Werten der einen Variablen kleine Werte der anderen Variablen gehören. Im Fall $\tilde{r} = 0$ nennt man die beiden Variablen *unkorreliert*. Das bedeutet, dass zwischen den beiden Variablen kein linearer Zusammenhang existiert. Dabei liegt jedoch die Betonung auf dem Wort **linear**, denn $\tilde{r} = 0$ schließt einen anderen funktionellen Zusammenhang nicht aus! Ist nämlich ein Zusammenhang durch eine gerade Funktion[28] gegeben, so kann ohne weiters $\tilde{r} = 0$ gelten, wie die folgende Beispielklasse zeigt.

Beispielklasse mit $\tilde{r} = 0$: Jede Punktwolke

$$\{(x_z, y_z),\ z \in \{-m, -m+1, ..., 0, ..., m-1, m\}\},$$

welche spiegelsymmetrisch zur y-Achse liegt, hat den Korrelationskoeffizient $\tilde{r} = 0$.

Ist nämlich mit den Werten $0 = \omega_0 < \omega_1 < ... < \omega_m$ und einer beliebigen Funktion $f : [0, \infty) \to \mathbb{R}$ auf folgende Weise eine Punktwolke $\{(x_z, y_z),\ z \in \{-m, -m+1, ..., 0, ..., m-1, m\}\}$ verknüpft: $(x_i, y_i) = (\omega_i, f(\omega_i)),\ i \in \{0, 1, ..., m\}$ und $(x_{-i}, y_{-i}) = (-\omega_i, f(\omega_i)),\ i \in \{1, ..., m\}$, dann gilt wegen $\bar{x}_{2m+1} = 0$, $\bar{y}_{2m+1} \in \mathbb{R}$ und $\omega_0 = 0$

$$\begin{aligned} s_{xy} &= \frac{1}{2m}\Big(\sum_{z=-m}^{m} x_z \cdot y_z - \bar{x}_{2m+1} \cdot \bar{y}_{2m+1}\Big) \\ &= \frac{1}{2m}\Big(\omega_0 \cdot f(\omega_0) + \sum_{i=1}^{m} \omega_i \cdot f(\omega_i) + \sum_{i=1}^{m} (-\omega_i) \cdot f(\omega_i)\Big) \\ &= \frac{1}{2m}\Big(\sum_{i=1}^{m} (\omega_i - \omega_i) \cdot f(\omega_i)\Big) \\ &= 0\,. \end{aligned}$$

[28] Eine solche ist durch die Eigenschaft $f(-x) = f(x)$ charakterisiert. Typische Beispiele sind die Polynome $f(x) = x^{2n},\ n \in \mathbb{N}$.

Kapitel 2

BEURTEILENDE STATISTIK

Die beurteilende Statistik beruht auf wahrscheinlichkeitstheoretischen Modellen. Sie umfasst folgende zwei Teilgebiete. Das

- Schätzen von Parametern, wobei man "Punktschätzer" und "Intervallschätzer" unterscheidet. Gebräuchlichere Bezeichnungen für letztere sind "Konfidenz- oder Vertrauensintervalle". Und das
- Testen von Hypothesen.

Wir geben zunächst eine exemplarische Einführung in das Schätzen von Parametern anhand des Schätzens des Umfangs einer durchnummerierten Grundgesamtheit.

2.1 SCHÄTZEN VON PARAMETERN

2.1.1 Exemplarische Einführung

Den Umfang einer durchnummerierten Grundgesamtheit zu schätzen, ist ein eindrucksvolles Beispiel für eine Situation, bei welcher eine Vielzahl vernünftiger Schätzer zur Verfügung steht. Aus dieser ist ein Schätzverfahren auszuwählen, welches "möglichst genau" ist. Wir gehen dabei von folgendem Anwendungsbeispiel aus.

Anwendungsbeispiel: Der Salzburger Jedermannlauf

Der "Salzburger Jedermannlauf" ist eine Breitensportveranstaltung, welche seit Jahren am Nationalfeiertag in der Landeshauptstadt organisiert wird.

Die Teilnehmer erhalten Startnummern $1, 2, ..., N$, wobei N die uns unbekannte Teilnehmerzahl ist. Nach Beendigung des Laufes erfolgt eine Verlosung von 25 Preisen. Dafür werden 25 Startnummern zufällig und ohne Zurücklegen aus der Menge der Startnummern jener Läufer gezogen, welche den Lauf beenden.

Von der Veranstaltung im Jahr 1994 wurden uns folgende 24 Nummern übermittelt.

616,	1436,	737,	11,	1133,	1003,	705,	139,	614,	665,	1057,	1076,
1070,	1075,	1382,	1384,	1394,	776,	650,	8,	688,	1065,	269,	195.

Versuchen Sie aus dieser Information die Teilnehmerzahl zu schätzen, indem Sie von der selbstverständlich nicht ganz realistischen Annahme ausgehen, dass die Startnummern lückenlos vergeben werden und alle Personen, welche eine Startnummer besitzen, auch den Lauf beenden und an der Verlosung teilnehmen.

Die zugehörige Idealisierung lässt sich durch folgendes Urnenmodell beschreiben.

Urnenmodell: Gegeben sei eine Urne mit N Jetons, welche von 1 bis N durchnummeriert sind. Der Parameter N sei unbekannt. n $(\leq N)$ Jetons werden zufällig und ohne (bzw. mit) Zurücklegen gezogen. $x_1, ..., x_n$ seien die Nummern der gezogenen Jetons. N ist mit Hilfe der beobachteten Werte $x_1, ..., x_n$ zu schätzen.

A) Punktschätzer beim Ziehen ohne Zurücklegen

Beispiel: Jemand wählt eine Zahl $N \in \{10, ..., 20\}$ und bestückt die Urne mit N Jetons. Wir sollen aufgrund einer Stichprobe vom Umfang $n = 5$ die gewählte Zahl N schätzen. Die Nummern der gezogenen Jetons sind

$$x_1 = ... \ , \ x_2 = ... \ , \ x_3 = ... \ , \ x_4 = ... \ , \ x_5 = ... \ .$$

Wir haben folgende Schätzer, das sind Vorschriften, den Schätzwert zu bestimmen, erarbeitet.

1. Der Mittelschätzer

Aufgrund des Gesetzes der großen Zahlen approximiert das Stichprobenmittel

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

für hinreichend große n den Erwartungswert $\mu = E(X_i) = \frac{N+1}{2}$ der Zufallsvariablen. D.h. es ist $\bar{X}_n \simeq \frac{N+1}{2}$ und daher $N \simeq 2\bar{X}_n - 1$. Diese für die Momentenmethode, die auf den englischen Statistiker *Karl Pearson* zurückgeht, typische Überlegung motiviert den Mittelschätzer

$$\bar{N} = 2\bar{X}_n - 1 \,.$$

Der entsprechende Schätzwert ist also $2 \cdot \frac{}{5} - 1 = ...$.

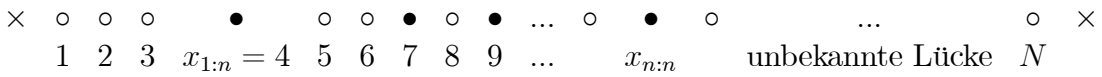
Im folgenden bezeichnen $X_{1:n} < X_{2:n} < ... < X_{n:n}$ die der Größe nach geordneten Stichprobenwerte. $X_{i:n}$ heißt dabei die *i-te Ordnungsstatistik*, $i \in \{1, ..., n\}$.

2. Der Medianschätzer: Der Einfachheit halber sei $n = 2m+1$ mit $m \in \mathbb{N}_0$. Ersetzt man nun beim Mittelschätzer das Stichprobenmittel $\bar{X}_n$ durch den Stichprobenmedian $X_{m+1:2m+1}$, so erhält man den Medianschätzer

$$\hat{N}_{m+1:2m+1} = 2 \cdot X_{m+1:2m+1} - 1 \,.$$

Der zugehörige Schätzwert ist also: $2 \cdot ... - 1 = ...$.

3. Die Lückenmethode [1]:



Motivation: Wegen $X_{n:n} \leq N$ schätzt das Stichprobenmaximum den Parameter N in aller Regel zu kurz. Um diesen Bias zu korrigieren, ersetzen wir die Länge $N - X_{n:n}$ der unbekannten "Lücke" durch

[1] In der nachstehenden Abbildung sind die beobachteten Werte durch • gekennzeichnet. So ist beispielsweise $x_{1:n} = 4$.

a) die bekannte Länge $X_{1:n} - 1$ der ersten Lücke und erhalten somit gemäß $N = X_{n:n} + N - X_{n:n} \simeq X_{n:n} + X_{1:n} - 1$ den Lückenschätzer

$$\tilde{N}_1 = X_{n:n} + X_{1:n} - 1\,.$$

Der entsprechende Schätzwert ist also $... + ... - 1 = ...$.

b) die durchschnittliche Länge

$$\frac{1}{n}[X_{1:n} - 1 + \sum_{i=2}^{n}(X_{i:n} - X_{i-1:n} - 1)] = \frac{1}{n}X_{n:n} - 1$$

der bekannten Lücken und erhalten somit gemäß $N = X_{n:n} + N - X_{n:n} \simeq X_{n:n} + \frac{1}{n}X_{n:n} - 1 = \frac{n+1}{n}X_{n:n} - 1$ den Maximumschätzer

$$\hat{N}_{n:n} = \frac{n+1}{n}X_{n:n} - 1\,.$$

Der entsprechende Schätzwert ist also $\frac{6}{5} \cdot ... - 1 = ...$.

Anmerkung 1: Der Medianschätzer und der Maximumschätzer gehören der folgenden Familie von Schätzern an

$$\hat{N}_{i:n} = \frac{n+1}{i} \cdot X_{i:n} - 1\ ,\ i \in \{1, ..., n\}\,.$$

Anmerkung 2: Mit der Vorstellung, es werden der Urne n Jetons mit einem Griff entnommen, ergibt sich die Verteilung der i-ten Ordnungsstatistik $X_{i:n}$

$$P\left(X_{i:n} = k\right) = \frac{\binom{k-1}{i-1}\binom{1}{1}\binom{N-k}{n-i}}{\binom{N}{n}}\,, \quad k \in \{i, ..., i+N-n\}\,,$$

also die Hypergeometrische Wartezeitverteilung mit den Parametern $i\,; N$ und n . Deren Erwartungswert und Varianz haben bekanntlich folgende Gestalt

$$\begin{array}{lll} (1) \quad E\left(X_{i:n}\right) & = & i \cdot \frac{N+1}{n+1} \\ (2) \quad V\left(X_{i:n}\right) & = & \frac{i(n+1-i)}{n+1} \cdot \frac{N+1}{n+1} \cdot \frac{N-n}{n+2} \end{array}$$

Folgerungen: $X_{i:n}$ schätzt offensichtlich zu kurz. Aus (1) lässt sich aber leicht ein "unverfälschter" Schätzer für N ermitteln. Indem man nämlich $X_{i:n}$ mit dem Faktor $\frac{n+1}{i}$ multipliziert und schließlich 1 abzieht,

erhält man den Schätzer $\hat{N}_{i:n}$ der oben angegebenen Familie. Dessen Erwartungswert ist tatsächlich

$$\begin{aligned} E(\frac{n+1}{i}X_{i:n}-1) &= \frac{n+1}{i}\cdot E\left(X_{i:n}\right)-1 \\ &= \frac{n+1}{i}\cdot i\cdot\frac{N+1}{n+1}-1 \\ &= N\,. \end{aligned}$$

Von allen Schätzern der Familie ist der Maximumschätzer $\hat{N}_{n:n}$ derjenige mit der kleinsten Varianz. Aufgrund von (2) gilt nämlich

$$\begin{aligned} V(\frac{n+1}{i}X_{i:n}-1) &= (\frac{n+1}{i})^2 V\left(X_{i:n}\right) \\ &= \frac{(n+1)^2}{i^2}\cdot\frac{i\,(n+1-i)}{n+1}\cdot\frac{N+1}{n+1}\cdot\frac{N-n}{n+2} \\ &= (\frac{n+1}{i}-1)\cdot(N+1)\cdot\frac{N-n}{n+2} \\ &\geq (\frac{n+1}{n}-1)\cdot(N+1)\cdot\frac{N-n}{n+2} \\ &= \frac{N+1}{n}\cdot\frac{N-n}{n+2} \\ &= V(\frac{n+1}{n}X_{n:n}-1)\,. \end{aligned}$$

Mit Hilfe von tiefgründigen Methoden lässt sich übrigens zeigen, dass $\hat{N}_{n:n}$ unter allen erdenklichen erwartungstreuen Schätzern jener mit kleinster Varianz ist.

Fallstudie: Anwendung der Statistik für Spionagezwecke[2]

Mindestens einmal in der Geschichte der Statistik wurde ein statistisches Verfahren für Spionagezwecke verwendet; und zwar im 2. Weltkrieg von den Alliierten zur Schätzung der deutschen Waffenproduktion.

Jedes deutsche Kriegsgerät, ob V-2-Rakete, Panzer oder Autoreifen, war während des Produktionsprozesses mit einer Seriennummer versehen worden. War beispielsweise die Gesamtanzahl der bis zu einem bestimmten Zeitpunkt hergestellten Mark-I-Panzer gleich N, so besaß jeder dieser Panzer eine

[2] Eine freie Übersetzung von Case Study 5.5.1 in [11]

Seriennummer zwischen 1 und N. Nun wurden den Alliierten im Verlauf der Kriegshandlungen einige Nummern

$$1 \leq X_{1:n} < ... < X_{n:n} \leq N$$

bekannt (entweder dadurch, dass Panzer zerstört oder erbeutet oder dass einschlägige Dokumente erbeutet wurden). Das Verfahren, das ursprünglich zur Schätzung von N angewendet wurde, war der Lückenschätzer

$$\begin{aligned} \tilde{N}_2 &= X_{n:n} + \frac{1}{n-1}\sum_{i=2}^{n}(X_{i:n} - X_{i-1:n} - 1) = X_{n:n} + \frac{1}{n-1}(X_{n:n} - X_{1:n}) - 1 \\ &= \frac{n}{n-1}X_{n:n} - \frac{1}{n-1}X_{1:n} - 1\,. \end{aligned}$$

Nach Ende des Krieges, als die Dokumente des deutschen Kriegsministeriums zugänglich wurden, fand man, dass die Schätzwerte für die Waffenproduktion, die auf statistischen Methoden beruhten, weit zuverlässiger waren, als jene, denen andere Informationen zugrunde lagen. So lag beispielsweise der mittels Seriennummern-Schätzer erhaltene Schätzwert $3\,400$ für die bis 1942 erzeugten deutschen Panzer dem tatsächlichen Wert sehr nahe. Der "offizielle" Schätzwert der Alliierten, der auf Informationen beruhte, welche vom Geheimdienst und von Spionageaktivitäten stammten, war hingegen mit $18\,000$ weit überhöht. Fehler dieser Größenordnung, vielfach in der offenbar sehr effektiven "Nazi-Propaganda" begründet, waren nicht ungewöhlich. Lediglich das sehr objektive Seriennummern-Verfahren war gegenüber derartig verfälschenden Einflüssen unempfindlich!

B) Intervallschätzer auf der Basis des Stichprobenmaximums $X_{n:n}$ für das Ziehen mit und ohne Zurücklegen. Als Vorüberlegung stellen wir zunächst die Verteilungsfunktion des Stichprobenmaximums $X_{n:n}$ bereit:

$$\begin{aligned} P(X_{n:n} \leq k) &= P(X_1 \leq k, ..., X_n \leq k) \\ &= \begin{cases} \frac{k^n}{N^n} & \text{mit} \\ \text{bei Ziehen} & \text{Zurücklegen} \\ \frac{k(k-1)\cdot...\cdot(k-n+1)}{N(N-1)\cdot...\cdot(N-n+1)} & \text{ohne} \end{cases} \\ &\leq (\frac{k}{N})^n, \quad k \in \{1, ..., N\}\,. \end{aligned}$$

Weiters sei $\lceil x \rceil$ die nächstgrößte ganze Zahl von x, also jener Wert $z \in \mathbb{Z} : x \leq z < x+1$. Demzufolge ist $x \leq \lceil x \rceil < x+1$. Für $c \in (0,1)$ gilt

somit

$$\begin{aligned} P\,(X_{n:n} < c \cdot N) &= P\,(X_{n:n} \leq \lceil c \cdot N \rceil - 1) \\ &\leq (\frac{\lceil c \cdot N \rceil - 1}{N})^n \\ &< (\frac{c \cdot N}{N})^n = c^n\,. \end{aligned}$$

Also gilt wegen $X_{n:n} \leq N$ und der entscheidenden Identität

$$\{c \cdot N \leq X_{n:n} \leq N\} = \{X_{n:n} \leq N \leq X_{n:n}/c\}$$

die Beziehung

$$P\,(X_{n:n} \leq N \leq X_{n:n}/c) = P\,(X_{n:n} \geq c \cdot N) > 1 - c^n\,.$$

Sei nun $0 < \alpha \ll 1$ (z.B. $\alpha = 0.1\,,\ \alpha = 0.05,\ \alpha = 0.01$) und $c = \sqrt[n]{\alpha}\,.$ Dann gilt

$$P\left(X_{n:n} \leq N \leq X_{n:n}/\sqrt[n]{\alpha}\right) > 1 - \alpha\,.$$

Also ist

$$\left[X_{n:n}, X_{n:n}/\sqrt[n]{\alpha}\right]$$

ein Intervall von zufälliger Lage (und Länge), das den unbekannten Parameter N mit einer Wahrscheinlichkeit von mindestens $1 - \alpha$ überdeckt.

Man sagt, $[X_{n:n}, X_{n:n}/\sqrt[n]{\alpha}]$ ist ein *Konfidenz- oder Vertrauensintervall* von mindestens $1 - \alpha$ %iger statistischer Sicherheit[3].

Anmerkung 3: Zur Konstruktion eines Konfidenzintervalls für den Parameter N kann anstelle des Stichprobenmaximums $X_{n:n}$ beispielsweise auch das Stichprobenminimum $X_{1:n}$ herangezogen werden. Auf dessen Basis erhält man für den Fall des Ziehens mit Zurücklegen analog zur obigen Vorgangsweise das Konfidenzintervall

$$[X_{1:n}, X_{1:n}/(1 - \sqrt[n]{1-\alpha}))\,.$$

[3] Als systematisch angewandtes statistisches Verfahren wurde das Konzept des Konfidenzintervalls vom polnischen Statistiker *Jerzy Neyman* (1894 – 1981) zu Beginn der 1930-er Jahre entwickelt. Es ist eine Frucht der etwa 10-jährigen intensiven Zusammenarbeit mit dem britischen Statistiker *Egon S. Pearson* (1895 – 1980) über die Theorie statistischer Tests.

Während die Länge des Konfidenzintervalls $[X_{n:n}, X_{n:n}/\sqrt[n]{\alpha}]$ mit wachsendem n gegen 0 geht, geht die des vorliegenden Konfidenzintervalls mit wachsendem n gegen ∞. Dies ist ein deutlicher Hinweis darauf, dass es ratsam ist, zur Konstruktion von Konfidenzintervallen Zufallsgrößen mit möglichst geringer Variabilität heranzuziehen.

2.1.2 Ausblick: Zur Korrektur einer Verfälschung bei Punktschätzern

In Abschnitt 1.4.1 und in den vorangehenden Ausführungen über Punktschätzer hatten die naheliegenden Kandidaten für Schätzer der entsprechenden Parameter, nämlich

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{für} \quad \sigma^2 \qquad \text{und} \qquad X_{n:n} \quad \text{für} \quad N$$

die Tendenz, den jeweiligen Parameter zu unterschätzen. In den beiden vorliegenden Fällen liegt also gemäß der nachstehenden schematischen Darstellung eine negative Verfälschung vor:

Schätzung	=	Wert des Parameters	+	Verfälschung	+	Zufallsschwankung

In beiden Fällen war es jedoch möglich, diese Verfälschung durch eine geeignete Modifikation dieser Kandidaten zu beheben. Das Resultat sind die entsprechenden unverfälschten Schätzer [4]

$$S_n^2 = \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{für} \quad \sigma^2 \quad \text{und} \quad \frac{n+1}{n} \times X_{n:n} - 1 \quad \text{für} \quad N\,.$$

[4] Der Begriffe *Verfälschung* (*bias*) stammt - wie viele andere Begriffe beim Schätzen von Parametern - vom englischen Statistiker *Sir Ronald Aylmer Fisher* $(1890-1962)$. Die systematische Untersuchung der sogenannten *unverfälschten oder erwartungstreuen Schätzer* (*unbiased estimators*) wurde in der Folge insbesondere vom schwedischen Mathematiker *Carl H. Cramér* $(1893-1985)$ vorangetrieben.

Es ist eine alte Tradition [5], das Schätzen von Parametern mit dem Zielschießen zu vergleichen, wobei folgende Entsprechungen gelten [6]:

Schätzen	**Zielschießen**
Parameter	Das Zentrum einer Zielscheibe
Verfälschung (*bias*)	Systemastische Abweichung
unverfälscht (*unbiased*)	keine systematische Abweichung
kleine Varianz	hohe Präzision

Der grundsätzliche Unterschied besteht jedoch darin, dass die Zielscheibe beim Zielschießen sichtbar und daher bekannt, der Parameter beim Schätzen jedoch unbekannt ist. Daher entspricht ein Schätzer eher einer intelligenten Lenkwaffe, welche sich ihr Ziel selbst sucht, als der Kugel eines Gewehrs.

Schießt man mit einem Gewehr auf eine Zielscheibe, so wird die systematische Abweichung durch den Höhenverlust der Kugel infolge a) der Schwerkraft und b) des Luftwiderstandes hervorgerufen. Im folgenden sei in knapper Form auf die Korrektur des durch die Schwerkraft hervorgerufenen Höhenverlusts eingegangen.

Angenommen,

die Kugel verlässt den Gewehrlauf mit einer Geschwindigkeit von v Metern pro Sekunde,

die Zielscheibe ist vom Ende des Gewehrlaufes x_0 Meter entfernt und

das Ende des Gewehrlaufes und die Mitte der Zielscheibe befinden sich auf gleicher Höhe, nämlich auf der Höhe von y_0 Metern.

Ferner sei $g \cong 9.81\ m/\sec^2$ die Erdbeschleunigung.

Bezeichnen nun

α den, in Bogenlänge angegebenen, *Anstell-* oder *Abgangswinkel* [7] des Gewehrs,

---

[5] Diese geht vermutlich auf den englischen Astronomen *John Herschel* $(1792 - 1872)$ zurück, der sich in einem Artikel aus dem Jahre 1869 mit der Genauigkeit beim Bogenschießen beschäftigt hat. Von ihm stammen übrigens auch die Begriffe *Positiv, Negativ* und *Schnappschuss* in der Photographie.

[6] Man vergleiche dazu die Ausführungen über *bias* und *variability* in Abschnitt *"Why randomize?"* in [14]

[7] Dies ist der von der Visierlinie und der sogenannten *Laufseelenachse* eingeschlossene Winkel.

Zur weiteren Begriffsklärung:

Die *Visierlinie* ist die durch *Kimme*, *Korn* und Zielobjekt bestimmte Gerade. Sie ist im vorliegenden Fall waagrecht.

x die horizontale Entfernung der Kugel vom Ende des Gewehrlaufes in Richtung Zielscheibe und

$y_\alpha(x)$ ihre Höhe als Funktion von x ,

so ist letztere aufgrund von Galileis Fallgesetz gleich

$$y_\alpha(x) = y_0 + \frac{1}{2\cos^2(\alpha)} x(\sin(2\alpha) - \frac{g}{v^2} \cdot x) \,.$$

Ist der *Anstellwinkel* gleich $\alpha = 0$, so ergibt sich demgemäß

$$y_0(x) = y_0 - \frac{g}{2v^2} \cdot x^2$$

und somit der Höhenverlust $\frac{g}{2v^2} \cdot x^2$. Dieser lässt sich dadurch vermeiden, dass man den *Anstellwinkel* α so wählt, dass $y_\alpha(x_0) = y_0$ gilt. Dies wird offensichtlich dadurch erreicht, dass man den letzten Term in der Formel für $y_\alpha(x)$ gleich Null setzt. Dementsprechend hat man den Anstellwinkel gleich

$$\alpha = \frac{1}{2} \arcsin(\frac{g}{v^2} \cdot x_0)$$

zu wählen. Dabei wird mit $\arcsin(.)$ der Hauptwert des Arcus-Sinus bezeichnet.

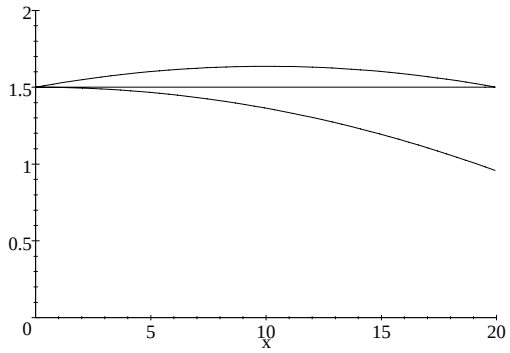


Abbildung der *Visierlinie* und der Bahnkurven der Kugel bei *Anstellwinkel* $\alpha = 0$ und bei adjustiertem *Anstellwinkel*

Die Laufseelenachse ist die durch den Gewehrlauf bestimmte Gerade.

Die Kimme ist eine $\vee$-fömige Kerbe am Visier des Gewehrs.

Das Korn ist der zugehörige komplementäre Teil des Visiers. Es hat demnach die Form $\wedge$.

2.1.3 Konfidenzintervalle für Wahrscheinlichkeiten und Anteilswerte

In diesem Abschnitt betrachten wir ein Zufallsexperiment mit zwei möglichen Ausgängen, die wir *"Erfolg"* und *"Misserfolg"* nennen. Die Wahrscheinlichkeit eines Erfolges sei für jede Durchführung des Experiments gleich $p \in (0,1)$. Es werden n solcher Zufallsexperimente durchgeführt, die Ergebisse $X_i,\ i \in \{1,...,n\}$ der einzelnen Experimente festgestellt und protokolliert. Dabei sei

$$X_i = 1 \quad \text{oder} \quad X_i = 0\,,$$

je nachdem ob beim i-ten Experiment Erfolg oder Misserfolg eintritt. Aufgrund des Versuchsergebnisses $(X_1,...,X_n)$ sei für die Wahrscheinlichkeit p ein Konfidenzintervall anzugeben.

Bezeichnen $S_n = \sum_{i=1}^{n} X_i$ die beobachtete Anzahl der Erfolge und $\hat{p}_n = \frac{S_n}{n}$ den zugehörigen Anteil der Erfolge. Aufgrund des empirischen Gesetzes der Großen Zahlen ist $\hat{p}_n$ ein naheliegender Schätzer für p. Diese Tatsache wird durch dessen theoretisches Gegenstück, nämlich das *Bernoulli*sche Gesetz der Großen Zahlen, bestätigt. Außerdem gilt aufgrund der Linearität des Erwartungswerts $E(\hat{p}_n) = p$, sodass $\hat{p}_n$ ein erwartungstreuer Schätzer für p ist. Da sich unter den im nachstehenden Urnenmodell präzisierten Voraussetzungen überdies zeigen lässt, dass $\hat{p}_n$ unter allen erdenklichen erwartungstreuen Schätzern jener mit kleinstmöglicher Varianz ist, ist $\hat{p}_n$ auch zur Konstruktion eines Konfidenzintervalls für p bestmöglich geeignet.

Wir unterscheiden zwei Klassen von Anwendungen:

In der Klasse A) laufen die Experimente stets unter den gleichen Bedingungen ab. Dementsprechend beinflussen einander die Ergebnisse der einzelnen Experimente nicht. Man sagt, die Experimente sind *unabhängig*.

In der Klasse B) bestehen die einzelnen Experimente darin, dass man einer Grundgesamtheit Elemente zufällig entnimmt und diese der Grundgesamtheit nicht wieder hinzufügt.[8] Die Größe p ist in diesem Fall der Anteil

[8] In der Qualitätskontrolle ist das Feststellen der Ergebnisse der einzelnen Experimente oft damit verbunden, dass die entnommenen Elemente zerstört werden. Es wäre daher absurd, die entnommenen Elemente der Grundgesamtheit wieder hinzuzufügen. In der Meinungsforschung ist dies deswegen unangebracht, da man ansonst ein- und dieselbe Person mehrfach befragen könnte.

der Elemente der Grundgesamtheit mit einer bestimmten Eigenschaft. Der Gesamtversuch besteht darin, dass man der Grundgesamtheit n Elemente entnimmt. Man nennt einen solchen Versuch eine *(Zufalls-)Stichprobe vom Umfang* n .

Die beiden Klassen von Anwendungen lassen sich bekanntlich durch folgendes Urnenmodell beschreiben: Eine Urne enthalte s schwarze und w weisse Kugeln, welche - von ihrer Farbe abgesehen - ununterscheidbar sind. Ferner bezeichne $N = s + w$ die Gesamtanzahl der Kugeln in der Urne.

Es werden n Kugeln zufällig und A) mit Zurücklegen bzw. B) ohne Zurücklegen gezogen. Für den Anteil $p = \frac{s}{N}$ der schwarzen Kugeln in der Urne ist mit Hilfe des Anteils $\hat{p}_n$ der schwarzen Kugeln in der Stichprobe ein Konfidenzintervall zu konstruieren.

