Inhaltsverzeichnis

Lizenz

Kapitel 1

Einführung

Was ist Statistik?

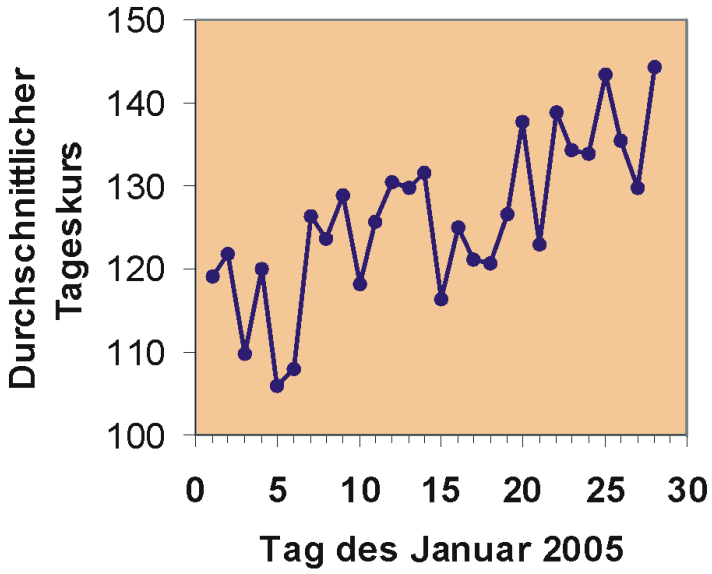

Abbildung 1: Durchschittliche Tageskurse der Firma Dachs AG im Januar 2005

Statistik begegnet uns überall im täglichen Leben:

- Die Lebenshaltungskosten sind gegenüber dem Vorjahr um 2 Prozentpunkte gestiegen.

- Im Januar 2005 erzielte die Firma Dachs im Durchschnitt die täglichen Aktienkurse, wie in der Grafik angegeben.
- Hochrechnung von Wahlergebnissen
- Wieviel Gewinn kann eine Lottogesellschaft auswerfen, damit ihr noch Überschuss bleibt?

Was haben diese Beispiele gemeinsam? Sie basieren auf Daten, und zwar sehr vielen Daten. In diese Daten wird Ordnung gebracht: Mit einer Grafik, mit Wahrscheinlichkeiten, mit Durchschnittsberechnungen, mit Vergleichen. Das ist angewandte Statistik.

Wir kommen damit zu einer Definition der Statistik, die relativ kurz und schnörkellos ist, aber im Wesentlichen alles sagt:

> **Statistik ist die Gesamtheit der Methoden, die für die Untersuchung von Massendaten angewendet werden können.**

Ziel der Statistik ist es also, Massendaten zu reduzieren und zu komprimieren, um Gesetzmäßigkeiten und Strukturen in den Daten sichtbar zu machen.

Anwendung im wirtschaftlichen Kontext

Die Lage der Unternehmen heute ist geprägt von Globalisierung, Konkurrenz und Kostendruck. Einsame Manager-Entscheidungen aus dem Bauch heraus führen häufig zum Ruin des Unternehmens. Die Analyse von Wirtschafts- und Unternehmensdaten erlaubt rationale und fundierte Unternehmensentscheidungen. In der Realität sind jedoch Informationen über Unternehmensprozesse nur teilweise bekannt. Gründe dafür sind beispielsweise

1. Die Informationen sind zu komplex, um vollständig erhoben zu werden. **Beispiel:** Der Papierverbrauch in einem großen Unternehmen hängt von vielen Faktoren ab, wie der Zahl der Kopien eines Schreibens, der Neigung der Mitarbeiter, sich alles ausdrucken zu lassen (E-Mails!), dem Umfang des Verteilers für bestimmte Schreiben etc. Man kann den Verbrauch nicht analytisch bestimmen.
2. Zukünftige Unternehmenszahlen sind nicht bekannt und müssen geschätzt werden, z. B. der Cash-Flow einer geplanten Investition für die Finanzierungsrechnung.

3. Umwelteinflüsse können nicht vorherbestimmt werden, etwa die „Gefahr“ einer Steuererhöhung oder die Akzeptanz eines neuen Produkts durch den Kunden.

In solchen Fällen können keine exakten Entscheidungsgrundlagen geliefert werden. Die resultierenden Lösungen sind „unscharf“. Ein Hilfsmittel für die Entscheidung unter Unsicherheit ist die Statistik.

Da in der heutigen informationsbasierten Gesellschaft eher zu viel als zu wenig Daten verfügbar sind, gewinnt die Statistik als Werkzeug der Entscheidungsfindung immer mehr an Bedeutung.

Einteilung der statistischen Methoden

1. **Deskriptive (beschreibende, empirische) Statistik:** Man untersucht ein Phänomen und fasst die Daten zusammen, ordnet sie, stellt sie grafisch dar. Auf wissenschaftliche Aussagen wird verzichtet.
2. **Induktive (schließende, folgernde, mathematische, analytische) Statistik:** Grundlage ist die Wahrscheinlichkeitstheorie. Ergebnisse der deskriptiven Statistik dienen häufig als Ausgangspunkt für verallgemeinernde Aussagen.

Die mathematische Statistik selbst ist wie die Wahrscheinlichkeitstheorie ein Teilgebiet der Stochastik.

Kapitel 2

Wahrscheinlichkeitsrechnung

Was ist **Wahrscheinlichkeit**?

Das weiß niemand. Sie ist ein Produkt menschlicher Bemühungen, Ereignisse in der Zukunft vorherzusagen. Sie soll eine Vorstellung über den Grad der Sicherheit vermitteln, mit der ein Ereignis auftritt. Jeder weiß, was es bedeutet, wenn ich sage: Die Wahrscheinlichkeit, eine Sechs zu würfeln ist größer als die Wahrscheinlichkeit, beim Skat einen Grand zu gewinnen. Aber trotzdem kann man Wahrscheinlichkeit nicht exakt definieren. So könnte man Wahrscheinlichkeitstheorie als Stochern im Nebel bezeichnen. Das hat aber nichts mit dem Begriff Stochastik zu tun!

Pizzaecken-Beispiel zum Begriff der Wahrscheinlichkeit

Harry und Paula gehen in die Pizzeria. Sie sind frisch verliebt. Paula bestellt sich eine Pizzaecke mit Salami und Harry eine mit Schinken. Dann tauschen sie jeweils eine Hälfte, wobei anzumerken ist, dass die Ecken sich in Rand- und Mittelstück teilen lassen. Obwohl Harry normalerweise Randstücke lieber mag, achtet er in seinem aktuellen Zustand nicht darauf. Und auch Paula gibt ihre Hälfte rein nach Zufall ab.

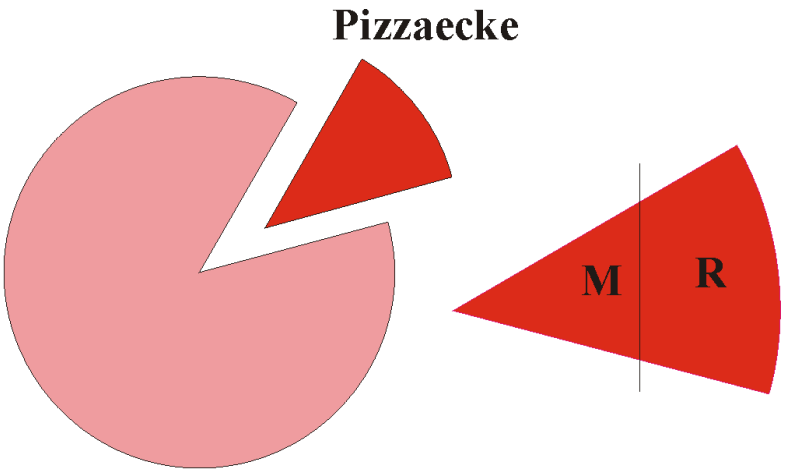

Abbildung 2: Pizzaecke

Wie groß ist eigentlich die Wahrscheinlichkeit, dass Harry zwei Randstücke auf dem Teller hat?

Die Meisten antworten richtig: 1/4.

Aber wieso beträgt die Wahrscheinlichkeit ausgerechnet 1/4?

Betrachten wir den Vorgang:

Bei **gleicher Ausgangslage (Bedingungskomplex)** kann der Versuch, zwei halbe Pizzaecken zufällig auszutauschen, **beliebig oft wiederholt** werden. Jeder Versuch hat einen **unsicheren Ausgang**. Es handelt sich hier um einen **Zufallsvorgang** (**Experiment**, **Versuch**).

Der Zufallsvorgang wird also beschrieben durch:

- Gleicher Bedingungskomplex
- Unsicherer Ausgang
- Beliebig oft wiederholbar

Ein bestimmtes Paar Eckhälften auf Harrys Teller ist ein **Ergebnis**. Ein Ergebnis wäre beispielsweise: Die erste Hälfte ist ein Randstück, die zweite Hälfte ist ein Mittelstück,

(R;M) oder kurz RM,

wobei das "linke" Stück von Harry stammt und das "rechte" von Paula.

Alle möglichen Paare fasst man in der **Ergebnismenge** Ω zusammen:

$\Omega = \{\text{RR, RM, MR, MM}\}$.

Ω ist also die Menge aller möglichen Ergebnisse, die bei einem Zufallsvorgang auftreten können. Führt man diesen Zufallsvorgang unendlich oft durch, müssten vermutlich in 25% aller Versuche zwei Randstücke resultieren, denn man könnte davon ausgehen, dass jedes Paar die gleiche Wahrscheinlichkeit hat, gezogen zu werden. Die Zahl der Ergebnisse, $|\Omega|$ genannt, ist also vier. Deshalb ist die Wahrscheinlichkeit für ein Paar Randstücke

$$P(RR) = \frac{1}{4} .$$

Wenn nun bei einem Versuch beispielsweise "RM" resultiert, ist das ein **Ereignis**.

Bei "RM" handelt es sich um ein **Elementarereignis**. Es ist ein Ereignis, das nur **ein** Element der Ergebnismenge enthält.

Es gibt auch kompliziertere, **zusammengesetzte** Ereignisse:

A: Mindestens ein Mittelstück: A = {RM, MR, MM}
B: Eine komplette Pizzaecke: B = {RM, MR}

Diese Ereignisse beinhalten mehrere Ergebnisse von Ω; ein Ereignis ist immer eine Teilmenge von Ω.

Die Wahrscheinlichkeit als theoretisches Konzept

Kurzer geschichtlicher Überblick

Es werden vermutlich schon so lange Wahrscheinlichkeiten angewendet, wie es den Homo Sapiens gibt. Am letzten Tag der Schlacht im Teutoburger Wald (9 n. Chr.) gab es ein Gewitter. Die Römer deuteten es als warnenden Hinweis von Merkur, des Gottes von Blitz und Donner. Die Germanen sahen es als Aufmunterung des Kriegsgottes Thor. Wie man weiß, hatten beide Parteien recht.

Im 17. Jahrhundert, dem Zeitalter des Rationalismus, befasste sich **Blaise Pascal** (1623 - 1662) systematisch mit Wahrscheinlichkeiten im Glücksspiel und begründete so die Wahrscheinlichkeitsrechnung als eigenständige Disziplin.

Jakob Bernoulli (1654 - 1705) befasste sich ebenfalls mit Fragen der diskreten Wahrscheinlichkeiten und gab vermutlich das erste Buch über Wahrscheinlichkeitsrechnung heraus.

Mit **Abraham de Moivre** (1667 - 1754) und **Pierre Simon Laplace** (1749 - 1827) wurde bereits die Normalverteilung entwickelt und von **Carl Friedrich Gauß** (1777 – 1855) weiter bearbeitet.

Richard Edler von Mises (1883 - 1953) lieferte wertvolle Beiträge zur Schätzung von Wahrscheinlichkeiten und zur mathematischen Statistik.

1933 schlug der russische Mathematiker **Andrej Nikolajewitsch Kolmogorow** (1903 - 1987) eine **axiomatische Definition der Wahrscheinlichkeit** vor, auf der die heutige Wahrscheinlichkeitstheorie basiert. Diese Definition ist eine Anwendung der Maßtheorie.

Ergebnisse und Ereignisse

Das heutige Konzept der Wahrscheinlichkeitsrechnung präsentiert sich folgendermaßen:

Gegeben ist die **Ergebnismenge** (Ereignisraum,Stichprobenraum) Ω **eines Zufallsvorgangs**. Diese Menge enthält alle möglichen Ergebnisse, die ein Zufallsvorgang hervorbringen kann. Je nach Art des Zufallsvorgangs muss man verschiedene Ergebnismengen betrachten:

Ω enthält **endlich** viele Ergebnisse.

Beispiele:

- Zufallsvorgang 1x Würfeln. $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Zufallsvorgang: Augenfarbe der nächsten Person, die bei einem Casting vorspricht. $\Omega = \{\text{blau, grün, braun}\}$.

Ω enthält **abzählbar unendlich** viele Ergebnisse.

Beispiele:

- Zufallsvorgang: Zahl der Autos, die eine Stunde lang ab 12 Uhr bei einer Fahrzeugzählung an einer bestimmten Zählstelle vorbeifahren. $\Omega = \{0, 1, 2, 3, ...\}$.
- Zufallsvorgang: Zahl der Anforderungen an einen Server innerhalb einer Stunde. $\Omega = \{0, 1, 2, ...\}$.

> Man kann zwar die Ergebnisse durchzählen, aber es kann keine vernünftige Obergrenze angegeben werden, deshalb lässt man die Obergrenze offen.

Ist Ω weder abzählbar noch abzählbar unendlich, so enthält Ω **überabzählbar** viele Ergebnisse. Man könnte auch sagen, die Ergebnismenge ist ein Intervall der reellen Zahlen.

Beispiele:

- Zufallsvorgang: Eine erwachsene Person wird gewogen (in kg). $\Omega = \{x|30 \leq x \leq 200; x \in R\}$.
- Zufallsvorgang: Cash-Flow eines Unternehmens (in €). $\Omega = R$.

  Cash-Flow bezeichnet übrigens die Differenz Einnahmen - Ausgaben, bzw. präziser: Einzahlungen - Auszahlungen.

  Hier können die Ergebnisse nicht mehr abgezählt werden. Ein beliebig kleines Intervall der Ergebnismenge enthält unendlich viele Elemente. Was ist das nächstgrößere Element von 50 kg: 51 kg, 50,01 kg oder 50,000000001 kg? Im Intervall [50, 51] sind also unendlich viele Elemente.

  > Man könnte hier einwenden, dass doch beispielsweise Cash-Flow als kleinste Einheit Cent hat, also doch eigentlich abzählbar ist. Das stimmt natürlich, aber bei sehr vielen, nah zusammenliegenden Elementen vereinfacht man die Analyse, indem man die Menge als stetig annimmt. Man spricht hier von Quasistetigkeit.

Hat ein Zufallsvorgang ein konkretes Ergebnis erbracht, ist ein **Ereignis** eingetreten. Es gibt einfache Ereignisse, die lediglich ein Ergebnis enthalten, so genannte **Elementarereignisse** und es gibt komplexere Ereignisse, die sich aus mehreren Ergebnissen zusammensetzen. **Ein Ereignis A ist immer eine Teilmenge der Ergebnismenge** Ω.

Da Ereignisse Mengen sind, können alle Operationen der **Mengenalgebra**, die mit der **Booleschen Algebra** (auch Schaltalgebra) gleichgesetzt werden kann, angewendet werden. Grundlegende Operationen für Mengen der Booleschen Algebra sind $\bar{}$ (||nicht|| als Komplement), $\cap$ und $\cup$. Alle anderen Operationen können daraus hergeleitet werden.

Alle interessierenden Ereignisse fasst man nun in einer so genannten **Ereignismenge** (**Ereignissystem**) **E** zusammen. **E** ist also eine Menge von

Teilmengen. Damit diese Menge mit der Booleschen Algebra bearbeitet werden kann, muss sie entsprechende Forderungen erfüllen:

- Wenn das Ereignis A in **E** enthalten ist, muss auch sein Komplement $\bar{A}$ enthalten sein.
- Wenn A und B enthalten sind, muss auch A $\cup$ B enthalten sein (Man kann ausrechnen, dass dann auch A $\cap$ B enthalten ist).
- Es muss das ||Null-Element|| Ø enthalten sein (Das impliziert, dass auch ||1-Element|| Ω , welches das Komplement von Ø ist, enthalten ist).

Die umfassendste Ereignismenge ist die Potenzmenge $\boldsymbol{P}$, die alle Teilmengen von Ω enthält.

Beispiel einer Potenzmenge:

Zufallsvorgang: Aus einer Urne mit einer blauen (b), einer roten (r) und einer gelben (g) Kugel wird eine Kugel gezogen. Wir interessieren uns für die Farbe der Kugel.

Ergebnismenge: Ω = {g, b, r}

Potenzmenge: $\boldsymbol{P}$ = {Ø, {r}, {g}, {b}, {r, g}, {r, b}, {g, b}, {r, g, b}}

Ausgehend von dieser Konstellation hat **Kolmogorow** mit seinen Axiomen ein **Wahrscheinlichkeitsmaß** konstruiert, d.h. eine Abbildung der Ergebnismenge Ω auf die Menge der reellen Zahlen im Intervall [0;1]:

F: $\Omega \rightarrow R$; A $\rightarrow$ P(A)

Eine Funktion P, die jedem Ereignis A aus **E** eine reelle Zahl zuordnet, heißt Wahrscheinlichkeit, wenn sie folgende Axiome erfüllt:

Axiome der Wahrscheinlichkeiten:

Gegeben sind zwei Ereignisse A,B $\subset$ Ω.

1. $P(A) \geq 0$. **Nichtnegativität**
2. $P(\Omega) = 1$. **Normiertheit**
3. $P(A \cup B) = P(A) + P(B)$, falls A und B disjunkt sind. **Additivität**

Dieses Axiomensystem kann nur auf endlich viele Ereignisse angewendet werden. Für unendlich viele Ereignisse A_i (i = 1, 2, ...) erhält man statt der

endlichen Ereignismenge die **σ-Algebra**. Sie enthält alle geforderten Eigenschaften der Ereignismenge auf unendlich viele Ereignisse A_i ausgeweitet. Hier wird das 3. Axiom entsprechend angepasst:

> 3. Sind die Ereignisse A_i sämtlich paarweise disjunkt, ist bei ihrer Vereinigung
>
> $$P(A_1 \cup A_2 \cup A_3 ...) = P(A_1) + P(A_2) + P(A_3) + ...,$$
>
> (σ-Additivität).

Berechnung der Wahrscheinlichkeit eines Ereignisses

Es müssen nun noch die Ereignisse mit Wahrscheinlichkeiten ausgestattet werden. Auf welche Weise das geschehen soll, ist in den Axiomen nicht angegeben. Es gibt hier verschiedene Verfahren. Man erhält schließlich die Wahrscheinlichkeitsverteilung.

Wie ordnen wir den Ereignissen am besten Wahrscheinlichkeiten zu?

Betrachten wir im Pizzaecken-Beispiel das Ereignis A: Mindestens ein Mittelstück. Es ist A = {RM, MR, MM}. A belegt in Ω drei von vier möglichen Ergebnissen, also ist die Wahrscheinlichkeit P(A) = 3/4. Diese Vorgehensweise entspricht der **Klassischen Wahrscheinlichkeitsauffassung**. Man bezeichnet sie als **Symmetrieprinzip** oder **Prinzip nach LAPLACE**:

Jedes Ergebnis ist gleich häufig. |A| ist die Zahl der Ergebnisse, die durch A belegt werden (Anzahl der günstigen Ergebnisse), $|\Omega|$ ist die Zahl aller möglichen Ergebnisse. Es ist

$$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{4}\,.$$

Das Symmetrieprinzip hat allerdings den Nachteil, dass es nicht bei allen Zufallsvorgängen angewendet werden kann, z.B. bei unendlich vielen Ergebnissen. Oft ordnet man auch Ergebnissen unterschiedliche Wahrscheinlichkeiten zu, z.B.

Zufallsvorgang: Wetter von heute.
Ergebnismenge Ω = {schön, schlecht}.
P(‖schön‖) = 0,6, P(‖schlecht‖) = 0,4.

Wie kommt man auf diese Wahrscheinlichkeiten 0,4 und 0,6? Man hat in diesem Fall etwa die Wetteraufzeichnungen der letzten 100 Jahre ausgewertet und hat festgestellt, dass der Anteil der schönen Tage 60 % betrug. Wir haben hier eine Anwendung der **Statistischen Wahrscheinlichkeitsauffassung**: Man führt ein Zufallsexperiment sehr oft durch. Mit steigender Zahl der Versuche nähert sich der Anteil der Versuche, die das Ereignis A hervorgebracht haben, der „wahren " Wahrscheinlichkeit P(A), formal ausgedrückt

$$P(A) = \lim_{n \to \infty} \frac{n(A)}{n} ,$$

mit n(A) als Zahl der Versuche, die das Ereignis A hervorgebracht haben. Man bezeichnet diesen Zusammenhang als **Gesetz der großen Zahlen**. Er liefert die Begründung, dass man unbekannte Wahrscheinlichkeiten mit Hilfe von empirischen Beobachtungen schätzen kann, wobei hier gilt: Viel hilft viel!

Bei manchen Fragestellungen versagen die beiden obigen Wahrscheinlichkeitskonzepte. Z.B. bei Ereignissen, die sehr selten auftreten, für die man also auch keine Versuchsreihen zur Verfügung hat, etwa die Wahrscheinlichkeit für den Erfolg eines neu auf dem Markt platzierten Produkts. Es möchte beispielsweise ein Unternehmen ein neues Spülmittel auf den Markt bringen. Es steht vor der Alternative, Fernsehwerbung einzusetzen oder nicht. Es ist mit den Ereignissen konfrontiert: Wenn Fernsehwerbung eingesetzt wird, ist das Spülmittel ein Erfolg/kein Erfolg. Wenn keine Fernsehwerbung eingesetzt wird, ist das Spülmittel ein Erfolg/kein Erfolg. Für diese vier Ereignisse sollen Wahrscheinlichkeiten ermittelt werden. Da man keine verlässlichen Informationen darüber hat, wird man aus dem Bauch heraus, eventuell unter Berücksichtigung ähnlicher Erfahrungen bestimmte Wahrscheinlichkeiten zuordnen. Dieses Vorgehen entspricht der **Subjektiven Wahrscheinlichkeitsauffassung**.

Da Ereignisse als Mengen definiert sind, kann man auch in vielen Fällen Ereignisse und ihre Wahrscheinlichkeiten in **Venn-Diagrammen** veranschaulichen. Die Wahrscheinlichkeit ist dann die Fläche der entsprechenden Menge. Manchmal ist es hilfreich, das Venn-Diagramm maßstabsgetreu auf kariertes Papier abzutragen, indem die Mengen rechteckig dargestellt werden.

Pizzeria-Beispiel zur Berechnung von Wahrscheinlichkeiten

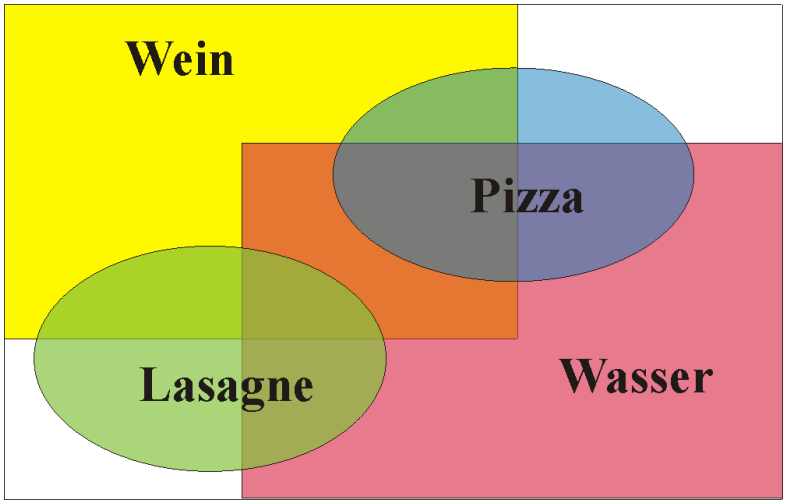

Abbildung 3: Aufteilung der Gäste nach Bestellung

Jetzt schauen wir uns in der Pizzeria etwas genauer um: Der Inhaber Carlo Pommodore ist ein mitleidiger Mensch und duldet auch arme Gäste, die sich nichts bestellen. Deshalb ist das Lokal mit seinen 50 Gästen eigentlich schon überfüllt. 20 Personen haben sich Pizza bestellt und 10 Lasagne. Das Essen ist so reichlich, dass niemand zwei Mahlzeiten bestellt. 40 Gäste trinken Wein und 20 Gäste trinken Mineralwasser, aber 15 trinken Wasser und Wein.

Wir ziehen zufällig einen Gast aus der fröhlich lärmenden Menge. Wie groß ist die Wahrscheinlichkeit, einen Pizza-Esser zu erhalten?

Wir haben $|\Omega| = 50$ verschiedene Ergebnisse. Man kann davon ausgehen, dass jeder Gast die gleiche Wahrscheinlichkeit hat, gezogen zu werden.

Wir definieren nun die Ereignisse:

A: Der Gast isst Pizza; B: Der Gast isst Lasagne;
C: Der Gast trinkt Wein; D: Der Gast trinkt Wasser.

Nach dem Symmetrieprinzip ist

$$P(A) = \frac{|A|}{|\Omega|} = \frac{20}{50} = \frac{2}{5}\,,$$

$$P(B) = \frac{10}{50} = \frac{1}{5} ,$$

$$P(C) = \frac{4}{5}$$

und $P(D) = \frac{2}{5}$.

Wir können berechnen:

Wahrscheinlichkeit, dass jemand Wasser **und** Wein trinkt:

$$P(C \cap D) = \frac{|C \cap D|}{|\Omega|} = \frac{15}{50} = \frac{3}{10} .$$

Wahrscheinlichkeit, dass ein zufällig ausgewählter Gast kein Wasser trinkt ($_{\bar{D}}$):

$$P(\bar{D}) = \frac{|\bar{D}|}{|\Omega|} = \frac{50-20}{50} = 1 - \frac{20}{50} = \frac{3}{5} = 1 - P(D) .$$

Anteil der Leute, die Wasser **oder** Wein trinken:

$P(C \cup D) = P(C) + P(D) - P(C \cap D) = \frac{40}{50} + \frac{20}{50} - \frac{15}{50} = \frac{45}{50} = \frac{9}{10}$.

Diese Beziehung gilt immer für zwei Ereignisse!

Wahrscheinlichkeit, dass ein Gast Pizza oder Lasagne isst:

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{20}{50} + \frac{10}{50} - 0 = \frac{30}{50}$.

Die Mengen A und B sind disjunkt.

Wahrscheinlichkeit, dass der zufällig ausgewählte Gast kein Wasser oder keinen Wein trinkt:

$P(\bar{C} \cup \bar{D}) = P(\bar{C}) + P(\bar{D}) - P(\bar{C} \cap \bar{D})$.

Hier ist die direkte Berechnung der Wahrscheinlichkeit analog zu oben umständlich. Man verwendet am besten die

DE MORGANsche Regel:

$P(\bar{C} \cup \bar{D}) = P(\overline{C \cap D}) = 1 - P(C \cap D) = 1 - \frac{15}{50} = \frac{35}{50} = 0,7$.

Was gelernt werden muss

Ein Ereignis A ($A \subset \Omega$) :

$$0 \leq P(A) \leq 1.$$

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

$$P() = 0.$$

Zwei Ereignisse A und B (A,B $\subset \Omega$) :

A und B sind im allgemeinen nicht disjunkt, also ist die Wahrscheinlichkeit, dass A **oder** B eintritt, nach dem **Additionssatz für zwei Ereignisse:**

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

Falls A und B **disjunkt** sind, ist

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

DE MORGANsche Regeln:

$$P(\bar{A} \cup \bar{B}) = P(\overline{A \cap B})$$

und

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$$

Für drei Ereignisse A_i (i=1, 2, 3) aus Ω gilt analog zu obigen Überlegungen:

$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$.

Mehrere Ereignisse A_i (i endlich oder unendlich):

Sind die Ereignisse A_i sämtlich paarweise disjunkt, ist bei ihrer Vereinigung

$P(A_1 \cup A_2 \cup A_3 ...) = P(A_1) + P(A_2) + P(A_3) + ...$

Übung

Zeigen Sie anhand eines Venn-Diagramms die Gültigkeit einer der DeMorganschen Regeln.

Stochastische Unabhängigkeit

Ein häufiges Untersuchungsobjekt in der Statistik ist, ob verschiedene Ereignisse **abhängig** oder **unabhängig** voneinander sind, d.h. ob das Zustandekommen eines Ereignisses durch ein anderes begünstigt wird. So untersucht man beispielsweise in der Marktforschung, ob Status und Bildung eines Konsumenten die Ausgaben für eine bestimmte Zeitschrift beeinflussen.

Beispiel zum Begriff der stochastischen Unabhängigkeit

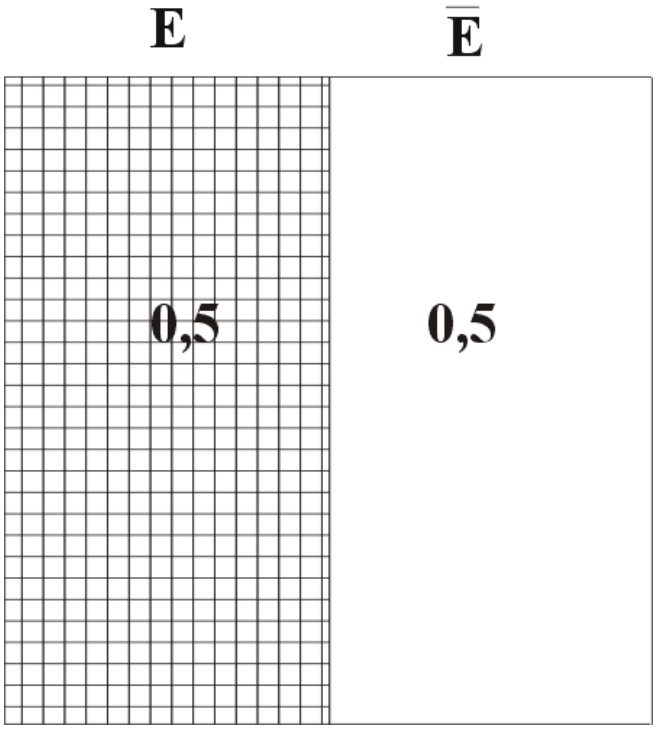

Abbildung 4: Grafik 1: Die Ereignisse: Studentin wohnt bei den Eltern - Die Studentin wohnt woanders

Eine umfangreiche Marketingstudie über Zahnputzgewohnheiten von Konsumenten hat ergeben, dass 50 % der Studierenden einer kleinen Hochschule bei ihren Eltern wohnen. Ebenso, dass 50 % der Studierenden Zahnpasta mit roten Streifen und 50 % andersfarbige Zahnpasta bevorzugen .

Betrachten wir den **Zufallsvorgang**: Eine Studentin kommt in einen Laden und kauft Zahnpasta. Es seien definiert die Ereignisse:

E: Die Studentin wohnt bei ihren Eltern.
R: Die Studentin kauft Zahnpasta mit roten Streifen.

Frage: Hat der Wohnort der Studentin einen Einfluss auf die Farbpräferenz?

Vermutlich nein, die Ereignisse E und R sind **stochastisch unabhängig**, d.h. in wahrscheinlichkeitstheoretischer Hinsicht unabhängig.

Wir interessieren uns zunächst für den Wohnort der Studierenden. In der Grafik 1 ist die Ergebnismenge nach dem Wohnort aufgeteilt.

Frage: Wieviel Prozent der Studierenden, die bei ihren Eltern wohnen, werden voraussichtlich Zahnpasta mit roten Streifen kaufen?

Da sich bei Unabhängigkeit der Ereignisse die Studierenden in Bezug auf ihre Farbpräferenz gleichmäßig auf die Wohnorte verteilen, werden wohl 50 % der Rotkäufer bei ihren Eltern wohnen und 50 % woanders. D.h. 50 % von 50 % der Studierenden wohnen bei ihren Eltern **und** bevorzugen rote Zahnpasta. Es gilt also:

$$P(R \cap E) = 0{,}5 \cdot 0{,}5 = 0{,}25.$$

Die Grafik 2 zeigt, wie sich bei Unabhängigkeit der Variablen Wohnort und Farbpräferenz die Wahrscheinlichkeiten der Farbpräferenz auf die Wohnorte aufteilen.

Ist nun beispielsweise P(E) = 40 % und P(R) = 60 %, ergibt sich bei Unabhängigkeit die Aufteilung wie in der Grafik 3, denn auch hier müssten 60 % der „Nesthocker" und 60 % der „Nestflüchter" gleichermaßen Zahnpasta mit roten Streifen kaufen.

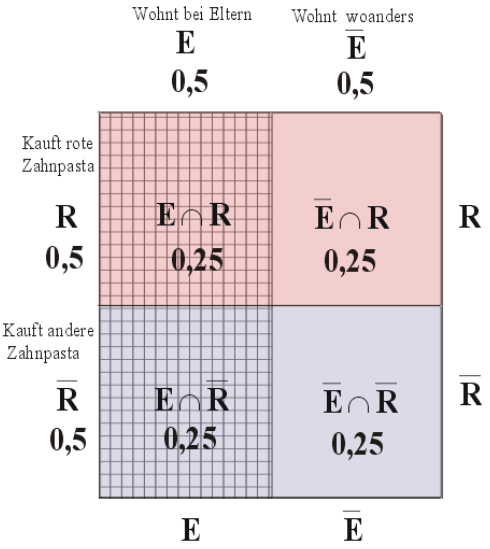

Abbildung 5: Grafik 2: Die Ereignisse Wohnort und Farbe der Zahnpasta durchmischen sich

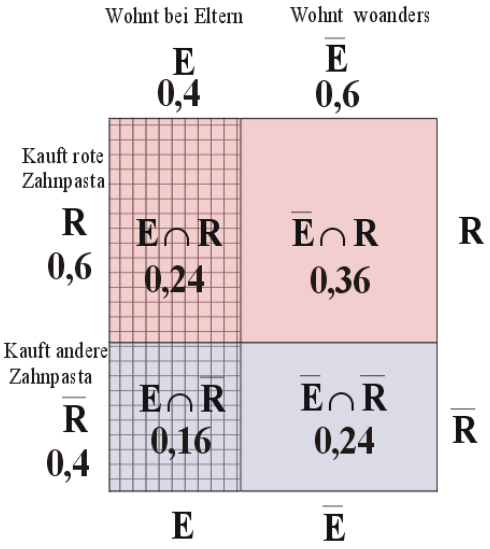

Abbildung 6: Grafik 3: Die Ereignisse: Studentin wohnt bei den Eltern - Die Studentin wohnt woanders

Beispiel zum Begriff der stochastischen Abhängigkeit

Oben haben wir den Fall betrachtet, dass zwei Ereignisse unabhängig sind. Im Allgemeinen muss man aber davon ausgehen, dass Ereignisse, die man gemeinsam analysiert, abhängig sind.

Im Rahmen der Marketingstudie wurden Daten eines Gesundheitsamtes in Musterstadt verwendet, die die Zahngesundheit von Schulkindern betraf. Man weiß aus dieser Studie, dass 50 % der Schulkinder Karies haben und 50 % der Schulkinder sich regelmäßig die Zähne putzen.

Wir betrachten den Zufallsvorgang: Es wird ein Schulkind zufällig ausgewählt.

Wir definieren als Ereignisse

Z: Das Schulkind putzt sich regelmäßig die Zähne.
K: Das Schulkind hat Karies.

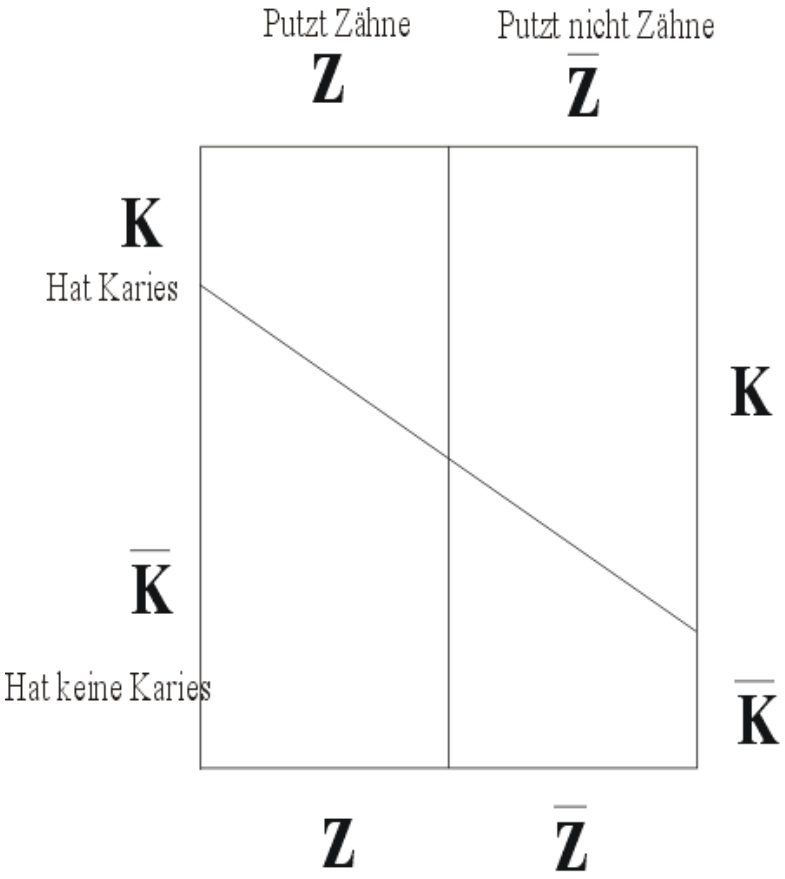

Abbildung 7: Grafik 4: Aufteilung von Zähneputzern und Kariesfällen

Ist nun

$$P(Z \cap K) > P(Z \cap \bar{K})$$

oder $P(Z \cap K) < P(Z \cap \bar{K})$?

Ist also die Wahrscheinlichkeit, ein Kind zu erhalten, das sich regelmäßig die Zähne putzt **und** Karies hat, größer als die Wahrscheinlichkeit, ein Kind zu erhalten, das sich regelmäßig die Zähne putzt **und** keine Karies hat, oder ist es umgekehrt, oder sind vielleicht die Wahrscheinlichkeiten gleich?

Es ist vermutlich

$$P(Z \cap K) < P(Z \cap \bar{K}),$$

denn Zähneputzen und Karies sind bekanntlich nicht unabhängig voneinander zu betrachten. Also sind Z und K stochastisch abhängige Ereignisse. Wir werden vermutlich eine Aufteilung der gemeinsamen Wahrscheinlichkeiten erhalten, die ähnlich der Grafik 4 ist. Besonders groß sind P(Z ∩ K) und P(Z ∩ K).

Die gemeinsamen Wahrscheinlichkeiten können allerdings nicht mit unseren Informationen bestimmt werden, sie hängen von der Stärke der Abhängigkeit ab.

Bei **stochastisch abhängigen Ereignissen** interessiert man sich häufig für das bedingte Auftreten eines Ereignisses, z.B. für die **bedingte Wahrscheinlichkeit**

$$P(K|\bar{Z}),$$

dass ein zufällig ausgewähltes Schulkind Karies hat, wenn man weiß, dass es sich nicht regelmäßig die Zähne putzt.

Bedingte Wahrscheinlichkeiten

Beispiel

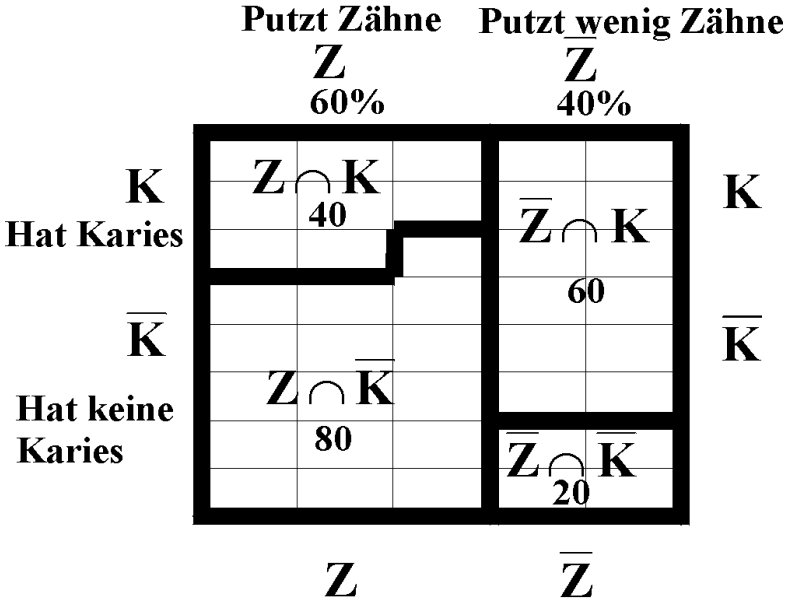

Abbildung 8: Wie hängen Kariesfälle und Zahnputzgewohnheit zusammen?

Einige Jahre später wurde in der Grundschule von Musterdorf zu Forschungszwecken wieder an 200 Kindern eine Reihenuntersuchung zur Zahngesundheit durchgeführt. Jetzt putzten sich 60 % der Kinder regelmäßig die Zähne. Von diesen Kindern hatten 40 Karies. Bei den Zahnputzmuffeln hatten 60 Kinder Karies.

Wir wollen ein maßstabsgetreues Venndiagramm konstruieren. Jedes Kästchen steht für 5 Kinder. Es sind

$$P(Z) = 0{,}6; \qquad P(\bar{Z}) = 0{,}4;$$

$$P(Z \cap K) = 0{,}2; \qquad P(Z \cap \bar{K}) = 0{,}4;$$

$$P(\bar{Z} \cap K) = 0{,}3; \qquad P(\bar{Z} \cap \bar{K}) = 0{,}1.$$

Wir interessieren uns nun für die bedingte Wahrscheinlichkeit, dass ein Kind Karies hat, wenn bekannt ist, dass es sich die Zähne putzt:

$$P(K|Z).$$

In andere Worte gekleidet: **Der Anteil der Kinder mit Karies an den Kindern, die sich regelmäßig die Zähne putzen**.

Es gilt für die bedingte Wahrscheinlichkeit

$$P(K|Z) = \frac{P(K \cap Z)}{P(Z)}.$$

Wie ist diese Wahrscheinlichkeit zu verstehen?

Es werden zunächst alle Kinder, die sich regelmäßig die Zähne putzen, in die Aula geschickt. Aus diesen 120 Kindern wird nun zufällig eins ausgewählt. Mit welcher Wahrscheinlichkeit hat dieses Kind Karies? Wir betrachten also 120 zahnputzende Kinder, davon haben 40 Kinder Karies.

Genau diese Vorgehensweise ist das Prinzip der bedingten Wahrscheinlichkeiten!

Es ergibt sich: $P(K|Z) = \frac{40}{120} = \frac{1}{3}$.

Ein Drittel der zähneputzenden Kinder hat Karies: Dann haben natürlich zwei Drittel der zähneputzenden Kinder keine Karies. Wir sehen sogleich, dass die obige Rechnung die schon bekannte Formel

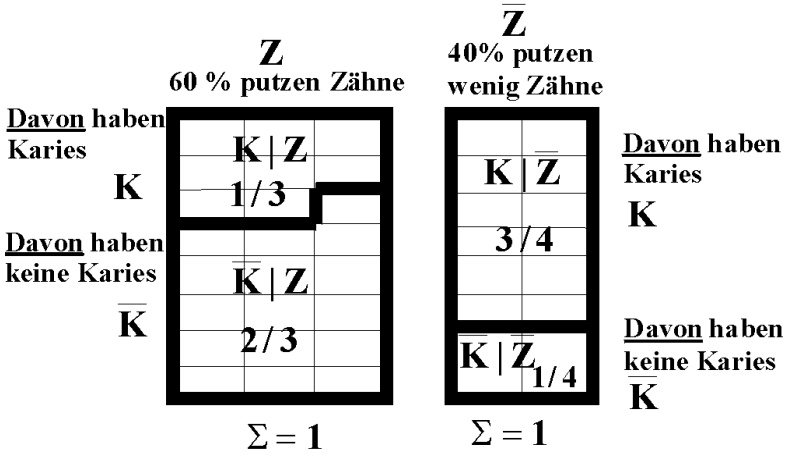

Abbildung 9: Wie teilen sich die Kariesfälle bezüglich der Zahnputzgewohnheiten auf?

$$P(K|Z) = \frac{P(K \cap Z)}{P(Z)} = \frac{\frac{40}{200}}{\frac{120}{200}} = \frac{40}{120} = \frac{1}{3},$$

darstellt. Entsprechend erhalten wir

$$P(\bar{K}|Z) = \frac{P(\bar{K} \cap Z)}{P(Z)} = \frac{\frac{80}{200}}{\frac{120}{200}} = \frac{2}{3},$$

$$P(K|\bar{Z}) = \frac{P(K \cap \bar{Z})}{P(\bar{Z})} = \frac{\frac{60}{200}}{\frac{80}{200}} = \frac{3}{4},$$

$$P(\bar{K}|\bar{Z}) = \frac{P(\bar{K} \cap \bar{Z})}{P(\bar{Z})} = \frac{\frac{20}{200}}{\frac{80}{200}} = \frac{1}{4}.$$

Vergleichen Sie das Venndiagramm mit dem vorhergehenden! Wieso unterscheiden sich beide Diagramme?

Übung

Es ist bekannt, dass die Aktienkurse des Unternehmens Dachs an 55% aller Börsentage gestiegen sind.

Ereignisse: K_1: Der Kurs steigt am ersten Tag K_2: Der Kurs steigt am zweiten Tag

Man hat folgende Gesetzmäßigkeit der Kursentwicklung festgestellt: In 40 % aller Beobachtungen stieg der Kurs am ersten Tag und am zweiten Tag, in 15 % der Beobachtungen stieg der Kurs am ersten Tag und fiel am zweiten Tag. Dagegen fiel in 15 % der Beobachtungen der Kurs am ersten Tag und stieg am zweiten Tag. An den restlichen Tagespaaren fiel der Kurs an beiden Tagen.

1. Stellen Sie die gemeinsamen Wahrscheinlichkeiten im Venndiagramm grafisch dar.
2. Sind die Ereignisse K_1 und K_2 stochastisch unabhängig? (Begründen Sie die Antwort formal mit Hilfe der Wahrscheinlichkeitstheorie.)
3. Am heutigen Tag ist der Kurs gestiegen.
   - Mit welcher Wahrscheinlichkeit wird er morgen steigen (Gesucht: $P(K_2|K_1)$)?
   - Mit welcher Wahrscheinlichkeit wird er dagegen fallen?

   Mit welcher Wahrscheinlichkeit wird der Kurs morgen steigen, wenn er heute gefallen ist?

Bayessches Theorem

Häufig liegen die Informationen über zwei Ereignisse nur als bedingte Wahrscheinlichkeiten vor. Wie kann man sie weiter verwenden?

Beispiel für zwei Ereignisse

Ein bekannter Vergnügungspark verbraucht täglich große Mengen an Glühbirnen für die Dekoration der Stände. Damit die Verbrauchskosten nicht so hoch werden, setzen sich die Glühbirnen nur zu 60% aus Markenware und zu 40 % aus markenfreier Ware zusammen. Aufgrund langjähriger Beobachtungen weiß man, dass von den Marken-Glühbirnen pro Monat 5% defekt werden. Jedoch werden von den markenfreien Glühbirnen monatlich 10% defekt.

Zunächst wollen wir das Gegebene grafisch (Grafik 5) darstellen: Wenn von den Markenglühbirnen 5 % defekt werden, bleiben 95% heil. 5% ist also **Anteil** der defekten Glühbirnen an den Markenglühbirnen, d.h. es handelt sich um die bedingte Wahrscheinlichkeit P(D|M) usw.

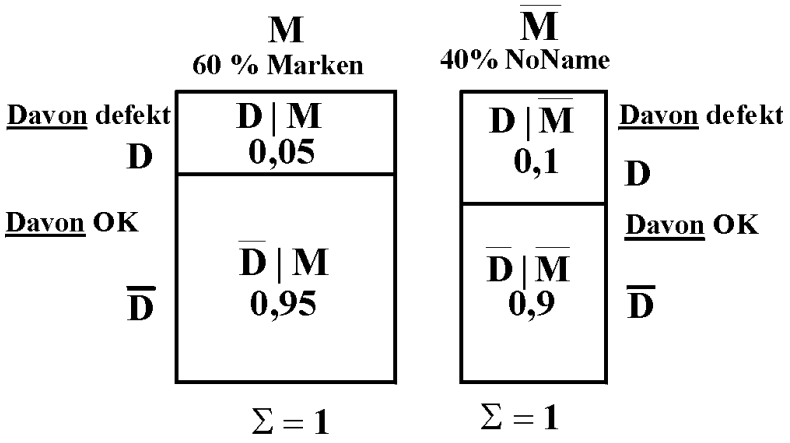

Abbildung 10: Grafik 5

Der Betreiber des Vergnügungsparks braucht für die Kostenplanung des nächsten Sommers die Information, wie groß der Anteil der Markenglühbirnen an den defekten Glühbirnen ist, d.h. er sucht P(M|D). Das bedeutet: **Alle defekten Glühbirnen eines Tages werden in einem Korb gesammelt. Es wird eine Glühbirne zufällig entnommen**. Mit welcher Wahrscheinlichkeit erhält man eine Markenbirne?

Wir wissen, dass gilt:

$$P(M|D) = \frac{P(M \cap D)}{P(D)}$$

.

Leider sind aber die Komponenten des Bruchs unbekannt. Wir werden nun eine Methode finden, sie doch zu berechnen.

Zunächst suchen wir den Zähler P(M ∩ D): Wir kennen P(D|M). Bekanntlicherweise berechnet es sich als

$$P(D|M) = \frac{P(M \cap D)}{P(M)}$$

.

Also ist der gesuchte Zähler auch in P(D|M) enthalten und kann ganz einfach durch Auflösung der Gleichung berechnet werden als

$$P(M \cap D) = P(D|M)P(M)$$

.

also

$$P(M \cap D) = 0{,}05 \cdot 0{,}6 = 0{,}03$$

.

Jetzt fehlt noch der Nenner P(D). Betrachten wir das Venndiagramm Grafik 6. D setzt sich aus den Schnittmengen D ∩ M und D ∩ $_{\bar{M}}$ zusammen.

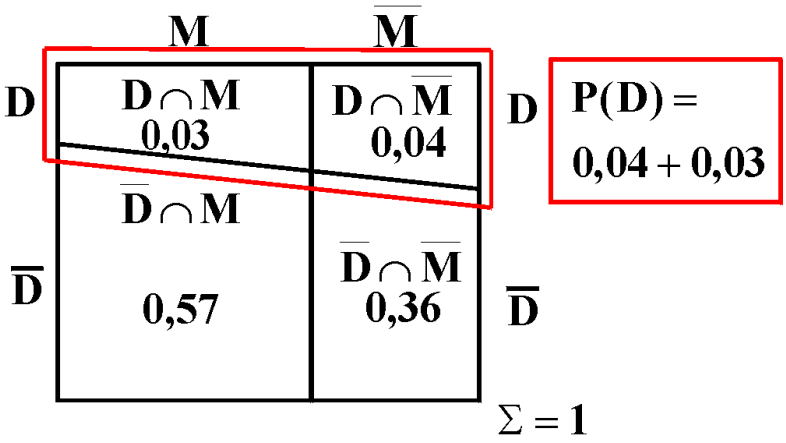

Abbildung 11: Grafik 6

Die gesamte Wahrscheinlichkeit von D ist also die Summe

$$P(D) = P(M \cap D) + P(\bar{M} \cap D)$$

.

eine Erkenntnis, die man auch als Satz der totalen Wahrscheinlichkeit bezeichnet, und das gibt, wie wir oben gesehen haben,

$$P(D) = P(D|M)P(M) + P(D|\bar{M})P(\bar{M})$$

,

in unserem Beispiel

$$P(D) = 0{,}05 \cdot 0{,}6 + 0{,}1 \cdot 0{,}4 = 0{,}07$$

.

Es sind also 7% aller Glühbirnen defekt.

Die gesuchte bedingte Wahrscheinlichkeit ist nun

$$P(M|D) = \frac{P(M \cap D)}{P(D)} = \frac{P(D|M)P(M)}{P(D|M)P(M) + P(D|\bar{M})P(\bar{M})}$$

,

Diese Formel wird als Bayessches Theorem bezeichnet.

Die gesuchte Wahrscheinlichkeit beträgt

$$P(M|D) = \frac{0{,}03}{0{,}07} = 0{,}4286$$

.

Diese Wahrscheinlichkeit fällt deshalb so überraschend hoch aus, weil 50% mehr Markenbirnen als markenfreie verwendet werden. Entsprechend ist der Anteil der markenfreien Glühbirnen an den defekten 0,5714.

Wir wollen nun mehr als zwei Ereignisse analysieren.

Beispiel für mehr als zwei Ereignisse

Eine Spedition beschäftigt drei LKW-Fahrer, die Herren Ahorn, Behorn und Zehorn. Ahorn fährt 50% aller Fuhren, Behorn 20% und Zehorn 30%. Aus Erfahrung weiß man, dass Ahorn bei 10% aller Fahrten eine Beule verursacht, Behorn bei 15% aller Fahrten und Zehorn bei 20% aller Fahrten (Grafik 7).

Wir definieren die Ereignisse:

F_1: Ahorn ist gefahren, F_2: Behorn ..., F_3: Zehorn ...
B: Eine Beule wurde gefahren.

Wir wollen zuerst das Gegebene festhalten: Wenn Ahorn in 10 % aller Fahrten eine Beule fährt, wickelt er die restlichen 90 % ohne Schaden ab usw.

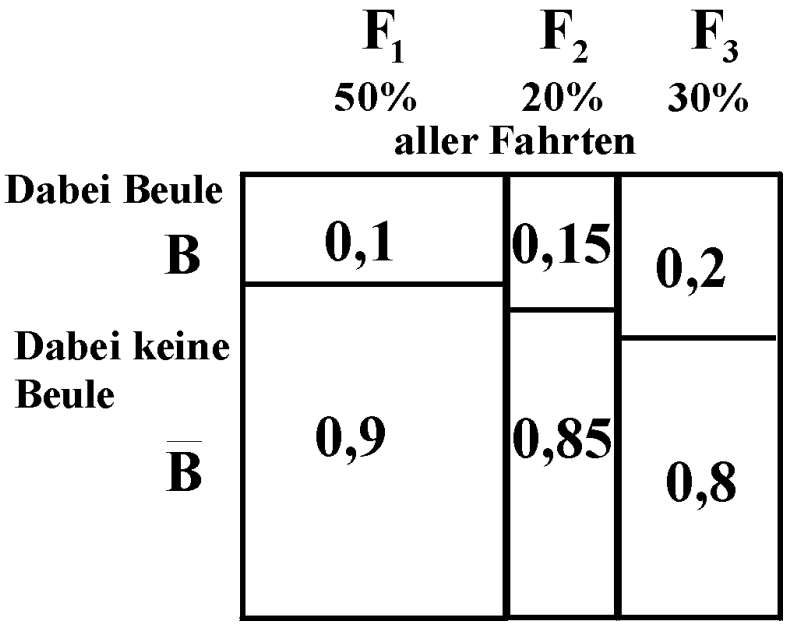

Abbildung 12: Grafik 7

Man interessiert sich für die Wahrscheinlichkeit, dass Ahorn gefahren ist, wenn wieder ein Mal eine Beule in einem LKW auftaucht, d.h. für $P(F_1|B)$.

Es ist wieder

$$P(F_1|B) = \frac{P(F_1 \cap B)}{P(B)}$$

.

Nach dem Multiplikationssatz der Wahrscheinlichkeiten muss

$$P(F_1 \cap B) = P(B|F_1)P(F_1)$$

sein, also

$$P(F_1 \cap B) = 0{,}1 \cdot 0{,}5 = 0{,}05$$

.

Aber wie erhalten wir P(B)? Auch hier gilt wieder der Satz von der totalen Wahrscheinlichkeit, z.B.:

$$P(F_1 \cap B) = P(B|F_1) \cdot P(F_1)$$

.

Wir erhalten dann für P(B)

$$P(B) = P(F_1 \cap B) + P(F_2 \cap B) + P(F_3 \cap B)$$

$$= P(B|F_1)P(F_1) + P(B|F_2)P(F_2) + P(B|F_3)P(F_3) \ ,$$

also

$$P(B) = 0{,}1 \cdot 0{,}5 + 0{,}15 \cdot 0{,}2 + 0{,}2 \cdot 0{,}3 = 0{,}05 + 0{,}03 + 0{,}06 = 0{,}14$$

.

Unsere gesuchte Wahrscheinlichkeit beträgt

$$P(F_1|B) = \frac{P(F_1 \cap B)}{P(B)} = \frac{0{,}05}{0{,}14} = 0{,}3571$$

.

Entsprechend sind

$$P(F_2|B) = \frac{0{,}03}{0{,}14} = 0{,}2143$$

und

$$P(F_3|B) = \frac{0{,}06}{0{,}14} = 0{,}4286$$

.

Also hat Zehorn mit größter Wahrscheinlichkeit die Beule gefahren.

Wir fassen nun das Gelernte dieser Seite zusammen:

Theoretische Erkenntnisse

Zwei Ereignisse A und B aus Ω:

Sind zwei Ereignisse A und B **stochastisch unabhängig**, ist ihre gemeinsame Wahrscheinlichkeit gleich dem Produkt der Einzelwahrscheinlichkeiten:

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

Man beachte: Ereignisse sind grundsätzlich **nicht** als unabhängig zu betrachten!

Die bedingten Wahrscheinlichkeiten für A und B sind

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

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

Allgemeiner Multiplikationssatz der Wahrscheinlichkeiten:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

.

Theorem von BAYES:

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

.

Verallgemeinerung für m Ereignisse A_i (i =1,...,m):

Diese m Ereignisse **zerlegen** die Ergebnismenge, d.h. sie sind disjunkt und füllen Ω aus. Enthält Ω noch ein Ereignis B, so schneidet B mindestens ein Ereignis A_i, und B ist dann

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_m \cap B)$$

.

Es gilt hier das **Bayessche Theorem**:

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

.

Übung:

Was ist P(A|B), falls A und B disjunkt sind?
Was ist P(A|B), falls A und B stochastisch unabhängig sind?

Lösungen der Übungen

Beispiel mit den Kursverläufen

1. Darstellung der verschiedenen Wahrscheinlichkeiten

	K_1 0,55	$\overline{K_1}$ 0,45
K_2 0,55 aus Summe der Zeile	$K_1 \cap K_2$ 0,4	$\overline{K}_1 \cap K_2$ 0,15
$\overline{K}_2$ 0,45 aus Summe der Zeile	$K_1 \cap \overline{K}_2$ 0,15	$\overline{K}_1 \cap \overline{K}_2$ 0,3

2. Bei stochastischer Unabhängigkeit müsste die gemeinsame Wahrscheinlichkeit gleich dem Produkt der Einzelwahrscheinlichkeiten sein.

$$P(K_1 \cap K_2) = 0{,}4$$

,

aber

$$P(K_1) \cdot P(K_2) = 0{,}55 \cdot 0{,}55 \neq 0{,}4$$

.

Also sind die Ereignisse stochastisch abhängig.

3. Es ist

$$P(K_2|K_1) = \frac{K_1 \cap K_2}{K_1} = \frac{0{,}4}{0{,}55}$$

und

$$P(\overline{K}_2|K_1) = \frac{K_1 \cap \overline{K}_2}{K_1} = \frac{0{,}15}{0{,}55}$$

4.

$$P(K_2|\overline{K}_1) = \frac{\overline{K}_1 \cap K_2}{\overline{K}_1} = \frac{0{,}15}{0{,}45}$$

Übungen zu Theoretische Erkenntnisse

Lösung: 0; P(A).

Kombinierte Zufallsvorgänge (insbesondere **wiederholte oder mehrfache Versuche**).

Allgemeines

Beispiele für kombinierte Zufallsvorgänge:

- Eine Münze werfen, dann einmal würfeln.
- Aus einer Urne ohne Zurücklegen 3 Kugeln ziehen.
- Aus einer Lostrommel 10 Gewinner ziehen.
- Gewinnspiel: Aus drei Toren eines wählen. Falls richtiges Tor, Wahl zwischen zwei Umschlägen.
- 5x auf ein Ziel schießen.

Beispiel für die formale Definition

Es sollen nacheinander drei Zufallsexperimente durchgeführt werden. Die Wahrscheinlichkeit, dass beim ersten Versuch das Ereignis A, beim zweiten Versuch das Ereignis B und beim dritten Versuch das Ereignis C resultiert, wird bezeichnet als $P(A^{(1)} \wedge B^{(2)} \wedge C^{(3)})$. A, B und C können verschiedenen Ergebnismengen entstammen! Der hochgestellte Index kann unter Umständen weggelassen werden.

Beispiel für unabhängige Versuche

Wir betrachten den Zufallsvorgang: Wir werfen zuerst eine Münze und würfeln dann.

Die beiden Versuche haben jeweils die Ergebnismenge

$$\Omega_M = \{\text{Wappen (W); Zahl (Z)}\} \text{ bzw. } \Omega_W = \{1,2,3,4,5,6\}$$

Es ergibt sich für diesen kombinierten Versuch die Ergebnismenge Ω^* als kartesisches Produkt von Ω_M und Ω_W :

$$\Omega^* = \{(W; 1), (W; 2), (W; 3), \ldots, (W; 6), (Z; 1), (Z; 2), \ldots, (Z; 6)\}.$$

Ω^* hat 12 Elemente. Jedes Element hat die selbe Wahrscheinlichkeit, gezogen zu werden.

Wir suchen nun die Wahrscheinlichkeit für das Ereignis A*: Es wird erst Wappen geworfen und dann mindestens Fünf (F) gewürfelt:

Das Ereignis A* = $W^{(1)} \wedge F^{(2)}$ belegt in Ω* 2 Elemente. Wir erhalten dann für die Wahrscheinlichkeit nach dem Symmetrieprinzip

$$P(A^*) = P(W^{(1)} \wedge F^{(2)}) = \frac{2}{12} = \frac{1}{6}$$

Würfeln und Münzwurf sind jedoch stochastisch unabhängig und die Wahrscheinlichkeit muss nicht umständlich über die Ergebnismenge ermittelt werden. Also ist dann

$$P(A^*) = P(W^{(1)}) \cdot P(F^{(2)}) = \frac{1}{2} \cdot \frac{2}{6} = \frac{1}{6}$$

Übung

Sie würfeln 3 mal. Mit welcher Wahrscheinlichkeit erhalten Sie zuerst zwei mal Sechs und dann höchstens Zwei?

Lösung: $\frac{1}{108}$.

Wiederholte Versuche können aber oft stochastisch abhängig sein.

Aus einer Urne mit 2 roten und 1 schwarzen Kugeln sollen zwei Kugeln ohne Zurücklegen gezogen werden.

Das zweite Ergebnis ist vom ersten natürlich nicht mehr unabhängig, weil sich je nach erster gezogener Kugel der Inhalt der Urne ändert. Es sei: R: eine rote Kugel wird gezogen und S: eine schwarze Kugel wird gezogen.

Wir wollen zuerst die Ergebnismenge der abhängigen Versuche analysieren. Nummerieren wir die beiden roten Kugeln in R_1 und R_2. Man kann dann bei zwei mal ziehen folgende Ergebnisse erhalten:

$$\Omega^* = \{(R_1; R_2), (R_1; S), (R_2; R_1), (R_2; S), (S; R_1), (S; R_2)\}$$

Ω* hat insgesamt 6 Ergebnisse.

Wir definieren das Ereignis A: Zuerst wird eine rote (R), dann eine schwarze Kugel (S) gezogen, also A = $R^{(1)} \wedge S^{(2)}$.

Es gibt in Ω* zwei Ergebnisse, die A betreffen, also ist die Wahrscheinlichkeit

$$P(A) = \frac{2}{6} = \frac{1}{3} .$$

Dieses Beispiel war einfach. Aber kann jetzt bei abhängigen Versuchen auch die Wahrscheinlichkeit für das kombinierte Ereignis unter Verzicht auf die vollständige Darstellung der Ergebnismenge bestimmt werden?

Bei stochastisch abhängigen Versuchen können die Wahrscheinlichkeiten nicht mehr ohne weiteres als Produkt der Einzelwahrscheinlichkeiten der Ereignisse bestimmt werden. Man kann aber sukzessiv den Multiplikationssatz der Ereignisse anwenden, der von den bedingten Wahrscheinlichkeiten bekannt ist: P(A∩B) = P(A)·P(B|A). Die Wahrscheinlichkeit, dass beim ersten Mal A und beim zweiten Mal B resultiert, ist also

$$P(A^{(1)} \wedge B^{(2)}) = P(A^{(1)}) \cdot P(B^{(2)}|A^{(1)})$$

Es ist nach der obigen Formel

$P(A)$	$= P(R^{(1)})$	$\cdot P(S^{(2)}\|R^{(1)})$	
$P(R^{(1)} \cap S^{(2)}) =$			
$=$	$\frac{2}{3}$	$\cdot\frac{1}{2}$	$= \frac{1}{3}$
	Beim ersten Versuch sind 3 Kugeln in der Urne; zwei sind rot	Beim zweiten Versuch sind noch 2 Kugeln in der Urne; eine ist schwarz.	

