von der Lippe, Peter

Book — Accepted Manuscript (Postprint)

Deskriptive Statistik

Kapitel 1: Gegenstand und Grundbegriffe der Statistik

1. Gegenstand der Statistik

Statistik ist die Lehre von Methoden zur Gewinnung, Charakterisierung und Beurteilung von zahlenmäßigen Informationen über die Wirklichkeit (Empirie).

Zu den Bestandteilen dieser Definition sei folgendes bemerkt:

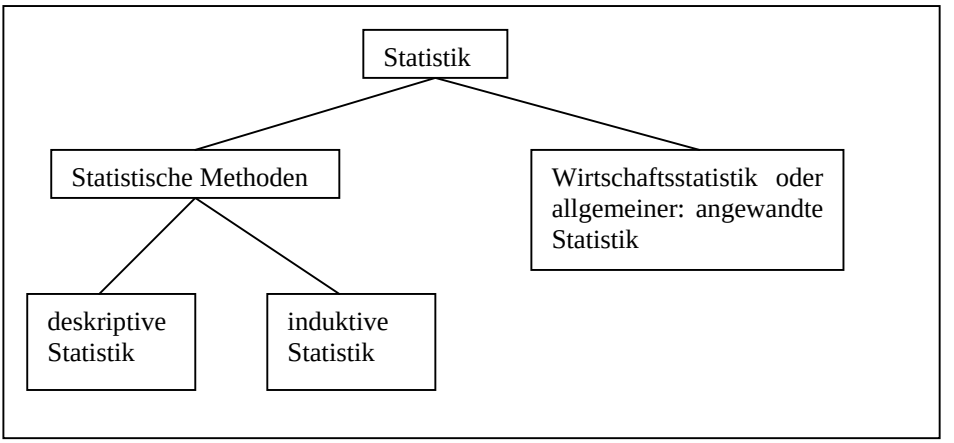
- **Information** ist hier im sehr weiten Sinne gemeint. Alle Sachverhalte, die zähl- oder meßbar und systematisch zu beobachten sind, können Gegenstand der Statistik sein. Dabei ist zu berücksichtigen, dass "messbar" im Alltagssprachgebrauch meist enger definiert wird, als in der Statistik (vgl. Abschn. 3).
- Unter **Gewinnung** von Information wird neben der eigentlichen Datenerhebung auch die Operationalisierung und Systematisierung von Konzepten verstanden, was v.a. ein Problem der Wirtschaftsstatistik[1] ist, sowie die Planung der Datenerhebung (design of experiments, design of surveys). Gegenstand der Statistik ist somit nicht nur die Auswertung bereits vorliegender Daten.
- Unter **Charakterisierung** soll hier zunächst die graphische und tabellarische Darstellung von Daten sowie die Berechnung von zusammenfassenden, den empirischen Sachverhalt beschreibenden Kennzahlen, verstanden werden. Dies ist primär Gegenstand der Deskriptiven Statistik. Die Möglichkeiten der statistischen Auswertung gehen aber über das weit hinaus, was üblicherweise im Rahmen der Deskriptiven Statistik an Methoden bereitgestellt wird (z.B. Multivariate Analyse).
- **Beurteilung** kann z.B. erfolgen durch Schlüsse auf der Basis unvollständiger Informationen (z.B. Schlüsse von der Stichprobe auf die ihr zugrundeliegende Grundgesamtheit), bzw. allgemeiner auf der Basis

[1] Vgl. hierzu: v.d. Lippe, Peter, Wirtschaftsstatistik, in dieser Reihe (UTB Bd. 209).

unsicherer Informationen unter Anwendung der Wahrscheinlichkeitsrechnung. Das ist Gegenstand der Induktiven Statistik.

Mithin ergibt sich die traditionelle Aufteilung der "Statistik", wie sie in Übersicht 1.1 dargestellt ist.

Übersicht 1.1 Aufbau des Faches Statistik



Leider wird in vielen Lehrbüchern und ganzen Studiengängen der Inhalt des Faches Statistik fast ausschließlich auf die Induktive Statistik eingeengt.

Statistik kann für empirische Untersuchungen nützlich sein bei:

a) der Berechnung zusammenfassender Kennzahlen im Rahmen der Deskriptiven Statistik,
b) Fragen der Verallgemeinerungsfähigkeit statistischer Aussagen (Induktive Statistik) und bei
c) der Beurteilung der Aussagefähigkeit statistischer Daten und Ergebnisse aufgrund der ihnen zugrundeliegenden Erhebungen und Konzepte (Wirtschaftsstatistik).

Die Statistik dient damit drei Zwecken, nämlich deskriptiven (Beschreibung, Bestandsaufnahme), analytischen (Verallgemeinerung, Erklärung) und operativen (Entscheidung) Zwecken. Sie kann auf allen drei Stufen der empirischen Arbeit eingesetzt werden, nämlich bei der:

- Formulierung von Hypothesen, Modellbildung, concept formation
- Planung und Durchführung von Erhebungen und Versuchen und
- bei der Überprüfung von Hypothesen.

Die Methoden der Statistik sind allgemein anwendbar, d.h. sie sind nicht beschränkt auf bestimmte inhaltliche Fragestellungen. Dies heißt aber nicht notwendig, dass sie auch in jedem Fall sinnvoll angewendet werden.

Es ist deshalb wohl das wichtigste Ziel der Beschäftigung mit der Disziplin "Statistik", beurteilen zu lernen, bei welcher Fragestellung und auf welche Daten eine bestimmte statistische Methode sinnvoll angewendet werden kann.

Die Frage, welche Folgerungen aus empirischen Daten gezogen werden sollten ist kein Gegenstand der Statistik, weshalb die Statistik auch offen ist für jede Art von Mißbrauch. Für diese Frage ist aber die Aussagefähigkeit der Daten wesentlich (bzw. sollte es sein) und es gibt erhebliche Unterschiede bei Anwendern der Statistik hinsichtlich der Bereitschaft und Fähigkeit, sich hiermit auseinanderzusetzen.

2. Einheiten, Masse, Merkmal

Def. 1.1: Einheit, Masse

a)	Statistische Einheiten (Elemente, Merkmalsträger) sind Träger von Informationen, bzw. Eigenschaften, die im Rahmen einer empirischen Untersuchung von Interesse sind.
b)	Eine statistische Masse (Kollektiv, Population) ist eine hinsichtlich sachlicher, räumlicher und zeitlicher Kriterien sinnvoll gebildete Gesamtheit von statistischen Einheiten.
c)	Unter dem Umfang einer Masse versteht man die Anzahl ihrer Einheiten (Elemente).

Bemerkungen zur Def. 1.1:

1. Beispiel für Einheiten sind Personen, Personengruppen, Fälle bzw. Ereignisse (z.B. Verurteilung, Eheschließung, Erkrankung), Gegenstände (Kranfahrzeuge oder Gebäude bei einer Gebäudezählung), Wirtschaftszweige, Regionen usw. Zu unterscheiden sind evtl. Erhebungs-, Zähl-, Darstellungs- oder Auswertungseinheiten.

2. Das Begriffspaar "Masse - Einheit" entspricht dem Begriffspaar "Menge - Element" aus der Mathematik.

3. Eine Masse muss sachlich, zeitlich und räumlich eindeutig definiert (abgegrenzt) sein. Dies kann erfolgen durch Aufzählung der Einheiten oder durch Angabe eines Prinzips, nach dem über die Zugehörigkeit eines Elements zur Masse entschieden wird, d.h. durch Identifikationsmerkmale (die zu unterschieden sind von den Untersuchungsmerkmalen [Def. 1.2]).

4. Die sachliche Abgrenzung von Massen und Einheiten kann schwierig sein (Einzelheiten hierzu gehören in die Wirtschaftsstatistik). Hierfür zwei Beispiele:

   - für Einheiten: Sollen Unternehmen, Betriebe, Arbeitsstätten oder fachliche Einheiten der Erhebung zugrundegelegt werden?
   - für Massen: Soll "Bevölkerung" im Sinne der Wohnbevölkerung, der ortsanwesenden Bevölkerung oder der Staatsangehörigkeit (rechtlicher Bevölkerungsbegriff) definiert werden?

Arten von Massen (zu unterscheidende Begriffspaare):

a) Nach dem Umfang bzw. der Vollständigkeit der untersuchten Masse (Objektmenge) unterscheidet man zwischen **(Grund-)gesamtheit** (oder Kollektiv, Population) und **Teilgesamtheit**.

   Eine Teilgesamtheit kann durch **Auswahl** von Einheiten aus der Grundgesamtheit oder durch eine **begriffliche Ausgliederung** entstehen. Im zweiten Sinne, d.h. durch Begriffshierarchien (Systematiken, Klassifikationen) mit Oberbegriffen und Unterbegriffen ist z.B. die Erwerbsbevölkerung eine Teilmasse der Wohnbevölkerung. Wird die Teilgesamtheit durch Zufallsauswahl gewonnen, so spricht man von einer **Stichprobe**. Von vielen Autoren wird jede Art der Teilerhebung (auch eine nicht-zufällige Auswahl) als Stichprobe bezeichnet (ein Sprachgebrauch, dem wir uns nicht anschließen wollen).

b) Nach der Verweildauer der beobachteten Einheiten einer Masse unterscheidet man **Bestandsmassen** (stocks) und **Bewegungsmassen** (vgl. Kapitel 12):

c) Man kann auch zwischen **realen** und **hypothetischen Gesamtheiten** unterscheiden.

   Reale (beobachtete) Gesamtheiten, die allein Gegenstand der Deskriptiven Statistik sind, sind stets endlich. Hypothetische, d.h. durch Abstraktion gebildete Massen sind dagegen meist unendlich (z.B. die Menge aller möglichen Würfe mit einer Münze). Sie stehen in der Induktiven Statistik im Vordergrund des Interesses.

Def. 1.2: Merkmal

Ein **Merkmal** ist eine Eigenschaft einer statistischen Einheit, die bei einer statistischen Untersuchung interessiert. Es hat endlich und unendlich viele **Merkmalsausprägungen** (mögliche Realisationen, Modalitäten). Ein Merkmal ist somit eine Menge von Merkmalsausprägungen. Ein **Merkmalswert** ist eine an einer statistischen Einheit ermittelte Merkmalsausprägung. Werden Merkmalswerten Zahlen zugeordnet, so spricht man auch von **Variablen**

Bemerkungen zu Def. 1.2:

1. Es sind Identifikationsmerkmale zur Abgrenzung einer Masse (vgl. Bem. 3 zu Def. 1.1) und Untersuchungsmerkmale, die Gegenstand einer Erhebung sind, zu unterscheiden. Nur letztere sind in Def. 1.2 gemeint.

2. Ein Merkmal stellt eine Abbildung der Masse, d.h. der Menge empirischer Einheiten in der Menge der Merkmalsausprägungen dar. Jeder statistischen Einheit wird eine und nur eine Ausprägung zugeordnet (Ausnahme: häufbare Merkmale). Im folgenden wird i.d.R. ein Merkmal mit großen Buchstaben (etwa X) und eine Merkmalsausprägung mit kleinen Buchstaben (etwa x_i) symbolisiert.

3. Gegenstand der Statistik sind stets Aussagen über Massen in bezug auf bestimmte Merkmale (z.B. Aussagen über die Einkommensverteilung der Haushalte in der Bundesrepublik Deutschland 1990), nicht Aussagen über einzelne Einheiten und auch nicht Aussagen ohne Bezugnahme auf genau definierte Merkmale der Einheiten.

   Einheiten (z.B. Haushalte) interessieren nur in ihrer Eigenschaft als Merkmalsträger, d.h. nicht in ihrer Totalität (mit "allen" ihren Kennzeichen) und Individualität (Statistik ist nie eine Einzelfalluntersuchung; im Zuge der Auswertung interessieren nur anonyme Daten).

4. Ein Merkmal muss **operational definiert** sein, d.h. es muss bei der Beobachtung einer Einheit entscheidbar sein, welche Merkmalsausprägung vorliegt. Wird das Merkmal "Bildung" gemessen durch die Anzahl der Jahre des Schulbesuchs, so ist die Bildung damit zwar sehr eng und nicht notwendig auch sachgerecht, dafür aber operational definiert. Unterscheidet man dagegegen die Ausprägungen "hochgebil-det", "gebildet", "ungebildet" (nach welchen Kriterien?), so liegt keine operationale Definition vor.

5. Bei quantitativen Merkmalsausprägungen ist der Begriff **Variable** üblich. Man kann dann mehrere Merkmalsausprägungen zu einer Klasse (**Größenklasse**) zusammenfassen.

Beispiele für Merkmale und Merkmalsausprägungen:

Merkmal:	Merkmalsausprägungen:
Alter (operational definiert als: Anzahl der vollendeten Jahre)	15 Jahre (einzelne Ausprägung) 10 bis unter 20 Jahre (Altersklasse)
Staatsangehörigkeit	deutsch, französisch, englisch usw.
Geschlecht *)	männlich, weiblich

*) ein dichotomes Merkmal, d.h. mit zwei Ausprägungen

Arten von Merkmalen:

a) Nach dem Informationsgehalt der Merkmalsausprägung wird von einigen Autoren unterschieden: qualitative, komparative (intensitätsmäßige) und quantitative Merkmale (vgl. Bem. 3 zu Übersicht 1.3).

b) Wenn bei mehreren Einheiten nicht die Summe der Merkmalswerte, sondern nur ein durchschnittlicher Merkmalswert sinnvoll interpretierbar ist, spricht man von einem **intensiven** Merkmal (z.B. Intelligenz), andernfalls von einem **extensiven** Merkmal (z.B. Einkommen).

c) Nach der Art der Messung kann man **manifeste** (direkt beobachtbare) und **latente Merkmale** unterscheiden. Letztere werden indirekt gemessen, bzw. konstruiert. In diesem Sinne schließt man von (manifesten) Meinungsäußerungen (opinions) auf latente Einstellungen (attitudes) oder von der (manifesten) Fähigkeit bestimmte Aufgaben zu lösen und Fragen zu beantworten auf die "dahinterstehende" "Intelligenz" (als ein latentes Konstrukt) usw.

d) Hinsichtlich der Anzahl möglicher Merkmalsausprägungen kann man bei quantitativen Merkmalen (Variablen) zwischen diskreten und stetigen Merkmalen unterscheiden (vgl. Def. 1.3).

Def. 1.3: diskret und stetig

Eine Variable X mit den Ausprägungen $x_1,x_2,...,x_m$ heißt diskret, wenn X nur endlich viele oder abzählbar unendlich viele reelle Werte x_j annehmen kann, und in jedem endlichen Intervall $a < x < b$ der reellen Zahlengeraden nur endlich viele Werte liegen können. Gilt entsprechend "überabzählbar unendlich viele Werte", so liegt eine stetige (kontinuierliche) Variable vor.

Bemerkungen zur Def. 1.3:

1. Diskret sind alle Merkmale, denen ein Zählvorgang zugrundeliegt; bei stetigen Merkmalen (Variablen) ist es ein Meßvorgang, der beliebig genau ist bzw. (theoretisch) beliebig genau sein könnte.

2. In der Deskriptiven Statistik treten i.d.R. nicht stetige Merkmale als solche auf, sondern nur in Form klassierter Daten (Kap. 3), d.h. aufeinanderfolgende Ausprägungen des Merkmals werden im Zuge der Erhebung oder Aufbereitung der Daten in Größenklassen (Intervallen) zusammengefaßt. Diese Klassenbildung (Klassierung) ist natürlich auch bei diskreten Daten durchführbar und üblich. Nach Klassierung kann ein stetiges Merkmal wie ein diskretes Merkmal behandelt werden.

3. Der methodisch entscheidende Unterschied ist: die Realisationsmöglichkeiten einer diskreten Variable sind isolierte Zahlen (nicht notwendig ganze Zahlen), bei einer stetigen Variable dagegen Intervalle (evtl. infinitesimal kleine Intervalle).

Beispiel 1.1:

Die Pizzeria P (des Eigentümers P) hat zwei Lokale (L_1 und L_2), bei denen man Mittag- und Abendessen (M,A) einnehmen kann, wobei es jedoch jeweils nur die folgenden Gerichte gibt: Pizza, Spaghetti, Ravioli und Canneloni. Es ergab sich, dass von den 4764 Gästen der Pizzeria insgesamt 5000 Gerichte im letzten Monat (April) wie folgt bestellt wurden:

	L_1		L_2		
	M	A	M	A	insgesamt
Pizza	400	600	600	800	2400
Sonstige	700	1100	400	400	2600
Summe	1100	1700	1000	1200	5000

1. Wieviele Merkmale werden in dieser Statistik dargestellt, wie heißen sie und welche Merkmalsausprägungen werden in der Tabelle dargestellt?

2. Was (Masse, Einheit, Merkmal usw.) ist im Falle dieser Statistik

   a) die in der Statistik mitgezählte Pizza, die Herr Schulze am 16.April zum Abendessen im Lokal L_1 gegessen hat?
   b) die Angabe "Pizza"?
   c) die Angabe des Eigentümers P der Pizzeria?
   d) die Zahl 5000?
   e) das Lokal L_2?
   f) die insgesamt 1700 + 1200 = 2900 Gerichte, die abends ausgegeben wurden?
   g) die insgesamt 2800 Gerichte im Lokal L_1?
   h) der Monat April?

Lösung 1.1:

zu 1:

Es werden drei qualitative (artmäßige, kategoriale) Merkmale mit jeweils zwei Merkmalsausprägungen (dichotome Merkmale) dargestellt, nämlich

Merkmal	Merkmalsausprägungen
Art des Gerichts	Pizza, Sonstige*)
Art des Lokals**)	Lokal L_1 , Lokal L_2
Zeit der Mahlzeit	Mittagessen (M), Abendessen (A)

*) nur zwei Ausprägungen in der Tabelle aber bei der Erhebung Unterscheidung nach den oben genannten vier Gerichten,

**) gemeint ist quasi die Zweigstelle (ein Betrieb) des Unternehmens P.

zu 2:

a) das ist eine Einheit (eine der 5000 in der Masse gezählten Einheiten, d.h. einzelnen Gerichte). Ferner sind jeweils

c und h) Identifikationsmerkmale (zur Abgrenzung der Masse),

b und e) Merkmalsausprägungen,

f und g) Teilgesamtheiten (Teilmassen), 2900 bzw. 2800 sind die jeweiligen Umfänge dieser Teilgesamtheiten und d) ist der Umfang der Masse (Gesamtmasse).

3. Messen, Skalen

a) Messung

Es gilt im folgenden den Begriff der Messung zu definieren. Hierfür und für die damit zusammenhängende Unterscheidung verschiedener Skalenarten ist aus der Schulmathematik der Begriff der Relation vorauszusetzen. Von den Relationen, die auf die Elemente m_1 , m_2 ,..., m_n einer Menge M definiert sein können interessieren vor allem zwei zweistellige (binäre) Relationen für jeweils zwei der Elemente von M, etwa für m_1 und m_2, nämlich:

1. die **Äquivalenzrelation** (=): Sie ist symmetrisch, d.h. aus $m_1=m_2$ folgt $m_2=m_1$, reflexiv ($m_1=m_1$) und transitiv (d.h. aus $m_1=m_2$ und $m_2=m_3$ folgt $m_1=m_3$) und
2. die **Ordnungsrelation** (>,<): Sie ist dagegen asymmetrisch, irreflexiv und transitiv.

Def. 1.4: Messung

Mit den folgenden zwei vorangestellten Definitionen kann man den Begriff der Messung definieren.

1. Es sei A eine Menge von empirischen Objekten und es seien $R_1,R_2,...R_s$ auf A definierte Relationen. Dann heißt die Menge
   A = $(A,R_1,R_2,...R_s)$ ein empirisches relationales System oder ein **empirisches Relativ**.
2. Ist X eine Menge von Zahlen oder Vektoren und sind $S_1,S_2,...S_m$ auf X definierte Relationen, dann heißt **X** = $(X,S_1,S_2,...,S_m)$ ein **numerisches Relativ** (oder numerisches relationales System). **X** ist eine Zahlenmenge (z.B. die Menge der natürlichen Zahlen oder die Menge der reellen Zahlen).

Damit erhält man die folgende **Definition**:

Unter einer **Messung** versteht man die Abbildung eines empirischen Relativs in ein numerisches Relativ, d.h. die Zuordnung von Zahlen zu Merkmalsausprägungen, so dass die für die Merkmalsausprägungen der empirischen Objekte geltenden Relationen auch für die hierfür verwendeten Zahlen gelten.

Bemerkungen zur Def. 1.4:

1. Empirische Objekte können Personen, Unternehmungen, Maschinen usw., kurz alle Arten von statistischen Einheiten sein. Diese stehen hinsichtlich einer oder mehrerer Eigenschaften in Beziehung zueinander, z.B. in Abb. 1.1 die Personen P_1 (links) und P_2 (rechts) hinsichtlich des Merkmals Körpergröße oder Haarfarbe. Das Vorhandensein eines empirischen Relativs, d.h. einer Menge empirischer Objekte, die in bezug auf bestimmte Eigenschaften in beobachtbaren Relationen zueinander stehen, ist Voraussetzung für eine Messung. Man beachte: Gegenstand der Messung sind nicht Objekte (also z.B. Personen) sondern immer nur Eigenschaften (Personen interessieren nicht in ihrer Einzigartigkeit, Gesamtheit, Individualität, sondern immer nur als Träger von bestimmten, genau definierten Eigenschaften).

Abb. 1.1: Empirisches Relativ

2. Abb. 1.1 stellt ein Beispiel für ein empirisches Relativ dar, denn:
   - die Menge A enthält die Personen P_1 und P_2 (linke und rechte Person)
   - die Personen P_1 und P_2 stehen hinsichtlich der Eigenschaft "Körpergröße", "Alter", "Haarfarbe" usw. in Beziehung zueinander und diese Beziehungen sind zweifelsfrei beobachtbar.

3. Man kann an diesem Beispiel leicht demonstrieren, was "Messung" bedeutet:
   - Es soll zunächst die Körpergröße K betrachtet werden: Für K soll die Ordnungsrelation "ist größer als" definiert werden. Empirisch (in Abb. 1.1) läßt sich die Relation "P_1 ist größer als P_2" beobachten. Eine sinnvolle Messung wäre dann $k_1 = 250$ (k_1 ist der Meßwert des Objekts P_1 auf der Skala K für das Merkmal K) und $k_2 = 35$ da $250 > 35$ (es soll noch nicht an eine Messung mit der Maßeinheit cm gedacht werden). Ebenso sinnvoll wäre aber auch $k_1 = 17{,}2$ und $k_2 = -3{,}25$, weil ja 17,2 ebenfalls größer ist als -3,25. Jedes Zahlenpaar, bei welchen k_1 größer ist als k_2 ist gleichermaßen "sinnvoll", weil hier bei der Körpergröße nur eine Ordnungsrelation (**Ordinalskala**) betrachtet werden soll.
   - Betrachten wir sodann die Haarfarbe H der beiden Personen von Abb. 1.1: Normalerweise kann man nicht mehr als die "qualitative" Unterscheidung zwi-

schen blond und dunkelhaarig treffen. Das bedeutet, dass für das Merkmal Haarfarbe nur die Äquivalenzrelation festgestellt werden kann: die Haarfarbe von P_1 ist anders (nicht z.B. besser oder mehr) als die von P_2 (es liegt "nur" eine **Nominalskala** vor). Folglich wären z.B. die Zahlen (Codierungen) $h_1 = 3$ und $h_2 = 8$ oder auch $h_1 = 7{,}3$ und $h_2 = -1{,}8$ gleich sinnvolle Messungen des Merkmals Haarfarbe.

4. Man beachte, dass in Def. 1.4 über die Anzahl n von empirischen Objekten keine Aussage gemacht wird. Für die Anzahl z_i der Stellen der Relation R_i muss natürlich gelten $z_i < n$. Im Beispiel wurde unterschieden:
   R_1: "ist größer als" ist eine zweistellige Relation bezüglich der Eigenschaft Körpergröße
   R_2: Haarfarbe: beispielsweise "ist blond" ist eine einstellige Relation bezüglich der Eigenschaft Haarfarbe.

5. Der anschauliche Begriff "Messen" oder "**messbar**" (im Alltagssprachgebrauch) ist enger als der hier verwendete. Man versteht darunter das wiederholte Anlegen eines Maßstabes (z.B. ein Meter). In diesem Sinne ist z.B. die Länge einer Wand 5m, weil der Metermaßstab fünfmal angelegt (addiert) werden kann. Diese Art Messung setzt eine metrische Skala voraus.

6. Eine **Dimension** ist eine durch eine Zahl zu beschreibende Eigenschaft (z.B. Länge), wobei dieser Zahl meist eine Maßeinheit (z.B. Meter) beigegeben ist. Eindimensionalität ist durch Transitivität gekennzeichnet. Gilt für die Eigenschaft X bei drei Objekten $x_1 < x_2$ und $x_2 < x_3$ aber $x_3 < x_1$ so ist dies ein Hinweis darauf, dass die Eigenschaft X mehrdimensional ist, also mehrere Zahlen und nicht nur eine Zahl zur Charakterisierung eines Objekts hinsichtlich X erforderlich sind.

b) Skalenarten

Mit einer **Skala** wird die Zahlenmenge (das numerische Relativ) definiert, die (das) zur Bezeichnung von Merkmalsausprägungen verwendet werden kann. Für die Zwecke der Statistik ist die Unterscheidung von fünf Skalentypen gem. Übers. 1.2 ausreichend.

Anmerkungen zu Übers. 1.2:

1. Kennzeichnend für den Skalentyp ist:

a) welche Rechenoperationen mit der Skala "sinnvoll" und "sinnlos" sind

Übersicht 1.2: Skalentypologie

Skala (Name)	definiert ist zusätzlich	zulässige Transformation	Beispiel	Mittelwerte[a)]
Nominalskala	Äquivalenzrelation	ein-eindeutige	Postleitzahlen, Steuerklasse	$\bar{x}_M$
Ordinalskala	Ordnungsrelation	streng monoton steigend	Windstärke (Beaufort)	$\bar{x}_M$, $\tilde{x}_{0,5}$
Intervallskala	Maßeinheit u. Nullpunkt [b)]	Linear $y_v = a + bx_v$	Temperatur in Grad Celsius	$\bar{x}_M$, $\tilde{x}_{0,5}$, $\bar{x}$
Ratio- bzw. Verhältnisskala	natürl. Nullpunkt (Maßeinheit noch willkürlich)	proportional $y_v = bx_v$ $(a = 0)$[c)]	Temperatur in Kelvin, Körpergröße	alle Mittelwerte
Absolutskala	auch natürliche Maßeinheit	identisch $y_v = x_v$ $(b = 1)$	Häufigkeiten	alle Mittelwerte

a) Zur. Notation vgl. Bem. 5.

b) beides (Nullpunkt und Maßeinheit) noch willkürlich.

c) d.h. der Nullpunkt ist nicht mehr willkürlich (er kann nicht durch $a \neq 0$ verschoben werden), wohl aber die Maßeinheit (weshalb $b \neq 1$ sein kann). Man kann sinnvoll Verhältnisse x_1 / x_2 (Proportionen, engl. "ratios") bilden (denn $y_1 / y_2 = x_1 / x_2$).

b) welche Transformationen "zulässig" (im noch definierten Sinne) sind.

Zwei Beispiele zeigen, wie man mit Hinweis auf a) und b) darlegen kann welcher Skalentyp vorliegt:

- Postleitzahlen sind Zahlenwerte einer Nominalskala (bei der Zahlen nur stellvertretend für Namen sind):

  a) es macht keinen Sinn zu sagen, der Mittelwert aus den Städten Oberhausen (Postleitzahl 42) und Münster (44) sei Essen (43) und
  b) die Zahlen -7 , 2432 und 489 könnten den gleichen Zweck erfüllen wie die Postleitzahlen 42,43 und 44;

- Zensuren sind Messungen auf einer Ordinalskala:

  a) die Abstände zwischen den Zensuren (Noten) 1,2,3,4 und 5 sind nicht (oder nicht notwendig) gleich und
  b) die durch eine monotone (ordnungserhaltende) Transformation gewonnenen Zahlen -2, 10, 13, 28 und 64 könnten den gleichen Zweck erfüllen.

2. Metrik:

Die ersten beiden Skalen der Übers. 1.2 (Nominal- und Ordinalskala) heißen auch "topologische Skalen", die letzten drei "metrische Skalen". Eine "Metrik" auf die Menge der Zahlen x_1, x_2, x_3 definieren (was bei der Nominal- und Ordinalskala noch nicht erfolgt) heißt ein Maß $d(x_i,x_j)$ [d: anschaulich interpretiert eine Distanz] festzulegen, für das gilt:

(1)	$d(x_i,x_i) = 0$ (i,j,k = 1,2,...)
(2)	$d(x_i,x_j) = d(x_j,x_i)$ und
(3)	$d(x_i,x_j) + d(x_j,x_k) \geq d(x_i,x_k)$ (Dreiecksungleichung).

Größenmäßig ordnen lassen sich nicht nur metrisch-, sondern auch ordinal skalierte Merkmale. Bei letzterer sind aber die Abstände nicht definiert. Nominalskalierte Merkmale lassen sich nicht nach der Größe, sondern nur nach dem Sachzusammenhang ordnen, was in Klassifikationen oder Systematiken erfolgt (z.B. Berufssystematik, Klassifikation der Wirtschaftszweige usw.).

3. Skalentyp und Merkmalsarten. Man kann unterscheiden:

- klassifikatorische (qualitative) Merkmale oder "Attribute" mit einer abzählbaren Menge nur artmäßig unterschiedener Ausprägungen bei einer Nominalskala,
- komparative (intensitätsmäßig abgestufte) Merkmale bei einer Ordinalskala und
- metrische (quantitative) Merkmale im Falle einer metrischen Skala.

Der Begriff qualitativ wird auch für die ersten beiden Merkmalstypen verwendet. Klassifikatorische Merkmale verlangen z.T. Methoden unterschiedlicher Art (z.B. Verhältniszahlen statt Mittelwerte oder in der Induktiven Statistik spielt das Begriffspaar "homograd"/"heterograd" eine wichtige Rolle). Ferner sind zwei spezielle Typen klassifikatorischer Merkmale von besonderem Interesse:

- **Dichotome Merkmale** (Alternativ-, binäre Merkmale): sie haben zwei Ausprägungen, die mit 0 und 1 codiert werden können. Im Zusammenhang mit der Assoziationsmessung (Kap. 7) oder mit Verhältniszahlen (Kap. 9) wird hierauf besonders eingegangen.
- **Häufbare Merkmale**: ein Merkmal ist **häufbar**, wenn eine Einheit gleichzeitig mehrere Ausprägungen realisieren kann (z.B. Beruf, Beschäftigung, Studienfach); normalerweise sind Merkmale **nicht-häufbar** (z.B. Alter, Geschlecht).

4. Bemerkungen zu den "zulässigen Transformationen" in Übers. 1.2:

Unter einer Transformation der Skala X in die Skala Y versteht man eine Funktion f, die jedem Wert von x_v einen (und nur einen) Wert y_v zuordnet. "Zulässig" heißt, dass die Skala Y den gleichen Informationsgehalt hat, wie die Skala X. Man kann verschiedene Arten von (zunehmend spezieller werdenden) Transformationen unterscheiden:

- Im einfachsten Fall einer **ein-eindeutigen (bijektive) Transformation** bedeutet X in Y zu transformieren nur, dass zwei verschiedenen Werten x_1 und x_2 zwei verschiedene Werte y_1 und y_2 zugeordnet werden. Eine speziellere Transformation ist die

- **streng monoton steigende Transformation:** ist $x_1 < x_2$, so muss auch $y_1 < y_2$ sein (bei "monoton" wäre auch zugelassen y_1 ú y_2; der Zusatz "steigend" ist notwendig, wenn die Reihenfolge erhalten bleiben soll, denn "fallend" würde bei $x_1 < x_2$ zu $y_1 > y_2$ führen).

- Die noch speziellere **lineare Transformation** (Lineartransformation) liegt vor, wenn gilt:

  (1.1) $y_v = a+bx_v$

  mit a und b als reellen Konstanten (das wird anhand der "Umrechnung" einer Temperaturmessung von "Grad Celsius" [x_v] in "Grad Fahrenheit" [y_v] in Beispiel 1.2 demonstriert).

- Eine **proportionale Transformation** liegt dann vor, wenn man in Gl. 1.1 a = 0 setzt (was bedeuetet, dass der Nullpunkt festliegt und nicht verändert werden darf); ein Beispiel hierfür ist die Umrechnung von Währungseinheiten, etwa von DM in US-$.

- Wird noch spezieller b = 1 gefordert, so dass $y_v = x_v$, so liegt eine **identische Transformation** vor, d.h. die Skala x kann überhaupt nicht mehr verändert werden, insbesondere ist wegen b = 1 auch die Maßeinheit fest (man würde im Alltagssprachgebrauch wohl gar nicht mehr von einer "Transformation" sprechen).

Wie man hieran sieht, stellen die Skalen der Übers. 1.2 eine **Hierarchie** dar hinsichtlich der zulässigen Transformationen und Rechenoperationen: speziellere Transformationen heißt mehr Informationsgehalt der Skala und damit mehr Möglichkeiten der sinnvollen rechnerischen Verarbeitung der Daten.

5. Bemerkungen zu den Mittelwerten in Übers. 1.2 und zur praktischen Bedeutung der Skalentypologie:

Auf die in Spalte 5 der Übers. 1.2 angegebenen zulässigen Mittelwerte wird in Kap. 4 eingegangen. Zu den Abkürzungen:

$\bar{x}_M$ = Modus
$\tilde{x}_{0,5}$ = Median
$\bar{x}$ = arithmetisches Mittel (Durchschnitt)

Es ist gerade am Beispiel der Mittelwerte die praktische Bedeutung der Unterscheidung der Skalenarten leicht einzusehen:

- Bei einer Nominalskala ist bereits der Modus sinnvoll anzuwenden, bei einer Ordinalskala Modus und Median, bei einer Intervallskala Modus, Median und arithmetisches Mittel usw.
- Ferner gilt:

Was bei gegebenen Daten jeweils sinnvoll zu berechnen ist, ist abhängig von a) der Qualität der Daten, insbesondere der Skalenart (je höher die Skalenart, desto mehr Mittelwerte sind sinnvoll zu berechnen) und b) der Fragestellung (Beispiel: Geschwindigkeiten [dort ist nicht das arithmetische, sondern das harmonische Mittel anzuwenden]).

6. Nicht eingegangen werden kann hier auf die **Skalierungsverfahren**. Hierbei geht es um Methoden, mit denen man in der Praxis komplexe und subjektive Eigenschaften messen kann, wie z.B. die Intensität von Sinneswahrnehmungen oder Gefühlen, die Messung der Intelligenz, des Sozialstatus, der Lebensqualität usw., die v.a. für Anwendungen der Statistik in der Psychologie und Soziologie von Interesse sind.

Beispiel 1.2:

Man zeige, dass eine Temperaturmessung in "Grad Celsius" auf einer Intervallskala, eine Messung in "Kelvin" dagegen auf einer Ratioskala erfolgt.

Lösung 1.2:

Man kann den Nachweis auf zwei Arten führen, indem man zeigt:

a) welche Rechenoperationen bei Messungen in "Grad Celsius" [^oC] bzw. in "Kelvin" [K] sinnvoll sind und
b) welche Transformationen zulässig sind.

zu a)
Man kann bei ^oC (anders als bei K) nicht sagen: 20^oC ist "doppelt so warm" wie 10^oC, wohl aber, dass der "Abstand" zwischen 10^oC und 20^oC genauso groß ist wie derjenige zwischen 20^oC und 30^oC.

zu b)
Die Temperaturmessung von "Grad Celsius" [x_v] ist von der gleichen Qualität, wie die Messung in "Grad Fahrenheit" [y_v]. Die Umrechnung erfolgt nach der Formel $y_v = 32 + 1{,}8x_v$, die eine Lineartransformation gem. Gl. 1.1 darstellt mit a = 32 und b = 9/5; so entsprechen sich beispielsweise 0^oC und 32^oF. Es gelten die folgenden Umrechnungen:

^oC	0	10	20	30	40
^oF	32	50	68	86	104

Das erklärt auch das Problem unter a) denn

- $20^oC/10^oC = 2$ aber $68^oF/50^oF = 1{,}36$, andererseits gilt aber
- die Abstände zwischen 10^oC, 20^oC, 30^oC usw. sind jeweils 10^oC und die entsprechenden Abstände zwischen 50^oF, 68^oF, 86^oF usw. sind ebenfalls gleich, nämlich 18^oF, wobei $18 = \frac{9}{5} \cdot 10$

Die Größe a bewirkt eine Verschiebung des Nullpunkts (Translation), der bei ^oC und ^oF willkürlich ist, bei der absoluten Temperaturmessung in Kelvin dagegen unveränderlich -273^oC ist und die Größe b bewirkt eine Streckung (wenn b > 1) oder Stauchung (b < 1) des Maßstabs.

Das Beispiel zeigt übrigens auch

- dass es nicht "in der Natur der Sache" liegt, mit welcher Skala ein Merkmal gemessen wird, sondern dass dies allein vom Stand der Messtechnik abhängt und
- dass durch Transformationen das Skalenniveau nicht gesteigert werden kann (die Intervallskala ^oC wird durch die Transformation in ^oF nicht zu einer Skala anderen Typs).

Kapitel 2: Daten, Maßzahlen und Axiomatik

1. Daten

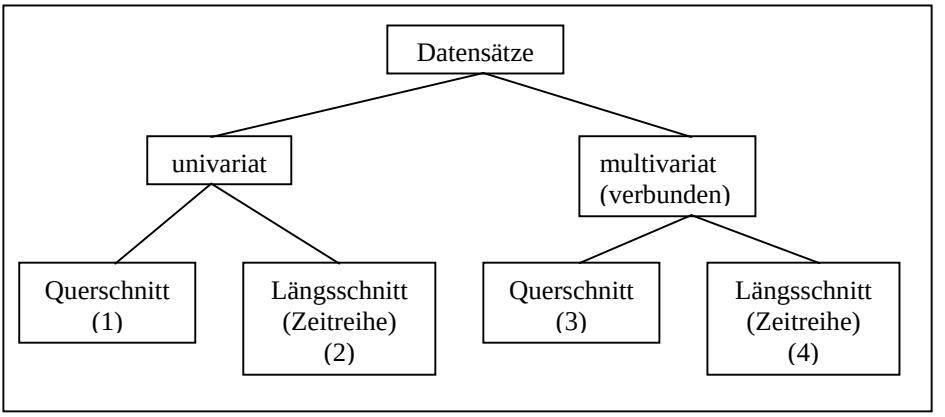
Def. 2.1: Daten, Datensatz

Statistische Daten sind der Ausgangspunkt weitergehender statistischer Auswertungen. Es sind Zahlenangaben über Merkmalsausprägungen, die an Einheiten beobachtet bzw. "gemessen" worden sind. Alle sachlich zusammengehörigen und einer statistischen Auswertung zugrundezulegenden Daten bilden einen Datensatz.

Übersicht 2.1 zeigt, welche Arten von Datensätzen sich unterscheiden lassen. Danach bestimmt sich auch die Art der sinnvoll zu berechnenden beschreibenden Maßzahlen, bzw. allgemeiner, der anzuwendenden statistischen Methoden. Die hier benutzte Terminologie ist leider nicht einheitlich und auch (insbesondere, was die Unterscheidung zwischen Quer- und Längsschnitt betrifft) nicht unproblematisch[1].

[1] So ist z.B. besonders der Begiff "Längsschnittsanalyse" im Rahmen der Bevölkerungsstatistik für eine spezielle Art, zeitlich geordnete Sachverhalte zu beschreiben, reserviert und deshalb auch nicht gleichbedeutend mit Zeitreihenanalyse. Es ist andererseits z.B. in der Ökonometrie üblich, von Längsschnitt in dem oben gemeinten Sinne zu sprechen. Hinzukommt, dass für das, was hier "Querschnitt" genannt wurde überhaupt keine allgemein anerkannte Bezeichnung existiert.

Übersicht 2.1. Arten von Datensätzen



Die in Übers. 2.1 getroffenen Unterscheidungen betreffen die Fragen, ob

- an einer Einheit ein Merkmal (univariat) oder mehrere Merkmale (multivariat) beobachtet worden sind; je nachdem wird eine Einheit v im Datensatz repräsentiert durch einen Skalar x_i (Ausprägung des Merkmals X, die bei der Einheit v beobachtet wurde) oder durch ein m-Tupel, den Zeilenvektor $[x_{1v}\ x_{2v}\ ...x_{mv}]$, womit die Ausprägungen gemeint sind, die bei der Einheit (z.B. Person) v hinsichtlich der Merkmale $X_1, X_2,...,X_m$ beobachtet worden sind;

- die Zahlenangaben, die im Datensatz zusammengefaßt sind, in einer zeitlichen Reihenfolge geordnet (datiert) sind, oder ob die Reihenfolge der Daten keine Rolle spielt.

Kennzeichnend für Daten, die als Zeitreihe vorliegen, ist es, dass die zeitliche Reihenfolge der Daten wesentlich ist, während es für "Querschnitts"-Daten (besser: undatierte Daten) i.d.R. unerheblich ist, ob sie ungeordnet vorliegen oder in irgendeiner Weise geordnet sind. Solche Daten (insbesondere Daten des Typs 1 in Übers. 2.1) können evtl. in einem anderen Sinne geordnet sein: Neben der zeitlichen Reihenfolge der Daten kann die Anordnung der Größe nach von Belang sein. Man spricht dann von einer (der Größe nach) geordneten Reihe.

Daten vom Typ 1, die in den folgenden Kapiteln (Kap. 3 - 6) behandelt werden, können in vier Formen vorliegen, nämlich als:

1. Einzelwerte (Einzelbeobachtungen)
   a. ungeordnete Reihe (Urliste) $x_1, x_2,...,x_n$
   b. (der Größe nach) geordnete Reihe $x_{(1)}, x_{(2)},...,x_{(n)}$
2. Häufigkeitsverteilung
3. klassierte Daten (klassierte Verteilung).

Im Fall 1a liegen die Daten in der Reihenfolge ihrer Erhebung vor und bei 1b gilt $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Die Zahl (j) in dieser Zahlenfolge ist die Rangzahl (der Rang) der Beobachtung $x_{(j)}$ [vgl. Def. 7.17]. Sie bedeutet, dass (mindestens) j Beobachtungen kleiner oder gleich $x_{(j)}$ sind.

Für statistische Berechnungen ist es vor allem entscheidend, ob die Daten als Einzelwerte vorliegen oder nicht. Man unterscheidet danach zwischen einer gewogenen und einer ungewogenen Berechnung einer Maßzahl (vgl. Def. 2.2):

ungewogen	Daten liegen als Einzelwerte vor,
gewogen (gewichtet)	Daten liegen als Häufigkeitsverteilung (Gewichtung mit Häufigkeiten) oder klassiert vor.

Bevor diese Unterscheidungen im Kap. 3 näher erläutert werden, mag das folgende Beispiel, in dem ein und der gleiche Datensatz in allen vier Formen dargestellt wird, ausreichen, um die hier eingeführten Begriffe zu erläutern.

Beispiel 2.1:

Die folgenden Daten sind gegeben als

1. ungeordnete Reihe: 1,1,2,3,8,2,4,3,1,2,4,1,4,2,3,9,1,1,3,9
2. geordnete Reihe: 1,1,1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,8,9,9
3. Häufigkeitsverteilung (nach Merkmalsausprägungen sortiert):

x_i	n_i
1	6
2	4
3	4
4	3
8	1
9	2
$\sum$	20

   Die Größen x_i sind die Merkmalsausprägungen, die n_i sind die absoluten Häufigkeiten, deren Summe n = 20 ist.

4. klassierte Daten (d.h. mit Größenklassen, die als halboffene, nicht notwendig gleich breite Intervalle abgegrenzt sind)

von...bis unter...	absolute Häufigkeit
0 - 3	10
3 - 6	7
6 - 9	1
9 - 12	2
$\sum$	20

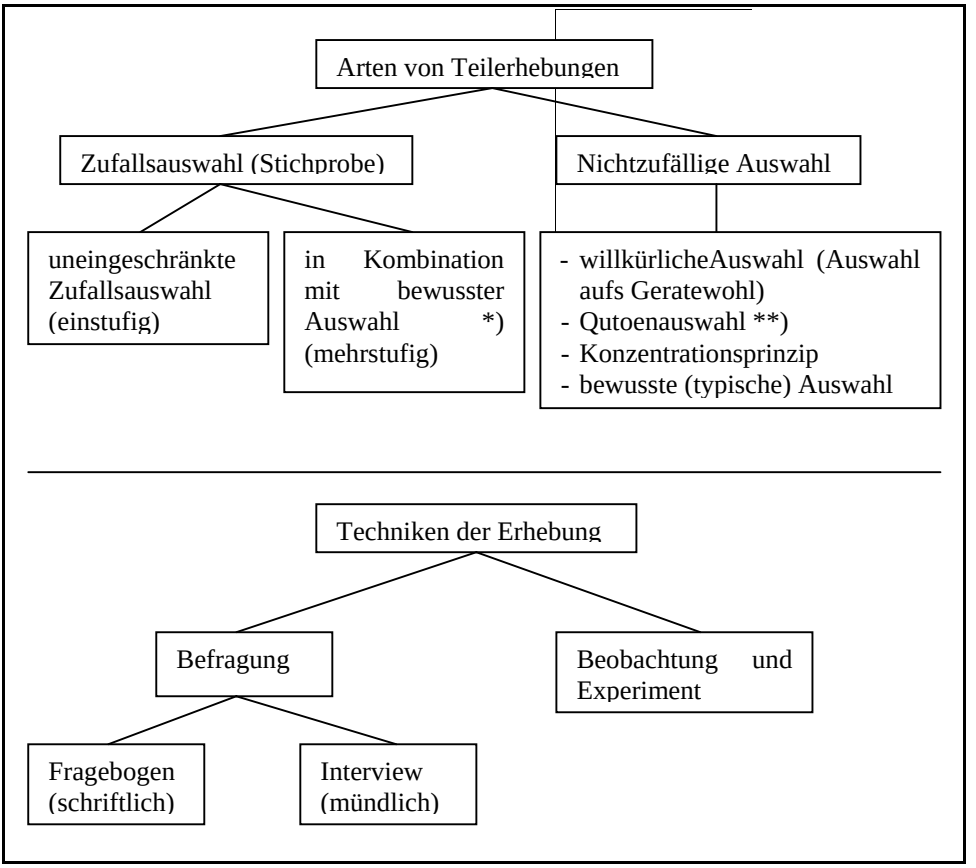
2. Methoden der Datengewinnung

In Übersicht 2.2 sind einige Begriffe zusammengestellt, die zum Verständnis statistischer Methoden bekannt sein sollten und hier nur kurz erläutert werden.

Man unterscheidet meist drei Stufen statistischer Arbeit: die **Erhebung**, **Aufbereitung** (in Gestalt von Tabellen und Graphiken) und **Analyse** (Auswertung) statistischer Daten. Aufbereitung ist die geeignete Darstellung des kompletten Datensatzes, während es bei der Auswertung um eine Interpretation der Daten, meist durch Berechnung zusammenfassender Kennzahlen (Maßzahlen) geht (vgl.Abschn.3).

Unter einer **Erhebung** versteht man jede systematische Gewinnung von statistischen Daten. Dies kann eine speziell für statistische Zwecke veranstaltete Datenbeschaffung sein (Primärerhebung oder Primärstatistik) oder auch durch Rückgriff auf ursprünglich für andere Zwecke angelegte Unterlagen erfolgen (Sekundärstatistik).

Übersicht 2.2: Methoden der Datengewinnung



*) Geschichtete Stichprobe, Klumpenauswahl (z.B. area sample) usw.

**) "Repräsentativer Bevölkerungsquerschnitt" (übliches Verfahren der Markt-, Meinungs- und Umfrageforschung)

Zu unterscheiden sind ferner (vgl. Übers. 2.2):

- Total- und Teilerhebungen
- verschiedene Techniken der Erhebung.

Eine **Teilerhebung** liegt vor, wenn nur n von N Einheiten der Grundgesamtheit (Masse) erhoben werden ($n < N$).

Häufig sind schon auf der Basis einer solchen Teilmenge des Umfangs n der Grundgesamtheit ausreichend genaue und sichere statistische Aussagen möglich. Eine Stichprobe ist oft nicht nur kostengünstiger als eine Totalerhebung, sondern auch die einzig vertretbare Erhebung (z.B. bei der Qualitätskontrolle).

Für die Deskriptive Statistik ist es ohne Belang, ob es sich um Stichproben- oder um Grundgesamtheits-Daten handelt. Die Theorie von Teilerhebungen, insbesondere von Stichproben, ist Gegenstand der Induktiven Statistik.
An dieser Stelle sollte jedoch bereits festgehalten werden:

- nicht jede Teilerhebung ist auch eine Zufallsauswahl (Stichprobe) und
- eine zufällige Auswahl ist streng von einer willkürlichen Auswahl zu unterscheiden.

In Übersicht 2.2 sind einige Arten von Teilerhebungen genannt, auf die hier nicht näher eingegangen wird. Einzelheiten hierzu sind - wie gesagt - im Rahmen der Induktiven Statistik zu behandeln.

Jede Teilerhebung ist mit einem Auswahlfehler verbunden, weil nicht alle Einheiten der Grundgesamtheit erfasst werden. Nur bei einer Stichprobe ist dieser Fehler ein Zufallsfehler, der mit der Wahrscheinlichkeitsrechnung abgeschätzt werden kann. Das gilt, weil das Auswahlverfahren die Zufallsauswahl ist, d.h. weil jede Einheit der Grundgesamtheit eine vor Ziehung der Stichprobe bekannte Wahrscheinlichkeit hat (bei Ziehung von n Elementen aus einer Urne von N Elementen "mit Zurücklegen" ist z.B. die Auswahlwahrscheinlichkeit bei uneingeschränkter Zufallsauswahl n/N). Willkürliche Auswahl bedeutet im Unterschied zur zufälligen Auswahl, dass keine Kenntnisse über Auswahlwahrscheinlichkeiten vorliegen. Eine Stichprobe zu ziehen kann auf erhebliche praktische Schwierigkeiten stoßen, so dass Ersatzverfahren oft angewendet werden, wie z.B. das in Übers. 2.2 genannte Quotenverfahren.
Durch Ausnutzung von Kenntnissen über die Grundgesamtheit kann eine Stichprobe oft auf effizientere und rationellere Art gezogen werden, als im beschriebenen Fall einer Ziehung aus einer Urne (uneingeschränkte Zufallsauswahl). Das kann z.B. durch mehrstufige Auswahl (z.B. zuerst Auswahl von Gemeinden, dann von Haushalten in den Gemeinden) geschehen. Geschichtete Stichproben und Klumpenstichproben kann man als Spezialfälle einer zweistufigen Auswahl betrachten.

In entwickelten Ländern werden die meisten Erhebungen als Befragungen der Erhebungseinheiten durchgeführt. Schriftliche und mündliche (Interview) Erhebungen haben jeweils ihre spezifischen Vor- und Nachteile. Beobachtungen und Experimente sind nicht auf naturwissenschaftliche Anwendungen der Statistik beschränkt. Eine Verkehrszählung durch Notieren der vorbeifahrenden Fahrzeuge ohne die Fahrer anzuhalten und zu befragen oder auch eine Statistik durch systematische Aufzeichnungen über Zeitungsannoncen (z.B. Stellenangebote) stellt z.B. eine Beobachtung dar.

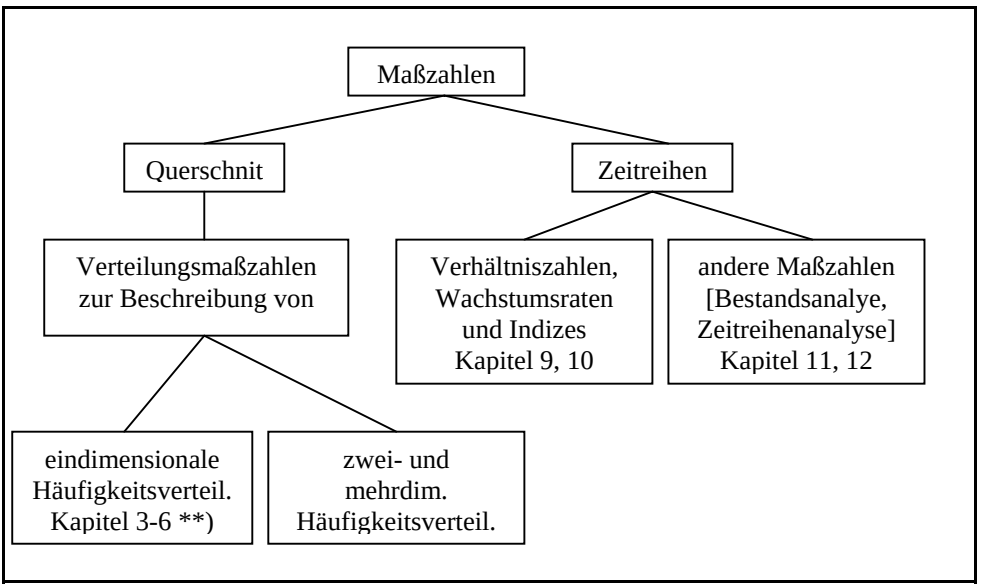
3. Maßzahlen, Eigenschaften und Axiome

a) Maßzahlen

Deskriptive Statistik ist im wesentlichen die Lehre von der Konstruktion von Maßzahlen (Kennzahlen), die einer zusammenfassenden Beschrei-

bung von Daten durch eine Zahl dienen. Das wohl bekannteste Beispiel einer statistischen Maßzahl ist ein Mittelwert.
Ziel dieser Beschreibung durch Maßzahlen ist die summarische Charakterisierung und der Vergleich von Datensätzen. Von dieser summarischen, d.h. die Information verdichtenden Beschreibung von Daten (z.B. einer Häufigkeitsverteilung) durch eine Zahl ist die vollständige Beschreibung durch Tabellen und Graphiken zu unterscheiden.
In Übersicht 2.3 werden die in der Deskriptiven Statistik gebräuchlichen Klassen von Maßzahlen unterschieden.

Übersicht 2.3: Arten von Maßzahlen



*) Viele, aber nicht alle Methoden sind auf Zeitreihen bezogen. Bestimmte Verhältniszahlen, wie Gliederungs- und Beziehungszahlen beziehen sich aber auf Querschnittsdaten.

**) Die Berechnung vieler in den Kap. 3 bis 6 dargestellten Maßzahlen ist nicht auf eindimensionale Häufigkeitsverteilungen beschränkt. Sie werden auch auf andere Arten von Daten angewandt, z.B. zeitliche Mittelwerte.

Maßzahlen können (müssen aber nicht) anschaulich interpretierbar sein. Häufig nehmen sie Zahlenwerte an, die den Merkmalsausprägungen nicht entsprechen (z.B. eine gebrochene Zahl während das Merkmal nur ganzzahlige Werte vorsieht). Eine statistische Masse kann in der Regel nur durch mehrere Maßzahlen hinreichend charakterisiert werden. Angewandt auf empirische Daten (z.B. eine empirische Häufigkeitsverteilung) nehmen Maßzahlen stets endliche Zahlenwerte an (was für analog konstruierte Maßzahlen in der induktiven Statistik nicht gelten muss).

Def. 2.2: Maßzahl

a) Eine Funktion f, die den reellen Beobachtungswerten $x_1,x_2,...,x_n$ des Merkmals (der Variablen) X eine reelle Zahl M zuordnet,

(2.1) $f: \mathbb{R}^n \rightarrow \mathbb{R}, \quad M = f(x_1,x_2,...,x_n)$

heisst (ungewogene) **Maßzahl** (Kennzahl), sofern sie bestimmten Axiomen genügt.

b) Entsprechend ist eine gewogene Maßzahl eine Funktion g, die den reellen Beobachtungswerten $x_1,x_2,...,x_m$ des Merkmals X und den dazu korrespondierenden Gewichten $g_1,g_2,...,g_m$ eine reelle Zahl G zuordnet:

(2.2) $g: \mathbb{R}^{2m} \rightarrow \mathbb{R}, \quad G = g[(x_1,g_1),(x_2,g_2),...,(x_m,g_m)].$

Bemerkungen zu Def. 2.2:

1. Welche Maßzahl im Einzelfall sinnvoll zu berechnen ist, hängt ab von

- dem Aussagezweck: soll z.B. die Streuung oder die Schiefe einer Häufigkeitsverteilung bestimmt werden?
- der "Sachlogik": welcher Mittelwert ist z.B. zur Bestimmung einer durchschnittlichen Wachstumsrate oder einer durchschnittlichen Geschwindigkeit adäquat?
- der Art der Daten (Skalenniveau): bestimmte Maßzahlen verlangen z.B. ein quantitatives Merkmal oder zusätzlich positive Merkmalswerte usw.

Durch "Aussagezwecke" im obigen Sinne wird eine Klasse von Maßzahlen (z.B. Mittelwerte) bestimmt. Die "Sachlogik" soll Maßstäbe zur Auswahl aus einer gegebenen Klasse liefern (z.B. Wahl des geometrischen- oder des harmonischen Mittels).

2. Als Gewichte $g_1,g_2,...,g_m$ im Rahmen einer Gewichtung der Beobachtungswerte $x_1,x_2,...,x_m$ können die (relativen) Häufigkeiten $h_1,h_2,...,h_m$ benutzt werden. Es ist aber auch denkbar, den einzelnen m Werten x_i mit anders definierten m Werten $g_1,g_2,...,g_m$ jeweils ein unterschiedliches "Gewicht" zu verleihen. Dass die Werte $g_1,g_2,...,g_m$ den Beobachtungswerten $x_1,x_2,...,x_m$ jeweils ein unterschiedliches "Gewicht" verleihen, ist unmittelbar einsichtig, wenn sie als Faktoren auftreten, also die Maßzahl so konstruiert ist, dass in ihr die Produkte $x_1g_1,...,x_m\,g_m$ auftreten.

3. Von den Gewichten wird üblicherweise gefordert, dass sie auf 1 normiert sind, d.h. dass gilt $g_1+g_2+...+g_m = 1$. Die Normierung der Gewichte ist zu unterscheiden von der im nächsten Abschnitt behandelten Normierung (des Wertebereichs) einer Maßzahl.

4. Man bezeichnet eine Maßzahl auch als Stichprobenfunktion oder Statistik (statistic), wenn der Datenvektor aus Stichprobenbeobachtungen besteht und sie zu inferenzstatistischen Zwecken (also in der Induktiven Statistik) konstruiert wird. In diesem Fall müssen zur Beurteilung einer Maßzahl auch inferenzstatistische Eigenschaften ("Gütekriterien", wie z.B. Erwartungstreue, Konsistenz usw.) herangezogen werden, die im Rahmen der Deskriptiven Statistik nicht betrachtet werden. Es wird in der Deskriptiven Statistik auch nicht näher spezifiziert, ob die Daten einer endlichen Grundgesamtheit oder einer Teilgesamtheit entnommen sind. Im Vordergrund stehen hier bei der Konstruktion von Maßzahlen deskriptive Aspekte der Datenreduktion. Sie haben aber durchaus eine eigenständige Bedeutung, weil viele wirtschaftliche und soziale Daten aus Grundgesamtheiten oder nicht zufälligen Teilgesamtheiten stammen.

b) Axiome, axiomatische Betrachtung

1. Was sind Axiome?

Axiome sind grundlegende, ohne Beweis anerkannte, bzw. geforderte Aussagen eines Wissenschaftsbereichs, aus denen andere Aussagen abgeleitet werden. In der Deskriptiven Statistik sind mit Axiomen bestimmte formal oder inhaltlich motivierte Eigenschaften von Maßzahlen gemeint, die es erlauben, eine Maßzahl als "sinnvoll" (meaningful) zu akzeptieren, bzw. als "sinnlos" (meaningless) zu verwerfen.

Es soll im folgenden unterschieden werden zwischen Forderungen, die

1. von **allen** Maßzahlen der Deskriptiven Statistik erfüllt werden sollen,
2. solchen, die von **einer** ganzen **Klasse** von Maßzahlen gefordert werden, um jeweils eine Maßzahl als sinnvoll zu bezeichnen,
3. **Eigenschaften,** die eine konkrete Maßzahl einer solchen Klasse von Maßzahlen darüber hinaus hat und die für ihre Interpretation von großer Bedeutung sein können, aber nicht notwendig sind, um die Maßzahl als sinnvoll zu bezeichnen.

Mit Axiomen ist im folgenden der Fall 2 gemeint.

Axiome sind formale Kriterien, die eine Klasse von Maßzahlen insgesamt erfüllt, wodurch sich diese Klasse auch unterscheidet von einer anderen Klasse von Maßzahlen.

Zu 3:

Es ist z.B. die Eigenschaft, zwischen dem kleinsten und den größten Wert zu liegen, fundamental für alle "Mittelwerte". Ein "Mittelwert", der dieses Axiom nicht erfüllt, also größer als der größte oder kleiner als der kleinste Zahlenwert der Daten sein kann, dürfte kaum als "sinnvoll" empfunden werden. Dagegen ist z.B. nicht von allen Mittelwerten zu fordern, dass sich positive und negative Abweichungen von diesem Mittelwert ausgleichen, was z.B. für das arithmetische Mittel gilt. Eine solche Eigenschaft von allen Mittelwerten, also von einer ganzen Klasse von Maßzahlen als Axiom zu fordern, würde auch bedeuten, dass dann das arithmetische Mittel der einzige sinnvolle Mittelwert wäre.

Zu 1:

Formale Eigenschaften, die für statistische Maßzahlen aller Art häufig gefordert werden sind beispielsweise:

a) **Stetigkeit:**
Geringfügige (infinitesimale) Veränderungen in den Komponenten des Beobachtungsvektors $[x_1,x_2,...,x_n]$ eines stetigen Merkmals X sollen sich nicht sprunghaft auf die entsprechende Maßzahl auswirken.

b) **Sensitivität, Robustheit:**
Hierbei geht es um die Frage ob eine Maßzahl resistent ist gegenüber Ausreißern und außergewöhnlichen Daten.

c) **Maßeinheit, Normierung des Wertebereichs:**
Häufig wird gefordert, dass eine Maßzahl die gleiche Maßeinheit besitzt, wie die mit ihr beschriebene Variable X oder aber dass sie "dimensionslos" (ohne Maßeinheit) ist. Im zweiten Fall wird der Wertebereich der Maßzahl meist auf ein bestimmtes Intervall "normiert". Übliche Normierungen einer Maßzahl M sind

$0 \le M \le 1$ oder $-1 \le M \le +1$.

Durch eine geeignete Lineartransformation lässt sich jede Maßzahl auf einen bestimmten Wertebereich normieren (Abschn. c).

d) Aggregation, Zerlegung:

Eine Maßzahl kann sich beziehen auf eine Gesamtheit G oder auf deren Teilgesamtheiten G_i (i=1,2,...,r). Dabei wird angenommen, dass die Gesamtheit (Menge) G in Teilmengen G_i "zerlegt", bzw. umgekehrt, die Gesamtheit aus diesen Teilmengen "zusammengefügt" (aggregiert) werden kann.

Eine vollständige Zerlegung (Partition) bedeutet, dass die Vereinigung der Teilmengen G_i die Menge G ist, und dass die Teilmengen G_i paarweise disjunkt sind (ihre Schnittmenge jeweils leer ist).

Wenn M die Maßzahl für die Gesamtheit ist und M_i die entsprechende Maßzahl für die i-te Teilgesamtheit ist, dann sollte $M = f(M_1,...,M_r)$ eine einfach zu interpretierende Funktion sein (z.B. ein gewogenes arithmetisches Mittel).

2. Vorteile der axiomatischen Betrachtungsweise

Die axiomatische Betrachtungsweise hat in der Deskriptiven Statistik große Vorteile:

- Sie erlaubt es, eine Klasse von Maßzahlen generell zu charakterisieren und gegen eine andere Klasse abzugrenzen (z.B. das "Wesen" der Streuung [Kap.5] zu definieren und z.B. Streuungsmaße gegen Disparitätsmaße [Kap.6] abzugrenzen).
- Mit ihr ist es möglich, Eigenschaften (auch weniger offensichtliche) von Maßzahlen systematisch herauszuarbeiten und Kriterien zur Auswahl und Beurteilung von Maßzahlen anzugeben; zu solchen interessierenden Eigenschaften gehört z.B. auch das "Verhalten" einer Maßzahl bei Änderungen an den Daten (etwa Hinzufügen oder Streichen bestimmter Meßwerte) in Form von sog. "Proben" (z.B. "Ergänzungsprobe" bei der Konzentrationsmessung, oder "Rundprobe" in der Indextheorie).
- Sie kann auch die Konstruktion neuer Maßzahlen[2)] anregen.

Die axiomatische Betrachtungsweise ist bisher am weitesten und erfolgreichsten bei Indexzahlen und Konzentrationsmaßen angewandt worden. Im Falle von Indexzahlen ist auch neben der Betrachtung formaler

---

2) In diesem Sinne nutzte z.B. Irving Fisher seine "Proben" (oder "tests") als "finders of formulae". Entsprechend sind in der Indextheorie der "best linear index" (Theil) oder der Integralindex (Divisia) als solche Maßzahlen konstruiert worden die bestimmten Axiomen genügen.

Axiome (in der sog. "formalen Theorie der Indexzahlen") die Untersuchung inhaltlicher (sachlicher) Forderungen an Indexzahlen üblich ("ökonomische Theorie der Indexzahlen").

c) Normierung von Maßzahlen

Wenn eine Maßzahl M den minimalen Wert M_u und den maximalen Wert M^o annimmt, so kann man leicht aus M durch eine Lineartransformation eine auf einen bestimmten Wertebereich normierte Maßzahl M^* erhalten. So erhält man z.B. - wie leicht zu beweisen ist - eine Maßzahl M^*, die zwischen $M_u,^*$ als kleinstem und M^{o*} als größtem Wert schwankt, mit der folgenden Lineartransformation:

$$(2.3) \quad M^* = M_u,^* + (M - M_u) \frac{M^{o*} - M_u^*}{M^o - M_u}$$

Praktisch besonders bedeutsam sind die beiden folgenden Normierungen:

a) Normierung der Maßzahl M zur Maßzahl M^* auf den Wertebereich $0 \le M^* \le 1$ mit Gl. 2.3a und

b) Normierung auf $-1 \le M^* \le +1$ mit Gl. 2.3b:

$$(2.3a) \quad M^* = \frac{M - M_u}{M^o - M_u} \qquad (2.3b) \quad M^* = \frac{2(M - M_u)}{M^o - M_u} - 1$$

Beispiel 2.2:

Angenommen, eine Maßzahl M schwankt zwischen ¼ als Untergrenze und ½ als Obergrenze. Wie kann man aus M eine Maßzahl M^* so bilden, dass M^* nur Werte zwischen

a) 0 und 1

b) -1 und +1 annimmt?

Lösung 2.2:

Wegen $M_u = 1/4$ und $M^o = 1/2$ erhält man im Fall

a) nach Gl. 2.3a: $M^* = 4M - 1$ und im Fall

b) nach Gl. 2.3b: $M^* = 8M - 3$.

Kapitel 3: Eindimensionale Häufigkeitsverteilungen

In Kapitel 2 wurden bereits Unterscheidungen hinsichtlich der Datenarten vorgenommen, auf die hier aufgebaut werden soll (vgl. Übersicht 3.1). Entscheidend auch für die in den folgenden Kapiteln behandelte Berechnung von Maßzahlen (gewogene und ungewogene Ansätze) ist danach die Unterscheidung in klassierte und unklassierte Daten und bei letzteren zwischen gruppierten Daten und Einzelbeobachtungen.

1. Unklassierte Daten

a) Häufigkeitsverteilung

Def. 3.1: Häufigkeiten

Seien $x_1,x_2,...x_m$ (gruppierte Daten) die m realisierbaren Ausprägungen eines diskreten Merkmals X, dann heißt die Anzahl der Beobachtungseinheiten mit der i-ten Ausprägung

(3.1) $n_i = n(x_i)$ absolute Häufigkeit (i = 1,2,...,m),

und mit $n = \Sigma n_i$ (Gesamthäufigkeit, Umfang der Beobachtungsgesamtheit) der Quotient

(3.2) $h_i = h(x_i) = \frac{n_i}{n}$ relative Häufigkeit

der i-ten Ausprägung des Merkmals X.

Bemerkungen und Folgerungen:

1. Offensichtlich ist n_i eine natürliche Zahl mit $n_i \geq 1$ und für die relativen Häufigkeiten gilt (bei nichthäufbaren Merkmalen):

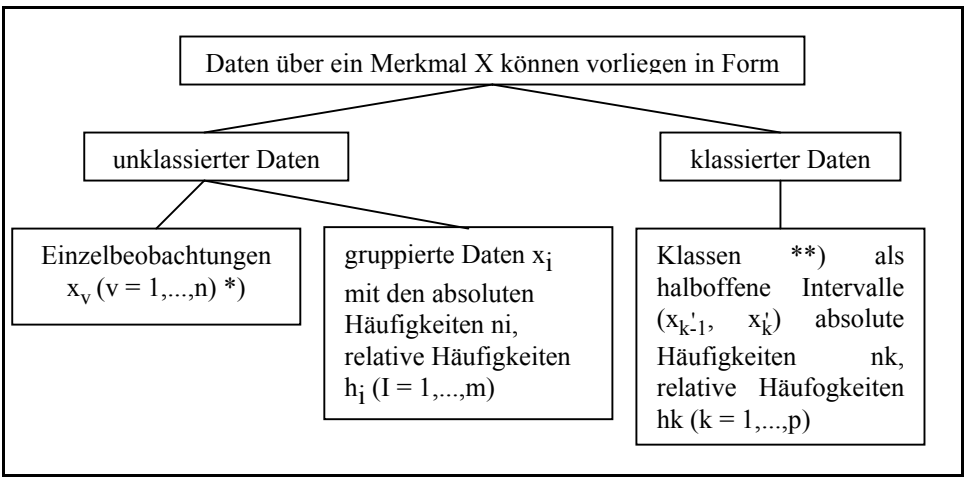
$0 \le h_i \le 1$ und (wegen $n = \Sigma n_i$) $\Sigma h_i = 1$.

Die mit 100 multiplizierten relativen Häufigkeiten heißen prozentuale Häufigkeiten.

2. Bei einem mindestens ordinalskalierten Merkmal sollten die Merkmalsausprägungen der Größe nach geordnet und numerisch codiert sein, so dass gilt $x_1 < x_2 < ... < x_m$.
   Eine solche Anordnung der Werte x_i ist erforderlich um kumulierte Häufigkeiten und die empirische Verteilungsfunktion (vgl. Def. 3.3) zu bestimmen. Eine Codierung (Zuordnung von Zahlen zu Merkmalsausprägungen) ist für die statistische Analyse nicht immer notwendig, erleichtert diese aber erheblich.

3. Bei den Merkmalswerten $x_1,x_2,...x_n$ der n Einheiten, die alle verschieden sind (Einzelbeobachtungen), ist $n_v = 1$ und $h_v = 1/n$ für alle Werte von v. Die Summe $\Sigma x_v = \Sigma x_i n_i = S$ heißt **Merkmalssumme**. Sie ist nur bei extensiven Merkmalen sinnvoll interpretierbar. Dichotome Merkmale werden zweckmäßig wie folgt codiert: $x_1 = 0$ und $x_2 = 1$, so dass die Merkmalssumme n_2 ist.

Übersicht 3.1: Daten



*) In späteren Abschnitten (insbes. im Kap. 8) wird gelegentlich auch x_i anstelle von x_v verwendet.

*) Es sei verabredet, dass x_k' die Obergrenze der k-ten Klasse (d.h. der k-ten der p aneinander grenzenden Größenklassen) ist, so dass x_{k-1}' die Obergrenze der (k-1)-ten Klasse und damit die Untergrenze der k-ten Klasse ist.

Def. 3.2: Häufigkeitsverteilung

Das m-Tupel $[(x_1,n_1),(x_2,n_2),...,(x_m,n_m)]$ heißt absolute Häufigkeitsverteilung und entsprechend ist $[(x_1,h_1),(x_2,h_2),...,(x_m,h_m)]$ die (relative) Häufigkeitsverteilung eines Merkmals X.

Sie ist eine Zuordnung von Häufigkeiten (n_i oder h_i) zu den Ausprägungen x_i (i = 1,2,...,m) des Merkmals X und zeigt, wie sich die n Einheiten über die möglichen Werte von X "verteilen". Häufigkeitsverteilungen können tabellarisch (Häufigkeitstabelle) oder graphisch dargestellt werden. Die Art der graphischen Darstellung hängt von der Skala des Merkmals X ab.

b) Tabellen und Graphiken

Für die praktische Anwendung der Statistik (weniger dagegen für die wissenschaftliche Beschäftigung mit Statistik) spielt die Gestaltung von Tabellen und "aussagefähigen" und eindrucksvollen Graphiken (meist mit einer speziell hierfür entwickelten Software) eine große Rolle. Es kann hier auf diese Gegenstände nur sehr kurz eingegangen werden.

Tabellen:

Für die Gestaltung von Tabellen gibt es Normen (DIN-Normblatt 55301). Eine Tabelle ist eine geordnete Zusammenstellung der Ergebnisse statistischer Erhebungen oder Berechnungen mit **Zeilen** (waagrecht) und **Spalten** (senkrecht), wobei in den so gebildeten **Tabellenfächern** i.d.R. Häufigkeiten eingetragen werden. Eine Tabelle hat stets eine Überschrift und eine Quellenangabe. Sie kann auch Fußnoten haben. Zeilen und Spalten können numeriert werden. Die **Überschrift** soll enthalten: Dargestellte Tatbestand, räumliche und zeitliche Abgrenzung der Erhebungsmasse. **Fußnoten** können Erklärungen zu Zahlen in einzelnen Tabellenfächern oder ergänzende Hinweise zu Textangaben enthalten.

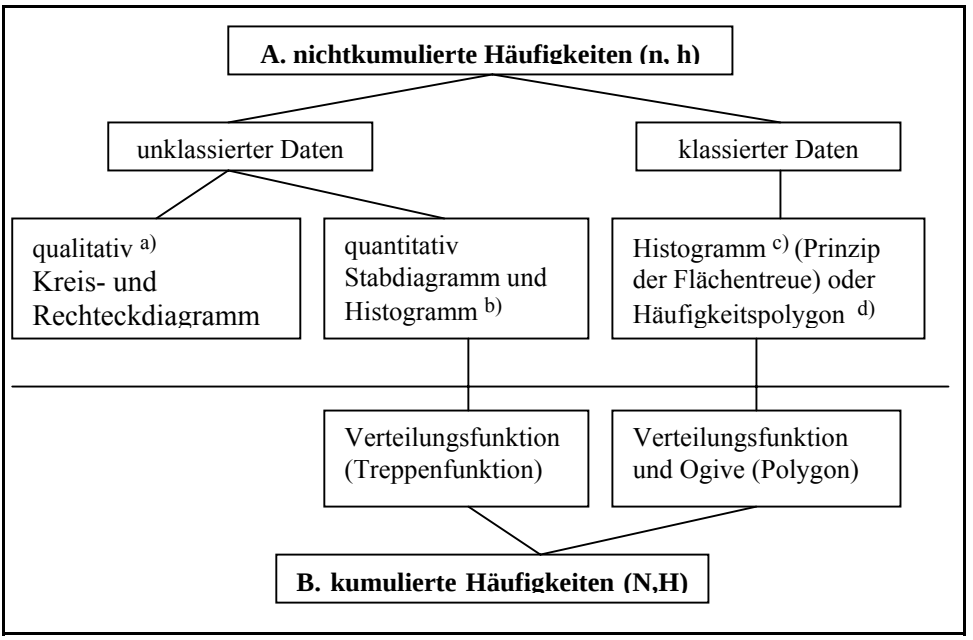
Graphiken:

Es gibt eine Fülle von Gestaltungsmöglichkeiten für graphische Darstellungen. Abgesehen von Piktogrammen (Bildgraphiken mit Verwendung anschaulicher Symbole) und Kartogrammen handelt es sich jedoch meist um Varianten der in Übers. 3.2 genannten Diagramme.

1. **Bei qualitativen Merkmalen** ist eine Reihenfolge der Merkmalsausprägungen nicht definiert, so dass die Häufigkeitsverteilung (bzw. in

diesem Fall, die **Struktur**) des Merkmals X am besten durch ein **Kreisdiagramm** (engl. pie chart) dargestellt wird (vgl. Beispiel 3.1 und Abb. 3.1 links). Eine Alternative ist das **Rechteckdiagramm** (z.T. auch Flächendiagramm genannt, Abb. 3.1 rechts).
Die Winkel a_i der Kreissegmente (und damit die Flächen der Kreissektoren) bzw. die Höhen der Rechtecke des Rechteckdiagramms sind proportional zu den Häufigkeiten. Damit verhalten sich auch die Flächen zueinander wie die relativen, bzw. absoluten Häufigkeiten und es gilt $a_i = 360°h_i$. Die Reihenfolge der Sektoren (Kreissegmente) kann beliebig gewählt werden (für X wird ja nur eine Nominalskala vorausgesetzt).

Übersicht 3.2: Graphische Darstellung von Häufigkeiten



a) kategorial, nominalskaliert;
b) in diesem Fall Stäbe, Säulen oder (nicht notwendig aneinander angrenzende) Blöcke gleicher Breite;
c) bei gleichen Breiten (äquidistante Klassen) ist die Höhe und die Fläche sowie bei ungleichen Breiten die Fläche der aneinander angrenzenden Blöcke proportional zur absoluten oder relativen Häufigkeit;
d) lineare Verbindung der Blockmitten (auch Kurvendiagramm genannt);
e) kumulierte Häufigkeiten (Summenhäufigkeiten) gem. Def. 3.3 (bei Resthäufigkeiten [Def. 3.4] erhält man jeweils fallende Treppenkurven).

Vergleicht man die Verteilungen mehrerer Massen A und B unterschiedlichen Umfangs (n_A,n_B) bezüglich des Merkmals X miteinander, so wird der Unterschied des Umfangs durch entsprechend unterschiedlich große Flächen der Kreise, bzw. Rechtecke zum Ausdruck gebracht. Da die Kreisfläche $F_A=r_A,^2p$ ist (und F_B entsprechend), muss für die Quadrate der Radien gelten $r_A^2 / r_B^2 = n_A / n_B$.

2. Bei **quantitativen diskreten Merkmalen** empfiehlt sich die Darstellung eines **Stabdiagramms** (andere Ausdrücke hierfür sind: **Balken-, Block-** oder **Säulendiagramm**, allgemein: **Histogramm**). Die Höhen der Stäbe bzw. Säulen sind proportional zu den relativen oder absoluten Häufigkeiten (Abb. 3.2). Bei ordinalskalierten Merkmalen sind ihre Abstände nicht eindeutig und bei qualitativen Merkmalen wäre auch ihre Reihenfolge nicht eindeutig.

Beispiel 3.1:

a) Im Jahr 1989 ergaben sich für die Bundesrepublik Deutschland folgende Anteile der einzelnen Wirtschaftsbereiche an der Gesamtzahl der Erwerbstätigen (Quelle: Gutachten SVR 1990):

Staat und Private Haushalte	19,8%
Dienstleistungsunternehmen	18,0%
Handel und Verkehr	18,7%
Warenproduzierendes Gewerbe	39,8%
Land- und Forstwirtschaft, Fischerei	3,7%

Veranschaulichen Sie die Struktur der Erwerbstätigen durch ein Kreis- und Rechteckdiagramm!

b) An einer Straßenkreuzung wurden an 128 Tagen die Unfallzahlen (Anzahl der Unfälle an einem Tag) gemessen.

Anzahl der Verkehrsunfälle (x_i)	0	1	2	3	4
Anzahl der Tage (n_i)	13	26	38	32	19

Stellen Sie die Häufigkeitsverteilung durch ein Stabdiagramms dar!

Lösung 3.1:

a) Zum Kreis- und Rechtecksdiagramm vgl. Abb. 3.1. Die Winkel des Kreisdiagramms sind:

Staat und Private Haushalte (Staat)	71° 17'
Dienstleistungsunternehmen (Dienst)	64° 48'
Handel und Verkehr (H+V)	67° 20'
Warenproduzierendes Gewerbe (Gew)	143° 17'
Land- und Forstwirtschaft, Fischerei (L+F)	13° 19'

b) Zum Stabdiagramm (oder Blockdiagramm) vgl. Abb. 3.2.

c) Summenhäufigkeiten

Def. 3.3: Summenhäufigkeit, Verteilungsfunktion

a) Die Summe N_i der absoluten Häufigkeiten n_j (j=1,2,...,i) aller Merkmalsausprägungen x_j eines mindestens ordinalskalierten Merkmals, die kleiner oder gleich x_i sind,

$$(3.3) \quad N_i = N(x_i) = n\,(X \le x_i) = \sum_{j=1}^{i} n_j$$

heißt absolute kumulierte Häufigkeit (**absolute Summenhäufigkeit**).

Abb. 3.1: Kreis- und Rechteckdiagramm

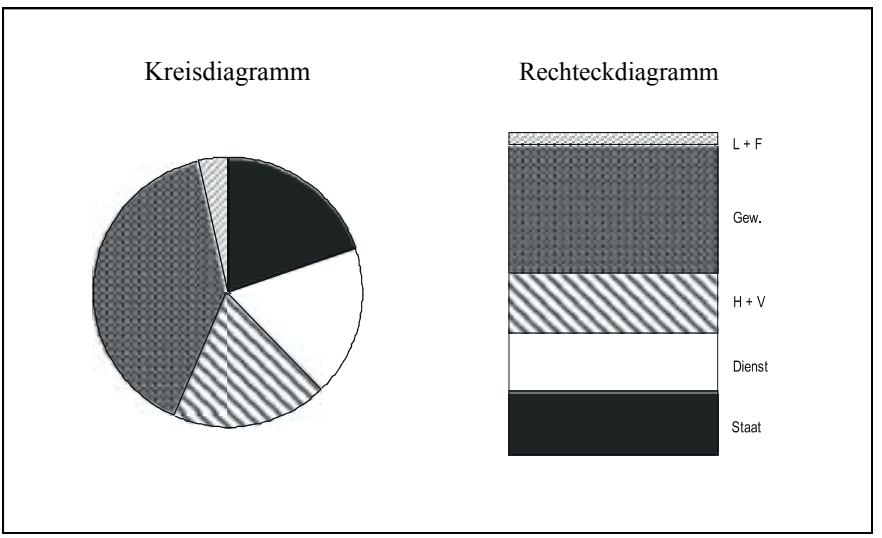
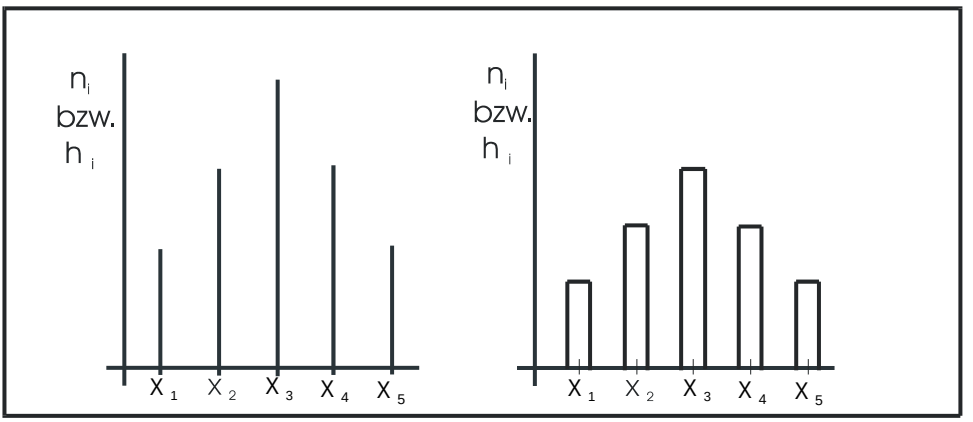


Abb. 3.2: Stabdiagramm



b) Entsprechend heißt

$$(3.4) \quad H_i = H(x_i) = h(X \le x_i) = \sum_{j=1}^{i} h_j = \frac{N_i}{n}$$

relative kumulierte Häufigkeit (relative Summenhäufigkeit).

c) Die Funktion

$$(3.5) \quad H(x) = \begin{cases} 0 & \text{für } x < x_1 \\ H_j & \text{für } x_j \le x < x_{j+1} (j = 1,2,...,m-1) \\ 1 & \text{für } x \mid x_m \end{cases}$$

der reellen Variable X heißt (empirische) **Verteilungsfunktion** oder (relative) Summenhäufigkeitskurve des diskreten Merkmals X.

d) Die Funktion x = G(H) ist die **inverse Verteilungsfunktion**.

Bemerkungen zu Def. 3.3:

1. Wie man leicht sieht, gilt: $N_1 = n_1$, $N_2 = n_1 + n_2$, $N_3 = n_1 + n_2 + n_3$ usw. und schließlich $N_m = \Sigma\, n_i = n$ (mit i=1,2,...,m). Ferner ist $H_m = 1$.
2. Im Fall von n verschiedenen **Einzel**beobachtungen $x_1,...,x_n$ gilt: $N_i = i$ und $H_i = i/n$.
3. Die empirische Verteilungsfunktion H(x) gibt die Summe der relativen Häufigkeiten aller Merkmalswerte an, die kleiner oder gleich x sind. Während die einzelnen Summenhäufigkeiten H_i dargestellt werden als Stäbe, bei denen die Häufigkeiten h_j "gestapelt" werden (stacked histogram) und zwischen den Stäben Lücken sind, hat

der Graph der Funktion H(x) die Gestalt einer Treppe. H_i ist geeignet für unklassierte Daten, H(x) auch für klassierte Daten.

4. H(x) ordnet jedem Wert x die bis dahin (also für $X \le x$) erreichte Summenhäufigkeit H zu. Mit der inversen Funktion G kann man für bestimmte Werte von H (etwa H = 1/2 oder H = 1/4) den Wert x bestimmen (was im Falle von 1/2 und 1/4 die Quartile Q_1 und $Q_2 = \tilde{x}_{0,5}$ [vgl. Kap. 4] sind).

5. Weniger gebräuchlich ist die Darstellung der absoluten Summenhäufigkeiten N_j als absolute Summenhäufigkeitskurve. Es gilt N(x) = nH(x) für jedes x.

Eigenschaften der empirischen Verteilungsfunktion H(x):

a) Aus $x < y$ folgt $H(x) < H(y)$ (H ist monoton nichtfallend).

b) $0 \le H(x) \le 1$ (mit $H(-\infty) = 0$ und $H(\infty) = 1$ wenn der Definitionsbereich nicht beschränkt ist).

c) H(x) ist für alle $x \in \mathbb{R}$ definiert und eine rechtsseitig stetige Funktion.

d) H(x) hat als Treppenfunktion Sprungstellen bei $x_1, x_2, ..., x_m$. Die Größe der Sprünge beträgt $h_i = H(x_i) - H(x_{i-1})$.

e) Unter bestimmten Voraussetzungen (X ein nichtnegatives Merkmal) ist die Fläche oberhalb von H(x) das arithmetische Mittel (vgl. Abb. 4.3).

Def. 3.4: Resthäufigkeit

Die Summe N_i^- der absoluten Häufigkeiten n_j (j = i+1, i+2, ..., m) aller Merkmalsausprägungen eines mindestens ordinalskalierten Merkmals, die größer als x_i sind,

$$(3.6) \quad N_i^- = N^-(x_i) = n(x > x_i) = \sum_{j=i+1}^{m} n_j = n - N_i$$

heißt **absolute Resthäufigkeit**. Entsprechend heißt

$$H_i^- = 1 - H_i \qquad \textbf{(relative) Resthäufigkeit}$$

und die analog zu Gl. 3.5 definierte Funktion

$$H^-(x) = 1 - H(x) \qquad \textbf{relative Resthäufigkeitsfunktion}.$$

Die Resthäufigkeiten spielen in statistischen Anwendungen eine weniger wichtige Rolle. Es ist leicht zu sehen, dass gilt:

1. Die in Abb. 4.3 schraffierte Fläche unter $\bar{H}(x)$ ist gleich der Fläche oberhalb von H(x) und damit gleich dem arithmetischen Mittel.
2. Ferner ist $N_i + \bar{N}_i = n$, $H_i + \bar{H}_i = 1$ und $H(x) + \bar{H}(x) = 1$.

Beispiel 3.2:

Stellen Sie anhand der Daten des Bsp. 3.1b die Summenhäufigkeiten H_i und die Resthäufigkeiten $\bar{H}_i$ graphisch dar.

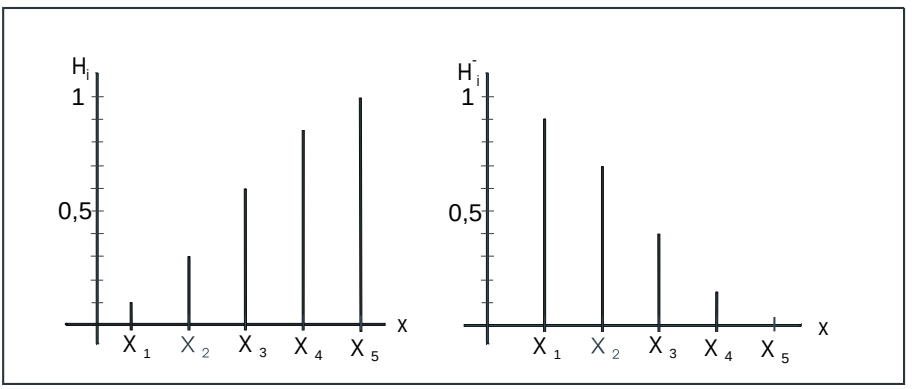
Lösung 3.2:

Zur Übersichtlichkeit wird folgende Arbeitstabelle aufgestellt:

Nr. i	Anzahl der Verkehrsunfälle (x_i)	Anzahl der Tage (n_i)	Anteil der Tage (h_i)*)	relative Summenhäufigkeit H_i	Resthäufigkeit $\bar{H}_i$
1	0	13	0,1	0,1	0,9
2	1	26	0,2	0,3	0,7
3	2	38	0,3	0,6	0,4
4	3	32	0,25	0,85	0,15
5	4	19	0,15	1,0	0

*) gerundet

Abb. 3.3: Summenhäufigkeiten H_i und Resthäufigkeiten $\bar{H}_i$ für das Beispiel 3.2



2. Klassierte Daten

a) Größenklassen

Notwendigkeit der Klassenbildung

Häufig enthält das Datenmaterial so viele unterschiedliche Merkmalsausprägungen, dass eine Darstellung sämtlicher Beobachtungen nach Art eines Stabdiagramms wenig aufschlußreich wäre (vgl. Abb. 3.4 wo die absoluten Häufigkeiten n_i jeweils nur 1 oder 2 betragen). In diesem Fall ist eine Klassenbildung (Klassierung) zu empfehlen um die Gestalt der Verteilung besser erkennbar zu machen.
Das gilt v.a. bei stetigen Merkmalen wie Gewicht, Körpergröße, Alter, Länge von Schrauben etc., die zumindest theoretisch beliebig genau gemessen werden können und bei quasistetigen Merkmalen wie Einkommen, Vermögen, Sparguthaben etc. Aber auch bei diskreten Merkmalen wie z.B. Punktzahlen in einer Klausur, IQ-Werte, Stückzahlen etc. kann eine unklassierte Verteilung sehr unübersichtlich sein, wie das Beispiel 3.3 (Abb. 3.4) zeigt.

Intervallabgrenzung

Für ein mindestens ordinalskaliertes Merkmal X lassen sich (beidseitig) offene oder geschlossene Intervalle wie folgt abgrenzen:
(a,b) soll bedeuten $a < x < b$ (offenes Intervall) und
[a,b] soll bedeuten $a \le x \le b$ (geschlossenes Intervall).
In diesen Fällen entstehen jedoch dann Unklarheiten, wenn x die (Grenz- oder Eck-) Werte x = a und x = b annimmt. Eine widerspruchsfreie Intervallabgrenzung ist jedoch möglich mit (a,b]:
$a < x \le b$ (oder mit [a,b), also wenn gilt $a \le x < b$).
Ist x_k,' die Obergrenze der k-ten Größenklasse, dann ist x'_{k-1} die Untergrenze dieser Klasse. Damit lassen sich die Begriffe der Def. 3.5 definieren:

Def. 3.5: Klassierung

a) In einer klassierten Verteilung wird die Variable X in p Intervalle (**Klassen**) $(x'_{k-1}, x'_k]$ eingeteilt (linksseitig offene Intervalle) mit k = 1,2,..,p wobei x'_k die Obergrenze der k-ten Größenklasse ist (vgl. Bem. Nr. 4).

b) Die Differenz $b_k = x'_k - x'_{k-1}$ heißt **Klassenbreite**
und die Größe $m_k = ½(x'_{k-1} + x'_k)$ heißt **Klassenmitte** der k-ten Klasse.

c) Die Anzahl n_k der betrachteten Einheiten, die in die k-te Klasse "fallen" $[n_k = n\ (x'_{k-1} < x \leq x'_k)]$, ist die absolute **Klassenhäufigkeit** (der k-ten Klasse) und die Anteile $h_k = n_k/n$ sind die relativen Klassenhäufigkeiten.

d) Die Quotienten $h_k^* = h_k/b_k$ [k=1,2,...,p] sind **Häufigkeiten je Klassenbreite**. Mit $b_k \rightarrow 0$ wird X zu einer stetigen Variable und das Häufigkeitspolygon zu einem kontinuierlichen Kurvenzug (zur Dichtefunktion).

Bemerkungen zu Def. 3.5:

1. Geht man davon aus, dass sich die Merkmalsträger in den Klassenmitten konzentrieren, so können die Daten wie gruppierte Daten behandelt werden. Durch Klassierung wird ein stetiges Merkmal quasi zu einem diskreten (Diskretionierung). Klassenbildung ist eine Transformation, bei der Information (nämlich die Verteilung innerhalb der Klasse) verloren geht.

2. Die **Klassenmitten** m_k sind i.d.R. nicht identisch mit den wahren Klassenmittelwerten, es sei denn die Einheiten verteilen sich gleichmäßig innerhalb einer Klasse. m_k heißt auch "Präsumptivwert" (also Schätzwert des [wahren] **Klassenmittelwerts**).

3. Für die Entscheidung über Anzahl und Breite der Klassen lassen sich keine formalen Kriterien angeben. Hierbei sind auch Manipulationen möglich, d.h. das gleiche Datenmaterial kann optisch sehr unterschiedlich wirken (vgl. Abb. 3.5). Es sollte dabei berücksichtigt werden:

   - Zweck der Untersuchung
   - Meßgenauigkeit beim Merkmal X
   - Streuung der Merkmalswerte
   - Anzahl der Erhebungs- bzw. Darstellungseinheiten.

   Ungleiche Klassenbreiten empfehlen sich, wenn die Merkmalsausprägungen sehr unterschiedlich dicht liegen. Zu kleine Klassen lassen Meßfehler zu stark hervortreten, zu große Klassen verdecken wiederum Charakteristiken der Verteilung. Klassen sollten in jedem Fall so gebildet werden, dass keine leeren Klassen auftreten. Im allgemeinen wird man mit 5 bis 20 Klassen auskommen.

4. Häufig sind die erste und p-te Klasse nicht geschlossen. Mit solchen sog. offenen Randgruppen treten Schwierigkeiten bei der graphischen Darstellung und der Berechnung bestimmter Mittelwerte auf.

5. Üblicher als die obige Abgrenzung $(x'_k, x'_{k+1}]$ sind in der Praxis rechtsseitig offene Intervalle $[x'_k, x'_{k+1})$ "von ... bis unter ...". Die Abgrenzung $(x'_k, x'_{k+1}]$ wurde jedoch in Def. 3.5 gewählt um mit der Verteilungsfunktion (Def. 3.3) konsistent zu sein.

b) Graphische Darstellungen

Häufigkeitsverteilung bei einem klassierten Merkmal:

Graphisch wird eine klassierte Verteilung durch das Histogramm dargestellt. Es setzt sich aus Rechtecken über den Klassenbreiten b_k zusammen, deren Flächen proportional zu den Klassenhäufigkeiten h_k sind (**Prinzip der Flächentreue**). Daraus folgt, dass die Höhen der Rechtecke die Häufigkeitsdichten h_k^* sind. Häufigkeiten werden also zweidimensional (durch Flächen) repräsentiert, was für die Beurteilung schwieriger ist: man kann leicht Unterschiede in der Höhe von Blöcken feststellen aber nicht immer eindeutig die Fläche von Rechtecken vergleichen, es sei denn ein Rechteck ist in beiden Dimensionen (Höhe und Breite) größer als das andere. Gelegentlich verbindet man auch die Mitten der oberen Rechteckseiten eines Histogramms miteinander, wobei dieser Polygonzug auf der x-Achse im Wert $x_1' - \frac{1}{2}b_1$ beginnt und mit $x_p' + \frac{1}{2}b_p$ endet. Abb. 3.6 zeigt dieses **Häufigkeitspolygon** (dessen Gesamtfläche über der Abszisse gleich derjenigen des Histogramms ist) für ein fiktives Beispiel mit p = 5 Klassen.
Bei infinitesimal kleinen Klassenbreiten geht die Darstellung des Histogramms bzw. Häufigkeitspolygons über in eine **Dichtefunktion** eines stetigen Merkmals (die Dichte ist ein stetiger Kurvenzug). Sie ist nicht für die Deskriptive, sondern nur für die Induktive Statistik von Bedeutung.

Beispiel 3.3:

Die Messwerte für das Körpergewicht X von n = 25 Personen (in kg) seien: 63, 61, 70, 81, 72, 74, 69, 62, 75, 79, 77, 80, 86, 76, 78, 70, 80, 77, 73, 66, 85, 67, 83, 82, 71. Man stelle diese Daten dar als

- Stabdiagramm, also unklassiert (Abb. 3.4),
- klassierte Verteilung mit folgender Klasseneinteilung (Abb. 3.5):

a) einheitliche Klassenbreiten von jeweils 5 von 60 bis 90kg: beginnend mit "60 bis einschließlich 65" (also (60,65]), bis (85,90];
b) einheitliche Klassenbreiten von jeweils 10.

Lösung 3.3:

Das Stabdiagramm der Abb. 3.4 ist wenig sinnvoll, denn es ist kaum erkennbar, dass die Daten eine Verteilung darstellen. Bei den beiden Klasseneinteilungen (Abb. 3.5) zeigt sich, dass bei ein und demselben Datensatz eine unterschiedliche Gestalt der Verteilung möglich ist. Bei einer Klassierung mit einer einheitlichen Klassenbreite von $b_k = 5$ (Abb. 3.5a), also p = 6 Klassen treten die Charakteristiken der Verteilung recht gut hervor. Dagegen scheint eine Wahl von p = 3 Klassen zu grob zu sein. So kommt in Abb. 3.5b nicht zum Ausdruck, dass sich die Meßwerte in den Intervallen (70, 80] und (80, 90] in der oberen bzw. in der unteren Hälfte häufen.

Abb. 3.4: Daten des Beispiels 3.3 als Stabdiagramm

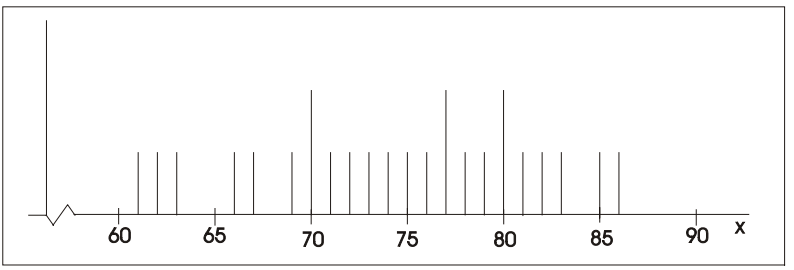
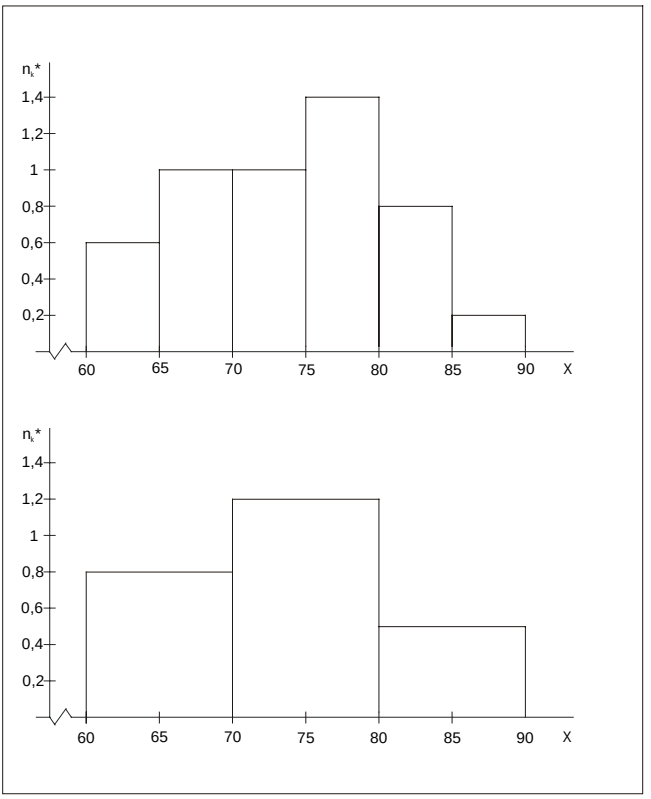


Abb. 3.5: Verschiedene Klasseneinteilungen für die Daten des Beispiels 3.3



Summenhäufigkeiten bei klassierter Verteilung:

Die Definitionen 3.2 und 3.3 für das diskrete Merkmal X gelten analog auch für ein klassiertes Merkmal X. Die Größen N_k bzw. H_k sind die absoluten bzw. relativen kumulierten Klassenhäufigkeiten. Letztere betragen $H_0 = 0$ vor der Untergrenze x_0' der ersten Klasse (denn $x_{1-1}' = x_0'$) und $H_p = 1$ nach der Obergrenze x_p' der letzten (p-ten) Klasse.

Analog zum Häufigkeitspolygon (Abb. 3.6) der Klassenhäufigkeiten kann man auch einen Polygonzug der Summenhäufigkeiten (kumulierten Klassenhäufigkeiten) bestimmen. Man nennt diese Kurve **Ogive** (Abb. 3.7 für das Beispiel der Abb. 3.6). Die Ogive[1] H(x) ist eine Näherung der exakten empirischen Verteilungsfunktion, die sich i.d.R. nicht angeben läßt, weil die Verteilung der Merkmalsträger (Einheiten)

[1] Die Ogive ist die lineare Verbindung der Treppenabsätze der Verteilungsfunktion H(x). Die Steigung ist dabei jeweils die Größe h_k^*. Um die Symbolik nicht zu kompliziert zu machen, soll die Ogive (oder "approximierende Verteilungsfunktion") auch H(x) genannt werden.

innerhalb der Klassen nicht bekannt ist. Sie ist unter der Annahme einer Gleichverteilung (Gleichhäufigkeit jeder Ausprägung in der Klasse) oder einer symmetrischen Verteilung innerhalb der Klassen konstruiert und deshalb eine stückweise lineare Funktion (Polygonzug) mit den genannten Eigenschaften von H(x).

Abb. 3.6: Histogramm und Häufigkeitspolygon p = 5 Klassen

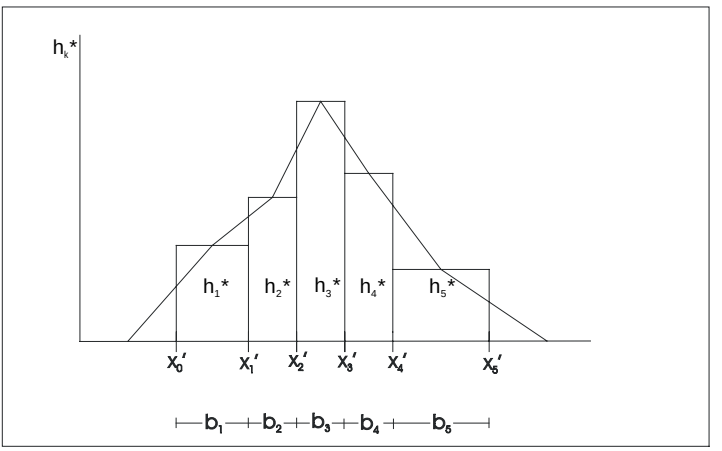
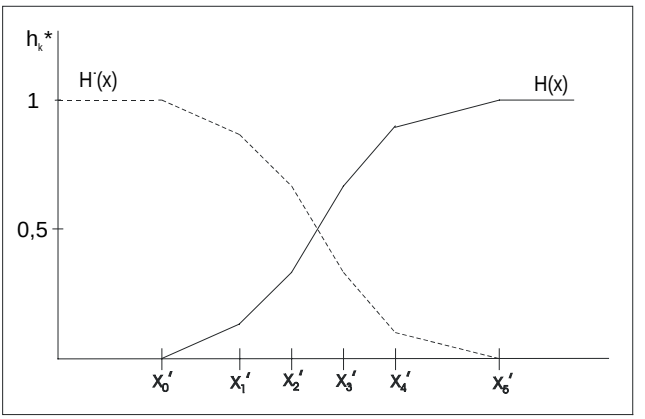


Abb. 3.7: Ogive H(x) und Resthäufigkeitsfunktion $H^-(x)$ für das Beispiel der Abb. 3.6



Beispiel 3.4:

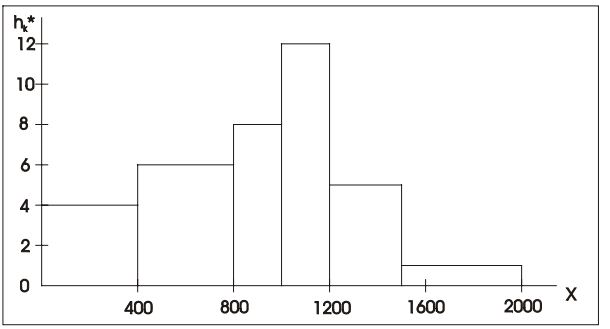
Gegeben sei die folgende Verteilung der Verdienste in einem Betrieb, für welche die Dichten h_K^*, sowie die Summen- und Resthäufigkeitskurve (also H(x) und $H^-(x)$) zu bestimmen und zu zeichnen sind:

Angaben		Lösung			
von...bis unter	h_k	Breite b_k	Dichte h_k^*	kumulierte Häufigk. H(x)	$H^-(x)$
0 - 400	0,16	400	4*)	0,16	0,84
400 - 800	0,24	400	6	0,40	0,60
800 - 1000	0,16	200	8	0,56	0,44
1000 - 1200	0,24	200	12	0,80	0,20
1200 - 1500	0,15	300	5	0,95	0,05
1500 - 2000	0,05	500	1	1	0

*) In dieser Spalte ist jeweils der 10000-fache Wert verzeichnet, d.h. die Angabe 4 in der ersten Zeile ist zu verstehen als $4 \cdot 10^{-4} = 0{,}16/400$. Die Dichte ist stets $h_k^* = h_k / b_k$. Es gilt $H^-(x) = 1 - H(x)$ und $H^-(x)$ ist 0,84 an der Stelle x = 400 und 1 bei x=0.

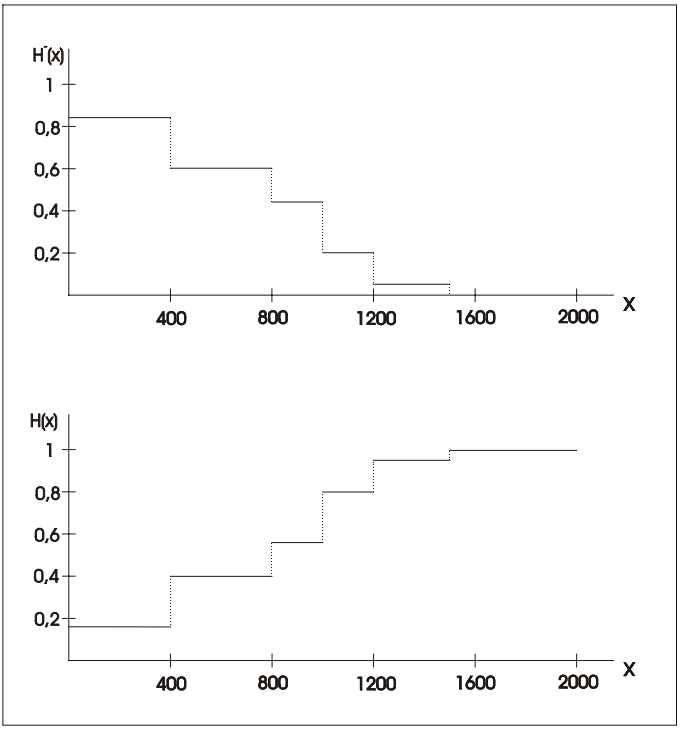
Abb. 3.8 zeigt die klassierte Verteilung mit den Höhen h_k^* und Abb. 3.9 die kumulierten Häufigkeiten H (Verteilungsfunktion) und Resthäufigkeiten H^- (gestrichelte Linie).

Abb. 3.8: Häufigkeitsverteilung der klassierten Verteilung des Beispiels 3.4



Man beachte, dass die Verteilungsfunktion (anders als bei einer diskreten Verteilung oder bei einer klassierten Verteilung mit gleich breiten Klassen) nicht einfach ein Aufeinanderstapeln der Stäbe bzw. Blöcke der Häufigkeitsverteilung sein kann. Denn das würde darauf hinauslaufen, die Höhen h_k^* und nicht (wie es richtig ist) die relativen Häufigkeiten h_k zu addieren.

Abb. 3.9: Verteilungsfunktion und Resthäufigkeitskurve (Beispiel 3.4)



Kapitel 4: Mittelwerte und andere Lagemaße

1. Eigenschaften von Mittelwerten

Verteilungsmaßzahlen sollen bestimmte Charakteristika einer Häufigkeitsverteilung hervorheben und durch eine Zahl kennzeichnen. Im Falle der Mittelwerte gilt es das Niveau zu charakterisieren, also die allgemeine Größenordnung der Messwerte, bzw. die "zentrale Tendenz" (Lage des Zentrums) einer Häufigkeitsverteilung durch Angabe eines "typischen Werts" (Stellvertreter-Bedeutung des Mittelwerts). Mittelwerte sind bei weitem die bekanntesten statistischen Größen, deren Aussage allgemeinverständlich ist: ein Mittelwert ist, wie schon der Name sagt, ein mittlerer Wert, der einen Datensatz durch eine einzige Zahl "zusammenfasst".

Mittelwerte sind nicht die einzigen, wohl aber die bekanntesten und elementarsten Verteilungsparameter. Andere Aspekte einer Häufigkeitsverteilung sind Streuung, Schiefe, Wölbung, Konzentration und Disparität. Zu den entsprechenden Verteilungsmaßzahlen vgl. Kap. 5 und 6. Diese Maßzahlen sollten bei der Interpretation eines Mittelwerts mitberücksichtigt werden. So ist z.B. ein Mittelwert, isoliert betrachtet, immer dann nicht sehr aussagefähig, wenn die Einzelwerte sehr unterschiedlich sind, d.h. die Streuung groß ist.

Def. 4.1: Mittelwertaxiome

Mittelwerte M sind Verteilungsmaßzahlen, die unter Berücksichtigung des Skalenniveaus die folgenden Axiome M1 bis M5 erfüllen:

M1 **Einschrankung**: Es gilt bei der Größe nach geordneten Einzelwerten $x_{(1)} \leq M \leq x_{(n)}$ bzw. bei Merkmalsausprägungen $x_1 \leq M \leq x_m$.

M2 **Ergänzung**: Tritt zu den n Beobachtungswerten $x_1,x_2,...,x_n$ mit dem Mittelwert $M(x_1,...,x_n) = M_n$ ein weiterer Wert x_{n+1} hinzu, so soll für den "neuen" Mittelwert $M(x_1,...,x_{n+1}) = M_{n+1}$ gelten:
wenn $x_{n+1} \le M_n$ dann $M_{n+1} \le M_n$
wenn $x_{n+1} \ge M_n$ dann $M_{n+1} \ge M_n$

M3 **Transformation**: Für den Mittelwert M^* der transformierten Beobachtungswerte $x_v^* = f(x_v)$ soll gelten: $M^* = f(M)$. Dabei ist f eine auf dem Skalenniveau des Merkmals X zulässige Transformation.

M4 **Monotonie**: Bei den Merkmalen X und Y mit den Beobachtungsvektoren (Vektoren der Beobachtungswerte) **x** und **y** soll die Mittelwertfunktion monoton zunehmen in bezug auf die Beobachtungswerte bzw. Merkmalsausprägungen.
Für $\mathbf{x} \ge \mathbf{y}$ (vgl. Bem. Nr. 4) gilt $M(\mathbf{x}) \ge M(\mathbf{y})$.

M5 **Unabhängigkeit von den absoluten Häufigkeiten**: Für ein reelles k und mit den Vektoren **x** der Merkmalsausprägungen und **n** der absoluten Häufigkeiten gilt $M(\mathbf{x},\mathbf{n}) = M(\mathbf{x},k\cdot\mathbf{n})$ (d.h. eine Ver-k-fachung der absoluten Häufigkeiten verändert den Mittelwert nicht).

Bemerkungen zu Def. 4.1:

1. Im Axiom M1 wird von einem Mittelwert gefordert, dass er zwischen dem kleinsten und dem größten Beobachtungswert (einschließlich) liegt. Das entspricht auch dem umgangssprachlichen Verständnis des Wortes "Mittel"- Wert. Aus M1 folgt, dass für $x_v = c$ (für alle v =1,...,n) M = c ist. Sind alle Beobachtungen gleich (d.h. gleich der Konstanten c), dann ist auch der Mittelwert gleich dieser Konstanten c und auch ganz offensichtlich der charakteristische Wert.

2. Die in M2 gestellte Forderung gewinnt an Bedeutung bei der Untersuchung der "Robustheit" einer Maßzahl. Ein Hilfsmittel hierzu kann die sog. Einflusskurve (influence curve, Inf(...)) sein:

   Def. 4.1a:

   $Inf(M_n,x_{n+1}) = (n+1)(M_{n+1} - M_n)$ [Einflusskurve]
   welche die Abhängigkeit der (n+1)-fachen Mittelwertdifferenz bei Hinzukommen einer (n+1)-ten Beobachtung (d.h. dem Merkmalswert x_{n+1}) von diesem Wert x_{n+1} beschreibt. Bei Darstellung der einzelnen Mittelwerte wird auf dieses Konzept zurückgegriffen.

3. Eine häufig betrachtete Transformation ist die Lineartransformation (Gl. 4.1). Axiom M3 bedeutet dann, dass aus

(4.1) $x_v^* = a + bx_v$ (b>0) folgt
(4.2) $M^* = a + bM$,

was eine problematische Forderung ist. Dies generell zu fordern [oder auch für andere, nichtlineare Transformationen], wie es P.J. Bickel und E.L. Lehmann[1] tun, halten wir nicht für sinnvoll, da Gl. 4.1 eine höchstens auf dem Intervallskalenniveau zulässige Transformation ist. Gl. 4.2 wird vom arithmetischen, nicht aber vom geometrischen und harmonischen Mittel erfüllt.
Andererseits wäre es nicht sinnvoll, wenn ein Mittelwert translationsinvariant (a hätte keinen Einfluss auf M^*) oder skaleninvariant (b hätte keinen Einfluß auf M^*) wäre.
In einer weiteren Bedingung fordern Bickel und Lehmann übrigens (4.3) $M(-\mathbf{x}) = -M(\mathbf{x})$ [entspricht a = 0 und b = -1 in Gl. 4.2], d.h. ändern alle Beobachtungswerte ihr Vorzeichen, so soll auch der Mittelwert sein Vorzeichen ändern. Als Mittelwerte sollten jedoch auch solche Kennzahlen zugelassen werden, die nur sinnvoll bei positiven Merkmalswerten zu berechnen sind, wie z.B. das geometrische Mittel.

4. Unter $\mathbf{x} \geq \mathbf{y}$ verstehen wir, dass $x_v \geq y_v$ (v = 1,...,n) ist, wobei für mindestens ein v $x_v > y_v$ gilt. Für die im Abschnitt b) dargestellten Mittelwerte gilt strenge Monotonie, d.h. es gilt unter den genannten Voraussetzungen $M(\mathbf{x}) > M(\mathbf{y})$. Dies sind Mittelwerte, die in bezug auf Unterschiedlichkeit der Merkmalswerte reagibler sind. Da ein Lokalisationsparameter das Niveau der Messwerte widerspiegeln soll, wäre es nicht sinnvoll, Mittelwerte zuzulassen, für die M4 nicht gilt. Denn das würde bedeuten, dass obgleich kein x-Wert kleiner als ein korrespondierender y-Wert ist, der Mittelwert der x-Werte kleiner als der Mittelwert der y-Werte sein kann.
Für den Fall, dass **y** aus **x** durch eine zulässige Transformation hervorgeht, ist mit M3 auch M4 erfüllt.

5. Mit M5 wird sichergestellt, dass ein Mittelwert nicht abhängig ist von der Anzahl der Beobachtungen. Es ist daher unerheblich, ob ein Mittelwert mit absoluten oder mit relativen Häufigkeiten berechnet wird.

6. Es mag plausibel erscheinen, dass im Falle einer symmetrischen Verteilung ein "Mittelwert" im "Zentrum der Symmetrie" liegt. Obgleich diese Forderung vom arithmetischen Mittel, Median (Zentralwert) und Modus erfüllt wird, ist es nicht sinnvoll, diese Eigenschaft generell zu fordern. Denn dann - sofern nicht alle Beob-

[1] Descriptive Statistics for Nonparametric Models, Part II: Location, in: The Annals of Statistics Vol 3 (1975), S.1045 - 1069.

achtungswerte identisch sind - liegen das geometrische und harmonische Mittel links vom Zentrum.

2. Spezielle Mittelwerte für metrisch skalierte Merkmale

In diesem und im folgenden Abschnitt werden die gebräuchlichsten Mittelwerte und ihre Eigenschaften vorgestellt. Das arithmetische, geometrische und harmonische Mittel erfordern eine metrische Skala, der Median mindestens eine Ordinalskala und der Modus nur eine Nominalskala. Einige weitere Mittelwerte werden im Zusammenhang mit speziellen Fragestellungen in anderen Kapiteln behandelt[2]. Wie gezeigt werden kann, erfüllen alle in diesem Abschnitt behandelten Mittelwerte die Axiome M1 bis M5.

a) Das arithmetische Mittel

Def. 4.2: arithmetisches Mittel

Die Maßzahl

(4.4)	$\bar{x} = \frac{1}{n} \sum x_v$	(Berechnung aus Einzelbeobachtungen) [**ungewogenes arithmetisches Mittel**]

oder

(4.5)	$\bar{x} = \frac{1}{n} \sum x_i n_i$	(Berechnung aus Merkmalsausprägungen)
	$= \sum x_i h_i$	[**gewogenes arithmetisches Mittel**]

heißt arithmetisches Mittel.

Hinweis:
Die Namen "gewogenes" und "ungewogenes" arithmetisches Mittel sollten nicht den Eindruck entstehen lassen, dass es sich um zwei verschiedene Mittelwerte handelt. Es sind nur Bezeichnungen für zwei Arten der Berechnung des gleichen arithmetischen Mittels, je nachdem, in welcher Form die Daten gegeben sind: einmal als Einzelbeobachtungen, zum anderen als gruppierte Daten (d.h. wenn Daten als Häufigkeitsverteilung vorliegen). Der Fall klassierter Daten wird im folgenden behandelt unter Nr.6.

---

2 Etwa der schwerste Wert und der Scheidewert im Kapitel 6.

Eigenschaften und Interpretation des arithmetischen Mittels:

1. Man kann leicht zeigen, dass $\bar{x}$ die Axiome M1 bis M5 erfüllt.
2. Schwerpunkt- oder Ausgleichseigenschaft

Satz 4.1: Schwerpunkteigenschaft

Es gilt

a) $\sum x_v - \bar{x} = 0$ bzw.

b) $\sum(x_i - \bar{x})h_i = 0$.

Beweis

a) $\Sigma(x_v - \bar{x}) = \Sigma x_v - n\,\bar{x} = 0$ da wegen (4.4) $\Sigma x_v = n\bar{x}$ gilt.

b) ist analog zu zeigen (es ist auch $\Sigma(x_i - \bar{x})\, n_i = 0$).

Beispiel 4.1 macht deutlich, warum hier vom "Schwerpunkt" gesprochen wird (vgl. Abb. 4.1). Die Summen der negativen (S_1) und positiven Abweichungen (S_2) der Merkmalswerte vom arithmetischen Mittel sind absolut gleich, nämlich $S_1 = -27$ und $S_2 = +27$.
Der **Schwerpunkt** ist bekanntlich der Quotient aus der Summe der "Momente", d.h. der Produkte aus Gewichten n_i und Hebelarmen x_i, also $\Sigma x_i\, n_i$ und der Summe der Gewichte Σn_i. In Abb. 4.1 erscheinen die Größen n_i als Gewichte (Symbol •) an einer Balkenwaage. Wird diese im Punkt $\bar{x} = 7$ unterstützt, so befindet sie sich im Gleichgewicht. Die Gleichheit von S_1 und S_2 bedeutet, dass sich negative und positive Abweichungen ausgleichen (**Ausgleichseigenschaft** des arithmetischen Mittels): Angenommen die Beträge 0,5,10,15 und 20 seien Einkommensbeträge (z.B. in 100 DM), dann wäre 7 (also 700 DM) genau der Betrag, den alle verdienen würden, wenn der Gesamtverdienst (die Summe aller Verdienste) auf alle n Einheiten gleich verteilt würde, d.h. wenn alle gleich viel verdienen würden, und S_1 und S_2 wären die Summen der unter-, bzw. der überdurchschnittlichen Einkommen.
Wird die **Merkmalssumme** $S = \Sigma x_i\, n_i$ auf n Einheiten (im Bsp. 4.1 ist $S = 70$ und $n = 10$) zu gleichen Teilen (jeweils 1/n-tel von S) verteilt, so erhält jede Einheit den Betrag $\bar{x} = S/n$. Das ist die Ausgleichs- oder Ersatzwerteigenschaft des arithmetischen Mittels. Folgerung:

> Das arithmetische Mittel bleibt unverändert bei solchen "Umverteilungen" zwischen den Merkmalswerten, bei denen die Merkmals**summe** konstant bleibt.

3. Minimumeigenschaft

Satz 4.2:

Die Funktion $Q(M) = \sum (x_v - M)^2$ besitzt ein Minimum an der Stelle $M = \bar{x}$, d.h. für alle $M \neq \bar{x}$ ist: $\sum (x_v - M)^2 > \sum (x_v - \bar{x})^2$.

Beweis:

$$\text{Aus} \quad \frac{dQ(M)}{dM} = 2 \sum (x_v - M)(-1) = 0$$

folgt $\Sigma x_v - nM = 0$ und somit $M = (\sum x_v)/n = \bar{x}$.
Da die zweite Ableitung nach M positiv ist, nämlich $dQ^2(M)/dM^2 = 2n > 0$, ist die Behauptung bewiesen. Die Summe der quadrierten Abweichungen hat kein Maximum [sie kann also beliebig groß sein], sondern nur ein Minimum.

Anmerkungen zum Beweis:

1. Der Beweis lässt sich auch mit dem Steinerschen Verschiebungssatz $\Sigma (x_v-a)^2 = \Sigma (x_v-\bar{x}) + n(\bar{x}- a)^2$ führen [Kap.5].
2. Wie man aus der notwendigen Bedingung $-2 \Sigma (x_v - M) = 0$ sieht, folgt die Schwerpunkteigenschaft aus der Minimumeigenschaft.

Interpretation der Minimumeigenschaft:

Deutet man x_v als Summe einer (systematischen) Niveaukomponente m, die additiv überlagert ist von einer Fehlerkomponente u_v, also $x_v = m + u_v$, so ist $\bar{x} = m$ derjenige Wert der Niveaukomponente, um den die Abweichungen bzw. Fehler $u_v = x_v - m$ geringst möglich streuen (vgl. Varianz, Kapitel 5).

4. Lineartransformation, Voraussetzungen hinsichtlich der Skalen

Satz 4.3:

Das arithmetische Mittel erfüllt das Mittelwertaxiom M3 für lineare Transformationen. Aus Gl. 4.1 folgt in Verbindung mit Def. 4.2 für das arithmetische Mittel der transformierten Werte
$x_v^* = a + b\, x_v$:

(4.6)	$\bar{x}^* = a + b\bar{x}$	(a,b reelle Zahlen).

Der Beweis ist elementar.
Damit ist auch gezeigt, dass das arithmetische Mittel bei Merkmalen, die auf einer Intervallskala [die ja invariant ist gegenüber linearen Transformationen] gemessen sind, anwendbar ist.

Sofern b > 0 ist, hat das Merkmal X^* die selben Skaleneigenschaften wie X. Somit ist die Lineartransformation eine "zulässige" Transformation.
Eine häufig benutzte Transformation ist die **Zentrierung**, d.h. die Bildung von Abweichungen vom arithmetischen Mittel, "zentrierter" Werte (sog. "deviation scores"):

$$z_v = x_v - \bar{x} \qquad (a = -\bar{x}\,,\ b = 1),$$

deren arithmetisches Mittel wegen Satz 4.3 Null ist.

5. Aggregationseigenschaft

Der folgende Satz läßt sich verallgemeinern für alle Mittelwerte, die ein Spezialfall des Potenzmittels sind.

Satz 4.4:

Wird eine Gesamtmasse (Gesamtheit) vom Umfang n zerlegt in g disjunkte Teilmassen (Teilgesamtheiten) mit den Umfängen n_1, n_2,...,n_g, so dass $n = n_1 + n_2 + ... + n_g$, dann ist das arithmetische Mittel
mit

$$(4.7) \qquad \bar{x} = \sum_{j=1}^{g} \bar{x}_j h_j \qquad (j = 1,2,...,g)$$

gegeben, wobei $\bar{x}_j$ das arithmetische Mittel der j-ten Teilgesamtheit ist:

$$(4.8) \qquad \bar{x}_j = \frac{1}{n_j} \sum_k x_{jk} \ . \qquad (k = 1,2,...,n_j),$$

Dabei ist $h_j = n_j / n$ der Anteil der j-ten Teilgesamtheit am Gesamtumfang n, so dass der Gesamtmittelwert ein gewogenes arithmetisches Mittel der Teilmittelwerte ist.

Beweis

Mit S_j als Merkmalssumme der j-ten Teilgesamtheit $S_j = \sum_k x_{jk} = n_j \bar{x}_j$

ist die Merkmalssumme S der Gesamtmasse mit $S = \sum_j S_j = \sum_j \sum_k x_{jk}$

bzw. unter Berücksichtigung von Gl. 4.8 durch $S = \sum_j n_j \bar{x}_j$ gegeben.

Wegen $\bar{x} = S/n$ folgt hieraus unmittelbar Gl. 4.7.

6. Klassierte Daten

Sofern die **Klassenmittelwerte** $\bar{x}_k$ (k = 1,2,...,p) bekannt sind, berechnet man den Gesamtmittelwert $\bar{x}$ gem. Gl. 4.9:

$$(4.9) \quad \bar{x} = \sum_k \bar{x}_k h_k \qquad 1 \le k \le p \ .$$

Andernfalls verwendet man die **Klassenmitten** m_k (vgl. Kap. 3, Def. 3.5 b) und erhält den **geschätzten** Gesamtmittelwert m (als Schätzung von $\bar{x}$) mit:

$$(4.10) \quad m = \hat{\bar{x}} = \sum_k m_k h_k$$

Im allgemeinen wird m von $\bar{x}$ verschieden sein. Die Näherung wird um so besser sein, je mehr sich die Beobachtungswerte (symmetrisch) um die Klassenmitten m_k verteilen.

Klassierte Daten stellen eine spezielle Form der in Satz 4.4 beschriebenen Zerlegung dar. Die Teilgesamtheiten sind aufgrund von aneinander angrenzenden Intervallen auf der x-Achse definiert (es ist dann g = p). Teilmassen können auch aufgrund anderer Kriterien (statt des x-Wertes) gebildet werden, wie z.B. nach dem Merkmal Geschlecht, Religion usw. Vgl. Beispiel 4.2 für eine Demonstration der Schätzung eines arithmetischen Mittels aus einer klassierten Verteilung.

Beispiel 4.1:

Man berechne das arithmetische Mittel für:

a) die folgenden 10 Merkmalswerte 0,5,0,10,15,5,0,20,10,5;
b) die folgende Häufigkeitsverteilung:

i	x_i	n_i
1	0	3
2	5	3
3	10	2
4	15	1
5	20	1

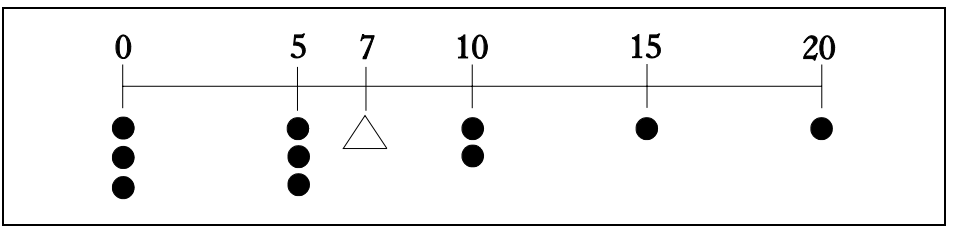
Lösung 4.1:

Es ist unschwer zu erkennen, dass es sich in beiden Fällen um die gleichen Daten handelt. Für das ungewogene arithmetische Mittel erhält man (0+5+0+10+15+5+0+20+10+5)/10 = 70/10 = 7 und für das gewogene arithmetische Mittel natürlich den gleichen Wert 7. Man berechnet das gewogene arithmetische Mittel am besten aus einer Arbeitstabelle, in der man in einer Spalte die Größen $x_i n_i$ bestimmt, deren Summe dann der Zähler von $\bar{x}$ ist (vgl. Abb. 4.1).

Abb. 4.1: Erläuterung des Begriffs "Gewicht" und der Schwerpunkteigenschaft von $\bar{x}$

i	x_i	n_i	$x_i n_i$	$x_i-\bar{x}$	$(x_i-\bar{x})n_i$
1	0	3	0	-7	-21
2	5	3	15	-2	-6
3	10	2	20	+3	+6
4	15	1	15	+8	+8
5	20	1	20	+13	+13
S	-	10	70	-	0

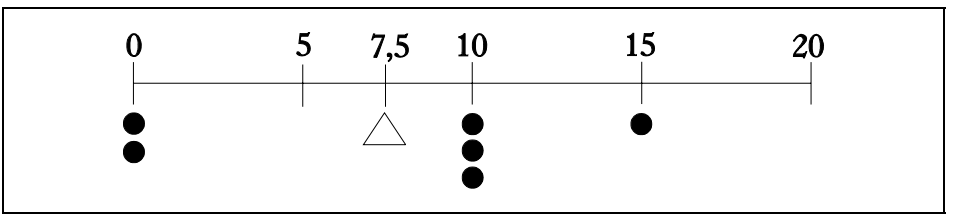
$\bar{x} = 70/10 = 7$



Variante von Bsp. 4.1: vgl. folgende Abbildung der Balkenwaage

i	x_i	n_i	$x_i n_i$
1	0	2	0
2	10	3	30
3	15	1	15

$\bar{x} = 45/6 = 7{,}5$ Im Punkt 7,5 befindet sich die Balkenwaage im Gleichgewicht.



Beispiel 4.2:

Berechnen Sie das arithmetische Mittel der Bruttomonatsverdienste von 150 Beschäftigten einer Unternehmung gemäß der folgenden Tabelle:

Klasse Nr. k	Bruttomonats-verdienst	Anzahl n_k
1	600 - 1000	4
2	1000 - 1400	11
3	1400 - 1800	19
4	1800 - 2200	31

Klasse Nr. k	Bruttomonats-verdienst	Anzahl n_k
5	2200 - 2600	37
6	2600 - 3000	22
7	3000 - 3400	18
8	3400 - 3800	8

Lösung 4.2:

Da die Klassenmittelwerte nicht bekannt sind, wird hier über die Klassenmitten m_k der Gesamtmittelwert m nach der Formel $m = \sum m_k h_k$ berechnet.

Ergebnis: m = 2305,60 DM (= durchschnittl. Bruttomonatsverdienst).

k	m_k	h_k	$m_k\, h_k$
1	800	0,027	21,6
2	1200	0,073	87,6
3	1600	0,127	203,2
4	2000	0,207	414,0
5	2400	0,247	592,8
6	2800	0,147	411,6
7	3200	0,120	384,0
8	3600	0,053	190,8
S			2305,6

Mittelwerte und Interpolation:

Das gewogene arithmetische Mittel von zwei Zahlen x_1 und x_2 (oder y_1 und y_2) stellt eine lineare Interpolation dar (Abb. 4.2).

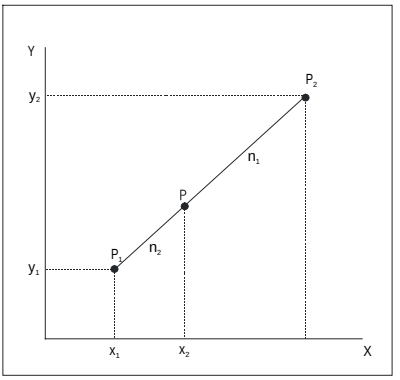
Der Punkt $P(\bar{x},\bar{y})$ teilt die Strecke P_1P_2 zwischen den Punkten $P_1(x_1,y_1)$ und $P_2(x_2,y_2)$ im Verhältnis n_2/n_1 auf (vgl. Abb. 4.2, Teil a), denn

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2}{n_1 + n_2} \qquad \text{und} \qquad \bar{y} = \frac{n_1 y_1 + n_2 y_2}{n_1 + n_2}$$

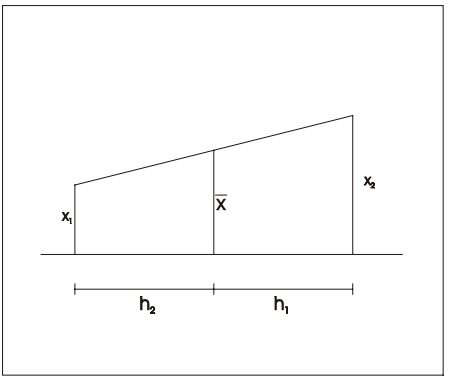
Abb. 4.2, Teil b zeigt: $\bar{x}$ ist die lineare Interpolation zwischen den Höhen x_1 und x_2, denn $(x_2-x_1)/(h_1+h_2) = (\bar{x}-x_1)/h_2$ und $h_1 + h_2 = 1$.

Abb. 4.2: Lineare Interpolation

a) eine zweidimensionale Betrachtung (Variablen X,Y)



b) eindimensionale Betrachtung (h_1, h_2 als relative Häufigkeit)



"Mittelwerte der Pythagoräer"

Das ungewogene arithmetische Mittel zweier Zahlen x_1 , x_2 basiert auf der folgenden Forderung bezüglich des Verhältnisses zweier Abstände

$$\frac{x_2 - M}{M - x_1} = \frac{x_2}{x_2} = \frac{x_1 - M}{M - x_2} = \frac{x_1}{x_1} = 1 \quad \Rightarrow \quad M = \bar{x} = \frac{x_1 + x_2}{2}$$

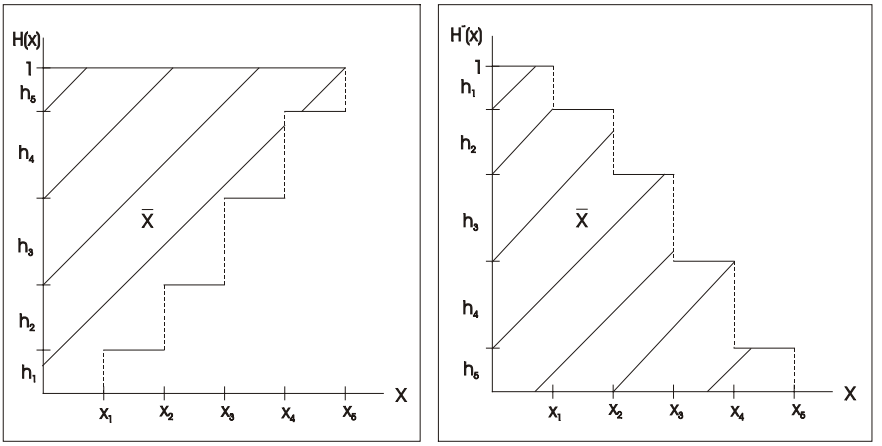
d.h. durch $\bar{x}$ wird die zwischen x_2 und x_1 gebildete Strecke genau halbiert.