Bezeichnen $B_{n,p}$ die Binomialverteilung mit den Parametern n und p und $H_{n,N,s}$ die Hypergeometrische Verteilung mit den Parametern n, N und s. Dann gelten für Verteilung, Erwartungswert und Varianz von S_n in den Fällen A) respektive B) bekanntlich

$$\begin{array}{lll} S_n \sim B_{n,p} & , \; E(S_n) = np & \text{und} \quad V(S_n) = np(1-p) \\ S_n \sim H_{n,N,s} & , \; E(S_n) = n\frac{s}{N} & \text{und} \quad V(S_n) = n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n-1}{N-1}) \, . \end{array}$$

Wir gehen im folgenden stets davon aus, dass der Stichprobenumfang n so groß ist, dass im Fall A) die Normalapproximation der Binomialverteilung und im Fall B) die Normalapproximation der Hypergeometrischen Verteilung gerechtfertigt ist. Im Fall B) haben wir wegen $n \leq N$ zudem vorauszusetzen, dass die Grundgesamtheit N (um einiges) größer als n ist.

A) Konfidenzintervalle für Wahrscheinlichkeiten

A0) Ermittlung eines Konfidenzintervalls: Unter der genannten Voraussetzung gilt für $z > 0$

$$P\left(\left|\frac{S_n - np}{\sqrt{npq}}\right| \leq z\right) \cong 2\Phi\left(z\right) - 1 \, .$$

Seien nun $\alpha \in (0,1)$ und $z_{1-\alpha/2} = \Phi^{-1}(1-\frac{\alpha}{2})$. Dann gilt für $z = z_{1-\alpha/2}$ wegen $\frac{S_n - np}{\sqrt{npq}} = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}$

$$P\left(\left|\frac{\hat{p}_n - p}{\sqrt{p\left(1-p\right)/n}}\right| \leq z_{1-\alpha/2}\right) \cong 1 - \alpha \, .$$

Da die Diskriminante der zur nachstehenden quadratischen Ungleichung gehörigen quadratischen Gleichung gleich

$$(\hat{p}_n + \frac{z^2_{1-\alpha/2}}{2n})^2 - (1 + \frac{z^2_{1-\alpha/2}}{n})\hat{p}_n^2 = \frac{z^2_{1-\alpha/2}}{n}(\hat{p}_n\,(1-\hat{p}_n) + \frac{z^2_{1-\alpha/2}}{4n})$$

ist, gilt

$$\begin{aligned}\left|\tfrac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}\right| \le z_{1-\alpha/2} \quad &\Longleftrightarrow \quad (\hat{p}_n - p)^2 \le \tfrac{z^2_{1-\alpha/2}}{n} p\,(1-p) \\ &\Longleftrightarrow \quad p^2(1 + \tfrac{z^2_{1-\alpha/2}}{n}) - 2p(\hat{p}_n + \tfrac{z^2_{1-\alpha/2}}{2n}) + \hat{p}_n^2 \le 0 \\ &\Longleftrightarrow \quad p \in [p_n^-, p_n^+]\ .\end{aligned}$$

Dabei sind

$$p_n^{\pm} = \frac{1}{1 + \frac{z^2_{1-\alpha/2}}{n}}(\hat{p}_n + \frac{z^2_{1-\alpha/2}}{2n} \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n) + \frac{z^2_{1-\alpha/2}}{4n}})$$

die beiden Lösungen der entsprechenden quadratischen Gleichung.
Daher ist $[p_n^-, p_n^+]$ ein $(1-\alpha)\cdot 100\%$-iges Näherungskonfidenzintervall für p, welches wir - entsprechend der Angelsächsischen Bezeichnung - *Score-Konfidenzintervall* [9] nennen.

Da n groß im Verhältnis zu $z^2_{1-\alpha/2}$ ist - andernfalls hätte die Normalapproximation nicht verwendet werden dürfen - wird in der Praxis $z^2_{1-\alpha/2}/n$ zumeist vernachlässigt. Dementsprechend erhält man für die Endpunkte $p_n^{\pm}$ des obigen Konfidenzintervalls folgende Approximationen

$$\hat{p}_n^{\pm} = \hat{p}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)}\,.$$

Das in der Praxis zumeist verwendete, sogenannte *Wald'sche Näherungskonfidenzintervall* [10] für die gesuchte Wahrscheinlichkeit p ist daher

$$[\hat{p}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)},\ \hat{p}_n + \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)}]\,.$$

[9] Mit seinem Score-Konfidenzintervall hat der US-amerikanische Statistiker *Edwin B. Wilson* (1879 – 1964) im Jahre 1927 die inbesondere durch *Jerzy Neyman* vorangetriebe Arbeit über Konfidenzintervalle eingeleitet.

[10] Dieses Näherungskonfidenzintervall wurde im Jahre 1943 von *Abraham Wald* (1902 – 1950) vorgeschlagen. Der aus Klausenburg (Kolozsvár, Cluj) im heutigen Rumänien stammende Statistiker war von 1933 – 1938 an dem damals von *Oskar Morgenstern* geleiteten Österreichischen Institut für Konjunkturforschung in Wien beschäftigt. Nach seiner Flucht in die U.S.A. war er im Zusammenhang mit der Qualitätssicherung bei der Waffenproduktion maßgeblich daran beteiligt, die sequentielle Statistik zu entwickeln.

Anmerkung 4: Seien α und n fest und $z = z_{1-\alpha/2}$. Dann genügen die Endpunkte der Score-Konfidenzintervalle $[p_n^-, p_n^+]$, welche lediglich vom Schätzer $\hat{p}_n \in [0,1]$ abhängen, der Gleichung

$$(\hat{p}_n - p)^2 = \frac{z^2}{n} p(1-p)\,.$$

Der geometrische Ort aller Punkte, die dieser Gleichung genügen, ist eine Ellipse, welche wir für unseren Zweck *Score-Ellipse* nennen wollen. Die Endpunkte der Wald'schen Näherungskonfidenzintervalle genügen der Gleichung

$$(\hat{p}_n - p)^2 = \frac{z^2}{n} \hat{p}_n(1-\hat{p}_n)\,.$$

Der geometrische Ort aller Punkte, die dieser Gleichung genügen, ist ebenfalls eine Ellipse. Wir nennen sie *Wald'sche Ellipse*. Score-Ellipse und Wald'sche Ellipse haben - dem Umstand entsprechend, dass bei ihnen die Rollen von p und $\hat{p}_n$ vertauscht sind - folgende gemeinsamen Eigenschaften:

- sie haben den Mittelpunkt $(\frac{1}{2}, \frac{1}{2})$ und
- beeinhalten die Punkte $(0,0)$ und $(1,1)$.

Ihre Tangenten in den beiden Punkte $(0,0)$ und $(1,1)$ sind jedoch

- im Fall der Score-Ellipse parallel zur Ordinate und
- im Fall der Wald'schen Ellipse parallel zur Abszisse.

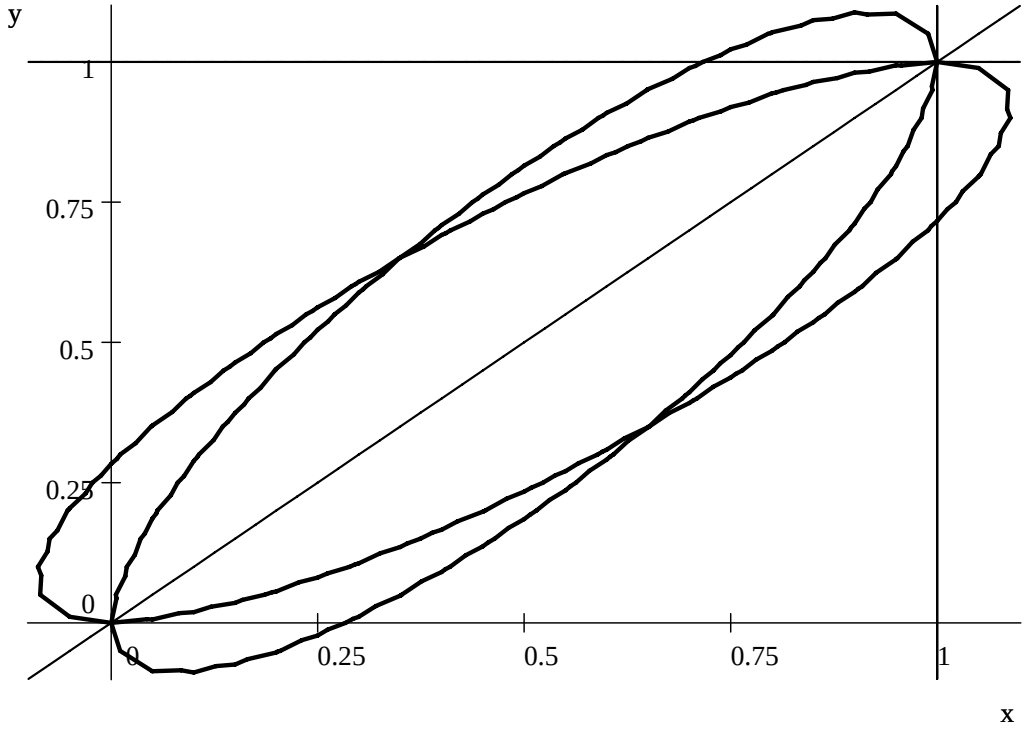


Abbildung: Score-Ellipse und Wald'sche Ellipse für $z = 2$ und $n = 10$

A1) Folgerungen aus der konkreten Gestalt des Näherungskonfidenzintervalls: Aus der konkreten Gestalt des Näherungskonfidenzintervalls lassen sich folgende Sachverhalte ablesen.

a) Wie im Fall des Konfidenzintervalls $[X_{n:n}, X_{n:n}/\sqrt[n]{\alpha}]$ des letzten Abschnitts hängen Lage und Länge des vorliegenden Konfidenzintervalls vom Zufall ab.

b) Die Länge des Konfidenzintervalls ist im vorliegenden Fall das Doppelte des sogenannten *Fehlers*[11]

$$\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)}\,.$$

[11] Der durch $\pm\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)}$ beschriebene Bereich wird in der angelsächsischen Literatur *margin of error* genannt.

Dieser hängt ab

b1) via $z_{1-\alpha/2} = \Phi^{-1}(\frac{1}{2}+\frac{1-\alpha}{2})$ von der vorgegebenen statistischen Sicherheit $1-\alpha$.

b2) via $1/\sqrt{n}$ vom Stichprobenumfang und

b3) via $\sqrt{\hat{p}_n\,(1-\hat{p}_n)}$ vom Schätzwert $\hat{p}_n$.

Zu b1) Da $\beta \longmapsto \Phi^{-1}(\beta)$ eine streng monoton wachsende Funktion ist, wächst $z_{1-\alpha/2} = \Phi^{-1}(\frac{1}{2} + \frac{1-\alpha}{2})$ mit der statistischen Sicherheit. Für die klassischen Werte α gelten beispielsweise

α	$(1-\alpha)\,100\,\%$	$z_{1-\alpha/2}$
0.1	90 %	1.645
0.05	95 %	1.960
0.01	99 %	2.576

Zu b2) Will man beispielsweise bei gleicher statistischer Sicherheit und gleichem Schätzwert $\hat{p}_n$ die Länge des Konfidenzintervalls halbieren, so muss man den Stichprobenumfang vervierfachen.

Zu b3) Da sich die Ungleichung

$$\sqrt{\hat{p}_n\,(1-\hat{p}_n)} \leq \frac{\hat{p}_n + 1 - \hat{p}_n}{2} (= \frac{1}{2})$$

zwischen dem geometrischen Mittel $\sqrt{\hat{p}_n\,(1-\hat{p}_n)}$ und dem arithmetischen Mittel $\frac{1}{2}$ an einem Thaleskreis mit Mittelpunkt $(\frac{1}{2}, 0)$ und Radius $r = \frac{1}{2}$ veranschaulichen lässt, ist $y = \sqrt{\hat{p}_n\,(1-\hat{p}_n)}$ die positive Lösung der zugehörigen Kreisgleichung

$$(\hat{p}_n - \frac{1}{2})^2 + y^2 = (\frac{1}{2})^2 \,. \qquad (0)$$

Demgemäß verhält sich die Länge des Konfidenzintervalls. Sie ist also für $\hat{p}_n \cong \frac{1}{2}$ groß und für $\hat{p}_n \cong 0$ oder $\hat{p}_n \cong 1$ verschwindend klein.

A2) Anmerkungen zur Versuchsplanung: Bei der Versuchsplanung hat der Statistiker in Zusammenarbeit mit dem Fachwissenschaftler bzw. in Verhandlung mit dem Auftraggeber einer Umfrage 1) die statistische Sicherheit $1-\alpha$ und 2) den Stichprobenumfang n festzulegen.

Im Fall, dass man auf Resultate von einschlägigen Voruntersuchungen oder auf Ergebnisse vergleichbarer Studien zurückgreifen kann, kann unter

Umständen der Stichprobenumfang beträchtlich verringert oder die statistische Sicherheit erhöht werden.

Für Beratungen bei der Versuchsplanung ist bekanntlich die Bedingung

$$P(|\hat{p}_n - p| \leq \varepsilon) \geq 1 - \alpha$$

entscheidend. Bei vorgegebenen ε und α erhält man nach Anwendung der Normalapproximation den Stichprobenumfang n gemäß

$$\begin{array}{rcl} \frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{p\,(1-p)} \leq \varepsilon \iff n & \geq & (\frac{z_{1-\alpha/2}}{\varepsilon})^2 p\,(1-p) \\ & \cong & (\frac{z_{1-\alpha/2}}{\varepsilon})^2 \cdot c \quad \text{mit} \quad c \in (0, 1/4]\,, \end{array}$$

wobei

$$c = 1/4 \quad \text{oder} \quad c = p_0(1-p_0) \ \text{ mit } \ 0 < p_0 < \frac{1}{2}$$

je nachdem, ob es keine Vorinformation hinsichtlich p gibt oder ob aufgrund von Vorinformationen mit gutem Grund angenommen werden kann, dass $p \in (0, p_0) \cup (1-p_0, 1)$ ist.

A3) Angabe der Parametermenge, für welche die Normalapproximation gerechtfertigt ist: Nach Abschätzung des Stichprobenumfangs lässt sich schließlich jene Menge der Parameter $p \in (0,1)$ angeben, für welche die Normalapproximation der Binomialverteilung angebracht ist. Dies geschieht mit Hilfe der Faustregel

$$n\,p(1-p) \geq 9$$

und der Darstellung von $y^2 = p(1-p)$ in der zu Punkt b3) angeführten Kreisgleichung. Demgemäß ist die Anwendung der Normalapproximation für alle $p \in (0,1)$ gerechtfertigt, welche

$$\left| p - \frac{1}{2} \right| \leq \frac{1}{2}\sqrt{1 - \frac{36}{n}}$$

erfüllen. (Damit diese Menge nicht leer ist, ist stets von einem Stichprobenumfang $n > 36$ auszugehen.)

A4) Durchführung eines Urnenversuchs und Interpretation

Ein Konfidenzintervall ist - wie gesagt - so konstruiert, dass die Wahrscheinlichkeit, dass das Konfidenzintervall den wahren Parameter p *überdeckt*, (ungefähr) $1-\alpha$ ist.

Die nachstehende Illustration ist Chapter 21: *What is a Confidence Interval?* in [13] entnommen.

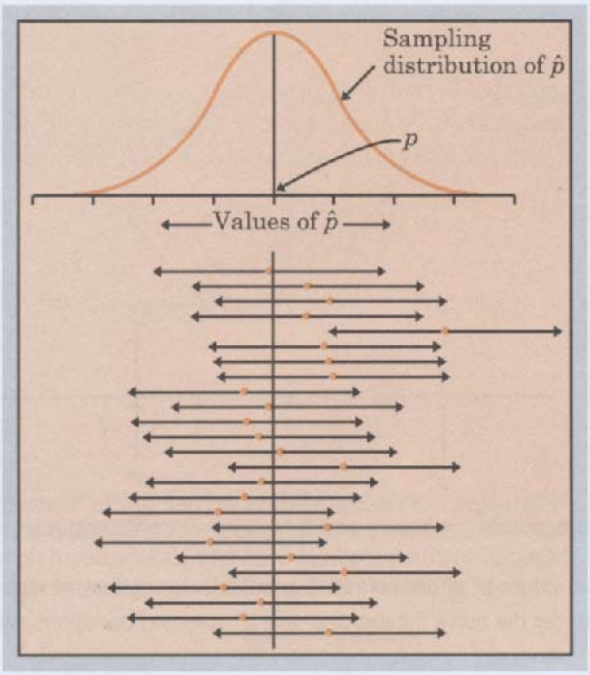


Figure 21.4 Twenty-five samples from the same population give these 95% confidence intervals. In the long run, 95% of all such intervals cover the true population proportion, marked by the vertical line.

Ein gutes Verständnis für die Wirkungsweise von Konfidenzintervallen vermittelt etwa folgender Urnenversuch, für den wir $\alpha = 0.05$ wählen.

Eine Urne enthalte 100 Kugeln, von denen $s \in [35, 65]$ schwarz und $w = 100 - s$ weiss sind. (Von der Farbe abgesehen seien die Kugeln ununterscheidbar.) Eine Stichprobe bestehe darin, der Urne $n = 40$ Kugeln zufällig und mit Zurücklegen zu entnehmen.

Wir ziehen 20 solcher Stichproben und ermitteln für jede das zugehörige Konfidenzintervall für den Anteil $p = \frac{s}{100}$ der schwarzen Kugeln in der Urne. Wegen $1 - \alpha = \frac{95}{100} = \frac{19}{20}$ ist zu erwarten, dass 19 der 20 Konfidenzintervalle den Anteil p überdecken.

Ein solcher Versuch ist unter *"Kurze Einführung in die Praxis und Theorie der Stichproben in Form eines Gesprächs"* in [28] dokumentiert.

B) Konfidenzintervalle für Anteilswerte

Da im Fall B) des Ziehens ohne Zurücklegen S_n gemäß der Hypergeometrischen Verteilung $H_{n,N,s}$ verteilt ist, ist unter der Annahme, dass n und $N\,(\geq n)$ hinreichend groß sind, die Normalapproximation der Hypergeometrischen Verteilung zu berücksichtigen. Der Umstand, dass deren Varianz aus der der zugehörigen Binomialverteilung durch Multiplikation mit dem Faktor $1-\frac{n-1}{N-1}<1$ hervorgeht, schlägt sich im Wald'sche Näherungskonfidenzintervall

$$\left[\hat{p}_n-\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)}\sqrt{1-\frac{n}{N}},\ \hat{p}_n+\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}_n\,(1-\hat{p}_n)}\sqrt{1-\frac{n}{N}}\right]$$

für den Anteilswert $\frac{s}{N}$ dadurch nieder, dass der Schätzer für die Standardabweichung mit der Quadratwurzel aus $1-\frac{n}{N}$ multipliziert wird. Diese Näherung von $1-\frac{n-1}{N-1}$ nennt man übrigens *Endlichkeitskorrektur.*

Die Verringerung der Varianz bewirkt - wie aus dem Folgenden ersichtlich ist - auch eine Verringerung des Stichprobenumfangs.

$$\begin{aligned}\frac{z_{1-\alpha/2}}{\sqrt{n}}\sqrt{p\,(1-p)}\sqrt{1-\tfrac{n}{N}}\leq\varepsilon \iff n &\geq c\left(\tfrac{z_{1-\alpha/2}}{\varepsilon}\right)^2\left[\frac{1}{1+\frac{c}{N}\left(\frac{z_{1-\alpha/2}}{\varepsilon}\right)^2}\right]\\ &\left(\leq c\left(\tfrac{z_{1-\alpha/2}}{\varepsilon}\right)^2\right).\end{aligned}$$

Diese fällt jedoch umso geringer aus, je größer N ist. Bei Meinungsumfragen aus einer sehr großen Grundgesamtheit fällt diese Verringerung nicht ins Gewicht, sodass eine Umfrage unter den Wahlberechtigten im Bundesland Salzburg kaum billiger als eine solche in den USA ist.

Anmerkung 5: In der Markt- und Meinungsforschung bedient man sich üblicherweise keiner Zufallsstichprobe sondern einer sogenannten *Quoten-* oder *Anteilsstichprobe.* Für eine solche wird die Bevölkerung nach zweckmäßigen Merkmalen wie Geschlecht, Altersgruppe und Wohngebiet (Regionen, Landgemeinden, Klein- Mittel- und Großstädten) in *Schichten*[12] eingeteilt. Die aus der amtlichen Statistik bekannten Anteile N_j/N dieser Schichten werden in der Stichprobe proportional nachgebildet, sodass bei einer Stichprobe

[12] *Schichten* (*strata*) sind Teilgesamtheiten der Grundgesamtheit, die einander ausschließen und zusammen die Grundgesamtheit ergeben.

von ingesamt n Befragten die Stichprobenumfänge $n_j \cong n \cdot \frac{N_j}{N}$, $j \in \{1, ..., k\}$, auf die einzelnen Schichten entfallen. Bezeichnet $\hat{p}_{n_j}$ den für die Schichte j ermittelten Schätzer, dann ist

$$\hat{p}_n = \sum_{j=1}^{k} \frac{n_j}{n} \cdot \hat{p}_{n_j}$$

der *Schätzer der Anteilsstichsprobe*. Im namentlich in der amtlichen Statistik bevorzugten Idealfall wird innerhalb jeder Schicht eine Zufallsstichprobe durchgeführt. Hinsichtlich der in der Markt- und Meinungsforschung geübten Praxis bei der Durchführung des Quotenverfahrens sei beispielsweise auf [15], Seite 97 ff verwiesen.

2.1.4 Ausblick 1: Geometrie der Score-Ellipse

Sei $c = \frac{z^2_{1-\alpha/2}}{n}$. Der Ausgangspunkt für die Ermittlung des Score-Konfidenzintervalls war die Beziehung

$$(p - \hat{p}_n)^2 = c \cdot p(1-p)$$

(vgl. Abschnitt 2.1.3, Anmerkung 4) bzw. im Hinblick auf (0)

$$(p - \hat{p}_n)^2 + c\,(p - \frac{1}{2})^2 - \frac{c}{4} = 0\,.$$

Für $x := p - \frac{1}{2}$ und $y := \hat{p}_n - \frac{1}{2}$ ergibt sich daraus die Gleichung

$$y^2 - 2xy + (1+c)x^2 - \frac{c}{4} = 0\,,$$

wobei wir im Weiteren stets $c \in (0, 2]$ annehmen, was für unsere Anwendung keine Einschränkung darstellt, zumal typischerweise bereits $z^2_{1-\alpha/2} \leq 2$ ist und im Hinblick auf die Anwendbarkeit der Normalapproximation zumindest $n > 36$ verlangt ist.

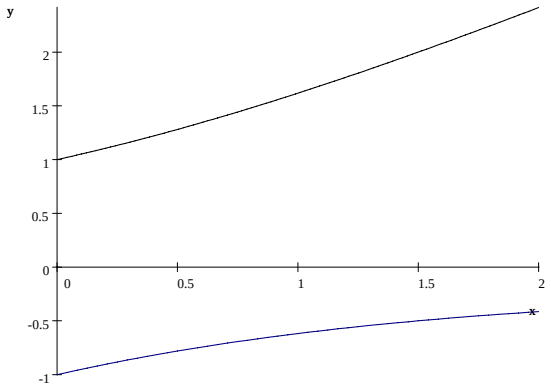
Behauptung 1 [13]**:** (a) Die Funktionen

$$k_a(c) = \frac{c}{2} + \sqrt{1 + (\frac{c}{2})^2} \quad \text{und} \quad k_b(c) = \frac{c}{2} - \sqrt{1 + (\frac{c}{2})^2}\,,\ c \in [0, 2]$$

[13] Hinsichtlich eines Beweises sei auf die entsprechende Übungsaufgabe zum Thema Quadriken in den Übungen "Geometrie für Lehramt" verwiesen.

sind streng monoton wachsend und es gelten

$$1 \le k_a(c) \le 1+\sqrt{2} \quad \text{und} \quad -1 \le k_b(c) \le 1-\sqrt{2}\,.$$



(b) Die Geradengleichungen der beiden Achsen der Score-Ellipse sind

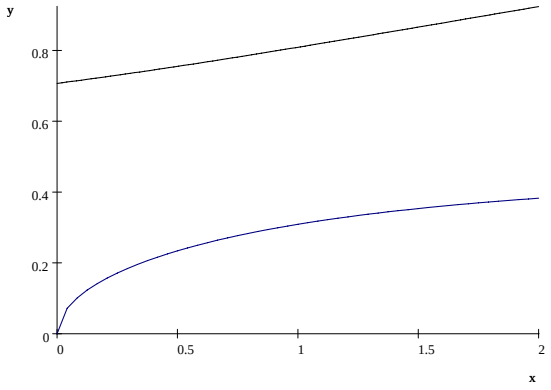
$$f_a(p) = \frac{1}{2} + k_a(c)(p-\frac{1}{2}) \quad \text{bzw.} \quad f_b(p) = \frac{1}{2} + k_b(c)(p-\frac{1}{2})\,.$$

(c) Die Länge ihrer großen und kleinen Halbachse ist

$$a(c) = \frac{1}{2}\sqrt{1+\frac{c}{2}+\sqrt{1+(\frac{c}{2})^2}} \;\text{ bzw. }\; b(c) = \frac{1}{2}\sqrt{1+\frac{c}{2}-\sqrt{1+(\frac{c}{2})^2}}\,,\; c \in [0,2]\,,$$

wobei auch die Funktionen $c \mapsto a(c)$ und $c \mapsto b(c)$ streng monoton wachsend sind und es gelten

$$\frac{1}{\sqrt{2}} \le a(c) \le \sqrt{\frac{1+\frac{1}{\sqrt{2}}}{2}} \quad \text{und} \quad 0 \le b(c) \le \sqrt{\frac{1-\frac{1}{\sqrt{2}}}{2}}\,.$$



Anmerkung 1: Für den Wert $c = 0$ degeneriert die Ellipse zur Strecke $(0,0)$ - $(1,1)$.

Anmerkung 2: Das Quadrat Exzentrizität $e(c)$ der Ellipse ist

$$e^2(c) = \frac{1}{2}(1 + \frac{c}{2}) \in [\frac{1}{2}, 1] .$$

2.1.5 Ausblick 2: Vergleich des Score-Konfidenzintervalls mit dem Wald'schen Approximationsintervall

Seien $\alpha_n = (1 + \frac{z^2}{n})^{-1}$ und $z = z_{1-\alpha/2}$, $\alpha \in (0,1)$. Dann ist der Mittelpunkt des Score-Konfidenzintervalls gemäß

$$\tilde{p}_n = \frac{p_n^+ + p_n^-}{2} = \alpha_n \cdot \hat{p}_n + (1 - \alpha_n) \cdot \frac{1}{2} = \frac{\hat{p}_n + z^2/2n}{1 + z^2/n} .$$

ein gewichtetes Mittel aus $\hat{p}_n$ und $\frac{1}{2}$ und es gilt

$$\frac{1}{2} - \tilde{p}_n = \alpha_n(\frac{1}{2} - \hat{p}_n) . \tag{1}$$

Das bedeutet, dass die Lage des Wald'schen Approximationsintervalls im Vergleich zu der des Score-Intervalls zu den Rändern 0 und 1 hin verschoben ist. In der Tat ist es nur dann zur Gänze im Intervall $[0,1]$ enthalten, wenn gilt

$$\hat{p}_n \in [1 - \alpha_n, \alpha_n] .$$

Die Länge $\frac{2z}{\sqrt{n}} \alpha_n \sqrt{\hat{p}_n (1 - \hat{p}_n) + \frac{1}{4}\frac{z^2}{n}}$ des Score-Intervalls ist genau dann größer als die des Wald'schen Approximationsintervalls, nämlich $\frac{2z}{\sqrt{n}} \sqrt{\hat{p}_n (1 - \hat{p}_n)}$, wenn gilt

$$\left| \hat{p}_n - \frac{1}{2} \right| > \frac{1}{2} \sqrt{1 - \frac{1}{2 + z^2/n}} .$$

(Dies trifft insbesondere für die Extremfälle $\hat{p}_n = 0$ und $\hat{p}_n = 1$ zu, für welche sich das Wald'sche Approximationsintervall auf je einen Punkt reduziert.)

Aus den beiden genannten Gründen ist die Überdeckungswahrscheinlichkeit des Wald'schen Approximationsintervalls für p nahe 0 oder 1 und kleine n deutlich kleiner als der Sollwert $1 - \alpha$.