Diese Regel läßt sich auch auf mehr als zwei Ereignisse erweitern:

Beispiel

Aus einer Urne mit 10 roten (R) und 5 schwarzen (S) Kugeln sollen ohne Zurücklegen nacheinander drei rote Kugeln gezogen werden. Die Wahrscheinlichkeit dafür ist

$$P(R^{(1)} \cap R^{(2)} \cap R^{(3)}) = \frac{10}{15} \cdot \frac{9}{14} \cdot \frac{8}{13}$$

Für mehr als zwei Ereignisse kann der allgemeine **Multiplikationssatz der Wahrscheinlichkeiten** angewendet werden. Er gilt auch für Ereignisse, die nicht aus einer gemeinsamen Ergebnismenge stammen:

$$\begin{aligned} & P(A^{(1)} \wedge A^{(2)} \wedge \cdots \wedge A^{(m)}) \\ = \; & P(A^{(1)}) \cdot P(A^{(2)}|A^{(1)}) \cdot P(A^{(3)}|A^{(1)} \wedge A^{(2)}) \cdot ... \\ & \cdot P(A^{(m)}|A^{(1)} \wedge A^{(2)} \wedge ... \wedge A^{(m-1)}). \end{aligned}$$

Falls die A(i) (i = 1, 2, ... ,m) stochastisch unabhängig sind, ist natürlich wieder

$$P(A^{(1)} \wedge A^{(2)} \wedge \cdots \wedge A^{(m)}) = P(A^{(1)}) \cdot P(A^{(2)}) \cdot \cdots \cdot P(A^{(m)})$$

.

Je nachdem, wie die Problemstellung ist, gibt es für die Berechnung von Wahrscheinlichkeiten kombinierter Zufallsvorgänge also verschiedene Möglichkeiten:

1. Wir bestimmen alle Elemente von Ω^*, falls das möglich und durchführbar ist. Dann wenden wir das Symmetrieprinzip an.
2. Wir überlegen uns, beispielweise mit Hilfe der Kombinatorik, die Zahl der Elemente in Ω^* und wenden dann das Symmetrieprinzip an.
3. Wir verwenden den allgemeinen Multiplikationssatz der Wahrscheinlichkeiten und können vielleicht sogar stochastische Unabhängigkeiten ausnützen.

Urnenmodelle

Bei wiederholten Versuchen greift man häufig auf das so genannte Urnenmodell zurück: Dieses Modell funktioniert im Prinzip folgendermaßen: Eine Urne enthält N viele Kugeln, die sich voneinander unterscheiden lassen. Es werden n viele Kugeln gezogen. Man interessiert sich für die Zahl von Kugeln mit einem bestimmten Merkmal unter den n gezogenen.

Wir unterscheiden grundsätzlich

- das **Urnenmodell mit Zurücklegen**: Eine Kugel wird gezogen und wieder zurückgelegt
- das **Urnenmodell ohne Zurücklegen**: Eine Kugel wird gezogen und nicht wieder zurückgelegt

Viele Zufallsvorgänge, speziell die wiederholter Versuche, können auf das Urnenmodell zurückgeführt werden. Den Anfänger mag die Vorstellung, eine Kugel zu ziehen und wieder zurückzulegen, eigenartig anmuten, aber so kann man unabhängige Versuche modellieren: Betrachten wir den Zufallsvorgang, zwei mal zu würfeln, so kann man stattdessen auch aus einer Urne mit 6 verschiedenen Kugeln zwei mal jeweils eine ziehen und wieder zurücklegen.

Kombinatorik

Wir haben eine Urne mit N Kugeln gegeben. Es sollen n Kugeln gezogen werden. Wir befassen uns nun mit der Zahl der möglichen Ergebnisse bei wiederholten Versuchen. Hier müssen wir die verschiedenen Arten der Anordnung gezogener Kugeln im Urnenmodell berücksichtigen.

Zur Verdeutlichung dieser Aufgabenstellung betrachten wir eine Urne mit 3 Kugeln A, B, C. Es sollen n = 2 Kugeln gezogen werden. Wie viel verschiedene Paare würden wir erhalten?

Wir unterscheiden die Aufgabenstellungen

Mit Wiederholung - Mit Berücksichtigung der Reihenfolge

Die Buchstaben werden mit Zurücklegen gezogen; ein Buchstabe kann also mehrmals im Paar auftauchen. Es kommt auf die Reihenfolge der Buchstaben an. Es sind folgende verschiedene Paare möglich:

(A,A), (A,B), (A,C), (B,A), (B,B), (B,C), (C,A), (C,B), (C,C).

Es gibt insgesamt N^n viele verschiedene Ergebnisse, wie man leicht sieht.

Mit Wiederholung - Ohne Berücksichtigung der Reihenfolge

Es sind folgende verschiedene Paare möglich:

(A,A), (A,B), (A,C), (B,B), (B,C), (C,C).

Es gibt insgesamt $\binom{N+n-1}{n}$ viele verschiedene Ergebnisse.

Ohne Wiederholung - Mit Berücksichtigung der Reihenfolge

Die Buchstaben werden ohne Zurücklegen gezogen; ein Buchstabe kann nur einmal im Paar auftauchen. Es sind folgende verschiedene Paare möglich:

(A,B), (A,C), (B,A), (B,C), (C,A), (C,B).

Es gibt insgesamt $\frac{N!}{(N-n)!}$ viele verschiedene Ergebnisse.

Ohne Wiederholung - Ohne Berücksichtigung der Reihenfolge

Es sind folgende verschiedene Paare möglich:

(A,B), (A,C), (B,C).

Es gibt insgesamt $\binom{N}{n}$ viele verschiedene Ergebnisse.

Übungsbeispiel

Aus vier Personen Anna (A), Balduin (B), Cäcilie (C), Dagobert (D) werden zwei zum Geschirrspülen ausgelost, wobei eine Person abspült und eine abtrocknet.

Handelt es sich um ein Modell mit oder ohne Zurücklegen? Theoretisch wäre auch ein Modell mit Zurücklegen denkbar. Da das aber als unfair empfunden wird, gehen wir vom Modell ohne Zurücklegen (M. o. Z.) aus.

- Mit welcher Wahrscheinlichkeit erwischt es zuerst Cäcilie und dann Balduin (Ereignis E)?

Hier kommt es auf die Reihenfolge der gezogenen „Kugeln“ an.

Methode a: Direkt über die Ergebnismenge

Die Ergebnismenge ergibt $\Omega^* =$

-	(A,B)	(A,C)	(A,D)
(B,A)	-	(B,C)	(B,D)
(C,A)	(C,B)	-	(C,D)
(D,A)	(D,B)	(D,C)	-

Jedes Paar hat die gleiche Wahrscheinlichkeit, gewählt zu werden. Es gibt insgesamt $|\Omega^*| = 12$ verschiedene Paare.

$$P(E) = P((C, B)) = \frac{1}{12}$$

Methode b: Über die Zahl der Ergebnisse

Es handelt sich um ein Modell ohne Zurücklegen mit Beachtung der Reihenfolge. Es gibt

$$\frac{N!}{(N-n)!} = \frac{4!}{(4-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{2} = 12$$

verschiedene Paare. Es gibt nur ein Ergebnis für das Ereignis E. Es ist also

$$P(E) = \frac{|E|}{|\Omega^*|} = \frac{1}{12}$$

Methode c: Über den Multiplikationssatz der Wahrscheinlichkeiten

$$P(C^{(1)} \cap B^{(2)}) = \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12}$$

- Mit welcher Wahrscheinlichkeit müssen die zwei Männer abwaschen (Ereignis F)?

Methode a:

Es ist F = {(B,D), (D,B)}. Dieses Ereignis belegt in Ω^* zwei Elemente. Also ist

$$P(F) = \frac{2}{12} = \frac{1}{6}$$

Methode b:

M.o.Z, ohne Beachtung der Reihenfolge. Es gibt

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot}{(1 \cdot 2)(1 \cdot 2)} = 6$$

verschiedene Paare . Es ist also $P(F) = \frac{1}{6}$

Methode c:

$P(F) = \frac{2}{4} \cdot \frac{1}{3} = \frac{1}{6}$.

Kapitel 3

Zufallsvariablen

Beispiel zum Begriff der Zufallsvariablen

Die fränkische Druckerei Printzig nennt 10 multifunktionelle Hochleistungsdrucker ihr eigen. Drei Drucker sind von der Firma Alpha, zwei sind von Beta, vier von Gamma und einer stammt von der Firma Delta. Da die Drucker auch von Kunden bedient werden, fallen sie aufgrund unsachgemäßer Handhabung häufig aus. Man hat festgestellt, dass alle Drucker in gleichem Maße anfällig sind. Wegen der Gewährleistung wird bei jedem Ausfall ein Wartungstechniker der betreffenden Firma geholt. Die Kosten für die Wiederherstellung eines Druckers hängen vom Hersteller ab, wobei die Drucker der Firma Gamma in der Reparatur am billigsten sind.

Am liebsten ist es natürlich Herrn Printzig, wenn ein Drucker mit den geringsten Reparaturkosten ausfällt.

Überlegen wir:

Welche Ergebnismenge gehört zu dem Zufallsvorgang: Ein Drucker fällt zufällig aus?

Mit welcher Wahrscheinlichkeit entstehen Herrn Prinzig die geringsten Kosten?

Wir erhalten die Ergebnismenge

$$\Omega = \{A_1, A_2, A_3, B_1, B_2, G_1, G_2, G_3, G_4, D_1\},$$

wobei z.B. B_2 Drucker Nr. 2 der Firma Beta bedeutet. G sei das Ereignis, die geringsten Reparaturkosten zu haben. Jeder Drucker hat die gleiche Wahrscheinlichkeit, auszufallen. Dann ist nach dem Symmetrieprinzip

$$P(G) = \frac{\text{Zahl der G-Drucker}}{\text{Zahl aller Drucker}} = \frac{|G|}{|\Omega|} = \frac{4}{10} = 0,4$$

Die Kosten für die Reparatur eines Druckers betragen je nach Hersteller wie folgt:

Hersteller	Alpha	Beta	Gamma	Delta
Kosten (Euro)	50	60	30	100

Überlegen wir: Wieviel muss Herr Printzig pro Ausfall im Durchschnitt bezahlen?

Ordnen wir nun der Ergebnismenge die entsprechenden Kosten zu:

A_1	A_2	A_3	B_1	B_2	G_1	G_2	G_3	G_4	D_1
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
50	50	50	60	60	30	30	30	30	100

Ω hat 10 Ergebnisse und jedes Elementarereignis hat die Wahrscheinlichkeit 1/10. Jeder Drucker fällt dann auch mit der Wahrscheinlichkeit 1/10 aus. Die durchschnittlichen Reparaturkosten sind also

$$\begin{aligned} &50 \cdot \frac{1}{10} + 50 \cdot \frac{1}{10} + 50 \cdot \frac{1}{10} + 60 \cdot \frac{1}{10} + 60 \cdot \frac{1}{10} + ... + 100 \cdot \frac{1}{10} \\ &= 50 \cdot \tfrac{3}{10} + 60 \cdot \tfrac{2}{10} + 30 \cdot \tfrac{4}{10} + 100 \cdot \tfrac{1}{10} \\ &= \tfrac{150}{10} + \tfrac{120}{10} + \tfrac{120}{10} + \tfrac{100}{10} = \tfrac{490}{10} = 49 \text{ Euro} \end{aligned}$$

Wir haben soeben eine Zufallsvariable konstruiert und zwar, indem wir allen Ergebnissen von Ω eine Zahl zugeordnet haben.

Den Durchschnitt konnten wir erst berechnen, nachdem wir die Drucker mit einer Zahl versehen hatten. Man kann je nach Interesse den Elementarereig-

nissen beliebige Zahlen zuordnen. So könnten für die laufende Wartung wieder ganz andere Kosten gelten. Nur die Ergebnismenge ist festgelegt. Man könnte nun die Wahrscheinlichkeit berechnen, dass bei einem Ausfall 60 Euro fällig werden: Es gibt 10 Elementarereignisse und zwei davon entsprechen 60 Euro. Also beträgt diese Wahrscheinlichkeit 2/10.

Wir bezeichnen eine Zufallsvariable mit einem großen Buchstaben. Die Werte, die eine Zufallsvariable annehmen kann, nennt man Ausprägung. Eine bestimmte Ausprägung kennzeichnen wir mit einem Kleinbuchstaben. Nennen wir unsere Zufallsvariable "Reparaturkosten" X. Wir fassen jetzt die verschiedenen Wahrscheinlichkeiten der Zufallsvariablen X in einer Wahrscheinlichkeitstabelle zusammen. Herr Printzig hat 4 mal die "Chance", 30 Euro zu bezahlen, also ist die Wahrscheinlichkeit, dass X = 30 ist, gleich 4/10, usw.

Wahrscheinlichkeitstabelle:

	x_1	x_2	x_3	x_4
Ausprägung x_i	30	50	60	100
Wahrscheinlichkeit $f(x_i)$	0,4	0,3	0,2	0,1

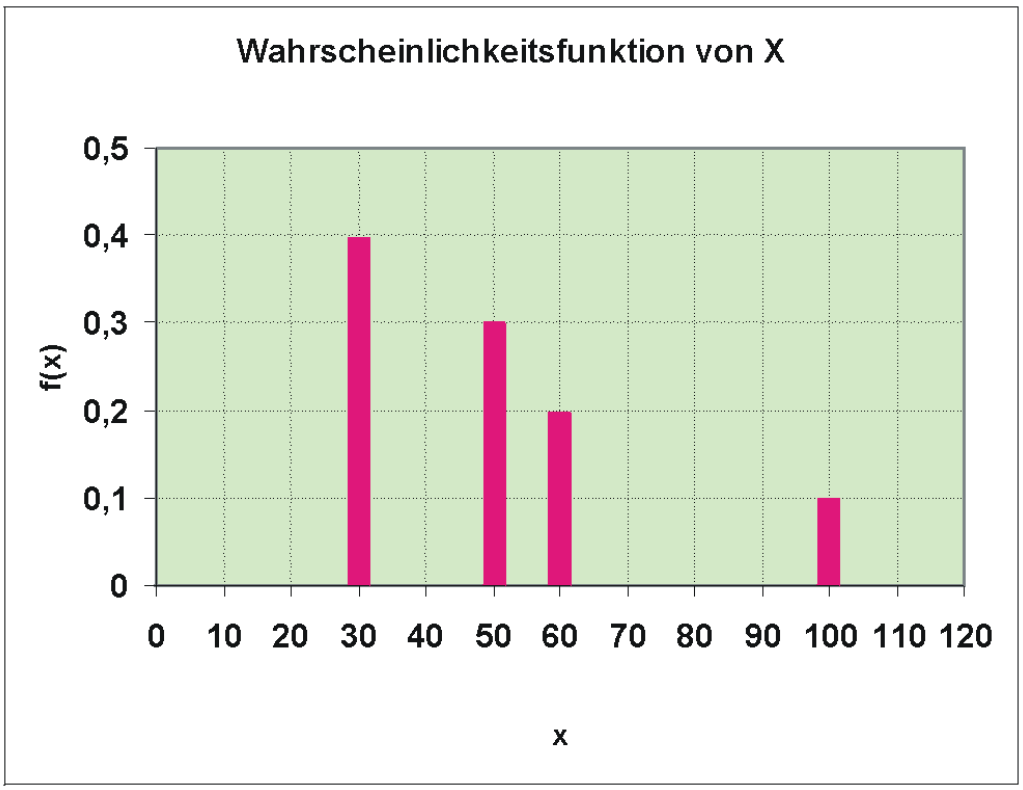

Abbildung 13: Wahrscheinlichkeitsfunktion von X: Reparaturkosten

f(x) bezeichnet die zur bestimmten Ausprägung x gehörende Wahrscheinlichkeit. Es ist beispielsweise

$$P(X = 60) = f(x_3) = f(60) = 0{,}2,$$

aber

$$P(X = 70) = f(70) = 0,$$

denn für X = 70 existiert kein Ergebnis.

Die Summe aller Wahrscheinlichkeiten ist

$$\sum_{i=1}^{m} f(x_i) = 1$$

Man kann diese Wahrscheinlichkeiten auch grafisch als Stabdiagramm darstellen.

Man sieht, dass an den x-Stellen 30, 50, 60 und 100 die Wahrscheinlichkeitsfunktion die Werte 0,4, 0,3, 0,2 und 0,1 annimmt, aber an allen sonstigen Werten von x Null ist.

Wie groß ist nun aber die Wahrscheinlichkeit, dass Herr Printzig höchstens 50 Euro bezahlen muss?

$$P(X \leq 50) = P(X = 30) + P(X = 50) = 0{,}4 + 0{,}3 = 0{,}7.$$

Das kann man auch aus der Graphik ersehen: Es ist die Summe der ||Stäbchen|| für $x \leq 50$.

Mit welcher Wahrscheinlichkeit muss Herr Printzig weniger als 100 Euro zahlen? Gefragt ist hier nach $P(X < 100)$. Ein Blick auf die Grafik verrät uns, dass gilt

$$P(X < 100) = P(X \leq 60) = P(X = 30) + P(X = 50) + P(X = 60) = 0{,}4 + 0{,}3 + 0{,}2 = 0{,}9.$$

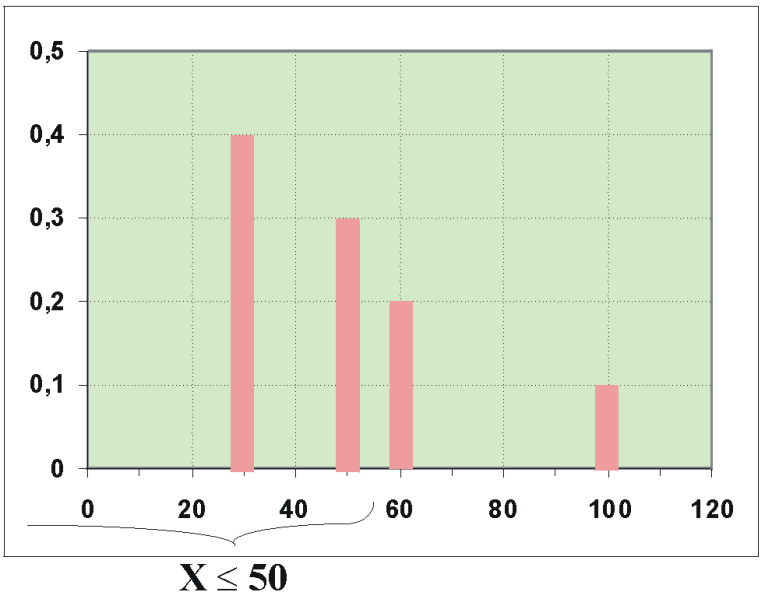

Abbildung 14

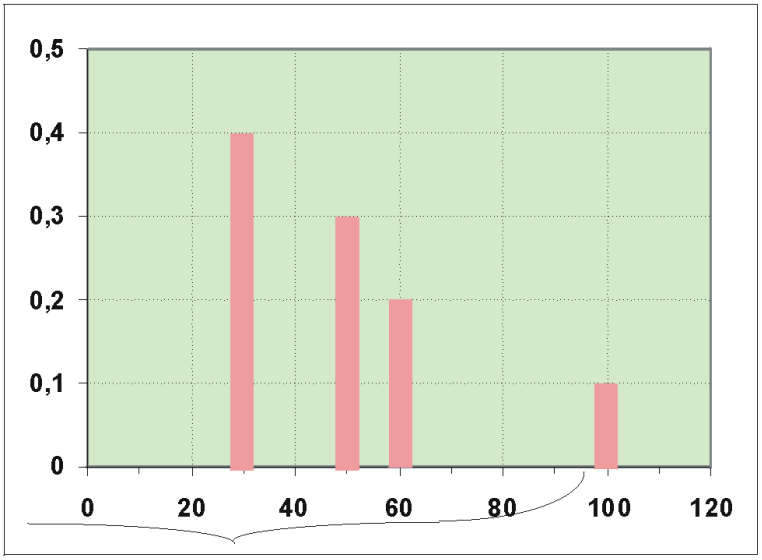

Abbildung 15

Wieviel ist nun $P(30 < X \leq 60)$?

Man kann hier wieder die "Stäbchenmethode" anwenden:

$$P(30 < X \leq 60) = 0{,}3 + 0{,}2 = 0{,}5.$$

Es gibt aber auch eine Rechenregel, die man mit Hilfe der Grafik leicht erkennt:

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

also

$$P(30 < X \leq 60) = P(X \leq 60) - P(X \leq 30) = 0{,}9 - 0{,}4 = 0{,}5.$$

Die Wahrscheinlichkeiten $P(X \leq a)$ einer bestimmten Ausprägung a von X bilden die Verteilungsfunktion von X, die die Wahrscheinlichkeitsverteilung von X in eindeutiger Weise beschreibt. Das ist eine Festlegung, die die Statistiker als sinnvoll erachten. Die Verteilungsfunktionen werden grossbuchstabig als F(a) bezeichnet. Meist wird statt a das Symbol x verwendet. Wir wollen die Verteilungsfunktion konstruieren, indem wir die obige Graphik zu Hilfe nehmen und für einzelne Stützwerte x die Verteilungsfunktion berechnen.

Wie groß ist z.B. $P(X \leq 10)$? Es ist $P(X \leq 10) = F(10) = 0$.

Ebenso sind $P(X \leq 15) = 0$ und $P(X \leq 20) = 0$.

Es ist also $F(a) = 0$ für alle Werte von a mit $-\infty < a < 30$.

Als nächstes untersuchen wir $P(X \leq 30)$:

$P(X \leq 30) = F(30) = 0{,}4$. Ebenso sind $P(X \leq 30{,}1) = 0{,}4$ und $P(X \leq 49{,}99999) = 0{,}4$.

Die Verteilungsfunktion hat also den Wert $F(a) = 0{,}4$ für $30 \leq a < 50$.

Es gilt weiter: $P(X \leq 50)$, $P(X \leq 59)$, ... $P(X< 60)$ sind, siehe Graphik: 0,4 + 0,3 = 0,7.

...

Schließlich ist die Wahrscheinlichkeit $P(X \leq 100)$ oder auch $P(X \leq 110)$, $P(X \leq 1000)$ usw... gleich 1.

Wir können die Wahrscheinlichkeiten zusammenfassen in der Verteilungsfunktion

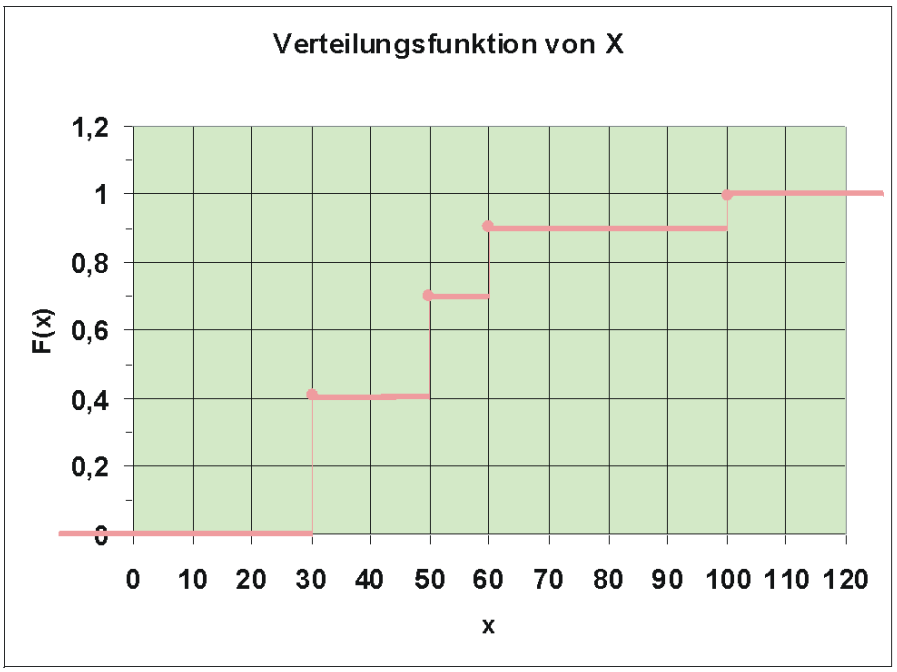

Abbildung 16: Verteilungsfunktion von X: Reparaturkosten

$$P(X \leq a) = F(a) = \begin{cases} 0 & \text{für } a < 30 \\ 0,4 & \text{für } 30 \leq a < 50 \\ 0,7 & \text{für } 50 \leq a < 60 \\ 0,9 & \text{für } 60 \leq a < 100 \\ 1 & \text{für } a \geq 100 \end{cases}$$

Man sieht, dass diese Verteilungsfunktion grafisch eine Treppenfunktion darstellt. Die Punkte links an den Stufen zeigen an, dass der Funktionswert dieser Stufe genau zum Punkt a gehört.

Man kann hier auch die Wahrscheinlichkeiten der Grafik entnehmen, z.B. ist P(X $\leq$ 70) = 0,9.

Besonders interessiert man sich bei einer Zufallsvariable für zwei Kennwerte, Parameter genannt, die die Zufallsvariable genauer beschreiben.

Einer ist der durchschnittliche Wert, den die Zufallsvariable „auf lange Sicht“ annimmt, wenn der Zufallsvorgang „sehr oft“ durchgeführt wird. Dieser Parameter wird Erwartungswert EX genannt, also der Wert, den man langfristig erwarten kann. Wir hatten ihn schon oben ermittelt als

$$EX = 50 \cdot \frac{3}{10} + 60 \cdot \frac{2}{10} + 30 \cdot \frac{4}{10} + 100 \cdot \frac{1}{10} = 49$$

die durchschnittlichen Reparaturkosten.

Ein weiterer Parameter ist die Streuung der X, ein Maß, wie stark die einzelnen Werte von X von EX abweichen, also 30-49, 50-49, 60-49, 100-49. Da z.B. 100 viel seltener auftritt als 30, gewichtet man auch diese Abweichungen mit ihrer Wahrscheinlichkeit. Eine Quadrierung sorgt dann einerseits dafür, dass sich positive und negative Abweichungen nicht aufheben, andererseits für eine überproportionale Berücksichtigung von besonders starken Abweichungen. Man erhält im Ergebnis als durchschnittliche quadratische Abweichung der X-Werte von EX die Varianz

$$\begin{aligned} \text{Var } X &= (30-49)^2 \cdot 0,4 + (50-49)^2 \cdot 0,3 \\ &+ (60-49)^2 \cdot 0,2 + (100-49)^2 \cdot 0,1 \\ &= 361 \cdot 0,4 + 1 \cdot 0,3 + 121 \cdot 0,2 + 2601 \cdot 0,1 = 429 \end{aligned}$$

wobei zu beachten ist, dass sich hier als Einheit Euro^2 ergibt.

Die Wurzel der Varianz ist die Standardabweichung; man könnte sie salopp als mittlere Abweichung der Ausprägungen vom Durchschnitt bezeichnen. Sie beträgt in unserem Beispiel etwa 20,71.

Allgemeine Darstellung einer Zufallsvariablen

Gegeben ist ein Zufallsvorgang mit der Ergebnismenge Ω. Jedem Element aus Ω wird eine reelle Zahl x zugeordnet:

$$\Omega \to \mathbb{R}, \quad \omega \mapsto X(\omega)$$

.

Die Elemente von X sind Realisationen, Ausprägungen, Werte. Die Verteilung der Zufallsvariablen kann festgelegt werden mit ihrer Verteilungsfunktion F, definiert als

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

.

Es gilt für die Verteilung jeder Zufallsvariablen:

- F(x) ist für alle $x \in \mathbb{R}$ definiert.
- $0 \leq F(x) \leq 1$.
- F(x) ist monoton steigend, also $x_1 < x_2 \to F(x_1) \leq F(x_2)$
- F(x) ist rechtsseitig stetig.
- $P(a < X \leq b) = P(X \leq b) - P(X \leq a)$.

Eine Zufallsvariable ist diskret, wenn sie in jedem beschränkten Intervall nur endlich viele Ausprägungen annehmen kann. Die diskrete Zufallsvariable kann endlich oder abzählbar unendlich viele Werte x_i (i = 1,2,..., m bzw. i = 1,2,...) annehmen.

Beispiele

- Zahl der Schadensleistungen, die in einem Jahr bei einer Versicherung auftreten
- Kinderzahl von Konsumenten
- Zahl der defekten Kondensatoren in einem Fertigungslos

Ihre Wahrscheinlichkeitsfunktion ist

$$P(X = x) = f(x) = \begin{cases} f(x_i) & \text{für } x = x_i \\ 0 & \text{sonst} \end{cases}$$

Es gilt

$$\sum_i f(x_i) = 1 \ .$$

Die Verteilungsfunktion $P(X \leq a) = F(a)$ ist die Summe aller Wahrscheinlichkeiten $f(x_i)$ für $x_i \leq a$.

Der Erwartungswert einer Zufallsvariablen ist der Durchschnitt des Auftretens ihrer Realisationen. Bei einer diskreten Zufallsvariablen beträgt er

$$EX = \sum_i x_i f(x_i) \ ,$$

falls EX existiert, d.h. nicht unendlich wird.

Die Varianz einer diskreten Zufallsvariablen berechnet sich als

$$\text{Var } X = \sum_i (x_i - EX)^2 f(x_i) \ .$$

Nach dem sog.Verschiebungssatz ist auch

$$\text{Var } X = (\sum_i x_i^2 f(x_i)) - (EX)^2 \ ,$$

im Beispiel:

$$\begin{aligned} \text{Var } X &= 30^2 \cdot 0{,}4 + 50^2 \cdot 0{,}3 + 60^2 \cdot 0{,}2 + 100^2 \cdot 0{,}1 - 49^2 \\ &= \quad 360 + 750 + 720 + 1000 - 2401 = 429 \; . \end{aligned}$$

Beispiel eines Zeitungskiosks

Dichtefunktion

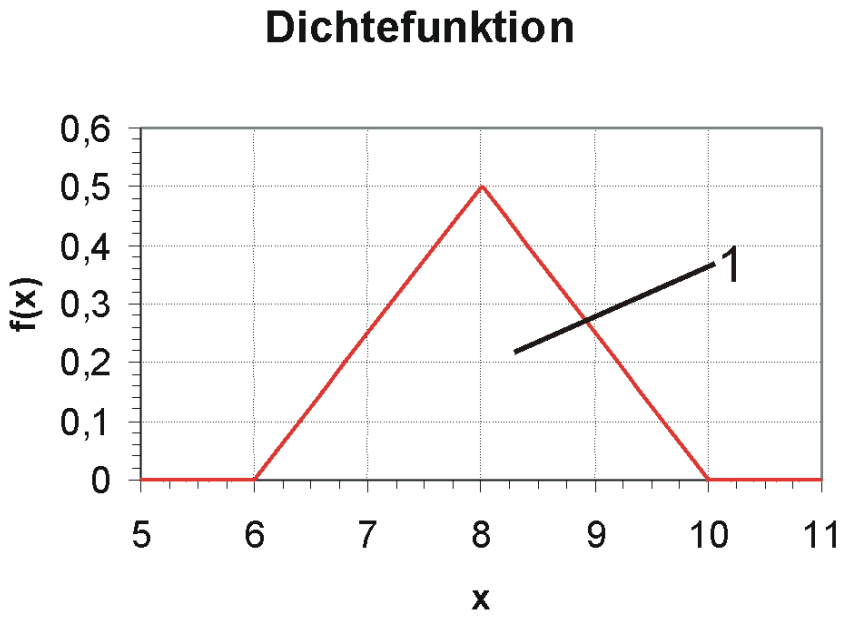

Abbildung 17: Dichtefunktion von X

Die Zufallsvariable X: ||An einem Tag verkaufte Menge an Tageszeitungen (in 100) eines Zeitungskiosks|| lässt sich beschreiben mit der (in diesem Fall frei erfundenen) **Dichtefunktion**

$$f(x) = \begin{cases} \frac{1}{4}x - \frac{3}{2} & \text{für } 6 \leq x \leq 8 \\ \frac{5}{2} - \frac{1}{4}x & \text{für } 8 < x \leq 10 \\ 0 & \text{sonst} \end{cases} \; .$$

Diese Zufallsvariable X ist nun **stetig**, d.h. sie hat in jedem Intervall a $\leq$ X $\leq$ b unendlich viele Ausprägungen.

Eine Analyse der Grafik zeigt, dass diese Dichtefunktion symmetrisch bezüglich 8 ist, was die Berechnung von Wahrscheinlichkeiten sehr erleichtert.

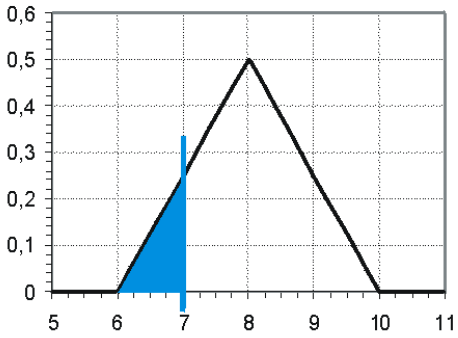

Abbildung 18: W', dass X höchstens 7 ist

Wir wollen nun die Wahrscheinlichkeit bestimmen, dass an einem Tag höchstens 700 Zeitungen verkauft werden, also $\mathbf{P(X \leq 7)}$. Wenn wir analog zu der diskreten Zufallsvariablen vorgehen, wo wir "die Summe der Stäbchen" ermittelten, müsste die Wahrscheinlichkeit P(X $\leq$ a) hier "unendlich viele Stäbchen", also eine **Fläche** ergeben.

Wir berechnen die Dreiecksfläche mit Hilfe der Geometrie:

$$P(X \leq 7) = \text{Breite des Dreiecks } \cdot \text{Höhe des Dreiecks } \cdot \frac{1}{2}$$

$$= 1 \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} \, .$$

Es ist übrigens auch

$$P(X < 7) = \frac{1}{8} \, ,$$

denn bei einer stetigen Zufallsvariablen ist $\mathbf{P(X = x) = 0}$, da es als unmöglich angesehen wird, genau einen bestimmten Wert x zu "treffen". Man

betrachtet also bei einer stetigen Zufallsvariablen nur Wahrscheinlichkeiten der Art $\mathbf{P(X \leq x)}$ o.ä.

Es ist $\mathbf{P(X \leq 8) = 0{,}5}$, wie man der Grafik sofort entnimmt.

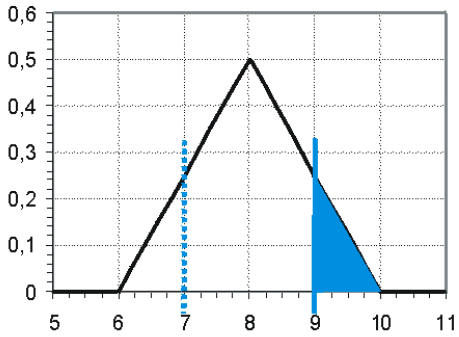

Abbildung 19: W', dass X mindestens 9 ist

$P(X \geq 9) = \frac{1}{8}$, denn wie man sieht, ist die Fläche von $\mathbf{P(X \geq 9)}$ genau gleich der Fläche $\mathbf{P(X \leq 7)}$.

Außerdem ist $P(X \leq 9) = 1 - P(X \geq 9) = \frac{7}{8}$.

Bestimmen wir die Wahrscheinlichkeit eines Intervalls. Es ergibt

$\mathbf{P(8 < X \leq 9) = P(X \leq 9)}$ **- P(X ≤ 8) = 0,875 - 0,5 = 0,375**,

wenn man die Rechenregel für $\mathbf{P(a < X \leq b)}$ anwendet.

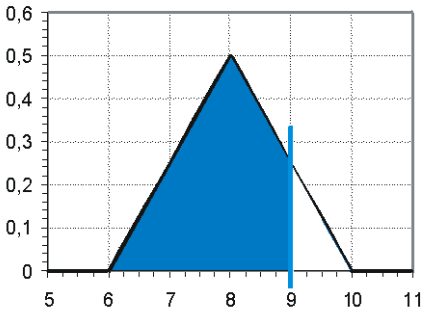

Abbildung 20: W', dass X höchstens 9 ist

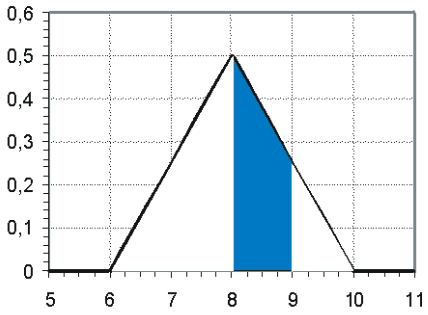

Abbildung 21: W', dass X zwischen 8 und 9 liegt

Verteilungsfunktion

Man kann Wahrscheinlichkeiten von X auch als **Verteilungsfunktion** darstellen. Sucht man die Wahrscheinlichkeit P(X $\leq$ a), muss also das Integral von -& infin; bis a berechnet werden:

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

Bei unserem Beispiel sind wir mit verschiedenen Bereichen konfrontiert:

1. a $<$ 6

$$P(X \leq a) = F(a) = \int_{-\infty}^{a} 0\,dx = 0$$

2. 6 $\leq$ a $\leq$ 8

$$\begin{array}{rcc} F(a) & = & \int_{-\infty}^{6} 0\,dx + \int_{6}^{a} \left(\frac{1}{4}x - \frac{3}{2}\right) dx \\ & = & 0 + \left[\frac{x^2}{8} - \frac{3}{2}x\right]_{6}^{a} \\ & = & \frac{a^2}{8} - \frac{3}{2}a - \left(\frac{6^2}{8} - \frac{3}{2}6\right) = \frac{a^2}{8} - \frac{3}{2}a + \frac{9}{2} \end{array}$$

3. 8 $<$ a $\leq$ 10

$$\begin{array}{rcc} F(a) & = & \int_{-\infty}^{6} 0\,dx + \int_{6}^{8} \left(\frac{1}{4}\cdot x - \frac{3}{2}\right) dx + \int_{8}^{a} \left(\frac{5}{2} - \frac{1}{4}x\right) dx \\ & = & 0 + \left[\frac{x^2}{8} - \frac{3}{2}\cdot x\right]_{6}^{8} + \left[\frac{5}{2}\cdot x - \frac{x^2}{8}\right]_{8}^{a} \\ & = & \left(\frac{64}{8} - \frac{3}{2}\cdot 8\right) - \left(\frac{36}{8} - \frac{3}{2}\cdot 6\right) + \left(\frac{5}{2}\cdot a - \frac{a^2}{8}\right) - \left(\frac{5}{2}\cdot 8 - \frac{64}{8}\right) \\ & = & -\frac{a^2}{8} + \frac{5}{2}\cdot a - \frac{23}{2} \end{array}$$

4. a $>$ 10

$$F(a) = 1$$

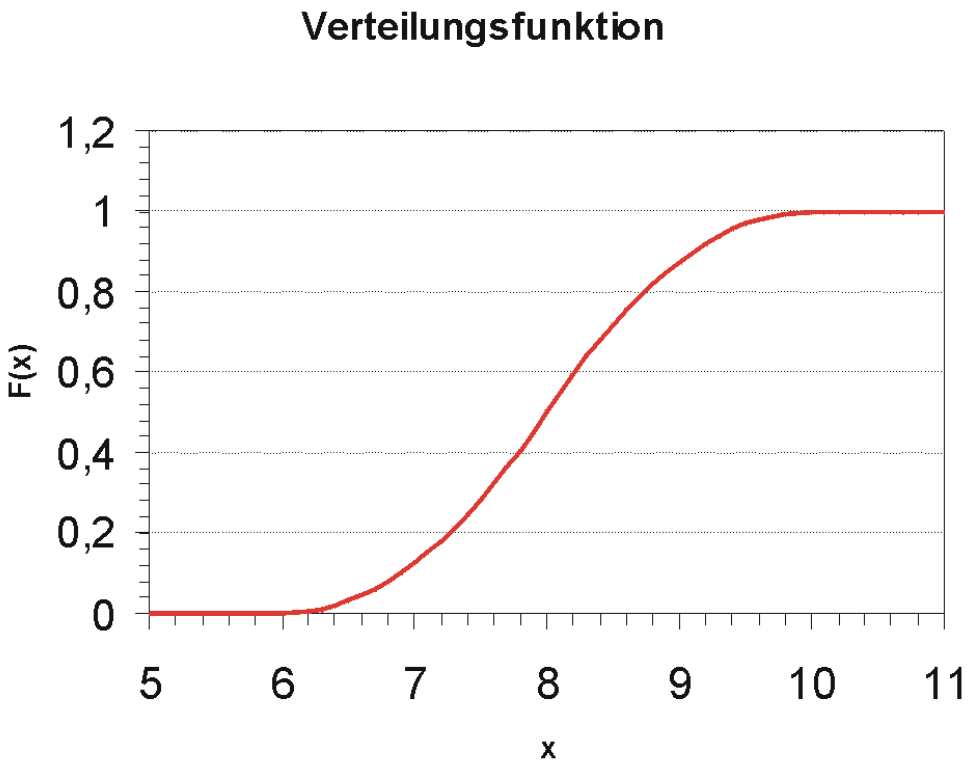

Abbildung 22: Verteilungsfunktion von X

$$P(X \leq a) = F(a) = \begin{cases} 0 & \text{für } a < 6 \\ \frac{a^2}{8} - \frac{3}{2}a + \frac{9}{2} & \text{für } 6 \leq a \leq 8 \\ -\frac{a^2}{8} + \frac{5}{2}a - \frac{23}{2} & \text{für } 8 < a \leq 10 \\ 1 & \text{sonst} \end{cases}$$

Wir erhalten beispielsweise durch Einsetzen in F(x)

$$P(X \leq 7) = F(7) = \frac{7^2}{8} - \frac{3}{2} \cdot 7 + \frac{9}{2} = \frac{1}{8} \ ,$$

$$P(X \leq 9) = F(9) = -\frac{9^2}{8} + \frac{5}{2} \cdot 9 - \frac{23}{2} = \frac{7}{8} \ .$$

Quantil

Das Quantil x(p) gibt die Ausprägung x an, die zu einem bestimmten Verteilungswert p = F(x) gehört.

Beispiele

x(0,875) = 9, d.h. zur Wahrscheinlichkeit 0,875 gehört der x-Wert 9.

Ebenso ist x(0,5) = 8. D.h. 8 ist der Median, also wurden an 50% aller Tage höchstens 800 Zeitungen verkauft.

Übung

Bestimmen Sie P(6,25 < X < 8,75). Mit welcher Wahrscheinlichkeit wurden an den 50% besten Tagen mindestens 900 Zeitungen verkauft? Gesucht ist hier P(X > 9| X > 8).

Was Sie speziell über stetige Zufallsvariablen wissen sollten

Eine stetige Zufallsvariable kann in jedem beschränkten Intervall unendlich viele Ausprägungen annehmen. Ihre Verteilung lässt sich durch eine Dichtefunktion f(x) beschreiben. f(x) ist keine Wahrscheinlichkeit, sondern eine Dichte.

- Die Verteilungsfunktion ist

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

- Es gilt: P(X = a) = 0.
- Wegen P(X = a) = 0 ist P(X ≤ a) = P(X a) = P(X ≥ a)
- Die Dichtefunktion f(x) ist die erste Ableitung der Verteilungsfunktion, falls diese an der Stelle x differenzierbar ist.
- Die Dichtefunktion f(a) kann auch größer als 1 werden.
- Ausgehend von $P(X \leq x) = p$ ist das p-Quantil x(p) der Wert x, der zu einer gegebenen Wahrscheinlichkeit p gehört. Speziell x(0,5) ist der Median.

- Der Erwartungswert einer stetigen Zufallsvariablen ist analog zu oben

$$EX = \int_{-\infty}^{\infty} x \cdot f(x)dx,$$

  falls EX existiert, d.h. nicht unendlich wird.

- Ihre Varianz ist

  $\text{Var } X = \int_{-\infty}^{\infty}(x - EX)^2 \cdot f(x)dx$

wobei auch hier der Verschiebungssatz angewendet werden kann:

$$\text{Var } X = \left(\int_{-\infty}^{\infty} x^2 f(x)dx\right) - (EX)^2$$

Bei symmetrisch verteilten Zufallsvariablen ist im Allgemeinen der Erwartungswert der Zufallsvariablen gleich dem Median.

In unserem Beispiel ist also EX = 8, denn die Verteilung ist symmetrisch. Das bedeutet, dass im Durchschnitt pro Tag 800 Zeitungen umgesetzt werden.

Wendet man die gegebene Formel für EX auf unser Beispiel an, so erhält man:

$$\begin{aligned}\text{E } X &= \int_{-\infty}^{6} x \cdot 0\,dx + \int_{6}^{8} x \cdot \left(\tfrac{x}{4} - \tfrac{3}{2}\right) dx + \int_{8}^{10} x \cdot \left(\tfrac{5}{2} - \tfrac{x}{4}\right) dx + \int_{10}^{\infty} x \cdot 0\,dx \\ &= \left[\tfrac{x^3}{12} - \tfrac{3x^2}{4}\right]_6^8 + \left[\tfrac{5x^2}{4} - \tfrac{x^3}{12}\right]_8^{10} = 8\end{aligned}$$

Entsprechend gilt für die Varianz:

$$\begin{aligned}\text{Var } X &= \int_{-\infty}^{6} x^2 \cdot 0\ dx + \int_{6}^{8} x^2 \cdot \left(\tfrac{x}{4} - \tfrac{3}{2}\right) dx + \int_{8}^{10} x^2 \cdot \left(\tfrac{5}{2} - \tfrac{x}{4}\right) dx \\ &+ \int_{10}^{\infty} x^2 \cdot 0\,dx - 64 \\ &= \left[\tfrac{x^4}{16} - \tfrac{3x^3}{6}\right]_6^8 + \left[\tfrac{5x^3}{6} - \tfrac{x^4}{16}\right]_8^{10} - 64 = \tfrac{2}{3} \approx 0{,}7\end{aligned}$$

Beispiel: Eingehende Anrufe bei Fernsehabstimmungen

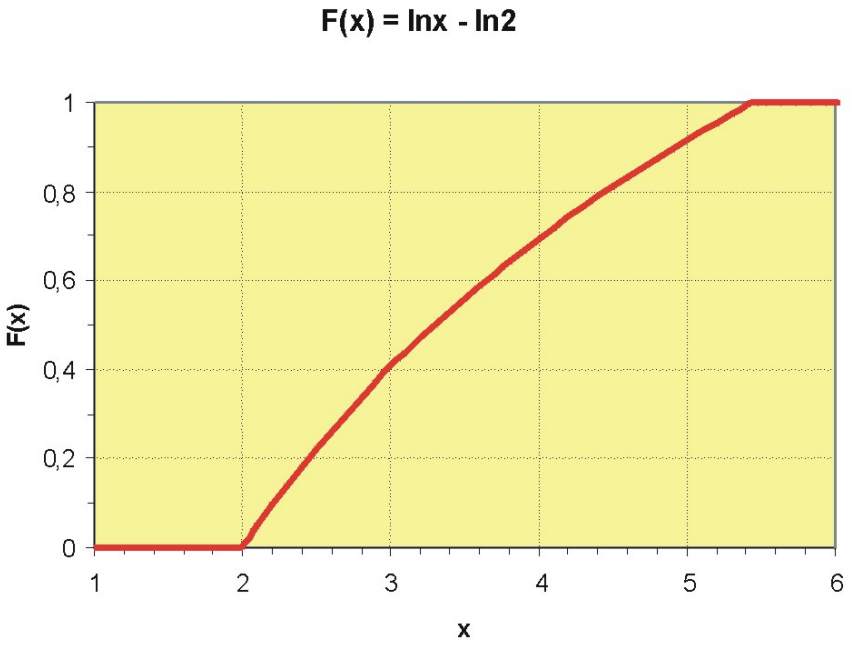

Abbildung 23: Verteilung von lnx - ln2

Während einer Fernsehsendung wurden die Zuschauer aufgefordert, telefonisch abzustimmen. Die Leitungen wurden um 14 Uhr freigeschaltet. Dann konnten die Zuschauer bis ca. 17.30 Uhr anrufen. Für die eintreffenden Anrufe ergab sich näherungsweise die Verteilungsfunktion der stetigen Zufallsvariablen X: Zeitpunkt, an dem ein Anruf eintrifft, wie folgt:

$$F(x) = \begin{cases} 0 & \text{für } x < 2 \\ \ln x - \ln 2 & \text{für } 2 \leq x \leq 2e \\ 1 & \text{für } x > 2e \end{cases}$$

Sei jetzt $\omega \in \Omega$ ein beliebiger Anruf.

Wir wollen nun bestimmen

1. die Dichtefunktion f(x)

2. die Wahrscheinlichkeit dass bis höchstens 15 Uhr der Anruf ω eingegangen ist.
3. die Wahrscheinlichkeit, dass zwischen 15 und 16 Uhr der Anruf ω eingegangen ist.
4. die Uhrzeit, zu der 90% aller Anrufe eingetroffen sind
5. den Median
6. den Erwartungswert
7. die Varianz

Die Grafik der Verteilung F(X) zeigt den typischen Verlauf einer logarithmischen Funktion.

1. Dichtefunktion f(x)

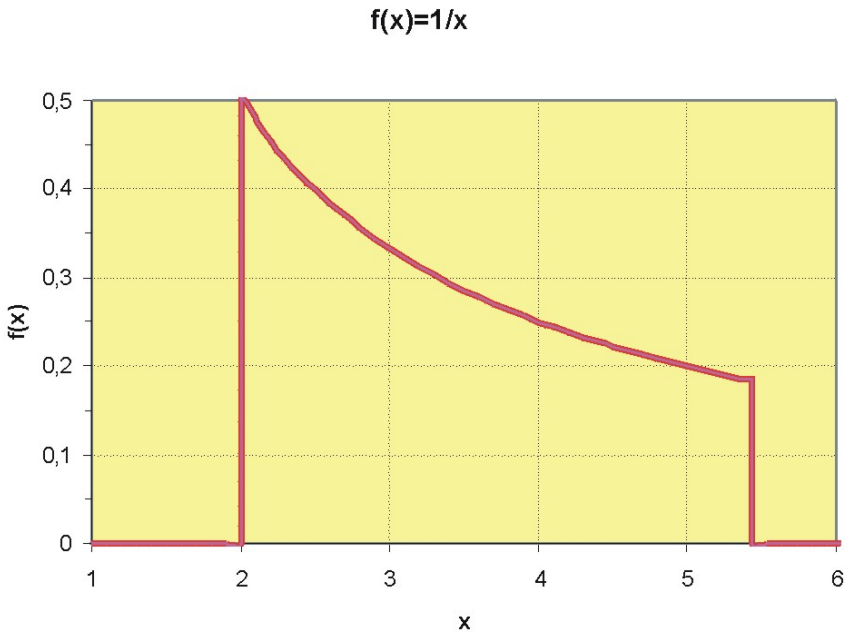

Abbildung 24: Dichtefunktion von lnx - ln2

Die Dichtefunktion ist immer die erste Ableitung der Verteilungsfunktion: f(x) = F'(x).

Unsere Verteilungsfunktion ist abschnittsweise definiert. Wir müssen bereichsweise ableiten (dass die Funktion an den Knickstellen möglicherwei-

se nicht differenzierbar ist, tut im Allgemeinen nicht weh, Hauptsache, die Fläche ergibt 1).

Bereich x < 2: $F(x) = 0 \rightarrow f(x) = 0$
Bereich $2 \leq$ x $\leq$ 2e: $F(x) = \ln x - \ln 2 \rightarrow f(x) = \frac{1}{x}$
Bereich x > 2e: $F(x) = 1 \rightarrow f(x) = 0$

Wir wollen jetzt f(x) noch ordentlich angeben:

$$f(x) = \begin{cases} \frac{1}{x} & \text{für } 2 \leq x \leq 2e \\ 0 & \text{sonst} \end{cases}$$

Betrachten wir mal die Dichtefunktion: Man sieht hier deutlich, dass die meisten Anrufe in den ersten 1,5 Stunden nach Freischalten eingelaufen sind. Danach flaut die Zahl der Anrufe allmählich ab.

"'2. Wahrscheinlichkeit, dass bis höchstens 15 Uhr der Anruf ω eingegangen ist"'

Gesucht ist P(X $\leq$ 3). In der Dichtefunktion ist das die Fläche von 2 bis 3. Diese Fläche ist das Selbe wie der Funktionswert F(3). Wir erhalten

$P(X \leq 3) = \ln 3 - \ln 2 = 1{,}0986 - 0{,}6931 = 0{,}4055$

Man kann also sagen, dass in einer Stunde ab Freischalten der Leitungen 40% der Anrufe eingegangen sind.

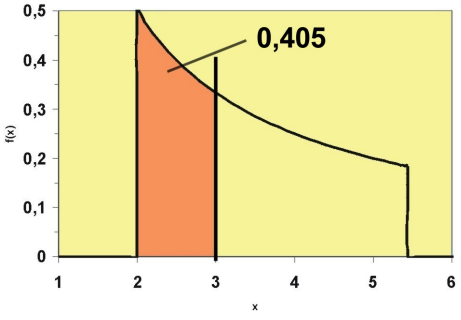

Abbildung 25: Fläche der Dichtefunktion für P(X<3)

Abbildung 26: Verteilungsfunktion für P(X<3)

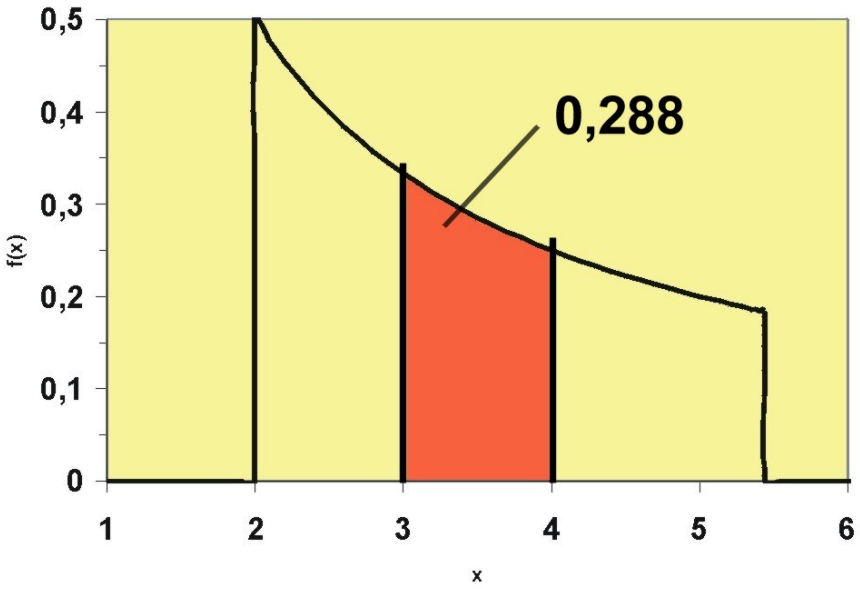

Abbildung 27: Fläche der Dichtefunktion für P(3 < X < 4)

"'3. Wahrscheinlichkeit, dass zwischen 15 und 16 Uhr der Anruf ω eingegangen ist"'

Gesucht ist hier P(3 ≤ X ≤ 4). Wir wissen schon, dass speziell für stetige Zufallsvariablen (bei diskreten muss man noch zwischen < und ≤ unterscheiden) gilt: P(3 ≤ X ≤ 4) = P(X ≤ 4) - P(X ≤ 3). Wir erhalten dann

$$\begin{aligned} P(3 \leq X \leq 4) &= F(4) - F(3) \\ &= \ln 4 - \ln 2 - (\ln 3 - \ln 2) \\ &= \ln 4 - \ln 3 \\ &= 1{,}3863 - 1{,}0986 = 0{,}2877 \end{aligned}$$

4. Uhrzeit, zu der 90% aller Anrufe eingetroffen sind

Hier ist die Wahrscheinlichkeit 0,9 gegeben und wir suchen den X-Wert, der zu dieser Wahrscheinlichkeit passt, also P(X ≤ ?) = 0,9. Gesucht ist also das 90%-Quantil. Wir bilden den Ansatz

F(?) = 0,9 oder etwas professioneller: F(x(0,9)) = 0,9, also

$$\ln x - \ln 2 = 0{,}9 \rightarrow \ln x = \ln 2 + 0{,}9 \rightarrow x = \exp(\ln 2 + 0{,}9) \approx 4{,}91$$

,

d.h. etwa um 16.55 waren 90% der Anrufe eingegangen.

5. Median

Der Median ist das 50%-Quantil. Es ergibt sich also analog zu oben:

$$\ln x - \ln 2 = 0{,}5 \rightarrow x \approx 3{,}30$$

6. Erwartungswert

Der Erwartungswert der Zufallsvariablen X wird bei einer stetigen Zufallsvariablen integriert:

$$EX = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Wir müssen hier wieder bereichsweise vorgehen und bestimmen zunächst mal die Teilintegrale:

Bereich x < 2: $\int_{-\infty}^{2} x \cdot 0\, dx = 0$
Bereich $2 \leq$ x $\leq$ 2e: $\int_{2}^{2e} x \cdot \frac{1}{x}\, dx = \int_{2}^{2e} 1\, dx = [x]_{2}^{2e} = 2e - 2 = 3{,}44.$
Bereich x $>$ 2e: $\int_{2e}^{\infty} x \cdot 0\, dx = 0$

Wir müssen nun die Teilintegrale addieren und erhalten

$$EX = 0 + 3{,}44 + 0 = 3{,}44$$

.

Es kam also ein Anruf im Durchschnitt um 15.30 an.

7. Varianz

Die Varianz berechnet sich nach der Formel

$$\text{Var } X = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x) dx \right) - (EX)^2$$

.

Analog zu oben erhalten wir

$$\left(\int_2^{2e} x^2 \cdot \tfrac{1}{x}\, dx\right) - (EX)^2 = \left(\int_2^{2e} x\, dx\right) - (2e-2)^2$$
$$= \left[\tfrac{x^2}{2}\right]_2^{2e} - 3{,}44^2 = \tfrac{(2e)^2}{2} - \tfrac{2^2}{2} - 3{,}44^2 = 0{,}9681$$

.

Mit der Ungleichung von Tschebyschew oder Biennaymé-Tschebyschew kann man Wahrscheinlichkeiten einer Zufallsvariablen mit unbekannter Verteilung abschätzen. Benötigt werden als Information der Erwartungswert und die Varianz der Zufallsvariablen, die im Allgemeinen geschätzt werden müssen.

Die Ungleichung lautet folgendermaßen:

$$P(|X - EX| \geq \epsilon) \leq \frac{\operatorname{Var} X}{\epsilon^2}$$

.

Besser kann man sich die Beziehung vorstellen, wenn man die Betragsungleichung ausschreibt :

$$P(X \leq EX - \epsilon \vee X \geq EX + \epsilon) \leq \frac{\operatorname{Var} X}{\epsilon^2}$$

Diese Abschätzung ist naturgemäß sehr grob und kann manchmal nichtssagende Ergebnisse liefern.

Beispiel

Es ist bekannt, dass ein Kaffeeautomat im Durchschnitt 250 ml Kaffee ausschenkt mit einer Varianz von 100 ml^2. Eine Tasse gilt als korrekt befüllt, wenn ihr Inhalt nicht mehr als 30 ml vom Durchschnitt abweicht. Der Anteil der inkorrekt befüllten Tassen beträgt höchstens

$$P(|X - 250| \geq 30) \leq \frac{100}{30^2} = \frac{1}{9}.$$

bzw.

$$P(X \leq EX - 30 \vee X \geq EX + 30) \leq \frac{100}{30^2} = \frac{1}{9}$$

.

Umgekehrt gilt dann auch

$$P(EX - \epsilon < X < EX + \epsilon) > 1 - \frac{\operatorname{Var} X}{\epsilon^2}$$

bzw.

$$P(|X - EX| < \epsilon) > 1 - \frac{\operatorname{Var} X}{\epsilon^2}$$

.

Also wäre der Anteil der korrekt befüllten Tassen mindestens 8/9.

Beispiel für mehrdimensionale Zufallsvariablen

y_j / x_i	0	5	10	15	$P(X = X_i)$
0	0,00	0,00	0,10	0,30	0,4
5	0,00	0,05	0,05	0,10	0,2
10	0,20	0,15	0,05	0,00	0,4
$P(Y = Y_j)$	0,2	0,2	0,2	0,4	1,0

Abbildung 28: Gemeinsame Wahrscheinlichkeit von Qualitätskontrolle X und Reklamationskosten Y

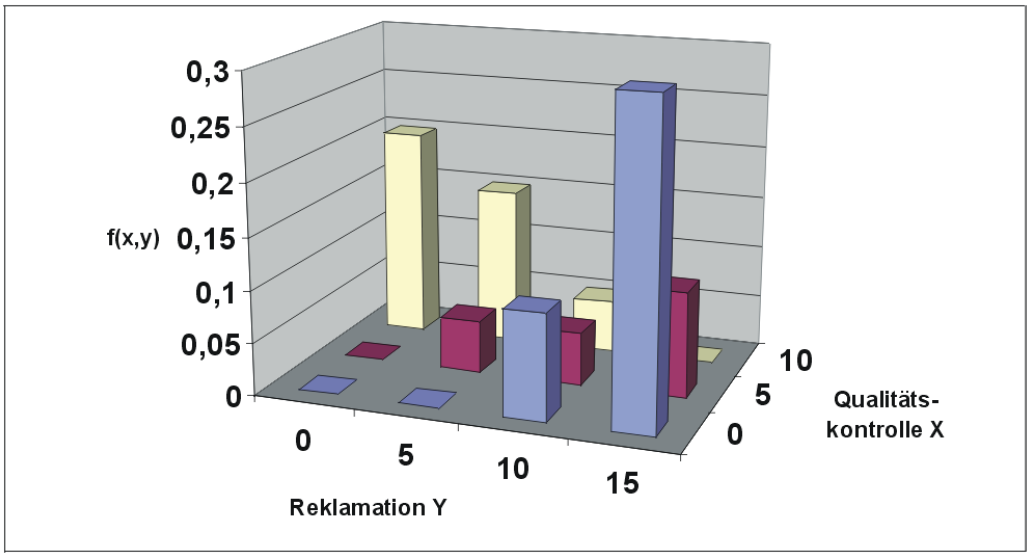

Abbildung 29: Gemeinsame Wahrscheinlichkeit von Qualitätskontrolle X und Reklamationskosten Y

In einer Studie über Total Quality Management wurde eine umfangreiche Befragung bei europäischen Produktionsbetrieben durchgeführt. Man erfasste unter anderem den Aufwand für Qualitätskontrolle während der laufenden Produktion, anteilig zu den Produktionskosten, und die Aufwendungen für Reklamationen, anteilig am Umsatz.

Wir definieren die Zufallsvariablen:

X: Anteilige Kosten der Qualitätskontrolle [%]. Y: Anteilige Kosten der Reklamationen [%].

Es ergibt sich unten die **gemeinsame Wahrscheinlichkeitstabelle** mit der iten Zeile (i = 1, ... , n) und jten Spalte (j = 1, ... , m), die auch grafisch dargestellt ist. Man sieht, wie bei steigendem Aufwand der Qualitätskontrolle die Ausgaben für die Reklamationen sinken.

Die **gemeinsame Wahrscheinlichkeit** $P(X = 5 \wedge Y = 10) = 0{,}05$ wird bezeichnet als $f_{X,Y}(5;10)$.

Die spalten- bzw. zeilenweisen Summen der gemeinsamen Wahrscheinlichkeiten ergeben die **Randwahrscheinlichkeiten** oder auch **Einzelwahrscheinlichkeiten** der Zufallsvariablen X bzw. Y.

Es ergeben sich also für diese beiden Variablen die Wahrscheinlichkeitsverteilungen

x_i	0%	5%	10%
$f_X(x_i)$	0,4	0,2	0,4

y_j	0%	5%	10%	15%
$f_Y(y_j)$	0,2	0,2	0,2	0,4

Die Einzelwahrscheinlichkeit berechnet sich beispielsweise als

$$P(X = x_1) = f_X(x_1) = \sum_{j=1}^{m} f_{X,Y}(x_1; y_j) \quad ,$$

also hier

$$P(X = 0) = f_X(0) = 0 + 0 + 0{,}1 + 0{,}3 = 0{,}4 \quad .$$

Stochastische Unabhängigkeit

Falls X und Y stochastisch unabhängig sind, ist

$$f_{X,Y}(x_i; y_j) = f_X(x_i) \cdot f_Y(y_j)$$

.

Beispiel:

Z.B. ist $P(X = 0 \wedge Y = 0) = 0$, aber $P(X = 0) \cdot P(Y = 0) = 0{,}4 \cdot 0{,}2 \neq 0$.

Also sind X und Y stochastisch abhängig. Es genügt schon, wenn die Unabhängigkeitsvoraussetzung für ein Paar nicht erfüllt ist.

Kovarianz

Man interessiert sich bei gemeinsam verteilten Variablen im allgemeinen auch dafür, inwieweit zwischen diesen Variablen ein Zusammenhang besteht. In

unserer Wahrscheinlichkeitstabelle des Beispiels der Qualitätskontrolle stehen beispielsweise links unten und rechts oben die größeren Wahrscheinlichkeiten, also scheinen niedrige Ausprägungen von X eher mit hohen Ausprägungen von Y und hohe Ausprägungen von X eher mit niedrigen Ausprägungen von Y einherzugehen.

Wahrscheinlichkeitstabelle des Beispiels von oben
Gemeinsame Wahrscheinlichkeit von Qualitätskontrolle X und Reklamationskosten Y

x y	0	5	10	15	f_x
0	0,00	0,00	0,10	0,30	0,4
5	0,00	0,05	0,05	0,10	0,2
10	0,20	0,15	0,05	0,00	0,4
f_y	0,2	0,2	0,2	0,4	1,0

Ein Maß für beispielsweise einen linearen Zusammenhang zweier Zufallsvariablen X und Y ist die Kovarianz covXY. Sie ist für diskrete Zufallsvariablen definiert als

$$covXY = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i - EX)(y_j - EY)f_{xy}(x_i; y_j)$$

bzw. wegen des Verschiebungssatzes

$$covXY = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i \cdot y_j \cdot f_{xy}(x_i; y_j) - EX \cdot EY$$

Es ergibt für unser Beispiel

$$EX = 0 \cdot 0{,}4 + 5 \cdot 0{,}2 + 10 \cdot 0{,}4 = 5$$

und

$$EY = 0 \cdot 0{,}2 + 5 \cdot 0{,}2 + 10 \cdot 0{,}2 + 15 \cdot 0{,}4 = 9$$

und damit die Kovarianz

$$
\begin{aligned}
covXY = \quad & (0-5)(0-9)\cdot 0 + (5-5)(0-9)\cdot 0 + (10-5)(0-9)\cdot 0{,}1 \\
& +(0-5)(5-9)\cdot 0 + (5-5)(5-9)\cdot 0{,}05 + (10-5)(5-9)\cdot 0{,}15 \\
& +(0-5)(10-9)\cdot 0{,}1 + (5-5)(10-9)\cdot 0{,}05 + (10-5)(10-9)\cdot 0{,}05 \\
& +(0-5)(15-9)\cdot 0{,}3 + (5-5)(15-9)\cdot 0{,}1 + (10-5)(15-9)\cdot 0 \\
= \quad & 0 + 0 + (-5)\cdot 0{,}1 + (-30)\cdot 0{,}3 + 0 + 0 + 0 + 0 \\
& +(-45)\cdot 0{,}2 + (-20)\cdot 0{,}15 + 5\cdot 0{,}05 + 0 = -21{,}25
\end{aligned}
$$

Eine positive Kovarianz deutet daraufhin, dass eher ein proportionaler Zusammenhang zwischen X und Y besteht, eine negative Kovarianz dagegen, dass eher ein umgekehrt proportionaler Zusammenhang zwischen X und Y besteht.