Mit entsprechenden Forderungen an Verhältnisse von Abständen (Strecken) wurden in der Antike die sog. "Mittelwerte der Pythagoräer", nämlich $\bar{x}$, $\bar{x}_G$, $\bar{x}_H$ und $\bar{x}_A$ hergeleitet (vgl. Bemerkungen zu den entsprechenden Mittelwerten).

arithmetisches Mittel und Summenhäufigkeitsfunktion:

Es ist leicht zu sehen, dass bei einem nichtnegativen Merkmal X das arithmetische Mittel die Fläche zwischen der Ordinate (Häufigkeitsachse) und der Summenhäufigkeitskurve H(x) ist (d.h. die schraffierte Fläche in Abb. 4.3), bzw. die Fläche unterhalb der Resthäufigkeitskurve $H^-(x)$.

Abb.4.3: Das arithmetische Mittel als Fläche



b) Das geometrische Mittel

Das geometrische Mittel ist weit weniger bekannt und gebräuchlich als das arithmetische Mittel. Es wird vor allem im Zusammenhang mit Wachstumsfaktoren (vgl. Kap.9) benötigt.

Def. 4.3: geometrisches Mittel

Die Maßzahl

$$(4.11) \quad \bar{x}_G = \left(\prod_{v=1}^{n} x_v \right)^{1/n} \qquad \text{(bei Einzelbeobachtungen, "ungewogen"),}$$

(das Produktzeichen P bedeutet $Px_v = x_1 x_2 ... x_n$), bzw.

$$(4.12) \quad \bar{x}_G = \prod_{i=1}^{m} x_i^{h_i} \qquad \text{(gruppierte Daten, "gewogen")}$$

heißt geometrisches Mittel (der positiven Merkmalswerte $x > 0$).

Beispiel 4.3:

Man berechne das geometrische Mittel für

a) die folgenden Beobachtungen: 5,5,5,10,10,15,20,20;
b) die folgende Häufigkeitsverteilung:

i	x_i	n_i
1	5	3
2	10	2
3	15	1
4	20	2

($n = \Sigma n_i = 8$).

Lösung 4.3:

Es ist offensichtlich, dass es sich bei diesem Zahlenbeispiel in beiden Fällen um die gleichen Daten handelt. Man erhält:

für das ungewogene geometrische Mittel $(5\cdot5\cdot5\cdot10\cdot10\cdot15\cdot20\cdot20)^{1/8} = (75.000.000)^{1/8} = 9{,}64679$

für das gewogene geometrische Mittel natürlich den gleichen Wert, nämlich $(5^3 10^2 15^1 20^2)^{1/8} = 5^{3/8} 10^{2/8} 15^{1/8} 20^{2/8}$. Das arithmetische Mittel ist größer, nämlich 90/8 = 11,25.

Eigenschaften und Interpretation des geometrischen Mittels:

1. Das geometrische Mittel erfüllt die Axiome M1 bis M5.

2. Aus Def. 4.3 folgt unmittelbar

$$(4.13)\quad \log \bar{x}_G = \frac{1}{n}\sum \log(x_v) = \overline{\log(x)}$$

   und entsprechend bei gruppierten Daten, so dass der Logarithmus des geometrischen Mittels gleich ist dem arithmetischen Mittel der logarithmierten Merkmalswerte. Das geometrische Mittel wird deshalb auch logarithmisches Mittel genannt.

3. Die der Schwerpunkteigenschaft des arithmetischen Mittels entsprechende Eigenschaft des geometrischen Mittels ist die folgende Eigenschaft, die auch gelegentlich "Einseigenschaft" genannt wird, d.h. der Ausgleich relativer Größen (im Verhältnis zu $\bar{x}_G$). Es gilt:

   a) $P(x_v / \bar{x}_G) = (\bar{x}_G)^{-n} P x_v = (\bar{x}_G)^{-n} (\bar{x}_G)^n = 1$, bzw.
   b) $P(x_i / \bar{x}_G)^{h_i} = 1$.

   Hierauf beruht die Eignung von $\bar{x}_G$ zur Mittelung von Wachstumsfaktoren (vgl. Kap. 9). Dies ist die wichtigste Anwendung des geometrischen Mittels in der Statistik. $\bar{x}_G$ hat für Vervielfachungsgrößen Be-

deutung und immer dann Sinn, wenn das Produkt der Einzelwerte eine sinnvolle Größe ist.

4. Während $\overline{x}$ invariant ist gegenüber "Umverteilungen" dergestalt, dass die **Merkmalssumme** konstant ist, gilt dies für $\overline{x}_G$ hinsichtlich des **Produkts** der Merkmalswerte. Eine Verdoppelung (Erhöhung um 100%) von x_1 wird in diesem Sinne durch eine Halbierung (Senkung um 50%) von x_2 "ausgeglichen", so dass das ungewogene $\overline{x}_G$ unverändert bleibt.

   Man beachte:
   In diesem Beispiel hat sich x_1 um 100% erhöht und x_2 um 50% verringert. Ein Ausgleich findet mithin zwischen den Wachstumsfaktoren, nicht zwischen den Wachstumsraten statt. Nicht richtig wäre somit z.B., dass sich eine 10%ige Zunahme durch eine 10%ige Abnahme ausgleiche (Beispiel 4.4).

5. In Abb. 4.4 liegen alle Kombinationen von x_1 und x_2, die zu dem gleichen Wert des geometrischen Mittels $\overline{x}_G = \sqrt{x_1 x_2}$ führen auf der eingezeichneten Hyperbel. Entsprechend ist die Gerade AB der geometrische Ort aller Kombinationen x_1,x_2, die zum gleichen arithmetischen Mittel führen. Die Abb. 4.4 zeigt auch, dass stets gilt $\overline{x}_G \leq \overline{x}$, wobei $\overline{x}_G = \overline{x}$ genau dann gilt, wenn $x_1 = x_2$. Der gleiche Zusammenhang wird auch mit dem Höhensatz der Planimetrie gezeigt (Abb. 4.5) wonach $x_1 x_2 \leq \overline{x}^2$ ist.

6. Das geometrische Mittel kommt in der Ökonomie in Gestalt der Cobb-Douglas-Produktionsfunktion vor. Das Produktionsergebnis ist danach ein gewogenes geometrisches Mittel der Einsatzmengen der Produktionsfaktoren. Der allgemeineren CES-Funktion entspricht das Potenzmittel.

Abb. 4.4

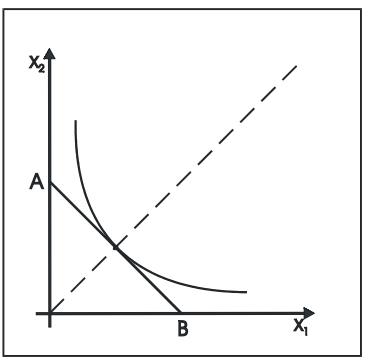
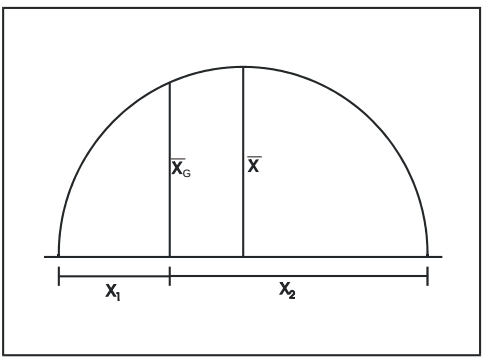


Abb. 4.5



7. Minimumeigenschaft

Satz 4.5:

Die Funktion $Q^*(M) = \sum[\log(x_v/M)]^2$ besitzt ein Minimum an der Stelle $M = \bar{x}_G$.

Beweis : Da $Q^*(M) = S\ [\log(x_v) - \log(M)]^2$ kann der Beweis analog Satz 4.2 geführt werden. Das Minimum ist an der Stelle $\log(M) = \overline{\log(x)}$, was nach Gl. 4.13 nichts anderes ist als $\log \bar{x}_G$. Es mag von Interesse sein, die Minimumeigenschaft des arithmetischen und des geometrischen Mittels an einem Beispiel zu vergleichen (vgl. Beispiel 4.5 und Abb. 4.6).

8. Deutet man x_v als ein Produkt einer (systematischen) Niveaukomponente m, die multiplikativ überlagert ist von einer Fehlerkomponente u_v, also $x_v = mu_v$, so ist $\bar{x}_G = m$ derjenige Wert der Niveaukomponente, um den die relativen Fehler $u_v = x_v/m$ geringst möglich streuen (vgl. Varianz, Kap. 5).

9. Zulässige Transformationen (erforderliche Skalen)

$\bar{x}_G$ erfüllt M3 für proportionale (linear-homogene) Transformationen. Aus $x_v^* = bx_v$ folgt

$$\bar{x}_G^* = \sqrt[n]{\prod x_v^*} = \sqrt[n]{\prod(bx_v)} = (b^n \prod x_v)^{1/n} = b\bar{x}_G.$$

Damit ist auch gezeigt, dass das geometrische Mittel nur bei Merkmalen, die mindestens auf einer Ratioskala gemessen sind, anwendbar ist.

10. Aggregation und klassierte Verteilung

Wird - wie in Satz 4.4 - eine Gesamtmasse (Gesamteinheit) vom Umfang n zerlegt in g disjunkte Teilmassen (Teilgesamtheiten), so ist $\bar{x}_G$ das mit den relativen Häufigkeiten h_j gewogene geometrische Mittel der geometrischen Mittel $\bar{x}_{Gj}$ der Teilgesamtheiten. Hieraus folgt für eine klassierte Verteilung:
Bei Kenntnis der geometrischen Mittel $\bar{x}_{Gk}$ (k = 1,2,...,p) der Klassen ist $\bar{x}_G$ als gewogenes geometrisches Mittel der (geometrischen) Klassenmittelwerte zu berechnen. Verwendet man statt dessen die Klassenmitten m_k, so kann $\bar{x}_G$ über- oder unterschätzt werden.

Die drei folgenden Eigenschaften sind für die Interpretation des geometrischen Mittels nicht sehr ergiebig, sie seien aber der Vollständigkeit halber aufgeführt:

11. Ersatzwerteigenschaft
    Ähnlich wie das arithmetische Mittel kann auch das geometrische Mittel $\overline{x}_G$ in einem bestimmten Sinn als Ersatzwert aufgefaßt werden. Die **Ersatzwert- oder Hochrechnungseigenschaft** von $\overline{x}_G$ bezieht sich auf das Gesamtmerkmalsprodukt Px_v mit $\overline{x}_G^n = Px_v$.

12. Das gewogene arithmetische Mittel zweier Zahlen x_1, x_2 stellt eine Interpolation gemäß einer Exponentialfunktion dar: $\overline{x}_G = x_1^{1-a}\ x_2^a = x_1(x_2/x_1)^a \quad 0 \leq a \leq 1$ mit $a = h_2$ im Unterschied zur linearen Interpolation $x_1 + a(x_2 - x_1)$ beim arithmetischen Mittel.

13. Das ungewogene geometrische Mittel zweier Zahlen x_1, x_2 basiert auf folgender Forderung bezüglich des Verhältnisses zweier Abstände:

$$\frac{x_2}{M} = \frac{M}{x_1} = \frac{x_2 - M}{M - x_1} \quad \text{so dass } M = \overline{x}_G \text{ denn } \overline{x}_G / x_1 = x_2 / \overline{x}_G.$$

Beispiel 4.4:

Diplom-Kaufmann K aus E erhält im Jahre 1989 eine Gehaltserhöhung um 20%. Wegen der schlechten Geschäftslage im Jahre 1990 muss er jedoch 1990 eine Gehaltssenkung um 20% hinnehmen. Er verdient jetzt (Richtiges ankreuzen):

☐ weniger als ☐ mehr als ☐ genausoviel wie vor der Gehaltserhöhung.

Lösung 4.4:

Nach einer einfachen Überlegung wird man „weniger“ ankreuzen, weil ja die Gehaltssenkung von 20% auf der Basis des gestiegenen Gehalts einen höheren Betrag ausmacht, als die vorhergehende Gehaltserhöhung. Es ergibt sich nömlich, wenn man einmal von DM 3000 ausgeht: Gehalt vor 1989: DM 3000, nach der Erhöhung: $3000 \cdot 1{,}2 = 3600$ DM und nach der Gehaltssenkung: $3600 \cdot 0{,}8 = 2880$ DM < 3000 DM.

Hinweis auf Kapitel 9:

Das Beispiel zeigt auch, warum es nicht sinnvoll ist, Wachstumsraten arithmetisch zu mitteln. Danach ergäbe sich als mittlere jährliche Wachstumsrate $x = ½[+20\% + (-20\%)] = 0\%$, was falsch ist, weil K ja nicht genauso viel verdient, wie vor der Gehaltserhöhung.

Der richtige Ansatz ist das geometrische Mittel der Wachstumsfaktoren $\sqrt{1,2 \cdot 0,8} = \sqrt{0,96} = 0,9798 = 1 - 0,0202$. Hätte K jedes Jahr eine (konstante) Gehaltssenkung von 2,02% "erlitten", so wäre er zum gleichen Gehalt von DM 2880 gelangt, denn $3000 \cdot 0,9798 \cdot 0,9798 = 2880$. Die mittlere jährliche Wachstumsrate beträgt also - 0,0202, d.h. -2,02%.

Beispiel 4.5:

Man bestimme und zeichne die Funktionen Q (gem. Satz 4.2) und Q^* (gem. Satz 4.5) für die folgenden Daten
$x_1 = 10$, $x_2 = 15$ und $x_3 = 20$
und für die folgenden Werte von M: 10,12,14,15,16,18,20.

Lösung 4.5:

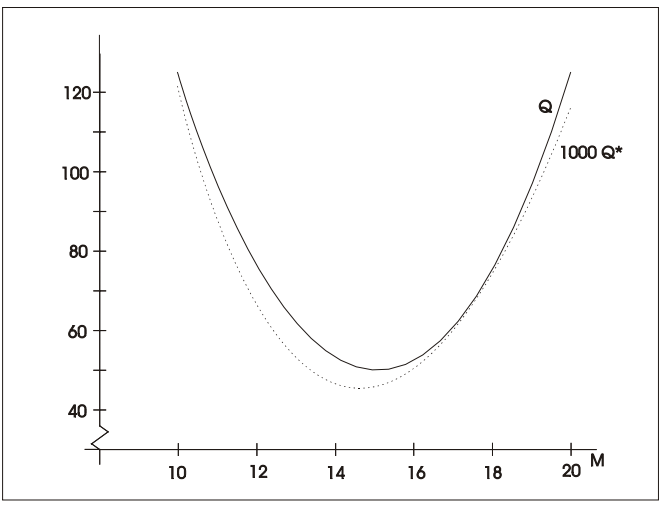
M	Q	1000 Q^*
10	125	121,6
12	77	64,9
14	53	46,2
14,42	51	45,7*)
15	50*)	46,6
16	53	51,8
18	77	73,5
20	125	106,2

*)Minimum

Die Mittelwerte sind $\bar{x} = 15$ und $\bar{x}_G = 14,4224$.
Die genannten Funktionen lauten:
$Q = Q(M) = s\,(x_v - M)^2$
$Q^* = Q^*(M) = s\,[\log(x_v/M)]^2$

Man erkennt, dass die Funktion Q ihr Minimum an der Stelle M = 15 und die Funktion Q^* ihr Minimum an der Stelle M = 14,42 hat. Q ist symmetrisch um 15, Q^* ist aber nicht symmetrisch um 14,42.

Abb. 4.6



c) Das harmonische Mittel

Das harmonische Mittel hat in der Praxis eine ziemlich geringe Bedeutung. Es gibt jedoch Fälle, bei denen dieser weitgehend unbekannte Mittelwert angewendet werden *muss*, weil jede andere Art der Mittelwertbildung zu unsinnigen Ergebnissen führen würde. Das bekannteste Beispiel ist die Mittelung von Geschwindigkeiten, bzw. allgemeiner von Beziehungszahlen (Kap. 9). Das harmonische Mittel spielt auch eine Rolle bei den Indexzahlen (Kap. 10).

Def. 4.4: harmonisches Mittel

Die Maßzahl

(4.14) $\bar{x}_H = \dfrac{n}{\sum \dfrac{1}{x_v}}$ bei Einzelbeobachtungen, "ungewogen"

(4.15) $\bar{x}_H = \dfrac{n}{\sum \dfrac{n_i}{x_i}} = \dfrac{1}{\sum \dfrac{h_i}{x_i}}$ bei gruppierten Daten, "gewogen"

heisst harmonisches Mittel ($x \neq 0$).

Beispiel 4.6: Mittelung von Geschwindigkeiten

Ein Flugzeug legt für den Flug von A nach B und zurück insgesamt 4800 km zurück. Aufgrund von Gegenwind kann das Flugzeug auf dem Hinweg nur eine Geschwindigkeit von 600 km/h erreichen, auf dem Rückweg jedoch eine Geschwindigkeit von 800 km/h. Mit welcher Durchschnittsgeschwindigkeit ist das Flugzeug unterwegs?

Lösung 4.6:

Die Durchschnittsgeschwindigkeit wird mit dem harmonischen Mittel berechnet:

$$\bar{x}_H = \frac{n}{\sum \frac{n_i}{x_i}} = \frac{4800}{\frac{2400}{600} + \frac{2400}{800}} = \frac{2}{\frac{1}{600} + \frac{1}{800}} = 685{,}71$$

In diesem Fall ist das gewogene harmonische Mittel gleich dem ungewogenen, weil die beiden Strecken gleich lang sind.

Eine nach dem arithmetischen Mittel berechnete Durchschnittsgeschwindigkeit von 700 km/h wäre nicht richtig gewesen, denn das Flugzeug benötigt für die 2400 km des Hinflugs 4 Std. und für die 2400 km des Rückflugs 3 Std. Somit war das Flugzeug 7 Std. unterwegs um 4800 km zurückzulegen, woraus eine Durchschnittsgeschwindigkeit von 685,71 km/h folgt.

Eigenschaften und Interpretation des harmonischen Mittels:

1. Wie man leicht sieht, gilt:

   der reziproke Wert von $\overline{x}_H$ ist das arithmetische Mittel der reziproken Werte (also der Werte $1/x_v$).

   Für die Berechnung ist es deshalb sinnvoll, zunächst den reziproken Wert $1/\overline{x}_H$ zu berechnen.

2. Man kann zeigen, dass $\overline{x}_H$ als Spezialfall des Potenzmittels alle Axiome M1 bis M5 erfüllt.

3. Das harmonische Mittel wird stets dann sinnvoll angewendet, wenn es gilt, Verhältniszahlen (siehe Kap. 9) zu mitteln, bei denen die im **Zähler** stehende Größe eine Konstante c ist und die im Nenner stehende Größe variabel ist. Das harmonische Mittel der Größen c/x_1, c/x_2,..., c/x_n beträgt nämlich $c/\overline{x}$, was sinnvoll interpretierbar ist. Entsprechend ist das arithmetische Mittel immer dann anzuwenden, wenn die im **Nenner** stehende Größe eine Konstante ist (vgl. Übers. 4.1).

Übersicht 4.1: Anwendung des arithmetischen und des harmonischen Mittels

Es ist anzuwenden das	
harmonische Mittel	**arithmetische Mittel**
bei einer Verhältniszahl) mit*	
$\frac{c}{x_v}$	$\frac{x_v}{c}$
Z konstant, N variabel	**Z variabel, N konstant**
bzw. bei einer Transformation von x_v in	
$y_v = \frac{c}{x_v}$	$z_v = \frac{x_v}{c} = \frac{1}{c} x_v$
(nichtlineare Transformation)	**(lineare Transformation)**

*) Eine Verhältniszahl (vgl. Kap. 9) ist ein Quotient mit einem Zähler Z und Nenner N. Die Größe c bezeichnet hier eine Konstante.

Zu welchen Ergebnissen eine arithmetische oder harmonische Mittelung führt sei an einem Beispiel mit n = 3 Beobachtungen gezeigt:

	Art der Mittelung	
	arithmetisch	harmonisch
y_v	$\bar{y} = (c/x_1 + c/x_2 + c/x_3)/3$	$\bar{y}_H = c/\bar{x}_H$ [sinnvoll]
z_v	$\bar{z} = \bar{x}/c$ [sinnvoll]	$\bar{z}_H = 3/(c/x_1 + c/x_2 + c/x_3)$

Man erkennt, dass die Größen $\bar{y}_H$ und $\bar{z}$ sinnvolle Größen sind, die Mittelwerte $\bar{z}_H$ und $\bar{y}$ dagegen keine sinnvollen Größen sind. Im speziellen Fall von n = 2 Einzelwerten gilt übrigens folgender Zusammenhang zwischen dem arithmetischen, harmonischen und geometrischen Mittel:

(4.16) $\bar{y} = ½(c/x_1 + c/x_2) = c\bar{x}/\bar{x}_G,^2$ und $\bar{z}_H = \bar{x}_G,^2/c\bar{x}$.

Drei Beispiele für die Anwendung des harmonischen Mittels

a) **Geschwindigkeiten**

Die Geschwindigkeit v auf einer gegebenen Strecke der Länge s ist umso größer (geringer), je kürzer (länger) die hierfür benötigte Zeit t ist, denn v = s/t. Die Durchschnittsgeschwindigkeit, die erreicht wird, wenn eine Strecke s zweimal befahren wird (hin und zurück) mit den Zeiten t_1 und t_2 beträgt dann

$2s/(t_1 + t_2) = 2/(t_1/s + t_2/s) = 2/(1/v_1 + 1/v_2)$.

Das ist aber genau das harmonische Mittel der beiden Geschwindigkeiten $v_1=s/t_1$ und $v_2=s/t_2$. Diese Überlegung lässt sich für

a) n gleichlange Teilstrecken der Länge s
b) unterschiedlich lange Strecken $s_1,s_2,...,s_m$, bei denen die Geschwindigkeiten $v_1,v_2,...,v_m$ betragen,

verallgemeinern.

zu a: Ermittelt werden soll die Durchschnittsgeschwindigkeit v der Einzelgeschwindigkeiten v_v (v = 1,2,...,n) für die Gesamtstrecke ns. Hierzu kann man von den Identitäten t = ns/v und $t = \Sigma t_v = \Sigma (s/v_v)$ ausgehen. Aus beiden Gleichungen folgt $v = ns/(\Sigma s/v_v) = n/\Sigma (1/v_v)$, was nichts anderes ist, als das ungewogene harmonische Mittel der n Einzelgeschwindigkeiten v_v.

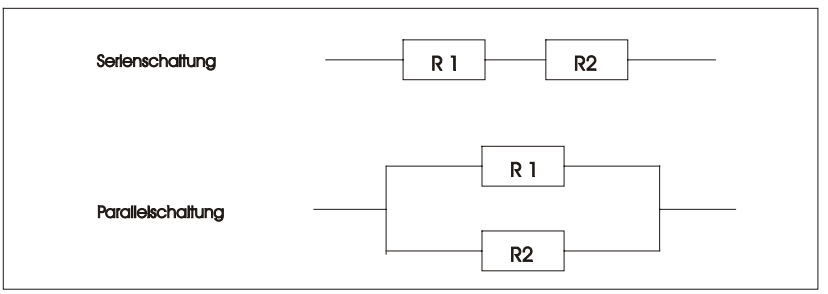
zu b: In diesem Fall ist $t = \Sigma t_i = \Sigma (s_i/v_i)$ (da für jede der m Teilstrekken gilt $v_i = s_i / t_i$) und $v = \Sigma s_i / \Sigma (s_i / v_i)$, d.h. das mit den Strecken s_i gewogene harmonische Mittel der Geschwindigkeiten v_i.
Die Durchschnittsgeschwindigkeit ist somit gleich dem harmonischen Mittel der Einzelgeschwindigkeiten. Das arithmetische Mittel führt (abgesehen vom Fall lauter gleicher Einzelgeschwindigkeiten) zu einer überhöhten mittleren Geschwindigkeit.

b) **Preise und Ausgaben**
Geldausgaben y_i sind Produkte aus Mengen q_i und Preisen $p_i = y_i / q_i$. Werden Preise bei konstanten Mengen (Nenner) gemittelt, so ist das arithmetische Mittel anzuwenden, sind die Mengen (aufgrund von Substitutionen) dergestalt veränderlich, dass die Ausgaben y (Zähler) konstant sind, so ist das harmonische Mittel anzuwenden. Ausgaben und gegebenenfalls auch Mengen sind aggregierbar, so dass $Y = \Sigma y_i$ und $Q = \Sigma q_i$. Gesucht ist ein mittlerer Preis (P) für den gilt $Y = P \cdot Q$. Es ist nun zu unterscheiden:

1. gleiche Mengen: sei $q_i = q$ für alle i = 1,...,n Waren, dann ist $P = Y/Q = \Sigma y_i / nq = q\Sigma p_i / nq = \Sigma p_i / n$ das ungewogene arithmetische Mittel der Preise;
2. gleiche Ausgaben: $y_i = y$, dann ist P das ungewogene harmonische Mittel der Preise $P = Y/Q = ny/\Sigma (y/p_i) = n/\Sigma (1/p_i)$.

c) **Widerstände** (vgl. Abb. 4.7)
Die Unterscheidung zwischen arithmetischer und harmonischer Mittelung tritt auch auf bei der Schaltung von Widerständen (für den Widerstand R gilt R = U/I mit U = Spannung, I = Stromstärke). Ein Widerstand, der den zwei Widerständen gleichwertig ist, ist der sog. "Ersatzwiderstand" R_E. Dabei gilt:
bei Serien- (Reihen-) Schaltung $R_E = R_1 + R_2$
bei Parallelschaltung $1/R_E = 1/R_1 + 1/R_2$
(entsprechende Zusammenhänge gelten bei mehr als zwei Widerständen).

Abb. 4.7: Schaltung von Widerständen



Weitere Eigenschaften des harmonischen Mittels:

1. Zulässige Transformationen, Skalen
   $\overline{x}_H$ erfüllt M3 für proportionale Transformationen: Aus $x_v^* = bx_v$ folgt mit $\overline{x}_H^* = n/S\ (1/x_v^*) = b\overline{x}_H$. Die Berechnung des harmonischen Mittels setzt somit eine Ratioskala voraus.

Die folgenden Eigenschaften seien nur kurz erwähnt:

2. Der Schwerpunkteigenschaft des arithmetischen Mittels entspricht beim harmonischen Mittel eine Ausgleichseigenschaft der reziproken Werte.

   Satz 4.6:

   Es gilt

   a) $\Sigma(1/x_v - 1/\overline{x}_H) = \Sigma(x_v^{-1} - \overline{x}_H^{-1}) = 0$ oder mit n multipliziert $\sum \frac{1}{x_v}(x_v - \overline{x}_M) = 0$ [3]

   b) $\Sigma\ (1/x_i - 1/\overline{x}_H)h_i = 0$

   Beweis

   a) $\Sigma\ x_v^{-1} - n/\overline{x}_H = 0$ da nach Definition $\overline{x}_H = n/\Sigma\ x_v^{-1}$ ist;

   b) gilt entsprechend.

3. Minimumeigenschaft: Die Funktion $\Sigma\ (c/x_v - c/M)^2$ mit der reellen Konstante c besitzt ein Minimum an der Stelle $M = \overline{x}_H$. Beweis: analog den Ausführungen beim arithmetischen Mittel.

4. Schwerpunkt- und Minimumeigenschaft kann man auch wie folgt darstellen. Es gilt:
   $\sum x_v^r\ (x_v - M) = 0$
   $\sum x_v^r\ (x_v - M)^2 = \text{Min}$ (bzgl. M)
   für folgende Mittelwerte:
   $M = \overline{x}_H$ (harmonisch): r = -1
   $M = \overline{x}$ (arithmetisch): r = 0
   $M = \overline{x}_A$ (antiharmonisch): r = 1 (Def. 4.5)

5. Aggregation und klassierte Daten
   Wird eine Gesamtmasse vom Umfang n zerlegt in g disjunkte Teilmassen mit den Umfängen $n_1, n_2, ..., n_g$ $(n = n_1 + n_2 + ... + n_g)$ so gilt:

   (4.17) $\overline{x}_H = n(n_1/\overline{x}_{H1} \cdot + ... + n_g/\overline{x}_{Hg})^{-1}$

   mit den (harmonischen) Teilmittelwerten $\overline{x}_{H1}, \overline{x}_{H2}, ..., \overline{x}_{Hg}$. Dieser Zusammenhang ist entsprechend anwendbar bei klassierten Daten.

---

[3] Zu einer inhaltlichen Interpretation dieser Eigenschaft des harmonischen Mittels vgl. Neubauer, W., Statistische Methoden, Frankfurt/M. 1991, S.62.

6. Von einer harmonischen Reihe $\frac{1}{1}, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ spricht man, weil jedes Glied das harmonische Mittel der benachbarten Glieder ist. Diese Reihe ist übrigens ein interessantes Beispiel dafür, dass eine Reihe nicht notwendig konvergiert, wenn ihre Glieder gegen Null streben.

7. Ersatzwerteigenschaft: Die Summe der reziproken Merkmalswerte $\Sigma\, 1/x_v$ kann aus dem reziproken Wert des harmonischen Mittels $\overline{x}_H$ gemäß $\Sigma\, 1/x_v = n\overline{x}_H^{-1}$ bestimmt werden.

8. Das gewogene harmonische Mittel zweier Zahlen x_1 , x_2 ist darstellbar als Interpolation gemäß der Funktion $\overline{x}_H = \frac{x_1 x_2}{x_1 a + x_2(1-a)}$ mit $0 \leq a = h_2 \leq 1$.

9. Das ungewogene harmonische Mittel zweier Zahlen x_1 , x_2 resultiert aus der Forderung $(x_2 - M)/(M - x_1) = x_2 / x_1$ oder $(x_1 - M)/(M - x_2) = x_1 / x_2$ oder $x_2 /(x_2 - M) = (M - x_1) / x_1$ für das Teilungsverhältnis einer Strecke $x_1 - x_2$ (weitere Erläuterungen hierzu vgl. Abb. 4.9). Denn dann ist $M = 2\, x_1 x_2 /(x_1+x_2) = \overline{x}_H$.

Zu weiteren Ausführungen über die die Anwendung von $\overline{x}_H$ und Zusammenhänge zwischen $\overline{x}$ und $\overline{x}_H$ vgl. Kap. 9 (Aggregation von Verhältniszahlen) und Kap. 10 (Zeitumkehrprobe bei Indexzahlen).

d) Quadratisches und antiharmonisches Mittel

Von sehr geringer Bedeutung für die Praxis sind das quadratische Mittel $\overline{x}_Q$ und das antiharmonische Mittel $\overline{x}_A$, die deshalb hier nur kurz zusammen behandelt werden und sinnvoll nur bei positiven Merkmalswerten berechnet werden.

Def. 4.5: quadratisches- und antiharmonisches Mittel

a) Das quadratische Mittel wird aus Einzelwerten ("ungewogen") mit

$$(4.18)\quad \overline{x}_Q = +\sqrt{\frac{\Sigma x_v^2}{n}}$$

bzw. bei gruppierten Daten (Merkmalsausprägungen, "gewogen") mit

$$(4.19)\quad \overline{x}_Q = \sqrt{\Sigma\, x_i^2\, h_i}$$

berechnet.

b) Die Maßzahl

$$(4.20)\quad \overline{x}_A = \frac{\overline{x}_Q^2}{\overline{x}}$$

heißt antiharmonisches Mittel.

Eigenschaften und Interpretation:

1. Das arithmetische Mittel der transformierten Werte $x_v^* = (x_v / \sqrt{\bar{x}})^2$ ist das antiharmonische Mittel der Werte x_v. Da dies eine monotone Transformation ist, gelten die vom arithmetischen Mittel erfüllten Axiome M1, M2 und M4 auch für $\bar{x}_A$. Da $\bar{x}_Q$ und $\bar{x}$ nur von den relativen, nicht den absoluten Häufigkeiten abhängen, erfüllt $\bar{x}_A$ auch M5. Das quadratische Mittel $\bar{x}_Q$ läßt sich als Potenzmittel darstellen, so dass x_Q alle Axiome M1 bis M5 erfüllt.
2. Für die transformierten Beobachtungswerte $x_v^* = bx_v$ gilt $\bar{x} = b\bar{x}_Q$ sowie $\bar{x}_A^* = b\bar{x}_A$, so dass $\bar{x}_Q$ und $\bar{x}_A$ bei Merkmalen, die auf einer Ratioskala gemessen sind, anwendbar sind.
3. Das quadratische Mittel tritt in der Streuungsmessung auf: die Standardabweichung ist das quadratische Mittel der Abweichungen vom arithmetischen Mittel.
4. Das ungewogene quadratische Mittel kann als $(\sqrt{1/n})$-fache euklidische Distanz zwischen den Ortsvektoren x_1 , x_2 ,...,x_n im n-dimensionalen orthogonalen Koordinatensystem aufgefaßt werden. Bei nur zwei Beobachtungen x_1 , x_2 gilt für diesen Abstand $\sqrt{x_1^2 + x_2^2} = \bar{x}_Q\sqrt{2}$

Nochmals: Mittelwerte der Pythagoräer

Es wurde bereits an den jeweiligen Stellen auf die formale Beschreibung der drei schon den Pythagoräern bekannten Mittelwerte $\bar{x}$, $\bar{x}_G$ und $\bar{x}_H$ eingegangen. Danach gelten für die Teilung der Strecke $x_2 - x_1$ (wenn $x_2 > x_1$)[4] in die beiden Streckenabschnitte $M-x_2$ und x_1-M (vgl. auch Abb. 4.8) die in Übersicht 4.2 aufgeführten Beziehungen.

Übersicht 4.2:

Mittel		
arithmetisch	$\frac{x_2 - M}{M - x_1} = \frac{x_2}{x_2} = \frac{½(x_2-x_1)}{½(x_2-x_1)} = 1$	wenn $M = \bar{x}$
geometrisch	$\frac{x_2 - M}{M - x_1} = \frac{x_2}{M} \geq 1$	wenn $M = \bar{x}_G$
harmonisch	$\frac{x_2 - M}{M - x_1} = \frac{x_2}{x_1} > 1$	wenn $M = \bar{x}_H$

antiharmonisch	$\frac{x_2 - M}{M - x_1} = \frac{x_1}{x_2} < 1$	wenn $M = \overline{x}_A$,

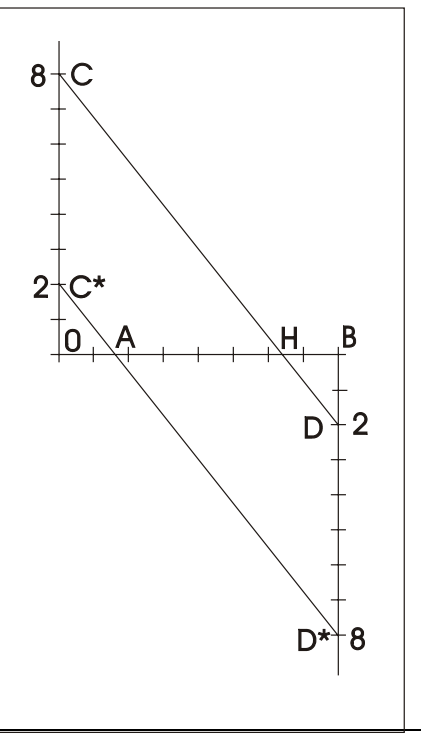
Aus dieser Betrachtung wird auch die Bezeichnung "anti"-harmonisch verständlich.

Die geometrische Veranschaulichung dieser Forderungen soll mit einem einfachen Zahlenbeispiel in Abb. 4.8 erfolgen. Die Strecke OB ist die Differenz der beiden Einzelwerte $x_2 = 8$ und $x_1 = 2$. Es ist also $OB = x_2 - x_1 = 6$. Trägt man

- in O senkrecht nach oben die Strecke x_2 ($OC = x_2 = 8$) und in B senkrecht nach unten die Strecke x_1 ($BD = x_1 = 2$) an, so teilt die lineare Verbindung zwischen C und D die Strecke OB in die Abschnitte $OH = x_2 - \overline{x}_H = 8 - 3{,}2 = 4{,}8$ und $HB = \overline{x}_H - x_1 = 3{,}2 - 2 = 1{,}2$ (**harmonisches Mittel**).
- in O senkrecht nach oben die Strecke x_1 ($OC^* = x_1 = 2$) und in B senkrecht nach unten die Strecke x_2 ($BD^* = x_2 = 8$) an, so teilt die lineare Verbindung zwi-schen C^* und D^* OB in die Abschnitte $OA = x_2 - \overline{x}_A = 1{,}2$ und $AB = \overline{x}_A - x_1 = 4{,}8$ (**antiharmonisches Mittel**).

Abbildung 4.8: harmonisches und antiharmonisches Mittel

Das arithmetische Mittel (Punkt M) teilt OB in genau zwei Hälften. Beim geometrischen Mittel $\overline{x}_G$ (Punkt G [$\overline{x}_G = 4$, so dass $OG = x_2 - \overline{x}_G = 8 - 4 = 4$]) gilt genauso wie beim harmonischen Mittel $(x_2 - M) / (M - x_1) > 1$ (weshalb die Punkte G und H rechts von M liegen), während beim antiharmonischen Mittel diese Relation < 1 ist (Punkt A liegt links von M).



[4] Die Formeln für $x_1 > x_2$ erhält man durch Vertauschung von x_1 und x_2.

e) Das Potenzmittel

Das sog. Potenzmittel ist eine Verallgemeinerung der in diesem Abschnitt behandelten Mittelwerte, so dass das harmonische, geometrische, arithmetische und quadratische Mittel jeweils Spezialfälle des Potenzmittels darstellen.

Def. 4.6: Potenzmittel

Die folgende Klasse von Mittelwerten

$$(4.21)\quad \bar{x}_{P,r} = \left[\frac{1}{n}(x^r,_1 + x^r,_2 + ...+ x^r,_n)\right]^{1/r} = \left(\frac{1}{n}\sum x_v^r\right)^{1/r}$$

(ungewogene Berechnung), bzw.

$$(4.22)\quad \bar{x}_{P,r} = (x_1^r h_1 + x_2^r h_2 + ...+ x_m^r h_m)^{1/r} = \left(\sum x_i^r h_i\right)^{1/r}$$

(gewogene Berechnung)

heißt Potenzmittel.

Bemerkungen zu Def. 4.6:

1. Es ist unschwer zu erkennen, dass die folgenden Mittelwerte Spezialfälle des Potenzmittels sind:

   $r = -1$: harmonisches Mittel $\bar{x}_{p,-1} = \bar{x}_H$
   $r \to 0$: geometrisches Mittel $\bar{x}_{p,0} = \bar{x}_G$
   $r = +1$: arithmetisches Mittel $\bar{x}_{p,1} = \bar{x}$
   $r = +2$: quadratisches Mittel $\bar{x}_{p,2} = \bar{x}_Q$.

   Es ist klar, dass r=0 nicht unmittelbar auf Gl. 4.22 angewendet werden kann, weil dies einen unbestimmten Ausdruck ergeben würde, bzw. für $\log(\bar{x}_{P,r})=\log(n^{-1}\Sigma\, x_v^r)/r$ den ebenfalls unbestimmten Ausdruck 0/0 liefern würde. Mit der l´Hospitalschen Regel (Differenzieren von Zähler [also $\log(n^{-1}\ \Sigma\ x_v^r)$] und Nenner [also r] dieses Ausdrucks nach r) lässt sich jedoch zeigen, dass $\log(\bar{x}_{P,0})$ der Logarithmus des geometrischen Mittels ist.

2. Man kann zeigen, dass $x_{p,r}$ mit wachsendem r monoton zunimmt, so dass die Ungleichung von Cauchy gilt:

   $$(4.23)\quad \bar{x}_H \le \bar{x}_G \le \bar{x} \le \bar{x}_Q$$

Die Gleichheit gilt dann und nur dann, wenn alle Beobachtungen x_v (v=1,2,...,n) gleich sind. Dass $\overline{x}_{p,r}$ mit wachsendem r monoton zunimmt, ergibt sich aus folgender Überlegung: Setzt man $f(x) = x^r$, was ja "nur" eine monotone Transformation ist, so ist $f(\overline{x}_{P,r}) = (\overline{x}_{P,r})^r = (1/n)\ \Sigma f(x_v) = (1/n)\ \Sigma x_v^r$, was mit wachsendem Exponenten r wächst (wie auch $\overline{x}_{p,r}$).

3. Mittelwerte und Lageparameter für nicht notwendig metrisch skalierte Merkmale

a) Zentralwert (Median)

Def. 4.7: Median

Das Merkmal X sei mindestens ordinalskaliert. Dann gilt für den Zentralwert (Median) $Z = \tilde{x}_{0,5}$

a) Bei Einzelbeobachtungen

$$(4.24)\quad Z = \tilde{x}_{0,5} = \begin{cases} x_{((n+1)/2)} & \text{falls n ungerade} \\ \tfrac{1}{2}[x_{(n/2)} + x_{([n/2]+1)}] & \text{falls n gerade} \end{cases}$$

Der Median ist der Wert, der in einer der Größe nach geordneten Reihe $x_{(1)} \le x_{(2)} \le \ldots x_{(n)}$ in der Mitte, d.h. an der ½(n+1)- ten Stelle steht (bzw. die Interpolation zwischen dem n/2-ten Wert und dem darauf folgenden Wert an der Stelle n/2 + 1).

b) Bei gruppierten Werten (Häufigkeitsverteilung) gilt entsprechend für den Median

$$(4.25)\quad \tilde{x}_{0,5} = \begin{cases} x_i & \text{falls } H_{i-1} < 0{,}5 \text{ und } H_i > 0{,}5 \\ \tfrac{1}{2}(x_i + x_{i+1}) & \text{falls } H_i = 0{,}5 \end{cases}$$

c) bei klassierten Daten wird der Median aus der Summenhäufigkeitskurve bestimmt (zur Interpolation vgl. Gl. 4.26, Bem. Nr. 7).

Bemerkungen zu Def. 4.7

1. Allgemein ist der Median der Merkmalswert, der die Daten in genau zwei gleiche Teile teilt: mindestens 50% der Merkmalswerte sind kleiner oder gleich Z und mindestens 50% aller Merkmalswerte sind größer oder gleich Z. Falls n gerade ist, kommt bei Einzelwerten somit jeder Wert im Intervall von $x_{(n/2)}$ bis zum folgenden Wert $x_{([n/2]+1)}$ als

Kandidat für den Median in Betracht. Entsprechend erfüllt für $H_i = 0,5$ bei gruppierten Daten jeder Wert im Intervall (x_i,x_{i+1}) diese Bedingung. Um eine eindeutige Bestimmung von Z zu gewährleisten, setzt man häufig (wie hier) in beiden Fällen Z jeweils als Intervallmitte fest.

2. Häufig interessiert nicht oder nicht nur der Durchschnitt im Sinne des arithmetischen Mittels, sondern vielmehr die Orientierung an der Mitte einer Verteilung. So bedeutet "Mittelmäßigkeit" im allgemeinen, dass hinsichtlich eines bestimmten Kriteriums etwa gleichviele Individuen besser und schlechter sind. Ähnlich ist bei ökonomischen Größen wie Einkommen, Vermögen und auch z.T. bei Preisen die Mitte der Verteilung häufig aufschlußreicher, als das arithmetische Mittel, da dies bei asymmetrischen Verteilungen nicht unbedingt ein "repräsentativer" oder "typischer" Wert zu sein braucht.

3. Während das arithmetische Mittel sehr reagibel gegenüber Ausreißern ist, ist der Median sehr robust.

4. Minimumeigenschaft

   Der Median hat die folgende Eigenschaft, die jedoch hier nicht bewiesen werden soll (man findet den Beweis aber in einigen Lehrbüchern der Statistik):

Die Summe der absoluten Abweichungen des Medians von den Beobachtungen x_v ist minimal.

   Während das arithmetische Mittel die Summe der quadrierten Abweichungen $\Sigma(x_v - M)^2$ minimiert, besitzt der Median diese Eigenschaft hinsichtlich der Summe der absoluten Abweichungen, d.h. $\Sigma|x_v - \tilde{x}_{0,5}|$ ist minimal. Beide Eigenschaften haben auch bei der Konstruktion von Streuungsmaßen eine Bedeutung.

5. Transformation

   Der Median erfüllt M3 für alle streng monotonen Transformationen, welche die Reihenfolge der Merkmalswerte nicht ändern. Bei nicht streng monotonen Transformationen läßt sich der Median der transformierten Daten im allgemeinen nicht dadurch bestimmen, dass man auf den Median der Ursprungswerte dieselbe Transformation anwendet. Mithin setzt eine Anwendung des Medians ein mindestens ordinal skaliertes Merkmal voraus.

6. Aggregation

   Im allgemeinen läßt sich der Median einer Gesamtheit nicht aus den Medianen von Teilgesamtheiten bestimmen. Es ist daher erforderlich, aus allen geordneten Reihen der Teilgesamtheiten, eine **einzige** geordnete Reihe der aus den Teilgesamtheiten bestehenden Gesamtheit zu bilden.

7. Klassierte Daten

   Bei klassierten Daten erfüllt der Median $\tilde{x}_{0,5}$ näherungsweise die Bedingung $H(x \leq \tilde{x}_{0,5}) = 0,5$. Ein Näherungswert für den Zentralwert wird durch Interpolation bestimmt. Man ermittelt hierzu zunächst die **Medianklasse** (d.h. die Klasse, in die der Median "fällt") k mit der Eigenschaft (mit x'_r = Klassen**obergrenze** der Klasse r):

   $H_{k-1} \leq 0,5 < H_k$, so dass gilt $x'_{k-1} \leq \tilde{x}_{0,5} < x'_k$.

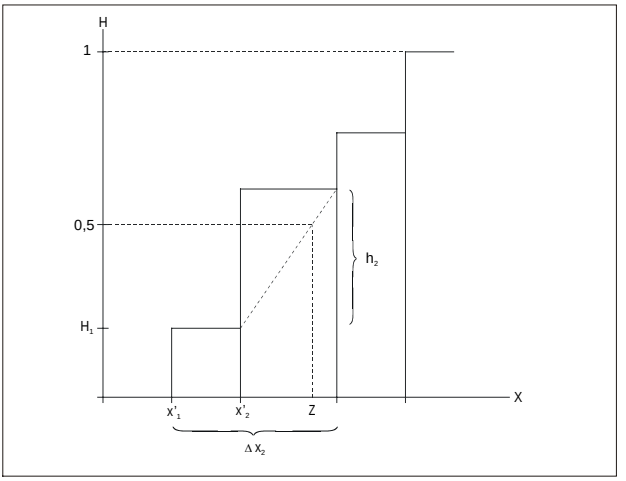
   Der Näherungswert für den Median lässt sich dann als Summe aus der Klassenuntergrenze x'_{k-1} der Medianklasse und einem Anteil der Klassenbreite b_k der k-ten Klasse (Medianklasse) definieren, der durch den Proportionalitätsfaktor $(0,5 - H_{k-1})/h_k$ gegeben ist (vgl. auch Abb. 4.9): Interpolation des Medians:

   (4.26) $\tilde{x}_{0,5} = x'_{k-1} + b_k(0,5 - H_{k-1})/h_k$

   Diese Formel ist unschwer aus einem "Dreisatz" zu entwickeln, denn offensichtlich gilt für den Betrag z, der zur Untergrenze x'_{k-1} der Medianklasse hinzuzuaddieren ist, gem. Abb. 4.9:
   $z/(½ - H_{k-1}) = b_k/h_k$, so dass $\tilde{x}_{0,5} = x'_{k-1} + z$.

Der Einfachheit halber unterscheiden wir hier in der Notation nicht zwischen dem unbekannten Median und seinem Näherungswert. Mit Gl. 4.26 erhält man eine umso bessere Approximation, je mehr sich die Verteilung innerhalb der Medianklasse k einer Gleichverteilung (im Sinne gleicher Häufigkeiten aller Werte innerhalb der Klasse) nähert. Z läßt sich analog zu Gl. 4.26 auch aus den kumulierten absoluten Häufigkeiten N_k bestimmen.

Abb. 4.9: Bestimmung des Medians mit Interpolation aus der Summenhäufigkeitskurve



Beispiel 4.7:

Bei einer Untersuchung der Lebensdauer (in Jahren) von 600 Fernsehgeräten eines bestimmten Typs ergaben sich die im folgenden aufgeführten Werte, deren Median (Zentralwert) zu bestimmen ist.

k	x_k *)	n_k	h_k	H_k
1	0 - 2	21	0,035	0,035
2	2 - 4	178	0,297	0,332
3	4 - 6	255	0,425	0,757
4	6 - 8	123	0,205	0,962
5	über 8	23	0,038	1,000

*) Lebensdauer von - bis unter - Jahre

Lösung 4.7:

Aus der Spalte H_k ist zu erkennen, dass k = 3 die Medianklasse ist, weil die bis zu Beginn der dritten Klasse erreichte kumulierte Häufigkeit 33,2% ist, während sie am Ende dieser Klasse 75,7% beträgt. Es gilt also mit der Interpolationsformel (Gl. 4.26):
Untergrenze der Medianklasse 4, Breite $b_k = 2$, relative Häufigkeit h_k = 0,425 und bei Beginn der Medianklasse $H_{k-1} = 0{,}332$. Also ist $\tilde{x}_{0,5} = 4 + 2(0{,}5 - 0{,}332)/0{,}425 = 4{,}791$.

Beispiel 4.8: Der optimale Standort des Stehgeigers von Budapest

László Varga möchte den Bewohnern der Häuser A bis D der Bartók-Straße seine Sonate in d-moll op. 125 zu Gehör bringen. Dabei wünscht der Tonkünstler, dass alle 25 Familien möglichst gleich gut die Gelegenheit haben, die Sonate zu hören und zu würdigen.
Wo sollte sich Varga hinstellen, wenn die Straße wie folgt aussieht (Im i-ten Haus, das x_i Meter vom Beginn der Straße entfernt ist, wohnen n_i Familien):

x_i	n_i	N_i
0	3	3
10	4	7
20	1	8
30	2	10
35	3	13
50	2	15

Man erkläre anhand dieser Aufgabe auch die "Minimumeigenschaft" des Medians!

Lösung 4.8:

Um von allen potentiellen Hörern (und Zahlern) gleich weit entfernt zu sein, sollte sich Varga in den Medianpunkt stellen. Der Median hat die geringste Summe der absoluten Abweichungen. Er beträgt in diesem Fall $\tilde{x}_{0,5} = 20$, da bei n = 15 Familien der Median der 8-te [8=(15+1)/2] Wert ist. Das arithmetische Mittel beträgt 21,667. Mit d_{Zi} sollen die Abstände (absolute Abweichungen) vom Median bezeichnet werden, mit d_{ai} die Abstände vom Wert x = 22, mit d_{bi} die Abweichungen von x = 18 und mit d_{ci} die Abstände von x = 19:

		x = z = 20		x = 22		x = 18		x = 19	
x_i	n_i	d_{Zi}	$d_{Zi}n_i$	d_{ai}	$d_{ai}n_i$	d_{bi}	$d_{bi}n_i$	d_{ci}	$d_{ci}n_i$
0	3	20	60	22	66	18	54	19	57
10	4	10	40	12	48	8	32	9	36
20	1	0	0	2	2	2	2	1	1
30	2	10	20	8	16	12	24	11	22
35	3	15	45	13	39	17	51	16	48
50	2	30	60	28	56	32	64	31	62
$\sum$			225		227		227		226

Wie man sieht, ist die Summe der Abstände vom Median (20) kleiner als von den benachbarten Positionen (x = 18, 19 oder 22).

b) Quantile

Quantile teilen eine Verteilung in Abschnitte gleicher Häufigkeit ein, so beispielsweise:

- zwei **Terzile** T_1, T_2 in drei Abschnitte mit den relativen Häufigkeiten $h_1 = 1/3$ (für das Intervall $x < T_1$), $h_2 = 1/3$ (für das Intervall $T_1 \le x < T_2$) und $h_3 = 1/3$ (für $x \ge T_2$)
- drei **Quartile** Q_1, $Q_2 = \tilde{x}_{0,5}$ (Median) und Q_3 in vier Abschnitte mit den relativen Häufigkeiten $h_1 = ... = h_4 = 1/4$
- vier **Quintile** in fünf Abschnitte mit $h_1 = ... = h_5 = 0,2$
- neun **Dezile** in zehn Abschnitte mit $h_1 = ... = h_{10} = 0,1$.

Das allgemeine Konzept ist das des Quantils (oder p-Quantils), wovon Terzile, Quartile, Quintile und Dezile Spezialfälle sind. Es handelt sich dabei um Lageparameter im allgemeinen Sinne. Nur das zweite Quartil, das identisch mit dem Median ist, kann als Mittelwert bezeichnet werden.

Def. 4.8: Quantil

Das Merkmal X sei mindestens ordinalskaliert. [c] bedeutet "ganze Zahl, die kleiner oder gleich c ist" (Gaußklammer).
Dann heißt die Maßzahl

$$\tilde{x}_p = \begin{cases} x_{([np+1]),} & \text{wenn np nicht ganzzahlig ist} \\ \tfrac{1}{2}(x_{[np]}+x_{[np+1]}), & \text{wenn np ganzzahlig ist} \end{cases} \tag{4.27}$$

p-Quantil ($0 < p < 1$).

Spezielle Quantile und Anzahl der Intervalle			
Zentile (Perzentile)	100	Quintile	5
Vigintile	20	Quartile	4
Dezile	10	Terzile	3

Bemerkungen zu Def.4.8:

1. Def. 4.8 ist eine Verallgemeinerung von Def. 4.7 für den Zentralwert, der das Quantil für p = 0,5 darstellt.

2. Die p-Quantile teilen die Daten in zwei Gruppen auf. Unterhalb von $\tilde{x}_p$ befinden sich höchstens 100·p% der Merkmalswerte und oberhalb von $\tilde{x}_p$ befinden sich höchstens 100(1-p)% der Merkmalswerte. Sofern np ganzzahlig ist, gelten diese Prozentsätze exakt. Gl. 4.27 kann ohne Schwierigkeiten auch bei gruppierten Daten angewendet werden.

3. Genau dann, wenn np nicht ganzzahlig ist, sind Quantile realisierte Merkmalswerte. Für ganzzahlige Werte np ist die Intervallmitte zwischen den beiden benachbarten Werten $x_{(np)}$ und $x_{(np+1)}$ der geordneten Reihe zu betrachten.

4. Klassierte Daten:
   Bei klassierten Daten können p-Quantile geometrisch leicht gedeutet werden. Ein p-Quantil ist derjenige Wert, bei dem die empirische Verteilungsfunktion H(x) den Wert p annimmt. Es kann im allgemeinen nur näherungsweise durch Interpolation bestimmt werden. Die Berechnungsformel ist analog zu Gl. 4.26. Ist k die Klasse, in die das p-Quantil $\tilde{x}_p$ fällt, dann ist $\tilde{x}_p$ näherungsweise durch

   (4.26a) $\tilde{x}_p = x'_{k-1} + b_k(p - H_{k-1})/h_k$

   gegeben.

5. Die p-Quantile sind Funktionswerte der inversen Verteilungsfunktion G[H(x)] = G(H) (vgl. Kap.3, Def. 3.3). So erhält man beispielsweise den Median $Z = \tilde{x}_{0,5}$ durch die Lösung der Gleichung Z = G(0,5) und die beiden Quartile Q_1 und Q_3 sind durch die Funktionswerte Q_1 = G(0,25) und Q_2 = G(0,75) der Funktion G(H) gegeben.

6. Aus Quantilen lassen sich durch Mittelung Lageparameter herleiten, etwa die

   Quartilsmitte $\frac{1}{2}(Q_1+Q_3)$

oder entsprechend die Dezilsmitte ½(D_1+D_9). Ein Grenzfall ist

(4.28) $\bar{x}_m = ½(x_{min}+x_{max})$

bekannt als der "Mittelpunkt" [midpoint oder midrange] (des Streubereichs), der auch "midrange" genannt wird. Man kann $\bar{x}_m$ herleiten aus einer Minimierungsaufgabe: $\bar{x}_m$ ist das Minimum der maximalen Abstände der Werte x_v von einem festen Wert c, was in Aufgabe 4.10 demonstriert wird:

$$\max_v |x_v - \bar{x}_m| \le \max_v |x_v - c|.$$

Im Unterschied zur Quartilsmitte ist $\bar{x}_m$ sehr schnell durch extreme Werte verzerrt. Deshalb kann $\bar{x}_m$ keine Alternative zu weniger sensiblen Mittelwerten sein.

Beispiel 4.9:

Die Untersuchung eines Einzelhändlers über die Umsatzzahlen U_v (DM pro Tag) der 10 im Sortiment befindlichen Radiotypen ergab die folgenden Werte U_v: 350, 420, 780, 890, 975, 1010, 1170, 1230, 1680, und 1910

a) Man bestimme die Quartile Q_1 (das untere Quartil) und Q_3 (oberes Quartil) sowie die Quartilsmitte! Unterscheiden sich Median und Quartilsmitte?
b) Man verfahre entsprechend, wenn sich die Statistik nur auf (die ersten) acht Radiotypen erstreckt.

Lösung 4.9:

a) zehn Radiotypen
Unteres Quartil Q_1:
p = 0,25. Dann ist np = 2,5 (nicht ganzzahlig!), so dass nach Def. 4.8 x_p der Wert $x_{([np+1])}$ ist. Nun ist np + 1 = 3,5 . Die Gaußklammer bedeutet "ganze Zahl, die kleiner oder gleich 3,5 ist", also die Zahl 3. Folglich ist $Q_1 = x_{(3)} = 780$.
Oberes Quartil Q_3:
p = 0,75. Die entsprechende Betrachtung führt zu $Q_3 = x_{(8)} = 1230$. Die Quartilsmitte ist folglich 1005. Bei der Bestimmung des Zentralwerts (Medians) ist Def. 4.8 mit p = 0,5 anzuwenden: der Median ist somit ½ [$x_{(np)} + x_{(np+1)}$] = ½ [$x_{(5)} + x_{(6)}$] = ½ (975 + 1010) = 992,5 (im Unterschied zur Quartilsmitte, die 1005 beträgt).

b) acht Radiotypen

Q_1: In diesem Fall ist np = 8·0,25 = 2 also ganzzahlig, so dass Q_1 als Mittelwert aus $x_{(np)}$ und $x_{(np+1)}$, also aus $x_{(2)}$ = 420 und $x_{(3)}$ = 780 zu bestimmen ist (Q_1 ist also 600);
Q_3: das obere Quartil ist das ungewogene Mittel aus $x_{(6)}$ = 1010 und $x_{(7)}$ = 1170 also 1090. Die Quartilsmitte ist ½(600+1090) = 845 (der Median ist dagegen ½(890+975) = 932,5 also nicht identisch mit der Quartilsmitte).

Beispiel 4.10:

Man zeige für die folgenden Daten x: 10,12,15,18,22, dass für die midrange $\max_v x_m$ (Gl. 4.28) gilt: $\max |x_v - \max_v x_m| \leq \max |x_v - c|$ (mit c als einer von $\max_v x_m$ verschiedenen Konstanten).

Lösung 4.10:

In diesem Fall gilt $\bar{x}$ = ½(10+22) = 16. Die Abweichungen der einzelnen Werte von verschiedenen Werten für c betragen:

Daten	Abweichungen von c								
x_v	c=12	c=13	c=14	c=15	c=16	c=17	c=18	c=19	c=20
10	2	3	4	5	6	7	8	9	10
12	0	1	2	3	4	5	6	7	8
15	3	2	1	0	1	2	3	4	5
18	6	5	4	3	2	1	0	1	2
22	10	9	8	7	6	5	4	3	2
$\max \|x_v - c\|$	10	9	8	7	6	7	8	9	10

Wie man sieht ist die geringste Maximalabweichung 6. Das ist die Abweichung von c = $\bar{x}_m$ = 16.

c) Modus (dichtester Wert, häufigster Wert)

Def. 4.9: Modus

Existiert bei einer diskreten Variable (einem diskreten Merkmal) X mit den Merkmalsausprägungen x_i genau ein Merkmalswert $\bar{x}_M$, so dass

$$(4.29) \quad h(x = \bar{x}_M) = \max_i h(x_i),$$

so ist dieser Wert der Modus $\bar{x}_M$ [oder D] (oder der Modalwert, der dichteste- bzw. häufigste Wert).
Der Modus ist derjenige Merkmalswert, der in einer Häufigkeitsverteilung am häufigsten vorkommt.

Bemerkungen zu Def. 4.9:

1. Während zur Bestimmung des Medians aus Einzelwerten eine geordnete Reihe gebildet werden muss, sind zur Ermittlung des Modus die Einzelwerte zu gruppieren. Die Bestimmung des Modus ist auch sinnvoll, wenn eine Klassierung vorgenommen wurde (d.h. bei einer klassierten Verteilung). Die Lage (der Wert) des Modus (der Modalklasse) kann dann aber durch unterschiedliche Arten der Klassenbildung verschoben werden (vgl. auch Bem. Nr. 6).

2. Mit dem Modus verbindet man eine gewisse Vorstellung von Normalität und Üblichkeit. Mit "normal" ist meist der häufigste Wert (der Gipfel einer Verteilung) gemeint. Bei unklassierten Daten ist der Modus immer ein realisierter Wert, was bei den anderen Lageparametern nicht notwendig der Fall zu sein braucht, so dass er sich durch hohe Realitätsnähe auszeichnet.

4. Transformation und Skaleneigenschaft: Es gilt bei einer eineindeutigen Transformation $x^*_i = f(x_i)$ für den Modus der transformierten Werte $D^* = f(D)$. Der Modus ist also schon bei nur nominalskalierten Merkmalen ein sinnvolles Lagemaß.

4. Die Bestimmung des Modus macht wenig Sinn, wenn die Verteilung nicht eingipflig (unimodal) ist. Bei einer U-förmigen Häufigkeitsverteilung nennt man ein lokales Minimum auch Antimodus.

5. Ein extremer Fall der Abstandsmessung ist die Verabredung, den Wert $d_v = d(x_v,x)$ (d bedeutet Distanz zwischen x_v und x) wie folgt zu definieren:
   $d_v = 0$ wenn $x_v = x$ und $d_v = 1$ wenn $x_v \neq x$
   (x ist ein beliebiger Wert für das Merkmal X [eine beliebige Merkmalsausprägung])
   Der Modus ist dann derjenige Wert von x, für den gilt $\Sigma d_v = \min_x$, denn ist $x = \bar{x}_M$, so wird $d_v = 0$ häufiger als bei allen anderen Werten für x vergeben.
   Das zeigt erneut, dass man Mittelwerte aufgrund von Extremwerten von "Gesamtabständen" definieren kann: das arithmetische Mittel ist wegen Satz 4.2 derjenige Wert x, für den die Größe $\Sigma(x_v - x)^2$, also die Summe der quadrierten Abstände minimal ist.

6. Klassierte Daten
   Bei einer klassierten Verteilung begnügt man sich im allgemeinen mit einer Angabe der Modalklasse d, also der Klasse, deren (relative) Häufigkeit die größte der Häufigkeitsverteilung ist. Als Modus gilt dann die Klassenmitte m_d der Modalklasse. Seltener bestimmt man den Modus durch Interpolation. Sofern die modale Klasse d sowie die beiden angrenzenden Klassen die gleiche Breite b_d haben, ist der Modus:

(4.30) $\bar{x}_M = D = x'_d + (A \cdot b_d)/(A+B)$

mit der Klassenuntergrenze x'_d der Modalklasse und den beiden Differenzen von relativen Häufigkeiten der benachbarten Klassen $A= h_d - h_{d-1}$ und $B= h_d - h_{d+1}$. Der Modus D rückt damit näher an die jeweils stärker besetzte der beiden benachbarten Klassen (in Abb.4.10 an die Klasse d+1).

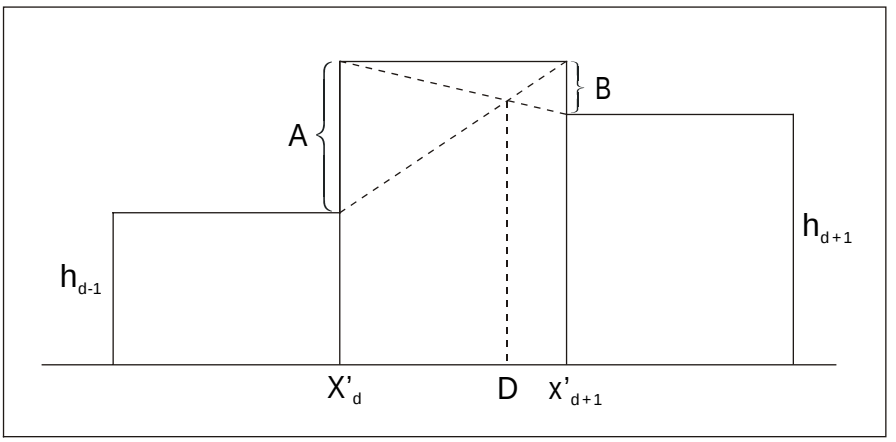
Beispiel 4.11:

Man bestimme den Modus nach Gl. 4.30 für das Beispiel 4.2!

Lösung 4.11:

Die Modalklasse ist die Klasse 5. Die Klassenbreiten der 5-ten und angrenzenden Klassen sind b_d = 400DM. Die Untergrenze der Modalklasse beträgt 2200DM. Ferner ist A = (37-31)/150 und B = 15/150, so dass man $\overline{x}_M$ = D = 2200 + (6/21)400 = 2314,29 erhält.

Abb.4.10: Interpolation des Modus



Kapitel 5: Streuung, Schiefe, Wölbung

1. Streuungsbegriff und Eigenschaften von Streuungsmaßen

a) Begriff der Streuung (Dispersion)

Streuungsmaße sind einmal beschreibende Statistiken von Häufigkeitsverteilungen und zum anderen auch bedeutsam für die Beurteilung statistischer Berechnungen. Sie dienen

1. der Charakterisierung der Variabilität eines Merkmals oder, gleichbedeutend, der Ausbreitung einer Häufigkeitsverteilung und der Homogenität einer statistischen Masse, d.h. der Ähnlichkeit ihrer Einheiten, bzw. der Distanz zwischen ihnen;
2. der Beurteilung der Güte einer Schätzung (z.B. aufgrund einer Stichprobe) oder der Treffsicherheit einer Prognose, sowie der Messung von Konzepten wie Risiko, Zuverlässigkeit und Fehleranfälligkeit.

zu 1:
Streuungsmaße sind wichtige Ergänzungen zu den Mittelwerten, die die zentrale Tendenz einer Verteilung widerspiegeln sollen. Bei geringer Streuung ist ein Mittelwert eher ein typischer Wert einer Verteilung, als bei einer starken Variabilität der Daten.

Für bestimmte Probleme kann die Streuung (Dispersion) einer Häufigkeitsverteilung sogar wichtiger sein als der Mittelwert. Bei zwei Garnsorten A und B kann z.B. das Garn A zwar eine größere mittlere Reißfestigkeit als das Garn B haben, trotzdem kann B dem Garn A vorgezogen werden, weil die Variabilität des Garns A im Vergleich zu Garn B derart groß ist, dass der Anteil der Fadenbrüche aufgrund eines häufigeren Unterschreitens der kritischen Reißfestigkeit beim Garn A nicht akzeptabel ist.

zu 2:
Die Streuung ist auch von Bedeutung für die Beurteilung der Treffsicherheit einer statistischen Prognose, die in der Regel auf einem stochastischen Modell beruht, und sie ist generell von Bedeutung für die Stichprobentheorie.

Wie in der Induktiven Statistik zu zeigen sein wird, ist der für eine gegebene Genauigkeit und Sicherheit der Schätzung erforderliche Stichprobenumfang eine Funktion eines Streuungsmaßes der Grundgesamtheit. Man kann sich dies leicht anhand des folgenden Extremfalles klar machen: Sind alle Einheiten der Grundgesamtheit in bezug auf das Merkmal X gleich (ist also die Streuung der Variablen X in der Grundgesamtheit Null), so genügt ein Stichprobenumfang von n=1, also einer Einheit, um mit Sicherheit und ohne Fehler Aussagen über die Grundgesamtheit machen zu können. Häufig werden, wie hier, die Begriffe Streuung und Dispersion als synonym betrachtet. Bei einigen Autoren (z.B. Ferschl) wird der Begriff Dispersion jedoch eingeschränkt auf relative Streuungsmaße (auch Streuungskoeffizienten genannt), die dimensionslos sind und als Quotient aus einem absoluten Streuungsmaß und einem Lagemaß gebildet werden (vgl. Abschn. 1d).

Beispiel 5.1 soll das Konzept der Streuung anhand verschiedener Häufigkeitsverteilungen veranschaulichen (Abb. 5.1). Dabei wird jeweils die Varianz berechnet, ein Streuungsmaß, das jedoch erst an späterer Stelle (Abschn. 2) definiert wird.

Beispiel 5.1:

Gegeben seien die folgenden drei Häufigkeitsverteilungen mit jeweils gleichem arithmetischen Mittel $\bar{x} = 3$ und zunehmender Streuung (was anhand der Abb. 5.1 leicht zu beurteilen ist, da alle Verteilungen symmetrisch sind):

Verteilung A

x_i	h_i
3	1,0

Verteilung B

x_i	h_i
2	0,2
3	0,6
4	0,2

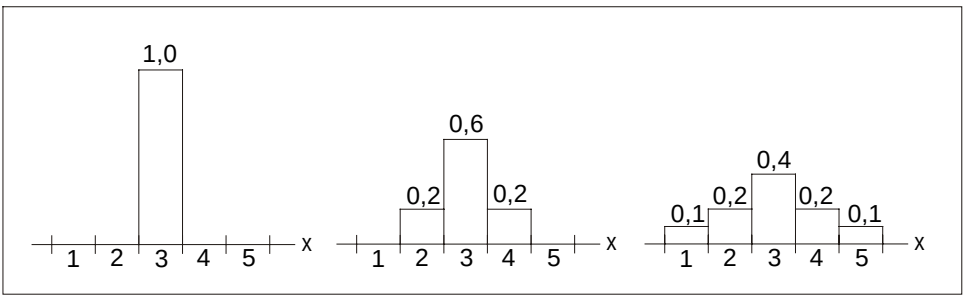
Verteilung C

x_i	h_i
1	0,1
2	0,2
3	0,4
4	0,2
5	0,1

Lösung 5.1:

Es ist ganz offensichtlich, dass in Abb.5.1 die Streuung von links nach rechts zunimmt. Verteilung A ist eine sog. Einpunktverteilung; alle Merkmalswerte sind gleich und die Streuung ist deshalb Null. Die Streuung, gemessen an der Varianz ist bei A: 0, bei B: 0,4 und bei C: 1,2.

Abb. 5.1



b) Konstruktion von Maßen der absoluten Streuung

Es gibt **drei Konstruktionsprinzipien** nach denen die gebräuchlichen Maße der absoluten Streuung gebildet werden, **wenn** das Merkmal X **metrisch skaliert** ist, also das Konzept des Abstands sinnvoll ist[1]. Ein Streuungsmaß kann danach berechnet werden als Maßzahl aus:

1. Abständen der Merkmalswerte von einem Lageparameter, z.B. von einem Mittelwert (nach diesem Prinzip sind die folgenden Streuungsmaße konstruiert: durchschnittliche Abweichung, Medianabweichung, Varianz und Standardabweichung),

[1] Zu Streuungsmaßen für nicht-metrisch skalierte Merkmale vgl. Exkurs am Ende von Abschnitt 3.

2. dem Abstand zweier Ordnungsstatistiken (Beispiele: Spannweite [range] oder mittlerer Quartilsabstand),
3. Abständen der Merkmalswerte untereinander (z.B. Ginis Maß).

Streuungsmaße, die Mittelwerte der Abstände der Beobachtungen von einem Mittelwert darstellen (Konstruktionsprinzip Nr. 1) unterscheiden sich nach

- der Art des Mittelwerts von dem die Abstände gemessen werden und
- nach der Art des Mittelwerts mit dem die Abstände gemittelt werden.

Übersicht 5.1 zeigt diese Zusammenhänge für einige besonders gebräuchliche Streuungsmaße.

Übersicht 5.1: Streuungsmaße nach dem Konstruktionsprinzip Nr. 1

Abweichung vom	Mittel der Abweichungen	Streuungsmaß
arithmet. Mittel	quadratisches Mittel	Standardabweichung
arithmet. Mittel*)	arithmetisches Mittel	Varianz
Median**)	arithmetisches Mittel	durchschn. Abweich.
Median**)	Median	Medianabweichung

*) quadrierte Abweichungen vom arithmetischen Mittel
**) absolute Abweichungen vom Median (Zentralwert)

Man kann sich aufgrund des Schemas der Übers. 5.1 auch weitere Streuungsmaße vorstellen, z.B. bei Verwendung des harmonischen Mittels. Da sich wegen der Schwerpunkteigenschaft des arithmetischen Mittels positive und negative Abweichungen gegenseitig aufheben, so dass stets gilt $\sum(x_v - \bar{x}) = 0$, ist eine arithmetische Mittelung nur bei anders definierten Abständen sinnvoll. Man kann

- absolute Abweichungen $\Sigma |x_v - \bar{x}|$, oder aber
- quadrierte Abweichungen $\Sigma (x_v - \bar{x})^2$ bilden,

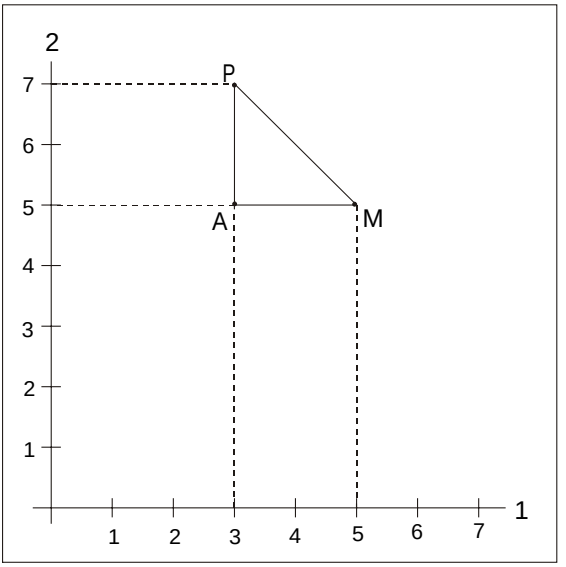
denn in beiden Fällen ist die Summe nichtnegativ und auch nicht notwendig Null (womit sie ohne jede Aussagefähigkeit wäre).

Der erste Weg ist von Laplace vorgeschlagen worden und wird bei der durchschnittlichen Abweichung benutzt, der zweite Weg geht auf Gauß zurück und wird bei der Varianz angewendet.

Beide Prinzipien sind auch bei der Konstruktion von **Distanzmaßen** in der Statistik gebräuchlich: die Summe absoluter Abweichungen wird benutzt bei der sog. City-Block-Distanz und die (bei der Standardabweichung verwendete) Wurzel aus der Summe der quadrierten Abweichungen ist die euklidische Distanz.

Geht man bei n Beobachtungen von einem n-dimensionalen Koordinatensystem aus, etwa zur graphischen Veranschaulichung von n=2 (Abb. 5.2), so sind die Daten darstellbar als ein Punkt (in Abb. 5.2 der Punkt P) und das arithmetische Mittel ebenfalls als ein Punkt (in Abb. 5.2 der Punkt M mit den Koordinaten 5 und 5).

Abb. 5.2: Euklidische- und City-Block-Distanz



In der Abb. 5.2 ist von den Werten $x_1 = 3$ und $x_2 = 7$ ausgegangen worden, so dass $\bar{x} = 5$ und $\Sigma |x_v - \bar{x}| = |3-5| + |7-5| = 4$ (das ist die Gesamtlänge des Weges von P über A nach M) und $\Sigma (x_v - \bar{x})^2 = 8$, was das Quadrat der euklidischen Distanz zwischen M und P ist. Während diese Distanz quasi die Luftlinie zwischen M und P darstellt, basiert die City-Block-Distanz auf der Vorstellung einer schachbrettartig aufgebauten Stadt, in der man von einem Punkt zum anderen durch Abschreiten rechtwinkliger Staßenzüge gelangt. Eine graphische Darstellung ist natürlich auch bei n=3 möglich. Dass Streuungsmessung und Distanzmessung miteinander verwandt sind, ist kein Zufall, denn die Streuung soll ja Ausdruck der Homogenität oder Heterogenität eines Datensatzes sein. Beide Distanzen, die euklidische- und die City-Block-Distanz sind Spezialfälle der Minkowski-Distanz :

$$d_{PM} = \left[\sum_{v=1}^{n} |x_v - \bar{x}|^r \right]^{\frac{1}{r}}$$ nämlich für r=1 und r=2.

Weitere Bemerkungen zu den Konstruktionsprinzipien:

1. Da bei allen diesen Konstruktionsprinzipien eine Abstandsmessung vorliegt, können die üblicherweise verwendeten Streuungsmaße nur für mindestens intervallskalierte Merkmale sinnvoll gebildet werden. Es gibt aber auch Streuungsmaße wie z.B. die Entropie und die im nachfolgenden Exkurs genannten Maße (z.B. Diversität), die nicht in dieses Schema der drei Konstruktionsprinzipien passen und deshalb auch keine metrische Skala voraussetzen.

2. Die Konstruktionsprinzipien hängen untereinander durchaus zusammen. So ist z.B. die Varianz aus quadrierten Abständen der Merkmalswerte vom arithmetischen Mittel gebildet (Konstruktionsprinzip Nr.1). Man kann die Varianz aber auch als ein Vielfaches der Summe der quadrierten Abweichungen der Merkmalswerte untereinander (Prinzip Nr. 3) darstellen (vgl. hierzu Satz 5.1).

3. Das Prinzip Nr. 2 wird auch bei der Messung der Schiefe angewendet.

4. Wenn auch das erste Konstruktionsprinzip am häufigsten angewandt wird, so hat doch das zweite mit der Verbreitung der explorativen Datenanalyse an Bedeutung gewonnen. Eine interessante Verknüpfung der beiden letzten Konstruktionsprinzipien stellt ein sog. Ginilike-Streuungsmaß dar, das resistent gegenüber Ausreißern ist.

c) Axiomatik absoluter Streuungsmaße

Absolute Streuungsmaße (S) sind Verteilungsmaßzahlen, die unter Berücksichtigung des Skalenniveaus die Axiome S1 bis S4 erfüllen.

S1	Ein absolutes Streuungsmaß S soll den Wert Null annehmen, falls $x_1 = x_2 = ... = x_n = \bar{x}$ gilt, d.h. wenn alle Merkmalswerte identisch sind.
S2	Sofern mindestens zwei Merkmalswerte x_i und x_j voneinander verschieden sind, ist $S > 0$ $(i,j = 1,2,...,n)$.
S3	Ersetzt man den Beobachtungswert x_k aus der Folge der Beobachtungen x_v $(v = 1,2,...,n)$ durch den neuen Wert x_p, so dass die Summe der absoluten Abweichungen von x_p von allen übrigen Werten größer ist als die Summe der absoluten Abweichungen von x_k von allen übrigen Werten, so soll das Streuungsmaß S nicht abnehmen.

S4 Invarianz gegenüber Verschiebungen des Nullpunkts (Translationen) aber nicht gegenüber Maßstabsänderungen: Falls S die Maßeinheit der Merkmalswerte $x_1,x_2,...,x_n$ hat, soll für die Streuung S_y der mit $y_v = a+bx_v$ transformierten Variablen X gelten: $S_y = |b|S_x$, $|b| > 0$. Für ein absolutes Streuungsmaß mit der quadrierten Maßeinheit der Merkmalswerte soll dann gelten $S_y = b^2S_x$.

Bemerkungen zu den Axiomen:

1. Ein Streuungsmaß S sollte nichtnegativ sein ($S \geq 0$), denn die "Streuung" ist ein durch ihr Ausmaß, nicht durch Ausmaß und Richtung zu kennzeichnender Tatbestand. Nach den Axiomen S1 und S2 ist ein Streuungsmaß S dann, und nur dann Null, wenn alle beobachteten Werte x_v gleich sind (bzw. [trivialer Fall] bei n = 1). Eine Obergrenze ist für S nicht vorgesehen, d.h. die Streuung ist nichtnegativ aber auch betragsmäßig nicht beschränkt, wenn nicht besondere Einschränkungen bezüglich der X-Variable gemacht werden. Mit den Merkmalswerten $x_1 = x$ und $x_2 = x + n$ ist z.B. die Varianz als Streuungsmaß $n^2 h_1 h_2$, was mit wachsendem n über alle Grenzen zunimmt.

2. Axiom S3 geht von der intuitiven Vorstellung der "Streuung" aus: Je mehr die Beobachtungswerte voneinander differieren, desto größer sollte ein Streuungsmaß sein. Würde man anstelle der obigen Formulierung von S3 auf Abweichungen von einem Mittelwert abstellen, dann entstünde das Problem der Wahl eines geeigneten Mittelwerts. Ein so formuliertes Axiom würde zu stark auf eines der genannten Konstruktionsprinzipien von Streuungsmaßen Bezug nehmen.

3. Axiom S4 bedeutet, dass ein Streuungsmaß invariant sein soll gegenüber

   a) Verschiebungen des Nullpunkts (Translation) mit der Größe a, d.h. es soll "verschiebungsinvariant" sein,

   b) Veränderungen der Skaleneinheit (Maßstabsänderung) in dem Sinne, dass eine Ver-b-fachung ($b \neq 0$) der Beobachtungswerte zu einem b-fachen Streuungsmaß führt, falls S die Maßeinheit der Merkmalswerte hat.

4. Die Axiome S3 und S4 stellen auf mindestens intervallskalierte Merkmale ab. Bei Streuungsmaßen, die für nominal- oder ordinalskalierte Merkmale konzipiert sind, gelten sie nicht.

d) Relative Streuung

Def. 5.1: Relative Streuung

Die Maße der relativen Streuung (S_r) sind definiert als Quotienten eines absoluten Streuungsmaßes S und eines Lokalisationsmaßes M (wenn $M \neq 0$),

$$(5.1) \quad S_r = \frac{S}{M}$$

sofern S die Maßeinheit der Merkmalswerte hat.

Bemerkungen zu Def. 5.1:

1. Dass absolute Streuungsmaße S die Maßeinheit der Merkmalswerte haben sollten, gewährleistet, dass ein relatives Streuungsmaß dimensionslos ist.

2. Relative Streuungsmaße lassen sich sinnvoll nur für mindestens intervallskalierte Merkmale interpretieren. Vergleicht man Häufigkeitsverteilungen, bei denen sich die Größenordnung der Merkmalswerte stark unterscheidet, dann sind sie aussagefähiger als absolute Streuungsmaße. Beim Konzept der relativen Streuung wird die Größenordnung der Merkmalswerte durch Maße der zentralen Tendenz widergespiegelt. Deshalb soll die für absolute Streuungsmaße im Axiom S4 geforderte Verschiebungsinvarianz hier gerade nicht erfüllt sein. Ein relatives Streuungsmaß besitzt demzufolge Eigenschaften, die bei Disparitätsmaßen als Axiome gefordert werden (vgl. Kapitel 6).

2. Varianz und Standardabweichung

a) Berechnung und Eigenschaften

Varianz und Standardabweichung sind die bekanntesten und am häufigsten benutzten Streuungsmaße. Sie sind aus quadrierten Abständen der Merkmalswerte vom arithmetischen Mittel gebildet (oben Konstruktionsprinzip Nr.1 genannt). Man kann die Varianz aber auch in ein Vielfaches

der Summe der quadrierten Abweichungen der Merkmalswerte untereinander umformen (vgl. Satz 5.1).

Def. 5.2: Varianz und Standardabweichung

a) Die Varianz s^2 eines mindestens intervallskalierten Merkmals X ist, wenn sie aus den einzelnen Merkmalswerten $x_1,x_2,...,x_n$ berechnet wird (ungewogener Ansatz), gegeben durch

$$(5.2) \quad s^2 = \frac{1}{n} \Sigma (x_v - \bar{x})^2 \qquad v = 1,2,...,n$$

und wenn sie aus einer Häufigkeitsverteilung (nicht aber bei klassierter Verteilung), d.h. aus den Merkmalsausprägungen $x_1,x_2,...,x_m$ berechnet wird (gewogener Ansatz), gilt

$$(5.3) \quad s^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2 n_i = \Sigma (x_i - \bar{x})^2 h_i . \qquad i = 1,2,...,m$$

b) Die positive Quadratwurzel aus der Varianz heißt Standardabweichung s

$$(5.4) \quad s = + \sqrt{s^2} .$$

Bemerkungen zur Def. 5.2:

1. Die Varianz s^2 und die Standardabweichung s erfüllen die Axiome S1 und S2. Gilt für alle v (v=1,2,...,n) $x_v = \bar{x}$, so folgt $s^2 = s = 0$. Falls es auch nur einen Wert x_v gibt, der nicht identisch mit $\bar{x}$ ist, so folgt hieraus $s^2 > 0$.

2. Die Gültigkeit des Axioms S3 ergibt sich unmittelbar aus Satz 5.1, nach dem die Varianz s^2 als die $(1/n^2)$-fache Summe der Abweichungsquadrate $(x_i-x_j)^2$, i<j, dargestellt werden kann.

3. Mit $y_v = a + bx_v$ für alle v und $b \neq 0$ ist die Varianz s^2_y des zum Merkmal (zur Variablen) Y transformierten Merkmals X durch

$$s_y^2 = \frac{1}{n} \Sigma (y_v - \bar{y})^2 = \frac{1}{n} \Sigma [a+bx_v - (a+b\bar{x})]^2 = b^2 s_x^2$$

und die Standardabweichung s_y durch $s_y = |b| s_x$ gegeben. Mithin ist das Axiom S4 erfüllt.

4. Die Standardabweichung s ist das quadratische Mittel der Abweichungen $(x_v - \bar{x})$ der Merkmalswerte vom arithmetischen Mittel. Sie ist

besonders anschaulich im Falle [annähernd] normalverteilter Variablen (Lage der Wendepunkte) [Induktive Statistik].

5. Nach dem Verschiebungssatz (Satz 5.2) kann die Varianz bei Einzelbeobachtungen auch wie folgt geschrieben werden:

$$(5.5) \quad s^2 = \frac{1}{n} \Sigma\, x_v^2 - \bar{x}^2$$

bzw. bei einer Häufigkeitstabelle:

$$(5.6) \quad s^2 = \frac{1}{n} \Sigma\, x_i^2\, n_i - \bar{x}^2 = \Sigma\, x_i^2\, h_i - \bar{x}^2$$

Um die Auswirkungen von Rundungsfehlern zu begrenzen, empfiehlt sich insbesondere bei der Entwicklung von Computerprogrammen s^2 nach Gl. 5.5 bzw. 5.6 zu berechnen.

6. Der Verschiebungssatz gem. Bem. Nr. 5 ist ein Spezialfall des Steinerschen Verschiebungssatzes. Die Varianz s^2 läßt sich danach auch mittels der folgenden Beziehung berechnen:

$$(5.7) \quad s^2 = \frac{1}{n} \Sigma\, (x_i - c)^2\, n_i - (\bar{x} - c)^2 .$$

Hierbei ist c eine beliebige reelle Zahl. Der erste Summand auf der rechten Seite von Gl. 5.7 ist die um c berechnete Varianz, die man mit $s^2{}_c$ bezeichnen kann. Zwischen s^2 (oder s_x^2) und s_c^2 besteht nach Gl.5.7 die folgende Beziehung:

$$(5.7a) \quad s_x^2 = s_c^2 - (\bar{x} - c)^2 \qquad \text{(Steinerscher Verschiebungssatz).}$$

Aus Gl. 5.7a ist unmittelbar zu erkennen:

- die Minimumeigenschaft des arithmetischen Mittels (s_x^2 ist minimal für $\bar{x} = c$) und
- mit c = 0 erhält man Gl. 5.5 als Spezialfall des Verschiebungssatzes von Steiner.

7. Eine weitere Darstellungsart der Varianz wird im Satz 5.3 angegeben.

8. Ersetzt man das arithmetische Mittel $\bar{x}$ in s^2 durch einen anderen Mittelwert M, so spricht man von der **mittleren quadratischen Abwei-**

chung $s_M^2 = n^{-1} \Sigma (x_v - M)^2$. Aus der Minimumeigenschaft des arithmetischen Mittels folgt, dass $s^2 \leq s_M^2$.

9. In Satz 5.4 wird der Einfluß einer zusätzlichen Beobachtung auf die Varianz untersucht. Danach reduziert die neue Beobachtung x_{n+1} die Varianz, falls ihr Abstand vom arithmetischen Mittel das $\sqrt{(n+1)/n}$-fache der ursprünglichen Standardabweichung unterschreitet. Ein Ausreißer kann dagegen die Varianz über alle Grenzen hinaus wachsen lassen, was zeigt, dass s^2 keine resistente Maßzahl der Streuung ist.

10. Eine wichtige Eigenschaft der Varianz ist die im Satz 5.5 gezeigte **Streuungszerlegung**. Danach läßt sich die Varianz s^2 der Variablen X für die aus r Teilgesamtheiten mit den Umfängen $n_1,n_2,...,n_r$ zusammengesetzten Gesamtheit zerlegen in eine externe Varianz s_{ext}^2 und eine interne Varianz s_{int}^2, so dass gilt

(5.8) $\quad s^2 = s_{ext}^2 + s_{int}^2 \,.$

Die externe und die interne Varianz sind jeweils gewogene Mittelwerte. Und zwar ist die externe Varianz

(5.9) $\quad s_{ext}^2 = \Sigma\, h_k(\bar{x}_k - \bar{x})^2 \quad \text{mit } h_k = \dfrac{n_k}{n}$

ein gewogenes Mittel der quadrierten Abstände zwischen den r Mittelwerten der Teilgesamtheiten und dem Gesamtmittelwert $\bar{x}$.
Die interne Varianz ist demgegenüber das gewogene Mittel der Varianz s_k^2 der Teilgesamtheiten

(5.10) $\quad s_{int}^2 = \Sigma\, h_k s_k^2$

mit den relativen Häufigkeiten h_k als Gewichte.

Die in Satz 5.5 dargestellte Varianzzerlegung ist eine Verallgemeinerung der Berechnung der Varianz aus einer Häufigkeitsverteilung nach Gl. 5.3. Setzt man die Subskripte k und i gleich, so dass $\bar{x}_k = x_i$ und r=m, so geht (5.8) in (5.3) über, da dann annahmegemäß alle n_i Einheiten mit der Merkmalsausprägung x_i gleich $\bar{x}_k$ sind (mit $n_i \geq 0$) und dann die Varianzen s_k^2 verschwinden und damit auch s_{int}^2).
Die Bedeutung des Satzes 5.5 ergibt sich ferner daraus, dass mit ihm das Verhalten der Varianz im Falle von Zerlegung und Aggregation (vgl. auch Bem. 11) ersichtlich wird.

- Bei der **Zerlegung** wird versucht, zu einer Kausalinterpretation zu gelangen, indem man die Beiträge verschiedener "Variationsquellen" zur Gesamtvarianz ermittelt (vgl. z.B. das Bestimmtheitsmaß in Kap. 7).
- Bei der **Aggregation** geht es nicht nur darum, die Varianz für die Gesamtheit zu berechnen, sondern auch zu zeigen, in welchem Maße ihr Wert von der Struktur der Gesamtmasse (d.h. den Anteilen der Teilmassen an der Gesamtmasse) abhängt.

Zwar hat die Varianz nicht die Aggregationseigenschaft, die in Kapitel 2 gefordert wurde, da sie zusätzlich noch den Summanden s_{ext}^2 enthält. Jedoch läßt die Zerlegung in eine externe und interne Varianz Interpretationsmöglichkeiten zu, wie sie bei einer entsprechenden Zerlegung anderer Maßzahlen der Streuung nicht gegeben wären, weshalb die Varianz auch anderen, einfach konstruierten und anschaulicher zu interpretierenden Streuungsmaßen (z.B. der durchschnittlichen Abweichung) vorgezogen wird.

11. Klassierte Daten

Liegen die Daten als klassierte Verteilung mit den Größenklassen k=1,2,...,r vor, so ergibt sich die Gesamtvarianz aufgrund der Streuungszerlegung mit

$$(5.11) \quad s^2 = \Sigma\, h_k(\bar{x}_k - \bar{x})^2 + \Sigma\, h_k s_k^2 = s_{ext}^2 + s_{int}^2$$

wobei s_k^2 die Varianz innerhalb der k-ten Klasse ist.
Häufig sind aber bei einer klassierten Verteilung die Einzelwerte nicht mehr bekannt. Während sich die (wahren) arithmetischen Mittelwerte $\bar{x}_k$ der Klassen durch die Klassenmitten m_k approximieren lassen, hat man für die Varianzen s_k^2 keine unmittelbar naheliegenden Näherungsgrößen. Sofern die Streuung innerhalb der Klassen im Vergleich zur Streuung zwischen den Klassen vernachlässigbar gering ist, kann man

$$(5.11a) \quad s_m^2 = \Sigma\, (m_k - m)^2\, h_k$$

(mit m als wahrem oder geschätzten Gesamtmittelwert) als Näherung für die Varianz s^2 verwenden. Im allgemeinen wird man sich in der Praxis mit der Näherung (5.11a) zur Abschätzung der Varianz einer klassierten Verteilung zufrieden geben müssen. Spezielle Korrekturen sehen Informationen über die Verteilung der Merkmalswerte innerhalb der Klassen vor.

Die beiden bekanntesten Korrekturmöglichkeiten sind:

a) Korrektur bei Gleichverteilung innerhalb der Klassen

Sofern die Merkmalswerte innerhalb der Klassen gleichverteilt sind, wird die Varianz s^2 durch s_m^2 unterschätzt. In diesem Fall ist die Varianz durch

(5.13) $$s^2 = \Sigma\, n_k \left[(m_k - m)^2 + \frac{b_k(n_k^2 - 1)}{12n_k^2} \right]$$

mit der Klassenbreite b_k der k-ten Klasse (k = 1,2,...,r) gegeben.

b) Sheppard-Korrektur (Dreiecksverteilung innerhalb der Klassen)

Falls sich die Merkmalswerte innerhalb der Klassen auf die Klassenmitten konzentrieren oder um die Klassenmitte dreiecksverteilt sind, wird bei gleichen Klassenbreiten b die Korrektur von Sheppard empfohlen. Danach ist

(5.14) $$SK = \frac{b^2}{12}$$ (SK = Sheppard-Korrektur)

von s_m^2 subtrahiert.

12. Häufig verwendet man auch die Formel

$$\hat{\sigma}^2 = \frac{1}{n-1} \Sigma\, (x_v - \bar{x})^2 \qquad (v = 1,2,...,n)$$

als Varianz anstelle der Gl. 5.2

$$s^2 = \frac{1}{n} \Sigma\, (x_v - \bar{x})^2$$

Die Formel für $\hat{\sigma}^2$ ist der (erwartungstreue) Schätzwert für die Varianz der Grundgesamtheit (die σ^2 genannt werden soll) aufgrund der Daten einer Stichprobe, während Gl. 5.2 die Varianz der Daten der Stichprobe (oder nach dem gleichen Muster gerechnet, die Varianz der Grundgesamtheit) wiedergibt.

Beispiel 5.2:

Man berechne Varianz und Standardabweichung aus der folgenden Tabelle der durchschnittlichen Bruttomonatsverdienste männlicher Angestellter in Industrie und Handel nach ("alten") Bundesländern.

Bundesland	Verdienst	Bundesland	Verdienst
Schleswig Holstein	3986	Berlin (West)	4348
Niedersachsen	4081	Nordrhein Westfalen	4408
Saarland	4158	Hessen	4428
Bayern	4246	Baden Württemberg	4509
Bremen	4254	Hamburg	4766
Rheinland Pfalz	4285		

(Quelle: Statist. Jahrb. 1989, S.490)

Lösung 5.2:

Bundesland	x_v	$x_v - \bar{x}$	$(x_v - \bar{x})^2$	x_v^2
Schleswig Holstein	3.986	-329	108.241	15.888.196
Niedersachsen	4.081	-234	54.756	16.654.561
Saarland	4.158	-157	24.649	17.288.964
Bayern	4.246	-69	4.761	18.028.516
Bremen	4.254	-61	3.721	18.096.516
Rheinland-Pfalz	4.285	-30	900	18.361.225
Berlin(West)	4.348	33	1.089	18.905.104
Nordrhein-Westfalen	4.408	93	8.649	19.430.464
Hessen	4.428	113	12.769	19.697.184
Baden Württemberg	4.509	194	37.636	20.331.081
Hamburg	4.766	451	203.401	22.714.756
Summe	47.469	4*	460.572	205.306.567

* wegen Rundungsfehler 4 statt 0 (denn der wahre Mittelwert ist 4315,3636 und nicht 4315, wie in den Spalten $x_v - \bar{x}$ und $(x_v - \bar{x})^2$ gerechnet wurde).

$n = 11$

$\bar{x} = 47469/11 = 4315$

$s^2 = \frac{1}{n} \Sigma(x_v - \bar{x})^2 = 460572/11 = 41870{,}182$ und $s = 204{,}62$DM

andere Berechnungsweise (Gl. 5.6):

$s^2 = \frac{1}{n} \Sigma x_v^2 - \bar{x}^2 = 205306567/11 - (4315{,}3636)^2 = 41870{,}363$ DM^2,

(Variationskoeffizient [Gl. 5.51] V = 0,047 also 4,7%)

Beispiel 5.3:

Beispiel für eine Lineartransformation

Die 200 Beschäftigten einer Arbeitsstätte erhalten einen monatlichen Durchschnittslohn von 2.200 DM mit einer Standardabweichung von s = 800 DM. Aufgrund einer Lohnverhandlung soll das Monatsgehalt jedes Beschäftigten um 10% angehoben werden, und es soll in Zukunft jedes Jahr jedem Beschäftigten ein Urlaubsgeld in Höhe von 960 DM gewährt werden. Wie ändern sich Mittelwert, Standardabweichung und Varianz der Gehälter der Beschäftigten?

Lösung 5.3:

Es liegt eine Lineartransformation vor: die bisherigen Gehälter x_v (v=1,2,...,200) werden zu den "neuen" Gehältern y_v transformiert nach Maßgabe der Transformation $y_v = a + bx_v$ mit a = 960/12 = 80 und b = 1,1.

Man erhält damit das arithmetische Mittel $\bar{y} = 2500$ die Varianz $s_y^2 = (1,1)^2(800)^2 = (880)^2 = 774400$ und die Standardabweichung $s_y = 880$.

b) Sätze über die Varianz

Satz 5.1: Abweichungsquadrate

Die Varianz s^2 läßt sich als die durch n^2 geteilte Summe der Abstandsquadrate aller Merkmalswerte untereinander darstellen:

$$(5.15) \quad s^2 = \frac{1}{n^2} \sum_{i<j} (x_i - x_j)^2 \qquad (\text{mit } i,j = 1,2,...,n)$$

Beweis:

Ausgehend von Gl. 5.5 nehmen wir folgende Umformung vor

$$s^2 = n^{-1} \sum x_i^2 - (n^2)^{-1} (\sum x_i)^2 = (n^2)^{-1} [\, \sum n x_i^2 - (\sum x_i)^2 \,]$$

$$= \left[\sum (n-1) x_i^2 - \sum_{i} \sum_{\substack{j \\ i \neq j}} x_i x_j \right] / n^2 = \frac{1}{n^2} \left[\sum (n-1) x_i^2 - \sum_{i<j} 2 x_i x_j \right],$$

woraus (5.15) unter Verwendung der binomischen Formel für die Summenglieder direkt folgt.

Die Gleichung 5.15 ist gleichbedeutend mit

$$(5.16) \quad s^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2$$

Die obige Doppelsumme stellt die Summe aller Elemente der folgenden Matrix dar

$$\begin{bmatrix} (x_1 - x_1)^2 & (x_1 - x_2)^2 & \dots & (x_1 - x_n)^2 \\ (x_2 - x_1)^2 & (x_2 - x_2)^2 & \dots & (x_2 - x_n)^2 \\ \vdots & \vdots & \vdots & \vdots \\ (x_n - x_1)^2 & (x_n - x_2)^2 & \dots & (x_n - x_n)^2 \end{bmatrix}$$

Satz 5.2: Verschiebungssatz

Bei Einzelwerten (v=1,2,...,n) kann man die Varianz s^2 in der Form

$$(5.5) \quad s^2 = \frac{1}{n} \Sigma x_v^2 - \bar{x}^2$$

darstellen bzw. im Falle einer Häufigkeitsverteilung

(5.6) $s^2 = \Sigma\, x_i^2 h_i - \bar{x}^2$.

Beweis:

Ausgehend von Gl. 5.3 erhält man für Einzelwerte

$$s^2 = \frac{1}{n}\,\Sigma\,(x_v^2 - 2x_v\bar{x} + \bar{x}^2) = \frac{1}{n}\,[\,\Sigma x_v^2 - 2\bar{x}\Sigma\, x_v + n\bar{x}^2]$$

woraus wegen $\bar{x} = (Sx_v)/n$ Gl. 5.5 folgt. Gl. 5.6 ist analog zu beweisen.

Satz 5.3

Die Varianz ist auch darstellbar als

(5.17) $s^2 = \frac{1}{n}\sum (x_v - \bar{x})x_v$ bzw.

(5.18) $s^2 = \frac{1}{n}\sum (x_i - \bar{x})x_i h_i$.

(Eine ähnliche Beziehung gilt auch für die Kovarianz; vgl. Gl. 7.17)

Beweis:

Aus Gl. 5.5 folgt mit $\bar{x} = (\Sigma x_v)/n$

$$s^2 = \frac{1}{n}\left[\Sigma x_v^2 - \bar{x}\,\Sigma\, x_v\right] = \frac{1}{n}\,\Sigma\, x_v(x_v - \bar{x})$$ also Gl.5.17.

In gleicher Weise zeigt man Gl. 5.18.

Satz 5.4: Einfluss einer zusätzlichen Beobachtung

Sei s_{n+1}^2 die Varianz der um die Beobachtung x_{n+1} erweiterten Daten und x_{n+1} das arithmetische Mittel der Werte $x_1, x_2, \ldots, x_n, x_{n+1}$ ($s_n^2 = s^2$ ist die Varianz und $\bar{x} = \bar{x}_n$ das arithmetische Mittel der ursprünglichen Zahlenfolge $x_1, x_2, \ldots, x_n$), dann ist:

(5.19) $s_{n+1}^2 = \frac{1}{n+1}\sum_{v=1}^{n+1}(x_v - \bar{x}_{n+1})^2$.

Die sog. ***Sensitivitätsfunktion*** SF der Varianz s^2, die mit

(5.20) $SF(\bar{x}_{n+1}, s_n^2) = (n+1)(s_{n+1}^2 - s_n^2)$

definiert ist und die (n+1)-fache Veränderung (nicht notwendig Zunahme) der Varianz darstellt, ist durch den Ausdruck

$$(5.21) \quad SF = \frac{n(x_{n+1} - \bar{x})^2}{(n+1)} - s_n^2$$

gegeben.

Beweis:

Aus der Definition von s_{n+1}^2

$$s_{n+1}^2 = \frac{1}{n+1}\left[(x_{n+1} - \bar{x}_{n+1})^2 + \Sigma\,(x_v - \bar{x}_{n+1})^2)\right]$$

erhält man

$$(n+1)s_{n+1}^2 = (x_{n+1} - \bar{x}_{n+1})^2 + \Sigma[(x_v - \bar{x}) + (\bar{x} - \bar{x}_{n+1})]^2$$

(v = 1,2,...,n) und wegen der Schwerpunkteigenschaft des arithmetischen Mittels

$$(n+1)s_{n+1}^2 = (x_{n+1} - \bar{x}_{n+1})^2 + \Sigma(x_v - \bar{x})^2 + \Sigma(\bar{x} - \bar{x}_{n+1})^2$$

Daraus erhält man nach einigen Umformungen

$(n+1)s_{n+1}^2 = n(x_{n+1} - \bar{x})^2/(n+1) + ns^2{}_n$ und damit Gl. 5.21.

Interpretation:

Die Varianz $s^2 = s_n^2$ verändert sich umso mehr durch einen hinzukommenden Wert x_{n+1}, je stärker x_{n+1} vom bisherigen Mittelwert abweicht. Man sieht ferner, dass die Varianz durch einen hinzukommenden Wert nicht notwendig größer werden muss. Dies sei anhand des folgenden Beispiels demonstriert.

Beispiel 5.4:

Der Zusammenhang der Gl. 5.21 soll anhand der folgenden Merkmalswerte verifiziert werden:

ursprüngliche Beobachtungen (n=4): 2,3,5,6;
(damit ist $\bar{x}_n = 4$ und $s^2{}_n = s^2{}_4 = 10/4 = 2{,}5$)
der neu hinzukommende Wert sei nun

a) $x_{n+1} = x_5 = 4$ bzw.
b) $x_{n+1} = x_5 = 9$.

Lösung 5.4:

Nach Gl. 5.21 gilt

$$(n+1)(s_{n+1}^2 - s_n^2) = \frac{n}{n+1}(x_{n+1} - \bar{x})^2 - s_n^2$$

Für das Zahlenbeispiel erhält man im Fall a): $s_{n+1}^2 = 2$ und damit für SF nach Gl. 5.20: $(n+1)(s_{n+1}^2 - s_n^2) = 5(2-2,5) = -2,5$
und für SF nach Gl. 5.21

$$\frac{n(x_{n+1} - \bar{x})^2}{n+1} - s_n^2 = 4(4-4)^2/5 - 2,5 = -2,5$$

also das gleiche Ergebnis, wie nach Satz 5.4 ja auch zu erwarten war und was übrigens auch demonstriert, dass die Varianz durch Hinzutreten eines weiteren Merkmalswerts nicht notwendig größer werden muss. Dies ist jedoch im Fall b) der Fall. Dort gilt

$s_{n+1}^2 = 30/5 = 6$ und für Gl. 5.21 erhält man die eingesetzten Zahlen:
$5(6-2,5) = (4/5)(9-4)^2-2,5$.

Das bestätigt auch Bem. 9 zu Def. 5.2, wonach es darauf ankommt ob $x_{n+1} - \bar{x}$ größer oder kleiner ist als $S_n \sqrt{(n+1)/n}$. Im Falle a ist die Differenz kleiner als $\sqrt{2,5} \cdot \sqrt{5/4} = 1,77$ nämlich Null (die Varianz verringert sich) und im Fall b ist sie 9 - 4 = 5 und damit größer als 1,77 und die Varianz vergrößert sich. Träte als fünfter Wert der Wert $x_{n+1} = 5,77$ hinzu, so würde die Varianz gleich bleiben: $s_{n+1}^2 = s_n^2 = 2,5$. Ist x_{n+1} kleiner (etwa 4) so wird sie kleiner, ist x_{n+1} größer (etwa 9) so wird sie größer.

Satz 5.5: Streuungszerlegung

Gegeben seien r Teilgesamtheiten $G_1, G_2, ..., G_r$ der Gesamtheit G, die eine Zerlegung bilden (r = 1,2,..,r). Die Umfänge der Teilgesamtheiten seien n_k mit $\Sigma n_k = n$ (so dass $h_k = n_k/n$). Die Beobachtung x_i sei jeweils ein Element der k-ten Teilgesamtheit ($x_i \in G_k$). Die Teilgesamtheiten haben die arithmetischen Mittel $\bar{x}_k$ und die Varianzen

$$(5.22) \quad s_k^2 = \frac{1}{n_k} \sum_{x_i \in G_k} (x_i - \bar{x}_k)^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (x_i - \bar{x}_k)^2$$

Dann gilt für die Varianz s^2 der Gesamtheit G

$$(5.12) \quad s^2 = \Sigma h_k s_k^2 + \Sigma h_k (\bar{x}_k - \bar{x})^2 .$$

Beweis:

Die Varianz s^2 läßt sich in der Form

$s^2 = \frac{1}{n} \sum_{k}^{r} \sum_{i=1}^{n_k} (x_i - \bar{x})^2$ schreiben, was sich durch Nullergänzung zu

$s^2 = \frac{1}{n}\sum_k \sum_i [(x_i - \bar{x}_k)^2 + (\bar{x}_k - \bar{x})]^2$ umformen läßt.

Ausmultiplizieren und Summieren ergibt dann wegen der Schwerpunkteigenschaft des arithmetischen Mittels

$$s^2 = \frac{1}{n}\sum_k \sum_i (x_i - \bar{x}_k)^2 - \frac{2}{n}\sum_k \left[(\bar{x}_k - \bar{x}) \sum_i (x_i - \bar{x}_k) \right] + \frac{1}{n}\sum_k \sum_i (\bar{x}_k - \bar{x})^2$$

Der erste Summand ergibt die interne Varianz, der zweite verschwindet wegen der Schwerpunkteigenschaft von $\bar{x}_k$ und der dritte ist die externe Varianz.

Beispiel 5.5:

Streuungszerlegung

Die Firma B – GmbH & Co KG hatte Probleme mit ihrer Belegschaft und ließ eine Untersuchung durch einen BWL-Professor und einen Statistiker durchführen:

1. Der BWL-Prof. stellte nach längerer intensiver Forschungsarbeit fest, dass die Unternehmung (nicht: das Unternehmen) B ein komplexes soziotechnisches System mit einer Triade von Subsystemen ist, in welchem eine nicht näher bestimmte Anzahl von Wirtschaftssubjekten mit einer signifikant differierenden Wirkungsintensität dergestalt operieren, dass emotionale Dysfunktionalitäten virulent waren, über deren Ausmaß aber ohne weitere hochkomplexe Evaluation kaum genauere Aussagen möglich sind und über die strategisch und situativ zu entscheiden ist.

2. Der Statistiker stellte fest, dass in den drei Betrieben des Unternehmens die insgesamt 2000 Beschäftigten sehr unterschiedlich verdienten, so dass in der Belegschaft Unfrieden herrschte. Er ermittelte die folgenden Zahlen:

Betrieb	Anzahl der Beschäftigten	Durchschnitts-verdienst ($\bar{x}$)	Standardab-weichung
1	500	2400	400
2	800	3000	600
3	700	2800	500

Man bestimme Mittelwert und Varianz der Verdienstverteilung des gesamten Unternehmens!

Lösung 5.5:

Mittelwert $\bar{x} = 2400 \cdot 0{,}25 + 3000 \cdot 0{,}4 + 2800 \cdot 0{,}35 = 2780$
interne Varianz s_{int}^2 :
$\Sigma h_j \cdot s_j^2 = 0{,}25 \cdot (400)^2 + 0{,}4 \cdot (600)^2 + 0{,}35 \cdot (500)^2 = 271500$

(Standardabweichung $s_{int.}$ 521,06; sie liegt zwischen 400 und 600)
externe Varianz s^2_{ext}:
$0{,}25\cdot(2400-2780)^2 + 0{,}4\cdot(3000-2780)^2 + 0{,}35\cdot(2800-2780)^2 = 55600$.
Gesamtvarianz 271500 + 55600 = 327100 (Standardabw. 571,93).

3. Andere Maße der absoluten Streuung

a) Durchschnittliche Abweichung und Medianabweichung

Def. 5.3: durchschnittliche- und Medianabweichung

a) Mit $a_1,a_2,...,a_n$ seien die absoluten Abweichungen der Merkmals**werte** $x_1,x_2,...,x_n$ eines mindestens intervallskalierten Merkmals X vom Median $\tilde{x}_{0,5}$ bezeichnet

(5.23) $a_v = |x_v - \tilde{x}_{0,5}| \quad v=1,2,...,n$

und $a_1,a_2,...,a_m$ seien die entsprechenden absoluten Abweichungen der Merkmals**ausprägungen** $x_1,x_2,...,x_m$

(5.24) $a_i = |x_i - \tilde{x}_{0,5}| \quad i=1,2,...,m.$

Dann ist das arithmetische Mittel der absoluten Abweichungen vom Median

(5.25) $d_x = \frac{1}{n} \Sigma a_v$ bei Einzelwerten bzw.

(5.26) $d_x = \Sigma a_i h_i$ bei Häufigkeitsverteilungen

die **durchschnittliche Abweichung** (vom Median). Üblich ist auch die Bezeichnung mittlere - oder **mittlere absolute Abweichung** (mean absolute deviation) .

b) Der Median (Zentralwert) der n absoluten Abweichungen a_v heißt **Medianabweichung** m_x. Bei Einzelwerten ist m_x der (n+1)/2 - te Wert, bzw. der Mittelwert aus dem n/2 - ten und dem folgenden Wert in einer der Größe nach geordneten Folge der absoluten Abweichungen a_v:

(5.27) $m_x = \begin{cases} a_{(n+1)/2} & \text{falls n ungerade} \\ ½[a_{(n/2)} + a_{(n/2+1)}] & \text{falls n gerade} \end{cases}$

c) Ein selteneres, in erster Linie in der Technik angewandtes Streuungsmaß ist a_{max}, die **maximale** absolute **Abweichung** a_v. Da das Maximum ein Grenzfall des

Potenzmittels ist, kann man auch die maximale Abweichung als Streuungsmaß nach dem Konstruktionsprinzip Nr. 1 auffassen.

Bemerkungen zu Def. 5.3:

1. Ebenso wie die Varianz und die Standardabweichung sind die mittlere absolute Abweichung und die Medianabweichung Maßzahlen, die nach dem ersten Konstruktionsprinzip, also unter Verwendung der Abstände der Beobachtungswerte von einem Lagemaß gebildet werden.

2. Da in d_x und m_x die absoluten Abweichungen der Merkmalswerte bzw. -ausprägungen vom Median einbezogen werden, sind beide Maßzahlen nichtnegativ.
   Gilt $x_1 = x_2 = ... = x_n$, so ist $\tilde{x}_{0,5} = x_v$ für alle v=1,2,...,n und somit $d_x = m_x = 0$, so dass d_x und m_x das Axiom S1 erfüllen.

3. Sobald x_i und x_j ungleich sind, so dass $|x_i - \tilde{x}_{0,5}| > 0$ oder/und $|x_j - \tilde{x}_{0,5}| > 0$ muss auch $d_x > 0$ sein.
   Mithin gilt das Axiom S2 für d_x. Allerdings sind entartete Fälle denkbar, in denen m_x Axiom S2 nicht erfüllt (Beispiel 5.7 am Ende dieser Bemerkungen).

4. Verhalten von m_x und d_x bei Lineartransformationen $y_v = a+bx_v$:
   $|y_v - \tilde{y}_{0,5}| = |a + bx_v - (a + b\tilde{x}_{0,5})| = |bx_v - b\tilde{x}_{0,5}| = |b|\, a_v$.

   Daraus folgt $d_y = |b|\, d_x$ und $m_y = |b|\, m_x$, womit die Gültigkeit des Axioms S4 gezeigt ist.

5. Verschiedentlich wird auch anstelle von d_x die weniger übliche mittlere absolute Abweichung um $\bar{x}$ verwendet, die wir d_x^* nennen wollen:

$$(5.28) \qquad d_x^* = \begin{cases} 1/n \sum |x_i - \bar{x}| & \text{bei Einzelwerten} \\ \sum |x_i - \bar{x}| h_i & \text{bei Häufigkeitsverteilungen} \end{cases}$$

   Aus der Minimumeigenschaft von $\tilde{x}_{0,5}$ folgt (5.29) $d_x \le d^*,_x$.

6. Im Vergleich zur Standardabweichung gilt (5.30) $d_x \le d^*,_x \le s$.

7. Der Vorteil der durchschnittlichen Abweichung ist ihre besondere Anschaulichkeit: d_x ist die durchschnittliche Entfernung einer Beobachtung vom Median. Die Standardabweichung ist zwar auch eine mitt-

lere Abweichung, aber das quadratische Mittel ist weniger allgemeinverständlich.

8. Als Faustregel gilt bei eingipfligen, der Normalverteilung ähnlichen Verteilungen $d_x = 0{,}8s$. Gegenüber der Varianz und der Standardabweichung spielt die mittlere absolute Abweichung wegen der analytischen Unhandlichkeit absoluter Abweichungen nur eine untergeordnete Rolle. Insbesondere in der induktiven Statistik dominieren Varianz und Standardabweichung. In die Diskussion gekommen ist die mittlere absolute Abweichung jedoch durch eine Untersuchung von Tukey (1960). Danach ist d_x bei "Verunreinigungen" durch "schlechte" Daten der Standardabweichung überlegen.

9. Die Medianabweichung m_x wird in der explorativen Datenanalyse (EDA) als "Hilfsskalenschätzer" verwendet (In der angloamerikanischen Literatur wird "scale" in diesem Zusammenhang für den Begriff Streuung benutzt). Als Analogon zum Median bei den Lokalisationsmaßen gilt m_x als besonders robust.

Beispiel 5.6:

Es sei die folgende Altersverteilung von n = 9 Personen gegeben (in Einzelwerten): 21, 25, 34, 39, 43, 52, 64, 72, 80. Das arithmetische Mittel $\bar{x}$ dieser Verteilung beträgt 47,78. Berechnen Sie die mittlere absolute Abweichung um den Median und um das arithmetische Mittel sowie die Medianabweichung!

Lösung 5.6:

Für den Median erhält man $\tilde{x}_{0.5} = 43$, woraus sich eine mittlere absolute Abweichung (vom Median) von $d_x = 16{,}56$ ergibt, denn die absoluten Abweichungen betragen:

$|21 - \tilde{x}_{0.5}| = |21 - 43| = 22;$
$|25 - 43| = 18;\ |34 - 43| = 9;\ |39 - 43| = 4;\ |43 - 43| = 0;$
$|52 - 43| = 9;\ |64 - 43| = 21;\ |72 - 43| = 29;\ |80 - 43| = 37$

und die Summe dieser (zur Berechnung von m_x bereits der Größe nach geordneten) absoluten Abweichungen beträgt 0 + 4 + 9 + 9 + 18 + 21 + 22 + 29 + 37 = 149. Die durchschnittliche (mittlere) absolute Abweichung beträgt somit 149/9 = 16,556. Daraus ergibt sich auch, dass die Medianabweichung $m_x = 18$ ist.
Die mittlere absolute Abweichung um $\bar{x}$ beträgt $d^*_{,x} = 17{,}09$.

Beispiel 5.7:

Gegeben seien die Merkmalswerte $x_1 = x_2 = x_3 = 0$ und $x_4 = 1$.
Liegt eine Streuung vor? Wie groß ist die Medianabweichung?

Lösung 5.7:

Es liegt offensichtlich eine Streuung vor, da nicht alle vier Beobachtungen identisch sind. Der Median ist $\tilde{x}_{0,5} = 0$ und man erhält die Abweichungen $a_1 = a_2 = a_3 = 0$ und $a_4 = 1$. Die Medianabweichung ist damit Null, also $m_x = 0$, obgleich eine gewisse Streuung vorliegt.

b) Spannweite, Quartilsabstand und Quantilsabstände

Die im folgendenen definierten Streuungsmaße sind weniger gebräuchlich. Es sind Streuungsmaße, die nach dem zweiten Konstruktionsprinzip aus dem Abstand zweier Ordnungsstatistiken gebildet worden sind.

Def. 5.4: Spannweite, Quartilsabstand, Quantilsabstände

a) Die Differenz zwischen dem größten Beobachtungswert $x_{(n)}$ und dem kleinsten $x_{(1)}$ heißt **Spannweite** R (range, Wertebereich, Variationsbreite):

$$(5.31) \quad R = x_{(n)} - x_{(1)}$$

(Die Berechnung von R ist nur bei Einzelwerten, nicht bei Häufigkeitsverteilungen üblich).

b) Der **Quartilsabstand** $Q_{0,25}$ (Interquartilsabstand IQR) ist die Differenz zwischen dem dritten und ersten Quartil (Gl. 5.32) und der **mittlere Quartilsabstand** $\bar{Q}_{0,25}$ (Semiquartilsabstand) ist durch Gl. 5.33 gegeben:

$$(5.32) \quad Q_{0,25} = Q_3 - Q_1 \quad \text{und} \quad (5.33) \quad \bar{Q}_{0,25} = ½\, Q_{0,25}$$

c) Der **Quantilsabstand** (Interquantilsabstand) Q_p ist die Differenz zwischen dem (1-p)-Quantil $\tilde{x}_{1-p}$ und dem p-Quantil $\tilde{x}_p$,

$$(5.34) \quad Q_p = \tilde{x}_{1-p} - \tilde{x}_p \quad \text{mit } 0 < p < 0{,}5 \,,$$

Analog zu Gl.5.36 heißt dann die Maßzahl (5.35) $\bar{Q}_p = ½Q_p$ **mittlerer Quantilsabstand** (Semiquantilsabstand).

Beispiel 5.8:

Man berechne die Spannweite und den mittleren Quartilsabstand für das Beispiel 5.6!

Lösung 5.8:

R = 80 - 21 = 59. Die Quartile sind (mit Interpolation) $Q_1 = 29{,}5$ (der 2,5-te Wert), $Q_2 = 43$ (= Median) und $Q_3 = 68$ (der 7,5-te Wert). Folglich ist der Interquartilsabstand 68 - 29,5 = 38,5 und der mittlere Quartilsabstand 38,5/2 = 19,25.

Bemerkungen zu Def. 5.4:

1. Aufgrund der Differenzenbildung ist unmittelbar einsichtig, dass alle Maßzahlen das Axiom S1 erfüllen. Axiom S2 wird ganz offensichtlich von der Spannweite erfüllt, nicht aber notwendig auch von Q_p und damit den anderen obigen Maßzahlen (vgl. hierzu Bsp. 5.9). Auch Axiom S3 muss nicht notwendig vom mittleren Quartilsabstand erfüllt sein (vgl. für ein konstruiertes, extremes Beispiel Bsp. 5.10).

2. Da die Spannweite nur von den beiden extremen Werten eines Datensatzes abhängt, nutzt sie die in den Daten enthaltene Information unzureichend aus und reagiert äußerst empfindlich auf Ausreißer. Daher wird R als Streuungsmaß kaum verwendet. Allerdings hat die Spannweite eine gewisse Bedeutung bei Ausreißertests und in der statistischen Qualitätskontrolle.

3. Der Quartilsabstand gibt den Bereich an, in den 50% der mittleren Beobachtungswerte fallen. Einerseits wird dadurch ein beträchtlicher Teil der Informationen eines Datensatzes "verschenkt". Gerade deshalb ist aber diese Maßzahl sehr robust, weshalb sie in der explorativen Datenanalyse (vgl. Exkurs über Boxplots) verwendet wird.

4. In der Form $Q_{0,25} = (Q_3-Q_2)+(Q_2-Q_1)$ wird der Quartilsabstand in eine "rechtsseitige" Streuung $(Q_3 - Q_2)$ rechts vom Median $(\tilde{x}_{0,5} = Q_2)$ und eine "linksseitige" Streuung (Q_2-Q_1) aufgespalten, so dass $\bar{Q}_{0,25}$ als Mittelwert der beiden Abstände interpretiert werden kann. Diese rechts- und linksseitige Streuung wird auch zur Konstruktion eines Schiefemaßes herangezogen.

5. Bei normalverteilten Daten ist der Quartilsabstand gleich dem 0,6745-fachen Wert der Standardabweichung.

6. Da die folgenden Streuungsmaße jeweils Spezialfälle des Potenzmittels von Abweichungen darstellen gilt

(5.36) $d_x \le s \le R.$

7. Offensichtlich ist der (mittlere) Quantilsabstand eine Verallgemeinerung des (mittleren) Quartilsabstands. Neben dem Quartilsabstand bieten sich somit eine Reihe weiterer spezieller Streuungsmaße an:

Bekanntlich ist es möglich, Verteilungen in beliebig viele gleichhäufig besetzte Abschnitte (Quantile) zu zerlegen. Bei einer Vier-Teilung spricht man von einer Zerlegung in drei Quartile (Q_1, Q_2, Q_3, wobei Q_2 dem Median entspricht). Hieraus ist der Interquartilsabstand (IQR) abzuleiten. Mit der gleichen Betrachtungsweise könnte man ein Streuungsmaß auf der Basis der vier Quintile, die wir Q_1^*, Q_2^*, Q_3^*, Q_4^*, welche die Verteilung in fünf Abschnitte zerlegen, konstruieren. Die Differenz zwischen dem vierten und dem ersten Quintil wäre dann der **Quintilsabstand**. Näherungswerte für verschiedene Q_p stellen die in der explorativen Datenanalyse verwendeten "spreads" (spr) dar.

Beispiel 5.9:

Gegeben seien die sieben Beobachtungswerte 2,3,3,3,3,3,4. Man bestimme die Spannweite und den mittleren Quartilsabstand.

Lösung 5.9:

Die Spannweite ist R = 4 - 2 = 2. Aber der Quartilsabstand ist in diesem (sehr konstruierten) Beispiel Null (und damit auch der mittlere Quartilsabstand), da $Q_3 = Q_1 = 3$; offenbar erfüllt also dieses Streuungsmaß nicht notwendig die Forderung S2.

Beispiel 5.10:

Gegeben seien die folgenden sieben Beobachtungen 2,3,4,5,6,7,7. Der Wert 3 werde durch den Wert 10 ersetzt. Man bestimme die Spannweite und den mittleren Quartilsabstand für die Reihen:

a) 2,3,4,5,6,7,7 und
b) 2,4,5,6,7,7,10.

Erfüllt der mittlere Quartilsabstand das Axiom S3 ?

Lösung 5.10:

Die Spannweite beträgt

im Fall a) R = 7 - 2 = 5
im Fall b) R = 10 - 2 = 8

Die Voraussetzung des Axioms S3 ist, dass der Ersatz des Wertes 3 durch den Wert 10 zu einer größeren Summe der absoluten Abweichungen der

einzelnen Beobachtungswerte untereinander führt. Das ist in diesem Beispiel gegeben. Die genannte Summe beträgt
im Fall a) 50 und im Fall b) 64.
Die Spannweite wird, wie es Axiom S3 verlangt, nicht kleiner, sondern sogar größer. Der mittlere Quartilsabstand ist aber

im Fall a) $\bar{Q}_{0,25} = ½(7-3) = 2$ und
im Fall b) $\bar{Q}_{0,25} = ½(7-4) = 1,5$.

Exkurs: Boxplot

Ein Boxplot ist eine komprimierte graphische Darstellung eines Datensatzes, die von Tukey (1977) eingeführt worden ist. Eine Box, die durch das erste und dritte Quartil begrenzt und durch den Median geteilt wird, vermittelt einen Überblick über die mittleren 50% der Beobachtungen eines Datensatzes (die Breite der Box ist beliebig). Um die Verhältnisse eines Datensatzes an seinen äußeren Enden einschätzen zu können, werden sog. Zäune (whiskers, adjacent values) festgelegt, deren Abstand von den Quartilen Q_1 und Q_3 durch gestrichelte senkrechte Linien dargestellt wird. Und zwar sind die Zäune durch diejenigen Beobachtungen festgelegt, die gerade noch innerhalb des durch die Quartile und den Quartilsabstand (IQR,$Q_{0,25}$) definierten Intervalls $[w_e,w_u]$ mit

(5.37) $w_e = Q_1 - ½\, Q_{0,25}$ und (5.38) $w_u = Q_3 + ½\, Q_{0,25}$

liegen. Darüber hinaus liegende Beobachtungen sind als mögliche Ausreißer "verdächtig". Sie können einzeln durch ein * als outside values (outlayers) gesondert gekennzeichnet werden. Aus einem Boxplot lassen sich rasch Informationen über die Lokalisation (Lage des Median), Streuung (Höhe der Box) und die Schiefe (Vergleich der beiden Hälften der Box oder der Längen der gestrichelten Linien) eines Datensatzes sowie über evtl. vorliegende Ausreißer ("deviante Beobachtungen") gewinnen.

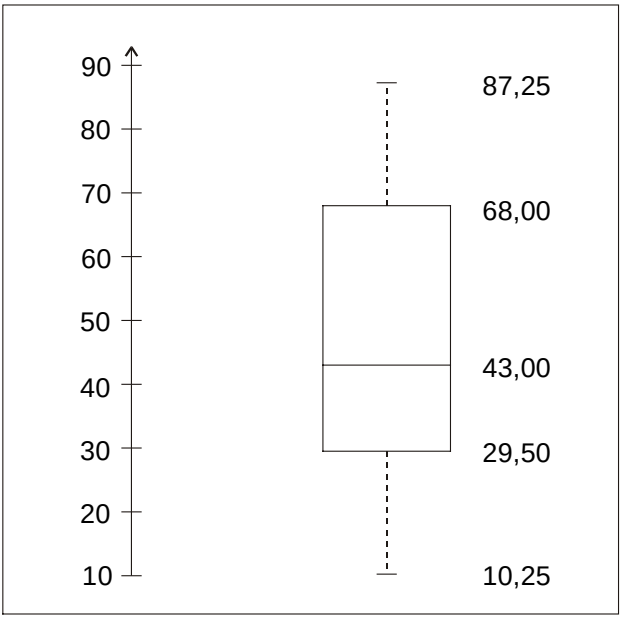
Beispiel 5.11:

Man zeichne den Boxplot für das Beispiel 5.6 (bzw. 5.8).

Lösung 5.11:

Im Beispiel 5.6 (bzw. 5.8) erhält man $Q_1 = 29,5$ $Q_2 = 43$ und $Q_3 = 68$. Ferner ist $Q_{0,25} = 38,5$ und die Zäune liegen bei $w_e = 29,5 - 19,25 = 10,25$ und $w_u = 68 + 19,25 = 87,25$. Es gibt also in diesem Beispiel keine ausreißerverdächtige Werte, weil der kleinste Wert 21 und der größte 80 ist. Ein Bild des Boxplots ist Abb. 5.3.

Abb. 5.3: Boxplot (Beispiel 5.11)



c) Ginis Dispersionsmaß (Ginis mittlere Differenz)

Ginis Dispersionsmaß (nicht zu verwechseln mit dem Disparitätsmaß von Giordano Gini [vgl. Kap. 6]) basiert auf dem dritten der drei Konstruktionsprinzipien von Streuungsmaßen. Es wird also aus den Abständen aller Beobachtungswerte untereinander gebildet.

Def. 5.5: Ginis Streuungsmaß

Für die Merkmalswerte $x_1,x_2,...,x_n$ eines metrisch skalierten Merkmals X ist Ginis Dispersionsmaß (auch mittlere Differenz genannt) gegeben durch

$$(5.39) \quad S_G = \frac{2}{n(n-1)} \sum_{v<w} |x_v - x_w|$$

(bei Einzelwerten v,w = 1,2,...,n) und bei einer Häufigkeitsverteilung durch

$$(5.40) \quad S_G = \frac{2}{n(n-1)} \sum_{i<j} |x_i - x_j| n_{ij} \; .$$

Beispiel 5.12:

Man berechne S_G nach Gl. 5.40 und S^*_G (vgl. Bem. 3 zu Def. 5.5) gem. Gl. 5.42 für die Daten

a) des Beispiels 5.10,
b) des Beispiels 5.6, bzw.5.8.

Lösung 5.12:

zu a) Die Summe der absoluten Abweichungen ist im Beispiel 5.10 bei 7 Werten bereits mit 50 bzw. 64 angegeben worden. Das ist die Doppelsumme

$$\sum_{v<w}|x_v - x_w| = \frac{1}{2}\sum_v\sum_w|x_v - x_w| \quad (\text{mit } v < w)$$

(die Berechnung dieser Summe wird in Teil b demonstriert)
Folglich ist bei 21 Paarvergleichen $S_G = 50/21 = 2{,}381$
bzw. $S_G = 64/21 = 3{,}048$ und

$S^*_G = 100/49 = 2{,}041$ bzw. $S^*_G = 128/49 = 2{,}612$ (weil $49 = 7^2$).

zu b) Datensatz von Beispiel 5.6: 21,25,34,39,43,52,64,72,80. Die folgende Matrix enthält die absoluten Differenzen:

	21	25	34	39	43	52	64	72	80	Summe
21	-	4	13	18	22	31	43	51	59	241
25		-	9	14	18	27	39	47	55	209
34			-	5	9	18	30	38	46	146
39				-	4	13	25	33	41	116
43					-	9	21	29	37	96
52						-	12	20	28	60
64							-	8	16	24
72								-	8	8
80									-	-
Summe										900

Die vollständige Matrix ist symmetrisch mit Nullen in der Hauptdiagonalen. Sie hat insgesamt $9^2 = 81$ Elemente, darunter $9(9-1)/2 = 36$ oberhalb der Hauptdiagonalen. Folglich ist $S_G = 900/36 = 25$ und $S^*_G = 1800/81 = 22{,}22 = (8/9)S_G$.

Bemerkungen zu Def. 5.5:

1. Offenbar erfüllt Ginis Dispersionsmaß das Axiom S1. Dass S_G auch Axiom S2 erfüllt, ergibt sich aus der analogen Betrachtung bei der Varianz (Satz 5.1). Auch Axiom S3 ist unmittelbar einsichtig.

2. Der Faktor 2/n(n-1) ist der reziproke Binomialkoeffizient $\binom{n}{2}$. Dies ist die Anzahl der Paarvergleiche ohne Berücksichtigung der Anordnung [Reihenfolge] von zwei verschiedenen Beobachtungswerten x_v und x_w (bzw. x_i und x_j). Mithin ist S_G das arithmetische Mittel der absoluten Abweichungen der Merkmalswerte untereinander, wobei jede Differenz einmal gerechnet wird.

3. Man kann auch analog zur Varianz eine mittlere Differenz aus allen n^2 möglichen Differenzen zwischen zwei Beobachtungen x_v und x_w, bzw. x_i und x_j bilden:

$$(5.41) \quad S_G^* = \frac{1}{n^2} \sum_{v=1}^{n} \sum_{w=1}^{n} |x_v - x_w|.$$

   Wegen der zusätzlichen Mittelung über die n Null-Differenzen für v = w in S_G^* gilt

$$(5.42) \quad S_G = \frac{n}{n-1} S_G^*.$$

4. Aus der Darstellung der Varianz in Abhängigkeit von den Abständen der Beobachtungswerte (Satz 5.1) untereinander erkennt man, dass

$$\sum_{v<w} (x_v - x_w) = \sum_{v=1}^{n} \sum_{w=1}^{n} (x_v - x_w) = 0$$

   ist, weshalb sich die Wahl der **absoluten** Werte der Abweichungen als sinnvoll erweist.

5. Während die Varianz von extremen Abweichungen aufgrund der Quadrierung sehr stark beeinflusst wird, zeichnet sich Ginis Dispersionsmaß durch eine größere Resistenz gegenüber Ausreißern aus. Deshalb hat S_G in neuerer Zeit wieder eine gewisse Beachtung als ein Konzept zur Konstruktion von "Gini-like" Lokalisations- und Dispersionsmaßen gefunden.

6. Auf einen Zusammenhang zwischen S_G^* und dem Disparitätsmaß von Gini (D_G) wird in Kap. 6 eingegangen.

d) Entropie

Die im folgenden dargestellte Entropie (E) kann als Streuungsmaß für kategoriale (qualitative) Daten aufgefaßt werden, weil sie einige Eigenschaften besitzt, die für ein Streuungsmaß zu fordern sind. Sie eignet sich für nominalskalierte Merkmale, weil sie nur von den relativen Häufigkeiten, nicht aber von den Merkmalswerten abhängig ist. Gerade deshalb

reagiert E aber nicht auf Transformationen der Merkmalswerte und Veränderungen des Wertebereichs von X, was aber bei quantitativen Merkmalen im Widerspruch zur anschaulichen Vorstellung der "Streuung" steht. Die Entropie E spielt auch eine Rolle bei der Disparitätsmessung.