Abschließend bieten wir eine zukunftsträchtige Verbesserung des Score-Intervalls an. Wegen

$$p(1-p) = \frac{1}{4} - (\frac{1}{2} - p)^2$$

und (1) ist

$$\begin{aligned}
\tilde{p}_n\,(1-\tilde{p}_n) &= \frac{1}{4} - (\frac{1}{2} - \tilde{p}_n)^2 \\
&= \frac{1}{4} - \alpha_n^2(\frac{1}{2} - \hat{p}_n)^2 \\
&= \frac{1}{4}\left(1-\alpha_n^2\right) + \alpha_n^2(\frac{1}{4} - (\frac{1}{2} - \hat{p}_n)^2) \\
&= \alpha_n^2\,\hat{p}_n\,(1-\hat{p}_n) + \frac{1}{4}\left(1-\alpha_n^2\right)\,.
\end{aligned}$$

Also gilt für das $n/(2z)^2$-Fache des Quadrats der Länge des Score-Konfindenzintervalls

$$\begin{aligned}
\alpha_n^2\,(\hat{p}_n\,(1-\hat{p}_n) + \frac{1}{4}\frac{z^2}{n}) &= \tilde{p}_n\,(1-\tilde{p}_n) - \frac{1}{4}(1-\alpha_n^2(1+\frac{z^2}{n})) \\
&= \tilde{p}_n\,(1-\tilde{p}_n) - \frac{1}{4}\,(1-\alpha_n) \\
&= \tilde{p}_n\,(1-\tilde{p}_n) - \frac{1}{4\,(n/z^2+1)} \\
&\leq \frac{n}{n+z^2}\,\tilde{p}_n\,(1-\tilde{p}_n)\,,
\end{aligned}$$

wobei die Ungleichung wegen

$$\begin{aligned}
\Delta &= n\cdot\tilde{p}_n\,(1-\tilde{p}_n) - \left(n+z^2\right)(\tilde{p}_n\,(1-\tilde{p}_n) - \frac{z^2}{4\,(n+z^2)}) \\
&= z^2(\frac{1}{4} - \tilde{p}_n\,(1-\tilde{p}_n)) \geq 0
\end{aligned}$$

gilt. Daher hat das Score-Konfidenzintervall in Abhängigkeit von $\tilde{p}_n$ die Form

$$p_n^{\pm} = \tilde{p}_n \pm \frac{z}{\sqrt{n}}\sqrt{\tilde{p}_n\,(1-\tilde{p}_n) - \frac{1}{4\,(n/z^2+1)}}\,,$$

wobei die Approximation

$$\tilde{p}_n^{\pm} \cong \tilde{p}_n \pm z\sqrt{\frac{\tilde{p}_n\,(1-\tilde{p}_n)}{n+z^2}}\,,$$

welche ähnlich handlich wie die Wald'sche ist, im Unterschied zu dieser jedoch die Überdeckungswahrscheinlichkeit auch für kleine n geringfügig erhöht (vgl. [25]). Dieser Vorteil der letztgenannten Approximation führt dazu, dass sie die Wald'sche Approximation in jüngster Zeit zu ersetzen beginnt.

2.2 TESTEN VON HYPOTHESEN (Teil 1): Wahrscheinlichkeiten und Anteilswerte

Eine Auswahl von Zitaten

"Ad destruendum sufficit unum."[14]

Galileo Galilei, Dialogo, Florenz 1632

Der theoretisch arbeitende Naturforscher ist nicht zu beneiden, denn die Natur, oder genauer gesagt: das Experiment, ist eine unerbittliche und wenig freundliche Richterin seiner Arbeit. Sie sagt zu einer Theorie nie "ja", sondern im günstigsten Fall "vielleicht", in den meisten Fällen aber einfach "nein". Stimmt ein Experiment zur Theorie, bedeutet es für letztere "vielleicht", stimmt es nicht, so bedeutet es "nein". Wohl jede Theorie wird einmal ihr "nein" erleben, die meisten Theorien schon bald nach ihrer Entstehung.

Albert Einstein, Leiden 1922[15]

"... als Grundlage unserer Handlungen sollten wir die bestgeprüfte Theorie vorziehen. Mit anderen Worten, es gibt keine "absolute Sicherheit"; doch da wir wählen müssen, ist es vernünftig, die bestgeprüfte Theorie zu wählen."

Sir Karl Popper, 1973[16]

[14] Sinngemäß übersetzt: *Um etwas zu widerlegen, bedarf es nur eines einzigen Gegenbeispiels.*

[15] *Einstein, A.*: Briefe. Diogenes Verlag, Zürich 1981 (S. 19 - 20)

[16] *Popper, K.*: Objektive Erkenntnis. Hoffmann und Campe, Hamburg 1973 (S. 14, 39 ff)

Wir betrachten ausschließlich

Hypothesen hinsichtlich Wahrscheinlichkeiten bzw. Anteilswerten bei Alternativexperimenten

Fragestellung: Aus einer Urne mit schwarzen und weißen Kugeln werden zufällig und mit Zurücklegen n Kugeln gezogen. Der Ausfall sei $X_i = 1\,(0)$, falls die i-te gezogene Kugel schwarz (weiß) ist. p sei der Anteil der schwarzen Kugeln in der Urne. Auf der Basis der n Beobachtungen soll eine Entscheidungsregel angegeben werden, die es erlaubt, sich zwischen zwei widersprechenden Hypothesen (Theorien)

Hypothese H_0 (*Nullhypothese*)
Hypothese H_1 (*Alternativhypothese*)

bezüglich der Urnenzusammensetzung zu entscheiden.

Dabei werden wir hier zumeist den Fall eines großen Stichprobenumfangs behandeln. Für den Fall eines kleinen Stichprobenumfangs sei auf die Übungsaufgaben verwiesen.

Hinsichtlich der Formulierung der Hypothesen unterscheiden wir folgende Fälle, wobei stets $p_0 \in (0,1)$ vorausgesetz wird.

1) $\boxed{\begin{array}{l} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{array}}$

2a) $\boxed{\begin{array}{l} H_0 : p = p_0 \\ H_1 : p > p_0 \end{array}}$ 2b) $\boxed{\begin{array}{l} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{array}}$

2c) $\boxed{\begin{array}{l} H_0 : p = p_0 \\ H_1 : p < p_0 \end{array}}$ 2d) $\boxed{\begin{array}{l} H_0 : p \geq p_0 \\ H_1 : p < p_0 \end{array}}$

3a) $\boxed{\begin{array}{l} H_0 : p = p_0 \\ H_1 : p = p_1 \end{array}}$ 3b) $\boxed{\begin{array}{l} H_0 : p \leq p_0 \\ H_1 : p \geq p_1 \end{array}}$

mit $p_0 < p_1 \leq 1$

Wir werden die Fälle 1), 2a), 2b) und 3a) behandeln, da einerseits die Fälle 2c) und 2d) ganz analog zu den Fällen 2a) und 2b) sind und andererseits die Behandlung von Fall 3b) Projekt 13 vorbehalten sei. Für die Fälle 2a) und 3a) geht man davon aus, dass die Urnenzusammensetzung so gewählt wurde, dass entweder H_0 oder H_1 zutrifft.

Zumal bekanntlich $\widehat{p}_n = \frac{1}{n} S_n$, $S_n = \sum_{i=1}^{n} X_i$ der Schätzer für p schlechthin ist, werden wir unsere Entscheidung auf der Basis von $\widehat{p}_n$ bzw. von S_n treffen.

1) $\boxed{\begin{array}{c} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{array}}$

Unter der Annahme der Gültigkeit von H_0 gilt $S_n \sim B_{n,p_0}$ und somit $E(S_n) = n \cdot p_0$. Es liegt somit nahe, sich zugunsten H_1 zu entscheiden, wenn S_n von np_0 "stark abweicht", d.h. wenn gilt $|S_n - np_0| > c$, wobei der *kritische Wert* $c > 0$ geeignet zu wählen ist.

Die Wahl von c geschieht so, dass eine wahrscheinlichkeitstheoretische Interpretation möglich ist: Wir wählen das sogenannte *Signifikanzniveau*[17]; das ist eine Zahl $0 < \alpha \ll 1$ (etwa $\alpha = 0.05$ oder $\alpha = 0.01$) und bestimmen $c = c_\alpha$ so, dass gilt

$$P_{H_0}(|S_n - np_0| > c_\alpha) \leq \alpha \quad (\cong \alpha) \ .$$

Damit ist die Entscheidungsregel so gewählt, dass die Wahrscheinlichkeit einer Fehlentscheidung zugunsten H_1, obwohl H_0 zutrifft, $\cong \alpha$ ist. Ist nun n groß genug, dass die Normalapproximation der Binomialverteilung anwendbar ist (eine Faustregel hiefür ist $np_0(1-p_0) \geq 9$), so gilt

$$\begin{aligned} P_{H_0}(|S_n - np_0| > c_\alpha) &= P_{H_0}\left(\frac{|S_n - np_0|}{\sqrt{np_0(1-p_0)}} > \frac{c_\alpha}{\sqrt{np_0(1-p_0)}}\right) \\ &\cong 2(1 - \Phi(\frac{c_\alpha}{\sqrt{np_0(1-p_0)}})) \overset{!}{\leq} \alpha \end{aligned}$$

und somit

$$c_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{np_0(1-p_0)} \, .$$

In den nachstehenden Zahlenbeispielen wird die Rechenarbeit der jeweils folgenden Anwendungsbeispiele vorweggenommen.

Zahlenbeispiel: $n = 929, \ p_0 = \frac{3}{4}, \ \alpha = 0.05$.

Demgemäß sind $np_0 = 929 \cdot \frac{3}{4} = 696.75$, $\sqrt{np_0(1-p_0)} = \sqrt{929 \cdot \frac{3}{4} \cdot \frac{1}{4}} = 13.20$ und $\Phi^{-1}(1 - \frac{\alpha}{2}) = \Phi^{-1}(0.975) = 1.96$. Somit ist $c_{0.05} = 1.96 \cdot 13.20 = 25.872$.

[17] Der Begriff "*signifikant*" wurde im Jahre 1885 vom irischen Statistiker *Francis Y. Edgeworth* (1845 – 1926) geprägt und bedeutet so viel wie "es besteht tatsächlich ein Unterschied".

Anwendungsbeispiel: Gregor Mendels 3. Versuch über Pflanzenhybriden [18]

"3. Versuch: Farbe der Samenschale. Unter 929 *Pflanzen brachten* 705 *violette Blüten und graubraune Samenschalen;* 224 *hatten weiße Blüten und weiße Samenschalen. Daraus ergibt sich ein Verhältnis* $3.15 : 1$ *."*

Aus einer Reihe solcher Versuche zieht Mendel folgenden Schluss:

"In der ersten Generation der Hybriden treten nebst den dominierenden Merkmalen auch die rezessiven in ihrer vollen Eigentümlichkeit wieder auf, und zwar in dem entschieden ausgesprochenen Verhältnis $3 : 1$*, so dass aus je* 4 *Pflanzen dieser Generation* 3 *den dominierenden und eine den rezessiven Charakter haben."*

Man formuliere die Hypothesen H_0 und H_1 und berechne den P-Wert für den oben angegebenen Versuch.

In diesem Fall sind tatsächlich $n = 929$ und $p_0 = \frac{3}{4}$. Da das Versuchsergebnis $s_{929} = 705$ ist, ist die beobachtete Abweichung $|705 - 696.75| = 8.25$. Diese ist kleiner als der ermittelte kritische Wert $c_{0.05} = 25.872$, sodass keine Ursache besteht, die Hypothese H_0, dass die Wahrscheinlichkeit, dass eine Tochterpflanze bei Selbstbefruchtung den Phänotyp "violette Blüten und graubraune Samenschalen" besitzt, gleich $\frac{3}{4}$ ist, zu verwerfen.

Da die Hypothese H_0 eine Folgerung aus der Mendelsschen Theorie des Erbmechanismus' ist, könnte man Versuchsergebnisse dieser Art - und es gibt tatsächlich ungeheuer viele solche - so deuten, dass sie diese Theorie "bestätigen" oder "stützen", sodass es angebracht erscheint, die Theorie als "Arbeitshypothese" zu verwenden. Dies umsomehr, als man diesen Mechanismus inzwischen auf mulekularer Ebene zu verstehen glaubt.

Anmerkung: Die Alternativhypothese $H_1 : p \neq \frac{3}{4}$ ist in diesem Anwendungsbeispiel zu eng gefasst und müsste durch folgende Hypothese ersetzt werden. H_1': Der Mechanismus der Vererbung ist kein Zufallsmechanismus der in H_0 spezifizierten Art bzw. gar kein Zufallsmechanismus.

2a) $\boxed{\begin{array}{c} H_0 : p = p_0 \\ H_1 : p > p_0 \end{array}}$

[18] Gregor *Mendel* (1822 – 1884): Versuche über Pflanzenhybriden (1866). Erschienen in Ostwald's Klassiker der Exakten Wissenschaften, Band 6, Neue Folge, Vieweg, Braunschweig 1970

Es wird wieder davon ausgegangen, dass der Stichprobenumfang so groß ist, dass Normalapproximation der Binomialverteilung angebracht ist.

$$\begin{aligned} P_{H_0}\left(S_n > c_\alpha\right) &= P_{H_0}\left(\frac{S_n - np_0}{\sqrt{np_0\left(1-p_0\right)}} > \frac{c_\alpha - np_0}{\sqrt{np_0\left(1-p_0\right)}}\right) \\ &\cong 1 - \Phi(\frac{c_\alpha - np_0}{\sqrt{np_0\left(1-p_0\right)}}) \stackrel{!}{\leq} \alpha \end{aligned}$$

Letzteres bedeutet, dass c_α folgendermaßen zu wählen ist

$$c_\alpha \cong np_0 + \Phi^{-1}\left(1-\alpha\right) \cdot \sqrt{np_0\left(1-p_0\right)}\,.$$

Zahlenbeispiel: $n = 6590,\ p_0 = \frac{1}{2},\ \alpha = 0.05$

Demgemäß sind $np_0 = 6590 \cdot \frac{1}{2} = 3295\,,\ \sqrt{np_0\left(1-p_0\right)} = \sqrt{6590 \cdot \frac{1}{2} \cdot \frac{1}{2}} \cong 40.59$ und $\Phi^{-1}\left(1-\alpha\right) = \Phi^{-1}\left(0.95\right) = 1.645\,.$ Folglich ist $c_{0.05} = 3295 + 1.645 \cdot 40.59 \cong 3362$ [19].

Anwendungsbeispiel: Zum Geburtenanteil der Geschlechter

Im Bundesland Salzburg gab es im Jahr 1992 6590 Neugeborene. Davon waren

3375 Buben und 3215 Mädchen.

Ist dieses Ergebnis mit der Annahme verträglich, dass die Wahrscheinlichkeit der Geburt eines Buben und eines Mädchen jeweils gleich $1/2$ ist?

In diesem Fall sind $n = 6590$ und $p_0 = \frac{1}{2}\,.$

($X_i = 1$ ... das Geschlecht des i-ten Neugeborenen ist männlich. Die Voraussetzung, dass die Zufallsvariablen X_i unabhängig sind, ist bestenfalls dann erfüllt, wenn unter den Neugeborenen keine eineiigen Zwillingen sind.)

Bei der Formulierung der Alternativhypothese $H_1 : p > \frac{1}{2}$ geht man nicht bloß von der Gültigkeit der Mendelschen Theorie der Vererbung aus, sondern darüber hinaus von einer aus Vorinformationen genährten Theorie, dass die Wahrscheinlichkeit, dass das Geschlecht eines Neugeborenen männlich ist, größer als $\frac{1}{2}$ ist.

Da das Versuchsergebnis $s_{6590} = 3375 > c_{0.05} = 3362$ ist, ist die Hypothese H_0 zugunsten der Hypothese H_1 zu verwerfen.

[19] Für die Wahl $\alpha = 0.01$ wäre $\Phi^{-1}\left(0.99\right) = 2.326$ und somit $c_{0.01} \cong 3389\,.$

Anmerkung: Im folgenden berechnen wir den sogenannten *P-Wert* (das *beobachtete Signifikanzniveau*). Das ist die unter der Annahme von H_0 ermittelte Wahrscheinlichkeit des Ereignisses, dass die betrachtete Zufallsvariable[20] einen Wert annimmt, der mindestens so extrem ist, wie der beobachtete:

$$\begin{aligned} P_{H_0}(S_{6590} \geq 3375) &= P_{H_0}(S_{6590} - 3295 \geq 3375 - 3295) \\ &= P_{H_0}\left(\frac{S_{6590} - 3295}{\sqrt{6590 \cdot \frac{1}{2} \cdot \frac{1}{2}}} \geq \frac{80}{40.589}\right) \cong 1 - \Phi(1.971) \cong 1 - 0.9757 \\ &= 0.0243\,. \end{aligned}$$

Je kleiner der P-Wert ist, desto stärker ist die durch die Daten gegebene Evidenz gegen H_0.[21]

2b) $\boxed{\begin{array}{c} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{array}}$

Typische Beispiele für diesen Fall entstammen der Qualitätssicherung, wobei $0 < p_0 \ll 1$ ist. Im Regelfall ist der Stichprobenumfang klein, sodass die Normalaproximation der Binomialverteilung unangebracht ist.

Ein Anwendungsbeispiel aus der Qualitätssicherung

In einer Fabrik werden Dinge erzeugt, die durch qualitative Merkmale wie etwa Farbe, Oberflächenbeschaffenheit usw. charakterisiert sind. Man möchte verhindern, dass der Ausschussanteil p eine vorgegebene Schranke $p_0\,(= 0.04)$ übersteigt. Dazu werden laufend Stichproben von jeweils Umfang $n\,(= 20)$ gezogen und die Ergebnisse in einer Qualitätsregelkarte eingetragen.

Überschreitet die Anzahl S_n der Ausschussstücke in der Stichprobe eine gewisse Schranke c, so wird man schließen, dass die Produktion nicht zufriedenstellend ist, dass also $p > p_0$ gilt. Dies kann freilich ein Fehlschluss sein. Deshalb wählt man die Schranke $c = c_\alpha$ so, dass die Wahrscheinlichkeit eines Fehlschlusses, d.h. einer Entscheidung zugunsten $p > p_0$, obwohl $p \leq p_0$ zutrifft, gering ist; dass also für ein kleines $\alpha \in (0,1)$ gilt:

$$P_{B_{n,p_0}}(S_n > c_\alpha) = \sum_{k=c_\alpha+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leq \alpha\,.$$

[20] Im konkreten Fall ist diese S_{6590}. Im vorangehenden Beispiel ist diese $|S_{929} - 696.75|$, sodass der zugehörige P-Wert $P(|S_{929} - 696.75| \geq 8.22) \cong 0.532$ ist.

[21] Da der P-Wert im konkreten Fall $P(S_{6950} \geq 3375) \cong 0.0243 < 0.05 = \alpha$ ist, wird die oben getroffene Entscheidung bestätigt.

$c_{0.05}$ heißt Warngrenze und $c_{0.01}$ Kontrollgrenze.

Um die Wahrscheinlichkeit eines Fehlschlusses weiter zu reduzieren, wird im Fall, dass S_n die Warngrenze überschreitet, eine Zusatzstichprobe gezogen.

Man entnehme für das konkrete Beispiel Warn- und Kontrollgrenze einer Tabelle von Binomialverteilungen und trage diese Grenzen in der DGQ-Qualitätsregelkarte ein.

Obwohl für dieses Beispiel die Nullhypothese $H_0' : p = p_0$ zunächst keine adäquate Beschreibung ist, gehen wir von dieser aus und wählen zu einem vorgegebenen $0 < \alpha \ll 1$ den kritischen Wert $c_\alpha \in \mathbb{N}_0$ so, dass gilt

$$P_{H_0'}(S_n > c_\alpha) = \sum_{k=c_\alpha+1}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \overset{!}{\leq} \alpha \,.$$

Diese Vorgangsweise ist wegen der nachstehenden Behauptung gerechtfertigt, zumal aus dieser

$$P_{B_{n,p_0}}(S_n > c_\alpha) = \sup\left\{P_{B_{n,p}}(S_n > c_\alpha)\,,\ p \leq p_0\right\}$$

folgt. Die Größe $P_{B_{n,p_0}}(S_n > c_\alpha)$ nennt man das *Produzentenrisiko.*

Behauptung: Sei $c \in \{0, ..., n-1\}$ fest. Dann ist die durch

$$F_c(p) = \sum_{k=0}^{c} \binom{n}{k} p^k (1-p)^{n-k}$$

definierte Funktion $F_c : [0,1] \mapsto [0,1]$ streng monoton fallend.

Beweis: Wegen $\binom{n}{k+1}(k+1) = n\binom{n-1}{k} = \binom{n}{k}(n-k)$ gilt für $p \in (0,1)$

$$\begin{aligned}
\frac{dF_c(p)}{dp} &= \sum_{k=1}^{c} \binom{n}{k} k p^{k-1} (1-p)^{n-k} - \sum_{k=0}^{c} \binom{n}{k} (n-k)\, p^k (1-p)^{n-k-1} \\
&= \sum_{k=0}^{c-1} \binom{n}{k+1} (k+1)\, p^k (1-p)^{n-k-1} - \sum_{k=0}^{c} \binom{n}{k} (n-k)\, p^k (1-p)^{n-k-1} \\
&= \sum_{k=0}^{c-1} \left[\tbinom{n}{k+1}(k+1) - \tbinom{n}{k}(n-k)\right] \cdot p^k (1-p)^{n-k-1} - \\
&\quad - n\tbinom{n-1}{c} p^c (1-p)^{n-1-c} \\
&= -n\binom{n-1}{c} p^c (1-p)^{n-1-c} < 0\,.
\end{aligned}$$

3a) $\boxed{\begin{array}{c} H_0 : p = p_0 \\ H_1 : p = p_1 \end{array}}$ mit $p_0 < p_1 \leq 1$

Bei diesem Fall beschränken wir uns auf die Behandlung zweier grundverschiedener Beispiele. Anhand des Beispiels $p_0 = \frac{1}{3}$, $p_1 = \frac{2}{3}$, dessen Behandlung aufgrund der Symmetrie besonders einfach ist, werden die für den vorliegenden Fall typischen Begriffe vorgestellt. Das Beipiel $p_0 = \frac{3}{4}$, $p_1 = 1$ wird uns anschließend Gelegenheit geben, dem Wesen des Testens statistischer Hypothesen näher zu kommen.

Beispiel: Planung eines Modellversuchs

Um Schüler von der Funktionstüchtigkeit statistischen Schließens zu überzeugen, führt ein Lehrer folgenden Versuch durch. Er lässt den Schülern die Wahl, eine Urne entweder zu $1/3$ mit roten und zu $2/3$ mit weißen Kugeln zu bestücken. (Hypothese H_0) oder zu $2/3$ mit roten und zu $1/3$ mit weißen (Hypothese H_1). Um herauszufinden, welche der beiden Alternativen die Schüler gewählt haben, will der Lehrer zufällig und mit Zurücklegen eine ungerade, feste Anzahl n von Kugeln ziehen und sich für die Hypothese H_1 bzw. H_0 entscheiden, falls die Mehrzahl der gezogenen Kugeln rot bzw. weiß ist.

Er möchte den Stichprobenumfang n möglichst klein wählen, jedoch so, dass die Wahrscheinlichkeit für jede der möglichen Fehlentscheidungen ≤ 0.01 ist.

Die beiden Hypothesen H_0 und H_1 sind gemäß der Angabe durch folgende Urnenzusammensetzungen beschrieben.

$$H_0 \ : \ \left(\tfrac{1}{3}, \tfrac{2}{3}\right) \doteq \boxed{\bullet \ \circ \ \circ}$$

$$H_1 \ : \ \left(\tfrac{2}{3}, \tfrac{1}{3}\right) \doteq \boxed{\bullet \ \bullet \ \circ}$$

Bezeichne im folgenden S_n die Anzahl der roten Kugeln in der Stichprobe. Dann entscheiden wir uns

$$\begin{array}{l} \text{zugunsten} \quad H_1\,, \quad \text{wenn} \quad S_n > \frac{n}{2} \quad \text{und} \\ \text{zugunsten} \quad H_0\,, \quad \text{wenn} \quad S_n < \frac{n}{2}\,. \end{array}$$

Es gibt zwei Arten von Fehlentscheidungen[22]: Den

Fehler erster Art,	wenn man sich zugunsten	H_1	entscheidet,
	obwohl	H_0	zutrifft, und den
Fehler zweiter Art,	wenn man sich zugunsten	H_0	entscheidet,
	obwohl	H_1	zutrifft.

Da S_n unter der Hypothese H_0 gemäß einer Binomialverteilung $B_{n,1/3}$ verteilt und unter der Hypothese H_1 gemäß einer Binomialverteilung $B_{n,2/3}$ verteilt ist, ist die Wahrscheinlichkeit eines

Fehlers erster Art gleich	$P_{B_{n,1/3}}(S_n > n/2)$	und die eines
Fehlers zweiter Art gleich	$P_{B_{n,2/3}}(S_n < n/2)\,.$	

Aufgrund der Symmetrie der beiden Hypothesen gilt im vorliegenden Fall

$$P_{B_{n,2/3}}(S_n < n/2) = P_{B_{n,1/3}}(S_n > n/2)\,.$$

Damit reicht es im Folgenden, nur die Wahrscheinlichkeit des Fehlers erster Art zu betrachten. Wir tun dabei so, als würde die zu ermittelnde Lösung n groß genug sein, um die Anwendung der Normalapproximation der Binomialverteilung zuzulassen.

$$\begin{aligned} P_{B_{n,1/3}}(S_n &> n/2) = \\ &= P_{B_{n,1/3}}\left(\frac{S_n - \frac{n}{3}}{\sqrt{n\cdot\frac{1}{3}\cdot\frac{2}{3}}} > \frac{n\left(\frac{1}{2}-\frac{1}{3}\right)}{\sqrt{n\cdot\frac{1}{3}\cdot\frac{2}{3}}}\right) \\ &\cong 1 - \Phi(\sqrt{n}\frac{\frac{1}{2}-\frac{1}{3}}{\sqrt{\frac{1}{3}\cdot\frac{2}{3}}}) \\ &= 1 - \Phi(\sqrt{\frac{n}{8}}) \stackrel{!}{\leq} \alpha\,. \end{aligned}$$

Dies ist gleichbedeutend mit

$$n \geq 8\left[\Phi^{-1}(1-\alpha)\right]^2\,.$$

[22] In der angelsächsichen Literatur spricht man vom *type I error* bzw. *type II error*.

Für $\alpha = 0.01$ und damit $\Phi^{-1}(1-\alpha) = \Phi^{-1}(0.99) = 2.326$ ergibt dies wegen $n \geq 8 \cdot 2.326^2 \cong 43.282$

$$n \geq 45\,.$$

Damit ist wegen $45 \cdot \frac{2}{3} \cdot \frac{1}{3} = 10 \geq 9$ die Faustregel bestätigt und damit nachträglich die Anwendung der Normalapproximation der Binomialverteilung gerechtfertigt.

Im Folgenden wird ein Modellversuch vorgestellt, welcher den indirekten Schluss sowohl in seiner - jedem Mathematiker vertrauten - "reinen Form" als auch in seiner - in den Anwendungen vorwiegenden - "stochastischen Form" enthält.

Ein Modellversuch mit der "reinen" und der "stochastischen Form" des indirekten Schlusses

Wir gehen davon aus, dass wir wissen, dass eine Urne gemäß einer der beiden folgenden Urnenzusammensetzungen bestückt ist.

$$H_0 \;:\; \left(\tfrac{3}{4}, \tfrac{1}{4}\right) \;\doteq\; \boxed{\bullet\ \bullet\ \bullet\ \circ}$$

$$H_1 \;:\; (1,0) \;\doteq\; \boxed{\bullet\ \bullet\ \bullet\ \bullet}$$

Um zu entscheiden, welche der beiden Urnenzusammensetzungen tatsächlich vorliegt, ziehen wir aus der Urne n Kugeln zufällig und mit Zurücklegen. Wir entscheiden uns

zugunsten H_1, wenn alle gezogenen Kugeln rot sind und
zugunsten H_0, wenn mindestens eine der gezogenen Kugeln weiß ist.

Im vorliegenden Fall ist ein Fehler zweiter Art, also eine Entscheidung zugunsten H_0, obwohl H_1 zutrifft, mit Sicherheit ausgeschlossen, denn unter der Annahme, dass die Urne nur mit roten Kugeln bestückt ist (und die gezogenen Kugeln ordungsgemäß zurückgelegt werden), kann ihr keine weiße Kugel entnommen werden.

Ein Fehler erster Art ist hingegen keineswegs auszuschließen. Bezeichnet nämlich E_n das Ereignis, dass alle n gezogenen Kugeln rot sind[23], so ist

[23] Bezeichnet - wie im ersten Modellversuch - S_n die Anzahl der roten Kugeln in der Stichprobe, so ist das Ereignis $E_n = \{S_n = n\}$.

die Wahrscheinlichkeit eines Fehlers erster Art gleich

$$P_{H_0}(E_n) = (\frac{3}{4})^n \,.$$

Wählt man, wie im obigen Beispiel $\alpha = 0.01$, so ist diese Wahrscheinlichkeit genau dann ≤ 0.01, wenn

$$n \geq \frac{\ln 0.01}{\ln(3/4)} \cong 16.008 \,,$$

also wenn $n \geq 17$ ist.

Das folgenden Anwendungsbeispiel ist wieder der Genetik, einem der schönsten und ergiebigsten Anwendungsgebiete der Wahrscheinlichkeitsrechnung, entnommen. Um die Denkstruktur des Testens statistischer Hypothesen einzuüben, führen wir die Überlegungen in dessen Einkleidung nochmals aus.

Anwendungsbeispiel: Test auf Mischerbigkeit von Erbsen mittels Selbstbefruchtung

Von einer Erbsenpflanze soll herausgefunden werden, welcher der beiden Genotypen Aa oder AA dem Phänotyp A : violett-roter Blüten zugrunde liegt. Dabei ist das Allel A : violett-rote Blüten dominant und das Allel a : weiße Blüten rezessiv.