Korrelationskoeffizient

Ist die Kovarianz Null, sind die Zufallsvariablen unkorreliert, sonst korreliert.

Die Kovarianz ist nicht normiert. Ein normiertes Maß für den linearen Zusammenhang stellt der **Korrelationkoeffizient** nach **BRAVAIS-PEARSON** $\rho_{\mathrm{X,Y}}$ dar, der definiert ist als

$$\rho_{XY} = \frac{\mathrm{cov}XY}{\sqrt{\mathrm{var}X}\sqrt{\mathrm{var}Y}}$$

.

Es gilt für den Korrelationskoeffizienten ρ_{xy} :

$$-1 \leq \rho_{XY} \leq 1$$

.

Ist ρ_{XY} 1 oder -1, besteht ein exakter linearer Zusammenhang zwischen X und Y.

Sind X und Y stochastisch unabhängig, ist covXY und damit $\rho_{\mathrm{X,Y}}$ gleich Null. Der Umkehrschluss ist nicht zulässig, da eine nichtlineare Abhängigkeitsstruktur zwischen X und Y bestehen kann, die vom Korrelationskoeffizienten nicht erfasst werden kann.

Beispiel:

Wir berechnen zunächst die Varianz von X als

$$\mathrm{var}X = (0-5)^2 \cdot 0{,}4 + (5-5)^2 \cdot 0{,}2 + (10-5)^2 \cdot 0{,}4 = 20$$

und entsprechend die Varianz von Y als

$$\mathrm{var}Y = 34$$

.

Damit erhalten wir

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}X}\sqrt{\mathrm{var}Y}} = \frac{-21{,}25}{\sqrt{20}\sqrt{34}} = -0{,}8149$$

.

Bedingte Wahrscheinlichkeiten von Zufallsvariablen"'

Auch für Zufallsvariablen sind bedingte Wahrscheinlichkeiten angebbar, nämlich

die bedingte Wahrscheinlichkeit einer Zufallsvariablen als

$$P(X \leq x_i | X \leq x_k) = \frac{P(X \leq x_i \wedge X \leq x_k)}{P(X \leq x_k)}$$

und die bedingte Wahrscheinlichkeit zweier Zufallsvariablen

$$P(X \leq x_i | Y \leq y_j) = \frac{P(X \leq x_i \wedge Y \leq y_j)}{P(Y \leq y_j)}$$

.

Entsprechendes gilt für $\geq$ und $=$.

Ebenso gilt:

Wenn X und Y stochastisch unabhängig sind, ist

$$P(X \leq x_i \wedge Y \leq y_j) = P(X \leq x_i) \cdot P(Y \leq y_j)$$

für **alle** i,j.

Beispiele:

$$P(Y \geq 15|Y \geq 5) = \frac{P(Y \geq 15 \wedge Y \geq 5)}{P(Y \geq 5)} = \frac{P(Y \geq 15)}{P(Y \geq 5)} = \frac{0{,}4}{0{,}8} = 0{,}5$$

.

||Die Hälfte aller Unternehmen mit Reklamationskosten hatte mindestens 15% Aufwand.||

$$P(Y \geq 5|X = 10) = \frac{P(Y \geq 5 \wedge X = 10)}{P(X = 10)} = \frac{0{,}15 + 0{,}05 + 0}{0{,}4} = 0{,}5$$

.

||Die Hälfte aller Unternehmen mit sehr viel Qualitätskontrolle hatte Reklamationskosten.||

Funktion einer Zufallsvariablen

Lineare Transformation einer Zufallsvariablen

Der Student Bert hat eine kleine schicke Appartementwohnung, die er hin und wieder säubern muss. Die Intervalle der Reinigungsaktionen sind unterschiedlich und lassen sich folgendermaßen beschreiben: Die Zeit in Wochen, die nach der letzten Säuberungsaktion verstrichen ist, wird als Zufallsvariable X bezeichnet. Die Intervalle verteilen sich folgendermaßen:

Zahl der Wochen bis zur nächsten Putzaktion x_i	0	1	2	3	4	5
Wahrscheinlichkeit $f(x_i)$	0,1	0,2	0,2	0,3	0,1	0,1

X hat den Erwartungswert EX =2,4 und die Varianz 2,04. Rechnen Sie das zur Übung selber nach.

Wenn Bert putzen muss, hängt der Aufwand in Stunden von der Zahl der Wochen ab, die er seine Wohnung vernachlässigt hat. Er braucht jedesmal ca. 1 Stunde für das Bad und einmal Durchsaugen. Für die restlichen Arbeiten muss er pro verstrichener Woche noch eine halbe Stunde Arbeitszeit hinzugeben. Morgen kommen seine Eltern zu Besuch. Mit welcher Wahrscheinlichkeit muss Bert heute 2 Stunden putzen? Wie lange putzt er durchschnittlich jedes Mal?

Hier überlegen wir uns zunächst mal, dass die Putzzeit von der vorherigen ||Karenzzeit|| X abhängt. Sie ist also auch eine Zufallsvariable. Man könnte sie so darstellen:

$$Y = 1 + 0,5 \cdot X$$

Wie ist nun Y verteilt? Y hängt direkt von X ab und wir erhalten die Wahrscheinlichkeitstabelle

Zahl der Wochen bis zur nächsten Putzaktion x_i	0	1	2	3	4	5
Aufgewendete Putzzeit y_i	1	1,5	2	2,5	3	3,5
Wahrscheinlichkeit $f(y_i)$	0,1	0,2	0,2	0,3	0,1	0,1

Man kann sofort sehen, dass Bernd mit einer Wahrscheinlichkeit von 20% 2 Stunden putzen wird.

Wir wollen nun Erwartungswert und Varianz von Y ermitteln. Der Erwartungswert berechnet sich wie gewohnt als

$$\begin{aligned} EY &= \sum_i y_i \cdot f(y_i) = 1 \cdot 0,1 + 1,5 \cdot 0,2 + 2 \cdot 0,2 \\ &+ 2,5 \cdot 0,3 + 3 \cdot 0,1 + 3,5 \cdot 0,1 \\ &= 0,1 + 0,3 + 0,4 + 0,75 + 0,3 + 0,35 = 2,2 \end{aligned}$$

.

Das bedeutet er putzt durchschnittlich 2,2 Stunden.

Die Varianz ergibt sich analog als

$$\begin{aligned} varY &= \sum_i y_i^2 \cdot f(y_i) - (EY)^2 = 1^2 \cdot 0,1 + 1,5^2 \cdot 0,2 + 2^2 \cdot 0,2 \\ &+ 2,5^2 \cdot 0,3 + 3^2 \cdot 0,1 + 3,5^2 \cdot 0,1 - 2,2^2 \\ &= 0,1 + 0,45 + 0,8 + 1,875 + 0,9 + 1,225 - 2,2^2 = 0,51 \end{aligned}$$

Schön wäre es allerdings, wenn man die Parameter der Verteilung etwas einfacher ausrechnen könnte. Y hat die schöne Eigenschaft, dass es eine **lineare Transformation** von X ist der Art

$$Y = a + bX$$

.

Bei linearen Transformationen wie oben gilt

$$EY = a + b \cdot EX$$

und

$$varY = b^2 \cdot varX$$

.

Rechnen wir nach:

$$EY = 1 + 0,5 \cdot EX = 1 + 0,5 \cdot 2,4 = 1 + 1,2 = 2,2$$

und

$$varY = 0,5^2 \cdot varX = 0,25 \cdot 2,04 = 0,51$$

.

Standardisierung

Eine spezielle lineare Transformation ist die Standardisierung einer Zufallsvariablen X durch

$$Z = \frac{X - EX}{\sqrt{varX}}$$

.

Man kann nämlich Z so umformen:

$Z = \frac{X}{\sqrt{varX}} - \frac{EX}{\sqrt{varX}} = a + bX$

mit $b = \frac{1}{\sqrt{varX}}$ und $a = -\frac{EX}{\sqrt{varX}}$, denn Erwartungswert und Varianz von X sind Konstanten.

Es ist dann EZ = 0 und varZ = 1.

Nichtlineare Funktion einer Zufallsvariablen

Lakonisch könnte man sagen: Eine nichtlineare Funktion ist eine Funktion, die nicht linear ist. Man kann sie also nicht in der Form Y = a + bx schreiben. Beispiele sind etwa

$$Y = X^2, \quad Y = \sin X, \quad Y = \sqrt{X}$$

Hier kann man die Parameter im Allgemeinen nur über die Verteilung der Zufallsvariablen bestimmen.

Beispiel

Es hat sich herausgestellt, dass der Aufwand an Putzmitteln (ml pro qm) in Abhängigkeit von der verstrichenen Zeit quadratisch steigt mit der Funktion

$Y = 2 + 1 \cdot X^2$

Zahl der Wochen bis zur nächsten Putzaktion x_i	0	1	2	3	4	5
Aufgewendete Putzmittel y_i	2	3	6	11	18	27
Wahrscheinlichkeit $f(y_i)$	0,1	0,2	0,2	0,3	0,1	0,1

Hier kann man Erwartungswert und Varianz von Y nur mit den bekannten Formeln ermitteln, etwa

$$\begin{aligned} EY &= \sum_i y_i \cdot f(y_i) = 2 \cdot 0,1 + 3 \cdot 0,2 + 6 \cdot 0,2 \\ &+ 11 \cdot 0,3 + 18 \cdot 0,1 + 27 \cdot 0,1 \\ &= 0,2 + 0,6 + 1,2 + 3,3 + 1,8 + 2,7 = 9,8 \end{aligned}$$

.

Lineare Funktionen mehrerer Zufallsvariablen

Zwei Variablen

Gegeben sind zwei Zufallsvariablen X_1 und X_2 mit den Verteilungsparametern EX_1, $varX_1$ und EX_2, $varX_2$. Außerdem sind die beiden Zufallsvariablen korreliert mit der Kovarianz $covX_1X_2$. Es wird eine Zufallsvariable

$$Y = b_0 + b_1X_1 + b_2X_2$$

gebildet. Analog zu oben errechnet sich der Erwartungswert von Y durch

$$EY = b_0 + b_1EX_1 + b_2EX_2$$

.

Die Varianz von Y setzt sich wieder aus den Einzelvarianzen der Zufallsvariablen zusammen. Hinzu kommt noch die Kovarianz:

$$varY = b_1^2varX_1 + b_2^2varX_2 + 2b_1b_2 \cdot covX_1X_2$$

.

Wenn die zwei Zufallsvariablen X_1 und X_2 stochastisch unabhängig sind, ist ihre Kovarianz Null. Dann reduziert sich die Formel für die Varianz auf

$$varY = b_1^2 varX_1 + b_2^2 varX_2$$

.

Beispiel

Die Versorgung mit Getränken in einem Fußballstadion mittlerer Größe wird bei Spielen von einem Gastronomieunternehmen betrieben. Man weiß aus Erfahrung, dass die Zahl der verkauften Bierbecher von der Zahl der vorbestellten Eintrittskarten abhängt, und zwar in unterschiedlicher Weise von einheimischen und auswärtigen Besuchern. Es sei X_1: Zahl der bestellten Karten von Einheimischen und X_2: Zahl der bestellten Karten von Auswärtigen.

Es hat sich herausgestellt, dass $EX_1 = 10.000, EX_2 = 1000$ und $varX_1 = 2000, varX_2 = 300$ sind.

Zudem sind X_1 und X_2 korreliert, denn je interessanter ein Spiel, desto mehr Einheimische und Auswärtige schauen das Spiel an. Es ist $\mathrm{cov}X_1X_2 = 400$.

Die Zahl der verkauften Getränke lässt sich angeben als

$$Y = 100 + 2X_1 + 3X_2$$

.

Es ist hier

$$EY = 100 + 2 \cdot 10000 + 3 \cdot 1000 = 23100$$

und

$$varY = 2^2 \cdot 2000 + 3^2 \cdot 300 + 2 \cdot 2 \cdot 3 \cdot 400 = 15500$$

Mehr als zwei Variablen

Gegeben sind n Zufallsvariablen X_i (i = 1, ..., n) mit den Erwartungswerten EX_i, den Varianzen $varX_i$ und den paarweisen Kovarianzen $\mathrm{cov}X_1X_2$, $\mathrm{cov}X_1X_3$, ..., $\mathrm{cov}X_{n-1}X_n$. $\mathrm{cov}X_iX_j$ (i < j; i = 1, ..., n-1; j = i+1, ..., n). Es sei

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n = b_0 + \sum_{i=1}^{n} b_iX_i$$

.

Dann erhalten wir für den Erwartungswert

$$EY = b_0 + b_1EX_1 + b_2EX_2 + ... + b_nEX_n = b_0 + \sum_{i=1}^{n} b_iEX_i$$

.

Die Varianz von Y können wir als Summe der Varianzen und paarweisen Kovarianzen ermitteln als

$$varY = \sum_{i=1}^{m} b_i^2 varX_i + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} b_ib_j covX_iX_j$$

.

und, falls die Zufallsvariablen X_i stochastisch unabhängig sind, als Varianz

$$varY = \sum_{i=1}^{m} b_i^2 varX_i$$

.

Kapitel 4

Ausgewählte Verteilungen

Bei den ausgewählten Verteilungen handelt es sich um theoretische Zufallsverteilungen. Das sind Verteilungen, deren Form durch eine allgemein bekannte Funktion beschrieben wird. Oftmals kann beobachtet werden, dass die Verteilung bestimmter Zufallsvariablen annähernd durch eine theoretische Verteilung dargestellt werden kann, z. B. das Gewicht von Hähnchen einer Geflügelzucht ist meistens annähernd normalverteilt. Meist haben diese Verteilungen bestimmte Vorzüge, sie können leicht berechnet werden, und man kann auch wahrscheinlichkeitstheoretische Folgerungen ziehen. Hier bekannt ist bereits die Dreiecksverteilung.

Binomialverteilung

Das **Urnenmodell mit Zurücklegen** bestimmt die binomialverteilte Zufallsvariable.

Gegeben ist eine Urne mit zwei Sorten Kugeln. Man spricht von einer dichotomen (griech: zweigeteilten) Grundgesamtheit. Es sind insgesamt N Kugeln in der Urne und M Kugeln der ersten Sorte. Der Anteil der Kugeln erster Sorte ist also

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

,

$(0 \leq \theta \leq 1)$. Es werden n Kugeln mit Zurücklegen gezogen. Es ist die Zufallsvariable definiert:

X: Anzahl der Kugeln 1. Sorte unter den n gezogenen Kugeln.

Beispiele für binomialverteilte Zufallsvariablen

- In einer Urne befinden sich 3 schwarze und 12 weiße Kugeln. Es werden fünf Kugeln gezogen, wobei jede Kugel sofort wieder zurückgelegt wird (Modell mit Zurücklegen). Wir definieren X als Zahl der weißen Kugeln bei n = 5 Entnahmen.
- 10 mal Würfeln. X: Zahl der Würfe mit einer Augenzahl von mindestens 5.
- Einem sehr großen Fertigungslos von Kondensatoren werden 10 Kondensatoren entnommen. Erfahrungsgemäß sind 15% der Kondensatoren schadhaft. X: Zahl der schadhaften Kondensatoren.
- In einer Schulklasse mit 30 Schülern und Schülerinnen wird täglich ein Kind per Los zum Tafeldienst bestimmt. X: Zahl der Tage, die Paula innerhalb von n = 40 Tagen Tafeldienst machen musste.

Exkurs

Beispiel: Sie würfeln 5 mal. Mit welcher Wahrscheinlichkeit erhalten Sie zweimal Sechs?

Offensichtlich handelt es sich bei diesem Problem um ein Urnenmodell mit Zurücklegen. Es wäre beispielsweise die Wahrscheinlichkeit, dass die ersten zwei Würfe Sechs ergeben:

$$\theta = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^3 = 0{,}01608$$

.

Es gibt aber noch mehr Möglichkeiten, zwei Sechsen zu erhalten, nämlich:

(FFFSS), (FFSFS), (FFSSF), (FSFFS), (FSFSF), (FSSFF), (SFFFS), (SFFSF) und (SFSFF).

Hier bedeuten S: eine Sechs wird gewürfelt, F: keine Sechs wird gewürfelt. Es gibt insgesamt

$$\binom{5}{2} = \frac{5 \cdot 4}{1 \cdot 2} = 10$$

verschiedene Möglichkeiten, zwei Sechsen zu erhalten. Wir erhalten für die gesamte Wahrscheinlichkeit P(X = 2), dass bei fünf Versuchen genau zwei Sechsen resultieren:

$$P(X=2) = \binom{5}{2}\left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^3 = 10 \cdot 0{,}01608 = 0{,}1608.$$

Formale Darstellung

Die Zufallsvariable X ist **binomialverteilt mit den Parametern n und** θ. Ihre Wahrscheinlichkeitsfunktion lautet **(0** $\leq \theta \leq$ **1)**

$$P(X=x) = b(x|n;\theta) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{falls } x = 0,1,\ldots,n \\ 0 & \text{sonst.} \end{cases}$$

Der Binomialkoeffizient berechnet sich als

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad k,n \in N, \quad k,n \geq 0.$$

Siehe auch in der Wikipedia: Binomialkoeffizient

Die Verteilungsfunktion P(X $\leq$ a) = B(a|n; θ) ergibt sich als Summe der Wahrscheinlichkeiten einer diskreten Zufallsvariablen, wie in Zufallsvariablen oder Diskrete Zufallsvariablen erläutert.

Wie man der obigen Formel entnehmen kann, ist zur Berechnung der Wahrscheinlichkeiten die Kenntnis von N und M nicht erforderlich, es genügt die Bekanntheit von θ .

Weitere Kennwerte der Binomialverteilung sind

EX = n $\cdot\theta$ und varX = n $\cdot$ $\theta\cdot$(1 - θ) .

Beispiel: Verkehrszählung

Der Anteil der LKWs an den Kraftfahrzeugen auf deutschen Autobahnen soll für unser Beispiel 20% betragen. Im Rahmen einer Verkehrszählung an einer Auffahrt der Autobahn werden während einer Stunde 5 einfahrende Fahrzeuge zufällig erfasst.

1. Mit welcher Wahrscheinlichkeit befinden sich 2 LKWs in einer Stichprobe?
2. In wieviel Prozent der Stichproben befanden sich mindestens 2 LKWs in einer Stichprobe?

Es handelt sich offensichtlich um ein Modell mit Zurücklegen, denn ein Fahrzeug kann theoretisch auch mehrmals diese Auffahrt nehmen. Da wir die Fahrzeuge in LKW und Nicht-LKW unterscheiden, ist die betrachtete Grundgesamtheit dichotom (zwei Sorten Kugeln in der Urne). Wir definieren als Zufallsvariable X: Zahl der LKWs bei fünf gezählten Fahrzeugen.

X ist also binomialverteilt mit den Parametern n = 5 und $\theta = 0{,}2$ (20%), in Kurzschreibweise

$$X \sim b(x|5;0{,}2)$$

.

Wir werden zunächst die Wahrscheinlichkeitsfunktion von X bestimmen:

X = 0	$\binom{5}{0} \cdot \left(\frac{1}{5}\right)^0 \cdot \left(1-\frac{1}{5}\right)^{5-0} = 1 \cdot 1 \cdot \left(\frac{4}{5}\right)^5 = \frac{1024}{3125}$	0,32768
X = 1	$\binom{5}{1} \cdot \left(\frac{1}{5}\right)^1 \cdot \left(\frac{4}{5}\right)^4 = 5 \cdot \frac{4^4}{5^5} = 5 \cdot \frac{256}{3125}$	0,4096
X = 2	$\binom{5}{2} \cdot \left(\frac{1}{5}\right)^2 \cdot \left(\frac{4}{5}\right)^3 = 10 \cdot \frac{4^3}{5^5} = 10 \cdot \frac{64}{3125}$	0,2048
X = 3	$\binom{5}{3} \cdot \left(\frac{1}{5}\right)^3 \cdot \left(\frac{4}{5}\right)^2 = 10 \cdot \frac{4^2}{5^5} = 10 \cdot \frac{16}{3125}$	0,0512
X = 4	$\binom{5}{4} \cdot \left(\frac{1}{5}\right)^4 \cdot \left(\frac{4}{5}\right)^1 = 5 \cdot \frac{4}{5^5} = 5 \cdot \frac{4}{3125}$	0,0064
X = 5	$\binom{5}{5} \cdot \left(\frac{1}{5}\right)^5 \cdot \left(\frac{4}{5}\right)^0 = 1 \cdot \frac{1}{3125} \cdot 1$	0,00032

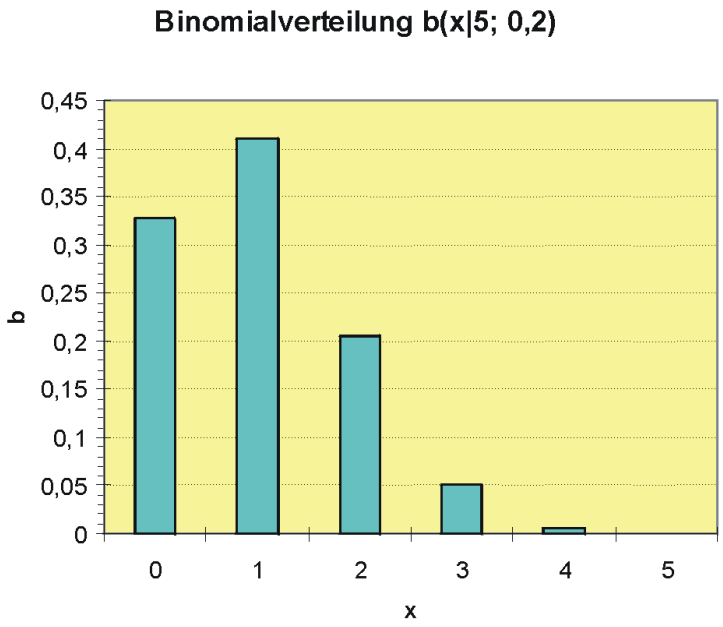

Abbildung 30: Wahrscheinlichkeitsfunktion der Binomialverteilung
mit n = 5 und $\theta = 0{,}2$

Wir erhalten dann die Wahrscheinlichkeitstabelle

x_i	0	1	2	3	4	5
$b(x_i\|5;0{,}2)$	0,32768	0,4096	0,2048	0,0512	0,0064	0,00032

Wir können also die gesuchten Wahrscheinlichkeiten aus der Tabelle ablesen

1. $P(X = 2) = 0{,}2048$
2. $P(X \geq 2) = 1$ - $P(X \leq 1) = 1$- $(0{,}3277 + 0{,}4096) = 0{,}2627$

Eigenschaften der Binomialverteilung

Bei einem Urnenmodell mit Zurücklegen und zwei Sorten Kugeln (dichotome Grundgesamtheit) ist die Zahl der Kugeln erster Sorte bei n Entnahmen **immer** binomialverteilt.

Bei einem relativ kleinen Anteil θ ist die Verteilung rechtsschief (bzw. linkssteil), da die Wahrscheinlichkeit für ein kleines x groß ist. Bei einem relativ großen Anteil θ ist die Verteilung linksschief, da die Wahrscheinlichkeit für ein großes x eher groß ist.

Ist $\theta = 0{,}5$, ist die Verteilung symmetrisch bezüglich $x = \frac{n}{2}$.

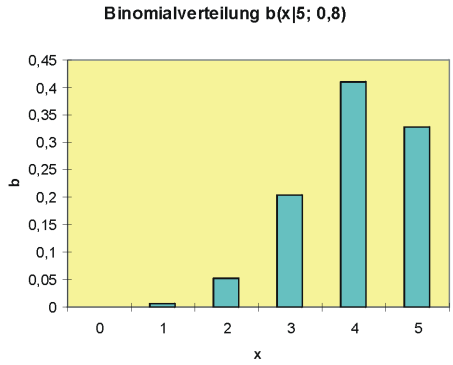

Abbildung 31: Wahrscheinlichkeitsfunktion der Binomialverteilung mit n = 5 und θ = 0,8

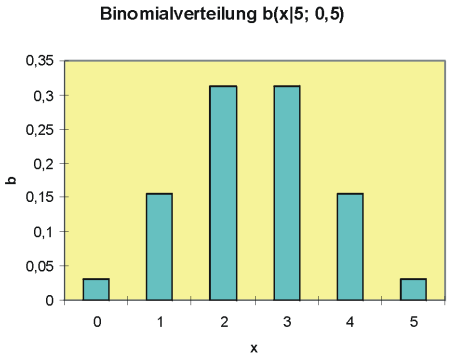

Abbildung 32: Wahrscheinlichkeitsfunktion der **symmetrischen** Binomialverteilung mit n = 5 und θ = 0,5

Bemerkung

Bei großem n wird die Berechnung der Binomialkoeffizienten ein numerisches Problem, das allerdings beispielsweise mit der Stirling-Formel gelöst werden kann. Bei der Berechnung von Verteilungswerten kann allerdings die Addition der Wahrscheinlichkeiten sehr umständlich werden. Unter Umständen kann man die Funktionswerte der Binomialverteilung durch die Poissonverteilung oder auch durch die Normalverteilung approximieren.

Siehe auch in der Wikipedia: Binomialverteilung

Das **Urnenmodell ohne Zurücklegen** bestimmt die hypergeometrisch verteilte Zufallsvariable.

Gegeben ist eine Urne mit zwei Sorten Kugeln. Man spricht von einer dichotomen (griech: zweigeteilten) Grundgesamtheit. Es sind insgesamt N Kugeln in der Urne und M Kugeln der ersten Sorte. Der Anteil der Kugeln erster Sorte ist also

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

,

$(0 \leq \theta \leq 1)$. Es werden n viele Kugeln ohne Zurücklegen gezogen. Es ist die Zufallsvariable definiert:

X: Anzahl der Kugeln 1. Sorte unter den n gezogenen Kugeln.

Beispiele für Hypergeometrische Verteilungen

- In einer Urne befinden sich 3 schwarze und 12 weiße Kugeln. Es werden fünf Kugeln ohne Zurücklegen gezogen (Modell ohne Zurücklegen). Wir definieren X als Zahl der weißen Kugeln bei n = 5 Entnahmen.
- Einem Fertigungslos von 100 Kondensatoren werden 10 Kondensatoren entnommen. Erfahrungsgemäß sind 15% der Kondensatoren schadhaft. X: Zahl der schadhaften Kondensatoren unter den 10 gezogenen.

Eine Zufallsvariable X ist "'hypergeometrisch verteilt mit den Parametern N, M und n, wenn ihre Wahrscheinlichkeitsfunktion lautet

$$P(X = x) = h(x|N; M; n) = \begin{cases} \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}} & \text{für x = 0, 1, ... , n} \\ 0 & \text{sonst} \end{cases}$$

Die Verteilungsfunktion P(X ≤ a) = H(a|N; M; n) ergibt sich als Summe der Wahrscheinlichkeiten einer diskreten Zufallsvariablen, wie in Zufallsvariablen oder Diskrete Zufallsvariablen erläutert.

Weitere Kennwerte der hypergeometrischen Verteilung sind Erwartungswert und Varianz,

$$EX = n \cdot \frac{M}{N}$$

und $varX = n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right) \frac{N-n}{N-1}$.

Der letzte Bruch wird Korrekturfaktor genannt; er korrigiert die Varianz bei einem Modell ohne Zurücklegen. Wir können leicht sehen, dass für eine sehr große Grundgesamtheit (N) dieser Faktor etwa 1 wird. Bei einer großen Grundgesamtheit kann man also das Modell ohne Zurücklegen durch ein Modell mit Zurücklegen annähern.

Beispiel:

Von den sechs Rettichen, die eine Marktfrau auf dem Wochenmarkt verkauft, sind vier holzig. Der Student Paul sucht sich 4 Rettiche aus. Man könnte sich nun fragen: Mit welcher Wahrscheinlichkeit erwischt er alle holzigen?

Hier haben wir es unzweifelhaft mit einem **Modell ohne Zurücklegen** zu tun. Da wir holzige und nicht holzige Rettiche vor uns haben, ist die betrachtete Grundgesamtheit dichotom (zwei Sorten Kugeln in der Urne).

Wir definieren als Zufallsvariable X: Zahl der holzigen Rettiche bei n = 4 Entnahmen.

X ist also hypergeometrisch verteilt mit den Parametern N = 6, M = 4 und n = 4, in Kurzschreibweise

$$X \sim h(x|N; M; n) = h(x|6; 4; 4)$$

.

Wir werden zunächst die Wahrscheinlichkeitsfunktion von X bestimmen:

X = 0	$\frac{\binom{4}{0}\cdot\binom{6-4}{4-0}}{\binom{6}{4}} = \frac{1\cdot 0}{15}$	0
X = 1	$\frac{\binom{4}{1}\cdot\binom{2}{3}}{\binom{6}{4}} = \frac{4\cdot 0}{15}$	0
X = 2	$\frac{\binom{4}{2}\cdot\binom{2}{2}}{\binom{6}{4}} = \frac{6\cdot 1}{15}$	$\frac{6}{15}$
X = 3	$\frac{\binom{4}{3}\cdot\binom{2}{1}}{\binom{6}{4}} = \frac{4\cdot 2}{15}$	$\frac{8}{15}$
X = 4	$\frac{\binom{4}{4}\cdot\binom{2}{0}}{\binom{6}{4}} = \frac{1\cdot 1}{15}$	$\frac{1}{15}$

Überlegen Sie sachlogisch, warum die ersten beiden Wahrscheinlichkeiten Null sind.

Der Student Paul wird also mit einer Wahrscheinlichkeit von 1/15 alle vier holzigen Rettiche erwischen.

Bemerkung

Werden M oder N groß, wird die Berechnung der Binomialkoeffizienten ein numerisches Problem, das allerdings beispielsweise mit der Stirling-Formel gelöst werden kann. Da der Unterschied zwischen einem Modell ohne Zurücklegen und mit Zurücklegen bei großem N unerheblich wird (ob man bei einer Entnahme 10000 oder 10001 Kugeln in der Urne hat, macht zahlenmäßig wenig aus), kann man bei großem N auch näherungsweise ein Modell mit Zurücklegen (siehe hierzu Binomialverteilung) verwenden. Häufig ist auch N unbekannt, hier kann das Modell ohne Zurücklegen gar nicht berechnet werden.

Wir betrachten eine poissonverteilte Zufallsvariable X mit den Ausprägungen 0, 1, 2, ...

Typische Beispiele für eine poissonverteilte Zufallsvariable sind:

- Es betreten in einer Minute durchschnittlich $\lambda = 2$ Kunden einen Kassenschalter. Wir definieren als X: Zahl der Kunden, die während einer bestimmten Minute an den Bankschalter kommen.
- Die Studentin Paula kauft sich in der Cafeteria ein Stück Rührkuchen. Wir definieren als X: Zahl der Rosinen in diesem Kuchenstück. Der Bäcker rechnet bei 20 Stück Kuchen mit 100 Rosinen. X ist also poissonverteilt mit dem Parameter $\lambda = 5$.
- Wir definieren als X: Zahl der Schadensfälle einer Versicherung im nächsten Jahr. Man weiß, daß pro Jahr durchschnittlich 500 000 Schadensfälle auftreten. Der Parameter ist hier $\lambda = 500\ 000$.

Man geht also typischerweise von den folgenden Fragestellungen aus: Anzahl des Auftretens eines Phänomens in einer Zeit- , Gewichts- oder sonstigen Einheit. Die Zufallsvariable X ist **poissonverteilt** mit dem Parameter λ.

Ihre Wahrscheinlichkeitsfunktion lautet ($\lambda > 0$)

$$P(X = x) = p(x|\lambda) = \begin{cases} \frac{e^{-\lambda} \cdot \lambda^x}{x!} & \text{für x} = 0, 1, \ldots \\ 0 & \text{sonst} \end{cases}$$

Die Verteilungsfunktion $P(X \leq a) = Px(a|\lambda)$ ergibt sich als Summe der Wahrscheinlichkeiten einer diskreten Zufallsvariablen, wie in Zufallsvariablen oder Diskrete Zufallsvariablen erläutert.

Es gilt bei der Poissonverteilung: EX = varX = λ.

Die Poissonverteilung ist **reproduktiv**: Eine Summe von n stochastisch unabhängigen poissonverteilten Zufallsvariablen X_i (i = 1, ... , n), mit jeweils dem Parameter λ_i, ist wiederum poissonverteilt, und zwar mit dem Parameter

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

Beispiel:

Von den mundgeblasenen Gläsern einer Glashütte ist bekannt, dass im Durchschnitt 0,2 Fehler pro Glas auftreten.

Es ist die diskrete Zufallsvariable X: ||Die Zahl der Unreinheiten in einem Glas|| annähernd poissonverteilt:

$$X \to p(x|0{,}2)$$

.

a) Mit welcher Wahrscheinlichkeit hat ein Glas genau einen Fehler?

$$P(X = 1) = \frac{e^{-0,2} \cdot 0,2^1}{1!} = 0,2 \cdot e^{-0,2} = 0,1637$$

b) Mit welcher Wahrscheinlichkeit hat ein Glas mindestens zwei Fehler?

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - \left(\frac{e^{-0,2} \cdot 0,2^0}{0!} + \frac{e^{-0,2} \cdot 0,2^1}{1!} \right)$$

$$= 1 - e^{-0,2} - 0,1637 = 1 - 0,8187 - 0,1637 = 0,0175.$$

c) Mit welcher Wahrscheinlichkeit enthalten drei Gläser zusammen mindestens zwei Fehler? Man geht davon aus, dass die Fehler der Gläser stochastisch unabhängig sind.

Man definiert als neue Zufallsvariable Y = X_1 + X_2 + X_3, mit X_1 als Zahl der Fehler des ersten Glases usw. Es ist dann $\lambda_y = 0,2 + 0,2 + 0,2 = 0,6$ und

$$P(Y \geq 2) = 1 - P(Y \leq 1) = 1 - \left(\frac{e^{-0,6} \cdot 0,6^0}{0!} + \frac{e^{-0,6} \cdot 0,6^1}{1!} \right)$$

$$= 1 - (e^{-0,6} + 0,6 \cdot e^{-0,6}) = 0,1219.$$

Was ist die Normalverteilung?

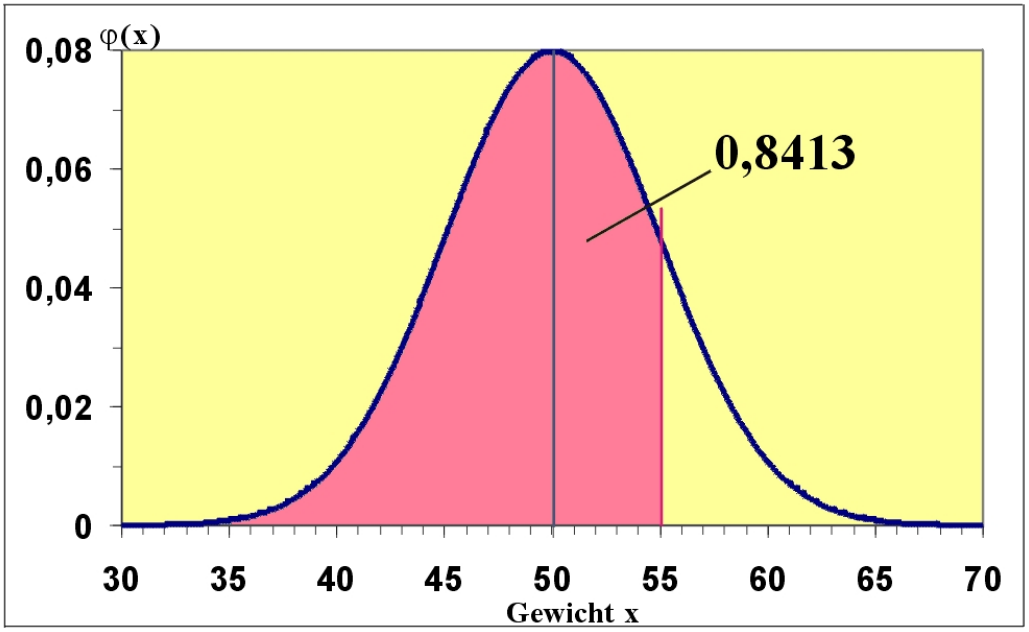

Abbildung 33: Normalverteilung des Gewichts von Eiern (g)

Beispiel:

Auf einer Hühnerfarm mit sehr vielen Hühnern werden eine Woche lang die einzelnen Eier gewogen. Definieren wir die Zufallsvariable X: Gewicht eines Eis in Gramm. Es stellt sich heraus, daß ein Ei im Durchschnitt 50 g wiegt. Der Erwartungswert EX ist daher 50. Außerdem sei bekannt, dass die Varianz varX = 25 g^2 beträgt. Man kann die Verteilung des Gewichts annähernd wie in der Grafik darstellen. Man sieht, daß sich die meisten Eier in der Nähe des Erwartungswerts 50 befinden und daß die Wahrscheinlichkeit, sehr kleine oder sehr große Eier zu erhalten, sehr klein wird. Wir haben hier eine Normalverteilung vor uns. Sie ist typisch für Zufallsvariablen, die sich aus sehr vielen verschiedenen Einflüssen zusammensetzen, die man nicht mehr tren-

nen kann, z.B. Gewicht des Huhns, Alter, Gesundheit, Standort, Vererbung usw.

Die Dichtefunktion der Normalverteilung ist definiert als

$$\phi_x(x|\mu;\sigma^2) = \frac{1}{\sqrt{2\cdot\pi\cdot\sigma^2}}\ \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \text{ für x } \in \mathbb{R}\ ,$$

wobei $E(X) = \mu$ und $varX = \sigma^2$ ist. Man sagt, X ist normalverteilt mit den Parametern μ und σ^2, in Symbolschreibweise

$$X \sim \phi_x(x|\mu;\sigma^2)$$

oder kürzer $X \sim N(\mu;\sigma^2)$.

In unserem Beispiel ist $X \sim N(50;25)$.

Die Normalverteilung ist symmetrisch bezüglich μ. Die Verteilung P(X $\leq$ a) von X ist wieder die Fläche unter dem Graph der Dichtefunktion. Sie wird bezeichnet als

$$P(X \leq a) = \Phi_x(a|\mu;\sigma^2) \text{ für alle a } \in \mathbb{R}\ .$$

Beispielsweise beträgt die Wahrscheinlichkeit, dass ein Ei höchstens 55 g wiegt, 0,8413. Das entspricht der roten Fläche in der Abbildung.

Das Integral der Dichtefunktion kann nicht analytisch berechnet werden. Die Werte der Verteilungsfunktion liegen i.a. tabellarisch vor. Es besteht nun das Problem, daß für jeden Wert von μ und σ^2 eine eigene Tabelle vorliegen müsste. Hier ist hilfreich, daß die aus X standardisierte Zufallsvariable Z wiederum normalverteilt ist und zwar mit den Parametern 0 und 1. Es kann jede beliebige Normalverteilung standardisiert werden. Mit Hilfe der standardisierten Zufallsvariablen wird dann die gesuchte Wahrscheinlichkeit bestimmt.

Standardnormalverteilung

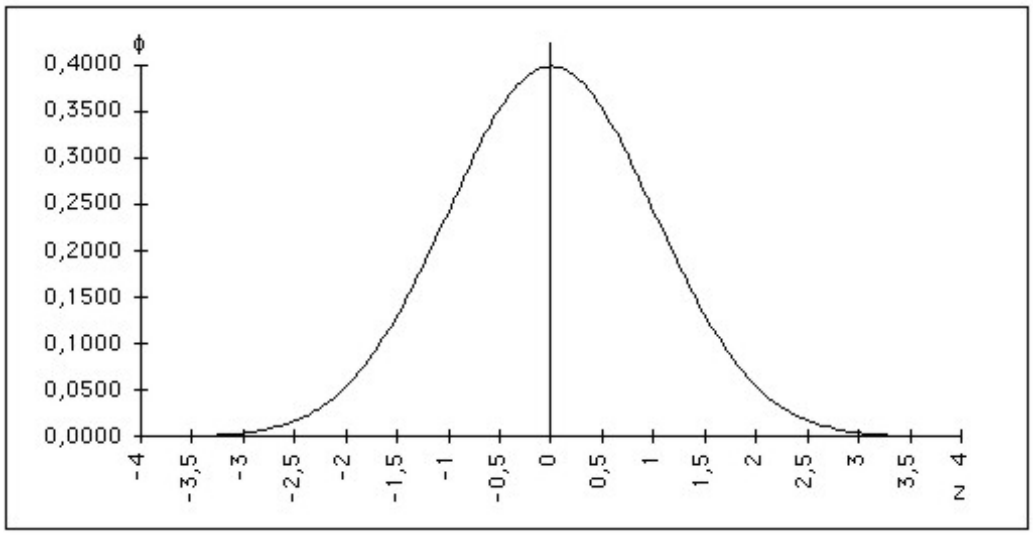

Abbildung 34: Dichtefunktion der Standardnormalverteilung

Man definiert also eine neue Zufallsvariable

$$Z = \frac{X - EX}{\sqrt{varX}} = \frac{X - \mu}{\sigma}.$$

Diese Zufallsvariable Z ist normalverteilt mit EZ = 0 und varZ = 1. Ihre Dichtefunktion ist in der folgenden Grafik dargestellt. Es ist also $Z \sim N(0;1)$.

Die Dichtefunktion von Z ist

$$\phi_z(z|0;1) = \frac{1}{\sqrt{2 \cdot \pi}} exp(-\frac{z^2}{2}) \text{ für z } \in \mathbb{R}$$

Ihre Verteilung, die man auch kurz als Φ(z) bezeichnet, ist (z const.)

$$P(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} \phi(u) du.$$

Verteilungswerte

Es ist beispielsweise die Wahrscheinlichkeit

$$P(Z \leq 0,44) = 0,6700$$

und

$$P(Z > 1,64) = 0,0505.$$

Wir wollen nun den Anteil der Eier mit höchstens 55 g bestimmen, also P(X $\leq$ 55). Wir standardisieren:

$$Z = \frac{z - \mu}{\sigma} = \frac{55 - 50}{\sqrt{25}} = 1 \, .$$

Es ist dann

$$P(X \leq 55) = P(Z \leq 1) = \Phi(1) = 0,8413.$$

Der Wert 0,8413 der Verteilungsfunktion wird in der Normalverteilungstabelle ermittelt. Der folgende Ausschnitt aus der Tabelle soll die Vorgehensweise verdeutlichen: In der ersten Spalte der Tabelle sind die zwei ersten signifikanten Stellen der Ausprägung z angegeben, in der ersten Tabellenzeile die zweite Nachkommastelle, so dass sich beispielsweise z = 1,00 zusammensetzt aus 1,0 + 0,00. Wo sich Zeile und Spalte des betreffenden Z-Wertes kreuzen, steht die gesuchte Wahrscheinlichkeit.

z	0,00	0,01	0,02
0,0	5000	5040	5080
0,1	5398	5438	5478
0,2	5793	5832	5871
0,3	6179	6217	6255
0,4	6554	6591	6628

0,5	6915	6950	6985
0,6	7257	7291	7324
0,7	7580	7611	7642
0,8	7881	7910	7939
0,9	8159	8186	8212
1,0	8413	8438	8461
1,1	8643	8665	8686
1,2	8849	8869	8888

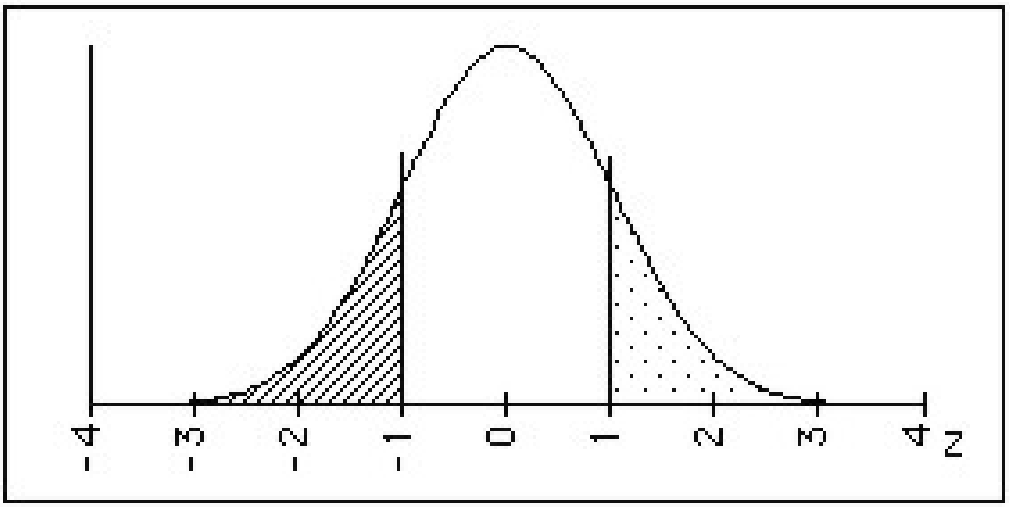

Abbildung 35

Der errechnete Wert z kann gerundet werden, falls die errechneten Stellen die Zahl der Stellen des tabellierten z-Wertes übertreffen. Da die Verteilung von Z symmetrisch bezüglich $\mu = 0$ ist, genügt die Tabellierung der Verteilungswerte ab z = 0 bzw. Φ(z) = 0,5. Es gilt, wie man auch anhand der Grafik leicht sieht:

$$P(Z \leq -z) = P(Z \geq z) = 1 - P(Z \leq z) = 1 - \Phi(z)$$

bzw.

$$\Phi(-z) = 1 - \Phi(z)$$

.

Beispiel:

$$\begin{array}{rcccl} P(Z \leq -1) & = & P(Z \geq 1) & = & 1 - P(Z \leq 1) \\ & = & 1 - \Phi(1) & = & 1 - 0,8413 \quad = 0,1587 \end{array}$$

Quantil

Häufig sucht man zu einer gegebenen Wahrscheinlichkeit p den dazugehörigen z-Wert z(p). Er wird als **p-Quantil** bezeichnet.

Es gilt also:

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

.

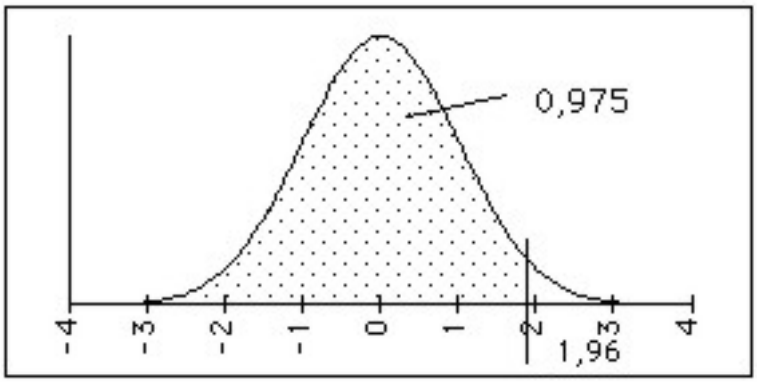

Abbildung 36: 97,5%-Quantil der Standardnormalverteilung

Beispielsweise ist z(0,975) = 1,96. Es ist also hier die Wahrscheinlichkeit 0,975 **gegeben** und der dazugehörige z-Wert wird **gesucht**. Man sucht in der Tabelle die Wahrscheinlichkeit 0,9750 und bestimmt dann am Rand den betreffenden z-Wert 1,96.

Liegt p zwischen zwei Tabellenwerten, genügt es, als p den Tabellenwert zu verwenden, der p am nächsten liegt.

Beispiel:

Gesucht: z(0,9)

	näher bei	0,9	
Wahrscheinlichkeit Φ	0,8997		0,9015
z-Wert oder Quantil	1,28		1,29

Es ist also $z(0,9) \approx 1,28$.

Für eine Normalverteilung mit μ und σ^2 berechnet sich das p-Quantil als

$$x(p) = \mu + \sigma \cdot z(p).$$

Beispiel:

Wie schwer sind höchstens die 2/3 leichtesten Eier? Gesucht ist also x(0,67):

$$x(p) = 50 + 5 \cdot z(0,67) = 50 + 5 \cdot 0,44 = 52,2.$$

Das schwerste der 67% leichtesten Eier wog also 52,2g .

Übung zur Berechnung von $\Phi_z(z)$

Schraffieren Sie die gesuchte Wahrscheinlichkeit in der Grafik und berechnen Sie die gesuchten Werte:

P(Z $\leq$ 0,51) P(Z $\leq$ 2,0) = P(Z $\leq$ - 0,51)

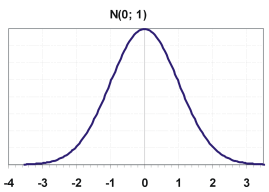

Abbildung 37

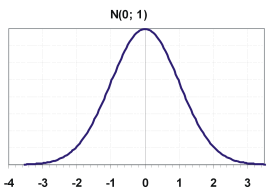

Abbildung 38

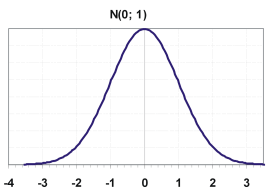

Abbildung 39

P(1,5 ≤ Z ≤ 2,35) P(- 0,8 ≤ Z ≤ 1,05) P(Z ≥ -0,89)

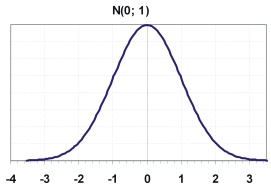

Abbildung 40
P(Z ≤ -1,68 ∪ Z ≥ 2)

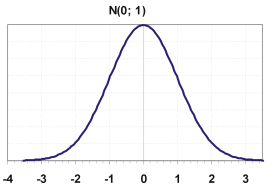

Abbildung 41
P(Z ≤ -1,96 ∪ Z ≥ 1,96)

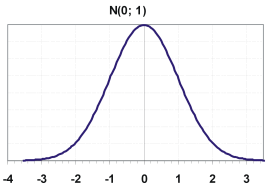

Abbildung 42
P(Z ≤ -5)

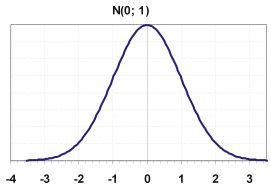

Abbildung 43

z(0,975)

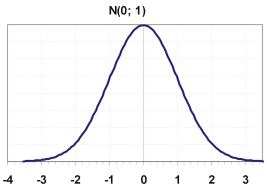

Abbildung 44
z(0,8)

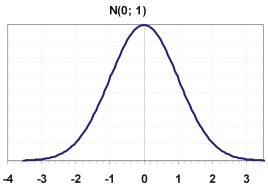

Abbildung 45

z(0,2)

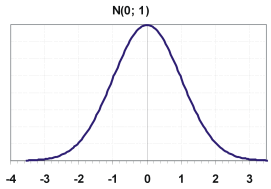

Abbildung 46

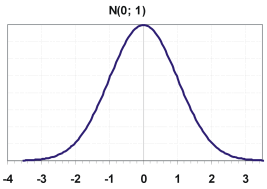

Abbildung 47

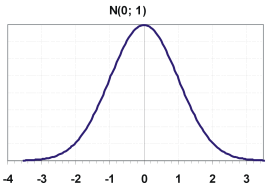

Abbildung 48

Übungen zum Eier-Beispiel

1. Wie groß ist die Wahrscheinlichkeit, daß ein Ei höchstens 60 g wiegt?
2. Wieviel Prozent der Eier wiegen höchstens 50 g?
3. Wie groß ist die Wahrscheinlichkeit, daß ein Ei mindestens 45 g wiegt?
4. Wieviel Prozent der Eier liegen zwischen 45 und 55 Gramm?
5. Mit welcher Wahrscheinlichkeit wiegt ein Ei genau 53 Gramm?
6. Welches Mindestgewicht haben die 30% schwersten Eier?

Lösungen:

Übung zur Berechnung von $\Phi_z(z)$

a) 0,6950 b) 0,9772 c) 0,3050 d) 0,0574 e) 0,6412 f) 0,8133 g) 0,0693 h) 0,05 i) 0 j) 1,96 k) 0,84 l) -0,84

Zentraler Grenzwertsatz

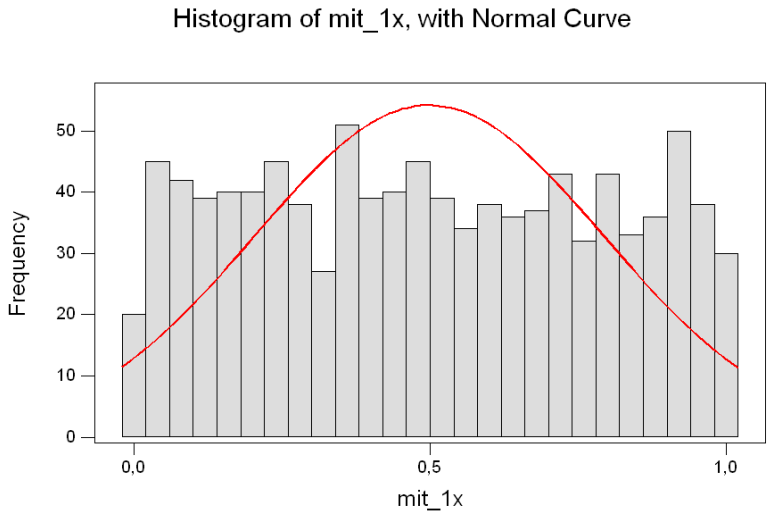

Abbildung 49: Histogramm einer gleichverteilten Zufallsvariablen

Gegeben sind die stochastisch unabhängigen Zufallsvariablen X_i (i = 1, 2,...). Die Verteilungen der Summen Y_i

$$Y_1 = X_1 \ , \ Y_2 = X_1 + X_2 \ , \ ..., \ Y_n = X_1 + X_2 + ... + X_n \ , \ ...$$

streben mit wachsendem n gegen die Normalverteilung. Als Faustregel gilt, daß die Verteilung einer Summe von mehr als 30 stochastisch unabhängigen Zufallsvariablen schon sehr gut annähernd mit der Normalverteilung bestimmt werden kann (n > 30).

Diese Regel ermöglicht zum einen die Bestimmung von Wahrscheinlichkeiten unbekannt verteilter Zufallsvariablen, zum anderen kann die Bestimmung kompliziert zu berechnender Wahrscheinlichkeitswerte mit der Normalverteilung angenähert (approximiert) werden.

Als Beispiel wurden je 1000 Zufallszahlen von im Intervall [0;1] gleichverteilten Zufallsvariablen erzeugt. Der Graph ihrer Dichtefunktion bildet ein Rechteck. Das Histogramm der Zufallszahlen lässt bei 1000 Werten deutlich

das Rechteck erkennen. Bei der Summe von zwei gleichverteilten Variablen zeichnet sich die unimodale symmetrische Struktur schon deutlich ab, wobei zu bemerken ist, dass die Summe von zwei gleichverteilten Zufallsvariablen eine Dreiecksverteilung ergibt. Bei 31 Variablen ist die Näherung zur Normalverteilung schon sehr ausgeprägt.

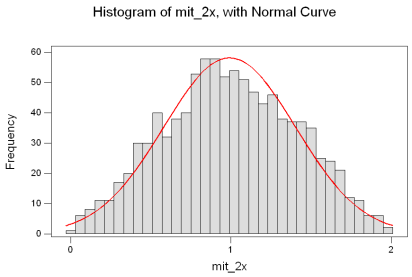

Abbildung 50: Histogramm der Summe von zwei gleichverteilten Zufallsvariablen

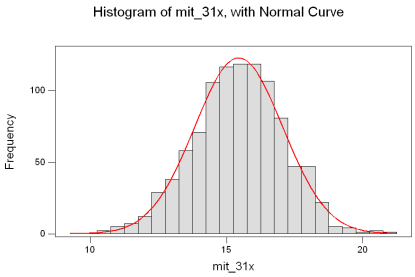

Abbildung 51: Histogramm der Summe von 31 gleichverteilten Zufallsvariablen

Linearkombinationen normalverteilter Zufallsvariablen

Gegeben sind n normalverteilte Zufallsvariablen X_i (i = 1, ... , n), mit $X_i \sim N(\mu_i;\sigma_i^2)$. Die Linearkombination (lineare Funktion)

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_nX_n = a_0 + \sum_{i=1}^{n} a_iX_i$$

ist ebenfalls normalverteilt (Reproduktivität der Normalverteilung), und zwar mit dem Erwartungswert

$$EY = a_0 + \sum_{i=1}^{n} a_iEX_i = a_0 + \sum_{i=1}^{n} a_i\mu_i$$

und, falls die X_i (i = 1, ... , n) stochastisch unabhängig sind, mit der Varianz

$$\operatorname{var} Y = \sum_{i=1}^{n} a_i^2 \operatorname{var} X_i = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

.

Da die Varianz jedoch echt größer Null sein muss, muss zudem $a_j \neq 0$ für mindestens ein $j \in \{1, \ldots, n\}$ gefordert werden.

Verteilung des Stichprobendurchschnitts

Sind speziell die n Zufallsvariablen X_i (i = 1, ... , n) sämtlich normalverteilt mit gleichem μ und gleichem σ^2, ist die Linearkombination X mit $a_0 = 0$, $a_1 = a_2 = ... = a_n = 1/n$, also

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

normalverteilt dem Erwartungswert

$$E\bar{X} = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

und, falls die Xi (i = 1, ... , n) stochastisch unabhängig sind, mit der Varianz

$$\operatorname{var} \bar{X} = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

.

Beispiel

Die Firma Ziemlich& Unbekannt produziert die Güter Ix und Ypsi. Die monatliche Produktionsmenge schwankt zufällig, so dass für die produzierten Mengen die Zufallsvariablen definiert werden: X und Y [ME]. Man weiß:

$X \sim N(20;5)$ und $Y \sim N(100;10)$.

Es wird vermutet, dass X und Y stochastisch unabhängig sind.

Wir interessieren uns für die monatlichen Gesamtkosten K in Crœtos (C):

Die monatlichen Fixkosten betragen a = 10.000 C, die variablen Kosten für X: b = 500 C und für Y: c = 200 C.

Die monatlichen Gesamtkosten können also dargestellt werden als

$$K = a + bX + cY = 10000 + 500X + 200Y.$$

Wie ist also K verteilt? Wegen der Reproduktivitätseigenschaft der Normalverteilung müsste K wieder normalverteilt sein. Seine Parameter sind

$$EK = a + b\,EX + c\,EY = 10.000 + 500{\cdot}20 + 200{\cdot}100 = 40.000$$

und

$$varK = b^2 varX + c^2 varY = 500^2{\cdot}5 + 200^2{\cdot}10 = 1.650.000.$$

Also ist $K \sim N(40.000;\ 1.650.000)$.

Mit welcher Wahrscheinlichkeit entstehen der Firma Gesamtkosten von mindestens 42.000 C?

Es ergibt sich

$$\begin{array}{rcccl} P(K \geq 42000) & = & 1 - P(K \leq 42000) & = & 1 - \Phi_z(\frac{42000-40000}{\sqrt{1650000}}) \\ & = & 1 - \Phi_z(1{,}57) & = & 1 - 0{,}9418 = 0{,}0582. \end{array}$$

χ^2-Verteilung

Beispiel

Wir haben 3 normalverteilte, paarweise stochastisch unabhängige Zufallsvariablen X_1, X_2 und X_3 gegeben mit den Erwartungswerten μ_1, μ_2 μ_3 und den Varianzen ${\sigma_1}^2$, ${\sigma_2}^2$,${\sigma_3}^2$. Wir standardisieren diese Variablen und erhalten 3 standardnormalverteilte Zufallsvariablen Z_1, Z_2 und Z_3,

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1},\ Z_2 = \frac{X_2 - \mu_2}{\sigma_2},\ Z_3 = \frac{X_3 - \mu_3}{\sigma_3}\ .$$

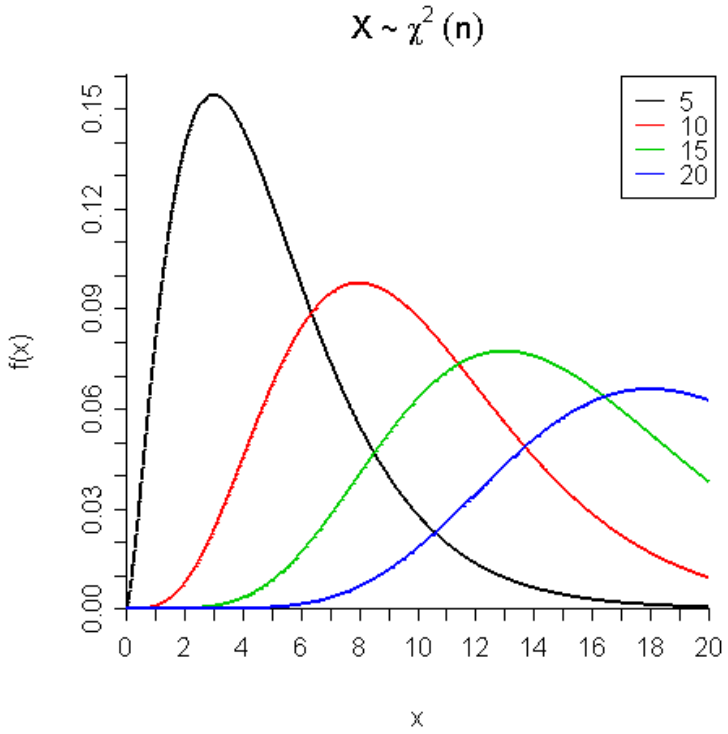

Abbildung 52: Dichtefunktion der χ^2-Verteilung mit ausgewählten Freiheitsgraden

Nun werden die standardnormalverteilten Zufallsvariablen quadriert und aufsummiert. Wir erhalten eine neue Zufallsvariable

$$Y = Z_1^2 + Z_2^2 + Z_3^2 \ .$$

Y ist χ^2-verteilt mit 3 Freiheitsgraden.

Allgemein

Es gilt: Die Summe von m quadrierten, stochastisch unabhängigen, standardnormalverteilten Zufallsvariablen ist χ^2-verteilt mit m Freiheitsgraden.

Man sieht anhand der Grafik, dass sich die Dichtefunktion mit wachsenden Freiheitsgraden einer symmetrischen Kurve nähert.

Die Wahrscheinlichkeit wird bezeichnet als $P(Y \leq a) = f_Y(a|n)$. Das p-Quantil ist $\chi^2(p;n)$.

Die Verteilungsfunktion der χ^2-Verteilung kann nicht analytisch ermittelt werden. Numerische Berechnungen können beispielsweise aus Tabellenwerken, etwa Tabelle der χ^2-Verteilung ersehen werden. Da Y für jeden Freiheitsgrad eine eigene Verteilung besitzt, sind in kleineren Tabellen wie oben nur Quantile nach Freiheitsgraden und ausgewählten Wahrscheinlichkeiten aufgeführt. Es ist z.B. das 95%-Quantil (Spalte) der χ^2-Verteilung mit 3 Freiheitsgraden (Zeile)

$f_Y(0{,}95;3) = 7{,}81$. Das bedeutet, die Wahrscheinlichkeit $P(y \leq 7{,}81) = 0{,}95$.

Gilt $n > 30$, ist

$$Z = \sqrt{2X} - \sqrt{2n-1}$$

näherungsweise standardnormalverteilt.

Nähere Erläuterungen zur χ^2-Verteilung, beispielsweise ihre Dichtefunktion, findet man bei Wikipedia. Da die Dichtefunktion jedoch nicht für die Berechnung der Verteilungswerte unmittelbar verwendet werden kann, wird sie hier nicht angeführt.

Beispiele:

Sei Y χ2-verteilt mit 10 Freiheitsgraden. Es ist

- $P(Y \leq 15,99) = 0,9$
- $P(Y > 3,94) = 1 - P(Y \leq 3,94) = 1 - 0,05 = 0,95$
- $P(3,25 \leq Y \leq 20,48) = P(Y \leq 20,48) - P(Y \leq 3,25) = 0,975 - 0,025 = 0,95$
- 10%-Quantil von Y : $\chi^2(0,1;10) = 4,87$
- 95%-Quantil von Y : $\chi^2(0,95;10) = 18,31$

Sei Y χ2-verteilt mit 61 Freiheitsgraden. Gesucht ist $P(Y \leq 98)$. Hier ist die Zahl der Freiheitsgrade $k > 30$. Es wird eine neue Zufallsvariable $X = \sqrt{2Y}$ gebildet. X ist näherungsweise normalverteilt wie $N(\sqrt{2k-1};1) = N(11;1)$. $P(Y \leq 98)$ entspricht also $P(X \leq \sqrt{2 \cdot 98}) = P(X \leq 14)$

Es ist $\Phi_X(14|11;1) = \Phi_X(\frac{14-11}{1}) = \Phi_X(3) = 0,9987$.

Bemerkung

Die χ^2-Verteilung ist **reproduktiv**, d.h. die Summe von zwei stochastisch unabhängigen χ^2-verteilten Zufallsvariablen mit m und n Freiheitsgraden ist wieder χ^2-verteilt mit m+n Freiheitsgraden.

Die χ^2-Verteilung ist eine so genannte Stichprobenverteilung.

Übung

1. Die Zufallsvariable X ist χ^2-verteilt mit 12 Freiheitsgraden.
   (a) Bestimmen Sie die Wahrscheinlichkeit, dass X kleiner als 6,30 ist.
   (b) Bestimmen Sie die Wahrscheinlichkeit, dass X mindestens 18,55 beträgt.
   (c) Bestimmen Sie das 5%-Quantil der Verteilung.

   Die Zufallsvariable Y ist χ^2-verteilt mit 40 Freiheitsgraden.
   (a) Bestimmen Sie die Wahrscheinlichkeit, dass Y kleiner als 40 ist.
   (b) Bestimmen Sie das 95%-Quantil der Verteilung.