Def. 5.6: Entropie

Für die Häufigkeitsverteilung (x_j,h_j) eines Merkmals X ist die Maßzahl E als Entropie von X definiert ($h_j > 0$):

$$(5.43) \quad E = \Sigma\, h_j\, ld(1/h_j)$$

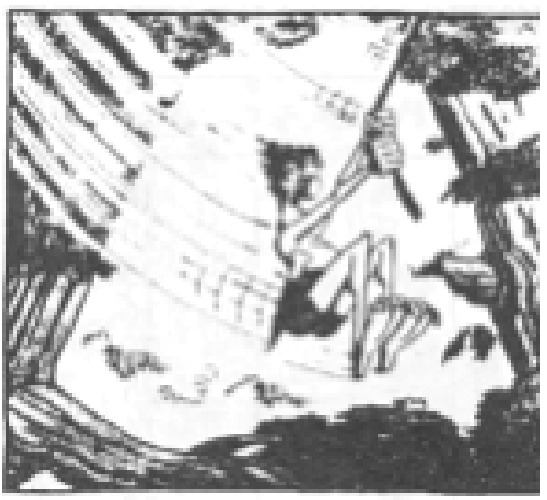
mit ld(x) als Logarithmus zur Basis 2 (logarithmus dualis); insbesondere gilt $ld(h_j) = \log(h_j)/\log(2) = 3{,}3219 \cdot \log(h_j)$. Die folgende Berechnungsformel ist äquivalent zu (5.43):

$$(5.44) \quad E = -\Sigma\, h_j\, ld(h_j).$$

Vor einer Darstellung der Interpretation und Eigenschaften der Entropie soll die Berechnung von E an einem Beispiel demonstriert werden. Das Konzept der Entropie stammt aus der Nachrichtentechnik, in der es den Gehalt einer Information quantifizieren soll.

Beispiel 5.13:

Schauspieler S (Rollenfach: jugendlicher Naturbursche) ist auf Tarzan-Filme spezialisiert. Gelegentlich spielt er aber auch in Krimis, Heimat- und Actionfilmen mit. Sein Produzent führte folgende Statistik über die Filme, an denen S mitgewirkt hat:



Art des Films	Anz. d. Filme mit S 1989	1990
Tarzan	4	6
Krimis	2	2
Heimat	1	1
Action	1	3
Summe	8	12

a) Hat die Unterschiedlichkeit der vielfältigen (bzw. vierfältigen) schauspielerischen Betätigung von S zu- oder abgenommen?

b) Man berechne die Entropien für die beiden Jahre!

c) Wie groß wären die Entropien, wenn S an jeweils gleich vielen Filmen jedes Typs mitgewirkt hätte?

Lösung 5.13:

a) Die Beantwortung dieser Frage dürfte ohne Berechnung eines für Nominalskalen geeigneten Streuungsmaßes schwierig sein.

b) Entropie 1989
$E_{1989} = (1/2)ld(2) + (1/4)ld(4) + (1/8)ld(8) + (1/8)ld(8) =$
$= ½\cdot 1 + ¼\cdot 2 + ¼\cdot 3 = 1{,}75.$

Entropie 1990
$E_{1990} = (1/2)ld(2) + (1/6)ld(6) + (1/12)ld(12) + (1/4)ld(4) =$
$= ½\cdot 1 + (1/6)\cdot 2{,}58496 + (1/12)\cdot 3{,}58496 + ¼\cdot 2 = 1{,}7296.$

c) Bei einer Gleichverteilung hätte S 1989 an zwei Filmen und 1990 an drei Filmen jedes Typs mitwirken müssen. Die Entropien wären dann gewesen $E^*_{1989} = E^*_{1990} = 4\cdot[¼\cdot ld(4)] = ld(4) = 2$.
Das Streuungsmaß E ist also in beiden Jahren durch die Anzahl der Ausprägungen des nominalskalierten Merkmals "Art des Films" nach oben begrenzt mit ld(4) = 2.

Bemerkungen zu Def. 5.6:

1. Im Falle der Einpunktverteilung (d.h. alle Merkmalswerte sind gleich) ist $E = 1\cdot ld(1) = 0$. Falls $m > 1$ ist $E > 0$, so dass beide Teile des Axioms S1 erfüllt sind. Sobald auch nur zwei unterschiedliche Beobachtungen auftreten ist $E > 0$, denn $½ld(2) = 1/2$. Somit erfüllt die Entropie auch das Axiom S2.

2. Im Unterschied zu allen bisher behandelten Maßzahlen führt bei E die Berechnung aus einer Häufigkeitsverteilung ("gewogene" Berechnung) nicht zum gleichen Ergebnis, wie die Berechnung aus Einzelwerten ("ungewogene" Berechnung), sondern notwendig zu einem kleineren Zahlenergebnis. Dies wird im Beispiel 5.14 demonstriert.
Wegen $h_j \geq 1/n$ gilt auch $ld(1/h_j) \leq ld(n)$, so dass $h_j\cdot ld(1/h_j) \leq h_j\cdot ld(n)$ für $n_j > 1$. Damit kann $E = \Sigma\ h_j\cdot ld(1/h_j)$ (gewogene Berechnung) nicht größer sein als $E_o = \Sigma(1/n)\cdot ld(n) = ld(n)$ (ungewogene Berechnung).

3. Satz 5.6 zeigt, dass $E_o = ld(n)$ auch die Obergrenze der Entropie darstellt. Mithin gilt

(5.45) $\quad 0 \leq E \leq ld(n)$

Nach Satz 5.6 nimmt E bei gegebener Anzahl m von Merkmalsausprägungen ein Maximum an, wenn jede Merkmalsausprägung gleich häufig ist.

4. E erfüllt die Axiome S3 und S4 nicht, da die Häufigkeiten nicht von einer Transformation der Merkmalswerte berührt werden und E nur von den Häufigkeiten, nicht aber von den Merkmalswerten abhängt.

5. Ersetzt man die relativen Häufigkeiten h_j durch die Merkmalsanteile q_j, so läßt sich E auch als Konzentrationsmaß verwenden (vgl. Kap. 6).

6. Besonders vorteilhaft ist das Verhalten der Entropie bei Aggregation und Zerlegung: Sind bei einer klassierten Verteilung jeweils alle n_k Beobachtungen innerhalb der Klasse k unterschiedlich (hat jede also die Häufigkeit $1/n_k$), so erhält man die Entropie E_k innerhalb der k-ten Klasse mit

$$(5.46) \qquad E_k = \sum_{l=1}^{n_k} \frac{1}{n_k} \operatorname{ld}(n_k) = \operatorname{ld}(n_k) \qquad (l = 1,2,...,n_k) \ .$$

Für die Gesamtentropie (auf der Grundlage von i=1,2,...,n Einzelwerten) gilt dann

$$(5.47) \qquad E_{ges} = \sum_{i=1}^{n} \frac{1}{n} \operatorname{ld}(n) = \operatorname{ld}(n).$$

Damit erhält man die folgende Streuungszerlegung nach Art der Varianzzerlegung (nach Satz 5.5)

(5.48)	$E_{ges} = E_{ext} + E_{int}$	mit
(5.48a)	$E_{ext} = \Sigma\, h_k \cdot \operatorname{ld}(1/h_k)$	und
(5.48b)	$E_{int} = \Sigma\, h_k\, E_k$.	

Die in Bsp. 5.15 verifizierte Zerlegung der Entropie gem. Gl. 5.48 ist wie folgt zu beweisen: Die Summe von E_{ext} und E_{int} gem. Gl. 5.49 und 5.50 beträgt

$\Sigma\, h_k \cdot \operatorname{ld}(n/n_k) + \Sigma h_k \cdot \operatorname{ld}(n_k) = \Sigma\, h_k\, [\operatorname{ld}(n/n_k) + \operatorname{ld}(n_k)] =$
$\Sigma\, h_k\, [\operatorname{ld}(n) - \operatorname{ld}(n_k) + \operatorname{ld}(n_k)] = \operatorname{ld}(n)\, \Sigma\, h_k = \operatorname{ld}(n) = E_{ges}$.

Beispiel 5.14:

Berechnen Sie die Entropie E für die

a) folgenden Einzelwerte : 2,3,3,4,4,4
b) folgende Häufigkeitsverteilung (vgl. auch Bsp. 6.3):

x_j	2	3	4
h_j	1/6	1/3	½

Es ist unschwer zu erkennen, dass es sich in beiden Fällen um die gleichen Daten handelt.

Lösung 5.14

a) Auf die Zahlenwerte für die Beobachtung kommt es nicht an. Es liegen sechs Beobachtungen vor, so dass sich E gem. Gl. 5.43 wie folgt errechnet:

$E = (1/6)ld(6) + ... + (1/6)ld(6) = ld(6) = log(6)/log(2) = 0{,}77815/0{,}3010 = 2{,}585.$

b) $E = (1/6)ld(6) + (1/3)ld(3) + (1/2)ld(2) = 2{,}585/6 + 1{,}585/3 + 1/2 = 1{,}459$. Das ist, wie in Bem. Nr. 2 dargelegt, kleiner als die Berechnung aus Einzelwerten, die zu 2,585 führte.

Beispiel 5.15:

Gegeben sei die folgende Häufigkeitsverteilung (vgl. Bsp. 5.14):

Klasse	Beobachtungen	n_k	h_k
1	$x_1 = 2$	1	1/6
2	$x_2=3$, $x_3=4$	2	1/3
3	$x_4=3$, $x_5=3$, $x_6=3$	3	1/2

Man verifiziere anhand dieses Beispiels den Satz über die Aggregation bzw. Zerlegung der Entropie (vgl. oben Bem. Nr. 6)

Lösung 5.15:

Die externe Entropie ist im Bsp. 5.14 bereits berechnet worden mit dem Ergebnis $E_{ext} = (1/6)ld(6) + (1/3)ld(3) + (1/2)ld(2) = 1{,}459$.

Um die interne Entropie zu errechnen sind zunächst die Entropien innerhalb der drei Klassen zu berechnen. Man erhält:

$E_1 = ld(1) = 0$
$E_2 = ld(2) = 1$ und
$E_3 = ld(3) = log(3)/log(2) = 1{,}58496.$

Die interne Entropie ist dann

$$E_{int} = \Sigma h_j E_j = 1/3 + (1/2)ld(3) = 1{,}9183.$$

Die Summe aus externer und interner Entropie ist dann

$$E_{ges} = (1/6)ld(6) + (1/3)ld(3) + 1/2 + 1/3 + (1/2)ld(3) = 2{,}58496 = ld(n) = ld(6).$$

Satz 5.6:

Die Entropie E nimmt bei Gleichverteilung ($h_j = 1/m$, $j= 1,2,...,m$) der m Merkmalswerte ihren maximalen Wert ld(m) an.

Beweis:

Die Maximierung von E unter der Nebenbedingung $Sh_j = 1$ (mit j = 1, 2,...,m) läßt sich mittels der Lagrange-Funktion

$L = - \Sigma h_j \, ld(h_j) - \lambda (\Sigma h_j - 1)$ durchführen. Man erhält dann

(*) $\partial L/\partial h_j = -ld(h_j) - ld(e)/h_j - \lambda = 0$ (j=1,2,...,m) und

(**) $\partial L/\partial \lambda = \Sigma h_j - 1 = 0$

Gemäß den m Gleichungen (*) muss für λ stets gelten: $\lambda = -ld(h_j) - ld(e)/h_j$, was nur möglich ist, wenn alle h_j gleich sind. In Verbindung mit (**) folgt daraus, dass für alle j gelten muss $h_j = 1/m$. Aus den Bedingungen zweiter Ordnung folgt, dass an der Stelle ($h_1 = 1/m$, $h_2 = 1/m$,... , $h_m = 1/m$) in der Tat das Maximum von E liegt. Die Entropie nimmt
dann den Wert $-ld(1/m) = ld(m)$ an.

Exkurs: Dispersionsindex und Diversität[2]

Die Anzahl möglicher Paarvergleiche einer Einheit mit $x = x_i$ mit jeweils einer Einheit mit $x =,/ x_i$ ist $n_i(n-n_i)$. Folglich ist die Anzahl der Paare bei denen sich beide Beobachtungswerte unterscheiden $\Sigma n_i(n-n_i)$ und infolge von Satz 5.6 ist dieser Ausdruck dann maximal, wenn für alle i gilt:

$$n_i = \frac{n}{m} = k \qquad \text{und} \qquad \Sigma k = mk = n^2 \frac{m-1}{m},$$

also eine Rechteckverteilung vorliegt, so dass man mit

$$(5.49) \quad S_D = \frac{m}{(m-1)n^2} \sum_{i=1}^{m} n_i (n-n_i) = \frac{m}{m-1} (1- \Sigma h_i^2)$$

den **Dispersionsindex** von Hammond und Householder (1962) erhält, der lediglich eine **Nominalskala** voraussetzt. Offenbar ist $S_D = 0$, wenn eine Einpunktverteilung vorliegt und $S_D = 1$, wenn alle Häufigkeiten n_i gleich sind, so dass $0 \le S_D \le 1$ gilt.

Man beachte, dass hier ein Konzept der Variabilität vorliegt, dass keinen Abstandsbegriff voraussetzt, sondern sich, ähnlich wie die Entropie, nur daran orientiert, in welchem Maße bestimmte Merkmalsausprägungen gehäuft auftreten. S_D nimmt aber auch nicht Bezug auf den Modus. Das Maß S_D (oder D) ist geeignet um **Strukturen** (Gliederungen nach einem qualitativen Merkmal) zu beschreiben hinsichtlich der Abweichung von einer Rechteckverteilung, z.B. die Arbeitsteilung und deren Veränderung oder die unterschiedliche Struktur der Warenkörbe bei Preisindizes.

---

[2] Hinweise auf die in diesem Abschnitt behandelten Streuungsmaße verdanke ich Herrn Prof. Dr. Piesch.

Im Falle einer **Ordinalskala** sind ebenfalls keine Abstände definiert, auf die ein Streuungsmaß Bezug nehmen könnte, wohl aber die Summenhäufigkeit. Darauf beruht die Diversität.

Für die lediglich eine **Ordinalskala** voraussetzende **Diversität** S_{D^*} oder D^* gilt:

$$(5.50) \quad S_{D^*} = D^* = \frac{4}{m\text{-}1} \sum_{i=1}^{m-1} H_i\,(1\text{-}H_i)$$

wobei H_i die Summenhäufigkeit darstellt.

Bei einer Rechteckverteilung ($n_i = n/m$ für alle $i = 1,2,...,m$) erreicht der Ausdruck $2\sum_{i=1}^{m-1} H_i\,(1\text{-}H_i)$ mit $H_i = \frac{i}{m}$ und $H_m = 1$ den Wert $\frac{m^2\text{-}1}{3m}$. Bei einer Nominalskala ist dies der maximale Wert. Im Falle einer Ordinalskala kann man allerdings von einer extremeren Situation mit $h_1 = 1/2$, $h_2 = h_3 = ...= h_{m-1} = 0$ und $h_m = 1/2$ ausgehen. Man erhält dann bei geradzahligem m für $2\sum_{i=1}^{m-1} H_i\,(1\text{-}H_i)$ den Wert $\frac{m\text{-}1}{2}$, was größer ist als $\frac{m^2\text{-}1}{3m}$ $= \frac{m\text{-}1}{2} \cdot \frac{2(m+1)}{3m}$ sobald $m > 2$. So erklärt sich Gl. 5.50

4. Maße der relativen Streuung

Mit Def. 5.1 ist die relative Streuung definiert als Relation mit einem Maß der absoluten Streuung im Zähler und ein (hierzu passender) Mittelwert im Nenner. Der Vorteil dieses Verhältnisses ist vor allem die Maßstabsunabhängigkeit der so gemessenen Streuung, so dass damit auch Streuungen verschiedener Häufigkeitsverteilungen vergleichbar sind. Das bei weitem bekannteste Maß der relativen Streuung ist der Variationskoeffizient (Def. 5.7), der auf der Basis der Standardabweichung konstruiert ist. Man kann auch andere Streuungsmaße zur Bildung von Maßen der relativen Streuung heranziehen, z.B. den mittleren Quartilsabstand beim Quartilsdispersionskoeffizienten (Def. 5.8).

Def. 5.7: Variationskoeffizient

Der folgende durch Gl. 5.51 definierte Ausdruck V ist bekannt als Variationskoeffizient:

$$(5.51) \quad V = \frac{s}{\bar{x}} \quad .$$

Bemerkungen zu Def. 5.7:

1. Der mit 100 multiplizierte Wert von V drückt die Streuung (gemessen als Standardabweichung) in Prozent des Durchschnitts (arithmetischen Mittels) aus. Die Standardabweichung (s) mißt zwar eine Art durchschnittlichen Abstand zwischen den Beobachtungswerten und dem arithmetischen Mittel $\bar{x}$, sie wird aber nicht ins Verhältnis zu $\bar{x}$ gesetzt, was dazu führt, dass die Streuungen von zwei verschiedenen Beobachtungsreihen nicht miteinander verglichen werden können. Dies folgt auch daraus, dass die Standardabweichung s nicht maßstabsunabhängig ist (nach Axiom S4). Zur Herstellung einer Vergleichbarkeit wird die Standardabweichung s in Beziehung zu $\bar{x}$ gesetzt.

   Eine Standardabweichung von 5 bei einem $\bar{x}$ von 500 erscheint nicht sehr hoch. Ist $\bar{x}$ allerdings nur 10, dann ist die Standardabweichung von 5 sehr hoch. Deswegen ist es bei einem Vergleich von Streuungen verschiedener Beobachtungsreihen sinnvoll s auf $\bar{x}$ zu beziehen bzw. den Variationskoeffizienten zu berechnen, der im ersten Fall 1% und im zweiten 50% beträgt.

2. Der Variationskoeffizient lässt sich sinnvoll nur für verhältnisskalierte (ratioskalierte) Merkmale, mit einem positiven Mittelwert interpretieren.

3. Die Dimensionslosigkeit des Variationskoeffizienten ermöglicht auch einen Vergleich von Verteilungen mit verschiedenen Maßeinheiten (km, DM, Jahr).

4. Der Variationskoeffizient kann auch als Maß der Ungleichheit (Disparität) aufgefaßt werden (Kap. 6). Das folgende Beispiel zeigt auch, wie sich der Variationskoeffizient durch eine Lineartransformation verändert.

Beispiel 5.16:

Wie ändert sich in Aufg. 5.3 der Variationskoeffizient? Interpretieren Sie das Ergebnis!

Lösung 5.16:

In Aufgabe 5.3 erhielt man: $\bar{x} = 2200$, $s_x = 800$ so dass $V_x = 0{,}364$ ist, $\bar{y} = 2500$, $s_y = 880$, womit sich V_y errechnet zu $V_y = 0{,}352$.

Der Variationskoeffizient verringert sich also, was allein auf den festen "Sockelbetrag" von 80,- DM für das Urlaubsgeld zurückzuführen ist. Das arithmetische Mittel vergrößert

sich um 13,64%. Die Standardabweichung, auf die allein die "lineare" Erhöhung aller Gehälter um 10%, nicht aber der Sockelbetrag einen Einfluß hat, vergrößert sich dagegen nur um 10%.

Def. 5.8: Quartilsdispersionskoeffizient

Setzt man den mittleren Quartilsabstand $\bar{Q}_{0,25} = ½(Q_3 - Q_1)$ als Maß der absoluten Streuung ins Verhältnis zum Wert $½(Q_1 + Q_3)$, den man als eine Art Mittelwert interpretieren kann, so erhält man QD, den Quartilsdispersionskoeffizient

$$(5.52) \quad QD = \frac{Q_3 - Q_1}{Q_3 + Q_1} .$$

Bemerkungen zu Def. 5.8:

1. Der Quartilsdispersionskoeffizient kann auch mit dem Median ($\tilde{x}_{0,5} = Q_2$) berechnet werden, man erhält dann

   $$(5.52a) \quad QD^* = \frac{Q_3 - Q_1}{Q_2} .$$

2. Auf der Basis des Medians lassen sich auch andere Maße der relativen Streuung konstruieren, etwa

   $$(5.52b) \quad RD = \frac{d_x}{Q_2} = \frac{d_x}{\tilde{x}_{0,5}}$$

   eine relativierte durchschnittliche Abweichung.

3. Die relative Streuung QD ist nicht zu verwechseln mit einem auf Quartile beruhenden Schiefemaß.

5. Momente

Momente sind sehr allgemeine Kennzahlen von Verteilungen. Viele Maßzahlen von Häufigkeitsverteilungen können als spezielle Momente angesehen werden. Übersicht 5.2 stellt ausgehend vom Begriff des Moments um a (vgl. Def. 5.9) die Zusammenhänge zwischen Momenten verschiedenen Typs dar.
Verteilungen lassen sich außer durch Lage- und Streuungsmaße auch durch andere Gestaltparameter, wie Schiefe und Wölbung charakterisieren. Diese Kennzahlen werden i.d.R. aus Momenten abgeleitet.

Def. 5.9: Momente

a) Momente sind Mittelwerte der (k-ten Potenz der lineartransformierten) Größen

$$\left[\frac{x_v - a}{b}\right]^k$$

mit b = s (Standardabweichung) und a = $\bar{x}$ (arithmetisches Mittel) erhält man ***standardisierte Momente***.

b) Mit b = 1 und der beliebigen reellen Konstanten a erhält man das k-te ***Moment um a***:

- bei Einzelwerten (ungewogene Berechnung)

$$(5.53) \quad m_{k(a)} = \frac{1}{n} \Sigma (x_v - a)^k \qquad \text{[k-tes Moment um a] und}$$

- bei Häufigkeitsverteilungen (gewogene Berechnung)

$$(5.54) \quad m_{k(a)} = \frac{1}{n} \Sigma (x_i - a)^k n_i = \Sigma (x_i - a)^k \, h_i \, .$$

c) Spezialfälle: ***Anfangs-Momente*** (oder Momente um Null) und ***zentrale Momente*** sind Spezialfälle des Moments um a (Übers. 5.2).

d) Von geringerer Bedeutung sind absolute Momente: analog Gl. 5.53 ist das k-te ***absolute Moment*** um a definiert als

$$(5.53a) \quad m^*_{k(a)} = \frac{1}{n} \Sigma |x_v - a|^k \qquad \text{[k-tes absolutes Moment um a].}$$

Bei einer geraden Zahl k sind die absoluten Momente gleich den "gewöhnlichen Momenten" [= Momente im Sinne von b) bzw. c)]

e) In der Induktiven Statistik spielen auch ***faktorielle Momente*** eine Rolle; $\frac{1}{n} \Sigma x_v (x_v - 1)(x_v - 2) \ldots$.

Bemerkungen zu Def. 5.9:

1. Man sieht leicht, dass das erste Anfangsmoment m_1 das arithmetische Mittel $\bar{x}$ ist und dass das zweite zentrale Moment z_2 die Varianz s^2 ist.

2. Das dritte zentrale Moment hat einen Zahlenwert von Null, wenn eine symmetrische Verteilung vorliegt. Es hat einen positiven Wert bei einer linkssteilen Verteilung und ist negativ bei einer rechtssteilen Verteilung, weshalb z_3 zur Konstruktion von Maßzahlen der **Schiefe** (Asymmetrie) verwendet wird. Dieser Zusammenhang gilt für alle ungeraden zentralen Momente, bis auf das erste z_1, das wegen der Schwerpunkteigenschaft von $\bar{x}$ immer Null ist.

3. Das vierte zentrale Moment z_4 charakterisiert die **Wölbung** einer Verteilung. Die Werte dieses, sowie aller anderen höheren geraden Momente nehmen nämlich um so mehr zu, je höher die Wölbung einer Verteilung ist.

4. Zentrale Momente und Anfangsmomente verhalten sich bei einer proportionalen Transformation der Variablen X zur Variablen Y mit $y_v = bx_v$ $(b \neq 1)$ wie folgt:

$$(5.56) \quad z_{k(y)} = b^k \cdot z_{k(x)} \,.$$

Das folgt aus $z_{k(y)} = \frac{1}{n} \Sigma (bx_v - b\bar{x})^k = \frac{1}{n} \Sigma b^k (x_v - \bar{x})^k = b^k [\frac{1}{n} \Sigma x_v - \bar{x}^k]$.

5. Momente sind nicht unabhängig vom Ursprung der Skala für die Variable X. Deshalb unterscheiden sich auch Momente um a, Anfangsmomente und zentrale Momente. Anders dagegen die sog. **Kumulanten** k_r: außer k_1 sind sie invariant gegenüber Translationen (Verschiebungen des Nullpunkts) gem. $y_v = a + x_v$. Bei proportionalen Transformationen $(y_v = bx_v)$ gilt Gl. 5.56.

6. Unter bestimmten, in der Praxis fast immer gegebenen Bedingungen ist eine Häufigkeitsverteilung durch die Folge ihrer Momente eindeutig bestimmt.

7. Bei zwei und mehr Variablen ist das **Produktmoment** eine Verallgemeinerung. Bei zwei Variablen X und Y ist das zentrale Produktmoment bekannt als Kovarianz und das standardisierte Produktmoment als Korrelationskoeffizient (vgl. Kap. 7).

8. Zum Begriff "Moment": die Analogie zur physikalischen Terminologie ist berechtigt. Die Varianz s^2 ist in der Tat das Trägheitsmoment bei Rotation um das arithmetische Mittel, wobei dieses als Schwerpunkt (in einer Dimension) der Quotient aus Summe der Momente ($\Sigma x_j h_j$, also Kraft h_j mal Hebelarm x_j) und Summe der Gewichte (Σh_j) ist.

Übersicht 5.2: Momente

allgemein: Mittelwerte der Größen

$$\left[\frac{x_v - a}{b}\right]^k$$

Spezialfälle

b = 1
a als beliebige reelle Konstante
Momente um a

(5.53) $m_{k(a)} = \frac{1}{n}\sum_{v=1}^{n}(x_v - a)^k$

[k-tes Moment um a] *)

b = s
a = $\bar{x}$
standardisierte Momente ms_k

$ms_1 = 0$; $ms_2 = 1$

(Mittelwert und Varianz einer standardisierten Variabel sind 0 und 1)

Anfangsmoment
a = 0

(5.54) $m_k = \frac{1}{n}\sum x_v^k$

(5.54a) $m_k = \sum x_i^k h_i$

zentrales Moment
a = $\bar{x}$

(5.55) $z_k = \frac{1}{n}\sum(x_v - \bar{x})^k$

(5.55a) $z_k = \sum(x_i - \bar{x})^k h_i$

weitere Spezialfälle:

$m_0 = 1$
$m_1 = \bar{x}$

$z_1 = 0$ (Schwerpunkteigenschaft!)
$z_2 = s^2$ (Varianz)

*)(ungewogene Berechnung, die gewogene Berechnung erfolgt analog, vgl. Def. 5.9)

Zusammenhänge zwischen Anfangs- und zentralen Momenten

$z_2 = m_2 - (m_1)^2$ (Verschiebungssatz für die Varianz).Analog folgt:
$z_3 = m_3 - 3m_1m_2 + 2(m_1)^3$ $z_4 = m_4 - 4m_1m_3 + 6(m_1)^2m_2 - 3(m_1)^4$ usw.

$$\text{allgemein: } z_k = \sum_{i=0}^{k} \binom{k}{i} m_{k-i}\,(-m_1)^i$$

Daraus folgt: Anfangsmomente sind nicht verschiebungsinvariant

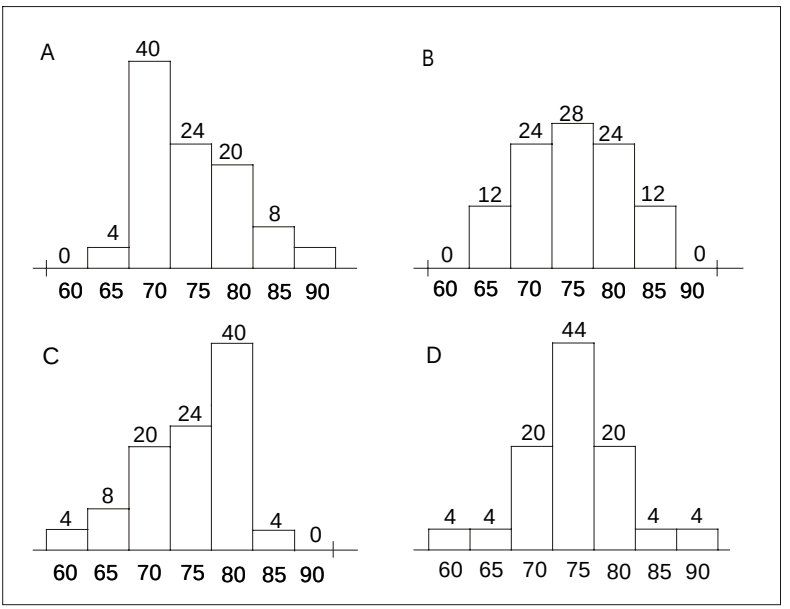
$m_i(c) = \sum (x_v + c)^i \neq m_i = \sum x_v^i$ wohl aber zentrale Momente

$z_i(c) = \sum \left[(x_v + c) - (\bar{x} + c)\right]^i = z_i$, denn $m_1(c) = \bar{x} + c$.

Beispiel 5.17:

Die folgenden vier Häufigkeitsverteilungen (A,B,C,D) haben jeweils das gleiche arithmetische Mittel ($\bar{x}$ = 75) und die gleiche Varianz (s^2 = 36). Gleichwohl ist ihre Gestalt hinsichtlich Schiefe und Wölbung sehr unterschiedlich (vgl. Abb. 5.4). Dies wird auch deutlich, wenn man die dritten und vierten zentralen Momente berechnet (Beispiel entnommen aus K.Stange, Angewandte Statistik, Bd. 1, S.87). Die Gesamthäufigkeit n ist jeweils 100, so dass man die relativen Häufigkeiten leicht ablesen kann.

Abb. 5.4: Häufigkeitsverteilungen des Bsp.5.17



x_j	absolute Häufigkeit			
	A	B	C	D
60	0	0	4	4
65	4	12	8	4
70	40	24	20	20
75	24	28	24	44

x_j	absolute Häufigkeit			
	A	B	C	D
80	20	24	40	20
85	8	12	4	4
90	4	0	0	4

Berechnen Sie die dritten und vierten und zentralen Momente für die vier Häufigkeitsverteilungen!

Lösung 5.17:

Verteilung	A	B	C	D
z_3	150	0	-150	0
z_4	3600	2700	3600	5100

6. Schiefemaße

a) Begriff der Schiefe

Mit der Schiefe (skewness) soll der Grad der Asymmetrie einer Häufigkeitsverteilung gemessen werden. Schiefe ist die Abweichung von der symmetrischen Verteilung eines metrisch skalierten Merkmals. Asymmetrie (= Schiefe) hat zwei Formen: Linkssteilheit und Rechtssteilheit. Die folgenden Begriffe werden synonym verwendet:

linkssteil	=	rechtsschief
rechtssteil	=	linksschief .

Linkssteilheit bedeutet, dass sich die Masse der Merkmalsträger am unteren Ende einer Häufigkeitsverteilung konzentriert. Sie wird auch oft ähnlich interpretiert wie "Ungleichheit" (Disparität), es gibt jedoch Unterschiede zur Disparität im Sinne der Statistik (vgl. Kap. 6).
Die Abb. 5.4 veranschaulicht die Begriffe Symmetrie, Linkssteilheit und Rechtssteilheit. Die Verteilungen B und D sind symmetrisch (sie unterscheiden sich jedoch durch die Wölbung), die Verteilung A ist linkssteil (positive Schiefe) und die Verteilung C ist rechtssteil (negative Schiefe).
Schiefe kann auch für die Datenanalyse als störend empfunden werden: es kann dann z.B. schwierig zu entscheiden sein, welcher Lageparameter das Niveau einer Verteilung (die Größenordnung der Merkmalswerte) ange-

messen beschreibt und welche Beobachtung als Ausreißer anzusehen ist. Deshalb gilt es nicht nur, die Schiefe zu messen, sondern auch gegebenenfalls Methoden anzuwenden, um diese zu beseitigen (Abschn. c).

Viele natürliche Erscheinungen sind symmetrisch verteilt (z.B. Körpergröße, Körpergewicht aber auch z.B. der Intelligenzquotient), während "soziale" Erscheinungen häufig linkssteil verteilt sind. Das gilt besonders für Einkommen und Vermögen. Es gibt zahlreiche Modelle zur Erklärung einer linkssteilen Einkommensverteilung.
Nach dem zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung ist eine Größe X, die sich als Summe sehr vieler stochastisch unabhängiger (zum Begriff der Unabhängigkeit vgl. Kap. 7) Einflußfaktoren u_j darstellen läßt $X = u_1 + u_2 + u_3 + ... + u_n = \Sigma u_j$, asymptotisch (mit wachsendem n) normalverteilt. Ist Z dagegen ein Produkt solcher Einflussfaktoren v_j also $Z = v_1 v_2 v_3 ... v_n$ so ist Z asymptotisch logarithmisch-normalverteilt. Die Normalverteilung ist eine symmetrische Verteilung, die logarithmische Normalverteilung eine linkssteile Verteilung. Man spricht auch vom "Gesetz der proportionalen Effekte", d.h. die Einzeleinflüsse verstärken sich gegenseitig, bzw. schwächen sich gegenseitig ab, in der Art, wie es im Matthäus-Evangelium beschrieben wird:

> *"Denn wer da hat, dem wird gegeben, dass er die Fülle habe, wer aber nicht hat, von dem wird auch genommen, das er hat" (Matth.13, Vers 12).*

Linkssteilheit und Rechtssteilheit sind zwei entgegengerichtete Abweichungen von der Symmetrie. Wenn eine Verteilung nicht symmetrisch ist, dann kann sie nur entweder linkssteil oder rechtssteil sein. Schiefemaße sollen Richtung und Ausgeprägtheit der Abweichung von der Symmetrie messen. Zu einer exakteren Begriffsbildung gelangt man durch die folgende Definition:

Def. 5.10: Achsensymmetrie

Die Häufigkeitsverteilung des metrisch skalierten Merkmals X heißt symmetrisch bezüglich des Medians $\tilde{x}_{0,5}$, falls für **alle** Werte einer reellen Konstante c gilt

$$h(\tilde{x}_{0,5} - c) = h(\tilde{x}_{0,5} + c) \quad c > O \; . \tag{5.57}$$

Dabei ist $h(\tilde{x}_{0,5} - c)$ die relative Häufigkeit der Merkmalsausprägung $x_c = \tilde{x}_{0,5} - c$ und $h(\tilde{x}_{0,5} + c)$ ist entsprechend definiert. Eine Verteilung ist schief oder asymmetrisch, wenn Gl. 5.57 nicht gilt.

Bemerkungen zu Def. 5.10

1. Die Definition ist eine exakte Beschreibung der Bedingungen, unter denen eine Häufigkeitsverteilung achsensymmetrisch um den Median ist. Sie ist nicht notwendig auch operational zur Herleitung eines Schiefemaßes. Nur dann, wenn eine Häufigkeitsverteilung einen Me-

dian dergestalt besitzt, dass die möglichen Ausprägungen der Variable X jeweils links und rechts gleich weit entfernt von $\tilde{x}_{0,5}$ liegen, kann Symmetrie gem. Gl. 5.57 definiert und verifiziert werden (Bsp. 5.18).

2. Man beachte, dass Gl. 5.57 für die kumulierten Häufigkeiten impliziert:

(5.58)	$h(x \le \tilde{x}_{0,5} - c) = h(x \ge \tilde{x}_{0,5} + c) \quad (c \ne 0)$ oder äquivalent: $H(\tilde{x}_{0,5} - c) = 1 - H(\tilde{x}_{0,5} + c)$ für jedes $c \ge 0$.

Die Größe H(x) ist die bis x erreichte kumulierte relative Häufigkeit (Fläche unter der Häufigkeitsverteilung). Beispiel 5.19 zeigt, dass Achsensymmetrie Flächengleichheit links und rechts von bestimmten Punkten unter der Häufigkeitsverteilung impliziert, nicht aber umgekehrt (aus Flächengleichheit [bei ausgewählten Intervallen] folgt nicht Achsensymmetrie).
Gl. 5.58 führt zu der Def. 5.11, die zwar bei einer stetigen Verteilung keine Schwierigkeiten bereiten würde, im diskreten Fall aber nicht anwendbar sein kann (Bsp. 5.19).

3. Im Satz 5.7 wird gezeigt, dass die Definition der Symmetrie mit Def. 5.10 - als Achsensymmetrie um $\tilde{x}_{0,5}$ bzw. um $\bar{x}$ - im Einklang steht mit einer Momentschiefe von Null.

4. Ausgehend von der Symmetrie im Sinne von Gl. 5.57 (Achsensymmetrie) könnte man definieren, dass Linkssteilheit dann entsteht, wenn ein Merkmalsträger mit dem Merkmalswert $\bar{x} + c$ (mit $c > 0$) ausgetauscht wird gegen einen Merkmalsträger mit dem Wert $\bar{x} - c$. Entsprechend wäre Rechtssteilheit zu definieren. Bei dieser Art, eine linkssteile Verteilung zu "erzeugen" verringert sich jedoch das arithmetische Mittel $\bar{x}$ zum neuen Wert $\bar{x} - 2c/n$. Entsprechend vergrößert sich das arithmetische Mittel beim Übergang zu einer rechtssteilen Verteilung zu $\bar{x} + 2c/n$.

Beispiel/Lösung 5.18:

Ein Beispiel für die Anwendbarkeit von Def. 5.10 wäre eine Gesamtheit mit den n = 9 Beobachtungen 10,15,15,20,20,20,25,25,30. Der Median ist 20 und die Häufigkeitsverteilung ist nach Def. 5.10 symmetrisch, da bei c = 5 gilt:

$$h(15) = h(\tilde{x}_{0,5} - 5) = h(\tilde{x}_{0,5} + 5) = h(25) = 2/9$$

und bei c = 10 ist entsprechend

$$h(\tilde{x}_{0,5} - 10) = h(\tilde{x}_{0,5} + 10) = 1/9.$$

Die Definition wäre aber z.B. nicht anwendbar auf eine Häufigkeitsverteilung mit dem Median 20, wenn zwar die Beobachtung x=17 nicht aber x=23 vorkommt.

Beispiel 5.19:

Die Einzelbeobachtungen 10,16,20,20 bilden eine Häufigkeitsverteilung, die nach anschaulicher Vorstellung nicht symmetrisch, sondern rechtssteil ist. Man überprüfe die Symmetriedefinition gem. Gl.5.58 und berechne das dritte zentrale Moment!

Lösung 5.19:

Das arithmetische Mittel beträgt $\bar{x} = 16{,}5$ und der Median (mit Interpolation) $\tilde{x}_{0,5} = 18$. Gilt Gl. 5.58 für **jedes** c, so müßte es auch für c = 0 gelten. Man erhält damit aber die für den Zentralwert stets erfüllte Gleichung: $h(x < \tilde{x}_{0,5}) = h(x > \tilde{x}_{0,5}) = ½$, natürlich auch hier, ohne dass deshalb die Verteilung symmetrisch wäre.
Symmetrie wäre nach Gl. 5.58 auch angezeigt für c = 1, denn

$$h(x \le \tilde{x}_{0,5} - 1) = h(x \le 17) = h(x \ge \tilde{x}_{0,5} + 1) = h(x \ge 19) = ½.$$

Die Bezugnahme auf Flächengleichheit zur Definition der **Symmetrie** (vgl. Def. 5.11) macht nur Sinn, wenn Gl. 5.58 für **jedes** c gilt. In diesem Beispiel gilt Gl. 5.58 z.B. nicht für c = 4. Denn

$$h(x \le \tilde{x}_{0,5} - 4) = h(x \le 14) = 1/4 > h(x \ge \tilde{x}_{0,5} + 4) = h(x \ge 22) = 0,$$

was nach Gl. 5.58b (Def. 5.11) auf Rechtssteilheit hinweist, die [gemessen an der Momentschiefe] auch tatsächlich gegeben ist.
Das dritte zentrale Moment beträgt hier nämlich

$$z_3 = ¼ [(10 - 16{,}5)^3 + (16 - 16{,}5)^3 + 2(20 - 16{,}5)^3] = -189/4 = -47{,}25.$$

Dieses Beispiel macht den Hintergrund der folgenden Def. 5.11 deutlich.

Def. 5.11: Schiefe

Eine Verteilung ist symmetrisch wenn **für alle** c gilt

(5.58) $h(x \le \tilde{x}_{0,5} - c) = h(x \ge \tilde{x}_{0,5} + c)$,
oder äquivalent: $H(\tilde{x}_{0,5} - c) = 1 - H(\tilde{x}_{0,5} + c)$.

und sie ist entsprechend

linkssteil wenn (5.58a)	$h(x \le \tilde{x}_{0,5} - c) < h(x \ge \tilde{x}_{0,5} + c)$
rechtsteil wenn (5.58b)	$h(x \le \tilde{x}_{0,5} - c) > h(x \ge \tilde{x}_{0,5} + c)$

für **mindestens ein c** gilt.

Bemerkungen zu Def.5.11:

Zur Erläuterung vgl. Bsp. 5.19. Diese Definition der Schiefe legt den Gedanken nahe, dass bestimmte Quantile jeweils vom Median gleich weit entfernt sein müssen, wenn Symmetrie herrscht. So ist z.B. für die Abstände von Quartilen zu fordern

bei Symmetrie: $Q_3 - Q_2 = Q_2 - Q_1$
bei Linkssteilheit: $Q_3 - Q_2 > Q_2 - Q_1$ und
bei Rechtssteilheit: $Q_3 - Q_2 < Q_2 - Q_1$.

Auf diesem Gedanken beruhen der Quartils- und [allgemeiner] der Quantilskoeffizient als Schiefemaß (vgl. Def. 5.12, Teil b).

Satz 5.7:

a) Bei einer Häufigkeitsverteilung, die im Sinne der Def. 5.10 symmetrisch ist gilt für das dritte zentrale Moment $z_3 = 0$.

b) Wird ausgehend von der symmetrischen Verteilung
mit $2(h_o+h_1) = 1$ und $z_3 = 0$

x_i	h_i
$\bar{x} - c$	h_1
$\bar{x}$	$2h_0$
$\bar{x} + c$	h_1

die relative Häufigkeit a (mit$0 < a < h_1$)
1. von $\bar{x}$+c auf $\bar{x}$-c
2. von $\bar{x}$-c auf $\bar{x}$+c
"verlagert" so entsteht eine linkssteile (Fall 1) bzw. rechtssteile Verteilung (Fall 2):

Fall 1:

x_i	h_i
$\bar{x} - c$	$h_1 + \alpha$
$\bar{x}$	$2h_o$
$\bar{x} + c$	$h_1 - \alpha$

Fall 2:

x_i	h_i
$\bar{x} - c$	$h_1 - \alpha$
$\bar{x}$	$2h_o$
$\bar{x} + c$	$h_1 + \alpha$

mit den folgenden Momenten
1) Mittelwert
$\bar{x}^* = \bar{x} - 2c\alpha$ (Fall 1) $\qquad \bar{x}^{**} = \bar{x} + 2c\alpha$ (Fall 2)
2) drittes zentrales Moment:
$z_3^* = 2c^3\alpha(6h_1 - 1 - 8\alpha^2) \qquad z_3^{**} = - z_3^*$

$z_3^* > 0$ (linkssteil) $z_3^{**} < 0$ (rechtssteil)

Beweis: Elementar durch Ausmultiplizieren. Die Ausdrücke für die Momente werden im folgenden in einem Beispiel verifiziert.

Beispiel 5.20

Ausgehend von der symmetrischen Verteilung (mit c = 10, $\bar{x}$ = 20, $h_o = h_1 = 1/4$) soll mit
α = 0,05 die im obigen Satz beschriebene "Erzeugung" einer links- bzw. rechtsteilen Verteilung demonstriert werden. Wie hängt die Schiefe vom Parameter α ab?

Lösung 5.20

Ausgangsverteilung

x_i	h_i
$\bar{x}$-c=10	h_1=1/4
$\bar{x}$=20	$2h_o$=1/2
$\bar{x}$+c=30	h_1=1/4

Fall 1

x_i	h_i
10	0,3
20	0,5
30	0,2

Fall 2

x_i	h_i
10	0,2
20	0,5
30	0,3

das arithmetische Mittel und das dritte zentrale Moment z_3 sind dann jeweils:

$\bar{x} = 20$ $\bar{x}^* = 19$ $\bar{x}^{**} = 21$

$z_3 = 0$ $z_3^* = 48$ $z_3^{**} = -48$

in der Tat ist mit $\bar{x}$=20, c=10, α=0,05
$\bar{x}^* = x,^- - 2c\alpha$ und $\bar{x}^{**} = \bar{x} + 2c\alpha$
$z_3^* = 2c^3\alpha(6h_1 - 1 - 8\alpha^2)$ und $z_3^{**} = - z_3,^*$.

Mit c = 10 und h_1 = ¼ gilt: $z_3^* = 2c^3\alpha(6h_1 - 1 - 8\alpha^2) = 2000\alpha(½ - 8\alpha^2)$
Für verschiedene Werte von α erhält man für z_3^*

α	0	0,05	0,10	0,15	0,20	0,25
$z_3,^*$	0	48	84	96	72	0

Die Funktion $z_3^* = 2000\alpha(½ - 8\alpha^2)$ hat ein Maximum an der Stelle
$\alpha = \sqrt{5/24} = 0{,}14434$ und beträgt an dieser Stelle $z_3^* = 96{,}225$.

Fechnersche Lageregel

Bei Häufigkeitsverteilungen gilt in der Regel, wenn sich die Daten über die x-Achse ausreichend dicht verteilen (und eine eingipflige Verteilung vorliegt) die Lageregel von Fechner:

- bei symmetrischer Verteilung gilt: $x = \tilde{x}_{0.5} = \bar{x}_M$
- bei linkssteiler (rechtsschiefer) Verteilung:
  arithmetisches Mittel ($\bar{x}$) > Median ($\tilde{x}_{0.5}$) > Modus ($\bar{x}_M$)
- bei rechtssteiler (linksschiefer) Verteilung:

arithmetisches Mittel ($\bar{x}$) < Median ($\tilde{x}_{0.5}$) < Modus ($\bar{x}_M$)
Also gilt

(5.59) linkssteil: $\bar{x}_M < \tilde{x}_{0.5} < \bar{x}$ rechtssteil: $\bar{x} < \tilde{x}_{0.5} < \bar{x}_M$.

Ausnahmen sind denkbar, insbesondere dann, wenn nur wenige Einzelwerte vorliegen (Bsp. 5.22). Auf der Basis der Fechnerschen Lageregel lassen sich auch Schiefemaße konstruieren.

Beispiel 5.21:

Man verifiziere die Fechnersche Lageregel für das Beispiel 5.17!

Lösung 5.21:

Im Bsp. 5.17 erhält man die folgenden Lageparameter

Verteilung	Modus	Median*)	arithmet. Mittel	Urteil**)
A	70	71,25	75	linkssteil $\bar{x}_M < \tilde{x}_{0.5} < \bar{x}$
B	75	75	75	symmetrisch
C	80	73,75	75	rechtssteil***)
D	75	75	75	symmetrisch

*) mit Interpolation.
**) mit der Fechnerschen Lageregel (Gl. 5.59); eine Beurteilung aufgrund der Momentschiefe erfolgt in Bsp. 5.25.
***) Allerdings ist hier der Median nicht größer, sondern kleiner als das arithmetische Mittel.

Beispiel 5.22:

Gegeben seien die Werte 10,16,17,20,22. Bestimmen Sie den Median und das arithmetische Mittel! Ist die Verteilung symmetrisch?

Lösung 5.22:

Es gilt $\bar{x} = \tilde{x}_{0,5} = 17$. Die Häufigkeitsverteilung ist gemessen an der Gleichheit von $\bar{x}$ und $\tilde{x}_{0,5}$ symmetrisch. Das dritte zentrale Moment beträgt aber $z_3 = -192/5 = -38,4$, so dass die Verteilung danach rechtssteil ist.

b) Schiefemaße

Angesichts der Schwierigkeiten, Schiefe zu definieren überrascht es nicht, dass sich Schiefemaße nicht auf eine anerkannte Axiomatik stützen

können und dass die Messung der Schiefe auf verschiedenen Konzepten beruht. Man kann von Schiefemaßen aber mindestens fordern, dass sie, wie auch andere Gestaltmaße (Formmaße), wie etwa die Streuung oder Wölbung

1. invariant sind gegenüber Translationen
2. die Richtung der Asymmetrie korrekt anzeigt: die Konvention über das Vorzeichen eines Schiefemaßes SK ist
   SK = 0 wenn die Verteilung symmetrisch ist
   SK > 0 wenn sie linkssteil ist (positive Schiefe)
   SK < 0 wenn sie rechtssteil ist (negative Schiefe).

Alle in Def. 5.12 präsentierten Schiefemaße erfüllen diese Forderungen. Schiefemaße sind in der Regel nicht beschränkt, so dass meist gilt:
$-\infty < SK < +\infty$.

Man kann Schiefemaße entwickeln auf der Basis

- der ungeraden zentralen Momente, wobei man zur Rechenvereinfachung von z_3 ausgeht (Konzept der Momentschiefe);
- der Abstände gewisser Lageparameter untereinander (nach dem in Def. 5.11 präsentierten Symmetriebegriff) oder auf der Basis
- der Fechnerschen Lageregel.

Def. 5.12: Schiefemaße

a) Die von Bowley und Fisher eingeführte **Momentschiefe** (Momentkoeffizient der Schiefe) lautet:

$$(5.60) \quad SK_M = \frac{z_3}{s^3} \qquad (\text{zu } z_3 \text{ vgl. Gl. 5.55}).$$

b) Als Quantilskoeffizient der Schiefe wird bezeichnet:

$$(5.61) \quad SK_{Q,\,p} = \frac{(\tilde{x}_{1-p} - Q_2) - (Q_2 - \tilde{x}_p)}{\tilde{x}_{1-p} - \tilde{x}_p} \quad (p < ½)$$

wobei $Q_2 = \tilde{x}_{0.5}$ = Median; der bekannteste spezielle Koeffizient (p = ¼) ist der Quartilskoeffizient der Schiefe (nach Yule und Bowley):

$$(5.62) \quad SK_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \quad .$$

c) Auf der Fechnerschen Lageregel beruhen die folgenden, von (Yule und) Pearson vorgeschlagenen Schiefmaße ($\bar{x}_M$ = Modus):

$$(5.63)\quad SK_{P1} = \frac{\bar{x} - \bar{x}_M}{s} \qquad \text{und} \qquad (5.64)\quad SK_{P2} = \frac{3(\bar{x} - \tilde{x}_{0,5})}{s}$$

Das zweite Pearsonsche Schiefemaße ist meist vorzuziehen, weil oft der Modus schwer zu bestimmen ist; (vgl. Bsp. 5.23).

Bemerkungen zu Def. 5.12

1. Die auf verschiedenen Konzepten beruhenden Schiefemaße sind nicht miteinander vergleichbar, d.h. es ist denkbar, dass eine Verteilung, die gemessen an der Momentschiefe linkssteil ist, nach anderen Schiefemaßen als symmetrisch oder rechtssteil beurteilt wird.

2. Liegt Symmetrie im Sinne der Def. 5.10 bzw. 5.11 vor, so sind alle Schiefemaße Null. Die Umkehrung dieses Satzes ist jedoch nicht zulässig.

3. Die Division durch s bzw. s^3 soll sicherstellen, dass das Schiefemaß dimensionslos ist, also nicht abhängig von der Maßeinheit der Variablen X ist.

4. Es gilt häufig mit guter Näherung $\bar{x} - \bar{x}_M \approx 3\,(\bar{x} - \tilde{x}_{0,5})$ worauf die Pearsonschen
Schiefemaße beruhen.

5. Als Schiefemaße werden in der Literatur auch Maßzahlen präsentiert, wie $(Q_3 + Q_1) / (Q_3 - Q_1)$, was nichts anderes ist als der reziproke Quartilsdispersionskoeffizient, oder die Größe $(Q_3 - Q_2)/(Q_2 - Q_1)$, die hinsichtlich des Vorzeichens offensichtlich die Anforderungen an ein Schiefemaß erfüllt.

6. Eine Variante des Quantilskoeffizienten der Schiefe ist:

$$(5.61a)\quad SK_{Q;0,2} = \frac{(\tilde{x}_{0,8} - Q_2) - (Q_2 - \tilde{x}_{0,2})}{\tilde{x}_{0,8} - \tilde{x}_{0,2}} \qquad (p = 0,2)$$

auf der Basis von Quintilen. Dieser Quintilskoeffizient der Schiefe ist nicht zu verwechseln mit der Quintilenschiefe q, die ein Disparitätsmaß (vgl. Kap. 6) ist. Sie lautet (in der Symbolik von Kap. 6)

$$(5.65)\quad q = \Sigma|h_j - 0,2|$$

und läßt sich mit den Angaben von Beispiel 5.24 errechnen (vgl. dort).

6. Aus der ersten Schreibweise von SK_Q in Gl. 5.62 ist bereits erkennbar, dass die Beziehung $-1 \le SK_{Q,p} \le +1$ gilt. Man kann ferner zeigen, dass gilt $-3 \le SK_{P2} \le +3$.

Wird p in $SK_{Q,p}$ nicht zu klein gewählt, so ist der mittlere Teil und nicht der extrem niedrige oder hohe Abschnitt der Häufigkeitsverteilung schiefebestimmend und $SK_{Q,p}$ ist dann relativ resistent gegenüber Ausreißern.

Beispiel 5.23:

Durch Variation des Beispiels 5.18 sind die folgenden drei Häufigkeitsverteilungen entstanden, die von A nach C [wie eine graphische Darstellung zeigen würde] im anschaulichen Sinne immer linkssteiler werden:

Verteilung A

x_j	n_j
15	3
20	4
25	1
30	1

Verteilung B

x_j	n_j
15	5
20	2
25	1
40	1

Verteilung C

x_j	n_j
15	6
30	1
60	1

Das arithmetische Mittel ist jeweils 20. Man bestimme die Momentschiefen SK_M der drei Häufigkeitsverteilungen!

Lösung 5.23:

Zur Berechnung des zweiten und dritten zentralen Moments bei der Verteilung A wird die folgende Arbeitstabelle aufgestellt:

$x_j - \bar{x}$	$(x_j - \bar{x})^2$	$(x_j - \bar{x})^2 n_j$	$(x_j - \bar{x})^3$	$(x_j - \bar{x})^3 n_j$
15 - 20= - 5	25	75	-125	-375
+5	25	25	125	+125
+10	100	100	1000	+1000
		$\sum$ 200		$\sum$ 750

Daraus ergibt sich:

Verteilung A: $z_2 = s^2 = 200/9$, $z_3 = 750/9$, so dass $SK_M = +0{,}7955$

In entsprechender Weise errechnet man:

Verteilung B: $z_2 = s^2 = 550/9$, $z_3 = 7500/9$ und $SK_M = +1{,}7444$
Verteilung C: $z_2 = s^2 = 1850/8$, $z_3 = 64250/8$ und $SK_M = +2{,}2838$.

Die Schiefekoeffizienten sind im Einklang mit der anschaulichen Vorstellung, dass die Verteilungen von A über B nach C zunehmend linkssteiler werden.

Beispiel 5.24:

Gegeben ist die folgende Verteilung der Haushaltsnettoeinkommen*)

Einkommen		relative Häufigkeit		Anteil**)
von...bis unter...		h_i	H_i	q_i
0	1000	0,1	0,1	
1000	1400	0,1	0,2	0,1
1400	1500	0,05	0,25	
1500	2000	0,15	0,4	0,12
2000	2400	0,1	0,5	
2400	2600	0,1	0,6	0,2
2600	3400	0,15	0,75	
3400	3800	0,05	0,8	0,24
3800	4800	0,1	0,9	
über 4800		0,1	1	0,34

*) Die Zahlenangaben sind so gewählt, dass sich bestimmte Quantile leicht bestimmen lassen. Damit ergibt sich eine stark vereinfachte Darstellung der Einkommensverteilung, wie sie in dieser Form 1988 in der Bundesrepublik bestand.

**) Anteil des ersten,...,fünften Quintils am Gesamteinkommen aller Haushalte.

Berechnen Sie den Quartils- und den Quintilskoeffizient der Schiefe, die Quintilenschiefe sowie die Schiefemaße von Pearson! Das arithmetische Mittel beträgt 2650 DM.

Lösung 5.24:

a) Quartilskoeffizient der Schiefe: Quartile $Q_1 = 1500$, $Q_2 = 2400$ und $Q_3 = 3400$; damit ist der Koeffizient (1000-900)/(1000+900) = +0,053.

b) Quintilskoeffizient der Schiefe:
Quintile: $Q_1^* = 1400$, $Q_4^* = 3800$, so dass man für den Koeffizienten erhält [(3800-2400)-(2400-1400)]/(3800+1400) = +0,167.

c) Quintilenschiefe: Es sind die absoluten Abweichungen der Größen 0,1; 0,12; 0,20; 0,24 und 0,34 von 0,2 zu bilden und zu addieren. Das Ergebnis ist dann q = 0,36 (ein Maß der Disparität [vgl. Kap.6]).

d) Schiefemaße von Pearson: der Modus ist hier schwer zu bestimmen, weil er abhängig ist von der Klasseneinteilung; sinnvoller ist es SK_{P2} zu bestimmen (allerdings ist aus den Angaben auch die Standardabweichung nicht zuverlässig zu bestimmen). Der Zähler von SK_{P2} ist positiv, weil der Zentralwert (Median) mit $\tilde{x}_{0,5} = 2400$ kleiner ist als das arithmetische Mittel (das mit $\bar{x} = 2650$ angegeben ist). Auch nach der Lageregel von Fechner ist die Verteilung linkssteil.

Exkurs: Schiefediagramm

Ein Schiefediagramm ist eine graphische Darstellung zur Beurteilung der Asymmetrie, was evtl. aufschlußreicher sein kann als die Berechnung einer summarischen Maßzahl. Ausgehend vom Median $\tilde{x}_{0,5} = Q_2$ werden die Abstände ausgewählter Quantile vom Median in einem rechtwinkligen Koordinatensystem wie folgt eingetragen:

Abszisse $\tilde{x}_{1-p} - \tilde{x}_{0,5}$ Ordinate $\tilde{x}_{0,5} - \tilde{x}_p$.

Mit dem Beispiel 5.25 soll die Konstruktion des Schiefediagramms demonstriert werden. Dieses Beispiel ist kennzeichnend für eine ausgesprochen rechtssteile Verteilung, bei der die Punkte überwiegend (in Abb. 5.5 sogar alle Punkte) rechts unterhalb der 45º-Linie liegen. Die 45º-Linie ist Ausdruck der Gleichheit $\tilde{x}_{1-p} - \tilde{x}_{0,5} = \tilde{x}_{0,5} - \tilde{x}_p$, also der Symmetrie gem. Def.5.11 (vgl.Bem. 2 zu Def. 5.11). Im Falle einer linkssteilen Verteilung liegen die Punkte dagegen überwiegend oberhalb der 45°-Linie, da für viele p gilt:

$\tilde{x}_{1-p} - \tilde{x}_{0,5} < \tilde{x}_{0,5} - \tilde{x}_p$.

Beispiel 5.25:

Man bestimme das Schiefediagramm für ausgewählte Werte von p für das Beispiel 5.6!

Lösung 5.25:

Die vorgegebenen Beobachtungswerte des Beispiels (21, 25, 34, 39, 43, 52, 64, 72, 80) sowie bestimmte Quantile sind im folgenden untereinander aufgelistet (der Median beträgt 43):

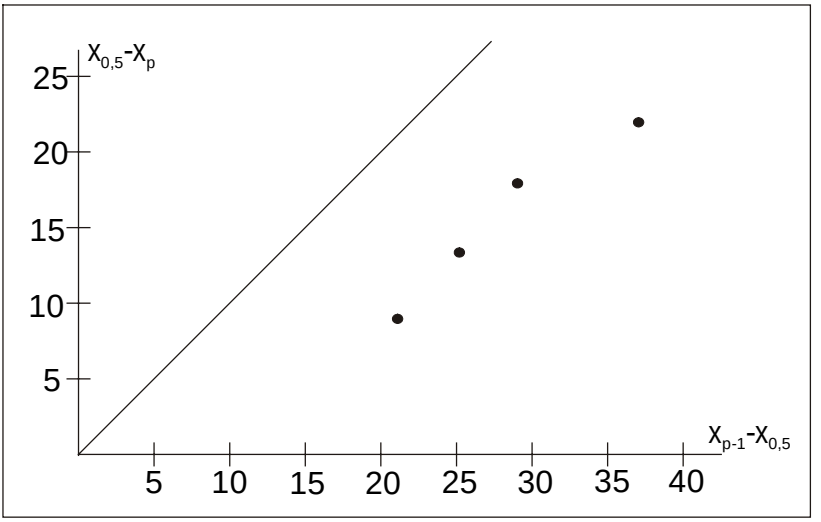
Wert	Quantil	$\tilde{x}_{0,5} - \tilde{x}_p$	Wert	Quantil	$\tilde{x}_{1-p} - \tilde{x}_{0,5}$
21	$=\tilde{x}_{1/9}$	43-21=22	80	$=\tilde{x}_{8/9}$	80-43=37
25	$=\tilde{x}_{0,2}{}^{+}$	43-25=18	72	$=\tilde{x}_{0,8}{}^{+}$	72-43=29
29,5*	$=\tilde{x}_{0,25}=Q_1$	43-29,5=13,5	68*	$=\tilde{x}_{0,75}=Q_3$	68-43=25
34	$=\tilde{x}_{1/3}$	43-34=9	64	$=\tilde{x}_{2/3}$	64-43=21

$^{+}$ erstes, bzw. fünftes Quintil

* interpoliert

Die Punkte mit den Koordinaten (37,22), (29,18), (25,13½), und (21,9) bilden das Schiefediagramm (Abb. 5.5).

Abb. 5.5: Schiefediagramm für das Beispiel 5.25



c) Symmetrisierende Transformationen

Zur Beseitigung einer Asymmetrie in den Daten werden bestimmte Transformationen empfohlen, deren bekannteste die Potenztransformation ist.

Def. 5.13: Potenztransformation

Die Variable X wird in die Variable Y nach Maßgabe einer Potenztransformation transformiert, wenn gilt

$$(5.66) \quad y_v = \begin{cases} (x_v + c)^p & \text{für } p \neq 0 \\ \ln(x_v + c)^p & \text{für } p = 0 \end{cases}$$

Bemerkungen zu Def. 5.13

1. Man spricht auch von einer Leiter der Transformationen (ladder of powers), weil der (nur durch trial and error zu findende) Parameter p beliebige Werte annehmen kann. Wichtige Spezialfälle sind (bei c=0):

   p = -1 $\quad y = 1/x$
   p = 1/2 $\quad y = \sqrt{x}$ (Wurzeltransformation)
   p = 1 $\quad y = x$ (Lineartransformation)

2. Ist p < 1, so werden die größeren Werte von X stärker reduziert (gestaucht) als die kleineren Werte. Diese Transformation eignet sich für linkssteile Verteilungen von X (die Verteilung von Y kann dann [nahezu] symmetrisch sein). Für p > 1 gilt dann das Umgekehrte.

3. Negative Werte von p führen zu einer Vertauschung der Reihenfolge, d.h. ist $x_1 < x_2$ dann ist $y_1 > y_2$.

4. Die Konstante c wird eingeführt (und so bemessen) damit die Variable Y nicht negativ wird.

5. Beispiel 5.26 zeigt für einige Werte von x und p die Wirkung der Transformation.

Beispiel/Lösung 5.26:

Wirkung der Potenztransformation

x	Transformierte Werte y bei					
	p = -½	p = 0	p = ½	p = 0,9	p = 1,5	p = 2
10	0,3162	2,3026	3,1623	7,9433	31,6228	100
20	0,2236	2,9957	4,4721	14,8227	89,4427	400
30	0,1826	3,4012	5,4772	21,3506	164,3168	900
40	0,1581	3,6889	6,3246	27,6601	252,9822	1600

Man erkennt: gleichen Abständen zwischen den x-Werten entsprechen nicht mehr gleiche Abstände zwischen den y-Werten. Letztere werden mit zunehmenden x absolut kleiner bei $|p| < 1$ und größer wenn gilt $|p| > 1$. Ist p negativ, so vertauscht sich die Reihenfolge.

7. Wölbung

Verteilungen können sich danach unterscheiden, wie sehr sich die Merkmalswerte in der Mitte oder an den Enden häufen, bzw. je steiler ihr Verlauf in der Umgebung des Medians ist. Man spricht dann von verschiedenen Arten und Stärken der Wölbung [synonym: Kurtosis, Exzess] einer (meist symmetrischen und eingipfligen) Verteilung. In Abb. 5.4 ist beispielsweise die Verteilung B flacher (schwächer) und die Verteilung D steiler (stärker) gewölbt. Die Wölbung ist ein ähnlicher Aspekt einer Häufigkeitsverteilung wie die Streuung, gleichwohl aber hiervon zu unterscheiden: denn die Verteilungen B und D in Abb.5.4 haben, gemessen an der Standardabweichung s, die gleiche Streuung, aber eine unterschiedliche Wölbung.

Maße der Wölbung W (Kurtosis) sollten als Formmaßzahlen (Gestaltparameter) - wie die Schiefemaße - invariant sein gegenüber linearen Transformationen: bei $y_v = a + b_x$ soll gelten $W_y = W_x$ ($b \neq 0$ [bei der Schiefe ist $b > 0$ zu fordern])

Wie in Abschnitt 5 bereits bemerkt, kann mit Hilfe des vierten zentralen Moments z_4 eine Maßzahl W_M der Wölbung gebildet werden, die angibt, inwieweit sich die Wölbung einer bestimmten Verteilung von der einer Normalverteilung unterscheidet.

Def. 5.14: Wölbungsmaße

a) Beim Wölbungsmaß W_M wird das vierte zentrale Moment durch die quadrierte Varianz (denn $(s^2)^2 = s^4$) geteilt:

$$(5.67)\quad W_M = \frac{z_4}{s^4} - 3 .$$

b) Weniger bekannt sind Wölbungsmaße auf der Basis von Quantilen, etwa ein Quantilskoeffizient W_Q der Wölbung:

$$(5.68)\quad W_Q = 1 - \frac{\tilde{x}_{1-p} - \tilde{x}_p}{\tilde{x}_{1-q} - \tilde{x}_q}$$

mit $0 < q < p < ½$.

Bemerkungen zu Def. 5.14:

1. Man sieht, dass in W_M das vierte zentrale Moment relativiert (und damit maßstabsunabhängig) wird durch s^4. Es läßt sich zeigen, dass der Ausdruck z_4 / s^4 für die Normalverteilung den Wert 3 annimmt. Deshalb gilt:

   - $W_M = 0$ bei der Normalverteilung, bzw. einer Häufigkeitsverteilung die genauso gewölbt ist wie die Normalverteilung (man sagt dann, sie sei ***mesokurtisch***),
   - $W_M > 0$ bei Häufigkeitsverteilungen, die vergleichsweise steiler als die Normalverteilung gewölbt sind (***leptokurtisch*** = hochgewölbt, spitz),
   - $W_M < 0$ bei Häufigkeitsverteilungen, die vergleichsweise flacher als die Normalverteilung gewölbt sind (***platykurtisch*** = flachgewölbt).

   In den Beispielen 5.27 und 5.28 wird die Vorgehensweise zur Berechnung der Wölbung W_M dargestellt.

2. Der Momentkoeffizient der Kurtosis W_M hat, wie gezeigt werden kann, den folgenden Wertebereich: $-2 < W_M < \infty$.

3. Ein Beispiel (Spezialfall) für den Quantilskoeffizient W_Q der Kurtosis wäre

   $$(5.68a)\quad W^*_{,Q} = 1 - \frac{Q_3 - Q_1}{Q_4^* - Q_1^*}$$

   mit $q = 0{,}2$ und $p = 0{,}25$ unter Verwendung des ersten und dritten Quartils (Q_1 und Q_3) sowie des ersten und vierten Quintils Q_1^* und Q_4^*. Im Beispiel 5.29 wird die Berechnung dieses Wölbungsmaßes gezeigt.

Beispiel 5.27:

Man bestimme die Momentkoeffizienten der Schiefe und Wölbung für die vier Verteilungen des Beispiels 5.17. Ändern sich Schiefe und Wölbung, wenn man zu allen Merkmalswerten die Zahl 5 addiert?

Lösung 5.27:

Für die zentralen Momente erhält man

Verteilung	z_3	z_4
A	+150	3600
B	0	2700
C	-150	3600
D	0	5100

Da die Standardabweichung bei allen Verteilungen s = 6 ist, ergibt sich für die:

	Momentschiefe z_3/s^3	Wölbung $(z_4/s^4) - 3$
Verteilung A:	+0,694	-(8/36) = - 0,2222
Verteilung B:	0	-33/36 = - 0,9167
Verteilung C:	-0,694	-0,2222
Verteilung D:	0	+0,9352

Schiefe und Wölbung ändern sich nicht, wenn man zu allen Merkmalswerten die Zahl 5 addiert, weil sich dann auch das arithmetische Mittel um 5 erhöht und nur **zentrale** Momente verwendet werden.

Beispiel 5.28:

Durch Variation des Beispiels 5.18 sind die folgenden Verteilungen A bis D erzeugt worden, die anschaulich im Zentrum zunehmend steiler verlaufen (zunehmende Wölbung):

B

x_j	n_j
10	1
15	2
20	3
25	2
30	1

C

x_j	n_j
10	1
15	1
20	5
25	1
30	1

D

x_j	n_j
10	0
15	1
20	7
25	1
30	0

Die Verteilung A besteht aus den Einzelwerten 0,5,10,15,20,25,30, 35 und 40. Alle Verteilungen haben das gleiche arithmetische Mittel von $\bar{x} = 20$.

Lösung 5.28:

	A	B	C	D
z_2	1500/9	300/9	250/9	50/9
z_4	442500/9	22500/9	21250/9	1250/9
$z_4/s^4 - 3$	-1,23	-0,75	+0,06	+1,5

Beispiel 5.29:

Man berechne und interpretiere die Kurtosis W_Q^* gem. Gl. 5.68a für das Bsp. 5.24!

Lösung 5.29:

Die Quintile sind in diesem Fall $Q_1^* = 1400$ und $Q_4^* = 3800$ und für die Quartile erhält man $Q_1 = 1500$ und $Q_3 = 3400$, so dass Gl.5.68a ergibt $W_Q^* = 1 - (3400-1500) / (3800-1400) = 1- 1900/2400 = 0{,}21$. Es ist stets $(Q_3 - Q_1) \leq (Q_4^* - Q_1^*)$, so dass der von 1 zu subtrahieren-de Quotient nicht größer sein kann als 1. Verabredet man die folgenden Strecken unter der Häufigkeitsverteilung:

$A = Q_3 - Q_1$ $\qquad$ $C = Q_1 - Q_1^*$ und

$B = Q_4^* - Q_3$ $\qquad$ $D = Q_4^* - Q_1^*$

so ist offenbar $D = A + B + C$ und $W_Q^* = (B+C)/D$. Bei zunehmender Wölbung wird die zentrale Strecke A kleiner und die Ausläufer B und C werden (relativ zu A) größer. Dann muss W_Q^* zunehmen.

Kapitel 6: Konzentrations- und Disparitätsmessung

1. Konzentrationsbegriff und Konzentrationsmessung

a) Absolute und relative, statische und dynamische Konzentration

Konzentration im wirtschaftlichen Sinne kann zweierlei bedeuten:

1. eine Ballung von Verfügungsmacht, Marktanteilen o.ä. auf wenige Einheiten (z.B. Marktbeherrschung) und
2. die Existenz erheblicher Größenunterschiede zwischen den Einheiten ("Ungleichheit").

Einmal wird auf die absolut geringe Anzahl der wirtschaftlichen Einheiten abgestellt (Anzahlaspekt der Konzentration), im anderen Fall auf die Ungleichheit der auf die Einheiten entfallenden Anteile am gesamten Merk-

malsbetrag (Disparitätsaspekt der Konzentration). Die statistischen Maße der absoluten Konzentration (Konzentration im engeren Sinne) berücksichtigen beide Aspekte, die der Disparität (oder: relativen Konzentration) nur den zweiten.

Beispiele

Eine Aussage im Sinne der relativen Konzentration ist zum Beispiel:

- 1,7% der Bevölkerung haben mehr als 70% des Produktivvermögens (sowohl die Merkmalsträger [Bevölkerung] als auch die auf sie entfallenden Merkmalbeträge [Anteile am Produktivvermögen] sind relativiert [Prozentangaben]).

Eine Aussage im Sinne der absoluten Konzentration ist dagegen

- auf einem bestimmten Markt haben nur 3 Unternehmen (die drei größten Unternehmen) zusammen einen Marktanteil von 90% (die Merkmalsträger sind in absoluter Zahl angegeben, es kommt auf die absolut geringe Anzahl an).

Im Falle der sog. egalitären Verteilung (Verteilung der Merkmalssumme auf n Einheiten zu gleichen Anteilen 1/n) wird der Unterschied besonders deutlich. Die relative Konzentration (oder: Disparität) ist jeweils Null, die absolute Konzentration dagegen umso größer, je kleiner n ist.

Neben der soweit dargestellten statischen Betrachtung eines Konzentrationszustands wird auch häufig eine **dynamische Konzentrationsmessung** gefordert, in der die Veränderung der Verteilung (des Konzentrationsmerkmals) im Zeitablauf geeignet beschrieben wird. Die Darstellung dieser Art von Mobilität macht Verlaufsanalysen erforderlich und führt zu sehr komplizierten Modellen der **Unternehmensdemographie**, die Zu- und Abgänge, Wachstum, Diversifikation, Aktivitätsverlagerung usw. berücksichtigen müßten (auch unter Berücksichtigung von Methoden des Kapitels 12) und auch aus Gründen des Datenschutzes kaum für empirische Untersuchungen angewandt werden können, so dass z.Zt. noch statische oder komparativ-statische Betrachtungen dominieren.

b) Konstruktion von Konzentrations- und Disparitätsmaßen

1. Notwendigkeit von Gedankenexperimenten

In der wirtschaftlichen Realität sind absolute und relative Konzentration nicht zwei streng unterschiedene Erscheinungen, sondern zwei in der Regel gemeinsam auftretende Aspekte eines Vorgangs. Neugründungen, Fu-

sionen, Auflösungen oder -teilungen, ungleiches Größenwachstum usw. berühren meist beide Arten von Konzentration und damit auch beide Arten von statistischen Maßzahlen gleichzeitig, wenngleich häufig in unterschiedlicher Weise. Dem steht jedoch nicht entgegen, dass man modellmäßige Vorgänge konstruieren kann, die sich isoliert auf einen der beiden Aspekte der Konzentration auswirken. Das geschieht bei der Entwicklung einer Axiomatik. In Form von sog. "Proben" werden die Auswirkungen solcher Vorgänge auf die Maßzahlen der Konzentration und Disparität untersucht.

Auch die Unterscheidung zwischen einer statischen und dynamischen Betrachtung ist nur eine gedankliche Abstraktion. In der Realität entwickeln sich Konzentrationszustände aus Konzentrationsprozessen und es ist fraglich, ob eine Momentaufnahme den Sachverhalt überhaupt hinreichend beschreiben kann. Wegen der Komplexität dieser Prozesse ist es auch hier notwendig, in Gedankenexperimenten isoliert den Effekt einfacher Veränderungen an den Daten "durchzuspielen".

2. Daten, im folgenden verwendete Symbole

Grundlage konzentrationsstatistischer Betrachtung sind die im folgenden definierten Größen n, h_i, q_i, bzw. c_i:

Def. 6.1: Anteile, Merkmalsanteile

Es sollen die folgenden Symbole verabredet werden:

1. Die Anzahl der **Merkmalsträger** ist n, ihre Anteile an der Gesamtheit der Merkmalsträger $h_i = 1/n$ (bei Einzelbeobachtungen) bzw. $h_i = n_i /n$ bei gruppierten oder klassierten Daten

2. Die Merkmalsanteile, d.h. die Anteile an dem **Merkmalsbetrag** (Merkmalssumme der zu verteilenden Größe) lauten q_i. Wenn besonders hervorgehoben werden soll, dass die Merkmalsanteile sich auf in einer bestimmten Weise geordnete Merkmalsträger beziehen kann statt q_i auch das Symbol c_i verwendet werden (vgl. Gl. 6.2).

Die folgenden Betrachtungen beziehen sich auf Einzelwerte. Unterschiede für den Fall klassierter Daten werden an späterer Stelle behandelt.

3. Konzentrationsmaße

Aufgabe der Deskriptiven Statistik ist es auch hier, Zustände

a) graphisch, bzw. tabellarisch
b) durch eine zusammenfassende Maßzahl (ein "Konzentrationsmaß") zu beschreiben.

Konzentrationsmaße, in deren Berechnung alle Anteile q_i eingehen, heißen summarische, solche dagegen, die nur einen Teil der Anteile q_i berücksichtigen heißen diskrete Konzentrationsmaße (vgl. auch Abschn. 6).

4. Axiomatik

Die Entwicklung einer Axiomatik erfolgt, wie gesagt, durch Gedankenexperimente in Form von sog. "Proben". So beschreibt z.B. die "Ergänzungsprobe" (Hinzufügen von Einheiten, deren Merkmalsbeträge jeweils Null sind) eine isolierte fiktive Vergrößerung der Disparität. Entsprechend untersucht die "Proportionalitätsprobe" (Aufteilung jeder Einheit in k genau gleich große Einheiten) die Auswirkungen des reinen Anzahleffekts auf die statistischen Maßzahlen. Maße der Konzentration und Disparität sollen auf diese beiden Proben unterschiedlich reagieren. Sie sind also auch durch ihre im Rahmen einer Axiomatik (zu der die beiden genannten "Proben" gehören) geforderten Eigenschaften unterschieden.

Übersicht 6.1

Darstellung	(absolute) Konzentration	(relative Konzentration) Disparität
a) graphisch	Konzentrationskurve	Lorenzkurve
b) Maße - summarisch - diskret	K_R(Rosenbluth)-Index*) Herfindahl-Index K_H concentration ratios	D_G Gini-Koeffizient **) Variationskoeffizient Schutz-Koeffizient***

*) aus der Konzentrationskurve entwickelt,

**) aus der Lorenzkurve entwickelt.

***) auch maximaler Nivellierungssatz von Lindahl genannt

5. Wirtschaftsstatistische Fragen

Bei der Durchführung einer Konzentrationsmessung im wirtschaftlichen Bereich sind eine Reihe wirtschaftsstatistischer Fragen zu klären: Wahl eines geeigneten Konzentrationsmerkmals (z.B. Umsatz oder Beschäftigtenzahl als Maß der Unternehmensgröße) oder z.B. die "Abgrenzung des relevanten Marktes" (d.h. der zugrundegelegten Masse)[1].

[1] Ausführlicher zu diesen Aspekten vgl. von der Lippe, P.: "Wirtschaftsstatistik" UTB Bd. 209, 4. Aufl., S. 230f.

6. Anforderungen an ein Konzentrationsmerkmal

Ein Konzentrationsmerkmal muss sein

- nichtnegativ
- extensiv und
- mindestens intervallskaliert.

7. Extremfälle der Konzentration

Die Entwicklung eines Axiomensystems wird durch die Existenz von zwei klar definierten Extremfällen der Konzentration erleichtert. Es ist insbesondere möglich, so Konzentrations-, bzw. Disparitätsmaße zu normieren. Es sind die folgenden zwei Extremsituationen zu unterscheiden:

a. **egalitäre Verteilung** (Einpunktverteilung oder extreme Gleichverteilung), die Situation der minimalen Disparität: jeder der n Merkmalsträger hat den gleichen Merkmalsbetrag x und damit auch den gleichen Merkmalsanteil

$$q_i = \frac{1}{n} \qquad i=1,2,...,n.$$

wobei diese Anteile auch in einem Spaltenvektor **q** zusammengefaßt werden.

Bei einer Disparitäts von Null wird auch mißverständlich von "Gleichverteilung"[2] geprochen.

b. **vollkommene Ungleichheit** (extreme Ungleichverteilung, maximale Konzentration, Zweipunktverteilung): ein Merkmalsträger vereinigt die gesamte Merkmalssumme auf sich, sein Anteil q ist somit 1 und die Anteile q der übrigen n-1 Merkmalsträger sind demnach alle jeweils Null.

Def. 6.2: Disparitäts- und Gleichheitsmaß

Ist D ein Disparitätsmaß, so ist G=1-D ein Gleichheitsmaß

Die im nächsten Abschnitt besprochenen Axiome für Konzentrations- und Disparitätsmaße lassen sich oft leichter formulieren, wenn auf Gleichheit anstatt Disparität abgestellt wird.

---

[2] Der beschriebene Fall der egalitären Verteilung heißt in der Statistik auch **Einpunkt-verteilung** (alle Einheiten haben die gleiche [und damit einzige] Merkmals-ausprägung, deren prozentuale Häufigkeit 100% ist). **Gleichverteilung** heißt dagegen: jede Merkmalsausprägung kommt gleich häufig vor (z.B. die Verteilung der Augenzahl beim Würfeln [eine Wahrscheinlichkeits- nicht Häufigkeitsverteilung]).

2. Eigenschaften von Konzentrations- und Disparitätsmaßen

a) Axiome

Es sind Eigenschaften zu unterscheiden

a) die sowohl Konzentrations- als auch Disparitätsmaße haben sollten und
b) solche, bei denen sich Konzentrations- und Disparitätsmaße unterscheiden.

Ist speziell ein Konzentrationsmaß gemeint, so wird es hier K genannt, ein Disparitätsmaß im allgemeinen Sinne heißt entsprechend D und bei einer Aussage, die sich sowohl auf Konzentrations- als auch auf Disparitätsmaße bezieht, soll das entsprechende Maß C genannt werden.

zu a) Eigenschaften von Konzentrations- und Disparitätsmaßen

K1: Unabhängigkeit von der Maßeinheit des Konzentrationsmerkmals

Ein Konzentrations- oder Dispariatätsmaß C soll invariant sein bei proportionaler Transformation: Ist $y_i = bx_i$ $(b > 0)$, so ist $C(y) = C(x)$.

K2: Verschiebungsprobe (Transfer)

Wird ein Betrag d mit $0 < d < h/2$ transferiert von einem Merkmalsträger i (mit dem Merkmalsbetrag $x_{(i)}$) zum Merkmalsträger j mit $x_{(j)} = x_{(i)} - h$ also $x_{(j)} < x_{(i)}$, so soll C abnehmen (regressiver [egalisierender, negativer, d.h. die Konzentration verringernder] Transfer).

Die Umkehrung sollte entsprechend bei einem progressiven [positiven] also die Konzentration (und damit auch das Konzentrationsmaß) erhöhenden Transfer ("von arm zu reich") gelten, d.h. das Konzentrationsmaß sollte dann steigen.

K3: (Verschiebung, Niveauänderung)

Sei $y_i = a + x_i$, dann ist bei egalitärer Verteilung des Merkmals X die Konzentration des Merkmals Y gleich, also

$C(y) = C(x)$

und in den sonstigen Fällen soll gelten

$$C(y) = \begin{cases} < C(x) & \text{wenn } a > 0 \quad \text{(abnehmende Konzentration)} \\ > C(x) & \text{wenn } a < 0 \quad \text{(zunehmende Konzentration)} \end{cases}$$

zu b) Eigenschaften, bei denen sich Konzentrations- und Disparitätsmaße unterscheiden

Im Unterschied zu den Axiomen K1 bis K3 wird bei den Axiomen K4 und K5 postuliert, dass Konzentrations- und Disparitätsmaße auf die in den "Proben" unterstellten Vorgänge unterschiedlich reagieren. Die Proportionalitätsprobe (Axiom K4) beschreibt den reinen Anzahleffekt, weshalb Disparitätsmaße hierauf nicht reagieren sollen. Die Forderung K5 ist damit motiviert, dass man gedanklich eine immer größer werdende Ungleichheit durch Hinzufügen von "Nulleinheiten" entstehen lassen kann.

K4: (Proportionalitätsprobe):

Ersetzt man jeden einzelnen Merkmalsträger i mit dem Anteil q_i am Merkmalsbetrag durch $k > 1$ gleich große Merkmalsträger mit den Anteilen q_i/k, so soll für das neue Disparitätsmaß D* gelten:

$D^* = D(\mathbf{q})$ (Disparität bleibt unverändert)

und für das "neue" Konzentrationsmaß K* im Vergleich zum "alten"

$K^* = \frac{K}{k} = \frac{1}{k} K(\mathbf{q})$ (Fall der Dekonzentration).

Entsprechend soll im "umgekehrten" Fall einer Fusion von k gleich großen Einheiten zu einer Einheit gelten $D^*=D$ und $K^*=kK$ (Fusion).

K5: (Ergänzungsprobe, Nullergänzung):

Fügt man einer Verteilung m Einheiten, deren Merkmalsbeträge jeweils Null sind ("Nullträger") hinzu, so soll gelten

$K^* = K$ und $D^* > D$.

K6: Wertebereiche

Die folgenden Wertebereiche sollen gelten: für

Konzentrationsmaße	$\frac{1}{n} \le K \le 1$
Disparitätsmaße	$0 \le D \le 1 - \frac{1}{n}$.

b) Erläuterungen zu den Axiomen

1. Die Axiome K1 bis K3 werden sowohl für Konzentrationsmaße, als auch für Disparitätsmaße gefordert. Das Axiom K1 besagt, dass ein Konzentrations- und ein Disparitätsmaß unabhängig sein soll von der Maßeinheit der zu verteilenden Größe (des Konzentrationsmerkmals). Die "Ungleichheit" der Einkommensverteilung wird nicht davon berührt, ob das Einkommen in DM oder in Pfennigen gerechnet wird. Sie ist in beiden Fällen gleich groß.

   Diese Invarianz gegenüber proportionalen Transformationen (auch "Bresciani-Turoni-Bedingung" genannt) wird dadurch sichergestellt, dass zur Berechnung von Konzentrations- und Disparitätsmaßen nur Merkmals**anteile**, nicht absolute Merkmals**beträge** benutzt werden.

2. Das Axiom K2 (auch "Transfer", "Pigou-Dalton-Effekt" oder [wegen K3] mißverständlich "Verschiebungsprobe" genannt) wird nicht selten nur für Disparitätsmaße gefordert. Bei einem Transfer "von reich zu arm" soll die Disparität abnehmen (negativer -, regressiver - oder egalisierender Transfer). Umgekehrt soll die Disparität zunehmen, wenn d von j auf i transferiert wird (positiver - oder progressiver Transfer).

   Transfers sind stets Umverteilungen bei gleichbleibendem gesamtem Merkmalsbetrag und wegen $d < h/2$ auch bei gleichbleibender Rangfolge der Merkmalsträger (unter denen der Transfer stattfindet). Die Bedingung $d < h/2$ wird deshalb eingeführt ("Überholverbot").
   Es ist auch üblich, eine Reaktion auf einen Transfer für Konzentrations-, nicht nur für Dispariatätsmaße zu fordern, oder aber zu unterscheiden:
   bei einem regressiven Transfer soll K abnehmen, D dagegen strikt abnehmen.
   Gelegentlich wird der Transfer auch "bewertet" in dem Sinne, dass sich ein positiver Transfer auf ein Konzentrations- oder Disparitätsmaß stärker erhöhend auswirken soll, als sich ein betragsmäßig gleich großer negativer Transfer verringernd auswirkt ("**starke Verschiebungsprobe**").

3. In der Literatur wird K3 häufig nur für Disparitätsmaße gefordert. Anders als die in K1 postulierte gleiche **relative** Änderung (z.B. Erhöhung aller Einkommen um 20%, so dass $b = 1{,}2$ ist) soll sich eine gleiche **absolute** Änderung der Merkmalsbeträge durchaus auf die Disparität auswirken. Nach Axiom K3 bedeutet eine für alle Merkmalsträger gleich große Zunahme ($a>0$) der Merkmalsbeträge eine Verringerung der Disparität und entsprechend eine Abnahme ($a<0$) eine Vergrößerung der Disparität.

4. Der mit der Proportionalitätsprobe beschriebene Anzahleffekt soll sich auf ein Disparitätsmaß nicht auswirken, die Konzentration soll sich jedoch proportional zu einer Verringerung (Fusion) der Anzahl der Merkmalsträger vergrößern, bzw. zu einer Vergrößerung (Dekonzentration) der Anzahl der Merkmalsträger verringern.

5. Hinter der Ergänzungsprobe (Axiom K5) steht - wie gesagt - die Vorstellung, dass man sich eine "Ungleichverteilung" durch Hinzufügen von Nullträgern aus der Gleichverteilung entstanden denken kann. In gleicher Weise wird eine bestehende "Ungleichheit" durch Hinzufügen von Nullträgern vergrößert.
   Der Vorgang der Nullergänzung soll dagegen ein Maß der absoluten Konzentration K unverändert lassen, weil die Anzahl derjenigen Merkmalsträger, die sich bisher die Merkmalssumme aufteilten und deren Anteile $q_i>0$ sind, gleich bleiben.
   Mit dem Axiom K5 ist noch nicht gesagt, um wieviel sich D bei Hinzufügen (Wegnehmen) von Nulleinheiten vergrößert (verringert). Eine solche Aussage wird mit der folgenden spezielleren Fassung des Axioms getroffen, die wir Axiom K5a nennen wollen

   Axiom K5a:

   Manchmal wird auch gefordert, dass sich ein Disparitätsmaß in einem bestimmten Ausmaß vergrößert. Fügt man einer Verteilung m=(k-1)n Nullelemente hinzu, so dass der Vektor der Merkmalsanteile $\mathbf{q} = [q_1...q_n]$ in den Vektor $\mathbf{q}_e = [q_1...q_n\ 0...0]$ mit kn Elementen übergeht, so sollte für ein Disparitätsmaß D und für ein Gleichheitsmaß G = 1 - D gelten

$$(6.1) \qquad 1 - D(\mathbf{q}_e) = \frac{1 - D(\mathbf{q})}{k}$$

$$(6.1a) \qquad G(\mathbf{q}_e) = \frac{G(\mathbf{q})}{k}\ .$$

   Das bedeutet z.B., dass durch eine Verdoppelung (k=2) der Anzahl der Merkmalsträger das Gleichheitsmaß halbiert wird. Es handelt sich aber nicht einfach um eine Verdoppelung der ursprünglichen Einheiten (was ein Anzahleffekt im Sinne von K4 wäre), sondern um ein Hinzufügen von n "Nullträgern" zu n Einheiten, deren Merkmalsbeträge größer oder gleich Null sind.

6. Es wird nicht nur gefordert, dass die oben genannte Grenzen gelten, sondern auch, dass sie in genau definierten Situationen angenommen werden. Danach gilt bei den bereits dargestellten extremen Zuständen:

maximale Konzentration	$K_{max} = 1$ und $D_{max} = 1 - \frac{1}{n}$
egalitäre Verteilung	$K_{min} = \frac{1}{n}$ und $D_{min} = 0$.

Man beachte, dass die ***minimale*** (nicht aber die maximale) Konzentration und die ***maximale*** (nicht aber die minimale) Disparität von n abhängig ist. Deshalb kann man auch nicht einfach fordern: $0 \leq D,K \leq 1$, auch nicht wenn man bedenkt, dass 1/n "gegen Null strebt", zumal n bei Konzentrationsuntersuchungen oft nicht sehr groß ist.

- *Konzentration:* War ein Markt bisher von 4 gleich großen Unternehmen beherrscht, so würde ein Übergang zu 3 gleich großen Unternehmen die Disparität nicht berühren (da stets alle Unternehmen gleich groß sind, bleibt D=0), wohl aber die Konzentration von 1/4 auf 1/3 erhöhen (die wegen D=0 minimale Konzentration ist also von n abhängig: je kleiner n desto größer K_{min}), denn im Sinne der absoluten Konzentration ist die Konzentration auf 3 statt 4 Unternehmen natürlich eine Steigerung

- *Disparität:* Sie kann nur maximal sein, wenn einer alles und der Rest (n - 1 Einheiten) nichts bekommt. Es ist klar, dass diese Ungleichheit umso größer ist je größer "der Rest" ist. Es ist deshalb sinnvoll für den Fall, dass eine Einheit die gesamte Merkmalssumme auf sich vereint danach zu differenzieren, wie groß n ist (die Obergrenze von D ist deshalb abhängig von n).

3. Messung der (absoluten) Konzentration

a) Konzentrationskurve und Rosenbluth-Index

Ordnung nach abnehmender Größe

Die Daten liegen in Form von Einzelbeobachtungen $x_1,x_2,...,x_n$ vor. Da Konzentration bedeutet, dass eine geringe Anzahl von Merkmalsträgern einen großen Anteil an der Merkmalssumme auf sich vereinigt, werden die Merkmalsbeträge nach abnehmender Größe geordnet (d.h. die Merkmalsträger werden nach abnehmenden Merkmalsbeträgen geordnet), so dass

$x^{(1)} \geq x^{(2)} \geq ... \geq x^{(n)}$ (abnehmende Merkmalsbeträge).

Bei der Disparitätsmessung ist es dagegen üblich, von einer Ordnung nach zunehmender Größe auszugehen, so dass gilt

$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ (zunehmende Merkmalsbeträge).

Konzentrationsraten, Konzentrationskurve und Rosenbluth-Index

Wir sind jetzt in der Lage, Konzentrationsraten, deren graphische Darstellung, die Konzentrationskurve und den hieraus abgeleiteten Rosenbluth-Index als Konzentrationsmaß (Maß der absoluten Konzentration) zu definieren.

Def. 6.3 : Konzentrationsraten, -kurve, Rosenbluth-Index

a) Wird der Merkmalsanteil des i-ten Merkmalsträgers

$$(6.2) \quad c_i = \frac{x^{(i)}}{\Sigma x^{(i)}} = \frac{x^{(i)}}{\Sigma x_j}$$

genannt (oder äquivalent q_i; vgl. Def. 6.1) dann ist

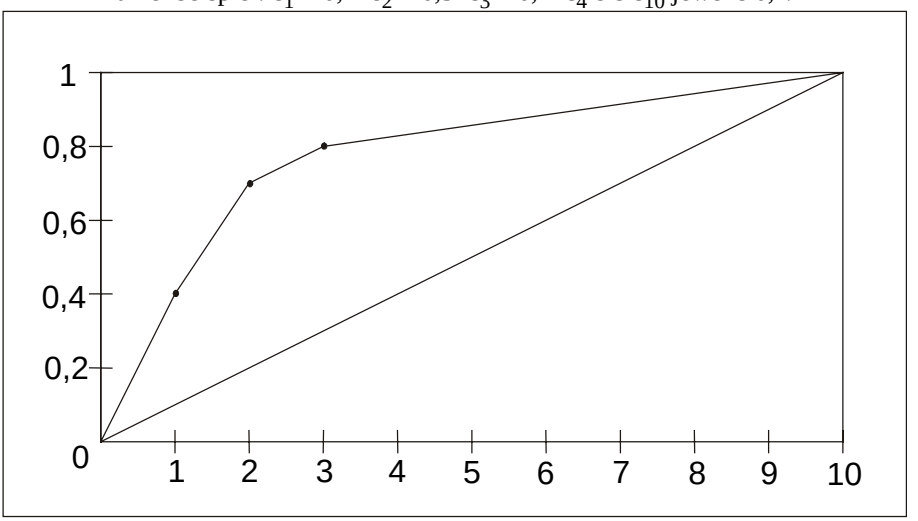
$$(6.3) \quad C_i = \sum_{j=1}^{i} c_j \qquad C_0 = 0 < C_1 < \ldots < C_n = 1$$

der kumulierte Anteil der i größten Merkmalsträger und heißt **Konzentrationsrate** (concentration ratio).

b) Zeichnet man die geordneten Paare (i, C_i) in ein kartesisches Koordinatensystem ein und verbindet man die Punkte mit den Koordinaten (0,0), $(1,C_1)$,..., (n,1), so heißt der daraus resultierende Polygonzug **Konzentrationskurve** (vgl. Abb. 6.1).

Abb. 6.1: Konzentrationskurve

Zahlenbeispiel: $c_1 = 0{,}4$ $c_2 = 0{,}3$ $c_3 = 0{,}1$ c_4 bis c_{10} jeweils 0,2/7



c) Der **Rosenbluth-Index** (Konzentrationskoeffizient von Rosenbluth) ist wie folgt definiert:

$$(6.4) \quad K_R = \frac{1}{2\Sigma i c_i - 1} \qquad \frac{1}{n} \le K_R \le 1 \qquad (i=1,2,...,n).$$

Bemerkungen zur Def. 6.3:

1. Zwischen den Ordnungsstatistiken $x^{(i)}$ und den bisher betrachteten Ordnungsstatistiken $x_{(i)}$ gilt die Beziehung
   $x^{(i)} = x_{(n-i+1)} \quad i = 1,2,...,n.$

2. Die Konzentrationsrate C_r ist der Anteil, den die r größten Merkmalsträger auf sich vereinigen. Wenn z.B. das Merkmal Umsatz betrachtet wird ist C_5 der Umsatzanteil der fünf größten Unternehmen. Konzentrationsraten C_i sind diskrete Konzentrationsmaße, weil sie nur einen Teil der Merkmalsanteile c_i berücksichtigen. Sinnvoll kann nur $r < n$ gewählt werden, weil

   - (absolute) Konzentration bedeutet, dass eine geringere Anzahl r von Merkmalsträgern einen kumulierten Merkmalsanteil C_r hat.
   - für r = n notwendig gilt $C_r = 1$, was keine Information liefert.

   Die Wahl von r ist willkürlich. Die Konzentrationsrate ist darüberhinaus zwar einfach und anschaulich, sie nutzt aber die in der Konzentrationskurve steckende Information nicht voll aus. Bei Untersuchungen über die Konzentration von Unternehmen in bezug auf Umsätze; Marktanteile usw. ist es üblich (z.B. in den Untersuchungen der Monopolkommission) die Konzentrationsraten C_3, C_6, C_{10} zu betrachten.

3. Die Konzentrationskurve ist eine graphische Darstellung der Konzentrationsraten. Aus $x^{(i)} \ge x^{(i+1)}$ folgt

   $$(6.5) \quad c_i \ge c_{i+1} \text{ und } C_i < C_{i+1} \ .$$

   Die Steigung der Konzentrationskurve beträgt im Intervall von i-1 bis i genau c_i und kann wegen (6.5) nicht zunehmen. Im Fall der egalitären Verteilung, bei dem für alle i gilt $c_i = 1/n$ ist die Konzentrationskurve eine die Punkte (0,0),...,(n,1) verbindende Gerade $C_i = i/n$ mit der Steigung 1/n. In allen anderen Fällen verläuft sie als Polygonzug oberhalb dieser Diagonalen konkav, so dass gilt

$$(6.6) \quad C_i \geq \frac{i}{n} \qquad (i = 1,2,...,n).$$

Im Falle der maximalen Konzentration ist $C_1 = 1$ und $C_i = 0$ (für $1 < i < n$).

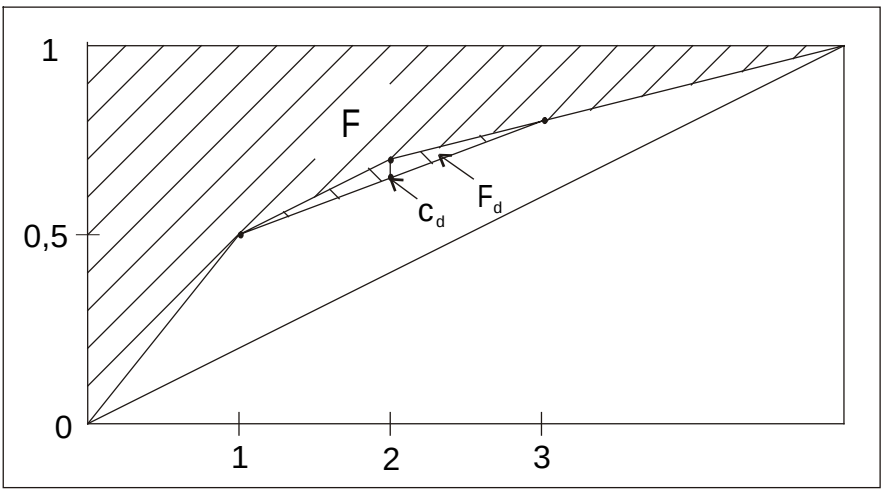
4. Der Rosenbluth-Index stellt ein Maß für die Wölbung der Konzentrationskurve dar. Man kann zeigen, dass K_R invers von der Fläche F (schraffierte Fläche in Abb. 6.2 bis 6.4) abhängt:

$$(6.7) \quad F = \Sigma i c_i - ½ \qquad (½ \leq F \leq n/2) \qquad \text{und}$$

$$(6.7a) \quad K_R = \frac{1}{2F}$$

Aus Gl. 6.7 ist ersichtlich, dass die Reihenfolge der Einheiten nicht beliebig ist. Bei der Konzentrationsmessung werden, wie gesagt, die Merkmalsträger stets nach abnehmenden Merkmalsanteilen c_i geordnet.

Abb. 6.2: Konzentrationskurve und Fläche F (vor und nach negativem Transfer zwischen Einheit 2 und 3)



5. Durch die proportionale Transformation $y_i = b \cdot x_i$ verändern sich die Merkmalsanteile c_i nicht, so dass K_R das Axiom K1 erfüllt.

6. Um zu zeigen, dass K_R das Axiom K2 erfüllt, betrachten wir ohne Beschränkung der Allgemeinheit einen negativen (egalisierenden) Transfer des Betrages d und damit des Anteils

$$c_d = \frac{d}{\Sigma x_i} < \frac{1}{2}(c_i - c_j)$$

von i zu j = i+1. Dadurch wird C_i um c_d geringer und C_{i+1} bleibt gleich. Die Konzentrationskurve verläuft dann zwischen den Punkten (i-1 , C_{i-1}) und (i+1,C_{i+1}) flacher (vgl. Abb. 6.2). Dadurch vergrößert sich die ursprüngliche Fläche F um die Fläche F_d. Mit geometrischen Überlegungen läßt sich zeigen, dass die Fläche F_d dem Merkmalsanteil c_d entspricht. Offensichtlich gilt dies auch bei einem Transfer von i auf j (bei j > i+1). Wegen der in Gl. 6.7a dargestellten inversen Beziehung zwischen F und K_R nimmt K_R ab. Analog nimmt K_R bei positivem Transfer zu.

7. Für $y_i = a+x_i$ erhält man die "neuen" Konzentrationsraten (von y statt von x)

$$(6.8) \qquad C_i,^* = \frac{a(i/n) + C_i \bar{x}}{a + \bar{x}} = \frac{a}{a + \bar{x}} \frac{i}{n} + \frac{\bar{x}}{a + \bar{x}} C_i$$

$C_i,^*$ ist also ein gewogenes Mittel aus dem bisherigen Wert C_i und dem Wert i/n bei egalitärer Verteilung. Ist die Ausgangsverteilung die egalitäre Verteilung, so bleibt diese erhalten. In allen anderen Fällen gilt:

- Für a > 0 ist $C_i^* < C_i$ für i < n, so dass die neue Konzentrationskurve unterhalb der bisherigen verläuft, was bedeutet, dass K_R abnimmt.
- Für a < 0 ist entsprechend $C_i^* > C_i$, wenn i < n ist, so dass K_R zunimmt (das Gewicht $a/(a+\bar{x})$ ist negativ).

Somit erfüllt K_R auch das Axiom K3.

8. Auch Axiom K4 ist, wie leicht zu sehen ist, erfüllt. Eine Dekonzentration mit k = 2 verändert die Konzentrationskurve wie in Abb. 6.3 gezeigt. Den Konzentrationsraten C_i sind jetzt die Abszissenwerte 2i zugeordnet, so dass die neue Fläche $F^*=2F$ ist. Daraus folgt für den neuen Rosenbluth-Index $K_{R^*} = (2F^*)^{-1} = ½K_R$. Bei einer Fusion von je zwei Einheiten mit gleichem Merkmalsanteil gilt $K_{R^*} = 2K_R$, was sich ebenfalls aus Abb. 6.3 ergibt, wenn man diese von rechts nach links liest.

9. Axiom K5 ist offensichtlich ebenfalls erfüllt, da dies nur darauf hinausläuft, dass das Rechteck über die Punkte (n,0) und (n,1) hinaus

verlängert wird (Abb. 6.4). Es ist klar, dass sich dadurch der Rosenbluth-Index nicht verändert.

Abb. 6.3: Wirkung einer Dekonzentration (Axiom K4)

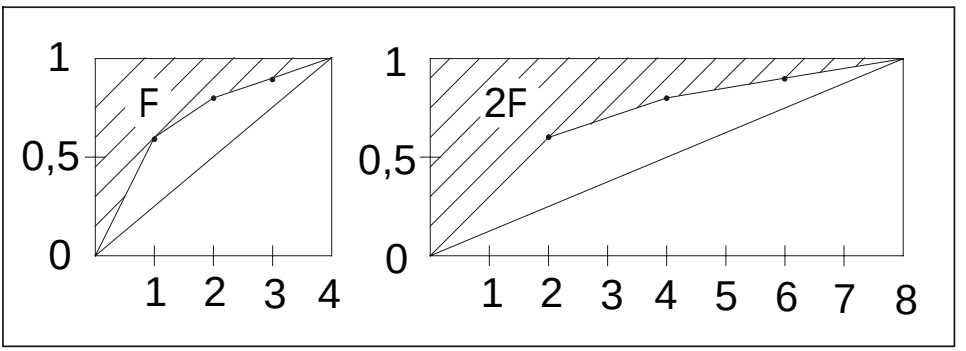
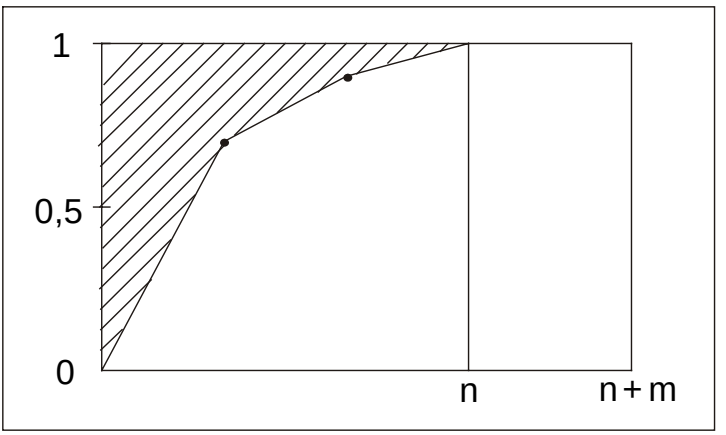


Abb. 6.4: Konzentrationskurve bei Nullergänzung von m Einheiten



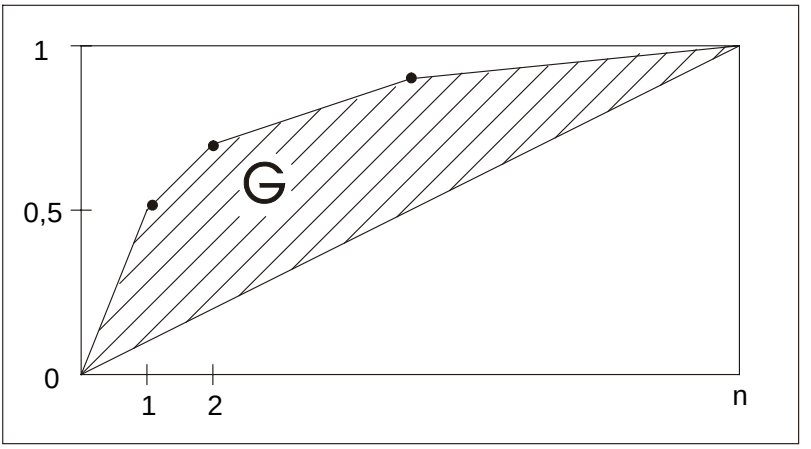
10. Die maximale Konzentration ist erreicht, wenn der größte Merkmalsträger die gesamte Merkmalssumme auf sich vereinigt. Man erhält dann für die Fläche F den Wert ½ und somit $K_R = 1$. Entsprechend gilt bei egalitärer Verteilung ($c_i=1/n$) $F = (1+2+...+n)/n - ½$ also $2F = n$ und somit $K_R = 1/n$.

11. Der Rosenbluth-Index hängt, wie gesagt, invers von der Größe der Fläche F ab. Dies legt den Gedanken nahe, ein Maß aus der Fläche G (vgl. Abb. 6.5) zwischen der Diagonalen und der Konzentrationskurve

$$G = n/2 - F \qquad 0 \leq G \leq (n-1)/2$$

zu konstruieren, analog zur Herleitung des Gini-Koeffizienten aus der Lorenzkurve (vgl. Abschn. 4).

Abb. 6.5: Fläche G als Konzentrationsmaß



Man kann aber leicht sehen, dass eine mit Gl. 2.3 auf das Intervall von 1/n bis 1 normierte Fläche G, also das Flächenmaß

$$G_n = \frac{1+2G}{n} \qquad \frac{1}{n} \le G_n \le 1$$

als Maß der absoluten Konzentration unbrauchbar ist, da es die Axiome K4 und K5 nicht erfüllt.

Beispiel 6.1:

An einem Markt bieten fünf Unternehmen an. Ihre Marktanteile sind:

1) $c_1 = 0{,}6$; $c_2 = 0{,}2$; $c_3 = 0{,}1$; $c_4 = 0{,}06$ und $c_5 = 0{,}04$. Man zeichne die Konzentrationskurve und berechne die Fläche F und K_R.

2) Der Marktanteil $c_d = 0{,}04$ wird von Unternehmen 2 auf Unternehmen 3 "übertragen".
Man zeichne wieder die Konzentrationskurve und berechne F und K_R.

3) Angenommen, jedes Unternehmen werde (ausgehend von Situation 1) in zwei gleich große Unternehmen "dekonzentriert". Man zeichne wieder die Konzentrationskurve und berechne F und K_R.

Lösung 6.1:

1) Ordinatenwerte der Konzentrationskurve: $C_1 = 0{,}6$; $C_2 = 0{,}8$; $C_3 = 0{,}9$; $C_4 = 0{,}96$ und $C_5 = 1$. Für die Fläche erhält man $F_1 = 1{,}74 - ½ = 1{,}24$ und für den Rosenbluth-Index $K_{R1} = 0{,}40323$.

2) Nach dem Transfer gilt $C_2 = 0{,}76$. Die anderen Werte bleiben unverändert. Ferner ist $F_2 = 1{,}28$ und $K_{R2} = 0{,}390625$, die Konzentration nimmt also ab (wie im Axiom K2 gefordert).

3) Es gilt jetzt: $C_1 = 0{,}3$; $C_2 = 0{,}6$; $C_3 = 0{,}7$;...;$C_8 = 0{,}96$, $C_9 = 0{,}98$ und $C_{10} = 1$. Schließlich ist $F_3 = 2{,}48 = 2F_1$ und $K_{R3} = 0{,}2016 = ½K_{R1}$.

b) Herfindahl-Index

Ein sehr einfaches und das am meisten verbreitete Maß der absoluten Konzentration ist der Herfindahl-Index, auf dessen Eigenschaften hier kurz eingegangen werden soll.

Def. 6.4: Herfindahl-Index

Die Summe der quadrierten Merkmalsanteile $q_1,q_2,...,q_n$

$$(6.9) \quad K_H = \Sigma\, q_i^2 = \mathbf{q'q} \qquad \frac{1}{n} \le K_H \le 1$$

heißt Herfindahl-Index oder Konzentrationsmaß von Herfindahl (Symbol: statt K_H auch einfach H). Man beachte, dass die Daten hier Einzelbeobachtungen sind. Für gruppierte und klassierte Daten ist K_H gem. Gl. 6.10 zu berechnen.

Interpretation und Eigenschaften des Herfindahl-Index:

1. Der Herfindahl-Index ist ein summarisches Maß der (absoluten) Konzentration. Man erkennt auch leicht, dass es auf die Reihenfolge der Merkmalsanteile (Anteilswerte) q_i nicht ankommt und dass man K_H auch als ein mit den Anteilen q_i gewogenes **arithmetisches** Mittel der Anteile q_i auffassen kann. Daraus folgt auch, dass das Axiom K1 erfüllt ist.

2. Der Herfindahl-Index ist auch in der Form
   $K_H = \Sigma(x_i / \Sigma x_i)^2 = \Sigma(x_i,^2 / n^2\, x,^{-2})$ darstellbar. Daraus folgt auch der bekannte Zusammenhang zwischen K_H und dem Variationskoeffizienten V

   $$(6.10) \qquad K_H = \frac{V^2 + 1}{n}.$$

   Der Zähler V^2+1 charakterisiert die Verteilungsungleichheit und der Nenner mißt den Anzahleffekt. Mithin steigt (sinkt) die absolute Konzentration, wenn sich bei

gegebener relativer Streuung die Anzahl n der Einheiten verringert (vergrößert). Es ist auch plausibel von steigender Konzentration im ökonomischen Sinne zu sprechen, wenn sich die Disparität erhöht und die Anzahl der Merkmalsträger kleiner wird. Wie man daran erkennt, hängt die Aussage der absoluten Konzentration nicht allein von der Anzahl n ab, sondern auch von der Unterschiedlichkeit der Größe der Einheiten:
So kann z.B. für n = 2, bei einem Duopol, K_H alle Werte zwischen ½ und 1 annehmen, je nachdem, welchen Marktanteil q (bzw. p = 1 - q) das erste (bzw. zweite) Unternehmen hat. Die quadratische Funktion $K_H = q^2 + p^2 = 1 - 2q(1-q)$ hat an der Stelle p = q = ½ ihr Minimum.

3. Fusionen von Einheiten vergrößern die absolute Konzentration (aber nicht notwendig auch die Disparität), weil sie die Anzahl der Einheiten verringern.

   Die konzentrationserhöhende Wirkung des Fusionierens der ersten m < n Einheiten ist im Falle des Herfindahl-Index leicht zu sehen, da sich K_H vergrößert wegen: $(\Sigma q_j)^2 > \Sigma q_j^2$ (j=1,2,...,m). Eine Fusion von jeweils k gleich großen Einheiten im Sinne der Proportionalitätsprobe (Axiom K4) bedeutet im Falle von k=2 einen Übergang vom Datenvektor 1: $[q_1\ q_1\ q_2\ q_2\ ...q_n\ q_n]$ zum Vektor 2: $[2q_1\ 2q_2\ ...\ 2q_n]$. Wie man leicht sieht, gilt $K_{H1} = 2\Sigma q_i^2$ und $K_{H2} = 4\Sigma q_i^2$, so dass sich K_H, wie im Axiom K4 gefordert, verdoppelt. Das Axiom ist für beliebiges k bei Fusionen und Dekonzentrationen erfüllt.

4. Der Herfindahl-Index erfüllt auch die übrigen Axiome. Ein negativer Transfer des Betrags d von i zu i+1 (mit den Merkmalsbeträgen x_i und $x_{i+1} = x_i$-h) im Sinne des Axioms K3 verringert K_H um den Betrag $2d(h-d)/(\Sigma x_i)^2$. Entsprechend vergrößert sich K_H bei einem gleich großen positiven Transfer. Es gilt also Axiom K2 (schwache Verschiebungsprobe), nicht aber die starke Verschiebungsprobe. Axiom K5 ist ganz offensichtlich erfüllt, da sich K_H durch Nullergänzung nicht verändert. Mit der im Axiom K3 geforderten Transformation $y_i = a + x_i$ verändert sich der quadrierte Variationskoeffizient wie folgt

   (6.11) $$V_Y^2 = \frac{\bar{x}^2\, V_X^2}{(a+\bar{x})^2} \quad \text{und damit } V_Y = \frac{\bar{x}\, V_X}{a+\bar{x}}\ ,$$

   so dass gilt $V_Y < V_X$, wenn a < 0 und $V_Y > V_X$, wenn a > 0, so dass der Herfindahl-Index auch das Axiom K3 erfüllt.

5. Offensichtlich gilt $1/n \le K_H \le 1$, wobei die Grenzen genau in den für Konzentrationsmaße beschriebenen Extremzuständen angenommen werden. In diesen Fällen ist $K_H = K_R = s$, wobei s die Steigung der Konzentrationskurve ist. Bei egalitärer Verteilung gilt

$$(6.12)\quad K_H = \sum \left(\frac{1}{n}\right)^2 = \frac{1}{n} \quad .$$

Hier lässt sich auch leicht der Anzahleffekt zeigen: Ein Übergang von n gleich großen Einheiten zu n-1 gleich großen Einheiten bedeutet (analog zu Gl. 6.13) eine Vergrößerung von $K_H = 1/n$ zu $K_H^* = 1/(n-1)$.

6. Weitere Hinweise:
   a) K_H nimmt in der Regel recht niedrige Werte an. Die amerikanischen Fusionsricht-linien stellten früher (1968) auf die concentration ratios bezogen auf die vier größten Unternehmen (also die Größe C_4) ab, neuerdings auf den Herfindahl-Index

	US-Fusionsrichtlinien	
Konzentrationsgrad	1968	1982
niedrig	$C_4 < 0{,}5$	$K_H < 0{,}1$
mittelhoch	$0{,}5 < C_4 < 0{,}7$	$0{,}1 < K_H < 0{,}18$
hoch	$C_4 > 0{,}7$	$K_H > 0{,}18$

Man kann zeigen, dass die angegebenen Wertebereiche sich in etwa entsprechen. Der Herfindahl-Index, aber auch die concentration ratios (oder "kritische Konzentrationskurven") werden auch zur Messung der Wettbewerbsintensität benutzt. So gilt z.B. nach §22 GWB als kritische Konzentration: $C_1 > 1/3$, $C_3 > 1/2$ und $C_5 > 2/3$.

   b) Das Konzept des Herfindahl-Index läßt sich verallgemeinern zu

$$(6.13)\quad K_\alpha = (\Sigma\, q_i^\alpha)^{1/(\alpha-1)} \quad (\alpha > 0)$$

Es gilt $K_2 = K_H$, Strebt α gegen 1 so geht K_α in den Exponentialindex über (vgl. Def. 6.5), wächst α über alle Grenzen ("gegen unendlich"), so strebt K_α gegen C_1, denn mit zunehmendem α wird den größeren Einheiten ein immer größeres Gewicht verliehen. Es gilt die Mittelwertungleichung $K_0 = 1/n \le K_1 = E \le K_2 = K_H \le K_\infty = C_1$

Beispiel 6.2:

Man berechne den Herfindahl-Index für Beispiel 6.1.

Lösung 6.2:

$K_{H1} = 0{,}6^2 + 0{,}2^2 + 0{,}1^2 + 0{,}06^2 + 0{,}04^2 = 0{,}4152;$
$K_{H2} = 0{,}6^2 + 0{,}16^2 + 0{,}14^2 + 0{,}06^2 + 0{,}04^2 = 0{,}4104$ und
$K_{H3} = 2(0{,}3)^2 + 2(0{,}1)^2 + 2(0{,}05)^2 + 2(0{,}03)^2 + 2(0{,}02)^2 = 0{,}2076 = K_{H1}/2.$

c) Exponentialindex und Entropie

Die folgenden beiden Konzentrationsmaße werden nur kurz eingeführt. Die Eigenschaften der Maße werden nicht im einzelnen diskutiert.

Def. 6.5: Exponentialindex

(6.14) $E = q_1^{q_1} \cdot q_2^{q_2} \ldots q_n^{q_n} = \Pi q_i^{q_i}$ mit $\frac{1}{n} < E < 1$ (i=1,2,...,n)

Der Exponentialindex E ist ein mit den Anteilen q_i gewogenes geometrisches Mittel der Anteile q_i. Daraus folgt $E \leq K_H$, da K_H ein entsprechendes arithmetisches Mittel ist. Offensichtlich ist $ld(E) = - K_E$ (K_E = Entropie, vgl. Def. 6.6).

Def. 6.6: Entropie

Die analog dem Streuungsmaß definierte Entropie lautet bei der Konzentrationsmessung:

(6.15) $K_E = \sum q_i \, ld\left(\frac{1}{q_i}\right) = - \sum q_i \, ld(q_i)$ (Entropie)

Häufig werden auch die von Theil vorgeschlagenen Maße der Redundanz herangezogen, weil K_E eher als Dekonzentrationmsmaß angesehen werden kann (K_E sinkt [steigt] mit zunehmender [abnehmender] absoluter Konzentration). Es gilt:

(6.16a) $K_T = ld(n) - K_E$ (absolute Redundanz)

(6.16b) $K_T^* = \frac{K_T}{ld(n)}$ (relative Redundanz)

Hierbei ist ld(x) der logarithmus dualis (Logarithmus zur Basis 2) der Größe x. Es gilt ldx = logx/log2 = 3,32193logx.

Bemerkungen zur Entropie (Def.6.6)

1. K_E (und entsprechend auch K_T und $K_T,^*$) hängen allein von den Größen q_i ab, sie erfüllen also Axiom K1. Bei vollständiger Konzentration ist also $K_E = -1 \cdot ld(1) = 0$ [statt 1] und bei egalitärer Verteilung gilt dem-nach $K_E = -\Sigma(1/n)ld(1/n) = ld(n)$ [statt 1/n], weshalb auch anstelle von K_E die Größe $K_T,^*$, eine Lineartransformation von K_E, als Konzentrationsmaß vorgeschlagen wird. Man kann K_E - wie gesagt - auch eher als Dekonzentrationmsmaß verstehen, wie die folgende Überlegung zeigt:

Ausgehend von der egalitären Verteilung steigt K_E wenn Egalität zwischen n+m statt n Einheiten besteht (n,m>0), denn ld(n+m) > ld(n). Aus diesem Grunde erfüllt K_E auch nicht Axiom K4. Es ist aber offensichtlich, dass K5 erfüllt ist weil sich K_E (anders als K_T und $K_{T,}^{*}$) nicht durch Nullergänzung verändert.

2. Bemerkenswerte Vorzüge der Entropie sind ihre einfache Verallgemeinerungsfähigkeit auf zwei und mehr Konzentrationsmerkmale und ihre günstigen Aggregationseigenschaften. Was letztere betrifft, so können diese leicht an dem folgendem Beispiel (der Zusammenhang kann natürlich verallgemeinert werden; vgl. auch Beispiel 6.3) demonstriert werden:

   Angenommen die Einheiten (Unternehmen) 1 und 2 fusionieren und ebenso die Einheiten 3 und 4, so erhält man ausgehend von der "alten" Entropie

$$\text{(6.17a)} \qquad K_E^{(1)} = q_1 ld\left(\frac{1}{q_1}\right) + q_2 ld\left(\frac{1}{q_2}\right) + q_3 ld\left(\frac{1}{q_3}\right) + q_4 ld\left(\frac{1}{q_4}\right)$$

   die "neue" Entropie

$$\text{(6.17b)} \qquad K_E^{(2)} = (q_1+q_2) ld\left(\frac{1}{q_1+q_2}\right) + (q_3+q_4) ld\left(\frac{1}{q_3+q_4}\right).$$

   Zwischen $K_E^{(1)}$ und $K_E^{(2)}$ besteht folgender Zusammenhang

$$\text{(6.17c)} \qquad K_E^{(1)} = K_E^{(2)} + K_E^{(3)} \text{ (Additionstheorem der Entropie),}$$

   wobei $K_E^{(3)}$ wie folgt definiert ist

$$\text{(6.17d)} \qquad K_E^{(3)} = (q_1+q_2)\left\{\frac{q_1}{q_1+q_2} ld\left[\frac{q_1+q_2}{q_1}\right] + \frac{q_2}{q_1+q_2} ld\left[\frac{q_1+q_2}{q_2}\right]\right\}$$
$$+ \ (q_3+q_4)\left\{\frac{q_3}{q_3+q_4} ld\left[\frac{q_3+q_4}{q_3}\right] + \frac{q_4}{q_3+q_4} ld\left[\frac{q_3+q_4}{q_4}\right]\right\}$$

   Die Ausdrücke innerhalb der äußeren Klammern stellen die Konzentrationen innerhalb der neuen durch Fusion (von jeweils zwei Einheiten) entstandenen Gesamtunternehmen dar. Somit kann $K_E^{(3)}$ als "interne" Konzentration bezeichnet werden.
   Sind die fusionierten Unternehmen gleich groß (wenn also z.B. gilt $q_1 = q_2$ und entsprechend $q_3 = q_4$ so sind die Ausdrücke innerhalb der inneren Klammern ld(2) = 1 und mithin $K_E^{(3)}$ = ld(2)=1 und entsprechend gilt bei der Fusion von jeweils k Un-

ternehmen, die gleich groß sind $K_E^{(3)} =$ ld (m), so dass die Entropie durch die Fusion von $K_E^{(1)}$ auf $K_E^{(2)}$ um $K_E^{(3)} =$ ld(k) sinkt. Das bedeutet, dass (wie bereits gesagt) die Entropie das Axiom K4 (Proportionalitätsprobe) nicht erfüllt: wird jede Einheit in k gleich große Einheiten aufgeteilt, so erhöht sich K_E um ld (k) - statt sich zu verringern - (vgl. auch Beispiel 6.3). Axiom K4 würde dagegen bei Dekonzentration $K^* = K/k$ statt $K_E^{(1)} = K_E^{(2)} + ld(k)$ und bei Fusion $K^* = kK$ statt $K_E^{(2)} = K_E^{(1)} - ld(k)$ verlangen.

Das Additionstheorem (Gl. 6.17c) ist danach (analog einer Varianzzerlegung) wie folgt zu lesen

(6.17c) Gesamtentropie = externe Entropie + interne Entropie

$$K_E^{(1)} = K_E^{(2)} + K_E^{(3)}$$

Beispiel 6.3

a) Drei Unternehmen teilen sich einen Markt. Ihre Marktanteile sind $q_1 = 1/2$, $q_2 = 1/3$, $q_3 = 1/6$. Man berechne die Entropie.

b) Berechnen Sie die Entropie, wenn sich die drei Unternehmen des Teils a in jeweils
- zwei
- drei

gleich große Unternehmen aufspalten.

c) Man berechne die Entropie sowie die absolute und relative Redundanz für die folgenden Konzentrationszustände (Vektoren der Marktanteile) $Q_A = [0{,}3\ 0{,}7]$ und $Q_B = [0{,}15\ 0{,}15\ 0{,}35\ 0{,}35]$.

Lösung 6.3:

a) $K_E = 1{,}45915$

b) bei zwei: $K_E = 1{,}45915 + ld(2) = 1{,}45915 + 1 = 2{,}45915$
bei drei: $K_E = 1{,}45915 + ld(3) = 1{,}45915 + 1{,}58496 = 3{,}04411$.
(für die relative Redundanz erhält man aber: $K_T^* = K_T/ld4 = K_T/2 = 0{,}11871/2 = 0{,}05936$ [wie es Axiom K4 fordert]).

c) bei Q_A: $K_E = 0{,}88129$, $K_T = 1 - 0{,}88129 = 0{,}11871 = K_T^*$
bei Q_B $K_E = 1{,}88129$, $K_T = 2 - 1{,}88129 = 0{,}11871$
(also unverändert, wie es eigentlich von einem Disparitätsmaß, nicht aber von einem Konzentrationsmaß, gefordert wird.)

4. Messung der relativen Konzentration (Disparität)

a) Lorenzkurve und Gini-Koeffizient

aa) Berechnung bei Einzelbeobachtungen

Bei der Berechnung des Gini-Koeffizienten ist es nützlich zu unterscheiden, ob der Berechnung Einzelbeobachtungen oder eine klassierte Verteilung zugrunde liegen. Es soll deshalb zunächst der Fall geordneter Einzelbeobachtungen betrachtet werden.

1. Definition bei Einzelbeobachtungen

Wie bei der Konzentrationsmessung gibt es auch bei der Disparitätsmessung eine graphische Darstellung und eine summarische Maßzahl (vgl. Übers. 6.1). Bei der Konzentrationsmessung wird grundsätzlich von Einzelbeobachtungen ausgegangen, während für die Disparitätsmessung auch die klassierte Verteilung und als theoretisches Modell die stetige Verteilung (vgl. Abschn. 6b dieses Kapitels) von Interesse ist.

Def. 6.7: Lorenzkurve und Gini-Koeffizient bei Einzelbeobachtungen

a) Lorenzkurve

Wird der Merkmalsanteil des i-ten Merkmalsträgers bei einer Ordnung nach **zunehmender** Größe

$$(6.18)\quad q_i = \frac{x_{(i)}}{\Sigma x_{(j)}} = \frac{x_{(i)}}{\Sigma x_j} \qquad (i,j = 1,2,...,n)$$

genannt (vgl. Def. 6.3), dann ist

$$(6.19)\quad Q_i = \Sigma\, q_j \qquad (j = 1,2,...,i)$$

der kumulierte Anteil der i kleinsten Merkmalsträger am Merkmalsbetrag. Die lineare Verbindung der Punkte $P_i(H_i,Q_i)$ mit den kumulierten relativen Häufigkeiten H_i und den kumulierten Anteilen Q_i im H-Q-Koordinatensystem heißt Lorenzkurve ($H_0 = Q_0 = 0$ und $H_n = Q_n = 1$). Für die H_i gilt im Fall von Einzelbeobachtungen:

$H_i = \frac{i}{n}$.

b) Gini-Koeffizient

Die Größe

$$(6.20)\quad D_G = \sum_{i=1}^{n} \frac{2i-n-1}{n} q_i \qquad (0 \le D_G \le 1 - \frac{1}{n})$$

heißt Disparitätskoeffizient von Gini (oder einfach Gini-Koeffizient).

2. Einführendes Beispiel für die Lorenzkurve

Beispiel 6.4:

In einer islamischen Familie mit 4 Kindern sind 2 männlich und 2 weiblich. Das an die Kinder zu vererbende Vermögen von 1200 Dinare soll getreu nach den Regeln des Korans vermacht werden:

> *"Und wenn die Geschwister Männer und Frauen sind, so soll ein Mann so viel erhalten wie zwei Frauen" (Sure 4, Vers 175)*

Man bestimme die sich bei Befolgung des islamischen Erbrechts ergebende Lorenzkurve und den Gini-Koeffizienten!

Lösung 6.4:

Die beiden Töchter bekommen jeweils 200 (zusammen 400) und die beiden Söhne jeweils 400 (zusammen also 800) Dinare.
Die Anteile sind $h_1 = h_2 = h_3 = h_4 = 1/4$ (wobei 1 und 2 die beiden Töchter und 3 und 4 die beiden Söhne symbolisieren) und $q_1=q_2=1/6$ (so dass $Q_2 = q_1 + q_2 = 1/3$) ferner ist $q_3 = q_4 = 1/3$ so dass $q_3 + q_4 = 2/3$. Man sieht leicht, dass $D_G = 1/6$ ist.

3. Bemerkungen zur Lorenzkurve

1. Die Reihenfolge der Messwerte (nach zunehmenden Anteilen q_i) ist wesentlich.

2. Die Größen H_i und Q_i sind nicht unabhängig voneinander. Vielmehr gilt wegen $Q_i = \Sigma\, q_j$ (mit $j = 1,2,...,i$) und $q_j = x_{(j)} / n\overline{x} = h_j(x_{(j)} / \overline{x})$, so dass zwischen den Anteilen q_j und h_j folgender Zusammenhang besteht:

$$(6.21)\quad \frac{q_j}{h_j} = \frac{x_{(j)}}{\overline{x}}$$

Der Bruch q_j / h_j ist die Steigung der Lorenzkurve, die somit proportional ist zu $x_{(j)}$ (denn $1/\overline{x}$ ist eine Konstante): vgl. Bem. 6.

3. Die Lorenzkurve stellt den funktionalen Zusammenhang zwischen den kumulierten Anteilen am Merkmalsbetrag Q_i und den kumulierten relativen Häufigkeiten H_i dar. Es gilt

   $$Q_i = L(H_i).$$

   Die Funktion L heißt Lorenzkurve.

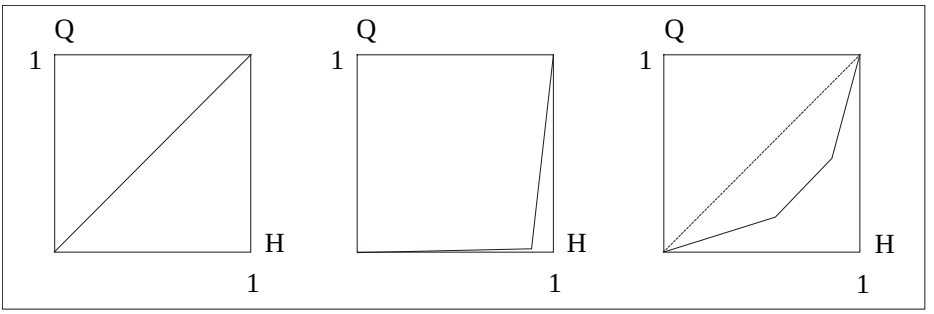
4. Über die Extremfälle der Disparität läßt sich hinsichtlich der Gestalt von L folgendes aussagen (Abb. 6.6).

   a) Bei egalitärer Verteilung ist die Funktion L eine Gerade (die 45-Grad-Linie). Es gilt wegen $q_i = 1/n$ für alle i=1,2,...,n die Gleichung $Q_i=H_i$. Die Lorenzkurve ist dann mit der Geraden zwischen den Punkten $P_0(0,0)$ und $P_n(1,1)$, der Gleichverteilungsgeraden, identisch (Abb. 6.6, links).

   b) Bei vollkommener Ungleichheit ist die Lorenzkurve aus der Strecke von $P_0(0,0)$ bis $P_{n-1}(n-1/n,0)$ und dem Punkt $P_n(1,1)$ zusammengesetzt (Abb. 6.6, Mitte).

   c) In allen realistischen Fällen liegt die Lorenzkurve zwischen diesen beiden Extremfällen, als konvexer Polygonzug unterhalb der Gleichverteilungsgeraden (Abb. 6.6, rechts).

Abb. 6.6: Lorenzkurve und extreme Fälle von Disparität



5. Man kann zeigen, dass die Fläche zwischen der Lorenzkurve und der Gleichverteilungsgeraden, die auch Konzentrationsfläche[2] genannt wird

   $$F = \sum \frac{2i-n-1}{2n} q_i$$

[2] Der Begriff ist aber auch üblich für die Fläche oberhalb der Konzentrationskurve (vgl. Gl. 6.7).

beträgt.
Daraus folgt, dass der Gini-Koeffizient das Verhältnis zwischen F und der Dreiecksfläche unterhalb der Gleichverteilungsgeraden ist (diese Dreiecksfläche beträgt ½), d.h. es gilt

$$(6.22) \qquad D_G = \frac{F}{1/2} = 2F \quad .$$

Bei egalitärer Verteilung ist F=0 und somit auch D_G=0; Bei vollkommener Ungleichheit ist F = ½ - ½(1/n) und somit D_G=(n-1)/n. Es gilt für D_G also die im Axiom K6 geforderte Einschrankung.

6. Die Steigung der Lorenzkurve zwischen den Punkten $P_{i-1}(H_{i-1}, Q_{i-1})$ und $P_i(H_i,Q_i)$ beträgt wegen Gl. 6.21

$$(6.23) \qquad \frac{q_i}{h_i} = \frac{x_{(i)} / n\bar{x}}{1/n} = \frac{x_{(i)}}{\bar{x}} \qquad \text{(Steigung der Lorenzkurve)}$$

Die Steigung der Lorenzkurve ist also nichtnegativ und von 0 bis unendlich monoton steigend, d.h. sie nimmt mit zunehmendem x und H zu.

Daraus folgt unmittelbar:

- Die Steigung erreicht den Wert 1, d.h. die Lorenzkurve verläuft parallel zur Gleichverteilungsgeraden, wenn $x_{(i)} = x,^-$, d.h. wenn das Einkommen (der Merkmalsbetrag) der i-ten Einheit gleich dem Durchschnittseinkommen ist.
- Die Lorenzkurve verläuft also zunächst flacher (bei Beziehern unterdurchschnittlicher Einkommen), dann ab $x_{(i)} = x,^-$ steiler als die Gleichverteilungsgerade (bei Beziehern überdurchschnittlicher Einkommen).
- Weil die Steigung der Lorenzkurve nicht abnehmen kann, kann die Lorenzkurve die Gleichverteilungsgerade nicht schneiden; das folgt auch aus $H_i \geq Q_i$,($H_i = Q_i$ gilt außer bei egalitärer Verteilung nur für i=0 und i=n), denn

$$(6.24) \quad H_i = \frac{i}{n} \leq \frac{\sum_{j=1}^{i} x_j}{n\bar{x}} = \frac{i}{n}\frac{\bar{x}_i}{\bar{x}} = Q_i$$

wobei $\bar{x}_i$ der mittlere Merkmalsbetrag der ersten i Merkmalsträger ist (nicht zu verwechseln mit $x_{(i)}$, dem Merkmalsbetrag der i-ten

Einheit), der wegen der Reihenfolge der Merkmalsträger notwendig stets kleiner ist als der Mittelwert $\bar{x}$, der sich auf alle n Merkmalsträger bezieht.