Dies wird mit Hilfe des Phänotyps einer gewissen Anzahl n (z.B. $n = 17$) von durch Selbstbefruchtung aus der vorliegenden Erbsenpflanze gewonnenen Tochterpflanzen erreicht.

Unter der Hypothese H_0 : Die Pflanze ist mischerbig (Genotyp Aa) liefert Selbstbefruchtung aufgrund der Mendelschen Gesetze, dass der Genotyp einer Tochterpflanze eine Binomialverteilung mit den Parametern $n = 2$ und $p = \frac{1}{2}$, d.h. die Verteilung $P = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$ auf der Menge $\{AA, Aa, aa\}$ besitzt[24]. Somit besitzt eine Tochterpflanze mit der Wahrscheinlichkeit $\frac{3}{4}$ den Phänotyp A und mit der Wahrscheinlichkeit $\frac{1}{4}$ den Phänotyp a.

Unter der Hypothese H_1 : Die Pflanze ist reinerbig (Genotyp AA) liefert Selbstbefruchtung Tochterpflanzen vom Genotyp AA, also vom Phänotyp A.

[24] Es wird tatsächlich gelegentlich die Meinung vertreten, dass es unter anderem seine Kenntnisse aus Wahrscheinlichkeitsrechnung waren, die *Gregor Mendel* $(1822 - 1884)$ zu seiner bahnbrechenden Theorie über den Mechanismus der Vererbung angeregt haben.

Zu H_1 : Wenn nur eine der n Tochterpflanzen den Phänotyp a hat, so kann die fragliche Pflanze nicht reinerbig sein. Somit ist die Hypothese H_1 mit Sicherheit auszuschließen.

Zu H_0 : Wenn hingegen alle n (z.B. $n = 17$) Tochterpflanzen den Phänotyp A haben, was wir als Ereignis E_n bezeichnen, so ist dies durchaus mit beiden Hypothesen verträglich. Unter der Hypothese H_0 ist ein solches Resultat allerdings unwahrscheinlich: Die zugehörige Wahrscheinlichkeit $P_{H_0}(E_n)$ ist nämlich

$$P_{H_0}(E_n) = (\frac{3}{4})^n \, (= (\frac{3}{4})^{17} < 0.01) \, .$$

Dies spricht zwar gegen die Hypothese H_0 - und man mag die Hypothese H_0 verwerfen - mit Sicherheit auszuschließen, wie oben H_1 , ist die Hypothese H_0 jedoch nicht. Verwirft man H_0 , obwohl H_0 zutrifft, so ist dies eine Fehlentscheidung.

Anmerkung zur Entsprechung

Testen von Hypothesen - Rechtssprechung

Zwischen dem Testen statistischer Hypothesen und der Rechtssprechung - insbesondere im Strafrecht - gibt es folgende Entsprechungen:

Testen von Hypothesen	**Rechtssprechung**
Nullhypothese	Unschuldsvermutung
Alternativhypothese	strafrechtlicher Tatbestand

In der statistischen Praxis wird eine *Nullhypothese* zugunsten einer *Alternativhypothese*[25] nur dann verworfen, wenn Beobachtungsergebnisse bzw. Ergebnisse eines Experiments, die mindestens so "extrem" wie die vorliegenden sind, unter der Annahme der Nullhypothese äußerst unwahrscheinlich sind.

Im Strafrecht ist stets von der *Unschuldsvermutung* auszugehen, d.h. von der Annahme, dass die eines *strafrechtlichen Tatbestands*[26] beschuldigte Person, unschuldig ist. Ein Schuldspuch ist nur dann zulässig, wenn die aufgrund

[25] Die Formulierung der beiden Hypothese erfolgt, wie aus den behandelten Beispielen ersichtlich ist, auf der Grundlage einer Theorie in einen bestimmten fachwissenschaftlichen Zusammenhang.

[26] Strafrechliche Tatbestände oder Delikte sind entweder Vergehen oder Verbrechen.

der Beweismittel[27] erhaltenen Ergebnisse der Unschuldsvermutung in hohem Maß widersprechen. Der Beweis gilt dann als erbracht, wenn die Verwirklichung des strafrechtlichen Tatbestands mit hoher Wahrscheinlichkeit als erwiesen angenommen werden kann.

[27] Beweismittel sind Indizien oder ein Geständnis.

2.3 TESTEN VON HYPOTHESEN (Teil 2): 2×2-Kontingenztafel

2.3.1 Bedingte Wahrscheinlichkeiten und Unabhängigkeit - Wiederholung

Gegeben sei eine Urne mit N - abgesehen von Farbe und Markierung - gleichartigen Kugeln. Davon sind

a ... schwarz und markiert
b ... weiß und markiert
c ... schwarz und unmarkiert
d ... weiß und unmarkiert

mit $a+b+c+d=N$.

Experiment: Eine Kugel wird zufällig gezogen. Die vier möglichen Elementarereignisse sind

A ... die gezogene Kugel ist schwarz
A^c ... die gezogene Kugel ist weiß
B ... die gezogene Kugel ist markiert
B^c ... die gezogene Kugel ist unmarkiert

Frage: Wie groß ist - unter Laplace-Annahme - die Wahrscheinlichkeit, dass die gezogene Kugel schwarz ist ?

Zur Beantwortung dieser Frage erstellen wir folgendes Schema

	A	A^c	
B	a	b	$a+b$
B^c	c	d	$c+d$
	$a+c$	$b+d$	N

Antwort a) Ohne weitere Information: $P(A)=\frac{a+c}{N}$

Antwort b) Mit der Vorinformation, dass die gezogene Kugel markiert ist (B eingetreten ist) [28]:

$$P(A/B)=\frac{a}{a+b}=\frac{\frac{a}{N}}{\frac{a+b}{N}}=\frac{P(A\cap B)}{P(B)}$$

[28] Wir setzen damit stillschweigend voraus, dass es überhaupt markierte Kugeln gibt, d.h. dass $a+b>0$ ist.

Definition 1: Seien $A, B \subseteq \Omega$ zwei Ereignisse und sei $P(B) > 0$. Dann heißt

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

die durch B bedingte Wahrscheinlichkeit von A.

Beispiel 1: Im Folgenden betrachten wir eine Urne, für die die Anzahl der schwarzen markierten Kugeln $a = 200$, die der schwarzen unmarkierten Kugeln $c = 300$ und die der weißen Kugeln $b + d = 1000$ ist. Damit ist die Anzahl der schwarzen Kugeln $a + c = 500$ und die Gesamtanzahl der Kugeln $N = 1500$. Daher ist

$$P(A) = \frac{500}{1500} = \frac{1}{3}.$$

Wir unterscheiden weiters drei unterschiedliche Aufteilungen der weißen Kugeln in markierte und unmarkierte Kugeln.

	A	A^c	
B	200	b	$200 + b$
B^c	300	$d = 1000 - b$	$1300 - b$
	500	1000	1500

Fall (i): $b = 800$, $d = 200$. In diesem Fall ist $P(A/B) = \frac{200}{1000} = \frac{1}{5} < P(A)$, d.h. das Ereignis B benachteiligt das Eintreten von A.

Fall (ii): $b = 200$, $d = 800$. In diesem Fall ist $P(A/B) = \frac{200}{400} = \frac{1}{2} > P(A)$, d.h. das Ereignis B begünstigt das Eintreten von A.

Fall (iii): $b = 400$, $d = 600$. In diesem Fall, bei welchem die Aufteilung der Menge der weißen Kugeln in markierte und unmarkierte proportional zu der bei den schwarzen Kugeln erfolgt - ist $P(A/B) = \frac{200}{600} = \frac{1}{3} = P(A)$, d.h. das Ereignis B beeinflusst das Eintreten von A nicht.

Aufgund von Definition 1 und $P(A/B) = P(A)$ gilt offensichtlich

$$P(A \cap B) = P(A) \times P(B) \qquad (1)$$

Definition 2: Zwei Ereignisse A und B heißen stochastisch unabhängig, wenn (1) gilt.

Anmerkung 1:

(a) Äquivalent zur obigen Bedingung ist

$$P(B) = 0 \quad \text{oder} \quad P(B) > 0 \text{ und } P(A/B) = P(A)\,.$$

Man beachte, dass bei der von uns gewählten Definition auf die Fallunterscheidung $P(B) = 0$ resp. > 0 verzichtet werden kann.

(b) Sei, wie gehabt, $P(B) = \frac{a+b}{N} > 0$. Dann sind wegen $P(A/B) - P(A) = \frac{a}{a+b} - \frac{a+c}{N} = \frac{a\times d - b\times c}{(a+b)N}$ die Ereignisse A und B genau dann stochastisch unabhängig, wenn gilt

$$a \times d - b \times c = 0\,.$$

Da diese Differenz bekanntlich die Determinante der Matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ ist, ist die stochastische Unabhängigkeit der beiden Ereignisse A und B im vorliegenden Fall gleichbedeutend mit der linearen Abhängigkeit der Vektoren

$$\begin{pmatrix} a & b \end{pmatrix} \text{ und } \begin{pmatrix} c & d \end{pmatrix}.$$

Anmerkung 2: Geht man davon aus, dass es zwei Urnen gibt, die folgendermaßen mit schwarzen und weißen Kugeln bestückt sind

a	...	Anzahl der schwarzen Kugeln in Urne 1
b	...	Anzahl der weißen Kugeln in Urne 1
c	...	Anzahl der schwarzen Kugeln in Urne 2
d	...	Anzahl der weißen Kugeln in Urne 2,

entspricht also - im Vergleich zum obigen Modell -

einer markierten Kugel	eine Kugel aus Urne 1	und
einer unmarkierten Kugel	eine Kugel aus Urne 2,	

so lässt sich die Größe $a \times d - b \times c$ wegen

$$\begin{aligned} a \times d - b \times c &= (a+b)(c+d)[\frac{a}{a+b} - \frac{c}{c+d}] && (2)\\ &= N(a+b)[\frac{a}{a+b} - \frac{a+c}{N}] && (3)\\ &= N[a - (a+b)\frac{a+c}{N}] && (3')\end{aligned}$$

auch als Maß für den Unterschied in der Bestückung der beiden Urnen bzw. als Maß für die Abweichung zweier Verteilungen interpretieren. Nämlich als Maß für die Abweichung der Verteilungen

$$\boxed{\frac{a}{a+b}}\boxed{\frac{b}{a+b}} \quad \text{und} \quad \boxed{\frac{c}{c+d}}\boxed{\frac{d}{c+d}} \quad \text{(gemäß (2))}$$

bzw. als Maß für die Abweichung der Verteilungen

$$\boxed{\frac{a}{a+b}}\boxed{\frac{b}{a+b}} \quad \text{und} \quad \boxed{\frac{a+c}{N}}\boxed{\frac{b+d}{N}} \quad \text{(gemäß (3))},$$

wobei man letztere dadurch erhält, dass man die Inhalte der beiden Urnen zu einer Urne vereinigt.

2.3.2 Assoziationsmaße

Wegen

$$P(A \cap B) = P(B) \times P(A) \quad \Longleftrightarrow \quad \frac{a}{N} = \frac{a+b}{N} \times \frac{a+c}{N}$$

und (3) gilt

$$a{\times}d - b{\times}c = \begin{cases} < 0 & \Longleftrightarrow \text{ das Ereignis } B \text{ benachteiligt das Eintreten von } A \\ = 0 & \Longleftrightarrow \text{ das Ereignis } B \text{ beeinflusst das Eintreten von } A \text{ nicht} \\ > 0 & \Longleftrightarrow \text{ das Ereignis } B \text{ begünstigt das Eintreten von } A\,. \end{cases}$$

Es liegt daher nahe, durch geeignete Normierung dieser Größe einen Koeffizienten zu definieren, der den Grad der Abhängigkeit der Ereignisse A und B misst.

Wir betrachten im folgenden zwei Koeffizienten dieser Art. Beide stammen vom schottischen Statistiker *George Udny Yule* $(1871 - 1951)$. Der im Artikel *"On the correlation of total pauperism with proportion of outrelief"* im Jahre 1895 eingeführte *Q-Koeffizient* hat zahlreiche Anwendungen in den Sozialwissenschaften. Der im Artikel *"On the methods of measuring the association between two variables"* im Jahre 1912 vorgeschlagene Φ-Koeffizient wird vorwiegend in der Psychologie und in den Erziehungswissenschaften verwendet. Wir werden in Abschnitt 2.3.3, B2) den Φ-Koeffizienten verwenden, da dieser die Normalapproximation der Hypergeometrischen Verteilung erlaubt.

Definition 3: Sei $a \times d + b \times c > 0$.[29] Dann heißt die Größe

$$Q = \frac{a \times d - b \times c}{a \times d + b \times c}$$

Q-Koeffizient.

Der *Q-Koeffizient* besitzt offensichtlich den Wertebereich $[-1, +1]$, wobei gilt

$$Q = \begin{cases} -1 \iff a \times d = 0 & \text{d.h. wenn es keine schwarzen markierten } \textbf{oder} \\ & \text{keine weißen unmarkierten Kugeln gibt} \\ 0 \iff a \times d - b \times c = 0 & \\ +1 \iff b \times c = 0 & \text{d.h. wenn es keine schwarzen unmarkierten } \textbf{oder} \\ & \text{keine weißen markierten Kugeln gibt.} \end{cases}$$

Definition 4: Sei - wie oben - $a \times d + b \times c > 0$. Dann heißt die Größe

$$\rho = \frac{a \times d - b \times c}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Korrelationskoeffizient (Φ*-Koeffizient*).

Behauptung: Der Φ-Koeffizient hat die Eigenschaft

$$-1 \leq \rho \leq 1\,,$$

wobei gilt

$$\rho = \begin{cases} -1 \iff a + d = 0 & \text{d.h. wenn es keine schwarzen markierten } \textbf{und} \\ & \text{keine weißen unmarkierten Kugeln gibt} \\ 0 \iff a \times d - b \times c = 0 & \\ +1 \iff b + c = 0 & \text{d.h. wenn es keine schwarzen unmarkierten } \textbf{und} \\ & \text{keine weißen markierten Kugeln gibt.} \end{cases}$$

Beweis: Diese Behauptung folgt unmittelbar aus der leicht nachzuprüfenden Darstellung der Differenz der Quadrate von Nenner und Zähler von ρ

$$(a+b)(a+c)(b+d)(c+d) - (a \times d - b \times c)^2 = N \times [(a+d)bc + (b+c)ad]\,.$$

[29] Diese Bedingung besagt, dass mindestens eine der beiden Diagonalen der $2 \times 2-$ Tafel besetzt ist.

Als Produkt der positiven Zahl N und der nichtnegativen Zahl $(a+d)bc+(b+c)ad$ ist diese nichtnegativ und offensichtlich

$$\begin{array}{llll} \text{im Fall} & ad-bc>0 & \text{genau dann} = 0 & \text{wenn} \quad b+c=0 \\ \text{im Fall} & ad-bc<0 & \text{genau dann} = 0 & \text{wenn} \quad a+d=0 \end{array}$$

ist.

Anmerkung 3: Aus der leicht nachzuprüfenden Beziehung

$$\begin{aligned} (a+b)(a+c)(b+d)(c+d)-(ad+bc)^2 \;&=\; ad\left[(a+d)(b+c)+b^2+c^2\right]+ \\ &\quad +bc\left[(a+d)(b+c)+a^2+d^2\right] \end{aligned}$$

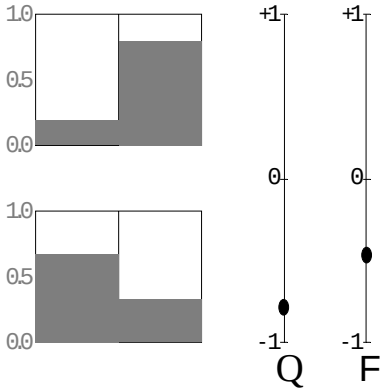
folgt $\sqrt{(a+b)(a+c)(b+d)(c+d)} \geq a\times d+b\times c$. Daher ist der Absolutbetrag des Φ-*Koeffizienten* stets kleiner oder gleich dem des Q-*Koeffizienten.*

Anmerkung 4: Bezieht man sich auf das Urnenmodell von Anmerkung 2, so lässt sich die Homogenität der Zusammensetzungen der beiden Urnen mit Hilfe des Φ-Koeffizienten (bzw. des Q-Koeffizienten) wegen (2) wie folgt mit Hilfe der Anteile $\frac{a}{a+b}$ und $\frac{c}{c+d}$ der schwarzen Kugeln in Urnen 1 bzw. Urne 2 charakterisieren

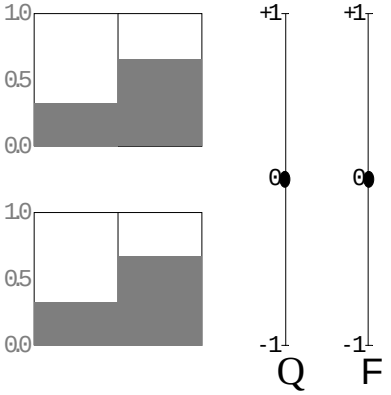
$$\rho = \begin{cases} <0 & \Longleftrightarrow \quad \frac{a}{a+b} < \frac{c}{c+d} \\ =0 & \Longleftrightarrow \quad \frac{a}{a+b} = \frac{c}{c+d} \\ >0 & \Longleftrightarrow \quad \frac{a}{a+b} > \frac{c}{c+d}\,. \end{cases}$$

Abschließend geben wir eine Darstellung der Histogramme der bedingten Verteilungen und die Q- und Φ-Koeffizienten für die Fälle (i), (ii) und (iii) aus Beispiel 1 an.

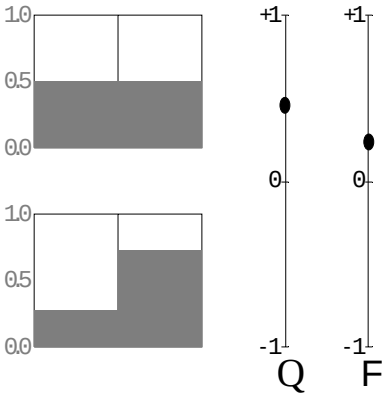
Beispiel 1 (Fortsetzung): $a = 200,\ c = 300,\ d = 1000 - b$



$b = 800 : Q = -0.71,\ \rho = -0,4$



$b = 400 : Q = 0,\ \rho = 0$



$b = 200 : Q = 0.45,\ \rho = 0.21$

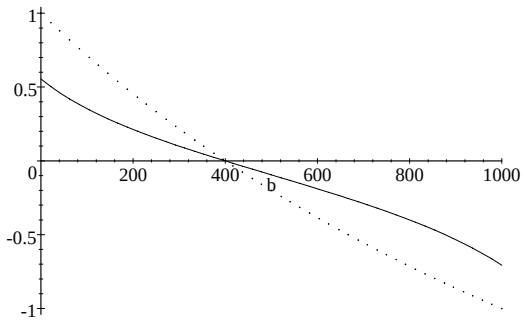


Abbildung: Abbildung der Funktionen $b \mapsto Q(200, b, 300, 1000-b)$ (punktiert) und $b \mapsto \rho(200, b, 300, 1000 - b),\ b \in [0, 1000]$

2.3.3 Statistische Verfahren

In zwei Urnen befinden sich jeweils schwarze und weiße Kugeln. Und zwar in der in Anmerkung 2 beschriebenen Zusammensetzung

	schwarze Kugeln	weiße Kugeln	
Urne 1	a	b	$a+b$
Urne 2	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d\,.$

Der zugehörige Φ-Koeffizient sei ρ .

A) Punktschätzer für den Φ-Koeffizienten

Im Folgenden wird ein naheliegendes Verfahren angegeben, um ρ zu schätzen.

Zu diesem Zweck wird aus Urne 1 eine Stichprobe vom Umfang $n \leq a+b$ und aus Urne 2 eine Stichprobe vom Umfang $N-n \leq c+d$ entnommen - und zwar jeweils durch Ziehen ohne Zurücklegen. Es seien

X ... die Anzahl der schwarzen Kugeln in der ersten Stichprobe und
S ... die Anzahl der schwarzen Kugeln in beiden Stichproben.

Durch diese Stichproben wird der obigen 2×2- Tafel für die beiden Urnen die folgende 2×2- Tafel für die beiden Stichproben nachgebildet:

	schwarze Kugeln	weiße Kugeln	
Stichprobe aus Urne 1	X	$n - X$	n
Stichprobe aus Urne 2	$S - X$	$N + X - (n + S)$	$N - n$
	S	$N - S$	N.

Der Φ-Koeffizient der Stichprobe ergibt sich durch Berücksichtigung von $(3')$ und indem man $a = X,\ a + b = n,\ a + c = S$ (somit $b + d = N - S$) und $c + d = N - n$ in Definition 4 einsetzt.

Somit ist

$$\hat{\rho}_{(n,N)}(X, S) = \frac{N(X - n\frac{S}{N})}{\sqrt{n\, S(N - S)(N - n)}}\,.$$

ein naheliegender Schätzer des Φ-Koeffizienten ρ, wobei der Umfang n der Stichprobe aus Urne 1 und der Umfang N der Gesamtstichprobe Parameter sind.

Beispiel 2: Für das in Abschnitt 2.3.4 angegebene Anwendungsbeispiel sind $n = 139,\ N = 279,\ x = 17$ und $s = 48$. Der Schätzwert ist daher

$$\hat{\rho}_{(139,279)}(17, 48) = \frac{279\,\big(17 - 139 \times \frac{48}{279}\big)}{\sqrt{139 \times\ 48 \times (279 - 48) \times (279 - 139)}} = -\,0.13132\,.$$

B) Testen von Hypothesen

Im Folgenden soll eine Entscheidungsregel angegeben werden, die es erlaubt, sich zwischen zwei Hypothesen hinsichtlich der Homogenität der Zusammensetzung der beiden Urnen zu entscheiden. Der Umstand, dass es nur zwei Farben gibt, erlaubt es, die Hypothesen hinsichtlich des Anteils einer der beiden Farben zu fomulieren.

Die *Nullhypothese* bestehe stets darin, dass der Anteil der schwarzen Kugeln in beiden Urnen gleich ist.

Bei der *Alternativhypothese* unterscheiden wir folgende beiden Fälle

1) der Anteil der schwarzen Kugeln in den beiden Urnen ist verschieden,

2a) der Anteil der schwarzen Kugeln in Urne 1 ist kleiner als der in Urne 2,[30]

[30] Die Formulierung einer einseitigen Alternativhypothese setzt bekanntlich eine entsprechende fachspezifisch begründete Vermutung voraus.

2b) der Anteil der schwarzen Kugeln in Urne 1 ist größer als der in Urne 2.

Diese Hypothesen lassen sich bequem mit Hilfe der Φ-Koeffizienten ρ formal ausdrücken:

1) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}}$ 2a) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{array}}$ 2b) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{array}}$

Wir werden lediglich die Fälle 1) und 2a) behandeln, da der Fall 2b) analog zu Fall 2a) zu behandeln ist.

Das Prinzip, dessen man sich beim Testen bedient, ist, dass man von der tatsächlich beobachteten Anzahl s der schwarzen Kugeln in der Gesamtstichprobe ausgeht und prüft, ob deren Aufteilung auf beide Urnen - bis auf eine Zufallsschwankung - proportional ist. Man betrachtet somit folgende 4-Felder-Tafel, bei der nunmehr nicht nur die Zeilen- sondern auch die Spaltensummen fest sind.

	schwarze Kugeln	weiße Kugeln	
Stichprobe aus Urne 1	X_n	$n - X_n$	n
Stichprobe aus Urne 2	$s - X_n$	$N + X_n - (n + s)$	$N - n$
	s	$N - s$	N .

Unter Annahme der Nullhypothese H_0 und $S = s$ ist die Zufallsvariable X_n gemäß einer Hypergeometrischen Verteilung mit den Parametern n, N und s verteilt. Knapp dargestellt gilt

für die Verteilung der Zufallsvariablen X_n	$X_n \sim H_{n,N,s}$
und somit für deren Erwartungswert	$E(X_n) = n\frac{s}{N}$
und deren Varianz	$V(X_n) = n\frac{s}{N}(1 - \frac{s}{N})(1 - \frac{n-1}{N-1})$.

B1) Verfahren für kleine Stichpobenumfänge: Der Exakte Test von Fisher-Yates und Irwin [31]

Die oben beschriebene Vorgangsweise, welche die Statistiker *Fisher* und *Yates* und unabhängig von diesen *Irwin* im Jahre 1935 vorgeschlagen haben, lässt auch die Behandlung im Fall kleiner Stichprobenumfänge n zu.

1) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}}$

Sei x die tatsächlich beobachtete Anzahl der schwarzen Kugeln in der Stichprobe aus Urne 1. Nun ist unter der obigen Annahme der zu erwartende Wert von X_n gleich $n\frac{s}{N}$. Die Abweichung unseres Beobachtungswerts vom Erwartungswert ist somit $d = \left|x - n\frac{s}{N}\right|$. Weiters seien $u = \left\lfloor n\frac{s}{N} - d \right\rfloor$ und $o = \left\lceil n\frac{s}{N} + d \right\rceil$.

Der zugehörige P-Wert ist demnach die Wahrscheinlichkeit, dass X_n vom Erwartungswert um mindestens so viel abweicht wie beobachtet und ist demnach

$$P_{H_{n,N,s}}\left(\left|X_n - n\frac{s}{N}\right| \geq d\right) = \sum_{k=0}^{u} \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}} + \sum_{k=o}^{n} \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}\,.$$

2a) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{array}}$

Sei wieder x die tatsächlich beobachtete Anzahl der schwarzen Kugeln in der Stichprobe aus Urne 1. Dann ist der zugehörige P-Wert die Wahrscheinlichkeit, dass X_n kleiner oder höchstens gleich dem tatsächlich beobachteten Wert x ist, und somit

$$P_{H_{n,N,s}}\left(X_n \leq x\right) = \sum_{k=0}^{x} \frac{\binom{s}{k}\binom{N-s}{n-k}}{\binom{N}{n}}\,.$$

B2) Verfahren für große Stichprobenumfänge

Für große Stichprobenumfänge [32] n und $N - n$ benützen wir für das Testen der genannten Hypothesen den Schätzer für den Φ-Koeffizienten ρ als

[31] *Frank Yates* $(1902 - 1994)$, *Joseph Oscar Irwin* $(1898 - 1982)$, englische Statistiker

[32] als Faustregel kann im vorliegenden Fall $n\frac{s}{N}(1 - \frac{s}{N})(1 - \frac{n-1}{N-1}) \geq 9$ gelten.

Teststatistik. Da dieser Schätzer unter den oben getroffenen Annahmen die Form

$$\hat{\rho}_{(n,N)}(X_n, s) = \frac{X_n - n\frac{s}{N}}{\sqrt{n\frac{s}{N}(1-\frac{s}{N})(N-n)}} = \frac{1}{\sqrt{N-1}} \frac{X_n - n\frac{s}{N}}{\sqrt{n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n-1}{N-1})}}$$

hat und X_n gemäß einer Hypergeometrischen Verteilung $H_{n,N,s}$ verteilt ist, ist die Größe

$$\sqrt{N-1} \times \hat{\rho}_{(n,N)}(X_n, s) = \frac{X_n - n\frac{s}{N}}{\sqrt{n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n-1}{N-1})}}$$

die vermittels Erwartungswert und Varianz standardisierte Zufallsvariable von X_n und daher aufgrund des Zentralen Grenzverteilungssatzes[33] für große Stichprobenumfänge n annähernd Standardnormalverteilt. Diesen Umstand machen wir uns im Folgenden zu Nutze.

1) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}}$

Sei x die tatsächlich beobachtete Anzahl der schwarzen Kugeln in der Stichprobe aus Urne 1 und $d = \left| x - n\frac{s}{N} \right|$. Dann ist der zugehörige P-Wert unter Berücksichtigung der Stetigkeitskorrektur

$$P_{H_{n,N,s}} \left(\left| X_n - n\frac{s}{N} \right| \geq d - 0.5 \right) \cong 2(1 - \Phi(\frac{d - 0.5}{\sqrt{n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n-1}{N-1})}})) \,,$$

wobei $z \mapsto \Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung ist.

2a) $\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{array}}$

Sei wieder x die tatsächlich beobachtete Anzahl der schwarzen Kugeln in der Stichprobe aus Urne 1. Dann ist der zugehörige P-Wert unter Berücksichtigung der Stetigkeitskorrektur

$$P_{H_{n,N,s}} \left(X_n \leq x + 0.5 \right) \cong \Phi(\frac{x + 0.5 - n\frac{s}{N}}{\sqrt{n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n-1}{N-1})}}) \,.$$

[33] Stichwort: Normalapproximation der Hypergeometrischen Verteilung

2.3.4 Fallstudie

Vorform eines kontrollierten Experiments[34]

"1601 schickte die East Indian Company ihre ersten Expeditionen nach Indien. James Lancaster fuhr mit einem großen und drei kleinen Schiffen los. Jeder Seemann auf dem großen Schiff erhielt täglich drei Löffel Zitronensaft, auf den kleinen dagegen nicht. Am Kap der guten Hoffnung waren 110 der 278 Seemänner auf den drei kleinen Schiffen tot. Auf dem großen Schiff erkrankten nur wenige.