   Es sei U=X+Y.
   (a) Bestimmen Sie den Erwartungswert von U.
   (b) Bestimmen Sie die Wahrscheinlichkeit, dass U kleiner als 40 ist.

F-Verteilung

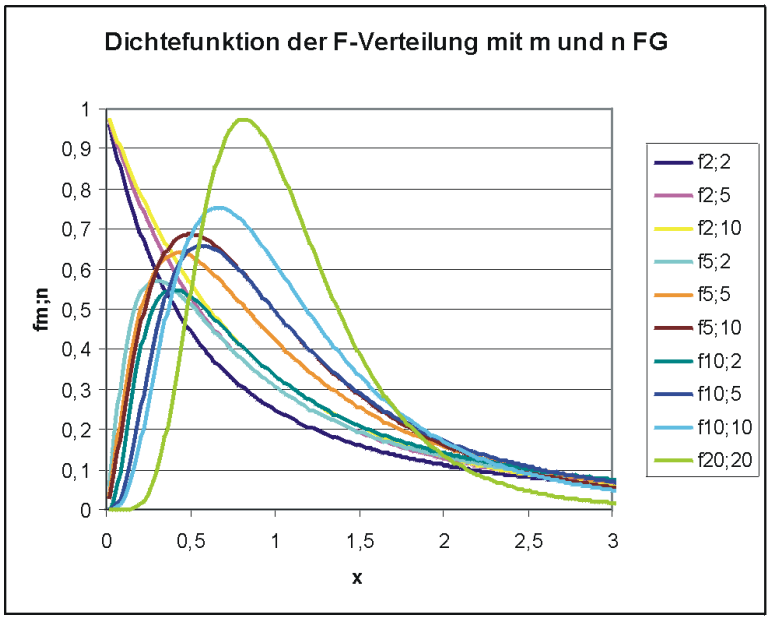

Abbildung 53: Dichtefunktion der F-Verteilung mit m und n Freiheitsgraden

Beispiel

Wir haben die drei standardnormalverteilten Zufallsvariablen von oben und vier weitere Z_4, Z_5, Z_6 und Z_7 gegeben. Alle Variablen sind wieder stochastisch unabhängig. Der Quotient

$$F = \frac{\frac{Z_1^2+Z_2^2+Z_3^2}{3}}{\frac{Z_4^2+Z_5^2+Z_6^2+Z_7^2}{4}}$$

ist dann F-verteilt mit 3 und 4 Freiheitsgraden.

Allgemein

Der Quotient aus zwei χ^2-verteilten Zufallsvariablen, jeweils geteilt durch ihre Freiheitsgrade, wobei die Zufallsvariable im Zähler m und die im Nenner

n Freiheitsgrade hat, ist F-verteilt mit m und n Freiheitsgraden. Einzelheiten dazu gibt es auch in der Wikipedia. Man schreibt

$$F \sim F_{m;n}$$

Die Wahrscheinlichkeit wird bezeichnet als $P(F \leq a) = f_F(a|m;n)$. Das p-Quantil ist F(p;m;n).

Auch die F-Verteilung liegt tabelliert vor und ist meistens nach ausgewählten Freiheitsgraden und Quantilen tabelliert. Eine nützliche Beziehung ist dabei $F(p;m;n) = \frac{1}{F(1-p;n;m)}$.

Für viele Freiheitsgrade kann man sich die Faustregel merken: Sind m und n größer als 30, kann man die Quantile näherungsweise mit der Standardnormalverteilung ermitteln:

$$F(p;m;n) \approx z(p) \ .$$

Die F-verteilung ist ebenfalls eine Stichprobenverteilung. Sie ist aber nicht reproduktiv.

t-Verteilung

Beispiel

Gegeben sind die standardnormalverteilten Zufallsvariablen von oben.

Der Quotient

$$t = \frac{Z_1}{\sqrt{\frac{Z_2^2+Z_3^2+Z_4^2+Z_5^2}{4}}}$$

ist t-verteilt mit 4 Freiheitsgraden.

Allgemein

Der Quotient aus einer standardnormalverteilten Zufallsvariablen und der Wurzel einer χ^2-verteilten Zufallsvariablen mit n Freiheitsgraden, geteilt durch ihre Freiheitsgrade, ist t-verteilt mit n Freiheitsgraden.

Die Wahrscheinlichkeit wird bezeichnet als $P(t \leq a) = f_t(a|n)$. Das p-Quantil ist t(p;n).

Die Dichtefunktion der t-Verteilung ist, ähnlich wie die der Standardnormalverteilung, symmetrisch bezüglich des Erwartungswertes 0. Es gilt daher für die Berechnung der Verteilungswerte:

$$P(t \leq -a) = P(t \geq a),$$

$a \in \mathbf{R}$.

Auch die t-Verteilung ist meistens nach Freiheitsgraden und ausgewählten Quantilen tabelliert: t-Verteilung

Für $n > 30$ kann man die Wahrscheinlichkeiten der t-Verteilung approximativ mit der Normalverteilung berechnen:

$t(p; n) \approx z(p)$.

Bemerkungen:

- Das Quadrat einer t-verteilten Zufallsvariablen ist F-verteilt.
- Die t-Verteilung ist eine Stichprobenverteilung
- Weitere Eigenschaften können in der Wikipedia nachgelesen werden.

Approximation: Approximation heißt Näherung, wie ja beispielsweise Alpha Proxima Centauri (eigentlich aus dem Drei-Sterne-System Alpha Centauri) der uns am nächsten gelegene Stern ist. Wir wollen also Verteilungswerte, bei deren Berechnung wir heftige Unlustgefühle entwickeln, mit Hilfe anderer Verteilungen annähern. Sie werden nun mit Recht einwenden, dass das ja heutzutage mit der Entwicklung schneller Rechner eigentlich überflüssig ist. Nun hat man aber nicht immer einen Computer dabei (etwa in einer Klausur) oder es fehlt die Software zur Berechnung. MS-Excel bietet zwar solche Funktionen, aber die Umsetzung ist etwas verquer, so dass häufig ein erhöhter Verstehensaufwand betrieben werden muss. Bei bestimmten Funktionswerten, wie großen Binomialkoeffizienten gehen schon mal Taschenrechner in die Knie.

Approximation diskreter Verteilungen durch diskrete Verteilungen

Die Wahrscheinlichkeitsfunktion der Hypergeometrischen Verteilung sieht so aus:

$$\frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

Haben wir als Anwendung eine Kiste mit 10 Ü-Eiern gegeben, von denen 3 den gesuchten Obermotz enthalten, kann man etwa die Wahrscheinlichkeit, bei 5 Versuchen zwei Obermotze zu erhalten, leicht errechnen - naja, relativ leicht.

Aber betrachten wir den Fall: In einer Sendung von 500 speziellen Chips sind 100 Stück defekt. Bei der Eingangskontrolle werden 20 Chips getestet. Wenn jetzt die Wahrscheinlichkeit verlangt wird, dass genau 10 defekte Chips gezogen werden, erhält man

$$\frac{\binom{400}{10} \cdot \binom{100}{10}}{\binom{500}{20}} .$$

Spüren Sie schon Unlustgefühle? Vielleicht können wir uns hier die Berechnung mit der Binomialverteilung erleichtern. Vergleichen wir die beiden Verteilungen, fällt auf, dass beide den gleichen Erwartungswert haben: EX = nθ. Nur in den Varianzen unterscheiden sie sich,

> Binomialverteilung: $varX = n\theta(1-\theta)$ und hypergeometrische Verteilung: $varX = n\theta(1-\theta)\frac{N-n}{N-1}$,

nämlich im Korrekturfaktor. Wird nun N sehr groß, ist der Korrekturfaktor fast Eins und wir erhalten approximativ die Varianz der Binomialverteilung. Wie groß ist jetzt ein großes N? Das kommt darauf an, wie genau wir die Näherung haben wollen. Für die Approximation der Hypergeometrischen Verteilung durch die Binomialverteilung gibt es mehrere empfohlene Faustregeln, je nach Geschmack der Autoren. Eine der einfacheren Faustregeln, die man sich auch einigermaßen merken kann, ist

$$h(x|N;M;n) \approx b(x|n;\cdot\frac{M}{N}), \text{ wenn } \frac{n}{N} < 0,05$$

ist. Da in unserem Beispiel diese Voraussetzungen erfüllt sind, berechnen wir die gesuchte Wahrscheinlichkeit als

$$\binom{20}{10} \cdot 0,8^{10} \cdot 0,2^{10} .$$

Wir haben also das Modell ohne Zurücklegen durch ein Modell mit Zurücklegen angenähert. Man könnte so argumentieren: Wenn etwa 10000 Kugeln in einer Urne sind, macht es kaum einen Unterschied, ob beim 2. Versuch noch 9999 oder 10.000 Kugeln übrig sind. Analoges gilt für die Zahl der Kugeln 1. Sorte. Deshalb genügt auch die Angabe des Anteils θ dieser Kugeln an der Gesamtheit der Kugeln:

$$\theta = \frac{M}{N} .$$

Noch eine Bemerkung: Stellt man sich allerdings bei der Berechnung dieser Binomialkoeffizienten ein bisschen dumm an, protestiert die Software, weil man einen Überlauf erhält. Man kann allerdings hier mit der Stirling-Formel noch etwas ausrichten. Oder man logarithmiert die Fakultäten.

Für sehr kleines θ (oder sehr kleines 1-θ) und sehr großes n ist die Binomialverteilung wiederum annähernd Poisson-verteilt. Es ist nämlich die Poissonverteilung die Grenzverteilung der Binomialverteilung für n $\to \infty$ und $\theta \to$ 0. Die Berechnung der Poissonverteilung ist einfacher als die Berechnung der Binomialverteilung. Eine Faustregel wäre hier etwa, dass eine binomialverteilte Zufallsvariable durch die Poisson-Verteilung angenähert werden kann, wenn $\theta \leq$ 0,05 und n $\geq$ 50 ist. Dann ist

$$b(x|N;M;n) \approx p(x|n \cdot \theta) .$$

Über den Umweg der Binomialverteilung kann dann auch die hypergeometrische Verteilung gegebenenfalls mit der Poisson-Verteilung approximiert werden:

$$h(x|N;M;n) \approx p(x|n;\frac{M}{N}), \text{ wenn } \frac{n}{N} \leq 0,05, \theta \leq 0,05 \text{ und } n \geq 50$$

ist.

Weiter unten folgt eine tabellarische Zusammenfassung ausgewählter Approximationen.

Approximation diskreter Verteilungen durch die Normalverteilung

Was ist nun aber, wenn wir wissen wollen, wie groß die Wahrscheinlichkeit ist, dass höchstens 15 defekte Chips gefunden werden: P(X $\leq$ 15)? Hier müssen wir auf die oben beschriebene Weise 15 Wahrscheinlichkeiten ermitteln und addieren. Spätestens hier wünscht man sich eine Möglichkeit, so etwas schneller errechnen zu können. Es wäre doch angesagt, wenn man da die Normalverteilung verwenden könnte.

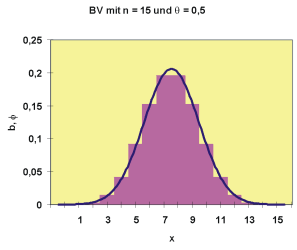

Abbildung 54: Binomialverteilung mit n = 15 und θ = 0,5 und darübergelegte Normalverteilungsdichte

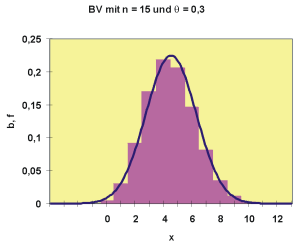

Abbildung 55: Binomialverteilung mit n = 15 und θ = 0,3 und darübergelegte Normalverteilungsdichte

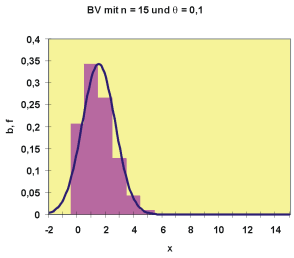

Abbildung 56: Binomialverteilung mit n = 15 und θ = 0,1 und darübergelegte Normalverteilungsdichte

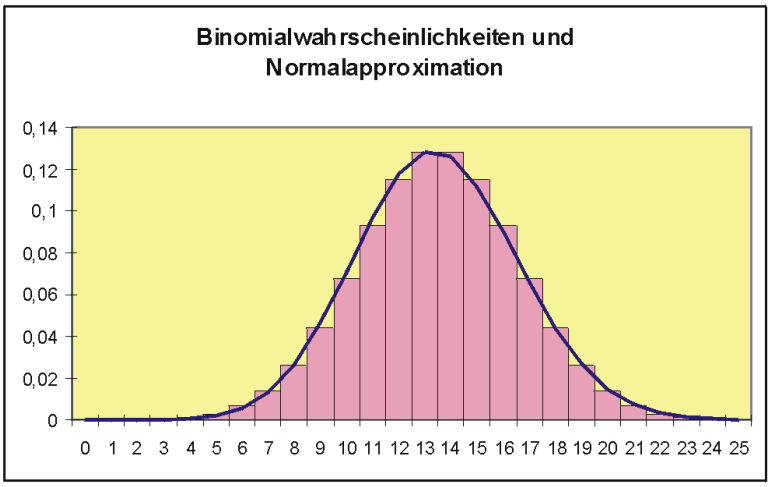

Abbildung 57: Binomialverteilung mit n = 45 und $\theta = 0{,}3$ und darübergelegte Normalverteilungsdichte

Vergleichen wir die Grafiken von den Binomialverteilungen. Es wurden hier die Wahrscheinlichkeiten als benachbarte Säulen dargestellt, was ja am optischen Erklärungswert nichts ändert.

Wir können deutlich erkennen, dass die Binomialverteilung für $\theta = 0{,}5$ symmetrisch ist. Hier passt sich die Normalverteilung am besten an. Je weiter θ von 0,5 abweicht, desto schlechter ist die Anpassung der Normalverteilung. Die so gut wie immer verwendete Faustregel ist, dass man mit der Normalverteilung approximieren darf, wenn

$$n > \frac{9}{\theta(1-\theta)}$$

ist. Dürfen heißt natürlich nicht, dass es sonst polizeilich verboten ist, sondern dass sonst die Anpassung unbefriedigend ist.

Eine Normalverteilung hat den Erwartungswert μ und die Varianz σ^2. Wie soll man diese Parameter bei der Approximation ermitteln? Nun wissen wir ja, dass der Erwartungswert der Binomialverteilung und ihre Varianz

$$EX = n\theta$$

und $varx = n\theta(1-\theta)$

sind, also nehmen wir doch einfach diese Parameter für die Normalverteilung, also

$$\mu = n\theta$$

und $\sigma^2 = n\theta(1-\theta)$.

Etwas fehlt uns noch: Wir nähern hier eine diskrete Verteilung durch eine stetige Verteilung an. Diskrete und stetige Verteilungen sind zwei völlig unterschiedliche Konzepte. Wir betrachten hier das Beispiel einer Binomialverteilung mit n = 45 und θ = 0,3.

Nähern wir P(X $\leq$ 12) = B(12|45;0,3) durch Φ(12|45·0,3; 45·0,3·0,7) an, wird nur die halbe Säule addiert, denn die stetige Verteilung kennt keine Säulen. Soll die ganze Säule einbezogen werden, müssen wir bis 12,5 gehen, also P(X $\leq$ 12) = B(12|45;0,3) durch Φ(12,5|45·0,3; 45·0,3·0,7).

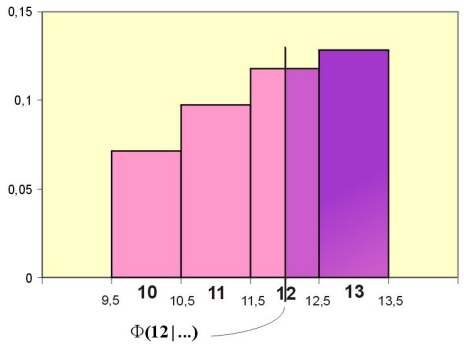

Abbildung 58: Wenn man mit der Normalverteilung P(X $\leq$ 12) berechnet, wird nur die halbe Säule addiert

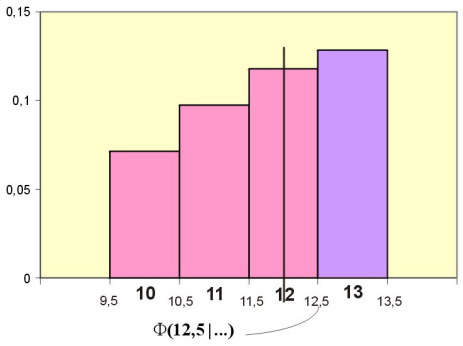

Abbildung 59: Wenn man mit der Normalverteilung P(X $\leq$ 12,5) berechnet, wird die ganze Säule addiert

Den addierten Wert 0,5 nennt man **Stetigkeitskorrektur**.

Speziell gilt für die Wahrscheinlichkeit **P(X = a)**:

P(X = a) = b(a|n;θ) $\approx$ Φ(a+0,5|nθ; nθ(1-θ)) - Φ(a -0,5|nθ; nθ(1-θ)).

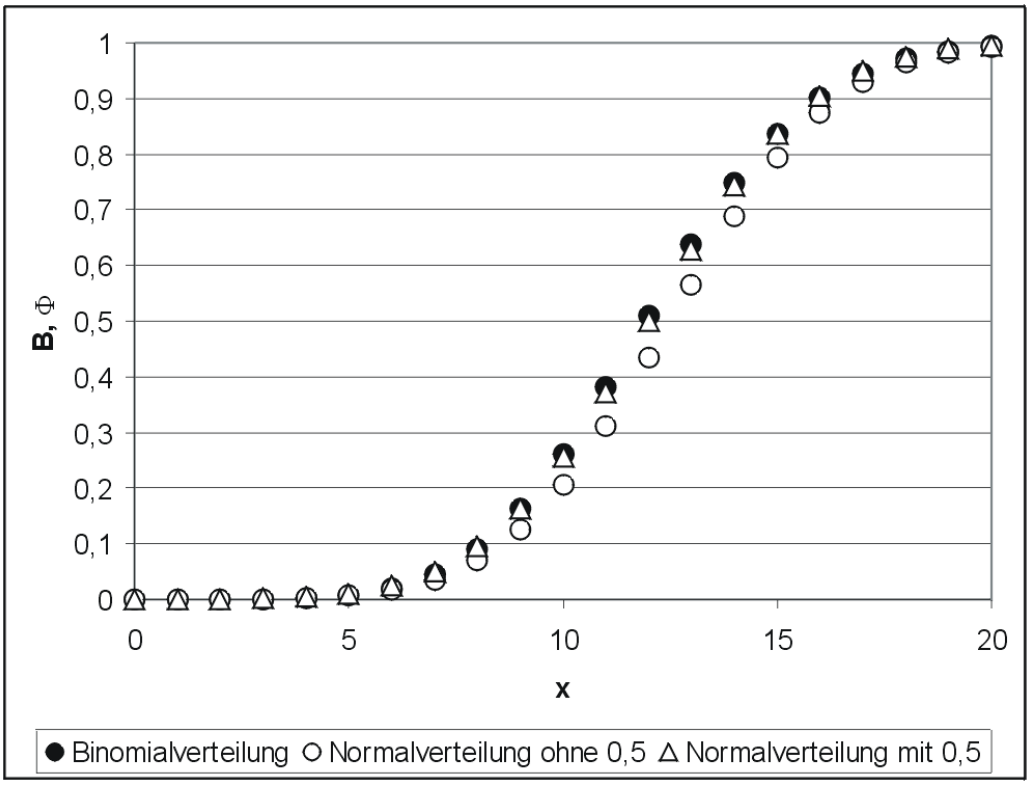

Abbildung 60

Approximation stetiger Verteilungen durch die Normalverteilung

Jetzt haben wir also auch noch stetige Funktionen, die wir mit der Normalverteilung annähern wollen. Was gibt es denn da für welche? Nun, welche die man oft braucht, etwa für Schätzen und Testen, als da wären die χ^2-Verteilung, die F-Verteilung und die t-Verteilung.

Nehmen wir uns doch mal die χ^2-Verteilung vor. Ein Blick auf ihre Dichtefunktion verrät, dass diese mit wachsendem n immer symmetrischer wird, sich also der Normalverteilung annähert. Wir wissen, dass die χ^2-Verteilung eine Summe von Zufallsvariablen, nämlich standardnormalverteilten, quadrierten, ist und wir erinnern uns (gell?), dass nach dem zentralen Grenzwertsatz sich die Verteilung einer Summe von Zufallsvariablen der Normalverteilung annähert. Betrachten wir die mit n Freiheitsgraden χ^2-verteilte Zufallsvariable X. Wir bilden eine neue Zufallsvariable

$$Y = \sqrt{2X}$$

Eine gängige Faustregel besagt für die Approximation für die Wahrscheinlichkeit P(Y $\leq$ y):

$$P(Y \leq y) \approx \Phi(y|\sqrt{2n-1}; 1) \ .$$

Die Dichtefunktion t-Verteilung dagegen hat eine ähnliche Form wie die Standardnormalverteilung, denn auch sie ist symmetrisch bezüglich der Null. Hier genügt eine einfache Faustregel: Wenn n $>$ 30 ist, kann man die Verteilungswerte der t-Verteilung annähernd mit Hilfe der Standardnormalverteilung bestimmen:

$t(x|n) \approx \Phi(x|0; 1)$.

Tabelle der Approximationen

Gesuchte Verteilung	**Approximation durch**		
$P(X \leq x)$	Binomial	Poisson	Normal
Binomial $B(x\|n\theta) \approx$	—	$P(x\|n\theta)$ falls $n \geq 50$ und $\theta \leq 0,05$	$\Phi(x + 0,5\|n \cdot \theta; n \cdot \theta \cdot (1-\theta))$ falls $n > \frac{9}{\theta(1-\theta)}$
Hypergeometrische $H(x\|N;M;n) \approx$	$B(x\|n\frac{M}{N})$ falls $\frac{n}{N} < 0,05$	über Binomialverteilung	$\Phi(x + 0,5\|n \cdot \frac{M}{N}; n \cdot \frac{M}{N} \cdot (1 - \frac{M}{N}) \cdot \frac{N-n}{N-1}$ falls $n > \frac{9}{\frac{M}{N} \cdot (1-\frac{M}{N})}$ und $\frac{n}{N} < 0,05$
Poisson $P(x\|\lambda) \approx$	—	—	$\Phi(x + 0,5\|\lambda;\lambda)$ falls $\lambda > 9$
χ^2-Verteilung $\chi^2(x\|n) \rightarrow$ $P(\sqrt{2X} \leq \sqrt{2x}) \approx$	—	—	$\Phi(\sqrt{2x}\|\sqrt{2n-1};1)$ falls $n > 30$
t-Verteilung $t(x\|n) \approx$	—	—	$\Phi(x\|0;1)$ falls $n > 30$
F-Verteilung $F(x\|m;n) \approx$	—	—	$\Phi(x\|0;1)$ falls $m > 30$ und $n > 30$

Kapitel 5

Deskriptive Statistik

Einführung

Die Verfahren der deskriptiven Statistik (beschreibende Statistik, empirische Statistik) haben als Grundlage die Erhebung bzw. Beobachtung von Daten. Es geht hier darum, diese Daten in geeigneter Weise zusammenzufassen, sie zu ordnen, sie grafisch darzustellen usw. Ziele der deskriptiven Statistik:

1. Die Daten einer empirischen Untersuchung möglichst übersichtlich zu präsentieren, so dass die wesentlichen Informationen schnell und optimal aufgenommen werden können. Beispiele: Tabellen, Säulendiagramme, Durchschnitte, Prognosen etc. Auf eine verteilungstheoretische Analyse wird verzichtet.
2. Man interessiert sich für die unbekannte Verteilung eines statistischen Merkmals, für Kennwerte der Verteilung usw. Da eine vollständige Erfassung dieses Merkmals meist zu teuer oder auch unmöglich ist, wird man sich auf eine Teilerhebung, eine Stichprobe, beschränken. Man schätzt nun mit Hilfe dieser Stichprobe die gesuchten Werte. Dabei versucht man, die Wahrscheinlichkeit einer Fehlschätzung miteinzubeziehen.

Analyse eines Merkmals

Die Analyse des Merkmals hängt u.a. davon ab, welche Informationen man wünscht:

- Verteilung: Ist sie symmetrisch oder schief, ein- oder mehrgipflig?
- Niveau der Daten, z.B. Durchschnitt, Median?
- Streuung der Einzelwerte: hoch oder niedrig?
- Sind mehrere Merkmale korreliert?

Definitionen in der deskriptiven Statistik

Beispiel:

Es wurden n = 7 Hunde befragt, wie gut ihnen das neue Fröhlix-Trockenfutter schmecke. Die Eingabe der Fragebögen in eine Datei ergab die unten folgende Liste. Anhand dieser Liste sollen Begriffe der deskriptiven Statistik erklärt werden.

Die Eigenschaften, die erhoben werden, sind die **Merkmale** (**statistische Variablen**) x, y, ... Das Objekt, dessen Eigenschaften erhoben (erfragt, gemessen) werden, ist die **Untersuchungseinheit** (**Merkmalsträger**). Die Menge aller statistischen Einheiten ist die **Grundgesamtheit** (statistische Masse). Die möglichen Werte, die eine Variable annehmen kann, sind die **Ausprägungen** (**Realisationen**). Die konkrete Merkmalsausprägung von x, die eine Untersuchungseinheit Nummer i aufweist, ist der **Merkmalswert** (**Beobachtungswert, Beobachtung**) x_i (i=1,2, ..., n).

Name	Geschlecht *Merkmal* 1=w, 2=m u	Rasse x	Alter *Merkmal* y	Note für Futter 1, ..., 5 *Ausprägungen* z
Rex ***Merkmalsträger***	2	Schäferhund	3	1
Rexona	1	Mischling	5	4 ***Merkmalswert***
Lassie	1	Collie	1	2
Hasso	2	Neufundländer	2	1
Strolchi ***Merkmalsträger***	2	Schnauzer	7	2

Susi	1	Spaniel	2	3
Waldi	2	Dackel	1 ***Merk-malswert***	5

Es sind die Ausprägungen des Merkmals

Note: 1,2,3,4,5

und die Ausprägungen des Merkmals

Geschlecht: 1,2.

Skalierung des Merkmals

Beispiel

Grundlage des Beispiels ist die Hundetabelle von oben. Der Student Paul leistet beim Hersteller von Fröhlix ein Praktikum ab. Er soll die Ergebnisse der Befragung präsentieren. Er fasst die Hundetabelle von oben zusammen und erhält u.a.

Durchschnittliches Alter eines Hundes:

$$\frac{1}{7}(3+5+1+2+7+2+1) = \frac{21}{7} = 3.$$

Ein befragter Hund war also im Durchschnitt 3 Jahre alt.

Durchschnittliches Geschlecht eines Hundes:

$$\frac{1}{7}(2+1+1+2+2+1+2) = \frac{11}{7} = 1,57.$$

Ein Hund hatte also im Durchschnitt 1,57 Geschlecht. ????? Würden Sie den Studenten Paul später in dieser Firma einstellen?

Es ist natürlich höherer Schwachsinn, vom Merkmal Geschlecht den Durchschnitt zu bilden. Man kann damit keinen Durchschnitt bilden, denn seine

Ausprägungen sind keine Zahlen. Geschlecht ist ein **qualitatives** Merkmal. Es ist anders **skaliert** als Alter.

Es gibt also Merkmale mit unterschiedlichen Messbarkeitsarten. Die Vorschrift für die Messbarkeit ist in einer **Skala** festgehalten.

Nominalskala

Merkmale wie

- Haarfarbe: braun, blond, ...;
- berufstätig ja/nein;
- Margarinemarke: Panorama, Botterama, ...

sind nominalsskaliert. Die Ausprägungen des nominalskalierten Merkmals können nicht **geordnet** werden, man kann sie nur **vergleichen** und **abzählen**. Es handelt sich um **qualitative** Merkmale. Erhalten die Ausprägungen Ziffern zugeordnet, handelt es sich nur um eine **Verschlüsselung** (Codierung): 1 = männlich, 2 = weiblich.

Ordinalskala

Zwischen den Ausprägungen des ordinalskalierten (rangskalierten) Merkmals existiert eine Beziehung der Form mehr oder weniger, , besser oder schlechter o.ä., also eine Art **natürlicher Reihenfolge**.

Beispiele

- Sterne eines Hotels: *; **; ***; ...
- Beurteilung eines Produktes durch einen Konsumenten: Sehr gut, eher gut, eher schlecht, ganz schlecht
- Noten: 1, 2, 3, 4, 5

Für die Ausprägungen läßt sich also eine **Rangordnung** feststellen, aber die Abstände zwischen den Rängen sind nicht festgelegt. So ist beispielsweise die Note Vier nicht doppelt so schlecht wie Zwei.

Metrische Skala

Die Abstände zwischen den Ausprägungen des metrisch skalierten (quantitativen) Merkmals können **gemessen** werden. Es handelt sich bei den Ausprägungen um **(reelle) Zahlen**.

Beispiele: Kinderzahl, Einkommen, Temperatur, ...

Die metrischen Variablen werden noch in diskret und stetig unterschieden:

Ein Merkmal ist **diskret** (=unterschieden), wenn man die Ausprägungen abzählen kann.

Beispiel

- Kinderzahl: 0, 1, 2, ... , 20.
- Mein ‖Einkommen‖, wenn ich falsch parke: 3 Euro (gesparte Parkgebühr) oder -10 Euro (Strafzettel).

Es gibt auch **abzählbar unendlich** viele Ausprägungen:

- Zahl der Ausfälle einer Maschine in einem Jahr: 0, 1, 2, ...

Ein Merkmal ist **stetig** (kontinuierlich), wenn sich in einem beschränkten Intervall der reellen Zahlen unendlich viele Ausprägungen (**überabzählbar viele**) befinden.

Beispiele: Wasserstand in einem Stausee; Gewicht eines Goldstücks; Temperatur; Körpergröße.

Bemerkung: Oft sind Merkmale eigentlich diskret, aber mit sehr vielen, nah beieinanderliegenden Ausprägungen, z.B. Einwohnerzahl, Preise (in Cents), usw. Hier definiert man das Merkmal zweckmäßigerweise als stetig, da es sich so besser analysieren läßt (**quasistetig**).

Übung

Wie sind die folgenden Merkmale skaliert?

- Täglicher Bierkonsum der Studentin Paula
  - - in Flaschen
  - - in Litern

Bekenntnis: 1= röm.-kath., 2 = evang., 3 = sonst

- Gewicht von Bernhardinern
- Aufgabe: schwer - leicht
- Zahl der zustehenden Urlaubstage
- Jeansmarke

Behandelt wird hier ein metrisch skaliertes Merkmal, von dem nur wenige verschiedene Beobachtungen vorliegen, beispielsweise das Gewicht von 10 Schlachthähnchen oder die abgefüllte Menge von Kakao in 6 ||250||-g Päckchen. Diese Konstellation wurde deshalb eigens hervorgehoben, weil sich damit viele Methoden der deskriptiven Statistik einfach erklären lassen.

Urliste

Beispiel

n = 10 ||Pfundschalen|| Erdbeeren wurden nachgewogen. Es ergab sich durch Nachwiegen die **Urliste**

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
480	500	510	450	400	490	505	510	480	480

mit dem Merkmal x: Gewicht eines Schälchens (g). Die Werte wurden in der Reihenfolge der Erhebung, also ungeordnet, als Urliste erfasst. Diese Art der Darstellung ist unübersichtlich und daher nur für wenige Beobachtungen geeignet.

Urlisten können auch mehrere, unterschiedlich skalierte Merkmale enthalten. Beispielsweise ist die Tabelle mit den Hunden eine Urliste.

Häufigkeitsverteilung

Liegt ein metrisch skaliertes Merkmal oder ein ordinalskaliertes Merkmal mit vielen Ausprägungen vor, kann man zunächst einmal die Urliste der Größe nach ordnen, um einen gewissen Eindruck zu erhalten.

Beispiel

Die Indizes in den eckigen Klammern bedeuten, dass die Beobachtungen der Größe nach geordnet wurden.

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
400	450	480	480	480	490	500	505	510	510

Man erkennt nun, dass über die Hälfte der Schälchen untergewichtig waren.

Allerdings ist das Sortieren mühsam, fehleranfällig und doch nicht sehr informativ. Mit dem **Zweig-Blätter-Diagramm** (**stem-and-leaf display**) kann man jedoch sowohl metrische Beobachtungen relativ leicht sortieren als auch eine erste Häufigkeitsverteilung erzeugen.

Zweig-Blätter-Diagramm

Beispiel:

Für das Jahr 2003 liegt das reale Wachstum des Bruttoinlandsprodukts für 38 europäische Staaten vor (© Statistisches Bundesamt, Wiesbaden 200 http://www.destatis.de/allg/d/impr/d_impr.htm)

> 4,7 1,1 3,9 -0,1 4,7 1,8 0,2 4,8 1,4 1,9 0,3 5,2 7,4 9,0 2,6 0,4 0,7
> 7,2 -0,8 0,3 0,7 3,7 -1,3 4,9 7,3 1,6 -0,5 4,0 4,2 2,3 2,4 2,9 5,8 4,8
> 2,9 2,1 4,7 2,0

Wir wollen die Daten ordnen und einen Eindruck von der Häufigkeitsverteilung gewinnen. Dazu werden wir jetzt ein Zweig-Blätter-Diagramm oder, für Anglophile, ein Stem-and-Leaf-Display erzeugen.

Zuerst wird der Zweig gemacht - aus den Einsern:

Dann hängen wir die Blätter an den Zweig, und zwar, indem wir von links nach rechts durch die Daten wandern: Der erste Wert ist 4,7. Das Blatt 7 wird an den Zweig 4 gehängt

−1	
−0	
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

Der zweite Wert ist 1,1, das Blatt 1 wird an die 1 gehängt

−1	
−0	
0	
1	1
2	
3	
4	7
5	
6	
7	
8	
9	

Schließlich erhalten wir

−1	
−0	
0	
1	
2	
3	
4	7
5	
6	
7	
8	
9	

Es folgen 3,9 -0,1 4,7 1,8 ...

−1	
−0	1
0	
1	18
2	
3	9
4	77
5	
6	
7	
8	
9	

Diese Prozedur war schnell erledigt. Wir bekommen schon einen guten Eindruck von der Verteilung der Beobachtungswerte. Kippen wir das Diagramm um 90°, erkennen wir eine Art Säulendiagramm. Außerdem können wir nun die Werte schnell der Größe nach sortieren. Wir erhalten nun unser Stengel-Blätter-Diagramm:

-1	3	-1	3
-0	185	-0	851
0	234737	0	233477
1	18496	1	14689
2	6349910	2	0134699
3	97	3	79
4	77890287	4	02777889
5	28	5	28
6		6	
7	423	7	234
8		8	
9	0	9	0

Für unsere Zwecke ist das Stem-and-Leaf-Display jetzt ausreichend. Ein Stem-and-Leaf-Display kann aber auch noch mehr Einzelheiten enthalten. Ist die Zahl der erhobenen Beobachtungen sehr groß, können die Werte in Klassen tabellarisch zusammengefaßt werden. Diese Art der Analyse erfolgt weiter unten.

Summenfunktion

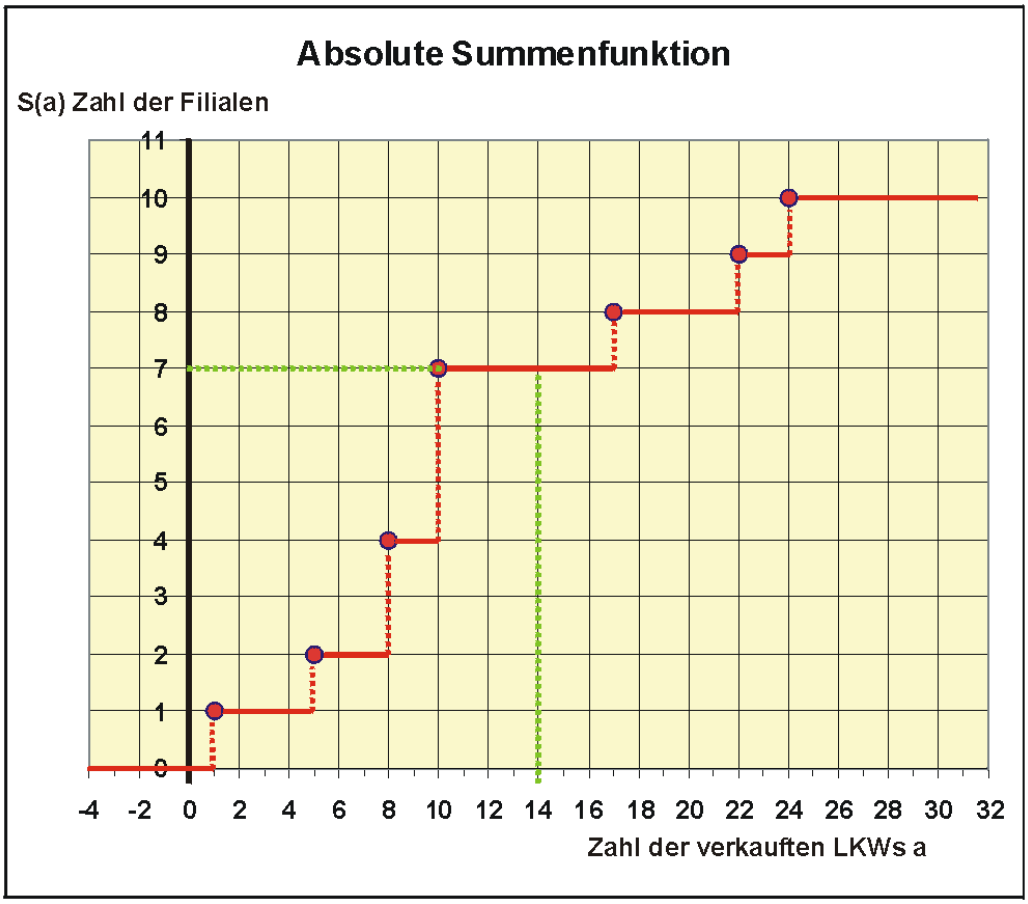

Abbildung 61: absolute Summenfunktion

Beispiel

Ein Autohaus hat von seinen n = 10 Filialen die Zahl der verkauften LKWs des letzten Jahres vorliegen. Es folgt die Urliste mit den x_i geordnet:

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
1	5	8	8	10	10	10	17	22	24

Wir wollen die absolute Summenfunktion S(a) bestimmen. S(a) gibt an, wieviel Beobachtungen $x_i \leq a$ sind:

Zum Beispiel:

- S(17) = 8, also sind 8 Beobachtungen höchstens 17

- S(8) = 4, also gibt es 4 Filialen, die höchstens 8 LKWs verkauft haben

Wir leiten nun die Summenfunktion her, von links nach rechts:

- Zum Beispiel: S(0,1) = 0, denn keine Filiale hat höchstens 0,1 LKW verkauft. Ebenso ist S(0,9) = 0, usw... also

S(a) = 0 für a $<$ 1.

- Zum Beispiel: S(1) = 1, denn genau eine Filiale hat höchstens einen LKW verkauft. Ebenso ist S(3) = 1, denn es hat auch eine Filiale höchstens drei LKWs verkauft. Ebenso S(4,9999) = 1 ..., also

S(a) = 1 für 1 $\leq$ a $<$ 5.

- Zum Beispiel:S(5) = 2, also

S(a) = 2 für 5 $\leq$ a $<$ 8.

usw... schließlich erhalten wir

S(a) = 10 für a $\geq$ 24.

Tragen wir die ermittelten Funktionswerte in die Grafik ein, sehen wir sofort, dass wir eine Treppenfunktion erhalten.

Die absolute Summenfunktion S(a) ist die **Zahl** der Beobachtungen xi $\leq$ a . Die relative Summenfunktion gibt stattdessen die **Anteile** der Beobachtungen an der Urliste an:

$$S^*(a) = \frac{S(a)}{n}$$

Der Informationswert der kumulierten Häufigkeit S(n) in der Grafik erschließt sich Ungeübten eher weniger. Aber man kann anhand der Grafik sofort Aussagen über die Verteilung machen. Man sieht beispielsweise sofort, daß z.B. 7 Werte kleiner als 14 sind, es haben also 70% der Filialen höchstens 14 LKWs verkauft.

Lageparameter

Der Lageparameter gibt an, auf welchem Niveau die Daten liegen.

Arithmetisches Mittel

Das arithmetische Mittel ist landläufig als ||Durchschnitt|| bekannt. Es ist eigentlich nur für metrisch skalierte Merkmale (Problem Notendurchschnitt) geeignet. Es berechnet sich als

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

Beispiel Pfundschalen Erdbeeren:

$$\begin{aligned} \bar{x} &= \tfrac{1}{10}(400 + 450 + 480 + 480 + 480 + 490 + 500 + 505 + 510 + 510) \\ &= \tfrac{4805}{10} = 480,5 \end{aligned}$$

.

Es waren die Schälchen also im Durchschnitt untergewichtig.

Median oder Zentralwert

Sind die Beobachtungswerte der Größe nach geordnet, also $x_{[1]}$, $x_{[2]}$, $x_{[3]}$, ... , $x_{[n]}$, ist der Median z die Stelle, die die Teilgesamtheit in zwei gleiche Hälften teilt. Er kann für **rang-** und **metrisch** skalierte Merkmale verwendet werden.

n ungerade

Beispiel für n = 7

Es wurden 7 Autofahrer nach ihren Fahrtkosten befragt. Es ergab sich für das Merkmal x: Monatliche Ausgaben für Benzin (in Euro) die Liste

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$
20	50	100	**170**	200	200	280

Es ist also der Median z = 170.

n gerade

Beispiel für n = 10 (Erdbeeren)

$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$		$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
400	450	480	480	480	**z**	490	500	505	510	510

Der Median liegt zwischen dem 5. und 6. Beobachtungswert. Man nimmt hier den mittleren Wert

$$z = \frac{1}{2}(480 + 490) = 485.$$

Wir berechnen also den Median so:

n ungerade: z ist der $\frac{n+1}{2}$te Wert $x_{[i]}$, also

$$z = x_{[\frac{n+1}{2}]}$$

n gerade: z liegt zwischen dem $\frac{n}{2}$ten und dem $\frac{n}{2} + 1$ten Beobachtungswert $x_{[i]}$, also

$$z = \frac{1}{2}(x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]})$$

Bemerkungen:

- Der Median kann für ordinal- und metrisch skalierte Werte verwendet werden.
- Bei **sehr großem und geradem n** kann man vereinfachend

  $$z = x_{[\frac{n}{2}]}$$

  setzen.

Vergleich Median - arithmetisches Mittel

Beispiel:

Eine Autozeitschrift hat n = 7 PKWs einer bestimmten Marke getestet. Unter anderem wurde auch untersucht, ob das Auto zuverlässig anspringt.

Es ergab sich die geordnete Urliste

1 1 1 1 1 2 14

Wir erhalten als durchschnittliche Zahl der Startversuche

$$\bar{x} = \frac{1}{7}(1+1+1+1+1+2+14) = \frac{21}{7} = 3$$

.

Wir würden hier also als Ergebnis erhalten: ||Ein PKW sprang im Durchschnitt erst nach 3 Versuchen an||. Irgendwie erscheint einem das nicht gerechtfertigt. Bis auf einen PKW, der offensichtlich einen Ausreißer darstellt, sprangen ja alle Fahrzeuge zuverlässig an.

Wir verwenden nun den Median als Lageparameter: Der Median ist der 4. Wert, also z = 1. Hier ist also der Median eher zutreffend, doch so ganz zufrieden sind wir auch nicht, denn immerhin gab es ja auch 2 und 14 Versuche.

Wir sehen also, dass bei Verwendung des Median sehr viel Information der Daten verloren geht, andererseits reagiert aber das arithmetische Mittel empfindlich auf Ausreißer in den Daten.

Es gibt aber auch Kompromisse zwischen diesen beiden Extremen, beispielsweise das getrimmte Mittel:

$$x_T = \frac{1+1+1+1+2}{5} = \frac{6}{5} = 1,2$$

.

Es werden in der geordneten Urliste links und rechts jeweils ein oder mehrere Werte gestrichen. Aus den restlichen Beobachtungen berechnet man dann ein arithmetisches Mittel. Dieser Mittelwert erscheint eher die Sachlage zutreffend zu beschreiben. Man nennt Parameter, die nur schwach auf Ausreißer reagieren, resistente Parameter. Neben dem getrimmten Mittel gibt es noch mehrere andere Ansätze.

Der Vergleich des Medians mit dem arithmetischen Mittel kann als Ausreißeranalyse verwendet werden. Weicht der Median auffällig vom arithmetischen

Mittel ab, sollten die Daten auf Ausreißer oder stark schiefe Verteilungen hin überprüft werden.

Weitere Lageparameter sind etwa der Modalwert, geometrisches Mittel oder harmonisches Mittel.

Varianz als Streuungsparameter

Der Lageparameter allein reicht für die Beschreibung einer Datenmenge nicht aus (analoges Problem wie bei Zufallsverteilungen). Information über die **Streuung** der Beobachtungswerte liefert ein **Streuungsparameter**. Es kommen verschiedene Kennwerte als Streuungsparameter in Betracht, beispielsweise die Varianz, die Spannweite, der Quartilsabstand und der Variationskoeffizient.

Varianz

Am häufigsten wird als Kennwert die Varianz verwendet, da sie wahrscheinlichkeitstheoretisch am besten zu untersuchen ist. Die Varianz sind die mittleren quadratischen Abweichungen der Einzelwerte x_i vom arithmetischen Mittel

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

Der Nenner n-1 wirkt vielleicht etwas befremdlich. Allerdings hat die Verwendung von n-1 statt n wahrscheinlichkeitstheoretische Vorzüge, wenn man die Varianz der Verteilung eines Merkmals mit s^2 schätzen möchte. Man nennt dieses Art der Varianz inferentielle Varianz.

Beispiel

Eine Firma möchte einen Kachelofen auf den Markt bringen, der für einen Komplettpreis zu erwerben ist. Für die Kalkulation dieses Preises benötigt die Firma Informationen über die Montagezeit für einen Kachelofen. Bei der Endmontage von 11 Kachelöfen ergaben sich die Zeiten

2,5 3 3 3,3 3,6 3 2,3 3 3,1 3,2 3

Die Varianz der Montagezeiten soll bestimmt werden. Nach der obigen Formel muss zunächst das arithmetische Mittel bestimmt werden:

$$\begin{aligned}\bar{x} &= \tfrac{1}{11}(2,5+3+3+3,3+3,6+3+2,3+3+3,1+3,2+3)\\ &= \tfrac{33}{11} = 3h\end{aligned}$$

.

Dann erhalten wir als Varianz

$$s^2 = \frac{1}{10}((2,5-3)^2+(3-3)^2+(3-3)^2+...+(3-3)^2)$$

$$\begin{aligned}&= \tfrac{1}{10}(0,25+0+0+0,09+0,36+0+0,49+0+0,01+0,04+0)\\ &= \tfrac{1,24}{10} = 0,124h^2\end{aligned}$$

.

Verzichtet man auf eine Schätzung, kann man auch die deskriptive Varianz

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

für die Beschreibung von statistischen Daten verwenden, was aber hier zur Vermeidung von Verwechslungen unterlassen wird.

Bei der manuellen Berechnung von s^2 ist es oftmals mühsam, erst die einzelnen Differenzen x_i - x zu bilden und dann zu quadrieren. Mit Hilfe des **Verschiebungssatzes** kann die laufende Differenzenbildung vermieden werden. Betrachten wir die Summe

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

.

Diese Summe lässt sich zerlegen in

$$Q = \sum_{i=1}^{n}x_i^2 - n\cdot\bar{x}^2$$

.

Setzt man den Ausdruck oben ein, erhält man für die Varianz

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

Beispiel:

$$\begin{aligned} s^2 &= \tfrac{1}{10}(2,5^2 + 3^2 + 3^2 + ... + 3^2 - 11 \cdot 3^2) \\ &= \tfrac{1}{10} \cdot (100,24 - 99) = 0,124h^2 \end{aligned}$$

.

Da die Varianz ein quadratischer Ausdruck ist, hat sie z.B. auch die Einheit h^2, wenn die x_i die Einheit h haben. Um die Varianz anschaulicher zu machen, kann man ihre Quadratwurzel, die Standardabweichung s betrachten:

Beispiel

$$s = \sqrt{0,124h^2} \approx 0,35h$$

,

also ca. 20 Minuten. Man könnte etwas flapsig sagen, dass die Montagezeit eines Ofens im Mittel 3 Stunden +/- 20 Minuten beträgt.

Auch die Varianz reagiert empfindlich auf Ausreißer. Es gibt hier resistente Streuungsparameter, die weiter unten behandelt werden.

In den letzten Abschnitten lernten wir, wie man Daten eines stetigen, metrischen Merkmals, die als Urliste vorlagen, analysiert. Wir wollen nun Daten untersuchen, die man in Häufigkeitstabellen zusammenfassen kann. Im Gegensatz zur obigen Urliste können hier die Daten übersichtlich grafisch dargestellt werden. Man unterscheidet im Wesentlichen Daten eines metrischen Merkmals mit wenigen verschiedenen Ausprägungen und große Mengen von Daten mit vielen verschiedenen Ausprägungen, die man in Klassen zusammenfasst.

Zu den **Merkmalen mit wenig verschiedenen Ausprägungen** gehören **nominal** skalierte, **ordinal** skalierte und **metrisch** skalierte Merkmale. Da sie nur **wenig Kategorien** haben, kann man sie in **Häufigkeitstabellen** zusammenfassen. Man nennt sie **häufbare Merkmale**.

Beispiele für Merkmale mit wenigen möglichen Ausprägungen:

- nominal skaliert: Augenfarbe von Studierenden
- ordinal skaliert: Note der Kundenzufriedenheit
- metrisch skaliert: Zahl der Autos in einem Haushalt

Bemerkung: Metrisch skalierte stetige Merkmale sind nicht unmittelbar häufbar, weil zu viele verschiedene Beobachtungen vorliegen.

Wenn man vorliegende Daten analysiert, wird man sich zunächst für die Verteilung des Merkmals interessieren:

Ist die Verteilung der Variablen einigermaßen symmetrisch oder stark schief? Sind Ausreißer in den Daten? Ist die Verteilung eingipflig oder mehrgipflig? Der Statistiker freut sich meistens über eine symmetrische Verteilung, weil man hier häufig die Vorteile der Normalverteilung ausnützen kann.

Werkzeuge für die Analyse sind hierbei die Häufigkeitstabelle, die Summenfunktion und diverse Grafiken, denn bei einem Merkmal mit wenig Ausprägungen können attraktive Diagramme erstellt werden.

Häufigkeitstabelle

Um eine Urliste von Beobachtungen eines Merkmals mit wenig Ausprägungen aufzubereiten, wird als erster Schritt der Analyse das Zählen des Auftretens der Realisationen stehen. Die Ergebnisse können in einer **Häufigkeitstabelle** zusammengefasst werden. Anhand der Daten eines nominalskalierten Beispiels wollen wir uns das Prinzip einer Häufigkeitstabelle ansehen.

Nominalskaliertes Merkmal

Beispiel

Es wurden 50 Personen telefonisch bezüglich gewisser Konsumpräferenzen befragt. Unter anderem erhob man den Familienstand. Es ist das Merkmal

> x: Familienstand - mit den Ausprägungen 1=ledig, 2=verheiratet, 3=geschieden, 4=verwitwet.

Es ergab sich die Urliste

2 2 1 2 3 3 1 2 3 2 3 4 4 1 2 1 1 2 3 2 1 2 2 1 2 2 2 1 4 2 2 4 3 1 2 2 1 3 2 3 1 2 2 3 2 2 2 1 3 3

Wir wollen nun die Daten in einer Häufigkeitstabelle zusammenstellen:

j	**Familienstand**	**absolute Häufigkeit**	**relative Häufigkeit**
1	ledig	12	0,24
2	verheiratet	23	0,46
3	geschieden	11	0,22
4	verwitwet	4	0,08
Σ		50	1,00

Es sind insgesamt n = 50 Untersuchungseinheiten erhoben worden. Die (absoluten) Häufigkeiten n_j (j = 1, ... , 4) verteilen sich auf m = 4 Kategorien (kategoriale Variable), wie in der Häufigkeitstabelle aufgelistet.

Wenn man sich für den Anteil der einzelnen Ausprägungen an der Gesamtheit interessiert, kann man auch die relativen Häufigkeiten bestimmen:

$$p_j = \frac{n_j}{n}$$

Es ist natürlich

$$\sum_{j=1}^{m} n_j = n$$

bzw. $\sum_{j=1}^{m} p_j = 1$

Für die Verteilung von Merkmalen mit wenig Ausprägungen kann man sehr ansprechende Grafiken erstellen.

Ordinalskaliertes Merkmal

Beispiel:

Bei der letzten Wiki-Matheklausur der Wikiversity ergaben sich die Noten wie folgt:

12 x 1, 15 x 2, 8 x 3, 3 x 4, 2 x 5

Hier erhält man die unten folgende Häufigkeitstabelle:

j	**Note** x_j	**absolute Häufigkeit** n_j	**relative Häufigkeit** p_j
1	sehr gut	12	12/40=0,3
2	gut	15	0,375
3	befriedigend	8	0,2
4	ausreichend	3	0,075
5	ungenügend	2	0,05
Σ		40	1

Auch hier bieten sich zur Veranschaulichung der Häufigkeiten Grafiken wie oben an.

Metrisch skaliertes Merkmal

Beispiel

Eine mainfränkische Weinbaustadt feiert ihr alljährliches Weinfest, bei dem auch die Winzerei Wavoma ihre Produkte anbietet. Sie verkauft Wein in Flaschen mit 0,5, 0,7, 1 und 2 Litern zu je 4, 5, 7 und 10 Euro. Es wurden am Sonntag Vormittag eingenommen (Merkmal x: Preis einer Flasche Wein (Euro)):

4 4 4 7 7 7 7 10 5 5 5 10 4 4 7 7 5 5 5 5 5 10 10 10 7

Wir erhalten die unten folgende Häufigkeitstabelle.

j	**Preis für eine Weinflasche** x_j	**absolute Häufigkeit** n_j	**relative Häufigkeit** p_j
1	4	5	5/25=0,2
2	5	8	0,32
3	7	7	0,28
4	10	5	0,2
Σ		25	1

Grafische Darstellungen

Eine weitere Art, Verteilungen eines Merkmals übersichtlich darzustellen, ist die grafische Darstellung. Mit hoher Aussagekraft der Grafik geht meist ein Informationsverlust einher, so daß die Grafik die Tabelle nicht ersetzen, sondern nur unterstützen kann.

Da Grafiken auf einen Blick informieren sollen, sollen sie nicht überladen sein. Häufig verwendet werden heute Piktogramme, d.h. Diagramme, die den Sachverhalt optisch anschaulich verdeutlichen.

Für beliebig skalierte Merkmale mit wenigen Ausprägungen bieten sich eine Vielzahl grafischer Darstellungen an, darunter insbesondere Stabdiagramm, Säulendiagramm, Kreisdiagramm. Diese Diagramme eignen sich nicht für Urlisten mit vielen verschiedenen Beobachtungswerten.

Übung: Warum nicht?

Stabdiagramm bzw. Säulendiagramm

Auf der ||x-Achse|| werden die verschiedenen Ausprägungen des Merkmals markiert. Dann werden die entsprechenden Häufigkeiten als Stab oder Säule senkrecht auf der Abszisse abgetragen.

Es sind hier anhand des obigen Beipiels bezüglich des Familienstandes die Säulendiagramme für die absoluten und relativen Häufigkeiten dargestellt. Wir sehen, dass die Struktur der Diagramme identisch ist.

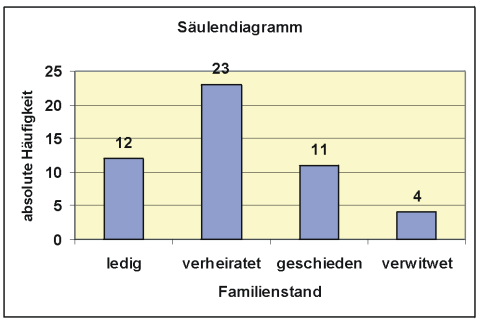

Abbildung 62: Absolute Häufigkeiten des Familienstandes

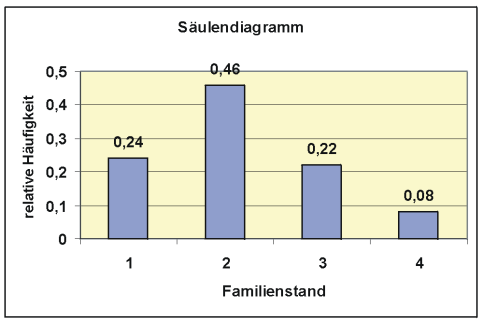

Abbildung 63: Relative Häufigkeiten des Familienstandes

Kreisdiagramm

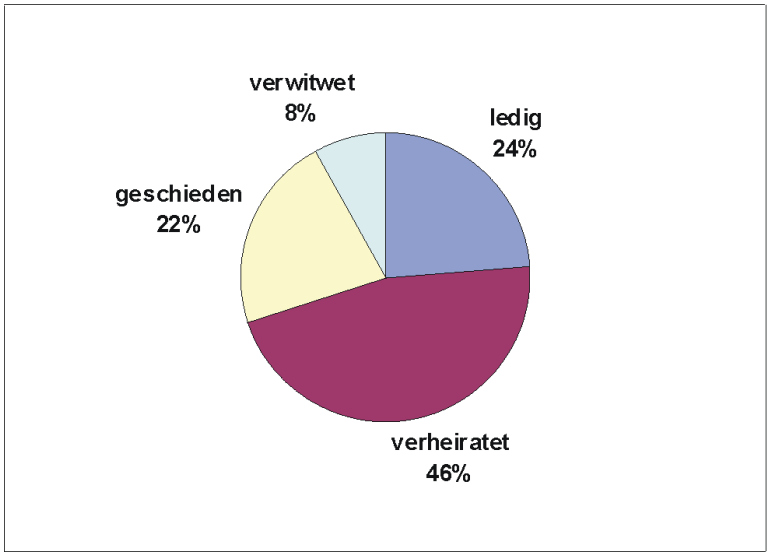

Abbildung 64: Kreisdiagramm: Relative Häufigkeiten des Familienstandes

Im Kreisdiagramm wird n als Gesamtfläche festgelegt. Die Häufigkeiten für die einzelnen Kategorien des Merkmals werden als "Tortenstücke" eingetragen, wobei deren Fläche proportional zur Häufigkeit ist. Der zur Häufigkeit n_j gehörende Winkel α_j eines Segments berechnet sich dann aus der Verhältnisgleichung

$$\frac{\alpha_j}{360} = \frac{n_j}{n}$$

Sollen zwei verschiedene Gesamtheiten mit verschiedenen Gesamthäufigkeiten n_I und n_{II} mittels zweier Kreisdiagramme verglichen werden, kann man die Flächen der Kreise proportional zu den n_I und n_{II} darstellen.

Für die Darstellung von Kreisdiagrammen gibt es heutzutage genügend Anwendersoftware, so dass eine genauere Erläuterung unterbleiben kann.

Summenfunktion

Man interessiert sich für Fragen wie "Wieviel % der Kunden gaben höchstens 5 Euro für eine Flasche Wein aus?" oder "Wieviel Einwohner Deutschlands sind mindestens 65 Jahre alt?". Man könnte nun die einzelnen Häufigkeiten einer Häufigkeitstabelle aufsummieren und so den Wert ermitteln, aber

einfacher ist es, schon in der Häufigkeitstabelle die Häufigkeiten (abs. oder rel.) laufend aufzuaddieren. Es ergeben sich die **Summenhäufigkeiten** als **kumulierte Häufigkeiten** Sj (absolut) bzw. Sj* (relativ) . Aus den Summenhäufigkeiten läßt sich dann einfach die Summenfunktion bestimmen.

Summenhäufigkeiten sind nur sinnvoll, wenn man das Merkmal nach Größe ordnen kann, also nur bei ordinal oder metrisch skalierten Merkmalen. Aus der Summenhäufigkeit kann man die Summenfunktion ermitteln.

Beispiel der verkauften Weinflaschen

j	**Preis für eine Weinflasche x_j**	**absolute Häufigkeit n_j**	**relative Häufigkeit p_j**	**absolute Summenhäufigkeit S_j**	**relative Summenhäufigkeit S^*_j**
1	4	5	5/25=0,2	5	0,20
2	5	8	0,32	13	0,52
3	7	7	0,28	20	0,80
4	10	5	0,2	25	1,00
Σ		25	1		

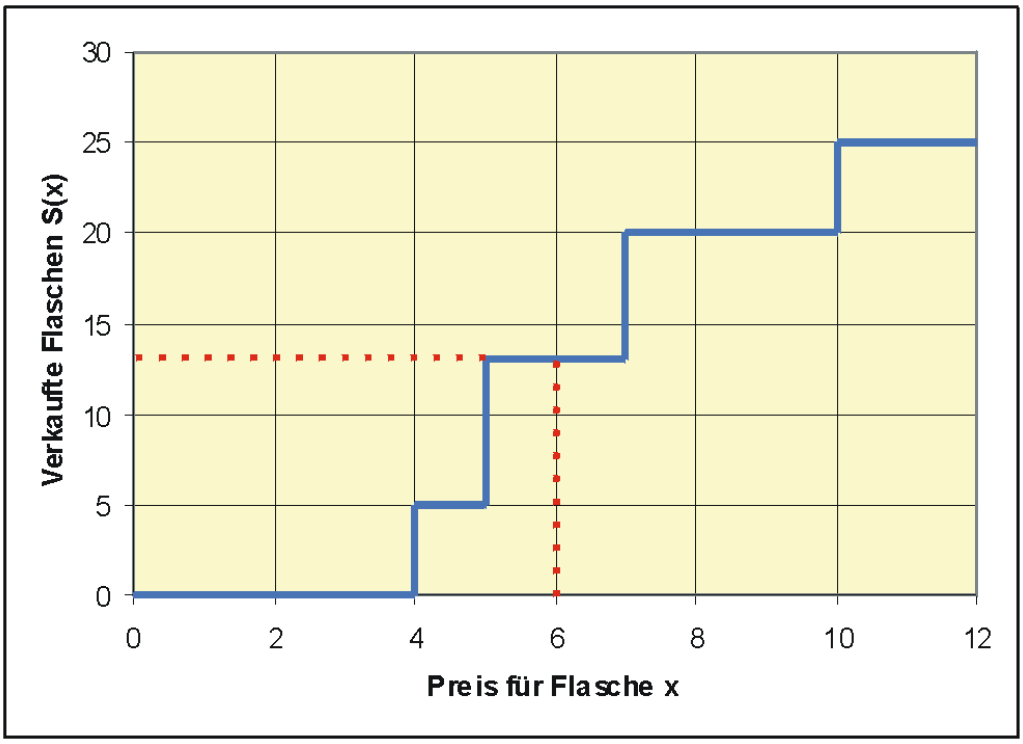

Abbildung 65: Summenfunktion

Für die Erstellung der Summenfunktion müssen die Beobachtungen der Urliste geordnet vorliegen. Die Häufigkeitsverteilung enthält alle Werte der Urliste geordnet. Analog zu oben kann man sich beispielsweise überlegen:

> 20 Kunden zahlten höchstens 7 Euro für eine Flasche, also S(7) = 20.

So können wir wieder wie vorher die Summenfunktion von links her aufbauen:

> 0 Kunden zahlten höchstens 2 Euro für eine Flasche, also S(2) = 0

usw.

Nun können wir die kumulierten Häufigkeiten auch aus der Grafik ablesen: z.B. S(6) = 13, es sind also 13 Flaschen zu einem Preis von höchstens 6 Euro verkauft worden.

Arithmetisches Mittel

Beispiel

Es wurden in einem Einkaufszentrum n = 20 Kunden bezüglich der Kinderzahl befragt. Wir erhielten die geordnete Urliste

0 0 0 0 1 1 1 1 1 2 2 2 2 2 3 3 3 4 5 5

Es resultierte die Häufigkeitsverteilung

j	**Zahl der Kinder x_j**	**absolute Häufigkeit n_j**	**relative Häufigkeit p_j**	$\mathbf{x_j n_j}$	$\mathbf{x_j p_j}$
1	0	4	0,2	0	0
2	1	5	0,25	5	0,25
3	2	5	0,25	10	0,5
4	3	3	0,15	9	0,45
5	4	1	0,05	4	0,2
6	5	2	0,1	10	0,5
Σ		20	1	**38**	**1,9**

Wir bestimmen das arithmetische Mittel als

$$\begin{aligned}\bar{x} &= \tfrac{1}{20}(0+0+0+0+1+1+1+1+1+2+2+2+2+2+3\\ &+ 3+3+4+5+5) = \tfrac{38}{20} = 1,9\end{aligned}$$

Wir können das Mittel aber auch so berechnen:

$$\bar{x} = \frac{1}{20}(4\cdot 0 + 5\cdot 1 + 5\cdot 2 + 3\cdot 3 + 1\cdot 4 + 2\cdot 5) = \frac{38}{20} = 1,9$$

was in Formelschreibweise ergibt

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{m} x_j \cdot n_j.$$

Ermitteln wir das arithmetische Mittel von Hand, können wir in der Häufigkeitstabelle die Summanden $x_j n_j$ in der jten Zeile eintragen und aufsummieren.

Alternativ können wir das arithmetische Mittel mit Hilfe der relativen Häufigkeit p_j ermitteln:

$$\bar{x} = \sum_{j=1}^{m} x_j \cdot p_j.$$

Zur Verdeutlichung ist auch diese Variante in der Häufigkeitstabelle aufgeführt.

Für ordinal- oder nominalskalierte Merkmale ist das arithmetische Mittel nicht geeignet.

Entsprechende Überlegungen gelten auch für die Varianz s^2 der Stichprobe.

Median

Beispiel mit den verkauften Weinflaschen

Wir haben die Urliste nun **geordnet**.