Da $\bar{x}_i < x_{(i)}$, ist die Steigung der Strecke P_0P_i mit $Q_i / H_i = \bar{x}_i / \bar{x}$ kleiner als die Steigung der Tangente im Punkt P_i der Lorenzkurve, die $q_i / h_i = x_{(i)} / \bar{x}$ beträgt.

7. Zwischen dem Gini-Koeffizienten D_G und dem Rosenbluth-Index K_R besteht folgender Zusammenhang

$$(6.25) \qquad D_G = 1 - \frac{1}{nK_R} \text{ und somit } K_R = \frac{1}{n(1-D_G)}$$

Dabei ist zu beachten, dass die Reihung der Merkmalsträger bei der Konzentrationsmessung nach abnehmender und bei der Disparitätsmessung nach zunehmender Größe erfolgt.

8. Eine Lineartransformation der Merkmalswerte mit $y_i=a+bx_i$ wirkt auf die Merkmalsanteile wie folgt ($q_i,^*$ ist ein gewogenes Mittel aus q_i und dem Wert i/n der Winkelhalbierenden [egalitäre Verteilung]),

$$q_i,^* = \left(\frac{a}{\bar{y}}\right) \cdot \frac{i}{n} + \left(\frac{b\bar{x}}{\bar{y}}\right) \cdot q_i$$

in Analogie zu Gl. 6.8, d.h. bei a,b >0 gilt $q_i \le q_i^* \le i/n$ (Mittelwerteigenschaft), so dass man für das Disparitätsmaß von Gini erhält

$$(6.26) \qquad D_G^* = \left(\frac{b\bar{x}}{\bar{y}}\right) D_G \, .$$

Bei proportionaler Transformation (Axiom K1) gilt a = 0 und wegen $\bar{y} = b\bar{x}$ auch $D_G^* = D_G$. Bei einer Niveauänderung im Sinne des Axioms K3 gilt b = 1 und folglich

$$(6.26a) \qquad D_G^* = \frac{\bar{x}}{a+\bar{x}} D_G \, .$$

so dass offensichtlich die Axiome K1 und K3 erfüllt sind. Bei a > 0 und b=1 rückt die Lorenzkurve in allen Punkten im Intervall $0 < H < 1$ näher an die Gleichverteilungsgerade heran.

Bei der Verschiebungsprobe (Axiom K2) gilt bei einem Transfer von i zu j mit $x_{(j)} < x_{(i)}$ für die kumulierten Anteile

$Q_j^* = Q_j + q_d$ und $Q_i^* = Q_i$

mit $q_d = d/n\bar{x}$, so dass die Lorenzkurve im Intervall $H_{j-1} < H < H_i$ näher an die Gleichverteilungsgerade heranrückt, wie es Axiom K2 bei einem negativen (egalisierenden) Transfer fordert.

10. Es ist leicht zu sehen, dass die Proportionalitätsprobe (Axiom K4) und die Ergänzungsprobe (Axiom K5) ebenfalls erfüllt sind. Die Zusammenhänge seien am Beispiel 6.5 demonstriert, gelten aber allgemein.

11. Ginis Disparitätsmaß steht auch im Zusammenhang mit Parametern der Pareto- Verteilung und der logarithmischen Normalverteilung, worauf jedoch aus Platzgründen nicht eingegangen werden kann.

12. Man kann D_G auch als einen speziellen Variationskoeffizienten auffassen. Zwischen D_G und Ginis "mittlere Differenz" S_{G*}

$$S_{G*} = \frac{1}{n^2} \sum_i \sum_k | x_i - x_k | \qquad i,k = 1,2,...,n ,$$

einem Steuungsmaß (vgl. Def. 5.5), besteht der folgende Zusammenhang:

$$(6.27) \qquad D_G = \frac{S_{G*}}{2\bar{x}} .$$

Daraus folgt übrigens auch, dass man D_G darstellen kann als

$$(6.27a) \qquad D_G = \frac{\Sigma\Sigma|q_i - q_k|}{2n} \qquad i,k = 1,...n.$$

13. Analog zu Gl. 6.27 gewinnt man auch ein Dispersionsmaß mit der durchschnittlichen Abweichung um $\bar{x}$ (d_x^* gem. Def. 5.3) als

$$(6.27b) \qquad \phi^* = \frac{d_x^*}{2\bar{x}}$$

das als Schutz-Koeffizient oder maximaler Nivellierungssatz (oder längste Lorenzkurvensehne) bekannt ist (vgl. Def. 6.13) und nach Gl. 6.44 nicht größer sein kann als der Gini-Koeffizient D_G.

14. Es ist offensichtlich, dass (unendlich viele) verschiedene Lorenzkurven trotz unterschiedlichen Verlaufs zum gleichen Flächenverhältnis D_G führen können, so dass das Maß D_G also einen Disparitätszustand nicht eindeutig abbildet.

15. Es ist bemerkenswert, dass Gini sein Disparitätsmaß D_G ursprünglich ohne Bezugnahme auf die Lorenzkurve entwickelte. Erst später erkannte er die Flächeninterpretation von D_G (Vgl. Bem. 5):

Die relativen Abstände

$$R_i = \frac{H_i - Q_i}{H_i}$$

zwischen der Lorenzkurve und Gleichverteilungsgeraden galten für Gini als Ausdruck der Disparität. Man kann zeigen, dass die Größe L^*, ein mit den kumulierten relativen Häufigkeiten H_i gewogenes arithmetisches Mittel der Größen R_i

$$L^* = \sum_{i=1}^{n} R_i \frac{H_i}{SH_i} = \frac{\Sigma(H_i - Q_i)}{SH_i} \quad (i = 1,2,...,n)$$

gegen den Disparitätskoeffizienten D_G strebt, d.h. dass gilt

$$D_G = \lim_{n \to \infty} L^* \qquad (i=1,2,...,n).$$

Aus diesem Grunde ist auch die Größe $L^* = \Sigma(H_i-Q_i)/\Sigma H_i$ bei hinreichend großer Anzahl n von Merkmalsträgern (oder von Klassen) näherungsweise der Gini-Koeffizient.

Beispiel 6.5:

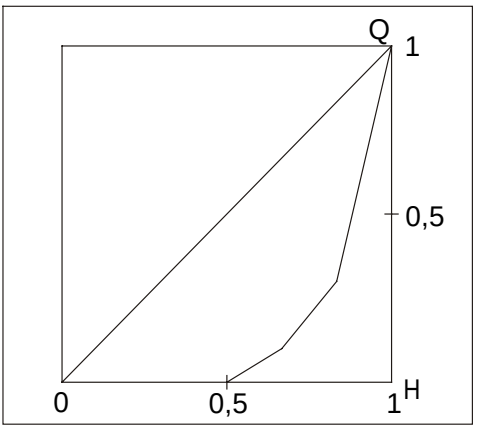
Ausgangspunkt sei eine Verteilung mit $q_1 < q_2 < q_3$ und $h_i = 1/3$ (i=1,2,3). Die *Aufspaltung jeder Einheit* in zwei gleich große Einheiten bewirkt, dass die neue Lorenzkurve die folgenden Punkte verbindet:

$P_1(0,0)$, $P_2(1/6,½q_1)$, $P_3(1/3,q_1)$, $P_4(½,q_1+½q_2)$, $P_5(2/3,q_1+q_2)$, $P_6(5/6,q_1+q_2+½q_3)$ und $P_7(1,1)$.

Die Punkte P_0, P_3, P_5 und P_7 sind identisch mit der bisherigen Lorenzkurve. Die Punkte P_2, P_4 und P_6 liegen jeweils auf der linearen Verbindung zwischen P_0 und P_3 usw. Die Lorenzkurve (und damit auch D_G) bleibt also unverändert.

Eine *Nullergänzung* durch 3 Nullträger auf der Basis der obigen Ausgangssituation bewirkt dagegen, dass die neue Lorenzkurve zwischen $0 \le H \le ½$ den Wert Q=0 annimmt (vgl. Abb. 6.7).

Abb. 6.7: Lorenzkurve bei Nullergänzung von n=3 Einheiten



Das Gleichheitsmaß $1\text{-}D_G = G$ betrug vor der Nullergänzung

$G_1 = 1/3\ (5q_1 + 3q_2 + q_3)$ und nach der Nullergänzung
$G_2 = 1/6\ (5q_1 + 3q_2 + q_3)$.

Es hat sich also halbiert wie es Axiom 5a bei k=2 also einer Ergänzung um $m = n(k-1) = 3(2-1) = 3$ Nullträgern verlangt.

bb) Berechnung bei gruppierten und klassierten Daten

Gegeben seien m Ausprägungen des Konzentrationsmerkmals bzw. m Klassen, so dass mit h_i die relativen Häufigkeiten und q_i die Anteile am Gesamtmerkmalsbetrag definiert sind (i = 1,...,m). Die kumulierten Anteile sind wieder $H_i = \Sigma h_j$ und $Q_i = \Sigma q_j$ (j=1,...,i).
Nach der Definition des Gini-Koeffizienten folgen zunächst einige weitere Bemerkungen zu den Eigenschaften der Lorenzkurve und dann zwei einführende Beispiele (Bsp. 6.6 und 6.7).

Def. 6.8: Lorenzkurve und Gini-Koeffizient bei klassierten Daten

a) Die lineare Verbindung der Punkte $P_i(H_i,Q_i)$ (i=0,1,...,m) mit $P_o(0,0)$ und $P_m(1,1)$ heißt Lorenzkurve.

b) Der Gini-Koeffizient D_G ist gegeben mit

(6.28) $D_G = 1 - \Sigma h_i(Q_i+Q_{i-1})$ oder (6.28a) $D_G = \Sigma q_i(H_i+H_{i-1}) - 1$

Diese Gleichungen gelten zunächst nur bei gruppierten Daten. Für die Berücksichtigung der Disparität innerhalb der Klassen bei klassierten Daten vgl. Bem. 4.

Bemerkungen zu Def. 6.8:

1. Die Bemerkungen im Abschnitt aa) gelten analog auch hier.

2. In Gl. 6.28 und 6.28a ist i=1 der Wert $H_{i-1} = H_{1-1} = H_0 = 0$ und entsprechend $Q_o = 0$ einzusetzen. Setzt man in Gl.6.28a für H_i den Wert i/n und entsprechend für H_{i-1} den Wert (i-1)/n ein, wie dies dem Fall der Einzelbeobachtungen entspricht, so folgt Gl. 6.20 aus Gl. 6.28a.

3. Wenn zwei und mehr aneinander angrenzende Größenklassen zusammengefaßt werden, etwa die Klassen i und i+1, so verläuft die "neue" Lorenzkurve flacher (näher an der Gleichverteilungsgeraden). Während sie vorher die Punkte P_{i-1}, P_i und P_{i+1} linear verband, stellt die Lorenzkurve jetzt eine Verbindung der Punkte P_{i-1} und P_{i+1} dar (vgl. Beispiel 6.7). Man kann leicht zeigen, dass sich die Fläche zwischen Lorenzkurve und Gleichverteilungsgerade dadurch verringert, und zwar um den Betrag $\frac{1}{2}(h_i q_{i+1} - h_{i+1} q_i)$, so dass gilt:

   D_G verringert sich um $(h_i q_{i+1} - h_{i+1} q_i)$ wenn die Größenklassen i und i+1 zusammengefaßt werden.

   Der Ausdruck $h_i q_{i+1} - h_{i+1} q_i$ ist übrigens deshalb nichtnegativ, weil die Steigung der Lorenzkurve monoton zunimmt, d.h. weil gilt $q_{i+1} / h_{i+1} \geq q_i / h_i$.

4. Für Ginis Dispersionsmaß existiert eine Zerlegung nach Art der Streuungszerlegung

   $$(6.29)\quad D_G = \left[1 - \sum h_i (Q_i + Q_{i-1})\right] + \left[\sum h_i^2 \frac{\overline{X}_i}{\overline{X}} D_G^{(i)}\right]$$

   wobei $D_G^{(i)}$ die Disparität innerhalb der i-ten Klasse darstellt, der gesamte Ausdruck in der zweiten eckigen Klammer also die interne Disparität ist (und die erste eckige Klammer die bisher allein betrachtete externe Disparität nach Gini).

5. Wegen des unter Nr. 3 dargestellten Zusammenhangs wird bei Aggregation (Zusammenlegung) von Größenklassen die Disparität abnehmen und bei Disaggregation (Aufspaltung) entsprechend zunehmen. Jede empirische, als Polgonzug dargestellte Lorenzkurve kann nur eine Näherung für die im Abschn. 6 definierten, und den Daten "eigentlich" zugrundeliegenden stetige Lorenzkurve sein. Ein Vergleich von Disparitäten (z.B. der Einkommensverteilungen von zwei Ländern) sollte deshalb nur bei gleich vielen Größenklassen erfolgen.

6. Aus Gl. 6.28 entwickelt sich leicht eine Formel für den Spezialfall, dass die Verteilung (z.B. die Einkommensverteilung) nur aus zwei Merkmalsausprägungen $x_2 > x_1$ besteht:

Einkommen	relative Häufigkeit	Anteil am Gesamtmerkmalsbetrag	Merkmalsbetrag
gering	h	q	x_1
groß	1-h	1-q	x_2

Die Lorenzkurve hat dann nur drei Punkte $P_0(0,0)$, $P_1(h,q)$ und $P_2(1,1)$ und Ginis Dispersionsmaß ist dann der senkrechte Abstand zwischen dem Punkt $P_1(h,q)$ und der Gleichverteilungsgeraden, also die Strecke von $P_1(h,q)$ bis $P_1^*(h,h)$.

(6.28b) $D_G = h - q$.

Diese Formel zeigt erneut, dass unterschiedliche Disparitätszustände zu gleichen Werten des Gini-Koeffizienten D_G führen können (eine häufig vorgebrachte Kritik an D_G): Haben z.B. in einem Land die 30% ärmsten Familien einen Anteil am Vermögen von nur 5% so ist die Vermögenskonzentration (genauer: Vermögensdisparität) genau so groß wie in einem anderen Land, in dem die 40% ärmsten Familien einen Anteil von 15% haben, denn $D_G = 0{,}3\text{-}0{,}05 = 0{,}4\text{-}0{,}15 = 0{,}25$.

Weiter kann man zeigen, dass in dieser Situation für den Variationskoeffizienten gilt

(6.28c) $V = \frac{D_G}{\sqrt{h(1-h)}} \leq 2D_G$.

Für die durchschnittliche Abweichung um $\bar{x}$ erhält man schließlich

$d_x^* = 2(x_2 - \bar{x})(1 - h)$,

so dass der Schutz-Koeffizient ϕ^* gem. Gl. 6.27b dann

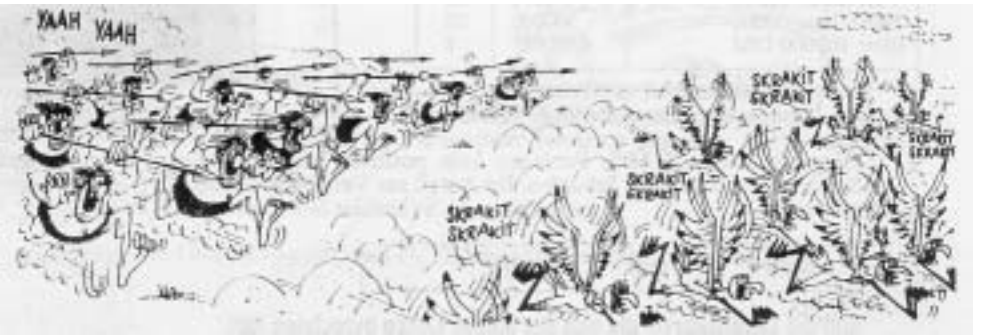
(6.28d) $\frac{d_x,^*}{2x,^-} = \phi^* = \frac{2x_2(1-h)}{2\bar{x}} - \frac{2\bar{x}(1-h)}{2\,\bar{x}} = h\text{-}q = D_G$

ist (da $x_2 / \bar{x} = 1\text{-}q$).

Beispiel 6.6:

Urmenschenproblem: 150 Angehörige eines primitiven Volksstammes (150 "Urmenschen") gehen auf die Jagd nach Federvieh. Ihre Beute beträgt 300 Wildgänse. Durch das an sich nur bei primitiven Völkern bekannte Gerangel um Geld, Gut und Prestige entstand eine etwas ungleiche Verteilung der Beute. Durch Eingreifen des Häuptlings konnte jedoch noch verhindert werden, dass jemand leer ausging. Es bekamen jeweils n_i Personen x_i Gänse:

x_i	1	2	3	4
n_i	60	45	30	15



Man zeichne die Lorenzkurve und berechne das Disparitätsmaß D_G von Gini. Wie sähe die Lorenzkurve aus, wenn jeder von der Beute gleichviel bekommen hätte?

Lösung 6.6:

Es sind 300 Gänse auf 150 "Urmenschen" zu verteilen; bei Gleichverteilung gilt also $\bar{x}$ = 300/150 = 2 Gänse für jeden. Es gibt also "Unterprivilegierte" mit nur x = 1 Gans und "Überprivilegierte" mit x = 3 oder gar x = 4 Gänsen. Für die Anteile h_i an den $n = \Sigma n_i = 150$ Urmenschen und die Anteile q_i an den $\Sigma\, x_i\, n_i = 300$ Gänsen gilt dann:

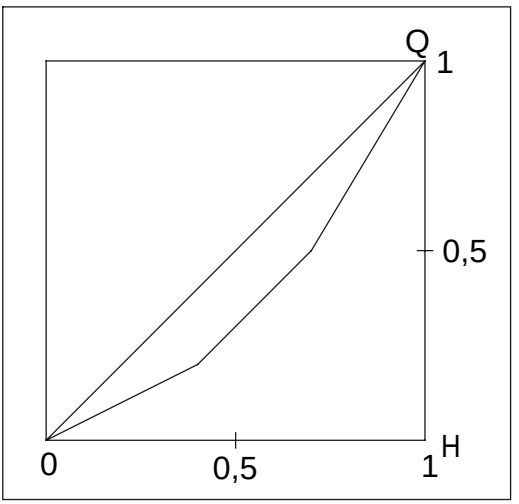
		x_i	1	2	3	4	Σ
A	Urmenschen	n_i	60	45	30	15	150
B	Gänse	$x_i n_i$	60	90	90	60	300
Anteile an A	$h_i = n_i / \Sigma\, n_i$		0,4	0,3	0,2	0,1	1,0
	$H_i = \Sigma\, h_j\ (j \le i)$		0.4	0,7	0,9	1	
Anteile an B	$q_i = x_i n_i / \Sigma\, x_i n_i$		0,2	0,3	0,3	0,2	1,0
	$Q_i = \Sigma\, q_j\ (j \le i)$		0,2	0,5	0,8	1	

Anteile an A = Anteile an den Urmenschen (an den Merkmalsträgern)
Anteile an B = Anteile an den Wildgänsen (an den Merkmalssumme).

Wie man sieht gilt $Q_i \le H_i$, was ja gerade die "Ungleichverteilung" ausmacht, denn die H_1 = 40% ärmsten Urmenschen haben nicht einen Anteil von 40%, sondern von weniger, nämlich nur 20% (denn Q_1 = 0,2) an der Beute. Entsprechend haben die 70% ärmsten nicht einen Anteil von 70%,

sondern nur von 50% usw. Die Lorenzkurve dieses Beispiels ist in der Abb.6.8 dargestellt.
Für Ginis Koeffizient erhält man in diesem Beispiel $D_G = 0,27$.

Abb.6.8: Lorenzkurve für das Beispiel 6.6



Beispiel 6.7:

Aus der Vermögenssteuerstatistik 1983 (Statistisches Jahrbuch der Bundesrepublik Deutschland 1989, S.451) erhält man folgende (stark zusammengefaßte) Daten zur Schichtung des Gesamtvermögens von unbeschränkt vermögenssteuerpflichtigen natürlichen Personen:

Gesamtvermögen	h_i	q_i	H_i	Q_i
unter 200000	0,2417	0,0481	0,2417	0,0481
200 - 500000	0,4458	0,2011	0,6875	0,2492
0,5 - 1 Mill	0,1898	0,1815	0,8773	0,4307
1 - 5 Mill	0,1088	0,2813	0,9860	0,7200
5 - 20 Mill	0,0119	0,1460	0,9980	0,8580
über 20 Mill	0,0020	0,1420	1,0	1,0

a) Zeichnen Sie die Lorenzkurve und berechnen Sie den Gini- und den Herfindahl-Koeffizienten D_G und K_H.

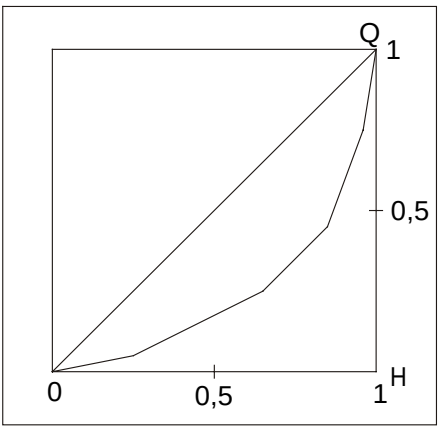
b) Wie ändert sich die Lorenzkurve und wie ändern sich und D_G und K_H wenn man die ersten beiden Größenklassen zusammenfaßt?

Lösung 6.7:

a) $D_G = 0,5512$ und $K_H = 0,3998$.

b) Die Lorenzkurve hat statt 7 Punkte (einschließlich $P_0(0,0)$ und $P_7(1,1)$) jetzt nur noch 6 Punkte, da die Verteilung nun nur noch 5 Größenklassen hat. Die bisherigen Punkte P_0 und P_2 werden linear verbunden. Gini- und Herfindahl-Index nehmen entsprechend ab: es gilt jetzt $D_G = 0{,}5241$ und $K_H = 0{,}4191$.

Abb. 6.9: Lorenzkurve für Beispiel 6.7



b) Der Variationskoeffizient als Disparitätsmaß

Der bereits als Maß der relativen Streuung eingeführte Variationskoeffizient $V = s/\bar{x}$ gilt auch als Maß der Disparität. Neben dem Variationskoeffzienten V wird auch das normierte Quadrat des Variationskoeffzienten als Disparitätsmaß benutzt.

Def. 6.9: normiertes Quadrat des Variationskoeffizienten

$$(6.30)\quad NV = \frac{V^2}{1+V^2} = 1 - \frac{1}{nK_H}$$

Interpretation und Eigenschaften

1. Für V gilt die folgende Beziehung zum Konzentrationsmaß von Herfindahl :

   $$(6.31)\quad V^2 = n\Sigma q_i^2 - 1 = nK_H - 1$$

   bei Einzelbeobachtungen und

   $$(6.31a)\quad V_2 + 1 = K_N / \bar{x}$$

bei gruppierten Daten, wobei K_N das arithmetische Mittel der Momentverteilung (vgl. auch Abschn. 6a, Gl. 6.35) ist. Mit $h_i = 1/n$ und $q_i = x_i / n\bar{x}$ erhält man Gl. 6.31 aus Gl. 6.31a.

Für V gilt bei den Extremsituationen der Disparität

- egalitäre Verteilung $V^2 = n\Sigma(1/n)^2 - 1 = 0$ und $NV = 0$
- vollkommene Ungleichheit $V^2 = n - 1$ und $NV = 1 - 1/n$.

NV ist also nur der auf den Wertebereich von 0 bis (n-1)/n eines Disparitätsmaßes normierter Variationskoeffizient und hat ansonsten die gleichen Eigenschaften wie V.

2. Man kann V nach Gl. 6.31, bzw. NV allein in Abhängigkeit von n und den Merkmalsanteilen q_i darstellen, so dass Axiom K1 erfüllt ist. Beide Maße erfüllen auch K2.

3. Offensichtlich ist auch das Axiom K3 erfüllt, denn bei der Lineartransformation $y_i = a + x_i$ gilt

$$(6.32)\quad V_y = \frac{\bar{x}}{a+\bar{x}} V_x \ ,$$

so dass $V_y < V_x$ wenn $a > 0$ und $V_y > V_x$ wenn $a < 0$.

4. NV erfüllt auch K4 und K5. Die Disparität, gemessen an V (und NV) steigt bei Nullergänzung (Hinzukommen von Nullträgern) aber nicht in dem in Axiom K5a geforderten Ausmaß.

c) Disparität und verwandte Konzepte

1. Disparität und relative Streuung

Die Verwendung des Variationskoeffizienten V als Disparitätsmaß oder der Größe $½S_{G*} / \bar{x}$ (vgl. Gl. 6.27) bzw. $d_x^* / 2\bar{x}$ (vgl. Gl. 6.27b) als Disparitätsmaß wirft die Frage nach dem Unterschied der beiden Konzepte "Disparität" und "relative Streuung" auf.
Es gibt Überschneidungen zwischen beiden Maßzahlenklassen in dem Sinne, dass Maßzahlen möglich sind, die ganz oder zum größten Teil die Axiome beider Klassen erfüllen (abgesehen von dem Wertebereich, also Axiom K6 sowie von dem Axiom K5a). Das bedeutet jedoch nicht, dass die Forderungen, die an die jeweilige Klasse von Maßzahlen gestellt werden, auf das gleiche hinauslaufen.

So wäre beispielsweise das Verhältnis von Spannweite und Median ein brauchbares Maß der relativen Streuung. Es sei SM genannt. Offenbar erfüllt SM auch das Axiom S2 (Vergrößerung der Streuung bei Hinzutreten eines Wertes, der größer (kleiner) als der bisher größte (kleinste) Wert ist, das im besonderen Maße den Gedanken der Streuung wiedergibt. Ähnlich eng verbunden mit der anschaulichen Vorstellung ist im Falle der "Disparität" das Axiom K2 (Transfer), das jedoch von SM nicht erfüllt sein muss. Ein Transfer im Sinne von Axiom K2 wirkt sich auch meist nicht aus auf die mittlere Abweichung, die gleichwohl Grundlage eines sinnvollen Maßes der Streuung sein kann.

2. Disparität und Schiefe

Im Sprachgebrauch von Politik und Wirtschaft wird Disparität ("Ungleichheit") auch häufig im Sinne einer linkssteilen Verteilung gebraucht: die große Mehrheit verdient wenig und einige wenige "Besserverdienende" verdienen vergleichsweise viel. Schiefe und Disparität sind zwei Aspekte einer Häufigkeitsverteilung die durchaus **Gemeinsamkeiten** haben, nämlich

1. beide sind invariant gegenüber proportionalen Transformationen, d.h. sie erfüllen das Axiom K1 und
2. beide sind unabhängig von der Anzahl der Merkmalsträger in dem Sinne wie dies bei der Proportionalitätsprobe (Axiom K4) postuliert wird.

Es gibt jedoch bedeutende **Unterschiede**:

1. Besonders auffallend ist, dass die Schiefe als Gestaltparameter notwendig verschiebungsinvariant ist, d.h. dass sie auf Niveauänderungen nicht reagiert und somit Axiom K3 nicht erfüllt.
2. Nicht in gleicher Weise unmittelbar einsichtig ist, dass die Schiefe nicht Axiom K2 erfüllt, also beispielsweise sich bei einem negativen (egalisierenden) Transfer nicht notwendig verringert.
3. Bei einer Nullergänzung im Sinne des Axioms K5 soll ein Disparitätsmaß zunehmen, die Schiefe kann jedoch auch abnehmen.

Die Punkte 2 und 3 seien kurz an Beispielen veranschaulicht:

zu 2:

Durch einen Transfer des Betrags d von i auf j ($x_{(i)} > x_{(j)}$) im Sinne des Axioms K2 verringert sich stets die Varianz von s^2 auf s^{2*} mit

$$s^{2*} = s^2 + \frac{2d}{n} \quad (x_{(j)} - x_{(i)} + d) < s^2$$

und damit auch die Standardabweichung. Anders verhält es sich jedoch mit dem dritten zentralen Moment z_3, das zwar abnehmen kann, möglicherweise aber in einem geringeren Maße als die dritte Potenz der Standardabweichung, so dass sich die Schiefe erhöht.

Als Beispiel sei die folgende linkssteile Ausgangsverteilung gewählt: $x_{(1)} = 100$ $x_{(2)} = 200$ und $x_{(3)} = 600$.
Das dritte zentrale Moment z_3 verringert sich von 6 auf 3,75 Millionen bei einem Transfer des Betrags 50 von Einheit 3 auf Einheit 1, die dann den Betrag von 150 hat. Dadurch steigt die Momentschiefe von 0,59517 auf 0,66547, d.h. die Verteilung ist trotz eines egalisierenden Transfers im höheren Maße linkssteil.

zu 3:

Gegeben sei die folgende linkssteile Ausgangsverteilung mit den Beträgen 100, 100, 100 und 200. Tritt zu den n = 4 Personen ein Nullträger hinzu, so wird die linkssteile Verteilung (Momentschiefe 1,1547) zu einer symmetrischen Verteilung. Die Disparität ist dagegen gemessen am Disparitätsmaß D_G von Gini von 0,15 auf 0,32 gestiegen. Wie man daran sieht, kann auch eine symmetrische Verteilung eine größere Disparität aufweisen, als eine linkssteile Verteilung.

Entgegen einer verbreiteten Vorstellung ist also "Ungleichheit" nicht adäquat operationalisiert mit der "Schiefe" (wobei meist an Linkssteilheit gedacht wird) im Sinne der deskriptiven Statistik.

5. Dominanzmaße: Entdeckung oligopolistischer Strukturen

Neben dem Herfindahl-Index K_H spielt der im folgenden dargestellte Linda-Index (nach Remo Linda, der ihn 1976 vorschlug) in den Arbeiten der Monopolkommission und der amtlichen Statistik eine große Rolle. Mit dem Linda-Index, bzw. genauer dem System von Linda-Indizes L_k (k=2,3,...,n-1), will man oligopolistische Strukturen erkennen, d.h. feststellen wieviele Unternehmen (z.B. die größten vier), die sog. "Oligopolgruppe", die übrigen Unternehmen (Mitläufer, Umfeld) "dominieren". Man spricht deshalb auch von einem Dominanzmaß. Neben dem Linda-Index sind auch andere Dominanzmaße vorgeschlagen worden, auf die hier aus Platzgründen nicht eingegangen werden kann. Es sollte jedoch vorweg erläutert werden:

1. die Unterscheidung zwischen summarischen und diskreten Konzentrationsmaßen und
2. das Konzept der Dominanz.

zu 1:

Summarische Konzentrationsmaße (wie K_H) beschreiben einen Markt mit n Teilnehmern (z.B. Anbietern) mit einer einzigen Maßzahl, ohne zwischen den n Unternehmen zu differenzieren. Diskrete Maße erlauben es dagegen auch Teile des Marktes (als m < n Unternehmen) zu betrachten und so oligopolistische Strukturen zu identifizieren.

zu 2:

Auch für Dominanzmaße gibt es eine Axiomatik, auf die hier jedoch nicht eingegangen werden kann. Es ist aber z.B. unmittelbar einsichtig, dass die Vorherrschaft der Oligopolgruppe umso größer ist, je

- kleiner (je weniger Unternehmen umfassend) die Gruppe ist
- geringer die Disparität zwischen den Unternehmen der Gruppe ist
- größer ihr kumulierter Marktanteil ist und je
- größer und weniger differenziert das Umfeld ist,

denn das sind Bedingungen für eine Interessensymmetrie und damit für gemeinsames Handeln der dominierenden Unternehmen. Daraus folgt auch, dass Dominanzmaße, die auf diese Bedingungen reagieren sollten, Eigenschaften von Konzentrations- und Disparitätsmaßen miteinander kombinieren.

Def. 6.10: Linda-Indizes

Das sog. oligopolistische Gleichgewicht der i größten Unternehmen EO_i (oder $EO_{i,k}$, weil es von i und k abhängt) ist ein Verhältnis von Marktanteilen (C_i, C_k sind Konzentrationsraten [Def. 6.3])

$$(6.33)\quad EO_i = \frac{C_i}{i} : \frac{C_k - C_i}{k-i} = \frac{C_i(k - i)}{i(C_k - C_i)}$$

mit $1 \le i \le k-1$ und $k = 2,3,...,n-1$ bei n Einheiten (z.B. Anbietern auf dem Markt). Der erste Klammerausdruck ist der obere, der zweite der untere Mittelwert. Geometrisch veranschaulicht ist der oligopolistische Kern die Stelle, an der die Konzentrationskurve den stärksten Knick hat. Der (k-te) Linda-Index L_k ist definiert als Mittelwert der durch k dividierten Größen EO_i

$$(6.34)\quad L_k = \frac{1}{k-1}\sum_{i=1}^{k-1} EO_i /k = k(k-1)\sum_{i=1}^{k-1} EO_i \quad .$$

Bemerkungen zu Def. 6.10:

1. Bei gleichen Marktanteilen von k Unternehmen ist $C_i = i/k$ für alle i so dass $EO_i = 1$ und $L_k = 1/k$. Nach oben ist L_k nicht beschränkt, so dass gilt $1/k \leq L_k < \infty$.

2. Sind die Marktanteile nicht gleich, so ist L_k eine Funktion von k mit mindestens einem Minimum. Der *Oligopolkern* (die k^* größten und dominierenden Unternehmen) ist der Wert $k = k^*$, bei dem L_k zum ersten Mal minimal ist. Das Minimum ist umso ausgeprägter, je größer die Disparität unter den k^* Kernunternehmen ist.

3. Man kann den unter Nr. 1 genannten Mangel beheben, indem man nur Unternehmen mit einem Marktanteil von mindestens δ in die Betrachtung einbezieht und L_k mit einem von δ abhängigen Faktor auf den Wertebereich $1/k \leq L_k^* \leq 1$ für einen modifizierten Index L_k^* normiert.

Beispiel 6.8:

Die Berechnung des Linda-Index soll demonstriert werden bei einem fiktiven Markt mit acht Unternehmen und den Marktanteilen c_i und den kumulierten Marktanteilen C_i:

	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8
c_i	0,40	0,20	0,15	0,10	0,08	0,04	0,02	0,01
C_i	0,40	0,60	0,75	0,85	0,93	0,97	0,99	1,00

Man stelle mit den Linda-Indizes L_k die Anzahl k^* der Unternehmen fest, die eine Oligopolgruppe bilden.

Lösung 6.8:

Die Berechnung von L_k soll für k = 2 und k = 3 ausführlich gezeigt werden und für die höheren Werte für k sind die Ergebnisse in einer Tabelle zusammengefaßt.

k = 2 (man erhält $L_2 = 1$)
$EO_1 = C_1/1 : (C_2-C_1)/(2-1) = 0,4/0,2 = 2$
dann ist $EO_1/k = 2/2 = 1 = L_2$

k = 3 ($L_3 = 0,7143$)
$EO_1 = C_1:(C_3-C_1)/(3-1) = 0,4: (0,35/2) = 0,4/0,175 = 2,2857$
$EO_1/k = 2,2857/3 = 0,7619$
$EO_2 = C_2/2 : (C_3-C_2)/(3-2) = 0,6/2 : 0,15/1 = 2$; $EO_2/k = 2/3$

und damit ist $L_3 = (EO_1/3 + EO_2/3)/(3-1) = (0{,}7619 + 2/3)/2$
$L_3 = 0{,}7143$

Der kleinste Wert von L_k ist, wie die folgende Tabelle zeigt, erreicht bei k=5 (denn $L_5 = 0{,}55901$), so dass man sagen kann, die fünf größten Unternehmen bilden eine Oligopolgruppe.

	Werte für EO_i bei							Linda-Index L_k
k	i=1	i=2	i=3	i=4	i=5	i=6	i=7	
4	2,67	2,4*)	2,5*)					0,65306
5	3,02	2,73	2,78	2,66				0,55901
6	3,51	3,24	3,41	3,54	4,65			0,61176
7	4,07	3,85	4,17	4,55	6,2*)	8,08		0,73613
8	4,67	4,5*)	5*)	5,67	7,97	11,11	14,14	0,94748

*) Auf Nachkommastellen wurde verzichtet, wenn nicht gerundet wurde. Die Angaben sind zu ungenau um L_k zu berechnen. Sie sollen nur eine Kontrolle ermöglichen, wenn man die Indizes "zu Fuß" (also Schritt für Schritt nach den angegebenen Formeln) berechnen möchte, was für Übungszwecke sehr zu empfehlen ist.

6. Zur Vertiefung des Verständnisses der Lorenzkurve

a) Momentverteilung und Häufigkeitsverteilung

Eine (diskrete) Häufigkeitsverteilung ist als Folge der Wertetupel (x_i,h_i) definiert. Treten die Merkmalsanteile q_i an die Stelle der relativen Häufigkeiten h_i, so spricht man von der Momentverteilung. Sie spielt nicht nur bei der Disparitätsmessung, sondern auch für die Betrachtung von Mittelwerten eine gewisse Rolle.

Def. 6.11: Momentverteilung, Scheidewert und Schwerster Wert

Die Wertetupel (x_i,q_i) heißen **Momentverteilung** (oder: Wertverteilung). Dabei sind die Größen q_i die mit

(6.2a) $q_i = \dfrac{x_{(i)}}{\Sigma x_{(i)}}$ bei Einzelwerten, bzw.

(6.2b) $q_i = \dfrac{n_i}{\Sigma x_i n_i}$ bei Häufigkeitsverteilungen

definierten Merkmalsanteile.

Der Median der Momentverteilung heißt **Scheidewert** (oder auch Medial) und der Modus der Momentverteilung ist der **Schwerste Wert** ($\bar{x}_T$).

Bemerkungen zu Def. 6.11:

1. Zwischen den Merkmalsanteilen q_i und den relativen Häufigkeiten h_i besteht allgemein die folgende Beziehung:

   (6.23) $$\frac{q_i}{h_i} = \frac{x_i}{\bar{x}}$$

   Bei Einzelwerten ist $x_{(i)}$ anstelle von x_i zu setzen und es gilt für alle i, dass $h_i = 1/n$. Dabei ist q_i/h_i die Steigung der Lorenzkurve zwischen den Punkten (H_{i-1},Q_{i-1}) und (H_i, Q_i). Daraus folgt:

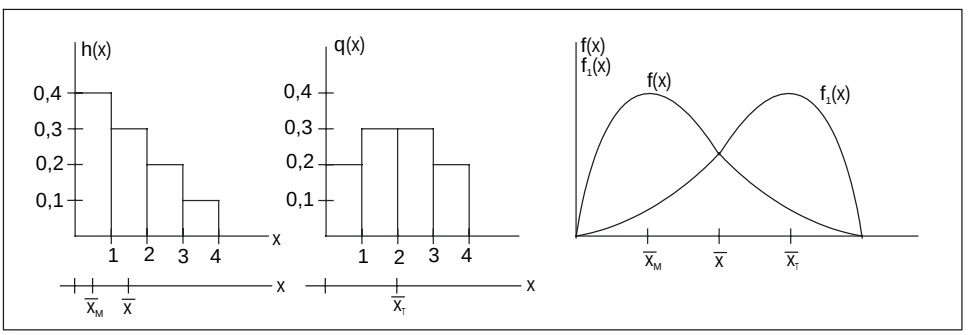
   wenn $x_i < \bar{x}$, dann $q_i < h_i$
   wenn $x_i = \bar{x}$, dann $q_i = h_i$
   wenn $x_i > \bar{x}$, dann $q_i > h_i$

   Entsprechend ist die Steigung der Lorenzkurve kleiner als 1, genau 1 oder größer als 1.

2. Für den Schwersten Wert $\bar{x}_T$ gilt im Verhältnis zum Modus (dichtester Wert) $\bar{x}_M$ deshalb $\bar{x}_T \geq \bar{x}_M$. Dies ergibt sich aus Bem. Nr. 1 wonach die Werte h_i ihr Maximum vor (links von) den Werten q_i erreicht haben müssen sowie aus dem in Abb. 6.10 dargestellten typischen Verlauf von Momentverteilung und Häufigkeitsverteilung.

3. Neben Modus und Median der Momentverteilung spielt auch das arithmetische Mittel K_N dieser Verteilung eine gewisse Rolle.

Abb. 6.10: Momentdichte $f_1(x)$ und Dichte $f(x)$



Es wurde als Disparitätsmaß vorgeschlagen (Niehaus 1955). Es gilt wegen Gl. 6.23 und in Verbindung mit dem quadratischen Mittel $\bar{x}_Q$ sowie dem Verschiebungssatz der Varianz (V = Variationskoeffizient)

$$(6.35) \qquad K_N = \Sigma x_i q_i = \frac{\Sigma x_i^2 h_i}{\bar{x}} = \frac{\bar{x}_Q^2}{\bar{x}} = \frac{s^2+\bar{x}^2}{\bar{x}} = \bar{x} + sV$$
$$= \bar{x}(V^2+1) \geq \bar{x} \ .$$

Das arithmetische Mittel der Momentverteilung kann also nicht kleiner als das arithmetische Mittel (der Häufigkeitsverteilung) sein. Nach Gl. 6.10 ergibt sich der folgende Zusammenhang zum Herfindahl-Index K_H

$$(6.36) \qquad K_H = \frac{K_N}{n\bar{x}} = \frac{K_N}{\Sigma x_i n_i},$$

so dass K_H auch als ein durch die Merkmalssumme normiertes arithmetisches Mittel der Momentverteilung interpretiert werden kann.

4. Der Ausdruck

$$(6.37) \qquad G_V = \frac{1}{1+V^2} = \frac{\bar{x}}{K_N} = \frac{1}{nK_N} \ ,$$

also das Verhältnis der Mittelwerte von Häufigkeits- und Momentverteilung, gilt als Gleichheitsmaß (vgl. auch Gl. 6.31a). G_V und K_H bilden ein Paar gleichmäßig normierter Maße (vgl. Abschn. 6d, Gl. 6.25a) und GV ist das zum normierten Quadrat des Variationskoeffizienten NV (gem. Gl. 6.30) gehörende Gleichheitsmaß.

Beispiel 6.9:

Man bestimme die Momentverteilung und deren Lageparameter für das Beispiel 6.6.

Lösung 6.9:

In der Lösung werden bereits Angaben gemacht, auf die erst an späterer Stelle (maximaler Nivellierungssatz) hingewiesen wird.

Häufigkeitsverteilung

x_i	h_i
1	0,4
2	0,3
3	0,2
4	0,1

Momentverteilung

x_i	q_i
1	0,2
2	0,3
3	0,3
4	0,2

Abstand Lorenzkurve-Gleichverteilungsgerade

x_i	H_i-Q_i
1	0,2
2	0,2
3	0,1
4	0,0

Modus $\bar{x}_M = 1$ Schwerster Wert $\bar{x}_T = 2{,}5$

Median = 1,333 Scheidewert (Median) = 2

$\bar{x} = 2$ $K_H = 2{,}5$.

Für die Häufigkeits- und Momentverteilung gilt also der typische Verlauf der Abb. 6.10. Der Abstand zwischen Gleichverteilungsgerade und Lorenzkurve ist maximal 2 an der Stelle $x = \bar{x}$ (bei H = 0,7 und Q = 0,5, also für die Gruppe, die genau die durchschnittliche Anzahl von Gänsen erhält, nämlich 2). Dieser Abstand ist der maximale Nivellierungssatz (vgl. Def. 6.13).
Weitere Kennzahlen und Zusammenhänge: Varianz der Häufigkeitsverteilung $s^2 = 1$, quadratisches Mittel $Q = \sqrt{5}$ (weshalb auch $K_N = Q^2/\bar{x} = 5/2 = 2{,}5$), Herfindahl-Index $K_H = K_N / n\bar{x} = 2{,}5/300 = 0{,}00833$. Man kann K_H auch aus Gl. 6.10 errechnen: $V^2 = 0{,}25$, so dass $K_H = 1{,}25/150$. Das Gleichheitsmaß G_V gem. Gl. 6.37 ist 0,8.

b) Stetige Lorenzkurve

Die folgende Betrachtung mag dazu beitragen, einige Zusammenhänge zu verdeutlichen, sie ist aber für die Anwendung auf empirische Daten nicht von Bedeutung, weil in diesem Falle eine stetige Variable stets in klassierter (und damit diskreter) Form vorliegt. An dieser Stelle werden einige Begriffe und Symbole aus der Induktiven Statistik vorausgesetzt. Leser, die hiermit nicht vertraut sind, können den Exkurs getrost überschlagen.
Für die stetige Variable X sei f(x) die Dichtefunktion und

$$\mu = \int_a^b xf(x)dx \qquad (a \le x \le b)$$

das arithmetische Mittel (das erste Anfangsmoment). Dann ist die erste Momentdichte gegeben mit

(6.38) $$f_1(x) = \frac{1}{\mu} xf(x) .$$

und

(6.38a) $F_1(x) = \int_a^x f_1(u)du$ ist die Momentverteilung analog zur

Verteilungsfunktion $F(x) = \int_a^x f(u)du$.

Die Zusammenhänge zwischen den Dichtefunktionen f(x) und $f_1(x)$ sowie den Verteilungsfunktionen F(x) und $F_1(x)$ sind aufschlußreich für das Verständnis der Lorenzkurve. Hierzu die folgenden Bemerkungen:

1.) Offensichtlich ist $f_1(x)$ eine Dichtefunktion, da

$$\int_a^b f_1(x)dx = \frac{\mu}{\mu} = 1$$

2.) Wegen Gl. 6.38 gilt $\mu f_1(x) = xf(x)$ nach der Produktregel der Differentiation

(6.39) $\mu f_1'(x) = x f'(x)+f(x),$

so dass das Maximum der Momentdichte $f_1(x)$, also der schwerste Wert $\bar{x}_T$ im Bereich fallender Werte der Dichtefunktion liegt, denn aus $f_1'(x) = 0$ folgt $f'(x) < 0$. Deshalb ist, wie bereits gesagt, der schwerste Wert $\bar{x}_T$ nicht kleiner als der dichteste Wert $\bar{x}_M$. In Abb. 6.10 ist der typische Verlauf der beiden Dichtefunktionen wiedergegeben.

3.) Aus (6.38) folgt

(6.40) $$F_1(x) = \frac{1}{\mu}\int_a^b uf(u)du = \mu_1 / \mu$$

wobei μ_1 das bis zur Stelle x erreichte arithmetische Mittel ist und $\mu_1 \leq \mu$ gilt. Daraus folgt, dass $F_1(x) \leq F(x)$, analog zur bekannten Beziehung $Q_i \leq H_i$ im diskreten Fall, gilt. Abgesehen von der Rechteckverteilung $F(x) = x(b-a)^{-1}$ gilt die Gleichheit von $F_1(x)$ und F(x) nur für die Punkte x = a und x = b.

Wir sind nun in der Lage, die Lorenzkurve für eine stetige Variable X zu definieren und diese Definition anhand eines Beispiels (Beispiel 6.10) zu erläutern.
Diese stetige Darstellung ist vor allem deshalb von Interesse, weil so die allgemeinen Eigenschaften von Disparitätsmaßen besser herausgearbeitet werden können. In allen empirischen Anwendungen ist dagegen bei der Disparitätsmessung stets von klassierten Daten oder Einzelbeobachtungen auszugehen, so dass die Lorenzkurve ein Polygonzug darstellt, den man als Annäherung an den "wahren" kontinuierlichen Kurvenverlauf auffassen kann.

Def. 6.12: Lorenzkurve im stetigen Fall

Die Funktion L(F), die im F(x), $F_1(x)$-Koordinatensystem alle Punkte mit den Koordinaten $L(F)=F_1(F)$ und F miteinander verbindet, heißt Lorenzkurve. Man erhält $F_1(F)$, indem man x in $F_1(x)$ ersetzt durch x = G(F), wobei G die inverse Verteilungsfunktion ist.

Zum Begriff der inversen Verteilungsfunktion (vgl. Def. 3.3,d):

Für einen Wert x des Merkmals X ist mit F = F(x) die kumulierte relative Häufigkeit gegeben. Man kann umgekehrt fragen, welchen x-Wert man erhält, wenn F einen bestimmten vorgegebenen Wert annehmen soll, etwa F = ½, dann ist der gesuchte x-Wert der Median $\tilde{x}_{0,5} = G(F=½)$.

Beispiel 6.10:

Mit diesem einfachen Beispiel soll die mit Definition 6.12 gegebene Vorschrift zur Herleitung der Lorenzkurve demonstriert werden. Gegeben sei die Dichtefunktion

$$f(x) = 3x^2/8 \qquad \text{für } 0 \le x \le 2.$$

Hieraus folgt $\mu = 1{,}5$ und $F(x) = 1/8\ x^3$.
Dann ist $f_1(x) = x^3/4$ und $F_1(x) = x^4/16$.
Löst man F(x) nach x auf, so erhält man $x = (8F)^{1/3}$. Dies ist die inverse Verteilungsfunktion.
Die Lorenzkurve lautet dann $L(F) = F_1(F) = (1/16)(8F)^{4/3} = F^{4/3}$.
Die Steigung der Lorenzkurve im Beispiel 6.11 ist

$$L'(F) = \frac{dL}{dF} = \frac{f_1(x)}{f(x)} = \frac{x^3}{4} \cdot \frac{8}{3x^2} = \frac{x}{1{,}5} = \frac{x}{\mu} \ge 0.$$

Ferner ist $L''(F) = [\mu f(x)]^{-1} > 0$,

d.h. die Steigung der Lorenzkurve ist nicht-negativ und monoton zunehmend.
Für das Disparitätsmaß von Gini D_G erhält man:

$$(6.41)\quad D_G = 1 - 2\int_0^1 L(F)dF$$

Man kann zeigen, dass für Ginis D_G auch gilt

$$(6.42)\quad D_G = \frac{1}{\mu}\int_a^b (F-F^2)dx$$

Im Beispiel 6.10 ist $D_G = 1/7$. Man verifiziert im obigen Beispiel übrigens leicht, dass gilt $L(F) \le F$, d.h. die Lorenzkurve ist konvex und dass für die Momentverteilung das arithmetische Mittel $\int_0^2 xf_1(x)dx = 1{,}6$ beträgt und größer ist als $\mu = 1{,}5$.

c) Schutz-Koeffizient (Maximaler Nivellierungssatz)

Im folgenden wird gezeigt, dass der senkrechte Abstand zwischen der Gleichverteilungsgeraden und der Lorenzkurve maximal ist an der Stelle $x = \mu$ und als Disparitätsmaß zu interpretieren ist. Es ist der Schutz-Koeffizient. Er wird auch maximaler Nivellierungssatz (von Lindahl) oder längste Lorenzkurvensehne genannt.

Def. 6.13: Schutz-Koeffizent

Der Schutz-Koeffizient oder maximale Nivellierungssatz von Lindahl ϕ^* ist mit

$$(6.43)\quad \phi^* = \frac{1}{\mu} F_\mu(\mu - \mu_1) = F_\mu\left(1 - \frac{\mu_1}{\mu}\right)$$

gegeben. Hierbei ist μ_1 der durchschnittliche Merkmalsbetrag derjenigen Teilgesamtheit, die einen unterdurchschnittlichen Merkmalsbetrag hat, also solange $x \le \mu$ ist.
Es ist der Anteil am Gesamtmerkmalsbetrag, der von der Teilgesamtheit mit überdurchschnittlichem Merkmalsbetrag an diejenige mit unterdurchschnittlichem Betrag umverteilt werden müßte, um zur Gleichverteilung (egalitären Verteilung) zu gelangen.

Im Beispiel 6.10 beträgt diese Größe $F(x=\mu) - F_1(x=\mu) = 27/64 - 81/256 = 0{,}1055$. Es gelten die folgenden Zusammenhänge:

1. Die senkrechte Distanz zwischen der Gleichverteilungsgeraden und der Lorenzkurve $\phi(F) = F - L(F)$ ist maximal an der Stelle $x = \mu$, an der die Steigung der Lorenzkurve 1 beträgt und sie ist dann:
   $\phi^* = \max \phi(F) = F_\mu - L(F_\mu)$ mit $F_\mu = F(x=\mu)$

2. Die maximale Distanz beträgt ϕ^* gem. Gl. 6.43.

Beweis

1. Aus $\phi'(F) = 1 - L'(F) = 1 - x/\mu = (\mu - x)/\mu$ und $\phi''(F) = -L''(F) = -\mu f(x)^{-1} < 0$ folgt wegen $\phi'(F) = 0$, dass das Maximum von ϕ an der Stelle $x = \mu$ gegeben ist.

2. Es gilt (am Beispiel der Einkommensverteilung) für die Durchschnittseinkommen der unterdurchschnittlich (μ_1) bzw. der überdurchschnittlich (μ_2) verdienenden Merkmalsträger

   $$\mu_1 = F_\mu^{-1} \int_a^{x=\mu} x f(x)dx \quad \text{und} \quad \mu_2 = (1-F_\mu)^{-1} \int_{x=\mu}^{b} x f(x)dx \quad \text{mit } F_\mu = \int_a^{x=\mu} f(x)\,dx$$

   Wegen $\mu = F_\mu \mu_1 + (1-F_\mu)\mu_2$ gilt

   $$\phi^* = F_\mu - L(F_\mu) = F_\mu - F_1(F_\mu) = F_\mu - \mu^{-1} \int_a^{x=\mu} x f(x)\,dx = F_\mu - \mu^{-1} F_\mu \mu_1.$$

3. Die Interpretation von ϕ^* als "Nivellierungssatz" wird aus der folgenden Überlegung deutlich:
   Wird der Anteil ϕ^* von den überdurchschnittlich verdienenden Einheiten, deren bisheriger Einkommensanteil $(1 - F_\mu)\,\mu_2 / \mu$ beträgt, an die

unterdurchschnittlich verdienenden Einheiten übertragen, deren Anteil bisher $F_\mu \mu_1 /\mu$ beträgt, so wird die
egalitäre Verteilung entstehen. Denn dann gilt für die Einkommensanteile der bisher überdurchschnittlich Verdienenden

$$\mu^{-1}(1-F_\mu)\mu_2 - F_\mu(1 - \mu_1/\mu) = 1-F_\mu$$

und der bisher unterdurchschnittlich Verdienenden

$$\mu^{-1}F_\mu\mu_1 + F_\mu(1 - \mu_1/\mu) = F_\mu.$$

Die Anteile an dem Merkmalsbetrag entsprechen jetzt genau den Anteilen der beiden Gruppen an der Gesamtzahl der Merkmalsträger, d.h. es ist $F_1=F$ (egalitäre Verteilung). Zu diesen Zusammenhängen vgl. auch Bsp. 6.12.

d) Gleichmäßig normierte Maße

Def. 6.14: gleichmäßig normierte Maße

Maße der Gleichheit G (vgl. Def. 6.2) und der Konzentration K, die sich bezüglich der Axiome K4 und K5 gleichmäßig ihrer unteren Grenze nähern und für die gilt

$$\text{(6.25a)} \quad K = \frac{1}{n(1-D)} = \frac{1}{nG} \quad \text{und damit} \quad G = \frac{1}{nK}$$

heißen gleichmäßig normierte Maße (nach Jöhnk).

Bemerkungen zu Def. 6.14:

1. Zwei Paare gleichmäßig normierter Maße sind beispielsweise
   1. Gini-Koeffizient D_G und Rosenbluth-Index K_R
   2. normierter Variationskoeffizient $NV = V^2 / (1+V^2)$ und Herfindahl-Index K_H
   denn es gilt

Disparitätsmaß	Gleichheitsmaß	Konzentrationsmaß	vgl. Gleichung
Gini D_G	$1-D_G$	$K_R = [n(1- D_G)]^{-1}$	Gl. 6.25
norm.Var. NV	$G_V = (V^2+1)^{-1}$	$K_H = (V^2+1)/n$	Gl. 6.10, 6.37

2. Die Gleichung KnG = 1 läßt sich umformen in

   $$\log(K) + \log(n) + \log(G) = 0.$$

   Dies erlaubt es, zwei Konzentrationszustände (K_1,K_2) in bezug auf den Anzahleffekt ($n_1 < n_2$ oder $n_1 > n_2$) und den Disparitätseffekt Unterschiedlichkeit von D_1 und D_2,
   bzw. G_1 und G_2) zu vergleichen:

   $$\text{(6.44)} \qquad \log(K_2/K_1) = \log(n_1/n_2) + \log(G_1 /G_2) .$$

Der erste Summand ist der Anzahleffekt (AE, absolute Komponente), der zweite der Verteilungs- oder Disparitätseffekt (DE, relative Komponente). Diese Zerlegung wird an dem folgenden Beispiel 6.11 demonstriert.

3. Aus der Zerlegung gem. Gl. 6.44 folgt, dass Veränderung der (absoluten) Konzentration sowohl von einem Anzahl-, als auch von einem Disparitätseffekt herrühren kann. Eindeutige Aussagen sind nur möglich, wenn sich n und G gleichsinnig verändern, d.h. wenn:

   n steigt und G steigt (D sinkt): dann sinkt K
   n sinkt und G sinkt (D steigt): dann steigt K .

Beispiel 6.11:

Ausgangsverteilung $n_1 = 2$ nach Nullergänzung $n_2 = 3$

Verteilung 1

Einheit	Anteil
1	0,8
2	0,2

Verteilung 2

Einheit	Anteil
1	0,8
2	0,2
3	0

a) Zeigen Sie die Wirkung auf das Konzentrationsmaß $K = K_R$ (Rosenbluth) und das Gleichheitsmaß $G = 1 - D_G$ sowie die Zerlegung des Konzentrationseffektes in den Anzahl- und Disparitätseffekt.

b) Das Beispiel ist dahingehend zu variieren, dass eine Verteilung Nr. 3 mit $n_3 = 4$ Einheiten aus der Ausgangsverteilung (Verteilung 1) durch Halbierung jeder Einheit entsteht.

Lösung 6.11:

Das Beispiel ist so konstruiert, dass der Disparitätseffekt (Teil a, Axiom K5) und der Anzahleffekt (Teil b, Axiom K4) jeweils in reiner Form zum Tragen kommen, was sich dann auch an der Zerlegungsformel (Gl. 6.44) zeigen muss.

a) $K_1 = 1/1,4 = 0,7143$ $G_1 = 0,7$ (da $D_G = 0,3$).
(für den Herfindahl-Index erhielte man $K_H = 0,68$ und für G_V den Wert $G_V = 1/1,36 = 0,7353$)
Die Nullergänzung verändert die Konzentration nicht (Axiom K5) $K_2 = K_1$, während die Disparität zunehmen muss. Es gilt $G_2 = 0,4667 = (2/3)G_1$.
Für die obige Zerlegungsformel (Gl. 6.42) erhält man
$\log(1) = \log(2/3) + \log(3/2) = -0,1761 + 0,1761 = 0.$
Anzahl- und Disparitätseffekt sind betragsmäßig gleich und heben sich auf.

b) Die neue Verteilung hat folgende Gestalt:

Einheit	1	2	3	4
Anteil	0,4	0,4	0,1	0,1

Es gilt dann $K_3 = 1/2,8 = ½K_1$ und $G_3 = G_1$. Dann ist

$\log(1/2) = \log(4/2) + \log(1) = -0{,}30103$.

Fazit: Die Halbierung der Konzentration ist ausschließlich auf den Anzahleffekt zurückzuführen, was ja auch bei dem gem. Axiom K4 (reiner Anzahleffekt) konstruierten Beispiel zu erwarten war.

Beispiel 6.12:

Gegeben sei das stetige Konzentrationsmerkmal X (z.B. das Einkommen), dessen Dichte lautet

$$f(x) = \begin{cases} 0{,}4 - 0{,}08x & \text{für } 0 \le x \le 5 \\ 0 & \text{sonst} \end{cases}$$

Man berechne die Momentdichte (und deren Maßzahlen, wie z.B. Mittelwerte), die Lorenzkurve und die anderen in diesem Exkurs dargestellten Größen.

Lösung 6.12:

- Verteilung von x:
  $f(x) = 0{,}4 - 0{,}08x$ (Dreiecksverteilung) $\mu = 5/3 = 1{,}667$
  der Median ist an der Stelle $x = 1{,}4645$ denn $F(x = 1{,}4645) = ½$,
  $F(x) = 0{,}4x - 0{,}04x^2 = F$,
  inverse Verteilungsfunktion $x = G(F) = 5 - 5\sqrt{(1-F)}$.

- Momentdichte $f_1(x)$ und Momentverteilung $F_1(x)$
  $f_1(x) = (x/\mu)f(x) = 0{,}24x - 0{,}048x^2$ $\qquad F_1(x) = 0{,}12x^2 - 0{,}016x^3$
  arithmetisches Mittel der Momentdichte $K_N = \int_0^5 x f_1(x)dx = 2{,}5$;
  Median der Momentdichte (= Scheidewert) 1,675 [Achtung: die Bestimmung des Scheidewerts verlangt wegen $F_1(x)$ die Lösung einer kubischen Gleichung]; der Modus der Momentdichte (d.h. der Schwerste Wert $\bar{x}_T$) ist $\bar{x}_T = 2{,}5$.

- Lorenzkurve
  $L(F) = F_1(F)$ wobei in $F_1(x)$ für x die inverse Verteilungsfunktion einzusetzen ist. Man erhält dann: $L(F) = 2(1-F)^{3/2} + 3F - 2$.
  Die Ableitung dieser Funktion ergibt die Steigung der Lorenzkurve. Es gilt:
  $dL(F)/dF = 3 - 3(1-F)^{½}$. Man erkennt sofort, dass dieser Ausdruck gleich ist der Relation x/μ, denn aus der inversen Verteilungsfunktion folgt
  $x = 5 - 5(1-F)^{½}$ und für μ gilt $\mu = 5/3$.

- Ginis-Disparitätsmaß D_G
  Man errechnet D_G aus Gl. 6.41. Für das Integral über L(F) in den Grenzen von 0 bis 1 erhält man (Integrationskonstante C) den Ausdruck:
  $-(4/5)(1-F)^{5/2} + 1{,}5F^2 - 2F + C$. Mit den Grenzen F=0 und F=1 ergibt sich dann für D_G der Zahlenwert $D_G = 0{,}4$.
  Das Disparitätsmaß G_V nimmt als das Verhältnis der Mittelwerte von Dichte f(x) und Momentdichte $f_1(x)$ den Wert $G_V = 2/3$ an.

- Abstand zwischen Gleichverteilungsgeraden und Lorenzkurve F - L(F)

Er beträgt $F - L(F) = 2(1-F) - 2(1-F)^{3/2}$. Dieser Abstand ist maximal an der Stelle F = 5/9 und er beträgt dort 8/27. Man erhält den Wert 5/9 auch, wenn man in F(x) für x den Wert $x = \mu = 5/3$ einsetzt, d.h. $F_\mu = 5/9$. Das verifiziert den Satz, dass der Abstand zwischen der Gleichverteilungsgeraden und der Lorenzkurve maximal an der Stelle $x = \mu$ ist.

- Maximaler Nivellierungssatz (Schutzkoeffizient) und Interpretation hiervon:

  Der maximale Nivellierungssatz nach Lindahl ist damit 8/27. Das Einkommen der unterdurchschnittlich verdienenden Einheiten μ_1 und der überdurchschnittlich verdienenden Einheiten μ_2 beträgt:

  $\mu_1 = \frac{1}{F_\mu} \int_0^\mu xf(x)dx = (9/5)(35/81)=7/9= 0{,}7778$. Entsprechend ist

  $\mu_2 = \frac{1}{1 - F_\mu} \int_\mu^5 xf(x)dx = (9/4)(100/81) = 25/9 = 2{,}7778.$

  Mit $\mu = 15/9 = 5/3 = 1{,}6667$ ist also $\mu_1 < \mu < \mu_2$.

  Die Anteile an der Gesamtheit der Merkmalsträger betragen für die
  unterdurchschnittlich Verdienenden $F_\mu = 5/9$
  überdurchschnittlich Verdienenden $1 - F_\mu = 4/9$.

  Die Anteile am Gesamtmerkmalsbetrag sind für die
  unterdurchschnittlich Verdienenden $F_{1\mu} = F_1(x=\mu) = 7/27$
  überdurchschnittlich Verdienenden $1 - F_{1\mu} = 20/27$.

  Der maximale Nivellierungssatz ist 8/27. Wird der Anteil 8/27 den unterdurchschnittlich Verdienenden zugeschlagen, so ist ihr Anteil dann
  $F_{1\mu} + 5/9 = 7/27 + 8/27 = 15/27 = F_1(x=\mu) = 5/9$.
  Entsprechend verbleibt bei den überdurchschnittlich Verdienenden ein Anteil von 20/27-8/27=4/9, wie es ihrem Anteil an der Bevölkerung entspricht.

- Noch zur Interpretation des Gini-Koeffizienten

  Würde man nur die beiden Gruppen unter- und überdurchschnittlich Verdienende unterscheiden und mit diesen beiden Klassen eine Lorenzkurve durch lineare Verbindungen der Punkte (0,0),(5/9,7/27) und (1,1) zeichnen, so würde D_G nach Gl. 6.29a den Wert 5/9 - 7/27 = 8/27 (= maximaler Nivellierungssatz) annehmen. Man sieht, dass die Lorenzkurve als Polygonzug nur eine Näherung an die "wahre" stetige Lorenzkurve sein kann und dass im ersten Fall D_G kleiner ist als im zweiten (denn 8/27 = 0,2963 < 0,4), so dass generell gilt:

  (6.44) $\phi^* \leq D_G$.

Kapitel 7: Zweidimensionale Häufigkeitsverteilungen

1. Regression und Korrelation

Gegenstand der in den Kapiteln 7 und 8 dargestellten Methoden ist der Zusammenhang von zwei (oder mehr) Merkmalen. Dabei sind im wesentlichen folgende Fragestellungen üblich:

- Wie lässt sich der Grad (die Intensität) des Zusammenhangs zwischen zwei (oder mehr) als wechselseitig abhängig (interdependent) angenommenen Merkmalen (bzw. Variablen) messen? Je nach Art der vorliegenden Skalen für X und Y spricht man von Korrelations-, Rangkorrelations-, Kontingenz- oder Assoziationsanalyse oder auch von "Korrelation" im allgemeinen Sinne.
- Besteht ein Zusammenhang zwischen zwei oder mehr Merkmalen dergestalt, dass es möglich ist Y aufgrund von X oder Y aufgrund X_1 und X_2 zu schätzen, d.h. eine Funktion für Y in Abhängigkeit anderer Variablen aus den Daten zu schätzen?

Es ist üblich, die erste Fragestellung als typisch für die Korrelationsanalyse anzusehen, während es bei der Regressionsanalyse um die zweite Fragestellung geht. Es gilt in erster Näherung:

Korrelation = Analyse der Stärke der Interdependenz (wechselseitigen Abhängigkeit) und **Regression** = Analyse der Art der Dependenz (Abhängigkeit einer Variablen von anderen Variablen).

Für eine erste Näherung mag diese Unterscheidung zwischen Regressions- und Korrelationsanalyse, zwei Methoden, die meist in einem Atemzug genannt werden, ausreichen. Bei genauerer Betrachtung (die zunächst noch zurückzustellen ist) zeigt sich, dass der Unterschied zwischen diesen beiden Methoden jedoch weniger in der Zielsetzung liegt, als in den Voraussetzungen, die hinsichtlich der Variablen getroffen werden. Die Unterscheidung zwischen Interdependenz und Dependenz ist im übrigen keineswegs klar. Sie ist deshalb nicht befriedigend. Es ist insbesondere ziemlich abwegig alle multivariaten Verfahren im Sinne dieser Unterscheidung klassifizieren zu wollen, was in manchen Lehrbüchern geschieht.

Die Bestimmung einer Funktion zur Beschreibung des Zusammenhangs zwischen Variablen, bzw. genauer, zur Schätzung einer Variablen Y, aufgrund ihrer Abhängigkeit von anderen Variablen ist Gegenstand der Regressionsanalyse. Eine "abhängige" Variable Y (auch Regressand genannt) wird durch eine oder mehrere "unabhängige" Variablen "erklärt". Einfache Regression bedeutet Y in Abhängigkeit einer unabhängigen Variable (eines Regressors) X zu schätzen. Wird Y durch eine Funktion mehrerer Regressoren X_1, X_2,...,X_p erklärt, so spricht man von multipler Regression. Hinter solchen Betrachtungen kann (muss aber nicht) eine Kausalvorstellung stehen.

Voraussetzung sowohl der Regressions- als auch Korrelationsanalyse ist die Beschreibung eines zweidimensionalen (bivariaten) Datensatzes. In diesem Kapitel soll deshalb zunächst gezeigt werden, wie bi- oder allgemeiner multivariate Datensätze geeignet tabellarisch und (wenn möglich) auch graphisch dargestellt werden können. Es werden dann Maße des Zusammenhangs, also Korrelationskoeffizienten vorgestellt und interpretiert.

2. Darstellung mehrdimensionaler Datensätze

a) Verbundene Beobachtungen, gemeinsame Verteilung

Hinsichtlich der Art der Daten soll im folgenden vorausgesetzt werden:

1. es liegen verbundene Beobachtungen von mehreren Merkmalen vor,
2. diese Merkmale sind metrisch skaliert (mindestens Intervallskala),
3. die Daten können Einzelbeobachtungen, gruppierte oder klassierte Merkmale sein.

Erläuterungen:

zu 1: Was mit verbundenen Beobachtungen gemeint ist, soll durch das einführende Beispiel (Bsp. 7.1) veranschaulicht werden. Im folgenden beschränken wir uns auf zwei Merkmale.

zu 2: Es ist hinsichtlich der Art (Skalen) der Merkmale zu unterscheiden zwischen folgenden Fällen (bei Beschränkung auf zwei Merkmale):

a) zwei auf gleichem Skalenniveau skalierte Merkmale, etwa beide intervallskaliert oder beide nominalskaliert, wie im Bsp. 7.1;

b) Merkmale auf unterschiedlichem Skalenniveau, etwa X nominalskaliert und Y intervallskaliert.

Es soll im folgenden zunächst der Fall a) betrachtet werden.

zu 3: Die Methoden dieses Kapitels werden (im Unterschied zu denen des Kapitels 8) in der Regel nicht für den Fall von Einzelbeobachtungen demonstriert.

Beispiel 7.1:

Bei drei Klausuren A, B und C wurde der Zusammenhang zwischen Geschlecht (Merkmal X) und Klausurleistung (Merkmal Y) untersucht. Es gab jeweils 200 Klausurteilnehmer, darunter 150 Männer und 50 Frauen und jede der Klausuren wurde von 70% der Teilnehmer bestanden (also von 140 Personen) und von 30% nicht bestanden. Man erhielt die folgenden Daten:

x_1 = männlich x_2 = weiblich y_1 = bestanden y_2 = nicht bestanden.

Klausur A

	y_1	y_2	Σ
x_1	105	45	150
x_2	35	15	50
Σ	140	60	200

Klausur B

	y_1	y_2	Σ
x_1	140	10	150
x_2	0	50	50
Σ	140	60	200

Klausur C

	y_1	y_2	Σ
x_1	90	60	150
x_2	50	0	50
Σ	140	60	200

a) Erläutern Sie die Art der Zusammenstellung der Daten und interpretieren Sie die Daten.

b) Worin besteht der Unterschied zwischen verbundenen und unverbundenen Beobachtungen?

Lösung 7.1:

a) Es handelt sich um Vierfeldertafeln, jeweils eine spezielle Form der zweidimensionalen Häufigkeitsverteilung (Kontingenztabelle). Die

Summenzeilen und -spalten stellen die Randverteilungen der Variablen X und Y dar (vgl. Def. 7.3).

b) Man sieht, dass Art und Stärke des Zusammenhangs in den drei Klausuren sehr unterschiedlich sind, obgleich die Randverteilungen gleich sind: das Beispiel soll v.a. zeigen, dass die Kenntnis der (eindimensionalen) Randverteilungen nicht ausreicht, um einen Zusammenhang zu beurteilen. Die Unterschiede der drei Klausuren werden deutlich, wenn man die "Durchfallquoten" von Männern und Frauen bei den drei Klausuren betrachtet (d.h. praktisch die bedingten Verteilungen - vgl. Def. 7.4 - des Merkmals Y untersucht):

	Anteil nicht-bestandener Klausuren in vH		
Klausur	Männer	Frauen	insgesamt
A	30%(45/150)	30%(=15/50)	30%(=60/200)
B	6,7%	100%	30%
C	40%	0%	30%

Im Falle der Klausur A besteht Unabhängigkeit der beiden Variablen X (Geschlecht) und Y (Klausurleistung)(vgl. Def.7.5).

Def. 7.1: Verbundene Beobachtungen

a) im Falle von Einzelbeobachtungen:
Wird jede Einheit v =1,2,...,n mit zwei Merkmalen, d.h. einem Tupel (x_v,y_v), mit drei Merkmalen [einem Tripel (x_v,y_v,z_v)] oder mit p Merkmalen (p-Tupel) beschrieben, so spricht man von verbundenen Beobachtungen (im Rahmen einer zwei-, drei-,..., p-dimensionalen Messung) [im folgenden Beschränkung auf p = 2 Dimensionen].

b) bei gruppierten Daten:
Das Merkmal X habe die Ausprägungen $x_1,x_3,...,x_m$ oder allgemein x_i (i=1,2,...,m) und das Merkmal Y habe die Ausprägungen y_j (j = 1,2,...,k). Dann ist n_{ij} die Anzahl der Einheiten mit den Ausprägungen $X = x_i$ **und** $Y = y_j$ (also die Anzahl gleicher Wertetupel). Wie im Falle der eindimensionalen Häufigkeitsverteilung n(...) eine Funktion ist, die einer Merkmalsausprägung eine absolute Häufigkeit zuordnet, so soll n(...) hier einer **Kombination** von Merkmalsausprägungen eine absolute Häufigkeit zuordnen:

(7.1) $n_{ij} = n(X = x_i \text{ und } Y = y_j)$ $(i = 1,...,m \text{ und } j=1,...,k)$.

Für die relativen Häufigkeiten gilt analog zur eindimensionalen Häufigkeitsverteilung

(7.2) $h_{ij} = n_{ij}/n \quad \text{mit} \quad n = \sum_i \sum_j n_{ij} = \sum_{i,j} n_{ij}$

c) bei klassierten Daten gilt b) analog.

Bemerkungen zu Def. 7.1:

1. Auf die in Gl. 7.2 erscheinende Doppelsumme wird in Def. 7.2 näher eingegangen. Für die mk (m Ausprägungen von X und k Ausprägungen von Y) relativen Häufigkeiten h_{ij} gilt analog zum eindimensionalen Datensatz
   $0 \le h_{ij} \le 1 \quad \text{und} \quad \Sigma\Sigma h_{ij} = 1.$
2. Ein erstes Beispiel für den Fall gruppierter Daten, wenn also gleiche Merkmalsausprägungen von X und Y gehäuft auftreten, ist das Beispiel 7.1, das im übrigen demonstriert, dass die (simultane) Betrachtung zweidimensionaler Messungen (Beobachtungen) nicht identisch ist mit der isolierten Betrachtung zweier eindimensionaler Messungen.
3. Es gibt folgende Darstellungen verbundener Beobachtungen:

	graphisch *)	tabellarisch
Einzelbeobachtungen	Streuungsdiagramm [oder Streudiagramm] (vgl. Beispiel 7.2)	Urliste von n Tupeln (Wertepaaren)
gruppierte und klassierte Daten	dreidimensionales Histogramm **)	Kontingenztafeln (vgl. Def. 7.2)

*) metrische Skala vorausgesetzt.

**) für eine zweidimensionale Häufigkeitsverteilung (je eine Achse für die Variablen X und Y und eine Achse für die Häufigkeiten).

Def. 7.2: Zweidimensionale Häufigkeitsverteilung

Eine zweidimensionale Häufigkeitsverteilung ist eine Zuordnung der gemeinsamen absoluten (n_{ij}) oder relativen (h_{ij}) Häufigkeiten zu den Ausprägungen x_i des Merkmals (der Variablen) X und y_j des Merkmals (der Variablen) Y nach Art nachfolgender Tabelle (Matrix). Bei kategorialen (nominalskalierten) Merkmalen spricht man auch von einer Kontingenztafel.

Zweidimensionale Häufigkeitsverteilung
(relative Häufigkeiten)

Merkmal X	Merkmal Y					
	y_1	y_2	...	y_j	...	y_k
x_1	h_{11}	h_{12}	...	h_{ij}	...	h_{1k}
x_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2k}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{ik}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_m	h_{m1}	h_{m2}	...	h_{mj}	...	h_{mk}

Zum Sprachgebrauch:

Der Begriff Kontingenztafel wird von vielen Autoren auch bei metrisch skalierten Variablen benutzt. Die absoluten oder relativen Häufigkeiten heißen auch gemeinsame Häufigkeiten und die gesamte Häufigkeitsverteilung auch **gemeinsame Häufigkeitsverteilung.** Die Größen x_i (i=1,2,...,m), bzw. y_j (j=1,2,...,k) können Merkmalsausprägungen (gruppierte Daten) oder Größenklassen (klassierte Daten) der Merkmale X und Y bezeichnen.

Bemerkungen zu Def. 7.2:

1. In der gleichen Art, wie die gemeinsamen relativen Häufigkeiten h_{ij} dargestellt wurden, lassen sich auch die (gemeinsamen) absoluten Häufigkeiten n_{ij} und die (gemeinsamen) relativen oder absoluten Summenhäufigkeiten (H_{ij}, bzw. N_{ij}) darstellen, wobei gilt:

$$(7.3) \qquad N_{ij} = \sum_{x \le i}^{n} \sum_{y \le j}^{n} n_{xy} \quad \text{und} \quad H_{ij} = N_{ij}/n.$$

2. Jeder mehrdimensionalen Häufigkeitsverteilung sind weitere Verteilungen zugeordnet (d.h. sie ergeben sich hieraus), nämlich die Randverteilungen und die bedingten Verteilungen (vgl. die nachfolgenden Definitionen).

Anhand des folgenden Beispiels 7.2 soll das Streuungsdiagramm (oder Streudiagramm, scatter diagram) erklärt werden.

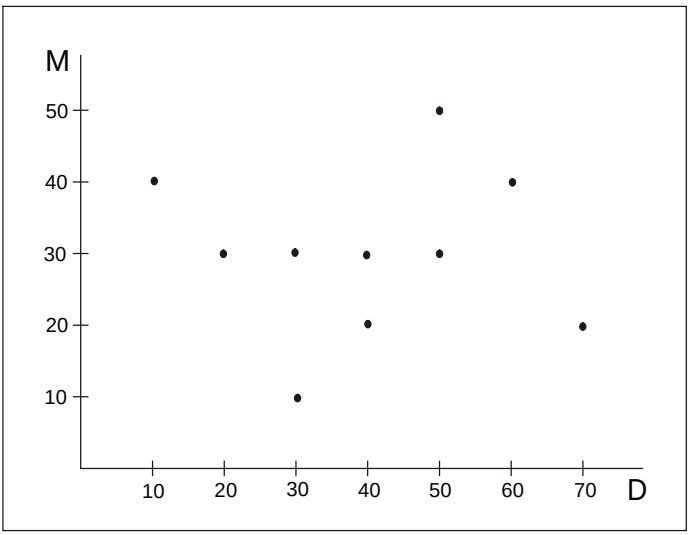
Beispiel 7.2: Streuungsdiagramm

König Egon der XIII, auch der "Labile" genannt, hatte zwei Mätressen, die Pompadur (D) und die Pompamoll (M), die miteinander heftig um die Gunst des Königs konkurrierten. Dass sie jeweils verschiedene Seiten des empfindsamen Gemüts des Königs ansprachen und für ihn deshalb komplementär waren, steht seit der These des berühmten Historikers H in allen Lehrbüchern. H's jüngerer Kollege h glaubt dies jedoch aufgrund einer seinerzeit von der Hofschranze S verfassten Notiz empirisch widerlegen zu können. Aus dieser Notiz geht hervor, wie Egon seine Freizeit (gemessen in Stunden) in den letzten 10 Wochen des Jahres 1789 auf die Damen aufteilte:

D	40	30	20	10	40	30	50	50	60	70
M	30	10	30	40	20	30	50	30	40	20

Zeichnen Sie das Streuungsdiagramm! (Die Aufgabe wird fortgesetzt.)

Abb.7.1: Streuungsdiagramm für das Beispiel 7.2



Lösung 7.2: vgl. Abb. 7.1

Abb. 7.1 zeigt das Streuungsdiagramm für dieses Beispiel. Es wird später gezeigt, dass die beiden Variablen D und M nicht miteinander korrelieren und deshalb die These von H, derzufolge mit einer hohen positiven Korrelation zu rechnen ist, vermutlich nicht aufrechtzuerhalten ist.

b) Aus der gemeinsamen Verteilung abgeleitete Verteilungen

Jeder gemeinsamen Verteilung sind zwei weitere Verteilungstypen zugeordnet, in dem Sinne, dass sie aus der gemeinsamen Verteilung hergeleitet sind:

1. Eine p-dimensionale Verteilung besteht aus einer gemeinsamen Verteilung und p jeweils (p-1)-dimensionalen Randverteilungen (Def. 7.3); im Falle von zwei Dimensionen (p=2) sind es also zwei eindimensionale Randverteilungen.
2. Es gibt ferner jeweils eindimensionale bedingte Verteilungen (Def. 7.4): Jeder Zeile und jeder Spalte der gemeinsamen Verteilung entspricht jeweils eine bedingte Verteilung (die stets eindimensional ist).

Def. 7.3: Randverteilungen, marginal distributions

Da die Ausprägung x_i bei den Kombinationen (x_i,y_1), (x_i,y_2),..., (x_i,y_k) also allen Merkmalskombinationen der i-ten Zeile der zweidimensionalen Häufigkeitsverteilung (Kontingenztabelle) vorliegt, ist die Randhäufigkeit $h_{i.}$ definiert als Zeilensumme

$$(7.4) \quad h_{i.} = \sum_{j=1}^{j=k} h_{ij} = \sum_{j} h_{ij} = h(X=x_i).$$

Die als Summen von Zeilen gebildeten Randhäufigkeiten $h_{1.}$, $h_{2.}$,...,$h_{m.}$ stellen die Randverteilung $h_x(x)$ der Variablen X dar.
Entsprechend bilden die als Summen von Spalten definierten Randhäufigkeiten $h_{.1}$,$h_{.2}$,...,$h_{.k}$ die Randverteilung $h_y(y)$ des Merkmals (der Variablen) Y, wobei gilt:

$$(7.5) \quad h_{.j} = \sum_{i=1}^{i=m} h_{ij} = \sum_{i} h_{ij} = h(Y=y_j).$$

Die Randverteilungen ausgedrückt in absoluten Häufigkeiten $n_x(x)$ mit den über k Spalten summierten absoluten Häufigkeiten einer Zeile

$$(7.4a) \quad n_{i.} = n_{i1} + n_{i2} + ... + n_{ik}$$

und die Randverteilung $n_y(y)$ mit den k absoluten Häufigkeiten $n_{.j}$ sind entsprechend definiert.

Die beiden Randverteilungen (in relativen Häufigkeiten) sind in der folgenden Tabelle besonders durch Einrahmung markiert:

Merkmal X	Merkmal Y: y_1	y_2	...	y_j	...	y_k	$\Sigma: h_x(x)$
x_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1k}	$h_{1.}$
x_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2k}	$h_{2.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{ik}	$h_{i.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_m	h_{m1}	h_{m2}	...	h_{mj}	...	h_{mk}	$h_{m.}$
$\Sigma: h_y(y)$	$h_{.1}$	$h_{.2}$	...	$h_{.j}$	...	$h_{.k}$	1

Die Summenspalte $h_x(x)$ ist die Randverteilung von X und die Summenzeile $h_y(y)$ ist die Randverteilung von Y.

Auch bei den Randverteilungen kann man - wie bei jeder Verteilung - unterscheiden zwischen der Häufigkeitsfunktion (-verteilung) und der **Verteilungsfunktion (Summenverteilung, kumulierte Häufigkeiten**) und das jeweils bei absoluten und bei relativen Häufigkeiten. Es ist ferner offensichtlich, dass alle Kennzahlen eindimensionaler Häufigkeitsverteilungen (z.B. Mittelwerte, Streuungsmaße, Schiefemaße usw.) auch für die Randverteilungen berechnet werden können (vgl. Abschn. 3).

Def. 7.4: bedingte Verteilung

Die durch Gl. 7.6 definierten bedingten relativen Häufigkeiten $h_{i\,|\,j}$ stellen die bedingte Häufigkeitsfunktion (-verteilung) von X, gegeben $Y = y_j$ dar, also die Spalte j als eine Verteilung:

$$(7.6) \quad h_{i\,|\,j} = \frac{h_{ij}}{h_{.j}} = \frac{n_{ij}}{n_{.j}} = h(x|Y=y_j) \; .$$

Analog ist die bedingte Häufigkeitsfunktion (-verteilung) von Y definiert durch die relativen Häufigkeiten der Ausprägung $y_1, y_2, ..., y_k$ (allgemein: y_j) "gegeben $X = x_i$" (oder: bedingt durch x_i, oder: wenn $X = x_i$)

$$(7.7) \quad h_{j\,|\,i} = \frac{h_{ij}}{h_{.i}} = \frac{n_{ij}}{n_{i.}} = h(y|X=x_i) \; .$$

Es ist eine Zeile (die i-te) als Häufigkeitsverteilung.

Bemerkungen zu Def. 7.4:

1. Es ist in diesem Fall nur üblich, von *relativen* (nicht absoluten) Häufigkeiten auszugehen. Es gilt, wie leicht zu zeigen ist:

$$\text{(7.6a)} \quad \sum_{i=1}^{i=m} h_{i\,|\,j} = 1 \qquad \text{und} \qquad \text{(7.7a)} \quad \sum_{j=1}^{j=k} h_{j\,|\,i} = 1 \quad .$$

2. Mit Gl. 7.6 werden aus den k Spalten (nach den Ausprägungen des Merkmals Y, j=1,2,...,k) und mit Gl. 7.7 aus den m Zeilen (i=1,2,...,m) der Matrix der gemeinsamen Verteilung jeweils Häufigkeitsverteilungen gebildet.

Def. 7.5: Unabhängigkeit

Unabhängigkeit lässt sich auf zwei Arten definieren:

1. Sind die k bedingten Verteilungen $h_{i\,|\,j}$ des Merkmals X bei allen Ausprägungen y_j (j = 1,2,...,k) des Merkmals Y identisch (und damit auch gleich der [unbedingten] Randverteilung von X), so sind X und Y unabhängig (analog gilt: Gleichheit der m bedingten Verteilungen $h_{j\,|\,i}$ bei allen Ausprägungen von X des Merkmals Y, bedeutet Unabhängigkeit von X und Y).
2. Im Falle der Unabhängigkeit ergeben sich die absoluten, bzw. relativen gemeinsamen Häufigkeiten aus den entsprechenden Häufigkeiten der Randverteilungen gemäß

$$\text{(7.8)} \quad n_{ij} = \frac{n_{i.}\, n_{.j}}{n} \qquad \text{bzw.} \qquad \text{(7.8a)} \quad h_{ij} = h_{i.} h_{.j} \quad .$$

Folgerungen:

1. Unabhängigkeit ist die stärkere Forderung als die später zu besprechende (vgl. Def. 7.7) Unkorreliertheit:

Satz 7.1:

Unabhängigkeit impliziert Unkorreliertheit aber nicht umgekehrt, d.h. Unkorreliertheit kann bestehen, obgleich die Variablen X und Y nicht unabhängig sind.

Beweis: siehe Bem. Nr.9 zu Def. 7.6.

2. Aus Teil 1 der Def. 7.5 folgt, dass bei Unabhängigkeit jeweils alle bedingten Verteilungen mit der Randverteilung identisch sind. Gilt Identität der k bedingten Verteilungen des Merkmals X, so sind auch die bedingten Verteilungen des Merkmals Y

für alle m Ausprägungen von X identisch, d.h. Unabhängigkeit ist eine symmetrische Relation: ist X unabhängig von Y, so ist auch Y unabhängig von X.

3. Beide Arten der Definition der Unabhängigkeit sind äquivalent, d.h. eine Art der Definition lässt sich jeweils aus der anderen folgern. Aus Folgerung 2 ergibt sich $n_{ij} / n_{i.} = n_{.j} / n$ und hieraus folgt Gl. 7.8.

4. Umgekehrt gilt: Aus Gl. 7.8a folgt

$$h_{i \mid j} = \frac{h_{ij}}{h_{.j}} = \frac{h_{i.}\, h_{.j}}{h_{.j}} = \frac{h_{ik}}{h_{.k}} = h_{i \mid k} = \frac{h_{i.}\, h_{.k}}{h_{.k}} = h_{i.} ,$$

d.h. bei Unabhängigkeit ist Gleichheit der bedingten Verteilungen von X gegeben (X bedingt durch $Y = y_j$ und X bedingt durch $Y = y_k$).

Beispiel 7.3:

Einer fehlgeschlagenen Intrige bei Hofe hat es Graf Giselher von Gelsenkirchen zu verdanken, dass er in einem Burgverlies schmachtet. Statt vor dem Verwaltungsgericht Gelsenkirchen zu klagen, (diese neuzeitliche Denkweise war Giselher noch vollkommen fremd) machte er sich daran, die meterdicke Wand zu durchbohren. Es gibt Tage, an denen er y=1, y=2 und y=3 Zentimeter der Wand wegschaben konnte. Über den Zeitaufwand X des Schabens (in Stunden) und die Zentimeterleistung Y des Verdünnens der Wand bestehen für 13 Tage Aufzeichnungen des Grafen:

X: Arbeitszeit	Y: Leistung 1cm	2cm	3cm
6Std.	1	2	0
8Std.	1	3	1
10Std.	1	2	2

Man bestimme sowohl die beiden Randverteilungen als auch die bedingten Verteilungen sowie die absoluten gemeinsamen Summenhäufigkeiten. Sind die beiden Merkmale X und Y unabhängig?

Lösung 7.3:

Randverteilungen			
x_i	$n_{i.}$	y_j	$n_{.j}$
6	3	1	3
8	5	2	7
10	5	3	3

Bedingte Verteilungen							
	y=1	**y=2**	**y=3**		**x=6**	**x=8**	**x=10**
x=6	1/3	2/7	0	**y=1**	1/3	0,2	0,2
x=8	1/3	3/7	1/3	**y=2**	2/3	0,6	0,4
x=10	1/3	2/7	2/3	**y=3**	0	0,2	0,4
von x (bedingt durch y)				von y (bedingt durch x)			

X und Y sind nicht unabhängig. Es genügt, ein Tabellenfeld der gemeinsamen Verteilung zu überprüfen. Nach Gl. 7.8 müsste für n_{12} gelten $n_{12} = n_{1.}\, n_{.2} /n = (3 \cdot 7)/13 = 1{,}615$ statt $n_{12} = 2$. Auch kein anderes Feld der gemeinsamen Verteilung erfüllt Gl. 7.8. So ist etwa $n_{22} = 3$ statt $(5 \cdot 7)/13 = 2{,}692$. Dass keine Unabhängigkeit gegeben ist, kann man auch daran erkennen, dass die bedingten Verteilungen verschieden sind. In der Aufgabe war auch verlangt, die absoluten Summenhäufigkeiten zu bestimmen. Die Größe N_{ij} ist die absolute Häufigkeit für $x \le x_i$ und $y \le y_j$

	y=1 cm	y=2 cm	y=3 cm
x=6 Std.	1	3	3
x=8 Std.	2	7	8
x=10 Std.	3	10	13

Die Häufigkeit 7 in dieser Tabelle bedeutet, dass es 7 Tage gibt, an denen Giselher bis zu höchstens 8 Stunden arbeitet (also 6 oder 8 Stunden) und dabei bis zu höchstens 2 cm der Wand abschabt.

3. Kennzahlen zur Beschreibung einer zweidimensionalen Verteilung

Einer gemeinsamen Verteilung h(x,y) der (mindestens intervallskalierten Variablen X und Y) sind jeweils eindimensionale Randverteilungen und bedingte Verteilungen zugeordnet. Jede der genannten Verteilungen lässt sich durch Kenngrößen (Maßzahlen, Parameter) beschreiben, die

- eindimensionalen Verteilungen (Randverteilungen, bedingte Verteilungen) durch Mittelwerte (auch Mediane), Varianzen etc.;
- zweidimensionale gemeinsame Verteilung durch die Kovarianz und den Korrelationskoeffizienten.

a) Kennzahlen der eindimensionalen Verteilungen und Regressionslinie

Übersicht 7.1 stellt die Zusammenhänge zwischen den Verteilungen und den im folgenden zu behandelnden beschreibenden Kennzahlen (Parameter) dar.
Es sollen zunächst die Parameter der eindimensionalen Verteilungen betrachtet werden (Nr. 1 und 2) und im nächsten Abschnitt die der gemeinsamen Verteilung:

1) Mittelwert und Varianz der Randverteilungen

Mittelwert $\bar{x}$ der Randverteilung $h_x(x)$

(7.9) $\bar{x} = \Sigma x_i h_{i.} = \sum_i x_i h_{ij}$

und die Varianz

(7.10) $s_x^2 = \Sigma x_i^2 h_{i.} - \bar{x}^2$.

Die entsprechenden Parameter der Randverteilung $h_y(y)$ sind analog definiert.

2) Parameter der bedingten Verteilungen

a) Die wichtigsten Parameter der bedingten (Häufigkeits-) Verteilungen sind die **bedingten Mittelwerte**

(7.11) $\bar{x}|y = \bar{x}(y_j) = \sum_{i=1}^{i=m} x_i h_{i|j}$

(7.12) $\bar{y}|x = \bar{y}(x_i) = \sum_{j=1}^{j=k} y_j h_{j|i}$.

b) Seltener ist die Berechnung der **bedingten Varianzen**

(7.12a) $s_x^2(y_j) = \sum_{i=1}^{i=m} [x_i - \bar{x}(y_j)]^2 h_{i|j} = \sum_{i=1}^{i=m} x_i^2 h_{i|j} - [\bar{x}(y_j)]^2$

und $s_y^2(x_i)$ analog, bzw. der bedingten Standardabweichungen $s_x(y_j)$ und
$s_y(x_i)$.

Mit den bedingten Varianzen wird die Streuung der Beobachtungen um die Regressionslinie (vgl. Def. 7.6) gemessen. Sie ist Teil der internen Streuung in einer Varianzzerlegung (vgl. Satz 7.4).

Übersicht 7.1

a) Zusammenhänge zwischen den Verteilungen

Ausgangspunkt: **Zweidimensionale** gemeinsame Verteilung, d.h.
eine Matrix mit m Zeilen für die Ausprägungen
x_i (i = 1,,2,...,m) und k Spalten für y_j (j = 1,2,...,k)
daraus abgeleitete **eindimensionale** Verteilungen

zwei Randverteilungen für X und Y: $h_x(x)$ mit den Häufigkeiten $h_{i.}$ und $h_y(y)$ mit den Häufigkeiten $h_{.j}$	m bedingte Verteilungen von Y (bedingt durch m Ausprägungen von X) und k bedingte Verteilungen von X (bedingt durch Y)

b) Beschreibende Kennzahlen* **der**

zweidimensionalen Verteilung	**eindimensionalen Verteilungen**	
Kovarianz s_{xy} und Korrelationskoeffizient r_{xy}	Randverteilungen: Mittelwerte und Varianzen	bedingte Verteilungen: bedingte Mittelwerte (Regressionslinie)

*) Nur die am häufigsten verwendeten Maßzahlen (Kennzahlen). Man kann natürlich auch z.B. die Schiefe der Randverteilungen bestimmen oder bedingte Varianzen berechnen.

Def. 7.6: empirische Regressionslinie

Die lineare Verbindung der bedingten Mittelwerte $\bar{x}|y$ ist die Regressionslinie (empirische Regressionslinie) der Variablen X. Entsprechend ist die lineare Verbindung der Punkte $P(x, \bar{y}|x)$ die Regressionslinie der Variablen Y.

Der Begriff Regressions"linie" soll deutlich machen, dass die Punkte nicht notwendig auf einer Geraden liegen müssen. Es sind also Regressionslinie und Regressionsgerade (Kap. 8) zu unterscheiden.

Selbst wenn die Regressionslinie eine Gerade ist, muss sie nicht identisch mit der Regressionsgeraden sein.

Beispiel 7.4:

Man bestimme und zeichne die Regressionslinien für das Bsp. 7.3!

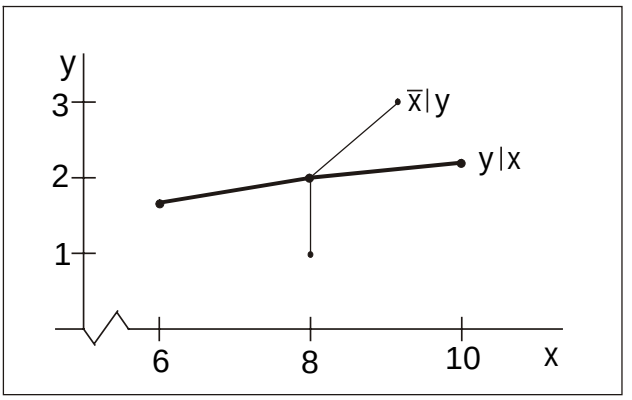
Lösung 7.4:

Bedingte Mittelwerte

von y:		von x:	
$\bar{y}\|x=6$	$= 1{,}67$	$\bar{x}\|y=1$	$= 8$
$\bar{y}\|x=8$	$= 2$	$\bar{x}\|y=2$	$= 8$
$\bar{y}\|x=10$	$= 2{,}2$	$\bar{x}\|y=3$	$= 9{,}33$

Zur graphischen Darstellung der Regressionslinien vgl. Abb. 7.2

Abb. 7.2: Regressionslinien (Beispiel 7.4 bzw. 7.3)



Beispiel 7.5:

Der Student S glaubt wieder einmal eine Recht-Klausur ganz astrein gelöst zu haben. Mit seiner Selbsteinschätzung (Variable X), die mehr oder weniger gefühlsmäßig und zufällig, weniger aus tiefer juristischer Einsicht erfolgt, liegt er zwar oft in der Tendenz ganz richtig. Die genaue Klausurnote (Y) erscheint ihm aber fast immer rätselhaft und unerklärlich. So wie es ihm geht, ergeht es jedoch auch seinen 35 Mitstudenten. Dass die Noten bei den Rechtklausuren irgendwie mysteriös sind, scheinen inzwischen fast alle zu glauben, wie die folgende Gegenüberstellung von X und Y für alle 36 Studenten zeigt:

		Variable Y				
		1	2	3	4	5
Variable X	1	1	2	3	2	0
	2	1	2	2	1	0
	3	0	1	2	2	1
	4	0	0	3	4	3
	5	0	2	1	1	2

Bestimmen Sie die empirischen Regressionslinien!

Lösung 7.5:

$(\overline{y}|x=1) = 2{,}75$ $(\overline{x}|y=1) = 1{,}5$
$(\overline{y}|x=2) = 2{,}5$ $(\overline{x}|y=2) = 2{,}714 = 19/7$
$(\overline{y}|x=3) = 3{,}5$ $(\overline{x}|y=3) = 2{,}727 = 30/11$
$(\overline{y}|x=4) = 4$ $(\overline{x}|y=4) = 3{,}1$
$(\overline{y}|x=5) = 3{,}5$ $(\overline{x}|y=5) = 4{,}167 = 25/6$

Beispiel 7.6:

Gegeben sei die folgende zweidimensionale Häufigkeitsverteilung (Angabe mit relativen Häufigkeiten) für welche die Regressionslinien sowie die Parameter der Randverteilungen zu bestimmen sind:

	y=2	y=3	y=4
x=2	0,2	0,5	0,1
x=3	0,1	0,1	0

Lösung 7.6:

Mittelwerte $\overline{x} = 2{,}2$ und $\overline{y} = 2{,}8$; Varianzen: $s_x^2 = 0{,}16$ und $s_y^2 = 0{,}36$. Die bedingten Mittelwerte von x lauten $\overline{x}|y=2 = 2{,}33$, $\overline{x}|y=3 = 2{,}167$ und $\overline{x}|y=4 = 2$. Diejenigen von y lauten $\overline{y}|x=2 = 2{,}875$ und $\overline{y}|x=3 = 2{,}5$. Beide Regressions**linien** sind Geraden (sie sind übrigens, wie später gezeigt wird, identisch mit den Regressions**geraden**).

b) Kovarianz und Korrelationskoeffizient

Die Kovarianz ist Ausdruck des Zusammenhangs zwischen zwei metrisch skalierten Variablen. Als Maß für den Grad (die Intensität) des Zusammenhangs ist sie im Unterschied zur Korrelation jedoch nicht geeignet, weil ihr Betrag von der Maßeinheit der Variablen X und Y abhängt. Über Kovarianz und Korrelation im Falle einer klassierten Verteilung vgl. Abschn. 3e dieses Kapitels.

1. Kovarianz

Def. 7.7: Kovarianz

Die Kovarianz ist als beschreibende Kennzahl einer zweidimensionalen Verteilung definiert als

$$(7.13)\quad s_{xy} = \frac{1}{n} \Sigma (x_v - \bar{x})(y_v - \bar{y}) \quad \text{bei n Einzelbeobachtungen}$$

bzw. bei gruppierten Daten mit absoluten gemeinsamen Häufigkeiten

$$(7.14)\quad s_{xy} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_i - \bar{x})(y_j - \bar{y}) n_{ij}$$

und mit relativen Häufigkeiten

$$(7.14a)\quad s_{xy} = \Sigma\Sigma (x_i - \bar{x})(y_j - \bar{y}) h_{ij}.$$

Zur Interpretation der Kovarianz (Bemerkungen zu Def. 7.7)

1. a) Die Kovarianz heisst auch zentrales **Produktmoment**: "zentral", weil die Mittelwerte $\bar{x}$ und $\bar{y}$ jeweils abgezogen werden von x_v bzw. y_v und "Produkt", weil diese Abweichungen miteinander multipliziert werden. Starke Abweichungen vom jeweiligen Mittelwert beeinflussen die Kovarianz stark, so dass diese empfindlich ist gegenüber Ausreißern.

   b) Ähnlich wie bei der Varianz kann auch hier unterschieden werden zwischen der Berechnung der Kovarianz

   - als beschreibende Statistik einer Stichprobe gem. Gl. 7.13, bzw. 7.14 und
   - als Schätzwert für die Kovarianz der Grundgesamtheit aufgrund der Stichprobe: dann ist durch n-1 statt durch n zu dividieren.

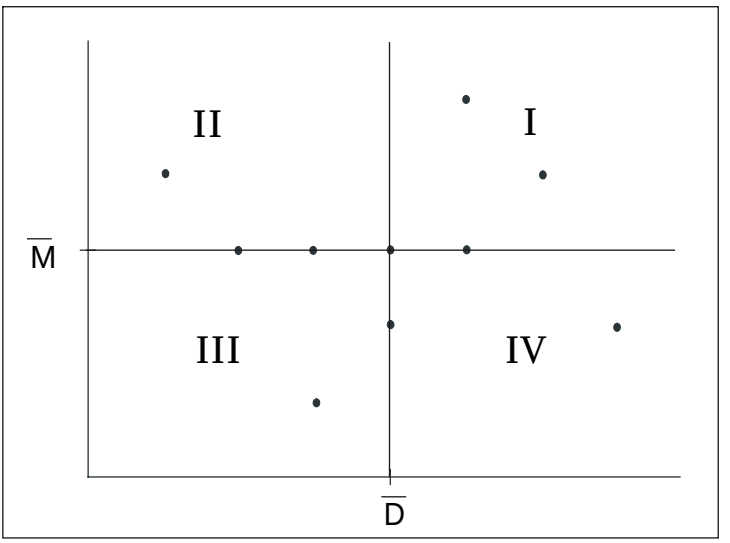
2. Es ist unmittelbar zu sehen, dass die Varianz als Spezialfall der Kovarianz aufgefaßt werden kann $s_{xx} = s_x^2$. Die Kovarianz ist außerdem symmetrisch, d.h. es gilt $s_{xy} = s_{yx}$.

3. Weil ein Produkt von Abweichungen gebildet wird, kann die Kovarianz positiv oder negativ sein. Es ist üblich, sich die Bedeutung der Mittelung über Abweichungsprodukte bei der Kovarianz anhand der Abb. 7.3 zu verdeutlichen.

Daten von Beispiel 7.2: D = Pompadur und M = Pompamoll; die arithmetischen Mittel sind $\overline{D} = 40$ und $\overline{M} = 30$. Das Streuungsdiagramm zeigt Unkorreliertheit der Variablen an.

- Bei Beobachtungen im Quadrant I sind die Abweichungen $(x_v-\overline{x})$ und $(y_v-\overline{y})$ **beide** jeweils positiv, so dass $(x_v-\overline{x})(y_v-\overline{y})>0$;
- Quadrant III $(x_v-\overline{x})(y_v-\overline{y}) > 0$ (weil beide Abweichungen negativ sind);
- Gegenläufig verhalten sich dagegen X und Y (d.h. negative Korrelation) in den Quadranten II und IV, weshalb dann das Produkt $(x_v-\overline{x})(y_v-\overline{y}) < 0$ also negativ ist.

Abb. 7.3: Streuungsdiagramm (Bsp. 7.2) zur Veranschaulichung des Konzepts der Kovarianz



Liegen die Punkte im Streuungsdiagramm hauptsächlich in den Quadranten I und III, so liegt eine positive Korrelation vor, liegen sie vorwiegend in den Quadranten II und IV, so ist die Korrelation negativ. Man kann ein Korrelationsmaß definieren, indem man einfach die Anzahl n_k (k = konkordant) der Punkte im Quadranten I und III mit der Anzahl n_d (d=diskonkordant) der Punkte im Quadrant II und IV vergleicht: **Fechners Korrelationskoeffizient** $r_F = (n_k-n_d)/(n_k+n_d)$.

4. Die Kovarianz ist betragsmäßig nicht beschränkt. Sie ist deshalb nicht geeignet für die Messung des Zusammenhangs zwischen zwei Vari-

ablen X und Y. Der Korrelationskoeffizient (vgl. Def. 7.8) ist dagegen die auf den Wertebereich von -1 bis +1 normierte Kovarianz.

5. Die Kovarianz ist nicht invariant gegenüber linearen Transformationen. Man kann leicht zeigen, dass die Kovarianz der transformierten Variablen $x^* = a + bx$ und $y^* = c + dy$ wie folgt mit der Kovarianz zwischen x und y zusammenhängt:

(7.15) $s_{x^*y^*} = bds_{xy}$ (Kovarianz bei Lineartransformation)

6. Auch für die Kovarianz gilt der **Verschiebungssatz**:

(7.13a) $s_{xy} = \frac{1}{n}\Sigma x_v y_v - \bar{x}\cdot\bar{y}$

bzw. bei gruppierten Daten mit absoluten Häufigkeiten n_{ij}

(7.14a) $s_{xy} = \frac{1}{n}\Sigma\Sigma x_i y_j n_{ij} - \bar{x}\cdot\bar{y}$

und relativen gemeinsamen Häufigkeiten h_{ij}

(7.14b) $s_{xy} = \Sigma\Sigma x_i y_j h_{ij} - \bar{x}\cdot\bar{y}$

oder

(7.16) $s_{xy} = \overline{xy} - \bar{x}\cdot\bar{y}.$

Hierin ist $\overline{xy}$ der Mittelwert des Produkts der x und y Werte und $\bar{x}\cdot\bar{y}$ ist das Produkt der Mittelwerte.

7. Die damit gegebene Beziehung zwischen dem Anfangsproduktmoment xy,¾ und dem zentralen Produktmoment s_{xy} führt auch wegen der Schwerpunkteigenschaft des arithmetischen Mittels zu folgenden Darstellungen der Kovarianz:

(7.17) $s_{xy} = \frac{1}{n}\Sigma(x_v - \bar{x})y_v = \frac{1}{n}\Sigma x_v(y_v - \bar{y}).$

Es genügt also, wenn eine der beiden Variablen zentriert ist.

8. Ein einfach zu zeigender Satz ist:

Satz 7.2:

Verschwindet eine der Varianzen (etwa $s_x^2 = 0$), so ist auch die Kovarianz Null (also $s_{xy} = 0$). Die Umkehrung des Satzes gilt nicht, d.h. $s_{xy} = 0$ ist verträglich mit $s_x^2 > 0$ und $s_y^2 > 0$.

Äquivalent ist die folgende Formulierung:

Die Kovarianz einer Variablen X mit einer Konstanten k ist stets Null, also $s_{xk} = 0$.
Beweis: Dieser Satz ergibt sich unmittelbar aus Gl. 7.17 und ist auch einfach zu interpretieren:

$s_x^2 = 0$ bedeutet, dass alle x-Werte gleich sind, d.h. alle Punkte des Streudiagramms liegen auf einer zur Ordinaten parallelen Geraden. Es liegt dann praktisch eine zu einer eindimensionalen Verteilung degenerierte zweidimensionale Verteilung vor. Soll X die Ursache für Y sein, was impliziert, dass eine nicht unerhebliche Kovarianz zwischen X und Y besteht, so muss X in einem gewissen Maße variieren. Man kann z.B. auch nicht die Ursächlichkeit eines Düngemittels für einen mehr oder weniger großen Ernteertrag nachweisen, d.h. den Zusammenhang "je mehr Dünger desto mehr Ernte" aufzeigen, wenn auf allen Äckern die gleiche Menge gedüngt wird. Man darf andererseits auch nicht folgern: die Variablen X und Y sind umso mehr korreliert, je stärker X und Y streuen. Eine hohe Korrelation ist mit den verschiedensten Varianzen (sofern diese nicht Null sind) verträglich. Das ergibt sich unmittelbar aus der Invarianz des Korrelationskoeffizienten gegenüber linearen Transformationen (vgl. Bemerkung 2 zu Def. 7.8).

9. Unabhängigkeit führt zu einer Kovarianz von Null, denn wenn Gl. 7.8 gilt, dann ist

$$\overline{xy} = \frac{1}{n^2}\sum_i \sum_j x_i n_{i.} y_j n_{.j} = \bar{x}\cdot\bar{y} \text{ und damit } s_{xy} = 0.$$

Die Umkehrung gilt jedoch nicht: die Kovarianz kann sehr wohl verschwinden, obgleich keine Unabhängigkeit besteht (vgl. Beispiele 7.7 und 7.8). Es gilt also der bereits erwähnte

Satz 7.1:

Unabhängigkeit führt zum Verschwinden der Kovarianz (und damit zu Unkorreliertheit). Die Umkehrung gilt jedoch nicht.

Beweis: s.o. und Bemerkungen zu Def. 7.5.

10. Ist Y linear abhängig von X (und damit auch umgekehrt X von Y), etwa dergestalt, dass $y_v = a + bx_v$, so folgt aus Gl. 7.13a $s_{xy} = bs_x^2$.
Ferner ist $s_y^2 = b^2 s_x^2$, so dass bei linearer Abhängigkeit gilt $(s_{xy})^2 = s_x^2 s_y^2$
Wie Def. 7.8. zeigt, bedeutet dies, dass der Korrelationskoeffizient dann den Wert 1 annimmt. In allen anderen Fällen gilt der folgende Satz:

Satz 7.3:

(7.18) $0 \le (s_{xy})^2 \le s_x^2 s_y^2$

Dieser Zusammenhang ist auch bekannt als Schwarzsche - Ungleichung oder Cauchy-Schwarzsche-Ungleichung.

Beweis:

Offenbar gilt wegen der Quadrierung
$0 \le \frac{1}{n}\Sigma[(y_v-\overline{y}) - (s_{xy}/s_y^2)(x_v-\overline{x})]^2$.

Die rechte Seite ergibt den Ausdruck $s_y^2 - 2s_{xy}^2 / s_x^2 + s_{xy}^2 / s_x^2$, der nicht-negativ sein muss, also $0 \le s_y^2 - s_{xy}^2 / s_x^2$. Das ergibt umgeformt Gl. 7.18.

Die Ungleichung lässt sich auch zeigen, wenn man y durch eine Regressionsgerade (vgl. Kapitel 8) erklärt: Es gilt dann $y_v = a + bx_v + u_v$ mit $u = 0$, so dass man erhält: $s_{xy} = bs_x^2 + s_{xu}$ und $s_y^2 = b^2 s_x^2 + s_u^2$. Da im Modell der Regressionsanalyse gilt $s_{xu} = 0$ und $s_u^2 > 0$, ist die Ungleichung 7.18 erfüllt.

11. Der folgende Zusammenhang zwischen der Kovarianz und der mittleren Differenz zwischen den Merkmalswerten ist für manche Betrachtungen (z.B. Rangkorrelation) von Interesse:

Satz 7.4:

Mit der Differenz $d_v = x_v - y_v$ des X-Werts und des Y-Werts der v-ten Einheit (Beobachtung) gilt für die Kovarianz:

(7.19) $s_{xy} = ½[(s_x^2 + s_y^2) + (\overline{x}^2 - \overline{y}^2) - (\Sigma d_v^2)/n]$

Beweis:

$\Sigma d_v^2 = \Sigma(x_v - y_v)^2 = \Sigma x_v^2 - 2\Sigma x_v y_v + \Sigma y_v^2$. Damit ist
$\Sigma x_v y_v = ½(\Sigma x_v^2 + \Sigma y_v^2 - \Sigma d_v^2)$ so dass
$s_{xy} = n^{-1}\Sigma x_v y_v - \overline{x}\cdot\overline{y} = ½(n^{-1}\Sigma x_v^2 + n^{-1}\Sigma y_v^2 - n^{-1}\Sigma d_v^2 - 2\overline{x}\cdot\overline{y})$, womit man nach einigen Umformungen Gl. 7.19 erhält.

Zu weiteren Bemerkungen über die Kovarianz vgl. auch Gl. 7.21ff.

Beispiel 7.7:

Man zeige, dass im Beispiel 7.2 zwar die Kovarianz verschwindet, gleichwohl aber keine Unabhängigkeit vorliegt!

Lösung 7.7:

Für die Variablen D (Pompadur) und M (Pompamoll) erhält man die folgenden Werte: ΣD = 400, ΣM = 300 und ΣMD = 12000, so dass man mit n = 10 erhält s_{DM} = 12000/10 - 40·30 = 0. Um zu zeigen ob M und D unabhängig sind ist aus den Daten eine zweidimensionale Verteilung herzuleiten. Man erhält:

	M=10	20	30	40	50	Σ
D=10	0	0	0	1	0	1
20	0	0	1	0	0	1
30	1	0	1	0	0	2
40	0	1	1	0	0	2
50	0	0	1	0	1	2
60	0	0	0	1	0	1
70	0	1	0	0	0	1
Σ	1	2	4	2	1	10

Es ist unschwer zu erkennen, dass die Häufigkeiten dieser Tabelle nicht mit denen übereinstimmen, die sich bei Unabhängigkeit gem. Gl. 7.8 ergäben: so ist z.B. die relative Häufigkeit für die Kombination D = 40 und M = 30 bei Unabhängigkeit 0,2·0,4 = 0,08, statt des empirischen Werts von 0,1. Bei Unabhängigkeit dürfte auch kein Tabellenfeld eine absolute Häufigkeit von Null ausweisen.
Beispiel 7.8 ist ein weiteres Beispiel dafür, dass die Kovarianz verschwinden kann, obgleich keine Unabhängigkeit vorliegt.

Beispiel 7.8:

Gegeben sei die folgende zweidimensionale Häufigkeitsverteilung (relative Häufigkeiten):

	Y=-2	Y=0	Y=1	Σ
X=4	1/8	1/4	1/8	1/2
X=5	3/16	1/16	1/4	1/2
Σ	5/16	5/16	3/8	1

Man zeige, dass hier (wie im Bsp. 7.7) die Kovarianz zwischen X und Y verschwindet, gleichwohl aber keine Unabhängigkeit vorliegt.

Lösung 7.8:

Bei Unabhängigkeit müßte die relative Häufigkeit h_{12} den Wert ½(5/16) = 0,15625 und nicht 1/8 = 0,125 annehmen. Es genügt, ein Tabellenfeld zu überprüfen, um festzustellen, dass keine Unabhängigkeit vorliegt.

2. Korrelationskoeffizient

Die meisten Bemerkungen zur Kovarianz gelten auch für den Korrelationskoeffizienten, der nur eine auf den Wertebereich von -1 bis +1 normierte Kovarianz darstellt.

Def. 7.8: Korrelationskoeffizient

Der Korrelationskoeffizient nach Bravais-Pearson (auch Produkt-Moment-Korrelationskoeffizient oder im folgenden einfach Korrelationskoeffizient genannt) ist das Verhältnis aus Kovarianz (vgl. Def. 7.7) und dem Produkt der Standardabweichungen

$$(7.20)\quad r_{xy} = \frac{s_{xy}}{s_x s_y} \quad .$$

Bem: Aus den unterschiedlichen Darstellungsmöglichkeiten der Kovarianz (Def. 7.7) und der Varianz ergeben sich auch unterschiedliche Berechnungsformeln für den Korrelationskoeffizienten.

Interpretation und Eigenschaften

1. Aus der Schwarzschen Ungleichung (Satz 7.3, Gl. 7.18) folgt die Einschrankung

   $$(7.20a)\quad -1 \le r_{xy} \le +1 \; .$$

   Der Korrelationskoeffizient ist also die durch das Produkt der Standardabweichungen (oder: das geometrische Mittel der Varianzen) auf den Wertebereich von -1 bis +1 normierte Kovarianz. Die Grenzen r = -1 und r = +1 werden gem. Bem. 10 zu Def. 7.7 erreicht, wenn y eine Lineartransformation von x ist, wenn also gilt y = a + bx. Die Punkte des Streuungsdiagramms liegen dann genau auf der Geraden y = a + bx. Das Vorzeichen des Korrelationskoeffizienten wird allein durch die Kovarianz bestimmt, weil der normierende Nenner $s_x s_y$ stets positiv ist.

2. Durch die Normierung auf den Wertebereich von -1 bis +1 ist der Korrelationskoeffizient (anders als die Kovarianz) maßstabsunabhängig, d.h. er ist unabhängig davon, in welcher Maßeinheit die Variablen X und Y gemessen sind und invariant gegenüber linearen Transformationen.
   Insbesondere gilt bei $x^* = a + bx$ und $y^* = c + dy$ für die Korrelation zwischen x^* und y^* in Relation zur ursprünglichen Korrelation

$$r_{x^*y^*} = \begin{cases} + r_{xy} & \text{wenn } bd > 0 \\ - r_{xy} & \text{wenn } bd < 0 \end{cases}$$

Der Korrelationskoeffizient r_{xy} ist die Kovarianz zwischen den standardisierten Variablen $x^* = (x - \bar{x}) / s_x$ und $y^* = (y - \bar{y}) / s_y$. Die Standardisierung ist eine spezielle Lineartransformation.

3. Die unter Nr. 2 genannte Eigenschaft bedeutet, dass r ein Maß des linearen Zusammenhangs ist. Mit $|r| = 1$ ist ein perfekter **linearer** Zusammenhang gegeben (die Punkte des Streuungsdiagramms liegen genau auf einer Geraden). Je nachdem, wie stark r betragsmäßig vom Wert 1 abweicht, weichen die Punkte mehr oder weniger von den in Kap. 8 behandelten Regressionsgeraden ab. Der Wert $r = 0$ bedeutet nicht, dass kein Zusammenhang zwischen den Variablen X und Y besteht, sondern nur, dass kein **linearer** Zusammenhang besteht (vgl. Beispiel 7.9). Er bedeutet insbesondere auch nicht notwendig Unabhängigkeit; denn nach Satz 7.1 impliziert Unabhängigkeit Unkorreliertheit aber nicht umgekehrt.

4. Gem. Satz 7.2 gilt: Das Verschwinden einer Varianz (etwa $s_x^2 = 0$) impliziert $r_{xy} = 0$. Das bedeutet: Wenn auch nur eine der beiden Variablen X und Y konstant ist, dann können sie auch nicht miteinander korreliert sein.

   Um also eine Korrelation zwischen zwei Variablen X und Y feststellen zu können, müssen beide Variablen streuen. Daraus kann jedoch nicht geschlossen werden, dass die Korrelation umso größer ist, je stärker die beiden Variablen streuen. Denn man kann durch eine Lineartransformation von X zu X^*, etwa $x^* = a + bx$ die Standardabweichung vergrößern (ver-b-fachen), ohne dass sich die Korrelation wegen der Invarianz gegenüber Lineartransformationen ändert ($r_{x^*y} = r_{xy}$).

Weitere Zusammenhänge zwischen Varianzen, Kovarianzen und Korrelationskoeffizienten

Für die folgende Darstellung ist es nützlich, die Notation etwas zu vereinfachen: statt mit s_{xy} soll die Kovarianz mit C(X,Y) bezeichnet werden; entsprechend ist $V(X) = s_x^2$ und

$R(X,Y) = r_{xy}$. Dann gilt:

(7.21) $C(X,Y+Z) = C(X,Y) + C(X,Z)$
(7.22) $V(X+Y) = V(X) + V(Y) + 2C(X,Y) = C(X,X+Y) + C(Y,X+Y)$
(7.23) $C(X+Y,X-Y) = V(X) - V(Y)$

Danach kann X+Y und X-Y unkorreliert sein, obgleich X mit Y korreliert ist.
Aus $C(X,Y-X) = C(X,Y) - V(X)$ folgt, wenn $V(X) = V(Y)$

(7.24) $R(X,Y-X) = -\sqrt{½(1-r_{xy})} \leq 0$ und

(7.25) $R(X,Y+X) = \sqrt{½(1+r_{xy})} \geq 0.$

Beispiel 7.9

Auf ihrer fünftägigen Erkundung des noch wenig erforschten, bisher für unbewohnt gehaltenen, aber bereits gut kartographierten Planeten "Amar" (persisch: Statistik) maßen zwei Astronauten jeweils an drei Zeitpunkten täglich Längengrad (X) und Breitengrad (Y). Dabei erhielten sie die folgenden Messwerte:

Tag	X	Y
Montag	8	5
	9	4
	12	3
Dienstag	15	4
	16	5
	17	8
Mittwoch	16	11
	15	12
	12	13
Donnerstag	9	12
	8	11
	7	8

Zu ihrem großen Staunen korrelierten X und Y nicht mit r = +1 miteinander, obgleich die beiden Astronauten von einer vorgegebenen Linie nicht abgewichen sind. Wie ist das zu erklären?

Lösung 7.9:

Man kann leicht nachrechnen, dass die Korrelation zwischen X und Y den Wert Null annimmt. Es liegt ein funktionaler, aber kein linearer Zusammenhang vor. Eine graphische Darstellung zeigt, dass die Beobachtungen auf einem Kreis im x,y-Koordinatensystem liegen (mit dem Mittelpunkt [$\overline{x}$ = 12 und $\overline{y}$ = 8] sowie dem Radius 5), so dass der funktionale Zusammenhang $(x - 12)^2 + (y - 8)^2 = 5^2$ gegeben ist.

c) Scheinkorrelation

Zu den bekanntesten Fehlinterpretationen der Korrelation gehört die falsche (sachlich nicht gerechtfertigte) kausale Interpretation. Wenn X und Y miteinander korrelieren so kann dies bedeuten, dass:

- X die Ursache von Y ist,
- Y die Ursache von X ist, wobei dies vom ersten Fall mit dem Korrelationskoeffizienten nicht zu unterscheiden ist;
- X und Y rein zufällig in einer entsprechend kleinen Stichprobe miteinander korrelieren, in der Grundgesamtheit jedoch nicht (ein Aspekt der Induktiven Statistik, der in der Deskriptiven Statistik nicht zu behandeln ist);
- Messfehler bei der Beobachtung der Variablen auftreten (wären X und Y fehlerfrei beobachtet, dann würden sie nicht miteinander korrelieren);
- X und Y nur deshalb miteinander korrelieren, weil sie gemeinsam abhängig sind von einer dritten Variablen Z (Scheinkorrelation) und mit Z (nicht direkt unter einander) in einer Kausalbeziehung stehen.

Wegen dieser Nichteindeutigkeit ist die Meinung sehr verbreitet, dass Korrelation und Kausalität nichts miteinander zu tun hätten. Das bedeutet allerdings, das Kind mit dem Bade auszuschütten. Hierauf soll an späterer Stelle eingegangen werden (Abschn. 5).

Def. 7.9: Scheinkorrelation, spurious correlation

Sind zwei Variablen X und Y nur deshalb hoch korreliert, weil sie gemeinsam abhängig sind von einer dritten Variablen Z, so spricht man von Scheinkorrelation.

Bemerkungen zu Def. 7.9:

1. Das beliebteste Beispiel für eine Scheinkorrelation ist die Korrelation zwischen Störchennestern und Geburten. Jeder weiß, dass dies nicht kausal interpretiert werden kann, also kein direkter Kausalzusammenhang besteht.
   Die "dahinterstehende" Variable ist die Urbanisierung und wirtschaftliche Entwicklung, der Übergang von der Agrar- zur Industriegesellschaft, was zum einen dazu geführt hat, dass den Fröschen (und damit auch Störchen) der Lebensraum genommen wurde und zum anderen dazu, dass die Familiengrößen kleiner wurden. Da es in diesem Fall sehr offensichtlich ist, dass nur eine Scheinkorrelation und keine "echte" (d.h. kausal zu interpretierende) Korrelation vorliegt, spricht man auch von **nonsense correlation** (vgl. auch Beispiel 7.10).

Ein weiteres Beispiel ist die Korrelation zwischen der Anzahl der Feuerwehrlöschzüge (X) und der Größe des Brandschadens (Y). Der dritte Faktor (Z) ist die Größe des Brandes (z.B. die Flammenmenge). Die falsche, kausale Interpretation würde lauten: je mehr Feuerwehrlöschzüge bei einem Brand eingesetzt werden, desto größer ist der Brandschaden; also ist der Feuerwehreinsatz die Ursache des Brandschadens.

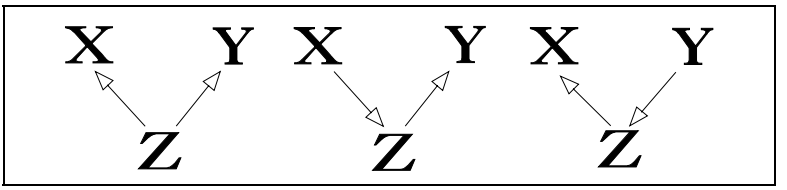
2. Es gibt aber auch viele Fälle einer Scheinkorrelation, bei denen es weniger offensichtlich ist, dass eine Kausalinterpretation nicht zulässig ist. Das ist sehr häufig bei der Korrelation von Zeitreihen, die einen gemeinsamen Trend haben, der Fall (die trendbereinigten Zeitreihen X^* und Y^* korrelieren dann weniger miteinander als die noch trendbehafteten Ursprungswerte X und Y). Sehr häufig tritt das in der Wirtschaftsstatistik bei der Korrelation mit Sozialproduktsgrößen oder allen wertmäßigen (und damit von der Inflation tangierten) Größen auf.
   Ein scherzhaftes Beispiel hierfür ist die hohe Korrelation zwischen den Preisen von kubanischem Rum und den Gehältern von amerikanischen Priestern. Die (in US-$ ausgedrückten) Rumpreise sind nicht gestiegen, weil die Priester so viel Rum nachfragten (eine Kausalinterpretation), sondern weil die Preise - wie die Gehälter - der Entwertung des US-$ unterlagen.

3. Häufig entsteht Scheinkorrelation auch durch Aggregation von Daten (Beispiel 7.11). Bei Disaggregation zeigt sich, dass sich die Korrelation verringert, d.h. dass sie bei Bezugnahme auf homogenere Gesamtheiten nicht gilt.

4. Bei nicht metrisch skalierten Merkmalen zeigt sich Scheinkorrelation durch ein Verschwinden (oder eine Verringerung) der Korrelation zwischen X und Y wenn der Einfluß der dritten Variablen Z "ausgeschaltet" wird d.h. der partiellen Korrelation (vgl. Kap. 8 u. 9).

5. Das Wirken einer dritten Variable Z geschieht bei Scheinkorrelation meist nach Art von Abb. 7.4 linkes Bild. Das Pfeilschema soll andeuten, dass X und Y gemeinsam "verursacht" werden von Z. Hinsichtlich der formalen Zusammenhänge zwischen den Korrelationskoeffizienten sind aber die drei Situationen der Abb. 7.4 nicht unterscheidbar. Auf Abb. 7.4 wird im abschließenden Abschn. 5 noch einmal eingegangen.

Abb. 7.4: Die drei Fälle von Scheinkorrelation zwischen X und Y (wobei jeweils gilt $r_{xy} = r_{xz}r_{zy}$)



Beispiel 7.10:

Während des italienischen Feldzuges im Zweiten Weltkrieg wurde eine positive Korrelation zwischen der Anzahl X der Propaganda-Flugblätter, die über den deutschen Linien abgeworfen wurden, und der Größe des von den Alliierten eroberten Gebietes (Y) bei einer bestimmten Stärke der Offensive (Z) festgestellt. Es sei

$r_{xy} = 0{,}8 \quad r_{xz} = 0{,}94 \quad$ und $\quad r_{yz} = 0{,}85$!

Beweist der hohe Wert von r_{xy}, dass man einen Krieg allein durch möglichst viele Propa-ganda-Flugblätter gewinnen kann?

Lösung 7.10:

Es besteht sicher kein direkter Kausalzusammenhang zwischen der Propaganda X und dem Kriegserfolg (Y). Die relativ hohe Korrelation von $r_{xy} = 0{,}8$ ist also nicht kausal im Sinne von $X \rightarrow Y$ oder gar $Y \rightarrow X$ zu deuten. Als dritter Faktor ist die Stärke der Offensive (Z) zu betrachten. Es ist wohl meist anzunehmen, dass je stärker die Offensive ist, desto eher ist sie erfolgreich ($Z \rightarrow Y$) und desto mehr ist sie auch mit entsprechenden Propagandafeldzügen verbunden ($Z \rightarrow X$). Für die Korrelation gilt in der Tat hier $r_{xy} = r_{xz}\, r_{zy} = 0{,}799$ (Die Zahlenangaben waren ja auch fiktiv). Das Ergebnis besagt, dass die partielle Korrelation $r_{xy.z}$ (vgl. Kap. 8) verschwindet.

Beispiel 7.11:

Gegeben seien die folgenden Daten über die Schuhgröße (X) und das Monatseinkommen in 1000 DM (Y):

x	35	36	37	38	41	42	43	44
y	2,8	2,5	2,0	2,5	4,5	5,5	5,2	3,8

Die Korrelation beträgt $r \approx 0{,}8$. Heisst dies, dass man deshalb mehr verdient, weil "man auf großem Fuß lebt"?

Lösung 7.11:

Es liegt ein typischer Fall von Scheinkorrelation vor. Angenommen, bei den ersten vier Personen handelt es sich um Frauen, die in der Regel eine kleinere Schuhgröße haben als Männer und häufig auch weniger verdienen. Die zweiten vier Personen seien Männer. Dann erhält man für die ersten vier Personen (also für die Frauen) für die Korrelation zwischen X und Y $r_{xy} = -0{,}54495$ und für die zweiten vier Personen (also die Männer)

r_{xy} = -0,40801, bei den beiden Gruppen zusammen aber r_{xy} = 0,8. Man beachte auch, dass sich das Vorzeichen ändert!

d) Bestimmtheitsmaß

Ist $y_v = z_v + u_v$, wobei Z irgendeine Funktion von X ist [allgemein $z_v = z(x_v)$], etwa eine lineare Regressionsfunktion $z_v = a + bx_v$ (s. Kap. 8), so gilt wegen Gl. 7.22 für die Varianz von Y:

$$s_v^2 = s_z^2 + s_u^2 + 2s_{zu}.$$

Verschwindet die Kovarianz s_{zu}, wie im Falle einer Regressions**geraden**, so lässt sich die Varianz von Y darstellen als Summe einer systematischen (durch X "erklärten") Varianz s_z^2 und einer (nicht erklärten) Residualvarianz s_u^2. Es gilt also bei einer mit der systematischen Komponente Z unkorrelierten Variable U stets die folgende **Varianzzerlegung**:

(7.26) $s_v^2 = s_z^2 + s_u^2$

Gesamtvarianz	erklärte Varianz	Residualvarianz

und nach Division durch s_v^2

$$(7.27)\quad 1 = \frac{s_z^2}{s_y^2} + \frac{s_u^2}{s_y^2} = B_{yx} + U_{yx},$$

worin B_{yx} die Bestimmtheit (von y durch x) und U_{yx} die Unbestimmtheit (von y durch x) darstellt.

Def. 7.10: Bestimmtheitsmaß

Der Ausdruck

$$(7.28)\quad B_{yx} = \frac{s_z^2}{s_y^2}$$

heißt Bestimmtheitsmaß (oder Bestimmtheit, coefficient of determination) und $U_{yx} = 1 - B_{yx}$ heißt Unbestimmtheit (oder coefficient of alienation). s_z^2 ist die erklärte Varianz von y und s_v^2 die Gesamtvarianz von y.

Bemerkungen zu Def. 7.10:

1. Da B_{yx} ein Varianzanteil ist, gilt $0 \le B_{yx} \le 1$. Die Bestimmtheit ist also nichtnegativ. Sie ist nicht notwendig symmetrisch, d.h. es muss nicht $B_{yx} = B_{xy}$ erfüllt sein.

2. Das Bestimmtheitsmaß liefert einen allgemeineren Zugang zur Messung des Zusammenhangs zweier metrisch skalierter Variablen als der Korrelationskoeffizient, der nur den Grad des **linearen** Zusammenhangs misst.

   Im Falle einer **linearen** Beziehung zwischen y und x gilt:

   1. Symmetrie: $B_{yx} = B_{xy}$
   2. Zusammenhang mit Korrelation: $B_{yx} = r^2_{xy}$,

   was bei nichtlinearer Beziehung nicht gewährleistet ist.

   Bei nichtlinearer Korrelation ist die Berechnung von r_{xy} (gem. Def. 7.8) wenig sinnvoll. Es kann aber $+\sqrt{B_{yx}}$ oder $+\sqrt{B_{xy}}$ als Korrelationskoeffizient berechnet werden. Diese Korrelation ist dann keine Produkt-Moment-Korrelation, sie kann nicht negativ sein und ist auch nicht notwendig symmetrisch. Ist der Zusammenhang nichtlinear, so macht es i.d.R. auch nicht viel Sinn zwischen positiver und negativer Korrelation zu unterscheiden.

3. Zum Begriff "**erklärte**" Varianz und zum Konzept der "Ursache" (X als Ursache für Y):
   "Erklärung" im Sinne der Statistik besteht stets in der Zerlegung der Varianz einer zu erklärenden Variable Y in einen Varianzanteil B_{yx}, der auf eine bekannte Variationsquelle X zurückgeführt werden kann und einen restlichen Anteil U_{yx}, der nicht auf eine explizit berücksichtigte, identifizierbare, "erklärende" Variationsquelle U zurückzuführen ist.

   Für diese "Erklärung" ist es nicht unwichtig, ob Beobachtungs- und Experimentdaten vorliegen: Nur im letzten Fall besteht a priori eine klare Unterscheidung, welche Einflußfaktoren zu X und welche zu U zu rechnen sind, weil eine Größe X bewußt und kontrolliert (bei Konstanz anderer Größen) variiert werden kann.

4. Im Unterschied zum Korrelationsverhältnis, das auf die Regressions**linie** Bezug nimmt, ist im Falle des Bestimmtheitsmaßes eine Regressions**funktion**, z.B. eine lineare Regressionsfunktion zu schätzen. Da dies erst im Kap. 8 gezeigt wird, soll hier kein Rechenbeispiel für das Bestimmtheitsmaß betrachtet werden.

e) Korrelationsverhältnis und Korrelationskoeffizient bei klassierter Verteilung

Es mag sein, dass bei Nichtlinearität des Zusammenhangs zwischen zwei metrisch skalierten Variablen X und Y der Produktmoment-Korrelationskoeffizient r_{xy} den Grad des Zusammenhangs unterschätzt und dass durch Berechnung eines nichtlinearen Korrelationskoeffizienten (vgl. Bem. Nr. 2 zu Def. 7.10) ein höherer Grad des Zusammenhangs festgestellt werden kann (wenn z.B. der Zusammenhang durch Annahme einer parabolischen statt linearen Beziehung besser erfaßt wird). Im Falle gruppierter oder klassierter Daten beschreibt die Regressionslinie den Zusammenhang zwischen X und Y in einer Weise, wie es durch keine (wie immer geartete) nichtlineare Regressionsfunktion besser geschehen könnte. Das Korrelationsverhältnis η(eta) misst die Güte der Anpassung der Regressions**linie** an die Daten und beruht auf der folgenden Varianzzerlegung.