Natürlich hat Lancaster über seine Erfahrungen berichtet. Die britische Admiralität hat jedoch diesen Bericht ignoriert. Kein Wunder: Es kam oft vor, dass von mehreren Schiffen eines von Skorbut verschont blieb. Der Kapitän des glücklichen Schiffes versuchte, sein Glück einem bestimmten Zusatz in der Diät zuzuschreiben, der bei den anderen Schiffen fehlte. Man hatte schon zahllose 'Wundermittel' gegen Skorbut. Dazu kam ein weiteres - Zitronensaft - hinzu."

Die Wirkung einer Behandlung wird also durch einen Vergleich ermittelt; dem Vergleich der Ergebnisse einer *Versuchsgruppe* und der einer *Kontrollgruppe.* Dabei spricht man von einem *kontrollierten Versuch*, wenn der Versuchsplaner bestimmt, welcher Proband (Patient, usw., allgemein: Einheit) eine Behandlung bekommt und welcher nicht. (Andernfalls spricht man von einer *Beobachtungsstudie.* So sind z.B. Studien über den Effekt des Rauchens üblicherweise Beobachtungsstudien, da Raucher und Nichtraucher vorgegebene Gruppen sind.)

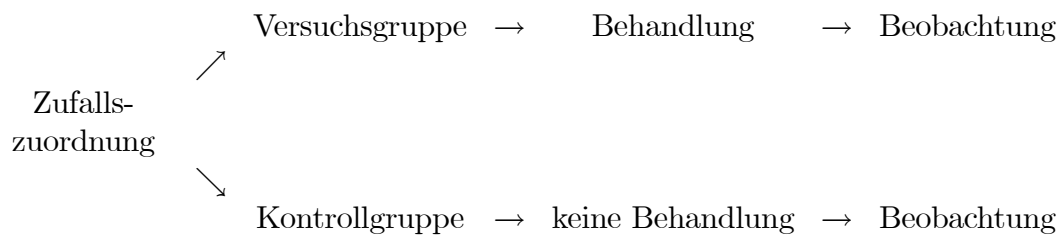
Selbstverständlich wäre ein Resultat wertlos, wenn ein erheblicher Unterschied in der Struktur der beiden Gruppen bestünde (wenn z.B. *Lancaster* kraftstrotzende Seeleute auf dem großen Schiff und gesundheitlich arg mitgenommene auf den anderen Schiffen versammelt hätte!). Es kommt also ersichtlich auf eine "gute" Zuordnung der Einheiten zu den beiden Gruppen an.

[34] Diese Fallstudie stammt - wie Beispiel 2 - aus [23].

Kontrollierte Versuche wurden seit der Mitte des 19. Jahrhunderts in großem Umfang in der landwirtschaftlichen Forschung in Großbritannien verwendet. Agrarier mußten versuchen, hinsichtlich Bodentyp, Fruchtbarkeit und anderen Faktoren äquivalente Gruppen von Einheiten (kleine Parzellen von Land) zu erhalten. Und dies ist insbesondere deswegen schwierig, weil man manche Faktoren von vornherein gar nicht kennt.

Der Durchbruch gelang in den 20er Jahren des vergangenen Jahrhunderts dem englischen Statistiker und Genetiker *Ronald Aylmer Fisher* $(1890 - 1962)$, der damals an der nun sehr traditionsreichen *landwirtschaftlichen Forschungsstation Rothamsted*[35] arbeitete. *Fisher* erkannte, dass äquivalente Gruppen am besten durch zufällige Zuordnung der Einheiten zu den Gruppen realisiert werden. Gerade so, wie eine (einfache) Zufallsstichprobe mit großer Wahrscheinlichkeit repräsentativ für die Grundgesamtheit ist, so ist eine Zufallsauswahl von (sagen wir) der Hälfte der verfügbaren Einheiten sehr wahrscheinlich zwei Gruppen (bestehend aus den ausgewählten bzw. nicht-ausgewählten Elementen) erzeugt, die in jeder Beziehung ähnlich sind.

Der einfachste *randomisierte kontrollierte Versuch* ist also von der Form



Die Zufallsstichprobe hat mit der Zufallszuordnung noch die weitere entscheidende Eigenschaft gemein: sie ist der Anwendung der Wahrscheinlichkeitsrechnung zugänglich.

[35] *Rothamsted Empirical Station* (in Harpenden, nördlich von London)

Beispiel 2 (Fortsetzung): Vitamin C und Erkältung

"Schon lange besteht der Glaube, dass Vitamin C gegen gewöhnliche Erkältung hilft. Dieser Volksglaube wurde im Allgemeinen nicht von Ärzten, Ernährungsfachleuten und Behörden geteilt.

1970 greift der doppelte Nobelpreisträger Linus Pauling in die Diskussion ein. In seinem Buch 'Vitamin C and the Common Cold' behauptet er, dass massive Dosen von Vitamin C Erkältungen verhindern oder mildern können. Insbesondere behauptet er, dass die regelmäßige Einnahme von einem Gramm Vitamin C pro Tag die Häufigkeit der Erkältungen um 45% reduziert und die Dauer der Krankheit um 60%. Der Chemiker Pauling wurde von Medizinern angegriffen, da seine Behauptungen auf sehr dürftiger Evidenz beruhten. Daraufhin publizierte er 1971 eine Übersicht über die einschlägige Evidenz. Er mußte zugeben, dass nur vier doppelblinde randomisierte kontrollierte Experimente dieses Problem untersuchten (bei einem doppelblinden Versuch wissen sowohl Patient wie auch Arzt nicht, ob der Patient der Versuchs- oder der Kontrollgruppe angehört). Die umfangreicheste Studie, die Pauling 1971 zur Verfügung stand, stammte vom Schweizer Arzt G. Ritzel (1961). Es war eine doppelblinde Studie von 279 französischen Skifahrern an einem Ski-Ort. Die 2 × 2-Tabelle in der folgenden Abbildung zeigt das Ergebnis.

	erkältet	*nicht erkältet*	
Vitamin C	17	122	139
Placebo	31	109	140
	48	231	279

Da es sich um einen randomisierten Versuch handelt, kommt jeder der $\binom{279}{139}$ Möglichkeiten, von 279 Leuten eine Versuchsgruppe von 139 auszuwählen, dieselbe Wahrscheinlichkeit zu.

Unter der Hypothese, dass kein Unterschied in der Wirkung von Vitamin C besteht, dürften Unterschiede bei der Aufteilung der Erkälteten auf die Versuchs- und Kontrollgruppe nur auf den Zufall zurückzuführen sein.

Der Alternative, dass Vitamin C doch erkältungshemmend wirkt, liegt die Frage nahe, wie gut die beobachtete Aufteilung

(von 17 Erkälteten in der Versuchsgruppe und 31 in der Kontrollgruppe) oder eine die Alternative noch besser stützende (von i bzw. $48-i$, $i < 17$) rein zufällig zu erklären ist. Das adäquate Maß dafür ist die Wahrscheinlichkeit dieses Ereignisses unter der Hypothese, dass kein Unterschied in der Wirkung von Vitamin C und Placebo besteht.

Da die Anzahl der Möglichkeiten, dass genau i der Erkälteten der Versuchsgruppe zufallen, $\binom{48}{i}\binom{231}{139-i}$ ist, ist die gefragte Wahrscheinlichkeit

$$P(X_{139} \leq 17) = \sum_{i=0}^{17} \frac{\binom{48}{i}\binom{231}{139-i}}{\binom{279}{139}} = 0.020523\,.$$

(Unter Berücksichtigung der Normalapproximation der Hypergeometrischen Verteilung erhält man übrigens

$$P(X_{139} \leq 17.5) \cong \Phi(-2.0312) = 0.02112\,.\,)$$

2.4 DER χ^2-TEST

2.4.1 Einleitung

Wir beginnen mit einer einfachen Anwendung des χ^2-Tests[36], der die nötigen allgemeinen Betrachtungen folgen.

Beispiel: Es sollte untersucht werden, ob ein bestimmter Würfel fair ist. Zu diesem Zweck wurde der Würfel $n = 600$ mal geworfen. Die Anzahlen b_j der Würfe mit der Augenzahl $j \in \{1, ..., 6\}$ waren wie folgt:

b_1	b_2	b_3	b_4	b_5	b_6
142	79	92	102	110	75

Kann auf Grund dieses Ergebnisses die Annahme (Hypothese H_0) aufrecht erhalten werden, dass der Würfel fair ist?

Seien $X_1, ..., X_n$ unabhängige, identisch verteilte Zufallsvariable mit Wertebereich $W_{X_1} = \{\omega_0, \omega_1, ..., \omega_m\}$ und Verteilung $P = (p_0, p_1, ..., p_m)$ mit positiven Wahrscheinlichkeiten. Weiters sei $\mathbb{B}_m^{(n)} = (B_0, B_1, ..., B_m)$ der Vektor der Ausfallshäufigkeiten

$$B_j = |\{\, i \in \{1, ..., n\} : X_i = \omega_j\}| \;, \quad j \in \{0, 1, ..., m\}\,.$$

(Somit ist $\sum_{j=0}^{m} B_j = n$). Dann gilt

- Der Zufallsvektor $\mathbb{B}_m^{(n)}$ ist Multinomialverteilt mit den Parametern n und $P = (p_0, p_1, ..., p_m)$, d.h.

  $$P\left((B_0, ..., B_m) = (k_0, ..., k_m)\right) = \binom{n}{k_0, ..., k_m} p_0^{k_0} \times ... \times p_m^{k_m}$$

  mit $(k_0, ..., k_m) \in \mathbb{N}_0^{m+1} : \sum_{j=0}^{m} k_j = n$.

- Somit sind die Zufallsvariablen B_j Binomialverteilt mit den Parametern n und p_j und es gelten daher

  $E(B_j) = n\,p_j$ und $V(B_j) = n\,p_j(1 - p_j)\,, \quad j \in \{0, ..., m\}$.

[36] Der χ^2-Test ist eine Erfindung des englischen Statistikers *Karl Pearson*. Sein bahnbrechender Artikel erschien im Jahr 1900. Der χ^2-Test ist der bekannteste Vertreter der sogenannten *Anpassungstests* (*Goodness-of-fit tests*).

Wir definieren nun ein "Abstandsmaß" des Zufallsvektors $\mathbb{B}_m^{(n)} = (B_0, B_1, ..., B_m)$ der Beobachtungen vom Vektor $\mathbb{E}_m^{(n)} = (np_0, np_1, ..., np_m)$ der zugehörigen Erwartungswerte, und zwar durch das n-fache des Erwartungswerts der Quadrate der relativen Abweichungen

$$\chi^2_m\,(n,P) = n \times \sum_{j=0}^{m} p_j (\frac{B_j/n - p_j}{p_j})^2 = \sum_{j=0}^{m} \frac{(B_j - n\,p_j)^2}{n\,p_j}\,.$$

Für den Spezialfall $m = 1$ können wir uns mit den uns zur Verfügung stehenden Mitteln davon überzeugen, dass der dadurch definierte "χ^2-Abstand", der übrigens keineswegs die Eigenschaften einer Metrik besitzt, statistisch brauchbar ist. Wegen $B_0 + B_1 = n$ und $n\,p_0 + n\,p_1 = n$ gilt nämlich

$$\begin{aligned}
\sum_{j=0}^{1} \frac{(B_j - n\,p_j)^2}{n\,p_j} &= \frac{((n - B_1) - (n - n\,p_1))^2}{n\,(1-p_1)} + \frac{(B_1 - n\,p_1)^2}{n\,p_1} \\
&= (B_1 - n\,p_1)^2\,(\frac{1}{n\,(1-p_1)} + \frac{1}{n\,p_1}) \\
&= (B_1 - n\,p_1)^2\,\frac{1}{n\,p_1(1-p_1)} \\
&= (\frac{B_1 - n\,p_1}{\sqrt{n\,p_1(1-p_1)}})^2\,.
\end{aligned}$$

Nun besagt der Satz von DeMoivre-Laplace

$$\lim_{n\to\infty} P\left(\frac{B_1 - n\,p_1}{\sqrt{n\,p_1(1-p_1)}} \le z\right) = \Phi(z)$$

mit $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\,\pi}} e^{-x^2/2}\,dx$, $z \in \mathbb{R}$. Somit gilt für $z > 0$

$$\lim_{n\to\infty} P\left((\frac{B_1 - n\,p_1}{\sqrt{n\,p_1(1-p_1)}})^2 \le z\right) = 2\,\Phi(\sqrt{z}) - 1$$

und daher wegen $\frac{d}{dz}\,(2\,\Phi(\sqrt{z}) - 1) = \frac{1}{\sqrt{2\,\pi\,z}} e^{-z/2}$

$$\lim_{n\to\infty} P\left((\frac{B_1 - n\,p_1}{\sqrt{n\,p_1(1-p_1)}})^2 \le z\right) = \int_0^z \frac{1}{\sqrt{2\,\pi\,x}} e^{-x/2} dx\,.$$

Die Funktion

$$f_1(x) = \frac{1}{\sqrt{2\,\pi\,x}} e^{-x/2}\,,\ x \in (0,\infty)\,,$$

ist der Spezialfall der Klasse der Dichtefunktionen der χ^2-*Verteilungen mit* m *Freiheitsgraden* für den Spezialfall $m = 1$.

Definition: Es sei $m \in \mathbb{N}$. Dann heißt die stetige Verteilung mit dem Träger $(0,\infty)$ und der Dichtefunktion

$$f_m(x) = \frac{x^{m/2-1} e^{-x/2}}{2^{m/2}\,\Gamma(m/2)}$$

χ^2-*Verteilung mit* m *Freiheitsgraden.* Dabei ist $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\,dx\,,\ \alpha > 0$, die sogenannte *Gamma-Funktion.*

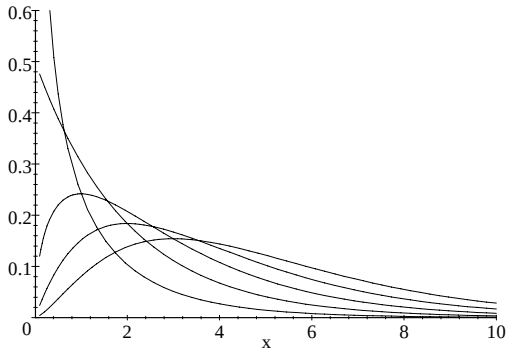


Abbildung der Dichtefunktionen $f_m(x)$ für $m = 1, 2, 3, 4$ und 5

Anmerkung 1: Erwartungswert und Varianz der χ^2-Verteilung mit m Freiheitsgraden sind m bzw. $2m$. Nun ist der Erwartungswert der χ^2-Statistik

$$E\left(\chi^2_m(n,P)\right) = m\,,$$

also tatsächlich gleich dem Erwartungswert der χ^2-Verteilung mit m Freiheitsgraden. Die zugehörige Varianz ist

$$V\left(\chi^2_m(n,P)\right) = 2m + \frac{1}{n}[\sum_{j=0}^{m} \frac{1}{p_j} - (m+2)^2 + 3]\,,$$

also nur asymptotisch gleich der Varianz $2m$ dieser Verteilung.

Der χ^2-Test beruht auf dem von *Karl Pearson* im Jahr 1900 veröffentlichten Satz, dessen Gültigkeit man aus dem behandelten Spezialfall für $m = 1$ und aus Anmerkung 1 vermuten kann.

Satz: Seien $m \in \mathbb{N}$ fest und $P = (p_0, p_1, ..., p_m)$ eine beliebige Wahrscheinlichkeitsverteilung mit positiven Werten p_j. Dann gilt

$$\lim_{n\to\infty} P\left(\chi^2_m(n,P) \le z\right) = \int_0^z f_m(x)\, dx\,.$$

Statistische Folgerung des Satzes von Pearson: Seien $X_1, ..., X_n$ unabhängige, identisch verteilte Zufallsvariable mit Wertebereich $W_{X_1} = \{\omega_0, \omega_1, ..., \omega_m\}$ und bezeichne $V_m = \{P' = (p'_0, p'_1, ..., p'_m) : p'_j \ge 0\ \forall\ j \in \{0, 1, ..., m\},\ \sum_{j=0}^m p'_j = 1\}$ die Menge der Wahrscheinlichkeitsverteilungen auf $\{\omega_0, \omega_1, ..., \omega_m\}$. Ferner sei $P = (p_0, p_1, ..., p_m) \in V_m$ derart, dass alle p_j positiv sind. Um die

- Hypothese $H_0 : X_1 \sim P = (p_0, p_1, ..., p_m)$

  gegen die

- Hypothese H_1 : Es gibt eine Verteilung $Q \in V_m \setminus \{P\}$ derart, dass $X_1 \sim Q$

auf einem Niveau $\cong \alpha$ mit $0 < \alpha \ll 1$ zu testen, kann man sich des folgenden (verteilungsfreien) Verfahrens bedienen. Man lehne H_0 zugunsten H_1 ab, wenn

$$\chi^2 = \sum_{j=0}^m \frac{(b_j - n\,p_j)^2}{n\,p_j} \ge \chi^2_{m,1-\alpha}$$

gilt. Dabei ist der kritische Wert $\chi^2_{m,1-\alpha}$ das $(1-\alpha)$-Quantil der Verteilungsfunktion der χ^2-Verteilung mit m Freiheitsgraden.

Im Folgenden ist die Tabelle der kritischen Werte $\chi^2_{m,1-\alpha}$ für $m \in$

$\{1, ..., 10\}$ und $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$ angegeben.

$m \setminus \alpha$	0.5	0.2	0.1	0.05	0.01
1	0.455	1.642	2.706	3.841	6.635
2	1.386	3.219	4.605	5.991	9.210
3	2.366	4.642	6.251	7.815	11.345
4	3.357	5.989	7.779	9.488	13.277
5	4.351	7.289	9.236	11.070	15.086
6	5.348	8.558	10.645	12.592	16.812
7	6.346	9.803	12.017	14.067	18.475
8	7.344	11.030	13.362	15.507	20.090
9	8.343	12.242	14.684	16.919	21.666
10	9.342	13.442	15.987	18.307	23.209

Anmerkung 2: Da der *Satz von Pearson* eine asymptotische Aussage ist, ist ein Niveau $\cong \alpha$ nur dann gesichert, wenn n "hinreichend groß" ist. Eine gebräuchliche Faustregel, dies zu überprüfen, ist $n \times \min\{p_0, p_1, ..., p_m\} \geq 5$.

Fortsetzung des einführenden Beispiels:

Da die Hypothese H_0, dass der Würfel fair ist, durch die Wahrscheinlichkeitsverteilung $P = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ beschrieben wird und der Stichprobenumfang $n = 600$ ist, ist $n \times \min\{p_0, p_1, ..., p_m\} = 600 \times \frac{1}{6} = 100$. Also ist die Faustregel bestens erfüllt, sodass wir den χ^2-Test bedenkenlos anwenden können. Der beobachtete χ^2-Wert ist gleich

$$\begin{aligned} \chi^2 &= \frac{(142-100)^2}{100} + \frac{(79-100)^2}{100} + \frac{(92-100)^2}{100} + \\ &\quad \frac{(102-100)^2}{100} + \frac{(110-100)^2}{100} + \frac{(75-100)^2}{100} \\ &= \frac{1}{100}[42^2 + 21^2 + 8^2 + 2^2 + 10^2 + 25^2] \\ &= 29.98\,. \end{aligned}$$

Die Anzahl der Freiheitsgrade ist $6 - 1 = 5$. Wegen $29.98 \geq \chi^2_{5,1-0.01} = 15.086 > \chi^2_{5,1-0.05} = 11.07$ ist die Hypothese H_0, dass der Würfel fair ist, nicht bloß auf dem Niveau $\alpha = 0.05$, sondern auch auf dem Niveau $\alpha = 0.01$ zu verwerfen. Der P-Wert ist übrigens $P(\chi^2_5 \geq 29.98) \cong 1.5 \times 10^{-5}$. Demgemäß ist die durch die Daten gegebene Evidenz gegen H_0 sehr stark.[37]

[37] *Ronald A. Fisher* hat maßgeblich dazu beigetragen, die Niveaus $\alpha = 0.05$ und

2.4.2 Anpassung von Modellen

Im Folgenden behandeln wir die Anpassung von Wahrscheinlichkeits-modellen. In diesem Fall ist eine Familie von Wahrscheinlichkeitsverteilungen gegeben, deren Parameter aus den Daten geschätzt wird. Anschließend wird mit Hilfe des χ^2-Tests untersucht, ob die durch den Schätzwert bestimmte Verteilung dem zugehörigen Histogramm gut angepasst ist.

Seien $X_1, ..., X_n$ unabhängige, identisch verteilte Zufallsvariable mit endlichem oder abzählbar unendlichem Wertebereich W_{X_1}. Ferner sei I ein offenes, nicht notwendigerweise endliches Intervall der Menge $\mathbb{R}$ der reellen Zahlen und

$$P_\theta\,,\ \theta \in I \subseteq \mathbb{R}$$

eine Familie von (diskreten) Wahrscheinlichkeitsverteilungen auf W_{X_1}.

Um die Nullhypothese

$H_0:\ \exists\,\theta \in I : X_1 \sim P_\theta$

gegen die Alternativhypothese

$H_1:\ \exists$ kein $\theta \in I : X_1 \sim P_\theta$

auf der Basis der Stichprobe $(x_1, ..., x_n)$ zu testen,

- ermittelt man zunächst den Schätzwert $\hat{\theta}_n = \hat{\theta}(x_1, ..., x_n)$ des Parameters θ mit Hilfe eines geeigneten Schätzers $\hat{\theta}_n$,
- macht nötigenfalls eine geeignete Klasseneinteilung $\{K_0, K_1, ..., K_m\}$ des Wertebereichs W_{X_1} der betrachteten Zufallsvariablen, sodass die Faust-regel $n \times \min\left\{P_{\hat{\theta}_n}(K_0), ..., P_{\hat{\theta}_n}(K_m)\right\} \geq 5$ erfüllt ist,
- berechnet den Wert χ^2 der zugehörigen χ^2-Statistik

$$\chi^2_{m-1}(n, P_{\hat{\theta}_n}) = \sum_{j=0}^{m} \frac{(B(K_j) - n\,P_{\hat{\theta}_n}(K_j))^2}{n\,P_{\hat{\theta}_n}(K_j)}$$

- und verwirft H_0 zugunsten H_1 auf dem Niveau $\cong \alpha$, sofern

$$\chi^2 \geq \chi^2_{m-d,1-\alpha}$$

ist.

$\alpha = 0.01$ populär zu machen. In einer Zeit, in welcher die Berechnung des P-Werts aufgrund mangelnder Rechenhilfen sehr mühsam war, war es auch naheliegend, die kritischen Werte für einige wenige Niveaus tabellarisch festzuhalten. Dies umso mehr, wenn die Prüfverteilung - wie z.B. die χ^2-Verteilung - noch dazu von einem Parameter abhängt.

Die Anzahl der Freiheitsgrade ist dabei gleich der Anzahl der Klassen vermindert um 1 und überdies vermindert um die Dimension d des zu schätzenden Parameters[38]. Das ist $(m+1)-1-d=m-d$.

Anwendungsbeispiel: Der radioaktive Zerfall

In diesem Anwendungsbeispiel wollen wir nachprüfen, dass - wie behauptet wird - die Verteilung der Anzahl der mit einem *Geiger-Müller*-Zähler in einem Zeitintervall vorgegebener Länge registrierten Szintillationen tatsächlich gut durch eine Poissonverteilung beschrieben wird.

E. Rutherford und *H. Geiger* haben bei ihrem klassischen, im Jahre 1910 durchgeführten Experiment eine Poloniumquelle benützt und die Szintillationen von 2608 disjunkten Zeitintervallen von je 7.5 Sekunden Dauer registriert.

Es bezeichnen

$b(j)$ ... die beobachteten Häufigkeiten der Zeitintervalle mit j Szintillationen und

$e_{\hat{\lambda}}(j)=2608\cdot\frac{\hat{\lambda}^j e^{-\hat{\lambda}}}{j!}$ ... die unter der Annahme der Poissonverteilung zu erwartenden Häufigkeiten, $j\in\{0,1,...,14\}$,

wobei $\hat{\lambda}$ der durch das Stichprobenmittel $\hat{\lambda}=\frac{1}{2608}\sum_{j=0}^{14} j\cdot b(j)=3.872$ geschätzte Parameter ist. Dies deswegen, weil eine gemäß einer Poissonverteilung mit dem Parameter λ verteilte Zufallsgröße X den Erwartungswert $E(X)=\lambda$ hat und das Stichprobenmittel der erwartungstreue Schätzer von λ mit kleinster Varianz ist.

j	0	1	2	3	4	5	6	7	8	9
$b(j)$	57	203	383	525	532	408	273	139	45	27
$e_{\hat{\lambda}}(j)$	54.3	210.3	407.1	525.3	508.4	393.7	254.0	140.5	68.0	29.2

j	10	11	12	13	14	≥ 15
$b(j)$	10	4	0	1	1	0
$e_{\hat{\lambda}}(j)$	11.3	4.0	1.3	0.4	0.1	0.1

Im vorliegenden Fall ist es daher naheliegend, die folgenden Klassen zu wählen, welche offensichtlich alle die genannte Faustregel erfüllen

$$K_0=\{0\},K_1=\{1\},...,K_{10}=\{10\}\quad\text{und}\quad K_{11}=\{11,12,...\}.$$

Der beobachtete χ^2-Wert ist 12.961.

[38] salopp ausgedrückt: überdies vermindert um die "Anzahl der zu schätzenden Parameter"

Da im vorliegenden Fall die Anzahl der Klassen 12 und die Dimension des zu schätzenden Parameters $d = 1$ ist, ist die Anzahl der Freiheitsgrade gleich 10 .

Wegen $P(\chi^2_{10} \geq 12.961) > 0.2$ gibt es im vorliegenden Fall daher tatsächlich keinen Grund, die Nullhypothese zu verwerfen, dass die zugrundeliegende Verteilung eine Poissonverteilung ist.

Anmerkung 3: Bei genauerer Analyse der Funktionsweise eines Geiger-Müller-Zählers müsste man dessen sogenannte Totzeit berücksichtigen. Zur Beschreibung dieses Phänomens sei auf die Abschnitte 12 und 16 in [20] verwiesen.

2.4.3 Test zweier Wahrscheinlichkeitsverteilungen auf Gleichheit (Test auf Homogenität)

Wir gehen im Folgenden von zwei Stichproben mit den Stichprobenumfängen n_0 und n_1 aus. $N = n_0 + n_1$ bezeichne den Gesamtumfang beider Stichproben.

Die erste Stichprobe sei durch n_0 unabhängige, identisch verteilte Zufallsvariable $X_{01}, ..., X_{0n_0}$ gegeben, deren Wertebereich $W = \{\omega_0, \omega_1, ..., \omega_m\}$ und deren Wahrscheinlichkeitsverteilung $P_0 = (p_{00}, p_{01}, ..., p_{0m})$ ist. $\mathbb{B}_m^{(n_0)} = (B_{00}, B_{01}, ..., B_{0m})$ bezeichne den zugehörigen Vektor der Ausfallshäufigkeiten

$$B_{0j} = |\{\, i \in \{1, ..., n_0\} : X_{0i} = \omega_j\}| \;, \quad j \in \{0, 1, ..., m\} \,.$$

Die zweite Stichprobe sei durch n_1 unabhängige, identisch verteilte Zufallsvariable $X_{11}, ..., X_{1n_1}$ mit demselben Wertebereich $W = \{\omega_0, \omega_1, ..., \omega_m\}$ und der Wahrscheinlichkeitsverteilung $P_1 = (p_{10}, p_{11}, ..., p_{1m})$ gegeben. $\mathbb{B}_m^{(n_1)} = (B_{10}, B_{11}, ..., B_{1m})$ bezeichne den zugehörigen Vektor der Ausfallshäufigkeiten

$$B_{1j} = |\{\, i \in \{1, ..., n_1\} : X_{1i} = \omega_j\}| \;, \quad j \in \{0, 1, ..., m\} \,.$$

Die beiden Vektoren sind in der folgenden $2 \times (m+1)$-Tafel zusammen mit den Zeilen- und Spaltensummen dargestellt.

Ausfall	ω_0	ω_1	...	ω_m	
erste Stichprobe	B_{00}	B_{01}	...	B_{0m}	n_0
zweite Stichprobe	B_{10}	B_{11}	...	B_{1m}	n_1
	$B_{00} + B_{10}$	$B_{01} + B_{11}$	...	$B_{0m} + B_{1m}$	N

Unsere Aufgabe sei es,

die Nullhypothese $H_0 : P_1 = P_0$ gegen die Alternativhypothese $H_1 : P_1 \neq P_0$ zu testen.