4 4 4 4 4 5 5 5 5 5 5 5 5 7 7 7 7 7 7 7 10 10 10 10 10

Der Median teilt die kleineren 50% der Datenwerte von den 50% größeren Werten ab. Also liegt hier der Median auf dem 13. Beobachtungswert.

Bei Daten in Häufigkeitstabellen liegen die Werte schon zwangläufig geordnet vor. Es muss nur die Kategorie gefunden werden, die den Median enthält.

j	**Preis für eine Weinflasche x_j**	**absolute Häufigkeit n_j**	**absolute Summenhäufigkeit S_j**
1	4	5	5
2	**5**	**8**	**13**
3	7	7	20
4	10	5	25
Σ		25	

Anhand der Summenhäufigkeiten können wir sehen, dass der 13. Wert gerade noch in der 2. Kategorie liegt. Diese Kategorie ist die **Einfallsklasse** des Medians.

Hier wollen wir die Berechnung der **Varianz eines häufbaren metrischen Merkmals** ansehen. Unsere Überlegungen laufen analog zum arithmetischen Mittel. Wir betrachten das

Beispiel mit den verkauften Weinflaschen

Aus der Urliste mit 25 Beobachtungen:

4 4 4 4 4 5 5 5 5 5 5 5 5 7 7 7 7 7 7 7 10 10 10 10 10

berechnen wir die Stichprobenvarianz aus

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

In dieser Formel ist x_i die i. Beobachtung aus der Urliste.

Analog zum arithmetischen Mittel eines Merkmals mit wenig Ausprägungen werden wir aber nicht die obige Formel für die Varianz verwenden, sondern die Vorteile der Häufigkeitstabelle nützen. Wir können nämlich die Stichprobenvarianz berechnen als

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{m}(x_j - \overline{x})^2 \cdot n_j,$$

wobei die x_j jetzt die verschiedenen Ausprägungen des Merkmals darstellen.

j	**Preis für eine Weinflasche x_j**	**absolute Häufigkeit n_j**	$x_j\ n_j$	$(x_j - x)^2$	$(x_j - x)^2 n_j$
1	4	5	20	5,5696	27,8480
2	5	8	40	1,8496	14,7968
3	7	7	49	0,4096	2,8672
4	10	5	50	13,2496	66,2480
Σ		**25**	**159**		**111,7600**

Zunächst benötigen wir den Mittelwert x. Er berechnet sich wie in Lageparameter als

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{m} x_j n_j = \frac{159}{25} = 6,36.$$

Wir erhalten nun

$$s^2 = \frac{1}{24} \cdot 111,7600 \approx 4,66.$$

Der Computer kann das leicht ermitteln. Möchten wir jedoch die Varianz händisch ausrechnen, finden wir den ||krummen|| Mittelwert als störend. Wir können natürlich auch hier den Verschiebungssatz anwenden. Es gilt nämlich für die benötigte Quadratsumme:

$$Q = \sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot n_j = (\sum_{j=1}^{n} x_j^2 \cdot n_j) - n \cdot \bar{x}^2.$$

Wir berechnen sukzessive in unserer Häufigkeitstabelle die $x_j{}^2$ und $x_j{}^2$ n_j und erhalten zunächst für Q

$$Q = 1123 - 25 \cdot 6,36^2 = 111,76$$

und für die Varianz

$$s^2 = \frac{111,76}{25-1} = 4,66.$$

j	**Preis für eine Weinflasche x_j**	**absolute Häufigkeit n_j**	**x_j n_j**	**x_j^2**	**$x_j^2 n_j$**

1	4	5	20	16	80
2	5	8	40	25	200
3	7	7	49	49	343
4	10	5	50	100	500
Σ		25	159		1123

Varianz als Streuungsparameter

Der Lageparameter allein reicht für die Beschreibung einer Datenmenge nicht aus (analoges Problem wie bei Zufallsverteilungen). Information über die **Streuung** der Beobachtungswerte liefert ein **Streuungsparameter**. Es kommen verschiedene Kennwerte als Streuungsparameter in Betracht, beispielsweise die Varianz, die Spannweite, der Quartilsabstand und der Variationskoeffizient.

Varianz

Am häufigsten wird als Kennwert die Varianz verwendet, da sie wahrscheinlichkeitstheoretisch am besten zu untersuchen ist. Die Varianz sind die mittleren quadratischen Abweichungen der Einzelwerte x_i vom arithmetischen Mittel

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

Der Nenner n-1 wirkt vielleicht etwas befremdlich. Allerdings hat die Verwendung von n-1 statt n wahrscheinlichkeitstheoretische Vorzüge, wenn man die Varianz der Verteilung eines Merkmals mit s^2 schätzen möchte. Man nennt dieses Art der Varianz inferentielle Varianz.

Beispiel

Eine Firma möchte einen Kachelofen auf den Markt bringen, der für einen Komplettpreis zu erwerben ist. Für die Kalkulation dieses Preises benötigt die Firma Informationen über die Montagezeit für einen Kachelofen. Bei der Endmontage von 11 Kachelöfen ergaben sich die Zeiten

2,5 3 3 3,3 3,6 3 2,3 3 3,1 3,2 3

Die Varianz der Montagezeiten soll bestimmt werden. Nach der obigen Formel muss zunächst das arithmetische Mittel bestimmt werden:

$$\begin{aligned} \bar{x} &= \tfrac{1}{11}(2,5+3+3+3,3+3,6+3+2,3+3+3,1+3,2+3) \\ &= \tfrac{33}{11} = 3h \end{aligned}$$

.

Dann erhalten wir als Varianz

$$s^2 = \frac{1}{10}((2,5-3)^2+(3-3)^2+(3-3)^2+...+(3-3)^2)$$

$$\begin{aligned} &= \tfrac{1}{10}(0,25+0+0+0,09+0,36+0+0,49+0+0,01+0,04+0) \\ &= \tfrac{1,24}{10} = 0,124h^2 \end{aligned}$$

.

Verzichtet man auf eine Schätzung, kann man auch die deskriptive Varianz

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

für die Beschreibung von statistischen Daten verwenden, was aber hier zur Vermeidung von Verwechslungen unterlassen wird.

Bei der manuellen Berechnung von s^2 ist es oftmals mühsam, erst die einzelnen Differenzen x_i - x zu bilden und dann zu quadrieren. Mit Hilfe des **Verschiebungssatzes** kann die laufende Differenzenbildung vermieden werden. Betrachten wir die Summe

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

.

Diese Summe lässt sich zerlegen in

$$Q = \sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2$$

.

Setzt man den Ausdruck oben ein, erhält man für die Varianz

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

Beispiel:

$$\begin{aligned} s^2 &= \tfrac{1}{10}(2,5^2 + 3^2 + 3^2 + ... + 3^2 - 11 \cdot 3^2) \\ &= \tfrac{1}{10} \cdot (100,24 - 99) = 0,124h^2 \end{aligned}$$

.

Da die Varianz ein quadratischer Ausdruck ist, hat sie z.B. auch die Einheit h^2, wenn die x_i die Einheit h haben. Um die Varianz anschaulicher zu machen, kann man ihre Quadratwurzel, die Standardabweichung s betrachten:

Beispiel

$$s = \sqrt{0,124h^2} \approx 0,35h$$

,

also ca. 20 Minuten. Man könnte etwas flapsig sagen, dass die Montagezeit eines Ofens im Mittel 3 Stunden +/- 20 Minuten beträgt.

Auch die Varianz reagiert empfindlich auf Ausreißer. Es gibt hier resistente Streuungsparameter, die weiter unten behandelt werden.

Metrische Merkmale mit vielen verschiedenen Ausprägungen

Klassierung

Liegen sehr viele verschiedene Beobachtungen eines metrisch skalierten Merkmals vor, ist es wenig sinnvoll, die Ausprägungen zu zählen. Hier müssen die

einzelnen Werte für die Häufigkeitstabelle zusammengefasst werden. Das geschieht in sogenannten Klassen.

Beispiel

Es liegen für 32 europäische Länder als Indikator für den Wohlstand die Zahlen der PKWs pro 1000 Einwohner vor:

31 43 65 152 156 247 264 266 280 289 295 332 341 351 357 365 400 421 422 423 438 451 452 456 489 494 514 516 541 557 591 641

Diese Vielzahl unterschiedlicher Werte ist unübersichtlich. Sie werden zu **Klassen** zusammengefasst, und zwar so:

Klasse 1	über 0 - bis 200	31 43 65 152 156
Klasse 2	über 200 bis 300	247 264 266 280 289 295
Klasse 3	über 300 bis 400	332 341 351 357 365 400
Klasse 4	über 400 bis 500	421 422 423 438 451 452 456 489 494
Klasse 5	über 500 bis 700	514 516 541 557 591 641

So dass wir dann die folgende Häufigkeitstabelle erhalten:

j	**Zahl der PKW pro 1000**	**Zahl der Länder absolute Häufigkeit n_j**	**relative Häufigkeit p_j**
1	über 0 - bis 200	5	5/32 =0,15625
2	über 200 bis 300	6	0,1875
3	über 300 bis 400	6	0,1875
4	über 400 bis 500	9	0,28125
5	über 500 bis 700	6	0,1875
Σ		32	1

Struktur von Klassen

Wir wollen anhand des Beispiels die Struktur von Klassen ansehen:

Es werden benachbarte Merkmalsausprägungen x_i zu einer Klasse zusammengefasst. Wir bezeichnen als

- Zahl der Klassen: m (m=5)
- Absolute der Beobachtungswerte in der Klasse j (j = 1, ..., m): n_j
- Relative Häufigkeit: $p_j = \frac{n_j}{n}$
- Klassenobergrenze: x_{oj} ; Klassenuntergrenze: x_{uj}
- Klassenbreite: $d_j = x_{oj} - x_{uj}$
- Klassenmitte: :$x'_j = \frac{x_{oj}+x_{uj}}{2}$

Bemerkungen

Die Beobachtungen sollen in einer Klasse möglichst gleichmäßig verteilt sein. Idealerweise haben alle Klassen dieselbe Breite, was aber nur bei gleichmäßiger Verteilung der Beobachtung zu empfehlen ist. Auf jeden Fall sollen keine leeren Klassen in der Mitte auftreten.

Für die empfehlenswerte Zahl von Klassen gilt die Faustregel $m \approx \sqrt{n}$. Die Zuordnung der Beobachtung zu einer Klasse muß eindeutig sein, also

nicht	10 - 11	11 - 12	12 - 13	...
sondern	10 - unter 11	11 - unter 12	12 - unter 13	...
oder	über 10 bis 11	über 11 bis 12	über 12 bis 13	...

Manchmal treten **offene Randklassen** auf.

Beispiel:

Größe der landwirtschaftlichen Betriebe in Bayern

Klasse $_j$	Größe des Betriebs(in ha)	...
1	höchstens 5	...
2	über 5 bis 10	...
3	über 10 bis 50	...
4	mehr als 50	...

Behandlung offener Randklassen

Bestimmte Verfahren wie beispielsweise Histogramme etc. verlangen einen Randwert für die oberste und unterste Klasse. Bei offenen Randklassen muß der äußere Randwert „erfunden“ werden.

1. Falls gleiche Klassenbreiten existieren, werden die Randklassen genauso breit gemacht.
2. Man verwendet als äußere Klassengrenze einen plausiblen Wert.

Grafiken

Der Klassiker einer Grafik für klassierte Daten ist das Histogramm, eine Entsprechung des Säulendiagramms. Man trägt auf der Abszisse die Klassen ab und errichtet über den Klassen Rechtecke, deren Fläche die absolute oder relative Häufigkeit beträgt.

Wir wollen nun für die PKW-Indikatordaten ein Histogramm konstruieren. Die Intervallbreiten und die Flächen der einzelnen Rechtecke sind bekannt, uns fehlt jedoch die Höhe einer Säule. Wir werden dieses Problem geometrisch angehen:

Es gilt Fläche = Höhe * Breite, bzw.

$$n_j = h_j \cdot d_j$$

,

also

$$h_j = \frac{n_j}{d_j}$$

.

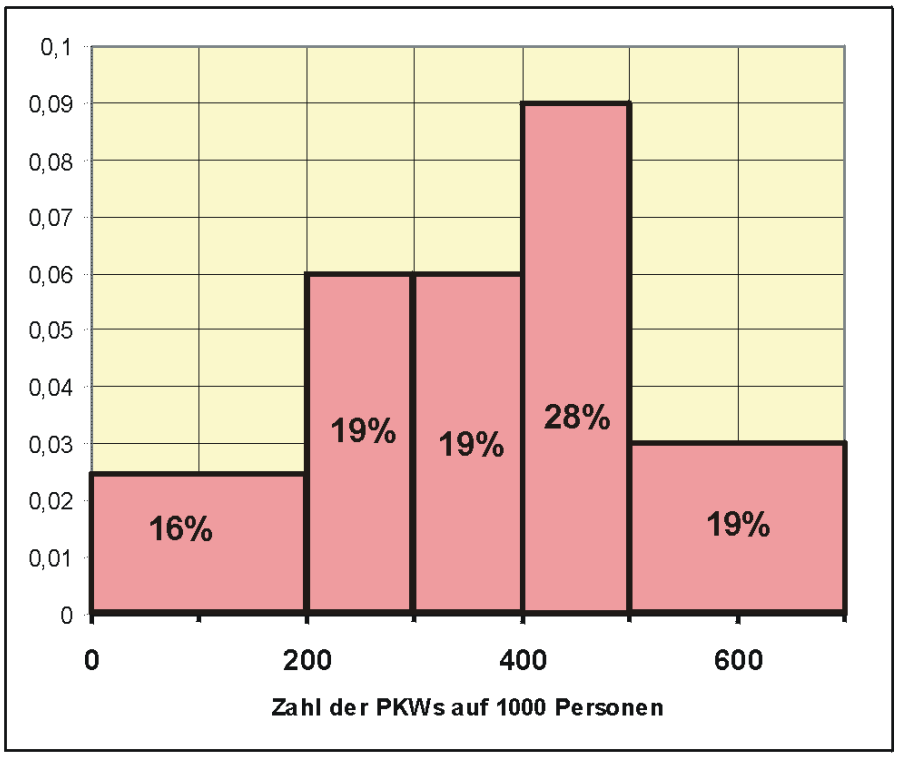

Abbildung 66: Histogramm der PKWS pro tausend Einwohner in Europäischen Ländern

j	**Zahl der PKW pro 1000**	**Zahl der Länder absolute Häufigkeit n_j**	**Klassenbreite d_j**	**Säulenhöhe $h_j = n_j/d_j$**
1	über 0 - bis 200	5	200 - 0 = 200	0,025
2	über 200 bis 300	6	100	0,06
3	über 300 bis 400	6	100	0,06
4	über 400 bis 500	9	100	0,09
5	über 500 bis 700	6	200	0,03

Üblicherweise wird beim Histogramm die Ordinate (y-Achse) weggelassen, weil sonst die Höhe der Säule als Häufigkeit gedeutet wird. Tatsächlich ist aber die Fläche der Säule die Häufigkeit. Es geht ja in der Grafik darum, einen optischen Eindruck von der Aufteilung der Daten zu bekommen. In unserem Beispiel wurde die Ordinate beibelassen, damit die Konstruktion des Histogramms deutlich wird. Man kann zur Unterstützung der Information noch die Häufigkeiten in die Säulen eintragen.

Bei Beobachtungen, die man zweckmäßigerweise **klassiert** zusammenfasst, ist eine Summenfunktion aus der Urliste schwierig zu erstellen und auch unhandlich.

Da hier das Merkmal als stetig angesehen wird, nähert man die grafische Darstellung der Verteilung durch ein Kurve an. Dabei wird folgendermaßen vorgegangen:

Um die absolute Summenfunktion zu erstellen, berechnet man für jede Klasse j die kumulierte Häufigkeit S_j. Dann trägt man die Wertepaare (x_{oj};S_j), also die Klassenobergrenze und Summenhäufigkeit in ein Diagramm ein und verbindet die Punkte geradlinig. Es ist der erste Punkt (x_{u1};0). Ab (x_{om};n) verläuft die Summenkurve horizontal.

PKW-Beispiel

Dazu fassen wir die benötigten Werte am besten wieder in einer Zahlentabelle zusammen: Wir benötigen die Klassenobergrenzen x_{oj} und die Summenhäufigkeiten S_j. Die Summenhäufigkeiten sind die kumulierten Häufigkeiten

$$S_j = \sum_{k=1}^{j} n_k$$

etwa $S_1 = 5$, $S_2 = 5 + 6 = 11$, $S_3 = 5 + 6 + 6 = 17$ ...

Klasse	**Merkmals-werte**	**Absolute Häufigkeit**	**Klassen-obergrenze**	**Absolute Summen-häufigkeit**
j	**x**	$\mathbf{n_j}$	$\mathbf{x_{oj}}$	$\mathbf{S_j}$
1	0 - bis 200	5	200	5
2	ü. 200 bis 300	6	300	11

3	ü. 300 bis 400	6	400	17
4	ü. 400 bis 500	9	500	26
5	ü. 500 bis 700	6	700	32
Σ		32		

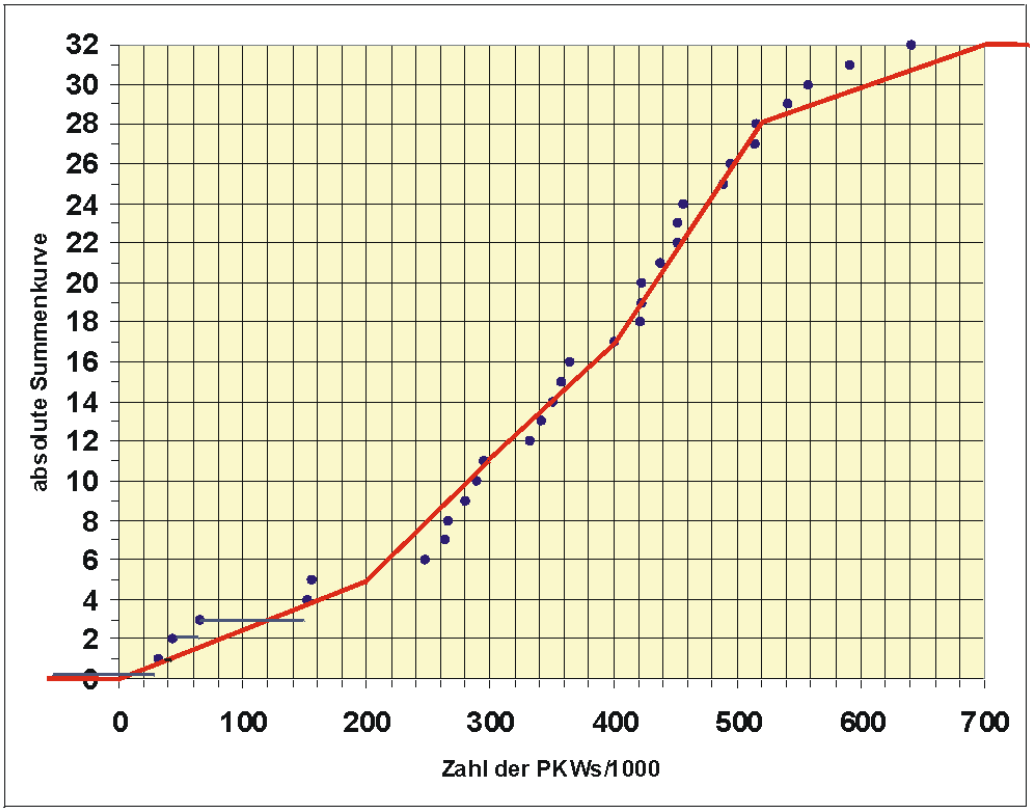

Abbildung 67: Absolute Summenkurve

Je gleichmäßiger die einzelnen Beobachtungen über die Klassen verteilt sind, desto besser passt sich die Summenkurve an die Summenfunktion der einzelnen Beobachtungen an.

In der Grafik ist die Summenkurve für das PKW-Beispiel angegeben. Zum Vergleich wurde die Summenfunktion der Urliste mit eingetragen, wobei aus Übersichtlichkeitsgründen nur bei den ersten Werten die Horizontale gezeigt wird. Man sieht, dass im Intervall 200 - 300 die Kurve die tatsächlichen Beobachtungen überschätzt, im Intervall 600 - 700 liegt die Kurve unter der tatsächlichen Summenfunktion.

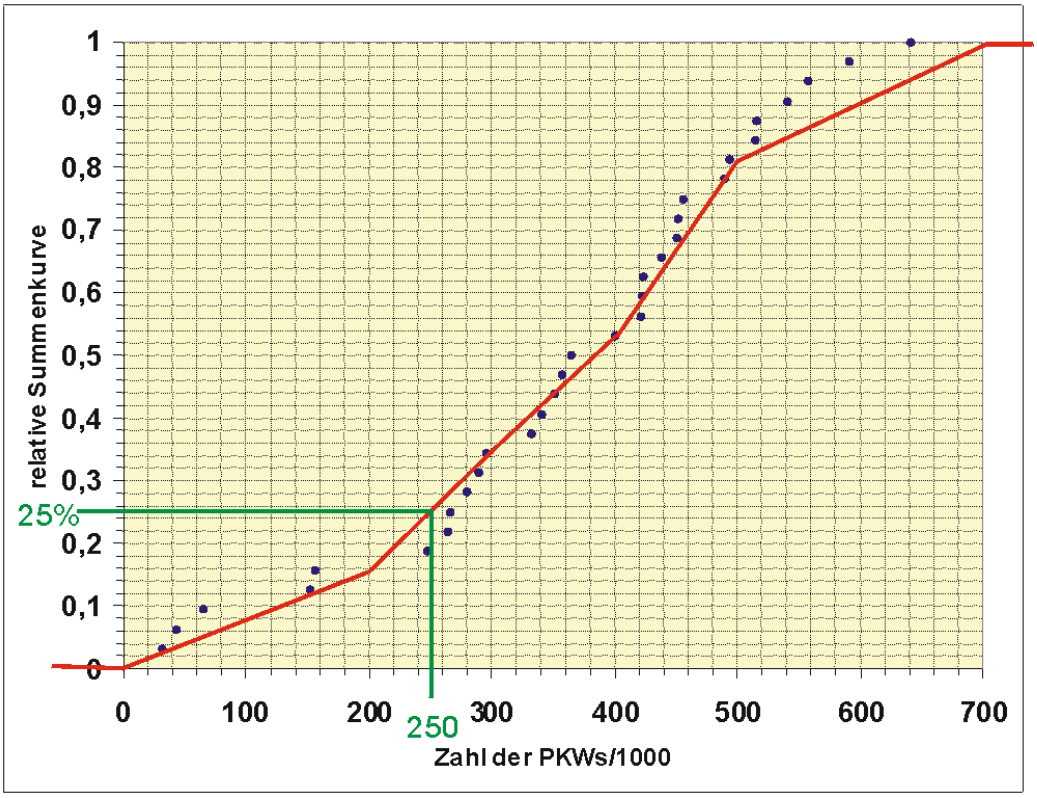

Abbildung 68: Relative Summenfunktion mit 25%-Quantil

Die Summenfunktion ist eine empirische Beschreibung der Verteilung des Merkmals in der Grundgesamtheit. Wie andere grafisch dargestellte Verteilungen ist auch sie vom optischen Informtionsgehalt her eher wenig instruktiv. Man kann aber Verteilungssaussagen grafisch ermitteln, z.B.

Bei der relativen Summenkurve wird statt der absoluten Häufigkeit S_j) die relative Summenhäufigkeit

$$S_j^* = \frac{S_j}{n}$$

verwendet. Die Form der Summenkurve bleibt erhalten.

Arithmetisches Mittel

Ist die Urliste gegeben, berechnet sich das arithmetische Mittel aus der bekannten Durchschnittsbildung der Beobachtungswerte. Sind jedoch die Informationen der Urliste nicht mehr verfügbar, kann man das arithmetische Mittel nur noch näherungsweise bestimmen. Man verwendet die Klassenmitte x_j' als Ersatz für die Merkmalsausprägung x_j in der Klasse j und nähert das arithmetische Mittel an als

$$\bar{x} \approx \bar{x}' = \frac{1}{n} \sum_{j=1}^{m} x'_j \cdot n_j$$

Die Klassenmitte soll das Niveau einer Klasse widerspiegeln. Das ist vor allem der Fall, wenn sich die Einzelwerte der Urliste gleichmäßig in einer Klasse verteilen. Sind die Einzelwerte mehrheitlich an einer Klassengrenze gelegen, gibt x_j' unter Umständen nicht mehr das Niveau korrekt wieder. Die optimale Aufteilung der Klassen sollte schon bei Klassenbildung berücksichtigt werden. Im Sinne einer einfachen Handhabung sollte x_j' eine Zahl sein, mit der man leicht rechnen kann, also z.B. 200 und nicht 199,5.

PKW-Beispiel

Es ergibt sich also als angenähertes arithmetisches Mittel

$$\bar{x}' = \frac{1}{32} \cdot 11750 = 367,1875$$

Klasse	**Intervall**	**Absolute Häufigkeit**	**Klassenmitte**	
j		**n_j**	**x_j'**	**x_j' n_j**
1	0 - bis 200	5	100	500
2	200 bis 300	6	250	1500
3	300 bis 400	6	350	2100
4	400 bis 500	9	450	4050
5	500 bis 700	6	600	3600
Σ		32		11750

Median

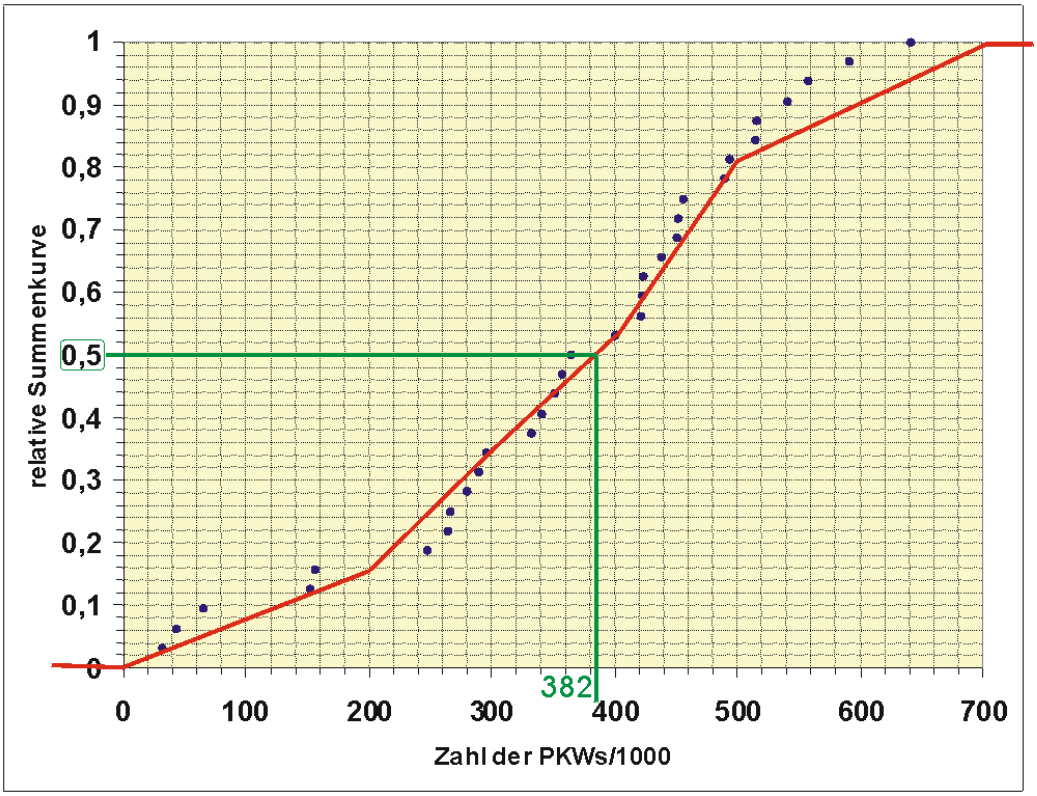

Abbildung 69: Grafische Ermittlung des Medians

Grafische Ermittlung

Hier bietet sich vor allem die grafische Ermittlung des Medians an:

Man bestimmt aus der absoluten (relativen) Summenkurve grafisch den Wert x, der zu n/2 (0,5) gehört.

Im Pkw-Beispiel wurde der Median aus der relativen Summenkurve grafisch ermittelt. Der x-Wert, der zu S*(X)=0,5 gehört, beträgt etwa 382. Es hatten also 50% der untersuchten Länder höchstens ca. 382 Fahrzeuge pro 1000 Einwohner.

Ist n klein, könnte man auch vom Ordinatenwert (n+1)/2 bei geradem n ausgehen.

Ermittlung mit der Häufigkeitstabelle

Man kann den Median auch näherungsweise durch lineare Interpolation aus der Häufigkeitstabelle ermitteln. Allerdings genügt im Allgemeinen auch die Klassenmitte der Einfallsklasse als Näherung für den Median, da ohnehin meistens keine Informationen über die Verteilung der Beobachtungen in den Klassen vorliegen.

Im PKW-Beispiel ergäbe die Näherung durch die Klassenmitte z' = 350.

Lineare Interpolation würde

$$x_{u3} + \frac{0,5 \cdot (x_{o3} - x_{u3})}{p_3} = 300 + \frac{0,5 \cdot 100}{0,5312} = 394,12$$

ergeben.

Einleitung

Liegen bei einem klassierten Merkmal keine Informationen über die Urliste mehr vor, können wir die Varianz des Merkmals analog zum arithmetischen Mittel mit den Klassenmitten näherungsweise berechnen. Wir erhalten für die Näherung $s^{2'}$

$$s^2 \approx s^{2\prime} = \frac{1}{n-1} \sum_{j=1}^{m} (x'_j - \bar{x}')^2 \cdot n_j \ ,$$

deren Exaktheit auch wieder von der Verteilung der einzelnen Werte in den Klassen abhängt. Verwenden wir statt der absoluten Häufigkeiten n_j die relativen p_j, berechnet sich die Varianz als

$$s^2 \approx s^{2\prime} = \frac{n}{n-1} \sum_{j=1}^{m} (x'_j - \bar{x}')^2 \cdot p_j \ .$$

Man kann auch im Fall der näherungsweisen Berechnung den Verschiebungssatz anwenden. Wir wollen ihn hier nur für absolute Häufigkeiten angeben. Für die Quadratsumme der zentrierten Klassenmittel gilt

$$\sum_{j=1}^{m}(x'_j - \bar{x}')^2 \cdot n_j = \sum_{j=1}^{m} x'^2_j \cdot n_j - n \cdot \bar{x}'^2,$$

so dass sich für die angenäherte Varianz ergibt

$$s^2 \approx s^{2\prime} = \frac{1}{n-1}(\sum_{j=1}^{m} x'^2_j \cdot n_j - n \cdot \bar{x}'^2 \ .)$$

PKW-Beispiel

Wie bei der Ermittlung des arithmetischen Mittels verwenden wir auch hier zweckmäßigerweise eine Tabelle. Es war das angenäherte arithmetische Mittel 367, 1875. Es wird zunächst die Varianz mit Hilfe der zentrierten Werte ermittelt:

Klasse	**Intervall**	**Absolute Häufigkeit**	**Klassenmitte**
j	**über ... bis ...**	**n_j**	**x_j**
1	0 - 200	5	100
2	200 - 300	6	250
3	300 - 400	6	350
4	400 - 500	9	450
5	500 - 700	6	600
Σ	–	32	–

Klasse			
j	**x_j' - x_j**	**$(x_j$' - $x_j)^2$**	**$(x_j$' - $x_j)^2 \cdot n_j$**
1	-267,19	71390,50	356952,48
2	-117,19	13733,50	82400,98
3	-17,19	295,50	1772,98
4	82,81	6857,50	61717,46
5	232,81	54200,50	325202,98
Σ	–	–	828046,88

Wir erhalten für die Varianz

$$s^2 = \frac{1}{32-1} \cdot 828046,88 = 26711,19$$

und für die Standardabweichung

$$s = \sqrt{26711,19} = 163,44$$

Mit dem Verschiebungssatz dagegen erhalten wir

Klasse	**Intervall**	**Absolute Häufigkeit**	**Klassenmitte**		
j	**über ... bis ...**	$\mathbf{n_j}$	$\mathbf{x_j'}$	$\mathbf{x_j'^2}$	$\mathbf{x_j'^2\ n_j}$
1	0 - 200	5	100	10000	50000
2	200 - 300	6	250	62500	375000
3	300 - 400	6	350	122500	735000
4	400 - 500	9	450	202500	1822500
5	500 - 700	6	600	360000	2160000
Σ		32			5142500

Wir erhalten für die Varianz

$$s^2 = \frac{1}{32-1}(5142500 - 32 \cdot 367,19^2) = 26711,19$$

Kapitel 6

Analyse mehrerer Merkmale

Deskriptive Analyse mehrerer Merkmale

Häufig interessiert man sich für mehrere Merkmale zugleich. Interpretiert man die Beobachtungen wieder als Stichprobe einer unbekannten Grundgesamtheit, könnte man fragen, ob die Variablen unabhängig sind oder, falls nicht, in welcher Beziehung sie zueinander stehen. So kann man beispielsweise etwa vermuten, daß zwischen Werbeausgaben und Umsatz eines Supermarktes ein positiver Zusammenhang besteht.

Korrelation zweier Merkmale

Für die Untersuchung der Beziehung zwischen mehreren Variablen muß grundsätzlich wieder nach Skalierung dieser Variablen unterschieden werden. Die Kovarianz bzw. der Korrelationskoeffizient für zwei Zufallsvariablen einer **Grundgesamtheit** sind uns bereits bekannt. Analog dazu gibt es in der deskriptiven Statistik die **(Stichproben)-Kovarianz** bzw. den **(Stichproben)-Korrelationskoeffizienten**.

Korrelationskoeffizient nach Bravais-Pearson

Es seien zwei Merkmale x und y zu beobachten. Bei einer Stichprobe im Umfang von n ergeben sich n viele Wertepaare $(x_i;y_i)$ (i = 1, ..., n).

Beispiel

Es soll untersucht werden, ob das Bevölkerungswachstum eines Landes mit der Fruchtbarkeitsrate (durchschnittliche Zahl der Geburten einer gebärfähigen Frau) zusammenhängt. Es wurden acht Länder zufällig ausgewählt und wir erhalten die Daten

Land	**Bevölkerungs-wachstum x**	**Fruchtbarkeits-rate y**
Ägypten	1,8	3
Türkei	1,1	2
Vereinigte Arabische Emirate	1,6	3
Jamaika	0,7	2
Mauritanien	2,9	5
Island	1	1,8
Tadschikistan	2,1	4,1
Gabun	2,4	4,7

Um sich einen Eindruck vom Zusammenhang der Daten zu verschaffen, tragen wir sie in einem Streudiagramm ab.

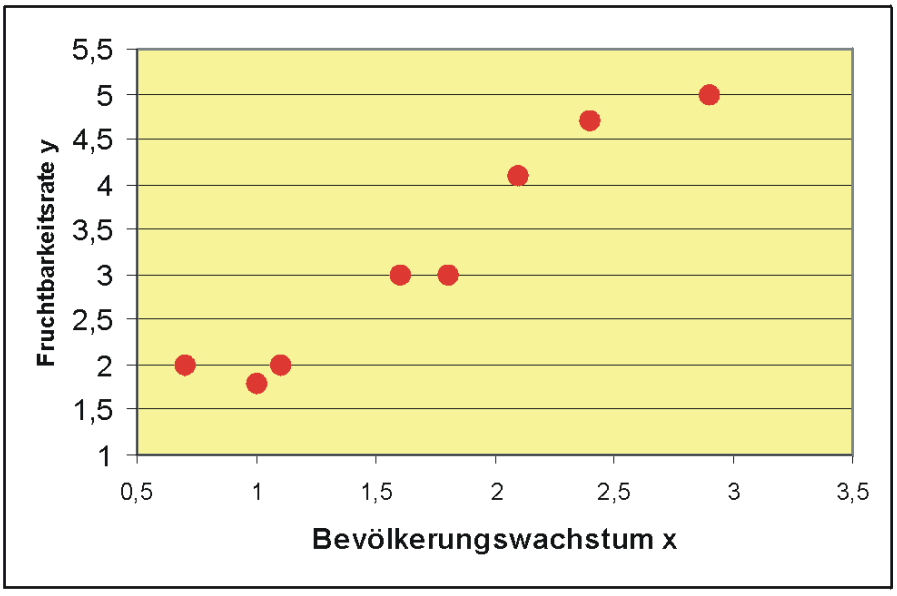

Abbildung 70: Streudiagramm zwischen Bevölkerungswachstum und Fruchtbarkeitsrate für 8 ausgewählte Länder

Man sieht hier schon ohne Analyse, dass offensichtlich mit steigender Fertilität auch das Bevölkerungswachstum zunimmt. Die gestreckte Punktwolke ist fast eine steigende Gerade, also besteht zwischen Fertilität und Bevölkerungswachstum ein annähernd linearer Zusammenhang. Die Merkmale sind offensichtlich stetig. Für metrisch skalierte Merkmale stellt der Korrelationskoeffizient r_{xy} oder kurz r nach Bravais-Pearson ein **Maß für die lineare Abhängigkeit** zweier statistischer Variablen dar:

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

wobei x_1 , x_2 , ..., x_n und y_1 , y_2 , ..., y_n die Messwerte der beiden Merkmale und $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ das arithmetische Mittel des Merkmals x sind, y entsprechend.

Analog zu oben kann auch hier wieder der Verschiebungssatz angewendet werden:

$$r = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n \cdot (\bar{x})^2) \cdot (\sum_{i=1}^{n} y_i^2 - n \cdot (\bar{y})^2)}}$$

Es gilt: Je näher |r| bei 0 ist, desto schwächer ist der ||lineare Zusammenhang||, d.h. die Korrelation. Man sieht an den folgenden Streudiagrammen, dass bei einem Korrelationskoeffizienten von 0,9 das Diagramm stark einer Geraden ähnelt. Je kleiner |r| wird, desto verwaschener wird die Gerade bis hin zur strukturlosen Punktwolke. Ist der Korrelationskoeffizient kleiner als Null, hat die Punktwolke eine fallende Tendenz.

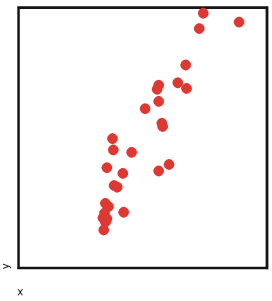

Abbildung 71: **r ≈ 0,9**

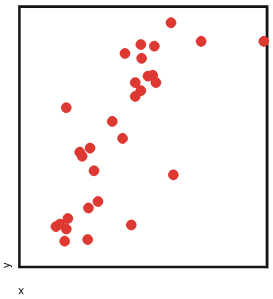

Abbildung 72: **r ≈ 0,7**

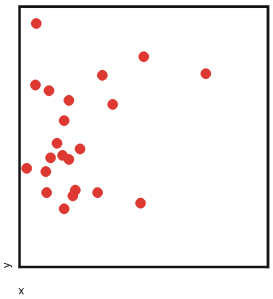

Abbildung 73: **r ≈ 0,2**

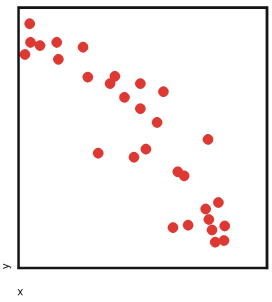

Abbildung 74: **r ≈ -0,9**

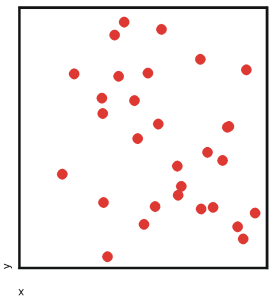

Abbildung 75: **Die Merkmale sind stochastisch unabhängig**

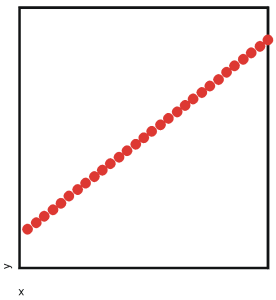

Abbildung 76: **r = 1; y = a + bx**

In der Grundgesamtheit ist bei stochastisch unabhängigen Zufallvariablen die Kovarianz und damit der Korrelationskoeffizient gleich Null. Bei einer Stichprobe stetiger Merkmale wird man aber so gut wie niemals einen Korrelationskoeffizienten erhalten, der genau Null ist. In unserem Beispiel mit den stochastisch unabhängigen Merkmalen wurden 30 Zufallszahlen zweier stochastisch unabhängiger Variablen erzeugt. Der errechnete Stichproben-Korrelationskoeffizient ergab jedoch -0,272. Die Frage ist nun, wie groß muss der errechnete Korrelationskoeffizient mindestens sein, damit man von einer vorhandenen Korrelation ausgehen kann? Hier kann man den Korrelationskoeffizienten statistisch testen, um nachzuprüfen, ob er groß genug ist.

Beispiel mit zentrierten Merkmalswerten

Wir wollen nun den Korrelationskoeffizienten des obigen Beispiels mit der Formel

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

ermitteln. Am besten ordnet man die Daten für die Berechnung in einer Tabelle an (siehe unten). Wir benötigen als Erstes den Mittelwert x:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8} \cdot 13{,}6 = 1{,}7 \ ,$$

entsprechend erhalten wir für y

$$\bar{y} = \frac{1}{8} \cdot 25{,}6 = 3{,}2 \ .$$

Wir wollen nun zuerst die Elemente x_i - x bestimmen, wir nennen diese zentrierten Werte von x hier x*:

$x_1^* = x_1$ - x = 1,8 - 1,7 = 0,1
$x_2^* = x_2$ - x = 1,1 - 1,7 = -0,6
...

Wir können nun die Formel von oben etwas kürzer schreiben als

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

Setzen wir die entsprechenden Spaltensummen der Tabelle ein, ergibt sich

$$r = \frac{6{,}47}{\sqrt{3{,}96 \cdot 11{,}22}} = 0{,}9706 \ .$$

Der Korrelationskoeffizient beträgt also 0,9706. x und y sind hochkorreliert: Wenn die Fruchtbarkeitsrate groß ist, wächst die Bevölkerung stark.

	BevW	**FrR**	x* = x - x	y* = y - y			
i	**x**	**y**	**x***	**y***	**x*y***	**x^{*2}**	**y^{*2}**
1	1,8	3	0,1	-0,2	-0,02	0,01	0,04
2	1,1	2	-0,6	-1,2	0,72	0,36	1,44
3	1,6	3	-0,1	-0,2	0,02	0,01	0,04
4	0,7	2	-1	-1,2	1,2	1	1,44
5	2,9	5	1,2	1,8	2,16	1,44	3,24
6	1	1,8	-0,7	-1,4	0,98	0,49	1,96
7	2,1	4,1	0,4	0,9	0,36	0,16	0,81
8	2,4	4,7	0,7	1,5	1,05	0,49	2,25
Σ	**13,6**	**25,6**	**0**	**0**	**6,47**	**3,96**	**11,22**

Beispiel mit Verschiebungssatz

Wir berechnen Korrelationskoeffizienten mit Hilfe des Verschiebungssatzes:

$$r = \frac{49{,}99 - 8 \cdot 1{,}7 \cdot 3{,}2}{\sqrt{(27{,}08 - 8 \cdot 1{,}7^2) \cdot (93{,}14 - 8 \cdot 3{,}2^2)}} = 0{,}9706$$

	BevW	**FrR**			
i	**x**	**y**	**xy**	**x^2**	**y^2**
1	1,8	3	5,4	3,24	9
2	1,1	2	2,2	1,21	4
3	1,6	3	4,8	2,56	9
4	0,7	2	1,4	0,49	4
5	2,9	5	14,5	8,41	25
6	1	1,8	1,8	1	3,24
7	2,1	4,1	8,61	4,41	16,81
8	2,4	4,7	11,28	5,76	22,09
Σ	**13,6**	**25,6**	**49,99**	**27,08**	**93,14**

Bemerkungen

- Der Korrelationskoeffizient nach Bravais-Pearson reagiert stark auf Ausreißer in den Beobachtungen. Daher sollten die vorliegenden Daten idealerweise normalverteilten Merkmalen entstammen.
- Aufgrund der Durchschnittsbildung ist er für ordinalskalierte Merkmale nicht zulässig.
- In der praktischen Anwendung werden bei Verwendung des Verschiebungssatzes die Produkte häufig sehr groß. Um Rundungsfehler zu vermeiden, zentriert man hier vor Berechnung des Korrelationskoeffizienten die Datenwerte zu xi* und yi* wie oben gezeigt.

Rangkorrelationskoeffizient nach Spearman

Für Variablen, die **stark von der Normalverteilung abweichen**, und auch **ordinalskalierte** Variablen, eignet sich der **Rangkorrelationskoeffizient nach Spearman-Pearson**. Hier werden die einzelnen Beobachtungen von x bzw. y der Größe nach geordnet. Jedem Wert wird seine Rangzahl zugewiesen. Es entstehen so n Paare mit Rangzahlen $\mathrm{rg}(x_i)$ und $\mathrm{rg}(y_i)$. Aus diesen Rängen wird der Korrelationskoeffizent nach Bravais-Pearson errechnet. Man erhält so den Korrelationskoeffizenten nach Spearman-Pearson:

$$r_{SP} = \frac{\sum_i (rg(x_i) - \overline{rg(x)})(rg(y_i) - \overline{rg(y)})}{\sqrt{\sum_i (rg(x_i) - \overline{rg(x)})^2}\sqrt{\sum_i (rg(y_i) - \overline{rg(y)})^2}}$$

.

Wenn alle Ränge verschieden sind, kann man die obige Form zu

$$r_{SP} = 1 - \frac{6 \sum_i d_i^2}{n \cdot (n^2 - 1)}$$

,

umformen mit $d_i = rg(x_i) - rg(y_i)$.

Liegen mehrere gleiche Merkmalswerte vor, handelt es sich um **Bindungen**. Die untere der beiden Formeln ist eigentlich nur korrekt anwendbar, wenn

keine Bindungen vorliegen. Meistens kann man jedoch zur Vereinfachung die Formel näherungsweise verwenden. Zur konkreten Berechnung von Bindungen soll das folgende Beispiel verwendet werden.

Beispiel: Evaluation einer Vorlesung

Es wurde eine Statistikvorlesung evaluiert. Die gesamten Daten sind unter Evaluation verfügbar. Es wurden hier 10 Studierende zufällig ausgewählt. Wir interessieren uns für die Frage, ob möglicherweise die Zufriedenheit der Leute mit der Vorlesung davon abhängt, ob die Vorlesung verständlich war. Es ergaben sich die Daten

Stoff verständlich	**Note für Vorlesung**
x	**y**
2	1
4	4
2	2
3	3
4	3
3	2
3	2
4	3
3	3
3	3

Es werden nun die Ränge ermittelt. Da mehrere Merkmalswerte gleich sind, liegen Bindungen vor, d.h. gleiche Werte bekommen gleiche Rangzahlen. Es gibt verschiedene Methoden, gleiche Rangzahlen zuzuweisen. Meistens werden mittlere Rangzahlen verwendet. Wir wollen für x die Rangzahlen ermitteln. Dazu ordnen wir die x-Werte der Größe nach und numerieren sie durch:

x aufsteigend geordnet	**Laufende Nummer**	**mittlerer Rang**	**Rangzahl**
2	1	$\frac{1+2}{2}$	1,5
2	2		1,5
3	3	$\frac{3+4+5+6+7}{5}$	5
3	4		5

3	5		5
3	6		5
3	7		5
4	8	$\frac{8+9+10}{3}$	9
4	9		9
4	10		9

Für die Ränge von y verfahren wir entsprechend, wie die unten folgende Tabelle zeigt. Nun können wir den Korrelationskoeffizienten nach Spearman-Pearson berechnen:

$$r_{SP} = \frac{\sum_i (rg(x_i) - \overline{rg(x)})(rg(y_i) - \overline{rg(y)})}{\sqrt{\sum_i (rg(x_i) - \overline{rg(x)})^2}\sqrt{\sum_i (rg(y_i) - \overline{rg(y)})^2}} = \frac{54{,}5}{\sqrt{70} \cdot \sqrt{70{,}5}} = 0{,}7758 \ ,$$

wobei sich für $\overline{rg(x)} = \frac{1}{10} \cdot 55 = 5{,}5$ ergibt, für rg(y) ebenfalls. Es scheint zwischen dem Verstehen des Statistikstoffs und der Gesamtzufriedenheit ein deutlich positiver Zusammenhang zu bestehen: Je besser der Stoff verstanden wurde, desto besser fiel tendenziell auch die Note aus.

x	**y**	**rg(x)**	**rg(y)**	**rg(x)*** = rg(x)-rg(x)
2	1	1,5	1	-4
4	4	9	10	3,5
2	2	1,5	3	-4
3	3	5	7	-0,5
4	3	9	7	3,5
3	2	5	3	-0,5
3	2	5	3	-0,5
4	3	9	7	3,5
3	3	5	7	-0,5
3	3	5	7	-0,5
		55	**55**	**0**

x	**y**	**rg(y)*** = rg(y)-rg(y)	**rg(x)*rg(y)***	**rg(x)*²**	**rg(y)*²**
2	1	-4,5	18	16	20,25

4	4	4,5	15,75	12,25	20,25
2	2	-2,5	10	16	6,25
3	3	1,5	-0,75	0,25	2,25
4	3	1,5	5,25	12,25	2,25
3	2	-2,5	1,25	0,25	6,25
3	2	-2,5	1,25	0,25	6,25
4	3	1,5	5,25	12,25	2,25
3	3	1,5	-0,75	0,25	2,25
3	3	1,5	-0,75	0,25	2,25
		0	**54,5**	**70**	**70,5**

Wir werden nun den Korrelationskoeffizienten zum Vergleich mit der vereinfachten Formel ermitteln:

$$r_{SP} = 1 - \frac{6 \sum_i d_i^2}{(n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot 31{,}5}{10 \cdot (100 - 1)} = 0{,}8091$$

.

Dieser Wert weicht etwas vom vorhergehenden ab.

x	**y**	**rg(x)**	**rg(y)**	**d_i= rg(x)- rg(y)**	**d_i^2**
2	1	1,5	1	0,5	0,25
4	4	9	10	-1	1
2	2	1,5	3	-1,5	2,25
3	3	5	7	-2	4
4	3	9	7	2	4
3	2	5	3	2	4
3	2	5	3	2	4
4	3	9	7	2	4
3	3	5	7	-2	4
3	3	5	7	-2	4
					31,5

Bemerkungen

- Wie beim Korrelationskoeffizienten nach Bravais-Pearson kann auch hier der Verschiebungssatz verwendet werden.
- Wird für die Berechnung des Korrelationskoeffizienten der Computer eingesetzt, sollte die vereinfachte Formel nicht verwendet werden, denn sie soll lediglich bei der Berechnung von Hand die Arbeit erleichtern - es sei denn, alle Rangzahlen sind verschieden.

Einfaches lineares Regressionsmodell

Einführung mit Beispiel einer Preis-Absatz-Funktion

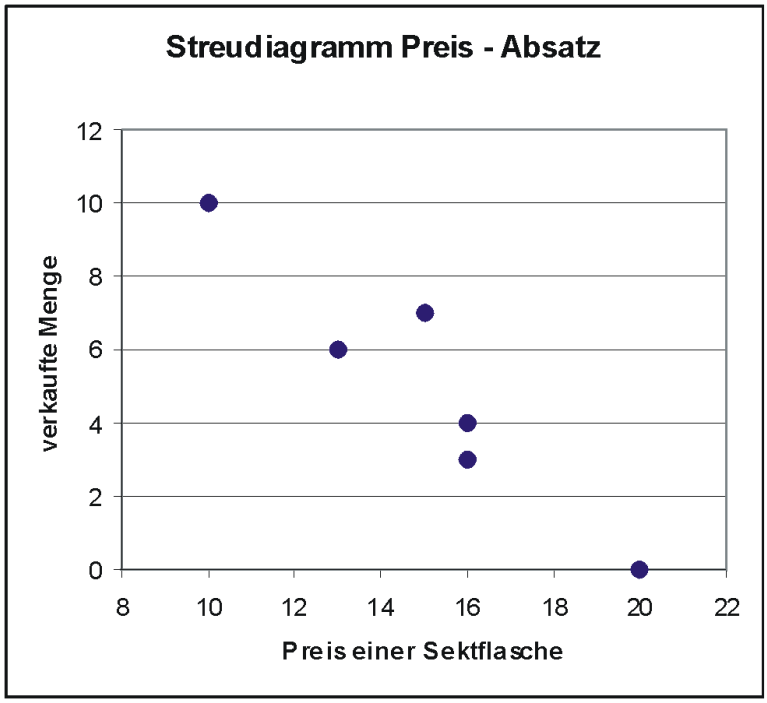

Abbildung 77: Streudiagramm von Preis und abgesetzter Menge an Sektflaschen

Eine renommierte Sektkellerei möchte einen hochwertigen Rieslingsekt auf den Markt bringen. Für die Festlegung des Abgabepreises soll zunächst eine Preis-Absatz-Funktion ermittelt werden. Dazu wurde in n = 6 Geschäften ein Testverkauf durchgeführt. Man erhielt sechs Wertepaare mit dem Ladenpreis x (in Euro) einer Flasche und die verkaufte Menge y an Flaschen:

Laden	**i**	1	2	3	4	5	6
Preis einer Flasche	x_i	20	16	15	16	13	10
verkaufte Menge	y_i	0	3	7	4	6	10

Modell

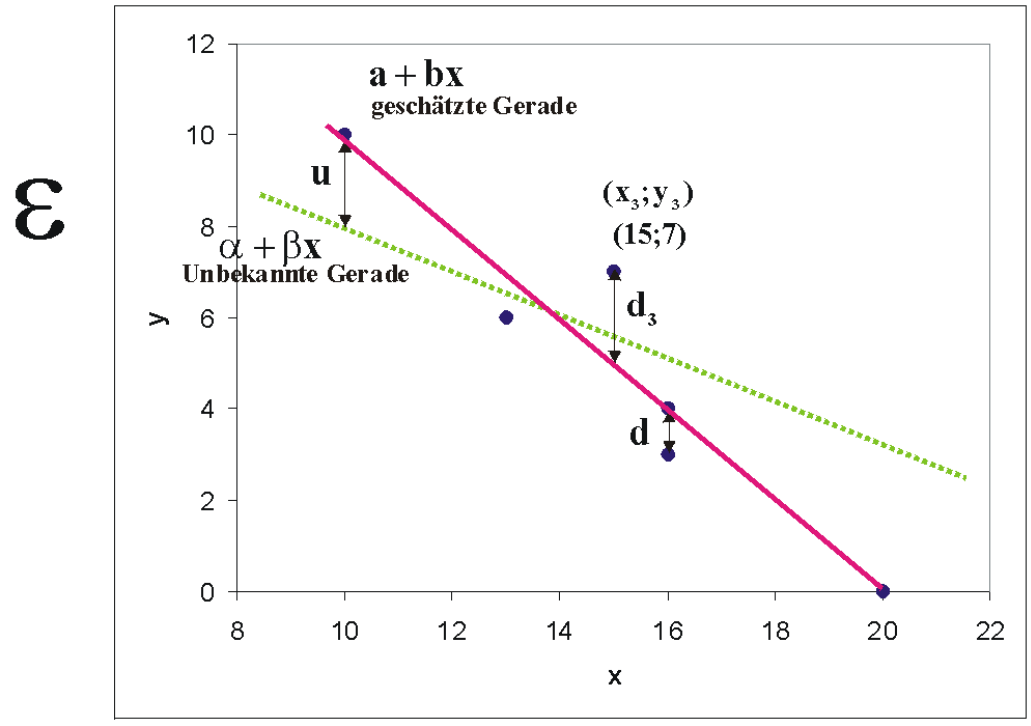

Abbildung 78: Wahre unbekannte und geschätzte Regressionsgerade

Man geht von folgendem statistischen Modell aus:

Man betrachtet zwei Variable, die vermutlich ungefähr in einem linearen Zusammenhang

$$y \approx \alpha + \beta x$$

stehen. Dabei sind x als unabhängige und y als abhängige Variable definiert. Man nennt auch x erklärende oder exogene Variable und y Zielvariable oder endogene Variable. Es existieren von x und y je n Beobachtungen x_i und y_i (i = 1, ... , n). Der funktionale Zusammenhang y = f(x) zwischen x und y kann nicht exakt festgestellt werden, da $\alpha + \beta$x von einer Störgröße u überlagert wird, die nichterfassbare Einflüsse (menschliches Verhalten, Messungenauigkeiten usw.) mit einschließt. Es ergibt sich also das Modell

$$y = \alpha + \beta x + u$$

mit den einzelnen Beobachtungen

$$y_i = \alpha + \beta x_i + u_i \ .$$

Da α und βx nicht bekannt sind, kann y auch nicht in die Komponenten α + βx und u zerlegt werden.

Es soll eine mathematische Schätzung für die Parameter α und β durch zwei Konstanten a und b gefunden werden, und zwar so, daß sich ergibt

$$y_i = a + bx_i + d_i$$

,

wobei d_i das Residuum bezeichnet, die Abweichung des beobachteten y- Wertes vom geschätzten. Es gibt verschiedene Möglichkeiten, die Regressiongerade zu schätzen. Man könnte eine Gerade so durch den Punkteschwarm legen, dass die Quadratsumme der Residuen, also der senkrechten Abweichungen d_i der Punkte von dieser **Ausgleichsgeraden** minimiert wird.

Beispiel zum Modell der Grundgesamtheit

Wöchentliche Düngergabe pro Pflanze (ml)									
	x1	x2	x3	x4	x5	x6	x7	x8	x9
	40	50	60	70	80	90	100	110	120
Ertrag (kg)	1,41	1,47	1,45	1,70	1,57	1,77	2,01	2,32	2,26
	1,44	1,56	1,64	1,58	1,56	2,11	1,79	2,02	2,30
	1,24	1,28	1,62	1,71	1,79	1,92	2,09	2,13	2,21
	1,22	1,35	1,40	1,46	1,78	1,91	2,08	2,16	2,33
	1,26	1,30	1,80	1,80	1,74	2,14	1,79	2,25	2,39
	1,37	1,33	1,57	1,83	1,59	1,90	1,81	1,91	2,18
	1,18	1,36	1,59	1,70	2,03	1,75	2,08	2,23	2,33
	1,56	1,49	1,60	1,75	1,74	1,87	2,13	2,07	1,99
$\alpha+\beta x$	1,4	1,5	1,6	1,7	1,8	1,9	2	2,1	2,2
Störgröße	u1	u2	u3	u4	u5	u6	u7	u8	u9
$u=y-(\alpha+\beta x)$	0,01	-0,03	-0,15	0,00	-0,23	-0,13	0,01	0,22	0,06
	0,04	0,06	0,04	-0,12	-0,24	0,21	-0,21	-0,08	0,10
	-0,16	-0,22	0,02	0,01	-0,01	0,02	0,09	0,03	0,01
	-0,18	-0,15	-0,20	-0,24	-0,02	0,01	0,08	0,06	0,13
	-0,14	-0,20	0,20	0,10	-0,06	0,24	-0,21	0,15	0,19
	-0,03	-0,17	-0,03	0,13	-0,21	0,00	-0,19	-0,19	-0,02
	-0,22	-0,14	-0,01	0,00	0,23	-0,15	0,08	0,13	0,13
	0,16	-0,01	0,00	0,05	-0,06	-0,03	0,13	-0,03	-0,21

Abbildung 79: Tabelle 1: Daten

In einem breit angelegten Versuch wird ein Flüssigdünger an in Nährlösung gezogenen Peperonis untersucht. Es wird wöchentlich jeder Pflanze eine bestimmte Menge Dünger verabreicht. Nach zwei Monaten wird der Gesamtertrag einer Pflanze gewogen. Die Abhängigkeit des Ertrags y (kg) von der Düngermenge x (ml) lässt sich beschreiben als

$$y = \alpha + \beta x + u = 1 + 0,01x + u \ ,$$

wobei natürlich nur der große Statistik-Gott diese Gerade kennt, wir können nur einzelne Versuche machen. In der Tabelle 1 sind für die Düngergaben 40, 50, ... ,120 ml für jeweils 8 Pflanzen die resultierenden Erträge aufgeführt. Man sieht, dass die Erträge um $\alpha + \beta \cdot x$ schwanken, was natürlich an der Störgröße $u = y - (\alpha + \beta \cdot x)$ liegt.

Betrachten wir die Störgröße bei einer Düngermenge von $x_3 = 60$ ml. Es wurden hier die Erträge von 150 Peperoni-Pflanzen erfasst. Wenn man die

Realisationen der Störgröße u_3 in einem Dotplot abträgt (Grafik 2), erkennt man, dass die Werte normalverteilt sein könnten. Zu x_3 gehört also eine eigene Wahrscheinlichkeitsverteilung der Störgröße, ebenso zu x_1, x_2 usw. In der Grafik 3 sind diese verschiedenen Verteilungen der u exemplarisch angedeutet.

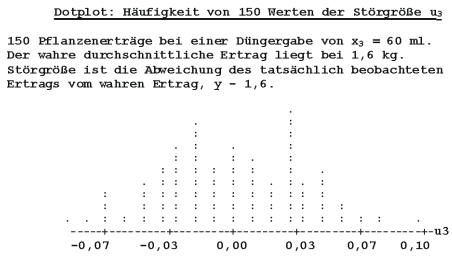

Abbildung 80: Grafik 2: Dotplot von 150 Realisationen der Störgröße bei x = 40

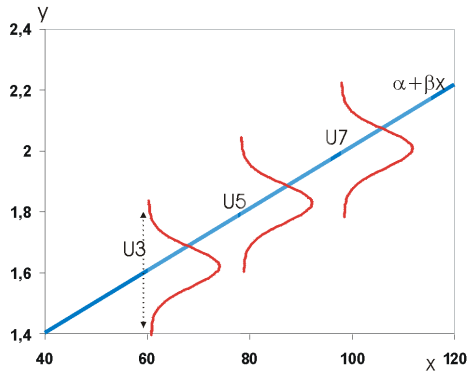

Abbildung 81: Grafik 3: Verteilung der Störgrößen auf der Regressionsgeraden

Axiome des linearen Regressionsmodells

Damit dieses Verfahren sinnvolle Ergebnisse liefert, wurden für das Lineare Regressionsmodell bestimmte verteilungstheoretische Annahmen getroffen. Wir gehen aus von der Beziehung

$$y_i = \alpha + \beta x_i + u_i \ .$$

und definieren die Störgröße u_i als Zufallsvariable. Die Annahmen des linearen Regressionsmodell sind

1. Alle u_i haben den Erwartungswert Null: $Eu_i = 0$, (i = 1, ..., n) .
2. Alle u_i haben die gleiche Varianz (Homoskedastizität): $varu_i = varu_j$ (i,j = 1, ... , n, i & ne; j) .
3. Die u_i sind sämtlich stochastisch unabhängig voneinander.

Der Sinn dieser Axiome wird weiter unten erläutert.

Minimierung

Die herkömmliche Methode, die sich auf der Basis der Axiome ergibt, ist die Minimum-Quadrat-Methode oder Methode der kleinsten Quadrate. Man minimiert also die summierten Quadrate der Residuen,

$$RSS = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \rightarrow min!$$

bezüglich a und b.

Wir multiplizieren die Klammer aus:

$$S = \sum_{i=1}^{n} (y_i - a - bx_i) \cdot (y_i - a - bx_i)$$

$$= \sum_{i=1}^{n} (y_i^2 - y_i a - y_i b x_i - a y_i + a^2 + a b x_i - y_i b x_i + a b x_i + b^2 x_i^2)$$

$= \sum_{i=1}^{n} (y_i^2 - 2 y_i a - 2 y_i b x_i + a^2 + 2 a b x_i + b^2 x_i^2)$
$= \sum_{i=1}^{n} y_i^2 - 2a \sum_{i=1}^{n} y_i - 2b \sum_{i=1}^{n} y_i x_i + n a^2 + 2ab \sum_{i=1}^{n} x_i + b^2 \sum_{i=1}^{n} x_i^2.$

Wir minimieren durch Ableiten

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} y_i + 2na + 2b \sum_{i=1}^{n} x_i,$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} x_i y_i + 2a \sum_{i=1}^{n} x_i + 2b \sum_{i=1}^{n} x_i^2,$$

und Nullsetzen, was ein wenig optisch geschönt die Normalgleichungen

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

ergibt.

Wir erhalten die gesuchten Regressionskoeffizienten als die Lösungen

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

und

$$a = \bar{y} - b\bar{x} \ ,$$

wobei $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ der Mittelwert, also der Durchschnitt der x-Daten ist, y entsprechend. Wegen des Verschiebungssatzes kann man b auch darstellen als

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

oder, nach Erweiterung des Bruchs durch 1/(n-1),

$$b = \frac{s_{xy}}{s_x^2}$$

mit s_{xy} als Kovarianz zwischen den $x_i$und y_i und $s_x{}^2$ als Varianz der x_i. Man nennt diese Schätzungen auch **Kleinste-Quadrate-Schätzer**, **KQ**- oder **OLS-Schätzer**.