Satz 7.5: Varianzzerlegung

Mit den m bedingten Mittelwerten $\bar{y}(x_i) = \bar{y}|x=x_i$ gem. Gl. 7.12 lassen sich die Abweichungen der n_i Beobachtungen in der i-ten Klasse bezüglich des Merkmals X (bzw. in den Fällen, in denen $X = x_i$ ist) vom Gesamtmittelwert $\bar{y}$ darstellen als

$$y_{pi} - \bar{y} = y_{pi} - \bar{y}(x_i) + \bar{y}(x_i) - \bar{y}$$

mit i = 1,2,...,m Ausprägungen oder Klassen bezüglich X und p = 1,2,...,n_i Beobachtungen in der i-ten Klasse (wenn X die Ausprägung x_i hat).

Damit gilt auch

$$(7.29)\quad \Sigma\Sigma y_{pi} - \bar{y} = \Sigma\Sigma[y_{pi} - \bar{y}(x_i)] + \Sigma n_{i.}[\bar{y}(x_i) - \bar{y}].$$

Summierung über alle m Ausprägungen oder Klassen der Variablen X und Quadrierungen der Abweichungen liefert

$$(7.30)\quad \Sigma\Sigma(y_{pi} - \bar{y})^2 = \Sigma n_{i.} s_y^2(x_i) + \Sigma n_i[\bar{y}(x_i) - \bar{y}]^2$$
$$SAQ_{tot} = SAQ_{int} + SAQ_{ext}$$

wobei die interne Varianz $SAQ_{int}/n = \Sigma h_{i.} s_y^2(x_i)$ ein gewogenes Mittel der bedingten Varianzen (gem. Gl. 7.12a) darstellt.

In Gl. 7.30 ist SAQ die Summe der Abweichungsquadrate. Die gesamte Summe der Abweichungsquadrate (SAQ_{tot}) lässt sich also in eine interne

und eine externe Summe der Abweichungsquadrate (SAQ_{int} und SAQ_{ext}) zerlegen. Dividiert man die linke und rechte Seite von Gl. 7.30 durch n, so erhält man die Varianzzerlegung:

$$(7.30a) \quad V(y_{tot}) = V(y_{int}) + V(y_{ext}) \qquad (V = SAQ/n = \text{Varianz}).$$

Beweis: Zu zeigen ist allein der Übergang von Gl. 7.29 zu Gl. 7.30. Er ergibt sich durch Ausmultiplizieren unter Berücksichtigung von $s_y^2(x_i) = n_i^{-1} \Sigma y_{pi}^2 - [\bar{y}(x_i)]^2$ und aus der Schwerpunkteigenschaft des arithmetischen Mittels. Die Zerlegung der Varianz (Gl. 7.30a) bzw. der Summe der Abweichungsquadrate (SAQ, Gl. 7.30) wird auch in den Beispielen 7.12 und 7.13 demonstriert.

Def. 7.11: Korrelationsverhältnis

Das Korrelationsverhältnis η (eta) ist der Ausdruck

$$(7.31) \quad \eta_{yx} = +\sqrt{\frac{SAQ_{ext}}{SAQ_{tot}}}$$

mit den Summen der Abweichungsquadrate SAQ gem. Gl. 7.30.

Bemerkungen zu Def. 7.11:

1. Das Korrelationsverhältnis ist nicht notwendig symmetrisch, d.h. in der Regel sind η_{yx} und η_{xy} nicht gleich (vgl. Beispiele 7.12 und 7.13). Auch kann mit η nicht zwischen positiver und negativer Korrelation unterschieden werden, denn η ist (wie das Bestimmtheitsmaß) aus einem Varianzanteil abgeleitet: η_{yx}^2 ist der Anteil der durch die Regres-sionslinie erklärten Varianz von Y und entsprechend wird in η_{xy}^2 eine Zerlegung der Varianz von X vorgenommen.

2. η ist i.d.R. verschieden vom (linearen) Korrelationskoeffizienten r_{xy} bei gruppierten (bzw. analog mit $\bar{x}_i$, $\bar{y}_j$ klassierten) Daten:

$$(7.32) \quad r_{xy} = \frac{n \Sigma\Sigma x_i y_i n_{ij} - (\Sigma x_i n_{i.})(\Sigma y_j n_{.j})}{\sqrt{[\Sigma(x_i-\bar{x})^2 n_{i.}]\,[\Sigma(y_j-\bar{y})^2 n_{.j}]}}$$

r_{xy}: linearer Korrelationskoeffizient bei gruppierten bzw. klassierten Daten

Im Zähler von Gl. 7.32 steht die n^2-fache Kovarianz und im Nenner die Wurzel aus dem Produkt der n-fachen Varianzen von X und Y. Die Berechnung des Korrelationskoeffizienten wird in Bsp. 7.12 und 7.13 demonstriert.

Bei klassierten Daten ist anstelle von x_i und y_j der jeweilige Klassenmittelwert der i-ten Klasse bezüglich X, bzw. der j-ten Klasse bezüglich Y einzusetzen.

2. Das Korrelationsverhältnis ist nicht unabhängig von der Anzahl der Klassen bzw. "Gruppen" (unterschiedliche Werte x_i bzw. y_j). Werden nur wenige Klassen unterschieden z.B. aus Gründen der Rechenvereinfachung im Beispiel 7.14, so ist die Berechnung von η wenig sinnvoll. Andererseits gilt: Bei vielen schwach besetzten Klassen können die bedingten Mittelwerte stark schwanken, weil in jeder Klasse nur wenige Beobachtungen vorliegen.

Beispiel 7.12:

Man verifiziere die Varianzzerlegung (Gl. 7.30) und berechne die Korrelationsverhältnisse und den Korrelationskoeffizienten r_{DM} für das Beispiel 7.2!

Lösung 7.12:

Es soll der Einfachheit halber nur von Dur (D) und Moll (M) gesprochen werden.

Daten zur Regressionslinie zur Schätzung von Dur:

wenn M	n_j	bed.Mittelw.	bedingte Varianz von Dur (D)
10	1	30	0 da $n_{.1} = 1$
20	2	55	½ $[(40-55)^2 + (70-55)^2] = 225$
30	4	35	¼ $[(20-35)^2 + (30-35)^2 +$ $(40-35)^2 + (50-35)^2] = 125$
40	2	35	½ $[(10-35)^2 + (60-35)^2] = 625$
50	1	50	0 da $n_{.5} = 1$

Man erkennt an der Folge der bedingten arithmetischen Mittelwerte, dass kein *linearer* Zusammenhang besteht (es ist auch $r_{DM}=0$).

Berechnung der Summe der Quadrate der Abweichungen SAQ für Dur:
da das Gesamtmittel $\bar{D} = 40$ ist

$SAQ_{tot} = (10-40)^2 + (20-40)^2 + 2(30-40)^2 + 2(40-40)^2 + 2(50-40)^2 + (60-40)^2 + (70-40)^2 = 3000$ (so dass wegen n = 10 für die Varianz gilt $s_D^2 = 300$).

$SAQ_{int} = 2\cdot225 + 4\cdot125 + 2\cdot625 = 2200;$

$SAQ_{ext} = (30-40)^2 + 2(55-40)^2 + 6(35-40)^2 + (50-40)^2 = 800.$

Man sieht, dass gilt $SAQ_{tot} = SAQ_{int} + SAQ_{ext} = 3000 = 2200 + 800$.

Für das Korrelationsverhältnis erhält man dann $\eta_{DM}^2 = SAQ_{ext}/SAQ_{tot} = 800/3000 = 4/15 = 0{,}267$ und folglich ist $\eta_{DM} = \sqrt{4/15} = 0{,}5164$.

Entsprechend erhält man mit der Regressionslinie zur Schätzung von Moll:

$SAQ_{tot} = 1200 \quad (s_M^2 = = 120)$

$SAQ_{int} = 2\cdot100 + 2\cdot25 + 2\cdot100 = 450$

$SAQ_{ext} = 100 + 200 + 50 + 200 + 100 + 100 = 750$

Es gilt wieder $SAQ_{tot} = SAQ_{int} + SAQ_{ext} = 1200 = 450 + 750$.

Korrelationsverhältnis ($\eta^2_{MD} = SAQ_{ext}/SAQ_{tot} = 750/1200 = 0{,}625$ und somit $\eta_{MD} = 0{,}79057$ (was ungleich $\eta_{DM} = 0{,}5164$ ist).

Beispiel 7.13:

(Die emanzipierte Fassung von Aufg. 7.2/7.12: Das Experiment einer Dreierbeziehung und ein neuerliches Beispiel für konkrete Lebenshilfe durch Statistik)

Nachdem Andrea (A) zwei Jahre mit Charlie (C) ging, haben sie sich `ne echt besitzhafte Identität aufgebaut, aus der sich A nun emanzipieren will. Sie ist jetzt mehr so auf Bernd (B) drauf, kann aber noch nicht total auf B einflippen. Und weil ihr bisheriger Typ C die Trennungsverarbeitung erst einmal konkret abgecheckt haben will und das, was zwischen A und B so läuft emotional noch nicht so auffangen kann, haben sie jetzt alle drei beschlossen, das Problem bis spätestens zum nächsten Jahr zu dritt ganz konkret aufzuarbeiten. In ihrer total fixierenden Art, mit der sie mit jeder Beziehungskiste umgeht hat A die folgenden Aufzeichnungen gemacht über die Tage, die sie im Monat mit B bzw. C verbracht hatte:

	Tage mit C			
Tage mit B	0-10	10-20	>20	Σ
0-10	0	2	4	6
10-20	1	2	0	3
>20	3	0	0	3
Σ	4	4	4	12

Man verifiziere die Varianzzerlegung (Gl. 7.30) und berechne die Korrelationsverhältnisse sowie den Korrelationskoeffizient r_{BC}! Als Klassenmittelwerte sind die Zahlen 5, 15 und 25 anzusetzen.

Lösung 7.13:

Während Beispiel 7.12 die Berechnung des Korrelationsverhältnisses bei gruppierten Daten demonstrieren soll, gilt es hier, Varianzzerlegung und Berechnung des Korrelatinsverhältnisses sowie des Korrelationskoeffizienten bei einer zweidimensionalen klassierten Verteilung zu zeigen.

Parameter der Randverteilungen

Bernd	Charlie
Mittel $\bar{B} = 12{,}5$	Mittel $\bar{C} = 15$
Varianz $s^2_B = 825/12$	Varianz $s^2_C = 800/12$,

so dass SAQ_{tot} 825 bzw. 800 ist.

Regressionslinie Bernd

wenn C	$n_{.j}$	bed.Mittelwert	bedingte Varianz der Regressionslinie v. Bernd
5	4	22,5	$\frac{1}{4}[(15-22,5)^2+ 3(25 - 22,5)^2] = 75/4$
15	4	10	$\frac{1}{4}[2(5-10)^2+2(5 -10)^2] = \frac{1}{4}100 = 25$
25	4	5	0 da alle 4 Beobachtungen gleich sind

Regressionslinie Charlie

wenn B	$n_{i.}$	bed. Mittelw.	bed. Varianz
5	6	130/6	133,33/6
15	3	35/3	22,222
25	3	5	0

$SAQ_{int} = 75 + 100 = 175$ $SAQ_{int} = 133,33 + 66,67 = 200$

$SAQ_{ext} = 4(22,5-12,5)^2+4(10-12,5)^2+4(5-12,5)^2 = 650$ $SAQ_{ext} = 600$ (SAQ_{tot}=800)

$SAQ_{tot} = 825 = SAQ_{ext} + SAQ_{int} = 650 + 175.$

Berechnung der Korrelationsverhältnisse:

Bernd: $\eta_{BC} = \sqrt{SAQ_{ext}/SAQ_{tot}} = \sqrt{650/825} = 0,8876$

Charlie: $\eta_{CB} = \sqrt{600/800} = \sqrt{3/4} = 0,8660$

Die beiden Korrelationsverhältnisse sind nicht gleich. Man kann an ihnen auch nicht erkennen dass eigentlich eine negative Korrelation besteht.

Berechnung des Korrelationskoeffizienten:

Kovarianz s_{BC}: $(5\cdot5\cdot0 + 5\cdot15\cdot2 + 5\cdot25\cdot4 + 15\cdot5\cdot1 + 15\cdot15\cdot2 + 15\cdot25\cdot0 + 25\cdot5\cdot3 + 25\cdot25\cdot0 + 25\cdot25\cdot0)/12 - 12,5\cdot15 = -700/12$ und da die Varianzen $s_B,^2 = 825/12$ und $s_C,^2 = 800/12$ betragen erhält man $r_{BC} = s_{BC}/s_B s_C = - 0,86164$.

4. Zusammenhang bei nicht metrisch skalierten Variablen

a) Maße des Zusammenhangs und Skalenniveaus (Übersicht)

In Abhängigkeit von dem Skalenniveau der Variablen X und Y gibt es zahlreiche Maße für den Grad des Zusammenhangs zwischen X und Y (vgl. Übers. 7.2). Man kann in allen Fällen auch von Korrelationen im weiteren Sinne sprechen und die bisher behandelte Korrelation als "**Maßkorrelation**" (Korrelation im engeren Sinne) bezeichnen. Im englischen Sprachgebrauch ist auch "association" ein entsprechender Oberbegriff, während Assoziation im engeren Sinne nur den Zusammenhang zweier dichotomer Merkmale bezeichnet. Auf einige der in Übersicht 7.2 genannten Maße des Zusammenhangs wird in den folgenden Abschnitten eingegangen.

Übersicht 7.2:
Maße des Zusammenhangs zwischen den Merkmalen X und Y

a) Fallunterscheidung nach dem Skalentyp der beiden Variablen

	Variable X		
Var. Y	nominal	ordinal	metrisch
nominal	3*)	5	4
ordinal	5	2*)(2a/2b)	
metrisch	4		1

*) weitere Fallunterscheidung unter b)

b) Fälle im einzelnen und Maßzahlen

1. Beide Variablen sind metrisch skaliert (*Fall 1*): **Maßkorrelation**
   - Produkt-Moment-Korrelationskoeffizient (Def. 7.8)
   - Korrelationsverhältnis (Def. 7.11)
   - Gelegentlich wird auch auf den älteren, kaum noch gebräuchlichen Korrelationskoeffizient r_F von Fechner in diesem Zusammenhang verwiesen (vgl. Bem. 3 zu Def. 7.7). Da es aber bei ihm nur auf die Vorzeichen der Abweichungen vom Schwerpunkt $(\overline{x},\overline{y})$ bzw. vom Medianpunkt $(\tilde{x}_{0,5}, \tilde{y}_{0,5})$ ankommt, paßt r_F eher zur Situation 2b.

2. Die ordinale Abstufung (*Fall 2*) kann
   2a) durch Rangplätze beschrieben werden (**Rangkorrelation**).
   2b) ohne Rangplätze bestehen:
   bei beiden Merkmalen X und Y werden die Ausprägungen allein ordinal unterschieden, etwa $x_1 < x_2 < x_3$ usw.; sie werden nicht mit [wie immer gefundenen] Zahlenwerten (z.B. Rangplätzen) codiert (**Rangassoziation** oder ordinale Assoziation [order association]).

   Bem.: Im Falle 2a) sind zwar nicht die Meßwerte x_1, x_2 usw. (Merkmalsausprägungen) metrisch skaliert, wohl aber die hierfür verwendeten Rangplätze.

3. Weitere Fallunterscheidung (*Fälle 3 und 4*)

   Nominalskala mit
   p > 2 Ausprägungen: **Polytomie (P)**,
   p = 2 Ausprägungen: **Dichotomie (D)**.
   Speziell im Falle einer Dichotomie kann man unterscheiden:

   D1: Echte Dichotomie (qualitative Unterscheidung),
   D2: Dichotomie mit einer zugrundeliegenden Rangordnung (z.B. Unterscheidung von gut/schlecht, positiv/negativ usw.),

D3: Dichotomie bei an sich zugrundeliegender Normalverteilung.

Y	X-Skala				
Skala	D1/D2	D3	P	M*)	O*)
D1/D2	A	*	K	PB	RB
D3	*	T	*	B	
P	K	*	K		

*) M = metrisch-skaliert O = ordinal-skaliert

A = Assoziationsmaße (Vierfelderkorrelation)
T = Tetrachorische Korrelation (tetrachoric correlation)
K = Kontingenzmaße
B = Biserielle Korrelation (biserial correlation, analog z.B. triseriell wenn Y als Trichotomie vorliegt)
PB = Punkt-biserielle-Korrelation (point biserial correlation)
RB = Rang-biserieller Korrelationskoeffizient.

b) Assoziation und Kontingenz

1. Allgemeines

Konstruktionsprinzipien für Zusammenhangsmaße

Man kann Assoziations- und Kontingenzmaße (Übersicht 7.2) konstruieren aufgrund folgender Überlegungen:

a) *Auf χ^2 basierende Maße*: Vergleich der Häufigkeiten n_{ij} für die Kombination (x_i,y_j) der Merkmale X und Y mit den zu erwartenden Häufigkeiten bei Unabhängigkeit [Def. 7.5] (sind sie gleich, so liegt kein Zusammenhang vor).
b) *Prädikationsmaße*: Liegt ein Zusammenhang zwischen X und Y vor (also keine Unabhängigkeit), so ist bei Kenntnis der Verteilung von X die Verteilung von Y (bzw. umgekehrt) "besser" vorauszusagen als ohne Kenntnis.
c) Man kann die *Häufigkeit konkodanter und diskonkodanter Merkmalskombinationen* vergleichen, was bei Nominalskalen ohne jede dahinterstehende Rangordnung nicht sinnvoll ist, wohl aber z.B. bei Dichotomien im Sinne von D2/D3 der Übers. 7.2 (also in der Assoziationsanalyse).
d) Speziell in der Assoziationsanalyse: Berechnung von r_{xy}, der Produkt-Moment-Korrelation (Def. 7.8) für mit 0 und 1 codierte Variablen X und Y ("Vierfelderkorrelation" [vgl. Def. 7.16]).

Normierungsprobleme

Während die Untergrenze aller Maße des Zusammenhangs eindeutig ist und durch Unabhängigkeit (Def. 7.5) gegeben ist, macht es Schwierigkeiten, einen "maximalen" Zusammenhang zu definieren und so Kontingenz- und Assoziationsmaße auf den Wertebereich von 0 bis 1 oder -1 bis +1 zu normieren.
Abhängigkeit (Unabhängigkeit) besteht in der Unterschiedlichkeit (Gleichheit) der bedingten Verteilungen. Aber:

- eine maximale Unterschiedlichkeit ist nicht eindeutig definiert;
- während Unabhängigkeit eine symmetrische Eigenschaft ist, muss dies für die Abhängigkeit nicht gelten.

Axiomatik

Für ein Assoziations- oder Kontingenzmaß (AK) sollte gelten

A1/K1 Das Maß AK sollte dann und nur dann den Wert AK = 0 annehmen, wenn die beiden Variablen unabhängig sind.

A2/K2 Bei einer genau definierten "maximalen Abhängigkeit" sollte AK die Obergrenze AK = +1 annehmen. Diese Obergrenze ist im Falle der Assoziation, nicht aber in dem der Kontingenz eindeutig definiert.

A3/K3 Das Maß AK sollte nicht invariant sein gegenüber einer Ver-k-fachung von einzelnen Zeilen oder Spalten.

A4/K4 Das Maß AK sollte von der Gesamtzahl n der Beobachtungen unabhängig sein.

Durch eine Ver-k-fachung der Häufigkeiten einer einzelnen Zeile (oder Spalte) ändert sich die entsprechende bedingte Verteilung nicht, wohl aber die Abhängigkeit zwischen den Merkmalen, was im Fall der Assoziation durch Hinweis auf die Regressionsanalyse gezeigt werden wird. Es gibt Maße, die invariant sind gegenüber einer solchen Veränderung der Häufigkeiten. Im Unterschied hierzu soll eine Ver-k-fachung **aller** Häufigkeiten nach A4/K4 das Maß AK nicht verändern.

Beispiel/Lösung 7.14:

Man könnte von "vollständiger Abhängigkeit" der Variablen Y von der Variablen X (oder von einem "maximalen Zusammenhang" zwischen den Variablen X und Y) sprechen, wenn aus der Verteilung von X eindeutig die Verteilung von Y hervorgeht und umgekehrt, wie dies bei der folgenden Tafel 1 der Fall ist. Es ist klar, dass dieses Konzept eines "maximalen Zusammenhangs" nur Sinn macht bei quadratischen Kontingenztafeln. Da für X und Y nur Nominalskalen vorausgesetzt werden, müßte der Grad der Abhängigkeit gleich bleiben wenn Zeilen und Spalten permutiert oder auch

zusammengefasst werden. Es müßten also die folgenden vier Tafeln jeweils die gleiche Kontingenz (als Grad der Abhängigkeit) aufweisen:

Tafel 1

	y_1	y_2	y_3	Σ
x_1	20	0	0	20
x_2	0	30	0	30
x_3	0	0	50	50
Σ	20	30	50	100

Tafel 2

	y_2	y_1	y_3	Σ
x_3	0	0	50	50
x_2	30	0	0	30
x_1	0	20	0	20
Σ	30	20	50	100

Kontingenzmaße betrachten i.d.R. die Tafeln 1 und 2 als gleichwertig und sie nehmen jeweils ihren maximalen Wert 1 an, nicht dagegen in den Fälle der Tafel 3 und 4.

Tafel 3

	y_1+y_3	y_2	Σ
x_3	50	0	50
x_1	20	0	20
x_2	0	30	30
Σ	70	30	100

Tafel 4

	y_1+y_3	y_2	Σ
x_2+x_3	50	30	80
x_1	20	0	20
Σ	70	30	100

3. Kontingenzmaße

a) auf den Vergleich mit der Unabhängigkeit basierende Maße

Sind die Merkmale X und Y Polytomien, also Nominalskalen mit jeweils zwei und mehr Ausprägungen, so kann man die Größe Chi-Quadrat (χ^2) berechnen. Sie beruht auf einem Vergleich der beobachteten Häufigkeiten n_{ij} mit den (bei Unabhängigkeit) zu erwartenden Häufigkeiten f_{ij}.

Def. 7.12: Chi-Quadrat

Mit den beobachteten Häufigkeiten n_{ij} ($n = \Sigma\Sigma n_{ij}$) und den bei Unabhängigkeit [Def. 7.5] zu erwartenden Häufigkeiten f_{ij} ist die Größe Chi-Quadrat (χ^2) definiert als

(7.33) $\chi^2 = \Sigma\Sigma(n_{ij} - f_{ij})^2/f_{ij}$ mit $f_{ij} = n_{i.}n_{.j}/n$ $\quad (i = 1,2,...,r \text{ und } j = 1,2,...,c)$

Bemerkungen zu Def. 7.12:

1. Die Größe χ^2 selber ist nicht als Kontingenzmaß geeignet, weil sie direkt von der Anzahl der Beobachtungen abhängt. Eine Ver-k-fachung aller Häufigkeiten führt auch zu einem k-fachen Wert von χ^2.

2. Einfluß hat auch die Anzahl r der Zeilen (rows) und c der Spalten (columns) der zugrundeliegenden Kontingenztafel (i = 1,2,...,r und j = 1,2,...,c).

3. Die Punkte 1 und 2 haben die Konstruktion einiger auf χ^2 basierender Kontingenzmaße angeregt (Def. 7.13), die wegen χ^2 stichprobentheoretisch gewisse Vorteile haben, andererseits aber kaum anschaulich interpretierbar sind. Es gibt deshalb auch Kontingenzmaße, die einem anderen Konzept folgen (Def. 7.14).

4. Aus der Definition von χ^2 (Gl. 7.33) folgt, dass χ^2 bei Unabhängigkeit den Wert Null annimmt und in allen anderen Fällen größer als Null ist.

5. Die Quadrierung in der Größe χ^2 ist damit zu motivieren, dass die Zeilen- und Spaltensumme der einfachen Abweichungen jeweils verschwindet, denn:

   (7.33a) $$\sum_i (n_{ij} - f_{ij}) = \sum_j (n_{ij} - f_{ij}) = \sum_i \sum_j (n_{ij} - f_{ij}) = 0$$

   Der Zusammenhang wird im Beispiel 7.15 demonstriert.

Def. 7.13: auf Chi-Quadrat beruhende Kontingenzmaße

(7.34) $\phi = \sqrt{c^2/n}$ Phi-Koeffizient

$\phi^2 = \chi^2/n$ heisst auch mittlere quadratische Kontingenz

(7.35) $C = \sqrt{c^2/(c^2 + n)}$ Kontingenzmaß von Pearson

so dass $C^2 = \phi^2/(\phi^2+1)$

(das ist das unnormierte Kontingenzmaß von Pearson, das maximal den Wert $C_{max} = \sqrt{(m-1)/m}$ mit m = min(r,c) annimmt; der auf den Wertebereich [0,1] normierte Kontingenzkoeffizient von Pearson lautet:

(7.35a) $$C^* = \frac{C}{C_{max}} = \sqrt{\frac{mc^2}{(m-1)(c^2 + n)}} \qquad 0 \le C^* \le 1$$

(7.36) $T^2 = \chi^2/[n\cdot\sqrt{(r-1)(c-1)}]$ T: Kontingenzmaß von Tschuprow

(7.37) $V^2 = \chi^2/[n \cdot \min(r-1,c-1)]$ V: Kontingenzmaß von Cramer

Folgerungen:

- bei einer quadratischen Tabelle (r=c) ist T = V
- bei einer Vierfeldertafel [Def. 7.15], d.h. bei r = c = 2 (Assoziation) ist T = V = ϕ und $C^* = \sqrt{2f^2/(f^2+1)}$ mit $0 \le C^* \le 1$.

Beispiel 7.15:

Man bestimme die Kontingenzmaße der Def. 7.13 für die vier Tafeln von Bsp. 7.14!

Lösung 7.15:

Die Bestimmung der Tafel bei Unabhängigkeit gem. Def. 7.5 (auch "Indifferenztabelle" genannt) und damit der Größe χ^2 wird für Tafel 1 ausführlich gezeigt:

empirische (beobachtete) Häufigkeiten n_{ij}

	y_1	y_2	y_3	Σ
x_1	20	0	0	20
x_2	0	30	0	30
x_3	0	0	50	50
Σ	20	30	50	100

erwartete (bei Unabhängigkeit) Häufigkeiten f_{ij}

	y_1	y_2	y_3	Σ
x_1	4	6	10	20
x_2	6	9	15	30
x_3	10	15	25	50
Σ	20	30	50	100

Damit erhält man die Abweichungen $n_{ij} - f_{ij}$

	y_1	y_2	y_3	Σ
x_1	16	-6	-10	0
x_2	-6	21	-15	0
x_3	-10	-15	25	0
Σ	0	0	0	0

(womit auch Gl. 7.33a verifiziert ist) χ^2 ergibt sich nun durch Summierung der Größen $(n_{ij} - f_{ij})^2/f_{ij}$ über alle Zeilen und Spalten.

Für Tafel 1 ist also $\chi^2 = 200$. Mit n = 100 und r=c=3 erhält man außerdem $\phi = \sqrt{2}$, T = V = 1 ferner $C = C_{max} = \sqrt{2/3}$.
Für Tafel 2 erhält man die gleichen Ergebnisse wie für Tafel 1 und für Tafel 3 ergibt sich $\chi^2 = 54{,}857$, $\phi = V = 0{,}74065$, $T = 0{,}62282$, $C = 0{,}59518$.
Für Tafel 4 ist $\chi^2 = 10{,}7143$ und $\phi = T = V = 0{,}32733$ und $C = 0{,}31109$.

b) <u>Prädikationsmaße der Kontingenz (Konzept der Fehlerreduktion)</u>

Eine Alternative zu der wenig anschaulichen Größe χ^2 ist eine Klasse von Maßzahlen, die auf dem Konzept der Fehlerreduktion beruhen und von Leo A. Goodman und William H. Kruskal (1954) in die Diskussion gebracht wurden. Danach sind X und Y dann "korreliert", wenn es gelingt, bei Kenntnis des Merkmalswertes x_v der v-ten Einheit, deren Wert y_v besser (mit geringerem Fehler, mit größerer Treffsicherheit) vorherzusagen als ohne Kenntnis von x_v. Aus dieser Definition des "Zusammenhangs" zwischen X und Y, bzw. der Abhängigkeit der Variablen Y von X

> Y hängt von X in dem Maße ab, in dem Kenntnis über X die Unsicherheit über Y reduziert

folgt auch, dass Zusammenhang nicht notwendig eine symmetrische Relation ist. Die Maße von Goodman und Kruskal nach dem Konzept der pro-

portionalen (=relativen) Fehlerreduktion (proportional reduction in error PRE) sind **asymmetrische Kontingenzmaße**. Sie setzen voraus:

1. ein Konzept (ein Maß) für den **Fehler** der Vorhersage unter Berücksichtigung des Skalenniveaus;
2. eine (i.d.R. für beide Fälle a und b identische) **Vorhersageregel** für die Vorhersage von Y durch X
   a) *ohne* Kenntnis von X (aus der Randverteilung von Y)
   b) aufgrund der (durch X) bedingten Verteilungen von Y (also *mit* Kenntnis von X) und
3. ein Maß für das Konzept der "**Fehlerreduktion**".

Die Konstruktion von Kontingenzmaßen aufgrund bestimmter Festlegungen zu diesen drei Punkten soll anhand des folgenden Beispiels demonstriert werden. Dargestellt werden die Koeffizienten

- *lambda (**λ**)* der *optimalen* (modalen) Vorhersage und
- *tau (**τ**)* der *proportionalen* Vorhersage

von Goodman und Kruskal, wobei sich λ und τ nur durch die Vorhersageregel (Punkt 2a und 2b) unterscheiden.

Beispiel 7.16:

Entwicklung und Demonstration der Maße von Def. 7.14
Anhand der folgenden 3x4 Kontingenztafel (mit absoluten Häufigkeiten n_{ij}) sollen die Maße λ und τ hergeleitet werden.

	y_1	y_2	y_3	y_4	Σ
x_1	9	1	2	13	25
x_2	6	19	6	15	46
x_3	6	5	10	8	29
Σ	21	25	18	36	100

Die Zahlen sind so gewählt (n=100), dass eine Umrechnung in relative Häufigkeiten sehr einfach ist.

Zu 1 und 2a:

Vorhersage von Y **ohne** Kenntnis von X (Vorhersageregel/-fehler)

a) optimale (modale) Vorhersage [λ]
   100 Einheiten, deren y-Wert unbekannt ist werden vollständig der häufigsten (modalen) Ausprägung y_4 zugeordnet: damit werden 36 Einheiten richtig und 64 Einheiten falsch zugeordnet. Der Vorhersagefehler ist somit:
   $E_1 = 1 - p_{.4} = 1 - \max_j p_{.j} = 0{,}64$.

b) proportionale Vorhersage [τ]

100 Einheiten, deren y-Wert unbekannt ist werden proportional zu den Häufigkeiten der Randverteilung zugeordnet, also 21 in y_1, 25 in y_2 usw. Der Vorhersagefehler ist somit:
$0{,}21 \cdot 0{,}79 + 0{,}25 \cdot 0{,}75 + 0{,}18 \cdot 0{,}82 + 0{,}36 \cdot 0{,}64 = 0{,}7314$
$F_1 = \Sigma p_{.j}(1 - p_{.j}) = 1 - \Sigma(p_{.j})^2 = 0{,}7314.$

Zu 2b:

Vorhersage von Y **mit** Kenntnis von X, also aufgrund der bedingten Verteilungen von Y (Vorhersageregel/-fehler)

a) optimale (modale) Vorhersage [λ]
Die Einheiten werden jeweils vollständig der bei gegebenem Wert von X häufigsten Ausprägung von Y zugeordnet. Der Vorhersagefehler ist somit:
bei $X = x_1$: 0,25 - 0,13 = 0,12,
bei $X = x_2$: 0,46 - 0,19 = 0,27 usw. insgesamt also
$E_2 = \Sigma_i [p_{i.} - \max_j (p_{ij})] = 1 - \Sigma_i \max_j (p_{ij}) = 0{,}58.$

b) proportionale Vorhersage [τ]
Die Einheiten werden proportional zu den Häufigkeiten der bedingten Verteilungen zugeordnet. Der Vorhersagefehler ist dann: bei $X = x_1$:
$9/25 \cdot 16/25 + 1/25 \cdot 24/25 + 2/25 \cdot 23/25 + 13/25 \cdot 12/25 = 0{,}592$
oder allgemein:
$f_1 = \sum_j (p_{1j}/p_{1.})(1 - p_{1j}/p_{1.}) = 1 - \sum_j (p_{1j}/p_{1.})^2,$

entsprechend bei $X = x_2$: $f_2 = 1 - \sum_j (p_{2j}/p_{2.})^2 = 0{,}689$ usw.

($f_3 = 0{,}7325$)

Der gesamte Fehler ist dann:

$F_2 = \Sigma_i p_{i.} - f_i = \Sigma_i p_{i.}[1 - \Sigma_j (p_{ij}/p_{i.})^2] = 1 - \Sigma_j (p_{ij})^2/p_i)$.

(im Beispiel $F_2 = 0{,}67737$).

Zu 3:

Konzept der proportionalen (relativen) Fehlerreduktion
a) optimale Vorhersage [λ] b) proportionale Vorhersage [τ]

Ergebnis:

λ_{xy}	= (E1 - E2)/E1	τ_{xy}	= (F1 - F2)/F1
	= (0,64 - 0,58)/0,64		= (0,7314 - 0,67737)/0,7314
	= 0,09375		= 0,07387

Def. 7.14: Kontingenzmaße von Goodman-Kruskal

Nach Umformungen erhält man für die asymmetrischen Kontingenzmaße λ und τ mit den relativen Häufigkeiten $p_{ij} = n_{ij}/n$ und mit den absoluten Häufigkeiten

a) Koeffizient der optimalen Vorhersage λ

aa) Vorhersage von Y durch X

$$(7.38) \qquad \lambda_{xy} = \frac{\Sigma_i \max_j(p_{ij}) - \max_j(p_{\cdot j})}{1 - \max_j(p_{\cdot j})} = \frac{\Sigma_i \max_j(n_{ij}) - \max_j(n_{\cdot j})}{n - \max_j(n_{\cdot j})}$$

ab) Vorhersage von X durch Y

$$(7.38a) \qquad \lambda_{yx} = \frac{\Sigma_j \max_i(p_{ij}) - \max_i(p_{i\cdot})}{1 - \max_i(p_{i\cdot})} = \frac{\Sigma_j \max_i(n_{ij}) - \max_i(n_{i\cdot})}{n - \max_i(n_{i\cdot})}$$

b) Koeffizient der proportionalen Vorhersage τ

ba) Vorhersage von Y durch X

$$(7.39) \tau_{xy} = \frac{\Sigma\Sigma p_{ij}^2 / p_{i\cdot} - \Sigma p_{\cdot j}^2}{1 - \Sigma p_{\cdot j}^2} = \frac{n\Sigma\Sigma n_{ij}^2 / n_{i\cdot} - \Sigma n_{\cdot j}^2}{n^2 - \Sigma n_{\cdot j}^2}$$

bb) Vorhersage von X durch Y

$$(7.39a) \tau_{yx} = \frac{\Sigma\Sigma p_{ij}^2 / p_{\cdot j} - \Sigma p_{i\cdot}^2}{1 - \Sigma p_{i\cdot}^2} = \frac{n\Sigma\Sigma n_{ij}^2 / n_{\cdot j} - \Sigma n_{i\cdot}^2}{n^2 - \Sigma n_{i\cdot}^2}$$

Bemerkungen zu Def. 7.14:

1. Es ist offensichtlich, dass λ und τ asymmetrisch sind, also die Abhängigkeit nicht vertauscht werden darf.

   Im Beispiel 7.16 erhält man
   $\lambda_{xy} = 0{,}09375$ und $\tau_{xy} = 0{,}07387$ aber
   $\lambda_{yx} = 0{,}26563$ und $\tau_{yx} = 0{,}11601$
   (Zwischenergebnisse bei λ_{yx}: E1 = 0,64 und E2 = 0,47

bzw. bei τ_{yx} F1 = 0,6418 und F2 = 0,56734)

2. Aus Gl. 7.39 folgt unmittelbar, dass τ bei Unabhängigkeit (wenn $p_{ij} = p_{i.}p_{.j}$ für alle Werte von i und j) den Wert Null annimmt. Man kann zeigen, dass dies auch für λ der Fall ist. Unabhängigkeit ist aber nur eine notwendige, nicht eine hinreichende Bedingung für das Verschwinden von λ.

   Die folgende Kontingenztafel

	y_1	y_2	y_3	Σ
x_1	20	40	10	70
x_2	10	15	5	30
Σ	30	55	15	100

   führt zu $\lambda_{xy} = \lambda_{yx} = 0$ obgleich keine Unabhängigkeit vorliegt (übrigens ist τ_{xy} = 0,002849 und τ_{yx} = 0,00433 also nicht Null).

   Das Beispiel ist so konstruiert, dass alle Zeilenmaxima in Spalte 2 und alle Spaltenmaxima in Zeile 1 sind.

3. Es ist leicht zu sehen, dass λ und τ dann den Wert 1 annehmen, wenn jede Zeile und jede Spalte einer quadratischen Kontingenztafel mit nur einer Häufigkeit besetzt ist, wie dies in Tafel 1 von Bsp. 7.15 der Fall ist [was oben als "vollständige Abhängigkeit" bezeichnet wurde].

4. Zu weiteren Eigenschaften von τ und λ (im Spezialfall der Assoziationsanalyse) vgl. Bem. 8 zu Def. 7.16.

5. Als ein Maß der proportionalen (relativen) Fehlerreduktion könnte man auch interpretieren:

   - das **Quadrat des Korrelationsverhältnisses**, wobei als Fehler E1 (oder F1) die Summe der Abweichungsquadrate SAQ_{tot} und als Fehler E2 (oder F2) SAQ_{int} auftritt;
   - das **Bestimmtheitsmaß** (im Falle linearer Regression) $B_{xy} = r_{xy}^2$.

   Das legt den Gedanken nahe, dass nicht λ (τ) sondern die Wurzel von λ (bzw. τ) eigentlich das Kontingenzmaß sein sollte.

3. Assoziation Vierfelderkorrelation

a) Dichotome Merkmale

Von Assoziation spricht man, wenn X und Y dichotome (binäre) Merkmale sind. Man kann dann wie folgt codieren:

$$x = \begin{cases} 0 & \text{wenn eine bestimmte Eigenschaft nicht vorhanden ist} \\ 1 & \text{wenn diese Eigenschaft (das Attribut) vorhanden ist} \end{cases}$$

und entsprechend ist Y als 0-1-Variable gegeben (d.h. es gibt nur die beiden Ausprägungen y=0 und y=1).

Man kann auch codieren: ja (1), nein (0) oder "+" (für 1) und "-" (für 0), um anzuzeigen, dass lediglich danach unterschieden wird, ob eine Eigenschaft (ein "Attribut") gegeben ist oder nicht gegeben ist.

Def. 7.15: Vierfeldertafel

Die zeilenweise (spaltenweise) Anordnung des dichotomen Merkmals X (Y) mit den Häufigkeiten a,b,c,d (statt $n_{11},n_{12},n_{21},n_{22}$) in der folgenden Art

	Variable Y		
Variable. X	y=1	y=0	Σ
x=1	a	b	a+b
x=0	c	d	c+d
Σ	a+c	b+d	n

heißt Vierfeldertafel (oder Assoziationstafel; vgl. Bsp. 7.1 für Tafeln dieser Art).

b) Bedingte Mittelwerte, Unabhängigkeit, Regressionsgeraden

1. Mit der Codierung als 0-1-Variablen sind Mittelwerte und bedingte Mittelwerte als relative Häufigkeiten (bzw. bedingte relative Häufigkeiten) zu interpretieren:

$$(7.40) \quad \bar{x} = \frac{1\cdot(a+b) + 0\cdot(c+d)}{n} = \frac{a+b}{n}$$

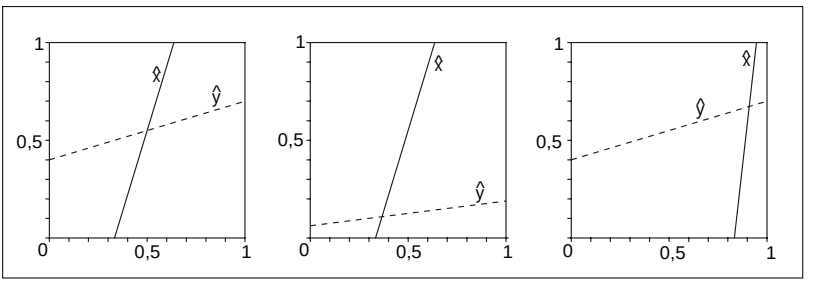
$$(7.41) \quad \bar{x}(y_1) = a/(a+c) \text{ und } \bar{x}(y_2) = b/(b+d)$$

   Entsprechendes gilt für den unbedingten Mittelwert $\bar{y}$ und für die bedingten Mittelwerte $\bar{y}(x_i)$.

2. Die Punkte $Px_1[a/(a+c),1]$ und $Px_2[b/(b+d),0]$ stellen im x,y-Koordinatensystem die bedingten Mittelwerte $\bar{x}(y_j)$ (mit j=1,2) dar. Ihre lineare Verbindung stellt die Regressionslinie zur Schätzung von x aufgrund

von y dar. Weil es sich nur um zwei Punkte handelt, ist es eine **Regressionsgerade** und es lässt sich zeigen, dass diese mit der im Kap. 8 eingeführten Regressionsgeraden $\hat{x}$ identisch ist. Entsprechend ist die Regressionsgerade $\hat{y}$ durch die Punkte $Py_1[1,a/(a+b)]$ und $Py_2[0,c/(c+d)]$ gegeben. In Abb. 7.5 sind die Regressionsgeraden für ein Zahlenbeispiel dargestellt.

Abb. 7.5: Regressionsgeraden für das Beispiel 7.17



3. Die Unterschiedlichkeit der bedingten Mittelwerte ist Ausdruck der Abhängigkeit der Merkmale untereinander. Bei Unabhängigkeit (Def. 7.5) gilt:

$$\bar{x}(y_1) = \bar{x}(y_2) = \bar{x} \text{ und } \bar{y}(x_1) = \bar{y}(x_2) = \bar{y}.$$

Hieraus folgt als Bedingung für die Unabhängigkeit

(7.42) $ad = bc.$

Weil die Merkmale X und Y hier nur zwei Ausprägungen haben, kann die Abhängigkeit nur linear sein, weshalb hier zwischen Unabhängigkeit und Unkorreliertheit (keine Assoziation) nicht unterschieden werden kann.

c) Normierung eines Assoziationsmaßes

Ein Assoziationsmaß A soll bei Unabhängigkeit den Wert Null annehmen. Die Obergrenze des Betrags von A ist dagegen nicht eindeutig. Besteht zwischen den Ausprägungen eine Ordnungsrelation etwa dergestalt, dass $x_1 > x_2$ und $y_1 > y_2$ (Dichotomien mit dahinterstehender Rangordnung), so kann auch sinnvoll zwischen positiver und negativer Assoziation unterschieden werden. Es sollte dann gelten:

(7.43) $-1 \le A \le +1.$

Man kann dann unterscheiden (nach M.G.Kendall):

total association	$b = c = 0$
total disassociation	$a = d = 0$
complete association	$b = 0$ oder $c = 0$

complete disassociation $a = 0$ oder $d = 0$

Es soll im folgenden von totaler bzw. vollständiger Assoziation bzw. Disassoziation gesprochen werden.

- Bei totaler Assoziation bzw. Disassoziation nehmen jeweils **beide** bedingten Mittelwerte (bedingte relative Häufigkeiten) einer Regressionsgeraden die Werte Null und Eins an.
- Bei vollständiger Assoziation bzw. bei vollständiger Disassoziation nimmt jeweils nur einer der beiden Endpunkte einer Regressionsgeraden die extremen Werte Null und Eins an.

Die Funktionen der beiden "Regressionsgeraden" lauten:

(7.44) $$\hat{y} = \frac{c}{c+d} + \left[\frac{a}{a+b} - \frac{c}{c+d}\right] x$$

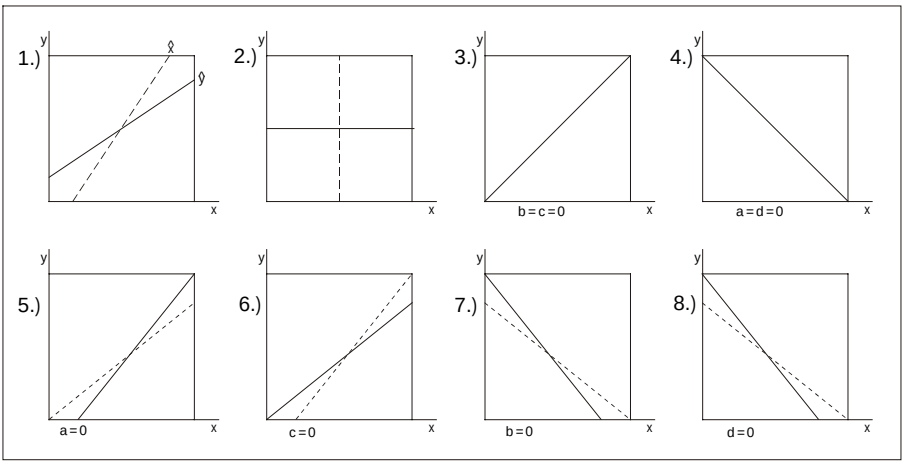
und

(7.45) $$\hat{x} = \frac{b}{b+d} + \left[\frac{a}{a+c} - \frac{b}{b+d}\right] y.$$

Zur Gestalt der Regressionsgeraden in den beschriebenen Fällen vergleiche man Abb. 7.6. Dabei zeigt sich, dass es durchaus sinnvoll ist, totale und vollständige Assoziation/Disassoziation als unterschiedlich anzusehen.

Abb.7.6: Dichotome "Regression"

1: allgemeiner Fall; 2: Unabhängigkeit; 3: totale Assoziation 4: totale Disassoziation, 5/6: vollständige Assoziation; 7/8: vollständige Disassoziation



d) Axiomatik für Assoziations- und Kontingenzmaße

Die unter c) getroffenen Aussagen über den Wertebereich eines Assoziationsmaßes führen zu dem Gedanken, auch hier eine Axiomatik anzugeben (vgl. Seite 222f). Für das Assoziationsmaß A sollte speziell die Forderung A2 noch wie folgt spezifiziert werden:

A2: Ist totale Assoziation gegeben, so sollte A = +1 sein. Bei geordneten Ausprägungen $x_1 > x_2$ und $y_1 > y_2$ sollte auch Assoziation und Disassoziation unterscheidbar sein und bei totaler Disassoziation den Wert A = -1 annehmen.

Die Axiome A1, A3 und A4 (vgl. Seite 222f) sind auch bei Kontingenzmaßen sinnvoll zu fordern. Ein Assoziationsmaß kann sehr wohl invariant sein gegenüber einer Ver-k-fachung aller Häufigkeiten, gleichwohl aber A3 nicht erfüllen (Das gilt z.B. für das Maß Q, vgl. Gl. 7.50).

Def. 7.16: Einige Assoziationsmaße

a) Die in Def. 7.13 zusammengestellten Kontingenzmaße haben im speziellen Fall einer Vierfeldertafel die folgende Gestalt:

$$(7.46)\quad \phi = \frac{ab - cd}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = \sqrt{\frac{c^2}{n}}$$

heisst Phi-Koeffizient oder Vierfelderkorrelation.
Es gilt $\phi = T = V$. Das Kontingenzmaß von Pearson ist speziell bei zwei dichotomen Merkmalen mit $C = \sqrt{\frac{f^2}{f^2+1}}$ kein sinnvoller Ausdruck.

b) Vorgeschlagen wurde auch

(7.47)	$\lvert xy\rvert = (ad - bc)/n^2$	Kreuzprodukt von Lazarsfeld
(7.48)	$cpr = ad/bc$	Kreuzproduktverhältnis
(7.49)	$\delta_x = a/(a+c) - b/(b+d)$	und
(7.49a)	$\delta_y = a/(a+b) - c/(c+d)$	(die sog. Anteilsdifferenzen)

Offenbar ist $\phi = \sqrt{d_x d_y}$ und die Anteilsdifferenzen δ_x und δ_y sind die Steigungen der Regressionsgeraden $\hat{x}$ und $\hat{y}$.

c) Auf dem Vergleich konkodanter und diskonkodanter Merkmalskombinationen (Paare) beruht das Assoziationsmaß Q von Yule

$$(7.50)\quad Q = \frac{ad - bc}{ad - bc} = \frac{cpr - 1}{cpr + 1} \quad .$$

sowie der Verbundenheitskoeffizient von Yule (coefficient of colligation):

$$(7.50a)\quad Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \quad ,$$

der nur von geringer Bedeutung ist, da er nur eine monotone Transformation von Q darstellt.

d) Ohne Bedeutung für die Assoziationsmessung sind die in Def. 7.14 definierten Kontingenzmaße von Goodman und Kruskal.

Bemerkungen zu Def. 7.16:

1. Bei einer Codierung von X und Y mit den Variablenwerten 0 und 1 ist das Kreuzprodukt |xy| die Kovarianz zwischen X und Y. Für Varianzen gilt $s_x^2 = (a+b)(c+d)/n^2$ und $s_y^2 = (a+c)(b+d)/n^2$, so dass ϕ nichts anderes ist als der Produkt-Moment-Korrelationskoeffizient r_{xy} für den Fall einer 0-1-Codierung von X und Y.

2. Bedeutet x = 1 (y = 1) das Auftreten des Ereignisses A (bzw. B) so ist |xy| die Abweichung von der stochastischen Unabhängigkeit |xy| = |AB| = P(AB) - P(A)P(B). Als Kovarianz hat das Kreuzprodukt den Wertebereich $-1/4 \leq |xy| \leq +1/4$. Die Schreibweise |xy| oder |AB| beruht darauf, dass diese Größe die Determinante der Vierfeldertafel (mit relativen statt absoluten Häufigkeiten, bzw. mit Wahrscheinlichkeiten) darstellt. Das Kreuzprodukt spielt in der "Latent Structure Analysis" eine große Rolle.

3. Yules Assoziationsmaß Q erfüllt die Axiome A1 und A4. Es nimmt aber die extremen Werte +1 bzw. -1 nicht nur im Falle totaler, sondern auch vollständiger Assoziation bzw. Disassoziation an. Q erfüllt also A2 nicht (vollständig) und auch A3 nicht.
   Die drei Vierfeldertafeln von Beispiel 7.17 führen alle zum gleichen Wert von Q = 5/9 und auch zum gleichen Kreuzproduktverhältnis cpr = 3,5, obgleich die Phi-Koeffizienten unterschiedlich sind.

4. Aus Gl. 7.46 folgt, dass bei Ver-k-fachung aller Häufigkeiten χ^2 auch k-mal so groß ist wie bisher, so dass χ^2 das Axiom A4 nicht erfüllt. χ^2 hat jedoch sehr günstige Aggregationseigenschaften, weil es in beliebige Teilsummen aufzuspalten ist. Außerdem sind χ^2 und alle auf χ^2 basierenden Maße invariant gegenüber Vertauschungen von Zeilen und Spalten.

5. Die mit den Axiomen A1 und A2 geforderte Normierung des Wertebereichs wird vom Kreuzprodukt nicht erfüllt: Die Grenzen des Intervalls $-1/4 \leq |xy| \leq +1/4$ werden nur erreicht bei totaler Assoziation **und** wenn die Varianzen von X und Y jeweils 1/4 (also maximal) sind, also a=d bzw. b=c beträgt.

6. Zum Kreuzproduktverhältnis: Es ist nichtnegativ und es gilt $0 \leq cpr \leq 1$ bei Disassoziation, cpr = 1 bei Unabhängigkeit und $1 < cpr < \infty$ bei Assoziation. Schon bei vollständiger Disassoziation ist cpr = 0.

7. Wegen der Abhängigkeit von χ^2 gilt für C die Einschrankung: $0 \le C \le \sqrt{1/2}$, mit der Obergrenze $\sqrt{1/2}$ bei totaler Assoziation oder Disassoziation.

8. Überträgt man die Kontingenzmaße von Goodman und Kruskal (Def. 7.14) auf den speziellen Fall der Assoziationsanalyse (r=c=2) so ergeben sich komplizierte Ausdrücke. Die Maße haben auch erhebliche Nachteile: λ kann auch ohne Unabhängigkeit den Wert Null annehmen, andererseits sind λ und τ zwar bei totaler, nicht aber bei vollständiger Assoziation 1, wie die folgenden zwei Tabellen zeigen:

	$y_1=1$	$y_2=0$	Σ
$x_1=1$	4	2	6
$x_2=0$	3	1	4
Σ	7	3	10

$\lambda_{xy} = \lambda_{yx} = 0$
dagegen $\phi = 0{,}0891$ und
$\tau_{xy} = 0{,}00794 = 1/126 = \tau_{yx}$

	$y_1=1$	$y_2=0$	Σ
$x_1=1$	7	1	8
$x_2=0$	0	2	2
Σ	7	3	10

$\lambda_{xy} = 2/3$, $\lambda_{yx} = 1/2$
(vollständige Assoziation)
$\tau_{xy} = 0{,}58333 = 7/12 = \tau_{yx}$

Man kann zeigen, dass τ bei einer Vierfeldertafel stets symmetrisch ist.

Beispiel 7.17:

Man berechne die Assoziationsmaße der Def. 7.16 (a-c) sowie die Regressionsgeraden für die folgenden drei Vierfeldertafeln:

Tafel 1

	y=1	y=0	Σ
x=1	70	30	100
x=0	40	60	100
Σ	110	90	200

Tafel 2

	y=1	y=0	Σ
x=1	7	30	37
x=0	4	60	64
Σ	11	90	101

Tafel 3

	y=1	y=0	Σ
x=1	70	30	100
x=0	4	6	10
Σ	74	36	110

Welche Besonderheiten gelten für das Assoziationsmaß Q sowie das Kreuzproduktverhältnis (cpr)? Wie reagiert χ^2 bei einer Verdoppelung aller Häufigkeiten der Tafel 1?

Lösung 7.17:

1) Als Muster jeweils die Berechnung für Tafel 1
Q = (70·60-30·40)/(70·60+30·40) = (4200-1200)/(4200+1200) = 5/9
cpr = 4200/1200 = 3,5
Phi-Koeffizient $\phi = (70\cdot60-30\cdot40)/\sqrt{100\cdot100\cdot110\cdot90} = 0{,}30151$
$\chi^2 = n\phi^2 = 200/11 = 18{,}182$ und $C = \sqrt{1/12} = 0{,}288675$
Regressionsfunktionen $\hat{y} = 0{,}4 + 0{,}3x$ und $\hat{x} = 1/3 + (10/33)y$
daraus folgt: Anteilswertdifferenzen $\delta_x = 10/33$ und $\delta_y = 0{,}3$ sowie für den Phi-Koeffizienten $\phi = \sqrt{d_x\, d_y} = \sqrt{30/330} = 0{,}30151$.

2) Berechnungen für Tafel 2 und Tafel 3:
$Q_2 = Q_3 = 5/9$ $(= Q_1)$

$cpr_2 = cpr_3 = 3{,}5$ (= cpr_1).

Es fällt also auf, dass Q und cpr invariant sind gegenüber proportionalen Transformationen einzelner Zeilen und Spalten, was für die übrigen Assoziationsmaße nicht gilt:

$\phi_2 = 0{,}195935$ ($\neq \phi_1 = 0{,}3015$) $\phi_3 = 0{,}183804$

$\chi^2_2 = 3{,}87746$, $\chi^2_3 = 3{,}71622$ und $C_2 = 0{,}192289$, $C_3 = 0{,}18078$

Regressionsgeraden und Anteilsdifferenzen:

Tafel 2	Tafel 3
$\hat{y} = 0{,}0625 + 0{,}1267x$	$\hat{y} = 0{,}4 + 0{,}3x$
$\hat{x} = 1/3 + (10/33)y$	$\hat{x} = 5/6 + 0{,}1126y$
$\delta_{y2} = 0{,}1267$ $\delta_{x2} = 10/33$	$\delta_{y3} = 0{,}3$ $\delta_{x3} = 0{,}1126$

3) Veränderung von χ^2 bei Verdoppelung der Häufigkeiten:

Tafel 1

	y=1	y=0	Σ
x=1	70	30	100
x=0	40	60	100
Σ	110	90	200

$\chi^2 = 200/11 = 18{,}1818$

$\phi^2 = 1/11$

Tafel 1a

	y=1	y=0	Σ
x=1	140	60	200
x=0	80	120	200
Σ	220	180	400

$\chi^2 = 400/11 = 36{,}3636$

$\phi^2 = 1/11$

Verdoppelung der Häufigkeiten führt auch zu einer Verdoppelung von χ^2, lässt aber den Phi-Koeffizienten unberührt.

Beispiel 7.18:

Man berechne ϕ sowie Goodman Kruskals λ für die folgenden Vierfeldertafeln!

	$y_1=1$	$y_2=0$	Σ
$x_1=1$	4	6	10
$x_2=0$	7	5	12
Σ	11	11	22

	$y_1=1$	$y_2=0$	Σ
$x_1=1$	4	7	11
$x_2=0$	6	5	11
Σ	10	12	22

Lösung 7.18:

Man erkennt, dass die rechte Tafel nur die transponierte linke Tafel ist. Für λ erhält man für die linke Tafel $\lambda_{xy} = 1/30 = 0{,}0333$ und $\lambda_{yx} = 1/30$. λ_{xy} der linken Tafel ist gleich λ_{yx} der rechten Tafel. Man erhält also stets den gleichen Wert 1/30 für λ. Das gleiche gilt für ϕ, das ohnehin symmetrisch und deshalb für die linke und rechte Tafel gleich ist, nämlich $\phi = 0{,}18257$.

c) Rangkorrelation, Zusammenhang bei ordinalskalierten Variablen

1. Rangkorrelation

Verzichtet man auf eine Metrik und nutzt nur die Ranginformation in den Tupeln (x_v, y_v) oder liegen nur ordinalskalierte Variablen vor, die in eine Rangskala transformiert werden können, so lassen sich Rangkorrelationskoeffizienten berechnen. Man kann dann nicht mehr von einem linearen Zusammenhang sprechen, sondern von einem Zusammenhang, der (bei Rangkorrelation) linear in den Rängen $R(x_v)$, $R(y_v)$ ist. Bei der Rangassoziation werden keine Ränge vergeben (Def. 7.17), der Zusammenhang ist dann monoton.

Def. 7.17: Rangtransformation

Der zweidimensionalen Variable (X,Y) mit den der Größe nach geordneten Ausprägungen x_v, y_v werden Rangzahlen mit $R(x_v) = v$ und $R(y_v) = v$ zugeordnet.

Bemerkungen zu Def. 7.17:

1. Die Definition der Rangzahlen setzt voraus, dass X und Y mindestens ordinalskaliert sind und alle Ausprägungen verschieden sind, so dass $x_{(v-1)} < x_{(v)} < x_{(v+1)}$ (und die Ordnung der y-Werte entsprechend definiert ist). Die Rangzahlen (Ränge) sind natürliche Zahlen.

2. Sind die Ausprägungen nicht jeweils alle verschieden, d.h. treten **Bindungen** (ties) auf, so wird jeweils k gleichen Werten das arithmetische Mittel der auf sie entfallenden Rangzahlen zugeordnet.

   Beispiel: Bei $x_{(1)} < x_{(2)} < x_{(3)} = x_{(4)} = x_{(5)} < x_{(6)} = x_{(7)}$ erhalten die 3te bis 5te Ausprägung alle den Rangplatz 4 und die 6te und 7te Ausprägung den Rang 6,5. Die Reihe der Rangzahlen lautet also 1, 2, 4, 4, 4, 6½, 6½. Diese Art Bindungen zu behandeln verändert die Rangsumme $\Sigma R(x_v)$ (und $\Sigma R(y_v)$ entsprechend) nicht, wohl aber die Summe der quadrierten Ränge $\Sigma[R(x_v)]^2$. Sie ist bei Auftreten von Bindungen kleiner als bei Ausprägungen, die alle unterschiedlich sind. Denn in $\Sigma[R(x_v)]^2$ tritt im Falle n gleicher Werte n-mal die Größe $\bar{x}^2$ auf und bei n unterschiedlichen Werten dagegen die Größe Σx_j^2 (j = 1,2,...,n). Es gilt als Konsequenz des Verschiebungssatzes der Varianz $\Sigma x_j^2 > n\bar{x}^2$, denn $n^{-1}\Sigma x_j^2 - \bar{x}^2 = s_x^2 > 0$.

Def. 7.18: Rangkorrelationskoeffizient nach Spearman

Der Korrelationskoeffizient nach Bravais-Pearson für Rangzahlen

$$(7.51)\quad R_{xy}^{S} = 1 - \frac{6\Sigma d_v^2}{n(n^2-1)} \qquad \text{mit } d_v = R(x_v) - R(y_v)$$

heisst Rangkorrelationskoeffizient nach Spearman (v = 1,2,...,n).

Bemerkungen zu Def. 7.18:

1. *Zur Größe d:* Hat z.B. die Einheit v (die v-te Beobachtung) bezüglich des Merkmals X den 5ten Platz und bei Y den 3ten Rang, so ist $R(x_v) = 5$ und $R(y_v) = 3$, so dass $d_v = 5 - 3 = 2$ ist.

2. *Zusammenhang mit Produkt-Moment-Korrelation:* Da die Ränge $R(x_1),...,R(x_n)$ natürliche Zahlen sind, gilt $\Sigma R(x_v) = 1 + 2 + ... + n = n(n+1)/2$ (v = 1,2,...,n). Den gleichen Ausdruck erhält man für $\Sigma R(y_v)$. Die arithmetischen Mittel sind also $\overline{R(x)} = \overline{R(y)} = (n+1)/2$. Das gilt auch, wenn Bindungen auftreten. Ferner gilt: $\Sigma[R(x_v)]^2 = \Sigma[R(y_v)]^2 = n(n+1)(2n+1)/6$, so dass man für die Varianzen $s^2_{R(x)}$ und $s^2_{R(y)}$ der Rang-reihen R(x), R(y) den Ausdruck $(n^2-1)/12$ erhält.

   Treten Bindungen auf, so werden die Varianzen geringer sein (vgl. Bem. 2 zu Def. 7.17).

   Nach Satz 7.4 ist die Kovarianz zwischen den Rängen $(n^2-1)/12 - (\Sigma d^2{}_v)/2n$, woraus sich mit Def. 7.8 (Gl. 7.20) für die Korrelation r_{xy} zwischen den Rängen Gl. 7.51 er-gibt. Der Rangkorrelationskoeffizient von Spearman ist also die Produkt-Moment-Korrelation zwischen den Rangzahlen $R(x_v)$, $R(y_v)$.

3. *Skalen*: Mit den Rangplätzen (Rängen) wird gerechnet wie mit metrisch skalierten Variablen, was eigentlich voraussetzt, dass die Abstände zwischen den Rängen gleich groß sind. Es wird also angenommen, dass zwar nicht die Ausprägungen der zu korrelierenden Merkmale X und Y selber, wohl aber die hierfür vergebenen Rangplätze metrisch skaliert sind.

4. Im Falle von Bindungen werden im Nenner $n(n^2-1)$ von Gl. 7.51 Korrekturen angebracht. Der Rangkorrelationskoeffizient ist dann gleichwohl meist kleiner als im Falle ohne Bindungen.

Axiomatik

Es ist von einem Rangkorrelationskoeffizienten R zu fordern:

R1: $-1 \le R \le 1$

R2: Bei vollständiger Übereinstimmung der Rangreihen, also $R(x_v) = R(y_v)$ für alle v = 1,2,...,n, soll R = +1 sein.

R3: Bei inverser Rangordnung $R(y_v) = (n+1) - R(x_v)$ soll R = -1 sein.

R4: Bei Unabhängigkeit von X und Y soll R = 0 gelten.

R5: Rangkorrelationskoeffizienten sollten invariant sein gegenüber monoton steigenden Transformationen der Rangzahlen.

Weitere Bemerkungen zu Spearmans Rangkorrelation

Man sieht leicht, dass R^s die Axiome R1 bis R3 erfüllt. Bei Übereinstimmung ist $d_v = 0$ für alle v = 1,2,...,n, so dass $\Sigma d_v^2 = 0$ und $R^s = 1$. Bei inverser Rangordnung ist $d_v = (n+1) - 2R(x_v)$, so dass $\Sigma d_v^2 = -n(n+1)^2 + 2n(n+1)(2n+1)/3 = n(n^2-1)/3$. Dies eingesetzt in Gl. 7.51 liefert $R^s = -1$. Ist $R^s = 0$, so ist der Mittelwert der Rangdifferenzen gleich der Summe der Varianzen, d.h. es gilt $(\Sigma d_v^2)/n = (n^2 -1)/6$ und die Kovarianz zwischen den Rängen ist Null. Wie man auch sieht, ist $(n^2-1)/6$ der halbe Betrag dessen, was sich für $(\Sigma d_v^2)/n$ bei $R^s = -1$ ergibt. Während die Situationen $R^s = +1$ und $R^s = -1$ eindeutig sind, ist $R^s = 0$ mit verschiedenen Konstellationen verträglich (vgl. Beispiel 7.19). Spearmans Rangkorrelationskoeffizient erfüllt auch nicht das Axiom R5. Eine Verdoppelung aller Rangzahlen führt zu einer Vervierfachung der Σd_v^2, wodurch sich R^s verändert (wobei R^s dann nicht mehr zwischen -1 und +1 liegen muss; die Ränge sind dann ja auch nicht mehr, wie in Def. 7.18 vorausgesetzt, die Folge natürlicher Zahlen).

Beispiel 7.19:

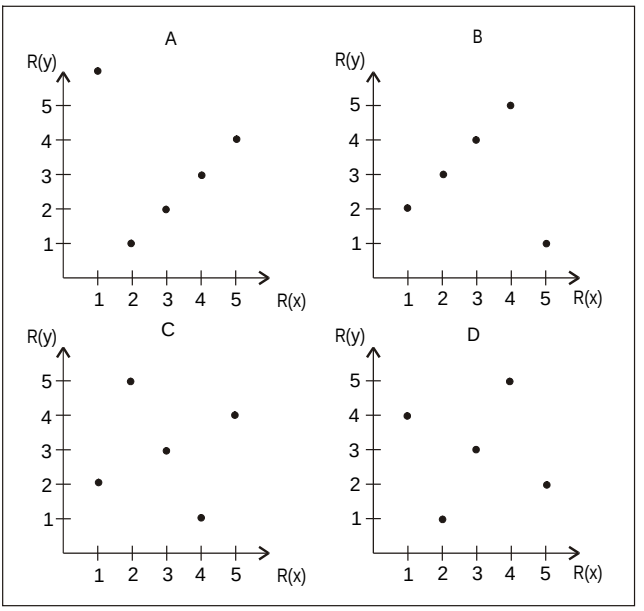
Man zeige, dass bei gegebener Rangfolge der X-Werte die folgenden vier Rangordnungen der Y-Werte jeweils zu $R^s = 0$ führen.

R(x)	Rangordnungen von y			
	A	B	C	D
1	5	2	2	4
2	1	3	5	1
3	2	4	3	3
4	3	5	1	5
5	4	1	4	2

Lösung 7.19:

In allen vier Fällen (A bis D)- vgl. Abb.7.7 - ist die Summe der quadrierten Rangdifferenzen 20, so dass bei n=5 gilt: $R^s = 1 - \frac{6 \cdot 20}{5(25-1)} = 1 - 1 = 0$.

Abb. 7.7: Fälle mit einer Rangkorrelation von R^S = 0 (Bsp 7.19)



2. Paarvergleiche, Bindungen, Rangassoziation

Die im Teil 1 betrachteten Daten waren vom folgenden Typ:

- Einzelbeobachtungen (n Einheiten),
- den Merkmalswerten werden Ränge zugeordnet (Rangtransformation Def. 7.17) und sie
- sind i.d.R. alle unterschiedlich, so dass die rangtransformierten Merkmalswerte $R(x_v)$, $R(y_v)$ natürliche Zahlen von 1 bis n sind,
- nur als Ausnahme gibt es auch "Bindungen" (Bem. 2 zu Def. 7.17).

Die Merkmale X und Y sind ordinalskaliert, aber durch die **Rangtransformation** entsteht mehr als eine Ordinalskala: Ränge sind äquidistant, die Merkmalsausprägungen, für die diese Ränge vergeben werden, sind es aber nicht. Es ist zweifelhaft, ob eine solche Anhebung des Skalenniveaus vertretbar ist. Wünschenswert ist demgegenüber

1. eine Datenanalyse ordinaler Merkmale, die nur die Information einer Ordinalskala ausnutzt, d.h. nur dass bei $x_i > x_j$ die Ausprägung x_i des Merkmals X "größer", "höher" o.ä. ist als x_j,
2. die Berücksichtigung häufigen (nicht nur als Ausnahme) Auftretens gleicher Merkmalsausprägungen (also von Bindungen).

Soll ohne Rangtransformationen "Zusammenhang" bei ordinalen Merkmalen definiert werden (Rangassoziation statt Rangkorrelation), so ist vom Paarvergleich auszugehen.

Def. 7.19: Paarvergleiche

a) Die folgende Tabelle soll ***Kontingenztabelle*** genannt werden, obgleich dieser Begriff speziell häufig für nominalskalierte Merkmale benutzt wird. Die Anzahl der Zeilen und Spalten ist nur beispielhaft. Die Zusammenhänge gelten allgemein:

	y_1	y_2	y_3	Σ
x_1	n_{11}	n_{12}	n_{13}	$n_{1.}$
x_2	n_{21}	n_{22}	n_{23}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Es gilt $y_1 > y_2 > y_3$ und $x_1 > x_2$. Die n Einheiten lassen n(n-1)/2 Paarvergleiche zu. Alle n_{ij} Einheiten sind jeweils bezüglich X **und** Y gleich (Bindungen).

b) ***Bindungen (Verknüpfungen, ties)***: Einheiten sind verknüpft, wenn sie bezüglich eines Merkmals oder beider Merkmale gleiche Ausprägungen haben.

c) Jeder der n(n-1)/2 Paarvergleiche ist ein Vergleich im Sinne eines der folgenden fünf Typen:

1. Konkordante Vergleiche:
n_{ij} Einheiten mit $X = x_i$ und $Y = y_j$ werden verglichen mit allen Einheiten, für die gilt: $X < x_i$ und $Y < y_j$. Die Anzahl der konkordanten Paarvergleiche ist $N_c = n_{11}(n_{22} + n_{23}) + n_{12}n_{23}$.

2. Diskonkordante Vergleiche:
N_d ist die Anzahl der Vergleiche von jeweils n_{ij} Einheiten mit solchen Einheiten, bei denen $X > x_i$ und $Y < y_j$. Für die Tabelle gilt
$N_d = n_{13}(n_{21} + n_{22}) + n_{12}n_{21}$.

3. Es gibt T_x in X verknüpften Vergleichen
$T_x = n_{11}(n_{12} + n_{13}) + n_{12}n_{13} + n_{21}(n_{22} + n_{23}) + n_{22}n_{23}$.

4. Bei T_y-Vergleichen liegen Bindungen bezüglich Y vor:
(Vergleiche mit den restlichen Einheiten einer Spalte):
$T_y = n_{11}\,n_{21} + n_{12}\,n_{22} + n_{13}n_{23}$.

5. Verknüpft in X und Y
sind T_{xy}-Vergleiche:
$T_{xy} = 1/2\ \Sigma\Sigma n_{ij}(n_{ij}-1) = 1/2[n_{11}(n_{11}-1) + n_{12}(n_{12}-1 + ... + n_{23}(n_{23}-1)]$.

d) Es gilt

$$n(n-1)/2 = N_c + N_d + T_x + T_y + T_{xy}. \quad (7.52)$$

Bemerkungen zu Def. 7.19:

1. Man kann N_c auch bestimmen durch Vergleiche mit Einheiten, für die gilt $X > x_i$ und $Y > y_j$ statt mit Einheiten, für die gilt $X < x_i$ und $Y < y_i$: Die Summen $n_{11}(n_{22} + n_{23}) + n_{12}n_{23}$ und $n_{23}(n_{12} + n_{11}) + n_{22}n_{11}$ sind gleich. Bei einer Vierfeldertafel ist $N_c = ad$.

2. Auch hier könnte man alternativ definieren: Vergleiche mit Einheiten, bei denen $X < x_i$ und $Y > y_j$. Man erhält N_d auch mit $n_{21}(n_{12} + n_{13}) + n_{22}n_{13} = n_{13}(n_{21} + n_{22}) + n_{12}n_{21}$. Bei einer Vierfeldertafel ist $N_d = bc$.

3. Bei einer Vierfeldertafel ist $T_x = ab + cd$ und $T_y = ac + bd$.

Def. 7.20: Maße der Rangassoziation

Auf der Anzahl konkordanter und diskonkordanter Paarvergleiche beruhen die folgenden Maße der Rangassoziation.

(7.53) $\gamma = (N_c - N_d)/(N_c + N_d)$ index of order association von Goodman/Kruskal.
[Es wird auch ω als Symbol verwendet; bei Vierfeldertafeln sind γ und Yules Q identisch].

(7.54a) $\tau_a = \frac{2(N_c - N_d)}{n(n-1)} = \frac{N_c - N_d}{\binom{n}{2}}$ (Kendalls τ) oder in einer anderen Version

(7.54b) $\tau_b = \frac{N_c - N_d}{\sqrt{(N_c+N_d+T_x)(N_c+N_d+T_y)}}$

[bei Vierfeldertafeln ist dies gleich dem Phi-Koeffizient, vgl. Def. 7.16, Gl. 7.46].

Bemerkungen zu Definition 7.20:

1. Sämtliche Werte in einem Tabellenfeld stellen untereinander verglichen Bindungen in X **und** Y dar. Es gilt bei $n = \Sigma\Sigma n_{ii}$, also bei einer symmetrischen Tabelle, die nur in der Hauptdiagonalen besetzt ist: $T_{xy} = \Sigma\Sigma[n_{ii}(n_{ii}-1)]/2$, so dass $n(n-1)/2 = N_c + T_{xy}$. Die Konsequenz ist, dass selbst in diesem Falle vollständiger Rangassoziation τ_a nicht den Maximalwert 1 annimmt, denn $N_c < \binom{n}{2}$ und $N_d = 0$, es sei denn, für alle i gilt $n_{ii} = 1$.

2. Der Koeffizient τ_b enthält entsprechende Korrekturen. Er ist betragsmäßig nie kleiner als τ_a. Ob er seinen Maximalwert erreicht, hängt auch davon ab, ob die Tabelle quadratisch und die Randverteilungen Gleichverteilungen sind.

3. Kendalls τ wird aus Einzelwerten (n Tupel für die n Einheiten) wie folgt berechnet
(7.55) $\tau_a = \Sigma\Sigma d_{ij}/[n(n-1)/2]$,

wobei d_{ij} den Wert +1 oder –1 annimmt, je nachdem, ob zwei verglichene Objekte (Einheiten) rangmäßig in X und Y gleich oder ungleich abgestuft sind und alle Einheiten mit jeder anderen Einheit verglichen werden.

4. Die Vergabe der Werte +1 für konkordante und und -1 für diskonkordante Paarvergleiche legt auch die folgende Schreibweise nahe:

   (7.56) $$\tau_a = \frac{\Sigma\Sigma a_{ij} b_{ij}}{\sqrt{\Sigma\Sigma a_{ij}^2 \Sigma\Sigma b_{ij}^2}}$$

   wobei $a_{ij} = +1$ wenn $R(x_i) > R(x_j)$ und $a_{ij} = -1$ wenn $R(x_i) < R(x_j)$ und b_{ij} entspre-chend für R(y) definiert ist. Konkordante Paarvergleiche ergeben $a_{ij}b_{ij} = +1$ und ent-sprechend ist $a_{ij}b_{ij}$ bei diskonkordanten Vergleichen -1. Gl. 7.56 erinnert an den Pro-dukt-Moment-Korrelationskoeffizienten.

5. Im Unterschied zu den Koeffizienten von Goodman und Kruskal sind die Maße der Def. 7.20 symmetrische Zusammenhangsmaße.

6. Der Koeffizient γ (Gl. 7.53) ist sehr beliebt und kann anders als Q für beliebig dimensionierte Kontingenztafeln berechnet werden. Außerdem kann γ im Sinne der proportionalen Fehlerreduktion interpretiert werden.

Beispiel 7.20:

Gegeben seien folgende Rangdaten (ohne Bindungen) von X und Y mit n = 4 Objekten (Einheiten) für die τ_a nach Gl. 7.55 und γ zu berechnen sind:

v	R(x)	R(y)
1	1	3
2	3	2
3	4	1
4	2	4

Lösung 7.20:

Die Objektpaare, die von X jeweils höher bewertet werden (niedrigerer Rang) als von Y sind 1 und 4: denn $R(x_1) < R(y_1)$ und $R(x_4) < R(y_4)$. Alle übrigen Objektpaare werden von X und Y unterschiedlich bewertet. Man erhält also die folgende Tabelle der Werte d_{ij} von Paarvergleichen, die nur oberhalb der Hauptdiagonalen ausgefüllt wird.

Objekte	1	2	3	4
1	-	-1	-1	+1
2		-	-1	-1
3			-	-1
4				-

Dann ist $\Sigma\Sigma d_{ij} = -4$, also bei n = 4 und n(n-1)/2 = 6, $\tau_a = -4/6 = -2/3$ oder in der Schreibweise von Gl. 7.53: $N_c = 1$, $N_d = 5$. Für γ erhält man (1-5)/(1+5) = -4/6, was mit τ_a identisch ist, weil keine Bindungen auftreten.

3. Spezialfall: ohne Bindungen

Die im letzten Abschnitt behandelten Konzepte lassen sich auch anwenden im Fall fehlender Bindungen (wie im Abschn. 1 angenommen). Man kann zeigen, dass dann gilt für Kendalls τ:

(7.57) $\tau_a = \frac{4\Sigma u_i}{n(n-1)} - 1,$

wobei $R(x_i) = i$ die Referenzrangreihe ist und u_i die Anzahl der in der nachfolgenden Reihe der Werte R(y) höheren Ränge als $R(y_i)$ ist. Ferner ist im Fall ohne Bindungen $\Sigma u_i = N_c$ und entsprechend die Summe niedrigerer Ränge $\Sigma v_i = N_d$ und $N_c + N_d = n(n-1)/2$. In diesem Fall ist Kendalls τ (τ_a) mit Goodman Kruskals γ identisch.

4. Konkordanzkoeffizient

Als Verallgemeinerung von Kendalls τ gilt der Konkordanzkoeffizient W von Kendall für den Vergleich von mehr als zwei Rangordnungen. Werden n Objekte durch m Personen rangmäßig beurteilt, so ergibt sich eine Matrix von Rangzahlen mit n Spalten und m Zeilen. Die zu erwartende Summe der Rangzahlen (Summe einer Spalte) eines jeden Objekts beträgt dann $M = n(n+1)/2$ und die tatsächlich erreichte Rangsumme des j-ten Objekts ($j = 1,2,..., n$) sei R_j. Gäbe es keinen Zusammenhang zwischen den m Rangordnungen, so wären die Rangsummen aller n Objekte jeweils gleich und somit $R_j = M$.

Def. 7.21: Konkordanzkoeffizient

Der Konkordanzkoeffizient mißt die Abweichung von der Gleichheit der Rangvergabe, d.h. er mißt, in welchem Maße sich die Rangordnungen unterscheiden. Er lautet:

(7.58) $W = 12\ \Sigma\ D_j^2/m^2n(n^2-1)$ für $D_j = R_j - M$ und $j=1,2,...,m$

d) Weitere Maße des Zusammenhangs

Auf einige weitere in Übers. 7.2 genannte Korrelationsmaße kann aus Platzgründen hier nicht eingegangen werden. Es soll jedoch kurz die punktbiserielle Korrelation hergeleitet werden.

Ist X dichotom (mit den Ausprägungen $x_o = 0$ und $x_1 = 1$ und den absoluten Häufigkeiten n_0 und n_1) und Y metrisch skaliert und sind $\bar{y}_0$ bzw. $\bar{y}_1$ die bedingten Mittelwerte von Y wenn $X = x_o$ bzw. $X = x_1$ ist, so gilt für die (zweipunktverteilte) Variable X:

Mittelwert: $\bar{x} = n_1/n = p_{1x} = 1 - p_{ox}$

Varianz: $s_x^2 = n_o n_1/n^2 = p_{ox}p_{1x}$

und für die Kovarianz zwischen X und Y

$s_{xy} = p_{1x}\bar{y}_1 - p_{1x}\bar{y} = p_{1x}(\bar{y}_1 - \bar{y})$, da $(1/n)\Sigma xy = p_{1x}\bar{y}_1$ ist.

Daraus ergibt sich für den Produkt-Moment-Korrelationskoeffizienten r_{xy} in diesem speziellen Fall einer dichotomen Variablen X:

$$(7.59)\quad r_{xy} = \frac{p_{1x}(\bar{y}_1-\bar{y})}{s_y\sqrt{p_{ox}p_{1x}}} = \frac{\bar{y}_1-\bar{y}}{s_y}\cdot\sqrt{\frac{p_{1x}}{p_{ox}}} \quad .$$

Das ist der punkt-biserielle Korrelationskoeffizient R_{pb}, der wegen $\bar{y} = p_{ox}\bar{y}_0 + p_{1x}\bar{y}_1$ auch mit unterschiedlichen Umformungen bekannt ist.

Def. 7.22: point biserial correlation

Der punkt-biserielle Korrelationskoeffizient R_{pb} ist definiert als

$$(7.60)\quad R_{pb} = (\bar{y}_1 - \bar{y}_0)\cdot\frac{s_x}{s_y}$$

mit $s_x = \sqrt{p_{0x}p_{1x}}$, wenn X zweipunktverteilt (dichotom) ist, d.h. die relativen Häufigkeiten betragen p_{0x} wenn $X = x_0$ und p_{1x} wenn $X = x_1$ und Y metrisch skaliert ist. In der umgekehrten Situation (Y dichotom) gilt entsprechend

$$(7.60a)\quad R_{pb} = (\bar{x}_1 - \bar{x}_0)\cdot(s_y/s_x) \quad \text{mit } s_y = \sqrt{p_{0y}p_{1y}}.$$

Bemerkungen zur Def. 7.22:

1. Wie aus der Herleitung von Gl. 7.59/60 hervorgeht, ist R_{pb} die Produkt-Moment-Korrelation für den speziellen Fall, dass eine der beiden Variablen dichotom ist. Gl. 7.60 hat eine gewisse Ähnlichkeit mit der Prüfgröße beim t-Test für zwei unabhängige Stichproben über den Unterschied zweier Mittelwerte.

2. Verabredet man s_o^2 für die bedingte Varianz von Y, wenn $X = x_o$ und s_1^2 entspre-chend, wenn $X = x_1$, so ist in Analogie zu Gl. 5.11 (Satz 5.5) $s_y^2 = (p_{0x}\, s_o^2 + p_{1x}\, s_1^2) + [p_{0x}\,(\bar{y}_0 - \bar{y})^2 + p_{1x}\bar{y}_1 - \bar{y})^2]$, worin der erste Ausdruck die interne und der zweite die externe Varianz ist, die auch zu $(\bar{y}_1 - \bar{y}_0)^2\ (p_{0x}\, p_{1x})^2 = (\bar{y}_1 - \bar{y}_0)^2\, s_x^2$ umgeformt werden kann. R_{pb} ist also auch ein analog dem Korrelationsverhältnis η konstruiertes Maß.

Hinweise auf weitere Korrelationskoeffizienten:

1. Der punktbiserielle Korrelationskoeffizient R_{pb} ist nicht zu verwechseln mit dem **biseriellen Korrelationskoeffizienten** R_b (vgl. Übers. 7.2), bei dem man voraus-setzt, dass das dichotome Merkmal X an sich normalverteilt ist und so dichotomisiert ist, dass für $-\infty < X \le x^*$ der Wert $X = x_0$ vergeben wurde und für $X > x^*$ der Wert $X = x_1$. Die Dichtefunktion der Normalverteilung hat an der Stelle x^* den Wert $f(x^*)$. Dann ist $R_b = [(\bar{y}_1-\bar{y})p_{1x}]/[s_y f(x^*)]$

2. Erwähnt sei auch der **tetrachorische Korrelationskoeffizient** mit zwei dichotomisierten, aber an sich normalverteilten Variablen. Er ist ein Assoziationsmaß und in exakter Form schwer zu berechnen. Meist wird er aufgrund des Kreuzproduktverhältnisses cpr aus Tabellen bestimmt.

3. Es gibt auch weitere, nicht im Abschnitt 4b behandelte Versuche, Kontingenzmaße auf der Basis des Konzepts der Fehlerreduktion zu konstruieren, wobei als "Fehler" auch Streuungsmaße benutzt werden, wie z.B. die Entropie.

5. Korrelation und Kausalität

Die Suche nach kausalen Erklärungen oder "Gesetzen" entspringt dem Bedürfnis, Regelmäßigkeiten zu finden. Praktisches Handeln ist unmöglich, ohne das Vertrauen darauf, dass unter ähnlichen Bedingungen auch ähnliches geschehen wird. Es gibt mithin pragmatische, aber auch theoretische Motive einer kausalen Betrachtung.

Gerade in den Wirtschaftswissenschaften laufen viele Kontroversen auf die Frage hinaus, ob eine Variable X (etwa Lohn- oder Geldmengensteigerung) "nur" eine passive Begleiterscheinung oder aber eine aktive (auslösende) Ursache für Y (etwa Inflation) ist. Es ist deshalb eine wichtige Frage, ob und wie empirisch (statistisch) festgestellt werden kann, ob Y die Wirkung von X oder nur eine passive Begleiterscheinung ist.

Im folgenden soll versucht werden, den Kausalbegriff zu definieren und auf zwei Mißverständnisse über das Verhältnis zwischen Kausalität und Korrelation einzugehen, nämlich die Aussagen:

- man könne Kausalität positiv beweisen und (das andere Extrem)
- Korrelation und Kausalität habe nichts miteinander zu tun.

Es ist vergeblich, nach einer Methode zu suchen, mit der man allein auf statistische Daten gestützt, "beweisen" kann, dass X die Ursache von Y ist und nicht nur eine Begleiterscheinung, weil eine Kausalaussage nicht verifiziert, sondern nur falsifiziert werden kann.

Es war das Ziel von Hume und Mill, axiomatisch oder konstruktiv zu einer Festlegung darüber zu gelangen, wie ein empirischer Befund beschaffen sein muss, um auf Kausalität induktiv schließen zu können. Dabei wurde von der Induktion dieselbe Sicherheit verlangt wie von der Deduktion. Ein solches Programm ist zum Scheitern verurteilt. Wie gezeigt wurde (einleitend zu Def. 7.9) beweist ein hoher Betrag der Korrelation r_{xy} nicht, dass X die Ursache für Y ist (oder umgekehrt Y für X), weil dies auch Ergebnis einer Scheinkorrelation sein kann. Hierauf wird auch in Kapitel 8 eingegangen (Gl. 8.37).

Kausalität "beweisen", hieße ausschließen zu können, dass irgendein anderer als der vermutete Kausalzusammenhang für das Zustandekommen der Beobachtungen verantwortlich ist. Das ist in einer positiven, direkten Art nicht möglich, wohl aber kann man indirekt vorgehen. Wie jede andere Hypothese kann die Kausalhypothese dadurch und **nur** dadurch (indirekt) geprüft werden, dass man feststellt, ob der empirische Befund

nicht evtl. im Widerspruch steht zu den bei Geltung der Hypothese zu erwartenden Beobachtungen. Dabei stellen sich aber zwei Fragen:

1. Kann man ein für das praktische Handeln ausreichend sicheres Urteil über eine Kausalhypothese erreichen?
2. Inwiefern können statistische Betrachtungen hierzu beitragen?

zu 1.:
Kausalität kann zwar als Allsatz ("Immer dann, wenn") streng genommen nie verifiziert sondern nur falsifiziert werden. Eine positive Aussage ist trotzdem aus praktischen Gründen oft notwendig. Sie kann aber stets nur vorläufig und unsicher sein. Das damit verbundene Problem kann auch ethischer Natur sein, nicht nur methodischer: Wieviele Menschen müssen z.B. durch Rauchen gestorben sein, bis die Hypothese der Schädlichkeit des Rauchens annehmbar ist?

zu 2.:
Mit der Korrelation r_{xy} kann die Existenz einer (kausalen) Beziehung nicht bewiesen werden, da r_{xy} stets die Summe direkter und indirekter (oder auch nur indirekter, wie bei der Scheinkorrelation!) Einflüsse zwischen X und Y ist.

Im Vorgriff auf Kapitel 8 sei folgendes Modell der multiplen Regression angenommen.

$$Y = b_{yx}X + b_{yz}Z + u_y$$

Hierbei sind alle Variablen standardisiert (also auch $s_x^2 = s_z^2 = 1$), die Koeffizienten b_{yx} und b_{yz} sind sind somit auch die standardisierten Regressionskoeffizienten und die Störgröße u_y ist mit den Regressoren X und Z nicht korrelliert. Man erhält dann

$$(7.61) \quad r_{xy} = b_{yx} + b_{yz}\, r_{xz} \ .$$

Hier misst b_{yx} den direkten und b_{yz} den indirekten (d.h. den über Z vermittelten) Einfluss von X auf Y.

Die in den drei Pfeilschemen der Abb. 7.4 dargestellten Kausalmodelle stellen sich dann wie folgt dar; z.B. beim linken Bild:

$$X = b_{xz}\, Z + u_x$$
$$Y = b_{yz}\, Z + u_y$$

Daraus folgt, wenn u_x und u_y mit Z und auch untereinander nicht korreliert sind,

$$(7.62) \quad r_{xy} = b_{xz}\, b_{yz} = r_{xz}\, r_{yz}.$$

Es gibt jetzt keinen (sich in b_{yx} ausdrückenden) direkten Einfluß von X auf Y, gleichwohl ist aber $|r_{xy}| > 0$.
Entsprechend erhält man bei einer Kausalstruktur, wie man sie im mittleren Teil der Abb. 7.4 dargestellt ist

$$Z = b_{zx}\, X + u_z \qquad \text{so dass } r_{xz} = b_{zx}$$

$Y = b_{yz} Z + u_y$ so dass $r_{yz} = b_{yz}$

woraus folgt, wenn u_y und u_z nicht miteinander korreliert sind

$r_{xy} = b_{yz} r_{xz} = r_{yz} r_{xz}$

wie Gl. 7.62, d.h. beide Pfeilschemen (und auch das dritte) der Abb. 7.4 sind hinsichtlich ihrer beobachtbaren Konsequenz, nämlich Gl. 7.62 nicht unterscheidbar. Sie sind aber unterscheidbar vom Regressionsmodell $x \rightarrow y \leftarrow z$, das zu gl. 7.61 führt. Betrachtungen dieser Art sind Gegenstand der Pfadanalyse, auf die hier nicht weiter eingegangen werden kann.

Aus dem gleichen Grunde ist es auch falsch anzunehmen, man könne über die Richtung der Kausalität (X als Ursache von Y statt Y als Ursache von X) damit entscheiden, ob z.B. die verzögerte Variable X_{t-1} mit Y_t stärker korreliert ist als Y_{t-1} mit X_t (dann $X \rightarrow Y$) oder umgekehrt (dann $Y \rightarrow X$). Hinzu kommt: Weder ist bei allen kausalen Vorgängen diese Asymmetrie von Ursache und Wirkung zu fordern, noch ist diese eindeutig aus der zeitlichen Reihenfolge oder aus den Ergebnissen der - diese Asymmetrie nicht voraussetzenden - Korrelationsanalyse zu folgern.

Andererseits gilt aber: wenn eine bestimmte Kausalstruktur angenommen wird (vgl. Abb. 7.4), dann müßten bestimmte Beziehungen für die Korrelation folgen, etwa (in den drei Fällen von Abb. 7.4) $r_{xy} = r_{xz} r_{yz}$. Trifft dies bei den empirischen Beobachtungen nicht zu, so könnte diese Kausalhypothese verworfen werden.

Das Konzept der Kausalität weist über die bloße Beobachtung hinaus, denn es muss Bezug nehmen auf eine theoretische Fundierung und zwar aus den folgenden drei Gründen:

1. Erklärung:

Oft wird unterschieden zwischen Gesetzmäßigkeit und (evtl. zufälliger) Regelmäßigkeit, je nachdem, ob die beobachteten Zusammenhänge deduktiv in Verbindung mit Sätzen einer Theorie gebracht werden können oder ob dies (noch) nicht der Fall ist. Vom Standpunkt der Beobachtung kann meist zwischen den beiden Arten der Regelmäßigkeit nicht unterschieden werden.

2. Sprachgebundenheit:

In jedem Fall muss sich die Erklärung der Sprache bedienen, d.h. ihre Gültigkeit ist nicht unabhängig davon, wie die Ereignisse oder Variablen bezeichnet und operationalisiert werden, die kausal verknüpft werden. Je enger z.B. das (ursächliche) Ereignis definiert wird, desto kleiner ist die Beobachtungsbasis und Geltungsdauer für den vermuteten Kausalzusammenhang: Im Extremfall mag man die historische Einmaligkeit des Ereignisses behaupten, weil so viele Aspekte einer Erscheinung als "wesentlich" erscheinen, dass die Vergleichbarkeit ausgeschlossen ist. Dann gibt es natürlich keine Möglichkeit, auf Regelmäßigkeiten zu schließen.

3. Modellbildung:

Das Kausalkonzept weist stets ins Unendliche: Jede Ursache hat unendlich viele Wirkungen und selbst wieder unendlich viele "tiefere" oder "letzte" Ursachen. Es ist eine triviale und nutzlose "Erkenntnis", dass alles irgendwie mit allem zusammenhängt. Eine Theorie hat deshalb die Aufgabe, zu einem überprüfbaren Modell zu gelangen durch Ausschluß bestimmter denkbarer Kausalbeziehungen einerseits und durch Festlegung der Art und der Richtungen der Verursachung zwischen ausgewählten Variablen andererseits (causal ordering). Das geschieht z.B. durch die Annahme linearer stochastischer Beziehungen nach Art der oben betrachteten Pfeilschemen von Abb. 7.4.

Die Kennzeichen der Kausalität lassen sich somit in folgender Definition zusammenfassen:

Def. 7.23: Kausalität

Eine Kausalbeziehung zwischen Ereignissen bzw. Variablen X und Y dergestalt, dass X die Ursache von Y ist, bedeutet:

1. Eine Änderung von X "bewirkt" systematisch auch eine Veränderung von Y (Produktionsaspekt).
2. Die Beziehung ist i.d.R. asymmetrisch, d.h. ist X die Ursache von Y, so ist nicht gleichzeitig Y die Ursache von X.
3. Für den Fall, dass ein Experiment nicht durchführbar ist, kann keine Methodik gefunden werden, die es erlaubt, ohne die Interpretationshilfe einer Theorie von bestimmten Daten auf einen Kausalmechanismus zu schließen, d.h. diesen zu beweisen. Man kann aber falsifizierend vorgehen und denkbare Kausalhypothesen ausschließen und zwar auch bei Beobachtungsdaten. Dabei ist auch der Vergleich von beobachteten und erwarteten Korrelationen bedeutsam.

Ob ein statistisch gemessener Zusammenhang kausal interpretiert werden darf oder nicht, ist somit i.d.R. nicht allein aus den Daten zu erkennen.

Bemerkungen zu Def. 7.22:

1. Es ist sinnvoll, zwischen Kausalität in bezug auf Ereignisse (0-1-Variablen; vgl. Kap. 9) und Kausalität in bezug auf Variablen zu unterscheiden.

2. Es wird nicht gefordert, dass ein eindeutiger (funktionaler) Zusammenhang besteht: aus X=x folgt Y=y, wohl aber sollte der mit dem Bestimmtheitsmaß gemessene Anteil der systematischen (mit X erklärten) Variation an der Gesamtvariation von Y (in einem nicht näher bestimmten Maße) beträchtlich sein.

3. Mit der Asymmetrie ist die kausal nicht interpretierbare Interdependenz (feed back) ausgeschlossen. Die zeitliche Folge von Ursache und Wirkung ist nur insofern bedeutsam, als sie eine Möglichkeit bietet, sich empirisch der Asymmetrie zu

vergewissern. Sie ist für sich genommen genauso wenig ein Beweis für Kausalität wie eine gelungene Prognose (post hoc ergo propter hoc - Fehlschluß). Die zeitliche Abfolge ist oft nur der einzige Anhaltspunkt für die Existenz einer Kausalkette. Sie empirisch nachzuweisen ist als solches bereits ein schwieriges methodisches Problem. Eine Methode zur Überprüfung von Kausalhypothesen kann, muss aber nicht notwendig, eine explizite Zeitvariable vorsehen.

4. Eine Methode, sich empirisch des Produktionsaspektes und der Asymmetrie zu vergewissern, ist das Experiment, d.h. die alleinige Variation von X bei Konstanz aller übrigen Einflüsse. Es ist offensichtlich, dass die Ursächlichkeit von X für irgendeine Wirkung von Y nur demonstriert werden kann, wenn X variiert. Ist X konstant, so ist immer Unkorreliertheit gegeben. Eine Konstante X scheidet als erkennbare Ursache (aber auch als Wirkung) aus: eine Konstante ist der Erklärung weder fähig noch bedürftig.

5. Eine Theorie hat eine Erklärung und ein falsifizierbares Modell (causal ordering) zu liefern. Auf die statistischen Methoden zur Behandlung solcher Modelle (z.B. Pfadanalyse) kann hier nicht eingegangen werden.

Kapitel 8: Regressionsanalyse

Die Regressionsanalyse beschäftigt sich mit der Schätzung funktionaler Beziehungen zwischen zwei oder mehreren metrisch skalierten Merkmalen. Sie steht im engen Zusammenhang mit der im Kap. 7 behandelten Korrelationsanalyse. Gegenstand der Regressionsanalyse ist die Art der Abhängigkeit zwischen den Variablen, z.B. die Beziehung zwischen dem Jahresumsatz (Y) und den Ausgaben für Werbung (X) einer Unternehmung (Einfachregression) oder zwischen dem Umsatz (Y) den Werbeausgaben (X_1) und den Wareneinkäufen (X_2) (mehrfache [multiple] Regression). Das Ziel kann dabei die Analyse (d.h. das bessere Verständnis des Kausalzusammenhangs) oder die Prognose von bestimmten Größen sein.

1. Lineare Einfachregression

a) Arten von Beziehungen

Bei der einfachen Regression wird die Beziehung zwischen **zwei** Variablen untersucht. Dabei wird davon ausgegangen, dass eine Variable Y von der anderen Variablen X abhängig ist. Daher auch die folgenden synonymen Bezeichnungen:

Variable Y: abhängige-, zu erklärende-, endogene Variable oder Regressand;
Variable X: unabhängige-, erklärende-, einflußausübende-, exogene (besser: vorherbestimmte) Variable, Prädiktor oder Regressor.

Es sind nun verschiedene Begriffspaare zu definieren, nämlich:

- funktionaler und stochastischer Zusammenhang
- einfache und multiple Regression
- lineare und nichtlineare Regression

Def. 8.1: Zusammenhang, Arten von Regressionsfunktionen

a) Ist Y funktional (deterministisch) abhängig von X, d.h. $y = f(x)$ [Y ist eine Funktion von X], so ist jedem Wert von X ein und nur ein Wert von Y zugeordnet. Bei einer stochastischen Beziehung ist diese Funktion, die **Regressionsfunktion**, von einer **Störgröße** (Restgröße, Residuum) U überlagert (i.d.R. additiv), so dass für die einzelne Beobachtung gilt $y_v = f(x_v) + u_v$. Nach der Art der Regressionsfunktion (d.h. des funktionalen Teils der stochastischen Beziehung) unterscheidet man:

b) einfache und multiple Regression:
Bei der **einfachen** Regression werden nur zwei Variablen X und Y betrachtet. Von **multipler** Regression spricht man, wenn es eine abhängige Variable Y und mehrere unabhängige Variablen $X_1, X_2, ..., X_p$ gibt.

c) lineare und nichtlineare Regression:
Eine Regressionsfunktion ist **linear** (in den Variablen und in den Parametern), wenn gilt

$y_v = a + bx_v$ [a und b heißen Regressionskoeffizienten] (einfache lineare Regression) oder

$y_v = b_o + b_1x_{1v} + b_2x_{2v} + ...+ b_px_{pv}$ (multiple lineare Regression, p Regressoren),

andernfalls ist sie **nichtlinear** (vgl. Abschn. 5).

Im folgenden soll zunächst nur die einfache lineare Regression betrachtet werden. Die Regressionsfunktion ist dann eine Gerade im x,y-Koordinatensystem des Streuungsdiagramms. Sie stellt den funktionalen (systematischen) Teil des Zusammenhangs zwischen den beiden Variablen X und Y dar. Die Beobachtungen streuen als Punkte, je nach Höhe der Korrelation r_{xy} (vgl. Def. 7.8) mehr oder weniger um die Regressionsgerade.

Bemerkungen zu Def. 8.1:

1. Ist Y eine Funktion von X, also y = f(x) so ist z.B. dem Wert $X = x_1$ **ein** Wert $Y = y_1$ zugeordnet. Ist die Beziehung dagegen stochastisch, so können einem Wert x_1 **mehrere** Werte von Y zugeordnet sein, die auch oberhalb oder unterhalb der Funktion f(x) streuen können, etwa

   $Y = y_{11} = f(x_1) + u_1 = y_1 + u_1 < y_1$ oder
   $Y = y_{12} = f(x_1) + u_2 = y_1 + u_2 > y_1$

   je nachdem, welchen Wert U annimmt, z.B. einen positiven ($u_2>0$) oder einen negativen Wert ($u_1<0$).

2. Die Störgröße U erlaubt es, Zusammenhänge zu betrachten, die nicht exakt einer Funktion folgen, wie dies bei praktisch allen empirischen Daten der Fall ist, sei es wegen der Fehlerhaftigkeit der Messung, sei es weil Y noch von anderen Größen als X determiniert wird. Damit wird es auch möglich, systematische Zusammenhänge bei Beobachtungsdaten festzustellen, d.h. dann, wenn im Unterschied zum Experiment nicht alle Einflußgrößen kontrolliert werden können. Das Residuum U ist Ausdruck aller sonstiger, d.h. anders als X nicht explizit berücksichtigter, nicht kontrollierter oder auch gar nicht bekannter Einflüsse auf Y (weshalb U auch zufällig variieren sollte).

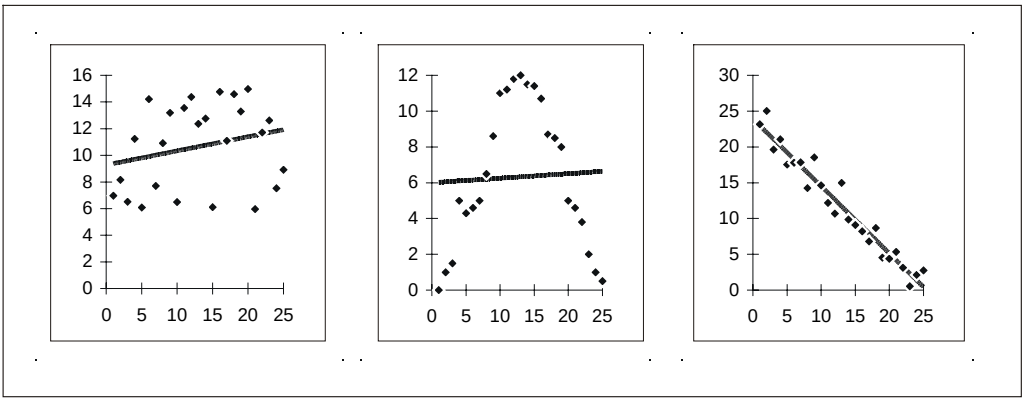
3. Es ist zu unterscheiden zwischen der rein deskriptiven Behandlung der Regression und dem stochastischen Modell, das hier erst später dargestellt wird (Abschn. 3c). Im ersten Fall sind keine Annahmen über

U erforderlich und die Schätzung einer Regressionsfunktion reduziert sich zu einem Problem der Kurvenanpassung, d.h. der Suche nach einer Kurve, die sich einer Punktwolke im Streuungsdiagramm bestmöglich anpaßt.

4. Vor jeder einfachen Regressionsanalyse sollte ein Streuungsdiagramm gezeichnet werden, denn das Streuungsdiagramm (vgl. Bsp. 7.2) erlaubt Rückschlüsse über

   - den jeweiligen Funktionstyp (lineare - oder nichtlineare einfache Regression), vgl. hierzu Abb. 8.1;
   - die Höhe der Korrelation (als Maß der Güte der Anpassung).

Abb. 8.1: Verschiedene Streuungsdiagramme



In Abb. 8.1 sind beispielhaft drei Streuungsdiagramme (mit Regressionsgeraden $\hat{y}$) gegenübergestellt. Wie leicht zu sehen ist, kann man aus der ersten (linken) Punktwolke auf keinen bzw. einen geringen positiven ($r = + 0{,}2408$) Zusammenhang, aus der zweiten Punktwolke auf einen parabolischen und aus der dritten Punktwolke auf einen beträchtlichen negativen ($r = - 0{,}9727$) linearen Zusammenhang der Variablen X und Y schließen (Ein Streuuungsdiagramm mit $r = 0$ ist auch in Abb. 7.1 und 7.3).

5. Die Beschränkung auf lineare Zusammenhänge ist vertretbar, weil

- viele ökonomische Beziehungen zumindest in guter Näherung durch lineare Funktionen dargestellt werden können,
- lineare Funktionen relativ einfach zu handhaben und zu interpretieren sind,
- auch nichtlineare Funktionen häufig linearisiert werden können.

b) Die Regressionsgeraden

Ist ein linearer Zusammenhang zwischen zwei Variablen zu vermuten, so kann dieser durch Schätzung der Regressionskoeffizienten a und b (oder c und d) näher spezifiziert werden.

Def. 8.2: Regressionsgerade

a) Die lineare Regressionsfunktion (Regressionsgerade) zur Bestimmung von Y (abhängige Variable) durch X (unabhängige Variable) lautet:

$$(8.1) \quad \hat{y}_v = a + bx_v$$

dabei ist $\hat{y}_v$ der Regresswert für die v-te Beobachtung (Einheit) mit v = 1,2,...,n und für die einzelne Beobachtung (xv,yv) gilt

$$(8.1a) \quad y_v = \hat{y}_v + u_v = a + bx_v + u_v,$$

d.h. die geschätzte Störgröße u_v für die v-te Beobachtung ist der senkrechte Abstand zwischen y_v und $\hat{y}_v$ im x,y-Koordinatensystem.

b) Die Größen a und b werden Regressionskoeffizienten genannt, wobei a den Ordinatenabstand und b die Steigung der Regressionsgeraden angibt. Es gilt, die Parameter a und b (mit der Methode der kleinsten Quadrate) sowie s_u^2 (Varianz der Störgröße) zu schätzen.

c) Der Zusammenhang zwischen abhängiger und unabhängiger Variable ist rein rechnerisch vertauschbar, d.h. neben der Regressionsgeraden nach Gl. 8.1 ist auch

$$(8.2) \quad \hat{x}_v = c + dy_v$$

zu berechnen, wobei dann für x_v gilt

$$(8.2a) \quad x_v = c + dy_v + v_v.$$

Die Störgröße V ist jeweils der waagrechte Abstand zwischen einem Beobachtungspunkt x_v und $\hat{x}_v$ im x,y-Koordinatensystem.

Bemerkungen zu Def. 8.2:

1. Die Regressionskoeffizienten a und b (bzw. c und d) werden mit der Methode der kleinsten Quadrate geschätzt.

2. Die lineare Einfachregression wird hier zunächst nur als bloße Rechentechnik im Rahmen der Deskriptiven Statistik betrachtet. Es ist dann auch zulässig, die Abhängigkeiten durch Vertauschung von X und Y umzukehren, d.h. statt der Regressionsgeraden $\hat{y}_v = a + bx_v$ die Regressionsgerade $\hat{x}_v = c + dy_v$ zu schätzen. Die Formel zur Berechnung von c (bzw. d) ergibt sich aus derjenigen für a (bzw. b) durch Vertauschung von X und Y. Man kann grundsätzlich zwei Regressionsgeraden berechnen, aber meist nur eine der beiden kausal interpretieren.

3. Ob eine Variable abhängig (endogen) oder unabhängig (exogen) ist, liegt für das Modell der Regressionsanalyse nicht in ihrem Inhalt, ihrer sachlichen Interpretation begründet, sondern allein darin, ob die Variable im Rahmen des Modells durch eine Gleichung "erklärt" wird (endogen) oder nicht (exogen). Eine unabhängige (exogene) Variable ist annahmegemäß nicht mit der Störgröße der Gleichung in der sie auftritt korreliert. Auf die Annahmen des zu schätzenden Modells der Regression wird erst im Abschn. 3c eingegangen.

c) Schätzung der Regressionskoeffizienten mit der Methode der kleinsten Quadrate

Die Regressionskoeffizienten sind mit einem eindeutigen und objektiven Verfahren so zu bestimmen, dass sich die Regressionsgerade der Punktwolke im Streuungsdiagramm bestmöglich anpaßt. Alle Kriterien zur Schätzung der Regressionskoeffizienten a und b (bzw. c und d) gehen von den Residuen u_v (bzw. v_v) aus. Ein solches Verfahren ist die Methode der kleinsten Quadrate. Bei ihr werden die Regressionskoeffizienten so bestimmt, dass die Summe der Quadrate der Abweichungen von der Regressionsgeraden ein Minimum annimmt.

Aus Gleichung (8.1a) folgt

$u_v^2 = (y_v - a - bx_v)^2$ mit v = 1,2,...,n und damit:

$\Sigma u_v^2 = \Sigma(y_v^2 - ay_v - bx_v y_v - ay_v + a^2 + abx_v - bx_v y_v + abx_v + b^2 x_v^2)$

$= \Sigma(y_v^2 - 2ay_v - 2bx_v y_v + a^2 + 2abx_v + b^2 x_v^2)$.

Mithin gilt:

$$(8.3) \quad \Sigma u_v^2 = \Sigma y_v^2 - 2a\Sigma y_v - 2b\Sigma x_v y_v + na^2 + 2ab\Sigma x_v + b^2\Sigma x_v^2$$

Die Summe Σu_v^2 ist bei gegebenen Beobachtungen (also auch gegebenen Werten Σx_v, Σx_v^2, Σy_v, Σy_v^2 und $\Sigma x_v y_v$) eine Funktion Q(a,b) von a und b und es gilt a und b so zu bestimmen, dass $\Sigma u_v^2 = Q(a,b)$ minimal ist. Hierzu ist lediglich Σu_v^2 gem. Gl. 8.3 partiell

nach a und b zu differenzieren und die beiden Ableitungen sind Null zu setzen, denn man kann leicht zeigen, dass Q(a,b) kein Maximum, sondern nur ein Minimum besitzt [Man kann eine "Regressionsgerade" auch beliebig weit außerhalb der Punktwolke des Streuungsdiagramms legen so dass Σu_v^2 beliebig groß werden kann].

Das ergibt die folgenden zwei Gleichungen:

1) $\frac{dQ}{da} = -2\Sigma y_v + 2na + 2b\Sigma x_v \stackrel{!}{=} 0$

2) $\frac{dQ}{db} = -2\Sigma x_v y_v + 2a\Sigma x_v + 2b\Sigma x_v^2 \stackrel{!}{=} 0$,

die umgeformt das System der ***Normalgleichungen*** darstellen:

(8.4a)	$an + b\Sigma x_v = \Sigma y_v$	1. Normalgleichung
(8.4b)	$a\Sigma x_v + b\Sigma x_v^2 = \Sigma x_v y_v$	2. Normalgleichung

Wird dieses Normalgleichungssystem nach a und b aufgelöst so erhält man als Schätzwerte für die Regressionskoeffizienten a und b:

(8.5a) $$a = \frac{\Sigma x_v^2 \Sigma y_v - \Sigma x_v \Sigma x_v y_v}{n\Sigma x_v^2 - (\Sigma x_v)^2}$$ (8.5b) $$b = \frac{n\Sigma x_v y_v - \Sigma x_v \Sigma y_v}{n\Sigma x_v^2 - (\Sigma x_v)^2}$$

Mit $n\Sigma x_v^2 - (\Sigma x_v)^2$ steht die n^2-fache Varianz der Variablen X im Nenner von a und b gem. Gl. 8.5 und die Steigung b ist das Verhältnis von Kovarianz und Varianz:

(8.6a) $$b = \frac{\Sigma(x_v - \bar{x})(y_v - \bar{y})}{\Sigma(x_v - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$.

Wie man leicht sieht, gilt aufgrund der ersten Normalgleichung:

(8.6b) $$a = \bar{y} - b\bar{x}$$

d.h. die Regressionsgerade $\hat{y}_v = a + bx_v$ verläuft durch den Schwerpunkt P(x,y) des Streuungsdiagramms.

Man kann also a und b bestimmen

1. aufgrund der beiden Normalgleichungen und
2. indem man zunächst b nach Gl. 8.6a und dann a gem. Gl. 8.6b berechnet.

Regressionsgerade $\hat{x} = c + dy$

Man erhält die entsprechenden Formeln zur Bestimmung von c und d, indem man in den Normalgleichungen (Gl. 8.4) bzw. in den Formeln (Gl. 8.6) für a und b einfach x und y vertauscht:

(8.4a*) $cn + d\Sigma y_v = \Sigma x_v$	(8.6a*) $d = \frac{s_{xy}}{s_y^2}$
(8.4b*) $c\Sigma y_v + d\Sigma y_v^2 = \Sigma x_v y_v$	(8.6b*) $c = \bar{x} - d\bar{y}$

denn aufgrund der ersten Normalgleichung (Gl. 8.4a*) verläuft auch die Funktion $\hat{x} = c + dy$ durch den Schwerpunkt [die beiden Regressionsgeraden schneiden sich also in diesem Punkt].

Die Zusammenhänge werden an einem einfachen Zahlenbeispiel demonstriert und anschließend einige Eigenschaften der Regressionsgeraden interpretiert.

Beispiel 8.1:

Gegeben seien die folgenden Daten für n = 4 Personen (Es ist klar, dass man in der Praxis nicht bei n = 4 Werten eine Regressionsgerade schätzen würde. Das Beispiel soll aber möglichst einfach und überschaubar sein):

x_v	2	3	7	8
y_v	4	5	10	5

Für die folgenden Berechnungen empfiehlt es sich, eine Arbeitstabelle aufzustellen:

	x_v	y_v	x_v^2	y_v^2	$x_v y_v$	$(x_v-\bar{x})^2$	$(y_v-\bar{y})^2$	$(x_v-\bar{x})(y_v-\bar{y})$
	2	4	4	16	8	$(2-5)^2=9$	$(4-6)^2=4$	$(-3)(-2)=6$
	3	5	9	25	15	$(3-5)^2=4$	$(5-6)^2=1$	$(-2)(-1)=2$
	7	10	49	100	70	$2^2=4$	$4^2=16$	8
	8	5	64	25	40	$3^2=9$	$(-1)^2=1$	-3
Σ	20	24	126	166	133	26	22	13

Es ist nützlich, zunächst die Parameter der Randverteilungen und der zweidimensionalen Verteilung gem. Kap. 7 zu berechnen:

Mittelwerte: $\bar{x} = \Sigma x_v/n = 20/4 = 5$

$\bar{y} = \Sigma y_v/n = 24/4 = 6$

Varianzen: $s^2_x = [\Sigma(x_v-\bar{x})^2]/n = 26/4 = 6{,}5$

$s^2_y = [\Sigma(y_v-\bar{y})^2]/n = 22/4 = 5{,}5$

Kovarianz: $s_{xy} = [\Sigma(x_v-\bar{x})(y_v-\bar{y})]/n = 13/4 = 3{,}25$

Berechnung der Regressionsgeraden $\hat{y} = a + b\cdot x$ und $\hat{x} = c + d\cdot y$:

Die Normalgleichungen lauten

für a und b:	für c und d:
$an + b\Sigma x_v = \Sigma y_v$	$cn + d\Sigma y_v = \Sigma x_v$
$a\Sigma x_v + b\Sigma x_v^2 = \Sigma x_v y_v$	$c\Sigma y_v + d\Sigma y_v^2 = \Sigma x_v y_v$
$4a + 20b = 24$	$4c + 24d = 20$
$20a + 126b = 133$	$24c + 166d = 133$

Daraus folgt: $a = 3{,}5$ und $b = ½$ sowie
$c = 32/22 = 1{,}4545$ und $d = 13/22 = 0{,}5909$

Ferner gilt: $b = s_{xy} / s_x^2 = 13/26 = ½$ und $d = s_{xy} / s_y^2 = 13/22 = 0{,}5909$.

Die Regressionsgeraden schneiden sich im Schwerpunkt, d.h.:

$$a = \bar{y} - b\cdot \bar{x} = 3{,}5 \quad \text{und} \quad c = \bar{x} - d\cdot \bar{y} = 32/22 = 1{,}4545.$$

Für die beiden Regressionsgeraden erhält man also:

$$\hat{y}_v = 3{,}5 + 0{,}5x_v \qquad \text{und} \qquad \hat{x}_v = 32/22 + (13/22)y_v.$$

Die Berechnung der Funktion Q(a,b)

$$Q(a,b) = \Sigma u_v^2 = \Sigma y_v^2 - 2a\Sigma y_v - 2b\Sigma x_v y_v + na^2 + 2ab\Sigma x_v + b^2\Sigma x_v^2$$

für dieses Beispiel führt zu

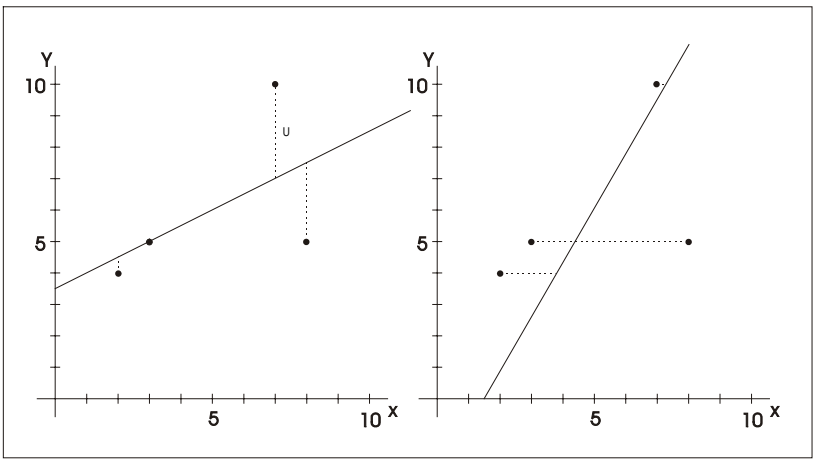
$$Q = \Sigma u_v^2 = 166 - 48a - 266b + 4a^2 + 40ab + 126b^2,$$

einer Funktion mit einem Minimum von $\Sigma u_v^2 = 15{,}5$. Die folgende Tabelle gibt einige Funktionswerte von Q(a,b) in der Nähe des Minimums an:

	b=0,3	b=0,4	b=0,5	b=0,6	b=0,7
a=3,3	22,30	17,72	15,66	16,12	19,10
a=3,4	21,38	17,20	15,54	16,40	19,78
a=3,5	20,54	16,76	15,50	16,76	20,54
a=3,6	19,78	16,40	15,54	17,20	21,38
a=3,7	19,10	16,12	15,66	17,72	22,30

Mit beispielsweise b = 0,5 gilt $Q(a,b=½) = 64,5 - 28a + 4a^2$, was eine Parabel im a,Q-Koordinatensystem ist. Man sieht, dass $Q = \Sigma u_v^2$ in der Tat an der Stelle a = 3,5 mit $\Sigma u_v^2 = 15,5$ ein Minimum hat.

Abb. 8.2: Streuungsdiagramm und Regressionsgeraden für Bsp. 8.1



Wie man sieht, schneiden sich die beiden Regressionsgeraden im Punkt S ($\bar{x}$=5, $\bar{y}$=6), dem Schwerpunkt.

Mit diesem Demonstrationsbeispiel lassen sich auch einige Eigenschaften der Methode der kleinsten Quadrate zeigen. Man kann z.B. die in Abb. 8.2 eingezeichneten (senkrechten) Residuen u_v und die waagrechten Residuen v_v bestimmen und die Residuen u_v der Regressionsgerade $\hat{y}_v = a + bx_v = 3,5 + ½x_v$ vergleichen mit den Abweichungen u_v,* um eine alternative Gerade, etwa um $\hat{y}_v^* = 3,4 + 0,6x_v$:

		Regressionsgerade $\hat{y}_v = 3{,}5 + 0{,}5x_v$			alternative Gerade $\hat{y}_v^* = 3{,}4 + 0{,}6x_v$		
x_v	y_v	$\hat{y}_v$	u_v	u_v^2	$\hat{y}_v^*$	u_v^*	$(u_v^*)^2$
2	4	4,5	-0,5	0,25	4,6	-0,6	0,36
3	5	5	0	0	5,2	-0,2	0,04
7	10	7	3	9	7,6	2,4	5,76
8	5	7,5	-2,5	6,25	8,2	-3,2	10,24
Σ 20	24	24	0	15,5	25,6	-1,6	16,4

Wie man sieht gilt: $\Sigma u_v = 0$ und $\Sigma y_v = \Sigma \hat{y}_v = 24$ (beides infolge der ersten Normalgleichung), während dies für eine andere Gerade nicht gelten muss. Abgesehen davon, dass die Summe der Abweichungsquadrate $\Sigma(u_v^*)^2$ mit 16,4 größer ist als $\Sigma u_v^2 = 15{,}5$ ist auch u_v mit x nicht korreliert, wohl aber u_v^* mit x (denn die Kovarianz zwischen u_v^* und x beträgt $(\Sigma x u^*)/n - \bar{u}^*\bar{x} = -2{,}65 - 5(-0{,}4) = -0{,}65$).

Für die Störgröße v in der Regression von x (abhängig) auf y (unabhängig) gilt:

$$v_v = x_v - \hat{x}_v = x_v - [32/22 + (13/22)y_v].$$

y_v	x_v	$\hat{x}_v$	v_v	v_v^2
4	2	3,818	-1,818	3,3058
5	3	4,409	-1,409	1,9855
10	7	7,364	-0,364	0,1322
5	8	4,409	+3,591	12,8946
Σ 24	20	20	0	18,3182

d) Regressionskoeffizienten und Korrelationskoeffizient

Nach Def. 7.8 gilt für den Korrelationskoeffizienten

$$(8.7) \quad r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} = \begin{cases} +\sqrt{bd} & \text{wenn } b,d > 0 \\ -\sqrt{bd} & \text{wenn } b,d < 0 \end{cases}$$

Weil das Vorzeichen von b und d allein durch die Kovarianz s_{xy} bestimmt wird, haben b und d stets das gleiche Vorzeichen. Der (lineare) Korrelationskoeffizient r_{xy} ist also das geometrische Mittel der Steigungen der beiden Regressionsgeraden.

Im Unterschied zu b und d ist r invariant gegenüber linearen Transformationen von X und Y (vgl. Bem. Nr. 2 zu Def. 7.8). Mit den Lineartransformationen

$x^* = p_1 + q_1 \cdot x$ und $y^* = p_2 + q_2 \cdot y$

erhält man für die Koeffizienten der Regressionsfunktion

$y^* = a^* + b^*x^*$ $\quad a^* = p_2 + q_2a - p_1b^*$ und $\quad b^* = b(q_2/q_1)$

$x^* = c^* + d^*y^*$ $\quad c^* = p_1 + q_1c - p_2d^*$ und $\quad d^* = d(q_1/q_2)$.

Die Regressionskoeffizienten a,b,c und d werden also von Maßstabsänderungen in den Skalen von X und/oder Y berührt, der Korrelationskoeffizient dagegen nicht.

e) Varianzzerlegung und Bestimmtheitsmaß (Determinationskoeffizient)

Die Regressionsfunktion $\hat{y} = a + bx$ bedeutet, dass die Streuung von Y wegen der linearen Abhängigkeit von X zum Teil durch die Streuung von X erklärt werden kann. Die Varianz von Y ist zu zerlegen in einen auf $\hat{y}$ (und damit auf X) zurückgehenden Teil und in eine Residualvarianz:

Man kann dies zeigen, indem man den Abstand (die Abweichung) eines
Datenpunkts vom Mittel zerlegt in zwei Abstände u_v und $\hat{y}_v - \bar{y}$:
$(y_v - \bar{y}) = (y_v - \hat{y}_v) + (\hat{y}_v - \bar{y})$ wobei $u_v = (y_v - \hat{y}_v)$.

Berücksichtigt man, dass gilt (vgl. Gl. 8.11ff.) $\Sigma y = \Sigma\hat{y}$, $= 0$ und $\Sigma\hat{y}_v u_v = \Sigma x_v u_v = 0$, so ergibt sich

$$(y_v - \bar{y}) = (\hat{y}_v - \bar{y}) + (u_v - \bar{u})$$
$$\quad T \qquad\quad E \qquad\quad R$$

Die beiden Differenzen E und R auf der rechten Seite dieser Identität deuten auf zwei Variationsquellen hin:

E: Die Differenz von dem durch die Regression geschätzten Wert (dem Regresswert) $\hat{y}_v$ und dem arithmetischen Mittel $\bar{y}$ ist verantwortlich für die durch die Regression "erklärte" Streuung.

R: Die Abweichung des beobachteten Wertes y_v von dem geschätzten Wert $\hat{y}_v$ (d.h. das Residuum) kann als "durch die Regression nicht erklärte Abweichung" bezeichnet werden.

Durch Quadrieren der Abstandsgleichung und anschließender Summation erhält man:

$$\Sigma(y_v - \bar{y})^2 = \Sigma(y_v - \hat{y}_v)^2 + 2\Sigma(y_v - \hat{y}_v)(\hat{y}_v - \bar{y}) + \Sigma(\hat{y}_v - \bar{y})^2$$
$$\Sigma T^2 \quad = \quad \Sigma R^2 \quad + \quad 2\Sigma RE \quad + \quad \Sigma E^2$$

Wie leicht zu sehen ist, fällt der mittlere Ausdruck auf der rechten Seite weg, denn wegen $\Sigma u_v = \Sigma u_v \cdot x_v = 0$ gilt

$$\Sigma(y_v - \hat{y}_v)(\hat{y}_v - \bar{y}) = \Sigma u_v(\hat{y}_v - \bar{y}) = \Sigma u_v(a + bx_v - \bar{y}) = 0 .$$

Somit gilt auch (Varianzzerlegung)

(8.8)	$\frac{1}{n}\Sigma(y_v - \bar{y})^2$	=	$\frac{1}{n}\Sigma(\hat{y}_v - \bar{y})^2$	+	$\frac{1}{n}\Sigma(y_v - \hat{y}_v)^2$
	totale Varianz	=	erklärte Varianz	+	Residualvarianz
	s_y^2	=	$s_{\hat{y}}^2$	+	s_u^2

Diese Gleichung legt den Gedanken nahe, die erklärte Varianz $s_{\hat{y}}^2$ und die Residualvarianz $s_u^2 = n^{-1}\Sigma(u_v - \bar{u})^2 = n^{-1}\Sigma(u_v - 0)^2 = (\Sigma u_v^2)/n$ durch die Gesamtvarianz zu dividieren. Die Anteile sind nach Def. 7.10 das **Bestimmtheitsmaß** B_{yx} (Bestimmtheit von Y durch X) und das **Unbestimmtheitsmaß**[1] $U_{yx} = 1 - B_{yx}$. Offenbar gilt $0 \leq B_{yx} \leq 1$ und entsprechend $0 \leq U_{yx} \leq 1$, weil B und U Varianz**anteile** darstellten.

Speziell für die einfache lineare Regression gilt für das Bestimmtheits - und Unbestimmtheitsmaß:

1. Symmetrie: $B_{yx} = B_{xy}$ mit $B_{xy} = \frac{s_{\hat{x}}^2}{s_x^2}$
2. Das Bestimmheitsmaß B_{yx} ist das Quadrat des Korrelationskoeffizienten r_{xy} ($B_{yx} = r_{xy}^2$).

Anders als beim linearen Korrelationskoeffizient r kann man mit dem Bestimmtheitsmaß r^2 nicht zwischen positiver und negativer Korrelation differenzieren.

Wegen $\hat{y}_v = a + bx_v$ und $\bar{y} = a + b\bar{x}$ erhält man für die erklärte Varianz $s_{\hat{y}}^2$ auch $s_{\hat{y}}^2 = b^2 s_x^2$ und weil $b = s_{xy}/s_x^2$

$$(8.9) \quad B_{yx} = \frac{s_{xy}^2}{s_x^2 \, s_y^2} = b \cdot d = r_{xy}^2 .$$

[1] Die Größe U (Unbestimmtheitsmaß) und u (Störgröße) sollte nicht verwechselt werden.

Entsprechend ist die mit X erklärte Varianz $s^2_{\hat{x}} = d^2 s^2_y$ und damit

$$B_{xy} = \frac{d^2 s_y^2}{s_x^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = b \cdot d = r_{xy}^2 = B_{yx}.$$

Für die Grenzfälle $r_{xy}^2 = 1$ und $r_{xy}^2 = 0$ gilt:

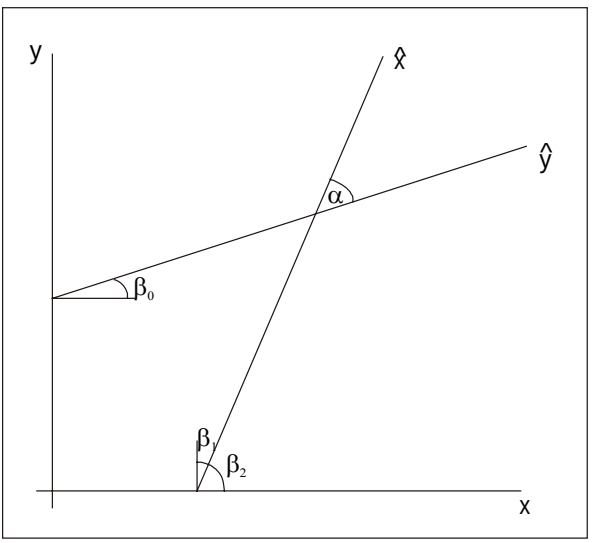
- $r_{xy},^2 = B_{yx} = 1 \qquad (U_{yx} = 0)$
  $u_v = 0$ und $v_v = 0$ für alle v = 1,2,...,n, alle n Beobachtungen liegen genau auf den Regressionsgeraden, die in diesem Fall "zusammenfallen" und lauten $\hat{y} = a + bx$ und $\hat{x} = (-a/b) + (1/b)y$

- $r_{xy},^2 = B_{yx} = 0 \qquad (U_{yx} = 1)$
  für jedes v ist der Regresswert $\hat{y}_v$ identisch mit dem Mittelwert $\bar{y}$ (die y Regressionsgerade verläuft also parallel zur x-Achse und hat die Steigung b = 0 [sie lautet $\hat{y} = \bar{y}$]) und für jedes v ist der Regresswert $\hat{x}_v$ identisch mit dem Mittelwert $\bar{x}$. Die Regressionsgerade $\hat{x}$ lautet $\hat{x} = \bar{x}$ und steht senkrecht auf der Regressionsgeraden $\hat{y} = \bar{y}$.

Man könnte vermuten, dass der Winkel α zwischen den Regressionsgeraden mit dem Korrelationskoeffizienten zusammenhängt, da $\alpha = 0^o$ bedeutet r = +1 oder r = -1 und $\alpha = 90^o$ impliziert r = ± 0. Der Zusammenhang ist jedoch nicht ganz so einfach. Man kann leicht zeigen, dass gilt

$$(8.10) \qquad \tan(\alpha) = \frac{1-bd}{b+d} = \frac{s_x s_y (1-r^2)}{r(s_x^2 + s_y^2)}$$

Abb. 8.3: Zur Geometrie der Regressionsgeraden



und dass die Steigung $\tan(\beta_2)$ der Regressionsgeraden $\hat{x}$ im x,y-Koordinatensystem betragsmäßig stets größer ist als die Steigung b der Regressionsgeraden $\hat{y}$, denn (vgl. Abb. 8.3) es gilt:

$\tan(\beta_o) = b = s_{xy}/s_x^2$;
$\tan(\beta_1) = d = s_{xy}/s_y^2$;
$\tan(\beta_2) = \cot(\beta_1) = 1/d$,

so dass sich die Behauptung über die Steigungen aus der Schwarzen Ungleichung ergibt.

Beispiel 8.2:

Man verifiziere Gl. 8.7 bis 8.9 für das Bsp. 8.1!

Lösung 8.2:

Für das Bsp. 8.1 erhält man: $r = 13/\sqrt{22 \cdot 26} = 0{,}54356$, da die Kovarianz 13/5 und die Varianzen $s_x^2 = 22/5$ und $s_y^2 = 26/5$ betragen. Die Steigungen sind $b = 0{,}5$ und $d = 13/22$, so dass $r = \sqrt{13/44} = 0{,}54356$. Ferner erhält man $\Sigma u_v^2 = 15{,}5$ und $\Sigma(y_v - \bar{y})^2 = 22$, so dass gilt $U_{yx} = 15{,}5/22 = 0{,}7045 = 1 - r_{yx}^2$ (es gilt $r_{yx}^2 = 1 - U_{yx} = 1 - 0{,}7045 = 0{,}2954 = (0{,}54356)^2$) und $\Sigma v_v^2 = 18{,}318$ und $\Sigma(x_v - \bar{x})^2 = 26$, so dass gilt $U_{xy} = 18{,}318/26 = 0{,}7045 = 1 - r_{xy}^2$. Man sieht, dass die Bestimmtheit hier 29,54% beträgt und dass $B_{yx} = B_{xy}$ das Quadrat des Korrelationskoeffizienten ist.

Beispiel 8.3:

Es sei X der Intelligenzquotient (IQ) des Vaters und Y der des Sohnes. Psychologen fanden heraus, dass der IQ (praktisch in allen Generationen) mit Mittelwert 100 und Standardabweichung 16,4 symmetrisch verteilt ist, und dass die Korrelation zwischen dem IQ des Vaters und des Sohnes $r_{xy} = +0{,}5$ ist:

a) Man bestimme die Regressionsgerade $\hat{y} = a + bx$.

b) Welcher IQ ist für den Sohn zu erwarten, wenn der Vater einen IQ von 75 (d.h. leichte Debilität) und welcher, wenn der Vater einen IQ von 130 (überragende Intelligenz) hat?

c) Kann man aufgrund der Ergebnisse schließen, dass durch die Vererbung ein unaufhaltsamer Trend zum Mittelmaß besteht, so dass es nach einigen Generationen nur noch Personen mit einem IQ von 100 gibt?

d) Wie ändert sich die Situation hinsichtlich der unter c) gegebenen Interpretation, wenn die Korrelation nicht $r_{xy} = +0{,}5$, sondern $r_{xy} = +0{,}25$ beträgt.