Da die zugehörigen Vektoren der Erwartungswerte der Beobachtungshäufigkeiten unter der Annahme der Nullhypothese offensichtlich die entsprechenden Vielfachen des Vektors $\hat{P} = (\frac{B_{00}+B_{10}}{N}, \frac{B_{01}+B_{11}}{N}, ..., \frac{B_{0m}+B_{1m}}{N})$, nämlich $\mathbb{E}_m^{(n_0)}=n_0\hat{P}$ und $\mathbb{E}_m^{(n_1)}=n_1\hat{P}$ sind, ist die zugehörige χ^2-Statistik wegen

$$B_{1j} - n_1\frac{B_{0j}+B_{1j}}{N} = -(B_{0j} - n_0\frac{B_{0j}+B_{1j}}{N}) \quad \text{und} \quad \frac{1}{n_0}+\frac{1}{n_1} = \frac{1}{n_0(1-\frac{n_0}{N})}$$

$$\begin{aligned}
\chi^2 &= \sum_{j=0}^{m}\sum_{i=0}^{1}\frac{(B_{ij}-E_{ij})^2}{E_{ij}} \\
&= \sum_{j=0}^{m}[\frac{(B_{0j}-n_0\frac{B_{0j}+B_{1j}}{N})^2}{n_0\frac{B_{0j}+B_{1j}}{N}} + \frac{(B_{1j}-n_1\frac{B_{0j}+B_{1j}}{N})^2}{n_1\frac{B_{0j}+B_{1j}}{N}}] \\
&= \frac{1}{n_0(1-\frac{n_0}{N})}\sum_{j=0}^{m}\frac{(B_{0j}-n_0\frac{B_{0j}+B_{1j}}{N})^2}{\frac{B_{0j}+B_{1j}}{N}}\,.
\end{aligned}$$

Da unter der Nullhypothese die Wahrscheinlichkeitsverteilung $\hat{P}$ als gegeben angenommen werden kann und die Stichprobenumfänge zudem fest vorgegeben sind (d.h. dass alle Spalten- und Zeilensummen unserer $2\times(m+1)$-Tafel fest sind), können von den $(m+1)^2$ Eintragungen nur m frei gewählt werden. Daher ist die Anzahl der Freiheitsgrade der χ^2-Statistik gleich m.

Bei gegebenem Singnifikanzniveau α ist die Nullhypothese H_0 zugunsten der Alternativhypothese H_1 demnach dann zu verwerfen, wenn der beobachtete Wert von $\chi^2 \geq \chi^2_{m,1-\alpha}$ ist.

Anmerkung zum Spezialfall m = 1 : Berücksichtigt man die in Abschnitt 2.3.3 verwendeten Bezeichungen, nämlich $n_0 = n$, $B_{00} = X_n$, $B_{00} + B_{01} = n$, $B_{10} + B_{11} = N - n$, $B_{00} + B_{10} = s$ und $B_{01} + B_{11} = N - s$ und daher

$$B_{01} - n\frac{B_{01}+B_{11}}{N} = -(X_n - n\frac{s}{N})\,,$$

so erhält die χ^2-Statistik folgende Form

$$\begin{aligned}\chi^2 &= \frac{1}{n(1-\frac{n}{N})}[\frac{(B_{00}-n\frac{B_{00}+B_{10}}{N})^2}{\frac{B_{00}+B_{10}}{N}}+\frac{(B_{01}-n\frac{B_{01}+B_{11}}{N})^2}{\frac{B_{01}+B_{11}}{N}}] \\ &= \frac{(X_n-n\frac{s}{N})^2}{n(1-\frac{n}{N})}[\frac{1}{\frac{s}{N}}+\frac{1}{1-\frac{s}{N}}] \\ &= \frac{(X_n-n\frac{s}{N})^2}{n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n}{N})}\,.\end{aligned}$$

Dies ist das $\frac{N}{N-1}$-Fache des Quadrats der in Abschnitt 2.3.3, B2) erhaltenen Teststatistik $\sqrt{N-1}\times\hat{\rho}_{(n,N)}(X_n,s)$, welche - für große n und $N-n$ - in guter Näherung $N(0,1)$-verteilt ist. Im Hinblick darauf, dass das Quadrat einer $N(0,1)$-verteilten Zufallsgröße eine χ^2-Verteilung mit einem Freiheitsgrad besitzt, wird die oben angegebene Regel, dass für die χ^2-Statistik einer $2\times(m+1)$-Tafel die Anzahl der Freiheitsgrade mit m zu wählen sei, für den Spezialfall $m=1$ als richtig bestätigt.

Die obige χ^2-Statistik ist prädestiniert dazu, die Nullhypothese $H_0:\rho=0$ gegen die zweiseitige Alternativhypothese $H_1:\rho\neq 0$ zu testen. Zum Unterschiede von der in Abschnitt 2.3.3, B2) verwendeten Statistik

$$\frac{X_n-n\frac{s}{N}}{\sqrt{n\frac{s}{N}(1-\frac{s}{N})(1-\frac{n-1}{N-1})}}$$

erlaubt sie es jedoch nicht, H_0 gegen eine der beiden einseitigen Alternativhypothesen $\rho<0$ oder $\rho>0$ zu testen.

Kapitel 3

PROJEKTE UND ÜBUNGSAUFGABEN

Projekt 1: Experiment
Wählen Sie ein Experiment aus der gegebenen Sammlung aus und führen Sie dieses durch.

Projekt 2: Sind Geburtstage gleichverteilt?
Beschaffen Sie sich von der in Projekt 3 angeführten Stelle die Geburtsdaten der im letzten Kalenderjahr im Bundesland Salzburg geborenen Kinder und fertigen Sie ein Histogramm für die Anzahl der in den einzelnen Monaten geborenen Kinder an.

Projekt 3: Körpergröße von Mädchen und Buben
Beschaffen Sie sich von der unten angegebenen Stelle Daten hinsichtlich der Körpergröße von Schulkindern - Mädchen und Buben getrennt - einer bestimmten Altersgruppe bzw. Schulstufe und fertigen Sie je ein *Histogramm* (eventuell auch ein *Stängel-Blatt-Diagramm*) an. Ermitteln Sie die besprochenen Kennwerte für Lage und Streuung und zeichnen Sie die zugehörigen *Kasten-Bilder*.

Referat 0/03: Statistik
Amt der Salzburger Landesregierung
Leiter: *Mag. Josef Raos*
Fanny von Lehnert-Straße 1
Postfach 527
5010 Salzburg
Tel. 8042-3525

Projekt 4: Fußballmeisterschaft
Erheben Sie von einer geeigneten Stelle die jeweiligen Resultate der Meisterschaftsspiele der österreichischen Fußball-Bundesliga der abgelaufenen Saison.
Stellen Sie je ein Histogramm für die Anzahl der Tore

a) der jeweiligen Heimmannschaft
b) der jeweiligen Gastmannschaft
c) insgesamt

dar. Stellen Sie einander die Anzahl der Tore der Heimmannschaft und der Gastmannschaft in einem *Streudiagramm* gegenüber.

Projekt 5: Programm 3 der Videoreihe Against all Odds: Inside Statistics
Stellen Sie die wesentlichen Vokabeln von Program 3: *Discribing Distributions: Numerical Description of Distributions* der Reihe Against all Odds: Inside Statistics zusammen und geben Sie eine kurze Inhaltsangabe. Präsentieren Sie den Film mit Hilfe der erarbeiteten Unterlagen in einer der nächsten Lehrveranstaltungsstunden.

Projekt 6: Die "Wiener Pictographie" von Otto Neurath [1]
Vermitteln Sie einen Eindruck von *Otto Neuraths* bildhaften Darstellungen im Bereich der Statistik.

Projekt 7: Schulbuchvergleich hinsichtlich Stichprobenvarianz
Untersuchen Sie die verschiedenen Lehrbücher dahingehend, ob für den Nenner der Stichprobenvarianz

$$\frac{1}{\text{Nenner}} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2$$

$n-1$ oder n gewählt wird und wie allenfalls die Wahl von $n-1$ gerechtfertigt wird.

Projekt 8: Lügen mit Statistik
Behandeln Sie Themen Ihrer Wahl aus [9] über Fehlerquellen bei statistischen Untersuchungen.

Projekt 9: Das Hubblesche Gesetz - Anwendungsbeispiel für eine homogene Regressionsgerade

[1] *Otto Neurath* $(1882-1945)$ war Mitglied des *Wiener Kreises.*

Behandeln Sie die Fallstudie 10.4.1 aus [11]

Projekt 10: Das Hooksche Gesetz - Anwendungsbeispiel für eine allgemeine Regressionsgerade

Die Feder einer Federwaage habe die Länge von d cm, wenn sie ohne Gewicht ist. Wenn ein Gewicht von x Kilogramm an den Haken am Ende der Feder gehängt wird, wird die Feder zur neuen Länge von y cm gestreckt. Aufgrund des Hookschen Gesetzes gilt (selbstverständlich für nicht allzu große Gewichte), dass der Längenzuwachs proportional zum Gewicht ist. Für die Länge y gilt somit die lineare Beziehung

$$y = k \cdot x + d\,.$$

Folgende Wertepaare (x_i, y_i) wurden ermittelt:

$(0, 439.00)$, $(2, 439.12)$, $(4, 439.21)$, $(6, 439.31)$, $(8, 439.40)$, $(10, 439.50)$.

Ermitteln Sie die Regressionsgerade und das Bestimmtheitsmaß.

Informieren Sie sich über das Hooksche Gesetz und führen Sie mit einer Federwaage eigene Versuche durch. Orientieren Sie sich an dem einschlägigen Beitrag in [4], Part II. Chapter 12

Projekt 11: Bestimmung der Erdbeschleunigung mit Hilfe von Pendeluhren

Christiaan Huygens $(1629 - 1695)$ ist es gelungen, die von ihm konstruierte Pedeluhr zur Bestimmung der Erdbeschleunigung g zu verwenden. Er hat nämlich zudem gefunden, dass die Schwingungsdauer T eines sogenannten mathematischen Pendels in folgender Weise von der Länge l des Pendels und der Erdbeschleunigung g abhängt:[2]

$$T = 2\pi\sqrt{\frac{l}{g}}\,.$$

Ist nun $\nu = \frac{1}{T}$ die Frequenz des Pendels (Anzahl der Schwingungen pro Sekunde), dann gilt demgemäß

$$\nu^2 = \frac{g}{(2\pi)^2} \times \frac{1}{l}\,.$$

Stellen Sie von $n \geq 2$ Pendeluhren deren Pendellängen l_i und Frequenzen ν_i fest und benützen Sie die obige Beziehung, um aus den Punktepaaren $(\frac{1}{l_i}, \nu_i^2)$, $i \in \{1, ..., n\}$, die Erdbeschleunigung g zu schätzen.

[2] T ... in Sekunden, l ... in Metern, g ... in Metern pro Sekundenquadrat

Hinweis: Eine faszinierende historische Darstellung finden Sie im Abschnitt *Ein Meßversuch* von [7].

Projekt 12: Der Spearmansche Rangkorrelationskoeffizient

a) Geben Sie einen Überblick über Leben und Werk von *Charles Spearman* [3] und

b) stellen Sie - ausgehend von Aufgabe M 15 - den Zusammenhang des Pearsonschen Korrelationskoeffizienten und Spearman's ρ dar.

c) Zeigen Sie, dass für den Wertebereich W_n der sogenannten *Hotelling-Pabst-Statistik* [4]

$$s_n = \sum_{i=1}^{n} (y_i - x_i)^2$$

gilt

$$W_n = 2 \cdot \begin{cases} \{0, ..., \binom{n+1}{3}\} & \text{für} \quad n \in \mathbb{N}\backslash\{1,3\} \\ \{0,1,3,4\} & \text{für} \quad n = 3 \end{cases}$$

und d) folgern Sie daraus, dass für alle $n \geq 4$ mit Ausnahme jener n, welche durch 2 aber nicht durch 4 teilbar sind, ρ den Wert 0 annehmen kann.

e) Beschreiben Sie ein einschlägiges statistisches Anwendungsbeispiel z.B. aus [6], Seite 553 ff.

Projekt 13: Die Capture-Recapture-Methode

Erarbeiten Sie die Capture-Recapture-Methode (Rückfangmethode) für das Ziehen ohne Zurücklegen anhand der bereitgestellten Unterlagen und präsentieren Sie die angegebenen Schätzer.

Projekt 14: Zur Planung von Versuchen

Ausgehend von der am Beginn von Abschnitt 2.2 beschriebenen Fragestellung zum Testen von Hypothesen bezüglich Wahrscheinlichkeiten und Anteilswerten bei Alternativexperimenten ist hinsichtlich der Formulierung der Hypothesen vom Fall

$$3b) \quad \boxed{\begin{array}{c} H_0 : p \leq p_0 \\ H_1 : p \geq p_1 \end{array}}$$

[3] Der englische Psychologe *Charles Spearman* $(1863 - 1945)$ entwickelte im Jahre 1904 seine Zwei-Faktoren-Theorie der Intelligenz.

[4] Der US-amerikanische Statistiker und Ökonom *Harold Hotelling* $(1895 - 1973)$ war Mitbegründer des ersten Instituts für Statistik in den Vereinigten Staaten an der University of North Carolina at Chapel Hill.

für allgemeine p_0 und p_1 auszugehen, $0 < p_0 < p_1 < 1$. Seien $\hat{p}_n$ die relative Häufigkeit der schwarzen Kugeln in einer Stichprobe vom Umfang n, $0 < \alpha, \beta;\ \alpha + \beta < 1$ und

$$p_\gamma = p_0 + \gamma(p_1 - p_0) \quad \text{mit} \quad \gamma \in (0, 1)\,.$$

Die Entscheidung zugunsten H_1 und H_0 werde dann getroffen, wenn $\hat{p}_n \geq p_\gamma$ bzw. $\hat{p}_n < p_\gamma$ ist. Dabei soll gewährleistet sein, dass die Wahrscheinlichkeit des Fehlers erster Art durch α und die des Fehlers zweiter Art durch β nach oben beschränkt ist.

Der Parameter γ ist so zu wählen, dass der dazu erforderliche Stichprobenumfang $n = n_{\alpha,\beta}(\gamma)$ minimal ist.

Projekt 15: Sex Bias in Graduate Admissions
Informieren Sie sich über *Simpson's Paradoxon* (z.B. in [22], vgl. auch Aufgabe M 26 und präsentieren Sie die Ergebnisse der Studie "Sex Bias in Graduate Admissions: Data from Berkeley" [3].

Projekt 16: Rothamsted Experimental Station
Referieren Sie über geographische Lage, Organisationsform, Geschichte und Aufgabenbereiche der *Rothamsted Experimental Station* in Harpenden, England (http://www.rothamsted.ac.uk/).

Projekt 17: Klimadiagramme
Referieren Sie über die in der Geographie üblichen *hygrothermischen Klimadiagramme* nach *Walter* und *Lieth*.

Projekt 18: Analyse von Daten aus dem Kombinationsfach
Analysieren Sie Daten, die in Zusammenhang mit Ihrem Kombinationsfach erhoben wurden und präsentieren Sie die Ergebnisse.

Übungsaufgaben mit vorwiegend statistisch-anwendungsorientiertem Charakter

1. Beschaffen Sie sich Informationen zu den verschiedenen Temperaturskalen hinsichtlich a) den Namensgeber der Skala, b) die Hintergünde für das Design der Skala, c) die Verbreitung der Skala und beschreiben Sie schließlich deren Umrechnung.

2. Zeigen Sie, dass nach dem Gregorianischen Kalender (Schaltjahr, wenn die Jahreszahl durch 4 teilbar ist, mit Ausnahme der Jahre, die durch 100 aber nicht durch 400 teilbar sind) der 13. eines Monats im langjährigen Durchschnitt häufiger auf einen Freitag fällt als auf irgend einen anderen Wochentag. Hinweis: Der 1. Jänner 2000 war ein Samstag.

3. Eine Firma führt mit 80 Bewerbern einen Eignungstest durch. Das Ergebnis der Punktebewertung sieht so aus:

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

   Bilde eine Klasseneinteilung für die Punktezahlen mit a) 10 Klassen, b) 5 Klassen! Gib die Häufigkeiten für die einzelnen Klassen in einer Tabelle an und zeichne ein Histogramm!

4. Bei einer Untersuchung der Druckfestigkeit von genormten Betonklötzen erhielt man folgende Werte (in N/cm^2):

3512	2256	2413	3100	2305	4110	3355	3522	2845	4444	3286	2874
3934	3895	3237	2560	2403	3080	2659	3159	3218	3453	4022	3571
2727	3610	2963	4346	3640	3590	2933	2708	3384	2394	3296	2237
3012	3944	3483	3394	3267	2453	2345	3718	3021	2325	3689	3277
3483	3424	3218	2737	2943	3120	4493	2796	2668	3463	3738	3865
2786	3375	2972	3178	2737	3541	4258	3787	3816	2953	3345	3110
3178	3453	3610	3864	3463	3139	3708	3090	3218	3267	3110	3512
2963	3473	2796	3669								

   a) Erstelle ein *Stängel-Blatt-Diagramm* mit geeigneter Klasseneinteilung und

   b) zeichne die zugehörige empirische Verteilungsfunktion.

5. **Schwangerschaftsdauern** (siehe [11], S. 407)

Die nachstehende Tabelle zeigt die in einem der *County General Hospitals* in den U.S.A. im Jahre 1978 erhobenen Dauern von 70 Schwangerschaften.

a) Erstellen Sie ein *Stängel-Blatt-Diagramm* für die ihrer Größe nach geordneten Daten, b) zeichnen Sie die zugehörige empirische Verteilungsfunktion und c) fertigen Sie ein Kastenbild an.

251	264	234	283	226	244	269	241	276	274
263	243	254	276	241	232	260	248	284	253
265	235	259	279	256	256	254	256	250	269
240	261	263	262	259	230	268	284	259	261
268	268	264	271	263	259	294	259	263	278
267	293	247	244	250	266	286	263	274	253
281	286	266	249	255	233	245	266	265	264

6. Am National Bureau of Standards in Washington, U.S.A., wurden in den Jahren $1962 - 1963$ Messungen des dortigen Standardgewichts NB 10 (Nominalwert von 10 Pond) mit einem der besten Meßgeräte und unter Gewährleistung von möglichst gleichbleibenden Bedingungen durchgeführt. Da alle Messwerte etwa 400 Mikropond unter dem Nominalwert lagen, war es vorteilhaft, die Differenzen der Messwerte vom Nominalwert in Mikropond anzugeben und als Stichprobenwerte zu verwenden. (So ist z.B. für einen Messwert von 9.999591 Pond diese Differenz $10 - 9.999591 = 0.000409$ Pond oder 409 Mikropond.) Bei 100 aufeinanderfolgenden Messungen ergaben sich folgende Stichprobenwerte (vgl. [4], § 6 Measurement Error).

409	400	406	299	402	406	401	403	401	403
398	403	407	402	401	399	400	401	405	402
408	399	399	402	399	397	407	401	399	401
403	400	410	401	407	423	406	406	402	405
405	409	399	402	407	406	413	409	404	402
404	406	407	405	411	410	410	410	401	402
404	405	392	407	406	404	403	408	404	407
412	406	409	400	408	404	401	404	408	406
408	406	401	412	393	437	418	515	404	401
401	407	412	375	409	406	398	406	403	404

a) Erstellen Sie ein *Stängel-Blatt-Diagramm* für die ihrer Größe nach geordneten Daten, b) zeichnen Sie die zugehörige empirische Verteilungsfunktion, c) fertigen Sie ein Kastenbild an und c) tragen Sie die Daten in das Wahrscheinlichkeitsnetz ein.

7. Changing the choice of classes can change the appearance of a histogram. Here is an example in which a small shift in the classes, with no change in the number of classes, has an important effect on the histogram. The data are acidity levels (measured by pH) in 105 samples of rainwater. Distilled water has pH 7.00. As the water becomes more acid, the pH goes down. The pH of rainwater is important to environmentalists because of the problem of acid rain (aus [14], S. 35).

   a) Make a histogram of pH with 14 classes, using class boundaries 4.2, 4.4, ..., 7.0. How many modes does your histogram show? More than one mode suggests that the data contain groupes that have different distributions.

   b) Make a second histogram, also with 14 classes, using class boundaries 4.14, 4.34, ..., 6.94. The classes are those from a) moved 0.06 to the left. How many modes does the new histogram show?

   c) Use your software's histogram function to make a histogram without specifying the number of classes and their boundaries. How does the software's default histogram compare with those in a) and b)?

   d) Make a normal quantile plot of these data.

4.33	4.38	4.48	4.48	4.50	4.55	4.59	4.59	4.61	4.61
4.75	4.76	4.78	4.82	4.82	4.83	4.86	4.93	4.94	4.94
4.94	4.96	4.97	5.00	5.01	5.02	5.05	5.06	5.08	5.09
5.10	5.12	5.13	5.15	5.15	5.15	5.16	5.16	5.16	5.18
5.19	5.23	5.24	5.29	5.32	5.33	5.35	5.37	5.37	5.39
5.41	5.43	5.44	5.46	5.46	5.47	5.50	5.51	5.53	5.55
5.55	5.56	5.61	5.62	5.64	5.65	5.65	5.66	5.67	5.67
5.68	5.69	5.70	5.75	5.75	5.75	5.76	5.76	5.79	5.80
5.81	5.81	5.81	5.81	5.85	5.85	5.90	5.90	6.00	6.03
6.03	6.04	6.04	6.05	6.06	6.07	6.09	6.13	6.21	6.34
6.43	6.61	6.62	6.65	6.81					

8. In 1798 the English scientist *Henry Cavendish* measured the density of the earth by careful work with a torsion balance. The variable recorded

was the density of the earth as a multiple of the density of water. Here are the Cavendish's 29 measurements (aus [14], S. 37):

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.58	

a) Present the measurements graphically by either a stemplot or a histogram and explain the reason for your choice.

b) Then briefly discuss the main features of the distribution. In particular, what is your estimate of the density of the earth based on these measurements?

c) Find $\bar{x}_n$ and s_n for these data.

9. Zu Beispiel 5 in Abschnitt 1.3.1: a) Ermitteln Sie das Stichprobenmittel $\bar{x}_{100}$, die Stichprobenvarianz s^2_{100} und die geordneten zentrierten und standardisierten Daten

$$\tilde{x}_{i:100} = \frac{x_{i:100} - \bar{x}_{100}}{s_{100}}\,,\ i \in \{1, ..., 100\}\,.$$

Sei nun

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})dt\,,\ x \in \mathbb{R}\,,$$

die Verteilunsfunktion der $N(0,1)$-Verteilung und Φ^{-1} die zugehörige Inverse. Dann gilt offensichtlich $\Phi^{-1}(\Phi(x)) = id(x)$. Darauf beruht das sogenannte *Wahrscheinlichkeitsnetz* (*normal quantile plot*) [5], in dem man die $\Phi^{-1}(i/n)$ gegen die $x_{i:n}$ aufträgt. Die Punktepaare

$$(x_{i:n}, \Phi^{-1}(i/n))\,,\ i \in \{1, ..., n\}\,,$$

liegen erwartungsgemäß annähernd auf einer Geraden, wenn die Daten aus einer Normalverteilten Grundgesamtheit stammen.

b) Zeichnen Sie die "Punktwolke"

$$(\tilde{x}_{i:100}, \Phi^{-1}(i/n))\,,\ i \in \{1, ..., 100\}\,,$$

[5] Dieses ist im Handel erhältlich; beispielsweise bei Fa. Schleicher & Schüll GmbH, Grimsehlstraße 23, D-37574 Einbeck

und vergleichen Sie diese mit der Geraden $y = x$.

c) Ermitteln Sie gemäß Abschnitt 1.6.2 das zugehörige (homogene) Bestimmtheitsmaß.

10. Präsentieren Sie das Kastenbild (boxplot) aus [23] unter Verwendung einschlägiger Software.

11. Eine Bakterienkultur wächst während der ca. 16 Tagstunden um 20 % pro Stunde und während der Nacht um nur 12 % pro Stunde. Berechne das durchschnittliche Wachstum pro Stunde! (Beispiel E 1418 aus [59])

12. Seien $ATX_i \in (0, \infty)$, $i \in \{0, 1, ..., 20\}$ die ATX-Werte[6] von 21 aufeinanderfolgenden Börsentagen, $r_i = ATX_i/ATX_{i-1}$, $i \in \{1, ..., 20\}$. Dann heißt

$$s_{20} = \sqrt{\frac{1}{19}\sum_{i=1}^{20}\left(\ln(r_i) - \mu_{\ln}\right)^2}$$

die *Volatilität*[7] pro Börsentag. Dabei ist

$$\mu_{\ln} = \frac{1}{20}\sum_{i=1}^{20}\ln(r_i) = \ln(\sqrt[20]{ATX_{20}/ATX_0})$$

die *durchnittliche logarithmische Tagesrendite.*

a) Entnehmen Sie der Homepage der Österreichischen Nationalbank (www.oenb.at) die ATX-Werte für die aufeinanderfolgenden Tagen eines Monats und ermitteln Sie die zugehörige Werte der durchschnittlicher logarithmischen Tagesrendite und der Volatilität.

b) Stellen Sie die Wertepaare $(i, ATX_i/ATX_0)$, $i \in \{0, ..., 20\}$ graphisch dar und vergleichen Sie die Werte ATX_i/ATX_0 mit den zugehörigen Funktionswerten der Funktionen

$$f_k(t) = e^{\mu_{\ln} t + k s_{20}\sqrt{t}}, \; t \in [0, 20], \quad k \in \{-1, 0, 1\}.$$

13. Bei einem Autorennen sind 5 Runden zu fahren. Die mittleren Geschwindigkeiten in den einzelnen Runden betragen für einen Fahrer: 183, 210, 201, 180, 182 (km/h). Wie groß ist die mittlere Geschwindigkeit für alle 5 Runden? (Beispiel E 1419 aus [59])

[6] *Austrian Trended Index* (ATX)

[7] abgeleitet vom Italienischen *volare* (fliegen), also "Flatterhaftigkeit", "Beweglichkeit".

14. a) Von zwei Körpern mit einem Gewicht von jeweils $12\,kg$ verdrängt der eine $2\,l$ und der andere $1\,l$ Wasser. Die Dichten der beiden Körper sind daher $6\,kg/dm^3$ bzw. $12\,kg/dm^3$. Berechnen Sie die durchschnittliche Dichte der beiden Körper (d.i. die Dichte jenes Körpers, welcher durch Zusammenfügen der beiden einzelnen Körper entsteht.)

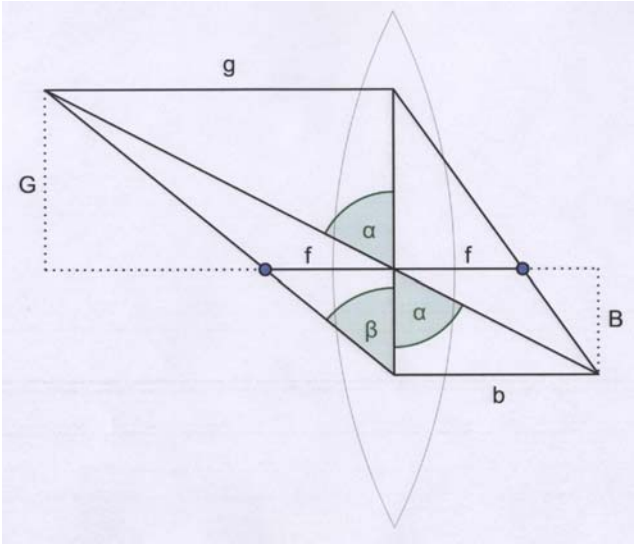
    b) Ermitteln Sie die mittlere Krümmung in den Scheitelpunkten einer Ellipse.

15. Für eine *bikonvexe Linse* bezeichne

    G ... die Gegenstandsgröße, g ... die Gegenstandsweite,

    B ... die Bildgröße, b ... die Bildweite und

    f ... die Brennweite.



a) Entnehmen Sie der obigen Abbildung - vermittels Ähnlichkeit zweier Dreiecke - zunächst die Beziehung (1_α) und leiten Sie mit deren Hilfe aus einer weiteren analog entnommenen Beziehung (1_β) die Abbildungsgleichung (2) her:

Abbildungsmaßstab	$\frac{G}{g}$	$=$	$\frac{B}{b}$	(1_α)
Abbildungsgleichung	$\frac{1}{f}$	$=$	$\frac{1}{g}+\frac{1}{b}$	(2)

b) Überzeugen Sie sich davon, dass $2f$ das harmonische Mittel von g und b ist.

16. Führen Sie eigene Versuche durch, um die Zahl π zu schätzen und die Beziehung $u = \pi \cdot d$ zwischen Kreisumfang u und Kreisdurchmesser d experimentell zu prüfen.

    Orientieren Sie sich dabei an Beispiel 2 von Abschnitt 1.6 .

17. Der Franziskanermönch, Mathematiker und Musiktheoretiker *Marin Mersenne* $(1588 - 1648)$, ein Zeitgenosse von *Galilei*, hat Fallversuche durchgeführt, um die Erdbeschleunigung g zu bestimmen. Die in seinem Werk *Harmonie Universelle* (Paris 1636) veröffentlichten Messergebnisse sind in der folgenden Tabelle dargestellt,

Falldauer	Fallhöhe
1/2	3
1	12
2	48
3	110
3 1/2	146 1/2

    wobei die Falldauern t_i in Sekunden und die Fallhöhen s_i in "*königlichen*" *Fuß* $(32.87\,cm)$ angegeben sind.

    Ermitteln Sie aus diesen Daten und der Beziehung $s = \frac{g}{2}t^2$ einen Schätzwert für g . Berücksichtigen Sie bei der Interpretation des Ergebnisses freilich, dass die Zeitmessung für *Mersenne* noch ein beträchtliches Problem war. Vgl. Sie dazu auch Projekt 11 .