Wir wollen nun für das obige Sektbeispiel die Regressionskoeffizienten bestimmen:

Preis einer Flasche	**verkaufte Menge**	**x_i- x**	**y_i- y**				
x_i	**y_i**	**x***	**y***	**x*y***	**x*x***	**y*y***	**$\hat{y}$**
20	0	5	-5	-25	25	25	0,09
16	3	1	-2	-2	1	4	4,02
15	7	0	2	0	0	4	5,00
16	4	1	-1	-1	1	1	4,02
13	6	-2	1	-2	4	1	6,96
10	10	-5	5	-25	25	25	9,91
90	**30**	**0**	**0**	**-55**	**56**	**60**	**30,00**

Wir berechnen in dem Beispiel zunächst die arithmetischen Mittel als die Koordinaten des Schwerpunktes der n Messwerte bzw. der Punktwolke.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6}\ 90 = 15\ ,$$

entsprechend :$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6}\ 30 = 5\ ,$

und dann die Regressionskoeffizienten

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-55}{56} = -0,98$$

als die Steigung der Regressionsgeraden

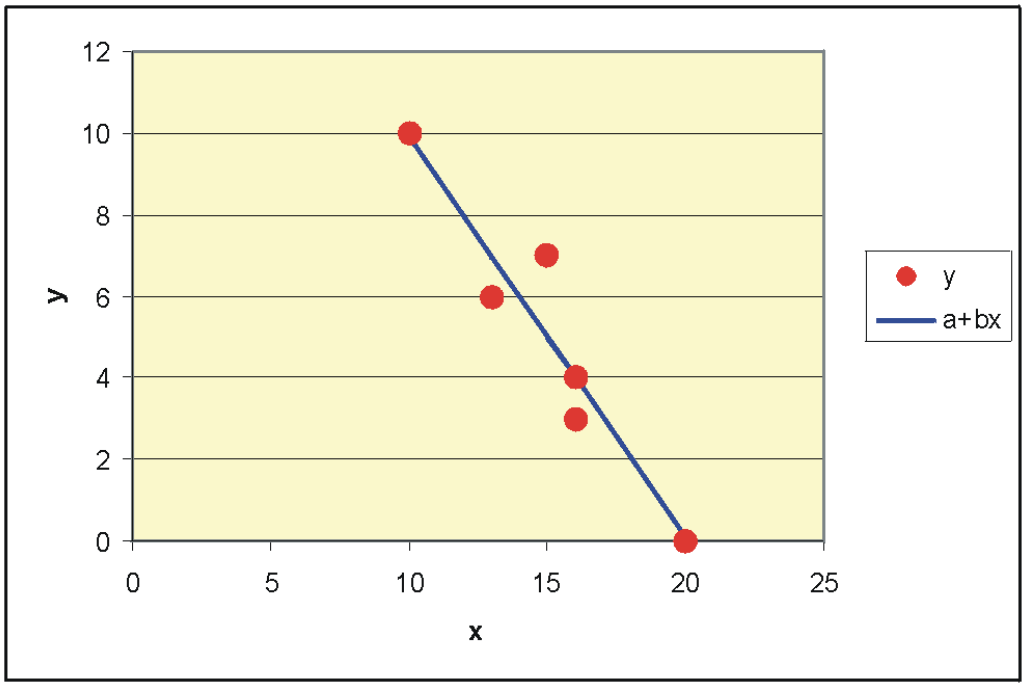

Abbildung 82: Regressionsgerade a+bx

und

$$a = \bar{y} - b\bar{x} = 5 + 0,98 \cdot 15 = 19,7$$

Die geschätzte Regressionsgerade lautet $\hat{y} = 19,73 - 0,98x$, so dass man vermuten kann, dass bei jedem Euro mehr der Absatz im Durchschnitt um ca. 1 Flasche sinkt.

Für die gegebenen x-Werte erhalten wir als Schätzungen $\hat{y}$

$\hat{y}_1 = a + bx_1 = 19{,}73 - 0{,}98 \cdot 20 = 0{,}09$ $\hat{y}_2 = a + bx_2 = 19{,}73 - 0{,}98 \cdot 16 = 4{,}02$ . . . $\hat{y}_6 = a + bx_6 = 19{,}73 - 0{,}98 \cdot 10 = 9{,}91$

Für die beobachteten Absatzwerte y bleibt das Residuum r_i übrig:

$y_1 = a + bx_1 + d_1 = \hat{y}_1 + d_1 \rightarrow d_1 = y_1 - \hat{y}_1 = 0 - 0{,}09 = -0{,}09$ $y_2 = a + bx_2 + d_2 = \hat{y}_2 + d_2 \rightarrow d_2 = y_2 - \hat{y}_2 = 3 - 4{,}02 = -1{,}02$ . . . $y_6 = a + bx_6 + d_6 = \hat{y}_6 + d_6 \rightarrow d_6 = y_6 - \hat{y}_6 = 10 - 9{,}91 = 0{,}09$.

Schätzung der Varianzen

Die Stichprobenvarianz der Residuen berechnet sich als:

$s^2 = \frac{1}{n-2} \sum_i d_i^2$

Man schätzt damit die Varianz der Störgröße u (eigentlich U!).

Gesetzmäßigkeiten

Bezüglich der Zielvariablen und der Residuen gilt:

- $\sum_i d_i = 0$ und damit $\bar{d} = 0$.

  Die Residuen sind im Mittel Null, sie enthalten also keine Information mehr.

- $\sum_i x_i d_i = 0$

  Die unabhängige Variable x und die Residuen sind orthogonal. Geometrisch bedeutet das, dass sie senkrecht aufeinander stehen. Sie sind daher unkorreliert. Die Residuen enthalten also keinerlei Information mehr, die in x steckt. Die Informationen aus x sind alle in a + bx. Nur, was von x nicht mehr erklärt werden kann, bleibt als Rest in d.

- $\bar{\hat{y}} = \frac{1}{n} \sum_i (y_i - d_i) = \frac{1}{n} \sum_i y_i - \frac{1}{n} \sum_i d_i = \bar{y}$.

  Das arithmetische Mittel der ŷ ist gleich dem Mittel von y.

Vorteile der Minimum-Quadrat-Methode:

- Positive und negative Abweichungen heben sich bei Summenbildung nicht auf.
- Große Residuen werden im Verhältnis stärker gewichtet als kleine.
- Der Durchschnitt der Residuen ist Null.
- Die Regressionskoeffizienten können mit einer Formel berechnet werden.

Nachteil der Minimum-Quadrat-Methode:

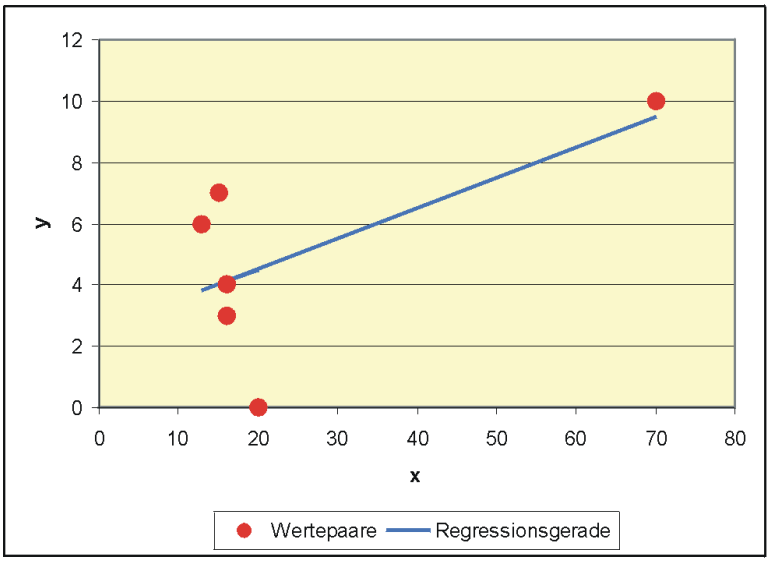

Abbildung 83: Ein Ausreißer bei x: Die Gerade wird nach oben gezogen

Nicht resistent gegenüber Ausreißern in den Daten.

Sekt-Beispiel:

Wegen eines Erhebungsfehlers wurde für x_6 statt 10 der Wert 70 eingetippt. Die neue Regressionsgerade ergibt sich als $\hat{y} = 2{,}51 + 0{,}10x$. Dieser Ausreißer beeinträchtigt das Ergebnis so sehr, dass sogar das Vorzeichen der Steigung umgedreht wird. Eigentlich sollte die Regressionsgerade durch die Punktwolke auf der linken Seite der Grafik führen und fallend sein. Der Ausreißer hebelt die Gerade regelrecht aus: Man spricht von einem High-Leverage-Value, also einem Wert mit großer Hebelkraft. Wir erkennen sofort, dass dieser Ausreißer die Analyse völlig wertlos gemacht hat. In dem speziellen Sachzusammenhang könnte man sogar einen fatalen Fehlschluss machen: Bei Luxusgütern sind steigende Preis-Absatz-Funktionen denkbar, weil ein hoher Preis statusfördernd ist. Man könnte also fälschlicherweise annehmen, dass dieser Zusammenhang auch hier gilt. Man würde also einen sehr hohen Preis festlegen und am Markt scheitern.

Bestimmtheitsmaß

Ein Kriterium für die Beurteilung der Güte einer Regressionsschätzung ist das Bestimmtheitsmaß. Die Begründung für dieses Maß leitet sich aus der

sog. Streuungszerlegung her. Die Gesamtvarianz von y läßt sich, ausgehend von der Beziehung

$$y_i = \hat{y}_i + d_i$$

zerlegen in die durch a + bx erklärte Varianz von y und die nicht erklärte Varianz:

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2 \ .$$

Einige Umformungen ergeben das Bestimmtheitsmaß

$$r^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

als Anteil der erklärten Streuung an der Gesamtstreuung von y. Es ist

$$r^2 = \frac{(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \ ,$$

wobei ersichlich ist, daß r^2 das Quadrat des Korrelationskoeffizienten von x und y darstellt. Mit dem Verschiebungssatz erhalten wir

$$r^2 = \frac{(\sum_{i=1}^{n} x_i y_i - n \cdot \bar{x} \cdot \bar{y})^2}{(\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n \cdot \bar{y}^2)}.$$

Es gilt:

$$0 \leq r^2 \leq 1$$

Je näher r^2 bei 1 ist, desto größer ist der Anteil der erklärten Streuung, desto besser wird y durch x erklärt. $r^2 = 0$ bedeutet, dass x und y unkorreliert sind, und $r^2 = 1$, dass x und y eine Gerade bilden.

Die Berechnung der Varianz der Residuen von Hand mit der Formel

$$s^2 = \frac{1}{n-2} \sum_i d_i^2$$

ist aufwendig, weil zuerst die Residuen ermittelt werden müssen. Eine vereinfachte Form leitet sich aus der Beziehung

$$r^2 = 1 - \frac{\sum_{i=1}^n d_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

her. Es ist dann nämlich

$$s^2 = \frac{1}{n-2}(1 - r^2) \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Sekt-Beispiel

Da hier die arithmetischen Durchschnitte glatte Werte sind, wollen wir das Bestimmtheitsmaß mit der Formel

$$r^2 = \frac{(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \ ,$$

ermitteln. Die Quadratsummen wurden oben in der Tabelle bereits ausgerechnet. Wir erhalten

$$r^2 = \frac{-55^2}{56 \cdot 60} = 0,9003 \ .$$

Man könnte also sagen, dass etwa 90% der Information in y von x stammen, die restlichen 10% haben andere Ursachen.

Anforderungen an das Regressionsmodell

Das Regressionsmodell kann nur optimale Ergebnisse liefern, wenn bestimmte Anforderungen erfüllt sind. Diese Anforderungen lassen sich aus dem Axiomensystem des klassischen linearen Regressionsmodells herleiten:

Die Residuen sollen nur rein zufällig streuen und keinerlei Systematik mehr enthalten, d.h. die Zielvariable y soll durch x vollständig erklärt werden. Systematik in den Residuen deutet daraufhin, daß das Modell möglicherweise falsch bestimmt wurde, d.h. es liegt ein Spezifikationsfehler vor.

Als bestes Mittel zur Überprüfung dieser Modellvoraussetzungen wird das (x;y)-Streudiagramm angesehen, das schnell einen optischen Eindruck von der Verteilung der Störgröße vermittelt.

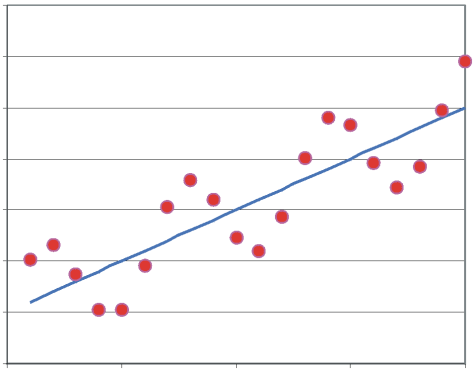

Abbildung 84: **Korrelierte Residuen**:
In den Residuen ist noch ein Schwingungskomponente, die man ev. mit dem Ansatz $y = a + b_1x + b_2\sin(x)$ einbinden könnte.

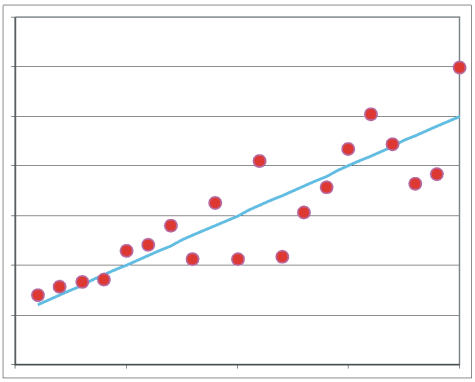

Abbildung 85: **Verschiedene Varianz der Residuen**:
Die linken Residuen schwanken schwächer als die rechten. Vermutlich sind zwei verschiedene Populationen gemischt worden.

Prognose

Ein Ziel der Regressionsanalyse ist die Prognose $\hat{y}_0$, d.h. man fragt danach, welchen Wert y annimmt, wenn ein bestimmtes x_0 vorgegeben ist:

$$\hat{y}_0 = a + bx_0$$

Sekt-Beispiel: Wieviel Flaschen Sekt werden im Durchschnitt verkauft, wenn der Preis auf $x_0 = 11$ Euros festgelegt wird? Es ergibt sich der Prognosewert

$$\hat{y}_0 = 19,7321 - 0,9821 \cdot 11 = 8,93$$

Das heißt jetzt aber nicht, dass in jedem Laden genau 8,93 Flaschen verkauft werden, was auch schwierig sein dürfte, sondern dass in einem Laden durchschnittlich 8,93 Flaschen abgesetzt werden.

Je weiter x_0 vom „Zentrum“ x der Daten entfernt ist, desto unverläßlicher werden die Prognosen - ihre Varianz wird immer größer. Deshalb sollte man sich bei einer Prognose nicht zu weit von den Daten entfernen.

Multiple Regression

Beispiel mit demografischen Daten ausgewählter Länder:

Row i	popgrow	fertil	explife	infmort	mort	birth
y	x_1	x_2	x_3	x_4	x_5	x_6
1	0,14	8,90	9,56	1,35	78,87	4,68
2	2,57	44,46	18,79	6,28	44,20	98,67
3	0,47	18,64	12,16	2,08	56,01	68,78
4	1,01	15,77	5,71	2,06	76,38	9,05
5	1,52	18,99	4,32	2,33	76,63	10,26
6	2,74	33,09	5,66	4,40	68,26	52,71
7	0,41	11,89	9,51	1,78	79,25	3,73
8	0,41	10,90	10,37	1,46	77,35	5,13
9	1,71	9,63	4,05	1,04	81,53	2,28
10	0,29	10,88	10,19	1,66	78,27	5,22

Erklärung der Variablen:

birth	Geburtenrate (pro 1000 Einwohner)
mort	Sterblichkeit (pro 1000 Einwohner)
popgrow	Wachstumsrate der Bevölkerung
fertil	Fertilität (Durchschn. Kinderzahl pro gebärfähiger Frau)
explife	Lebenserwartung (Jahre)

infmort	Kindersterblichkeit (pro 1000 Lebendgeburten)

Es wurden die demografischen Daten für n=10 zufällig ausgewählte Länder erhoben ([https://www.cia.gov/library/publications/the-world-factbook/index.html Quelle: Worldfact-Book der CIA])

Es soll nun das Bevölkerungswachstum *popgrow* erklärt werden. Es wird zunächst als erklärende Variable die Geburtenrate *birth* versucht:

$$popgrow = a + b \cdot birth$$

bzw. $y = a + bx$.

Wir erhalten die Regressionsgerade

$$popgrow = -0,104 + 0,0672 \cdot birth$$

mit einem Bestimmtheitsmaß von 66,4%. Die Information in *popgrow* wird also zu 66% durch *birth* erklärt, die restlichen 34% entstammen anderen Einflussgrößen. Wir machen einen zweiten Versuch und verwenden die Sterblichkeit als erklärende Variable:

$$popgrow = a + b \cdot mort + d \ .$$

.

Hier ergibt sich als Regressionsgerade

$$popgrow = 1,16 - 0,0032 \cdot mort + d$$

mit einem Bestimmtheitsmaß von ca. 0%. Dieses Ergebnis ist enttäuschend und auch das vorherige war nicht gerade berauschend. Jetzt versuchen wir mal was Innovatives: Wir machen einen Regressionsansatz mit zwei unabhängigen Variablen

$$popgrow = b_0 + b_1 \cdot birth + b_2 \cdot mort + d$$

bzw. $y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + d$.

Gesucht ist also die geschätzte **Regressionsebene**

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

.

Wir erhalten das Gleichungssystem

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + d_1,$$

$$y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + d_2,$$

$$y_3 = b_0 + b_1 x_{31} + b_2 x_{32} + d_3,$$

$$\ldots$$

$$y_{10} = b_0 + b_1 x_{10,1} + b_2 x_{10,2} + d_{10} \ .$$

Wir wollen nun die einzelnen Daten zu Matrizen zusammenfassen. Wir erhalten die (10x3)-Datenmatrix

$$\underline{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & & \vdots \\ 1 & x_{10,1} & x_{10,2} \end{pmatrix} = \begin{pmatrix} 1 & 8,90 & 9,56 \\ 1 & 44,46 & 18,79 \\ 1 & 18,64 & 12,16 \\ 1 & 15,77 & 5,71 \\ 1 & 18,99 & 4,32 \\ 1 & 33,09 & 5,66 \\ 1 & 11,89 & 9,51 \\ 1 & 10,90 & 10,37 \\ 1 & 9,63 & 4,05 \\ 1 & 10,88 & 10,19 \end{pmatrix}$$

und die Vektoren

$$\underline{y} = \begin{pmatrix} 0,0014 \\ 0,0257 \\ 0,0047 \\ 0,0101 \\ 0,0152 \\ 0,0274 \\ 0,0041 \\ 0,0041 \\ 0,0171 \\ 0,0029 \end{pmatrix},$$

$$\underline{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \quad \text{und} \quad \underline{d} = \begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_{10} \end{pmatrix}.$$

Mit diesen Matrizen können wir das Gleichungssystem in Matrixschreibweise darstellen als

$$\underline{y} = \underline{X}\underline{b} + \underline{d}$$

wobei Vektoren und Matrizen unterstrichen sind.

Auch hier wird die Quadratsumme der Residuen minimiert, um die Regressionskoeffizienten zu erhalten. Diese berechnen sich mit der Formel

$$\underline{b} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$$

.

Wir erhalten den Vektor der Regressionskoeffizienten

$$\underline{b} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y} = \begin{pmatrix} 0,556 \\ 0,089 \\ -0,117 \end{pmatrix} ,$$

also

$$popgrow = 0,556 + 0,089 \cdot birth - 0,117 \cdot mort$$

bzw.

$\hat{y} = 0,556 + 0,089x_1 - 0,117x_2$.

Das Bestimmtheitmaß ist hier 88,8%. Die Anpassung hat sich beträchtlich verbessert. Hier scheint das Zusammenwirken der beiden Regressoren mehr bewirkt zu haben als ||die Summe der Teile||.

Die Wurzel aus dem Bestimmtheitsmaß ergibt den **multiplen Korrelationskoeffizienten** r = 0,942. Der multiple Korrelationskoeffizient kann nur zwischen 0 und 1 liegen, wobei 1 wieder vollständige Korrelation bedeutet.

Die Regressionskoeffizienten 0,089 und 0,117 sind die **partiellen Ableitungen** der Regressionsebene. Man könnte die Koeffizienten so interpretieren: Steigt bei konstanter Sterblichkeit die Geburtenrate um einen Punkt, erhöht sich das Bevölkerungswachstum um ca. 0,1 Prozent. Steigt dagegen bei konstanter Geburtenrate die Sterblichkeit um einen Punkt, sinkt das Bevölkerungswachstum um ca. einen Punkt. Eine simultane Analyse der Regressionsebene bezüglich beider Regressionskoeffizienten ist kompliziert und meistens auch nicht sinnvoll interpretierbar. Die Analyse eines Regressionskoeffizienten bei Konstanthaltung der übrigen Regressoren nennt man eine **Ceteris-Paribus-Analyse**.

In der Regel ist die Berechnung der Regressionskoeffizienten im multiplen linearen Regressionsmodell so aufwendig, daß Computer eingesetzt werden müssen. Spezielle statistische Datenbanksysteme wie SPSS,SAS oder Minitab ermöglichen eine umfassende Regressionsanalyse.

Die Vor- und Nachteile der Minimum-Quadrat-Methode sind dieselben wie bei der Einfachregression: Es sei $x_{8,2} = 100$ statt 10,9. Man erhält

$$popgrow = 1,13 + 0,0031 \cdot birth - 0,0092 \cdot mort$$

mit einem Bestimmtheitsmaß von 0,7%.

Einführung

Zeitreihen sind Beobachtungen, die im Lauf der Zeit erhoben wurden. Bei der Analyse von Zeitreihen versuchen wir, die Beobachtungen durch den Faktor Zeit zu erklären. Wir suchen nach bestimmten Gesetzmäßigkeiten, nach denen diese Zeitreihen zustande kommen.

Für die optische Unterstützung stellen wir eine Zeitreihe als Streudiagramm dar. Um den Verlauf, die Entwicklung des Merkmals darstellen, können wir die Punkte zu einer Kurve (Polygonzug) verbinden.

Wir haben hier beispielsweise das Bruttoinlandsprodukt der Bundesrepublik Deutschland (Quelle: © Statistisches Bundesamt Deutschland 2005) der Quartale 2001 bis 2005 gegeben.

Stichtag	Mrz 01	Jun 01	Sep 01	Dez 01	Mrz 02	Jun 02
BIP	514,51	522,63	531,51	544,91	519,19	531,66

Stichtag	Sep 02	Dez 02	Mrz 03	Jun 03	Sep 03	Dez 03
BIP	546,06	551,9	524,4	533,59	550,76	556,12

Stichtag	Mrz 04	Jun 04	Sep 04	Dez 04	Mrz 05
BIP	537,36	547,85	557,21	564,82	539,78

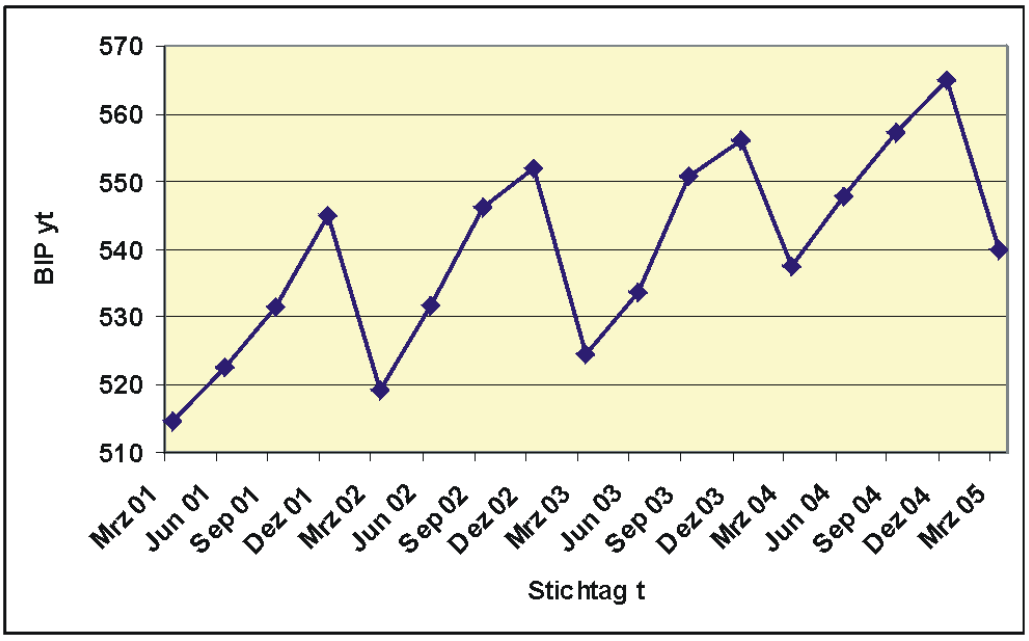

Abbildung 86: Zeitreihe des deutschen Bruttoinlandsprodukts (Milliarden €) für die Quartale der Jahre 2001 bis 2005

Modell der Zeitreihe

Die Zeitreihenanalyse erfordert die Konzipierung der Zeitreihe als Modell:

Wir betrachten einen Beobachtungszeitraum mit T vielen Zeitpunkten t. Zu einem Zeitpunkt t gehört die Beobachtung y_t des Merkmals y.

Da Zeitangaben häufig unhandlich bei der Berechnung sind (z. B. 1.3.1996), empfiehlt es sich, die Zeitpunkte durchzunummerieren, z.B. t = 1, 2, ... , n.

Beispiel Großhandel

Es liegen n = 60 Quartalsumsätze des Gartenbedarfsgroßhandels Rosalinde vor. Die Quartale sind durchnummeriert als t = 1, ... , 60. Es sind hier nur die ersten Beobachtungen wiedergegeben. Die komplette Zeitreihe befindet sich in Zeitreihe Rosalinde.

Stichtag zum Ende des Monats	Quartal	Umsatz in Mio. €	Linearer Trend
Mrz 90	1	52,19	42
Jun 90	2	48,69	44
Sep 90	3	49,28	46
	...	...	...

Stichtag zum Ende des Monats	Saisonaler Zyklus	Konjunktureller Zyklus	Restschwankung
Mrz 90	6,00	3,06	1,13
Jun 90	0,00	5,66	-0,96
Sep 90	-6,00	7,39	1,89
	...	...	...

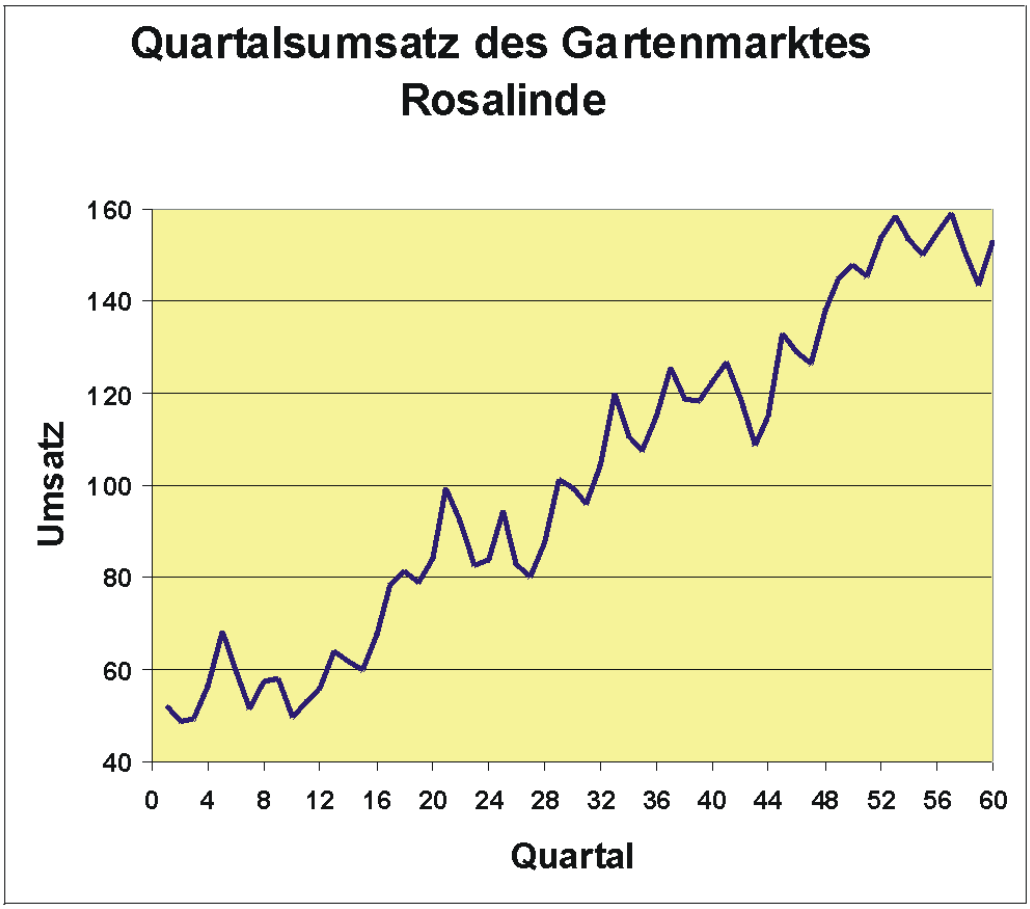

Abbildung 87: Zeitreihe der Quartalsumsätze des Großhandels Rosalinde

Wir sehen, dass die Tendenz der Umsätze steigend ist. Es scheint sich außerdem ein vermutlich konjunktureller Zyklus abzuzeichnen, der z. B. 1992 ein Tief und 1995 ein Hoch hatte. Und es ist deutlich ein einjähriger, saisonaler Zyklus zu erkennen, der auch aus der Tabelle ersichtlich ist.

Wir können also die Komponenten der Zeitreihe unterscheiden:

- Trend Q
- Konjunkturelle Schwankung K
- Saisonale Schwankung S
- Restschwankung r

Sind diese Komponenten unabhängig voneinander, gehen wir vom additiven Modell aus:

$$y = Q + K + S + r$$

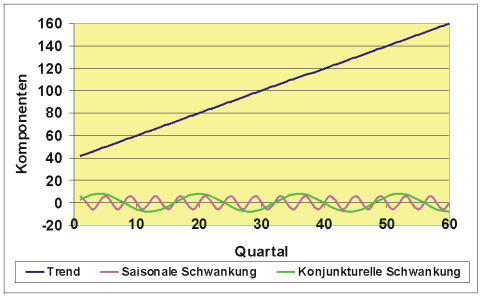

Abbildung 88: Zerlegung der Zeitreihe Rosalinde in die einzelnen Komponenten

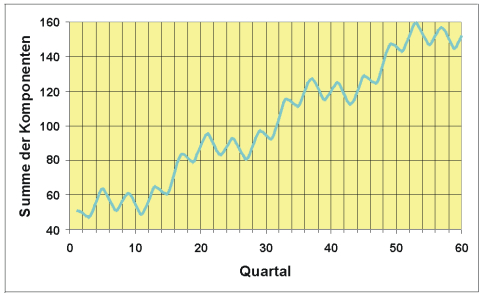

Abbildung 89: Summe der Zeitreihenkomponenten ohne Restschwankung

Oft überlagern sich mehrere zyklische Schwankungen. Es gibt hier spezielle Verfahren, die Zyklen zu identifizieren.

Ein Problem in der Zeitreihenanalyse ist die Wahl des richtigen Modells. Bei den einfacheren Modellen beschränkt man sich meist auf die Bestimmung einer glatten Komponente, die aus Trend und/oder konjunktureller Komponente gebildet wird, einer saisonalen Komponente und die Restschwankung.

Üblicherweise wird bei der Schätzung des Trends Q und der Saisonkomponente S so vorgegangen, dass zuerst der Trend Q bestimmt wird. Es wird dann y vom Trend bereinigt, d.h. von Beobachtungen y_t werden die Trendwerte Q_t abgezogen. Aus den resultierenden Restwerten wird dann die saisonale Komponente errechnet. Man kann auch beide Komponenten mit Hilfe der multiplen Regression auf einmal bestimmen.

Schätzung des Trends mit der Regressionsgerade

Wenn wir von einem **linear verlaufenden Trend** ausgehen können, schätzen wir ihn mit dem Regressionsmodell

$\hat{y}_t = a + bt$ bzw. $y_t = a + bt + d_t$ $(t = 1, 2, ..., T; y_t = y_1, y_2, ..., y_T)$

mit den Lösungen

$$b = \frac{\sum_{t=1}^{T}(t-\bar{t})(y_t-\overline{y})}{\sum_{t=1}^{T}(t-\bar{t})^2}$$

$$= \frac{\sum_{t=1}^{T} t \cdot y_t - T \cdot \bar{t} \cdot \overline{y}}{\sum_{t=1}^{T} t^2 - T \cdot \bar{t}^2}$$

$$= \frac{\sum_{t=1}^{T} t y_t - \frac{T(T+1)}{2}\overline{y}}{\frac{1}{12}(T^3-T)}$$

und

$$a = \overline{y} - b \cdot \bar{t}$$

$$= \overline{y} - b \cdot \frac{T+1}{2}$$

.

Die Trendwerte Q_t sind dann

$$Q_t = \hat{y}_t = a + bt$$

.

Beispiel Herrenbekleidung

Die monatlichen Aufträge für die letzten 3 Jahre eines Herstellers für Herrenbekleidung (in 1000 Stück) sind durch die unten folgende Zeitreihe in der Zeitreihe Herrenbekleidung gegeben, von der ein Ausschnitt vorliegt. Die Grafik zeigt, daß offensichtlich ein steigender linearer Trend mit saisonalem Jahreszyklus vorliegt.

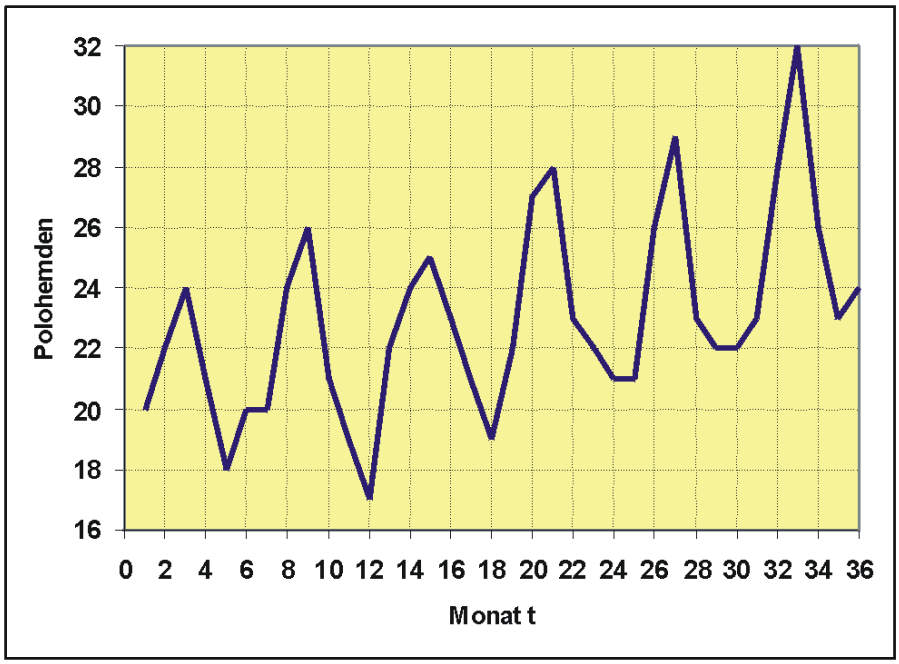

Abbildung 90: Monatliche Aufträge für Polohemden eines Herstellers für Herrenbekleidung

t	y_t	$t \cdot y_t$	t^2
1	20	20	1
2	22	44	4
3	24	72	9
4	21	84	16
...	...	...	...
666	828	15889	16206

Wir ermitteln zuerst die arithmetischen Durchschnitte:

$\overline{t} = \frac{666}{36} = 18,5$ und entsprechend $\overline{y}_t = 23$. Dann erhalten wir für den Regressionsansatz

$$\hat{y}_t = a + bt$$

die Regressionskoeffizienten nach dem Verschiebungssatz

$$b = \frac{15889 - 36 \cdot 18,5 \cdot 23}{16206 - 36 \cdot 18,5^2} = 0,1470$$

und

$$a = \overline{y} - b \cdot \overline{t} = 23 - 0,1470 \cdot 18,5 = 20,2810$$

Die geschätzten Trendwerte sind $\hat{y}_t = a + bt$, z.B.

$$\hat{y}_1 = 20,2810 + 0,1470 \cdot 1 \approx 20,43$$

,

$$\hat{y}_2 = 20,2810 + 0,1470 \cdot 2 \approx 20,57$$

,

usw.

Die Residuen sind

$$y_1 - \hat{y}_1 = 20 - 20,43 = -0,43$$

,

$$y_2 - \hat{y}_2 = 22 - 20,57 = 1,43$$

,

usw.

t	y_t	a + bt	d_t
1	20	20,43	-0,43
2	22	20,57	1,43
3	24	20,72	3,28
4	21	20,87	0,13
5	18	21,02	-3,02
6	20	21,16	-1,16
...	...	...	...
34	26	25,28	0,72
35	23	25,43	-2,43
36	24	25,57	-1,57

Liegt ein nichtlinearer Trendverlauf vor, kann auch ein nichtlinearer Regressionsansatz gewählt werden. Es können neben t auch andere exogene Variablen in das Modell aufgenomen werden.

Schätzung der Saisonkomponente

Gehen wir von dem additiven Modell

$$y_t = Q_t + S_t + r_t$$

aus, bleibt nach Schätzung der Trendkomponente Q noch die Abweichung

$$d_t = y_t - Q_t$$

übrig, die sich zusammensetzt aus

$$d_t = S_t + r_t$$

Wir nennen deshalb d_t auch den trendbereinigten Zeitreihenwert. Es soll nun noch die saisonale Komponente S_t ermittelt werden. Wir könnten etwa versuchen, diese zyklische Komponente mit einer Sinusfunktion zu schätzen.

Einfacher ist aber folgendes Vorgehen: Wir ermitteln die trendbereinigten Zeitreihenwerte d_t. Dann wird aus allen Werten d_t, die die gleiche Saison betreffen, ein arithmetischer Durchschnitt gebildet, der als Schätzung für die saisonale Komponente verwendet wird.

Beispiel Herrenbekleidung

Für die Januar-Saisonkomponente werden alle Januarwerte der d_t gemittelt:

$$S_{jan} = S_1 = S_{13} = S_{25} = \frac{-0,43 - 0,19 - 2,96}{3} = -1,19$$

usw.

$$r_t = y_t - Q_t - S_t$$

ergibt dann die nichterklärte Restschwankung.

Wir können jetzt eine Prognose für den Zeitpunkt T+k ermitteln als

$$\hat{y}_{T+k} = Q_{T+k} + S_{T+k},$$

wobei wir für S_t den Saisonwert für diejenige Saison wählen, die in T+k auftritt.

Beispiel für eine Prognose:

Wir wollen für März des 4. Jahres eine Prognose des Auftragseingangs machen. Es handelt sich um den Zeitpunkt t = 39.

Wir erhalten den Trend als

$$Q_{39} = 20,281 + 39 \cdot 0,147 = 26,014$$

und die Saisonkomponente als

$$S_3 = \frac{3,28 + 2,51 + 4,75}{3} = 3,51$$

.

Die Prognose errechnet sich nun als

$$26,014 + 3,51 = 29,524$$

.

Multiplikative Verknüpfung der Zeitreihen-Komponenten

Bisher wurde von einer additiven Überlagerung des Trends durch die Saisonkomponente ausgegangen, d.h. die Komponenten wurden als unabhängig angesehen. Häufig nehmen aber die zyklischen Schwankungen mit steigendem Trend zu. Es könnte hier beispielsweise das multiplikative Modell

$$y_t = Q_t \cdot K_t \cdot r_t$$

vorliegen. Wir können den Ansatz logarithmieren und erhalten

$$\log y_t = \log Q_t + \log S_t + \log r_t$$

Mit dem logarithmierten Ansatz führen wir die Zerlegung des Modells in seine Komponenten durch, wie oben beschrieben.

Schätzung der glatten Komponente mit gleitenden Durchschnitten

Lässt sich die Trendkomponente des Zeitreihenmodells offensichtlich durch keine funktionale lineare oder nichtlineare Beziehung darstellen, kann man eine **glatte Komponente** mit Hilfe gleitender Durchschnitte bestimmen.

Gleitende Durchschnitte ungeradzahliger Ordnung

Beispiel Hotelaufenthalte (G. D. 3. O)

In einem Kurhotel werden Ende April, Ende August und Ende Dezember die Zahl der Hotelaufenthalte festgehalten. Es wurde mit Ende Dezember begonnen.

Stichtag	t	Aufenthalte y_t
Dez 89	1	408
Apr 90	2	372
Aug 90	3	480
Dez 90	4	444
Apr 91	5	447
Aug 91	6	492
Dez 91	7	429
Apr 92	8	411
Aug 92	9	486
Dez 92	10	525
Apr 93	11	495

Zur Ermittlung des Trends wurden gleitende Durchschnitte 3. Ordnung gebildet:

$$\overline{y}_2 = \frac{1}{3} \cdot (408 + 372 + 480) = 420$$

$$\overline{y}_3 = \frac{1}{3} \cdot (372 + 480 + 444) = 432$$

$$\ldots$$

$$\overline{y}_9 = \frac{1}{3} \cdot (411 + 486 + 525) = 474$$

$$\overline{y}_{10} = \frac{1}{3} \cdot (486 + 525 + 495) = 502$$

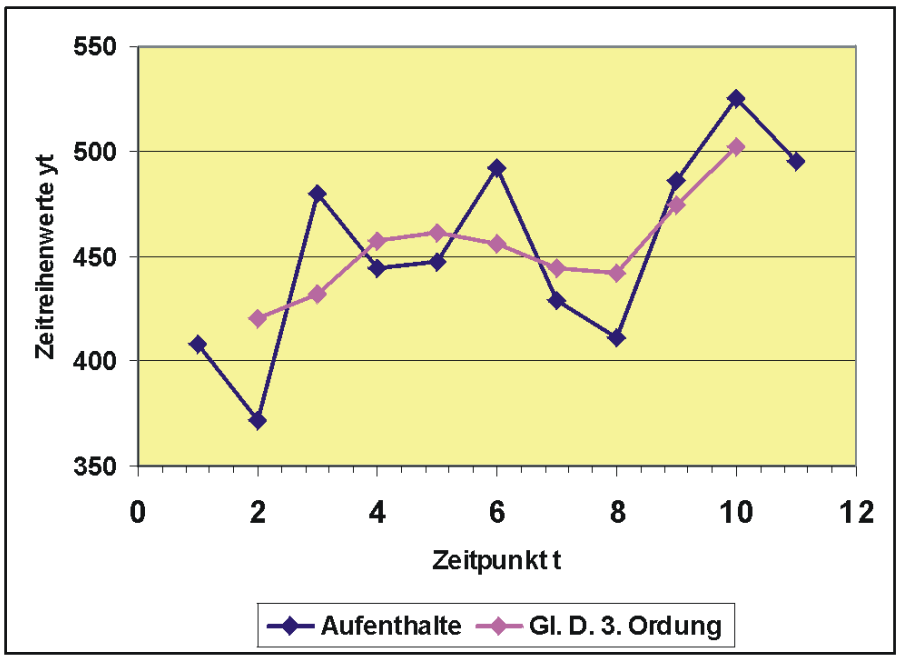

Abbildung 91: Gleitende Durchschnitte 3. Ordnung für die Zahl der Hotelaufenthalte

Stichtag	t	Aufenthalte y_t	y_t
Dez 89	1	408	
Apr 90	2	372	420
Aug 90	3	480	432
Dez 90	4	444	457
Apr 91	5	447	461
Aug 91	6	492	456
Dez 91	7	429	444
Apr 92	8	411	442
Aug 92	9	486	474
Dez 92	10	525	502
Apr 93	11	495	

Der Index t der Glättung y_t entspricht immer dem Beobachtungswert in der Mitte der beteiligten Zeitreihenwerte.

Man sieht, dass die gleitenden Durchschnitte die starken Schwankungen glätten und man den Trend, oder besser die glatte Komponente, besser erkennt.

Die Zahl der beteiligten Beobachtungen gibt die Ordnung des Durchschnitts an. Man berechnet einen gleitenden Durchschnitt 3. Ordnung folgendermaßen:

$$\overline{y}_2 = \frac{y_1 + y_2 + y_3}{3}$$

$$\overline{y}_3 = \frac{y_2 + y_3 + y_4}{3}$$

$$\ldots$$

$$\overline{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3}$$

$$\ldots$$

$$\overline{y}_{n-1} = \frac{y_{n-2} + y_{n-1} + y_n}{3}$$

Entsprechend ergeben sich gleitende Durchschnitte 5. Ordnung als

$$\overline{y}_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$$

$$\overline{y}_4 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5}$$

...

$$\overline{y}_t = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5}$$

...

usw.

Beispiel Hotelaufenthalte (G. D. 5. O)

Wir berechnen die gleitenden Durchschnitte 5. Ordnung als

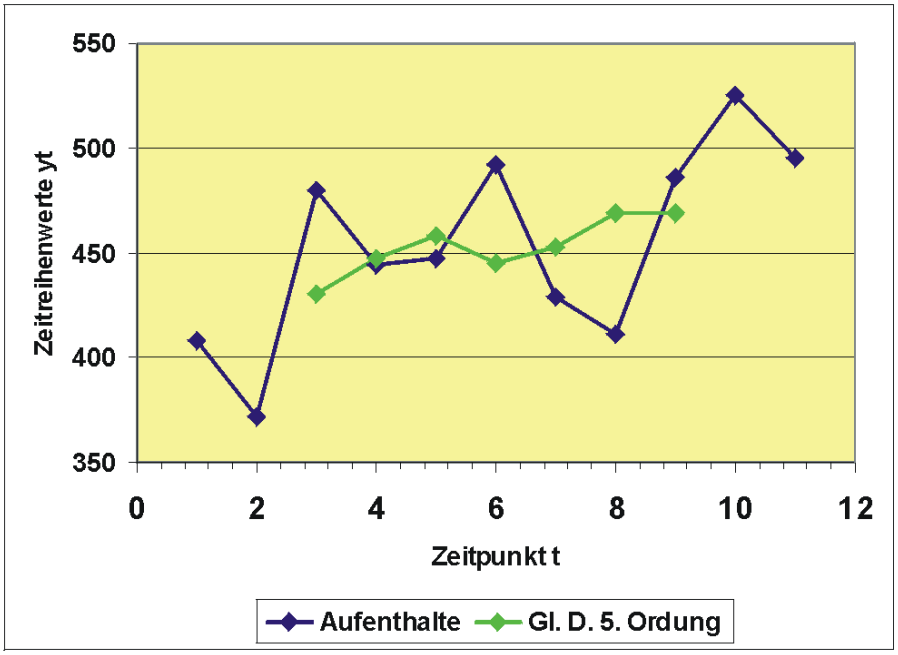

Abbildung 92: Gleitende Durchschnitte 5. Ordnung für die Zahl der Hotelaufenthalte

$$\overline{y}_3 = \frac{408 + 372 + 480 + 444 + 447}{5}$$

$$\overline{y}_4 = \frac{372 + 480 + 444 + 447 + 492}{5}$$

usw., also

Stichtag	t	Aufenthalte yt	y_t
Dez 89	1	408	
Apr 90	2	372	
Aug 90	3	480	430,2
Dez 90	4	444	447
Apr 91	5	447	458,4
Aug 91	6	492	444,6
Dez 91	7	429	453
Apr 92	8	411	468,6
Aug 92	9	486	469,2
Dez 92	10	525	
Apr 93	11	495	

Zur Prognose über den Beobachtungszeitraum hinaus sind gleitende Durchschnitte nicht so recht geeignet, da die Randwerte der Zeitreihe nicht geschätzt werden. Allerdings gibt es Verfahren, mit denen man diese Werte durch eine Gewichtung der benachbarten Werte ausfüllen kann.

Gleitende Durchschnitte geradzahliger Ordnung

Die Rechtfertigung für gleitende Durchschnitte als Schätzung für die glatte Komponente begründet sich darin, daß sich saisonale Schwankungen kompensieren, d.h. daß sich positive und negative Abweichungen der Saisonkomponente vom Trend aufheben. Das ist allerdings nur dann gegeben, wenn die Breite, also die Ordnung des gleitenden Durchschnitts genau einen Zyklus umfaßt und wenn alle Zyklen die gleiche Länge aufweisen. Sonst können gleitende Durchschnitte verzerrte Schätzungen liefern.

Die Breite saisonaler Schwankungen ist allerdings meist geradzahlig, z.B. Quartale, Monate usw. Es entsteht hier das Problem, daß bei der Durchschnittsbildung ein Schätzwert $y_{t+0,5}$ zwischen zwei Beobachtungswerte y_t und y_{t+1} zu liegen kommt, was meist unerwünscht ist. Man behilft sich hier so, daß man den Durchschnitt auf der Zeitachse um eine halbe Zeiteinheit nach rechts verschiebt. Die beiden Randwerte, die nur teilweise erfaßt werden, gehen mit einem Gewicht von 0,5 in den Durchschnitt ein,

z.B. statt

$$\overline{y}_{3,5} = \frac{y_2 + y_3 + y_4 + y_5}{4}$$

berechnet man

$$\overline{y}_4 = \frac{\frac{y_2}{2} + y_3 + y_4 + y_5 + \frac{y_6}{2}}{4}$$

Beispiel

Es liegen 12 Quartalswerte vor und es sollen gleitende Durchschnitte 4. Ordnung ermittelt werden.

t	yt	gleitender Durchschnitt 4. Ordnung
1	12	
2	8	
3	12	12
4	14	13
5	16	14,125
6	12	15,625
7	17	17,5
8	21	19
9	24	20,5
10	16	21,5
11	25	
12	21	

Z.B.

$$\overline{y}_3 = \frac{\frac{12}{2} + 8 + 12 + 14 + \frac{16}{2}}{4} = 12$$

$$\overline{y}_4 = \frac{\frac{8}{2} + 12 + 14 + 16 + \frac{12}{2}}{4} = 13$$

.

Exponentielle Glättung

Lässt eine Zeitreihe keinerlei systematisches Muster wie linearen Anstieg oder Ähnliches erkennen, kann man versuchen, mit der exponentiellen Glättung eine glatte Komponente nachzubilden. Insbesondere kann man damit eine Prognose für den Zeitpunkt T + 1 erhalten.

Das Verfahren wird beispielsweise in der Lagerhaltung verwendet, wenn es etwa darum geht, den Bedarf eines zu bestellenden Artikels im kommenden Jahr zu ermitteln. So hat etwa die Schweizer Armee mit der exponentiellen Glättung gute Erfolge bei der Ermittlung der benötigten Gewehre im folgenden Jahr gemacht.

Man geht von dem Ansatz aus, dass der gegenwärtige Zeitreihenwert immer auch von den vergangenen Werten beeinflusst wird, wobei sich der Einfluss abschwächt, je weiter der Wert in der Vergangenheit liegt.

Formales Modell

Gegeben ist eine Zeitreihe mit den Beobachtungen y_1, y_2, ..., y_t, ... zu den Zeitpunkten t. Im Zeitpunkt t wird für y_t ein geglätteter Schätzwert y^*_t errechnet, der sich als gewichteter Durchschnitt ergibt aus dem aktuellen Zeitreihenwert y_t und dem Schätzwert der Vorperiode y^*_{t-1}. Die Gewichtung wird durch den Glättungsfaktor α bestimmt, wobei $0 \leq \alpha \leq 1$ sein muss. Man erhält

$$y_t^* = \alpha \cdot y_t + (1 - \alpha) \cdot y*_{t-1} \quad .$$

Die Zeitreihe baut sich so rekursiv auf. Theoretisch ist die laufende Zeitreihe beim Zeitpunkt t bereits unendlich lang. Für die praktische Ermittlung des geglätteten Wertes wird man allerdings einen Startwert y^*_0 vorgeben und von da an die geglättete Zeitreihe ermitteln.

Baut man nun, beginnend bei y^*_0, die geglättete Zeitreihe auf,

$$y_1* = \alpha y_1 + (1-\alpha) y_0^* \ ,$$

$$y_2* = \alpha y_2 + (1-\alpha) y_1^* \ ,$$

$$y_3* = \alpha y_3 + (1-\alpha) y_2^* \ ,$$

...

erhält man, wenn man die Rekursivität auflöst,

$$y_t^* = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + ... + \alpha(1-\alpha)^{t-1} y_1 + (1-\alpha)^t y_0 \ .$$

Man sieht, wie wegen $\alpha < 1$ die Einflüsse der Vergangenheit immer mehr verschwinden.

Der Schätzwert y^*_t liefert den Prognosewert für den Zeitpunkt t+1. Liegt dann im Zeitpunkt t + 1 eine neue Beobachtung vor, kann die Prognose für t + 2 ermittelt werden usw.

Für die Wahl des Glättungsfaktors wird häufig 0,2 bis 0,3 empfohlen. Man kann aber auch mit Hilfe der Regressionsanalyse den Glättungsfaktor schätzen.

Einfaches Zahlenbeispiel

Es sind die Zeitreihenwerte y_1. ..., y_{10} gegeben, wie unten in der Tabelle aufgeführt. Diese Werte sollen exponentiell geglättet werden. Es wurde ein Glättungskoeffizient von $\alpha = 0{,}3$ gewählt und man benötigt einen Anfangswert, der hier $y_0^* = 19$ betragen soll. Wir beginnen

$$y_1* = 0,3 \cdot 20 + 0,7 \cdot 19 = 6 + 13,3 = 19,3$$

$$y_2* = 0,3 \cdot 18 + 0,7 \cdot 19,3 = 18,91$$

usw. In der Tabelle sind die Glättungen für ausgewählte Werte von α aufgeführt.

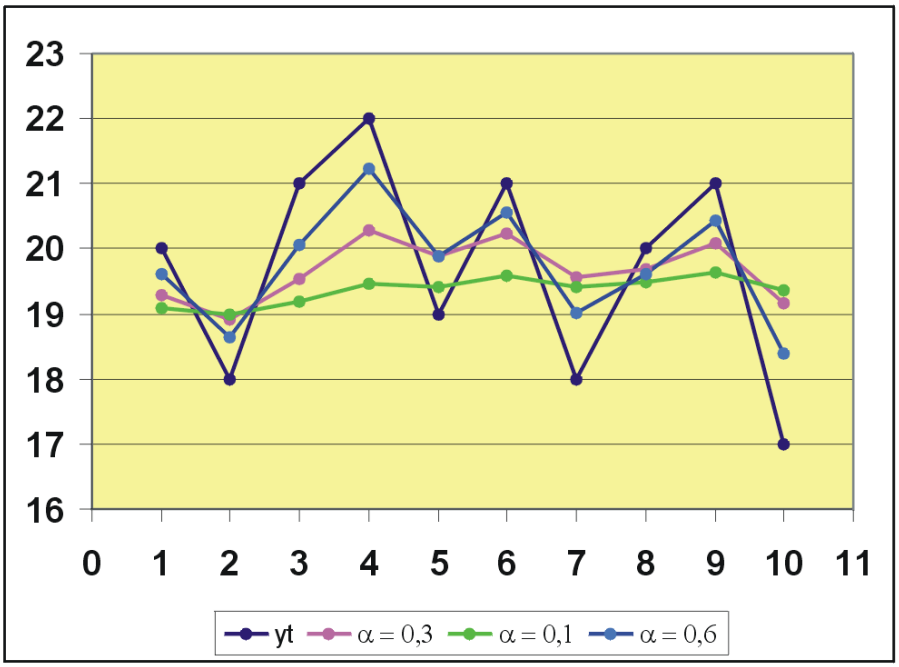

Abbildung 93: Zeitreihe mit exponentiell geglätteten Werten

t	y_t	y_t *	y_t *	y_t *
		$\alpha = 0,3$	$\alpha = 0,1$	$\alpha = 0,6$
0	-	19	19	19
1	20	19,3	19,1	19,6
2	18	18,91	18,99	18,64
3	21	19,54	19,19	20,06
4	22	20,28	19,47	21,22
5	19	19,89	19,42	19,89
6	21	20,23	19,58	20,56

7	18	19,56	19,42	19,02
8	20	19,69	19,48	19,61
9	21	20,08	19,63	20,44
10	17	19,16	19,37	18,38

Die Graphik zeigt die Glättung für $\alpha = 0{,}3$ und $\alpha = 0{,}7$. Man sieht, dass der kleinere Glättungsfaktor die Zeitreihe stärker glättet, denn hier geht der aktuelle Wert nur mit einem Gewicht von 0,3 ein, wogegen die ||mittleren|| Vergangenheitswerte mit 0,7 berücksichtigt werden.

Beispiel für den exponentiell geglätteten DAX

Es soll mit den monatlichen Durchschnittswerten des Aktienindex DAX für die Monate Januar 1977 bis August 1978 eine exponentielle Glättung berechnet werden. Die Daten liegen nebst den geglätteten Zeitreihenwerten in der Tabelle vor:

DAX-Werte und ihre exponentielle Glättung ($\alpha = 0{,}3$) Monat Zeitpunkt t DAX Vt Glättung y*t 1977 Jan 0 512,3 512,3 1977 Feb 1 496,2 507,5 1977 Mrz 2 509,8 508,2 1977 Apr 3 551,9 521,3 1977 Mai 4 539,9 526,9 1977 Jun 5 524,9 526,3 1977 Jul 6 530,3 527,5 1977 Aug 7 540,9 531,5 1977 Sep 8 541,3 534,4 1977 Okt 9 554,2 540,4 1977 NOV 10 557,5 545,5 1977 Dez 11 549,34 546,7 1978 Jan 12 549,4 547,5 1978 Feb 13 552,9 549,1 1978 Mrz 14 549,7 549,3 1978 Apr 15 532,1 544,1 1978 Mai 16 545,5 544,5 1978 Jun 17 553,0 547,1 1978 Jul 18 582,1 557,6 1978 Aug 19 583,1 565,2

Der erste Wert wird mit 512,3 als Startwert y^*_0 genommen. Wir verwenden einen Glättungsfaktor $\alpha = 0{,}3$.

Es ergeben sich die geglätteten Werte

$$y_1^* = 0,3 \cdot 496,2 + 0,7 \cdot 512,3 = 507,5 \ ,$$

$$y_2^* = 0,3 \cdot 509,8 + 0,7 \cdot 507,5 = 508,2 \ ,$$

$$y_3^* = 0,3 \cdot 551,9 + 0,7 \cdot 508,2 = 521,3 \ ,$$

...

Die Schätzung y^*_1 ist jetzt der Prognosewert für die Periode 2 und so weiter.

Exponentielle Glättung bei trendbehafteten Werten

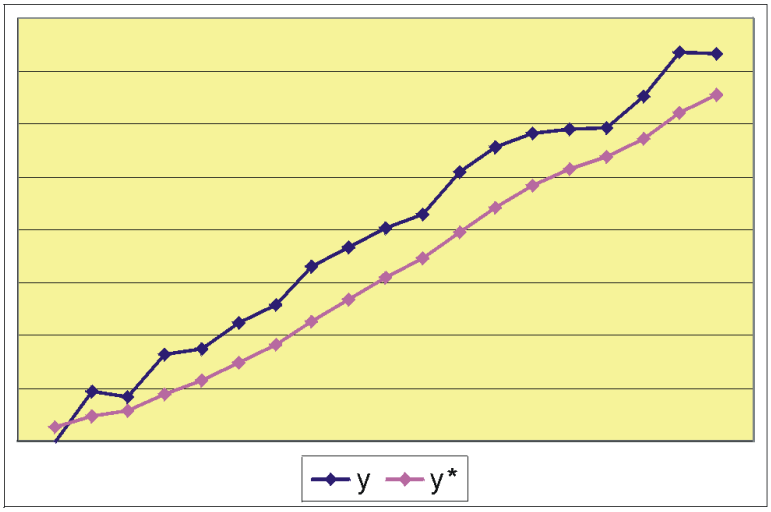

Abbildung 94: Die geglätteten Prognosewerte y* liegen systematisch unter den beobachteten trendbehafteten Zeitreihenwerten y

Die exponentielle Glättung ist dann ein empfehlenswertes Verfahren, wenn die Zeitreihenwerte einen chaotischen Eindruck machen und keinerlei Systematik erkennen lassen. Liegen allerdings Beobachtungen vor, die einen Trend beinhalten, d.h. die laufend steigen oder fallen, ||schleppen|| die geglätteten Werte ||hinterher||. Man sieht in der Grafik deutlich, wie die Schätzwerte immer systematisch unter den beobachteten Werten liegen.

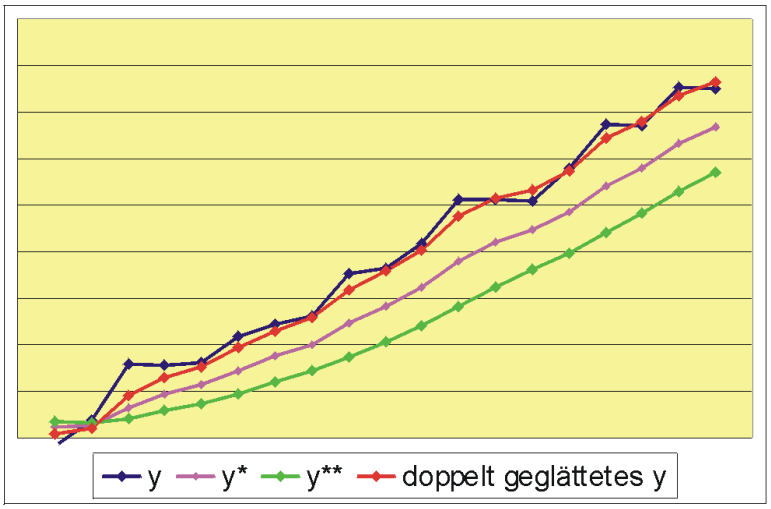

Abbildung 95: Mit doppelt geglätteten Zeitreihen erreicht man eine korrekte Prognose der trendbehafteten y-Werte

Eine zufriedenstellende Lösung für das Problem, daß bei einem steigenden (fallenden) Trend die Zeitreihenwerte systematisch unterschätzt (überschätzt) werden, bieten gleitende Durchschnitte zweiter Ordung. Hier werden die bereits einmal geglätteten Werte noch einmal einer Glättung unterzogen. Man erhält den Schätzwert y^{**}, der sich analog zu oben berechnet aus

$$y_t^{**} = \alpha \cdot y_t^* + (1 - \alpha) \cdot y_{t-1}^{**}$$

Für einen brauchbaren Prognosewert für Periode t+1 muss man dann bestimmen

$$\widehat{y}_{t+1} = 2 \cdot y_t^* - y_{t-1}^{**}$$

.

Abbildung 96: Graph der einfach geglätteten DAX-Werte.
(Copyright: Deutsche Bundesbank, Frankfurt am Main, Deutschland)

In der Grafik der Dax-Kurse liegen beispielsweise zwischen t = 7 und t = 12 die Schätzwerte immer systematisch unter den beobachteten Werten.

Kapitel 7

Maßzahlen

Die **Konzentration** befasst sich mit der Intensität, mit der sich ein Objekt auf eine vorgegebene Menge verteilt. Eine typische Aussage der Konzentrationsmessung wäre etwa: 20% der Menschen eines bestimmten Staates besitzen 90% des Vermögens. Demnach teilen sich die anderen 80% die restlichen 10%. Hier kann man von einer starken Konzentration sprechen.

Kino-Beispiel

Im Rahmen einer Controllinganalyse eines Kinos wurden die Besucherzahlen (Merkmal x) für die 5 angebotenen Spielfilme an einem Tag erfasst. Man erhielt die Tabelle

Filmtitel	Zahl der Besucher x
Rotkäppchen	25
Verliebt ins Abendrot	75
Leif Erikson	125
Söhne der Alhambra	250
Galaxy-Fighter	525

Definitionen

Es gibt verschiedene Verfahren zur Konzentrationsmessung. Man kann die Konzentration grafisch darstellen oder Kennwerte berechnen. Die Merkmalsbeträge x müssen aufsteigend geordnet vorliegen, also $x_{[1]} \leq x_{[2]} \leq ... \leq x_{[n]}$.

Für die Konzentrationsmessung werden neben der relativen Summenfunktion S_i* folgende Definitionen benötigt:

- Merkmalssumme $\sum_i x_i = n \cdot \overline{x}$
- Kumulierte Merkmalsbeträge $q_i = \sum_{k=1}^{i} x_{[k]}$
- Relative kumulierte Merkmalsbeträge $q_i* = \frac{q_i}{n\overline{x}}$

Grafik

Die Lorenzkurve ist eine grafische Darstellung der Konzentration:

Die Wertepaare (S_i*;q_i*) werden in einem Diagramm abgetragen. Das erste Wertepaar ist (0;0), das letzte (1;1). Es wird zwischen diesen beiden Wertepaaren die Winkelhalbierende des Koordinatensystems eingetragen. Alle Wertepaare (0;0), (S_1*;q_1*), ... , (1;1) werden geradlinig verbunden.

Tabelle

Die für die Lorenzkurve benötigten Zwischenwerte werde in der folgenden Tabelle aufgeführt. So ergibt sich beispielsweise für die kumulierten Merkmalsbeträge q_i

$$q_1 = 25$$

, $q_2 = 25 + 75 = 100$, $q_3 = 100 + 125 = 225$ usw.

Die relativen oder anteiligen Merkmalsbeträge errechnen sich durch Teilen des Gesamtmerkmalbetrags 1000, also

$$q_1* = \frac{25}{1000} = 0,025$$

usw.

Ebenso ermitteln wir die absolute Summenhäufigkeiten als Zahl der Filme, also

$$S_1 = 1$$

, $S_2 = 1 + 1 = 2$, $S_3 = 2 + 1 = 3$ ...

und wiederum die relative Summenhäufigkeit mit

$$S_1^* = \frac{1}{5} = 0,2$$

, $S_2^* = \frac{2}{5} = 0,4$, ...

Es wurde außerdem noch als Platzhalter die Zeile für i = 0 eingefügt.

i	Filmtitel	x_i	q_i	q_i*	S_i	S_i*
0		0	0	0	0	0
1	Rotkäppchen	25	25	0,025	1	0,2
2	Verliebt ins Abend-rot	75	100	0,100	2	0,4
3	Leif Erikson	125	225	0,225	3	0,6
4	Söhne der Al-hambra	250	475	0,475	4	0,8
5	Galaxy-Fighter	525	1000	1,000	5	1
Summe		1000				

So wurden beispielsweise 40% (S_2*) der Filme von nur 10% (q_2*) der Besucher angesehen.

Die Lorenzkurve ist eine grafisches Maß für das Ausmaß einer Konzentration. Je weiter die Kurve ||durchhängt||, desto größer ist die Konzentration. Unten sind die beiden extremen Situationen dargestellt, die gleichmäßge Aufteilung der Objekte auf die gesamte Menge und die vollständige Konzentration, bei

der ein Element alle Objekte auf sich vereint und alle anderen Elemente leer ausgehen.