Lösung 8.3:

a) Aus $r_{xy} = 0{,}5$ folgt wegen $s_x = s_y = 16{,}4$ für die Kovarianz $s_{xy} = r_{xy} s_x s_y = ½ \cdot (16{,}4)^2 = ½ s_x,^2$ und damit : $b = s_{xy} / s_x^2 = ½$. Aus $\bar{y} = a + b\bar{x}$ folgt dann wegen $\bar{x} = \bar{y} = 100$ und $b = ½$ für a der Wert: $a = 50$.
Die Regressionsgerade lautet also $\hat{y} = 50 + ½x$ (Sohn in Abhängigkeit vom Vater)

b) Die Fragen nach dem zu erwartenden IQ des Sohnes bei gegebenem IQ des Vaters laufen auf eine deterministische (d.h. funktionale) Interpretation der Regressionsfunktion hinaus. Man kann mit der Regressionsfunktion leicht die folgenden Werte nachrechnen:

Vater (x)	Sohn ($\hat{y}$)	$\hat{y}$ ist
75 (unter 100)	$50 + ½ \cdot 75 = 87{,}5$	größer als x
130 (über 100)	$50 + ½ \cdot 130 = 115$	kleiner als x

Man erkennt, dass grundsätzlich gilt: $\hat{y} > x$ wenn $x < 100$ (der Sohn ist intelligenter als der Vater, wenn dieser unterdurchschnittlich intelligent ist) und umgekehrt $\hat{y} < x$ wenn $x > 100$.

c) Das legt die (verfehlte) Interpretation, auf die hier bezug genommen wird nahe. Sie ist verfehlt, weil die Regressionsfunktion deterministisch interpretiert wird, wozu bei einer Bestimmtheit von nur 25% ($r^2 = 0{,}25$) kein Anlaß besteht. Zu einem ähnlichen (Fehl-) Schluß gelangten amerikanische Wirtschaftsforscher bei einer langfristigen Analyse von Unternehmensgewinnen). Man bezeichnet den dargestellten Zusammenhang auch als regression to the mean.

d) Die Regressionsgerade lautet jetzt: $\hat{y} = 75 + ¼\, x$ und die Regression zum Mittel (in der falschen Interpretation also der Trend zum Mittelmaß) ist noch schneller: denn ist $x = 75$, so ist (deterministisch interpretiert) $\hat{y} = 93{,}75$ statt 87,5 und bei $x = 130$ ist $\hat{y} = 107{,}5$ statt 115. Wegen $s_x = s_y$ ist generell bei diesem Beispiel die Regressionsfunktion

$$\hat{y} = 100(1 - r) + rx$$

und für die Abweichung vom Mittelwert gilt

$$(\hat{y} - 100) = r(x - 100).$$

Sie ist also bei gegebenen x umso kleiner (und damit die Geschwindigkeit der regression to mean umso größer) je kleiner r ist.

2. Bemerkungen zur Methode der kleinsten Quadrate

a) Eigenschaften der geschätzten Residuen

Aus den beiden Normalgleichungen (Gl. 8.4a und 8.4b) für die Bestimmung der Regressionsgerade $\hat{y} = a + bx$

$an + b\Sigma x_v = \Sigma y_v$	1. Normalgleichung
$a\Sigma x_v + b\Sigma x_v^2 = \Sigma x_v y_v$	2. Normalgleichung

ergeben sich die folgenden Eigenschaften der geschätzten Störgröße u (für v in der Regression $\hat{x} = c + dy$ gelten die folgenden Ausführungen analog), die ihrerseits gewisse Folgerungen erlauben:

- aus der ersten Normalgleichung folgt:
  1. Die Summe der Residuen (und somit auch der Mittelwert $\bar{u}$) ist Null

     (8.11) $\Sigma u_v = n\bar{u} = 0$

     oder äquivalent
  2. Die geschätzte Regressionsgerade verläuft durch den Schwerpunkt
  3. Die Regresswerte $\hat{y}$ und die beobachteten y-Werte sind in der Summe und damit auch im Mittel gleich

     (8.12) $\Sigma\hat{y}_v = \Sigma y_v$ und $\overline{\hat{y}} = \bar{y}$.

- aus der zweiten Normalgleichung folgt:
  4. Multipliziert man $y_v = a + bx_v + u_v$ mit x_v und summiert man über alle n Beobachtungen so erhält man
     $\Sigma x_v y_v = a\Sigma x_v + b\Sigma x_v,^2 + \Sigma x_v u_v$, so dass wegen der zweiten Normalgleichung U und X nicht miteinander korreliert sind:

     (8.13) $\Sigma x_v u_v = 0$ und $s_{ux} = r_{ux} = 0$.

  5. Multipliziert man $\hat{y}_v = a + bx_v$ mit u_v und summiert man über alle n Beobachtungen, so erhält man $\Sigma\hat{y}_v u_v = a\Sigma u_v + b\Sigma x_v u_v$, so dass wegen $\Sigma x_v u_v = \Sigma u_v = 0$ gilt:

     (8.14) $\Sigma\hat{y}_v u_v = r_{\hat{y}u} = 0$.

  6. Die entsprechende Betrachtung mit $y_v = a + bx_v + u_v$ führt zu $\Sigma y_v u_v = a\Sigma u_v + b\Sigma x_v u_v + \Sigma u_v^2$ und wegen $\Sigma x_v u_v = \Sigma u_v = 0$ zu

     (8.15) $\Sigma y_v u_v = \Sigma u_{yx}^2$ und $s_u^2 = \dfrac{\Sigma u_v^2}{n} = s_{uy}$.

     Hieraus folgt $r_{yu} = s_u/s_y$ und damit auch

     (8.16) $(r_{yu})^2 = \dfrac{s_u^2}{s_y^2} = 1 - (r_{xy})^2 = U_{xy}$

das Unbestimmtheitsmaß (Unbestimmtheit von Y durch X) als Bestimmtheit von Y durch U (und umgekehrt).

7. Aus Gl. 8.14 folgt mit $y_v = \hat{y}_v + u_v$ die Varianzzerlegung $s_y^2 = s_{\hat{y}}^2 + s_u^2$ sowie $s_{y\hat{y}} = s_{\hat{y}}^2$, so dass auch gilt

$$r_{y\hat{y}} = \frac{s_{y\hat{y}}}{s_{\hat{y}} s_y} = \frac{s_{\hat{y}}^2}{s_{\hat{y}} s_y} = \frac{s_{\hat{y}}}{s_y} = \sqrt{\frac{s_{\hat{y}}^2}{s_y^2}} = \sqrt{B_{xy}} = r_{xy} \text{ also}$$

$$\boxed{(8.17)\quad r_{y\hat{y}} = r_{xy}.}$$

Die Korrelation zwischen x und y ist also auch als Korrelation zwischen y und dem Regresswert $\hat{y}$ zu interpretieren, eine Interpretation, die sich auf die multiple Regression übertragen läßt.

8. Die Regressionsgerade $\hat{y} = a + bx$ verläuft durch die zwei Punkte $P_1(\bar{x},\bar{y})$, d.h. den Schwerpunkt und $P_2(\bar{x}_w, \bar{y}_w)$, wobei $\bar{x}_w$ und $\bar{y}_w$ gewogene Mittel der x_v- bzw. y_v-Werte sind mit den Gewichten $w_v = x_v/\Sigma x_v$.

b) Alternativen zur Minimierung der Summe der Quadrate der Abweichungen

Man kann sich alternative Schätzungen der Regressionskoeffizienten a und b der Regressionsfunktion $\hat{y} = a + bx$ (die folgende Betrachtung gilt übrigens ganz entsprechend auch für nichtlineare und multiple Regression) vorstellen, von denen jedoch nicht alle sinnvoll und eindeutig sind. Dargestellt werden im folgenden sechs Alternativen zur Methode der kleinsten Quadrate:

1.) Fordert man z.B. den Ausgleich positiver und negativer Abweichungen von der Regressionsfunktion, d.h. dass die Summe der Residuen den Wert Null annimmt, so erhält man <u>nur eine</u> Gleichung zur Bestimmung von a und b, nämlich die erste Normalgleichung.

Die Forderung $\Sigma u_v = 0$ führt zur Gleichung $a + b\bar{x} = \bar{y}$, die von allen Geraden erfüllt wird, die durch den Schwerpunkt gehen. Eine von Ihnen ist die Regressionsgerade nach der Methode der kleinsten Quadrate. Das Kriterium $\Sigma u_v = 0$ ist also in $\Sigma u_v^2 =$ Min impliziert, führt aber allein nicht zu einer eindeutigen Lösung.

2.) Die Minimierung der Abweichungen Σu_v scheitert daran, dass die Funktion $Q = \Sigma u_v = \Sigma y_v - na - b\Sigma x_v$ keine Extremwerte besitzt (bei ge-

gebenem $a = a_o$ oder $b = b_o$ ist die Funktion $Q(a_o,b)$ bzw. $Q(a,b_o)$ eine Gerade). Wie Beispiel 8.4 (Abb. 8.4) zeigt, ist die Lösung nicht eindeutig.

3.) Auch das Kriterium der Minimierung der absoluten Abweichungen
$\Sigma|u_v| = \Sigma\,|y_v - a - bx_v| = \text{Min!}$
führt nicht in jedem Fall zu einem eindeutigen und befriedigenden Ergebnis (Bsp. 8.4) und bereitet zudem beträchtliche Schwierigkeiten bei der Bestimmung der Regressionskoeffizienten.

4.) Ein sinnvolles und zugleich eindeutiges Kriterium ist jedoch die **orthogonale Regression**, d.h. die Minimierung der orthogonalen (statt senkrechten) Abstände der Datenpunkte zur (orthogonalen) Regressionsgeraden.

Mit der orthogonalen Regressionsgeraden $\hat{y}^o_v = A + Bx_v$ (statt $\hat{y}_v = a + bx_v$ für die Regression nach der Methode der kleinsten Quadrate) erhält man für die quadrierten orthogonalen Abstände $d_v^2 = u_v^2/(1+B^2)$. Zur Schätzung von B ist B aus der quadratischen Gleichung $B^2 s_{xy} + B(s_x^2 - s_y^2) - s_{xy} = 0$ zu bestimmen (wobei nur eine Lösung erkennbar sinnvoll ist) und für A gilt $A = \bar{y} - B\bar{x}$ (Schwerpunktbedingung). Beispiel 8.5 demonstriert den Rechengang. Die Steigung B liegt zwischen den beiden Steigungen b und d, wobei das Verhältnis der Varianzen s_y^2/s_x^2 von Bedeutung ist. Außerdem ist die orthogonale Regression symmetrisch, d.h. die Geraden $\hat{y}^o$ und $\hat{x}^o$ "fallen zusammen".

5.) Eine Regressionsgerade, die durch beiden Punkte $P_1(\bar{x}_1, \bar{y}_1)$ und $P_2(\bar{x}_2, \bar{y}_2)$ "läuft" hat die Parameter $a = (\bar{x}_2\bar{y}_1 - \bar{x}_1\bar{y}_2)/(\bar{x}_2 - \bar{x}_1)$ und $b = (\bar{y}_2 - \bar{y}_1)/(\bar{x}_2 - \bar{x}_1)$. Man kann also z.B. eine Regressionsgerade durch die Punkte $P_1(\bar{x}_1, \bar{y}_1)$ und $P_2(\bar{x}_2, \bar{y}_2)$ wobei $\bar{x}_1$ der Mittelwert der ersten (mit kleineren x-Werten) Hälfte der Daten und $\bar{x}_2$ der Mittelwert zweiten Hälfte der Daten ist (Regressionsgerade nach Wald). Oder man teilt die Daten in drei gleich umfangreiche Teilgesamtheiten nach Maßgabe zunehmender x-Werte und bestimmt eine Regressionsgerade, die durch die Punkte $P_1(Z_{x1},Z_{y1})$ und $P_3(Z_{x3},Z_{y3})$ läuft, wobei Z_x und Z_y die Mediane (Zentralwerte) sind. Dies sind Vorschläge, eine Regressionsgerade zu bestimmen, die "robuster" ist gegenüber Ausreißern als die mit der Methode der kleinsten Quadrate errechnete Regressionsgerade.

6.) Natürlich kann man sich viele weitere Bedingungen vorstellen, die zu einer eindeutigen und sinnvollen Schätzung einer Geraden, die sich den Daten anpaßt, führen würden. Vorgeschlagen wird insbesondere auch eine Gewichtung der quadrierten Abweichungen also

$$\Sigma w_v(u_v)^2 = \Sigma w_v(y_v - a - bx_v)^2 = \text{Min}$$

so dass die Regressionsfunktion nach der Methode der kleinsten Quadrate "nur" der Spezialfall gleicher Gewichte $w_1 = w_2 = ... = w_n$ aller Beobachtungen ist.

Beispiel 8.4 (vgl. Abb. 8.4):

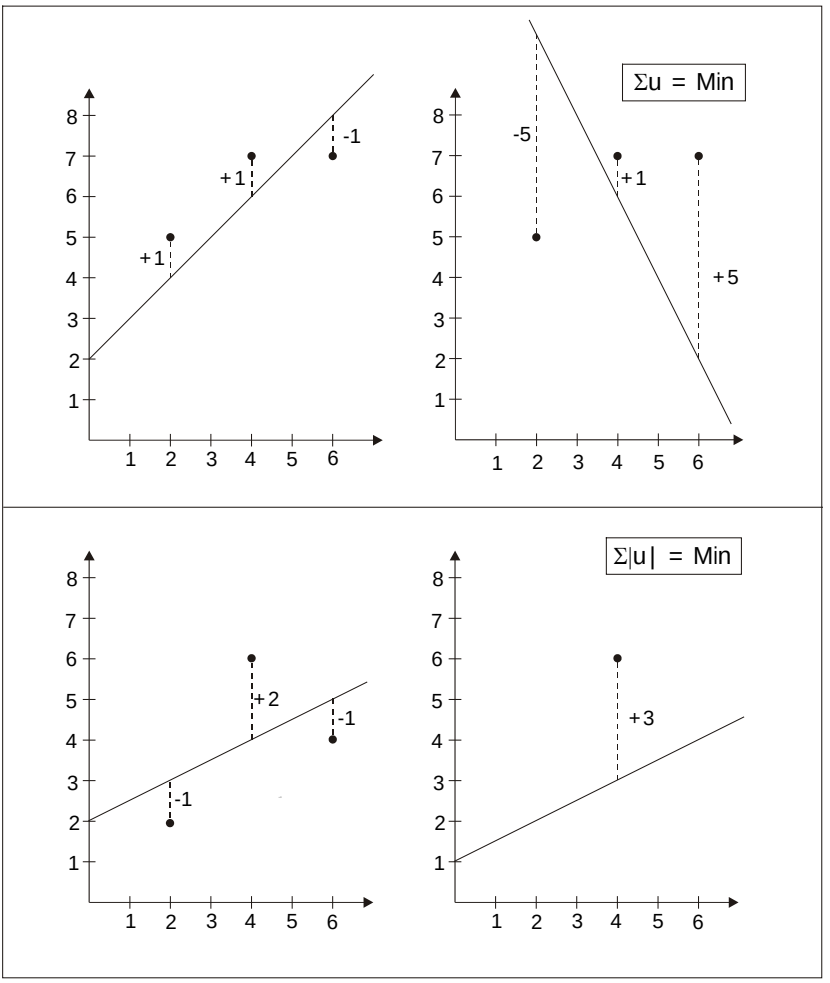
a) Gegeben seien die folgenden Daten (*Abb. 8.4 oben*)
(x,y): (2,5), (4,7) und (6,7)

Die Geraden $\hat{y} = 2 + x$ und $\hat{y} = 14 - 2x$ führen jeweils zu $\Sigma u_V = 1$ obgleich sie sich den Daten offensichtlich sehr unterschiedlich gut anpassen. Man erhält übrigens
für $\hat{y} = 2 + x$, $\Sigma|u_V| = 3$ und $\Sigma u_v^2 = 3$ dagegen
für $\hat{y} = 14 - 2x$ die Werte $\Sigma|u_V| = 11$ und $\Sigma u_v^2 = 51$,
die offensichtlich schlechter sind.

b) Gegeben seien die folgenden Daten (*Abb. 8.4 unten*)
(x,y): (2,2), (4,6) und (6,4).

Die Gerade $\hat{y} = 1 + 1/2x$ (Abb. 8.4 rechts unten) ist, gemessen am Kriterium $\Sigma|u_V| = \text{Min}$ besser ($\Sigma|u_V| = 3$) als die Gerade $\hat{y} = 2 + 1/2x$ (Abb. 8.4 links unten, $\Sigma|u_V| = 4$), nicht aber nach dem Kriterium der kleinsten Quadrate, denn man erhält $\Sigma u_v^2 = 9$ für die rechte und $\Sigma u_v^2 = 6$ für die linke Gerade.

Abb. 8.4: Alternativen zur Minimierung der Summe der Quadrate der (senkrechten) Abweichungen (vgl. Bsp. 8.4)



Beispiel 8.5:

Man bestimme die kleinste-Quadrate- und die orthogonale Regression sowie die orthogonalen (d_v) und die senkrechten (u_v) Residuen für die beiden Datensätze des Beispiels 8.4!

Lösung 8.5:

a.) Für die Daten (x;y): (2,5), (4,7) und (6,7) erhält man mit der Methode der kleinsten Quadrate: $\hat{y} = 13/3 + 1/2x$ und $\hat{x} = -11/2 + 3/2y$.

Wegen $s_x^2 = 8/3$ $\bar{x} = 4$ $s_{\bar{y}}^2 = 8/9$ $\bar{y} = 19/3$ und $s_{xy} = 4/3$ erhält man für die orthogonale Regression $\hat{y}^0 = 4{,}19259 + 0{,}53518x_v$ und somit die folgenden Residuen:

x	y	y	u	u^2	y^0	$(u^0)^2$
2	5	16/3 = 5,3	1/3	1/9	5,2629	0,06915
4	7	19/3 = 5,3	-2/3	4/9	6,3333	0,44444
6	7	22/3 = 7,3	1/3	1/9	7,4037	0,16766
Σ				2/3		0,67657

Man beachte, dass $\Sigma(u^0)^2 = 0{,}67657$ zwar größer ist als $\Sigma u^2 = 0{,}667$ (Σu^2 wird ja minimiert), nicht aber die Summe der quadrierten **orthogonalen** Abstände für die gilt: $\Sigma d_v^2 = \Sigma u_v^2/(1+b^2) = (2/3)/(5/4) = 0{,}533$ bei der Regressionsgeraden (kleinste Quadrate) bzw. $\Sigma d_v^2 = \Sigma(u^o{}_v)^2/(1+B^2) = 0{,}52593$ (dieses Σd_v^2 wird ja auch minimiert).

b) Für die Daten von Teil b des Beispiels 8.3 erhält man für die Regressionsfunktion $\hat{y} = 2 + ½x$ (kleinste Quadrate, die in Abb. 8.4 links unten eingezeichnete Gerade ist die Regressionsgerade) $\Sigma u_v^2 = 6$ und $\Sigma d_v^2 = 4{,}8$. Für die lineare orthogonale Regression erhält man $\hat{y}^o = x$ (also die 45^o Linie weil $s_x^2 = s_y^2$!) und $\Sigma u_v^2 = 8$ aber für die minimierte Größe $d_v^2 = ½\ 8 = 4$.

3. Ergänzungen zur linearen einfachen Regression

a) Standardisierte Variablen, gruppierte Daten

Def. 8.3: Zentrierung, Standardisierung

Die Transformation $x_v^z = x_v - \bar{x}$ und entsprechend $y_v^z = y_v - \bar{y}$ heißt Zentrierung und die Transformation $x_v^s = (x_v - \bar{x})/s_x = x_v^z/s_x$ (y_v^s ist analog definiert) stellt eine Standardisierung dar. Zum Begriff der Standardisierung vgl. Kap. 2. In einem anderen Sinne wird der Begriff in Kap. 9 gebraucht).

Folgerungen:

1. Regressionsgerade bei zentrierten Variablen x_v^z und y_v^z:

Die Gleichung $y_v = a + bx_v + u_v$ läßt sich wegen $\bar{y} = a + b\bar{x}$ umformen zu

(8.18) $y_v^z = bx_v^z + u_v$,

d.h. es verschwindet das Absolutglied a bei Zentrierung und b ist gem. Gl. 8.6a zu bestimmen, in diesem Fall also $b = \Sigma x_v^z \, y_v^z / \Sigma (x_v^z)^2$. Das Absolutglied a verschwindet und die Steigung b bleibt unverändert.
Das Verschwinden des Absolutglieds durch Verwendung zentrierter Variablen (d.h. durch Verschiebung des Koordinatensystems) ist nicht zu verwechseln mit der "**homogenen Regression**", bei der eine Regressionsgerade gesucht wird, die durch den Ursprung des (ursprünglichen) Koordinatensystems verläuft.

Es wird dann $\Sigma(u_v^*)^2 = \Sigma(y_v - b_H x_v)^2$ statt $\Sigma(u_v)^2 = \Sigma(y_v - a - bx_v)^2$ minimiert und das führt zur einen Normalgleichung $\Sigma x_v y_v = b_H \Sigma x_v^2$. Die Steigung bleibt nicht unverändert. Im Beispiel 8.1 erhält man die homogenen Regressionsgeraden $\hat{y}_v = 1{,}055 x_v$ (denn $b_H = \Sigma x_v y_v / \Sigma x_v^2 = 133/126 = 1{,}055$ während $b = 1/2$) und $\hat{x}_v = 0{,}8012 y_v$ (denn $d_H = \Sigma x_v y_v / \Sigma y_v^2 = 133/166$).

2. Regressionsgerade bei standardisierten Variablen x_v^S und y_v^S:

Die arithmetischen Mittel $\bar{x}^S$ und $\bar{y}^S$ der standardisierten Variablen sind Null und ihre Varianzen sind Eins. Bei Verwendung standardisierter Variablen folgt aus der ersten Normalgleichung der Regressionsfunktion

$\hat{y}^S = a_S + b_S x^S$ $\bar{x}^S = \bar{y}^S = 0$ dass $a_S = 0$.

Und für die zweite Normalgleichung erhält man

$a_S \Sigma x_v^S + b_S \Sigma (x_v^S)^2 = \Sigma x_v^S y_v^S = n r_{xy}$, woraus unter Berücksichtigung von $a_S = 0$ und $\frac{1}{n}\Sigma (x_v^S)^2 = \frac{1}{n}\Sigma [(x_v - \bar{x})/s_x]^2 = s_x^2 / s_x^2 = 1$ folgt

(8.19) $b_S = r_{xy}$.

In der Regressionsfunktion $\hat{y}_v^S = a_S + b_S x_v^S$ mit den standardisierten Variablen x_v^S und y_v^S ist also der Ordinatenabschnitt $a_S = 0$ und die Steigung b_S gleich dem Korrelationskoeffizienten. Die beiden durch den Ursprung verlaufenden Regressionsgeraden lauten

$$\hat{y}_v^S = r_{xy} \cdot x_v^S \qquad \text{und} \qquad \hat{x}_v^S = r_{xy} \cdot y_v^S.$$

Gl. 8.19 liefert auch den Grund dafür, dass die in Bsp. 8.3 Teil c und d dargestellte regression to the mean umso schneller erfolgt, je geringer r ist, denn bei standardisierten Variablen ist die Regressionsfunktion $\hat{y}^S = b_S \, x^S = r_{xy} x^S$, so dass die

Standardabweichung der Regresswerte $\hat{y}^S$ genau r_{xy} beträgt (die Standardabweichung von x^S ist wegen der Standardisierung natürlich 1). Mithin streut $\hat{y}^S$ um den Mittelwert 0 umso weniger, je kleiner betragsmäßig r_{xy} ist.

Bei **gruppierten Daten** erhält man mit der Notation von Def. 7.2

$$(8.20a) \qquad a = \frac{\Sigma x_i^2 n_{i.} \Sigma y_j n_{.j} - \Sigma x_i n_{i.} \Sigma x_i y_j n_{ij}}{n \Sigma(x_i^2 n_{i.}) - (\Sigma x_i n_{i.})^2}$$

$$(8.20b) \qquad b = \frac{n\Sigma x_i y_i n_{ij} - \Sigma x_i n_{i.} \Sigma y_i n_{.j}}{n \Sigma(x_i^2 n_{i.}) - (\Sigma x_i n_{i.})^2} = \frac{n^2 s_{xy}}{n^2 s_x^2}$$

für die Koeffizienten a und b der Regresssionsgerade $\hat{y}$ und entsprechend c und d für die Regressionsgerade $\hat{x}$.

Bei gruppierten Daten kann man auch die in Kapitel 7 dargestellten Regressionslinien bestimmen. Die bedingten Mittelwerte

$$\bar{y}|x_i = \frac{\Sigma_j y_j n_{ij}}{n_{i.}}$$

müssen nicht notwendig auf einer Geraden und schon gar nicht auf der Regressionsgeraden $\hat{y}$ liegen (und entsprechend müssen die Werte $\bar{x}|y_j$ nicht identisch sein mit $\hat{x}$).

Die Regressions***linien*** (vgl. Def. 7.6) sind i.d.R. von den Regressions***geraden*** verschieden.

b) Exkurs zum Regressionsmodell

Es wurde gelegentlich darauf hingewiesen, dass die rein deskriptive Behandlung der Regression, die allein in diesem Rahmen dargestellt werden soll, von dem stochastisch fundierten Modell der Regression zu unterscheiden ist. So ist es z.B. im Kontext dieses Modells nicht mehr möglich, die Abhängigkeiten zwischen X und Y einfach zu vertauschen.
Bei der modellmäßigen Erfassung des Zusammenhangs der beiden Variablen X und Y wird davon ausgegangen, dass für die wahre Regression (in der Grundgesamtheit) gilt:

$$(8.21) \qquad Y = \alpha + ßx + U.$$

Hierin ist U (und deshalb auch Y) eine Zufallsvariable, während x nichtzufällig schwankt[2] (die Varianz σ_u^2 in der Grundgesamtheit ist nicht Null). Die Parameter α und ß sind die Regressionskoeffizienten der "wahren" Regressionsfunktion der Grundgesamtheit, die es anhand der n beobachteten Wertepaare (x_v, y_v) einer Stichprobe zu schätzen gilt. Dabei sind die Werte y_v als Realisationen einer Zufallsvariable Y aufzufassen.

Es werden folgende Annahmen über das Modell und die nicht zu beobachtende Zufallsvariable U gemacht:

1. Die x_v - Werte (v=1,..,n) sind fest vorgegeben, d.h. sie sind frei von Meßfehlern und somit keine Zufallsvariablen.
2. Die y_v - Werte sind wegen der Überlagerung durch eine Störgröße U_v, deren Realisation u_v genannt sei: $y_v = \alpha + ßx_v + u_v$. Bekannt sind aber weder α und ß noch u_v. Vielmehr sind α, ß und u_v mit den Beobachtungen x_v, y_v zu schätzen. Die (Stichproben-) Schätzwerte heißen $a = \hat{\alpha}$ und $b = \hat{\beta}$; für u_v (wie bisher immer bezeichnet) wäre $\hat{u}_v$ eigentlich die bessere Bezeichnung.
3. Die zufälligen Störvariablen U_v sollen folgende Eigenschaften haben:
   a) $E(U_v) = 0 \quad (v=1,..,n)$
   b) $Var(U_v) = \sigma_u^2 = \sigma$
   c) die Größen $U_1,...,U_n$ sind stochastisch unabhängig und
   d) unabhängig identisch normalverteilt mit $E(U) = 0$ und $\sigma_u = \sigma$.

Bemerkungen zu Voraussetzung Nr. 3:

zu a)
Diese Forderung ist nicht damit zu verwechseln, dass die mit der Methode der kleinsten Quadrate **geschätzte** Störgröße u im Mittel Null ist. Man beachte, dass E(U) eine Aussage über einen Erwartungswert ist, der sich auf eine Wahrscheinlichkeitsverteilung bezieht. Es ist ja auch ein Stichprobenmittel von $\bar{x} = 0$ nicht gleichbedeutend damit, dass in der Grundgesamtheit $\mu = 0$ ist. Aus a) folgt auch, dass x und U nicht korreliert sind, was ebenfalls nicht damit zu verwechseln ist, dass der Regressor mit der **geschätzten** Störgröße nicht korreliert (zweite Normalgleichung).

zu b)
Diese Eigenschaft wird auch Homoskedastizität (Homoskedastie) genannt und sie besagt, dass die U_v, unabhängig von dem jeweiligen x-Wert, alle dieselbe Varianz σ_u^2 haben.

---

[2] Abweichend von der bisherigen Notation (X = Variable, x = konkreter Wert einer Variable) soll hier zwischen großen und kleinen Buchstaben in einer Weise unterschieden werden, wie das in der Induktiven Statistik üblich ist: Y = Zufallsvariable, x = nicht zufällige Variable oder Realisation der Zufallsvariable X.

zu c)

Dies bedeutet v.a. bei Zeitreihen, dass z.B. einem großen (kleinen) Wert von U_t nicht notwendig ein großer (kleiner) Wert U_{t+1} folgen darf, d.h. dass die Störvariablen nicht autokorreliert sein dürfen.

zu d)

Wenn es nicht von Interesse ist, eine Intervallschätzung oder einen Test für α, ß, σ oder die Korrelation durchzuführen, ist diese Annahme nicht erforderlich. Es wird oft übersehen, dass für eine rein deskriptive Behandlung der Regression alle obigen Annahmen über U nicht erforderlich sind.

Ist 3a nicht erfüllt, so sind a und b keine erwartungstreue Schätzer für α und ß. Von 3b und 3c wird Gebrauch gemacht, bei der Bestimmung der Varianzen und Kovarianzen der Stichprobenverteilungen von a und b und wäre 3d nicht gegeben, so wären die Schätzer a und b (für α und ß) bzw. s_u^2 für σ_u^2 nicht t- bzw. χ^2-verteilt.

Weitere Bemerkungen und Zusammenfassung:

- Man kann auch unterscheiden das "klassische" Modell der "Regression" $Y = \alpha + ßx + U$ (nicht-stochastischer Regressor x) vom Modell der Korrelation $Y = \alpha + ßX + U$ (stochastischer Regressor X).

- Es ist, wie gesagt, im stochastischen Modell nicht mehr möglich, die Abhängigkeiten zwischen X und Y einfach zu vertauschen, wie dies bei rein deskriptiver Betrachtung möglich ist, denn: gilt z.B. bei einem stochastischen Regressor X in der Regression $Y = \alpha_o + \alpha_1 X + U$ die Annahme $E(UX) = 0$, so kann für die Umkehrfunktion $X = ß_o + ß_1 Y + V$ wegen $ß_o = -\alpha_o/\alpha_1$ und $ß_1 = 1/\alpha_1$ sowie $V = -U/\alpha_1$ nicht auch zugleich $E(VY) = 0$ gelten. Vielmehr ist $E(VY) = -\alpha_1^{-2}E(UY) = -\alpha_1^{-2}E(U^2) = -\sigma_u^2/\alpha_1^2 < 0$.

- Die aufgrund der Stichprobe zu schätzenden Parameter sind bei einfacher Regression: α, ß und (was häufig vergessen wird) $\sigma^2 = \sigma_u^2$ (die Varianz der Störgröße) und die Schätzwerte sind a (für α), b (für ß) und $\hat{\sigma}^2 = \Sigma u^2/(n-2)$ für σ^2 der Varianzschätzer, was nicht zu verwechseln ist mit der Stichprobenvarianz $s^2 = \Sigma u^2/n$.

Übersicht 8.1: Größen und Parameter des Modells der einfachen (multiplen) Regression

Größen	beobachtbar	nichtbeobachtbar	Schätzwert
nichtzufällig[a]	$\mathbf{x}$[b] ($\mathbf{x}_1,...,\mathbf{x}_p$)	$\alpha,ß,\sigma_u,^2$ ($ß_o,ß_1,...,ß_p,\sigma_u,^2$)	a,b,$\hat{\sigma}_u,^{2}$[c] ($b_o,b_1,...,b_p,\hat{\sigma}_u,^2$)
zufällig	$\mathbf{y}$	Störgröße $\mathbf{U}$	$\hat{\mathbf{y}}$ und $\hat{\mathbf{u}}$ bzw. $\mathbf{u}$

(a) nichtzufällige Größen, bzw. zu schätzende Konstanten im Modell; zufällige Größen sind dagegen Y und U, wobei y_v eine Realisation von Y ist;

(b) mit $\mathbf{x}$,($\mathbf{x}_1$,$\mathbf{x}_2$,...), $\mathbf{y}$, $\mathbf{U}$, $\mathbf{u} = \hat{\mathbf{u}}$ und $\hat{\mathbf{y}}$ sind jeweils Vektoren (Spaltenvektoren mit n Zeilen) gemeint, also $\mathbf{x}' = [x_1\ x_2\ ...x_n]$ oder $\mathbf{x}_1' = [x_{11}\ x_{12}\ ...\ x_{1n}]$.

(c) Der Schätzer der Varianz der Störgröße $\sigma_u^2 = \Sigma u^2/(n-p)$ (bei p-1 Regressoren, also z.B. p=2 bei einfacher Regression) ist erwartungstreu, also $E(\hat{\sigma}_u^2) = \sigma_u^2 = \sigma^2$. Die oben betrachtete Größe $s^2 = \Sigma u^2/n$ ist aber nicht erwartungstreu.

4. Multiple lineare Regression

a) Beschreibung des Modells

Multiple Regression (Mehrfachregression) heißt Y auf den Einfluß nicht nur einer, sondern mehrerer erklärender Variablen $X_1,..,X_p$ zurückzuführen. Im linearen Fall bedeutet dies, das Modell:

$$(8.24)\quad Y_v = ß_0 + ß_1x_{1v} + ß_2x_{2v} + ... + ß_px_{pv} + U_v \quad (v=1,..,n)$$

mit Stichprobendaten durch die Gleichung

$$(8.25)\quad y_v = b_0 + b_1x_{1v} + b_2x_{2v} + ... + b_px_{pv} + u_v \quad (v=1,..,n)$$

zu schätzen, wobei

$$(8.26)\quad \hat{y}_v = b_0 + b_1x_{1v} + b_2x_{2v}.+... + b_px_{pv}$$

die (geschätzte) Regressionsfunktion ist.

Das Absolutglied b_0 ist für die Interpretation irrelevant und kann beseitigt werden durch Zentrierung (deviation scores) oder Standardisierung. Ein Beispiel für eine zweifache Regression ist die bereits erwähnte Abhängigkeit des Jahresumsatzes (Y) von den Ausgaben für Werbung (X_1) und den Wareneinkäufen (X_2). Das Streuungsdiagramm ist dann eine dreidimensionale Darstellung (Y, X_1, X_2) und der Ausdruck Punkt-"wolke" wäre jetzt (im Unterschied zur einfachen [bivariaten] Regression) erstmals korrekt. Bei zwei Regressoren ist die Funktion $y = y(x_1,x_2)$ eine Ebene. Der Regresswert $\hat{y}$ ist ein Punkt auf der Ebene. Die Beobachtungen (Datenpunkte) liegen auf, ober- oder unterhalb dieser Ebene. Die senkrechten [in Richtung der y-Achse] Abstände von dieser Ebene stellen die Störgröße dar.

Def. 8.4: multiple lineare Regression

a) Das Gleichungssystem (Gleichungen) 8.26 beschreibt eine multiple lineare Regression mit p Regressoren und für die v-te Beobachtung (v=1,2,...,n) gilt:

$$(8.26a)\quad \hat{y}_v = b_0 + b_1 x_{1v} + b_2 x_{2v} + ... + b_p x_{pv} = \sum_{i=0}^{p} b_i\, x_{iv} \qquad \text{mit } x_{0v} = 1$$

oder in anderer Schreibweise, z.B. bei drei Regressoren:

$$(8.26b)\quad \hat{y} = b_{y0.123} + b_{y1.23} x_1 + b_{y2.13} x_2 + b_{y3.12} x_3$$

b) Die Koeffizienten b_i oder $b_{yi.jk}$ heißen ***partielle Regressionskoeffizienten*** (als Maß für den isolierten Einfluß des Regressors X_i "bei Konstanz" von X_j, X_k usw.), da die partielle Ableitung der Regressionsfunktion $b_i = \beta y/\beta x_i$ ist, bzw. (synonym) multiple Regressionskoeffizienten.

c) Zur Beschreibung des Modells werden auch die folgenden Größen herangezogen: ***multiple Korrelation*** und ***multiple Bestimmtheit*** sowie die ***partielle Korrelation*** (vgl. Bem. Nr. 3 und Teile c und d dieses Abschnitts).

d) Unter ***Orthogonalität*** versteht man, dass (bei stochastischen Regressoren) die Regressoren $X_1,..,X_p$ stochastisch unabhängig sind. Wenn (im anderen Extremfall) zwischen den unabhängigen Variablen $X_1,..,X_p$ lineare Abhängigkeiten bestehen spricht man von offener ***Kollinearität*** (Multikollinearität). In diesem Fall ist eine Kleinste-Quadrate-Schätzung der Regressionskoeffizienten nicht möglich.

Bemerkungen zu Def. 8.4:

1. Gl. 8.26 stellt im p+1 dimensionalen Raum eine p - dimensionale Hyperebene dar. Im Fall einer zweidimensionalen Regression [zwei Regressoren X_1 und X_2] ist dies, wie gesagt, eine Ebene im anschaulichen Sinne (im dreidimensionalen [Y, X_1, X_2] Raum).

2. Es sollen die gleichen Modellannahmen wie bisher gelten und zusätzlich Kollinearität (Multikollinearität) möglichst ausgeschlossen sein, d.h. im Idealfall Orthogonalität herrschen. In der Praxis ist weder Orthogonalität noch offene Kollinearität üblich, sondern eine **verdeckte Kollinearität**, d.h. eine mehr oder weniger starke Korrelation zwischen den unabhängigen Variablen $X_1,..,X_p$. Bei offener Kollinearität ist eine Kleinste-Quadrate-Schätzung der Regressionskoeffizienten nicht möglich und bei starker verdeckter Kollinearität ist sie nur mit großem (und unbekanntem) Fehler möglich, also sehr unzuverlässig.

3. Während die Begriffe partielle und multiple Regressionskoeffizienten synonym sind, gilt dies keineswegs für die partiellen und multiplen Korrelationskoeffizienten. Der

   - **multiple Korrelationskoeffizient**, etwa $R_{y.123}$ ist die einfache Korrelation zwischen y und $\hat{y} = f_3(x_1,x_2,x_3)$; entsprechend wäre z.B. $R_{y.12}$ die Korrelation zwischen y und $\hat{y} = f_2(x_1,x_2)$ (es gilt übrigens $R_{y.12} \mid R_{y.123}$);
   - **partielle Korrelationskoeffizient**, etwa $r_{y1.23}$ ist die Korrelation zwischen Y und X_1 wobei der Einfluß von X_2 und X_3 in einem noch beschriebenen Sinne "ausgeschaltet" ist.

4. Das Kriterium für die beste Anpassung der Regressionsfunktion (Regressions- [hyper] -ebene) ist wieder die Minimierung der Quadrate der Abweichungen u_v.

Auf die Regressionsfunktion wird somit wieder die Methode der Kleinsten Quadrate angewendet, d.h. es sind die Werte $b_0,b_1,...,b_p$ zu bestimmen, die die Funktion $Q(b_0,b_1,...,b_p)$ minimieren:

$$Q(b_0,b_1,...,b_p) = \Sigma u_v^2 = \Sigma(y_v - \hat{y}_v)^2 = \text{Min}$$

mit $Q(b_0,b_1,...,b_p) = \Sigma u_v^2 = \Sigma(y_v - b_0 - b_1x_{1v} - ... - b_px_{pv})^2$.
Q ist partiell nach $b_o,b_1,..,b_p$ zu differenzieren und die Ableitungen sind Null zu setzen, womit man p+1 Normalgleichungen erhält. Im Falle von zwei Regressoren lautet das System der drei Normalgleichungen (jeweils summiert über v = 1,2,...,n):

$$\begin{array}{llllllll}
1. & b_o n & + & b_1\Sigma x_{1v} & + & b_2\Sigma x_{2v} & = & \Sigma y_v \\
2. & b_o\Sigma x_{1v} & + & b_1\Sigma x_{1v}^2 & + & b_2\Sigma x_{1v}x_{2v} & = & \Sigma x_{1v}y_v \\
3. & b_o\Sigma x_{2v} & + & b_1\Sigma x_{1v}x_{2v} & + & b_2\Sigma x_{2v}^2 & = & \Sigma x_{2v}y_v.
\end{array}$$

Man beachte, dass die Schätzwerte b_o, b_1, b_2 linear in y (und damit auch in u) sind, also lineare Schätzer für $ß_o$, $ß_1$, $ß_2$ sind.

b) Darstellung in Matrixschreibweise

Erheblich einfacher und übersichtlicher sind die Schätzgleichungen in Matrixschreibweise darzustellen. Dazu vereinbart man die Datenmatrix **X**, den Datenvektor **y**, den Vektor **u** der geschätzten Residuen und den Vektor **b** der zu schätzenden Regressionskoeffizienten wie folgt:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & . & x_{p1} \\ 1 & x_{12} & x_{22} & . & x_{p2} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & . & x_{pn} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_p \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_n \end{bmatrix}$$

Die Spaltenvektoren y und u haben jeweils n Zeilen (n Beobachtungen) **X** hat n Zeilen und p+1 Spalten und die Elemente x_{jv} (v = 1,2,...,n und j = 0,1,...,p). Der dummy - Regressor X_o (mit $x_{o1} = ... = x_{on} = 1$) wird erforderlich zur Schätzung des Absolutglieds. Das Gleichungssystem 8.26 stellt sich dann in Matrixschreibweise wie folgt dar:

(8.27) $\mathbf{y} = \mathbf{Xb} + \mathbf{u}$.

Das Matrizenprodukt **X'X** ist die quadratische Matrix der n-fachen Anfangsmomente bzw. Anfangsproduktmomente:

$$\begin{bmatrix} 1 & 1 & ... & 1 \\ x_{11} & x_{12} & ... & x_{1n} \\ . & . & . & . \\ . & . & . & . \\ x_{p1} & x_{p2} & ... & x_{pn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & ... & x_{p1} \\ 1 & x_{12} & ... & x_{p2} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1n} & ... & x_{pn} \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 & ... & \Sigma x_p \\ \Sigma x_1 & \Sigma x_1^2 & ... & \Sigma x_1 x_p \\ . & . & . & . \\ . & . & . & . \\ \Sigma x_p & \Sigma x_1 x_p & ... & \Sigma x_p^2 \end{bmatrix}$$

und man erkennt leicht, dass die Normalgleichungen in Matrixschreibweise lauten:

(8.28) $\mathbf{X'Xb} = \mathbf{X'y}$ (Normalgleichungen).

Die zu minimierende Residualquadratsumme ist:

$\Sigma u_v^2 = \mathbf{u'u} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = \mathbf{y'y} - \mathbf{b'X'y} - \mathbf{y'Xb} + \mathbf{b'X'Xb}$

Berücksichtigt man, dass **b'X'y** ein Skalar ist und deshalb $(\mathbf{b'X'y})' = \mathbf{y'Xb}$ ist, gilt weiter:

(8.29) $\mathbf{u'u} = \mathbf{y'y} - 2\mathbf{y'Xb} + \mathbf{b'X'Xb}$.

Durch partielle Differentiation von Gl. 8.29 nach dem Vektor **b** erhält man das Normalgleichungssystem der multiplen Regression, d.h. Gl. 8.28, in ausführlicher Schreibweise ohne den Summationsindex v

$$
\begin{array}{lllll}
b_0 n & + b_1\Sigma x_1 & + ... + & b_m\Sigma x_m & = \Sigma y \\
b_0\Sigma x_1 & + b_1\Sigma x_1^2 & + ... + & b_m\Sigma x_1 x_m & = \Sigma x_1 y \\
. & . & & . & . \\
b_0\Sigma x_p & + b_1\Sigma x_1 x_p & + ... + & b_m\Sigma x_m^2 & = \Sigma x_m y.
\end{array}
$$

Werden beide Seiten der Gl. 8.28 mit der Inversen $(\mathbf{X'X})^{-1}$ prämultipliziert, so gelangt man zum Schätz-Vektor der Regressionskoeffizienten:

(8.30) $\mathbf{b} = (\mathbf{X'X})^{-1}\,\mathbf{X'y}$.

Voraussetzung für die Berechnung von **b** nach Gl. 8.30 ist natürlich, dass die Inverse der Matrix **X'X** überhaupt existiert, was z.B. nicht der Fall ist bei offener Multikollinearität.

Folgerungen aus den Normalgleichungen:

1) Aus Gl. 8.29 folgt übrigens wegen $\mathbf{y'} = \mathbf{b'X'} + \mathbf{u'}$ und deshalb $\mathbf{y'Xb} = \mathbf{b'X'Xb}$
$\mathbf{u'u} = \mathbf{y'y} - 2\mathbf{y'Xb} + \mathbf{y'y}$
und damit die folgende Beziehung:

(8.31) $\mathbf{y'y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{u'u}$ (Varianzzerlegung) .

2) Aus Gl. 8.28 folgt wegen $\mathbf{y} = \mathbf{Xb} + \mathbf{u}$, dass das Produkt **X'u** einen Nullvektor ergeben muss also

(8.32) $\mathbf{X'u = 0}$

gilt (oder ausführlich $\Sigma u = \Sigma x_1 u = \Sigma x_2 u = ... = \Sigma x_p u = 0$).

3) Aus Gl. 8.31 und 8.32 folgt generell $s_{\hat{y}}^2 = s_{y\hat{y}}$, so dass ganz allgemein im Regressionsmodell gilt:

(8.17a) $r_{y\hat{y}} = R_{y.ij...} = \dfrac{s_{\hat{y}}}{s_y} = \sqrt{B_{y.ij...}}$

c) Multiple Korrelation und multiple Bestimmtheit

Die Güte des Gesamtzusammenhangs wird beschrieben mit dem multiplen Korrelationskoeffizient $R_y = R_{y.12}$ [oder $r_{y.12}$] (bei zwei Regressoren X_1 und X_2, was im folgenden als Beispiel dienen mag), bzw. mit dem multiplen Bestimmtheitsmaß R_y^2. Anders als bei der einfachen Regression macht

es hier (wie auch bei nichtlinearer Regression) keinen Sinn, zwischen positiver und negativer Korrelation zu unterscheiden, so dass R nie negativ sein kann.
Die Bezeichnung des multiplen Korrelationskoeffizienten mit $R_{y.12}$ macht deutlich, dass in der Regressionsfunktion Y durch X_1 und X_2 "erklärt" wird.

Def. 8.5: multiple Korrelation und Bestimmtheit

a) Der multiple Korrelationskoeffizient $R_{y.12}$ ist die einfache Korrelation zwischen y und $\hat{y} = b_{y1.2} \cdot x_1 + b_{y2.1} \cdot x_2$ also zwischen dem beobachteten Wert y und dem Regresswert $\hat{y}$

(8.33) $\quad R_{y.12} = r_{y\hat{y}}$.

b) Die multiple Bestimmtheit $B_{y.12}$ ist analog zum einfachen Bestimmtheitsmaß der Anteil der durch die multiple Regression erklärten Varianz der abhängigen Variable Y an der Gesamtvarianz von Y also:

(8.34) $\quad B_{y.12} = \frac{s^2_{\hat{y}}}{s^2_y}$ (mit $\hat{y}$ als Funktion von x_1 und x_2)

und analog bei mehr als zwei Regressoren; es ist nach Gl. 8.17a gleich $R^2_{y.12}$.

c) Die Größe $U_{y.12} = 1 - B_{y.12} = s^2_u/s^2_y$, wobei s^2_u die Varianz der Störgröße U ist, ist entsprechend das multiple Unbestimmtheitsmaß.

Auch für das multiple Bestimmtheitsmaß liegt der Wertebereich zwischen 0 (keine Erklärung) und 1 (alle Beobachtungen "liegen auf" der Regressionshyperebene, perfekte Erklärung von Y durch X_1, X_2, ..., X_p). Den multiplen Korrelationskoeffizienten $R_{y.12..p}$ zwischen Y und $X_1,..,X_p$ erhält man auch aus der positiven Wurzel des multiplen Bestimmtheitsmaßes.

d) Partielle Regressions- und Korrelationskoeffizienten

Der partielle Regressionskoeffizient erster Ordnung $b_{y1.2}$ ist der isolierte (partielle) Einfluss von X_1 auf Y in dem Sinne, dass der Einfluß

- von X_2 auf Y in Höhe von $b_{y2}x_2$ aus der einfachen Regression $y = b_{y2}x_2 + u$ und
- von X_2 auf X_1 in Höhe von $b_{12}x_2$ aus der einfachen Regression

$$x_1 = b_{12}x_2 + v$$

ausgeschaltet ist (wobei hier zentrierte Werte [deshalb kein Absolutglied] vorausgesetzt sind).
Dann ist $b_{y1.2}$ der Regressionskoeffizient in der einfachen Regression von $u = y - b_{y2}x_2$ (Regressand) auf $v = x_1 - b_{12}x_2$ (Regressor) und der partielle Korrelationskoeffizient $r_{y1.2}$ ist der (einfache) Korrelationskoeffizient zwischen u und v.

Man beachte, dass $b_{y1.2}$ und $b_{1y.2}$ verschiedene Größen sind, nicht aber $r_{y1.2}$ und $r_{1y.2}$. Es gilt ferner:

$$(8.35)\quad r_{ij.k} = \sqrt{b_{ij.k}b_{ji.k}},$$

analog zur einfachen (bivariaten) Regression.
Bei der einfachen Regressionsanalyse gab es nur einen Korrelationskoeffizienten r. Er steht auf der gleichen Ebene wie der multiple Korrelationskoeffizient $R_{y.12..}$, da auch bei einfacher Korrelation r als Korrelation zwischen y und $\hat{y}$ dargestellt werden kann.

e) Standardisierte Regressionskoeffizienten, Rekursionsformeln

1. Standardisierte Regressionskoeffizienten

Man beachte:

- Bei Kollinearität misst der Regressionskoeffizient $b_{ij.k}$ nicht den isolierten Einfluß von X_j (Regressor, "Ursache") auf X_i (Regressand, "Wirkung"), sondern einen "gemischten Einfluß von X_j ***und*** X_k, ohne dass man diese Einflüsse trennen könnte.
- Aussagen dergestalt, welche Einflußgröße die (kausal) wichtigere sei sind außerdem auch deshalb nur mit großer Vorsicht zu machen, weil die Regressionskoeffizienten maßstabsabhängig sind.

Es kann sehr wohl X_k ein bedeutsamerer (größerer) Einfluß auf Y sein, als X_i - auch wenn $b_{yi.k} > b_{yk.i.}$ gilt - einfach deshalb, weil die Varianz von X_i kleiner ist als die von X_k. Statt der mit Gl. 8.30 zunächst errechneten nichtstandardisierten Regressionskoeffizienten $b_{yi.k}$ sind die standardisierten Regressionskoeffizienten $c_{yi.k}$ (oft auch $ß_{yi.k}$ genannt, was dann aber mit den "wahren" Koeffizienten der Gl. 8.24 verwechselt werden könnte) zu berechnen nach der Formel:

$$(8.36)\quad c_{yi.k} = b_{yi.k}\,\frac{s_i}{s_y}\,,$$

wobei s^2 die entsprechenden Varianzen sind.

Man erhält den Vektor **c** der standardisierten Regressionskoeffizienten aus dem Normalgleichungssystem

$$(8.28a)\quad \mathbf{Rc} = \mathbf{r}$$

mit der Matrix **R** der Korrelationskoeffizienten (anstelle von **X'X** in Gl. 8.28 und dem Vektor **r** der (einfachen) Korrelationen zwischen den Regressoren und Y (anstelle von **X'y**). Bei zwei Regressoren gilt z.B.:

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} c_{y1.2} \\ c_{y2.1} \end{bmatrix} = \begin{bmatrix} r_{y1} \\ r_{y2} \end{bmatrix}$$

2. Rekursionsformeln

Der partielle Regressionskoeffizient erster Ordnung $b_{ij.k}$ läßt sich rekursiv aus den einfachen Regressionskoeffizienten (und entsprechend auch die partielle Korrelation erster Ordnung aus den einfachen Korrelationen) wie folgt bestimmen:

$$(8.37)\quad b_{ij.k} = \frac{b_{ij} - b_{ik}b_{ki}}{1 - b_{ik}b_{ki}}\,,$$

$$(8.38)\quad r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}\,.$$

Man kann die partiellen Regressionskoeffizienten auch auf Varianzen und Kovarianzen zurückführen. Dann gilt:

$$(8.37a)\quad b_{ij.k} = \frac{s_{ij}s_k^2 - s_{jk}s_{ik}}{s_j^2 s_k^2 - s_{jk}}\,,$$

$$(8.37b)\quad b_{ij.k} = \frac{s_i}{s_j} \cdot \frac{r_{ij} - r_{jk}r_{ik}}{1 - r_{jk}^2}\,.$$

Der zweite Bruch in Gl. 8.37b stellt den standardisierten Regressionskoeffizienten $c_{ij.k}$ (gem. Gl. 8.36) dar. Wie man sieht sind bei unkorrelierten Regressoren ($r_{jk} = 0$) die partiellen Regressionskoeffizienten $b_{ij.k}$ identisch mit den einfachen b_{ij}.

Für die Interpretation der partiellen Regressionskoeffizienten mag auch der folgende Zusammenhang von Interesse sein:

(8.37c) $b_{ij} = b_{ij.k} + b_{ik.j} \cdot b_{jk}$,

wonach der einfache Regressionskoeffizient b_{ij} die Summe eines direkten Einflusses von X_j auf X_i (gemessen als $b_{ij.k}$) und eines indirekten Einflusses von X_k über X_j auf X_i (mit $b_{ik.j} \cdot b_{jk}$) ist. Deshalb mißt auch b_{ij} nicht, wie man meinen möchte den "echten" Einfluß von x_j auf x_i sondern stets einen direkten **und** indirekten Einfluß von x_j auf x_i. Der indirekte Einfluß kann auch dominieren, wie das bei einer Scheinkorrelation der Fall ist. Eine entsprechende Beziehung gilt auch für die Korrelation.

Entsprechende Formeln existieren auch für die partiellen Regressions- und Korrelationskoeffizienten zweiter Ordnung, die rekursiv aus den Koeffizienten erster Ordnung zu bestimmen sind.
Für die rekursive Bestimmung der multiplen Bestimmtheit gilt:

$$(8.39)\quad R^2_{i.jk} = \frac{r^2_{ij} + r^2_{ik} - 2r_{ij}r_{ik}r_{jk}}{1 - r^2_{jk}} \quad .$$

Aus rekursiven Beziehungen ergibt sich übrigens auch:

$$(8.40)\quad 1 - R_{y.123} = (1 - R_{y.12})(1 - r_{y3.12}),$$

so dass $R_{y.123} \geq R_{y.12}$ oder allgemein die multiple Bestimmtheit bei p Regressoren nicht kleiner sein kann als bei p-1 Regressoren[3]. Aus Gl. 8.40 folgt übrigens auch die folgende Interpretation der partiellen Bestimmtheit im Sinne der proportionalen Fehlerreduktion ([PRE] vgl. Def. 7.14):

$$(8.40a)\quad r^2_{y3.12} = \frac{R^2_{y.123} - R^2_{y.12}}{1 - R^2_{y.12}}$$

wonach $r_{y3.12},^2$ die Zunahme der Bestimmtheit von Y durch den Regressor X_3 (oder die Abnahme der Unbestimmtheit!) ins Verhältnis zur Unbestimmtheit mit den beiden Regressoren X_1 und X_2 setzt. Es gilt $R^2_{y.123} > R^2_{y.12}$ es sei denn $r_{y3.12} = 0$.

Zusammenhänge dieser Art veranlassen auch manche Anwender mit verschiedenen Verfahren der stufenweisen Regression (stepwise regression) durch Hinzufügen oder Weglassen von Regressoren zu einer Regressionsgleichung mit möglichst großer Bestimmtheit und wenig Regressoren zu gelangen. Es kann hier nicht näher begründet werden, warum ein solches Vorgehen mit Skepsis zu betrachten ist. Der Hauptgrund hierfür ist die Kollinearität. Denn nur bei Orthogonalität (d.h. Abwesenheit von Kollinearität) ist die multiple Bestimmtheit eine Summe der einfachen Bestimmtesten:

---

[3] Dem Umstand wird oft durch Berechnung einer korrigierten Bestimmtheit $R,^{-2}$ Rechnung getragen: $R,^{-2} = 1-(n-1)(1-R^2)/(n-p)$ wobei $R,^{-} \mid R$

(8.40b) $R^2_{y.123} = r^2_{y1} + r^2_{y2} + r^2_{y3}$.

Aus Gl. 8.28a folgt auch Gl. 8.37b, wonach $c_{ij.k} = (r_{ij} - r_{jk}r_{ik})/(1 - r^2_{jk})$ und dass bei Orthogonalität ($r_{jk} = 0$) gilt $c_{ij.k} = r_{ij}$. Ferner wird erkennbar, dass die multiple Bestimmtheit nur bei Orthogonalität mit Gl. 8.40b als Summe partieller Bestimmtheiten darstellbar ist. Wegen der Standardisierung ($s_y = s_1 ... = s_p = 1$) gilt dagegen z.B. bei zwei korrelierten Regressoren:

(8.41) $R^2_{y.12} = c_{y1.2},^2 + c^2_{y2.1} + 2c_{y1.2}c_{y2.1}r_{12}$.

Zur Summe der partiellen Bestimmtheiten tritt also bei Kollinearität noch das doppelte Produkt der c-Koeffizienten mit der positiven oder negativen Korrelation r_{12} hinzu, das bei Orthogonalität wegfällt.

Beispiel 8.6:

Der Bierverkauf einer Brauerei hänge wie folgt von der erklärenden Variable "Ausgaben für Fernsehwerbung" X_1 und einer zweiten erklärenden Variable "Ausgaben für Zeitungswerbung" X_2 ab. Es ergibt sich die folgende (natürlich fiktive) Datenkonstellation:

v	y_v	x_{1v}	x_{2v}
1	0,8	1,5	1,0
2	1,0	4,0	1,5
3	1,4	4,5	3,0
4	1,6	7,0	5,0
5	1,7	8,5	4,5
6	2,1	8,5	5,5
7	2,1	11,5	7,0
8	2,5	12,0	8,5

Man bestimme die lineare Zweifachregression $\hat{y} = b_0 + b_1x_1 + b_2x_2$ und zeige, ob und in welchem Maße die Hinzunahme eines zweiten Regressors (X_2) zur Erklärung von Y beiträgt. Hilfsangaben:

$$\mathbf{(X'X)^{-1}} = \begin{bmatrix} 0{,}725175874 & -0{,}171375851 & 0{,}1403529 \\ -0{,}171375851 & 0{,}173451736 & -0{,}238957444 \\ 0{,}1403529 & -0{,}238957444 & 0{,}350478607 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} 13{,}2 \\ 109{,}15 \\ 69{,}65 \end{bmatrix}$$

Lösung 8.6:

Man erhält

$$\mathbf{b} = \mathbf{(X'X)^{-1}\,X'y} = \begin{bmatrix} 0{,}642226963 \\ 0{,}026709722 \\ 0{,}181288202 \end{bmatrix}$$

also: $\hat{y} = 0{,}642226963 + 0{,}026709722x_1 + 0{,}181288202x_2$. Das (unkorrigierte) multiple Bestimmtheitsmaß beträgt $R^2_{y.12} = 0{,}957045$. Es wird also 95,7% der Varianz der abgesetzten Biermenge Y durch die Zweifachregressionsfunktion erklärt.

Für die einfache Regression von Y auf X_1 erhält man $\hat{y} = 0{,}569682 + 0{,}1503126x_1$ und für die Bestimmtheit $R^2_{y.1} = 0{,}916971$. Das Beispiel enthält eine beachtliche Kollinearität, denn X_1 und X_2 korrelieren mit 0,969171. Deshalb ist auch die multiple Bestimmtheit $R^2_{y.12}$ sehr viel kleiner als die Summe der einfachen Bestimmtheiten $R^2_{y.1} = r^2_{y1} = 0{,}916971$ und $R^2_{y.2} = r^2_{y2} = 0{,}955287$. Der Erklärungsbeitrag von X_2 ist nach Gl. 8.40a die partielle Bestimmtheit und er beläuft sich auf nur 48%:
$r^2_{y2.1} = (R^2_{y.12}-R^2_{y.1})/(1-R^2_{y.1}) = 0{,}040074/0{,}083029 = 0{,}48262$, so dass die partielle Korrelation nur 0,6947 beträgt. Man kann den partiellen Korrelationskoeffizienten auch mit $r_{y2.1} = (r_{y2.1} - r_{y1}\, r_{12})/\sqrt{(1 - r^2_{y1})(1 - r^2_{12})}$ rekursiv berechnen.

5. Nichtlineare Regression

Oft lassen theoretische Erkenntnisse oder die Betrachtung des Streuungsdiagramms darauf schließen, dass ein nichtlinearer Zusammenhang zwischen den Variablen besteht und somit eine nichtlineare Regressionsfunktion bestimmt werden muss, worauf hier nur kurz eingegangen werden kann.
Eine Regressionsfunktion ist **nichtlinear**

- **in den Variablen**, wenn Regressoren wie etwa x, x^2, x^{-1}, e^x, x_1x_2, sin(x) oder log(x) auftreten
- **in den** Regressionskoeffizienten (**Parametern**), wenn in der Gleichung Koeffizienten auftreten wie b^2 oder b^{-1} usw.
- in den Variablen **und** in den Parametern, wenn beides zutrifft, wie etwa bei der logistischen Funktion als Regressionsfunktion $\hat{y}_i = k/[1+\exp(a-bx_i)]$. Man kann dann auch von **intrinsischer** Nichtlinearität sprechen.

Ist eine Funktion allein nichtlinear in den Variablen oder in den Parametern, nicht aber in beiden, so kann sie meist linearisiert werden durch eine Variablen- bzw. Parametersubstitution oder durch eine Variablentransformation. Übersicht 8.2 gibt hierzu einige Hinweise. Wenn sie nichtlinear in den Variablen **und** in den Parametern ist, so sind andere Methoden notwendig (z.B. eine Näherung durch eine Taylorreihenentwicklung, die dann meist nach einigen Gliedern abgebrochen wird).

1. Variablensubstitution

Ist z.B. eine parabolische (Polynom zweiten Grades) Regression zu schätzen ($\hat{y}_v = b_0 + b_1x_v + b_2x_v^2$) so kann man x durch x_1 und x^2 durch x_2 substituieren, also neue Variablen definieren *(redefining variables)***.** Das führt zu den folgenden Normalgleichungen:

$$
\begin{aligned}
b_0 n &+ b_1 \Sigma x_1 &+ b_2 \Sigma x_2 &= \Sigma y \\
b_0 \Sigma x_1 &+ b_1 \Sigma x_1^2 &+ b_2 \Sigma x_1 x_2 &= \Sigma x_1 y \\
b_0 \Sigma x_2 &+ b_1 \Sigma x_1 x_2 &+ b_2 \Sigma x_2^2 &= \Sigma x_2 y
\end{aligned}
$$

die äquivalent sind dem Gleichungssystem:

$$
\begin{aligned}
b_0 n &+ b_1 \Sigma x &+ b_2 \Sigma x^2 &= \Sigma y \\
b_0 \Sigma x &+ b_1 \Sigma x^2 &+ b_2 \Sigma x^3 &= \Sigma xy \\
b_0 \Sigma x^2 &+ b_1 \Sigma x^3 &+ b_2 \Sigma x^4 &= \Sigma x^2 y \,.
\end{aligned}
$$

Die Parabel $y = b_0 + b_1 x + b_2 x^2$ ist zwar nicht linear in x, wohl aber linear in den Variablen x_1 und x_2 und zu schätzen mit einer multiplen Regression mit den Regressoren $x_1 = x$ und $x_2 = x^2$.

Bei Nichtlinearität in den Variablen ist eine Linearisierung durch Variablensubstitution möglich. Das bedeutet, dass die Formeln der linearen Regression zur Bestimmung der Regressionskoeffizienten bei Verwendung der transformierten Variablen angewandt werden können und (im Unterschied zu vielen Fällen der Variablentransformation) auch die üblichen Annahmen über die (additive) Störgröße gültig bleiben.

Übersicht 8.2: Einige leicht linearisierbare Funktionen

Die Übersicht ist ähnlich aufgebaut, wie Übers. 9.5, in der zugleich Trendmodelle behandelt werden (vgl. auch Kap. 11). Allerdings ist hier x an die Stelle von t in Übers. 9.5 gesetzt worden. Der Parameter k ist wie in Übers. 9.5 ein Sättigungsniveau.

	Funktion f(x) y=f(x)	linearisiert: $y^*=a^*+b^*x^*$			
		y^*	a^*	b^*	x^*
1	$(a+bx)^\alpha$				
1a	α=1:Gerade y =a+bx	y	a	b	x
1b	α=-1: $\frac{1}{a+bx}$	$\frac{1}{y}$	a	b	x
1c	α=½: $\sqrt{a+bx}$	y^2	a	b	x
1d	Potenzfunktion bx^α	ln(y)	ln(b)	α	ln(x)
2	Parabel $a+bx+cx^2$ (1)				
3	$a\cdot\exp(bx^\alpha)$				
3a	α=1: ae^{bx} (2)	ln(y)	ln(a)	b	x
3b	α= -1: $ae^{b/x}$	ln(y)	ln(a)	b	$\frac{1}{x}$
4	$k+be^{cx}$ (3)				
5	$k+\frac{b}{c+x}$ (Hyperbel)				
5a	k+b/x (c=0)	y	k	b	x^{-1}
5b	b/(c+x) (k=0)	y^{-1}	c/b	b^{-1}	x
6	$\frac{k(x+a)}{x+b}$ (b>a)				
6a	kx/(x+b) (a=0)	y^{-1}	k^{-1}	b/k	x^{-1}
6b	x/(cx+b) (a=0,k=1)	y^{-1}	c	b	x^{-1}
6c	k(x+a)/x (b=0)	y	k	ak	x^{-1}
7	ln(y)=K-a/(b+x) (4)				
7a	ln(y)=K-a/x (b=0) (5)	ln(y)	K	-a	x^{-1}

(1) Linearisierbar als multiple lineare Regression mit Variablensubstitution $z = x^2$; man erhält dann: a+bx+cz. Entsprechend ist die im Beispiel 9.13 genannte logarithmische Parabel zu linearisieren mit $y^* = \ln(y)$ und $z = x^2$, also $y^* = a+bx+cz$.

(2) Oder in allgemeinerer Form $y = ar^x$, mit $r=e^b$ und damit $b = \ln(r)$.

(3) Oder $y=k+br^x$ mit $r=e^c$ (k: Sättigungsniveau); das ist die "modifizierte Exponentialfunktion". Mit k=0 erhält man den Fall 3a. Im Falle $k \chi 0$ ist aber die Schätzung schwierig. Mit einem Versuchswert für k kann man die lineare Funktion $\ln(y-k) = \ln(b) + cx$ schätzen.

(4) In Übers. 9.5 unter Nr. 9 (Johnson-Funktion); $k=e^K$ ist das Sättigungsniveau.

(5) Dies ist äquivalent mit Nr. 3b, denn $y = k\cdot e^{-a/x}$ ($k=e^K$) hat die gleiche Form wie $y = ae^{b/x}$. Die Funktion steigt S-förmig mit einem Wendepunkt bei x=a/2 auf ein Sätti-

gungsniveau von k. Sie ist im Unterschied zu der logistischen Funktion jedoch nicht punktsymmetrisch um x=a/2 weil bei diesem Wert von x die Ordinate y=k-2 und nicht y=k/2 ist.

2. Variablentransformation

Viele v.a. in den Parametern nichtlineare Funktionen lassen sich durch geeignete Transformationen, z.B. durch Logarithmieren in eine lineare Form bringen. Dadurch wird die nicht nur komplizierte, sondern häufig auch nicht eindeutige Lösung eines nichtlinearen Normalgleichungssystems umgangen.

Eine direkte Schätzung der Regressionsfunktion

(8.42) $y = b \cdot x^{\alpha} + u$

mit der Methode der kleinsten Quadrate führt zu den beiden (in den zu schätzenden Parametern α und b) nichtlinearen Normalgleichungen

(1) $\Sigma y x^{\alpha} = b \Sigma x^{\alpha+2}$ und
(2) $2b\alpha \Sigma y x^{\alpha-1} = (\alpha+2) b^2 \Sigma x^{\alpha+1}$.

Oder im Modell

(8.43) $y = e^{-bx} + u$

wäre der eine Parameter b durch die Normalgleichung

$$\Sigma\, x_i \cdot \exp(-2bx_i) = \Sigma\, y_i \cdot x_i \cdot \exp(-bx_i)$$

zu schätzen.

Offensichtlich kann die wegen ihrer konstanten Elastizität in der Ökonomie (als Modell der Kosten- oder Nachfragefunktion) sehr beliebte Potenzfunktion $y_i = b \cdot x_i^{\alpha}$ (gem Gl. 8.42) als nichtlineare Funktion durch Logarithmieren linearisiert werden, denn aus

(8.42a) $y_i = b \cdot x_i^{\alpha}$ folgt

(8.42b) $\ln(y_i) = \ln(b) + \alpha \cdot \ln(x_i)$.

Bei Variablentransformationen dieser Art ist jedoch auch die Störgröße in der Regression zu beachten. So ist z.B. Gl. 8.42b nur dann die linearisierte Schätzfunktion, wenn man das Modell

(8.42*) $y_i = b \cdot x_i^{\alpha} \cdot \exp(u_i)$

zugrundelegt, in dem die Störgröße (der Stör**faktor**) e^u multiplikativ auftritt (also umso größere Werte annimmt, je größer y ist). Entsprechendes

gilt bei Potenzfunktionen mit zwei Regressoren x_1 und x_2, etwa der Cobb-Douglas-Funktion $y = bx_1^{\alpha}x_2^{\beta}e^u$. Dagegen ließe sich das Modell

(8.42) $y_i = b \cdot x_i^{\alpha} + u_i$

nicht so einfach durch Logarithmieren linearisieren, weil die Störgröße U hier nicht in einer Form vorkommt, bei der sie nach Transformation als additives Störglied erscheint.

Auch bei Variablensubstitutionen ist die Störgröße zu beachten. Im Beispiel 8.7 wird auf dieses Problem (einer ungerechtfertigten deterministischen Interpretation der Regressionsgleichung, d.h. Vernachlässigung der Störgröße) eingegangen.

3. Intrinsische Nichtlinearität

Nicht linearisierbar durch Variablensubstitution oder -transformation sind Funktionen, die nichtlinear in den Variablen **und** in den Parametern sind. Das gilt z.B. für die in Übers. 9.5 unter Nr. 9 genannte logistische Funktion $y_i = k/[1+\exp(a-bx_i)]$.

Eine Linearisierung wäre z.B. dann möglich, wenn die Sättigungsgrenze k bekannt wäre (oder wenn man hierfür einen brauchbaren "Versuchswert" ansetzen könnte). Man erhält dann die Schätzgleichung $\ln[(k-y_i)/y_i] = a - bx_i$. Die logistische Funktion ist linear in den Variablen x und ln[(k-y)/y] (den sog. "Logits") und in den zu schätzenden Parametern a und b. Ist ein Parameter (wie etwa k) dagegen nicht als bekannt vorauszusetzen, so sind Funktionen dieser Art nicht einfach durch Transformationen (z.B. Logarithmieren) oder Variablensubstitution zu linearisieren.

Def. 8.6: Linearität einer Regression

Eine Regressionsfunktion ist linear (oder linearisiert), wenn die abhängige Variable y (bzw. die transformierte oder substituierte Variable y) im Mittel eine lineare Funktion von z_1, z_2 .... ist, wobei die Größen z_1, z_2 usw. **bekannte** Funktionen der Regressoren x_1, x_2 sind.

Erläuterung zu Def. 8.6:
So ist z.B. die Funktion
(8.43a) $y = e^{-bx+u} = e^{-bx}e^u$ oder äquivalent $\ln(y_i) = -bx_i + u_i$ linear in ln(y) und x. Linearisierbar ist auch die Funktion
(8.43b) $y = \exp(-\frac{b}{x} + u) = e^{-b/x}e^u$ oder äquivalent $\ln(y_i) = -b(1/x_i) + u_i$, weil 1/x eine **bekannte** Funktion von x ist.
Dagegen ist die unter Nr. 8 in Übers. 9.5 genannte Funktion
$y = \exp(k + br^x + u)$ mit $r = e^c$ (Gompertz-Funktion)

nichtlinear, weil r^x in $\ln(y) = k + br^x + u_i$ **keine bekannte** Funktion von x ist solange r nicht bekannt ist.

Beispiel 8.7:

a) Ist die nichtlineare Funktion (vgl. Nr. 6a in Übers. 8.2)

$$(8.44) \quad y_i = \frac{kx_i}{x_i+b}$$

zu linearisieren, indem man x und y durch die reziproken Variablen substituiert, also mit

(1) $\frac{1}{y_i} = \frac{1}{k} + \frac{b}{k}\frac{1}{x_i}$ oder ist

(2) $\frac{x_i}{y_i} = \frac{b}{k} + \frac{1}{k}x_i$ die "richtige" Schätzgleichung?

b) Man schätze die Funktion 8.44 mit den beiden unter a) genannten Schätzgleichungen für folgende Daten:

x	1	2	3	4	5	6	7
y	2,7	4,4	5,3	6,6	6,5	6,3	7,7

c) Kann man von den gleichen Eigenschaften der Störgröße wie bei linearer Regression ausgehen?

Lösung 8.7:

a) Die Gleichungen 1 und 2 sind nur dann äquivalente Umformungen von Gl. 8.44, wenn man die Störgröße nicht beachtet, wenn man also den Fehler begeht, die Regressionsfunktion deterministisch zu interpretieren. Die Parameter 1/k und b/k mit der Methode der kleinsten Quadrate zu schätzen setzt voraus, dass gilt:

(1) $\frac{1}{y_i} = \frac{1}{k} + \frac{b}{k}\frac{1}{x_i} + u_{1i}$

bzw. dass man rücktransformiert von der Gleichung

$$(8.44a) \quad y_i = \frac{kx_i}{x_i+b+w_{1i}}$$

mit $w_{1i} = kx_iu_{1i}$ als Störgröße ausgehen kann. Entsprechend läuft der Ansatz (2) darauf hinaus, dass gilt

(2) $\frac{x_i}{y_i} = \frac{b}{k} + \frac{1}{k} x_i + u_{2i}$ und damit rücktransformiert

(8.44b) $y_i = \frac{kx_i}{x_i+b+w_{2i}}$ mit $w_{2i} = ku_{2i}$ als Störgröße.

In keinem der beiden Fälle ist damit aber direkt das Modell

(8.44c) $y_i = \frac{kx_i}{x_i+b} + u_i$ mit u_i als Störgröße

geschätzt.

Es handelt sich also (und das wird oft übersehen) bei allen drei Gleichungen (8.44a bis 8.44c) um verschiedene Modelle, die nur bei einer deterministischen Interpretation der Regression als äquivalent angesehen werden können.

b) Die Daten stammen aus einem Lehrbuch von G. Ross[4] , der die Gl. 8.44 mit verschiedenen Methoden geschätzt hatte. So ergab z.B. eine direkte Schätzung von Gl. 8.44c mit nichtlinearen Normalgleichungen: k = 9,956 und b = 2,549.

Mit dem Ansatz (1) erhält man die (zurücktransformierten) Parameterschätzer k = 10,32046 und b= 2,79428 bei einer Bestimmtheit von r^2 = 0,98718 und mit dem Ansatz (2) k = 9,94663 und b= 2,57574 bei r^2 = 0,95197.

c) Anhand des Ansatzes 1 (Gl. 8.44a) soll gezeigt werden, dass die üblichen Eigenschaften einer Störgröße nicht gelten für die Abweichung zwischen y und dem inversen Schätzwert für y^{-1} nach der folgenden Schätzgleichung

(1) $y_i^{-1} = k^{-1} + (b/k)x_i^{-1} + u_{1i}$

Schätzt man Gl.(1) so erhält man das Ergebnis 1/k = 0,096895 und b/k = 0,270751 woraus sich die oben angegebenen Werte k und b zurückrechnen lassen. Die folgende Tabelle gibt die mit der Schätzgleichung errechneten Regresswerte $\left(\frac{\hat{1}}{y}\right) = \hat{1/y_i}$ an sowie die sich daraus ergebenden Schätzwerte $\hat{y}$ für y (mit $\hat{y} = 1/(\hat{y}^{-1})$. Für die Störgröße

$u_i = \frac{1}{y_i} - \left(\frac{\hat{1}}{y}\right)$

[4] Ross, Garvin J.S., Nonlinear Estimation, New York usw. (Springer Verlag) 1990.

gelten natürlich die in den Gl. 8.11ff genannten Eigenschaften (so ist etwa $\Sigma u_i = 0$), nicht aber für die Störgröße $w_i = y_i - \hat{y}_i$ (so ist Σw_i nicht 0 sondern 0,0174):

x_i (1)	y_i (2)	y_i^{-1} (3)	$\widehat{1/y_i}$ (4)	u_i (5)	$\hat{y}_i$ (6)	w_i (7)
1	2,7	0,37037	0,36765	0,00272	2,720006	-0,020006
2	4,4	0,22727	0,23227	-0,00500	4,305323	0,094677
3	5,3	0,18868	0,18715	0,00500	5,343439	-0,043439
4	6,6	0,15152	0,16458	-0,01307	6,075973	0,524027
5	6,5	0,15385	0,15105	0,00280	6,620534	-0,120534
6	6,3	0,15873	0,14202	0,01671	7,041256	-0,741256
7	7,7	0,12987	0,13557	-0,00570	7,376062	0,323938

Zwischen den Spalten bestehen folgende Beziehungen: (3) ist der reziproke Wert von (2) und (6) der reziproke Wert von (4). Ferner gilt (5) = (3)-(4) sowie (7)=(2)-(6).

Kapitel 9: Verhältniszahlen, Wachstumsraten und Aggregation

1. Arten von Verhältniszahlen

Absolute Größen, wie Summen oder Durchschnitte (Mittelwerte) usw. sind, isoliert betrachtet, meist wenig aussagefähig. Sie können oft erst im Vergleich mit anderen Größen richtig eingeschätzt werden und sie sind auch abhängig von Umfang und Struktur der Masse, auf die sie sich beziehen.

So mag es z.B. nicht überraschen, dass im Lande A die absolute Anzahl der Geburten größer ist als im Lande B, weil A auch mehr Einwohner hat als B. Vergleicht man A und B aber auf der Basis der Geburtenrate (Lebendgeborene/Wohnbevölkerung), so ist der unterschiedliche **Umfang** der Massen (d.h. der Länder A und B hinsichtlich der Einwohnerzahl) "ausgeschaltet". Noch aussagefähiger für die Darstellung "echter" Unterschiede hinsichtlich der Fruchtbarkeit ist ein Vergleich, der auch die unterschiedliche **Struktur** der beiden Länder hinsichtlich Alter und Geschlecht berücksichtigt. Entsprechend ist auch die absolute Zunahme des Sozialprodukts oft nur von geringem Interesse. Um die Entwicklung zu beurteilen ist es besser mit Wachstumsraten zu rechnen.

Verhältniszahlen dienen dem Vergleich von Massen, der Beschreibung von Strukturen (Verteilungen bzgl. "qualitativer" Merkmale) und der Cha-

rakterisierung der Dynamik einer Entwicklung. Sie sind neben Mittelwerten die am häufigsten auch von Nichtstatistikern benutzten statistischen Kennzahlen und in ihrer Aussage meist leicht verständlich. Sie werden aber auch sehr oft mißverstanden. Die Probleme mit ihnen sind mehr inhaltlicher als formaler Art.

Bei der Analyse nominalskalierter Variablen spielen Verhältniszahlen oft die Rolle, die Mittelwerte bei metrischen Variablen haben. Mit Messzahlen (eine spezielle Art von Maßzahlen) - aber auch mit den gerade bei Ökonomen sehr beliebten Wachstumsraten - werden Daten, die als Zeitreihe (vgl. Def. 11.1) vorliegen, auf sehr einfache Art beschrieben.

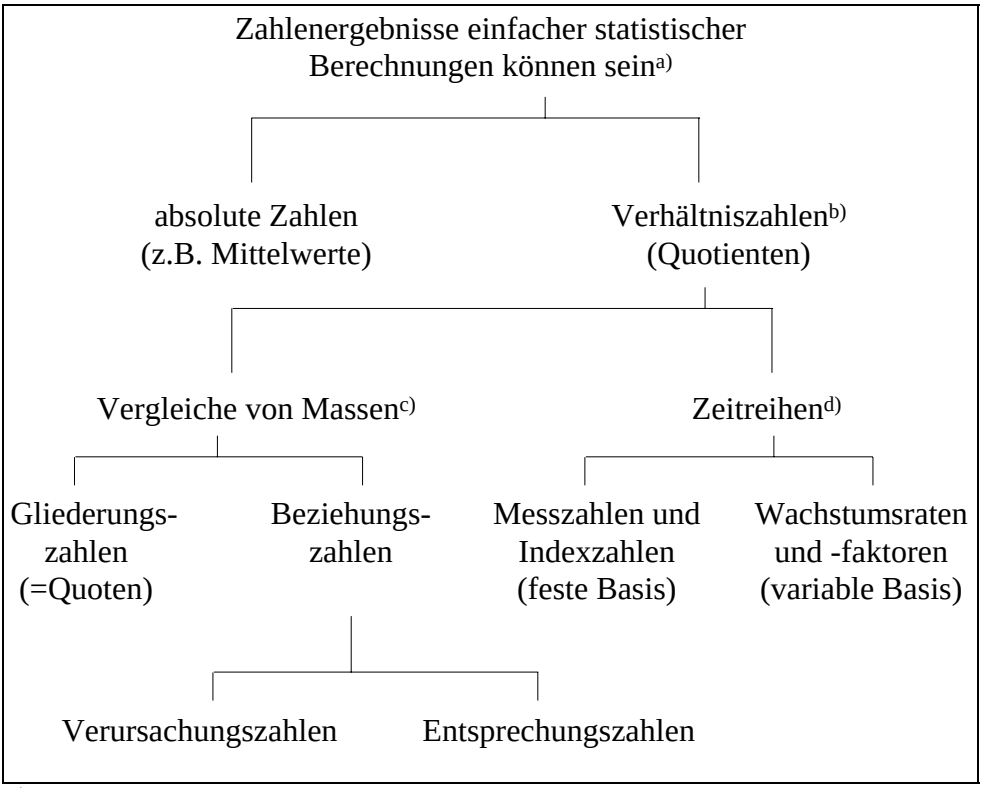
Def. 9.1: Verhältniszahlen

a) Kennzahlen, die als Quotient gebildet sind heißen **Verhältniszahlen**. Man unterscheidet zwischen Gliederungszahlen, Beziehungszahlen und Messzahlen, je nachdem, wie Zähler und Nenner des Quotienten definiert sind. Auch Wachstumsfaktoren und Wachstumsraten sind als Quotienten Verhältniszahlen im weiteren Sinne (vgl. Übers. 9.1).

b) Bei **Gliederungszahlen** G_i ist der Zähler eine Teilmenge des Nenners. Die Gesamtheit (Nennermenge) wird nach einem i.d.R. kategorialen (nominalskalierten) Merkmal in m Teilmassen zerlegt (vgl. Kap. 2 zum Begriff der Zerlegung). Mit dem Umfang n_i der i-ten Teilgesamtheit und $n = \Sigma n_i$, der Gesamtheit bzw. den Merkmalssummen S_i und $S = \Sigma S_i$ist eine Gliederungszahl wie folgt definiert

$$(9.1) \quad G_i = \frac{n_i}{n} \quad \text{oder} \quad G_i = \frac{S_i}{S} \qquad (i = 1,2,...,n).$$

Übersicht 9.1: Arten von Verhältniszahlen



a) Zahlenergebnisse statistischer Berechnungen können auch Schätzwerte für die Parameter eines Modells sein, z.B. Regressionskoeffizienten.
b) Die englischen Begriffe sind ratios (Verhältniszahlen), rates (Beziehungszahlen), proportions (Gliederungszahlen) und relatives (Messzahlen).
c) ohne Zeitbezug (Querschnittsdaten).
d) Darstellung eines zeitlichen Ablaufs.

Eine Gliederungszahl (Quote, Anteilswert) G_i ist "dimensionslos" (genauer: G_i hat keine Maßeinheit). In der Praxis wird G_i mit 100 multipliziert und hat dann die Maßeinheit "Prozent".

c) Bei **Beziehungszahlen** sind Zähler und Nenner Umfänge oder Merkmalssummen von selbständigen Massen, die jedoch in sinnvoller Beziehung zueinander stehen sollten. Die Beziehungszahl ist deshalb auch i.d.R. nicht dimensionslos. Je nachdem, ob die Zählermasse als von der Nennermasse "verursacht" gelten kann oder nicht unterscheidet man zwischen Verursachungszahlen und Entsprechungszahlen.

Übersicht 9.2: Beispiele für Gliederungs- und Beziehungszahlen

Typ: E = Entsprechungszahl
G = Gliederungszahl
V = Verursachungszahl

Name	Typ	Zähler	Nenner
Erwerbsquote	G	Erwerbspersonen	Wohnbevölkerung
Arbeitslosenquote	G	Arbeitslose	Erwerbspersonen
Produktivität[a)b)]	V	output	input
Geburtenrate	V	Lebendgeborene	Wohnbevölkerung
Bevölkerungsdichte	E	Wohnbevölkerung	Fläche
Arealitätsziffer	E	Fläche	Wohnbevölkerung
Arztdichte	E	Anzahl der Ärzte	Wohnbevölkerung
Rentabilität	V	Gewinn	Kapital
Umschlagshäufigkeit[c)]	V	Umsatz	Bestand
durchschn. Steueraufk.	V	Steueraufkommen	Wohnbevölkerung
Hektarertrag[b)]	V	Ernteertrag	landw.Anbaufläche
Belastungsquote[d)]	E	B(x<20,x>60)	B(20≤x≤60)

a) Produktivität allgemein. Bei der Arbeitsproduktivität ist z.B. der Output zu dividieren durch den Arbeitseinsatz also speziell dem Input des Faktors Arbeit.

b) Von einigen Autoren auch als Entsprechungszahlen bezeichnet.

c) z.B. Umschlagshäufigkeit (-geschwindigkeit) eines Lagers: Lagerbewegungen/ Lagerbestand.

d) Es bedeuten B(x < 20, x > 60) = Bevölkerung im Alter von unter 20 und über 60 Jahren und B(20 ≤ x ≤ 60) = Bevölkerung im Alter von 20 bis 60 Jahren.

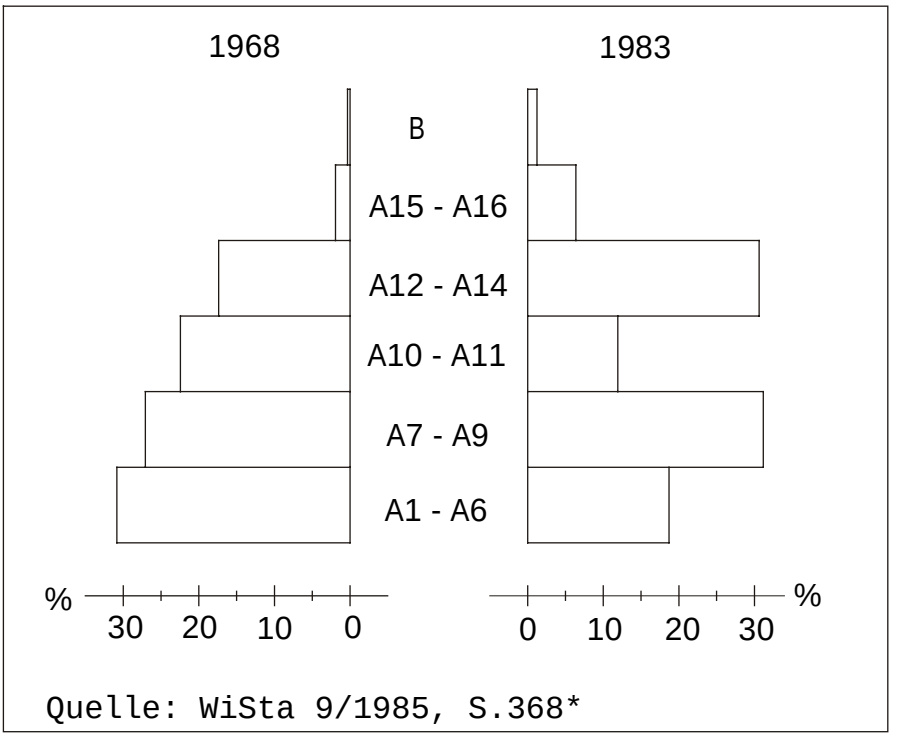
d) Eine **Messzahl** setzt einen (meist aktuellen) Wert y_t ins Verhältnis zum Basiswert y_0, wobei t die "Berichtsperiode" und 0 die (meist zurückliegende) "Basisperiode" (Referenzperiode) ist. Eine dem räumlichen Vergleich dienende Messzahl ist analog definiert. Auch Messzahlen sind wie Gliederungszahlen dimensionslos, weil Kenngrößen (Umfänge, Merkmalsbeträge) gleichartiger Massen ins Verhältnis gesetzt werden. Indexzahlen (Kap. 10) sind zusammengefaßte Messzahlen. Wachstumsraten und -faktoren werden in Def. 9.3 definiert.

Die Unterscheidung zwischen Gliederungs- und Beziehungszahlen (Übers. 9.1) wird auch mit Beispielen erläutert (Übers. 9.2).

Zum Sprachgebrauch:

Die Begriffe "Quote" und "Rate" werden in Politik, Wirtschaft und Journalismus oft völlig willkürlich verwendet. Ein krasser Fall von Verwirrung, die entstehen kann, ist Beispiel 9.2.
So wird z.B. oft von der "Arbeitslosenrate" gesprochen obgleich es sich dabei um eine Quote handelt. Oder die sog. "Belastungsquote" (vgl. Übers. 9.2) ist keine Quote, sondern eine Beziehungszahl.
Quoten sind stets Gliederungszahlen, wobei der Zähler eine echte Teilmenge des Nenners sein muss und "Raten" sind Beziehungszahlen. In diesem Sinne sind z.B. auch die Schuldenquote (Schuldenstand/Sozialprodukt) oder die Scheidungsquote (vgl. Bsp. 9.2) keine echten Quoten. Besonders inflationär wird der Begriff "Quote" bei betriebswirtschaftlichen Kennzahlen oder in der Finanzstatistik verwendet (z.B. Schuldendienstquote = Ausgaben für den Schuldendienst/Staatseinnahmen).

Abb. 9.1: Der "Stellenkegel" der Beamten 1968 und 1983



Quelle: WiSta 9/1985, S.368*

Beispiel / Lösung 9.1:

In den 70er Jahren vollzog sich im ganz besonderen Maße unter den Beamten der Bundesrepublik quasi ein "kollektiver Aufstieg", nicht nur durch eine allgemeine Erhöhung der Besoldung, sondern auch durch eine Veränderung der Struktur zugunsten der höheren Laufbahn- und Besol-

dungsgruppen[1*)], wie dies durch die folgende Tabelle gezeigt wird, für die geeignete Verhältniszahlen zu berechnen sind.

Als Verhältniszahlen eignen sich in diesem Fall besonders Meß- und Gliederungszahlen, die in der nachfolgenden Tabelle wiedergegeben sind.

Beamte und Richter nach Besoldungsgruppe*) am 2.10.1968 und am 30.6.1983 in Tausend**)

	Daten			Messzahl 1983	Quoten in vH	
	Besoldungs-gruppe	1968	1983	1968=100 (1)	1968 (2)	1983 (3)
1	B,R3-10,C4	4,3	19,3	448,8	0,34	1,23
2	A15-A16a)	24,4	100,2	410,7	1,93	6,38
3	A12-A14	220,1	481,1	218,6	17,37	30,61
4	A10-A11	284,5	187,5	65,9	22,45	11,93
5	A7-A9	334,1	489,9	142,8	27,07	31,17
6	A1-A6	390,9	293,7	75,1	30,85	18,70
	Summe	1267,3	1571,7	124,0b)	100c)	100c)

Quelle: Wirtschaft und Statistik Sept. 1985, S. 368*

*) Bekanntlich gliedert sich die Beamtenschaft in vier Laufbahngruppen (einfacher-, mittlerer-, gehobener- und höherer Dienst). Innerhalb jeder Laufbahngruppen gibt es verschiedene Besoldungsgruppen. A und B sind Besoldungsgruppen in der Verwaltung, R für Richter und C für Hochschullehrer. Die Höhe der Besoldung nimmt von der ersten bis zur sechsten Zeile ab.

**) ohne Personen in Ausbildung.

a) sowie R1/R2 und C2/C3 in der Richter- und Hochschullehrerbesoldung.

b) ein gewogenes arithmetisches Mittel der einzelnen Messzahlen, gewogen mit der Struktur gem. Spalte 2.

c) Die Summen können nur im Rahmen von Rundungsfehlern von 100 abweichen.

Die Berechnung von Gliederungszahlen (Spalten 2 und 3) und deren Darstellung in Abb. 9.1 zeigt deutlich die behauptete Strukturveränderung zugunsten der höheren Besoldungsgruppen. Das bestätigt sich auch durch die Messzahlen: man sieht dass bestimmte Gruppen weit mehr (z.B. um 348,8%) zunehmen als der Durchschnitt (+24%) und andere unterdurchschnittlich zunehmen bzw. sogar abnehmen (d.h. die Messzahl ist kleiner als 100%).

Die Abnahme der Anteile in Zeile 4 (Bes. Gr. A10/A11) und die Zunahme in Zeile 5 (Bes. Gr. A7-A9) scheinen der These vom Aufstieg durch "Strukturverbesserung" zu

1 vgl. Fußnoten zur folgenden Tabelle in der Lösung 9.1.

widersprechen. Dies zeigt sich aber aus der hier vorgenommenen Zusammenfassung von Besoldungsgruppen ohne Rücksicht auf Laufbahngruppen. Bei genauerer Betrachtung der Zahlen zeigt sich, dass die Strukturveränderungen weitgehend darin bestanden, dass sich 1985 sehr viele Beamte jeweils auf der für ihre Laufbahngruppe höchste Besoldungsgruppe befanden (insbesondere im gehobenen Dienst), während 1968 die Verhältnisse noch ausgeglichener waren:

einfacher Dienst

Bes. Gr.	1968	1983
A5"S"	33,2	51,5
A4	63,7	80,4
A3	74,5	17,3
A1/A2	30,5	2,4
S	201,9	151,6

mittlerer Dienst

Bes. Gr.	1968	1983
A9"S"	29,5	104,0
A7/A8	260,6	334,7
A5/A6	189,0	142,1
S	479,1	580,8

gehobener Dienst

Bes. Gr.	1968	1983
A13*)	5,0	101,9
A12	68,2	205,6
A11	197,2	99,0
A10	87,3	88,5
S	357,7	495,0

*) und A13"S" bis A15"S" (also Besoldungen für Beamte des gehobenen Dienstes die denen des höheren Dienstes entsprechen)

Die These vom kollektiven Aufstieg belegen auch Messzahlen nach den vier Laufbahn gruppen für 1983 (1968 = 100):

einfacher Dienst	75,1	unterdurchschnittliche
mittlerer Dienst	121,2	Zunahme (<24%)
gehobener Dienst	138,4	überdurchschnittliche
höherer Dienst	166,9	Zunahme (>24%)
insgesamt	124,0	

Das Beispiel zeigt auch, dass es mit einer oder einigen wenigen Verhältniszahlen oft nicht getan ist, um sich ein zutreffendes Bild zu machen und dass dieses Bild auch sehr davon abhängt, wie das Datenmaterial gegliedert ist.

Beispiel 9.2:

Der nebenstehende Text aus der Zeitung DER SPIEGEL (Nr. 37/1976) ist ein schönes Beispiel für einen Verwirr-**Text**, denn es ist sehr schwer, den wahren Sachverhalt aus dem Text zu "rekonstruieren":

a) Welche Größen werden miteinander verglichen?

b) In welchem Sinne wird das Wort "Quote" benutzt?

c) Was heißt "Unfallraten" mit Fahrzeugkilometer "korrelieren"?

d) Nennen Sie Fälle, in denen gleiche Größen (Berechnungen) mit unterschiedlichen Worten beschrieben werden (oder umgekehrt)!

Verwirr-Statistik

Einen neuerlichen Verfall der westdeutschen Verkehrssitten schienen die jüngsten Unfallziffern des Statistischen Bundesamtes in Wiesbaden zu belegen: Verglischen mit dem Juni des Vorjahres, sei im letzten Juni die Uahl der Verkehrstoten um 7,7 Prozent, die Zahl der Verletzten um 5,3 Prozent höher gewesen; im Vergleich der ersten Halbjahre 1975 und 1976 sei die Totenquote um 0,8 Prozent geringer, die Verletzenrate um 3,7 Prozent höher ausgefallen. Prompt fragt die „Frankfurter Allgemeine“: „Fahren wir wieder unvorsichtiger?“ Wohl nicht. Denn unberücksichitgt blieben bei den amtlichen Zahlen die im Vergleichszeitraum beträchtlich gestiegenen Kfz-Zulassungen und Fahrleistungen. Experten des ADAC machten sich die Mühe, das schiefe Bild aus Wiesbaden zurechtzurücken und die Unfallraten mit Fahrzeugkilometern zu korrelieren Im Juni-Vergleich ergab sich dananch, statt eines Anstiegs der Todesfälle um 7,7 Prozentpunkte, eine gegenüber 1975 um 0,4 Punkte niedrigere Zahl der tödlichen Unfälle; die absolut um 5,3 Prozent höhere Verletztenquote leigt bei korrekter Auswertung um 2,5 Prozent unter der des Juni 1975. Ähnliches gilt für den Halbjahresvergleich: 0,8 Prozent weniger Tote als im Vorjahr bedeuten in Wahrheit ein Absinken um vier Prozent; und die Verletzten-Rate lag, statt um 3,7 Prozent, nur um 0,4 Prozent höher.

DER SPIEGEL, Nr. 37/1976

Lösung 9.2:

a) Verglichen werden verschiedene Wachstumsraten und zwar

- im "Jahresvergleich" Juni 76 zu Juni 75 und im "Halbjahresvergleich" Januar bis Juni 76 zu Januar bis Juni 75,

- der Anzahl von Unfällen, Unfallverletzten und Unfalltoten mit und ohne Berücksichtigung der 1975 bzw. 1976 gefahrenen Entfernungen.

b) "Quote" ist ganz offensichtlich die relative Zunahme (also Wachstumsrate), wie z.B. bei "Toten-" oder "Verletztenquote", wobei letztere beim Halbjahresvergleich auch Verletztenrate genannt wird. Auch die

nicht definierte Unfallrate ist wohl so zu verstehen. Im Text sind Quote und Rate ganz nach Belieben und immer falsch verwendet worden, so dass der Text verwirrend und fast völlig unverständlich ist.

c) Gemeint ist offenbar, dass durch die Anzahl der 1975 bzw. 1976 gefahrenen Fahrzeugkilometer (die den Zahlenangaben zufolge um etwa 8% zugenommen hat) dividiert wurde.

d) Gleich gesetzt werden nicht nur Quoten und Raten, sondern auch Prozent, Prozentpunkte und "Punkte", ferner Tote, Verkehrstote, Totenquote, Todesfälle und tödliche Unfälle.

Fazit: Es ist unschön, wenn jemand schulmeisterlich Statistiken kritisiert und dabei selbst die einfachsten Begriffe der Statistik durcheinanderwirft.

2. Gliederungs- und Beziehungszahlen

a) Gliederungszahlen

Eigenschaften von Gliederungszahlen

Es ergibt sich als unmittelbare Folgerung aus Gl. 9.1:

(9.2) $0 \le G_i \le 1$ und

(9.3) $\Sigma G_i = 1$ $(i = 1,2,...,m)$.

Zu den Gliederungszahlen gehört auch die relative Häufigkeit h_i der Merkmalsausprägung x_i bzw. der i-ten Größenklasse der Variablen X (Def. 3.1). Gliederungszahlen beziehen sich aber auch auf kategorial gebildete Teilgesamtheiten, nicht nur auf metrisch skalierte Variablen und sie können auch Quotienten von Merkmalssummen sein, nicht nur von Häufigkeiten (Umfänge von Massen und Teilmassen).

Interpretationsprobleme

Häufig werden aus Gliederungszahlen Schlüsse gezogen, insbesondere über Entwicklungen, die nur in Verbindung mit entsprechenden Beziehungszahlen und deren Veränderung zulässig sind. Hierzu zwei Beispiele:

1. Man schließt auf ein besonders hohes Unfallrisiko auf dem Beifahrersitz eines PKW, weil die Zahl der Unfallopfer, gegliedert nach den vier Sitzen eines PKW, auf denen die verunfallte Person jeweils gesessen hat, im Falle des Beifahrersitzes am größten ist. Der Fehlschluß rührt daher, dass nicht berücksichtigt wird, wie oft die

einzelnen Sitze besetzt sind. So kann das Unfallrisiko auf einem der hinteren Sitze größer sein, wenn man berücksichtigt, dass diese nicht so oft besetzt sind.

2. Ein zunehmender Anteil Krebstoter an der Gesamtzahl der Sterbefälle muss nicht ein höheres Krebsrisiko bedeuten. Mögliche Gründe: die allgemeine Sterblichkeit sinkt, die Krebssterblichkeit aber unterdurchschnittlich oder die Altersstruktur (auch wegen des Rückgangs der Sterblichkeit) hat sich verändert.

Auf ein weiteres Interpretationsproblem, nämlich die Strukturabhängigkeit einer Verhältniszahl, wird an späterer Stelle gesondert hingewiesen. Auch hierfür ein Beispiel:

Eine Klausur gilt als schwerer, weil die Durchfallquote höher ist. Tatsächlich ist aber die Struktur der Klausurteilnehmer anders: es gab z.B. in der "schwierigeren" Klausur mehr Wiederholer, deren Durchfallquote oft größer ist als bei Studenten in ihrem ersten Klausurversuch.

Gliederungszahlen und Wahrscheinlichkeiten

Der Anteil der Knabengeburten an der Gesamtzahl der Geburten (eine Gliederungszahl) ist eine Schätzung für die Wahrscheinlichkeit einer Knabengeburt. Die Wahrscheinlichkeit ist wie eine Gliederungszahl eine Zahl zwischen 0 und 1. Sie bezieht sich aber nicht auf eine empirisch beobachtete und daher notwendig endliche Gesamtheit, sondern auf Ergebnisse eines prinzipiell beliebig (unendlich) oft wiederholbaren Zufallsversuch.

Zieht man in einer Lotterie bei zehn Losen 9 Nieten (Gewinn**quote** 1/10 also, 10%) so ist das natürlich kein Widerspruch zu einer Gewinn**wahrscheinlichkeit** von 20% (ein anderer Teilnehmer hat mehr Glück und erzielt eine Quote von 30%). Auch geeignet gebildete Raten (also Beziehungszahlen, z.B. die Todesrate) stehen im Zusammenhang mit Wahrscheinlichkeiten (z.B. der Sterbewahrscheinlichkeit).

b) Beziehungszahlen

Allgemeine Eigenschaften

1. Dimension:
   Anders als Gliederungszahlen und Messzahlen setzen Beziehungszahlen verschiedenartige Massen ins Verhältnis. Sie sind deshalb auch nicht dimensionslos, sondern haben eine Maßeinheit.

2. Umkehrbarkeit:
   Beziehungszahlen sind grundsätzlich umkehrbar wie das Beispiel der Bevölkerungsdichte und der Arealitätsziffer (Übers. 9.2) zeigt. Einmal

mißt man die Anzahl der Personen je Fläche (Bevölkerungsdichte), im anderen Fall (Arealitätsziffer) die durchschnittliche Fläche um eine Person.
Ein ökonomisches Beispiel für die Umkehrbarkeit ist die

$$\text{Kapitalproduktivität} = \frac{\text{Output (z.B. Inlandsprodukt)}}{\text{Kapitaleinsatz (Anlagevermögen)}}$$

Der reziproke Wert ist der (durchschnittliche) Kapitalkoeffizient. Die Kapitalproduktivität zeigt, wieviel DM Output mit einer eingesetzten DM Kapital zu erwirtschaften ist und der Kapitalkoeffizient zeigt wie groß der Kapitalbedarf ist (wieviel investiert werden muss) um eine DM Output zu erwirtschaften.

3. Zusammenhang mit Mittelwerten:

   Beziehungszahlen sind nichts anderes als Mittelwerte, wenn eine Merkmalssumme (Zähler) zu einer entsprechend abgegrenzten Personengesamtheit (Nenner) ins Verhältnis gesetzt wird. Ein Beispiel ist das durchschnittliche Steueraufkommen (Übers. 9.2) oder die Beziehungszahl Bierkonsum/Einwohner, der durchschnittliche Bierverbrauch (der dann allerdings nicht aufgrund der individuellen Verbrauchsmengen errechnet ist).

Alle Verhältniszahlen sind ferner Mittelwerte durch Aggregation (vgl. Kap. 9, Abschn. 5), d.h. sie sind in dem Sinne Mittelwerte dass eine auf die Gesamtmasse bezogene Verhältniszahl ein Mittel der entsprechenden Verhältniszahlen der Teilmassen ist (so ist z.B. die rohe Todesrate ein Mittel der altersspezifischen Todesraten).

Verursachungszahlen

1. Der Begriff "Verursachungszahl" ist problematisch weil er eine monokausale Verursachung suggeriert und der Quotient ist meist nicht eindeutig, denn

   a) es ist oft nicht klar welche Nennergröße zu nehmen ist und
   b) die erwähnte Umkehrbarkeit ist zu beachten.

   zu a)

   Ein bekanntes Problem ist die zutreffende Beschreibung eines Unfallrisikos mit PKWs (oder entsprechend die Messung des Flugrisikos) durch eine geeignete Maßzahl. Soll man die Anzahl der Unfälle mit PKWs dividieren durch

   - die Einwohnerzahl,

- den Bestand an PKWs,
- die Anzahl der gefahrenen Kilometer oder
- das Produkt dieser Größe mit der Anzahl der beförderten Personen (also die "Personenkilometer")?

zu b)

Man mag die Kapitalproduktivität als Verursachungszahl bezeichnen, wohl kaum aber den reziproken Kapitalkoeffizienten. Im Zähler der Kapitalproduktivität steht übrigens der Output *aller* Produktionsfaktoren, nicht nur der des Kapitals, weshalb diese Beziehungszahl auch nicht kausal interpretiert werden darf.

2. Verursachungszahlen sind meistens Verhältnisse von Bestands- und Bewegungsmassen, wie z.B. die

$$\text{Fruchtbarkeitsrate}^{a]} = \frac{\text{Lebendgeborene eines Jahres}}{\text{gebärfähige Frauen}^{b]} \text{ in der Mitte}^{c]} \text{ dieses Jahres}}$$

a] früher war die Bezeichnung Fruchtbarkeits"ziffer" üblicher, entsprechend sprach man früher von der Sterbeziffer oder der Geburtenziffer (heute: Todesrate, Geburtenrate [besser: Geborenenrate]).
b] Frauen im Alter zwischen 15 und 45 Jahren.
c] oder im Durchschnitt des Jahres.

Begriff der "Rate"

Eine "Rate" bezieht die Häufigkeit eines Ereignisses in einem Zeitintervall auf die durchschnittliche Anzahl der Einheiten, die (zu Beginn des Intervalls) dem Risiko des Ereignisses ausgesetzt waren.
Sie steht in einem Zusammenhang mit

- der Wachstumsrate und
- der Wahrscheinlichkeit.

Die Differenz zwischen der Geburten- und der Todesrate ist die "natürliche" (durch Geburt und Tod) Wachstumsrate der Bevölkerung.
Die altersspezifische Sterberate (als Verursachungszahl) z.B. der 60jährigen Männer könnte eine Schätzung der Sterbewahrscheinlichkeit sein, wenn die Sterbefälle 60-jähriger Männer des Jahres t bezogen wären auf den Bestand an Lebenden *zu Beginn* des Jahres t (üblich ist aber: in der *Mitte*, oder im *Durchschnitt* des Jahres t) .
Gemeint ist dabei die *einjährige* Sterbewahrscheinlichkeit 60-jähriger Männer, d.h. die (bedingte) Wahrscheinlichkeit das Alter 61 nicht mehr zu erreichen, *wenn* man (Bedingung!) das Alter von 60 erreicht hat. Die *unbedingte* Sterbewahrscheinlichkeit, irgendwann einmal zu sterben, egal in welchem Alter, ist natürlich immer 1.

Beispiel 9.3:

Für die Bundesrepublik ("alte" Bundesländer) galten 1990 etwa (gerundet) die folgenden Zahlen:

Ehescheidungen	140 Tausend	Wohnbevölkerung	60 Millionen
Eheschließungen	400 Tausend	Bestand an Ehen	15 Millionen
Lebendgeborene	600 Tausend		

Man berechne die folgenden Beziehungszahlen

1. Lebendgeborene je
   a) 1000 Einwohner
   b) 100 bestehende Ehen
2. Ehescheidungen je
   a) 1000 bestehende Ehen
   b) 100 geschlossene Ehen

und interpretiere die Ergebnisse!

Lösung 9.3:

Die vier Beziehungszahlen sind leicht zu berechnen:

1a) 10 also 1% (das ist die sog. rohe [allgemeine] Geburtenrate),

1b) 4 also 4%.

2a) 140/15 = 9,333 also 0,93%. Diese Größe **und** 2b) sind auch bekannt als Scheidungs-"quote",

2b) (140/400)100 = 35 (also 35%).

Interpretation (verbreitete Fehlinterpretationen):
Bei der Interpretation der Geburtenrate [Geburtenziffer] und der Scheidungsquote (die keine Quote, sondern eine Beziehungszahl ist) wird oft vergessen, dass es sich um die Geburten bzw. Scheidungen **eines Jahres** handelt. Die Interpretation von 1b): "von 100 Ehen **haben** nur vier ein Kind" widerspricht ganz offensichtlich der Erfahrung. Richtig wäre: nur vier (von 100) Ehepaare haben **in diesem Jahr** ein Kind **bekommen**. Auf der gleichen Linie liegt die Fehlinterpretation von 2a): noch nicht einmal 1% der Ehen endet vor dem Scheidungsrichter (diese Interpretation wäre nur zulässig, wenn die durchschnittliche Verweildauer in der Ehe ein Jahr wäre). Man kann auch nicht sagen [im Sinne von 2b)]: etwa jede 3-te Ehe (35%) wird geschieden, denn es sind die Scheidungen eines Jahres (von Ehen, die in einem Zeitraum von vielen früheren Jahren geschlossen wurden), die verglichen werden mit den Eheschließungen dieses gleichen Jahres. Eine solche Interpretation wäre aber zulässig, wenn der Bestand an Ehen dem Modell der stationären Bevölkerung (=Sterbetafelbevölkerung) entspräche und diese Scheidungsquote (gem. 2b) jedes Jahr 100 % (statt 35%) betragen würde (vgl. Kap. 12).

c) Partielle Assoziation

Gegeben seien drei dichotome Merkmale X,Y und Z. Unter der partiellen Assoziation erster Ordnung versteht man die Assoziation zwischen zwei Merkmalen (etwa X und Y) in einer der beiden Teilgesamtheiten bezüglich des dritten Merkmals (also wenn $Z = z_1 = 1$ und wenn $Z = z_o = 0$ ist) . Von Interesse sind die Beziehungen zwischen diesen beiden partiellen Assoziationen und der Assoziation von X mit Y in der Gesamtheit (die man auch partielle Assoziation nullter Ordnung nennen könnte) und deren Interpretation. Mit dem Kreuzprodukt (Gl. 7.47) als Assoziationsmaß erhält man:

$$(9.4) \quad |xy| = p_z \cdot |xy;z_1| + q_z \cdot |xy;z_0| + \frac{|xz| \cdot |yz|}{p_z q_z}$$

Hierin sind $|xy;z_1|$ und $|xy;z_0|$ die beiden partiellen Assoziationen, p_z ist der Anteil der Merkmalsträger mit $Z = z_1 = 1$ und $q_z = 1 - p_z$ der Anteil der Merkmalsträger mit $Z = z_0 = 0$.
Gl. 9.4 hat folgende Interpretation:
Die Assoziation zwischen X und Y in der (inhomogenen) Gesamtheit ist eine Summe der "**internen Assoziation**" (AI)

$$(9.4a) \quad AI = p_z \cdot |xy;z_1| + q_z \cdot |xy;z_0|$$

und der **marginalen Assoziation** AM, d.h. der auf die eindimensionalen Randverteilungen bezogenen Beziehung zwischen X und Y mit Z (der Kontrollvariablen):

$$(9.4b) \quad AM = \frac{|xz| \cdot |yz|}{p_z q_z} = \frac{|xz| \cdot |yz|}{s^2_z},$$

so dass man erhält

$$(9.4c) \quad |xy| = AI + AM.$$

Man beachte, dass die Kreuzprodukte |xy|, |xz| und |yz| jeweils Kovarianzen darstellen und dass die Varianzen der dichotomen Merkmale $s_x^2 = p_x q_x$ und s_y^2 und s_z^2 entsprechend definiert sind.
Für die Interpretation von besonderem Interesse sind die folgenden beiden Spezialfälle:

1) Scheinkorrelation AI = 0 (die Assoziation |xy| ist ausschließlich auf die Assoziation von X und Y mit Z zurückzuführen) und
2) AM = 0 (die Assoziation |xy| ist das gewogene Mittel der partiellen Assoziationen), weil X und Y nicht mit Z korreliert sind.

Im ersten Fall gilt für die Vierfelderkorrelationen (Phi-Koeffizienten gem. Gl. 7.46), die hier mit r bezeichnet werden sollen,

(9.5) $r_{xy} = r_{xz}\, r_{yz}$.

Das ist die Analogie zum Verschwinden der partiellen Korrelation bei metrisch skalierten Merkmalen ($r_{xy.z} = 0$, was kennzeichnend ist für Scheinkorrelation zwischen X und Y).

d) Allgemeine Interpretationsprobleme von Gliederungs- und Beziehungszahlen

In diesem Abschnitt sollen die folgenden Probleme im Umgang mit Gliederungs- und Beziehungszahlen betrachtet werden:

1. die Konstruktionsprinzipien von Beziehungszahlen:
   - Bildung homogener Massen
   - Ausscheiden unbeteiligter Massen
2. das Problem der Strukturabhängigkeit, d.h.
   - der Unterschied zwischen Verhältniszahlen (insbesondere Gliederungs- und Beziehungszahlen) kann allein strukturbedingt (und nicht "echt") sein,
   - das Simpson-Paradoxon
3. die Scheinkorrelation.

Die Probleme Nr. 2 und 3 sind Folge der Abhängigkeit der untersuchten Beziehung von einer "dritten" Variablen.

1. Konstruktionsprinzipien

Der oberste Grundsatz der Konstruktion von Beziehungszahlen (oder allgemeiner Maßzahlen) ist:

Äquivalenten (in den relevanten Aspekten gleichen) Sachverhalten ist ein gleicher Zahlenwert der Maßzahl zuzuordnen, nicht-äquivalenten Sachverhalten dagegen ein verschiedener Zahlenwert.

Ist für den Ernteertrag allein die Größe der Anbaufläche relevant, weil z.B. der Ertrag der Fläche proportional ist, so ist der Hektarertrag (Übers. 9.2) eine sinnvolle Beziehungszahl. Andere Einflußfaktoren, wie z.B. Klima, Bodenart, Bearbeitungstechnik [z.B. Düngung], oder Art [Rechtsform usw.] des landwirtschaftlichen Betriebs sollten für die Größe des Ertrags irrelevant sein.

Wenn das nicht der Fall ist, dann sollten (am Beispiel des Hektarertrags) nur Erträge von Flächen bei *gleichem* Klima, *gleicher* Bodenart usw. in die Berechnung der entsprechenden Verhältniszahlen einbezogen werden.

Häufig sind deshalb Zähler und Nenner einer Verhältniszahl zu spezifizieren, um die verglichenen Massen homogener zu machen, denn:
Nur homogene Massen repräsentieren einen einheitlichen Ursachenkomplex; sind die Massen inhomogen, so nivellieren sich charakteristische Unterschiede, die herauszustellen gerade die Aufgabe einer Beziehungszahl sein kann. Die Homogenisierung geschieht durch:

- Ausscheiden unbeteiligter Teile einer Masse:
  Ein Land A kann einen geringeren Bierverbrauch "pro Kopf" der Bevölkerung haben als das Land B, weil die Altersstruktur anders ist. Es kann sinnvoller sein, Säuglinge und Kinder aus der Nennermasse auszuscheiden und nicht die Gesamtbevölkerung, sondern die Bevölkerung ab eines bestimmten Alters als Bezugsgröße zu wählen.
- Bildung spezifischer Verhältniszahlen:
  Es ist z.B. in der Bevölkerungsstatistik üblich, viele Verhältniszahlen nach Alter und Geschlecht zu differenzieren, etwa die Erwerbsquoten: der Verlauf der altersspezifischen Erwerbsquoten der Frauen zeigt u.a. deutlich den Einfluß der Familienbildung (in einem Altersintervall von ca. 25 bis 40 Jahre) auf die Erwerbsbeteiligung von Frauen.

2. Strukturabhängigkeit und Simpson-Paradoxon

Wie an späterer Stelle (Abschn. 5d) gezeigt wird, ist eine Verhältniszahl stets ein gewogenes Mittel von speziellen ("spezifischen", d.h. für Teilgesamtheiten gebildeten) Verhältniszahlen des gleichen Typs. Folgerungen:

- Deshalb können sich zwei Verhältniszahlen allein durch die Gewichtung unterscheiden (***Struktureffekt***; Beispiel 9.4);
- ***Standardisierung*** (vgl. Abschn. 5d) ist ein Verfahren, um beim Vergleich von Verhältniszahlen den Struktureffekt auszuschalten;
- Aufgrund unterschiedlicher Gewichtung (Struktur) kann im Extremfall das ***Simpson Paradoxon*** auftreten (Beispiel 9.5).

Def. 9.2: Simpson Paradoxon

Die Tatsache, dass ein Mittelwert oder eine Verhältniszahl (z.B. eine Quote, ein Anteilswert) für eine Gesamtheit A größer sein kann, als für eine andere Gesamtheit B, obgleich diese Größe (Mittelwert oder Verhältniszahl) in allen Teilgesamtheiten von A kleiner ist als in denen von B, ist bekannt als "Simpson-Paradoxon" (nach Th. Simpson 1710 - 1761).

Beispiel 9.4:

(Strukturabhängigkeit): Die Sterbeziffer [Todesrate] (= Zahl der Gestorbenen je 1.000 Lebende) von Geistlichen ist viel höher (0,55%) als die der Bergarbeiter (0,15%) [fiktive Zahlen]:

Alters-klasse	Geistliche		Bergarbeiter	
	Lebende	Gestorbene	Lebende	Gestorbene
unter 50	100	10	900	90
über 50	900	540	100	60
insgesamt	1000	550	1000	150

Kann man aus den Angaben schließen, dass der Beruf des untertage arbeitenden Bergmanns "gesünder" ist, als der des Geistlichen?

Lösung 9.4:

Man kann natürlich nicht in der Weise schließen, wie dies die Fragestellung nahelegt. Der Grund ist die unterschiedliche Altersstruktur der Geistlichen und der Bergarbeiter. Der Anteil der unter 50-jährigen ist bei den Geistlichen 10% und bei den Bergarbeitern 90% (entsprechend sind die Anteile der über 50-jährigen 90% und 10%). Die altersspezifischen Todesraten (Sterbeziffern) sind für beide Berufe in beiden Altersklassen gleich, nämlich bei

den unter 50-jährigen : 10%
den über 50-jährigen : 60%.

Die rohen Todesraten sind (mit den Anteilen 10% und 90% für die Altersstruktur) gewogene arithmetische Mittelwerte:
für die Geistlichen: $0{,}55 = 0{,}1 \cdot 0{,}1 + 0{,}9 \cdot 0{,}6$
für die Bergarbeiter: $0{,}15 = 0{,}9 \cdot 0{,}1 + 0{,}1 \cdot 0{,}6$.

Beispiel 9.5:

(Simpson-Paradoxon): Das Beispiel 9.4 wird wie folgt modifiziert:

Alters-klasse	Geistliche			Bergarbeiter		
	Lebende	Gestorbene	Rate[a)]	Lebende	Gestorbene	Rate[a)]
< 50	100	10	0,10	600	80	0,13
≥ 50	900	540	0,60	400	280	0,70
S [b)]	1000	550	0,55	1000	360	0,36

a) Todesraten.

b) bzw. Durchschnitt.

Die altersspezifischen Todesraten der Bergarbeiter sind in allen (beiden) Altersklassen höher, als die der Geistlichen und trotzdem ist die rohe (gesamte) Todesrate der Bergarbeiter niedriger, als die der Geistlichen (Simpson-Paradoxon). Woran liegt das?

Lösung 9.5

Der Struktureffekt (Bergarbeiter sind jünger), der dahingehend wirkt, dass die rohe Todesrate der Bergarbeiter kleiner sein müßte, als die der Priester, wirkt dem echten Unterschied in der Sterblichkeit (Todesraten der Bergarbeiter in allen Altersklassen größer) entgegen.

3. Scheinkorrelation

Das Konzept der Scheinkorrelation ist in Def. 7.9 definiert (Korrelation zwischen zwei Variablen nur weil diese gemeinsam abhängig sind von einer dritten Variablen) und wird im folgenden Beispiel demonstriert.

Beispiel 9.6:

Einer (fiktiven) Statistik zufolge ergaben sich die folgenden Daten über die Unfallhäufigkeit von Männern und Frauen:

Autounfall	Männer	Frauen	Summe
wenigstens einmal	3.122	2.255	5.377
nie	3.958	4.695	8.653
Summe	7.080	6.950	14.030

Beispiel entnommen aus dem Lehrbuch von H. Zeisel, Say it with Figures (deutsche Übersetzung: Die Sprache der Zahlen, Köln, Berlin 1970, S. 126).

Kann man aufgrund dieser Zahlen schließen, dass Frauen bessere (sicherere) Autofahrer sind als Männer?

Lösung 9.6:

Eine Schlussfolgerung in der genannten Weise ist sehr verbreitet. Sie wird auch nahegelegt durch die Berechnung von Quoten. Betrachtet man die "Quote der mindestens einmal Verunfallten", so ist sie bei den Männern 3122/7080 = 0,441 (also 44,1%) und bei den Frauen 2255/6950 = 0,324, also nur 32,4% (für die Gesamtheit ist die Quote natürlich ein gewogenes Mittel der beiden Quoten; sie beträgt 5377/14030 = 0,383), so dass Frauen scheinbar die besseren Autofahrer sind als Männer.

Bei genauerem Hinsehen kann sich dies aber als Scheinkorrelation erweisen. H. Zeisel untersucht die "Fahrhäufigkeit" (>10.000 oder ≤ 10.000 Meilen im Jahr) als dritte, die Scheinkorrelation erzeugende Variable mit folgenden Zahlen:

	häufiges Fahren			seltenes Fahren		
Unfall	Männer	Frauen	Σ	Männer	Frauen	Σ
wen. Einmal	2.605	996	3.601	517	1.259	1.776
nie	2.405	919	3.324	1.553	3.776	5.329
Summe	5.010	1.915	6.925	2.070	5.035	7.105

Berechnet man jetzt die entsprechenden Quoten, so sind sie in den Teilgesamtheiten "häufiges-" und "seltenes Fahren" für Männer und Frauen jeweils gleich, was ein Zeichen dafür ist, dass die Korrelation zwischen Geschlecht und Fahrtüchtigkeit nur eine Scheinkorrelation ist.

Man kann den Nachweis der Scheinkorrelation auch mit Assoziationsmaßen führen. Der Zusammenhang von Gl. 9.4 stellt sich hier mit X: Geschlecht, Y: Unfallhäufigkeit und Z: Fahrhäufigkeit wie folgt dar:

(9.4) $|xy| = p_z \cdot |xy;z_1| + q_z \cdot |xy;z_0| + |xz| \cdot |yz| / p_z q_z$

$|xy| = 0{,}0291225$, $f = r_{xy} = 0{,}1198$ (Phi-Koeffizient)

$p_z = 6925/14030 = 0{,}4936$ und entsprechend $q_z = 0{,}5064$

$|xy;z_1| = -\,0{,}00002888$ ($f = -0{,}000129$ also » 0)

und $|xy;z_0| = -0{,}00006012$ ($f = -0{,}0003056$ also » 0).

Das Verschwinden der beiden partiellen Assoziationen ist ein Zeichen für Scheinkorrelation. Für die marginalen Beziehungen erhält man:
$|xz| = 0{,}108013$ ($f = r_{xy} = -0{,}4321$) und $|yz| = 0{,}067498$ ($r_{yz} = 0{,}2777$), so dass $r_{xy} \approx r_{xz}\, r_{yz}$. Ferner ist $AM = 0{,}029167 \approx |xy| = 0{,}0201225$ und $AI \approx 0$.

3. Messzahlen

Die Messzahl m_{0t} (d.h. zur Basis 0, Berichtszeit t) einer Variablen Y ist nach Def. 9.1 die Größe:

$$(9.6) \quad m_{0t} = \frac{y_t}{y_0},$$

bei der diskreten Zeitvariable t = 0,1,2,...,T bzw. die mit 100 multiplizierten Größen

$$(9.6a) \quad m_{0t}^{*} = 100 m_{0t} = 100 \frac{y_t}{y_0}.$$

Die Größe t kann, muss aber nicht "Zeit" bedeuten. Messzahlen können z.B. auch dem räumlichen Vergleich dienen, wenn 0 das Basisland und t das Vergleichsland ist.

Eigenschaften von Messzahlen:

1. Messzahlen haben die in Übers. 9.3 definierten und leicht zu verifizierenden Eigenschaften, die nicht unbedingt auch für Indexzahlen (Kapitel 10) gelten müssen.

2. Aus Gl. 9.6 folgt, dass die Bildung von Messzahlen eine Lineartransformation darstellt: y_t wird mit der Konstanten $(y_0)^{-1}$ multipliziert, d.h. die Entwicklung der Zeitreihe y_t (also der Folge $y_1, y_2,..., y_t$ wird in Einheiten [bzw. bei m_{0t}^{*} in Prozent] des Basiswerts y_0 dargestellt.

3. Das wohl wichtigste methodische Problem ist die ***Wahl der Basisperiode***, z.B. des Basisjahres: Die Regel ist, ein "Normaljahr" (ohne auffallende Besonderheiten) zu wählen damit die nachfolgende Entwicklung nicht unter- oder überzeichnet wird. Man kann sich an Beispielen leicht klar machen, welchen Effekt die Wahl eines extremen Jahres als Basisjahr hätte.

Übersicht 9.3: Eigenschaften von Messzahlen

Eigenschaft	Inhalt der Forderung
Identität	$m_{00} = m_{tt} = 1$ ($m^*_{00} = m^*_{tt} = 100$) Identität von Basis und Berichtsperiode
Dimensionalität	$m_{(a)0t} = ay_t/ay_0 = m_{0t} = y_t/y_0$ Unabhängigkeit von der Maßeinheit der Messwerte
Zeitumkehrbarkeit (Reversibiltät)	$m_{t0} = m_{0t}^{-1}$ Vertauschung von Basis- und Berichtsperiode ($m_{t0}m_{0t} = 1$)
Zirkularität (Transitivität, Verkettbarkeit)	für je drei Perioden 0,s und t gilt $m_{0t} = m_{0s}m_{st}$ (= Verkettung; Folgerung: $m_{st} = m_{0t}/m_{0s}$ [= Umbasierung])
Faktorumkehrprobe	ist für alle Perioden die Größe W das Produkt aus P und Q so gilt für die entsprechenden Messzahlen $m^W_{0t} = m^P_{0t} \cdot m^Q_{0t}$ *)

*) z.B. eine Wertmesszahl m^W_{0t} ist das Produkt aus Preis- und Mengenmesszahl.

4. Zweck der Messzahlenbildung ist die Vergleichbarmachung von Wachstumsvorgängen, die sich auf einem unterschiedlichen absoluten Niveau abspielen, z.B. der Staatsverbrauch und der meist fast dreimal so große Private Verbrauch in Beispiel 9.7.

5. Die Messzahlen $m^{(x)}_{0t} = x_t/x_0$ und $m^{(y)}_{0t} = y_t/y_0$ der Zeitreihen X_t und Y_t haben für alle Perioden t = 1,2,...,T die gleichen Wachstumsraten wie die Zeitreihen X_t und Y_t selber.

6. Die Messzahl m^z_{0t} der ***Beziehungszahl*** $z_t = y_t/x_t$ ist das Verhältnis (die Beziehungszahl) der Messzahlen m^y_{0t} und m_{0t},^{x} also $m^z_{0t} = m^y_{0t}/m^x_{0t}$. Praktische Bedeutung: Berechnung der Messzahlen der Produktivität aufgrund der Messzahlen von Output und Input ohne Rückgriff auf die absoluten Größen von Output und Input.
   Ist Z ein ***Produkt***, etwa $z_t = y_t \cdot x_t$ so gilt (Faktorumkehrprobe!) $m^z_{0t} = m^x_{0t} \cdot m^y_{0t}$.

Beispiel / Lösung 9.7:

Gegeben sind die Werte (zu jeweiligen Preisen) für den Privaten Verbrauch (PV) und den Staatsverbrauch (SV) in Mrd. DM in der Bundesrepublik Deutschland, für die Messzahlen zur Basis 1980 zu bilden sind (bei

der mit 100 multiplizierten Messzahl m_{0t},* schreibt man üblicherweise die an sich unsinnige "Gleichung" 1980 = 100):

	Daten		Messzahlen m_{0t} zur Basis 1980		m_{0t},* (für SV)
Jahr	PV	SV	PV	SV	1980=100
1980	840,78	297,79	1,0	1,0	100
1981	887,85	318,16	1,0560	1,0684	106,84
1982	918,05	326,19	1,0919	1,0954	109,5
1983	964,16	336,21	1,1467	1,1290	112,90
1984	1003,57	350,23	1,1936	1,1761	117,61
1985	1038,34	365,66	1,2350	1,2279	122,79
1986	1068,61	382,72	1,2710	1,2852	128,52
1987	1112,68	396,97	1,3234	1,3331	133,31
1988	1156,81	411,46	1,3759	1,3817	138,17

Quelle: Jahresgutachten des Sachverständigenrats 1989/90, Tab 25*

Man erkennt unschwer, dass der Staatsverbrauch von 1980 bis 1988 nur geringfügig stärker angestiegen ist (nämlich um 38,2%, Prozent von 1980), als der Private Verbrauch (der um 37,6% gestiegen ist). Beim Vergleich der Ursprungszahlen 411,46 und 1156,81 für 1988 mit 297,79 und 840,78 für 1980 wäre das auf einen Blick kaum möglich zu erkennen. Außerdem ist leicht zu sehen (vgl. Abb. 9.2), dass es auch Perioden gibt, in denen SV langsamer ansteigt, als PV, auch wenn über die ganze Zeitspanne (von 1980 bis 1988) SV stärker angestiegen ist, als PV.

Beispiel 9.8:

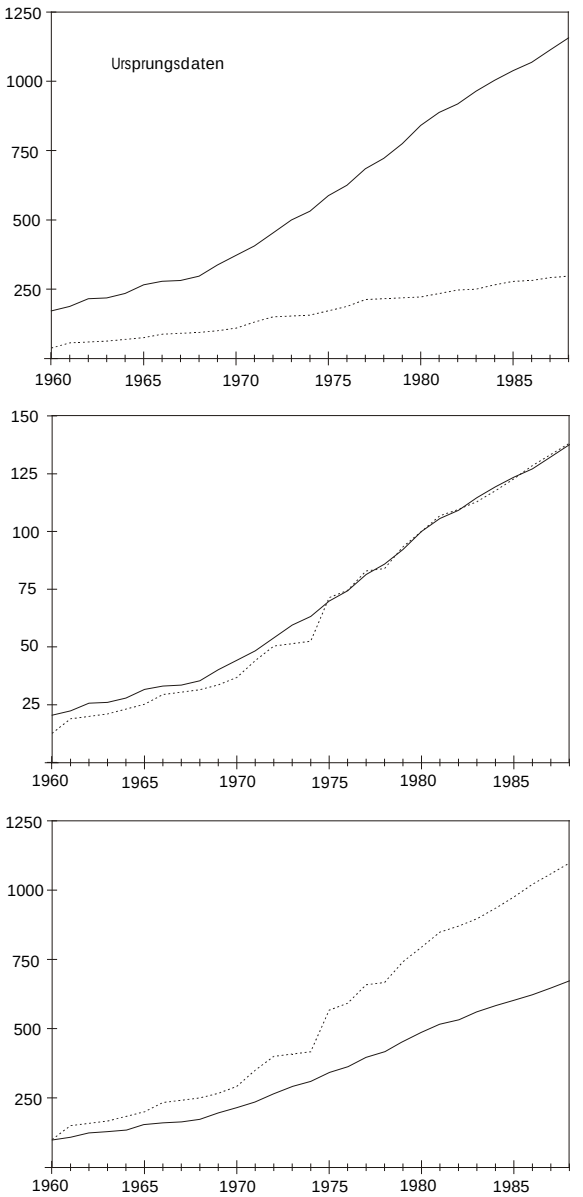
Die Messzahlen des Beispiels 9.7 zur Basis 1980 sind umzubasieren auf das "neue" Basisjahr 1985!

Lösung 9.8:

Eine Neuberechnung sämtlicher Messzahlen aus den Ursprungswerten ist nicht erforderlich. Zur Berechnung von Messzahlen $m_{85,t}$ aus den Messzahlen $m_{80,t}$ reicht es aus, alle Messzahlen $m_{80,t}$ durch $m_{80,85}$ zu dividieren, also bei PV alle Messzahlen durch 1,235 und bei SV durch 1,2279.
Für die mit 100 multiplizierten Messzahlen m^* gilt entsprechend die Formel: $m^*_{85,t} = (m^*_{80,t}/m^*_{80,85})\cdot 100$ für alle Jahre t = 80,81,...,88.

Abb. 9.2: Messzahlen und Ursprungsdaten von Privatem Verbrauch und Staatsverbrauch 1960 bis 1988

Privater Verbrauch: durchgezogene Linie, Staatsverbrauch: gestrichelte Linie
a) Ursprungsdaten b) Messzahlen 1980=100 c) Messzahlen 1960=100



Umbasierung und Verkettung:

Umbasierung (Basiswechsel) ist die Umkehrung der Verkettung. Mit den Perioden 0,s und t (etwa 1980, 1985 und 1990) bedeutet

Umbasierung: die bisherige Messzahl m_{0t} ist auf die neue Basis s umzustellen (um sie z.B. mit anderen Messzahlen der Basis s vergleichen zu können). Es ist also die Messzahl m_{0t} zu bestimmen.
Verkettung: zwei Messzahlenreihen zur Basis 0 und s sind zu einer langen Reihe zusammenzufügen (die Reihe mit der Basis 0 ist mindestens bis s geführt worden).

Lösung:

a) Messzahlen m_{0t},m_{st}:

Umbasierung:	$m_{st} = \frac{m_{0t}}{m_{0s}}$
Verkettung:	$m_{0t} = m_{0s} \cdot m_{st}$

(Bemerkung: dahinter steht der "Dreisatz" $m_{0t}/m_{0s} = m_{st}/m_{ss}$ wegen $m_{ss} = 1$)

b) Messzahlen m^*_{0t}, m^*_{st} (mit 100 multiplizierte Messzahlen):

Umbasierung:	$m^*_{st} = \frac{m_{0t}}{m_{0s}} \; 100$
Verkettung:	$m^*_{0t} = \frac{m_{0s} \cdot m_{st}}{100}$

4. Wachstumsraten und Wachstumsfaktoren

a) Wachstumsraten und Wachstumsfaktoren bei diskreter Zeit

Gerade in der Ökonomie ist es sehr verbreitet mit Wachstumsraten (relativen Zuwächsen) zu argumentieren. Die Rechenregeln für den Umgang mit Wachstumsraten (z.B. Berechnung einer mittleren Wachstumsrate, der Schluß von der Wachstumsrate des Preisniveaus P auf die Wachstumsrate der reziproken Kaufkraft K [K = 1/P], vgl. Beispiel 9.12) sind aber häufig nicht bekannt, so dass es angebracht erscheint, hierauf kurz einzugehen.