18. Zwei Kunstkritiker brachten 12 Gemälde nach ihrem Werte in eine

Rangreihe, welche in der nachstehenden Tabelle widergegeben ist

Gemälde	Kritiker 1	Kritiker 2
1	8	6
2	7	9
3	3	1
4	11	12
5	4	5
6	1	4
7	5	8
8	6	3
9	10	11
10	2	2
11	12	10
12	9	7

Ermitteln Sie den Spearmanschen Rangkorrelationskoeffiztienten.

19. Die Umsätze (in Mrd \$) und Beschäftigtenzahlen der zwölf größten Unternehmen der Fahrzeugbranche sind in der nachstehenden Tabelle dargestellt.

Unternehmen	Land	Umsatz	Beschäftigte
General Motors	USA	123.8	756.300
Ford Motor	USA	89.0	332.700
Toyota Motor	Japan	78.1	102.423
Daimler-Benz	BRD	57.3	379.252
FIAT	Italien	46.8	287.957
Volkswagen	BRD	46.0	265.556
Nissan Motor	Japan	42.9	138.326
Honda Motor	Japan	30.6	85.500
Renault	Frankreich	29.4	147.195
Chrysler	USA	29.4	126.500
Boeing	USA	29.3	159.100
Peugeot	Frankreich	28.4	156.800

Ersetzen Sie die Punktepaare (x_i, y_i) der Messwerte x_i und y_i durch die Punktepaare $\{rg(x_i), rg(y_i)\}$ der zugehörigen Rangzahlen (Rangzahl 1 für den kleinsten Messwert, Rangzahl 2 für den nächstgrößeren

Messwert, ...), $i \in \{1, ..., 12\}$, und berechnen Sie den Korrelationskoeffizienten für die Rangzahlen (Tabelle 6.1 und Beispiel 6.6 aus [10], vgl. dazu auch Aufgabe 11 der Übungsaufgaben mit vorwiegend mathematischem Charakter).

20. Nehmen Sie ein Bridgepaket und denken Sie sich die Karten in der folgenden Weise mit Nummern versehen:

	As,	2,	...,	10,	Bub,	Dame,	König
Treff (♣)	1,	2,	...,	10,	11,	12,	13,
Pik (♠)	14,	15,	...,	23,	24,	25,	26,
Karo (♢)	27,	28,	...,	36,	37,	38,	39,
Herz (♡)	40,	41,	...,	49,	50,	51,	52.

Nun wählen Sie zufällig eine Karte der Farbe Pik, Karo oder Herz und entfernen alle Karten mit höheren Werten aus dem Paket. Aus dem Restpaket ziehen Sie 10 Stichproben bestehend aus je 6 Karten. Die Stichproben sind jeweils zufällig und ohne Zurücklegen zu entnehmen und die Nummern der gezogenen Karten sind zu registrieren. Nach jeder Stichprobe legen Sie die gezogenen Karten ins Restpaket zurück und mischen dieses gründlich. Schließlich ermitteln Sie auf der Basis der 10 Stichproben je ein Konfidenzintervall für die Anzahl N der Karten des Restpakets mit 90%-iger statistischer Sicherheit und interpretieren das Resultat.

21. Dem statistischen Jahrbuch 1985 für die Bundesrepublik Deutschland kann man die Zahlen der untenstehenden Tabelle entnehmen. Berechne für jedes der angegebenen Jahre ein Vertrauensintervall zur Vertrauenszahl $\gamma = 0.99$ für die unbekannte Wahrscheinlichkeit, dass ein lebendgeborenes Kind ein Junge (Mädchen) ist, und vergleiche die Intervalle.

Jahr	Lebendgeborene	
	insgesamt	männlich
1981	624557	320633
1982	621173	319293
1983	594177	305255
1984	584157	300120

22. Es wurden 2000 österreichische Jugendliche über die Häufigkeit von Erkrankungen befragt. Dabei ergab sich folgende Tabelle

Erkrankungen im letzten Jahr		
nie	einmal	mehr als einmal
62%	24%	8%

Im Jahr 1973, als die Erhebung durchgeführt wurde, gab es in Österreich ca. 1 275 000 Jugendliche. Schätze die Gesamtanzahl der Jugendlichen, die im betreffenden Jahr a) nie b) einmal c) mehr als einmal erkrankten, mit einer Sicherheit von 95%! (Hinweis: Schätze zuerst den relativen Anteil und daraus die Anzahl!) (Beispiel 3205 aus [59])

23. Um die Anzahl der Fische in einem Teich zu bestimmen, werden 70 Fische gefangen, mit einem Band markiert und wieder freigelassen. Einige Tage später werden 100 Fische gefangen. Davon sind 12 markiert. Schätze die Gesamtzahl der Fische mit einer Sicherheit von 95%! (Beispiel 3206 aus [59])

24. Der Anteil p eines bestimmten Variablenwerts in einer großen Grundgesamtheit soll auf 1% mit einer Sicherheit von 95% geschätzt werden.

    a) Wie viele Versuche sind nötig, wenn kein Schätzwert für den Anteil bekannt ist?

    b) Nach einer Voruntersuchung mit 250 Versuchen ist bekannt, dass $p > 0.8$ ist. Wie groß muss nun die Hauptuntersuchung gewählt werden? Wie groß ist die Ersparnis im Vergleich zu a)?

25. Vor einer österreichischen Bundespräsidentenwahl gibt das Team eines der Kandidaten eine Umfrage mit einen Stichprobenumfang von $n = 1600$ möglichen Wählern in Auftrag, um herauszufinden, ob ihr Kandidat die absolute Mehrheit der Stimmen erreichen wird. Der Anteil der Stimmen, die bei der Wahl für diesen Kandidaten abgegeben wird, sei p.

    Wie groß müsste die Anzahl $c_{0.05}$ seiner Wähler in einer Zufallsstichprobe sein, damit des Wahrscheinlichkeit des Fehlers, sich zugunsten der Hypothese $H_1 : p \geq 1/2$ zu entscheiden, dass ihr Kandidat die absolute Mehrheit erreicht, obwohl in Wahrheit die Hypothese $H_0 : p < 1/2$ zutrifft, dass er diese nicht erreicht, höchstens $\alpha = 0.05$ ist.

Anmerkung: Bezeichne S_n die Anzahl der Stimmen für den fraglichen Kandidaten in der Stichprobe. Da der Umfang $N \cong 5\,000\,000$ der Grundgesamtheit sehr groß ist, ist für die Verteilung von S_n anstelle der Hypergeometrischen Verteilung die Binomialverteilung zu verwenden.

26. Modifizieren Sie das unter Fall 3a) behandelte Beispiel zur Planung eines Modellversuchs für die folgenden Hypothesen hinsichtlich der Anteile der roten bzw. weißen Kugeln in der Urne

$$H_0 : (\frac{1}{4}, \frac{3}{4}) \quad \text{und} \quad H_1 : (\frac{3}{4}, \frac{1}{4})$$

und führen Sie den zugehörigen Versuch durch.

27. In dem unter Fall 3a) behandelte Anwendungsbeispiel erfolgt der Test auf Mischerbigkeit von Erbsen mittels Selbstbefruchtung (d.h. die zu untersuchende Erbsenpflanze wird mit sich selbst gekreuzt.) Angenommen, für den Test steht eine reinerbige Erbsenpflanze mit weißen Blüten zur Verfügung, sodass die zu untersuchende Ebsenpflanze mit einer Erbsenpflanze mit Genotyp aa gekreuzt werden kann. Finden Sie ein Urnenmodell für den zugehörigen Test und führen Sie sowohl bei Wahl von H_0 als auch von H_1 je einen einen Modellversuch für $n = 11$ durch.

28. Beispiel 10.11 aus [51]

29. Beispiel 10.19 aus [51]

30. Beispiel 10.24 aus [51]

31. Beispiel 976 aus [55]

32. Beispiel 984 aus [55]

33. Beispiel 998 aus [55]

34. Beispiel 935 aus [57]

35. Beispiel 936 aus [57]

36. Beispiel 950 aus [57]

37. In den österreichischen Bestimmungen für Mahlprodukte[8] ist u.a. festgelegt, dass ein Schwarzbrot mit einem Normwert von 1 kg mindestens

    0.97 kg wiegen müsse. Bei der Kontrolle einer Bäckerei zieht die Marktbehörde eine Stichprobe von 10 Broten. Unterschreitet das mittlere Gewicht dieser Brote den genannten Grenzwert, so wird der Bäcker angezeigt.

    Angenommen, die zuständige Preiskommission wäre bei der Festlegung des Grenzwerts davon ausgegangen, dass das tatsächliche Gewicht von einem Schwarzbrot mit einem Normwert von 1 kg nomalverteilt ist (mit Erwartungswert 1 und einer gewissen Varianz σ^2) und dass die fälschliche Anzeige eines Bäckers nur mit Wahrscheinlichkeit von 0.01 erfolgen soll. Wie groß wäre dabei die Standardabweichung σ angenommen worden?

    Ein Schwarzbrot mit einem Normwert von 2 kg hat einen Grenzwert von 1.94 kg. Was ist in diesem Fall die Antwort auf die obige Frage?

38. Nahezu bis zum Ende des 19. Jahrhunderts war die Sterblichkeit im Zusammenhang mit chirurgischen Eingriffen sehr hoch. Das Hauptproblem dabei war die Infektion. Solang es kein Modell der Krankheitsübertragung gab, gab es auch kein Konzept der Sterilisierung, sodass viele Patienten an post-operativen Komplikationen starben.

    Der dringend nötige Durchbruch gelang dem britischen Arzt *Joseph Lister*. Diesem war bereits aufgefallen, dass geschlossene Wunden im Gegensatz zu offenen kein Eiter bilden. Er begann, Arbeiten von *Louis Pasteur* zu lesen, dem es in einer Reihe klassischer Experimente gelungen war nachzuweisen, welche Rolle Hefe und Bakterien bei der Fermentation spielen. *Lister* vermutete, dass die menschliche Infektion eine ähnliche organische Ursache habe und somit Eiter auf etwas in der Luft Vorhandenes zurückzuführen ist. Er begann schliesslich, Karbolsäure im Operationssaal als Desinfektionsmittel zu verwenden.

    Die folgende Tabelle zeigt die Ergebnisse von 75 von *Lister* durchgeführten Operationen. 35 davon geschahen ohne und 40 mit Verwendung von

[8] Verordnung vom 13. August 1982

Karbolsäure.

Karbolsäure \ Patient	überlebt	überlebt nicht	Summe
verwendet	34	6	40
nicht verwendet	19	16	35
Summe	53	22	75

Ermitteln Sie den zugehörigen Φ-Koeffizienten und testen Sie - unter geeigneten Annahmen über die Datenerhebung - die Hypothese H_0, dass die beiden Behandlungsmethoden keine unterschiedliche Wirkung haben gegen die Alternativhypothese H_1, dass die Verwendung von Karbol die Überlebenschance von Patienten steigert.

39. Der englische Statistiker *William Searly Gosset* (1876 – 1937), der unter dem Pseudonym "*Student*" für die Brauerei Guinness arbeitete, hat bei der Untersuchung des Stichprobenfehlers, der beim Zählen von Hefezellen mit einem Haemazytometer auftritt, nachstehende 400 Beobachtungswerte erhoben.

a) Erstellen Sie zunächst eine Strichliste,

b) vergleichen Sie das Histogramm der beobachteten Häufigkeiten mit dem Histogramm der mit Hilfe einer geeigneten Poissonverteilung angepassten Häufigkeiten, und zwar für die Klassen $K_0 = \{0\}$, $K_1 = \{1\}$, $K_2 = \{2\}, ...$, $K_9 = \{9\}$, $K_{10} = \{10, 11, 12, ...\}$ und

c) führen Sie schließlich einen χ^2-Test durch.

2	2	4	4	4	5	2	4	7	7	4	7	5	2	8	6	7	4	3	4
3	3	2	4	2	5	4	2	8	6	3	6	6	10	8	3	5	6	4	4
7	9	5	2	7	4	4	2	4	4	4	3	5	6	5	4	1	4	2	6
4	1	4	7	3	2	3	5	8	2	9	5	3	9	5	5	2	4	3	4
4	1	5	9	3	4	4	6	6	5	4	6	5	5	4	3	5	9	6	4
4	4	5	10	4	4	3	8	3	2	1	4	1	5	6	4	2	3	3	3
3	7	4	5	1	8	5	7	9	5	8	9	5	6	6	4	3	7	4	4
7	5	6	3	6	7	4	5	8	6	3	3	4	3	7	4	4	4	5	3
8	10	6	3	3	6	5	2	5	3	11	3	7	4	7	3	5	5	3	4
1	3	7	2	5	5	5	3	3	4	6	5	6	1	6	4	4	4	6	4
4	2	5	4	8	6	3	4	6	5	2	6	6	1	2	2	2	5	2	2
5	9	3	5	6	4	6	5	7	1	3	6	5	4	2	8	9	5	4	3
2	2	11	4	6	6	4	6	2	5	3	5	7	2	6	5	5	1	2	7
5	12	5	8	2	4	2	1	6	4	5	1	2	9	1	3	4	7	3	6
5	6	5	4	4	5	2	7	6	2	7	3	5	4	4	5	4	7	5	4
8	4	6	6	5	3	3	5	7	4	5	5	5	6	10	2	3	8	3	5
6	6	4	2	6	6	7	5	4	5	8	6	7	6	4	2	6	1	1	4
7	2	5	7	4	6	4	5	1	5	10	8	7	5	4	6	4	4	7	5
4	3	1	6	2	5	3	3	3	7	4	3	7	8	4	7	3	1	4	4
7	6	7	2	4	5	1	3	12	4	2	2	8	7	6	7	6	3	5	4

40. In der nachstehenden Tabelle sind die absoluten Häufigkeiten H_j der Tage dargestellt, an den in Oxford (England) in den Jahren $1925-1930$ ($N = 2191$ Tage) $j \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ Erdbeben (ohne Nachbeben) registiert wurden (aus [21], S. 70 f.).

j	0	1	2	3	4	5	6	7
H_j	685	792	467	160	68	13	5	1

a) Erstellen Sie ein zugehöriges Histogamm und b) vergleichen Sie dieses mit dem Histogamm einer Poissonverteilung P_λ, deren Parameter λ durch das Stichprobenmittel $\hat{\lambda}$ gegeben ist. c) Führen Sie einen χ^2-Test auf Anpassung der Daten durch diese Poissonverteilung durch.

41. Von den 442 BewerberInnen um die Zulassung zu einem Graduate-Program an der University of California, Berkeley, gab es in einem

Department folgende Ergebnisse hinsichtlich Zulassungsstatus und Geschlecht:

	zugelassen	nicht zugelassen	
Frauen	40	42	82
Männer	97	263	360
	137	305	442

Testen Sie die Hypothese H_0, dass es bezüglich der Zulassung keine geschlechtsspezifischen Unterschiede gibt, gegen die Hypothese H_1, dass es eine Diskriminierung zulasten der Frauen gibt, wie es die einschlägigen Daten der gesamten Universität zunächst nahegelegt hatten, wonach der Anteil (von 34.6 %) der zugelassenen weiblichen Bewerberinnen gegenüber dem der männlichen Bewerber (von 44.3 %) deutlich geringer war. Vgl. Sie dazu Aufgabe M 20 und Projekt 13.

Übungsaufgaben mit vorwiegend mathematischem Charakter

1. Verifizieren Sie Anmerkung 4 aus Abschnitt 1.4.1 und visualisieren Sie diese für den Fall $n = 2$ mit Hilfe von GeoGebra.

2. Seien p_i positive Gewichte und X_i, $i \in \{1, ..., n\}$ unabhängige Zufallsvariable mit Erwartungwert μ derart, dass gilt

$$V\left(\sqrt{p_i}X_i\right) = \sigma^2,\ i \in \{1, ..., n\},$$

wobei μ und σ^2 unbekannt sind. (a) Zeigen Sie, dass der Schätzer

$$\hat{\mu}_n = \frac{\sum_{i=1}^{n} p_i X_i}{\sum_{i=1}^{n} p_i}$$

für μ die Verallgemeinerung

$$\sum_{i=1}^{n} p_i \left(X_i - x\right)^2 = \sum_{i=1}^{n} p_i \left(X_i - \hat{\mu}_n\right)^2 + \sum_{i=1}^{n} p_i (\hat{\mu}_n - x)^2$$

des Steinerschen Verschiebungssatzes erfüllt und (b) folgern Sie daraus die Beziehung

$$E(\sum_{i=1}^{n} p_i \left(X_i - \hat{\mu}_n\right)^2) = (n-1)\sigma^2 .$$

3. Vergleich der Sensitivität der mittleren absoluten Abweichung und der Standardabweichung bei Erhöhung des Stichprobenmaximums: Seien $n \geq 2$, $d > 0$ und

$$x_{1:n} \leq ... \leq x_{n-1:n} \leq \begin{cases} x_{n:n} \\ x_{n:n} + d \end{cases} \quad \text{oder}$$

und bezeichne - jeweils im ersten bzw. zweiten Fall -

$\bar{x}_n$, $\bar{x}_n(d)$ ... das Stichprobenmittel
$\tilde{s}_n$, $\tilde{s}_n(d)$ ... die mittlere absolute Abweichung
s_n, $s_n(d)$ ... die Standardabweichung.

Zeigen Sie die Gültigkeit folgender Sachverhalte

$$\text{a)} \quad \tilde{s}_n(d) = \tilde{s}_n + \frac{d}{n}, \quad \text{b)} \quad \bar{x}_n(d) = \bar{x}_n + \frac{d}{n},$$

$$\text{c)} \quad s_n^2(d) = s_n^2 + \frac{d^2}{n}(1 + \frac{2n\,(x_{n:n} - \bar{x}_n)}{(n-1)\,d})$$

und

$$\text{d)} \quad \tilde{s}_n(d) - \tilde{s}_n < \sqrt{s_n^2(d) - s_n^2}\,.$$

4. Beweisen Sie Anmerkung 4 aus Abschnitt 1.4.2.

5. Seien $n_1, n_2 \in \mathbb{N}\backslash\{1\}$, $n = n_1 + n_2$ und $x_1, ..., x_{n_1}, x_{n_1+1}, ..., x_n$ Messwerte aus $\mathbb{R}$ und $\bar{x}_{n_1}$, $s_{n_1}^2$; $\bar{x}_{n_2}$, $s_{n_2}^2$; $\bar{x}_n$, s_n^2 jeweils Stichprobenmittel und Stichprobenvarianz von $x_1, ..., x_{n_1}$; $x_{n_1+1}, ..., x_n$ bzw. $x_1, ..., x_n$.

Zeigen Sie, dass für diese Größen folgende Beziehungen gelten:

$$\bar{x}_n = \frac{n_1}{n} \cdot \bar{x}_{n_1} + \frac{n_2}{n} \cdot \bar{x}_{n_2}$$

und

$$s_n^2 = \frac{n_1 - 1}{n-1} \cdot s_{n_1}^2 + \frac{n_2 - 1}{n-1} \cdot s_{n_2}^2 + \frac{n_1 n_2}{n\,(n-1)} \left(\bar{x}_{n_1} - \bar{x}_{n_2}\right)^2 .$$

6. Die verallgemeinerungsfähige Art, das α-Quantil (der Verteilung) von n Beobachtungswerten $x_1, ..., x_n \in \mathbb{R}$ zu definieren, erfolgt mit Hilfe der

Inversen der zugehörigen empirischen Verteilungsfunktion F_n. Seien nämlich $\alpha \in (0,1)$,

$$F_{n,-}^{-1}(\alpha) = \sup\{x \in \mathbb{R} : F_n(x) < \alpha\} \quad \text{und} \quad F_{n,+}^{-1}(\alpha) = \min\{x \in \mathbb{R} : F_n(x) > \alpha\}.$$

Dann stimmt

$$q_n(\alpha) := (1-\alpha)F_{n,-}^{-1}(\alpha) + \alpha F_{n,+}^{-1}(\alpha)$$

mit der Definition des α-Quantils $q_{\alpha,n}$ in Abschnitt 1.4.3 überein.

Zeigen Sie dies.

7. Bestimmen Sie a) die Fläche des flächengrößten Rechtecks mit vorgegebenem Umfang u und b) das Volumen des volumsgrößten Quaders, dessen Summe der Kantenlängen k ist.

8. **Die Babylonische Quadratwurzelapproximation** (vgl. z.B. [27] [9])

   Seien $0 < x < y < \infty$, und $a(x,y)$, $g(x,y)$ und $h(x,y)$ das arithmetische, geometrische bzw. harmonische Mittel von x und y. Seien weiters $x_0 := 1$ und $y_0 := c \in (1,\infty)$ und es gelten folgende Rekursionsformeln

   $$x_{n+1} := h(x_n, y_n) \quad \text{und} \quad y_{n+1} := a(x_n, y_n), \; n \in \mathbb{N}_0.$$

   Zeigen Sie, dass für alle $n \in \mathbb{N}$ gelten

   a) $1 < x_n < x_{n+1} < \sqrt{c} < y_{n+1} < y_n < c$ und b) $y_n - x_n < \dfrac{c-1}{2^n}$.

   c) Sei $n \in \mathbb{N}_0$. Zeigen sie, dass

   $$\kappa_n := \frac{y_n - \sqrt{c}}{y_n - x_n} = \frac{1}{1 + \frac{\sqrt{c}}{y_n}}$$

   ist und folgern Sie daraus

   $$\frac{1}{2} < \kappa_{n+1} < \kappa_n < \kappa_0 = \frac{1}{1+\frac{1}{\sqrt{c}}} < 1 \;\; \forall\, n \in \mathbb{N}.$$

[9] Eine Verschärfung von b) finden Sie in Abschnitt 3.3 von [39].

d) Sei schließlich $(y_n)_{n\in\mathbb{N}_0}$ die durch

$$y_n = \frac{1}{2}(y_{n-1} + \frac{c}{y_{n-1}})\,,\ n \in \mathbb{N} \quad \text{und} \quad y_0 = c \in (1,\infty)$$

definierte (streng monoton fallende und durch $\sqrt{c}$ nach unten beschränkte) Folge und $\varepsilon \in (0, c-1)$. Wie groß muss n mindestens gewählt werden, damit gilt

$$y_n - \sqrt{c} \leq \varepsilon\,?$$

9. Leiten Sie die *Cauchy-Schwarzsche Ungleichung* - einschließlich des Kriteriums für Gleichheit - durch geeignete Anwendung der Ungleichung zwischen dem geometrischen und dem arithmetischen Mittel her.

10. **Zur Hypergeometrischen Verteilung**

Seien $n, N, s \in \mathbb{N}$ mit $n, s < N$, $w = N - s$ und $m = N - n$. Weiters sei $X_1 \sim H_{n,N,s}$ und die Zufallvariablen X_2, X_3 und X_4 gemäß der nachstehenden Vierfelder-Tafel

	schwarz	weiß	gesamt
in der Stichprobe	X_1	X_2	n
nicht in der Stichprobe	X_3	X_4	m
gesamt	s	w	N.

Dann gelten

$$X_2 = n - X_1 \sim X_{n,N,w}\ ,\ X_3 = s - X_1 \sim X_{m,N,s} \text{ und } X_4 = m - X_3 \sim X_{m,N,w}\,.$$

und somit bekanntlich

$$E(X_1) = \frac{ns}{N}\,,\ E(X_2) = \frac{nw}{N}\,,\ E(X_3) = \frac{ms}{N} \text{ und } E(X_4) = \frac{mw}{N}$$

und

$$\sigma^2 = V(X_1) = \frac{swnm}{N^2(N-1)} = V(X_2) = V(X_3) = V(X_4)\,.$$

a) Verifizieren Sie Letzteres und b) zeigen Sie

$$4\frac{N-1}{N}\sigma^2 = \frac{1}{\frac{1}{4}\left(\frac{1}{E(X_1)} + \frac{1}{E(X_2)} + \frac{1}{E(X_3)} + \frac{1}{E(X_4)}\right)}\,,$$

dass also $4\frac{N-1}{N}\sigma^2$ das harmonische Mittel der Erwartungswerte $E(X_i)$, $i \in \{1,2,3,4\}$ ist.

11. Zwei gerade Gänge der Breite $b = 1$ verlaufen normal zueinander und münden in einer Ecke zusammen. Jemand möchte einen Stab in waagrechter Lage von einem Gang in den anderen befördern ohne ihn an der Ecke zu biegen oder zu knicken. Wie lang kann der Stab maximal sein ?

12. Es seien alle x-Koordinaten der Wertepaare $\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^n$ von 0 verschieden. Ermitteln Sie die Stelle $\hat{k}$ und den Wert $f(\hat{k})$ des Minimums der in Abschnitt 1.6.2 untersuchten Zielfunktion

$$f(k) = \sum_{i=1}^{n} (k \cdot x_i - y_i)^2 = \sum_{i=1}^{n} x_i^2 (k - \frac{y_i}{x_i})^2$$

mit Hilfe von Aufgabe M2 (a).

13. Gegeben seien Wertepaare $\{(x_1, y_1), ..., (x_n, y_n)\} \subseteq \mathbb{R}^n$ derart, dass $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 > 0$ ist. Leiten Sie die Gleichung der allgemeinen Regressionsgeraden durch Minimieren der Funktion

$$f(d, k) = \sum_{i=1}^{n} (k\, x_i + d - y_i)^2$$

mit Hilfe partieller Differentiation her.

Hinweis: Setzen Sie zunächst die partielle Ableitung nach d gleich 0 und setzen Sie den für d erhaltenen Ausdruck anschließend in die partielle Ableitung nach k ein.

14. Es seien $\sigma_x, \sigma_y, c \in (0, \infty)$ und σ_{xy} derart, dass gilt

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \in (-1, 1)\,.$$

Gehen Sie von der durch

$$(\frac{x}{\sigma_x})^2 - 2\rho \cdot \frac{x}{\sigma_x} \cdot \frac{y}{\sigma_y} + (\frac{y}{\sigma_y})^2 = c$$

gegebenen Ellipse aus und überzeugen Sie sich davon,

a) dass die Gleichungen der beiden Geraden, welche durch diejenigen Ellipsendurchmesser bestimmt sind, in deren Endpunkten die Tangenten parallel zu den Koordinatenachsen sind,

$$y = \frac{\sigma_{xy}}{\sigma_x^2} x \quad \text{und} \quad x = \frac{\sigma_{xy}}{\sigma_y^2} y$$

sind und b) dass für den Spezialfall $\sigma_x = \sigma_y = 1$ die beiden Achsen der Ellipse die Geraden $y = x$ und $y = -x$ sind und dass die Längen der großen und kleinen Halbachse

$$l_{1,2} = \sqrt{\frac{c}{1 \mp \rho}}$$

sind.

Hinweis: Die Gleichungen in a) entsprechen denen der beiden Regressionsgeraden in Abschnitt 1.6.4. Vergleichen Sie dazu auch Figure 10.3.4 "Galton's contour lines" in [11], Seite 445, wo man überdies Informationen zur Genese des Bergriffs der Regression findet.

15. a) Zeigen Sie, dass die Spezialisierung des *Pearsonschen Korrelationskoeffizienten*

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n) \cdot (y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y}_n)^2}}$$

für den Fall, dass die x_i und y_i Rangzahlen sind, genauer, dass gilt

$$\{x_1, ..., x_n\} = \{y_1, ..., y_n\} = \{1, ..., n\},$$

folgende Form besitzt

$$r = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\binom{n+1}{3}}\,.$$

Diese Größe heißt der *Spearmansche Rangkorrelationskoeffizient.*

Hinweis: Verwenden Sie die Identität $2xy = x^2 + y^2 - (y - x)^2$.

b) Zeigen Sie überdies

$$r = \frac{\sum_{i=1}^{n}((n + 1 - y_i) - x_i)^2}{\binom{n+1}{3}} - 1\,.$$

Hinweis: Zeigen Sie, dass diese Darstellung im Hinblick auf a) gleichbedeutend mit

$$\binom{n+1}{3} = \frac{1}{2}(\sum_{i=1}^{n}(n+1-(y_i+x_i))^2 + \sum_{i=1}^{n}(y_i-x_i)^2)$$

ist und verifizieren Sie Letztes mit Hilfe der Beziehung

$$x^2+y^2-(n+1)\,(y+x-\frac{n+1}{2}) = \frac{(n+1-(y+x))^2+(y-x)^2}{2}\,.$$

c) Geben Sie unter Zuhilfenahme von a) und b) Kriterien für $r=1$ und $r=-1$ an.

d) Ermitteln Sie für alle Datenpaare $(i,y_i)\in\{1,...,4\}^2$ mit $\{y_1,...,y_4\}=\{1,...,4\}$ den Wert von r. Für welche Datenpaare ist $r=0$?

16. Seien $X_1,...,X_n$ unabhängig identisch $B_{1,p}$-verteilte Zufallsvariable, $p\in(0,1)$. Dann ist der "Standardschätzer" für p die relative Häufigkeit der Erfolge

$$h_n = \frac{S_n}{n}\,.$$

Sei nun

$$W_n = \{\,\boldsymbol{\alpha}=(\alpha_1,...,\alpha_n)\in[0,1]^n : \sum_{i=1}^{n}\alpha_i = 1\,\}$$

die Menge aller Wahrscheinlichkeitsverteilungen auf der Menge $\{1,...,n\}$. Dann definiert

$$\hat{p}_{\boldsymbol{\alpha}} = \sum_{i=1}^{n}\alpha_i\cdot X_i\,,\ \boldsymbol{\alpha}\in W_n$$

eine Familie von Schätzern.