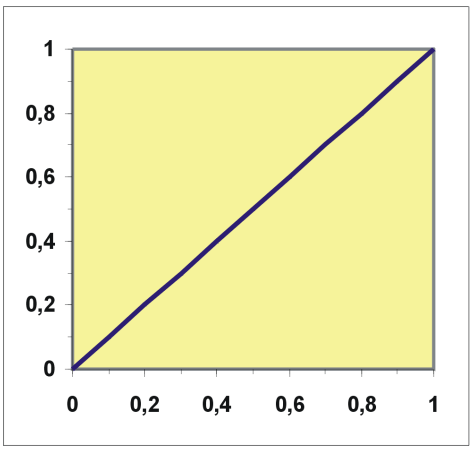

Abbildung 97: Lorenzkurve bei gleichmäßiger Aufteilung

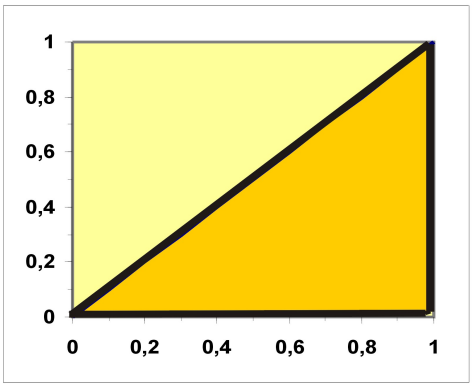

Abbildung 98: Lorenzkurve bei vollständiger Konzentration

Werden mehrere gleichartige Gesamtheiten gegenüberstellt, bieten die verschiedenen Lorenzkurven eine schnelle optische Vergleichsmöglichkeit. Siehe dazu auch das weiter unten folgende Beispiel mit den Agrarflächen in Bayern.

Ginikoeffizient

Als Ginikoeffizient G wird bezeichnet der Anteil der Fläche, die durch die Winkelhalbierende und die Lorenzkurve gebildet wird, an der Gesamtfläche unter der Winkelhalbierenden. Wenn vollkommene Konzentration besteht, ist die Fläche über der Lorenzkurve deckungsgleich mit dem Dreieck unter der Winkelhalbierenden. G ist dann 1. Bei fehlender Konzentration ist dann G=0.

Ermittlung des Ginikoeffizienten

Verbindet man die Punkte auf der Lorenzkurve mit den entsprechenden Punkten auf der Winkelhalbierenden, wird klar, dass wir es mit n vielen Trapezen zu tun haben, deren Flächen wir einzeln bestimmen und dann aufsummieren. Die Fläche eines Trapezes, wie in der Grafik angegeben, ermittelt man als

$$F = \frac{1}{2} \cdot (a + c) \cdot h$$

.

Wir wollen die Fläche F3 des Trapezes zwischen den Abszissenwerten (x-Achse) 0,4 und 0,6 ermitteln. Man sieht, dass das Trapez im Vergleich zur obigen Grafik gekippt vorliegt. Die Höhe h ist also die Differenz

$$S *_3 - S*_3 = 0,6 - 0,4 = 0,2$$

.

Wir fassen a als linke Senkrechte von F3 als a auf: Dann ist

$$a = 0,4 - 0,1 = 0,3$$

.

Entsprechend beträgt die rechte Seite c

$$c = 0,6 - 0,225 = 0,375$$

und wir erhalten als Fläche

$$F_2 = (0,3 + 0,375) \cdot 0,5 \cdot 0,2 = 0,0675$$

.

Allgemein: Die obige Fläche ergibt sich dann als

$$\sum_{i=1}^{n} (S *_i - S*_{i-1}) \cdot \frac{1}{2}((S *_i - q*_i) + (S *_{i-1} - q*_{i-1}))$$

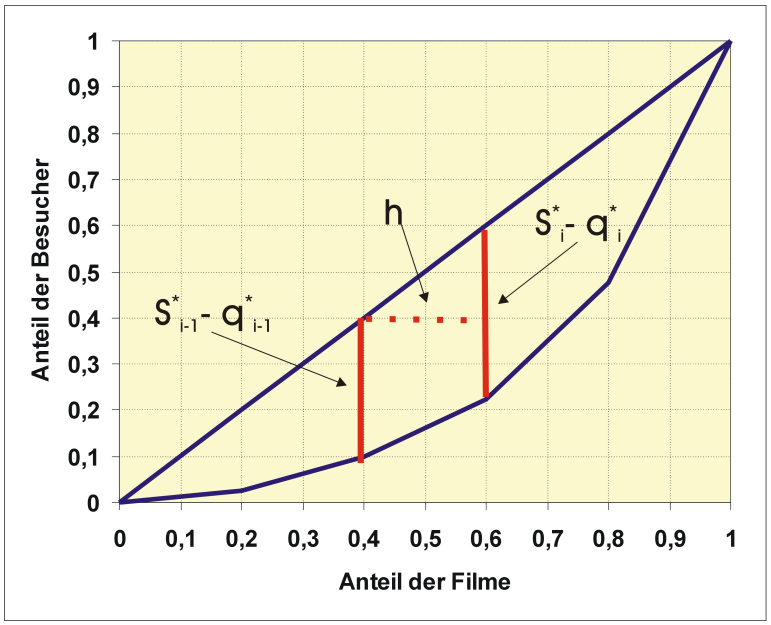

Abbildung 99: Ginikoeffizient: Ermittlung einer Trapezfläche für i=3

Es folgt beispielhaft die Berechnung des Gini in der Tabelle. Mit Tabellenkalkulation kann der Ginikoeffizient leicht ermittelt werden. Wir erhalten schließlich für den Ginikoeffizienten

$$G = \frac{0,235}{0,5} = 0,47$$

i	q^*_i	S^*_i	$h^*_i = S_i - S^*_{i-1}$	$a_i = S^*_i - q^*_i$	$c_i = S^*_{i-1} - q^*_{i-1}$	$0,5 \cdot (a_i + c_i)$	$0,5 \cdot (a_i + c_i) \cdot h_i$
–	0	0	-	-	-	-	-
1	0,025	0,2	0,2	0,175	0	0,0875	0,0175
2	0,1	0,4	0,2	0,3	0,175	0,2375	0,0475
3	0,225	0,6	0,2	0,375	0,3	0,3375	0,0675
4	0,475	0,8	0,2	0,325	0,375	0,35	0,07
5	1	1	0,2	0	0,325	0,1625	0,0325
Summe							0,235

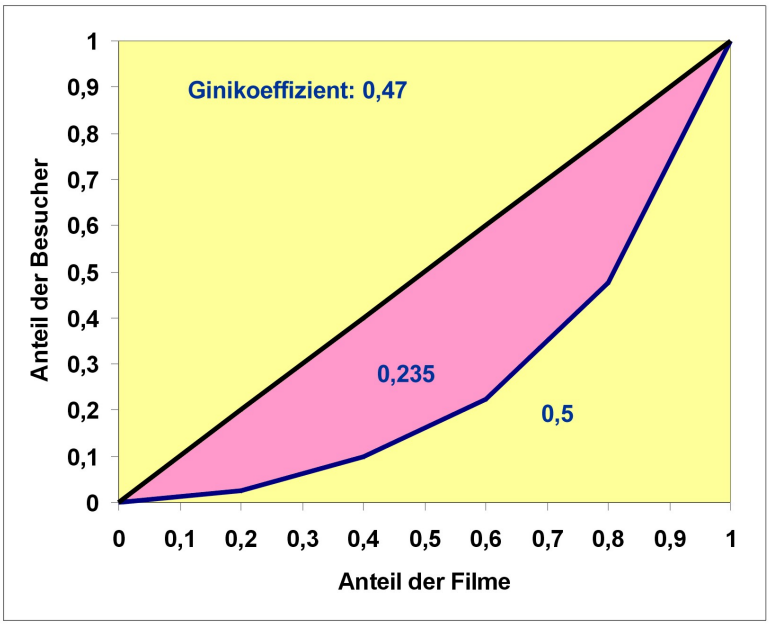

Abbildung 100: Ginikoeffizient

Metrisches Merkmal mit wenig möglichen Ausprägungen

Beispiel

Das interessierende Merkmal ist die Zahl der Autos in einem Haushalt. Es wurden 50 Haushalte befragt.

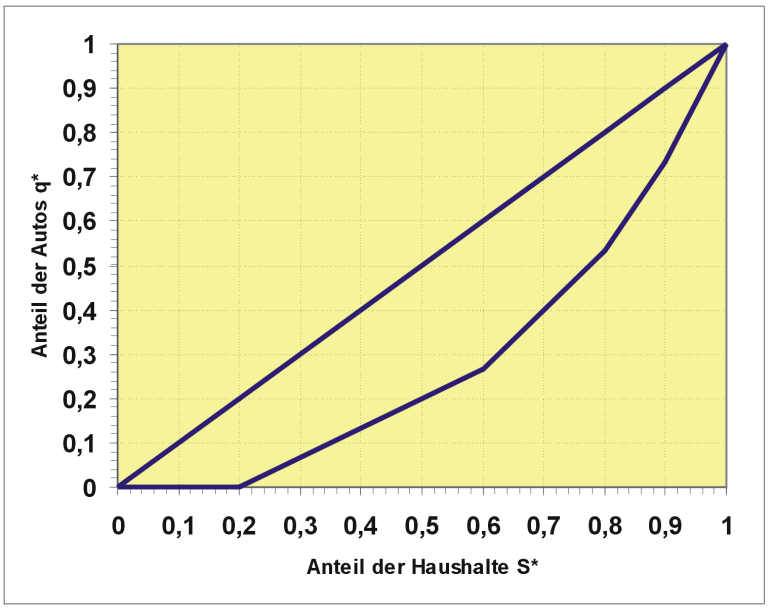

Abbildung 101: Lorenzkurve für die Verteilung der PKWs auf Haushalte

j	x_j	n_j	S_j	S_j*	$x_j n_j$	q_j	q_j*
1	0	10	10	0,2	0	0	0,00
2	1	20	30	0,6	20	20	0,27
3	2	10	40	0,4	20	40	0,53
4	3	5	45	0,9	15	55	0,73
5	4	5	50	1	20	75	1
Summe		50			75		

Lorenzkurve und der Ginikoeffizient berechnen sich im Prinzip wie oben, statt i wird hier der Index j verwendet. Der Merkmalsbetrag x_i wird durch x_j*n_j ersetzt.

Klassiertes Merkmal

Hier wird die Klassenmitte x'_j als Ersatz für den Merkmalswert x_j verwendet.

Beispiel

Landwirtschaftliche Nutzfläche	Zahl der Betriebe (1000)

von ... bis ... unter	1980	2003
2 - 10	112	43
10 - 20	78	34
20 - 30	34	18
30 oder mehr	20	36

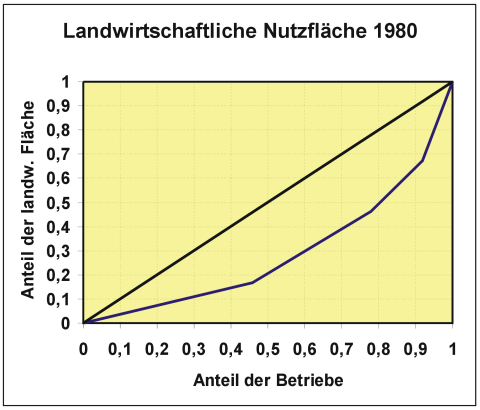

Abbildung 102: Lorenzkurve der Nutzfläche eines bayerischen Landwirtschaftsbetriebes im Jahr 1980

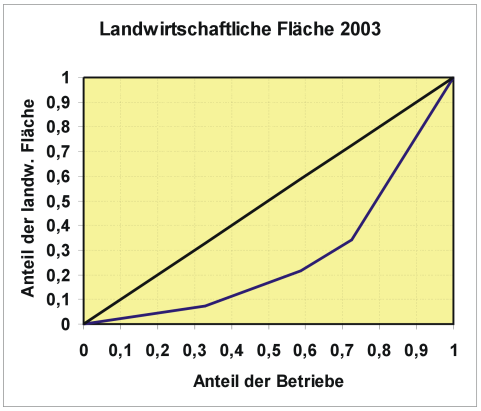

Abbildung 103: Lorenzkurve der Nutzfläche eines bayerischen Landwirtschaftsbetriebes im Jahr 2003

Klasse j von ... bis unter ...	Klassenmitte x_j	n_j	$x_j * n_j$	S_j	S_j^*	q_j	q_j^*
2 - 10	6	112	672	112	0,4590	672	0,1683
10 - 20	15	78	1170	190	0,7787	1842	0,4614
20 - 30	25	34	850	224	0,9180	2692	0,6743
30 - 100	65	20	1300	244	1,0000	3992	1,0000
Summe		244	3992				

Wir erhalten als Ginikoeffizient für das Jahr 1980 den Wert 0,43 und für das Jahr 2003 den Wert 0,46.

Kapitel 8

Schätzen und Testen

Der frühere Inhalt wurde in Mathematik: Statistik: Prinzip des Konfidenzintervalls und Mathematik: Statistik: Ausgewählte Konfidenzintervalle aufgegliedert. Hier wird demnächst etwas Allgemeines über Konfidenzintervalle stehen.

Beispiel mit Absatz von Kaffeepaketen

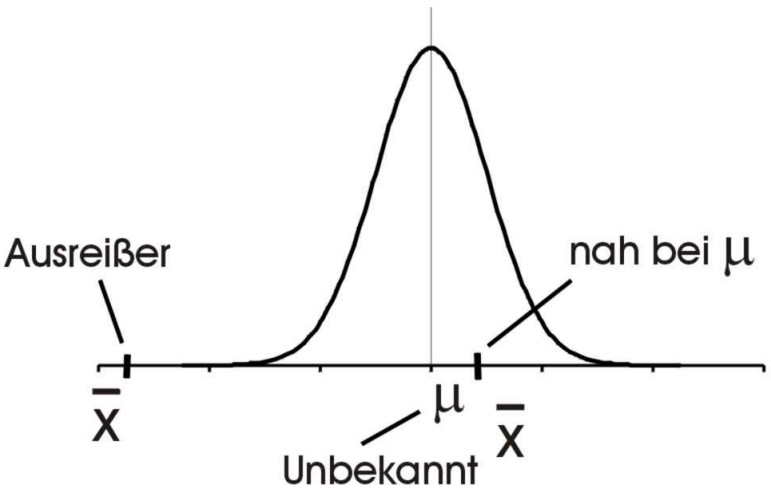

Abbildung 104: Lage einer Schätzung von μ

Beispiel:

Eine Kaffeerösterei möchte eine neue Röstanlage anschaffen. Um beurteilen zu können, ob die Firma den aufzunehmenden Kredit tilgen kann, braucht sie Informationen über den durchschnittlichen monatlichen Absatz an Kaffeepaketen. Pro Monat muss die Firma 20.000 € Annuität zahlen. Zusammen mit den Produktionskosten sollte sie im Durchschnitt auf einen Absatz von 100.000 Kaffeepaketen im Monat kommen. Die Frage ist nun, wird sie es schaffen?

Plausible Überlegungen zur Schätzung

Der durchschnittliche monatliche Absatz von Kaffeepaketen ist unbekannt. Wie könnte man den Durchschnitt ermitteln? Man könnte eine Stichprobe mit z.B. n = 50 Beobachtungen ziehen und versuchen, aus dem arithmetischen Mittel x auf den durchschnittlichen monatlichen Absatz der Grundgesamtheit zu schließen. Ist die Stichprobe groß genug, kann man vermuten, dass der Durchschnitt EX in der Grundgesamtheit, hier μ, in der Nähe von x liegen müsste. Meistens wird x in der Nähe von μ liegen, da aber x die Realisation einer Zufallsvariablen ist, kann in sehr wenigen Fällen x auch extrem weit von μ weg liegen, so daß man dann μ verkehrt einschätzt.

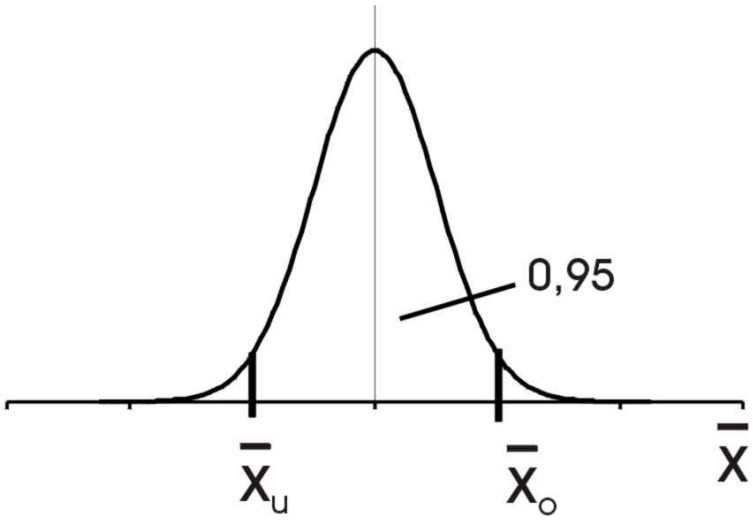

Abbildung 105: 95%-Intervall des durchschnittlichen monatlichen Absatzes

Wir betrachten nun den monatlichen Absatz von Kaffeepaketen (in 1000). Wir bezeichnen ihn als Zufallsvariable X. Es soll der monatliche durchschnittliche Absatz der Kaffeepäckchen geschätzt werden. Bekannt ist lediglich, dass

die Zahl der verkauften Kaffeepakete normalverteilt ist mit einer Varianz 200 [1000^2 Stück^2].

Wie sollen wir nun μ eingrenzen? Wir könnten etwa ein Intervall bestimmen, in dem z.B. 95% aller möglichen x-Werte liegen, also

$$P(\bar{x}_u \leq \bar{X} \leq \bar{x}_o) = 0,95 \; .$$

Damit man dieses Intervall berechnen kann, müssen Informationen über die Verteilung von X verfügbar sein. Es soll eine Stichprobe von n = 50 gezogen werden, d.h. es werden die verkauften Kaffeepakete der letzten 50 Monate erfasst:

$$\bar{X} = \frac{1}{50} \sum_{i=1}^{50} X_i \; .$$

Verteilung des Merkmals und der Schätzfunktion

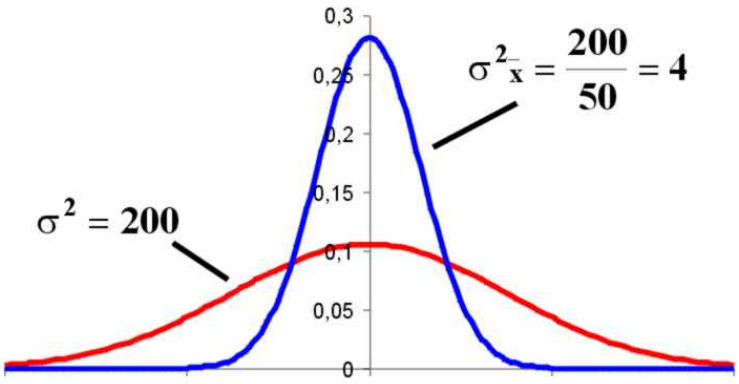

Abbildung 106: Vergleich: Normalverteilung der Zufallsvariablen Absatz X und der Zufallsvariablen Durchschnittlicher Absatz X

Die Zufallsvariable X in der Grundgesamtheit soll normalverteilt sein mit dem Durchschnitt EX = μ und der Varianz VarX = σ^2. Die Varianz soll bekannt sein.

Es wird eine Stichprobe vom Umfang n gezogen. Der Stichprobendurchschnitt X ist selbst eine Zufallsvariable und ist als lineare Transformation von X wiederum normalverteilt, und zwar mit den Parametern

$$E\bar{X} = \mu$$

und $var\bar{X} = \frac{\sigma^2}{n}$.

Hier ist

$$\frac{\sigma^2}{n} = \frac{200}{50} = 4$$

.

Herleitung des Intervalls

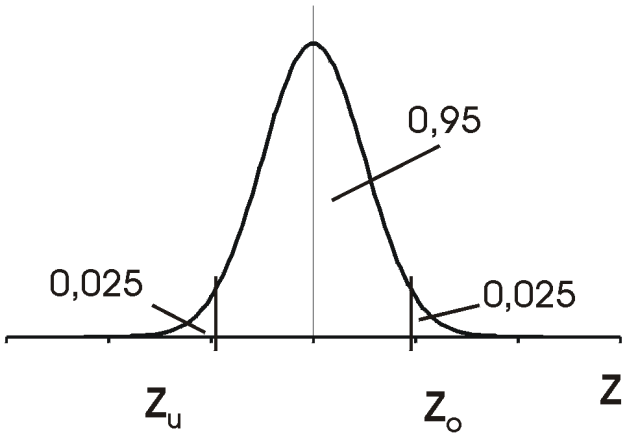

Abbildung 107: Ober- und Untergrenze der standardnormalverteilten Zufallsvariablen Z

Ausgegangen wird von

$$P\left(\bar{x}_u \leq \bar{X} \leq \bar{x}_o\right) = 0,95 \; .$$

Untergrenze u und Obergrenze o sollen nun bestimmt werden. Wir standardisieren zunächst

$Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}-\mu}{\sqrt{4}}$,

so dass sich analog zu oben

$$P\left(z_u \leq Z \leq z_o\right) = 0,95$$

ergibt. z_o ist hier das 0,975-Quantil der Standardnormalverteilung. Ein Blick in die Normalverteilungstabelle verrät uns, dass der z-Wert, der zur Wahrscheinlichkeit 0,975 gehört, 1,96 ist.

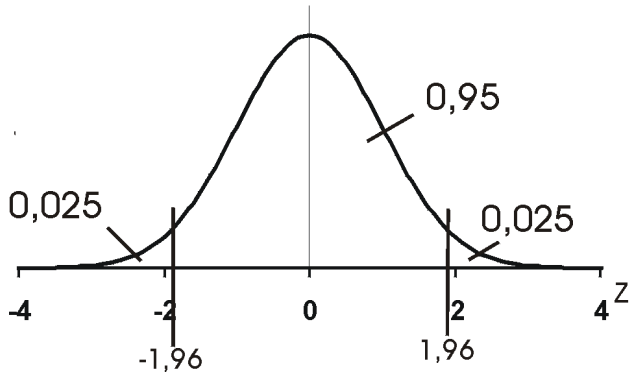

Abbildung 108: (1-$\alpha/2$)-Quantil der Standardnormalverteilung

Wir können jetzt das entsprechende Intervall für Z

$$P\left(-1,96 \leq Z \leq 1,96\right) = 0,95$$

angeben. Die Ungleichung wird bezüglich μ aufgelöst:

$$P\left(-1,96 \leq \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 1,96\right) = 0,95 \; .$$

$$P\left(-1,96 \leq \frac{\bar{X}-\mu}{\sqrt{4}} \leq 1,96\right) = 0,95 \; .$$

$$P\left(-1,96 \cdot 2 \leq \bar{X} - \mu \leq 1,96 \cdot 2\right) = 0,95 \; .$$

$$P\left(-\bar{X} - 1,96 \cdot 2 \leq -\mu \leq -\bar{X} + 1,96 \cdot 2\right) = 0,95 \; .$$

$$P\left(\bar{X} + 1,96 \cdot 2 \geq \mu \geq \bar{X} - 1,96 \cdot 2\right) = 0,95 \; .$$

$$P\left(\bar{X} - 1,96 \cdot 2 \leq \mu \leq \bar{X} + 1,96 \cdot 2\right) = 0,95 \; .$$

Dieses Intervall wird Zufallsintervall genannt, weil es von einer Zufallsvariablen (X) gebildet wird. Wir schreiben jetzt dieses Intervall mit Symbolen:

$$P\left(\overline{X} - z\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z\left(1 - \frac{\alpha}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \; .$$

.

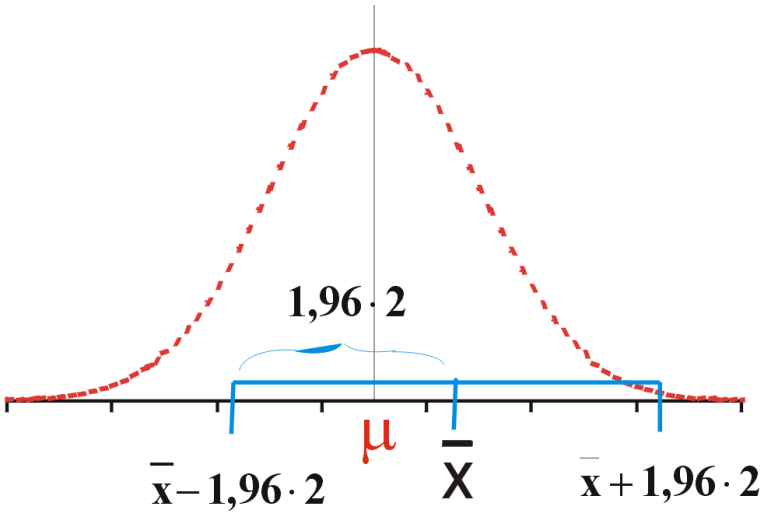

Abbildung 109: Breite des Konfidenzintervalls

Wir bezeichnen $1 - \alpha = 0,95$ als Konfidenzkoeffizient. $\alpha = 0,05$ dagegen ist die Irrtumswahrscheinlichkeit oder das Signifikanzniveau.

Die Breite des Intervalls ist hier

$$2 \cdot (2 \cdot 1,96) = 2 \cdot \frac{\sigma}{\sqrt{n}} \cdot z(0,975) = 7,84 \; .$$

Also schwankt eine X-Schätzung für μ mit einer 95%igen Wahrscheinlichkeit in einem Intervall der Breite von 7840 Kaffeepaketen, d.h. μ befindet sich mit einer 95%igen Wahrscheinlichkeit in diesem Intervall.

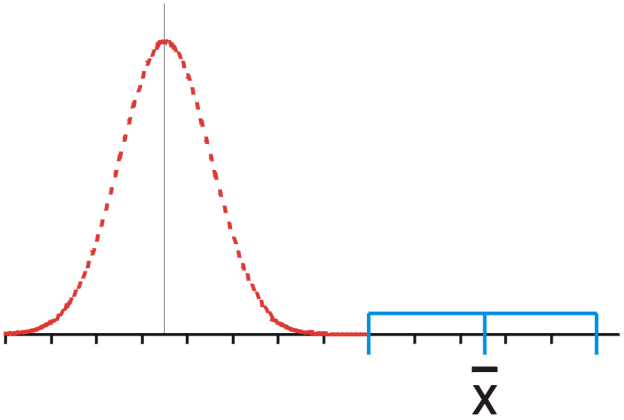

Abbildung 110: x liegt sehr weit vom wahren μ weg

Es kann aber passieren, dass die Schätzung x extrem daneben liegt. In der Grafik wurde mit x daneben gegriffen; dieser Fall durch diese restlichen 5% abgedeckt.

Konkretes 95%-Konfidenzintervall

Es liegt nun ein konkreter Schätzwert von x = 98 vor. Wir erhalten das Konfidenzintervall

$$[\bar{x} - 2 \cdot 1,96; \bar{x} + 2 \cdot 1,96] =$$

$$[98 - 2 \cdot 1,96; 98 + 2 \cdot 1,96] =$$

$$[98 - 3,92; 98 + 3,92] =$$

$$[94,08; 101,92] .$$

Entscheidung: μ kann bei einer Wahrscheinlichkeit von 95% unter 100 liegen, also kann der mittlere Umsatz unter 100.000 liegen. Deshalb sollte die Firma von dieser Investition absehen.

Was wäre, wenn man $[101; 108,84]$ erhalten hätte? Dann wäre eine dauerhafte Liqidität zu vermuten.

Einfluss der Varianz auf das Konfidenzintervall

Was wäre, wenn σ^2 statt 200 den Wert 5000 hätte? Dann wäre

$\bar{X} \to N(\mu; \frac{5000}{50} = 100)$.

Wir erhielten das Konfidenzintervall

$$[\bar{x} - 1,96 \cdot \sqrt{100}; \bar{x} + 1,96 \cdot \sqrt{100}] =$$

$$[98 - 19,6; 98 + 19,6] =$$

$$[78,4; 117,6] .$$

Das hieße, der wahre durchschnittliche Absatz läge mit einer Wahrscheinlichkeit von 95% zwischen 78 400 und 117 600 Päckchen. Dieses Intervall wäre eine sehr grobe Abschätzung. Mit so etwas kann man nicht mehr vernünftig planen.

Also wird das Konfidenzintervall mit steigender Varianz breiter, die Schätzungen werden schlechter, ungenauer. Hier könnte man als Abhilfe den Stichprobenumfang erhöhen.

Mindest erforderlicher Stichprobenumfang

Wie groß muss die Stichprobe mindestens sein, damit die Breite des Konfidenzintervalls höchstens 10 ist?

Die Breite des Konfidenzintervalls ist

$$2 \cdot 1,96 \cdot \sqrt{\frac{5000}{n}} \leq 10 \rightarrow$$

$$\sqrt{n} \geq \frac{2 \cdot 1,96 \cdot \sqrt{5000}}{10} = 27,71 \ .$$

Man müsste also mindestens n = 769 Monate erheben, über 64 Jahre!

90%-Konfidenzintervall

Es soll nun ein 90%-Konfidenzintervall für μ bestimmt werden.

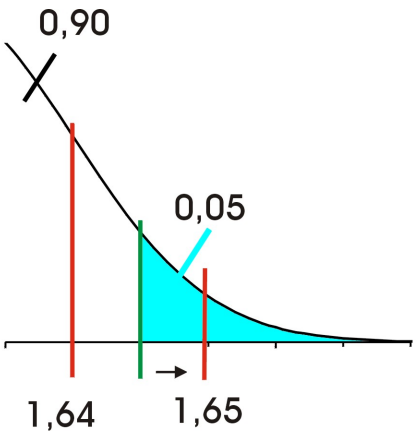

Abbildung 111: Wenn die vorgegebene Wahrscheinlichkeit zwischen zwei Quantile fällt, rückt man auf das äußere Quantil

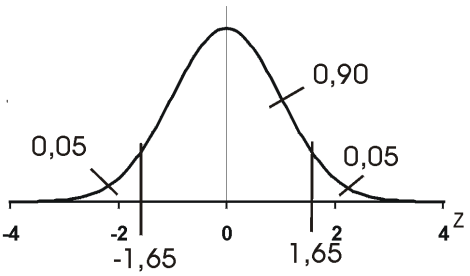

Abbildung 112: 90%-Konfidenzintervall

$$\left[\bar{x} - z(0,95) \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + z(0,95) \cdot \frac{\sigma}{\sqrt{n}}\right] =$$

$$[98 - 2 \cdot 1,65; 98 + 2 \cdot 1,65] =$$

$$[98 - 3,3; 98 + 3,3] =$$

$$[94,7; 101,3] \ .$$

Dieses Intervall ist schmäler als das 95%-Intervall.

Konfidenzintervalle für den Durchschnitt einer Grundgesamtheit

Es sei $X_1, ...X_n$ eine unabhängige Stichprobe aus der Grundgesamtheit. Der Stichprobenmittelwert ist:

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

und die Stichprobenvarianz:

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

Die observierten Werte dieser Stichprobenfunktionen deuten wir an mit $\bar{x}$, und s^2,.

Normalverteiltes Merkmal mit bekannter Varianz

Im obigen Beispiel war die Verteilung des Merkmals in der Grundgesamtheit bekannt und normalverteilt und die Varianz σ^2 war bekannt. Man erhält hier das 1-α-Konfidenzintervall für μ, den Durchschnitt des Merkmals in der Grundgesamtheit

$$[\bar{x} - z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}}; \bar{x} + z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}}] \; .$$

Normalverteiltes Merkmal mit unbekannter Varianz

Ist zwar das Merkmal in der Grundgesamtheit normalverteilt, aber die Varianz unbekannt, muss die Varianz des Merkmals durch s^2 geschätzt werden. Damit erhalten wir ein Zufallsintervall daß mit Wahrscheinlichkeit 1-α den Parameter enthält:

$$P(\bar{X} - t(1-\tfrac{\alpha}{2}; n-1)\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t(1-\tfrac{\alpha}{2}; n-1)\frac{S}{\sqrt{n}}) = 1-\alpha \; .$$

.

Daraus folgt das 1-α-Konfidenzintervall für den Durchschnitt μ des Merkmals in der Grundgesamtheit:

$$[\bar{x} - t(1-\tfrac{\alpha}{2}; n-1)\frac{s}{\sqrt{n}} \; ; \; \bar{x} + t(1-\tfrac{\alpha}{2}; n-1)\frac{s}{\sqrt{n}}] \; .$$

Das Quantil $t(1-\frac{\alpha}{2}; n-1)$ kommt jetzt aus einer t-Verteilung mit n-1 Freiheitsgraden. Die t-Verteilung hat eine ähnliche Form wie die Normalverteilung, ist aber etwas breiter. In der hier betrachteten Art (zentral) ist sie ebenfalls symmetrisch. Da sie verschiedene Freiheitsgrade hat, ist sie nur für ausgewählte Quantile tabelliert. Es gilt beispielsweise

t(0,975;4) = 2,776

und

t(0,025;4) = -2,776.

Merkmal mit unbekannter Verteilung und bekannter Varianz

Ist die Verteilung des Merkmals unbekannt, aber die Varianz σ^2 bekannt, kann man für EX des Merkmals X, das Konfidenzintervall

$$[\bar{x} - z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}}; \bar{x} + z(1-\tfrac{\alpha}{2})\frac{\sigma}{\sqrt{n}}] \; .$$

angeben, falls n groß genug ist (Faustregel $n > 30$).

Merkmal mit unbekannter Verteilung und unbekannter Varianz

Sind Verteilung und Varianz des Merkmals unbekannt, kann man für $n > 50$ das Konfidenzintervall für EX angeben als

$$[\bar{x} - z(1-\tfrac{\alpha}{2})\frac{s}{\sqrt{n}} \; ; \; \bar{x} + z(1-\tfrac{\alpha}{2})\frac{s}{\sqrt{n}}] \; .$$

Konfidenzintervalle für den Anteilswert einer dichotomen Grundgesamtheit

Modell mit Zurücklegen

Die Verteilung eines Merkmals einer dichotomen Grundgesamtheit lässt sich durch das Urnenmodell beschreiben. Man möchte den Anteilswert p, also den Anteil der Kugeln erster Sorte in der Urne bestimmen. Der Anteilswert wird geschätzt durch

$$\hat{p} = \frac{x}{n} \; ,$$

worin x der beobachtete Wert der Anzahl X der Kugeln erster Sorte in der Stichprobe ist.

Bei einem Urnenmodell mit Zurücklegen ist X binomialverteilt. Falls n groß genug ist (als Faustregel gilt: n > 100 und $n\hat{p}(1-\hat{p}) \geq 9$), erhält man das 1-$\alpha$-Konfidenzintervall für p durch eine Approximation der Binomialverteilung mit Hilfe der Normalverteilung:

$$\left[\hat{p} - z(1-\tfrac{\alpha}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;;\; \hat{p} + z(1-\tfrac{\alpha}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Exakt läßt sich das Konfidenzintervall mit den Verteilungswerten der Binomialverteilung bestimmen. Dafür muß zum Beispiel für eine untere Vertrauensgrenze ein Parameter p_u für die Binomialverteilung bestimmt werden, das so klein ist, daß die Wahrscheinlichkeit aus einer Binomialverteilung mit den Parametern n und p_u gerade x oder mehr Treffer zu erhalten höchstens die eingeräumte Irrtumswahrscheinlichkeit ist.

Für eine Alternative zu diesem Verfahren ist der Zusammenhang der Binomialverteilung mit der Betaverteilung nützlich. Eine untere Vertrauensgrenze für p_u liefert das α-Quantil der Betaverteilung mit den Parametern x und $n - x + 1$. Eine obere Vertrauensgrenze liefert das $1 - \alpha$-Quantil der Betaverteilung mit den Parametern $x + 1$ und $n - x$. Dabei handelt es sich nicht um zwei verschiedene Methoden, sondern nur um zwei verschiedene Suchverfahren nach einem geeigneten Parameter für die Binomialverteilung, so dass jeweils der einseitige Test für den Parameter der Binomialverteilung nicht zu Ablehnung führt. Weil Quantile der Betaverteilung durch eine Nullstellensuche in der unvollständigen Beta-Funktion bestimmt werden können, ist die Suchstrategie über die Betaverteilung schon dann leicht zugänglich, wenn man einen numerischen Zugang zur unvollständigen Betafunktion und ein allgemeines Verfahren zur Nullstellensuche zu Verfügung hat. Dies kann ein Vorteil gegenüber der Suche nach einem geeigneten Parameter der Binomialverteilung sein, für den das beobachtete x gerade nicht zur Ablehnung führt.

Die exakte Methode über die Suche nach einem geeigneten Parameter der Binomialverteilung so, dass ein einseitiger Test für die Beobachtung x gerade nicht zu Ablehung führt, ist nur für die Suche nach einer einseitigen Vertrauensgrenze unverfälscht. Ein unverfälschtes zweiseitiges Konfidenzintervall für den Parameter p der Binomialverteilung muss aus einem unverfälschten zweiseitigen Test für den Parameter p abgeleitet werden. Weil die Binomialverteilung außer für $p = 1/2$ nicht symmetrisch ist, genügt es nicht die Irrtumswahrscheinlichkeit α zu gleichen Teilen auf die beiden Enden der Verteilung aufzuteilen.

Modell ohne Zurücklegen

Bei einem Urnenmodell ohne Zurücklegen ist X hypergeometrisch verteilt. Falls die Bedingungen

$n > 9/(p \cdot (1-p))$, $n > 100$ $n/N \leq 0,05$ erfüllt sind, ist die Approximation der hypergeometrischen Verteilung durch die Normalverteilung brauchbar und man erhält das approximative $(1-\alpha)$-Konfidenzintervall für θ

$$\left[p - z\left(1-\frac{\alpha}{2}\right) \cdot \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}; p + z\left(1-\frac{\alpha}{2}\right) \cdot \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}} \right] .$$

.

Einführung an Hand eines Beispiels mit Wurstgläsern

Die Firma HappyWurscht stellt Wurstwaren her. Sie ist vor allem für ihre delikate Leberwurst in 250g-Gläsern bekannt. Diese werden durch eine Füllanlage mit der noch heißen, flüssigen Masse befüllt. Um Beanstandungen bezüglich der Füllmenge zu vermeiden, füllt man etwas mehr Masse als 250 g ein. Die Füllmenge schwankt immer leicht, aber es wird ein durchschnittliches Füllgewicht von 260g angestrebt. Die Qualitätssicherung soll die Einhaltung dieser Durchschnittsmenge überprüfen.

Überlegung zur Verteilung der Stichprobe

Es ist aber das durchschnittliche Füllgewicht eines Wurstglases unbekannt. Bekannt ist in diesem Beispiel lediglich, daß das Füllgewicht normalverteilt ist mit einer Varianz $\sigma^2 = 64$ [g^2].

Wie könnte man nun den Durchschnitt ermitteln? Man könnte eine Stichprobe mit z.B. n = 16 Beobachtungen ziehen und versuchen, aus dem arithmetischen Mittel x auf das durchschnittliche Füllgewicht der Grundgesamtheit zu schließen.

Wir betrachten nun das Füllgewicht eines Wurstglases. Wir bezeichnen es als Zufallsvariable X. Es soll geprüft werden, ob durchschnittlich 260g in einem Glas sind, d.h. ob EX = 260 ist.

Beträgt nun tatsächlich der wahre durchschnittliche Absatz der Grundgesamtheit $\mu_0 = 260$, kann man bei einer genügend großen Stichprobe vermuten, daß x in der Nähe von μ_0 liegen müßte. Meistens wird x in der Nähe von μ_0 liegen, da aber x die Realisation einer Zufallsvariablen ist, kann in sehr wenigen Fällen x auch extrem weit von μ_0 weg liegen, so daß man dann μ verkehrt einschätzt.

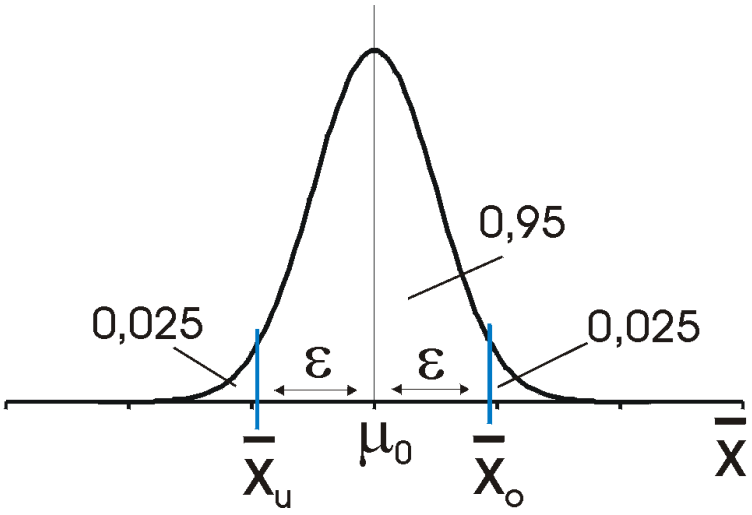

Abbildung 113

Man könnte aber ein Intervall um μ_0 bestimmen, in dem bei Vorliegen von μ_0 z.B. 95% aller möglichen x-Werte liegen, also

$$P(\bar{x}_u \leq \bar{X} \leq \bar{x}_o) = 0,95 \; .$$

Es wird dann eine konkrete Stichprobe genommen. Fällt x nicht in dieses Intervall $[\overline{x}_u; \overline{x}_o]$, ist x zu weit von μ_0 weg. Man geht dann davon aus, dass $\mu_0 \neq 260$ ist. Damit man dieses Intervall berechnen kann, müssen Informationen über die Verteilung von X verfügbar sein.

Ablauf eines Hypothesentests

Feststellung der Verteilung des Merkmals in der Grundgesamtheit

Die Zufallsvariable X: Füllgewicht eines Wurstglases ist normalverteilt mit einem unbekannten Erwartungswert μ und der bekannten Varianz varX = σ^2 = 64. Man interessiert sich für den Parameter μ.

Aufstellen der Nullhypothese

Man stellt die Nullhypothese H_0: $\mu = \mu_0 = 260$ auf, d.h. man behauptet, das wahre unbekannte durchschnittliche Füllgewicht in der Grundgesamtheit betrage $\mu_0 = 260$.

Festlegen des Nichtablehnungsbereiches für H_0

Zur Überprüfung der Hypothese soll eine Stichprobe im Umfang von n = 16 gezogen werden, die zu einer sog. Prüfgröße x zusammengefasst wird.

Der Stichprobendurchschnitt X ist selbst eine Zufallsvariable und ist als lineare Transformation von X wiederum normalverteilt und zwar mit den Parametern

$$E\bar{X} = \mu$$

und $var\bar{X} = \frac{\sigma^2}{n}$.

Bei Gültigkeit von H_0 ist also

$$\bar{X} \to N(\mu_0; \frac{\sigma^2}{n}) \ ,$$

hier

$$\bar{X} \to N(260; \frac{64}{16} = 4) \ .$$

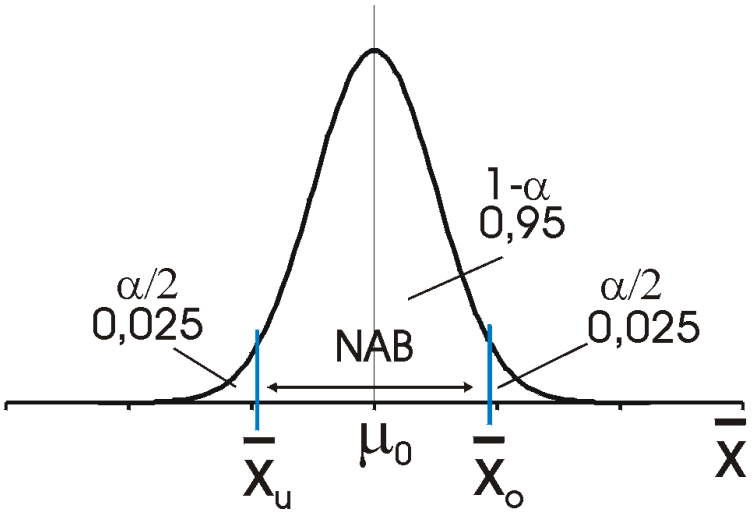

Abbildung 114

Nun wird der Bereich für x festgelegt, in dem die Nullhypothese nicht abgelehnt wird, der Nichtablehnungsbereich (NAB) [x_u; x_o]. Fällt die Prüfgröße x in diesem Bereich, wird H_0 nicht abgelehnt. Es soll sein

$$P(\bar{x}_u \leq \bar{X} \leq \bar{x}_o) = 0,95 = 1 - \alpha \ .$$

Wir nennen α das Signifikanzniveau oder den α-Fehler: Das ist die Wahrscheinlichkeit, dass die Nullhypothese H_0 abgelehnt wird, obwohl $\mu_0 = 260$ der wahre Parameter ist.

Bestimmung von [x_u ; x_o]:

Standardisiert man mit

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \ ,$$

können wir analog zu oben

$$P(z_u \leq Z \leq z_o) = 0,95$$

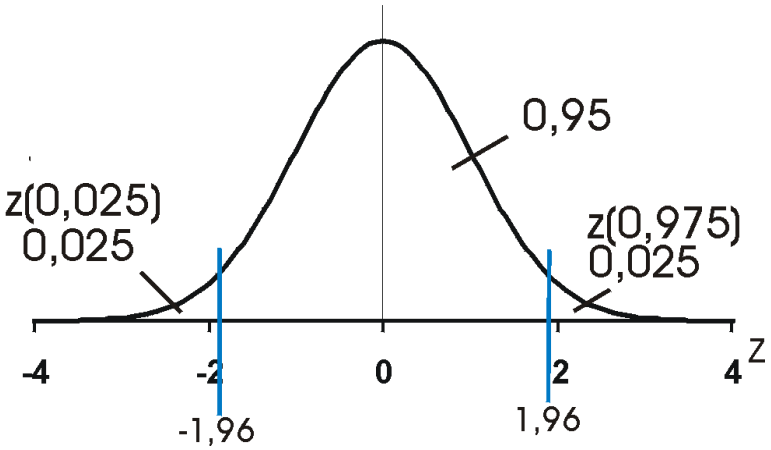

Abbildung 115

schreiben. Es ergibt als Intervall für Z:

$$[z_u; z_o] = [z(\alpha/2); z(1-\alpha/2);] =$$

$$[-z(1-\alpha/2); z(1-\alpha/2);] =$$

$$[-z(0,975); z(0,975)] = [-1,96; 1,96]$$

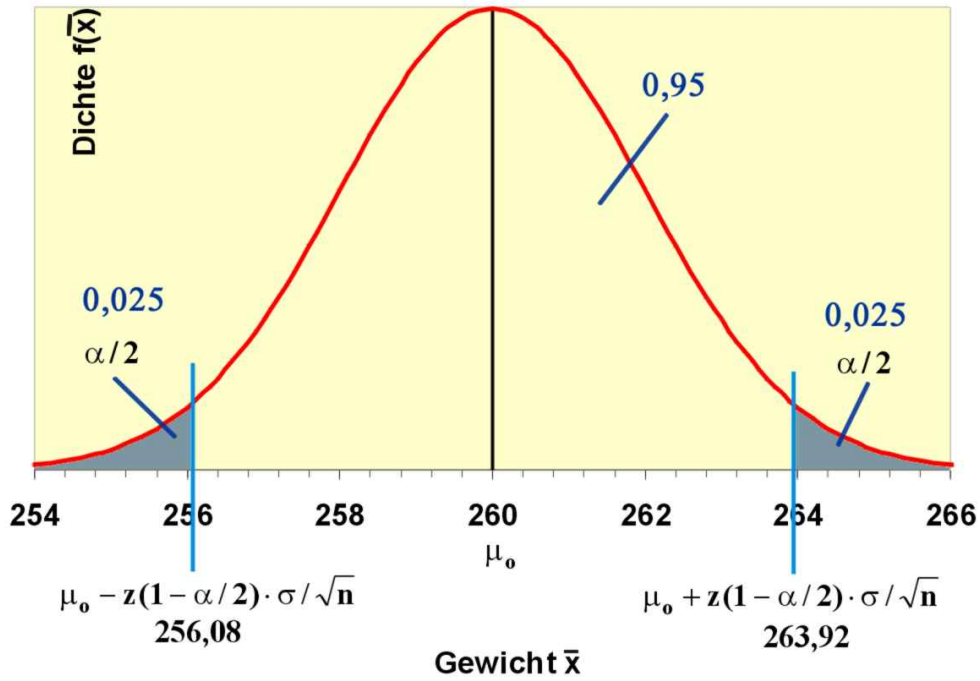

Abbildung 116: Nichtablehnungsbereich der Nullhypothese für[x

Es ist nun aber

$$\bar{x}_u = \mu_0 - z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

und $\bar{x}_o = \mu_0 + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$

so dass hier der Nichtablehnungsbereich für x

$$[\bar{x}_u; \bar{x}_o] =$$

$$[260 - 1,96 \cdot 2; 260 + 1,96 \cdot 2] =$$

$$[260 - 3,92; 260 + 3,92] =$$

$$[256,08; 263,92]$$

ist.

Wenn μ_0 tatsächlich 260 ist, würde x in 5% aller Stichproben in den Ablehnungsbereich

$$(-\infty; 256,08] \vee [263,92; \infty)$$

fallen.

Stichprobe erheben

Nach der Festlegung des Nichtablehnungsbereichs wird eine Stichprobe genommen. Es wurde hier der Inhalt von 16 Gläsern gewogen. Es ergab sich die Urliste

268 252 254 252 251 245 257 275 268 270 253 250 266 265 250 267

Es ist dann

$$\bar{x} = \frac{1}{16}(268 + 252 + ... + 267) =$$

$\frac{1}{16}(4144) = 259.$

Entscheidung treffen

Wir fällen nun die Entscheidung: Da x = 259 im Nichtablehnungsbereich liegt, wird H_0 nicht abgelehnt. Es wird davon ausgegangen, dass die Maschine die Gläser korrekt befüllt.

Eine äquivalente Vorgehensweise ist, man bestimmt zunächst die standardisierte Prüfgröße z:

$z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{259-260}{\frac{8}{\sqrt{16}}} = \frac{-1}{2} = -0,5.$

Der Nichtablehnungsbereich für Z ist [-1,96; 1,96]. Da z in den Nichtablehnungsbereich fällt, wird H_0 nicht abgelehnt.

Beide Vorgehensweisen liefern das gleiche Ergebnis.

Punkt- und Bereichshypothesen

In obigen Beispiel wurde für das wahre μ nur ein bestimmter Punkt getestet: H_0: $\mu = \mu_0$, also handelt es sich um eine Punkthypothese. Es könnte aber sein, dass der Hersteller einem Großabnehmer versichert hat, dass das durchschnittliche Füllgewicht mindestens 260 g beträgt. Es wird also hier genügen, zu prüfen, ob der Mindestwert erreicht wird. Es ist aber kein Problem, wenn die durchschnittliche Füllmenge größer als 260 ist.

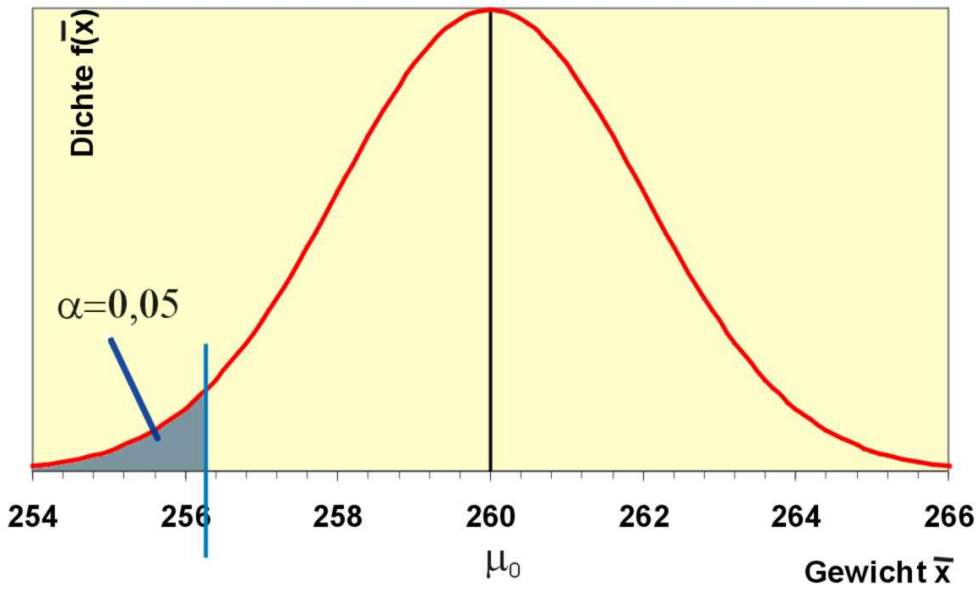

Abbildung 117: Ablehnungsbereich der Mindesthypothese H_0: $\mu \geq \mu_0 = 260$

Man stellt also als Arbeitshypothese auf: H_0: $\mu \geq \mu_0 = 260$. Wird die Prüfgröße x geringfügig kleiner als 260, kann das eine Zufallsschwankung sein. Aber wird x zu klein, muss H_0 abgelehnt werden. Da hier nur der Bereich links von μ_0 kritisch für die Ablehnung ist, wird das gesamte α links auf dem Zahlenstrahl plaziert, der kritische Wert für z ist also $z(\alpha) = -z(1-\alpha)$. Fällt z in den Ablehnungsbereich $(-\infty; -z(1-\alpha)]$, wird H_0 abgelehnt. Man geht dann davon aus, dass μ kleiner als μ_0 sein muss, dass also die Befüllung nicht ordnungsgemäß ist. Der kritische Wert für x ist hier

$$\bar{x}_{1-\alpha} = \mu_0 - z(1-\alpha) \cdot \frac{\sigma}{\sqrt{n}} ,$$

also

$$\bar{x}_{1-\alpha} = 260 - 1,65 \cdot \frac{8}{\sqrt{16}} = 256,7 \; .$$

Wenn die Stichprobe ein Durchschnittsgewicht von weniger als 256,7g ergibt, wird die Lieferung beanstandet.

Entsprechend erhält man unter der Hypothese H_0: $\mu \leq \mu_0$ für die Prüfgröße z den Ablehnungsbereich [z(1-a); ∞) bzw.

$$\bar{x}_{1-\alpha} = \mu_0 + z(1-\alpha) \cdot \frac{\sigma}{\sqrt{n}} \quad .$$

Fehler und Varianzen

Fehlerarten

Warum wird der α-Fehler als Fehler bezeichnet? Hier wollen wir uns zunächt mal überlegen, welche Fehler bei der Entscheidung überhaupt gemacht werden können?

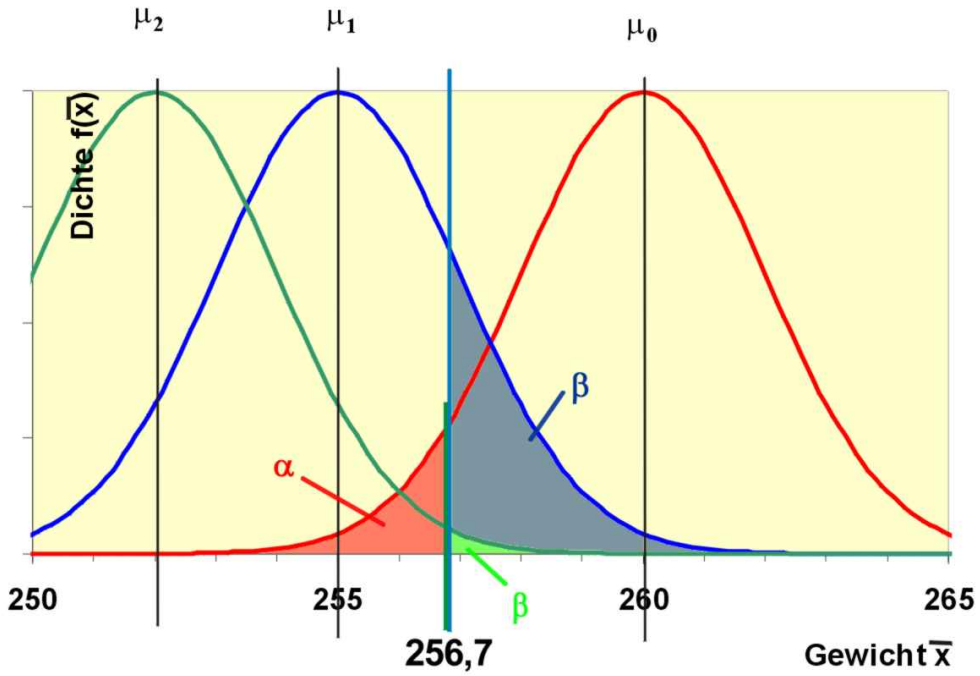

Abbildung 118: Vermischung der hypothetischen und tatsächlichen Verteilung

1. H_0 ist wahr, die Prüfgröße fällt aber in den Ablehnungsbereich (in α * 100% aller Stichproben). Hier würde man H_0 irrtümlicherweise ablehnen, obwohl H_0 wahr ist: α-Fehler oder Fehler 1. Art. In unserem Beispiel würde also die Lieferung möglicherweise zurückgewiesen werden, obwohl die Gläser korrekt befüllt worden sind.

1. H_0 ist falsch, die Prüfgröße fällt aber in den Nichtablehnungsbereich. In Wirklichkeit ist $\mu = \mu_1$, z.B. $\mu_1 = 255$ g. Jetzt ist bei unveränderter Varianz in Wahrheit der Stichprobendurchschnitt x verteilt wie

$$N(\mu_1; \frac{\sigma^2}{n}) = N(255; 4)$$

Unter dieser Verteilung beträgt die Wahrscheinlichkeit, dass H_0 (fälschlicherweise) nicht abgelehnt wird,

$$P(\bar{X} \geq 256,7) = 1 - \Phi_{\bar{x}}(256,7|255;4) \ ,$$

was sich einfach berechnen lässt als

$$1 - \Phi_z(\frac{256,7 - 255}{2}) = 1 - \Phi_z(0,85) = 0,1977.$$

Man würde also mit fast 20%iger Wahrscheinlichkeit irrtümlicherweise die Lieferung akzeptieren. Dieser Fehler ist der β-Fehler oder Fehler 2. Art.

Wenn in Wahrheit $\mu = \mu_2 = 252$ ist, beträgt der β-Fehler

$$P(\bar{X} \geq 256,7) = 1 - \Phi_{\bar{x}}(256,7|252;4) =$$

$1 - \Phi_z(\frac{256,7-252}{2}) = 1 - \Phi_z(2,35) = 0,0094.$

Hier ist die Wahrscheinlichkei einer irrtümlichen Ablehnung schon sehr klein.

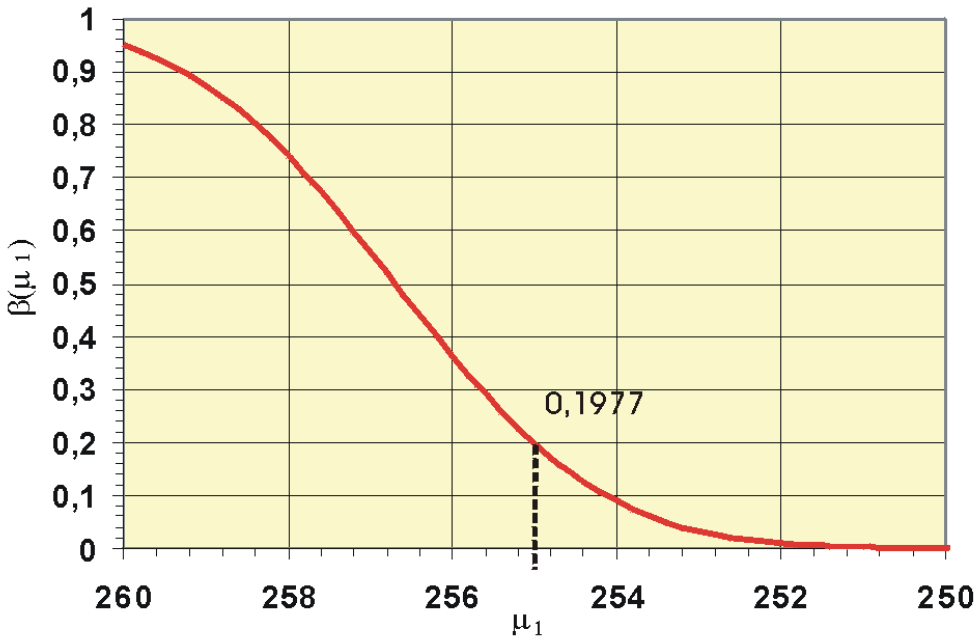

Abbildung 119: Operationscharakteristik zur Hypothese: $\mu \leq 260$

Der β-Fehler hängt also von μ_1 ab. Man kann den β-Fehler in Abhängigkeit von μ_1 als Funktion darstellen: $\beta = \mathrm{f}(\mu_1)$. Diese Funktion nennt man Operationscharakteristik. Der Wert 1 - β ist dagegen die Wahrscheinlichkeit, dass H_0

abgelehnt wird, wenn μ_1 der wahre Parameter ist. Man sieht an der Grafik, dass 1 - β für $\mu = 260$ gerade 0,05 ist. Das ist natürlich die Wahrscheinlichkeit, dass H_0 (hier fälschlicherweise) abgelehnt wird, wenn 260 tatsächlich der wahre Parameter ist.

Um die Wahrscheinlichkeit für eine falsche Entscheidung zu reduzieren, ist es wünschenswert, möglichst schnell in den Bereich $\beta \approx 0$ zu kommen. U. U. hilft eine Erhöhung des Stichprobenumfangs.

Eine Hypothese, die nicht abgelehnt ist, gilt nicht automatisch als angenommen, denn der β-Fehler ist i.a. unbekannt.

Wenn ein Test die Wahrscheinlichkeit der Annahme falscher Nullhypothesen möglichst reduziert, nennt man ihn trennscharf.

Breite des Nichtablehnungsbereichs

Es soll nun wieder die Punkthypothese H_0: $\mu = \mu_0$ betrachtet werden. Es ergab sich hier für x der Nichtablehnungsbereich [256,08; 263,92] mit einer Breite 7,84 g.

Änderung des Signifikanzniveaus

Welcher NAB ergibt sich für $\alpha = 0{,}01$? Wir errechnen das (1 - $\alpha/2$)-Quantil als

$$\alpha = 0,01 \rightarrow \alpha/2 = 0,005 \rightarrow 1 - \alpha/2 = 0,995 \rightarrow z(0,995) = 2,58$$

und erhalten den Nichtablehnungsbereich für x als

$$[260 - z(0,995) \cdot 2; 260 + z(0,995) \cdot 2] =$$

$$[260 - 2,58 \cdot 2; 260 + 2,58 \cdot 2] =$$

$$[260 - 5,16; 260 + 5,16] =$$

$$[254,84; 265,16]$$

.

Hier ist der Nichtablehnungsbereich breiter als für $\alpha = 0{,}05$: H_0 wird nur in 1% aller Stichproben fälschlicherweise abgelehnt. Hier hätte die Lieferfirma einen Vorteil.

Welcher NAB ergibt sich für $\alpha = 0{,}1$?

$$[260 - z(0,95) \cdot 2; 260 + z(0,95) \cdot 2] =$$

$$[260 - 1,65 \cdot 2; 260 + 1,65 \cdot 2] =$$

$$[260 - 2,30; 260 + 2,30] =$$

$$[257,70; 262,30]$$

.

Hier ist der Nichtablehnungsbereich schmäler, H_0 wird in 10% aller Stichproben fälschlicherweise abgelehnt.

Änderung der Varianz

Was passiert, wenn die Varianz $\sigma^2 = 256$ ist ($\alpha = 0{,}05$)? Man erhält hier für die Punkthypothese H_0: $\mu = \mu_0 = 260$ den NAB für x

$$[260 - 1,96 \cdot \sqrt{\frac{256}{16}}; 260 + 1,96 \cdot \sqrt{\frac{256}{16}}] =$$

$$[260 - 1,96 \cdot 4; 260 + 1,96 \cdot 4] =$$

$$[260 - 7,84; 260 + 7,84] =$$

$$[252,16; 267,84]$$

.

Die Breite des Nichtablehnungsbereichs ist hier 15,68g.

Für H;$_0$: $\mu \geq \mu_0$ ergibt sich dann entsprechend als kritischer Wert

$$260 - 1,65 \cdot 4 = 253,4$$

.

Die Grafik zeigt den Fall der Bereichshypothese mit einer Varianz von 16: Durch die große Varianz sind die Normalverteilungskurven sehr flach und durchmischen sich stark. Der Betafehler bei $\mu_1 = 255$ ist sehr groß. Eine vernünftige Kontrolle der Abfüllmaschine ist nicht mehr möglich.

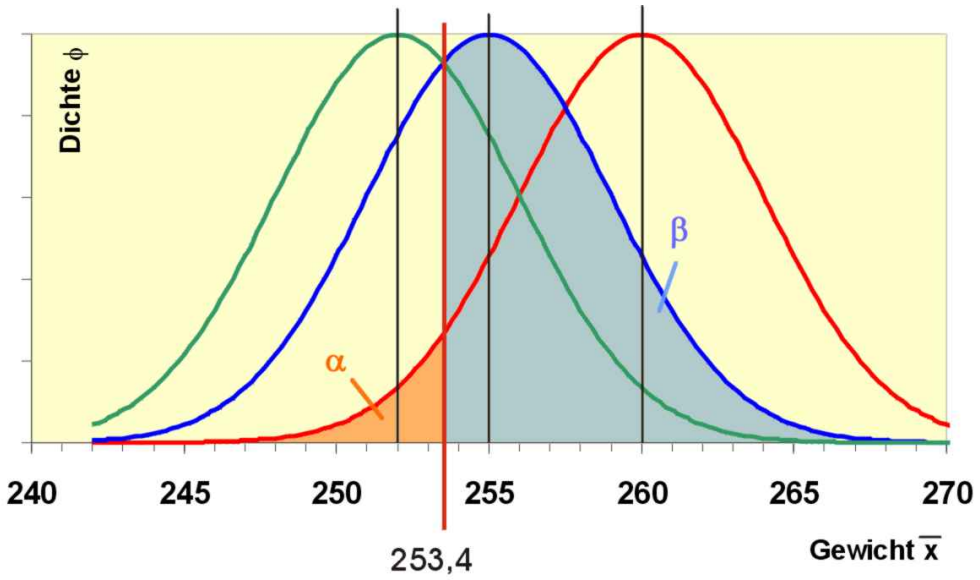

Abbildung 120: Der Betafehler bei großen Varianzen

Der Nichtablehnungsbereich wird mit wachsender Varianz breiter, der Test verliert an Trennschärfe.