Jede Wachstumsrate bezieht sich auf ein Intervall bestimmter Länge. Es ist unmittelbar einsichtig, dass die Verzinsung eines Kapitals mit 5% per anum (also jährlich) einen geringeren Zinsertrag bedeutet als eine monatliche Verzinsung von 5%. Mit einer Grenzbetrachtung bei infinitesimal kleinen (kurzen) Intervallen wird die Zeit zu einer stetigen Variable und die Formeln haben dann z.T. eine andere Gestalt als bei diskreter Zeitvariable.

Def. 9.3: Wachstumsrate und Wachstumsfaktor bei diskreter Zeit t

a) Mit der diskreten Zeitvariable t = 0,1,2,...,T erhält man für die Wachstumsrate und den Wachstumsfaktor (auch Gliedziffer oder Kettenindex genannt) der Zeitreihe y_t (d.h. der Zahlenfolge $y_0,y_1,...,y_t,...,y_T$) die folgenden Ausdrücke:

$$(9.7) \quad r_t = \frac{y_t - y_{t-1}}{y_{t-1}} = w_t - 1 \qquad (r_t\text{: Wachstumsrate}),$$

$$(9.8) \quad w_t = \frac{y_t}{y_{t-1}} = r_t + 1 \qquad (w_t\text{: Wachstumsfaktor}).$$

b) Für ein Wachstum mit konstanter Wachstumsrate [z.B. Verzinsung mit Zinseszins] r (r_t = r für alle t = 0,1,...,T) gilt:

$$(9.9) \quad y_t = y_o \cdot w^t = y_o \cdot (1+r)^t \qquad \text{(Wachstum mit konstanter Rate r).}$$

Bei variierenden Wachstumsraten r_t lautet die Wachstumsgleichung:

$$(9.10) \quad y_T = y_0(1+r_1)(1+r_2)...(1+r_T) = y_0 \prod_{t=1}^{T}(1+r_t) = y_0 \prod_{t=1}^{T} w_t \ .$$

c) Als mittlere Wachstumsrate r soll diejenige konstante Wachstumsrate bezeichnet werden, die über den gleichen Zeitraum von 0 bis T zum gleichen Wachstum von y_0 zu y_T geführt hätte, wie die tatsächlichen (unterschiedlichen) Wachstumsraten $r_1,r_2,...,r_T$. Daraus folgt, dass r aus dem geometrischen Mittel der Wachstumsfaktoren w_t zu berechnen ist

$$(9.11) \quad r = (w_1 w_2 ... w_T)^{1/T} - 1 = \left[\prod_{t=1}^{T} w_t\right]^{1/T} - 1 \ .$$

Bemerkungen zu Def. 9.3:

1. Drückt man die Wachstumsrate in Prozent aus, so ist der Wert r_t mit 100 zu multiplizieren (diese prozentuale Wachstumsrate sei p_t genannt: $p_t = 100 \cdot r_t$). Für das Wachstum mit konstanter Rate erhält man dann die vielen Kaufleuten bekannte Formel:
   $y_0(1 + p/100)^n$ für das Kapital y_n nach t = n Perioden (Jahren).

2. Gl. 9.9 ist die Lösung der Differenzengleichung
   $y_t = w \cdot y_{t-1}$ (t=1,2,...,T) mit dem Anfangswert y_0.

3. In der Ökonomie ist die ***halblogarithmische*** graphische ***Darstellung*** von Daten sehr beliebt (Ordinate: $\log(y_t)$ statt y_t und Abszisse t). Aus Gl. 9.9 folgt:

   (9.9a) $\log(y_t) = \log(y_o) + w \cdot t,$

   d.h. bei Wachstum mit konstanter Rate ist die Zeitreihe y_t in halblogarithmischer Darstellung eine Gerade. Zwei Zeitreihen x_t und y_t, die in bestimmten Intervallen (etwa $a \le t \le b$) in halblogarithmischer Darstellung parallel verlaufen, haben in diesem Intervall gleiche Wachstumsraten.

4. Der Wachstumsfaktor w_t ist der Faktor, mit dem y_{t-1} zu multiplizieren ist, um y_t zu erhalten. Man kann w_t als Messzahl mit variabler Basis betrachten und w_t ist unabhängig von der Wahl einer Basisperiode; d.h. man kann Wachstumsfaktoren (und damit auch Wachstumsraten) auch aus entsprechenden Messzahlen mit beliebiger aber gleicher Basis errechnen, da

   $$w_t = \frac{m_{0t}}{m_{0,t-1}} = \frac{m_{st}}{m_{s,t-1}}.$$

5. Während die Wachstumsrate r_t positive und negative Werte annehmen kann, gilt i.d.R. für den Wachstumsfaktor $w_t > 0$ (denn $w_t < 0$ wäre eine Abnahme um mehr als 100% und y wird i.d.R. als nichtnegativ angenommen). Aus Def. 9.3 folgt, dass für r_t und w_t stets gilt:

   $r_t = w_t - 1$ und $w_t = r_t + 1$. Hierzu folgende Beispiele:
   $r_t = +0{,}2$ (20% Zunahme) $w_t = 1{,}2$
   $r_t = +0{,}06$ (6% Zunahme) $w_t = 1{,}06$
   $r_t = -0{,}05$ (5% Abnahme) $w_t = 0{,}95$
   $r_t = -0{,}32$ (32% Abnahme) $w_t = 0{,}68$.

6. Man beachte, dass die Zeit diskret ist, d.h. es wird mit "Einheitsintervallen" (z.B. ein Monat, ein Quartal, ein Jahr usw.) gerechnet. Wird dies nicht beachtet, so entstehen typische Fehlschlüsse (weil trotz diskreter Zeit so gerechnet wird, wie es nur mit stetiger Zeit zulässig ist), auf die in den Beispielen 9.10 bis 9.12 hingewiesen wird.
7. Auf dem gleichen Fehlschluß beruht die Berechnung eines **arithmetischen** Mittels wenn es gilt, eine mittlere Wachstumsrate r zu berechnen. Damit wird r meist überschätzt. Wie Teil c von Def. 9.3 zeigt, ist vom **geometrischen** Mittel der Wachstums**faktoren** auszugehen (vgl. auch Bsp. 9.11).
8. Die Wachstumsraten-Transformation ist eine **nichtlineare** Transformation der Zeitreihe y_t (während die Bildung von Messzahlen eine lineare Transformation darstellt), d.h. dass die Gestalt der Zeitreihe r_t von derjenigen der Zeitreihe y_t durchaus verschieden sein kann. Das gilt v.a. für die Lage von Maxima und Minima, die in der Ökonomie (z.B. bei der Konjunkturdiagnose) auch "Wendepunkte" genannt werden (vgl. Bsp. 9.15).

Beispiel 9.9:

(Ein Beispiel für Wachstum mit konstanter Wachstumsrate bei diskreter Zeit: Ausbreitung des Vampirismus nach dem Wiedererscheinen von Graf Dracula. Hinweis für Leser, die mit dem Vampirismus nicht vertraut sind: wenn ein Vampir V einen Nichtvampir beißt und ihm das Blut aussaugt, dann wird dieser ebenfalls zum Vampir, aber V stirbt durch den Biss nicht, sondern er ist im Gegenteil darauf angewiesen weiteren Menschen das Blut auszusaugen.)

Der häufig von skurrilen Vorstellungen geplagte Statistiker L wird nach dem Besuch einer einschlägigen Filmvorführung den Alptraum nicht los, dass Graf Dracula von den Toten auferstehen könnte. Er geht davon aus, dass der "Durchschnittsvampir" pro Monat zwei Menschen das Blut aussaugt. Wie lange wird es dauern, bis nach Draculas Wiedererscheinen eine Bevölkerung vom Umfang

- einer Großstadt mit 700.000 Menschen
- der Bundesrepublik Deutschland (70 Millionen Menschen)

vollständig vom Vampirismus befallen sein wird?

Lösung 9.9:

Die Anzahl y_t der Vampire entwickelt sich wie folgt: $y_0 = 1 = 3^0$ (Start mit Dracula) $y_1 = 3 = 3^1$, $y_2 = 3^2 = 9$ usw. Somit gilt $w = 3$ (Wachstumsrate $r =$

2 also 200% jeden Monat) sowie $y_0 = 1$ (Dracula). Nun ist t zu bestimmen aus der Gleichung

- $700.000 = 3^t$ ($t = \log(700000)/\log(3) = 12{,}25$) bzw.
- $70.000.000 = 3^{t^*}$ ($t^* = \log(70000000)/\log(3) = 16{,}44$),

d.h. die Großstadt ist nach einem Jahr (12 Monate) und einer Woche (0,25 Monate) und die Bundesrepublik nach 16,44 Monaten vollständig vom Vampirismus befallen.

Beispiel 9.10:

Einem verbreiteten Missverständnis zufolge entspricht einer monatlichen Wachstumsrate von 0,5% einer jährlichen Wachstumsrate von 12·0,5 = 6%. Wie groß ist die tatsächliche jährliche Wachstumsrate bei einer monatlichen Wachstumsrate von

- 0,5%
- 5%
- 15%?

Lösung 9.10:

Die Rechnung: jährliche Wachstumsrate = 12·monatliche Wachstumsrate, die der gleichen Logik folgt, wie die Berechnung eines arithmetischen Mittels (Bsp. 9.11), ist nur bei kleinen Wachstumsraten annähernd richtig; je größer die monatliche Wachstumsrate, desto größer ist der Unterschied zur richtig berechneten jährlichen Wachstumsrate:

0,5%: $(1{,}005)^{12} - 1 = 0{,}06168$ also 6,17% statt 6%
5%: $(1{,}05)^{12} - 1 = 0{,}79586$ also 79,59% statt 60%
15%: $(1{,}15)^{12} - 1 = 4{,}35025$ also 435% statt 180% (=12·15%).

Folgerung aus Beispiel 9.10:

Der Grund für die Diskrepanz zwischen den beiden Rechenergebnissen ist in folgender Ungleichung zu finden:

(9.12) $w^t - 1 \geq (w-1)t = rt.$

Die Differenz zwischen w^t-1 und rt ist leicht zu bestimmen, wenn man berücksichtigt, dass w^t als binomiale Entwicklung

$$w^t = (1+r)^t = \sum_{i=0}^{t} \binom{t}{i} r^i = 1+rt + \left[\binom{t}{2} r^2 + \binom{t}{3} r^3 + \ldots + t r^{t-1} + r^t \right]$$

darzustellen ist. Sie ist mithin der Ausdruck in der eckigen Klammer.

Beispiel 9.10 zeigt, dass die Exponentialfunktion w^t von der linearen Entwicklung $rt + 1$ umso mehr abweicht, je größer w und damit auch r^2, r^3,... ist. Die Differenz $D(t) = (1+r)^t - rt - 1$ wächst auch mit t, denn es gilt:
$D(0) = D(1) = 0$, $D(2) = r^2$, $D(3) = r^3 + 3r^2$, $D(4) = r^4 + 4r^3 + 6r^2$ usw.

Beispiel 9.11:

a) Man zeige: der Mittelwert der Wachstumsraten
   - 3, 5, 4 und 8% ist nicht 5% sondern nur 4,9835%.
   - 30, 50, 40 und 80% ist nicht 50% sondern nur 48,8877%.

b) Gegeben seien die Daten 200, 260, 273, 284 und 307. Wie groß ist die mittlere Wachstumsrate?

c) Wie unterscheiden sich die korrekt gerechnete mittlere Wachstumsrate aus zwei Wachstumsraten r_1, r_2 von dem fälschlich angewendeten arithmetischen Mittel von Wachstumsraten?

Lösung 9.11:

a) Die mittlere Wachstumsrate r ist nach Def. 9.3 nicht das arithmetische Mittel 5% (vgl. auch Bsp. 4.4), sondern über das geometrische Mittel der Wachstumsfaktoren zu berechnen denn wegen:

   $y_4 = y_0 \cdot 1{,}03 \cdot 1{,}05 \cdot 1{,}04 \cdot 1{,}08 = y_0 \cdot (1+r)(1+r)(1+r)(1+r)$ ist
   $r = (1{,}03 \cdot 1{,}05 \cdot 1{,}04 \cdot 1{,}08)^{1/4} - 1 = 0{,}049835$ also 4,98%

   Man kann als Ergebnis festhalten:

   > Die mittlere Wachstumsrate ist nach Def. 9.3 aus dem **geometrischen** Mittel der Wachstums**faktoren** zu bestimmen, nicht aber als **arithmetisches** Mittel der Wachstums**raten**.

   Entsprechend ist zu rechnen
   $r = (1{,}3 \cdot 1{,}5 \cdot 1{,}4 \cdot 1{,}8)^{1/4} - 1 = 4{,}914^{1/4} - 1 = 0{,}4888$ also 48,89% und nicht $(30+50+40+80)/4 = 200/4 = 50\%$.

b) Es ist nicht nötig, die einzelnen Wachstumsraten auszurechnen. Für die Wachstumsfaktoren erhält man 260/200, 273/260, 284/273 und 307/284. Bildet man das Produkt dieser vier Brüche, so sieht man, dass sich einiges wegkürzt, so dass man einfach erhält:
   $r = (307/200)^{1/4} - 1 = 0{,}113$ (also 11,3%)
   oder allgemein: $(y_t / y_0)^{1/t} - 1$ (wenn y_0 der erste Wert und y_t der letzte Wert ist). Man beachte, dass der Bruch y_t/y_0 die Messzahl m_{0t} darstellt, so dass man für die mittlere Wachstumsrate r erhält

(9.11) $r = (y_t/y_0)^{1/t} - 1 = \sqrt{}_t,m_{0t} - 1$ (mittlere Wachstumsrate),

bzw. in Prozent:

(9.11a) $r = [(y_t/y_0)^{1/t} - 1]100$

c) Die korrekt aufgrund eines geometrischen Mittels der Wachstumsfaktoren berechnete mittlere Wachstumsrate aus den beiden Wachstumsraten r_1, r_2 soll $r = r_G$ und das arithmetische Mittel soll r_A genannt werden ($r_A = ½(r_1+r_2)$). Dann ist offensichtlich stets $r_A > r_G$, es sei denn $r_1 = r_2$ denn es gilt:

$$(1 + r_A)^2 - ¼(r_1 - r_2)^2 = (1 + r_G)^2.$$

Es kommt mithin auf die Unterschiedlichkeit von r_1 und r_2 an. Der Ausdruck $¼(r_1 - r_2)^2$ ist übrigens die Varianz der Wachstumsraten r_1 und r_2 also (s_r^2). Insbesondere bedeutet auch $r_A = 0$, nicht $r_G = 0$ denn es gilt: wenn $r_A = 0$ dann $r_1 = - r_2$ und $(1+r_G)^2 = 1 - ¼(2r_1)^2 = 1-r_1^2$, so erhält man z.B. bei zunehmender Varianz s_r^2:

$r_1 = 0{,}1$ und $r_A = 0$ dann $r_G = -0{,}0050126$
$r_1 = 0{,}2$ und $r_A = 0$ dann $r_G = -0{,}0202041$ oder
$r_1 = 0{,}3$ und $r_A = 0$ dann $r_G = -0{,}0460608$.

Beispiel 9.12:

Einer Zunahme des Preisniveaus von 10% entspricht eine Abnahme der Kaufkraft in Höhe von (Richtiges ankreuzen)

- 10%
- mehr als 10%
- weniger als 10%

Lösung 9.12:

Viele sind geneigt, 10% anzukreuzen. Dass dies falsch ist, erkennt man daran, dass bei dieser Logik eine Zunahme des Preisniveaus um 100% (also eine Verdoppelung der Preise, was als Inflationsrate durchaus vorkommen kann) ein Kaufkraftverlust von 100% bedeuten würde, was natürlich nicht möglich ist. Eine Verdoppelung der Preise bewirkt vielmehr eine Halbierung der Kaufkraft. Der Kaufkraftverlust beträgt also "nur" 50% und nicht 100%. Da die Kaufkraft (K) das reziproke Preisniveau (P) ist, d.h. es ist $K_t = 1/P_t$ (für jede Periode t), gilt für die Wachstumsfaktoren (bzw. Wachstumsraten) von P und K, w_P und w_K (bzw. r_P und r_K) bei diskreter Zeit (vgl. auch Übers. 9.5):

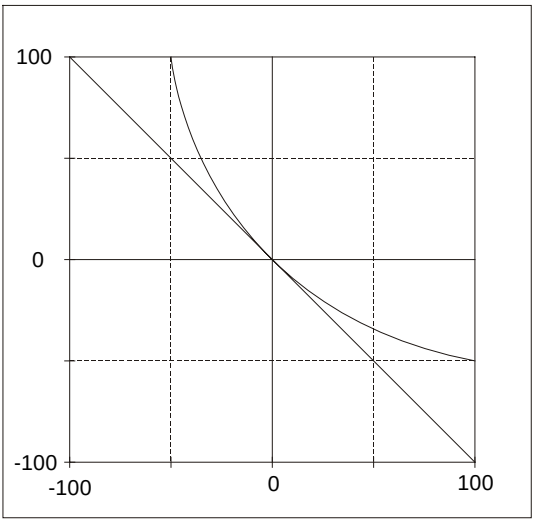
(9.13) $w_K = \frac{1}{w_P}$ und damit $r_K = \frac{1}{1 + r_P} - 1$ wenn $K = \frac{1}{P}$

bzw. wenn die Wachstumsraten in Prozent ausgedrückt sind:

(9.13a) $r_K^* = 10000/(100+r_P^*) - 100$.

Der Zusammenhang zwischen den Wachstumsraten von P und K ist also hyperbolisch wie die nebenstehende Abb. 9.3 zeigt. Beispiele für Gl. 9.13a: $r_P^* = 25\%$ dann ist $r_K^* = -20\%$ oder $r_P^* = 33.3\%$ (also 1/3) $r_K^* = -25\%$ (also 1/4) usw. Auch hier gilt: der Fehler, der dadurch entsteht, dass man bei diskreter Zeit so rechnet, wie mit stetiger Zeit (also $r_K^* = -r_P^*$, die eingezeichnete Gerade in Abb. 9.3) ist umso geringer, je kleiner die Wachstumsrate ist.

Abb. 9.3: Kaufkraft und Preisniveau



b) Wachstumsraten und Wachstumsfaktoren bei stetiger Zeit

Bei diskreter Zeit wurde die Zeitreihe und deren Wachstumsrate mit y_t und r_t (mit t=0,1,...,T als Periodenindex) bezeichnet. Wird die Zeit dagegen als stetige Variable ($t \in \mathbb{R}$) betrachtet, so ist eine Zeitreihen- und eine Wachstumsratenfunktion gegeben, die zur besseren Unterscheidbarkeit mit y(t) und r(t) bezeichnet werden sollen. Im Falle einer konstanten Wachstumsrate soll diese bei stetiger Zeit mit a bezeichnet werden (anstelle von r bei diskreter Zeit).

Der Übergang von $r_1,...,r_T$ der diskreten Folge $y_0,y_1,...,y_T$ gem. Def. 9.3 zur stetigen Wachstumsratenfunktion r(t) für die Zeitreihe y(t) erfolgt durch Verkürzung der Intervalle, auf die sich die Wachstumsraten beziehen. Mit

$$r(t,\Delta) = \frac{(y_t - y_{t-\Delta})/\Delta}{y_{t-\Delta}}$$

(man beachte jedoch, dass in diesem Fall t nicht mehr als diskret, sondern als stetig angenommen werden muss, weil D beliebig klein sein darf) ist die Wachstumsrate bezüglich des Intervalls der Länge D gegeben. Mit den Einheitsintervallen ($\Delta = 1$) erhält man Gl. 9.7 und der Grenzübergang

$$\lim_{\Delta \to 0} (r(t,\Delta)) = r(t)$$

liefert für die Funktion y(t) an jeder beliebigen Stelle t, etwa bei $t = t_o$ die Wachstumsrate r(t) von y(t). Sie ist nach Def. 9.4

$$r(t=t_o) = \frac{y'(t)}{y(t)}\Big|_{t_o} = \frac{dy/dt}{y}\Big|_{t_o}$$

Man kann y'(t) auch als absolutes - und r(t) = y'(t)/y(t) als relatives Wachstum bezeichnen.

Def. 9.4: Wachstumsrate bei stetiger Zeit

a) Die Wachstumsrate r(t) einer stetigen Funktion y = y(t) ist

$$(9.14)\quad r(t) = \frac{y'(t)}{y(t)} = \frac{dy/dt}{y} = \frac{d\ln(y)}{dt}$$

b) Bei konstanter Wachstumsrate r(t) = a (für jeden Wert von t) ist die stetige Zeitreihe y(t) gegeben mit

$$(9.15)\quad y(t) = y(0)\cdot e^{at} = y(0)\cdot\exp(at).$$

c) Bei variablen Wachstumsraten r(t) gilt entsprechend

$$(9.16)\quad y(T) = y(0)\cdot\exp\left(\int_0^T r(t)dt\right)$$

wenn r(t) eine im Intervall (0,T) stetige Funktion ist, so dass man für eine mittlere Wachstumsrate in Analogie zu Def. 9.3

$$(9.17)\quad a = \frac{1}{T}\int_0^T r(t)\,dt = \frac{\ln[y(T)/y(0)]}{T} = \frac{1}{T}\{\ln[y(T)] - \ln[y(0)]\}$$

erhält.

Bemerkungen zu Def. 9.4:

1. Im stetigen Fall hat der Begriff "Wachstumsfaktor" wenig Sinn (im Unterschied zur Wachstumsrate). Nach Gl. 9.14 erhält man die stetige Wachstumsrate indem man die erste Ableitung der Funktion y(t) bildet (y'(t) = dy/dt) und diese durch die Funktion y(t) teilt. Die Funktion r(t) ist die Wachstumsrate von y(t), d.h. für jeden Wert von t, etwa für $t=t_o$ stellt $r(t=t_o)$ die Wachstumsrate von $y(t=t_o)$ dar. Wegen

$$\frac{d\ln(y)}{dt} = \frac{1}{y}\frac{dy}{dt}$$

wird Gl. 9.14 auch logarithmische Ableitung von y(t) genannt.

2. Gl. 9.15 ist die Lösung der Differentialgleichung dy/dt = ay mit dem Anfangswert y(0). Nach Gl. 9.15 ist ferner ln[y(t)] = ln[y(0)] + at und die logarithmische Ableitung hiervon ist dln[y(t)]/dt = a. Der Übergang von Gl. 9.9 zu Gl. 9.15 ist auch leicht einsichtig zu machen, wenn man den sog. "Aufzinsungsfaktor" bei jährlicher, halbjährlicher, monatlicher Verzinsung betrachtet, denn es gilt dann:

$(1+r)^t$, $(1+\frac{r}{2})^{2t}$, $(1+\frac{r}{12})^{12t}$ usw., mit $\lim_{n\to\infty}(1+\frac{r}{n})^{nt} = e^{rt}$.

3. Aus Gl. 9.15 und 9.9 folgt, dass zwischen der Wachstumsrate α (stetige Zeit) und r (diskrete Zeit) die folgende Beziehung besteht:

(9.18) $e^{\alpha} = w = 1+r$, so dass gilt (9.19) $a = \ln(1+r)$.

Man erhält somit im Zusammenhang mit der Reihenentwicklung von e^a und ln(1+r) die folgenden Umrechnungen

$$(9.20)\quad r = e^{\alpha} - 1 = a + \frac{a^2}{2!} + \frac{a^3}{3!} + \frac{a^4}{4!} + \ldots$$

für die Umrechnung von a in r (so dass a < r) und

$$(9.21)\quad \alpha = \ln(1+r) = r - \frac{r^2}{2!} + \frac{r^3}{3!} + \frac{r^4}{4!} + \ldots$$

für die Umrechnung von r nach α.
Wie man sieht, gilt nur bei kleinen Wachstumsraten $r \approx \alpha$.

Beispiel 9.13:

Gegeben sei die Funktion $y(t) = \exp(a + bt + ct^2)$, die auch logarithmische Parabel genannt wird, da $\ln[y(t)] = a + bt + ct^2$.

a) Wie lautet die Funktion r(t) und welche Besonderheiten hat sie?

b) Wie lautet für die Funktion $\ln[y(t)] = 0{,}1 + 0{,}5t - 0{,}04t^2$ die
- Wachstumsrate r(3) (also r(t) für t=3)
- mittlere Wachstumsrate im Intervall $0 \le t \le 6$?

Lösung 9.13:

a) Man sieht an diesem Beispiel leicht, dass gilt

$$\frac{dy/dt}{y} = \frac{dln(y)}{dt}$$

denn die Ableitung dy/dt beträgt nach der Kettenregel mit $z = a + bt + ct^2$ $dy/dt = dy/dz \cdot dz/dt = e^z \cdot (b+2ct)$, so dass die Wachstumsrate $r(t) = b+2ct$ ist, also linear zu- oder abnimmt. Zum gleichen Ergebnis gelangt man durch die Ableitung von lny nach t.

b) Die Funktion y(t) ist in Abb. 9.4 (links oben y(t) und links unten r(t)) dargestellt. Sie steigt von $y(0) = e^{0,1} = 1,10517$ auf $y(6,25) = 5,2725$ und fällt dann auf $y(10) = e^{-2,9} = 0,05502$. Die Funktion $r(t) = 0,5 - 0,08t$ fällt linear von $r(0) = 0,5$ zu $r(6) = 0,02$. Sie nimmt bei $t = 3$ den Wert $r(3) = 0,26$ an und den Wert Null an der Stelle $t = 6,25$. An dieser Stelle durchläuft y(t) ein Maximum mit $y(t=6,25) = \exp(1,6625) = 5,2725$. Das Integral über r(t) in den Grenzen 0 bis 6 beträgt 1,56. Man beachte, dass gilt $y(0) = e^{0,1}$ und $y(6) = e^{1,66} = 5,2593$ und $y(0) \cdot e^{1,56} = y(6)$, so dass man nach Gl. 9.17 für die mittlere Wachstumsrate erhält

$$\alpha = \ln[y(T)/y(0)]/T = T^{-1} \{\ln[y(T)] - \ln[y(0)]\}$$
$$= \{\ln[e^{1,66}] - \ln[e^{0,1}]\}/6 = (1,66-0,1)/6 = 1,56/6 = 0,26.$$

Die mittlere stetige Wachstumsrate beträgt also 26% (dem entspräche eine mittlere Wachstumsrate im diskreten Fall von 29,69%, denn $(e^{0,26})^6 = (1,2969)^6 = 4,7588 = 5,2593/1,10517$. Die Größe 4,7588 ist genau das Verhältnis $y(6)/y(0) = e^{1,66}/e^{0,1} = e^{1,56}$.

Beispiel 9.14:

In einem stetigen Wachstumsmodell werden konstante Wachstumsraten in Höhe von 0,5%, 5% und 50% unterstellt. Wie groß sind die entsprechenden diskreten Wachstumsraten?

Lösung 9.14:

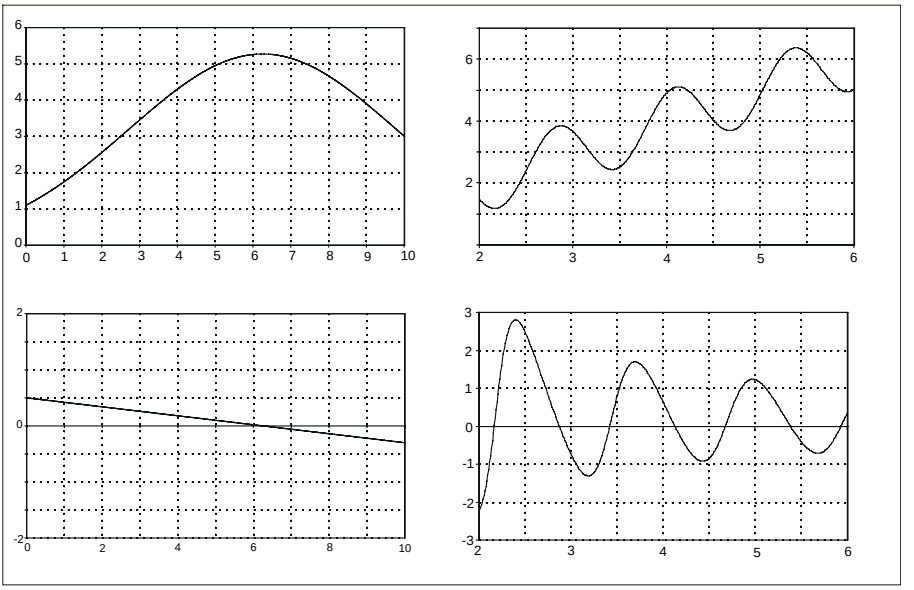
Wegen $r = e^a - 1$ gilt a = 0,005 entspricht r = 0,0050125 (der Unterschied zwischen r und a ist also gering), a = 0,05 entspricht r = 0,05127 und a = 0,5 entspricht r = 0,64872 (der Unterschied zwischen r und a ist beträchtlich).

Beispiel 9.15:

Man kann leicht zeigen, dass die Wachstumsratenfunktion andere Maxima und Minima besitzt, als die Zeitreihe y(t). Hierzu das Beispiel: y(t) = t + sin(5t). Wie lautet die Funktion der Wachstumsraten?

Abb. 9.4: Wachstumsraten für einige stetige Funktionen

Links: Logarithmische Parabel (Beispiel 9.13), Rechts: y(t) = t + sin(5t) (vgl. Beispiel 9.15), oberes Bild jeweils y(t) und unteres Bild r(t)



Lösung 9.15:

Die Ableitung von y(t) nach t lautet dy/dt = 1 + 5cos(5t), so dass r(t) = [1+5cos(5t)]/[t+sin(5t)]. Offensichtlich haben y(t) und r(t), wie Abb. 9.4 zeigt, unterschiedliche Extrema.

c) Weitere Bemerkungen zu Wachstumsraten

In diesem Exkurs soll noch einmal auf verbreitete Mißverständnisse im Umgang mit Wachstumsraten eingegangen werden und es werden einige Formeln zu bekannten Funktionen y(t) und deren Wachstumsraten r(t) zusammengestellt, die häufig als Modelle eines Wachstumsprozesses oder auch als Trendmodelle benutzt werden.

1. Die Wachstumsrate eines Produkts, Quotienten und Kehrwerts

Das Beispiel 9.12 hat gezeigt, dass oft falsche Vorstellungen über die Wachstumsrate einer Funktion von y(t) bestehen. Gerade in der Ökonomie

werden oft Größen als Produkte oder Quotienten (oder z.B. als Kehrwert) definiert:

- So ist z.B. der Umsatz (z) das Produkt aus Menge (x) und Preis (y) und viele sind geneigt, z.B. wie folgt zu rechnen: eine Zunahme des Preises um 7% und eine Zunahme der Menge um 3% bedeuten eine Umsatzzunahme von 10%, was nur bei stetiger Zeit gilt (ist t diskret so ist die Umsatzzunahme 1,07·1,03 - 1 = 0,1021 also nicht 10% sondern 10,21%);
- Das Realeinkommen ist der Quotient Nominaleinkommen/Preisniveau und oft wird wie folgt gerechnet: eine Zunahme des Nominaleinkommens z.B. um 20% bedeutet bei einer Preissteigerung von 6% eine Zunahme des Realeinkommens um 20 - 6 = 14%. Bei diskreter Zeit beträgt jedoch die Zunahme des Realeinkommens nicht 14% sondern nur 13,2%, denn 1,2/1,06 - 1 = 0,13208.

In Übersicht 9.4 sind die relevanten Formeln zusammengestellt.

Übersicht 9.4: Wachstumsraten von Produkten, Quotienten und Kehrwerten

	diskrete Zeit	stetige Zeit
Produkt $z = xy$	$w_z = w_x w_y$	$r_z(t) = r_x(t) + r_y(t)$
Quotient $z = \frac{x}{y}$	$w_y = \frac{w_x}{w_y}$	$r_z(t) = r_x(t) - r_y(t)$
Kehrwert $z = \frac{1}{y}$	$w_z = \frac{1}{w_y}$	$r_z(t) = - r_y(t)$

2. Einige Wachstumsfunktionen insbesondere Kurven mit einem Sättigungsniveau und deren Wachstumsraten

In Übersicht 9.5 sind einige Funktionen y(t) mit stetiger Zeitvariable t zusammengestellt, die bei Ökonomen von gewissem Interesse sind. Dies gilt insbesondere für solche Kurven, die einer Sättigungsgrenze k entgegenstreben (Nr. 4 bis 9).

Übersicht 9.5.: Wachstumsraten ausgewählter Funktionen

	Funktion y(t)	Ableitung $y = dy/dt$	Wachstumsrate $r(t)$ (1)
1	$(a+bt)^{\alpha}$ (2)	$ab(a+bt)^{\alpha-1}$	$ab(a+bt)^{-1}$
1a	$\alpha=1$: Gerade $y = a+bt$	b	$b/(a+bt) = r_G$
1b	$\alpha=-1$: $\frac{1}{a+bt}$	$-\frac{b}{(a+bt)^2}$	$-\frac{b}{a+bt} = -r_G$
1c	$\alpha=½$: $\sqrt{(a+bt)}$	$½b(a+bt)^{-1/2}$	$½r_G$
1d	Potenzfunktion bt^a	$b\alpha t^{\alpha-1}$	α/t (hyperbolisch)
2	Parabel $a+bt+ct^2$ (3)	$b+2ct$	$\frac{b+2ct}{a+bt+ct^2}$
3	$a\cdot\exp(bt^{\alpha})$	$y\cdot\alpha bt^{\alpha-1}$	$\alpha bt^{\alpha-1}$
3a	$\alpha=1$: ae^{bt} oder: ar^t mit $r = e^b$	yb $y\cdot\ln(r)$	b $b = \ln(r)$
3b	$\alpha=-1$: $ae^{b/t}$	$-yb/t^2$	$-b/t^2$
4	$k+be^{ct}$ oder $y = k+br^t$ mit $r=e^c$ (k: Sättigungsniveau)	cbe^{ct} strebt gegen 0, wenn $r<1$, $c=\ln r<0$	$-c\frac{k-y}{y} = -cR$ speziell für: $c = -1$, ist $r(t) = R$; Abb. 9.5
5	$k+\frac{b}{c+t}$ (Hyperbel)	$-\frac{b}{(c+t)^2}$	$-\frac{b}{k(c+t)^2+b(c+t)}$
6	$\frac{k(t+a)}{t+b}$ $(b>a)$	$\frac{k(b-a)}{(t+b)^2}$	$\frac{b-a}{(t+a)(t+b)}$
7	$\exp(K+br^t)$ mit $r=e^c$ oder: $\ln(y) = K+be^{ct}$; $k=e^K$ Sättigungsniveau $(b < 0)$	$ybcr^t$ strebt gegen 0, wenn $r<1$ $(c<0)$	$br^t\ln(r) = bce^{ct}$ [mit $\ln(r) = c < 0$] (vgl. Abb. 9.5)
8	$\frac{k}{1+e^{a-bt}}$ a,b,k>0 k Sättigungsniveau	$\frac{by(k-y)}{k}$	$b+ßy$ $(ß = -\frac{b}{k})$ (4) (vgl. Abb. 9.5)
9	$\ln(y) = K - \frac{a}{b+t}$ $k=e^K$ Sättigungsniveau	$\frac{ya}{(b+t)^2}$	$\frac{a}{(b+t)^2}$

(1) $r(t) = y'/y$

(2) In dieser allgemeinen Form liegt eine Polynomfunktion vom Grade a vor, wobei a Î Z, Z^+ ist. Bei a=½ liegt eine Wurzelfunktion und bei ganzzahligem a und a>0 eine Potenzfunktion vor (Fall 1d).
(3) Kann entsprechend verallgemeinert werden wie Funktion Nr.1.
(4) Kennzeichnend für die logistische Funktion: r(t) = f[y(t)] (f: linear).

Man beachte:
1. cy(t) und y(t) haben die gleiche Wachstumsrate r(t) [c: Konstante]
2. hat y(t) die Wachstumsrate r(t), so hat $[y(t)]^{-1}$ die Wachstumsrate -r(t)

Weitere Funktionen in den Beispielen 9.13 (logarithmische Parabel) und 9.15.

Sättigungsniveau:

Die Funktionen Nr. 4 ff haben Sättigungsniveaus (k), denen sich y(t) asymptotisch nähert und zwar [bei bestimmter Konstellation der Parameter] monoton (Nr. 4 bis 6) oder S-förmig (d.h. mit Wendepunkt, Nr. 7 - 9), weil die Ableitung dy/dt mit wachsendem t gegen 0 strebt.

Namen von Funktionen:

4: modifizierte Exponentialfunktion, 6: Törnquist-Funktion (für a=0 oder b=0 ergeben sich einfach linearisierbare Funktionen [Übers. 8.2], 7: Gompertz - Funktion, 8: logistische Funktion, 9: Johnson-Funktion (eine "logarithmische Hyperbel").

1. Die modifizierte Exponentialfunktion (Nr.4 in Übers. 9.5)

In Abb. 9.5 ist links oben die folgende Funktion dargestellt:
$y(t) = k + br^t$ mit k = 20, b= -16 und r = 0,95 (so dass c = ln(r) = -0,0513), also $y(t) = 20 - 16 \cdot 0{,}95^t$.

Die Funktion steigt von y(0) = k+b = 10+(-16) = 4 monoton zum Sättigungsniveau k=20 und zwar umso steiler, je kleiner r ist (für r=1 ist y(t) eine Parallele der Abszisse t). Sie

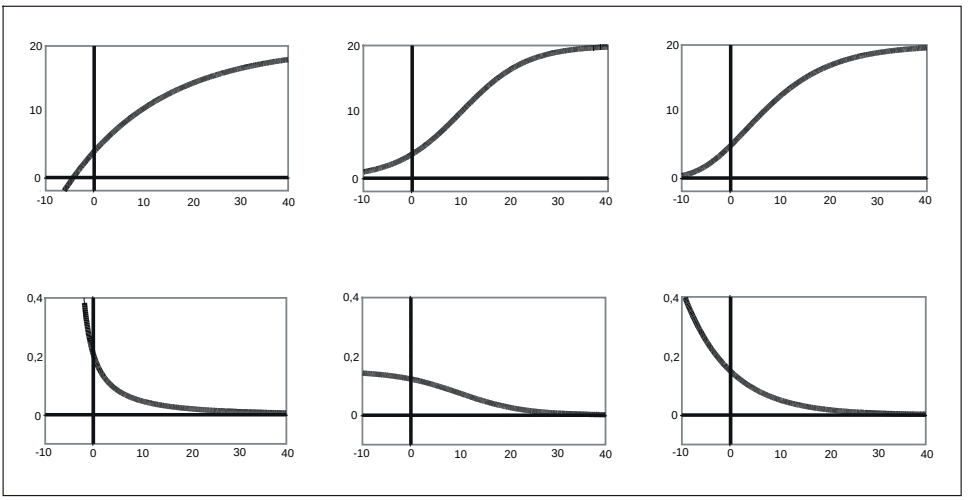
- nähert sich k monoton wenn r < 1 und zwar steigend wenn b < 0, fallend wenn b > 0
- entfernt sich monoton von k wenn r > 1, steigend wenn b > 0, fallend wenn b < 0.

Die Wachstumsrate (Abb. 9.5 links unten) ist $r(t) = (-\ln(r))[k-y(t)] / y(t) = 0{,}051294 \cdot 16 \cdot 0{,}95^t / [20+(-16 \cdot 0{,}95^t)]$und sie ist monoton fallend; $r(0) = 0{,}051294 \cdot (20-4)/4 = 0{,}20517$. Der Ausdruck

$R = [k - y(t)]/y(t)$

ist der relative Abstand der Funktion vom Sättigungsniveau k. Die modifizierte Exponentialfunktion ist also dadurch ausgezeichnet, dass die Wachstumsrate r dieser Größe R proportional ist (Proportionalitätsfaktor c). Die Logarithmen von R heißen **Logits**. Wie man leicht sieht erhält man mit k = 0 die einfache Exponentialfunktion. Dann ist R = -1 und die Wachstumsrate ist r(t) = (-c)R = c, was der Größe b in Zeile 3a der Übersicht 9.5 entspricht.

Abb. 9.5: Einige Funktionen mit einer Sättigungsgrenze und deren Wachstumsraten



2. Die logistische Funktion (Nr. 8 in Übersicht 9.5)

Sie hat, anders als die modifizierte Exponentialfunktion einen Wendepunkt. Als logistische Funktion bezeichnet man eine Funktion der folgenden Gestalt:

$$y(t) = \frac{k}{1 + e^{f(t)}}$$

wobei f(t) eine monoton fallende Funktion ist, z.B. die Gerade f(t) = a - bt (mit a,b,k > 0). In Abb. 9.5 (Mitte oben) wurde die folgende Funktion dargestellt

$$y(t) = \frac{20}{1 + \exp(1{,}5\text{-}0{,}15t)}$$

Es gilt: $y(0) = k/(1+e^a) = 20/(1+e^{1,5}) = 3{,}6485$.
Der Wendepunkt liegt bei t = a/b = 10 und dann ist y(t) = k/2 = 10. Die Kurve steigt umso steiler, je größer bei gegebenen k und a der Parameter b ist. Sie ist zentralsymmetrisch um den Wendepunkt.
Die logistische Funktion ergibt sich aus der Differentialgleichung dy/dt = by(k-y)/k.
Die Wachstumsrate (Abb. 9.5 Mitte unten) der logistischen Funktion ist linear abhängig vom erreichten Niveau y was die Schätzung der logistischen Funktion erleichtert:
r(t) = b - by/k = 0,15 - 0,0075y(t).
Sie fällt ab $r(0) = b - b/(1+e^a) = 0{,}15 - 0{,}15/(1+e^{1,5}) = 0{,}12264$ monoton. Die Logits L(t) haben einen linearen Trend: L(t) = ln(R) = ln{[k-y(t)]/y(t)} = a - bt.
Man erkennt an der Gestalt der Logit-Funktion übrigens auch die Symmetrie. So gilt z.B.
für y = k/2: L = ln(1) = 0, so dass t = a/b = 10
für y = k/4: L = ln(3) , so dass t = (a-ln3)/b = 2,676 und

für y = k/2: L = ln(1/3), so dass t = (a-ln3)/b = 17,324.

3. Die Gompertzfunktion (Nr. 7 in Übers. 9.5)

Auch sie hat einen Wendepunkt ist aber weniger bekannt als die logistische Funktion. In Abb. 9.5 (rechts oben) wurde dargestellt:

$y(t) = k \cdot \exp(b \cdot r^t) = 20 \cdot \exp(-1{,}4 \cdot 0{,}9^t)$.
k = 20, b = -1,4 und r = 0,9 so dass c = ln(r) = -0,10536.

Mit kleinerem Wert von r (etwa r=0,8) verläuft die Kurve steiler. Ist r > 1, so ist die Kurve fallend. Es gilt $y(0) = ke^{-b} = 20e^{-1,4} = 4{,}932$ (je größer betragsmäßig b ist, desto niedriger ist der Ordinatenabschnitt). Der Wendepunkt liegt bei $t_W = -\ln(|b|)/c = -\ln(1{,}4)/(-0{,}10536) = 3{,}1935$, wobei y stets $y(t_W) = k/e = 7{,}3576$ (mit k = 20) beträgt.
Die Wachstumsrate (Abb. 9.5 rechts unten) lautet $r(t) = b(\ln r)r^t = 0{,}147505 \cdot 0{,}9^t$, sie ist also ab r(0) = 0,1475 monoton fallend. Es gilt $\ln[k/y(t)] = br^t$, so dass c[ln(y) - ln(k)] = r(t), d.h. die Wachstumsrate ist proportional zur Differenz der Logarithmen von y und k (bzw. zu ln(y/k)).
Zusammenfassend kann festgestellt werden:

Funktion		die Wachstumsrate ist
4	modifizierte Exponentialfunktion	proportional zu $R = \frac{k-y}{y}$ relativer Abstand zum Sättigunsniveau
7	Gompertzfunktion	proportional zu $\ln\left(\frac{y}{k}\right) = \ln\left(\frac{1}{1+R}\right)$
8	logistische Funktion	linear abhängig von y/k $r(t) = b - b\frac{y}{k}$ die Logits L(t) = ln(R) haben einen linearen Trend L(t) = a - bt

5. Aggregationsprobleme

In diesem Abschnitt werden einige Probleme behandelt, die nicht beschränkt sind auf Verhältniszahlen, aber andererseits auch nicht ein eigenes Kapitel rechtfertigen dürften. Um den Begriff der Aggregation deutlich zu machen, ist auch auf Daten zurückzugreifen, die als Häufigkeitsverteilung vorliegen.

a) Begriff der Aggregation

Im Rahmen der Deskriptiven Statistik wird unter Aggregation meist eine Summenbildung verstanden. Liegen Daten als Häufigkeitsverteilung vor, so wird der Begriff benutzt für unterschiedliche Fragestellungen:

a) Aussagen über eine Variable X für verschiedene Teilgesamtheiten und Zusammenhänge zwischen den Aussagen (z.B. den Mittelwerten), die für die Gesamtheit und jene, die für die Teilmassen gelten. Dabei können diese Teilgesamtheiten aufgrund aneinandergrenzender Intervalle auf der x-Achse (Klassenbildung) oder aufgrund eines anderen (meist qualitativen aber auch quantitativen[1]) Merkmals gebildet sein (Zerlegung, vgl. S. 24). In diesem Sinne interessieren auch Zusammenhänge zwischen einer aggregierten Beziehungszahl X/Y und den Teilbeziehungszahlen x_j/y_j, wenn bei j = 1, 2, .., J Teilgesamtheiten gilt $X = \Sigma x_j$ und $Y = \Sigma y_j$.

b) Aussagen über eine Variable, die eine Summe darstellt (etwa Z = X + Y als ungewogene Linearkombination oder $Z = b_1 X_1 + b_2 X_2$ als [mit den Gewichten b_1 und b_2] gewogene Linearkombination).

Um den Unterschied deutlich zu machen, sollte man **im Fall a)** von **Aggregation von Verteilungen** (Summation über Einheiten bei einer Variablen) sprechen, und **im Fall b)** von einer **Linearkombination** (Summation über Variablen bei einer Gesamtheit). Im Fall a) wird über die Häufigkeiten n_i (auf der Ordinate) und im Fall b) über die Merkmalswerte (Abszisse) summiert.

Der Begriff Aggregation tritt auch in der Wirtschaftsstatistik auf, insbesondere in der Sozialproduktsrechnung. Gemeint ist damit die sog. "fundierte Schätzung" eines (inhomogenen) "Aggregats", die gerade nicht durch eine einfache Summenbildung möglich ist. Indexzahlen betrachten aggregierte Veränderungen. Die speziellen Probleme der Indizes (Kap. 10) im Unterschied zu einfachen Messzahlen (Def. 9.1 d) entstehen alle dadurch, dass Indizes Aggregate vergleichen.

Das Aggregationsproblem (im engeren Sinne) tritt auch in der Ökonometrie bei der Frage auf, ob und unter welchen Voraussetzungen eine funktionale Beziehung zwischen mikroökonomischen (auf das einzelne Wirtschaftssubjekt bezogene) Variablen, etwa X_v und Y_v (etwa eine Regressionsfunktion $\hat{y} = a + bx_v$) auch für aus diesen Größen durch

---

[1] Ein Beispiel für ein quantitatives Merkmal wäre: Zusammenhang zwischen der rohen und den altersspezifischen Todesraten.

Summation hervorgegangene Makrovariablen $X = \Sigma X_i$ und $Y = \Sigma Y_i$ gelten kann, und wie dann die entsprechenden Koeffizienten zu interpretieren sind.

b) Aggregation von Verteilungen

Bei der Aggregation von Verteilungen ist es im Unterschied zur Klassierung (Klassenbildung) nicht notwendig, dass die Teilgesamtheiten disjunkt sind. Die zulässigen Wertebereiche für die Variable X können sich auch überschneiden (vgl. Beispiel 9.16) oder gar identisch sein. Beispiele für das Auftreten dieses Problems sind die Schätzung einer "gemeinsamen" Einkommensverteilung auf der Grundlage der Lohn- und Einkommensteuerstatistik, die Verteilung der Arbeitsverdienste der Arbeitnehmer eines Unternehmens mit mehreren Betrieben oder die Schätzung von Bundesergebnissen auf der Basis der Bundesländer.

Über die Gestalt der so entstehenden "Mischverteilungen" lassen sich keine allgemeinen Aussagen machen. Durch Aggregation können z.B. mehrgipflige oder schiefe Verteilungen entstehen auch wenn alle zugrundeliegenden Verteilungen der Teilgesamtheiten unimodal und symmetrisch sind.
Allerdings gibt es für die Zusammenhänge zwischen den Verteilungsmaßzahlen der J Teilgesamtheiten (j = 1,2,...,J) und der entsprechenden Verteilungsmaßzahl der aggregierten Verteilung oft einfache Beziehungen (z.B. ein gewogenes Mittel).
Für Mittelwerte und Streuungsmaße sind sie in der Regel bei der Darstellung der entsprechenden Maßzahl erwähnt worden. So gilt z.B. für das arithmetische Mittel $\bar{x}$ und die Varianz s^2 der aggregierten Verteilung

(*) $\bar{x} = \Sigma g_j \bar{x}_j$ (Teilgesamtheit j = 1,2,...,J) und

(**) $s^2 = \Sigma(\bar{x}_j - \bar{x})^2 g_j + \Sigma s_j^2 g_j$

mit $g_j = n_j / n$, so dass $0 \leq g_j \leq 1$ und $\Sigma g_j = 1$.

Man beachte, dass die Formeln mit denen übereinstimmen, die für die Klassierung gelten.

Beispiel 9.16:

Gegeben seien die Häufigkeitsverteilungen für das Merkmal X für die beiden Teilgesamtheiten A und B (mit den absoluten Häufigkeiten n_{Ai} und n_{Bi}).

x_i	n_{Ai}	n_{Bi}
2	10	-
3	20	10
4	10	40
5	-	10

$\Sigma x_i n_{Ai} = 120$
$\Sigma x_i n_{Bi} = 240$

Bestimmen Sie die aggregierte Verteilung, sowie deren arithmetisches Mittel, Varianz und Schiefe.

Lösung 9.16:

x_i	n_i
2	10
3	30
4	50
5	10

$\Sigma x_i n_i = 360$

Es gilt $\bar{x}_A = 3$, $\bar{x}_B = 4$, $n_A = 40$, $n_B = 60$, so dass man erhält:
die Gewichte $g_A = 40/100 = 0{,}4$ und $g_B = 0{,}6$, ferner $\bar{x} = 3{,}6$ und für die Varianzen $s_A^2 = 1/2$, $s_B^2 = 1/3$. Die innere Varianz $\Sigma s_j^2 g_j$ beträgt somit 0,4 und für die äußere Varianz $\Sigma(\bar{x}_j - \bar{x})^2 g_j$ erhält man 0,24, so dass die Gesamtvarianz s^2 der aggregierten Verteilung 0,4 + 0,24 = 0,64 beträgt. Man sieht ferner, dass die Verteilungen in den Teilgesamtheiten A und B symmetrisch sind, die aggregierte Verteilung dagegen rechtssteil ist (Momentschiefe: $-0{,}168/0{,}8^3 = -0{,}328$). Betrachtet man die Merkmalssummen, so gilt einfach $\Sigma x_i n_i = \Sigma x_i n_{iA} + \Sigma x_i n_{iB} = 120+240 = 360$.

c) Verteilung einer Linearkombination

Gegeben sei eine Häufigkeitsverteilung der Variablen X und eine der Variable Y und gefragt ist nach der Verteilung der Größe X + Y, wobei jedoch die gemeinsame Verteilung von X und Y gegeben sein muss. Das folgende Beispiel möge dies veranschaulichen.

Beispiel 9.17:

Die Variablen X und Y besitzen die folgenden Verteilungen

Randverteilungen

x_i	n_i	y_j	n_j
2	10	2	20
4	12	3	10
5	8		

gemeinsame Verteilung

	y=2	y=3
x=2	8	2
x=4	9	3
x=5	3	5

$\overline{x} = 3{,}6$ $\overline{y} = 2{,}333 = 7/3$ $s_{xy} = 0{,}1667$ (Kovarianz)

$s_x^2 = 1{,}44$ $s_y^2 = 2/9$

Die Variable $Z_1 = X + Y$ ist eine ungewogene Linearkombination und die Variable $Z_2 = 0{,}8X + 0{,}2Y$ ist eine gewogene Linearkombination mit den Gewichten $b_1 = 0{,}8$ und $b_2 = 0{,}2$. Bestimmen Sie die Häufigkeitsverteilungen, (arithmetischen) Mittelwerte und Varianzen der Linearkombinationen Z_1 und Z_2!

Lösung 9.17:

Für die Häufigkeitsverteilungen der Linearkombinationen Z_1 und Z_2 erhält man:

z_{1h}	n_h	z_{2h}	n_h
4	8	2	8
5	2	2,2	2
6	9	3,6	9
7	6	4,2	3
8	5	4,4	3
		4,6	5
	30		30

Die Anwendung obiger Formeln für Mittelwert und Varianz einer Linearkombination führt zu:

ungewogene Linearkombination

$z_1 = \overline{x} + \overline{y} = 5{,}933 = 178/30$

$s_{z1}^2 = s_x^2 + s_y^2 + 2s_{xy} = 1{,}9956$

gewogene Linearkombination

$z_2 = 0{,}8\overline{x} + 0{,}2\overline{y} = 3{,}34667$

$s_{z2}^2 = 0{,}8^2 \cdot s_x^2 + 0{,}2^2 \cdot s_y^2 + 2 \cdot 0{,}8 \cdot 0{,}2 \cdot s_{xy} = 0{,}98.$

Man beachte, dass insbesondere die Formeln für die Varianzen der Linearkombinationen völlig verschieden sind von denen einer Aggregation im engeren Sinne (Abschn. b). Betrachtet man dagegen einfach die Merkmalssummen, so gilt $Sx_i\, n_i = 108$, $Sy_j\, n_j = 70$ und (wie in Abschn. b) $Sz_{1h} n_h = 178 = 108 + 70$.

d) Aggregation von Mittelwerten, Beziehungszahlen und Quoten, Struktureffekt und Standardisierung

Angenommen, für jede Teilgesamtheit j sei die Relation $Q_j = x_j/y_j$ (j=1,2,...,J) eine sinnvolle Beziehungszahl. Die für die Gesamtmasse definierte Beziehungszahl Q der Summen $X = Sx_j$ und $Y = Sy_j$ steht in folgender Beziehung zu den Teilbeziehungszahlen Q_j.

$$(9.22) \quad Q = \frac{X}{Y} = \sum Q_j g_{yj} \quad \text{mit} \quad g_{yj} = \frac{y_j}{\sum y_j} = \frac{y_j}{Y}$$

danach ist Q das **arithmetische Mittel** der Q_j, gewogen mit den Größen g_j, die Ausdruck der Struktur der Nennergröße sind.
Man kann Q auch darstellen als **harmonisches Mittel** der Q_j mit den Gewichten $g_{xj} = x_j /\Sigma x_j = x_j/X$, die die Struktur der Zählermasse darstellen

$$(9.23) \quad \frac{1}{Q} = \frac{Y}{X} = \sum \frac{1}{Q_j} g_{xj} \quad \text{mit } g_{xj} = \frac{x_j}{\sum x_j} = \frac{x_j}{X} \quad .$$

Wie leicht zu sehen ist, handelt es sich bei g_{xj} und g_{yj} jeweils um $\bar{x}$ (auf die Summe 1) normierte Gewichte.

Gl. 9.22 ist der Ausgangspunkt für eine Betrachtung

1. des bereits behandelten Struktureffekts und Simpson-Paradoxons (vgl. Def. 9.2) und
2. der Standardisierung von Beziehungszahlen.

Beispiel 9.4 demonstrierte den Struktureffekt von Beziehungszahlen. Es zeigte sich, dass die rohe Todesrate (d.h. die Todesrate für die Gesamtheit) nach Gl. 9.22 ein gewogenes arithmetisches Mittel der altersspezifischen Todesraten (d.h. der Größen Q_j für die beiden Alters-Teilgesamtheiten) darstellt. Vergleicht man rohe Todesraten, die unter Verwendung der gleichen Gewichte (also von Standardgewichten) errechnet sind, so spricht man von standardisierten Todesraten.

Def. 9.5: Struktureffekt, Standardisierung

Nach Gl. 9.22 ist eine aggregierte (für die Gesamtmasse errechnete) Beziehungszahl Q = X/Y das gewogene arithmetische Mittel der Teil-Beziehungszahlen $Q_j = x_j /y_j$ (j=1,1,...,J)

$$(9.22) \quad Q = \sum Q_j g_{yj} \quad .$$

Daraus folgt: Zwei Beziehungszahlen Q_A und Q_B für Gesamtheiten A und B, die sich jeweils in J Teilmassen gliedern lassen, können sich unterscheiden aufgrund unterschiedlicher

a) Teil-Beziehungszahlen Q_{Aj}, Q_{Bj}
b) Gewichte der Nennermasse g_{Ayj}, g_{Byj} .

Die Unterschiedlichkeit aufgrund von a) gilt als "echter" Unterschied, diejenige aufgrund von b) wird als Struktureffekt gedeutet. Um die echten Unterschiede herauszuarbeiten, vergleicht man nicht Q_A mit Q_B, sondern

$$(9.24) \quad Q_A^* = \sum Q_{Aj}\, g_j^* \quad \text{mit} \quad Q_B^* = \sum Q_{Bj} g_j^* ,$$

d.h. man vergleicht Beziehungszahlen, die unter Zugrundelegung der gleichen Gewichte (Standardgewichte) g_j^* berechnet sind. Die Größen Q^* heißen dann standardisierte Beziehungszahlen.

Bemerkungen zur Def. 9.5:

1. **Standardisierte und erwartungsgemäße Verhältniszahlen**
   So wie Verhältniszahlen bei gleicher Gewichtung (gleicher Struktur) verglichen werden können (d.h. standardisierte Zahlen) können auch Verhältniszahlen bei gleicher Sachkomponente verglichen werden, also etwa Q_A^0 mit Q_B^0 gem. Übersicht 9.6. Man spricht auch von "erwartungsgemäßen" Verhältniszahlen.

2. **Zerlegung in den Struktureffekt und den echten Effekt**
   Der Struktureffekt kann wie folgt dargestellt werden:

   $Q^D = Q_A - Q_B$

   ist die Differenz der globalen (Gesamt)-Beziehungszahlen oder Quoten und

   $Q_j^D = Q_{Aj} - Q_{Bj}$

   ist die Unterschiedlichkeit der Teil-Beziehungszahlen $1 \le j \le J$ oder (**Q_j,^{D} = Quotendifferenzen**).

   $G_j^D = g_{Ayj} - g_{Byj}$

   gibt die Unterschiedlichkeit der Gewichte wieder (**G_j^D = Gewichtsdifferenzen**).
   Im folgenden soll der Einfachheit halber das Subskript y weggelassen werden, also $G_j^D = g_{Aj} - g_{Bj}$. Dann gilt:

$$(9.25) \quad Q^D = \sum Q_{Aj} G_j^D + \sum g_{Bj} Q_j^D = S + E$$

Der erste Summand S ist Ausdruck der Unterschiedlichkeit der Gewichte (G_j^D). Er ist also der Struktureffekt, ausmultipliziert

(9.25a) $S = \sum Q_{Aj} g_{Aj} - \sum Q_{Aj} g_{Bj}$.

Der zweite Summand E ist Ausdruck der unterschiedlichen Teil-Beziehungszahlen oder Teilquoten (Q_j^D)bei gleicher Gewichtung mit g_{Bj}, also

(9.25b) $E = \sum Q_{Aj} g_{Bj} - \sum Q_{Bj} g_{Bj}$,

und somit ein Maß für den "echten" Unterschied.

Übersicht 9.6: Struktureffekt und echter Effekt bei Verhältniszahlen

(am Beispiel von Beziehungszahlen Q = X/Y)

	zu vergleichende Größen	Unterschiedlichkeit ist Ausdruck von
unbereinigter Vergleich	$Q_A = \sum_j Q_{Aj} g_{Aj}$ mit $Q_B = \sum_j Q_{Bj} g_{Bj}$ (j=1,2,..,n)	Struktureffekt **und** echter Effekt (Gl. 9.25)
standardisierte Verhältniszahlen	Q_A^* mit Q_B^* gem. Gl. 9.24; gleiche Strukturkomponente g_j^*; ungleiche Sachkomponenten Q_{Aj}, Q_{Bj}	einem echten Unterschied (bei $g_j^* = g_{Bj}$ ist die Differenz $Q_A^* - Q_B^* = E$ gem. Gl. 9.25b)
erwartungsgemäße Verhältniszahlen	$Q_A^0 = \sum Q_j^0 g_{Aj}$ mit $Q_B^0 = \sum Q_j^0 g_{Bj}$ gleiche Sachkomponente Q_A^0, ungleiche Strukturkomponenten g_{Aj},g_{Bj}	einem Strukturunterschied (wenn $Q_j^0 = Q_{Aj}$ ist die Differenz $Q_A^0 - Q_B^0 = S$ gem. Gl. 9.25a).

3. Standardisierung

Der Unterschied zwischen Q und Q^*, der nichtstandardisierten und der standardisierten Beziehungszahl, ist eine mit den Teil-Beziehungszahlen gewogene Summe der Differenzen zwischen den empirischen Gewichten g_j und den Standardgewichten g_j^*

$$(9.26) \quad Q - Q^* = \sum Q_j (g_j - g_j^*) ,$$

wobei die Differenzen wie G_j^D in Gl. 9.25 Unterschiede im Wägungs-

schema wiedergeben, so dass man die Differenz zwischen einer nicht-standardisierten Beziehungszahl Q und der entsprechenden standardisierten Beziehungszahl Q^* als Ausdruck des Struktureffekts deuten kann.

4. **Simpson-Paradoxon**

Das in Def. 9.2 beschriebene Simpson-Paradoxon, wonach z.B. ein Mittelwert für eine Gesamtheit A größer sein kann, als für eine andere B, obgleich er in allen Teilgesamtheiten von A kleiner ist als in denen von B, ergibt sich aus Gl. 9.25a (in Analogie zu Gl. 9.25):

$$(9.25a)\quad \bar{x}^D = \Sigma \bar{x}_{Aj} G_j^D + \Sigma g_{Byj} M_j^D \ .$$

Danach kann sehr wohl $\bar{x}^D = \bar{x}_A - \bar{x}_B > 0$ sein, obgleich für alle j gilt $M_j^D = \bar{x}_j - \bar{x}_j < 0$ (die Mittelwertdifferenzen M_j^D treten in Gl. 9.25a an die Stelle der Quotendifferen-zen Q_j^D von Gl. 9.25).

Beispiel 9.18:

Bekanntlich ist die Lohnquote der Anteil der Bruttoeinkommen aus unselbständiger Arbeit (L_t) am Volkseinkommen (Y_t). Dass diese Quote im Zeitablauf zunimmt ist schon deshalb zu erwarten, weil der Anteil der unselbständig Beschäftigten an den Beschäftigten zunimmt. Neben anderen "Bereinigungen" wird deshalb häufig eine Standardisierung dergestalt vorgenommen, dass man die Lohnquote berechnet, wie sie wäre, wenn sich dieser Anteil gegenüber einer Basisperiode (z.B. 1960) nicht verändert hätte. Wie ist eine solche Berechnung durchzuführen und was ist hieran kritisch anzumerken?

Lösung 9.18:

a) Berechnung: Es sei $Q_t = L_t/Y_t$ die "Lohnquote" zur Zeit t. Dann ist bei einer Anzahl U_t von Unselbständigen und E_t von Erwerbstätigen $l_t = L_t/U_t$ eine Art "Durchschnittslohn" und entsprechend $y_t = Y_t/E_t$ eine Art gesamtes (Arbeits- und Profit-) "Durchschnittseinkommen" für die Periode t. Es gilt also, dass die Lohnquote das Produkt einer "Pro-Kopf-Lohnquote" $q_t = l_t / y_t$ und einer Unselbständigen- oder "Abhängigen"-quote $a_t = U_t / E_t$ ist, denn:

$$Q_t = \frac{L_t}{Y_t} = \frac{l_t}{y_t} \cdot \frac{U_t}{E_t} = q_t a_t \quad \text{(unbereinigte Lohnquote).}$$

Die "bereinigte" (mit der Beschäftigtenstruktur des Basisjahres 0 gewogene, d.h. standardisierte) Lohnquote ist dann:

$$Q_t^* = \frac{l_t U_0}{y_t E_0} = q_t a_0 \quad \text{mit } q_t = \frac{l_t}{y_t} = \frac{L_t}{U_t}\frac{Y}{E_t}. \quad \text{(bereinigte Lohnquote)}$$

Abb. 9.6 stellt die unbereinigte Lohnquote Q_t (durchgezogene Linie) und die mit der Struktur a_o von 1960 bereinigte Lohnquote $Q_t,^*$ (gestrichelte Linie) dar.

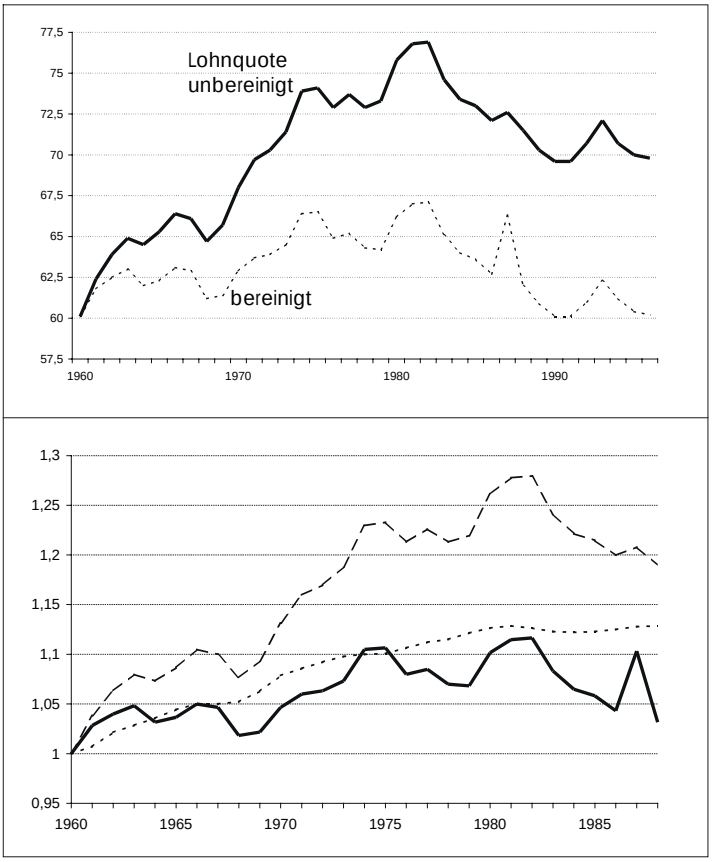
Offenbar ist

- $Q_t/Q_t^* = a_t/a_0$, d.h. die Veränderung (Messzahl) der Abhängigenquote (mittlere Linie in ***Abb. 9.7***) oder die Differenz $Q_t - Q_t^* = q_t(a_t - a_0)$ ist Ausdruck des Struktureffekts. Demgegenüber gilt
- $Q_t^*/Q_0 = q_t/q_0$ (untere durchgezogene Linie in Abb. 9.7) oder die Differenz $Q_t^* - Q_o = a_0(q_t - q_0)$ als echte Zunahme der Lohnquote (denn $Q_0^* = Q_0$).
- Der Gesamteffekt $Q_t/Q_0 = (Q_t,^*/Q_0)(Q_t/Q_t,^*) = (q_t/q_0)(a_t/a_0)$ ist dann das Produkt der beiden Effekte (obere gestrichelte Linie in Abb. 9.8), bzw. die Differenz $Q_t - Q_0 = (Q_t - Q_t,^*) + (Q_t,^* - Q_0) = q_t(a_t - a_0) + a_0(q_t - q_0) = S + E$ ist die Summe der beiden Effekte (die letzte Gleichung ist analog Gl. 9.25 gebildet).

Die Daten (Abb. 9.6,7) zeigen, dass die Abnahme der Lohnquote v.a. Ergebnis einer echten Veränderung der Verteilungsrelation ist. Das wird demonstriert am Vergleich von unbereinigter und bereinigter (standardisierter) Lohnquote (Abb. 9.6) und an der Abnahme von q_t/q_0 (untere Linie in Abb. 9.7). Die Abnahme der Lohnquote ist dagegen nicht auf einen Struktureffekt (a_t/a_0) zurückzuführen. Letzterer hätte eher im Gegenteil eine Zunahme der Lohnquote bewirken müssen.

b) Der kritische Punkt bei dieser Art von Betrachtung, die in der Statistik sehr verbreitet ist (so sind z.B. Preisindizes Verhältnisse von standardisierten Ausgaben), ist dass die Pro-Kopf-Lohnquote $q_t = l_t/y_t$ sich wahrscheinlich nicht so entwickelt hätte, wie sie sich tatsächlich entwickelt hat, wenn die Abhängigenquote konstant geblieben wäre. Es wird also eine Unabhängigkeit der Einkommensentwicklung von der Beschäftigtenstruktur unterstellt, die mit Sicherheit (was sich auch theoretisch begründen läßt) nicht gegeben ist. Außerdem ist die Beschäftigtenstruktur nicht der einzige Faktor, der die echte Lohnzunahme überlagert. Ein anderer Effekt ist z.B. die Arbeitszeitverkürzung. Danach wäre es auch gerechtfertigt, die Lohnquote nicht nur mit einer Standardstruktur hinsichtlich der Beschäftigten (unselbständig oder nicht), sondern auch hinsichtlich der Arbeitszeit zu standardisieren, d.h. die Lohnquote zu berechnen, die sich ergeben hätte, wenn die Produktivitätssteigerung voll in Lohnsteigerungen und nicht z.T. auch in Arbeitszeitverkürzungen weitergegeben wäre.

Abb. 9.6 (oben) und 9.7 (unten): Erläuterungen in der Lösung 9.18



e) Aggregation von Messzahlen und Wachstumsraten

1. Aggregation von Messzahlen

Die Variable Y_t, deren Messzahlen $m_{0t} = Y_t/Y_0$ zu betrachten sind, soll ein Aggregat darstellen, d.h. es gelte zu jeder Periode t jeweils $Y = Sy_j$. Wie man leicht sieht, gilt

$$(9.22a) \qquad m_{0t} = \frac{Y_t}{Y_0} = \frac{y_{1t}}{y_{10}}\frac{y_{10}}{Y_0} + \frac{y_{2t}}{y_{20}}\frac{y_{20}}{Y_0} + \ldots = \Sigma\, m_{0t}^{j}\, g_{j0}$$

wobei die Gewichte g_{j0} jeweils die Anteile y_{j0}/Y_0, also der Teilbeträge y_j am Gesamtbetrag Y zur Basiszeit sind. Gl. 9.22a ist in Analogie zu Gl. 9.22 zu sehen und man kann analog zu Gl. 9.23 auch die Messzahl eines Aggregats als harmonisches Mittel darstellen mit Gewichten g_{jt}, die sich auf die Berichtszeit beziehen, also

$$(9.23a) \quad \frac{1}{m_{0t}} = m_{t0} = \sum_j \frac{1}{m_{0t}^j} g_{jt} = \sum_j m_{0t}^j g_{jt} \quad .$$

Gl. 9.26 bzw. 9.25a ist entsprechend bei Differenzen anzuwenden.

Beispiel 9.19:

Es gilt die Gleichung: Inlandsprodukt (P) plus Außenbeitrag (A) ist das Sozialprodukt Y. Für ein Land mögen die folgenden Zahlen gelten (in Mrd DM):

Jahr	P	A	Y
1970	660	15	675
1990	2000	120	2120

Man verifiziere Gl. 9.22a!

Lösung 9.19:

Die Messzahlen des Inlandsprodukts und Außenbeitrags betragen $m_{0t}^P = 3{,}0303$ und $m_{0t}^A = 8$. Die Gewichte betrugen zur Basiszeit 660/675 und 15/675. Die Messzahl $m_{0t}^Y = 2120/675 = 3{,}14074$ ist ein hiermit gewogenes arithmetisches Mittel der einzelnen Messzahlen.

2. Aggregation von Wachstumsfaktoren und Wachstumsraten

a) Aggregation

Es ist unmittelbar einsichtig, dass für Wachstumsfaktoren Gl. 9.22a analog gilt, wobei der Wachstumsfaktor w_t des Aggregats $Y=\Sigma y_j$ als gewogenes arithmetisches Mittel der Wachstumsfaktoren der Summanden j =1,2,...J darzustellen ist

$$(9.22b) \quad w_t = \frac{Y_t}{Y_{t-1}} = \Sigma(y_{jt}/y_{j,t-1}) g_{j,t-1} = \Sigma w_{jt} g_{j,t-1}$$

wobei die Gewichte die Anteile $g_{j,t-1} = y_{j,t-1}/Y_{t-1}$ der Summanden am Gesamtmerkmalsbetrag in der Vorperiode sind. Natürlich kann man auch Gl. 9.23a analog anwenden, wenn die Gewichte $g_{j,t}$ die entsprechenden Anteile
zur Zeit t darstellen.

Ausgehend von den Wachstumsfaktoren w können mit r = w - 1 entsprechende Aussagen über die Wachstumsraten (r) gemacht werden.

$$(9.27) \quad r_t = \Sigma r_{jt} g_{j,t-1}.$$

b) Struktureffekt

Ist eine Größe Z eine Summe von X und Y, so gilt für die **Wachstumsrate** der Linearkombination Z nach Gl. 9.27:

$$r_z = r_x g_x + r_y g_y \quad \text{mit} \quad g_x = \frac{X_{t-1}}{Z_{t-1}} \quad \text{und} \quad g_y = \frac{Y_{t-1}}{Z_{t-1}}.$$

Entsprechend kann man die **Differenz von Wachstumsraten** einer Linearkombination zerlegen:

Die Differenz aufeinanderfolgender Wachstumsraten $r_{zt} - r_{z,t-1}$ ist darstellbar als die folgende Summe:

der erste Summand $[(r_{xt} - r_{x,t-1})g_{x,t-1} + (r_{yt} - r_{y,t-1})g_{y,t-1}]$ ist Ausdruck des Wachstumseffekts und der zweite Summand $[r_{x,t-1}(g_{x,t-1} - g_{x,t-2}) + r_{y,t-1}(g_{y,t-1} - g_{y,\,t-2})]$ ist Ausdruck des Struktureffekts.

Dies ist eine zu Gl. 9.25 analoge Betrachtung mit Wachstumsraten.

Man kann ferner die Wachstumsrate einer Linearkombination, etwa von Z = X + Y für den Zeitraum von 0 bis t zerlegen in einen Teil, der auf den isolierten Einfluß von X und einen Teil der auf den isolierten Einfluß von Y zurückgeführt werden kann. Mit den Gewichten $g_x = X_0/Z_0$ und $g_y = Y_0/Z_0$ erhält man

$$(9.28) \quad r_{z,t} = \frac{Z_t}{Z_0} - 1 = (g_x w_x + g_y) - 1 + (g_y w_y + g_x) - 1.$$

Interpretation:
Die erste Klammer stellt den isolierten Einfluß von X auf die Summe Z dar, denn $g_x w_x + g_y = (X_t + Y_0)/Z_0$. Entsprechend ist der zweite Klammerausdruck der isolierte Einfluss
von Y auf die Summe Z: $g_y w_y + g_x = (Y_t + X_0)/Z_0$. Berücksichtigt man, dass gilt $g_x + g_y =$
1 und r = w - 1, so ist Gl. 9.28 nur eine andere Darstellung von Gl. 9.27.
Betrachtungen dieser Art haben seinerzeit für die Statistik-Ausbildung von Ökonomen in der DDR eine große Bedeutung gehabt. In dem bis zur "Wende" 1989 als "Bibel" geltenden Statistik-Lehrbuch von Donda et al. wird dieser Zusammenhang über viele Seiten als "additive Faktorenanalyse mit Indizes" (d.h. Messzahlen) ausgeführt. Das folgende Beispiel 9.20 zeigt eine Anwendung dieser Art.

Beispiel 9.20:

Die Kosten Z eines Betriebes setzen sich aus Material- (X) und Arbeitskosten (Y) zusammen. Es mögen die folgenden Zahlen gelten:

Jahr	X	Y	Z
1970	75	25	100
1980	80	40	120
1990	100	80	180

a) Wie lässt sich die Zunahme der Kosten von 1980 bis 1990 um 50% von 120 auf 180 auf zunehmende Arbeits- und zunehmende Materialkosten verteilen?

b) Man zerlege die **Veränderung** der Wachstumsrate von 20% (1970-80) auf 50% (1980-90) in Wachstums- und Struktureffekt.

Lösung 9.20:

a) Nach Gl. 9.28 gilt mit $g_x = 2/3$ und $g_y = 1/3$. Die Zunahme von 50% ist dann: $0,5 = (2/3 \cdot 10/8 + 1/3) - 1 + (1/3 \cdot 2 + 2/3) - 1$. Die erste Klammer führt zu $7/6 = (100 + 40)/120$ und ist Ausdruck der Erhöhung der Materialkosten von 80 auf 100. Der zweite Klammerausdruck beträgt 4/3 und dies ist zugleich $(80 + 80)/120$, also die Wirkung der Erhöhung der Lohnkosten von 40 auf 80. Man kann also sagen, dass die Zunahme der Gesamtkosten um 50% zurückzuführen ist auf 1/6 Material- und 1/3 Arbeitskosten (denn $1/6 + 1/3 = 1/2$).

b) Die Differenz der Wachstumsraten von 0,5 und 0,2 (also 0,3) läßt sich zerlegen in einen Wachstumseffekt $(20/80 - 5/75)(2/3) + (40/40 - 15/25)(1/3) = 23/90$
und in einen Struktureffekt $(5/75)(2/3 - 3/4) + (15/25)(1/3 - 1/4) = 4/90$ [$23/90 + 4/90 = 0,3$].
Das Gewicht von X war zur Zeit t-1 (1980): 2/3, dagegen zur Zeit t-2 (1970): 3/4. Es ist von 1980 bis 1990 beständig zurückgegangen und entsprechend ist das Gewicht der Arbeitskosten gestiegen (von 1/4 1970 über 1/3 im Jahr 1980 bis 4/9 im Jahre 1990). Dies ist für den Struktureffekt verantwortlich, dessen Bedeutung mit 4/90 ganz wesentlich geringer ist als der Wachstumseffekt (23/90).

Kapitel 10: Indexzahlen

1. Gegenstand und Bedeutung von Indexzahlen

a) Definition von Indexzahlen

Indexzahlen spielen vor allem in ökonomischen Anwendungen der Statistik eine große Rolle. Auch Nichtfachleuten ist z.B. der **Preis**index für die Lebenshaltung (nicht Lebenshaltungs**kosten**index!) als Maß der allgemeinen Teuerung (der Kaufkraft des Geldes) bekannt. In diesem Kapitel geht es um die methodischen Probleme von Indexzahlen. Es wird zunächst eine möglichst allgemein gehaltene Definition versucht, die anschließend kommentiert wird.

Indexzahlen sind Maßzahlen für den Vergleich einer Gesamtheit von Erscheinungen. Der summarische Charakter unterscheidet Indexzahlen von Verhältniszahlen (insbesondere Messzahlen). Indizes sind Maße der aggregierten Veränderung.

1. Gegenstand des Vergleichs, Indizes und Messzahlen

Indizes werden also vorwiegend zur Darstellung einer zeitlichen Entwicklung verwendet, seltener zum räumlichen Vergleich. Indizes und Messzahlen

- haben gemeinsam den (zeitlichen) Vergleich und die dem Vergleich verschiedener Erscheinungen dienende Bezugnahme auf einen Basiswert (Wert zur Basisperiode), der gleich 100 gesetzt wird;

- unterscheiden sich dadurch, dass Messzahlen die Entwicklung einer Einzelerscheinung (z.B. der Preis einer Ware), Indizes dagegen die einer Gesamtheit von Erscheinungen (z.B. die Preisentwicklung für einen aus n Waren bestehenden "Warenkorb") darstellen.

Demzufolge sind auch die meisten gebräuchlichen Indexformeln (gewogene) Mittelwerte von Messzahlen und. Früher sprach man von Indizes und Generalindizes, so wie man heute zwischen Messzahlen und Indexzahlen unterscheidet.

2. Die "Gesamtheit" (das Aggregat)

der Erscheinungen ist kategorial (qualitativ) abgegrenzt. Es ist kennzeichnend für einen Index, dass das Kriterium, nach dem Einzelerscheinungen (z.B. Preismesszahlen von verschiedenen Gütern und Dienstleistungen) zusammengefaßt werden, qualitativer Art ist (z.B. Einzelhandelspreise, Preise für die Lebenshaltung, Grundstoffpreise usw.).

3. Die Art der Einzelerscheinungen

(bzw. im Falle des zeitlichen Vergleichs: der Meßziffern) entscheidet über die Art des Index. Ein Preisindex mißt als Mittelwert von Preismesszahlen die durchschnittliche Preisentwicklung eines Aggregats. Entsprechend mißt ein Mengenindex eine durchschnittliche Mengenentwicklung.

4. Index als Funktion

Ein ungewogener Preisindex P_{0t} zur Basis 0 und für die Berichtsperiode t (mit einem "Warenkorb" von n Waren) ist eine Abbildung von zwei n-dimensionalen reellen Datenvektoren $[p_{1t} ... p_{nt}]$ und $[p_{10} ... p_{n0}]$ in die reelle Zahl P_{0t} (= Preisindex), wobei p_{it} der Preis der i-ten Ware zur Berichtszeit und p_{i0} zur Basiszeit ist (mit i = 1,2,...,n). Ein (Preis-) Index ist also eine Funktion von (Preis-) Vektoren, die bestimmten, in der Indextheorie (Abschn. 3) definierten formalen und ökonomischen Kriterien genügt. Das Hauptproblem jeder Indexkonstruktion ist es also, die für zwei Perioden [Zeitpunkte oder -intervalle] 0 und t, bzw. beim regionalen Vergleich die für zwei Regionen gegebenen Vektoren sinnvoll durch eine Zahl zu vergleichen.

5. Der Begriff "Index"

wird auch verwandt für absolute Größen, nicht nur für Verhältniszahlen. Häufig wird darunter auch ein Mittelwert von nach irgendeinem Schema für verschiedene Einzelerscheinungen (Variablen) vergebenen Punktzahlen verstanden, so z.B. bei einem Level of Living Index (als Maß der Wohlfahrt), Status-Index (Messung der sozialen Schichtung) oder Konjunkturindex. Oft wird der Begriff "Index" auch im allgemeinen Sinne eines "Indikators" oder einer Maßzahl benutzt.

b) Heuristische Einführung der Preisindex-Formeln

Im folgenden soll gezeigt werden, wie eine einfache Überlegung in fünf Gedankenschritten zur Preisindexformel von Laspeyres führt. Dabei werden auch einige andere (ältere) Preisindexformeln sowie axiomatische Forderungen an eine "sinnvolle" Indexformel vorgestellt, die das Verständnis der folgenden Abschnitte erleichtern dürften.

1. Preisniveau als relative Größe

Bei der Messung eines Preisniveaus könnte man zunächst daran denken, einen Durchschnittspreis zu berechnen, etwa ein ungewogenes arithmetisches Mittel der einzelnen Preise

$$(10.1)\quad \overline{p}_t = \frac{\Sigma p_{it}}{n} \qquad (i = 1,2,...,n).$$

Als absolute Größe hat diese Zahl wenig Aussagefähigkeit, zumal die Festlegung des "Warenkorbes" der n Waren entscheidend ist dafür, wie groß der "Durchschnittspreis" ist. Ein Preisniveau kann entgegen dem Eindruck, den das Wort "Niveau" erweckt, keine absolute, sondern nur eine relative Größe sein.
Wie problematisch eine Bezugnahme auf absolute Größen ist, mag auch das folgende Zitat aus einem Rundfunkkommentar verdeutlichen (Hessischer Rundfunk 9.12.1970):

"Die 2-Personen-Haushalte von Rentnern und Sozialhilfeempfängern, ... müssen ... zusätzlich mindestens 20 DM im Monat aufwenden. Die Ausgaben für den privaten Verbrauch eines 4-Personen-Arbeitnehmer-Haushalts mit mittlerem Einkommen, ... haben sich bereits mit 60 DM verteuert. Bei den 4-Personen-Haushalten von Beamten und Angestellten mit höherem Einkommen...macht die Verteuerung sogar über 100 DM im Monat aus."
Das legt den Schluß nahe, Haushalte mit höherem Einkommen hätten unter der Inflation mehr zu leiden als solche mit niedrigem Einkommen. Es wird dabei aber völlig vergessen, dass die Warenkörbe der verschiedenen Haushaltstypen auch zur Basiszeit unterschiedlich teuer waren. Das Bemühen, zu einer evtl. einfacher verständlichen Aussage zu gelangen als die für Laien vielleicht komplizierte Aussage eines Indexes ist zwar anzuerkennen, aber der dadurch vermittelte Eindruck ist einfach falsch. Es stimmt nicht, dass Haushalte mit höherem Einkommen unter der Inflation mehr zu leiden haben als solche mit niedrigem Einkommen. Wie die entsprechenden Preisindizes zeigen, ist vielmehr das Gegenteil der Fall.

Auch dieses Beispiel zeigt, dass ein Preisniveau nur eine relative Größe sein kann. Es muss unter Berücksichtigung des Prinzips des reinen Preisvergleichs, d.h. eines Vergleichs unter sonst gleichen Umständen (gleiche Mengen, Qualitäten usw.) die aggregierte Preisveränderung gemessen werden. Das ist die Aufgabe eines Preisindexes.

2. Preisindex als Messzahl von Durchschnittspreisen

Ein zeitlicher Vergleich des Durchschnittspreises bezogen jeweils auf den gleichen Warenkorb durch einen Preisindex zur Basis 0 und für die Berichtsperiode t könnte erfolgen durch:

$$(10.2)\quad P_{0t} = \frac{\overline{p}_t}{\overline{p}_0} = \frac{\Sigma p_{it}}{\Sigma p_{i0}} = P^D \qquad \text{(Dutot-Preisindex)}$$

Hierbei ist $\overline{p}_0$ analog zu $\overline{p}_t$ gem. Gl.10.1 definiert.
Setzt man die Basisperiode 100 statt 1 (es sei 0 etwa das Basisjahr 1990, was dann mit der bekannten, aber an sich unsinnigen Schreibweise "1990 = 100" bezeichnet wird), dann ist diese, wie auch alle nachfolgenden Indexformeln mit 100 zu multiplizieren.

Aber auch Gl. 10.2, eine 1738 von Dutot vorgeschlagene Indexformel kann kein sinnvolles Maß für ein Preisniveau sein, weil dieser Index gegen die axiomatische Forderung der Kommensurabilität (Axiom P5 vgl. Abschn. 3b) verstößt. Der Dutot-Index kann unterschiedliche Preissteigerungen ausweisen, je nachdem, ob der Preisnotierung (bei m < n Waren) beispielsweise Pfund- oder Kilopreise zugrunde liegen. Angenommen der Warenkorb besteht aus nur zwei Waren (also n = 2) und die Ware 1 werde

1. in Kilopreisen oder aber
2. in Pfundpreisen

notiert. Es ist offensichtlich, dass in der Regel

$$P^{D1} = \frac{p_{1t} + p_{2t}}{p_{10} + p_{20}} \quad \text{nicht identisch mit} \quad P^{D2} = \frac{½p_{1t} + p_{2t}}{½p_{10} + p_{20}}$$

sein wird. Allgemein gilt:
Preissummen und Preisdurchschnitte sind nicht unabhängig von der Maßeinheit der Mengen, auf die sich die Preisnotierungen zu beiden Zeiten 0 und t beziehen. Ein Index, der der Forderung der Kommensurabilität genügen soll kann also nicht eine Messzahl von Durchschnitten sein (wohl aber - was allerdings etwas ganz anderes ist - ein Durchschnitt von Messzahlen, vgl. Nr. 4).

3. Preisindex als Messzahl von Durchschnittswerten

Ein Durchschnittswert von n Waren ist definiert als

$$(10.3)\quad \bar{w}_t = \frac{\Sigma p_{it} q_{it}}{\Sigma q_{it}} \qquad i = 1,2,...,.n$$

wobei q_{it} die Menge der i-ten Ware zur Berichtszeit darstellt. Der Durchschnittswert $\bar{w}_0$ für die Basisperiode ist entsprechend definiert.
Gegen einen mit Durchschnittswerten gebildeten Preisindex

$$(10.4)\quad P^W = \frac{\bar{w}_t}{\bar{w}_0}$$

also gegen eine Messzahl von Durchschnittswerten sind folgende Einwände zu erheben:

- Die Summe Σq_{it}, bzw. Σq_{i0} der Mengen ist häufig nicht definiert. Sie setzt eine gemeinsame Mengeneinheit (z.B. Kilogramm, Stück, Liter, m^2 usw.) voraus, was nur möglich ist bei einer Gruppe von untereinander ähnlichen Waren (wenn z.B. wie in der Außenhandelsstatistik verschiedene Waren unter einer Warennummer zusammengefaßt werden). Für einen Preisindex für die Lebenshaltung wäre diese Formel somit schon deswegen nicht geeignet, weil bei ca. 750 Verbraucherpreisen praktisch alle Mengenarten vorkommen, die überhaupt nicht zu einer Summe Σq_i addiert werden können.
- Der Index P^W der Gl. 10.4 ist nicht darstellbar als Mittelwert von Preismesszahlen p_{it}/p_{i0}, wie dies beim Dutot-Preisindex der Fall ist. Aus Gl. 10.4 erhält man

$$P^W = \frac{\Sigma p_{it} q_{it}}{\Sigma q_{it}} \cdot \frac{\Sigma q_{i0}}{\Sigma p_{i0} q_{i0}} = \sum \frac{p_{it}}{p_{i0}} \frac{p_{i0} q_{it}}{\Sigma p_{i0} q_{i0}} \frac{\Sigma q_{i0}}{\Sigma q_{it}}$$

  Wie man leicht sieht, addieren sich die Gewichte $\frac{p_{i0} q_{it}}{\Sigma p_{i0} q_{i0}} \frac{\Sigma q_{i0}}{\Sigma q_{it}}$ nicht notwendig zu 1, so dass der Preisindex P^W größer als die größte oder kleiner als die kleinste Preismesszahl sein kann. Auch deshalb kann P^W nicht als sinnvolle Preisindexformel akzeptiert werden.

Man beachte aber, dass

- der erste Faktor dieser Gewichte, nämlich $p_{i0}q_{i0}/\Sigma p_{i0}q_{i0}$ die Ausgabenanteile zur Basiszeit darstellt und diese Größen sich sehr wohl zu 1 addieren (i = 1,2,...,n) und dass
- sich auch der Dutot-Index als ein (mit Anteilen an der Preissumme zur Basiszeit) gewogenes arithmetisches Mittel der Preismesszahlen darstellen läßt, so dass gegen P^D nicht der Einwand erhoben werden

kann, dass P^D größer als die größte oder kleiner als die kleinste Preismesszahl werden kann.

4. Preisindex als Durchschnitt von Preismesszahlen

Um der Kommensurabilität zu genügen und auch sicherzustellen, dass ein Preisindex nicht größer als die größte oder kleiner als die kleinste der n Preismesszahlen werden kann, ist ein Preisindex als Mittelwert von Preismesszahlen zu konstruieren, etwa gemäß der Formel von Graf Carli

(10.5) $P^C = \frac{1}{n} \sum \frac{p_{it}}{p_{i0}}$ (Preisindex von Carli, 1764).

Dieser Preisindex ist ein ungewogenes arithmetisches Mittel der Preismesszahlen. Es ist, wie gesagt, ein großer Unterschied, ob ein Index als Messzahl von Durchschnitten (wie P^D und P^W) oder als Durchschnitt von Messzahlen konstruiert ist. Anstelle des arithmetischen Mittels können auch andere Mittelwerte zur Definition eines Indexes herangezogen werden, etwa das geometrische Mittel, wie dies Jevons vorschlug:

(10.6) $P^J = [(p_{1t}/p_{10})...(p_{nt}/p_{n0})]^{1/n}$.

5. Preisindex als gewogener Durchschnitt von Messzahlen

Gegen die Formeln von Carli oder Jevons, P^C oder P^J ist nur einzuwenden, dass keine Gewichtung vorliegt. Jede Preismesszahl wird als gleich "wichtig" betrachtet. Bei einem nach der Formel P^C oder P^J berechneten Preisindex für die Lebenshaltung erhielte also z.B. eine Mieterhöhung das gleiche Gewicht, wie die Preissteigerung bei Ölsardinen-Konserven, obgleich der Haushalt von einer Mieterhöhung ungleich mehr betroffen sein wird als von einer gleich großen Zunahme des Preises von Ölsardinen. Dass die Preissteigerung bei der Wohnungsmiete bedeutsamer ist als bei Ölsardinen liegt daran, dass der Haushalt in der Regel für die Miete wesentlich mehr ausgibt als für Ölsardinen. Es ist deshalb sehr sinnvoll, eine Gewichtung mit Ausgabenanteilen vorzunehmen. Die heutzutage für die Praxis wichtigsten Preisindizes sind jeweils mit Ausgabenanteilen gewogene Mittelwerte von Preismesszahlen:

1. Der Preisindex von Laspeyres P^L ist ein mit den Ausgabenanteilen der Basiszeit gewogenes arithmetisches Mittel der Preismesszahlen.
2. Der Preisindex von Paasche P^P ist ein mit den Ausgabenanteilen der Berichtszeit gewogenes harmonisches Mittel der Preismesszahlen.

zu 1:

Der Ausgabenanteil der i-ten Ware (i=1,2,...,n) beträgt zur Basiszeit $g_{i0} = p_{i0}q_{i0}/\Sigma p_{i0}q_{i0}$ (mit $g_{i0} \geq 0$ und $\Sigma g_{i0} = 1$).
Der Preisindex nach Laspeyres (P^L) ist definiert als arithmetisches Mittel der hiermit gewogenen Preismesszifferm, also als

$$P^L = \Sigma(p_{it}/p_{i0})g_{i0} \text{ oder}$$

$$(10.7)\quad P^L = \sum \frac{p_{it}}{p_{i0}} \frac{p_{i0}q_{i0}}{\Sigma p_{i0}q_{i0}} \qquad \text{Laspeyres-Preisindex Messzahlenmittelwertformel.}$$

Es ist leicht zu sehen, dass dies umgeformt werden kann zu einem Verhältnis von Preis-Mengen-Produkten (Aggregaten), die z.B. im Falle eines Verbraucherpreisindexes (wie der Preisindex für die Lebenshaltung oder der Einzelhandelspreisindex) Ausgaben darstellen (bei einem Erzeugerpreisindex entsprechend Einnahmen), wobei der "Laufindex" (das Subskript) i der Übersichtlichkeit halber meist weggelassen wird. Eine andere Darstellung von P^L ist also:

$$(10.8)\quad P^L = \frac{\Sigma p_{it}q_{i0}}{\Sigma p_{i0}q_{i0}} = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0} \qquad \text{Laspeyres-Preisindex Aggregatformel.}$$

Offensichtlich sind Gl. 10.7 und 10.8 formal identisch. Gl. 10.7 beschreibt jedoch deutlicher die praktische Vorgehensweise der laufenden (meist monatlichen) Berechnung von Laspeyres-Preisindizes in der amtlichen Statistik und zugleich die Vorzüge der Laspeyres-Formel für die Praxis:

- Jeweils monatlich neu zu bestimmen sind nur die n Preismesszahlen p_{it}/p_{i0} durch laufende monatliche Preisnotierungen;
- Solange das Basisjahr beibehalten wird, bleiben dagegen die Gewichte ($p_{i0}q_{i0}/\Sigma p_{i0}q_{i0}$), d.h. der "Warenkorb" konstant.

zu 2:

Der meist nur in seiner Aggregatformel (Gl. 10.9) bekannte Preisindex nach Paasche

$$(10.9)\quad P^P = \frac{\Sigma p_{it}q_{it}}{\Sigma p_{i0}q_{it}} = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t} \qquad \text{Paasche-Preisindex Aggregatformel}$$

ist, wie oben behauptet, als mit den Ausgabenanteilen zur Berichtszeit ($g_{it} = p_{it}q_{it}/\Sigma p_{it}q_{it}$) gewogenes harmonisches Mittel von Preismesszahlen darstellbar. Bekanntlich ist der reziproke Wert des harmonischen Mittels das arithmetische Mittel der reziproken Werte (hier der reziproken

Preismesszahlen, also der Größen p_{i0}/p_{it}). Mit den Gewichten g_{it} erhält man nämlich

$$\frac{1}{P^P} = \sum \frac{p_{i0}}{p_{it}} \frac{p_{it}q_{it}}{\Sigma p_{it}q_{it}} = \frac{\Sigma p_{io}q_{it}}{\Sigma p_{it}q_{it}} \quad .$$

Im Unterschied zur Laspeyres-Formel ist bei P^P der Warenkorb nicht konstant. Die Berechnung von P^P ist also aufwendiger als die von P^L. Ein Preisindex für die Lebenshaltung wird i.d.R. nach der Formel P^L berechnet. Die Bedeutung von P^P liegt dagegen auf einem anderen Gebiet (nämlich dem der Preisbereinigung).

Wie man sieht, führen einfache Überlegungen zu den in der Praxis der amtlichen Statistik besonders beliebten Indexformeln von Laspeyres und Paasche. Die Indexaussage von P^L und auch von P^P sollte ja auch intuitiv verständlich sein. Die Überlegungen zum Preisniveau sind entsprechend übertragbar auf andere Indizes (etwa auf Produktions-, Auftragseingangsindizes usw.).

Beispiel 10.1:

Der Warenkorb der Verbraucher eines Landes bestehe nur aus den vier Waren A,B,C und D. Gegeben seien die folgenden Preise und Mengen dieser Waren zur Basis- (0) und zur Berichtszeit (t):

	Preise (p)		Mengen (q)	
Ware	0	t	0	t
A	2	3	25	50
B	4	8	20	30
C	7	9	30	25
D	3	4	10	90

a) Es sollen zunächst nur die Preise beachtet werden: Berechnen Sie den Preisindex nach Dutot
   - mit den obigen Preisen
   - für den Fall, dass die Ware B in Pfund- statt bisher in Kilopreisen notiert wird.

b) Berücksichtigen Sie nun auch die Mengen: Berechnen Sie die Preis- und Mengenmesszahlen aller vier Waren, sowie die Durchschnittswerte zur Basis- und Berichtszeit und den Index P^W.

c) Berechnen Sie die Preisindizes P^L und P^P nach der Aggregatformel.

Lösung 10.1:

a) Bei den angegebenen Preisen gilt $\Sigma p_{i0} = 16$, so dass der Durchschnittspreis 4 beträgt und $\Sigma p_{it} = 24$ (Durchschnitt: 6). Der Preisindex nach Dutot ist dann $P^D = 6/4 = 24/16 = 1{,}5$. Wird Ware B in Pfund statt in Kilo notiert, so sind die Preise $p_{B0} = 2$ und $p_{Bt} = 4$. Die Preissummen sind dann $\Sigma p_{i0} = 14$ und $\Sigma p_{i0} = 20$ und der Dutot-Preisindex ist dann bei gleicher Preissteigerung wie bisher $P^D = 1{,}4286$ statt 1,5.

b) Man erhält die folgenden Zahlenergebnisse:

Ware	Messzahlen		Werte		Ausgabenanteile	
(i)	Preise	Mengen	0	t	0	t
A	1,5	2	50	150	0,135	0,154
B	2	1,5	80	240	0,216	0,246
C	1,286	2,5	210	225	0,568	0,231
D	1,333	2,25	30	360	0,081	0,369

Es gilt $\Sigma q_{i0} = 85$ und $\Sigma q_{it} = 195$, ferner für die Werte $\Sigma p_{i0}q_{i0} = 370$ und $\Sigma p_{it}q_{it} = 975$, so dass für die Durchschnittswerte gilt $\overline{w}_0 = 370/85 = 4{,}35$ und $\overline{w}_t = 975/195 = 5$, so dass man erhält $P^W = \overline{w}_t/\overline{w}_0 = 1{,}1486$. Der Index P^W ist damit kleiner als die kleinste Preismesszahl (diejenige der Ware C mit 1,286).

c) Für P^L erhält man $P^L = \Sigma p_{it}q_{i0}/\Sigma p_{i0}q_{i0} = 545/370 = 1{,}473$ (bzw. wenn das Basisjahr 100 gesetzt wird 147,3). Wie man leicht sieht ist P^L auch als Mittelwert der Preismesszahlen zu errechnen mit $P^L = 1{,}5 \cdot 0{,}135 + \ldots + 1{,}333 \cdot 0{,}081$ und $P^L = 1{,}473$ liegt "in der Mitte" der Preismesszahlen, die zwischen 1,286 und 2 schwanken. Entsprechend ist der Paasche Preisindex $P^p = \Sigma p_{it}q_{it}/\Sigma p_{i0}q_{it} = 975/665 = 1{,}466$. Man beachte, dass der Zähler von P^p und der Nenner von P^L bereits im Teil b) berechnet wurden. Hierbei handelt es sich nämlich um tatsächliche Ausgaben (oder allgemeiner: Werte), während die Größen $\Sigma p_{it}q_{i0}$ und $\Sigma p_{i0}q_{it}$ fiktive Ausgaben sind.

Beispiel 10.2:

Um seinen notleidenden staatlichen Dienstleistungsbetrieben finanziell auf die Sprünge zu helfen, plant ein Minister eine Gebührenerhöhung bei zwei von 20 Gebührenarten (A und B) und zwar um 50% (bei A) und um 80%

(bei B). Die Ausgabenanteile für die Dienstleistungen A und B waren bei den Konsumenten bisher (Basiszeit) 15% (bei A) bzw. 20% (bei B).

a) Wie groß ist der Preisindex nach Laspeyres?

b) Es ist nicht bekannt, wie die Verbraucher reagieren werden. Kann man deshalb keine Aussage über P^P machen, kann also P^P ganz beliebige Werte annehmen, oder kann der Preisindex nach Paasche einen bestimmten Höchstwert nicht über- und einen Mindestwert nicht unterschreiten?

Lösung 10.2:

a) Der Laspeyres-Preisindex ist hier nur mit der Messzahlen-Mittelwertformel (Gl.10.7) zu errechnen. Für die Preismesszahlen und für die Ausgabenanteile gilt nach obigen Angaben

Dienstleistung	Preismesszahl	Ausgabenanteil
A	1,5	0,15
B	1,8	0,20
alle übrigen	1,0	0,65

Somit ist $P^L = 1{,}5 \cdot 0{,}15 + 1{,}8 \cdot 0{,}2 + 0{,}65 = 1{,}235$.

b) Man kann sehr wohl einen Bereich für P^P angeben, denn auch der Paasche-Preisindex ist ein Mittelwert der Preismesszahlen. Wie immer die Ausgabenstruktur zur Berichtszeit aussehen mag, P^P kann nicht größer als 1,8 und nicht kleiner als 1 sein.

c) Kompromissformeln durch Mittelwertbildung

Bevor wir auf die Eigenschaften der Preisindexformeln von Laspeyres und Paasche näher eingehen und auch weitere Beispiele für die Anwendung dieser Formeln durchrechnen, soll hier gezeigt werden, wie das Thema "Indexzahlen" im besonderen Maße geeignet ist, eine statistische Methode mit mathematisch-formalen Überlegungen zu untersuchen (was in Abschn. 3 vertieft wird). Zwar wurden die meisten älteren Indexformeln aus anderen Überlegungen heraus entwickelt, es entstand jedoch schon früh eine formale Theorie der Indexzahlen (wobei meist Preisindizes im Vordergrund standen).

Gegenstand einer formalen Indexbetrachtung ist

1. die Suche nach "idealen" Mittelwerten und/oder Wägungsschemen,
2. der Versuch, "ideale" Indexformeln dadurch zu finden, dass man Mittelwerte von Indizes oder von Gewichten bildet und
3. die Entwicklung einer Axiomatik für Indexzahlen und die Suche nach Indexformeln, die solchen Axiomen genügen.

Auf den Punkt Nr. 3 soll in Abschnitt 3 eingegangen werden. Zu den ersten beiden Punkten dürften knappere Hinweise schon an dieser Stelle genügen:

zu 1:

Da Indizes stets Mittelwerte von Messzahlen sind und zwar in der Regel auch gewogene Mittelwerte, liegt der Gedanke nahe, verschiedene Arten von Mittelwerten und verschiedene Wägungsarten "auszuprobieren". Dabei bestehen auch Zusammenhänge zwischen den Mittelwerten und Wägungsarten, z.B. zwischen dem arithmetischen und dem harmonischen Mittel (vgl. Übers. 10.1 und 10.5).

Übersicht 10.1

Gewichte (Ausgabenanteile)	arithmet. Mittel	harmon. Mittel
Basisperiode $p_0q_0/\Sigma p_0q_0$	P^L (Laspeyres)	
Berichtsper. $p_tq_t/\Sigma p_tq_t$		P^P (Paasche)
"hybrid" 1 $p_0q_t/\Sigma p_0q_t$	P^P	
"hybrid" 2 $p_tq_0/\Sigma p_tq_0$		P^L

Die in diesem Schema nicht ausgefüllten Kombinationsmöglichkeiten müssen nicht notwendig völlig unsinnig sein und können in dem indirekten Sinne eines reziproken Mengenindexes als "Preisindizes" aufgefaßt werden.
So wurde z.B. als Preisindex von Palgrave ein arithmetisches Mittel der Ausgabenanteile der Berichtsperiode vorgeschlagen. Die sich daraus ergebende Indexfunktion

$$P^{PA} = \frac{\Sigma p_{it}q_{i0}v_i}{\Sigma p_{it}q_{it}} \quad \text{mit} \quad v_i = \frac{p_{it}q_{it}}{p_{i0}q_{i0}}$$

unterscheidet sich vom reziproken Paasche-Mengenindex $(Q^P)^{-1} = \Sigma p_{it}q_{i0}/\Sigma p_{it}q_{it}$ durch die im Zähler an den Mengen q_{i0} mit dem Faktor (Ausgabenverhältnis) v_i angebrachten Korrekturen. Man kann $q_{i0}v_i$ oder gleichbedeutend $p_{it}q_{it}/q_{i0}$ als eine die Preissteigerung berücksichtigende "korrigierte" Menge q^*_{io} auffassen. P^{PA} ist aber auch das harmonische Mittel der reziproken Preismesszahlen mit den Ausgabenanteilen der Berichtszeit, so dass die Interpretation als Preisindex auch wieder problematisch erscheint.

zu 2:

Einige besonders bekannte durch Mittelung gebildete "Kompromissformeln" sind in Übers. 10.2 zusammengestellt.

Übersicht 10.2

Mittelwert von	Art des Mittelwerts	
	arithmetisch	geometrisch
Indexformeln	$P^{DRO} = ½(P^L+P^P)$ [Drobisch 1871]	$P^F = \sqrt{P^L P^P}$ [Fisher 1922]
Gewichten	$\frac{\Sigma p_{it}(q_{i0} + q_{it})}{\Sigma p_{i0}(q_{i0} + q_{it})}$ [Bowley et al.1887]	$\frac{\Sigma p_{it}\sqrt{q_{i0}\, q_{it}}}{\Sigma p_{i0}\sqrt{q_{i0}\, q_{it}}}$ [Walsh 1901]

Besonders Irving Fisher ist bekannt für seine systematische Suche nach einer Indexformel, die nach bestimmten Kriterien "ideal" ist, wie etwa der von ihm gefundene "Idealindex"

(10.10) $P^F = \sqrt{P^L P^P}$ ("Idealindex" von I. Fisher).

Als geometrisches Mittel von Laspeyres- und Paasche-Preisindex hat der Idealindex interessante Eigenschaften, auf die an späterer Stelle noch eingegangen wird.
Aus der Mittelung folgt, dass die Preisindizes von Drobisch (P^{DRO}) und Fisher zwischen denen von Laspeyres und Paasche liegen müssen, es gilt also z.B. immer entweder $P^P \le P^F \le P^L$ oder $P^L \le P^F \le P^P$.
Man könnte natürlich auch ein harmonisches Mittel aus P^L und P^P berechnen (P^{HM}). Das geometrische Mittel aus P^{DRO} und P^{HM} ist P^F.

2. Indizes nach Laspeyres und Paasche

a) Preis- und Mengenindizes, Preisbereinigung

Die Preisindexformel von Laspeyres (und auch die von Paasche) wurde bereits schrittweise hergeleitet als Mittelwert von Preismesszahlen (Gl. 10.7). Man kann die Formeln für P^L und P^P in der Form der sog. "Aggregatformel" (Gl. 10.8 und 10.9) auch einführen, ausgehend von dem Gedanken, dass ein "Wertindex" nicht als Preisindex verwendet werden kann (vgl. Übers. 10.3).

Def. 10.1: Wertindex

Ein Wertindex ist eine Messzahl der tatsächlichen (nominalen) Ausgaben bzw. Einnahmen:

(10.11) $W_{0t} = \frac{\Sigma p_{it} q_{it}}{\Sigma p_{i0} q_{i0}}$ oder einfach $W_{0t} = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0}$

Bemerkungen zu Def. 10.1:

1. Im Beispiel der Verbraucherpreise ist ein Wertindex ein reines Ausgabenverhältnis, ein **Lebenshaltungskostenindex**. W_{0t} kann aber **kein Preisindex für die Lebenshaltung** sein (der nach der Formel P^L konstruiert ist). Denn W_{0t} wird nicht allein von der Preisentwicklung bestimmt, sondern auch durch die Mengenentwicklung. Zähler und Nenner des Wertindexes unterscheiden sich nicht nur durch die Preise, sondern auch durch die Mengen.

2. Die Lebenshaltungskosten können z.B. steigen obgleich alle Preise für die Lebenshaltung gleich bleiben, allein deshalb weil die Mengen steigen. Umgekehrt können trotz steigender Preise für die Lebenshaltung die Lebenshaltungskosten sogar sinken, weil die Haushalte "den Gürtel enger schnallen".

3. Es liegt also nahe, zu einem reinen Preisvergleich dadurch zu gelangen, dass man im Zähler und Nenner des Ausgabenverhältnisses mit den gleichen Mengen rechnet und dabei auch fiktive Ausgaben betrachtet. Es sind nun zwei Ansätze üblich (und natürlich noch viele weitere denkbar): man kann die Mengen der Basiszeit q_{i0} (Laspeyresansatz) oder die Mengen der Berichtszeit q_{it} (Paascheansatz) dem Ausgabenvergleich zugrunde legen (vgl. auch Übers. 10.3).

Def. 10.2: Preisindizes nach Laspeyres und Paasche

In ihrer Aggregatformel sind die Preisindizes von Laspeyres (P^L) und Paasche (P^P) gegeben durch

(10.8) $P^L = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0}$ (Laspeyres) und (10.9) $P^P = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t}$ (Paasche)

Bemerkungen zu Def. 10.2:

1. Zähler und Nenner der Preisindizes stellen jeweils Aggregate dar (im Falle von Verbraucherpreisindizes: Ausgabenaggregate). Der Nenner von P^L und der Zähler von P^P sind empirisch beobachtbare Aggregate (tatsächliche Ausgaben, bzw. allgemeiner Werte). Entsprechend sind der Nenner von P^P und der Zähler von P^L fiktive Aggregate.

2. Der Laspeyres-Preisindex beantwortet die Frage: Was würde heute (d.h. mit den Preisen zur Zeit t) der frühere (historische) Warenkorb der Basisperiode (mit den Mengen q_{i0}) kosten?

3. Der Paasche-Preisindex beantwortet die Frage: Was würde der heutige Warenkorb (d.h. mit den aktuellen Mengen q_{it}) heute mit den Preisen p_{it} mehr, bzw. weniger als damals (mit den Preisen p_{i0} zur Zeit 0) kosten?

Def. 10.3: Mengenindizes nach Laspeyres und Paasche

Durch Vertauschung von Preisen und Mengen in den Gleichungen 10.8 und 10.9 erhält man die entsprechenden Mengenindizes. Oder: in der gleichen Art, wie P^L und P^P gewogene Mittelwerte von Preismesszahlen sind, sind Q^L und Q^P gewogene Mittel von Mengenmesszahlen (vgl. Übers. 10.3). Die Mengenindizes lauten jeweils in der Aggregatformel:

$$(10.12)\quad Q^L = \frac{\Sigma q_t p_0}{\Sigma q_0 p_0} \quad \text{und} \quad (10.13)\quad Q^P = \frac{\Sigma q_t p_t}{\Sigma q_0 p_t}$$

Folgerung aus Def. 10.2 und 10.3

Zwischen den Preis- und Mengenindizes sowie dem Wertindex bestehen die folgenden leicht zu verifizierenden Gleichungen, die zugleich die Grundlage der Preisbereinigung (vgl. Def. 10.4) darstellen:

$$(10.14)\quad W_{0t} = P^L \cdot Q^P = P^P \cdot Q^L \,.$$

Man sieht leicht, dass sich z.B. im Produkt $P^P \cdot Q^L$ Zähler und Nenner entsprechend kürzen lassen, so dass gilt:

$$P^P_{0t} \cdot Q^L{}_{,0t} = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t} \frac{\Sigma p_0 q_t}{\Sigma p_0 q_0} = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0} = W_{0t} \,.$$

Def. 10.4: Wert, Volumen, Preisbereinigung

a) Ein Wert W_t, oder (anders genannt) ein "nominales" Aggregat, eine Größe "zu laufenden Preisen" ist eine Summe von Preismengenprodukten mit laufenden (aktuellen) Mengen und laufenden Preisen:

$$(10.15) W_t = \Sigma p_{it} q_{it} \quad \text{entsprechend} \quad W_0 = \Sigma p_{i0} q_{i0}$$

b) Ein Volumen V_t oder ein "reales" Aggregat, ist eine Preismengenproduktsumme mit laufenden (aktuellen) Mengen (q_{it}) "zu konstanten Preisen" der Basisperiode (zu den Preisen p_{i0}):

(10.16) $V_t = \Sigma p_{i0} q_{it}$ und $V_0 = W_0$

c) Deflationierung oder Realwert- oder Volumenrechnung ist die Aufgabe aus einem Wert ein Volumen (aus W_t die Größe V_t), bzw. aus einem Wertindex W_{0t} ein Mengenindex nach Laspeyres ($Q^L{}_{,0t}$) zu berechnen.

Bemerkungen zu Def. 10.4

1. Ein Wertindex (Def. 10.1) ist eine Messzahl von Werten im Sinne der Def. 10.4. Er stellt die wertmäßige Zunahme (nominale Steigerung) eines Aggregats (im Vergleich zur Basiszeit) dar.

2. Der Laspeyres-Mengenindex Q^L stellt die reale (volumenmäßige) Zunahme eines Aggregats dar (er ist praktisch ein "Volumenindex" [der Begriff wird jedoch auch anders gebraucht]) und soll die reine mengenmäßige Entwicklung darstellen, da Mengen in der Regel nicht in physischen Einheiten aggregierbar sind.

3. Ein Wert wird deflationiert indem man ihn durch einen entsprechend definierten (die gleichen Güter umfassenden) Preisindex nach Paasche dividiert:

(10.14a) $V_t = \frac{W_t}{P^P}$ und (10.14b) $Q^L = \frac{W_{0t}}{P^P}$.

Übersicht 10.3: Wertindex, Preis- und Mengenindizes

Wertindex W_{0t} (z.B. Lebenshaltungskostenindex)

$$W_{0t} = \frac{\Sigma p_t q_t}{\Sigma p_0 q_0}$$

Laspeyres Preisindex

$$P_{0t}^{L} = \frac{\Sigma p_t q_0}{\Sigma p_0 q_0}$$

Verwendung für:
spezielle Preisniveaus
(z.B. Preisindizes für die Lebenshaltung)

Paasche Preisindex

$$P_{0t}^{P} = \frac{\Sigma p_t q_t}{\Sigma p_0 q_t}$$

Verwendung für:
Preisbereinigung*)
(Deflationierung, z.B. des Sozialprodukts)

Vertauschung von Preisen und Mengen in den Formeln führt zu den entsprechenden Mengenindizes Q^L und Q^P also:

Laspeyres Mengenindex

$$Q_{0t}^{L} = \frac{\Sigma q_t p_0}{\Sigma q_0 p_0}$$

Paasche Mengenindex

$$Q_{0t}^{P} = \frac{\Sigma q_t p_t}{\Sigma q_0 p_t}$$

Eine andere Art der Herleitung von Mengenindizes:

Während P^L ein mit Ausgabenanteilen zur Basiszeit ($q_0p_0/\Sigma q_0p_0$) gewogenes arithmetisches Mittel von Preismesszahlen ist, ist Q^L ein analog gebildetes Mittel von Mengenmesszahlen. Entsprechendes gilt für den Zusammenhang von P^P und Q^P: der Preisindex P^P ist ein mit den Ausgabenanteilen der Berichtszeit gewogenes harmonisches Mittel der Preismesszahlen und Q^P ist ein entsprechendes Mittel der Mengenmesszahlen. Man kann auch den Wertindex W_{0t} als Mittelwert der Wertmesszahlen $p_{it}q_{it}/p_{i0}q_{i0}$ auffassen. Dabei führen beide Arten der Mittelung, diejenige nach Laspeyres (arithmetisches Mittel, Ausgabenanteile der Basiszeit) und diejenige nach Paasche (harmonisches Mittel, Ausgabenanteile der Berichtszeit) zum gleichen Ergebnis.

Wertindex als Indexprodukt:

Es gilt nun die folgende grundlegende Formel für die Preisbereinigung:

$$(10.14) W = P^L Q^P = P^P Q^L$$

*) Zur Preisbereinigung (Deflationierung) oder Realwert- oder Volumenrechnung vgl. Def. 10.4 Teil c): $W_t \rightarrow V_t$ und $W_{0t} \rightarrow Q_{0t}^L$ gem. Gl. 10.14a und 10.14b

Beispiel 10.3:

Diplom-Kaufmann K aus E und Gattin gehen leidenschaftlich gern ins Kino. Von Zeit zu Zeit schätzen sie etwas Bildendes im "Filmkunst", und sie lassen sich auch schon mal politisieren im "Alternativkino". Die Ausgaben des Ehepaares für Kinobesuche sind von 1988 bis 1990 nominal um 40vH und real dagegen nur um 25vH gestiegen. Für die Eintrittspreise der Kinos gelte:

Nr.	Kino	Preis	
		1988	1990
1	Filmkunst	15	12
2	Alternativkino	10	14
3	Kolossal-Kino	12	18
4	Bahnhofskino	20	24

a) Man berechne den Preisindex nach Laspeyres, wenn sich die Ausgabenanteile für Kinobesuche bei dem Ehepaar 1988 wie folgt verhalten: 2:3:2:1 (=Aufteilung der Ausgaben auf die vier Kinos).

b) Berechnen Sie den Preis- und Mengenindex nach Paasche!

Lösung 10.3:

a) Die Preismesszahlen m_i lauten $m_1 = 0{,}8$, $m_2 = 1{,}4$, $m_3 = 1{,}5$ und $m_4 = 1{,}2$. Die Ausgabenanteile ergeben sich aus der Angabe 2:3:2:1. Folglich sind die Preismesszahlen zu gewichten mit 2/8, 3/8, 2/8 und 1/8. Man erhält dann $P^L = 10/8 = 1{,}25$.

b) Aus den Angaben ist zu entnehmen, dass $W_{0t} = 1{,}4$ und $Q^L = 1{,}25$ ist. Daraus errechnet sich $P^P = 1{,}4/1{,}25 = 1{,}12$ und $Q^P = 1{,}4/1{,}25 = 1{,}12$.

Beispiel 10.4:

a) Angenommen, das Sozialprodukt sei (über mehrere Jahre) nominal (zu jeweiligen Preisen) um 50vH gestiegen, real (zu konstanten Preisen eines Basisjahres) aber nur um 25vH. Welchen Wert nimmt dann der Preisindex des Sozialprodukts (ein Preisindex nach Paasche) an?

b) Das wertmäßige Bruttosozialprodukt habe sich verdoppelt, das volumenmäßige (in Preisen von 1970) Sozialprodukt sei dagegen nur um 1/3 gestiegen. Der "Preisindex des Sozialprodukts" 1970 = 100 beträgt somit (Richtiges ankreuzen):

☐ 166,67 ☐ 150 ☐ 133,33 ☐ 66,67.

Lösung 10.4:

a) 1,5/1,25 = 1,2 also 20% Preissteigerung, nicht 50% - 25%= 25%.
b) 150.

d) Vergleich von Laspeyres- und Paasche-Formel

In Abschnitt 1b (dort Ziff.5) und in den Bemerkungen zu Def.10.2 sind bereits einige Gegenüberstellungen zwischen P^L und P^P vorgenommen worden, die nachfolgend etwas vertieft und an Beispielen weiter erläutert werden sollen.

1. Praktische wirtschaftsstatistische Aspekte

Der Laspeyres-Preisindex hat, wie bereits gesagt, den Vorteil, dass er durch sein gleichbleibendes Wägungsschema leicht monatlich zu berechnen ist. Bei einem konstanten Warenkorb sind nach der Messzahlenmittelwertformel monatlich nur die Preismesszahlen auszutauschen; die Gewichte bleiben bis zu einer Neuberechnung (vgl. Abschn. 4) unverändert. Die Kehrseite ist jedoch, dass der Warenkorb von Zeit zu Zeit im Zuge einer Neuberechnung des Indexes aktualisiert werden muss. Der Paasche-Preisindex wird vor allem zur Preisbereinigung verwendet, aber auch dort, wo laufend die jeweiligen Warenkörbe quasi als Nebenprodukt anfallen, wie z.B. in der Außenhandelsstatistik.

2. Zeitreiheninterpretation

Aufeinanderfolgende Werte des Paasche-Preisindexes P^P unterscheiden sich (anders als bei P^L) nicht nur durch die Preise, sondern auch durch die Mengen. Die Folge P_{01}, P_{02}, P_{03} hat nämlich die folgende Gestalt:

Paasche: P_{01}^P, P_{02}^P, P_{03}^P	Laspeyres: P_{01}^L, P_{02}^L, P_{03}^L
$\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$, $\frac{\Sigma p_2 q_2}{\Sigma p_0 q_2}$, $\frac{\Sigma p_3 q_3}{\Sigma p_0 q_3}$	$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$, $\frac{\Sigma p_2 q_0}{\Sigma p_0 q_0}$, $\frac{\Sigma p_3 q_0}{\Sigma p_0 q_0}$

Aufeinanderfolgende Werte des Paasche-Preisindexes P^P sind somit streng genommen keine Zeitreihe, denn sie sind (anders als bei P^L) nicht untereinander, sondern jeweils nur mit dem Basiswert vergleichbar.

3. Größenbeziehung zwischen P^L und P^P[Kovarianz zwischen Preis- und Mengenmesszahlen]

Der folgende zuerst von Ladislaus von Bortkiewicz gefundene Zusammenhang zeigt unter welchen Voraussetzungen der Laspeyres-Preisindex (was die Regel ist) größer als der Paasche Preisindex ist. Er ist leicht zu zeigen:

Mit den Ausgabenanteilen g_i zur Basiszeit (wie oben definiert als $g_i = p_{i0}q_{i0}/\Sigma p_{i0}q_{i0}$) sowie den Preismesszahlen $b_i = p_{it}/p_{i0}$ und Mengenmesszahlen $c_i = q_{it}/q_{i0}$ erhält man für die Laspeyres-Indizes

$P_{0t}^L = \Sigma g_i b_i$ und $Q_{0t}^L = \Sigma g_i c_i$.

Ferner gilt $W_{0t} = \Sigma b_i c_i g_i = P^L Q^P = P^P Q^L$, so dass man für die Kovarianz C von Preis- und Mengenmesszahlen erhält

(10.17) $C = \Sigma(b_i - P^L)(c_i - Q^L)g_i = Q^L(P^P - P^L)$.

Eine Umformung von Gl. 10.17 liefert

(10.17a) $P^L/P^P = 1 - C/W$ und aus Gl. 10.14 folgt ferner

(10.17b) $P^L/P^P = Q^L/Q^P$.

Ist also $P^L > P^P$ so ist auch $Q^L > Q^P$.

Da die Kovarianz stets das Produkt des Korrelationskoeffizienten und der Standardabweichungen ist, also $C = r_{bc}s_b s_c$ gilt $P^L = P^P$ wenn:

- die Kovarianz $C = 0$ und damit auch $r_{bc} = 0$
- die Preis- und/oder Mengenmesszahlen keine Streuung haben, also alle gleich sind ($s_b = 0$, bzw. $s_c = 0$), d.h. alle (Preise) Mengen im gleichen Verhältnis zu- oder abnehmen.

Die Standardabweichungen s_b und s_c werden i.d.R. umso größer sein, je weiter das Basisjahr zurückliegt, weshalb dann meist auch die Unterschiedlichkeit von P^L und P^P zunimmt.

Aus Gl. 10.17 folgt ferner:

(10.18) $W = P^L Q^L + C$.

Bei negativer Korrelation zwischen Preis- und Mengenmesszahlen (was die Regel ist) gilt also $P^L Q^L > W$ und $P^P Q^P < W$, was zugleich bedeutet, dass weder der Laspeyres-, noch der Paasche-Index die Faktorumkehrprobe (vgl. Abschn. 3c) erfüllt.

Negative Korrelation bedeutet $P^L > P^P$, was bei rationaler Substitution an einer gegebenen Nachfragekurve der Fall ist. Weichen die Haushalte einer Preissteigerung in der Weise aus, dass sie die nachgefragte Menge eines im Preis (stärker) gestiegenen Gutes reduzieren im Vergleich zu einem im Preis gesunkenen (oder weniger stark gestiegenen) Gut, so ist $P^L > P^P$. Wenn aber (kurzfristig) keine Substitution möglich ist (Mieten, Kfz usw.) oder sich die Nachfragekurven wegen Einkommenssteigerung "nach außen verschieben", so kann auch $P^L < P^P$ sein, wie dies für entsprechende Teilindizes des Preisindexes für die Lebenshaltung auch durch Parallelrechnungen von P^L und P^P in der amtlichen Statistik empirisch festgestellt wurde. Man kann diesen auch "Laspeyres-Effekt" genannten Sachverhalt, dass nämlich i.d.R. gilt $P^L > P^P$ auch mit der Theorie des Nutzenindexes begründen (vgl. Abschn. 3d und Beispiel 10.5).
In den Abschnitten 3d und 4 werden noch weitere Unterschiede und Gemeinsamkeiten der beiden Indizes gezeigt.

Beispiel 10.5

Die folgende Aufgabe demonstriert die Größenbeziehung $P^L > P^P$ und deren Begründung mit der "rationalen Substitution":
Gegeben seien die folgenden Preise und Mengen für zwei Waren A und B zu zwei Zeitpunkten sowie alternative Mengen zur Zeit t, nämlich entweder die Mengen q_{1t} oder q_{2t}:

Ware	Preise	
	0	t
A	20	40
B	45	30

Menge
0
60
40

Mengen zur	
q_{1t}	q_{2t}
40	80
60	30

Man beachte, dass die Ware A teurer und die Ware B billiger geworden ist und berechne P^L sowie zweimal P^P, einmal mit den Mengen q_{1t} und einmal mit den Mengen q_{2t}.

Lösung 10.5:

a) Mit den Mengen q_{1t}: $P^L = 3600/3000 = 1{,}2$ (also 120%) und $P^P = 3400/3500 = 0{,}9714$ (wenn das Basisjahr 100 gesetzt wird, ist also $P^P = 97{,}1$); mithin ist $P^L > P^P$, denn den Preismesszahlen 2 und 2/3 (für A und B) stehen die Mengenmesszahlen 2/3 und 1,5 gegenüber: die teurer gewordene Ware A wird weniger und die billiger gewordene Ware B wird mehr konsumiert (rationale Substitution). Die Lebenshaltungskosten sind um 13,3% gestiegen ($W_{0t} = 3400/3000 = 1{,}133$), die Preise - gemessen an P^L - dagegen um 20%.

b) Mit den Mengen q_{2t}: P^L bleibt hiervon unberührt und für P^P erhält man jetzt $P^P = 4100/2950 = 1{,}390$ und es gilt anders als oben : $P^L < P^P$. Die

teurer (billiger) gewordene Ware A (B) wird mehr (weniger) konsumiert. Die Lebenshaltungskosten sind um 36,7% gestiegen (W_{0t} = 4100/3000 = 1,3667).

Beispiel 10.6:

Der Haushalt des arbeitslosen Diplom-Kaufmanns K aus E (vgl. Bild) konsumiere nur zwei Waren A und B. Über Preise und Mengen sei folgendes bekannt

Ware	Preise		Mengen	
	0	t	0	t
A	10	12	30	20
B	20	16	15	q_{Bt}

Wie groß muss q_{Bt} sein, wenn $P^P < P^L$ und wenn $P^P > P^L$ sein soll? Interpretieren Sie das Ergebnis!

Lösung 10.6:

Die Zahlen sind so gewählt, dass $P^L = 1$ (also 100%) ist. Man sieht leicht, dass gilt:

- $P^P < 1$ verlangt $q_{Bt} > 10$ (Beispiel: $q_{Bt}= 30$ führt zu $P^P = 0,9$)
- und entsprechend $P^P > 1$ bedeutet $q_{Bt} < 10$ (Beispiel: $q_{Bt}= 10/3 = 3,33$ führt zu $P^P = 1,1$)

Weitere Bemerkungen zum Beispiel 10.6:

Die Ausgabenanteile zur Basiszeit für die Waren A und B betragen jeweils ½. Somit ist $P^L = ½·1,2 + ½·0,8$. Der Paasche-Preisindex kann auch als arithmetisches Mittel der Preismesszahlen, gewogen mit den "hybriden" (vgl. Übers. 10.1), bzw. "realen" Ausgabenanteilen $p_0q_t / \Sigma p_0q_t$ aufgefaßt werden. Mit $q_{Bt} = 10$ erhält man für die so definierten Ausgabenanteil ebenfalls ½. Ist dagegen $q_{Bt} > 10$, so ist der reale Ausgabenateil für die billiger gewordene Ware B größer als ½, für $q_{Bt} < 10$ ist er dagegen kleiner als ½. Der erste Fall beinhaltet die oben so bezeichnete "rationale Substitution". Man beachte, dass $P^P < P^L$ nicht verlangt, dass von der billiger gewordenen Ware B zur Zeit t absolut mehr (und von der teuerer gewordenen Ware A absolut weniger) konsumiert werden muss. Es reicht, dass die Mengenmesszahl von B größer ist als diejenige für A.

3. Theorie und Axiomatik der Indexzahlen

a) Formale und ökonomische Theorie der Indexzahlen

Bei jeder praktischen Indexberechnung in der Wirtschaftsstatistik sind die folgenden vier Probleme zu lösen:

1. Auswahl der Reihen, d.h. es muss festgelegt werden, welche Güter hinsichtlich Art und Qualität in den Index einzubeziehen sind. Die "Zusammenfassung" der Einzelreihen zu einem Index muss sachlich "sinnvoll" sein (Problem der Repräsentativität des Warenkorbs).
2. Wahl des Wägungsschemas (der Art der Gewichte). Ein Beispiel: Sollen Aktienkurse im Aktienindex mit dem Stammkapital oder mit den Börsenumsätzen oder mit dem Eigenkapital der Gesellschaften gewogen werden?
3. Wahl des Basisjahres: Es ist unmittelbar einleuchtend, dass man vermeiden sollte, ein extrem "gutes" oder "schlechtes" Jahr als Basisjahr auszuwählen. Die Regel ist, ein Normaljahr zu wählen (schon wegen saisonaler Schwankungen wird - auch bei einem monatlich zu berechnenden Index - i.d.R. nicht ein Basismonat sondern ein Basisjahr gewählt).
4. Wahl der Indexformel.

Aufgabe der **formalen Indextheorie** ist es, vor allem das Problem Nr. 4 zu lösen durch Entwicklung von Bewertungskriterien für Indexformeln. Dieser Teil der Indextheorie hat jedoch auch Grenzen. Die formale Theorie berücksichtigt nicht die inhaltliche (ökonomische) Interpretierbarkeit und Aspekte der praktischen Durchführbarkeit der Indexberechnung oder z.B. das Kriterium der Verständlichkeit der Indexaussage. Sie bedarf deshalb der Ergänzung durch eine **ökonomische Theorie der Indexzahlen**.

Die formale Theorie war zunächst eine systematische, aber kaum theoriegeleitete Suche nach einem "idealen Index". Schon 1922 stellte Irving Fisher 134 mögliche Indexformeln zusammen. Fisher begann bereits, axiomatisch vorzugehen und aus formalen Eigenschaften von Indexzahlen Gütekriterien (Postulate) zu entwickeln (sog. Proben). Man erkannte jedoch früh, dass Kriterien dieser Art häufig nur aus Plausibilitätserwägungen hergeleitet wurden, dass sie nicht widerspruchsfrei sind und dass sie ökonomische Abhängigkeiten zwischen Preisen, Mengen und Einkommen ignorieren. Letztere sind vor allem Gegenstand der ökonomischen Theorie der Indexzahlen. Konkrete Probleme, denen sie ihre Entstehung verdankt, traten in Inflationszeiten auf: Wie stark müssen die Einkommen steigen, um ein Absinken des Lebensstandards (verstanden als Realeinkommen) zu verhindern? Die Bezugsgröße "Lebensstandard" oder besser "Nutzen" (als Konstrukt der

Wirtschaftstheorie) ist kennzeichnend für die ökonomische Theorie der Indexzahlen, während sich die formale Indextheorie auf mathematische Eigenschaften der Indexfunktion beschränkt.

b) Axiomatik der Preisindexzahlen

In der folgenden Übers. 10.4 werden zunächst die fünf Axiome, denen ein Preisindex genügen sollte vorgestellt und sie werden dann anschließend ausführlich kommentiert.

Übersicht 10.4: Axiomensystem von Eichhorn und Voeller

Notation: Preis- und Mengenvektoren (jeweils n Komponenten [Waren]) $\mathbf{p}_0$, $\mathbf{q}_0$, $\mathbf{p}_t$, $\mathbf{q}_t$. Subskript für die Warenart: i = 1,2,...,n. Die Indexfunktion P: $\mathbb{R}^{4n} \rightarrow \mathbb{R}$ sollte danach die folgenden Axiome erfüllen:

P1: Monotonie	a)	$P(\mathbf{p}_0,\mathbf{p}_t^{*)}) > P(\mathbf{p}_0,\mathbf{p}_t)$ wenn $\mathbf{p}_t^* > \mathbf{p}_t$ und für mindestens eine Ware i gilt: $p_{it}^* > p_{it}$
	b)	$P(\mathbf{p}_0^*,\mathbf{p}_t) < P(\mathbf{p}_0,\mathbf{p}_t)$ analog: $\mathbf{p}_0^* > \mathbf{p}_0$ und $p_{i0}^* > p_{i0}$ für mindestens ein i (eine Ware)
P2: Lineare Homogenität[a)]		$P(\mathbf{p}_0,\lambda\mathbf{p}_t) = \lambda P(\mathbf{p}_0,\mathbf{p}_t)$ mit $\lambda \in \mathbb{R}_+$
P3: Identität[b)]		$P(\mathbf{p}_t,\mathbf{p}_0) = 1$ wenn $p_{it} = p_{i0}$ für alle i
P4: Dimensionalität		$P(\lambda\mathbf{p}_0,\lambda\mathbf{p}_t) = P(\mathbf{p}_0,\mathbf{p}_t)$ mit $\lambda \in \mathbb{R}_+$
P5: Kommensurabilität		$P(\mathbf{A}\mathbf{p}_0, \mathbf{A}\mathbf{p}_t, \mathbf{A}^{-1}\mathbf{q}_0, \mathbf{A}^{-1}\mathbf{q}_t) = P(\mathbf{p}_0,\mathbf{p}_t,\mathbf{q}_0,\mathbf{q}_t)$ mit $\mathbf{A} = \text{diag}(\alpha_1,\alpha_2,..., \alpha_n)$ und $\alpha_i > 0$,

a) Unter Homogenität vom Grade -1 versteht man die Forderung $P(\lambda\mathbf{p}_0,\mathbf{p}_t) = P(\mathbf{p}_0,\mathbf{p}_t)/\lambda = \lambda^{-1}P(\mathbf{p}_0,\mathbf{p}_t)$. Sie ist erfüllt, wenn P2 und P4 gelten.

b) Axiome P2 und P3 stellen zusammen sicher, dass die sog. Proportionalitätsprobe (vgl. unten Bemerkung Nr. 4) erfüllt ist.