Zeigen Sie:

a) alle Schätzer dieser Familie sind erwartungstreu,

b) für die Varianz jedes Schätzer $\hat{p}_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha}\in W_n$ gilt

$$\frac{p(1-p)}{n} \le V(\hat{p}_{\boldsymbol{\alpha}}) \le p(1-p)\,,$$

wobei

(i) die untere Schranke genau dann angenommen wird, wenn $\hat{p}_{\boldsymbol{\alpha}} = h_n$ ist und

(ii) die obere Schranke genau dann, wenn gilt $\hat{p}_{\boldsymbol{\alpha}} = X_i\,,\ i \in \{1, ..., n\}\,.$

17. Seien $X_1, ..., X_n$ unabhängig identisch $B_{1,p}$-verteilte Zufallsvariable, $p \in (0,1)$ und $h_n = S_n/n$ die relative Häufigkeit der Erfolge. Zeigen Sie, dass

$$\frac{n}{n-1} \cdot h_n\,(1 - h_n)$$

ein ertwartungtreuer Schätzer für die Varianz $p(1-p)$ der $B_{1,p}$-Verteilung ist.

18. Seien $\mu \in \mathbb{R}$, $\sigma^2 > 0$ und $(x_i, Y_i)_{i\in\{1,...,n\}}$ Paare, wobei die x_i feste positive Werte und die Y_i stochastisch unabhängige Zufallsvariable mit Erwartungswert $E(Y_i) = x_i \cdot \mu$ und Varianz $V(Y_i) = \sigma^2 > 0$ sind. Dann sind

$$\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{x_i} \quad \text{und} \quad \sum_{i=1}^{n} \frac{x_i^2}{\sum_{j=1}^{n} x_j^2} \cdot \frac{Y_i}{x_i}$$

zwei Schätzer für $\mu\,.$

Zeigen Sie, dass unter diesen Annahmen a) beide Schätzer erwartungstreu sind und dass b) der zweite Schätzer stets eine kleinere Varianz als der erste besitzt.

19. Bestätigen Sie, dass der Mittelschätzer

$$\hat{M} = 2 \cdot \bar{X}_n - 1$$

erwartungstreu ist und berechnen Sie dessen Varianz für den Fall des Ziehens mit Zurücklegen.

20. Um die Anzahl der von 1 bis N durchnummerierten Kugeln in einer Urne zu schätzen, wird eine Stichprobe vom Umfang $n = 2m + 1 \leq N\,,\ m \in \mathbb{N}\,,$ gezogen. Wie groß muss das Verhältnis $n/\,(N-1)$ des Stichprobenumfangs zum um 1 verminderten Umfang der Grundgesamtheit mindestens sein, damit die Varianz des Medianschätzers $2 \cdot X_{m+1:2m+1} - 1$ beim Ziehen ohne Zurücklegen kleiner ist als die Varianz des Mittelschätzers $2 \cdot \bar{X}_n - 1$ beim Ziehen mit Zurücklegen?

21. Seien $a > 0$ die Seitenlänge eines Quadrats und $X_i = a + \eta_i\,,\ i \in \{1,2,3\}$ drei Messergebnisse von a, wobei diese unabhängig und identisch verteilt sind und für die Fehler η_i gelte $E(\eta_i) = 0$, d.h. dass die Messergebnisse X_i erwartungstreue Schätzer für die Seitenlänge a sind. Weiters sei $\sigma^2 = V(\eta_i)\,,\ i \in \{1,2,3\}$.

    a) Zeigen Sie, dass $\hat{a}_1 = X_1^2$ kein erwartungstreuer Schätzer für die Fläche a^2 des Quadrats ist und ermitteln Sie dessen Bias.

    b) Geben Sie einen erwartungstreuen Schätzer $\hat{a}_2 = f(X_1, X_2)$ für a^2 an, der die beiden Messergebnisse X_1 und X_2 berücksichtigt.

    c) Ermitteln Sie einen erwartungstreuen Schätzer $\hat{a}_3 = g(X_1, X_2, X_3)$ für a^2, der alle drei Messergebnisse berücksichtigt und gleichmäßig kleinste Varianz besitzt.

22. Zeigen Sie, dass gemäß Anmerkung 3 von Abschnitt 2.1.1

$$[X_{1:n}, X_{1:n}/(1 - \sqrt[n]{1-\alpha}))$$

    für den Fall des Ziehens mit Zurücklegen ein Konfidenzintervall von $(1-\alpha)100\%$ iger Sicherheit für den Parameter N ist.

23. Es seien $\hat{p}_n \in [0,1]\,,\ p \in \mathbb{R}$ und $z > 0$. a) Lösen Sie die Ungleichung

$$\left| \frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \right| \leq z$$

    nach p auf, b) bestimmen Sie durch Spezialisierung von Aufgabe 14 a) die Gleichungen der Geraden, welche durch diejenigen Ellipsendurchmesser bestimmt sind, in deren Endpunkten die Tangenten parallel zu den Koordinatenachsen sind, c) überzeugen Sie sich davon, dass die Endpunkte dieser beiden konjugierten Ellipsendurchmesser

$$(\frac{1}{2},\frac{1}{2}) \pm (\frac{1}{2},\frac{1}{2}) \in \{(0,0),(1,1)\} \text{ und } (p,\hat{p}_n) = (\frac{1}{2},\frac{1}{2}) \pm \frac{1}{2}(\sqrt{1+\frac{z^2}{n}}, \frac{1}{\sqrt{1+\frac{z^2}{n}}})$$

    sind und c) fertigen Sie mittels GeoGebra eine graphische Darstellung der Wald'schen Ellipse in Abhängigkeit vom Parameter $c = z_{1-\alpha/2}^2/n \in (0,2]$ an.

24. a) Machen Sie sich die Aussagen von Behauptung 1 in Abschnitt 2.1.4 durch Darstellung der Score-Ellipse mittels GeoGebra in Abhängigkeit vom Parameter $c \in (0, 2]$ zugänglich und b) verifizieren Sie Anmerkung 2.

25. Das Ergebnis jeder Messung einer bestimmten Größe μ sei durch eine normalverteilten Zufallsvariable mit Erwartungswert μ und bekannter Varianz $\sigma^2 > 0$ beschrieben. Es werden n unabhängige Messungen durchgeführt und das Stichprobenmittel $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ der Messergebnisse $X_i\,,\ i \in \{1, ..., n\}$, ermittelt.

    Berechnen Sie - bei vorgegebenem $\alpha \in (0, 1)$ - mit Hilfe von $\bar{X}_n$ ein Konfidenzintervall für μ mit $(1-\alpha)\cdot 100\%$ iger statistischer Sicherheit.

    Hinweis: Machen Sie dabei von der Tatsache Gebrauch, dass $\bar{X}_n$ eine $N(\mu, \sigma^2/n)$-verteilte Zufallsvariable ist.

26. Zeigen Sie, dass die Länge des Score-Konfidenzintervalls genau dann größer als die des Wald'schen Approximationsintervalls ist, wenn gilt

$$\left|\hat{p}_n - \frac{1}{2}\right| > \frac{1}{2}\sqrt{1 - \frac{1}{2 + z^2_{1-\alpha/2}/n}}\,.$$

27. Seien $a, b, c, d \in \mathbb{N}_0$ mit $a \times d + b \times c > 0$. Dann besitzt die Größe

$$\begin{aligned}\kappa &= \frac{a + d - [\frac{(a+b)(a+c)}{N} + \frac{(d+c)(d+b)}{N}]}{N - [\frac{(a+b)(a+c)}{N} + \frac{(d+c)(d+b)}{N}]} \\ &= \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}\,,\end{aligned}$$

    welche κ-*Koeffizient* oder *Cohen's Kappa* genannt wird, den Wertebereich $[-1, 1]$, wobei gilt

$$\kappa = \begin{cases} -1 & \Longleftrightarrow \quad a + d = 0 \ \text{ und } \ b = c \\ 0 & \Longleftrightarrow \quad a \times d - b \times c = 0 \\ +1 & \Longleftrightarrow \quad b + c = 0\,. \end{cases}$$

    Ferner ist der Absolutbetrag des κ-Koeffizienten stets kleiner oder gleich dem des Φ-Koeffizienten.

    Hinweis: Die erste Tatsache ist eine Folgerung aus der Beziehung

$$(a + b)(b + d) + (a + c)(c + d) = 2(ad + bc) + (a + d)(b + c) + (b - c)^2\,,$$

die zweite eine solche aus der Ungleichung

$$\sqrt{xy} \leq (x+y)/2$$

zwischem dem geometrischen und arithmetischen Mittel.

28. a) **Chuquet-Mittel**[10]

Zeigen Sie, dass - ausgehend von der Urnenzusammensetzung

	schwarze Kugeln	weiße Kugeln
Urne 1	a	b
Urne 2	c	d
Urne 1 $\cup$ Urne 2	$a+c$	$b+d$

mit $a, c \in \mathbb{N}_0$ und $b, d \in \mathbb{N}$ - für das *Chuquet-Mittel* $\frac{a+c}{b+d}$ von $\frac{a}{b}$ und $\frac{c}{d}$ gilt

$$\frac{a}{b} < \frac{c}{d} \implies \frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}.$$

b) **Simpson's Paradoxon** (vgl. [22])

Situation A: Auf zwei Tischen befinden sich jeweils ein schwarzer und ein grauer Hut, in denen sich - in einer bestimmte Bestückung $a_i, b_i, c_i, d_i \in \mathbb{N}$, $i \in \{1, 2\}$ - schwarze und weiße Kugeln befinden.

Situation B: Man vereinigt die Inhalte der beiden schwarzen Hüte in einem schwarzen Hut und die der beiden grauen Hüte in einem grauen Hut.

schwarzer Hut	schwarze Kugeln	weiße Kugeln
Tisch 1	a_1 (5)	b_1 (6)
Tisch 2	c_1 (6)	d_1 (3)
Vereinigung	$a_1 + c_1$ (11)	$b_1 + d_1$ (9)

grauer Hut	schwarze Kugeln	weiße Kugeln
Tisch 1	a_2 (3)	b_2 (4)
Tisch 2	c_2 (9)	d_2 (5)
Vereinigung	$a_2 + c_2$ (12)	$b_2 + d_2$ (9)

[10] *Nicolas Chuquet* (1445 – 1488), französischer Arzt

Ziel sei es, sowohl in Situation A als auch in Situation B jenen Hut zu wählen, für welchen - unter Laplace-Annahme - die Wahrscheinlichkeit, eine schwarze Kugel zu ziehen, größtmöglich ist.

Angenommen, die Bestückung der Hüte auf den beiden Tischen erfülle

$$\frac{a_1}{b_1} > \frac{a_2}{b_2} \text{ und } \frac{c_1}{d_1} > \frac{c_2}{d_2}\,. \tag{3}$$

Demnach ist in Situation A die Wahrscheinlichkeit, eine schwarze Kugel zu ziehen, dann größtmöglich, wenn man sowohl von Tisch 1 als auch von Tisch 2 den schwarzen Hut wählt. Dann ist wohl auch in Situation B die Wahl des schwarzen Hutes optimal. Oder etwa nicht ?

b1) Überzeugen Sie sich davon, dass für die in Klammern stehende Bestückung der Hüte Bedingung (3) und zugleich

$$\frac{a_1 + c_1}{b_1 + d_1} < \frac{a_2 + c_2}{b_2 + d_2}$$

gilt (*Simpson's Paradoxon*). b2) Geben Sie andere Bestückungen mit derselben Eigenschaft an und b3) begründen Sie, wodurch dieses Paradoxon zustand kommt.

29. a) Zeigen Sie die Aussage

$$E[\chi^2_m\,(n, P)] = m$$

in Anmerkung 1 von Abschnitt 2.4.1.

Sei X eine reelle Zufallsvariable mit Erwarungswert μ und Standardabweichung $\sigma > 0$. Dann heißt die Größe

$$\beta_2 = E[(\frac{X - \mu}{\sigma})^4]$$

Wölbung (*kurtosis*, *peakedness*) der zugehörigen Verteilung. Die Verteilung heißt

flachgipfelig (*leptopkurtic*)	für	$\beta_2 < 3$
normalgipfelig (*mesokurtic*)	für	$\beta_2 = 3$
steilgipfelig (*platykurtic*)	für	$\beta_2 > 3\,.$

b) Ermitteln Sie die Wölbung der $N(0, 1)$-Verteilung und der Binomialverteilung $B_{n,p}$. Letzteres aus der Spezialisierung von $V\,[\chi^2_m\,(n, P)]$ für

$m = 1$. (Ein Vergleich der Wölbung der $N(0,1)$-Verteilung mit jener einer $B_{n,p}$-Verteilung für große n "bestätigt" die Gültigkeit des zentralen Grenzverteilungssatzes.)

c) Unterscheiden Sie die Art der "Gipfeligkeit" der $B_{n,p}$-Verteilung in Abhängigkeit von $p \in (0,1)$.

30. Zeigen Sie folgende

Behauptung zur Klasseneinteilung: Seien $b \geq 0$ und $c > 0$. Dann gilt für alle β, $\gamma \in (0,1)$

$$\frac{(b-c)^2}{c} \leq \frac{(b \cdot \beta - c \cdot \gamma)^2}{c \cdot \gamma} + \frac{(b(1-\beta) - c\,(1-\gamma))^2}{c(1-\gamma)}\,,$$

bzw., für $k = \frac{b}{c}$ und $\beta = \gamma + \varepsilon$, $|\varepsilon| < \min(\gamma, 1-\gamma)$, gleichbedeutend

$$(k-1)^2 \leq \frac{((k-1)\,\gamma + k \cdot \varepsilon)^2}{\gamma} + \frac{((k-1) \cdot (1-\gamma) - k \cdot \varepsilon)^2}{1-\gamma}\,.$$

Dabei gilt Gleichheit genau dann, wenn $b = 0$ oder $\varepsilon = 0$ ist.

Literaturverzeichnis

[1] **Bücher**

[2] *Bickel, P.J.* and *K.A. Doksum*: Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day Inc., San Franciso 1977

[3] *Bickel, P.J., Hammel, E.A.* and *J.W. O'Connell*: Sex Bias in Graduate Admissions: Data from Berkeley. Science, Vol. 187, S. 398 – 404 (1975)

[4] *Freedman, D., Pisani, R.* and *R. Purves*: Statistics. Norton & Co., New York 1978

[5] *Gonick, L.* and *W. Smith*: The Cartoon Guide to Statistics. Harper Perennial, New York 1993

[6] *Hartung, J., Elpelt, B. und K-H. Klösener*: Statistik: Lehr- und Handbuch der angewandten Statistik. Oldenbourg Verlag, München - Wien 1991

[7] *Koyré, A.*: Leonardo, Galilei, Pascal - Die Anfänge der neuzeitlichen Naturwissenschaft. Fischer Taschenbuch Verlag, Frankfurt am Main 1998

[8] *Krämer, W.*: Statistik verstehen: Eine Gebrauchsanweisung. Campus Verlag, Frankfurt - New York, 1999

[9] *Krämer, W.*: So lügt man mit Statistik. Campus Verlag, Frankfurt - New York, 1997

[10] *Kröpfl, B., Peschek, W., Schneider, E.* und *A. Schönlieb*: Angewandte Mathematik - Eine Einführung für Wirtschaftswissenschaftler und Informatiker. Carl Hanser Verlag, München - Wien 1994

[11] *Larsen, R.J.* and *M.L. Marx*: An Intoduction to Mathematical Statistics and its Applications. Prentice-Hall, Englewood Cliffs, New Yersey 1986 / 2006

[12] *Lehn, J.* und *H. Wegmann*: Einführung in die Statistik. Wissenschaftliche Buchgesellschaft, Darmstadt 1985

[13] *Moore, D.S.*: Statistics: Concepts and Controversies, W.H. Freeman & Co., San Francisco 1979 / New York 2001

[14] *Moore, D.S.* and *G.P. McCabe*: Introduction to the Practice of Statistics. W.H. Freeman & Co., New York 2004

[15] *Noelle-Neumann, E.*: Umfragen in der Massengesellschaft: Einführung in die Methoden der Demoskopie. Rowohlt Taschenbuch Verlag, Reinbeck bei Hamburg 1963

[16] *Rao, C.R.*: Was ist Zufall? : Statistik und Wahrheit. Prentice Hall, München 1995

[17] *Sexl, R., Raab, I.* und *E. Streeruwitz*: Materie in Raum und Zeit - Eine Einführung in die Physik, Band 3, Verlag Sauerländer, Aarau - Frankfurt am Main - Salzburg 1996

[18] *Silvey, S.D.*: Statistical Inference. Chapman and Hall, London - New York 1975

[19] *Tanur, J.M.* (Edt.): Statistics: A Guide to the Unknown. Holden-Day, San Fransico 1978

[20] *Topsoe, F.*: Spontane Phänomene - Stochastische Modelle und ihre Anwendungen. Vieweg & Sohn Verlagsgesellschaft, Braunschweig 1990

[21] *Schönwiese, Ch.-D.*: Praktische Statistik für Metheorologen und Geowissenschaftler. Gebrüder Borntraeger, Berlin-Stuttgart 2006

Zeitschriften

[22] *Gardner, M.*: On the Fabric of Intuitive Logic, and some Probability Paradoxes. In: Mathematics: An Introduction to its Spirit and Use. Scientific American, W.H. Freeman & Co., San Francisco 1978

[23] *Engel, A.*: Statistik in der Schule: Ideen und Beispiele aus neuerer Zeit. Der Mathematikunterricht, Jahrgang 28, Heft 1 (1982), S. 57 – 85

[24] *Kütting, H.*: Stochastisches Denken in der Schule - Grundlegende Ideen und Methoden. Der Mathematikunterricht, Heft 4 (1985), S. 87 – 106

[25] *Agresti, A.* and *B.A. Coull*: Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician **52** (1998), 119 – 126

[26] *Diepgen, R.*: Warum nur $n-1$ und nicht n? Erwartungstreue - leicht gemacht. Stochastik in der Schule **19** (1999), Nr. 1, S. 10 – 13

[27] *Hirscher, H.*: Mittelwertfolgen - oder: Mitten inmitten von Mitten. Der Mathematikunterricht, Jahrgang 50, Heft 5 (2004), S. 42 – 55

Lehrveranstaltungsunterlagen

[28] *Österreicher, F.*: Ausgewählte Kapitel der Statistik. Salzburg 1985

Unterlagen zur Lehrer/innen/fortbildung

[29] *Österreicher, F.*: Der Salzburger Jedermannlauf oder: Die Anwendung der Statistik für Spionagezwecke. Lehrer/innenfortbildungstag "West", Innsbruck 2008

[30] *Österreicher, F.*: Analyse zweidimensionaler Daten: Regression und Korrelation. Lehrer/innenfortbildungstag "West", Salzburg 2009

[31] *Österreicher, F.*: Konfidenzintervalle. Lehrer/innenfortbildungstag "West" Innsbruck 2010

Diplom-, Magister- und Masterarbeiten

[32] *Schwanninger, Ch.*: Anwendungsaspekte der Familie der Gammaverteilungen - mit einer statistischen Auswertung der Tagesniederschlagsmengen in Salzburg. Salzburg 1994

[33] *Weiß, M.*: Binomialverteilung und Normalapproximation: Grundlegendes und Hintergrundinformation für den Stochastikunterricht. Salzburg 1995

[34] *Jandl, M.*: Computereinsatz im Stochastikunterricht. Salzburg 1997

[35] Kolmberger, M.: Statistik in der Nußschale: "Ist unser Würfel fair?" Salzburg 1997

[36] *Radauer, Ch.*: Läufe in binären Zufallsfolgen - Beurteilung und Kompression von Zufallsdaten. Salzburg 1998

[37] *Fritz, F.*: Proportionen in Mathematik und Musik - Kunst und Wissenschaft ergänzen einander. Salzburg 1998

[38] *Dürager, H.-P.*: Stetige Modelle in der Stochastik. Diplomarbeit 2007

[39] *Reichenberger, S.*: Mittelwerte der Pythagoreer und babylonischer Wurzelalgorithmus. Diplomarbeit 2007

[40] *Oberrrauner, S.*: Licht und Schatten als zentrale Elemente in der Geschichte der Astronomie. Salzburg 2008

[41] *Hemetsberger, M.*: Nutzung der Kryptographie für den Unterricht - Cäsar- und Vigenére Code. Salzburg 2008

[42] *Huber, Ch.*: Die Radioaktivität als Hilfsmittel der Geochronologie - Von Röntgen bis DAREosDAT. Salzburg 2008

[43] *Guggenberger, A.*: Die Erfindung des Konfidenzintervalls und dessen frühe Anwendungen. Salzburg 2008

[44] *Müller, C.*: Normalapproximation der Hypergeometrischen Verteilung. Salzburg 2008

[45] *Eichbauer, F.*: Testen von Hypothesen - Eine Aufbereitung für den Unterricht. Salzburg 2009

[46] *Naderer, C.*: Log-optimale und semi-log-optimale Portfolios. Salzburg 2009

[47] *Erla, S.*: Zur Geschichte erwartungstreuer Schätzer. Salzburg 2009

[48] *Lackner, M.*: Die Lebensversicherungsmathematik von den Anfängen bis 1914. Salzburg 2009

[49] *Morocutti, U.*: Power laws - Wahrscheinlichkeitstheoretische Modelle und statistische Anwendungen. Salzburg 2010

[50] *Wiedecke, M.*: Erwartungswert und Varianz - Ein mathematischer und geschichtlicher Überblick. Salzburg 2010

für Allgemeinbildende Höhere Schulen

[51] *Bürger·Fischer·Malle·Kronfellner·Mühlbacher·Schlöglhofer*: Mathematik Oberstufe 3. Verlag Hölder Pichler Tempsky, Wien 1992

[52] *Bürger·Fischer·Malle·Kronfellner·Mühlbacher·Schlöglhofer*: Mathematik Oberstufe 4. Verlag Hölder Pichler Tempsky, Wien 1992

[53] *Novak·Bolhar-Nordenkampf·Schalk·Stemmer*: Mathematik Oberstufe 3. Reniets Verlag, Wien 1991

[54] *Novak·Bolhar-Nordenkampf·Schalk·Stemmer*: Mathematik Oberstufe 4. Reniets Verlag, Wien 1991

[55] *Reichel·Müller·Hanisch·Laub*: Lehrbuch der Mathematik 7. Verlag Hölder Pichler Tempsky, Wien 1992

[56] *Reichel·Müller·Hanisch·Laub*: Lehrbuch der Mathematik 8. Verlag Hölder Pichler Tempsky, Wien 1992

[57] *Szirucsek·Dinauer·Unfried·Schatzl*: Mathematik 7. Verlag Hölder Pichler Tempsky, Wien 1991

[58] *Szirucsek·Dinauer·Unfried·Schatzl*: Mathematik 8. Verlag Hölder Pichler Tempsky, Wien 1992

für Handelsakademien

[59] *Kronfellner, M. Peschek·Blasonig·Fischer·Kronfellner, J.*: Angewandte Mathematik 4. Verlag Hölder Pichler Tempsky, Wien 1997

[60] *Schneider·Thannhauser*: Mathematik Arbeitsbuch und Aufgabensammlung einschließlich Lösungen. Band 4 für den V. Jahrgang HAK. Rudolf Trauner Verlag, Linz 1999

[61] *Steiner·Weilharter*: Mathematik und ihre Anwendungen in der Wirtschaft. Band 4. Reniets Verlag, Wien 1998

für Höhere Technische Lehranstalten

[62] *Schärf*: Mathematik 2 für HTL. Oldenbourg Verlag, Wien 1993

[63] *Schärf*: Mathematik 3 für HTL. Oldenbourg Verlag, Wien 1994

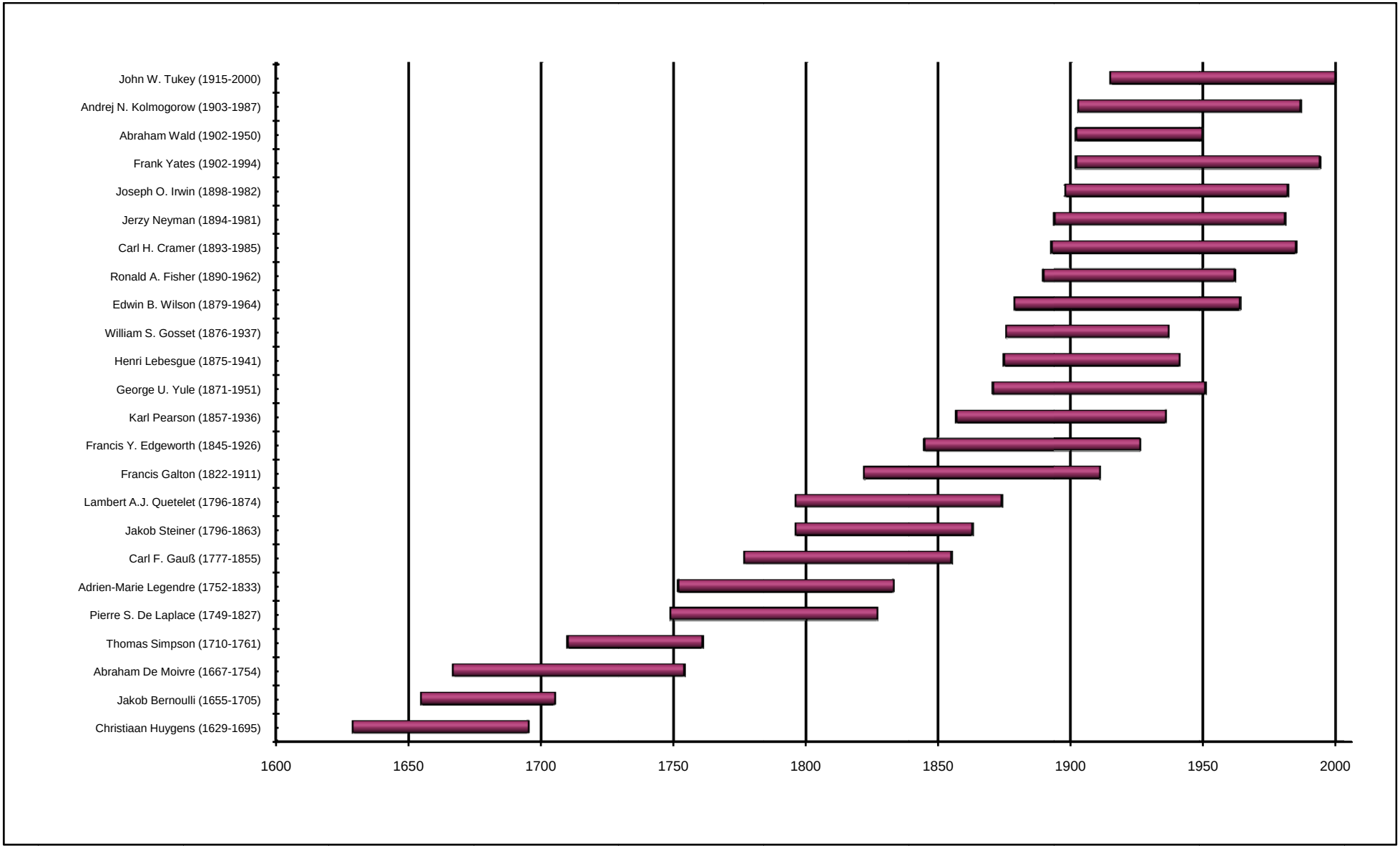
[64] *Schalk·Aubauer·Bolhar-Nordenkampf·Dorninger·Nöbauer·Plihal·Spitzer*: Mathematik 3 für HTL. Reniets Verlag, Wien 1988

Schulbücher aus Deutschland

[65] *Barth·Haller*: Stochastik Leistungskurs. Ehrenwirth Verlag, München 1983

[66] *Heigl·Feuerpfeil*: Stochastik Leistungskurs. Bayrischer Schulbuch Verlag, München 1987

[67] *Lambacher·Schweizer*: Stochastik Leistungskurs. Ernst Klett Schulbuchverlag, Stuttgart 1988


John W. Tukey (1915-2000)
Andrej N. Kolmogorow (1903-1987)
Abraham Wald (1902-1950)
Frank Yates (1902-1994)
Joseph O. Irwin (1898-1982)
Jerzy Neyman (1894-1981)
Carl H. Cramer (1893-1985)
Ronald A. Fisher (1890-1962)
Edwin B. Wilson (1879-1964)
William S. Gosset (1876-1937)
Henri Lebesgue (1875-1941)
George U. Yule (1871-1951)
Karl Pearson (1857-1936)
Francis Y. Edgeworth (1845-1926)
Francis Galton (1822-1911)
Lambert A.J. Quetelet (1796-1874)
Jakob Steiner (1796-1863)
Carl F. Gauß (1777-1855)
Adrien-Marie Legendre (1752-1833)
Pierre S. De Laplace (1749-1827)
Thomas Simpson (1710-1761)
Abraham De Moivre (1667-1754)
Jakob Bernoulli (1655-1705)
Christiaan Huygens (1629-1695)
1600
1650
1700
1750
1800
1850
1900
1950
2000