Änderung des Stichprobenumfangs

Was passiert, wenn der Stichprobenumfang jetzt 64 beträgt ($\alpha = 0{,}05$; $\sigma^2 = 64$)?

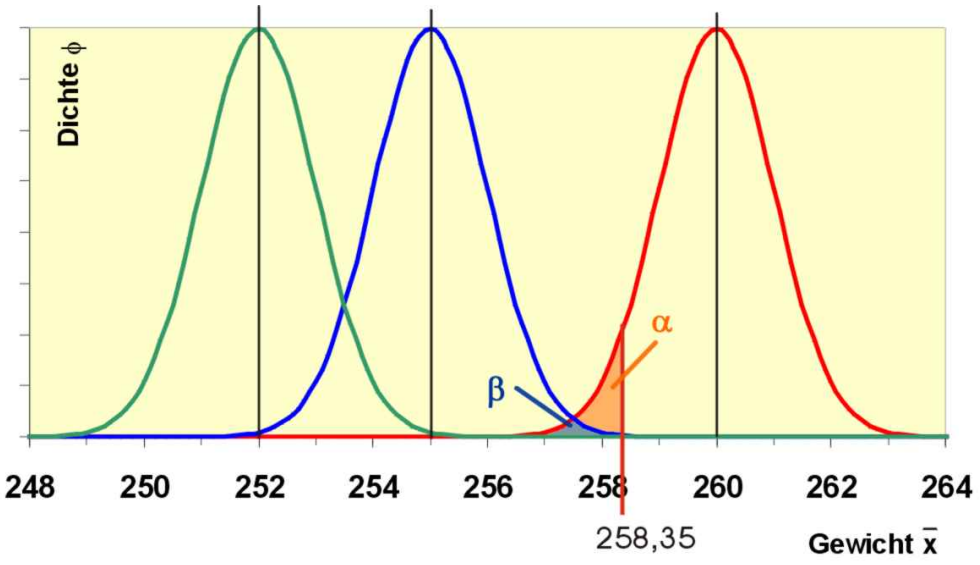

Abbildung 121: Der Betafehler bei kleinen Varianzen

$$[260 - 1,96 \cdot \sqrt{\frac{64}{64}}; 260 + 1,96 \cdot \sqrt{\frac{64}{64}}] =$$

$$[260 - 1,96 \cdot 1; 260 + 1,96 \cdot 1] =$$

$$[260 - 1,96; 260 + 1,96] =$$

$$[258,04; 261,96]$$

.

Hier hat der Nichtablehnungsbereich eine Breite von 3,92, denn durch den größereren Stichprobenumfang hat sich die Varianz von X verringert. Der NAB schrumpft bei steigendem Stichprobenumfang, der Test wird trennschärfer.

Mindest erforderlicher Stichprobenumfang

Wie groß muß die Stichprobe mindestens sein, damit die Breite des NAB für $\alpha = 0{,}05$ höchstens 10 beträgt?

Die Breite des NAB ist ja definiert durch

$$2 \cdot z(1 - \alpha/2) \cdot \frac{\sigma}{\sqrt{n}}$$

.

Es soll also hier sein

$$2 \cdot 1,96 \cdot \frac{64}{\sqrt{n}} \leq 10$$

.

Die Auflösung der Ungleichung nach $\sqrt{n}$ ergibt dann

$$\sqrt{n} \geq 2 \cdot 1,96 \cdot \frac{64}{10} = 25,088$$

und $(\sqrt{n})^2 = 629,41$.

Da wir nur ganze Wurstgläser analysieren können, brauchen wir einen Stichprobenumfang von mindestens 630 Gläsern.

Kann die Wurst mit dem Glas zusammen gewogen werden, stellt diese hohe Zahl kein Problem dar. Geht durch so eine Stichprobe allerdings die Zerstörung der Ware mit einher, etwas die lebensmitteltechnische Untersuchung einer Konservendose, muss man einen Kompromiss zwischen mangelnder Trennschärfe und Zerstörung der Ware finden.

Erwartungswert

1. Bekannte Verteilung und Varianz

Im einführenden Beispiel war die Verteilung des Merkmals in der Grundgesamtheit bekannt, also normalverteilt, und die Varianz ist bekannt. Die Prüfgröße

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

ist dann standardnormalverteilt. Wir erhalten die Entscheidungsregeln für eine gewählte Irrtumswahrscheinlichkeit α

- Ho: $\mu = \mu_0$ wird abgelehnt, falls z z(1-a/2) ist.
- Ho: $\mu \leq \mu_0$ wird abgelehnt, falls z > z(1-α) ist.
- Ho: $\mu \geq \mu_0$ wird abgelehnt, falls z < - z(1-α) ist.

2. Bekannte Verteilung und unbekannte Varianz

Häufig wird neben dem Erwartungswert die Varianz ebenfalls nicht bekannt sein, so dass man statt der Varianz in der Grundgesamtheit die Schätzung

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

verwendet. Wir erhalten nun bei normalverteilter Grundgesamtheit statt z die Prüfgröße

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}},$$

die t-verteilt mit n-1 Freiheitsgraden ist.

Die t-Verteilung hat eine ähnliche Form wie die Normalverteilung. In der hier betrachteten Art (zentrale t-Verteilung) ist sie ebenfalls symmetrisch bezüglich der Null. Da sie verschiedene Freiheitsgrade hat, ist sie nur für ausgewählte Quantile tabelliert. Es ist t(p;k) das p-Quantil der t-Verteilung mit k Freiheitsgraden.

Es gilt beispielsweise für die Zufallsavariable t mit 5 Freiheitsgraden:

$$P(t \leq 3,365) = 0,99$$

bzw. $t(0,99;5) = 3,365$.

Wir erhalten die Entscheidungsregeln

- H: $\mu = \mu_0$ wird abgelehnt, falls t t(1-α/2; n - 1) ist.
- H: $\mu \leq \mu_0$ wird abgelehnt, falls t > t(1-α; n - 1) ist.
- H: $\mu \geq \mu_0$ wird abgelehnt, falls t < - t(1-α n - 1) ist.
- Ist n > 30, können die Quantile der t-Verteilung durch die entsprechenden Quantile der Normalverteilung ersetzt werden.

3. Unbekannte Verteilung und bekannte Varianz

Ist die Verteilung des Merkmals X unbekannt, aber die Varianz varX bekannt, verwendet man bei einem n > 30 die standardnormalverteilte Prüfgröße

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Wir erhalten die Entscheidungsregeln analog zu 1.

4. Unbekannte Verteilung und unbekannte Varianz

Sind Verteilung und Varianz des Merkmals X in der Grundgesamtheit unbekannt, verwendet man für n > 50 die standardnormalverteilte Prüfgröße

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Wir verwenden die Entscheidungsregeln analog zu 1.

Anteilswert einer dichotomen Grundgesamtheit

Die Verteilung des Merkmals X einer dichotomen Grundgesamtheit lässt sich durch das Urnenmodell beschreiben. Man möchte den Anteilswert θ, also den Anteil der Kugeln erster Sorte in der Urne bestimmen. Der Anteilswert wird geschätzt durch

$$\hat{\theta} = p = \frac{x}{n}$$

,

wobei x die Zahl der Kugeln erster Sorte in der Stichprobe ist. Bei einem Urnenmodell mit Zurücklegen ist X binomialverteilt.

Falls

$$n > \frac{9}{\theta \cdot (1-\theta)}$$

können wir die Prüfgröße verwenden

$$z = \frac{x \pm 0,5 - n \cdot \theta_0}{\sqrt{n \cdot \theta \cdot (1-\theta)}}$$

- H_0: $\theta = \theta_0$ wird abgelehnt, falls

$$z = \frac{x + 0,5 - n \cdot \theta_0}{\sqrt{n \cdot \theta \cdot (1-\theta)}} < -z(1-\alpha/2)$$

,

(wenn die Prüfgröße z < 0 ist) oder

$$z = \frac{x - 0,5 - n \cdot \theta_0}{\sqrt{n \cdot \theta \cdot (1-\theta)}} > z(1-\alpha/2)$$

(wenn die Prüfgröße z > 0 ist) errechnet wird.

- H_0: $\theta \leq \theta_0$ wird abgelehnt, falls

$$z > z = \frac{x - 0,5 - n \cdot \theta_0}{\sqrt{n \cdot \theta \cdot (1-\theta)}} > z(1-\alpha)$$

ist.

- H_0: $\theta \geq \theta_0$ wird abgelehnt, falls

$$z = \frac{x + 0,5 - n \cdot \theta_0}{\sqrt{n \cdot \theta \cdot (1 - \theta)}} < -z(1 - \alpha)$$

ist.

Ist n zu klein, kann der Ablehnungsbereich mit Hilfe der F-Verteilung exakt bestimmt werden oder mit dem Prinzip des konservativen Testens festgelegt werden.

Test auf Varianz

Herleitung der Prüfgröße

Betrachten wir eine normalverteilte Grundgesamtheit. Die Schätzung für die Varianz ist hier

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

.

Wir wollen nun eine passende Prüfgröße für einen Varianztest herleiten. Die Summe von n quadrierten standardnormalverteilten Zufallsvariablen, jede mit dem Erwartungswert μ und der Varianz σ^2 ist χ^2-verteilt mit n Freiheitsgraden, also die Summe

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

.

Schätzt man

$$\hat{\mu} = \bar{x}$$

geht ein Freiheitsgrad verloren.

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

ist χ^2-verteilt mit n-1 Freiheitsgraden. Wir wollen nun diese Summe mit S^2 verquicken, um eine Prüfgröße für diesen Test zu erhalten. Es ist dann

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

$\frac{S^2 \cdot (n-1)}{\sigma^2}$

ebenfalls χ^2-verteilt mit n-1 Freiheitsgraden. Unter der Nullhypothese H_0: $\sigma^2 = {\sigma^2}_0$ ist dann

$$Y = \frac{S^2 \cdot (n-1)}{\sigma_0^2}$$

ebenfalls verteilt wie oben.

Wir wollen nun für H_0: $\sigma^2 = {\sigma^2}_0$ den Nichtablehnungsbereich für den Test angeben. Die Hypothese wird nicht abgelehnt, wenn die Prüfgröße y in das Intervall

$$[\chi^2(\frac{\alpha}{2}; n-1); \chi^2(1-\frac{\alpha}{2}; n-1)]$$

fällt, wobei χ^2(p;k) das p-Quantil der χ^2-Verteilung mit k Freiheitsgraden ist.

Die Nichtablehnungsbereiche für die Bereichshypothesen werden analog zu der Vorgehensweise bei Erwartungswerden festgelegt:

Bei der Mindesthypothese $H_0 : \sigma^2 \geq \sigma_0^2$ wird die Hypothese abgelehnt, wenn die Prüfgröße

$$Y < \chi^2(\alpha; n-1)$$

ist.

Bei der Höchsthypothese $H_0 : \sigma^2 \leq \sigma_0^2$ wird die Hypothese abgelehnt, wenn die Prüfgröße

$$Y > \chi^2(1-\alpha; n-1)$$

ist.

Beispiel für eine Punkthypothese

Ein großer Blumenzwiebelzüchter hat eine neue Sorte von Lilien gezüchtet. Die Zwiebeln sollen im Verkauf in verschiedenen Größenklassen angeboten werden. Um das Angebot planen zu können, benötigt der Züchter eine Information über die Varianz der Zwiebelgröße. Es wurden 25 Zwiebeln zufällig ausgewählt und gemessen. Man erhielt die Durchmesser (cm)

8 10 9 7 6 10 8 8 8 6 7 9 7 10 9 6 7 7 8 8 8 10 10 7 7

Es soll die Hypothese überprüft werden, dass die Varianz der Zwiebelgröße 3 cm^2 beträgt ($\alpha = 0{,}05$).

Die Nullhypothese lautet $H_0 : \sigma^2 = \sigma_0^2 = 3$

Nichtablehnungsbereich für die Prüfgröße y ist

$$[\chi^2(\frac{\alpha}{2}; n-1); \chi^2(1-\frac{\alpha}{2}; n-1)]$$

=

$[\chi^2(0,025; 24); \chi^2(0,975; 24)] = [12,40; 39,36]$.

Es ergab sich für die Stichprobe $\bar{x} = 8$ und $s^2 = \frac{42}{24} = 1,75$. Die Prüfgröße errechnet sich als

$$y = \frac{S^2 \cdot (n-1)}{\sigma_0^2} =$$

$\frac{1,75 \cdot 24}{3} = 14$.

Die Hypothese kann nicht abgelehnt werden.

Beispiel für eine Bereichshypothese

An einer Abfüllanlage werden Tagesdosen für ein sehr teures flüssiges Medikament in Plastikschälchen eingebracht. Da das Medikament hochwirksam ist, soll die Abweichung der Füllmenge vom Mittelwert möglichst wenig schwanken. Man weiß, dass die Füllmenge normalverteilt ist. Zur Kontrolle soll die Hypothese getestet werden, dass die Varianz höchstens 0,01 ml^2 beträgt. Eine Stichprobe von 20 Schälchen ergab den Mittelwert 0,5 und die Varianz 0,014.

Zu testen ist $H_0 : \sigma^2 \leq \sigma_0^2$.

Die Prüfgröße für H_0 ist $Y = \frac{S^2 \cdot (n-1)}{\sigma_0^2}$.

Die Hypothese wird abgelehnt, wenn $y > \chi^2(1-\alpha; n-1) = \chi^2(0,9; 19) = 27,20$ ist.

Die Stichprobe ergab

$y = \frac{0,014 \cdot 19}{0,01} = 26,6$

Die Hypothese wird nicht abgelehnt. Man geht davon aus, dass die Varianz der Füllmenge sich nicht verändert hat.

Vergleich zweier Varianzen

Wir haben es mit zwei verschiedenen Grundgesamtheiten zu tun. Wir interessieren uns dafür, ob die Varianzen dieser beiden Grundgesamtheiten gleich sind. Beide Merkmale dieser Grundgesamtheiten sollen normalverteilt sein.

Herleitung der Prüfgröße

Zu prüfen ist also die Hypothese: H_0: ${\sigma_1}^2 = {\sigma_2}^2$.

Geschätzt werden beide Varianzen wieder mit der Stichprobenvarianz

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

.

Es soll nun daraus eine Prüfgröße konstruiert werden. Wir wissen bereits, dass der Quotient

$$Y = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2}$$

χ^2-verteilt mit n-1 Freiheitsgraden ist. Eine Möglichkeit, zwei solche Zufallsvariablen zu verquicken, ist die F-Verteilung. Es ist nämlich der Quotient

$$f = \frac{\frac{Y_1}{n_1-1}}{\frac{Y_2}{n_2-1}} = \frac{\frac{\sum_{i=1}^{n_1}(X_{1i}-\bar{X}_1)^2}{(n_1-1)\sigma_1^2}}{\frac{\sum_{i=1}^{n_2}(X_{2i}-\bar{X}_2)^2}{(n_2-1)\sigma_2^2}}$$

F-verteilt mit n_1 - 1 und n_2 - 1 Freiheitsgraden. Wir müssen nun noch unsere Stichprobenvarianzen einpflegen und wir sehen, dass ja in Zähler und Nenner die Stichprobenvarianzen ${S_1}^2$ und ${S_2}^2$ schon dastehen. Also erhalten wir

$$f = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

$$= \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$$

Wir wollen diesen Quotienten nun mit der Nullhypothese in Verbindung bringen. Die Hypothese

$H_0 : \sigma_1^2 = \sigma_2^2$ lässt sich auch schreiben als $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ und es ist dann der Quotient der Prüfgröße unter H_0

$$f = \frac{S_1^2}{S_2^2} \cdot 1$$

.

Wenn die Nullhypothese wahr ist, sollte f nicht zu groß sein, aber auch nicht zu klein, weil sonst die Stichprobenvarianzen zu unterschiedlich wären. H_0 wird also nicht abgelehnt, wenn die Stichprobe f in den ||mittleren|| Bereich

$$[f(\frac{\alpha}{2}; n_1 - 1; n_2 - 1); f(1 - \frac{\alpha}{2}; n_1 - 1; n_2 - 1)]$$

fällt, wobei f(p;k_1;k_2) das p-Quantil der F-Verteilung mit k_1 und k_2 Freiheitsgraden ist.

Bereichhypothesen werden entsprechend aufgefasst:

$H_0 : \sigma_1^2 \leq \sigma_2^2$ lässt sich auch schreiben als $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \leq 1$.

Dieser Test wird abgelehnt, wenn

$$f > f(1 - \frac{\alpha}{2}; n_1 - 1; n_2 - 1)$$

,

wobei sich f wie oben berechnet.

Entsprechend wird $H_0 : \frac{\sigma_1^2}{\sigma_2^2} \geq 1$ abgelehnt, wenn

$$f < f(\frac{\alpha}{2}; n_1 - 1; n_2 - 1)$$

.

Beispiel

Bert und Berta haben im Fach Analysis ein Tutorium gehalten. Die Zeit, die die n_1 bzw. n_2 Studierenden für eine typische Klausuraufgabe benötigten, wurde festgehalten:

Tutorium von Bert: 8 3 4 4 10 9 2 9 Tutorium von Berta: 5 4 7 6 4

Beide Gruppen erzielten eine durchschnittliche Bearbeitungsdauer von 6 min. Ist aber auch die Varianz beider Gruppenleistungen gleich?

Wir wollen also nun bei einem Signifikanzniveau 0,05 die Nullhypothese testen, dass die Varianzen gleich sind.

Der Nichtablehnungsbereich für diesen Test ist

$$[f(\frac{\alpha}{2}; n_1 - 1; n_2 - 1); f(1 - \frac{\alpha}{2}; n_1 - 1; n_2 - 1)]$$

$= [f(0,025; 8; 5); f(0,975; 8; 5)] = [0,21; 6,76]$, wobei sich

$$f(0,025;8;5) = \frac{1}{f(0,975;5;8)} = \frac{1}{4,82} = 0,21$$

errechnet. Wir erhalten zunächst die Stichprobenvarianzen

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{8}((8-6)^2 + (3-6)^2 + (4-6)^2 + .. + (9-6)^2 = \frac{72}{8} = 9$$

und analog dazu

$$s_2^2 = 5,2$$

.

Die Prüfgröße errechnet sich nun als

$\frac{S_1^2}{S_2^2} \cdot 1 = \frac{9}{5,2} \cdot 1 = 1,73.$

Sie fällt in den Nichtablehnungsbereich und man kann die Hypothese nicht ablehnen.

Stochastische Unabhängigkeit

Die Beobachtungen zweier Merkmale X und Y liegen als gemeinsame klassierte Häufigkeitsverteilung vor mit n und m Kategorien und den dazugehörigen gemeinsamen Häufigkeiten n_{ij} (i = 1, ...,n; j = 1, ...,m) vor. Zur Prüfung der Hypothese H_0: „X und Y sind stochastisch unabhängig“ verwendet man die Prüfgröße

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}$$

Es soll jedes $\frac{n_{i.}n_{.j}}{n} \geq 5$ sein. Falls diese Forderung nicht gegeben ist, müssen so viele Zeilen und/oder Spalten zusammengefasst werden, bis die Vorgabe erfüllt ist.

Die Hypothese, dass X und Y stochastisch unabhängig sind, wird abgelehnt, wenn $\chi^2 > \chi^2(1 - \alpha; (m - 1)(n - 1))$ ist, als $(1-\alpha)$-Quantil der χ^2-Verteilung mit (m-1)(n-1) Freiheitsgraden.

Bemerkung: Dieser Test kann auch für die Prüfung der stochastischen Unabhängigkeit zweier Ereignisse verwendet werden. Man spricht hier von einem Vierfelder-Test.

Korrelation

Normalverteilung beider Merkmale

Die Merkmale X und Y sind normalverteilt. Es wird die spezielle Nullhypothese H_0: $\rho_{xy} = 0$ geprüft. Man schätzt den Korrelationskoeffizienten ρ mit dem Korrelationskoeffizienten r nach Bravais-Pearson und verwendet die Prüfgröße

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

H_0: $\rho_{xy} = 0$ wird abgelehnt, falls $t > t(1-\alpha/2; n - 2)$ ist.

Wird H_0 abgelehnt, geht man davon aus, dass X und Y korreliert sind. Sie sind dann auch stochastisch abhängig, so dass dieser Test im Ablehnungsfall auch die stochastische Unabhängigkeit erfasst. Bei Nichtablehnung können die Merkmale trotzdem abhängig sein, denn der Korrelationskoeffizient misst bekanntlich nur die lineare Abhängigkeit.

Wird H_0: $\rho_{xy} = \rho_0 \neq 0$ geprüft, hat r eine sog. nichtzentrale Verteilung, die nicht mehr ohne weiteres berechnet werden kann und nur noch näherungsweise mit der sog. Fisherschen Transformation angebbar ist.

Unbekannte Verteilung beider Merkmale

Die Merkmale X und Y sind beliebig verteilt. Es wird die spezielle Nullhypothese H_0: $\rho_{xy} = 0$ geprüft. Man schätzt den Korrelationskoeffizienten ρ mit dem Rangkorrelationskoeffizienten nach Spearman-Pearson r_{SP}.

Für $n > 10$ verwendet man die Prüfgröße

$$t = \frac{r_{SP}}{\sqrt{\frac{1-r_{SP}^2}{n-2}}}$$

H_0: $\rho_{xy} = 0$ wird abgelehnt, falls t $t(1-\alpha/2; n-2)$ ist.

Parameter der linearen Regression

Ausgegangen wird von der unbekannten Regressionsgeraden

$$y = \alpha + \beta x + u$$

und der Schätzung

$$y = a + bx + d$$

.

Die Störgröße u ist normalverteilt:

$$u \to N(0; \sigma^2).$$

Die Varianz der Störgröße σ^2 wird geschätzt mit

$$s^2 = \frac{1}{n-2} \sum_i (d_i - \bar{d})^2 = \frac{1}{n-2} \sum_i d_i^2$$

Es ist auch

$$\sum_i d_i^2 = (1 - r^2) \cdot \sum_i (y_i - \bar{y})^2$$

Steigungskoeffizent β

β wird geschätzt durch b. Unter H_0 ist $b \to N(\beta; \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2})$.

Verwendet wird die Prüfgröße

$$t = \frac{b - \beta_0}{\frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}}}$$

die unter H_0 t-verteilt ist mit n-2 Freiheitsgraden.

- H_0: $\beta = \beta_0$ wird abgelehnt, falls t t(1-α/2; n - 2) ist.
- H_0: $\beta \leq \beta_0$ wird abgelehnt, falls t > t(1-α/2; n - 2) ist.
- H_0: $\beta \geq \beta_0$ wird abgelehnt, falls t < - t(1-α/2; n - 2) ist.

In der Praxis wird meistens H_0: $\beta = 0$ getestet. Wird die Hypothese nicht abgelehnt, scheint x unerheblich für die Erklärung von y zu sein.

Absolutglied α

α wird geschätzt durch a. Unter H_0 ist

$$a \to N(\alpha_0; \frac{\sigma^2 \cdot \sum_i x_i^2}{\sum_i (x_i - \bar{x})^2})$$

Für den Test verwendet man die Prüfgröße

$$t = \frac{a - \alpha_0}{s} \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{\sum_i x_i^2}}$$

,

die unter H_0 t-verteilt ist mit n-2 Freiheitsgraden.

- H_0: $\alpha = \alpha_0$ wird abgelehnt, falls t t(1-α/2; n - 2) ist.
- H_0: $\alpha \leq \alpha_0$ wird abgelehnt, falls t > t(1-α/2; n - 2) ist.
- H_0: $\alpha \geq \alpha_0$ wird abgelehnt, falls t < - t(1-α/2; n - 2) ist.

In der Praxis wird meistens H_0: $\alpha = 0$ getestet. Wird die Hypothese nicht abgelehnt, geht die wahre Regressionsgerade möglicherweise durch den Nullpunkt des Koordinatensystems.

Kapitel 9

Übungsaufgaben

Kapitel 1

Aufgabe 1.1 - Mischanlage für Porzellanfabrik

Eine Porzellanfabrik erhält eine neue Mischanlage für spezielles Steingut. Diese muss eingerichtet und angepasst werden. Man geht davon aus, dass die Anlage in höchstens neun Tagen einsatzbereit ist. Wir definieren als Ereignisse

A: Es dauert mehr als 6 Tage, bis die Anlage einsatzbereit ist. B: Es dauert weniger als 8 Tage, bis die Anlage einsatzbereit ist.

1. Beschreiben Sie das Komplement zu A.
2. Beschreiben Sie die Schnittmenge zwischen A und B.
3. Sind A und B disjunkt?
4. Zeigen Sie, dass $(A \cap B) \cup (\bar{A} \cap B) = B$ ist.

Aufgabe 1.2 - Einrichtung der Mischanlage

Wir beziehen uns auf Aufgabe 1.1 Die Werksleitung vermutet für die Zahl der Tage, die benötigt werden, um die Anlage einzurichten, die Wahrscheinlichkeiten, wie in der folgenden Tabelle angegeben:

Zahl der Tage	5	6	7	8	9
Wahrscheinlichkeit	0,05	0,25	0,35	0,25	0,10

1. Geben Sie die Wahrscheinlichkeiten für A und B an.
2. Geben Sie die Wahrscheinlichkeit für die Schnittmenge von A und B an.
3. Geben Sie die Wahrscheinlichkeit für die Vereinigungsmenge von A und B an.
4. Jeder unproduktive Tag kostet die Firma 2000 Euro. Mit welcher Wahrscheinlichkeit muss die Firma mit Kosten von höchstens 12.000 Euro rechnen?

Aufgabe 1.3 Zustelldienst

Ein Zustellungsdienst beschäftigt Festangestellte und freie Mitarbeiter. 64% der Mitarbeiter sind fest angestellt. Eine Qualitätsanalyse ergab, dass 10% aller Zustellungen beanstandet wurden. Die Wahrscheinlichkeit, dass eine Sendung von einem festangestellten Mitarbeiter ausgeliefert wurde und beanstandet wurde, beträgt 6%.

Berta erhält eine Sendung. Mit welcher Wahrscheinlichkeit

1. wird die Sendung beanstandet?
2. stammt die Sendung von einem freien Mitarbeiter?
3. wird die Sendung beanstandet oder stammt von einem Festangestellten?
4. wird die Sendung nicht beanstandet oder stammt nicht von einem Festangestellten?
5. wird die Sendung beanstandet oder stammt nicht von einem Festangestellten?
6. wird die Sendung beanstandet oder nicht beanstandet?
7. wird die Sendung beanstandet, stammt aber nicht von einem Festangestellten?

Aufgabe 1.4 - 2x Würfeln

Sie würfeln zweimal.

1. Geben Sie die Ergebnismenge dieses Zufallsvorgangs an. Zweckmäßig ist eine matrixähnliche Anordnung.
2. Mit welcher Wahrscheinlichkeit erhalten Sie
   (a) beim ersten Wurf 1 und beim zweiten Wurf 5?
   (b) einen Pasch (2x die gleiche Augenzahl)?
   (c) 1 oder 5?
   (d) die Augenzahl 8?
   (e) mindestens die Augenzahl 7?

Aufgabe 1.5 - Münze 4x werfen

Eine Münze wird viermal geworfen. Es ist definiert: Z: Zahl liegt oben. K: Kopf liegt oben.

1. Stellen Sie die 16-elementige Ergebnismenge zusammen.
2. Geben Sie ein Beispiel für ein Ergebnis, ein Elementarereignis, ein zusammengesetztes Ereignis.
3. Es sind die Ereignisse definiert:

   A: Es treten zuerst zweimal Kopf, dann zweimal Zahl auf
   B: Es tritt höchstens zweimal Kopf auf
   C:Es tritt mindestens drei mal Zahl auf
   D: Es tritt einmal Kopf auf
   Ermitteln Sie die Wahrscheinlichkeit, dass

   1. A
   2. D
   3. nicht D
   4. B und C
   5. B oder C

6. A und C
7. nicht C und nicht D
8. nicht Kopf und nicht mindestens 3 mal Zahl
9. B ohne A
10. mindestens einmal Zahl

eintritt.

Aufgabe 1.6 - Aktiengewinne

Die Wertpapierabteilung einer Bank verwendet einen neuen speziellen Index zur Bewertung der zukünftigen Ertragsstärke eines Unternehmens. Eine erste Analyse ihrer Aktienportefeuilles hat ergeben, dass 75% der Aktien, deren Unternehmen als ertragsstark eingestuft worden waren, Kursgewinne einfahren konnten. Es wurden aber auch mit 30% der Aktien als ertragsschwach beurteilter Unternehmen Gewinne erzielt. Zur Vermeidung von Risiken setzten sich die Wertpapierfonds aus 80% Aktien als ertragsstark und 20% Aktien als ertragsschwach beurteilter Unternehmen zusammen.

1. Mit welcher Wahrscheinlichkeit kann von einer Aktie ein Kursgewinn erwartet werden?
2. Wieviel Prozent der Aktien mit Kursverlusten stammten tatsächlich von als ertragsschwach beurteilten Unternehmen?

Aufgabe 1.7 - Heulomat

Die Auto-Alarmanlage Heulomat heult erfahrungsgemäß bei 90% der Autoknacker, die sich am Auto zu schaffen machen. Leider heult sie auch bei 60% aller harmlosen Kollisionen, beispielsweise mit Spaziergängern. Man vermutet, dass insgesamt 80% aller Erschütterungen eines Autos harmlos sind.

1. In wie viel Prozent aller Fälle heult die Anlage **berechtigterweise**?
2. Wie groß ist die Wahrscheinlichkeit, dass die Anlage bei irgendeiner Erschütterung nicht heult?

Aufgabe 1.8 - Zugverspätung

Das Eisenbahnsystem des Staates Mobilia weist im Prinzip zwei Typen von Zügen auf: Den überregionalen Schnellzug ||Hypercity|| und die langsamere Regionalbahn ||Bummelzug||. Der Anteil der Hypercities am Fahrzeugbestand beträgt 20%. Man hat herausgefunden, dass 70% aller Hypercities verspätet sind, wogegen 80% aller Bummelzüge pünktlich ankommen.

Sie stehen am Bahnhof von Capitalis, der Hauptstadt von Mobilia, und sehen dem Treiben an den Bahnsteigen zu. Eine Lautsprecherdurchsage verkündet: ||Der Zug nach Metropolis fährt verspätet ein||.

1. Mit welcher Wahrscheinlichkeit handelt es sich um
   (a) einen Bummelzug?
   (b) einen Hypercity ?

   Wie groß ist der Anteil der Züge mit Verspätung?

Aufgabe 1.9 - Wand verkratzen mit Mülltonnen

In einem Mietshaus wird Dienstags die Mülltonne entleert. Bei 30% der Leerungen stellt Herr Löhlein die Mülltonne raus, bei 20% der Leerungen Frau Susemihl und bei 50% aller Leerungen Herr Feinbein. Eines Tages stellt der Vermieter fest, dass die Wand im Flur verschrammt ist. Er weiß, dass Herr Löhlein beim Mülltonne Tragen mit einer Wahrscheinlichkeit von 7%, Frau Susemihl mit einer Wahrscheinlichkeit von 8% und Herr Feinbein mit einer Wahrscheinlichkeit von 5% mit der Tonne an der Wand entlang kratzen.

1. Welcher Bewohner ist am „verdächtigsten“?
2. Mit welcher Wahrscheinlichkeit wird nächsten Dienstag die Wand verkratzt?
3. Nach jeder Schramme lässt der Vermieter die Wand weißen. Reicht etwa ein Anstrich pro Jahr?

Aufgabe 1.10 - Kaffeetassen

Frau Ahorn, Frau Behorn und Frau Zehorn bestellen nacheinander (in der Reihenfolge der Nennung) im Café Linde Kaffee. Zur Zeit sind noch 24 graue Tassen und 12 rosa Tassen heil. Die Tassen werden in der Reihenfolge der Bestellung zufällig ausgegeben.

1. Wie groß ist die Wahrscheinlichkeit, daß Frau Ahorn eine graue, Frau Behorn und Frau Zehorn eine rosa Tasse erhalten?
2. Wie groß ist die Wahrscheinlichkeit, daß Frau Zehorn eine rosa Tasse erhält?
3. Wie groß ist die Wahrscheinlichkeit, daß mindestens eine Kundin eine graue Tasse erhält?
4. Wie groß ist die Wahrscheinlichkeit, daß genau eine Kundin eine rosa Tasse erhält?
5. Es betreten 10 Kundinnen das Cafè. Wie groß ist die Wahrscheinlichkeit, daß mindestens 9 Kundinnen eine rosa Tasse erhalten? (Ansatz genügt)

Aufgabe 1.11 Kondensatoren

Einem Fertigungslos von 500 Kondensatoren werden fünf Kondensatoren zu Prüfzwecken entnommen. Aufgrund einer ungenauen Wicklung sind 100 schadhafte Kondensatoren im Fertigungslos. Mit welcher Wahrscheinlichkeit taucht kein einziger dieser schadhaften Kondensatoren in der Probe auf?

Aufgabe 1.12 - Schraubensortiment

Einem Heimwerkermarkt werden Schachteln mit Schraubensortimenten geliefert, die jeweils 30 kleine Schrauben, 20 mittlere Schrauben und 10 große Schrauben enthalten. Zu Kontrollzwecken werden den Schachteln Schrauben entnommen.

1. Es wird 3 Schachteln jeweils eine Schraube entnommen. Wie groß ist die Wahrscheinlichkeit,

(a) dass erst eine kleine, dann eine große, dann eine mittlere Schraube resultiert?

(b) dass mindestens eine große Schraube resultiert?

Es werden einer Schachte drei Schrauben (o. Z.) entnommen. Wie groß ist die Wahrscheinlichkeit, dass nur kleine und mittlere Schrauben gezogen werden?

Kapitel 2

Aufgabe 2.1 Münze 3x werfen

Eine Münze wird dreimal geworfen.

1. Geben Sie die acht-elementige Ergebnismenge für den Zufallsvorgang: „Eine Münze wird dreimal geworfen“ an (K: Kopf; Z:. Zahl).
2. Definiert ist die Zufallsvariable X: Anzahl von Kopf bei drei Würfen.

   (a) Bestimmen Sie die Wahrscheinlichkeitsfunktion von X.

   (b) Berechnen Sie den Erwartungswert und die Varianz von X.

   Der Zufallsvorgang ist die Grundlage für ein Glücksspiel. Eine Person zahlt einen Einsatz von 1 Euro. Sie wirft dreimal eine Münze. Für jeden Kopf erhält sie 60 Cents. Es sei die Zufallsvariable Y der Nettogewinn.

   (a) Geben Sie die Wahrscheinlichkeitsfunktion von Y an. Bestimmen Sie daraus E(Y) und VAR(Y).

   (b) Geben Sie Y in Abhängigkeit von X an.

   (c) Überlegen Sie, ob Y eine lineare Transformation von X ist.

   (d) Berechnen Sie gegebenenfalls die Parameter von Y mit Hilfe dieser Erkenntnis.

Aufgabe 2.2 - Urne mit Kugeln

In einer Urne befinden sich 3 rote und 7 blaue Kugeln. Der Urne werden 4 Kugeln ohne Zurücklegen entnommen.

1. Mit welcher Wahrscheinlichkeit erhalten Sie

(a) keine rote Kugel?

(b) mindestens 1 rote Kugel?

(c) vier rote Kugeln?

Es sei definiert X: Zahl der roten Kugeln bei n=4.

(a) Geben Sie für X die Wahrscheinlichkeitstabelle und die Verteilungsfunktion an.

(b) Tragen Sie die Verteilungsfunktion in ein Diagramm ein. Hinweis: Es genügt, wenn Sie für die Ordinate im Nenner 210 stehen lassen.

(c) Geben Sie Erwartungswert und Varianz von X an.

Aufgabe 2.3 - Buchladen

Eine Buchhandlung steht vor der Wahl, ein hochwertiges und sehr teures Faksimile einer mittelalterlichen Handschrift anzubieten. Die Marketingexperten eines beauftragten Instituts vermuten für die Verkaufszahlen X folgende Warscheinlichkeiten:

Verkaufszahl x	0	1	2	3	4	5	mehr als 5
Wahrscheinlichkeit	0,2	0,3	0,2	0,1	0,1	0,1	0

1. Zeichnen Sie die Verteilungsfunktion.
2. Bestimmen Sie die Wahrscheinlichkeit, dass
   (a) höchstens ein Buch
   (b) weniger als zwei Bücher
   (c) mindestens vier Bücher
   (d) mehr als ein, aber höchstens vier Bücher

   verkauft werden.

1. Bestimmen Sie die durchschnittliche Zahl von Bücher, die eine Buchhandlung verkaufen könnte, und die Varianz.

Aufgabe 2.4 - Bäckerei

Die Bäckerei Körnchen hat festgestellt, dass sich die Zahl der täglich verkauften Mischbrote annähernd durch die Zufallsvariable X (in 100) mit einer Dichtefunktion

$$f(x) = \begin{cases} ax & \text{für } 0 \leq x \leq 6 \\ 0 & \text{sonst} \end{cases}$$

beschreiben lässt.

1. An wie viel Prozent der Tage können höchstens 400 Brote verkauft werden?
2. An wie viel Prozent der Tage können mindestens 500 Brote verkauft werden?
3. An wie viel Prozent der Tage können zwischen 400 und 500 Brote verkauft werden?
4. An wie viel Prozent der Tage können genau 600 Brote verkauft werden?
5. Bestimmen Sie a so, dass f tatsächlich eine Dichtefunktion ist.
6. Bestimmen Sie analytisch Verteilungsfunktion, Erwartungswert und Varianz von X.
7. Geben Sie den Median der Verteilung an.
8. Wie viel Brote wurden mindestens an den 20% ||besten|| Tagen verkauft?

Aufgabe 2.5 - 2x Würfeln

Sie würfeln zweimal. Es ist die Zufallsvariable Y definiert als Summe der Augenzahlen der beiden Würfe.

1. Geben Sie Wahrscheinlichkeitstabelle und Verteilungsfunktion von Y an. Erstellen Sie jeweils eine Grafik.
2. Geben Sie die Wahrscheinlichkeit an,
   (a) dass die Summe der Augenzahlen genau 4 beträgt.

(b) dass die Summe der Augenzahlen genau 2,5 beträgt.

(c) dass die Summe der Augenzahlen mindestens 4 beträgt.

(d) dass die Summe der Augenzahlen mehr als 4 beträgt.

(e) dass die Summe der Augenzahlen mehr als 9,5 beträgt.

(f) dass die Summe der Augenzahlen höchstens 3 beträgt.

(g) dass die Summe der Augenzahlen mindestens 4 und höchstens 10 beträgt.

(h) dass die Summe der Augenzahlen mindestens 4 oder höchstens 10 beträgt.

(i) dass Y mehr als 6 und weniger als 8 beträgt.

Bestimmen Sie Erwartungswert und Varianz von Y

Aufgabe 2.6 - Gemeinsame Wahrscheinlichkeiten

Die gemeinsamen Wahrscheinlichkeiten der diskreten Zufallsvariablen X und y sind in der folgenden Wahrscheinlichkeitstabelle zusammengefasst:

X Y	**-2**	**-1**	**0**	**1**	$\mathbf{f_x x(x_i)}$
0	0,05	0,05	0,05	0,1	
1	0	0,1	0,2	0,05	
2	0	0	0,2	0,1	
3	0	0	0	0,1	
$\mathbf{f_y(y_j)}$					

1. Bestimmen Sie Verteilung, Erwartungswert und Varianz von X und Y.
2. Überprüfen Sie, ob X und Y stochastisch unabhängig sind.
3. Ermitteln Sie den Korrelationskoeffizienten von X und Y.

Aufgabe 2.7 - Rendite zweier Aktien

Die Studentin Berta möchte das Geld, das sie durch Programmieraufträge verdient hat, in Aktien anlegen. Ihr erscheinen die Newcomer Scheffel und Raff am aussichtsreichsten. Sie hat die Wahrscheinlichkeiten für die Renditen

(in Croetos), die die beiden Aktien gemeinsam abwerfen, in einer Renditetabelle zusammengefasst:

Scheffel	**Raff**	**Wahrscheinlichkeit**
X	**Y**	$\mathbf{f_{XY}}$
0	0	0,1
0	10	0,1
50	10	0,2
50	30	0,1
100	30	0,2
100	40	0,3

1. Geben Sie die gemeinsame Wahrscheinlichkeitstabelle von X und Y an.
2. Ermitteln Sie die durchschnittliche Rendite einer Aktie und ihre Varianz.
3. Ermitteln Sie den Korrelationskoeffizienten zwischen den Renditen.
4. Berta zahlt ihrem Anlageverwalter jährlich einmal 10 € und dann von der Rendite 1%. Wieviel muss sie ihrem Anlageverwalter jährlich im Durchschnitt zahlen, wenn sie Scheffel und Raff kaufen würde?

Kapitel 3

Aufgabe 3.1 - Abnahmekontrolle von Elektronik

Bei einer sehr großen Lieferung von hochwertigen elektronischen Bauteilen wird ein Ausschussanteil von 5% als akzeptabel angesehen. Bei der Abnahmekontrolle werden 15 Stück zufällig entnommen. Falls höchstens ein fehlerhaftes Stück auftritt, wird die Lieferung angenommen.

1. Bestimmen Sie die Wahrscheinlichkeit,
   (a) dass die Lieferung angenommen wird, wenn tatsächlich 5% Ausschuss vorliegen.
   (b) dass die Lieferung irrtümlicherweise abgelehnt wird, wenn tatsächlich 3% Ausschuss vorliegen.
   (c) dass die Lieferung irrtümlicherweise angenommen wird, wenn tatsächlich 10% Ausschuss vorliegen.

Wie groß muss die Stichprobe mindestens sein, damit die Wahrscheinlichkeit für eine irrtümliche Annahme der Lieferung bei 10% Ausschuss höchstens 10% beträgt? Verwenden Sie dazu die Binomialverteilungstabelle.

2. Oft vermeidet man die Abnahmeregel, dass eine Lieferung nur angenommen wird, wenn kein fehlerhaftes Stück auftritt, weil man diese Regel für zu streng hält. Beurteilen Sie diese Ansicht.

Aufgabe 3.2 - Bank nach 18 Uhr

Die Zahl der Kunden, die nach 18 Uhr während einer Stunde einen Bankschalter in einer Bankfiliale aufsuchen, ist poissonverteilt mit $\lambda = 10$.

1. Wie viele Kunden suchen stündlich im Durchschnitt einen Bankschalter auf?
2. Wie groß ist der Anteil der Stunden, in denen höchstens drei Kunden an einen Schalter kommen?
3. Wie groß ist der Anteil der Stunden, in denen mindestens zwei Kunden an einen Schalter kommen?

Aufgabe 3.3 - LKW-Versicherung

Die Zahl der Versicherungsfälle, die einer gewerblichen Haftpflichtversicherung durch einen LKW entstehen, ist annähernd poissonverteilt mit dem Parameter $\lambda = 2{,}5$.

1. Bei wie viel Prozent der LKWs muss die Versicherung in einem Jahr keinen Schadensersatz leisten?
2. Wie viel Prozent der LKWs verursachen mindestens drei Versicherungsleistungen?
3. Eine Firma betreibt für just in time Lieferungen drei LKWs. Verursacht keiner der LKWs Versicherungsleistungen, bekommt die Firma 2000 Euro gutgeschrieben, falls doch, ändert sich finanziell nichts für die Firma. Ist das Angebot der Versicherung Ihrer Meinung nach attraktiv für die Firma?

Aufgabe 3.4 - Batterienfunktion

Für die Tauglichkeitsprüfung eines MP3-Players wurde geprüft, wie lange man ihn mit einem Batteriensatz spielen kann. Es stellte sich heraus, dass die Funktionsdauer eines Batteriensatzes annähernd normalverteilt ist mit dem Erwartungswert von 200 Minuten und einer Standardabweichung von 20 Minuten.

1. Bestimmen Sie die Wahrscheinlichkeit, dass ein MP3-Player mit einem Batteriensatz höchstens zwei Stunden aushält.
2. Wie viel Prozent der MP3-Player schaffen mindestens 150 Minuten?
3. Mit welcher Wahrscheinlichkeit spielt ein MP3-Player zwischen zwei und dreieinhalb Stunden?
4. Bestimmen Sie d derart, dass der Anteil der MP3-Player, die zwischen μ - d und μ + d aushalten, 90% beträgt.

Aufgabe 3.5 - Küchenschaben

Eine Diplomarbeit über Küchenschaben hat ergeben, dass die Länge von Küchenschaben in einer bestimmten Altbauwohnung normalverteilt ist mit dem Erwartungswert 3 cm und der Varianz 4 cm^2. In der Nacht wird eine Schabe zufällig eingefangen.

Bestimmen Sie die Wahrscheinlichkeit, dass diese Schabe

1. (a) mindestens 5 cm
   (b) zwischen 2 und 5 cm
   (c) höchstens 1 cm
   (d) höchstens 2 oder mindestens 4 cm

lang ist.

Welche Mindestgröße haben die 10% größten Schaben?

Aufgabe 3.6 - Galapagos

Bei einer umfassenden Bestandsaufnahme von Großechsen auf einer Galapagosinsel stellte sich heraus, dass das Gewicht X dieser Echsen annähernd normalverteilt ist. 15,78% der Echsen wogen mehr als 120 kg. x(0,33) betrug 75.

1. Tragen Sie die Angaben in die Grafik ein, wobei die Eintragungen nicht exakt maßstabsgetreu sein müssen.
2. Wieviel wogen die Echsen im Durchschnitt?
3. Wieviel betrug die durchschnittliche quadratische Abweichung der Gewichte vom Mittel?

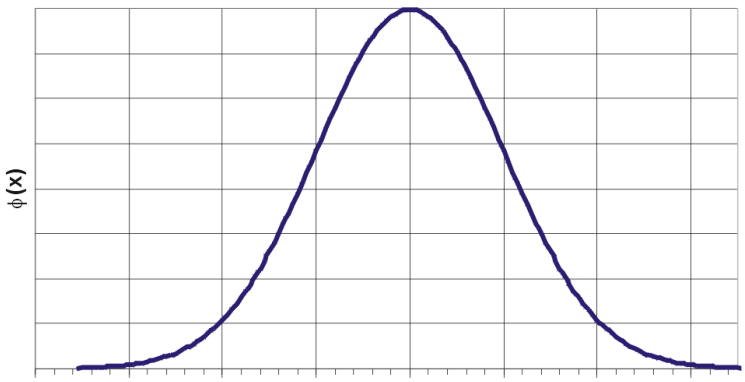

Abbildung 122

Aufgabe 3.7 - Nähfehler

Es ist bekannt, dass in einem Unternehmen, das Unterwäsche produziert, der Anteil von Spitzen-Damenunterhemden mit Nähfehlern etwa 10% beträgt. Der tägliche Output ist sehr groß. Es werden während eines Tages für die Warenkontrolle n=200 Hemdchen zufällig ausgewählt.

1. Bestimmen Sie die exakte Wahrscheinlichkeit, dass mindestens 15 Hemdchen Mängel aufweisen (nur Ansatz).

2. Berechnen Sie, falls möglich, die obige Wahrscheinlichkeit näherungsweise.
3. Mit welcher Wahrscheinlichkeit erhält man bei der Qualitätskontrolle mindestens 20 und höchstens 30 Hemdchen mit Fehlern?
4. Mit welcher Wahrscheinlichkeit erhält man bei der Qualitätskontrolle genau 20 fehlerhafte Hemdchen?
5. Ist es wahrscheinlicher, 19 bis 21 oder 23 bis 25 fehlerhafte Hemdchen zu erhalten?

Kapitel 4 "'

Kapitel 5

Aufgabe 5.1 - Hotelsterne

Eine Reiseveranstalter hat 9 Kunden nach ihrer Zufriedenheit mit dem Hotel befragt, das sie im letzten Urlaub hatten.

Kunde Nr. Sterne des Hotels Note des Kunden 1 * 3 2 *** 2 3 ** 2 4 ** 4 5 *** 1 6 ** 1 7 ** 3 8 **** 1 9 * 4

Ermitteln Sie den Rangkorrelationskoeffizienten der Sterne mit der Zufriedenheit

Aufgabe 5.2 - Solaranlagen

Eine Heizungsfirma hat in den letzten 8 Monaten jeweils x mal in der regionalen Tageszeitung inseriert. Sie konnte in diesen Monaten jeweils y viele Solaranlagen verkaufen.

Es ergab sich

Monat i 1 2 3 4 5 6 7 8 Inserate x 0 2 2 4 4 6 6 8 Solaranlagen y 6 6 8 8 12 8 16 16

1. Tragen Sie die Wertepaare in einem Streudiagramm ab.
2. Ermitteln Sie die Regressionsgerade y = a + bx und tragen Sie sie in das Diagramm ein.

3. Berechnen Sie die geschätzten Werte und die Residuen.
4. Berechnen Sie das Bestimmtheitsmaß.
5. Ermitteln Sie die Varianzen von y, $\hat{y}$ und der Residuen. Zeigen Sie, dass die Streuungszerlegung hier gilt und ermitteln Sie das Bestimmtheitsmaß als Anteil der durch $\hat{y}$ erklärten Streuung an der Gesamtstreuung von y.

Aufgabe 5.3 - Fair-Trade-Tee

Gegeben ist für die Jahre 1998 bis 2003 die Zahl der in der EU verkauften Tonnen Tee im fairen Handel.

Jahr Zeitpunkt x Menge y 1998 1 612 1999 2 842 2000 3 890 2001 4 1004 2002 5 1154 2003 6 1414

1. Ermitteln Sie eine Regressionsgerade, die die Entwicklung des Verkaufs im Lauf der Jahre beschreibt.
2. Berechnen Sie das Bestimmtheitsmaß.

Kapitel 6 "'

Kapitel 7

Aufgabe 7.1 - Tarifsystem

Eine Analyse der Kundenzufriedenheit eines großen Verkehrsbetriebes gab Anlass zu der Befürchtung, dass 75% der Fahrgäste das Tarifsystem nicht verstanden hätten.

1. 75% der Kunden haben das Tarifsystem nicht verstanden. Es wurden in einem zentral gelegenen U-Bahnhof zufällig 10 Personen befragt.
   (a) Mit welcher Wahrscheinlichkeit hat jeder die Tarifordnung verstanden?
   (b) Mit welcher Wahrscheinlichkeit haben genau 8 Personen die Tarifordnung verstanden?
   (c) Mit welcher Wahrscheinlichkeit haben mindestens 2 Personen die Tarifordnung nicht verstanden?

(d) Mit welcher Wahrscheinlichkeit haben an zwei aufeinanderfolgenden Tagen jeweils mindestens zwei Personen das Tarifsystem nicht verstanden, wenn die Befragungen stochastisch unabhängig waren.

Es wurden 100 Personen befragt.

(a) Mit welcher Wahrscheinlichkeit haben genau 75 Personen die Tarifordnung nicht verstanden?

(b) Mit welcher Wahrscheinlichkeit haben höchstens 75 Personen die Tarifordnung nicht verstanden?

(c) Es haben 70 Kunden angegeben, das System nicht verstanden zu haben. Überprüfen Sie die Hypothese ($\alpha = 0{,}05$), dass mindestens 75% die Tarifordnung nicht verstanden haben.

Aufgabe 7.2 - Kaviar

Ein Delikatessengroßhandel erhält eine umfangreiche Lieferung von 50-g-Schalen Kaviar. Es ist bekannt, dass die Füllmenge des Kaviars normalverteilt ist. Der Lieferant versichtert, dass sich in jeder Dose im Mittel mindestens 50 g Kaviar befänden. Es werden zu Prüfzwecken 6 Schälchen zufällig ausgewählt und geöffnet. Man erhält die Urliste

47 49 50 52 50 46

1. Prüfen Sie die Behauptung des Lieferanten ($\alpha = 0{,}1$).
2. Würde sich die Position des Lieferanten verschlechtern, wenn man ein Signifikanzniveau von 0,05 verwenden würde?

Kapitel 10

Statistik auf dem Computer

=Einfache Statistikprogramme=

Statistische Berechnungen mit der Tabellenkalkulation von Open Office

Auf vielen Rechnern findet sich mittlerweile das Programm Open Office.

Dort können Sie innerhalb der Tabellenkalkulation viele statistische Funktionen ausführen.

Öffnen Sie dazu ein neues Dokument in Open Office und wählen Sie bei der Art des Dokumentes *Tabellenkalkulation*.

Geben Sie Ihre Werte beispielsweise in der ersten Spalte A ein.

Über *Einfügen*, *Funktion* können Sie verschiedene Berechnungen auswählen, die Sie in einem freien Feld durchführen lassen. Ihre Werteliste markieren Sie für die Berechnung. Das Ergebnis erhalten Sie nach anklicken des grünen Häckchens neben der Eingabezeile.

Probieren Sie das ganze mit folgender Werteliste aus. Man kann sie mittels Zwischenspeicher direkt in die Tabellenkalkulation übernehmen. (Anmarkieren, mit Strg + C in den Zwischenspeicher holen, im ersten Feld der Tabellenkalkulation mit Strg + V wieder abladen.)

114,3 135,7 104,8 118,5 125,7 121,4 122,4 96,8 118,9 120 112,2 127,9 122,8 128,9 120,3

Versuchen Sie den Median, den Mittelwert, die Standardabweichung, die Varianz und den Maximalwert zu ermitteln. Dazu gehen Sie in ein leeres Feld am Ende ihrer Liste. Dann wählen Sie im Menü den Eintrag *Einfügen* und dort wieder *Funktionsliste*. Aus der Funktionsliste wählen Sie die statistischen Funktionen aus. Aus dem großen Angebot wählen Sie den *Median*.

Sie können auch direkt in die Eingabezeile folgendes eingeben:

=MEDIAN(A1:A15)

oder

=MITTELWERT(A1:A15)

Es müßten folgende Werte herauskommen:

Summe: 1790,6

Mittelwert 119,37

Median 120,3

Maximal 135,7

Standardabweichung 9,62

Varianz 92,6

Statistik mit Gnumeric "'

Gnumeric ist die Tabellenkalkulation unter Gnome. Es bietet bessere statistische Berechnungsmöglichkeiten als Excel. Siehe http://de.wikipedia.org/wiki/Gnumeric Siehe http://www.gnome.org/projects/gnumeric/

Statistische Berechnungen mit der Programmiersprache Gambas "'

Auf vielen Linuxrechnern findet sich mittlerweile die einfach zu lernende Programmsprache Gambas.

Dort kann man viele statistische Funktionen nachvollziehen. Es gilt das alte Motto: Habe ich es noch nicht programmiert, dann habe ich es noch nicht verstanden.

Im Gambas Wikibook sollen nach und nach eine Reihe von Statistikfunktionen im Quelltext erklärt und verfügbar gemacht werden.

Siehe http://de.wikibooks.org/wiki/Gambas:_Statistik

=Komplexere, professionelle Statistik-Software=

R (Windows, OS X, Linux) "'

R ist ein eine umfangreiche Statistiksoftware, genauer: eine *Programmierumgebung* für statistische Auswertungen. Im Funktionsumfang mit kommerziellen Softwarepaketen wie SPSS oder STATA durchaus vergleichbar (und stellenweise überlegen) werden eine Vielzahl statistischer Methoden und Routinen bereitgestellt. Der Programmaufbau mag zwar insbesondere für Anfänger etwas unübersichtlich sein, besticht jedoch durch zahlreiche Features: Vollständige Kontrolle über die Daten, Implementation einer grossen Anzahl an Analyse-Verfahren, flexible Graphikfähigkeiten, Systemunabhängigkeit, automatisierte Auswertungen, Schnittstellen zu vielen anderen Anwendungen und nicht zuletzt kostenlose Verfügbakeit könnten dazu führen, dass *R* sich im professionellen Bereich zum neuen Standard entwickelt.

Einsteiger können zudem auf graphische Bedienoberflächen zurückgreifen.

Siehe:

- R
  - http://www.r-project.org/
  - http://de.wikibooks.org/wiki/GNU_R

  Graphische Bedienung (GUIs):
  - Jaguar: http://stats.math.uni-augsburg.de/JGR/
  - R Commander: http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/
  - Das Statistiklabor: http://www.statistiklabor.de

SPSS (Windows, OS X, Linux (nur Server-Variante))

Insbesondere in den Sozial- und Verhaltenswissenschaften findet die kommerzielle Software SPSS grossen Zuspruch. Hauptvorteile sind einfache Bedienbarkeit für Anwender, die die Steuerung mit Menüs und Maus favorisieren. Tabellen und Graphiken sind für die Weiterverwendung in Office-Anwendungen optimiert und können nachträglich formatiert werden. Zudem gibt es Ergänzungspakete, die den Prozess der Datenerfassung erleichtern. Integriert sind die meisten gebräuchlichen Standardverfahren uni- und multivariater Statistik. Spezielle Anwendungen werden als Zusatzpakete vertrieben.

- Wikipedia über SPSS[1]
- http://www.spss.com/de/

STATA (Windows, OS X, Linux)

STATA ist eine Statistiksoftware, die bislang hauptsächlich im angloamerikanischen Raum Verbreitung gefunden hat. Wenngleich STATA ebenfalls über eine gut strukturierte Menübedienung verfügt, besticht die Software vor allem durch ihre an BASIC[2] erinnernde, relativ einfach zu erlernende Befehlssyntax und eine umfängliche, übersichtliche Integration weitreichender statistischer Verfahren.

- Wikipedia über STATA[3]
- http://www.stata.com

[1] http://de.wikipedia.org/wiki/spss
[2] http://de.wikipedia.org/wiki/BASIC
[3] http://de.wikipedia.org/wiki/stata

Kapitel 11

Literatur

- Grundlagen der Statistik, Bd.1, Beschreibende Verfahren von Jochen Schwarze, Verlag Neue Wirtschafts-Briefe (Januar 2001)
- Grundlagen der Statistik, Bd.2, Wahrscheinlichkeitsrechnung und induktive Statistik von Jochen Schwarze, Verlag Neue Wirtschafts-Briefe (Oktober 2001)
- Aufgabensammlung zur Statistik von Jochen Schwarze Verlag Neue Wirtschafts-Briefe (Januar 2002)
- Wahrscheinlichkeitsrechnung und schließende Statistik . Praxisorientierte Einführung. Mit Aufgaben und Lösungen von Günther Bourier
- Statistik-Übungen von Günther Bourier
- Beschreibende Statistik von Günther Bourier
- Stochastik Leistungskurs Lambacher Schweizer, Klett Verlag

  Didaktisch gut gemacht, mit sehr vielen Übungsaufgaben
- Stochastik Leistungskurs Lambacher Schweizer, Klett Verlag Lösungsheft
- Medizinische Statistik, Von Herbert Immich , Schattauer Verlag

  Leider nur noch antiquarisch, didaktisch hervorragend gemacht, mit sehr vielen praktischen medizinischen Beispielen.

[FISZ89] Marek Fisz, *Wahrscheinlichkeitsrechnung und mathematische Statistik*, VEB Deutscher Verlag der Wissenschaften, Berlin 1989, ISBN 3-326-00079-0

Für Fortgeschrittene. Wie ich finde, ein umfassendes Lehrbuch mit einer guten Einführung in die Wahrscheinlichkeitsrechnung und vielen guten Beispielen. Die Testverfahren sind meiner Meinung nach gut, aber sehr theoretisch, beschrieben.

Kapitel 12

Autoren

Edits	User
3	Mca
2	Peter.Stadler
13	ThePacker
1	Anitagraser
3	Berni
1	LeSpocky
2	Roland
1	Ragdoll
1	Carbidfischer
1	Wolfgang1018
1	Gymnasiallehrer
7	Gabriele Hornsteiner
111	Klartext
1	Fippu
2	Shogun
3	Klaus Eifert
1	Enrico
1	Octanitrocuban
1	LukasHager
21	Gabkdlly
5	Heuler06
14	Nijdam
3	NigheanRuach
1	Boehm
3	ASchumacher
1	Mathaxiom

1	Much89
2	Hartwigbaumgaertel
1	Moolsan
2	Af8
880	Philipendula
1	Fristu
2	Oliver Jennrich
1	Stefan Majewsky
1	Mceyran
4	JohannWalter
7	MichaelFrey
1	Hjn
6	RolandT
3	Max Vallender
3	Dominiklenne
1	Zero
2	Laudrin
15	Europol
1	KurtWatzka
3	Schmock
2	Elasto
1	Caustic
1	Produnis
1	Gardini
101	Dirk Huenniger

Kapitel 13

Bildnachweis

In der nachfolgenden Tabelle sind alle Bilder mit ihren Autoren und Lizenen aufgelistet.

Für die Namen der Lizenzen wurden folgende Abkürzungen verwendet:

- GFDL: Gnu Free Documentation License. Der Text dieser Lizenz ist in einem Kapitel diese Buches vollständig angegeben.
- cc-by-sa-3.0: Creative Commons Attribution ShareAlike 3.0 License. Der Text dieser Lizenz kann auf der Webseite http://creativecommons.org/licenses/by-sa/3.0/ nachgelesen werden.
- cc-by-sa-2.5: Creative Commons Attribution ShareAlike 2.5 License. Der Text dieser Lizenz kann auf der Webseite http://creativecommons.org/licenses/by-sa/2.5/ nachgelesen werden.
- cc-by-sa-2.0: Creative Commons Attribution ShareAlike 2.0 License. Der Text der englischen Version dieser Lizenz kann auf der Webseitehttp://creativecommons.org/licenses/by-sa/2.0/ nachgelesen werden. Mit dieser Abkürzung sind jedoch auch die Versionen dieser Lizenz für andere Sprachen bezeichnet. Den an diesen Details interessierten Leser verweisen wir auf die Onlineversion dieses Buches.
- PD: This image is in the public domian. Dieses Bild ist gemeinfrei.
- ATTR: The copyright holder of this file allows anyone to use it for any purpose, provided that the copyright holder is properly attributed.

Redistribution, derivative work, commercial use, and all other use is permitted.

Bild	Autor	Lizenz
1	Philipendula	GFDL
2	Philipendula	GFDL
3	Philipendula	GFDL
4	Philipendula	GFDL
5	Philipendula	GFDL
6	Philipendula	GFDL
7	Philipendula	GFDL
8	Philipendula	GFDL
9	Philipendula	GFDL
10	Philipendula	GFDL
11	Philipendula	GFDL
12	Philipendula	GFDL
13	Philipendula	GFDL
14	Philipendula	GFDL
15	Philipendula	GFDL
16	Philipendula	GFDL
17	Philipendula	GFDL
18	Philipendula	GFDL
19	Philipendula	GFDL
20	Philipendula	GFDL
21	Philipendula	GFDL
22	Philipendula	GFDL
23	Philipendula	GFDL
24	Philipendula	GFDL
25	Philipendula	GFDL
26	Philipendula	GFDL
27	User:Philipendula	GFDL
28	Philipendula	GFDL
29	Philipendula	GFDL
30	Philipendula	GFDL
31	Philipendula	GFDL
32	Philipendula	GFDL
33	Philipendula	GFDL
34	Philipendula	GFDL
35	Philipendula	GFDL
36	Philipendula	GFDL
37	Philipendula	GFDL
38	Philipendula	GFDL

39	Philipendula	GFDL
40	Philipendula	GFDL
41	Philipendula	GFDL
42	Philipendula	GFDL
43	Philipendula	GFDL
44	Philipendula	GFDL
45	Philipendula	GFDL
46	Philipendula	GFDL
47	Philipendula	GFDL
48	Philipendula	GFDL
49	Philipendula	GFDL
50	Philipendula	GFDL
51	Philipendula	GFDL
52	Claudia Hoffmann (Lady whiteadder) with R	PD
53	Philipendula	GFDL
54	Philipendula	GFDL
55	Philipendula	GFDL
56	Philipendula	GFDL
57	Philipendula 10:59, 13. Okt 2004 (UTC)	GFDL
58	Philipendula	GFDL
59	Philipendula	GFDL
60	Philipendula	GFDL
61	Philipendula	GFDL
62	Philipendula	GFDL
63	Philipendula	GFDL
64	Philipendula	GFDL
65	Philipendula	GFDL
66	Philipendula	GFDL
67	Philipendula	GFDL
68	Philipendula	GFDL
69	Philipendula	GFDL
70	Philipendula	GFDL
71	Philipendula	GFDL
72	Philipendula	GFDL
73	Philipendula	GFDL
74	Philipendula	GFDL
75	Philipendula	GFDL
76	Philipendula	GFDL
77	Philipendula	GFDL
78	Philipendula	GFDL
79	Philipendula	GFDL

80	Philipendula	GFDL
81	Philipendula	GFDL
82	Philipendula	GFDL
83	Philipendula	GFDL
84	User:Philipendula	GFDL
85	Philipendula	GFDL
86	Philipendula	GFDL
87	Benutzer:Philipendula	GFDL
88	Philipendula	GFDL
89	Philipendula	GFDL
90	Philipendula	GFDL
91	User:Philipendula	GFDL
92	Philipendula	GFDL
93	Philipendula	GFDL
94	Philipendula |other_versions	GFDL
95	Philipendula	GFDL
96	Philipendula	GFDL
97	Philipendula	GFDL
98	Philipendula	GFDL
99	Philipendula	GFDL
100	Philipendula	GFDL
101	Philipendula	GFDL
102	Philipendula	GFDL
103	Philipendula	GFDL
104	Philipendula	GFDL
105	Philipendula	GFDL
106	Philipendula	GFDL
107	Philipendula	GFDL
108	Philipendula	GFDL
109	Philipendula	GFDL
110	Philipendula	GFDL
111	Philipendula	GFDL
112	Philipendula	GFDL
113	Philipendula	GFDL
114	Philipendula	GFDL
115	Philipendula	GFDL
116	Philipendula	GFDL
117	Philipendula	GFDL
118	Philipendula	GFDL
119	Philipendula	GFDL
120	Philipendula	GFDL

121	Philipendula	GFDL
122	Benutzer:Philipendula	GFDL

Kapitel 14

GNU Free Documentation License

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or

XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it

has fewer than five), unless they release you from this requirement.

- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple

Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements."

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sublicense or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.