Bemerkungen zu den Axiomen:

1. Ein Axiomensystem grenzt eine mehr oder weniger weite Klasse von Indexfunktionen ab. Es ist z.T. eine "Geschmacksache" wie weit die Grenzen gezogen werden. Von den gleichen Autoren gibt es auch ein System mit vier Axiomen. Eine Axiomatik sollte widerspruchsfrei und unabhängig sein. Widerspruchsfreiheit ist gegeben, wenn es Formeln gibt, die in der Tat alle Axiome erfüllen. Lassen sich Formeln finden,

die je 4 der 5 Axiome erfüllen, das verbleibende fünfte aber nicht, dann ist die Unabhängigkeit bewiesen.

2. Die Axiome P1 bis P4 gelten auch für Preisindizes, die nur von den beiden Preisvektoren abhängen, also z.B. ungewogene Mittelwerte von Preismesszahlen, bei denen keine Mengenvektoren auftreten (mit denen gewichtet wird). Man beachte, dass generell keine Aussagen über die Gewichtvektoren $\mathbf{q}_0$ und $\mathbf{q}_t$ gemacht werden, wenn man von Axiom P5 absieht.

3. Nach Axiom P1 gilt: zunehmende Preise der Berichtsperiode (Basisperiode) müssen auch zu einer Zunahme (Abnahme) des Indexes führen, was offensichtlich eine sehr plausibel erscheinende Forderung ist. Eine spezielle Forderung ist die **Additivität**.

   Sie bedeutet in der Notation der Übersicht 10.4:

   Fall a) unterschiedliche Preise in der Berichtsperiode:
   $P(\mathbf{p}_0,\mathbf{p}_t^{*)}) = P(\mathbf{p}_0,\mathbf{p}_t) + P(\mathbf{p}_0,\mathbf{p}_t^{+})$ wenn für die Vektoren $\mathbf{p}_t^{*},\mathbf{p}_t$ und $\mathbf{p}_t^{+}$ gilt: $\mathbf{p}_t^{*} = \mathbf{p}_t + \mathbf{p}_t^{+}$ und entsprechend

   Fall b) unterschiedliche Preise in der Basisperiode:

   $[P(\mathbf{p}_0^{*},\mathbf{p}_t)]^{-1} = [P(\mathbf{p}_0,\mathbf{p}_t)]^{-1} + [P(\mathbf{p}_0^{+},\mathbf{p}_t)]^{-1}$ wenn für die Vektoren $\mathbf{p}_0^{*},\mathbf{p}_0$ und $\mathbf{p}_{0,}^{+}$ gilt: $\mathbf{p}_0^{*} = \mathbf{p}_0 + \mathbf{p}_0^{+}$.

   Man erkennt leicht, dass die Formulierung der Monotonieeigenschaft in Übers. 10.4 allgemeiner gehalten ist, dass also die Additivität ein Spezialfall hiervon ist. Im Beispiel 10.7 wird gezeigt, dass die Laspeyres- und die Paasche-Formel die Additivität erfüllen. Man sieht auch, dass Additivität lineare Homogenität (Axiom P2) impliziert (aber nicht umgekehrt). Der auf einer geometrischen Mittelung von P^L und P^P beruhende "Idealindex" von Fisher P^F erfüllt P2 (und auch P3 und damit auch die Proportionalitätsprobe), er ist aber nicht additiv (anders dagegen der arithmetisch gemittelte Index von Drobisch, der alle diese Axiome erfüllt).

4. P2 besagt, wenn z.B. gilt $\lambda = 1/N$ (bei N Personen), dass es irrelevant ist, ob sich eine Ausgabe auf alle N Personen bezieht, oder ob sie "pro Kopf" gerechnet ist. Aus der linearen Homogenität folgt in Verbindung mit der Forderung der Identität (Axiom P3) die sog. **Proportionalitätsprobe** (nach I. Fisher):
   Wenn sich alle Preise ver-λ-fachen, also für alle i = 1,2,...,n Waren gilt $p_{it} = \lambda p_{i0}$, dann soll der Preisindex den Wert λ annehmen. Es ist offensichtlich, dass z.B. der für die Praxis besonders bedeutsame Index nach Laspeyres diese Probe erfüllt. Steigen beispielsweise (verglichen mit der Basisperiode) alle n Preise um jeweils 20%, so ist $P^L = 1{,}2$ (also 120%).

5. Die trivial erscheinende Identitätsforderung P3 bedeutet, dass sich der Preisindex **nicht** ändert (100% beträgt), wenn sich **kein** Preis ändert. Ein Wertindex W_{0t} muss P3 nicht erfüllen. Man kann aber auch "umgekehrt" fordern, dass ein Index nicht 1 sein (bleiben) sollte, wenn **alle** Preise steigen. Das ist jedoch in der Monotonieforderung (P1) impliziert und wird von allen Indizes, die Mittelwerte von Preismesszahlen sind, erfüllt.

6. Gelten die Axiome P1 bis P3, so ist sichergestellt, dass der so konstruierte Preisindex die Eigenschaften eines Mittelwerts von Preismesszahlen hat, d.h. insbesondere dass er einen Wert annimmt, der zwischen der kleinsten und der größten Preismesszahl liegt.

7. P4 stellt die Unabhängigkeit von der Währungseinheit der Preisnotierung sicher (es ist irrelevant, ob z.B. die Preise in DM oder in Pfennigen oder in US$ notiert sind).

8. Entsprechendes leistet P5 hinsichtlich der Mengeneinheit, auf die sich die Preisnotierung zu den Zeiten 0 und t bezieht. Kommensurabilität bedeutet, dass ein Preisindex unabhängig davon ist, in welcher Mengeneinheit die Preise notiert sind. Wie bereits dargestellt, erfüllt der Index von Dutot (oder jeder andere auf Summen und Durchschnitte von Preisen beruhende Index) das Axiom P5 nicht.

   Die Diagonalmatrix $\mathbf{A}$, mit den Elementen α_i (wie oben definiert), bedeutet, dass sich z.B. der Preis der i-ten Ware verdoppelt ($\alpha_i=2$) , weil sich die zugrundeliegende Menge halbiert (z.B. Übergang von Pfund- zu Kilo-Preisnotierung). Es wird davon ausgegangen, dass sich die einzelnen α_i unterscheiden. Sind sie alle gleich ($\alpha_i = \alpha$ für alle i), so wäre dies eine sehr viel schwächere Forderung, nämlich die quantity dimensionality im Unterschied zur price dimensionality (P4), der auch Indizes genügen, die das Axiom P4, nicht aber P5 erfüllen, also z.B. der Dutot-Index.

9. Sind PI_1,..., PI_k Preisindizes, die alle Axiome dieses Axiomensystems erfüllen, dann ist in gewissen Fällen auch eine Funktion dieser Indizes, z.B. ein Potenzmittel der Preisindizes PI_1,..., PI_k ein Preisindex, der alle Axiome erfüllt.

10. Das Axiomensystem schließt sachlich (ökonomisch) gesehen ziemlich unsinnige Indexformeln nicht aus (es ist ja auch nur eine Grundlage für die **formale** Theorie der Preisindexzahlen) und Kriterien, wie ökonomische Interpretierbarkeit und Verständlichkeit der Indexaussage sind nicht maßgeblich. Aber was heißt "sachlich unsinnig"? Man kann hier verschiedene Maßstäbe ansetzen. Begnügt man sich mit der Mittelwerteigenschaft (vgl. Bem. 6), so wäre jeder Index "sinnvoll" der diese Axiome erfüllt, was m.E. zu weitgehend ist.

Beispiel 10.7:

Man zeige, dass der Laspeyres-Preisindex die Additivität erfüllt.

Lösung 10.7:

Wir beschränken uns auf einen Preisindex mit zwei Waren. Dann gilt mit den Preisdifferenzen d_1 und d_2 für die Laspeyres-Preisindizes:

$$P(\mathbf{p}_0,\mathbf{p}_t^{*)}) = \frac{(p_{1t} + d_1)q_{1t} + (p_{2t} + d_2)q_{2t}}{p_{10}q_{10} + p_{20}q_{20}}$$

$$P(\mathbf{p}_0,\mathbf{p}_t) = \frac{p_{1t}q_{1t} + p_{2t}q_{2t}}{p_{10}q_{10} + p_{20}q_{20}} \qquad P(\mathbf{p}_0,\mathbf{p}_t^{+}) = \frac{d_1q_{1t} + d_2q_{2t}}{p_{10}q_{10} + p_{20}q_{20}}$$

Man erkennt sofort, dass gilt $P(\mathbf{p}_0,\mathbf{p}_t^{*)}) = P(\mathbf{p}_0,\mathbf{p}_t) + P(\mathbf{p}_0,\mathbf{p}_t^{+})$.
In entsprechender Weise kann man auch zeigen, dass Additivität bei Erhöhung der Basispreise um d_1 bzw. d_2 erfüllt ist und dass der Paasche-Preisindex ebenfalls der Monotonieforderung im engeren Sinne der Additivität nachkommt.

c) Andere wünschenswerte Eigenschaften von Indexzahlen

Die folgenden drei Eigenschaften von Messzahlen werden von Indexzahlen nicht unbedingt erfüllt:

1. Zeitumkehrbarkeit
2. Zirkularität
3. Faktorumkehrbarkeit

Während Messzahlen diese Forderungen stets erfüllen sind sie bei Indizes, also aggregierten Messzahlen in der Regel nicht erfüllt.

zu 1: Zeitumkehrbarkeit

Hierunter versteht man dass die Vertauschung von Basis- und Berichtsperiode zum reziproken Preisindex führt

(10.19) $P_{0t}P_{t0} = 1$ (Zeitumkehrbarkeit).

Für den Laspeyres-Preisindex gilt im allgemeinen $P_{0t}^{L}\ P_{t0}^{L} > 1$ und für den Paasche-Preisindex $P_{0t}^{P}\ P_{t0}^{P} < 1$. Beide Indexformeln erfüllen also die Zeitumkehrbarkeit (time reversal) nicht.
Wie man leicht sieht, gilt jedoch

(10.20) $P_{0t}^{L}\ P_{t0}^{P} = P_{0t}^{P}\ P_{t0}^{L} = 1$.

In diesem Sinne ist die Laspeyres-Formel die "time antithesis" (Irving Fisher) der Paasche-Formel. Gl. 10.20 gilt entsprechend für Mengenindizes von Laspeyres und Paasche.
Die obige Feststellung, dass im allgemeinen - nämlich bei nicht zu großen Unterschieden zwischen den Warenkörben der Basis- und Berichtszeit - gilt $P^L_{0t}\ P^L_{t0} > 1$ und $P^P_{0t}\ P^P_{t0} < 1$, hängt mit der Art der Mittelwertbildung zusammen, wie im folgenden gezeigt wird. Bei ungewogenen Mittelwerten von n Messwerten $x_1,x_2,...,x_n$, etwa dem arithmetischen Mittel
$A(x_i)$, dem harmonischen Mittel $H(x_i)$ und dem geometrischen Mittel $G(x_i)$ gelten die folgenden, in Übersicht 10.5 zusammengestellten Beziehungen zu den entsprechenden Mittelwerten der reziproken Werte x_i^{-1}.

Übersicht 10.5

Zusammenhänge zwischen Mittelwerten der Messwerte und den Mittelwerten der reziproken Messwerte

$A(x_i^{-1}) > [A(x_i)]^{-1}$	$G(x_i^{-1}) = [G(x_i)]^{-1}$	$H(x_i^{-1}) < [H(x_i)]^{-1}$
arithmetisches Mittel	geometrisches Mittel	harmonisches Mittel

Ein Preisindex als **un**gewogenes geometrisches Mittel der Preismesszahlen würde also stets die Zeitumkehrprobe erfüllen.

Der Begriff Zeitumkehrprobe ist zu eng, weil Indizes z.B. auch für den räumlichen Vergleich benutzt werden. Im allgemeinen Sinne ist mit der Zeitumkehrprobe die Umkehrung der Vergleichsrichtung gemeint. Gerade im internationalen Vergleich ist der "reversal-test" auch im besonderen Maße motiviert: es gibt meist keinen Grund, ein bestimmtes Land als Basisland zu bevorzugen (Kriterium der "Basislandinvarianz"), während es im zeitlichen Vergleich die eindeutige zeitliche Abfolge ist, die es sinnvoll erscheinen läßt, 0 als Basis- und t als Berichtsperiode zu wählen und nicht umgekehrt.
Die Zeitumkehrprobe ist ein zweifelhaftes Kriterium, denn es ist unmittelbar einsichtig, dass eine Umkehrung der Vergleichsrichtung i.d.R. auch mit einer "Umkehrung" des Warenkorbs verbunden ist und warum sollte $P^L{}_{,t0} = (P^L{}_{,0t})^{-1}$ sein, wenn $P^L{}_{,0t}$ und $P^L{}_{,t0}$ Indizes
mit verschiedenen Warenkörben sind. Es ist deshalb auch nicht überraschend, dass Indizes, deren Wägungsschema durch Mittelwertbildung entstehen (vgl. Übers. 10.2), der Zeitumkehrprobe genügen.

zu 2: Zirkularität (Verkettbarkeit)

Mit dieser Forderung (auch Transitivität genannt, oder "Rundprobe" ["circular test" nach I. Fisher]) ist gemeint, dass für beliebige, aber verschiedene Perioden, etwa für $0 < s < t$ (die Reihenfolge ist nicht zwingend, es könnte also auch $0 > s > t$ sein) gelten soll:

(10.21) $P_{0t} = P_{0s}P_{st}$ (Verkettbarkeit).

Gl. 10.21 ist auch die Basis der Verkettung und Umbasierung von Indexzahlen (vgl. Abschn. 4). Ein Index, der nicht verkettbar ist, ist genau ge-

nommen auch nicht umbasierbar. Verkettbarkeit ist die strengere Forderung als Zeitumkehrbarkeit: Ist ein Index verkettbar, dann gilt auch die Zeitumkehrbarkeit, nicht aber umgekehrt. So erfüllt z.B. Fishers Idealindex P^F die Zeitumkehrbarkeit, nicht aber die Transitivität.

Aus Identität und Transitivität folgt Zeitumkehrbarkeit: Setzt man in Gl. 10.21 einfach t=0, so erhält man $P_{00} = P_{0s}P_{s0} = 1$.
Eine Produktbildung gem. Gl. 10.21 soll ganz allgemein gelten:

für $0 < m < n ... < r < s < t$ soll gelten $P_{0t} = P_{0m}P_{mn}...P_{rs}P_{st}$.

Weder P^L noch P^P erfüllen die Verkettbarkeit (vgl. Bsp. 10.7).
Der Grund weshalb Zirkularität eines Indexes gewünscht wird ist, dass dann

1) Veränderungsraten (Wachstumsfaktoren) unabhängig von der gewählten Basis sind. Gilt Gl. 10.21 so ist nämlich

$$\frac{P_{03}}{P_{02}} = \frac{P_{01}P_{12}P_{23}}{P_{01}P_{12}} = \frac{P_{13}}{P_{12}} = P_{23} \text{ da auch } P_{13}=P_{12}P_{23};$$

2) unmittelbar die Wachstumsrate (gegenüber der Vorperiode) abzulesen ist;

3) geltend gemacht wird, dass die in der gesamten Zeitreihe enthaltenen Information besser ausgenutzt werde als durch einen Zwei-Perioden-Vergleich;

4) der verkettbare Index (Kettenindex) laufend Veränderungen der Verbrauchsgewohnheiten berücksichtigen könne.

Die Forderung nach einem Kettenindex (chain based index im Unterschied zu fixed based index) ist gleichwohl nicht überzeugend, weil das Prinzip des reinen Preisvergleichs nicht erfüllt wird[1]. Beim internationalen Vergleich bedeutet die Transitivität der Paritäten: ein direkter Vergleich zweier Länder soll zum gleichen Ergebnis führen wie ein indirekter (über ein drittes Land), da sonst keine Eindimensionalität der Paritäten gegeben ist (Transitivität ist ja die Eigenschaft der Ordnungsrelation, also Bedingung dafür, dass Paritäten entlang einer Dimension angeordnet werden können).

Beispiel 10.8:

Es sei der in folgender Tabelle beispielhaft dargestellte Warenkorb, bestehend aus Wasser, Bier und Milch mit den Preisen (pro l) und die Pro-Kopf Verbräuchen (in l) in den Jahren von 1988 bis 1991 gegeben.

[1] vgl. von der Lippe, P.: Wirtschaftsstatistik, a.a.O.

	1988		1989		1990		1991	
	q	p	q	p	q	p	q	p
Wasser	300	0,80	310	0,85	315	0,90	320	0,95
Bier	220	1,80	220	1,90	225	2,00	230	2,10
Milch	110	1,20	120	1,30	130	1,40	135	1,45

Man zeige anhand dieses Beispiels, dass der Laspeyres- und der Paasche-Preisindex nicht verkettbar sind.

Lösung 10.8:

Man erhält folgende Laspeyres-Indizes:

$P_{01}^{L} = 1{,}0625$; $P_{12}^{L} = 1{,}0591$; $P_{23}^{L} = 1{,}0489$; $P_{03}^{L} = 1{,}180339$.

Setzt man nun diese Werte in Gleichung (10.21) ein, ergibt sich als Produkt der Verkettung $P_{03}^{*} = 1{,}0625 \cdot 1{,}0591 \cdot 1.0489 = 1{,}180320$. Da P_{03}^{*} verschieden von P_{03}^{L} (=1,180339) ist, erfüllt P^{L} und, wie leicht zu zeigen ist, auch P^{P} die Zirkularität nicht.

zu 3: Faktorumkehrbarkeit (Faktorumkehrprobe)

Die Faktorumkehrprobe ist vor allem damit motiviert, dass für einen einzelnen Wert und die entsprechenden Messzahlen jederzeit gilt, dass ein Wert (bzw. einer Wertmesszahl) das Produkt aus Menge und Preis ist. Aber, was für eine einzelne, die i-te Ware gilt, nämlich

$$\frac{p_{it}q_{it}}{p_{i0}q_{i0}} = \frac{p_{it}}{p_{i0}} \cdot \frac{q_{it}}{q_{i0}}$$

muss nicht notwendig auch für ein Aggregat von allen n Waren, also auf der Ebene der Indizes gelten. Die Zerlegbarkeit der Wertsteigerung eines Aggregats in eine Komponente des reinen Preis- und eine des reinen Mengeneinflusses, so dass für Indizes

(10.22) $W_{0t} = P_{0t}Q_{0t}$ (Faktorumkehrprobe)

gilt, ist ein Hauptproblem der Indextheorie.

Diese Zerlegbarkeit der Wertsteigerung in eine Preis- und Mengenkomponente ist das Ziel der Faktorumkehrprobe. Weder der Laspeyres- noch der Paasche-Preisindex erfüllen die Faktorumkehrprobe. Für Laspeyres-Indizes gilt

(10.18) $W = P^{L}Q^{L} + C$,

wobei - wie oben gezeigt - die Größe C die Kovarianz zwischen Preis- und Mengenmesszahlen darstellt. Es gilt jedoch Gl. 10.14 also $W = P^L Q^P = P^P Q^L$. In diesem Sinne ist die Laspeyres-Formel die "factor antithesis" (I. Fisher) der Paasche Formel. Fishers Ideal-Index P^F und der analog definierte Mengenindex Q^F erfüllen jedoch den "factor reversal" test (Faktorumkehrprobe).

d) Nutzenindex

Aus der mikroökonomischen Theorie ist das Problem bekannt, wie ein Haushalt bei gegebener Nutzenfunktion $U(q_1,...,q_n)$ und gegebenen Preisen $p_1,...,p_n$ einen bestimmten Nutzen U_0 mit minimalen Ausgaben erreichen kann. Die so im Haushaltsgleichgewicht eindeutig bestimmten Ausgaben R sind eine Funktion des Nutzens U_0 und des Preisvektors. Verschiedene Preisvektoren $\mathbf{p}_0$ und $\mathbf{p}_t$ führen zu unterschiedlichen Güterkombinationen (Mengenvektoren $\mathbf{q}_0$ und $\mathbf{q}_t$), die jedoch bei gleichem Nutzen auf einer Indifferenzkurve (bei n=2 Gütern), bzw. allgemein, einer Indifferenzfläche (n-1 dimensionale Hyperebene) liegen. Mit diesen Vorbemerkungen kann man den Nutzenindex definieren, der im Zentrum der ökonomischen Theorie der Preisindizes steht.

Def. 10.5: Nutzenindex

Der Nutzenindex (constant utility index, true cost of living index) ist das Verhältnis der bei verschiedenen Preisen für den gleichen Nutzen erforderlichen minimalen Ausgaben:

$$(10.23)\quad P^N_{0t}(U_0) = \frac{R(U_0,\mathbf{p}_t)}{R(U_0,\mathbf{p}_0)}$$

Bemerkungen zu Definition 10.5

1. Man beachte: Gl. 10.23 ist die Definition eines theoretischen Konstrukts, nicht aber eine operationale Rechenvorschrift, um den Nutzenindex empirisch zu bestimmen. Der Nutzenindex mißt die Veränderung der Kosten, die zur Aufrechterhaltung eines gegebenen Nutzenniveaus erforderlich sind (daher auch "true cost of living index"), was in der Regel bedeutet, dass Preis- und Mengenvektor veränderlich sind, während letzterer ja beim Laspeyres- und Paasche-Preisindex für die Vergleichsperioden gleich ist. Reiner Preisvergleich bedeutet bei P^L und P^P gleiche **Mengen**, bei P^N gleicher **Nutzen**. Die zum gleichen Nutzen führenden Mengen sind in verschiedenen Preissituationen nicht gleich: Preisänderungen lösen einen Substitutionseffekt aus. P^N mißt die Einkommensentschädigung für eine durch den Substitutionseffekt entstandene Ausgabenveränderung.

2. Da die Nutzenfunktion $U(\mathbf{q})$ des Gütervektors $\mathbf{q}$ und die hieraus abgeleitete Ausgabenfunktion $R(U,\mathbf{p})$ nicht empirisch bestimmbar sind, ist der Nutzenindex nur ein

theoretisches Konstrukt; P^N ist nicht tatsächlich nach Gl. 10.21 berechenbar. Insbesondere führt auch der Ausgabenvergleich auf der Basis des Nutzens U_t

$$(10.24)\quad P^N_{0t}(U_t) = \frac{R(U_t,\mathbf{p}_t)}{R(U_t,\mathbf{p}_0)}$$

i. d. R. nicht zum gleichen Ergebnis wie Gl. 10.23.

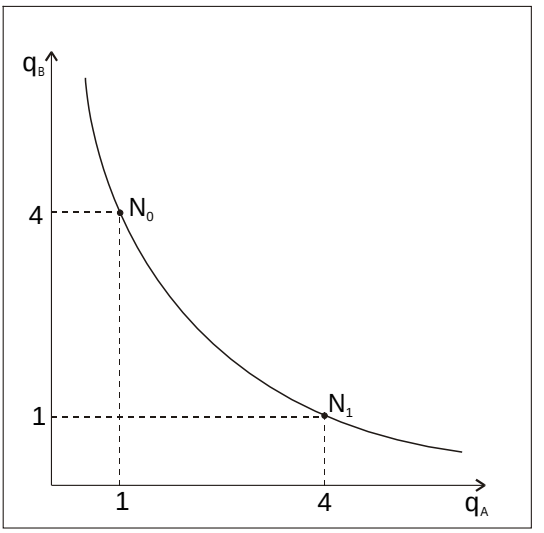
Beispiel 10.9:

Ein Haushalt habe eine Indifferenzkurve (vgl. Abb. 10.1) in der Art, dass er sich bezüglich der Kombinationen $q_A = 1$ und $q_B = 4$ zur Zeit t = 0 (Punkt N_0 auf der Indifferenzkurve) und $q_A = 4$ und $q_B = 1$ (zur Zeit t = 1, Punkt N_1) indifferent verhält. Für die Preise der beiden Güter möge zu den beiden Zeitpunkten gelten:

Gut	t=0	t=1
A	4	2
B	2	4

Berechnen Sie den Preisindex nach Laspeyres und nach Paasche sowie die Zunahme der Gesamtausgaben (Kostenindex)!

Abb. 10. 1



Lösung 10.9:

Die Zahlen und die Gestalt der Indifferenzkurve sind so gewählt, dass die Ausgaben für die Güterkombinationen N_0 und N_1 gleich (nämlich 12) sind. Der Kostenindex (Wertindex), der in diesem Fall zugleich ein Nutzenindex (P^N) ist (da N_0 und N_1 auf einer

Indifferenzkurve liegen) ist also 12/12 = 1. Für P^L erhält man 18/12 = 1,5 und für P^P = 12/18 = 2/3 so dass $P^L > P^N > P^P$. Die hier dargestellte Situation zeigt auch, warum bei einem (vom Ursprung aus gesehen) konvexen Verlauf der Indifferenzkurve $P^L > P^P$ sein muss.

Die Bilanzgerade zur Zeit t=0 tangiert die Indifferenzkurve in N_0. Sie hat die Funktion q_B

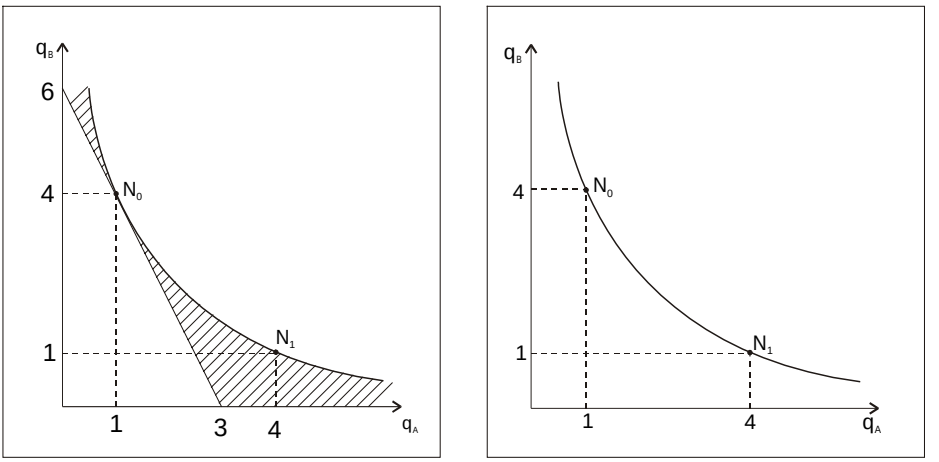
$= 6\text{-}2q_A$. Jede Güterkombination auf dieser Geraden führt zu Ausgaben in Höhe von 12 bei den Preisen des Preisvektors $\mathbf{p}_0$ also $\Sigma p_0 q_0 = 12$. Dass sich die Indifferenzkurve von der Bilanzgeraden (Iso-Ausgabenkurve) entfernt, was in Abb. 10.2 (links) durch Schraffur angedeutet wird, bedeutet dass bei gegebener Menge q_A die Menge q_B auf der Indifferenzkurve größer sein muss als auf der Bilanzgeraden, also $q_{Bt} > 6 - 2q_{At}$ so dass $\Sigma p_0 q_1 > \Sigma p_0 q_0$.

Entsprechendes gilt für die Ausgaben zur Zeit 1. Die konstante Ausgabe von 12 bei Preisen von t bedeuten $q_B = 3 - ½\ q_A$ (Geradenfunktion, Abb. 10.2, rechts). Es gilt $\Sigma\ p_1 q_1$

$< \Sigma p_1 q_0$, da ja der Punkt M (mit $q_A = 1$ und $q_B = 3 - ½ = 2{,}5$) unter dem Punkt N_0 (q_A=1
und $q_B = 4$) liegt.

Aus $\Sigma p_0 q_1 > \Sigma p_0 q_0$ und $\Sigma p_1 q_1 < \Sigma p_1 q_0$ folgt dass $P^P < P^L$, denn der Nenner von P^P ist
größer und der Zähler kleiner als der von P^L. Erheblich komplizierter wird die Betrachtung, wenn verschiedene Nutzenniveaus verglichen werden.
Es gilt dann $P^P < P_{0t}{}^N(U_t)$ und $P_{0t}{}^N(U_0) < P^L$.

Abb. 10.2



4. Besondere Rechenoperationen mit Indizes

a) Umbasierung und Verkettung

Von Zeit zu Zeit ist es notwendig einen Index von einer alten Indexbasis (0) auf eine neue (aktuellere) Indexbasis (s) umzustellen. Dabei ist zu unterscheiden:

a) Ist die Umstellung mit einer Revision der Reihenauswahl (d.h. bei einem Preisindex: des Preisvektors) und/oder des Wägungsschemas, also der Gewichtung verbunden, so spricht man von einer **Neuberechnung**.

b) Wird der Index dagegen nur von der Basis 0 auf die Basis s mit einer einfachen Rechenoperation (i.d.R. mit dem "Dreisatz") umgerechnet, so spricht man von **Umbasierung**.

Es wird im folgenden (nur zur Veranschaulichung) davon ausgegangen, dass für die Perioden gilt $0 < s < t$ (für die Formeln ist diese Annahme nicht nötig), denn die praktisch relevanten Fälle einer Umbasierung sind meist:

- die Umstellung eines Indexes von einer weiter zurückliegenden Basis (0) auf eine aktuellere (s) oder
- der Vergleich mehrerer Indizes mit verschiedener Basisperiode so, dass alle Indizes die gleiche Basis haben, meist diejenige des Indexes mit der neuesten Basis.

Def. 10.6: Umbasierung

Die Umbasierung eines Indexes (z.B. eines Preisindexes P) mit der Basis 0 auf die Basis s erfolgt mit

$$(10.25)\quad P_{st} = \frac{P_{0t}}{P_{0s}} \quad .$$

Bemerkungen zu Def. 10.6:

1. Werden die Indizes in Prozent ausgedrückt, so ist die rechte Seite der Gl. 10.25 mit 100 zu multiplizieren.

2. Da P_{0s} eine Konstante ist, ist die neue Indexreihe P_{st} ein konstantes Vielfaches (Proportionalitätsfaktor P_{0s}^{-1}) der alten Indexreihe P_{0t}. Denn Gl. 10.25 geht von der Annahme der Proportionalität aus. Löst man $P_{st}/P_{0t} = P_{ss}/P_{0s}$ nach P_{st} auf (unter Berücksichtigung von $P_{ss} = 1$) so erhält man Gl. 10.25.

3. Strenggenommen darf eine Umbasierung nach Gl. 10.25 nur durchgeführt werden, wenn die Zirkularität erfüllt ist, denn es ist leicht zu sehen, dass die Umbasierung nach Gl. 10.25 nur eine Umformung der Gl. 10.21 für die Verkettung darstellt.
   Bei einer Umbasierung eines (nicht verkettbaren) Laspeyres-Indexes, wird P_{st} aus Gl. 10.25 i.d.R. nicht mit dem direkt errechneten Ergebnis von P_{st}^{L} übereinstimmen (vgl. Bsp. 10.10). Setzt man in Gl. 10.25 die Laspeyres-Formel ein, so erhält man

$$\frac{P_{0t}}{P_{0s}} = \frac{\frac{\Sigma p_t q_0}{\Sigma p_0 q_0}}{\frac{\Sigma p_s q_0}{\Sigma p_0 q_0}} = \frac{\Sigma p_t q_0}{\Sigma p_s q_0} \quad \text{statt} \quad P_{st} = \frac{\Sigma p_t q_s}{\Sigma p_s q_s}$$

Die beiden Ergebnisse (Umbasierung und direkte Berechnung) werden sich nur dann wenig unterscheiden, wenn die Struktur der Warenkörbe zur Zeit 0 und zur Zeit s ähnlich ist.

Def. 10.7: Verkettung, splicing

Eine Verkettung von Indexwerten ist die Bildung einer langen Reihe zur Basis 0 aus mehreren verschiedenen (mit verschiedenen Basisperioden) sich überlappenden Indizes unter der Annahme der Proportionalität nach der folgenden Gleichung:

(10.25a) $P_{0t} = P_{0s}P_{st}$ (wobei meist gilt $0 < s < t$) oder
$P_{0t} = P_{0r}P_{rs}P_{st}$ (wenn $0 < r < s < t$)
oder von Periode zu Periode
$P_{0t} = P_{01}P_{12}....P_{t-1,t}$.

Soll die durch Verkettung errechnete Reihe mit einer originär aus den Daten für Preise und Mengen errechneten Reihe übereinstimmen, so ist Verkettbarkeit der Indexformel vorauszusetzen, was für die meisten Indexformeln aber nicht zutrifft.

Bemerkungen zu Def. 10.7:

1. Da Verkettung die Umkehrung der Umbasierung darstellt gelten die Bemerkungen zu Def.10.6 auch hier.

2. Die typische Aufgabenstellung, die eine Verkettung nahelegt ist in Beispiel 10.11 dargestellt. Mehrere Reihen werden i.d.R. zu einer einzigen Indexreihe zusammengefaßt, weil man an der Entwicklung über einen größeren Zeitraum interessiert ist, oder Bruchstellen vermeiden möchte.

3. In der Literatur wird gelegentlich von Verknüpfung gesprochen oder ein Unterschied zwischen Verknüpfung und Verkettung konstruiert, der jedoch weder formal, noch von der Fragestellung her sinnvoll gemacht werden kann. In jedem Fall wird eine Proportionalität aller Indexreihen angenommen, gleichgültig, ob man eine lange Reihe auf der Grundlage einer "alten" Basis errechnet (also eine alte Indexreihe fortführt), oder ob die lange Reihe auf der Grundlage einer "neuen" Basis gebildet werden soll (ob man also die neue Indexreihe zurückrechnet).

Beispiel 10.10:

Man basiere im Bsp. 10.8 den Laspeyres-Preisindex vom Basisjahr 0 auf das Basisjahr 1 um und vergleiche das Ergebnis nach Gl.10.25 mit dem aus den Daten direkt errechneten Ergebnis für $P^{L}_{,1t}$!

Lösung 10.10:

Ausgehend von Beispiel 10.8 erhält man folgende Laspeyres-Indizes:
$P^{L}{}_{,01} = 1{,}0625$; $P^{L}{}_{,02} = 1{,}1250$; $P^{L}{}_{,03} = 1{,}1803$; $P^{L}{}_{,13} = 1{,}1110$.
Wird gem. Gl. 10.25 auf die Indexbasis 1 umgestellt, so erhält man $P^{*}{}_{,13} = 1{,}1803/1{,}0625 = 1{,}1108$ statt 1,1110.

Beispiel 10.11:

Gegeben seien Indizes zur Basis 1980, 1985 und 1990

Jahr	Index A	Index B	Index C
1980	100		
1985	120	100	
1986	125	105	
1987		109	
1988		112	
1989		116	98
1990		118	100
1991			103
1992			105

Der Index A wurde ab 1986 nicht mehr fortgeführt, andererseits wurde der Index B nicht für die Zeit vor 1985 zurückgerechnet und der Index C nur für das Jahr 1989 zurückgerechnet. Berechnen Sie eine lange Indexreihe

a) zur Basis 1980 (Fortführung des alten Index A)
b) zur Basis 1990 (Rückrechnung des neuen Index C).

Lösung 10.11:

Die durch Verkettung errechneten Indexwerte haben jeweils das Symbol *.

zu a) Fortführung des Indexes A (oder B)

a1) (Verkettungsperiode jeweils das Basisjahr (also 1985, 1990):

Jahr	Index A*	(= Produkt·100)	Index B*	(=Produkt·100)
1986	126	(=1,20·1,05)		
1987	130,8	(=1,20·1,09)		
1988	134,4	(=1,20·1,12)		
1989	139,2	(=1,20·1,16)		
1990	141,6	(=1,20·1,18)		
1991	145,84	(=1,416·1,03)	121,54	(=1,03·1,18)
1992	148,68	(=1,416·1,05)	123,9	(=1,05·1,18)

a2) andere Möglichkeiten der Verkettung:

Liegt eine Überlappung der Indizes um mehr als eine Periode vor, so kann sich zeigen, dass die Annahme der Proportionalität evtl. gar nicht zutreffend sein muss und man könnte auch eine andere Periode der Verkettung zugrundelegen: Die Zunahme des Indexes A von 1985 auf 1986 beträgt 4,17% (von 120 auf 125), die des Indexes B dagegen 5% (von 100 auf 105). Wird das Jahr 1985 zur Verkettung von A und B benutzt, so muss die fortgeführte Reihe A* für 1986 demnach den Wert 126, statt 125 annehmen (126 ist 5% mehr als 120). Verkettet man den Index A und B zwecks Fortführung des Indexes A zu A* auf der Basis des Jahres 1986 (statt 1985) so ergäbe sich für den fortgeführten Index A*

Jahr	Index A*	statt oben	zu errechnen aus:
1987	129,76	130,8	109(125/105)
1988	133,33	134,4	112(125/105)
1989	138,10	139,2	116(125/105)
1990	140,48	141,6	118(125/105)

b) Rückrechnung des Indexes C durch Verkettung

(d.h. aufgrund der Proportionalität mit Index B in der Zeit 1985-1989 und aufgrund der Proportionalität mit Index A in der Zeit vor 1985)

Jahr	Index C*	(=Produkt 100)
1980	83,333	100(100/120)
1985	84,745	100(100/118)
1986	88,983	105(100/118)
1987	92,373	109(100/118)
1988	94,915	112(100/118)
1989	98,305	116(100/118)

Wie der tatsächliche Wert von Index C für 1989 (nämlich 98) zeigt, ist die Proportionalitätsannahme nicht gerechtfertigt. Man könnte auch hier eine alternative Rückrechnung aufgrund des Stands von 1989 (statt 1990) vornehmen.

b) Aggregation von und Zerlegung in Teilindizes

Ein Index sollte nach einer einfachen Formel in Teilindizes zerlegbar sein. Die Aggregation von Indizes kann an einem einfachen Beispiel gezeigt werden. Angenommen, es seien zwei Teilindizes (Sektorenindizes) zu bilden, der erste aus den Waren A und B, der zweite aus den Waren C und D. Dann gilt für die Laspeyres-Formel

$$P^L = (g_A + g_B)P^L{}_{,1} + (g_C + g_D)P^L{}_{,2} = g_1P^L{}_{,1} + g_2P^L{}_{,2}.$$

Der Gesamtindex ist also ein gewogenes Mittel der Sektorenindizes ($P^L{}_{,1}$ und $P^L{}_{,2}$), wobei als Gewichte die Summen der Ausgabenanteile (g) der in den Sektorenindizes zusammengefaßten Waren an den Ausgaben für alle Waren des Gesamtindexes auftreten. Für die Paasche-Formel ist entsprechend ein harmonisches Mittel zu verwenden mit den aggregierten Gewichten der Berichtsperiode.

Bei der Aggregation von Sektorenindizes zum Gesamtindex gelten also die gleichen Beziehungen (Art des Mittels und der Gewichtung) wie bei der Berechnung eines Indexes aus den Messzahlen. Dies soll im folgenden Beispiel (Bsp.10.12) verifiziert werden.

Beispiel 10.12:

In der oben angegebenen Weise (Zusammenfassung der Waren A und B in dem ersten - und der Waren C und D in dem zweiten Sektorenindex) ist im Beispiel 10.1 zu verfahren. Es sind Laspeyres- und Paasche-Preisindizes für das gesamte Aggregat (alle vier Waren) und für die beiden Teile (Sektoren) zu bilden.

Lösung 10.12:

Gesamtaggregate (vgl. Lösung 10.1) $P^L = 545/370 = 1{,}47297$ und $P^P = 975/665 = 1{,}466165$. Es soll nun gezeigt werden, wie sich die Gesamtindizes aus Teil- (Sektoren) indizes "zusammensetzen".

a) Teilaggregate (Sektorenindizes) nach Laspeyres
   Sektor 1: $P^L{}_{,1} = (3\cdot25 + 8\cdot20)/(2\cdot25 + 4\cdot20) = 235/130 = 1{,}80769$
   Sektor 2: $P^L{}_{,2} = (9\cdot30 + 4\cdot10)/(7\cdot30 + 3\cdot10) = 310/240 = 1{,}29167$

Die aggregierten Ausgabenanteile zur Basiszeit lauten 130/370 = 0,351 und 240/370 = 0,649, so dass gilt $P^L = 0{,}351\ P^L{}_{,1} + 0{,}649\ P^L{}_{,2} = 1{,}47297$.

b) Die entsprechende Berechnung nach Paasche: zunächst wieder die Berechnung der Sektorenindizes

Sektor 1: $P^P{}_1 = (3\cdot 50 + 8\cdot 30)/(2\cdot 50 + 4\cdot 30) = 390/220 = 1{,}7727$

Sektor 2: $P^P{}_2 = (9\cdot 25 + 4\cdot 90)/(7\cdot 25 + 3\cdot 90) = 585/445 = 1{,}3146$

aggregierte Ausgabenanteile zur Berichtszeit 390/975 = 0,4 und 585/975 = 0,6

$(P^P)^{-1} = 0{,}4(1{,}7727)^{-1} + 0{,}6(1{,}3146)^{-1} = 88/390 + 267/585 = 0{,}6820513 = (1{,}466165)^{-1}$, so dass $P^P = 1{,}466165$.

Kapitel 11: Einführung in die Zeitreihenanalyse

1. Gegenstand und Methoden der Zeitreihenanalyse

a) Zeitreihen und Zeitreihenanalyse

Die Zeitreihenanalyse beschäftigt sich mit Methoden zur Beschreibung von Daten, die zeitlich geordnet sind (d.h. als Zeitreihen vorliegen), bzw. allgemein, bei denen die Reihenfolge der Beobachtungen wesentlich ist.

Die Zeit hat eine natürliche, irreversible Ordnung. Wie immer das Kontinuum "Zeit" durch Zeitpunkte t_0, t_1,... strukturiert wird, es kann nie zweifelhaft sein, ob t_0 zeitlich vor oder nach t_1 kommt. Die Reihenfolge von Beobachtungen (Daten) ist eindeutig und im Falle der Zeitreihenanalyse für die statistischen Berechnungen auch wesentlich (zu berücksichtigen).

Zeitreihen als Daten werden auch in Kap. 9 und 10 (Meßzahlen, Indizes und Wachstumsraten) sowie in Kap. 12 (Zu- und Abgänge) betrachtet. Im Unterschied zu diesen Methoden geht es bei der Zeitreihenanalyse jedoch nicht um den Vergleich von Zuständen zu unterschiedlichen Zeitpunkten oder -intervallen, sondern um die Darstellung von Abläufen (Prozessen) zwischen Zuständen, d.h. um Vorgänge, die sich in Phasen gliedern lassen oder allgemein als Funktion der Variablen Zeit zu begreifen sind.

Das **Auswertungsziel** der Beschreibung von Abläufen hat, gerade für Ökonomen, große Bedeutung im Rahmen der folgenden praktischen Aufgaben:

- Prognose,

- Analyse, d.h. "Verstehen" im Sinne von Erkennen von Ursachen, u.a. auch durch Beschreibung der bisherigen Entwicklung,
- Kontrolle, d.h. Steuerung oder Regelung von Prozessen, um Abweichungen der Ist- von der Sollgröße auszugleichen.

Hierfür und insbesondere mit Blick auf ökonomische Zeitreihen ist das sog. "Komponentenmodell" der Zeitreihenanalyse ein wichtiges Instrument. Die traditionellen (älteren) und "einfacheren" Verfahren der Zeitreihenanalyse, die allein im folgenden behandelt werden können, gehen von dieser Modellvorstellung aus.

Def. 11.1: Zeitreihe, Ursprungswerte

Eine Folge von Beobachtungswerten y_t mit t = 1,2,...,T und einer natürlichen Ordnung dergestalt, dass die Werte in der Reihenfolge y_1, y_2,... beobachtet wurden, heißt "Zeitreihe". Die (meist diskrete) Variable t ist i.d.R. die Zeit, wobei die Werte t Zeitpunkte oder Zeitintervalle darstellen. Die noch nicht durch eine Zeitreihenanalyse bearbeiteten (z.B. transformierten) Beobachtungswerte y_t heißen "Ursprungswerte".

Bemerkungen zu Def. 11.1:

1. Man kann hinsichtlich der Variablen Y und der Variablen t verschiedene Arten von Zeitreihen unterscheiden:
   - Die Zeitvariable t wird normalerweise als diskret mit äquidistanten Einteilungen (gleichlange Intervalle) vorausgesetzt und die Datenpunkte im y - t - Koordinatensystem (Graph einer Zeitreihe) werden dann meist (nur zur Erleichterung der Augenführung) linear miteinander verbunden.
     Die Variable t kann aber auch stetig sein. Im zeitdiskreten Modell soll die Schreibweise y_t und im stetigen Modell y(t) gelten.
   - Die Beobachtungswerte y_t sind meist Ausprägungen einer metrisch skalierten Variable Y. Die Variable Y kann aber z.B. auch dichotom sein (0-1-Variable) und die Abgabe oder Abwesenheit eines *Impulses* bedeuten. Für die Analyse ist bei derartigen Prozessen dann der zeitliche Abstand zwischen je zwei Impulsen und nicht der Betrag von y (der ja nur 0 oder 1 sein kann) von Interesse.

2. Beispiele für Zeitreihen sind täglich oder stündlich gemessene Aktienkurse, jährliche Werte des Sozialprodukts, eine Fieberkurve oder

Messungen von Luftdruck und Temperatur, die auch kontinuierlich (stetig) erfolgen können.

3. Der **Graph einer Zeitreihe** kann evtl. bereits erste grobe Aufschlüsse über die Bestimmungsfaktoren der Entwicklung, über Regelmäßigkeiten und mögliche Brüche oder Ausreißer liefern und sollte Ausgangspunkt jeder Zeitreihenanalyse sein.

4. Wesentlich ist für eine Analyse ökonomischer Zeitreihen, dass die Variable Y zu allen Beobachtungszeitpunkten sachlich und räumlich gleich abgegrenzt ist. Ändert sich die Definition der Größe Y oder das Mess- bzw. Erhebungsverfahren, so ist ein **Bruch** gegeben und eine zusammenhängende Interpretation der Zeitreihe und ihrer Komponenten i.d.R. nicht mehr sinnvoll.

5. Unter **Konversion** soll eine Änderung der Periodizität der Daten verstanden werden.

   So können z.B. Monatsdaten zu Quartalsdaten konvertiert werden, indem man für den Quartalswert

   - den Wert des letzten (dritten) Monats des Quartals (z.B. bei Bestandsgrößen) oder
   - einen mittleren Wert (zweiter Monat) oder die Summe der drei Monatswerte eines Quartals heranzieht (z.B. bei Stromgrößen),
   - oder bei Bestandsgrößen ein chronologisches Mittel (Kap.12).

   Ähnlich lassen sich Quartalsdaten zu Jahresdaten konvertieren. Die Umkehrung (Vergrößerung der Periodizität), z.B. die Bestimmung von Monats- aus Quartalsdaten, verlangt Interpolationen von Daten , die i.d.R. nicht zu vertreten sind.

6. Die unter Nr. 5 beschriebene Konversion einer Zeitreihe ist eine spezielle **Transformation** der Ursprungswerte y_t (bzw. y(t)), nämlich eine abschnittsweise Aggregation (Summation) oder Mittelwertbildung.

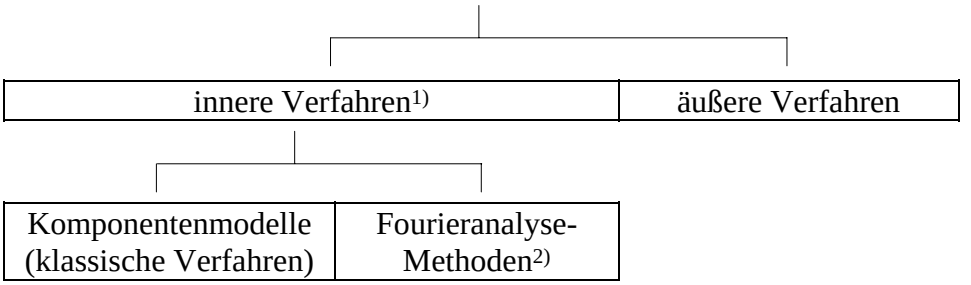
   Andere Transformationen sind algebraische Operationen (z.B. Differenzenbildung) oder funktionale Transformationen $z_t = f_d(y_t)$, bzw. $z(t) = f_s[y(t)]$ im diskreten, bzw. stetigen Modell. Transformationen werden durchgeführt, um die Varianz der transformierten Zeitreihe zu verringern oder um zu einer bestimmten Verteilung und Art der Überlagerung der Komponenten im transformierten Modell z_t bzw. z(t) zu gelangen. Sehr beliebt ist die Box-Cox-Transformation der Reihe y(t) in die Reihe z(t,k,c) wenn y(t) > -c:

$$z(t,k,c) = \begin{cases} \frac{1}{k}\{[y(t)+c]^k - 1\} & \text{wenn } k \neq 0 \\ \ln[y(t)+c] & \text{wenn } k = 0 \end{cases}$$

b) Methoden der Zeitreihenanalyse

Einfache **deskriptive (heuristische) Methoden** wie Graphiken, Vorjahresvergleiche, Berechnung von absoluten Zuwächsen (Differenzen) oder relativen Zuwächsen (Wachstumsraten) sowie Autokorrelationen usw. können hilfreich sein zur Beschreibung der Daten und für die Suche nach einem adäquaten Modell für das Zustandekommen dieser Daten. Weitergehende Erklärungen und auch Prognosen setzen i.d.R. eine **Modellbildung** voraus: Nach Art des Modells lassen sich verschiedene Zeitreihenanalyseverfahren unterscheiden (Übers. 11.1).

Übersicht 11.1: Grundlegende Ansätze der Zeitreihenanalyse[*)]*



*) Man unterscheidet auch univariate und multivariate Verfahren, je nachdem ob die zu beschreibende Zeitreihe y_t ein- oder mehrdimensional ist (ein Vektor mit den Daten y_{1t} , y_{2t} ,..); innere Verfahren sind nicht immer nur univariat. Hier sollen nur uni-variate Verfahren betrachtet werden.

1) Hierzu gehören auch die "Box-Jenkins-Methoden" der Zeitreihenanalyse auf die hier nicht näher eingegangen werden kann.

2) harmonische Analyse, Spektralanalyse.

Erläuterungen zur Übersicht 11.1:

1. Innere Methoden erklären eine Zeitreihe y_t allein als Funktion der Zeit oder früherer Werte y_{t-d} (Lag $d > 0$) der gleichen Zeitreihe; äußere Methoden ziehen auch andere Variablen (etwa x_t, x_{t-d}, z_t usw.) zur Erklärung heran und sind u.a. Gegenstand der Ökonometrie.

2. **Komponentenmodelle** interpretieren eine Zeitreihe y_t als Überlagerung einfacher Funktionen der Zeit, die formal aufgrund ihrer Peri-

odizität definiert sind und "Komponenten" genannt werden. Hierauf beruhende Verfahren der Zeitreihenanalyse eliminieren i.d.R. sukzessive, nicht simultan die einzelnen Komponenten, wobei die "Restkomponente" r_t meist als verfahrensbedingter Rest übrig bleibt (r_t ist auch meist nicht eine Zufallsvariable mit bestimmten a priori geforderten Eigenschaften).

3. "Fourieransätze" als Alternative zum "klassischen Verfahren" der sukzessiven Zerlegung einer Zeitreihe in Komponenten betrachten eine Zeitreihe (ohne die nichtzyklische [monotone] Trendkomponente) als Summe vieler Schwingungen verschiedener Frequenz und Amplitude, die "direkt" (simultan) geschätzt werden. Es gibt auch die Kombination klassischer und fourieranalytischer Methoden (ASA- und Berliner Verfahren). Auf Einzelheiten kann hier nicht eingegangen werden.

2. Das Komponentenmodell der Zeitreihenanalyse

Es wird a priori eine geringe Anzahl isolierbarer "Komponenten" vorausgesetzt, die jeweils einfache (und somit leichter [als die Ursprungswerte] inter- und extrapolierbare) Funktionen der Zeitvariable t sind. Diese Komponenten sind nicht nur durch formale Eigenschaften definiert, sondern auch inhaltlich (d.h. ökonomisch im Sinne länger- oder kürzerfristig wirkender Einflußfaktoren).

a) Beschreibung des Modells

Inhalt der Komponenten

Übersicht 11.2 stellt die seit K. Pearson üblicherweise in ökonomischen Zeitreihen unterschiedenen vier Komponenten dar. Formales Unterscheidungsmerkmal der Komponenten ist die Periodizität (Wellenlänge). Danach sollten von links nach rechts (von m_t bis r_t) Einflußfaktoren mit zunehmender Frequenz (abnehmender Wellenlänge) auftreten.

Der **Trend** als "glatte" Kurve ist Ausdruck von (relativ zur Länge der betrachteten Zeitreihe) "langfristigen" Einflußfaktoren die nicht periodisch sind. Die Bezeichnung des Trends mit m_t soll andeuten, dass der Trend formal als eine Folge bedingter *Mittelwerte* (bedingt durch die Zeitvariable t) aufgefaßt werden kann. **Zyklische** Komponenten, wie die Saisonkomponente und (früher meist auch) die Konjunktur sind demgegenüber (annähernd) periodische Funktionen f(t), wobei "periodisch" bedeutet $f(t + kp) = f(t)$ mit $k = 1,2,...$ und $p = 12$ Monate

(Saison) bzw. p = 4 oder p = 5 Jahre (Konjunktur, nicht genau festgelegte Periodenlänge).

Bemerkung zur Konjunktur

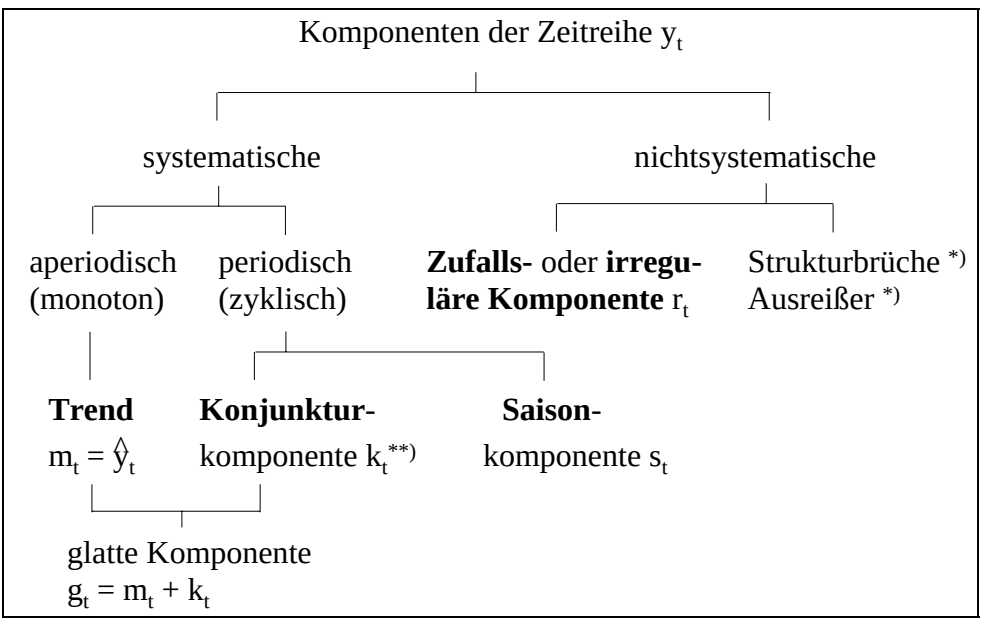
Früher wurde wie in Übers. 11.2 die Konjunktur meist als mittelfristige (Periode etwa 4 bis 5 Jahre, jetzt eher 2 bis 12 Jahre) aber nicht notwendig regelmäßige (zyklische) Schwingung angenommen. Heutzutage ist es üblich k_t nicht mehr als (annähernd) periodische Funktion anzusehen und die Konjunktur deshalb auch nicht mehr durch trigonometrische Funktionen, sondern lokal durch Polynome zu modellieren.

Kausalinterpretation

Die Komponenten sollen einige wenige (vermutete) Ursachenkomplexe darstellen, die auch kausal interpretiert werden können, auch wenn sie nur Funktionen der Zeitreihe sind. Es sollen nicht bloß Summanden, wie in Gl. 11.1 darstellen, die nicht zu interpretieren sind. Das ist übrigens ein Unterschied zwischen dem Komponentenmodell und anderen Verfahren, die ebenfalls eine additive Überlagerung nach Art der Gl. 11.1 voraussetzen.

Übersicht 11.2: Komponenten einer ökonomischen Zeitreihe

Die beobachteten Zeitreihenwerte y_t werden als Ergebnis des Zusammenwirkens folgender, den Ursachen nach verschiedener Komponenten aufgefaßt:



*) eigentlich keine Komponente

**) zur Konjunktur vgl. Bem. im Text oben

Komponente	ökonomische Beschreibung	formale Beschreibung
Trend	langfristige Niveauänderung, Wachstum [1)], Entwicklung, Grundtendenz	monoton steigende bzw. fallende Funktion, oder Polynom geringen Grades
Konjunktur	Schwankungen im Auslastungsgrad des Produktionspotentials	wurde früher (jetzt nicht mehr üblich) als zyklisch aufgefaßt
Saison	jahreszeitliche (z.B. Klima) und institutionelle Einflüsse[2)] (z.B. Steuer-, Ferientermine)	weitgehend regelmäßige Schwingung mit einer Periode (Wellenlänge) von (genau) einem Jahr

1) Wachstum (ökonomisch) = Zunahme des Produktionspotentials.

2) Es gibt Schwierigkeiten der Abgrenzung zwischen der Saisonkomponente und Kalenderunreglmäßigkeiten.

Danach ist beispielsweise:

- die Trendkomponente Resultat langfristig wirkender ("säkularer") Faktoren, wie z.B. Bevölkerungs- und Wirtschaftswachstum oder technischer Fortschritt;
- die Saisonkomponente oft Ausdruck des Einflusses der Jahreszeit (z.B. jahreszeitlich bedingte Schwankungen der Ernteerträge oder der Beschäftigung in der Bauwirtschaft), wenn dieser Einfluß wirksam ist (was z.B. nicht der Fall ist bei Löhnen und Gehältern die deshalb meist auch keine Saisonkomponente haben.

Dagegen hat die Zufallskomponente (irreguläre Komponente) entweder keine bekannten Ursachen oder sie ist Ergebnis (einer Vielzahl) einmaliger nicht vorhersehbarer (irregulärer) Einflüsse und wenig bedeutsamer Ereignisse (wobei jedoch Ereignisse, wie Streiks, Mißernten etc. eher "Ausreißer" darstellen).

Daten, Glatte Komponente

Die Berechnung der Saisonkomponente setzt unterjährige Daten voraus (z.B. Quartals-, Monatsdaten).

Trend- und Konjunktur sind nur unterscheidbar, wenn die Zeitreihe hinreichend lang ist (mindestens zwei Konjunkturzyklen umfaßt). Es ist deshalb oft sinnvoll, beide Komponenten zur sog. glatten Komponente g_t (oder: Trend-Zyklus-Komponente) zusammenzufassen.

Verknüpfung (Überlagerung) der Komponenten

Man unterscheidet zwei Grundmodelle:

1) *additive Verknüpfung*: wenn bei steigendem bzw. fallendem Trend die zyklischen Einflüsse gleich große Ausschläge besitzen, so dass die

Schwankungen k_t, s_t und r_t vom Niveau der Zeitreihe y_t unabhängig sind. Es gilt dann:

(11.1)	$y_t = m_t + k_t + s_t + r_t$	(additive Überlagerung)

2) *multiplikative Verknüpfung*: wenn z.B. die Schwankungen der zyklischen Komponenten, insbesondere der Saisonkomponente um den Trend mit steigendem (sinkendem) Niveau der Zeitreihe zunehmen (abnehmen):

(11.2)	$y_t = m_t \, k_t \, s_t \, r_t$	(mutliplikative Überlagerung)

Möglich sind auch "gemischte" Modelle, etwa der folgenden Art:
$y_t = m_t + k_t \, s_t \, r_t$ oder $y_t = m_t \, k_t + s_t + r_t$.
Durch Logarithmieren kann eine mutiplikative Verknüpfung in eine additive Verknüpfung transformiert werden.

Kritik am Komponentenmodell

Es ist notwendig, von bestimmten Hypothesen auszugehen, deren Geltung nur z.T. überprüfbar ist:

1. Existenz einer a priori gegebenen (und im Zeitablauf gleichbleibenden) endlichen Zahl isolierbarer und sachlich interpretierbarer Komponenten.
2. Die Komponenten sind Funktionen der Zeit und nur dieser (!).
3. Auch die Art der Überlagerung (additiv, multiplikativ, gemischt) muss a priori festgelgt werden.

Die Komponenten sollten "unabhängig" wirken, wenngleich nicht verlangt wird, dass sie statistisch unabhängig (Kap. 7) sind.
Bei vier Komponenten (Trend, Konjunktur, Saison, Rest) sind aus T Werten (Länge der Zeitreihe y_t) 4T Unbekannte zu schätzen. Schon daran ist zu erkennen, dass die Zeitreihenanalyse weitergehende Annahmen braucht (z.B. Trendfunktion linear) und nicht ohne gewisse subjektive Entscheidungen auskommt um zu einer Lösung zu kommen, die darin besteht, dass vier Funktionen der Zeit überlagert die Ursprungswerte y_t "reproduzieren". Man kann sich immer auch einen Satz von vier Funktionen vorstellen, der sich den Daten genauso gut anpaßt. Es gibt keine eindeutige, vollautomatische Zeitreihenanalyse und keine "wahren" Komponenten, z.B. keinen "wahren" Trend sowie meist auch keine ökonomische Theorie, derzufolge z.B. der Trend linear sein müßte. Es gibt deshalb auch meist kein Validierungskriterium, an dem gemessen ein Trend "besser" ist als ein anderer.

Ein wohl nicht überwindbarer Grundwiderspruch ist: Die Zeitreihe müßte möglichst lang sein, um mit großer Sicherheit die Komponenten (Einflüsse) nachweisen zu können. Ist sie aber sehr lang, dann ist es wahrscheinlich, dass sich Art und Wirkungsweise der Einflüsse im Zeitablauf geändert haben.

b) Methoden der Trendbestimmung

Im folgenden sollen zwei einfache Methoden zur Trendermittlung, bzw. zur Bestimmung der glatten Komponente behandelt werden. Es sind die Methode:

- der kleinsten Quadrate und
- der gleitenden Durchschnitte.

Es empfiehlt sich, z.B. dann Trendeinflüsse zu eliminieren, wenn Zeitreihen miteinander korreliert werden sollen. So ist z.B. die Produkt-Moment-Korrelation zwischen den beiden Zeitreihen mit einem linearen Trend $x_t = a + bt + u_t$ und $y_t = c + dt + v_t$ (mit den Störgrößen u_t und v_t)

$$r_{xy} = \frac{bds_t^2 + s_{uv}}{\sqrt{(b^2 s_t^2+s_u^2)(d^2s_t^2+s_v^2)}} ,$$

so dass eine hohe Korrelation zwischen zwei vom Trend bestimmten Zeitreihen x_t und y_t praktisch eine Scheinkorrelation aufgrund der Variablen t sein kann, wenn $s_{uv} \approx 0$,weil dann $r_{xy} \approx r_{xt}r_{yt}$, wobei $r_{xt} = bs_t / \sqrt{b^2s_t^2 + s_u^2}$ und $r_{yt} = ds_t /\sqrt{d^2s_t^2 + s_v^2}$ ist (s_t^2 ist die Varianz der Variablen t, die natürlich eine Funktion der Länge T der Zeitreihe ist. Gilt t = 1,2,...,T, so ist $s_t^2 = \frac{T^2-1}{12}$).

1. Trendberechnung mit der Methode der kleinsten Quadrate

Bei einem Trend mit einer mathematischen Funktion bestimmten Typs (lineare-, Exponential-, Potenzfunktion usw.) können die Parameter nach der Methode der kleinsten Quadrate bestimmt werden. Der Regressand (die abhängige Variable) ist wie in Kap. 8 die Variable Y mit den Beobachtungen y_t und an die Stelle der unabhängigen Variable X tritt bei der Trendfunktion die Zeit t. Die übrigen Komponenten mit den Meßwerten k_t, s_t und r_t stellen das Residuum dar.

Die Parameter a und b eines linearen Trends $\hat{y}_t = m_t = a + b\cdot t$ (t = 1,2,...,T) werden mit den Normalgleichungen

(11.3a) $aT+ b\Sigma t= \Sigma y_t$

(11.3b) $a\,\Sigma t + b\,\Sigma t^2 = \Sigma t y_t$

bestimmt. Die Größen können beliebig "datiert" werden, d.h. die Größe t kann Werte wie etwa 1988, 1989, 1990, 1991 oder die Werte 1,2,...,T annehmen.

Zweckmäßig ist es, für t die Werte ... -2, -1, 0, +1, +2,... zu vergeben, so dass $\Sigma t = 0$ (statt $T(T+1)/2$ wie bei t=1,2,...,T) ist. Das geschieht indem man von den ursprünglichen t-Werten 0,1,2,... den Wert $\bar{t} = (T+1)/2$ abzieht. Dann sind a und b direkt aus jeweils einer der beiden Normalgleichungen

(1) $aT + b\Sigma t = aT = \Sigma y_t$ so dass $a = \frac{\Sigma y_t}{T} = \bar{y}$

(2) $a\Sigma t + b\Sigma t^2 = b\Sigma t^2 = \Sigma t y_t$ so dass $b = \frac{\Sigma t y_t}{\Sigma t^2}$

zu bestimmen.

Andere Trendfunktionen können durch Variablensubstitution (z.B. Parabel, Hyperbel) oder Variablentransformation (z.B. Logarithmustransformation) linearisiert werden (etwa die Exponentialfunktion oder die Potenzfunktion als Trend). Sämtliche in den Übersichten 8.2 und 9.5 zusammengestellten Funktionen sind als Trendmodelle denkbar.

Anders als bei der Methode der gleitenden Durschschnitte ist es hier erforderlich, eine Trendfunktion anzunehmen, die dem **Typ** nach zu spezifizieren ist (es ist also z.B. zu entscheiden, ob ein linearer oder parabolischer Trend bestimmt werden soll) und die Methode der kleinsten Quadrate erlaubt es, die **Parameter** dieser Funktion empirisch zu bestimmen. Für die Frage nach dem Funktionstyp mag bei **Polynomen** als Trend der folgende (im Abschn. 3 dieses Kapitels näher erläuterte) Zusammenhang hilfreich sein:

> Bei einem Polynom q-ten Grades sind die q-ten Differenzen konstant und die (q+1)ten Differenzen verschwinden. Dem entspricht (bei stetiger Zeitvariable t) der aus der Algebra bekannte Satz, dass q-maliges Differenzieren eines Polynoms vom Grade q zu einer Konstanten führt.

Folgerung: Bei einem polynomialen Trend kann man durch wiederholtes Differenzieren (im zeitstetigen Fall) bzw. durch wiederholte Differenzenbildung (im zeitdiskreten Fall) das Polynom (den Trend) annullieren (vgl. hierzu Abschn. 3b). Dieses Vorgehen liefert aber keine Quantifizierung des Trends, wie z.B. die beiden folgenden Verfahren (Abschn. b und c). Es ist nur eine Hilfe zu Erkennung des Polynomgrades.

Demonstrationsbeispiel: Differenzieren, stetige Zeit:
Parabel (Polynom vom Grade 2): $y = a + bt + ct^2$. Die erste Ableitung nach t ist b+2ct, die zweite 2c und die dritte somit Null.

Differenzenbildung, diskrete Zeit:

	Zeit(t)	0	1	2	3	4
Gerade	10+2t	10	12	14	16	18
Parabel	$10+2t+3t^2$	10	15	26	43	66

Die ersten Differenzen $y_t - y_{t-1}$ sind bei der Geraden: 12-10 = 14-12 = 16-14 = 18-16 = 2, also konstant 2, so dass die zweiten Differenzen verschwinden.
Bei der Parabel (Polynom vom Grade 2) sind die ersten Differenzen 15-10=5, 26-15=11, 43-26=17, 66-43=23 und die zweiten Differenzen sind konstant 6, denn 11-5=6, 17-11= 6 und 23-17=6.

Beispiel 11.1:

Beispiel für die Berechnung eines linearen Trends: Der Umsatz y_t (in Mio. DM) eines Unternehmens entwickelte sich in den Jahren 1984-1992 wie folgt:

Jahr	1984	1985	1986	1987	1988	1989	1990	1991	1992
t_0	0	1	2	3	4	5	6	7	8
y_t	6	9	11	12	13	15	18	20	23

Berechnen Sie den linearen Trend $m_t = a + bt$.
Hilfsangaben: $\Sigma y_t = 127$, $\Sigma t = 36$, $n = 9$, $\Sigma y_t^2 = 2029$, $\Sigma t^2 = 204$, $\Sigma t y_t = 626$.

Lösung 11.1:

$m_t = 6{,}24 + 1{,}97t$

2. Trendberechnung mit der Methode der gleitenden Durchschnitte

Gleitende Durchschnitte sind eine Folge von arithmetischen Mitteln, die aus jeweils p aufeinanderfolgenden Werten y_t der Zeitreihe gebildet werden.

Def. 11.2: Gleitende Durchschnitte

a) Der dem Ursprungswert y_t zugeordnete gleitende symmetrische p-gliedrige Durchschnitt lautet bei ungeradzahligem $p = 2k+1$

$$(11.4) \quad \tilde{y}_t = \frac{1}{p} \cdot \sum_{h=-k}^{h=k} y_{t+h} \qquad (p = 2k+1, \text{ ungeradzahlig})$$

$$= \frac{y_{t-k} + y_{t-k+1} + ... + y_t + ... + y_{t+k-1} + y_{t+k}}{2k+1}$$

b) Bei geradzahligem p = 2k wäre der Durchschnitt

$(y_{t-k} +... + y_t + ... + y_{t+k-1})/ 2k$ der Periode t - ½ und $(y_{t-k-1} +... + y_t + ... + y_{t+k})/ 2k$ der Periode t + ½ zuzuordnen. Es liegt daher nahe einen ungewogenen Durchschnitt hieraus zu berechnen. Dieser der Periode t zugeordnete ***zentrierte*** gleitenden Durchschnitt lautet:

$$(11.5) \quad \tilde{y}^z_t = \frac{1}{p} \cdot \left[\sum_{h=-(k-1)}^{h=k+1} y_{t+h} + \frac{y_{t-k} + y_{t+k}}{2} \right] =$$

$$= \frac{½ \cdot y_{t-k} + y_{t-k+1} + ... + y_t + ... + y_{t+k-1} + ½ \cdot y_{t+k}}{2k}$$

Erläuterung zu Teil b der Definition:

Beispiel: p = 4 (gleitender viergliedriger Durchschnitt bei Quartalswerten mit t = 0, t = 1,...) k = 2

$\tilde{y}^z_t = (½ \cdot y_{t-2} + y_{t-1} + y_t + y_{t+1} + ½ \cdot y_{t+2})/4$
da p = 2k = 4 ist. Der erste Wert der Folge zentrierter gleitender Durchschnitte ist dann:

$\tilde{y}^z_2 = (½ \cdot y_{2-2} + y_{2-1} + y_2 + y_{2+1} + ½ \cdot y_{2+2})/4$

$= (½ \cdot y_o + y_1 + y_2 + y_3 + ½ \cdot y_4)/4$

und die nachfolgenden Werte sind:

$\tilde{y}^z_3 = (½ \cdot y_1 + y_2 + y_3 + y_4 + ½ \cdot y_5)/4$

$\tilde{y}^z_4 = (½ \cdot y_2 + y_3 + y_4 + y_5 + ½ \cdot y_6)/4$

$\tilde{y}^z_5 = (½ \cdot y_3 + y_4 + y_5 + y_6 + ½ \cdot y_7)/4$ usw.

Dabei gehen je zwei Werte an den Rändern verloren.

Beispiel 11.2:

Die Entwicklung des Umsatzes y_t (in Mio. DM) einer Brauerei in den Jahren 1989-1991 nach Quartalen sei (vgl. Abb. 11.1):

Jahr	1989				1990				1991			
Quartal	I	II	III	IV	I	II	III	IV	I	II	III	IV
t	1	2	3	4	5	6	7	8	9	10	11	12
y_t	2	6	8	6	6	10	12	10	10	14	16	14

Berechnen Sie:
a) die gleitenden Durchschnitte 3. Ordnung (zu p = 3 Perioden) und
b) zentrierte gleitende Durchschnitte zu p = 4 Quartalen!

Lösung 11.2:

a) Gleitender Durchschnitt ungerader Ordnung mit p = 3:

Die Gl. 11.4 ist im Falle von p = 3

$\tilde{y}_t = (y_{t-1} + y_t + y_{t+1})/3 = \frac{1}{3}\Sigma\, y_{t+h}$ mit h = -1,0,+1.

Es gilt also

$\tilde{y}_2 = (2+6+8)/3 = 16/3 = 5{,}33$ (t = 2 bedeutet: 1989, 2.Quartal)

$\tilde{y}_3 = (6+8+6)/3 = 20/3 = 6{,}67$ (t = 3 bedeutet: 1989, 3.Quartal)

usw. bis schließlich

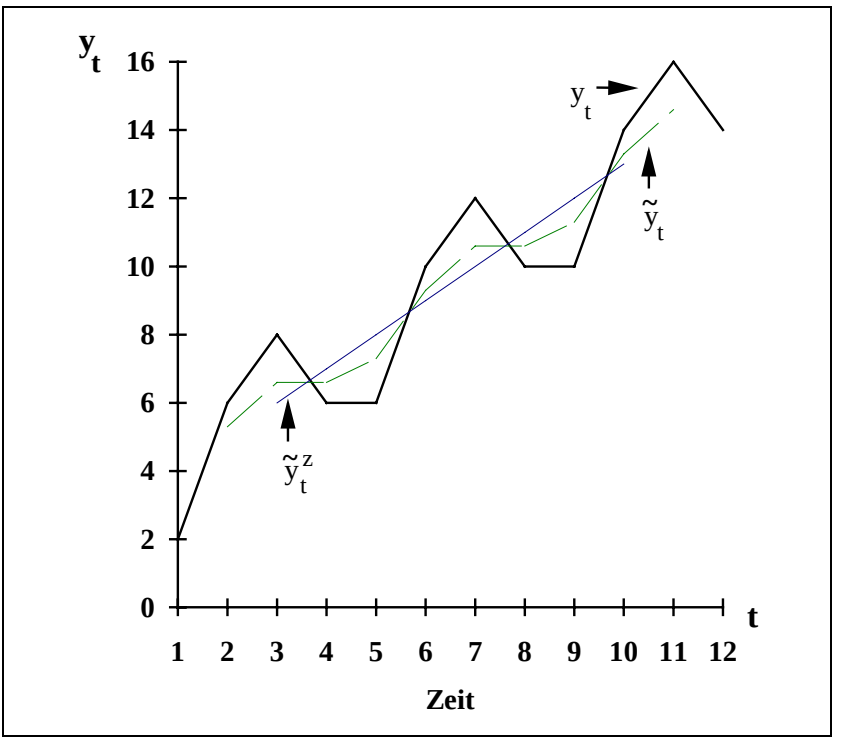
$\tilde{y}_{11} = (14+16+14)/3 = 44/3 = 14{,}67.$

Die folgende Tabelle enthält alle gleitenden Durchschnitte $\tilde{y}_t$ zur Ordnung p = 3 (von p = 3 Perioden) und die Ursprungswerte y_t sowie die unter b) zu berechnenden zentrierten 4-gliedrigen gleitenden Durchschnitte $\tilde{y}_t^z$:

Jahr und Quartal	t	y_t	$\tilde{y}_t$	$\tilde{y}_t^z$
89.1	1	2	*	*
89.2	2	6	5,33	*
89.3	3	8	6,67	6
89.4	4	6	6,67	7
90.1	5	6	7,33	8
90.2	6	10	9,33	9
90.3	7	12	10,67	10
90.4	8	10	10,67	11
91.1	9	10	11,33	12
91.2	10	14	13,33	13
91.3	11	16	14,67	*
91.4	12	14	*	*

Das Zeichen * soll bedeuten, dass hier verfahrensbedingt kein Wert zu berechnen ist, also Werte "verlorengehen".

Abb. 11.1: Ursprungswerte und Trends für Bsp. 11.2



b) Gleitender Durchschnitt gerader Ordnung

Bei einem gleitenden Durchschnitt gerader Ordnung p = 2k werden zur Zentrierung 2k+1 Perioden in die Mittelwertbildung einbezogen, wobei die erste und die letzte Periode nur mit dem halben Gewicht berücksichtigt wird.
Bei p = 2k = 4 ergibt sich für die zentrierten gleitenden Durchschnitte nach der allgemeinen Formel (Gl. 11.5):

$$\tilde{y}_t^z = [½y_{t-k} + y_{t-(k-1)} + .. + y_{t+(k-1)} + ½y_{t+k}]/2k$$

jeweils (für jedes t) ein Mittelwert aus fünf Werten:

$$\tilde{y}_t^z = 1/4[½y_{t-2} + y_{t-1} + y_t + y_{t+1} + ½y_{t+2}]$$

und somit im Beispiel für den ersten gleitenden Durchschnitt:

$$\begin{aligned}\tilde{y}_3^z &= 1/4[½y_{3-2} + y_{3-1} + y_3 + y_{3+1} + ½y_{3+2}]\\ &= 1/4[½y_1 + y_2 + y_3 + y_4 + ½.y_5]\\ &= (½\cdot 2 + 6 + 8 + 6 + ½\cdot 6)/4 = 24/4 = 6.\end{aligned}$$

Entsprechend sind die folgenden Werte:

$$\begin{aligned}\tilde{y}_4^z &= 1/4[½y_{4-2} + y_{4-1} + y_4 + y_{4+1} + ½y_{4+2}]\\ &= 1/4[½y_2 + y_3 + y_4 + y_5 + ½y_6] = (3+8+6+6+5)/4 = 28/4 = 7\end{aligned}$$

$$\begin{aligned}\tilde{y}_5^z &= 1/4[½y_{5-2} + y_{5-1} + y_5 + y_{5+1} + ½y_{5+2}]\\ &= 1/4[½y_3 + y_4 + y_5 + y_6 + ½y_7] = (4+6+6+10+6)/4 = 32/4 = 8\end{aligned}$$

usw. Man sieht, dass die Zahlen so gewählt sind, dass die zentrierten gleitenden Durchschnitte auf der Geraden $\tilde{y}_t^z = 3 + t$ liegen.

Folgerungen aus Def. 11.2:

- Am Anfang und Ende fallen beim gleitenden Durchschnitt jeweils k Glieder weg.
- Die Reihe der gleitenden Durchschnitte ist damit um 2k Glieder kürzer als die Reihe der Ursprungswerte (Daten).
- Der erste gleitende Durchschnitt fällt auf den k+1 ten Wert.
- Diese Zusammenhänge gelten allgemein, bei ungeradem p mit p = 2k+1 und bei zentrierten gleitenden p-gliedrigen Durchschnitten mit p = 2k.

	p = 2k+1 (ungerade)	p = 2k (gerade)
es fallen weg	k = (p-1)/2	k = p/2
der erste Wert	k+1 = (p+1)/2	k+1 = p/2 + 1

- Aufeinanderfolgende gleitende Durchschnitte haben 2k bzw. 2k-2 Glieder gemeinsam:
  ungerades p: p-1 = 2k Glieder,
  gerades p und zentriert: p + 1 - 3 = p - 2 = 2k - 2 Glieder.
- Gleitende Durchschnitte lassen sich somit auch rekursiv bestimmen:
  ungerades p (nicht-zentriert):

$$\tilde{y}_t = [y_{t-k} + y_{t+1-k} + ... + y_{t+(k-1)} + y_{t+k}]/p$$

$$\tilde{y}_{t+1} = [y_{t+1-k} + ... + y_{t-(k-1)} + y_{t+k} + y_{t+1+k}]/p$$

$$= \tilde{y}_t + [y_{t+k+1} - y_{t-k}]/p$$

gerades p und zentriert:

$$\tilde{y}^z_{t+1} = \tilde{y}^z_t + [(y_{t+k} + y_{t+1+k}) - (y_{t-k} + y_{t+1-k})]/2p.$$

Man kann dies leicht anhand des Beispiels 11.2 verifizieren. Dort galt:

- ungerades p = 3 = 2k+1 (also k=1):
  am Anfang und Ende fällt ein Glied weg (k=1); die Reihe der gleitenden Durchschnitte beginnt bei t=2 denn (p+1)/2 = 2.
- gerades p = 4 = 2k (k=2, zentrierte Quartalsdurchschnitte):
  am Anfang und Ende fallen zwei Glieder weg (k = 2 = p/2); die Reihe der zentrierten gleitenden Durchschnitte beginnt bei t=3 denn 3 = k+1 = (p/2)+1.

Man kann mit Beispiel 11.3 leicht erkennen, dass die Bestimmung des Trends bzw. der glatten Komponente und die Eliminierung einer regelmäßigen Schwankung genau dann gelingt, wenn die Gliederzahl des Zyklus mit der des gleitenden Durchschnitts übereinstimmt.

Beispiel 11.3:

a) Berechnen Sie für das Beispiel 11.2 die trendbereinigten Werte [vgl. Gl. 11.7] (mit dem zentrierten 4 - Quartals - gleitenden Durchschnitt als Trend).
b) Man berechne gleitende Durchschnitte mit p = 3 Perioden und die trendbereinigten Werte für die folgende Zeitreihe:

Zeit (t)	1	2	3	4	5	6	7	8
Zeitreihe (y_t)	6	4	11	12	10	17	18	16

c) Man interpretiere die Ergebnisse und vergleiche sie mit einer Trendberechnung nach der Methode der kleinsten Quadrate!

Lösung 11.3:

Es gilt im Teil a) die Abweichungen $d_t^z = y_t - \tilde{y}_t^z$ zu bestimmen und im Teil b) die gleitenden Durchschnitte $\tilde{y}_t$ zu berechnen und die Abweichungen $d_t = y_t - \tilde{y}_t$. Man erhält die Ergebnisse der nachfolgenden Tabelle.

Teil a)

t	y_t	$\tilde{y}_t^z$	d_t^z
3	8	6	2
4	6	7	-1
5	6	8	-2
6	10	9	1
7	12	10	2
8	10	11	-1
9	10	12	-2
10	14	13	1

Teil b)

t	y_t	$\tilde{y}_t$	d_t
1	6	(5)	1
2	4	7	-3
3	11	9	2
4	12	11	1
5	10	13	-3
6	17	15	2
7	18	17	1
8	16	(19)	-3

Man sieht, dass den konstruierten "Daten" im Fall a) ein Zyklus mit den Werten 2,-1,-2,1 zugrundeliegt, während im Fall b) die Daten entstanden, indem man zur Geraden 3+2t den Zyklus 2,1,-3 addierte. In beiden Fällen handelt es sich um einen regelmäßigen Zyklus, dessen Mittelwert Null beträgt.

Man beachte, dass die mit den gleitenden Durchschnitten in den Fällen a) und b) dieses Beispiels errechnete Gerade nicht identisch ist mit einer Trendgeraden nach der Methode der kleinsten Quadrate. Für diese erhält man:

für a) also die 12 Daten von Beispiel 11.2

$\hat{y}_t = 2{,}72727 + 1{,}04196t$ ($r^2 = 0{,}84$) statt $\tilde{y}_t^z = 3+t$

für b) also die 8 Daten von Beispiel 11.3 Teil b

$\hat{y}_t = 3{,}39287 + 1{,}85714t$ ($r^2 = 0{,}80$) statt $\tilde{y}_t = 3+2t$.

Dass beide Methoden nicht zur gleichen Gerade führen liegt an den unterschiedlichen Modellannahmen (vgl. Übers. 11.3). Der Methode der gleitenden Durchschnitte liegt ein lokales Trendmodell zugrunde, d.h. es werden jeweils p Beobachtungen durch einen Mittelwert angepaßt, während bei der Methode der kleinsten Quadrate der Stützbereich alle T Werte der Zeitreihe umfaßt (globales Modell).

Weitere Bemerkungen zur Methode der gleitenden Durchschnitte:

1. Ein regelmäßiger p-gliedriger Zyklus wird durch einen p-gliedrigen gleitenden Durchschnitt genau ausgeschaltet ["annulliert"] (vgl. Bsp. 11.3). Wird p nicht genau getroffen oder ist der Zyklus nicht regelmäßig, so tritt in jedem Fall eine Glättung ein, weil jeweils p Werte durch einen Mittelwert repräsentiert werden und dieser nicht kleiner (größer) sein kann, als der kleinste (größte) Wert $y_{t-k},....,y_{t+k}$.

2. Verkürzung der Reihe: Die Reihe der gleitenden Durchschnitte ist gegenüber der Reihe der Ursprungswerte am Anfang (historischer Rand) und [was problematischer ist; Prognose!] am Ende (aktueller Rand) um k Werte kürzer.

3. Wovon hängt es ab, wie stark die Reihe geglättet wird? Wie groß soll p gewählt werden? Die Zeitreihe wird i.d.R. umso mehr geglättet, je größer p gewählt wird. Als Faustregel für die Wahl der richtigen Periodenlänge p gilt:

   - p groß wählen, wenn die zu schätzende Trend- bzw. glatte Komponente schwach und die sie überlagernde Restkomponente (Zyklus) stark ausgeprägt ist

Übersicht 11.3: Gleitende Durchschnitte und kleinste Quadrate

	gleitende Durchschnitte	kleinste Quadrate
Trendfunktion	keine Funktion anzunehmen*), reines Glättungsverfahren	Typ der Trendfunktion muss a priori angenommen werden
Voraussetzungen bzgl. der Daten	Annahmen über Zykluslänge p erforderlich,	Zeitreihe muss nicht äquidistant sein
Trendmodell	lokales Trendmodell (Anpassung von jeweils p Werten)	globales Trendmodell (Anpassung aller T Beobachtungen)
Länge des Trends	Reihe der Trendwerte im Vergleich zu den Ursprungswerten am Anfang und Ende verkürzt	einheitliche Trendfunktion für alle Zeitpunkte von t=0 bis t=T

*) Im Gegensatz zur Trendbestimmung mit der Methode der kleinsten Quadrate braucht man bei der Methode des gleitenden Durchschnitts keine Vorkenntnisse über den möglichen Funktionstyp des Trends. Das Ergebnis ist dann eine "glatte" Kurve, für

die aber i.d.R. auch keine Funktion explizit in einem geschlossenen Ausdruck anzugeben ist.

- p klein wählen, wenn die glatte Komponente stark ausgeprägte Schwankungen hat, die nicht ausgemittelt werden sollen und die Restkomponente schwach ausgeprägt ist.

4. Verschiebung von "Wendepunkten" (im Sinne der Ökonomie, also Extremwerte): Ein zu starkes Glätten der Zeitreihe kann dazu führen, dass Wendepunkte verschwinden bzw. mit erheblicher Verzögerung angezeigt werden.

5. Das "Hintereinanderschalten" mehrerer gleitender Mittelwerte erzeugt einen gewogenen gleitenden Mittelwert mit entsprechend verlängertem Stützbereich (wie bereits bei der Herleitung von Gl. 11.5 gezeigt wurde).

   Beispiel: Die Reihe x_t wird geglättet mit $y_t = ½(x_t+x_{t-1})$ und anschließend wird diese Reihe mit $z_t = ½(y_t+y_{t-1})$ erneut geglättet. Das gleiche Ergebnis erhält man mit einer einmaligen Glättung mit dem gewogenen Mittel $z_t = ¼\ x_t + ½\ x_{t-t} + ¼\ x_{t-2}$. Die Gewichte c_i des dritten Mittels sind Produktsummen der Gewichte a_i und b_i der ersten beiden Mittel ($c_i = \Sigma_k a_i b_{i-k}$).

6. Bisher wurden nur gleitende ungewogene arithmetische Mittel betrachtet. Man kann auch **gewogene** arithmetische Mittel gleitend berechnen, etwa bei ungeradem p = 2k+1

$$\tilde{y}_t = [g_o\, y_{t-k} + g_1\, y_{t-k-1} + ... + g_{2k-1}\, y_{t+(k-1)} + g_{2k}\, y_{t+k}]/p,$$

   wobei die Gewichte g_i (i=0,1,...,2k) in der Summe 1, aber nicht notwendig alle positiv sein sollen. Man kann zeigen, dass ein polynomialer Trend nach der Methode der kleinsten Quadrate eine Folge gewogener gleitender Durchschnitte ist (vgl. Beispiel 11.4).
   Ein Beispiel für gewogene gleitende arithmetische Mittel ist das exponential smoothing. Man kann auch Mittelwerte anderen Typs verwenden, z.B. gleitende Mediane.

7. Durch die gleitende Mittelwertbildung oder Differenzenbildung kann $\tilde{y}_t$ auch bei einer reinen Zufallsfolge y_t (die nicht autokorreliert ist) einen Zyklus haben (sinusoidal verlaufen, autokorreliert sein). Dieser sog. "**Slutzki-Yule-Effekt**"

(Beispiel 11.5) kommt dadurch zustande, dass aufeinanderfolgende gleitende p-gliedrige (p ungerade) jeweils p-1 Glieder gemeinsam haben.

Beispiel 11.4:

Gegeben sei die folgende Zeitreihe:

t	-2	-1	0	+1	+2
y_t	10	12	15	17	16

Man berechne den Trend mit der Methode der kleinsten Quadrate und vergleiche die so erhaltenen fünf Trendwerte mit gewogenen arithmetischen Mitteln der fünf Ursprungswerte y_t der obigen Zeitreihe, wenn man die folgenden fünf Gewichtungsschemen benutzt:

	-2	-1	0	+1	+2
1	0,6	0,4	0,2	0	-0,2
2	0,4	0,3	0,2	0,1	0
3	0,2	0,2	0,2	0,2	0,2
4	0	0,1	0,2	0,3	0,4
5	-0,2	0	0,2	0,4	0,6

Was fällt bei der Betrachtung der fünf Gewichtungsschemen auf?

Lösung 11.4:

Das Beispiel soll Zusammenhänge zwischen der Methode der kleinsten Quadrate und der Methode der (gewogenen) gleitenden Mittelwerte demonstrieren.

Mit der Methode der kleinsten Quadrate erhält man bei den gegebenen Daten den linearen Trend $\hat{y} = 14 + 1{,}7 \cdot t$ und die Trendwerte:

-2	-1	0	+1	+2
10,6	12,3	14	15,7	14,4

Berechnet man den Mittelwert aus den Ursprungswerten mit dem Wägungsschema Nr. 1, so erhält man:
$0{,}6 \cdot 10 + 0{,}4 \cdot 12 + 0{,}2 \cdot 15 + 0.17 + (-0{,}2) \cdot 16 = 6 + 4{,}8 + 3 - 3{,}2 = 10{,}6.$

Mit dem Schema Nr.2 erhält man für das gewogene Mittel den Wert 12,3 usw. In der Reihenfolge der Wägungsschemen erhält man also genau die Werte der geschätzten Trendgeraden. Bei der Betrachtung der Gewichte fällt auf:

- Die Summe der Gewichte ist jeweils 1.
- Das dritte Wägungsschema ist das **un**gewogene arithmetische Mittel.
- Das fünfte Schema ist das "inverse" (Umkehr der Reihenfolge der Gewichte) erste, denn es gilt :

Schema 1:	0,6	0,4	0,2	0	-0,2
Schema 5:	-0,2	0	0,2	0,4	0,6

  und entsprechend ist das Schema 4 das "inverse" Schema 2.
- Das mittlere (dritte) Wägungsschema ist dagegen symmetrisch.

Beispiel 11.5:

Im Supermarkt S wurden an 11 Tagen die folgenden Mengen der Konserve K gekauft, die offenbar zufällig um den Wert 58,45 schwankten: 64, 57, 65, 58, 51, 77, 52, 45, 89, 46, 39. Man berechne gleitende dreigliedrige Durchschnitte $\tilde{y}_t$ und zeichne die Ursprungswerte y_t und die gleitenden Durchschnitte $\tilde{y}_t$. Was fällt bei dem Bild auf?

Lösung 11.5:

Die gleitenden Durchschnitte sind periodisch: 62, 60, 58, 62, 60, 58, 62, 60, 58, während die Ursprungswerte ziemlich regellos schwanken.

Beispiel 11.6:

Gegeben seien die Zeitreihen (zu diskreten Zeitpunkten t = 0,1,2,...,8)

$$y_1(t) = \sin\left(\frac{\pi}{4}t\right)$$

$$y_2(t) = \sin\left(\frac{\pi}{2}t\right)$$

Man bestimme nicht zentrierte gleitende Mittelwerte zu je zwei Perioden (die dann den Werten t = 0.5,1.5,2.5,...zugeordnet sind) von $y_1(t)$, $y_2(t)$ und $y_3(t) = y_1(t) + y_2(t)$

Lösung 11.6:

Aufgrund des folgenden trigonometrischen Zusammenhangs

$$\frac{1}{2}(\sin\alpha + \sin\beta) = \sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right)$$

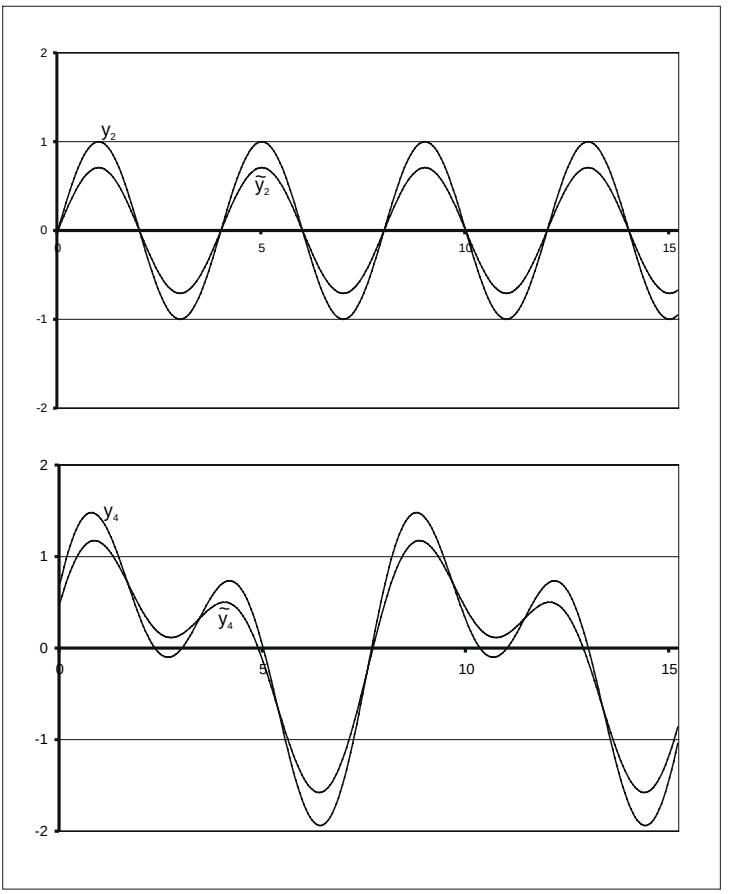
lässt sich zeigen, dass man die folgenden gleitenden Mittelwerte erhält

$$\tilde{y}_1(t) = \sin\left(\frac{3}{8}\pi\right) \sin\left(\frac{\pi}{4}t\right) = 0{,}92388 \cdot y_1(t) = A_1\, y_1(t)$$

$$\tilde{y}_2(t) = \sin\left(\frac{\pi}{4}\right) \sin\left(\frac{\pi}{2}t\right) = 0{,}70711 \cdot y_2(t) = A_2\, y_2(t)$$

Wie man sieht, entstehen durch gleitende Mittelwerte gedämpfte Schwingungen (Amplitude $A_1 < 1$, $A_2 < 1$). Die Zeitreihe $y_2(t)$ hat eine doppelt so große Frequenz (halb so lange Periode) wie $y_1(t)$ und sie wird etwas stärker gedämpft. Die Überlagerung $y_3(t)$ ist von beiden Zyklen geprägt.

Abb. 11.2: Zeitreihen und gleit. 2- Perioden Durchschnitte des Bsp. 11.6



Auch der gleitende Mittelwert $\tilde{y}_3(t) = \tilde{y}_1(t) + \tilde{y}_2(t)$ ist wieder eine gedämpfte Schwingung (vgl. Abb. 11.2). Im Abschnitt 3c wird der Gedanke der Überlagerung von Sinusschwingungen weiter verfolgt. Interessant ist

es festzuhalten, dass sich die gleitenden Mittelwerte nicht auf Frequenz und Phase sondern nur auf die Amplitude auswirken. Führt man eine Phasenverschiebung durch, etwa

$$y_4(t) = y_1(t) + \sin\left(\frac{\pi}{2}t + \frac{\pi}{4}\right)$$

so werden auch die gleitenden Mittelwerte von dieser Überlagerung geprägt. Man erhält

$$\tilde{y}_4(t) = \tilde{y}_1(t) + A_2 \sin\left(\frac{\pi}{2}t + \frac{\pi}{4}\right) \quad \text{statt} \quad \tilde{y}_3(t) = \tilde{y}_1(t) + A_2 \sin\left(\frac{\pi}{2}t\right)$$

c) Berechnung der Saisonkomponente

Die Saisonkomponente ist zu bestimmen, um eine Saisonbereinigung durchführen zu können. Dadurch können die (kurzfristigen) saisonalen Einflüsse von den längerfristigen Einflüssen isoliert werden, wodurch eine Änderung der mittel- bzw. langfristigen Entwicklung besser (bereinigt von kurzfristigen Einflüssen) zu erkennen ist.
Die im folgenden behandelten einfachen Verfahren heißen auch "Phasendurchschnittsverfahren" weil über gleichnamige Monate, Quartale usw. gemittelt wird.

1. Konstante Saisonfigur (Saisonnormale) bei additiver Überlagerung

Notation: Die Ursprungswerte y_t der Zeitreihe von T Perioden werden im folgenden doppelt indiziert. Statt y_t soll y_{js} geschrieben werden, wobei j = 1,2,...,J das Jahr und s = 1,2,...,n den Unterzeitraum bezeichnet, z.B. ein Quartal (n = 4) oder einen Monat (n = 12), so dass T = nJ.

Eine konstante (starre) Saisonfigur ist dann gegeben, wenn die Werte der Saisonkomponente für gleiche Unterzeiträume in jedem Jahr j identisch sind. Sie wiederholt sich in jedem Jahr in genau gleicher Weise. Im additiven Modell

$$(11.6) \quad y_{js} = m_{js} + k_{js} + s_{js} + r_{js}$$

gilt somit $s_{js} = s_s$ bei einem starren Saisonmuster.

Zieht man von den Ursprungswerten y_{js} den Trend bzw. die glatte Komponente $g_{js} = m_{js}+k_{js}$ (etwa in Gestalt von $\tilde{y}_{js}$ bei gleitenden Durchschnitten) ab, so erhält man die Reihe trendbereinigter Werte y^*_{js}, die nur noch aus saisonalen und irregulären Schwankungen besteht:

(11.7) $y^*_{js} = y_{js} - g_{js} = s_{js} + r_{js}$ (trendbereinigte Werte).

Die konstante und nicht-normierte Saisonkomponente oder "Saisonnormale" s_s ist das arithmetische Mittel aller trendbereinigter Werte, die dem gleichen Unterzeitraum s zugeordnet sind (über alle J Jahre gemittelt):

$$(11.8) \quad s_s = \frac{1}{J}\sum_{j=1}^{j=J} y^*_{js}$$

Die n einzelnen Werte s_s heißen auch Saisonkoeffizienten. Man kann statt arithmetische Mittel auch Mediane bestimmen.

Zur Begründung des Verfahrens:

Geht man vom Modell $y_{js} = y^*_{js} + u_{js}$ aus, wobei y^*_{js} die trendbereinigten Werte sind, so führt die Bestimmung von Konstanten k_s unter den Bedingungen

$$\sum_{j=1}^{J}(y^*_{js} - k_s)^2 = \text{Min (für jedes s)}$$

wegen der Minimumeigenschaft des arithmetischen Mittels zum Wert $k_s = s_s$ gem. Gl. 11.8. Ähnlich führen "Saisondummies" d_s, also Regressoren, für die gilt $d_s = 1$ sonst und wenn die Beobachtung in den Unterzeitraum s (s = 1, 2,..., n-1) fällt und $d_s = 0$ wenn s = n in der Regressionsgleichung $y_{js}{}^* = a + \sum_{s=1}^{n-1} b_s d_s + u_{js}$ zu einer Schätzung der Koeffizienten $b_1, b_2,...,b_{n-1}$ so dass $a+b_s = s_s$ für s = 1, 2,..., n-1 und $a = s_n$ jeweils das arithmetische Mittel gleichnamiger (für den gleichen Unterzeitraum s geltender) trendbereinigter Werte ist, wenn die Beobachtungen genau J Jahre umfassen. Die Beschränkung auf n-1 Dummies ist notwendig, damit keine lineare Abhängigkeit entsteht, d.h. die Matrix **X'X** (vgl. Kap. 8) invertierbar ist.

Beispiel 11.7:

Die Berechnung einer starren Saisonfigur durch Bildung eines arithmetischen Mittels der im folgenden gegebenen trendbereinigten Werte für gleichnamige Quartale (I bis IV) über J = 4 Jahre führt nach Gl. 11.8 zu den Werten:

$s_1 = (-7-1-6-2)/4 = -4,$

$s_2 = (3+5+1+3)/4 = +3$ usw.

Quartal	trendbereinigte Werte 1986	1987	1988	1989
I	-7	-1	-6	-2
II	+3	+5	+1	+3
III	+2	-1	+1	-2
IV	+8	+10	+9	+9

Saison-normale	Restkomponente 1986	1987	1988	1989	Σ
-4	-3	+3	-2	+2	0
+3	0	+2	-2	0	0
0	+2	-1	+1	-2	0
+9	-1	1	0	0	0

Zieht man die Saisonnormale von den trendbereinigten Werten ab, so erhält man nach Gl. 11.7 die Restkomponente.

Der mittlere Ausschlag der Saisonnormalen ist nicht Null, sondern $\bar{s} = 2 = 8/4$ wegen $\Sigma s_s = -4 + 3 + 0 + 9 = 8$. Die Saisonkomponente ist somit nicht "normiert".

Das arithmetische Mittel der n Saisonkoeffizienten beträgt

$$(11.9) \quad \bar{s} = \frac{1}{n}\sum_s s_s$$

wobei die Summe über alle n Unterzeiträume (s = 1,2,...,n) gebildet wird, also z.B. über die n = 4 Quartale oder n = 12 Monate.

Die auf einen Mittelwert von $\bar{s}_s^* = 0$ ***normierte Saisonnormale*** (oder "Saisonindex") $s_s,^*$ ist dann die Lineartransformation der nicht-normierten s_s (gem. Gl. 11.8) indem man von s_s den Mittelwert $\bar{s}$ dieser Saisonkoeffizienten (gem. Gl. 11.9) abzieht:

$$(11.10) \quad s_s^* = s_s - \bar{s} \qquad \text{(normierte Saisonnormale).}$$

Zieht man von den Ursprungswerten y_{js} den jeweils zugehörigen normierten Saisonkoeffizienten s_s^* ab, so erhält man die ***saisonbereinigte Zeitreihe***. Sie soll einen Eindruck vermitteln, wie sich die Zeitreihe (längerfristig) entwickelt hätte, wenn sie nicht von saisonalen Einflüssen überlagert worden wäre.

Die normierte Saisonnormale für die vier Quartale des Beispiels 11.4 beträgt:
für

Quartal I: -2 - 2 = -4 (da $\bar{s} = 2$)
Quartal II: 3 - 2 = 1

Quartal III: 0 - 2 = -2
Quartal IV: 9 - 2 = 5, was im Mittel Null ist.
Der normierte Wert s_s^* ist der Schätzwert für die typische Abweichung des s-ten Unterzeitraums. Die n Werte für s_s^* bilden die normierte Saisonfigur und sind im Mittel Null.

Beispiel 11.8:

Man bestimme die normierten Saisonkoeffizienten für die folgenden trendbereinigten Werte (n = 4 Quartale und J = 4 Jahre):

s	1987	1988	1989	1990
1	-7	-1	-6	-2
2	3	5	1	3
3	2	-1	1	-2
4	8	10	9	9

Lösung 11.8:

Nicht-normierte Saisonkoeffizienten:

$s_1 = -4, \quad s_2 = 3, \quad s_3 = 0, \quad s_4 = 9.$

Ihr Mittelwert beträgt (-4 + 3 + 0 + 9)/4 = 8/4 = 2, so dass man für die normierte Saisonnormale erhält:

$s_1^* = -6, \quad s_2^* = 1, \quad s_3^* = -2, \quad s_4^* = 7 \quad (\text{mit } \Sigma s_s^* = 0).$

2. Variable Saisonfigur bei multiplikativer Überlagerung

Bei multiplikativer Verknüpfung

(11.6a) $y_{js} = m_{js} \cdot k_{js} \cdot s_{js} \cdot r_{js}$

kann eine Saisonkomponente, die im Mittel 1 beträgt, bestimmt werden. Dabei ist statt der absoluten die relative (in bezug auf die glatte Komponente) Ausschlaghöhe für gleiche Unterzeiträume konstant.

Dividiert man die Ursprungswerte y_{js} durch die glatte Komponente g_{js}, so erhält man die trendbereinigten Werte y_{js}^* eines multiplikativen Modells. Durch arithmetische (nicht geometrische) Mittelung über gleichnamige Unterzeiträume (z.B. Quartale, Monate) erhält man die nicht-normierten Saisonkoeffizienten.
Diese und weitere Verfahrensschritte werden in Übers. 11.4 zusammengefaßt und dem Verfahren bei starrer Saisonfigur (additives Modell) gegenübergestellt.

Da die normierten Saisonkoeffizienten $s^*_{js} = s^*_s$ (für jedes Jahr j = 1, 2,..., J) im multiplikativen Modell im Mittel 1 sind, wird durch das Mittel $\bar{s}$ der nicht-normierten Saisonkoeffizienten s^*_{js} dividiert:

(11.10a) $$s^*_s = \frac{s_s}{\bar{s}} \qquad \text{mit } \bar{s} = \frac{1}{n}\sum s^*_{js}$$

(wobei summiert wird über s = 1,2,...,n).

Dividiert man die Ursprungswerte y_{js} durch den entsprechenden normierten Saisonkoeffizienten s^*_s, so erhält man die saisonbereinigten Werte.

Übersicht 11.4: Saisonbereinigung bei additiver und multiplikativer Überlagerung

	additives Modell (starre Saison)	muliplikatives Modell (variable Saison)
trendbereinigte[1] Werte y^*_{js} =	$= y_{js} - g_{js}$ Differenz, Gl.11.7	$= \frac{y_{js}}{g_{js}}$ Quotient
nichtnormierte Saisonnormale[2] s_s =	$\frac{1}{J}\sum_j y^*_{js}$ (arithmet. Mittel, Gl.11.8)	$\frac{1}{J}\sum_j y^*_{js}$ (arithmetisches Mittel)
normierte Saison s^*_s = (mit Mittel $\bar{s}$ nach Gl. 11.9)	$s_s - \bar{s}$ (mittlere nichtnormierte Saison abgezogen, [Gl.11.10])	$s_s / \bar{s}$ (Division durch mittlere nichtnormierte Saison, [Gl. 11.10a])
normiert auf[3]	$\bar{s}^*_s = 0$	$\bar{s}^*_s = 1$

1) "trendbereinigt" soll heißen: ohne Trend bzw. glatte Komponente $g_t = g_{js}$ (meist geschätzt als gleitender Durchschnitt $\tilde{y}_{js}$).

2) oder Saisonkoeffizienten (n Werte)

3) d.h. das arithmetische Mittel (aller n Werte) der normierten Saisonnormalen s^*_s (s = 1,2,...,n) beträgt 0 bzw. 1.

d) Integrierte Modelle

Man kann auch das Modell $y_t = g_t + s_t + u_t$ (mit dem Residuum u_t) in einem Akt schätzen, d.h. die glatte Komponente g_t und die Saisonkomponente s_t ***simultan***, statt ***sukzessiv*** schätzen, indem man die Methode der kleinsten Quadrate anwendet und Σu_t^2 als Funktion der zu schätzenden Parameter von g_t und s_t

- global, d.h. für die ganze Länge T der Zeitreihe (t = 1, 2, ..., T) minimiert oder
- lokal, d.h. Σu_t^2 für gleitende Stützbereiche der Länge p minimiert.

Die Komponenten g_t und s_t werden dabei meist wie folgt modelliert:

- g_t als Polynom geringen Grades, also
  $g_t = \Sigma a_i t^i$ (mit i = 0,1,...,p), etwa (p = 2) $a_0 + a_1 t + a_2 t^2$ und
- s_t als trigonometrische Funktion, d.h.
  $s_t = \sum_{j=1}^{j=q} [b_{1j} \cos(\lambda_j t) + b_{2j} \sin(\lambda_j t)]$ mit der Frequenz $\lambda_j = j \cdot 2\pi/P$ und
  $j = 1,2,...,q \leq ½P$
  wenn die Periode P ungeradzahlig ist bzw.

  $s_t = \sum_{j=1}^{q} [b_{1j} \cos(\lambda_j t) + b_{2j} \sin(\lambda_j t)] + b_{1q} \cos(\pi t)$ wenn die Periode P geradzahlig ist, weil bei q = ½P gilt $\sin\lambda qt = \sin(\pi t) = 0$ und deshalb b_{2q} nicht definiert ist.

3. Hinweise auf weiterführende Verfahren

a) Exponential Smoothing (exponentielles Glätten)

Exponentielles Glätten ist ein sehr einfaches und beliebtes Verfahren zur Bestimmung einer Ein-Schritt-Prognose (Prognose von y_{t+1} zur Zeit t). Es wird zunächst das bekannteste, sog. "einfache" Verfahren für Zeitreihen ohne Trend und ohne Saison behandelt.

1. Einfaches exponentielles Glätten:

Der zur Zeit t für t+1 ermittelte Prognosewert soll $\hat{y}_{t+1}$, und der dann später [zur Zeit t+1] beobachtete tatsächliche Wert y_{t+1} genannt werden. Der Ansatz ist rekursiv, d.h. man bestimmt $\hat{y}_{t+1}$ aufgrund von y_t und $\hat{y}_t$, $\hat{y}_{t+2}$ aufgrund von y_{t+1} und $\hat{y}_{t+1}$ usw.
Es gibt vier äquivalente Formulierungen des Prognoseansatzes, die den Sinn des Vorgehens deutlich machen und die im folgenden dargestellt werden.

1. Prognose als gewogenes Mittel aus den letzten Werten

Danach errechnet sich $\hat{y}_{t+1}$, der Prognosewert für die Periode t+1 als gewogenes arithmetisches Mittel aus y_t, dem tatsächlichen Wert der Periode t und $\hat{y}_t$, dem zur Zeit t-1 für t prognostizierten Wert, wobei die Gewichtung nach Maßgabe des frei zu wählenden Parameters α erfolgt, d.h. für $\hat{y}_{t+1}$ gilt:

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t \qquad (\text{mit } 0 < \alpha < 1). \tag{11.11}$$

Interessant sind die mit obiger Einschrankung für α ($0 < \alpha < 1$) an sich ausgeschlossenen zwei Extremfälle:

- $\alpha = 0$: $\hat{y}_{t+1} = \hat{y}_t$ (eine einmal gestellte Prognose wird unabhängig von der Erfahrung beibehalten)
- $\alpha = 1$: $\hat{y}_{t+1} = y_t$ (es wird quasi angenommen, dass morgen das eintreten wird, was heute eingetreten ist)

Gl. 11.11 erfordert eine Startbedingung [für t = 0]: $\hat{y}_o = y_o$. Das impliziert auch $\hat{y}_1 = y_o$.

2. Prognose als Mittel aller vergangenen Beobachtungen

Ersetzt man in Gl. 11.11 den prognostizierten Wert $\hat{y}_t$ durch den tatsächlichen Wert y_{t-1} und $\hat{y}_{t-1}$ (denn $\hat{y}_t = \alpha y_{t-1} + (1-\alpha)\hat{y}_{t-1}$), entsprechend hierin $\hat{y}_{t-1}$ durch $\hat{y}_{t-2}$ und y_{t-2} usw., so erhält man:

$$(11.12)\quad \hat{y}_{t+1} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + ... + \alpha(1-\alpha)^n y_{t-n}$$
$$+ (1-\alpha)^{n+1}\,\hat{y}_{t-n} = \sum_{i=0}^{i=n} \alpha(1-\alpha)^i\, y_{t-i} + (1-\alpha)^{n+1}\,\hat{y}_{t-n} \qquad (i = 0,1,...,n).$$

Wegen $0 < \alpha < 1$ ist $\lim_{n\to\infty}(1-\alpha)^{n+1} = 0$. Die Summe der unendlichen geometrischen Reihe $\Sigma\alpha(1-\alpha)^i = \alpha(1-\alpha)^0 + \alpha(1-\alpha)^1 + \alpha(1-\alpha)^2 + ...$ ist bekanntlich eins. Somit ist $\hat{y}_{t+1}$ ein gewogenes Mittel aller früheren tatsächlichen Werte y_{t-i}, gewogen mit $\alpha(1-\alpha)^i$, also mit exponentiell abnehmenden Gewichten [daher auch der Name exponentielle Glättung].

Gl. 11.12 zeigt zugleich die Grenzen des Verfahrens: da $\hat{y}_{t+1}$ ein gewogenes arithmetisches Mittel ist, kann es nicht größer als der größte und nicht kleiner als der kleinste vergangene Wert sein. Liegt kein Trend vor, d.h. besteht die Zeitreihe nur aus Schwankungen um ein konstantes Niveau a, so ist es gleichwohl plausibel $\hat{y}_{t-n}$ zu schätzen aus früher tatsächlich beobachteten Werten y_{t-i} und zwar so, dass y_{t-1} stärker ins Gewicht fällt als y_{t-2}, y_{t-2} stärker als y_{t-3} usw.

3. partielle Korrektur einer Fehlschätzung

Nach einer dritten Schreibweise des Prognoseansatzes, die man ebenfalls durch einfache Umformung von Gl.11.11 erhält, wird $\hat{y}_{t+1}$ durch Korrektur von $\hat{y}_t$ um den Prognosefehler $F = y_t - \hat{y}_t$ (oder genauer: um einen [durch α bestimmten] Teil des Prognosefehlers [daher "partielle" Korrektur]) nach der folgenden Regel gewonnen:

$$(11.13)\quad \hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t) = \hat{y}_t + \alpha F.$$

Mit $F = y_t - \hat{y}_t$ gilt

- bei einer Überschätzung (Prognose $\hat{y}_t$ lag zu hoch, höher als y_t, so dass $F < 0$ weil $\hat{y}_t > y_t$) wird die Prognose nach unten korrigiert, so dass $\hat{y}_{t+1} < \hat{y}_t$;
- bei einer Unterschätzung ($F > 0$ da $\hat{y}_t < y_t$) wird die Prognose nach oben korrigiert, so dass $\hat{y}_{t+1} > \hat{y}_t$.

Im Falle von $\alpha = 1$ wird der volle Fehler F adddiert [egal, wie gut oder schlecht die Prognose war], denn dann ist $\hat{y}_{t+1} = \hat{y}_t + F = y_t$, was darauf hinausläuft, den zuletzt beobachteten Wert als Prognosewert zu nehmen. Mit $\alpha = 0$ wird der Fehler überhaupt nicht korrigiert und es gilt $\hat{y}_{t+1} = \hat{y}_t$, auch wenn $\hat{y}_t$ falsch war, also von y_t abwich.

4. eponentielle Glättung und gleitende Mittelwerte:

Man kann Gl. 11.11 auch als sukzessive Schätzung einer Niveaukomponente a interpretieren. Die Zeitreihe y_t setzt sich danach aus einem Niveau a und einer Zufallskomponente u_t zusammen, so dass gilt:

(11.14) $y_t = a + u_t$.

Das Niveau a wird in jeder Periode t neu als a_t ("updating") geschätzt durch Mittelung der m früheren Werte von y:

$$(11.14a) \quad \hat{a}_{t-1} = \frac{y_{t-1} + ... + y_{t-m}}{m}$$

entsprechend gilt:

$$\hat{a}_t = \frac{y_t + ... + y_{t-m+1}}{m} \quad \text{und}$$

(11.15) $\hat{a}_t = \hat{a}_{t-1} + \alpha(y_t - \hat{a}_{t-1})$

wenn man $\alpha = 1/m$ und $y_{t-m} = \hat{a}_{t-1}$ setzt. Da $\hat{a}_{t-1}$ nach Gl. 11.14a definiert ist läuft die Annahme $y_{t-m} = \hat{a}_{t-1}$ darauf hinaus, dass:

$$(11.14b) \quad y_{t-m} = \hat{a}_{t-1} = \frac{y_{t-1} + ... + y_{t-m+1}}{m-1}$$

gelten soll, was bei hinreichend großem m und Geltung des Modells $y_t = a + u_t$ auch zutreffen dürfte. Denn dann dürfte auch
$(y_{t-1} + ... + y_{t-m})/m = (y_{t-1} + ... + y_{t-m+1})/(m-1)$ erfüllt sein.
Ersetzt man in Gl. 11.15 die Größe $\hat{a}_t$ durch $\hat{y}_{t+1}$ und $\hat{a}_{t-1}$ durch $\hat{y}_t$ so erhält man Gl. 11.13.

<u>Aussagen über die Wirkung von α:</u>

Ein großes α führt zu einer schnellen Abnahme der Gewichte:

	α	$\alpha(1-\alpha)$	$\alpha(1-\alpha)^2$	$\alpha(1-\alpha)^3$	$\alpha(1-\alpha)^4$
α	0,9	0,09	0,009	0,0009	0,00009
groß	0,8	0,16	0,032	0,0064	0,00128
α	0,3	0,21	0,147	0,1029	0,07203
klein	0,1	0,09	0,081	0,0729	0,06561

Ein großes α bedeutet ein kurzes Gedächtnis des Prozesses (die länger zurückliegenden Werte spielen eine geringere Rolle). Die Prognose reagiert schnell, auch auf vorübergehende Erscheinungen, sie paßt sich einem neuen Trend schnell an. Bei geringem α wird die Prognose auch von weiter zurückliegenden Daten stark beeinflußt und sie reagiert auf neue Entwicklungen sehr träge.

Beispiel 11.9

Die Hausfrau H wusste stets kulinarischen Genuß zu schätzen und entwickelte unterdessen ein Raumbedürfnis welches hienieden sonst nicht schicklich ist. Sie trachtete deshalb hinfort danach, durch Schlankheitsmittel ihre Proportionen auf ein gefälligeres Maß zu reduzieren. Dabei gebrach es ihr jedoch an der gebotenen Konsequenz, so dass ihr Gewicht y (in kg) stark schwankte und sich eine nachhaltige Reduktion nicht einstellen wollte, wie die folgenden Zahlen zeigen:

t	0	1	2	3	4	5	6
t^*	-3	-2	-1	0	+1	+2	+3
y_t	120	130	125	120	130	125	120

Man bestimme

a) Gleitende Durchschnitte zu jeweils 3 Perioden,
b) einen linearen Trend mit der Methode der kleinsten Quadrate und
c) einen Prognosewert für die Periode t = 7 (oder t* = 4) mit der Methode des exponentiellen Glättens (α = 0,2)!

Lösung 11.9

Die gleitenden Durchschnitte sind konstant 125. Mit der Methode der kleinsten Quadrate erhält man die nicht konstante, sondern leicht fallende Gerade: $y_t = 124{,}286 - 0{,}1786t^*$, bzw. (da $t^* = t-3$ und $124{,}686 + 3 \cdot 0{,}1786 = 125{,}222$) $y_t = 125{,}222 - 0{,}1786t$. Die Prognosewerte errechnen sich wie folgt gem. Gl. 11.11:

$\hat{y}_1 = y_o = 120$ (Startbedingung)

$\hat{y}_2 = \alpha y_1 + (1-\alpha)\hat{y}_1 = 0{,}2 \cdot 130 + 0{,}8 \cdot 120 = 122$

$\hat{y}_3 = \alpha y_2 + (1-\alpha)\hat{y}_2 = 0{,}2 \cdot 125 + 0{,}8 \cdot 122 = 122{,}6$ usw.

$\hat{y}_4 = 122{,}08$; $\hat{y}_5 = 123{,}664$; $\hat{y}_6 = 123{,}93$; $\hat{y}_7 = 123{,}145$

wobei die Prognose des Wertes 130 bei Betrachtung der Daten natürlich näher gelegen hätte. Wegen des geringen Wertes von α nähert sich die Folge der Prognosewerte nur allmählich dem den Daten zugrundeliegenden langfristigen Mittelwert von 125.

2. Exponentielles Glätten zweiter Ordnung

a) Rechengang

Beim *einfachen exponentiellen Glätten* (exponentielle Glättung *erster Ordnung*) galt $y_t = a + u_t$, so dass die Zeitreihe allein zufallsabhängig um ein bestimmtes Niveau a schwankt. Die Niveauschätzung a und damit $\hat{y}$ wird mit dem updating (11.15a) $\hat{a}_t = \alpha\, y_t + (1-\alpha)\, \hat{a}_{t-1} = \hat{y}_{t+1}$ laufend revidiert. Liegt dagegen ein linearer Trend vor, so soll das Modell

(11.16) $\quad y_{t+i} = a + i \cdot b + u_t$

(i= 1, 2, ..) lauten, wobei sowohl a als auch b unter Berücksichtigung vergangener Werte von y laufend neu zu schätzen sind als $\hat{a}$ und $\hat{b}$. Über dieses updating wirken die Daten y_t (und damit die Variable t) auf die Prognose ein. Der Prognoseansatz ist mithin

(11.16a) $\quad \hat{y}_{t+i} = \hat{a}_t + i \cdot \hat{b}_t$,

bzw. bei Ein-Schritt-Prognosen (i=1)

(11.17) $\quad \hat{y}_{t+1} = \hat{a}_t + \hat{b}_t$.

Zur rekursiven Bestimmung von $\hat{a}$ und $\hat{b}$ gibt es verschiedene Lösungsansätze, die sich hauptsächlich aufgrund der Anzahl zu verwendender Glättungsfaktoren unterscheiden. Beim *Zwei-Parameter-Glätten* nach Holt verwendet man für Grundwert (a) und Trendwert (b) jeweils eigene Glättungsfaktoren α und β ($0 \leq \alpha, \beta \leq 1$), wobei gilt

$$\text{(11.18)} \quad \hat{a}_t = \alpha\, y_t + (1-\alpha)\, \hat{y}_t$$

analog zu Gl. 11.11. Allerdings ist wegen der folgenden Gleichungen jeweils $\hat{a}_t$, $\hat{a}_{t-1}$,... hier nicht identisch mit den gleichnamigen Größen im einstufigen Fall. Über Gl. 11.18 beeinflussen die Daten y_t die Schätzung $\hat{a}_t$ (wenn $\alpha > 0$) und indirekt über $\hat{a}_t$ auch $\hat{b}_t$ wegen

$$\text{(11.19)} \quad \hat{b}_t = \beta\,(\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\,\hat{b}_{t-1}.$$

b) Interpretation

Aus den damit dargestellten Verfahrensschritten lassen sich zwei Gleichungen folgern, nämlich:

$$\text{(11.20)} \quad \hat{y}_{t+1} = \alpha \sum_{i=0}^{n} (1-\alpha)^i\, y_{t-i} + (1-\alpha)^{n+1}\, \hat{y}_{t-n} + \sum_{i=0}^{n} (1-\alpha)^i\, \hat{b}_{t-i}$$

und

$$\text{(11.20a)} \quad \hat{b}_t = \beta\, \Sigma\, (1-\beta)^i\, (\hat{a}_{t-i} - \hat{a}_{t-i-1}) + (1-\beta)^{n+1}\, \hat{b}_{t-(n+1)}$$

die im Vergleich mit Gl. 11.12 und bei Betrachtung extremer Werte für β, nämlich $\beta = 0$ und $\beta = 1$ für eine inhaltliche Interpretation nützlich sein dürften. Aus den Startbedingungen (vgl. Übers. 11.5) $\hat{a}_0 = y_0$ und $\hat{b}_0 = y_1 - y_0$ folgt generell $\hat{y}_1 = y_1$ und $\hat{y}_2 = 2y_1 - y_0$. Erst ab $\hat{y}_3$ wirkt sich aus, welche Werte für α und β angenommen werden.

$\beta = 0$

Ist $\beta = 0$, so ist $\hat{b}_t = \hat{b}_{t-1} = \ldots = \hat{b}_1 = \hat{b}_0 = y_1 - y_0$ und man erhält

$$\hat{y}_{t+1} = \alpha \sum_{i=0}^{n} (1-\alpha)^i\, y_{t-i} + (1-\alpha)^{n+1}\, \hat{y}_{t-n} + R$$

wobei das Restglied lautet

$$R = (y_1 - y_0) \sum_{i=0}^{n} (1-\alpha)^i = \frac{1-(1-\alpha)^{n+1}}{\alpha}\,(y_1 - y_0) \quad ,$$

d.h. es wird zu den Prognosewerten des einfachen exponentiellen Glättens praktisch (wenn $n \to \infty$) nur die Konstante $(y_1 - y_0)/(1-\alpha)$ addiert.

$\beta = 1$

Ganz allgemein gilt $\hat{a}_t - \hat{a}_{t-1} = \alpha(y_t - \hat{y}_t) + \hat{b}_{t-1}$. Wenn außerdem gelten soll $\beta = 1$, dann ist $\hat{b}_t = \hat{a}_t - \hat{a}_{t-1}$, so dass man erhält

$\hat{b}_0 = y_1 - y_0$

$\hat{b}_1 = \alpha(y_1 - \hat{y}_1) + \hat{b}_0 = y_1 - y_0$

$\hat{b}_2 = \alpha(y_2 - \hat{y}_2) + \hat{b}_1 = \alpha[y_2 - (2y_1 - y_0)] + (y_1 - y_0)$

$\hat{b}_3 = \alpha(y_3 - \hat{y}_3) + \hat{b}_2$ usw.,

so dass der Schätzung der Trendkomponente in diesem Fall eine Vorgehensweise in Analogie zu Gl. 11.13 (partielle Korrektur einer Fehlschätzung) zugrunde liegt, in ähnlicher Art übrigens wie dies für die Schätzung der Niveaukomponente $\hat{a}_t$ gilt. Denn aus den Gl. 11.17 bis 11.19 folgt

(11.18a) $\hat{a}_t = \hat{y}_t + \alpha\,(y_t - \hat{y}_t)$

(11.19a) $\hat{b}_t = \hat{b}_{t-1} + \alpha\beta\,(y_t - \hat{y}_t)$.

Zu einer zusammenfassenden Gegenüberstellung der Gleichungen für das exponentielle Glätten vgl. Übers. 11.5.

Beispiel 11.10:

Gegeben sei die folgende Zeitreihe mit einem ansteigenden Trend:

t	y_t	$\hat{a}_t$	$\hat{b}_t$	$\hat{y}$
0	14	14	2	14
1	16	16	2	16
2	17	17,7	1,91	18
3	17	18,83	1,68	19,61
4	19	20,06	1,54	20,51
5	-	-	-	21,6

Man bestimme Prognosewerte mit der zweistufigen Methode der exponentiellen Glättung (α, $\beta = 0{,}3$).

Lösung 11.10:

Für t=1 folgt daraus: $\hat{y}_1 = \hat{a}_0 + \hat{b}_0 = 14 + 2 = 16$,

$\hat{a}_1 = \hat{y}_1 + \alpha\,(y_1 - \hat{y}_1) = 16 + 0{,}3 \cdot (16-16) = 16$

und $\hat{b}_1 = \hat{b}_0 + \alpha\,\beta\,(y_1 - \hat{y}_1) = 2 + 0{,}3 \cdot 0{,}3 \cdot (16-16) = 2$

usw. Die Prognosewerte sind in der Tabelle enthalten. Man erhält am *aktuellen Rand* für den Zeitpunkt t=5:

$$\hat{y}_5 = \hat{a}_4 + \hat{b}_4 = 20{,}06 + 1{,}54 = 21{,}6$$

Nach Gl. 11.16a kann man eine Mehrschrittprognose etwa für t=10 (i=6) wie folgt berechnen:

$$\hat{y}_{10} = \hat{a}_4 + 6\,\hat{b}_4 = 20{,}06 + 6 \cdot 1{,}54 = 29{,}3.$$

Mit der (in diesem Fall nicht angemessenen) Methode des einstufigen exponentiellen Glättens ($\alpha = 0{,}3$) erhielte man folgende Werte

$\hat{y}_1 = 14$, $\hat{y}_2 = 14{,}6$, $\hat{y}_3 = 15{,}32$, $\hat{y}_4 = 15{,}824$ und als Prognose $\hat{y}_5 = 16{,}7768$.

Übersicht 11.5: Schätzgleichungen bei exponentieller Glättung

	Glättung erster Ordnung	*Glättung zweiter Ordnung*
Prognose	$\hat{y}_{t+1} = \hat{a}_t$ [1)]	$\hat{y}_{t+1} = \hat{a}_t + \hat{b}_t$
Rekursionsgl. für $\hat{a}_t$	$\hat{a}_t = \alpha\, y_t + (1-\alpha)\, \hat{y}_t$ [1)] $= \alpha\, y_t + (1-\alpha)\, \hat{a}_{t-1}$	$\hat{a}_t = \alpha\, y_t + (1-\alpha)\, \hat{y}_t$ [2)] $= \alpha\, y_t + (1-\alpha)\, (\hat{a}_{t-1} + \hat{b}_{t-1})$
	$\hat{a}_t - \hat{a}_{t-1} = \alpha\, (y_t - \hat{y}_t)$ [3)]	$\hat{a}_t - \hat{a}_{t-1} = \alpha\, (y_t - \hat{y}_t) + \hat{b}_{t-1}$
für $\hat{b}_t$		$\hat{b}_t = \beta\, (\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\, \hat{b}_{t-1}$ $= \hat{b}_{t-1} + \alpha\beta (y_t - \hat{y}_t)$ [4)]
Startbedingung	$\hat{a}_0 = y_0$	$\hat{a}_0 = y_0$ und $\hat{b}_0 = y_1 - y_0$

1) Wegen $\hat{y}_{t+1} = \hat{a}_t$ gilt auch $\hat{y}_t = \hat{a}_{t-1}$.

2) Dies ist nicht identisch mit Gl. 11.15a, denn $\hat{y}_t$ ist hier nicht gleich $\hat{a}_{t-1}$, sondern $\hat{a}_{t-1} + \hat{b}_{t-1}$, wie auch alle Größen $\hat{a}$ zahlenmäßig nicht identisch sind mit den entsprechenden Größen auf der linken Seite.

3) Zur Interpretation vgl. Gl. 11.13.

[4] Man kann die Rekursionsgleichung für $\hat{b}$ auch so auffassen, dass man mit den Differenzen $(\hat{a}_t - \hat{a}_{t-1})$ erneut eine Glättung durchführt (Gl. 11.12a).

b) Filter, Operatoren, Polynome

In diesem Abschnitt sollen kurz einige Werkzeuge der Beschreibung (und Modellwahl) von Zeitreihen sowie Folgerungen und Zusammenhänge unter Verwendung dieser Werkzeuge vorgestellt werden.

1. Filter:

Viele Rechenoperationen mit Zeitreihen kann man als Filterung der Zeitreihe auffassen. Ein "Filter" verwandelt eine Zeitreihe y_t (input) in eine transformierte Zeitreihe (output) z_t. Einfache lineare Filter sind z.B. gleitende Mittelwerte oder die Differenzenbildung (deren Output z_t lautet: $z_t = x_t - x_{t-1}$). Ein typischer nichtlinearer Filter ist die Bildung von Wachstumsraten mit $z_t = (x_t - x_{t-1})/x_{t-1}$.

2. Operatoren:

a) Verschiebungen der Variable t bewirkt der Backshift- oder Lag-Operator: $Ly_t = y_{t-1}$, $L^2y_t = y_{t-2}$ usw., der inverse Operator heißt Vorwärts- (V) oder Leadoperator $V^s y_t = L^{-s} y_t = y_{t+s}$.

b) Nicht auf t, sondern auf die Inputvariable wirken der Vorwärtsdifferenzenoperator (delta Δ) mit $\Delta y_t = y_{t+1} - y_t$, bzw. die Rückwärtsdifferenzen (nabla δ) $\delta y_t = y_t - y_{t-1}$. Hintereinanderausführen heißt Potenzieren des Operators

$$\Delta^2 y_t = \Delta y_{t+1} - \Delta y_t = \Delta(\Delta y_t) = y_{t+2} - 2y_{t+1} + y_t \ .$$

Man beachte, dass $\Delta^2 y_t$ nicht identisch ist mit $y_{t+2} - y_t$. Man kann Vor- und Rückwärtsdifferenzen auch für mehrere Perioden definieren, etwa $\delta(4)y_t = y_t - y_{t-4}$ oder $\delta(12)y_t = y_t - y_{t-12}$ beim Vorjahresvergleich mit Quartals- oder Monatsdaten, wovon man sich eine "automatische" Saisonbereinigung verspricht, weshalb man auch von saisonalen Differenzen spricht.

c) Zwischen den unter a) und b) genannten Operatoren bestehen Zusammenhänge. So sind beispielsweise δy_t und $(1-L)y_t$ äquivalent.

3. Lagpolynome:

Der Ausdruck $A_p(t) = a_0 + a_1t + a_2t^2 + ... + a_pt^p$ ist ein Polynom in t vom Grade p und A_p heißt Polynomoperator.

Ein autoregressives Schema (eine linear-rekursive Funktion) erhält man mit dem Lagpolynom

$A_p(L)y_t = (a_0 + a_1L + a_2L^2 + ... + a_pL^p)y_t = a_0y_t + a_1y_{t-1} + a_2y_{t-2} + ... + a_py_{t-p}$.

Lineare Filter kann man als Lagpolynome darstellen, so z.B.

- ein ungewogener gleitender dreigliedriger Durchschnitt
  $y_t = (a_{-1}L^{-1} + a_0L^0 + a_1L^1)y_t = 1/3(y_{t+1} + y_t + y_{t-1})$
  mit $a_{-1} = a_0 = a_1 = 1/3$ oder
- ein (Rückwärts) Differenzenfilter
  $y_t - y_{t-1} = (1-L)y_t$ mit $a_0=1$ und $a_1=-1$.

4. Zwei Folgerungen, die für die praktische Anwendung von Bedeutung sind, sollen hier kurz dargestellt werden:

   a) Ist y_t ein Polynom vom Grade $p > 0$ dann ist $\delta y_t = y_t - y_{t-1}$ ein Polynom vom Grade p-1, so dass die p-ten Differenzen $\delta^p y_t$ eine Konstante darstellen (Beispiel 11.9).

   b) **Satz:**

   Einem Polynom $y_t = A_p(t)$ ist eine linear rekursive Funktion $B_{p+1}(L)y$ äquivalent.

   Beispiel:

   p=1: der Funktion $y_t = a_0 + a_1t$ (Polynom vom Grade 1) ist das Lagpolynom $y_t = b_0 + b_1y_{t-1} + b_2y_{t-2} = 2y_{t-1} - y_{t-2}$ ($b_0 = 0$, $b_1 = 2$, $b_2 = -1$) mit den Anfangswerten $y_0 = a_0$ und $y_1 = a_0 + a_1$ äquivalent.
   Aus $y_t = 2y_{t-1} - y_{t-2}$ folgt übrigens $\delta y_t = y_t - y_{t-1} = y_{t-1} - y_{t-2} = \delta y_{t-1} =$ const und für die zweiten Differenzen $\delta^2 y_t = 0$ (also das Verschwinden der zweiten Differenzen).

   p=2: $y_t = a_0 + a_1t + a_2t^2$
   äquivalent ist $y_t = 3y_{t-1} - 3y_{t-2} + y_{t-3}$
   mit den Anfangswerten $y_0 = a_0$, $y_1 = a_0 + a_1 + a_2$ und $y_2 = a_0 + 2a_1 + 4a_2$. Hieraus folgt $\delta^3 y_t = 0$.

   Wie man sieht, führt die Verallgemeinerung für beliebiges p zum binomischen Satz. Offenbar sind autoregressive Schemen, wenn

deren Koeffizienten bestimmte Werte annehmen, als Polynome in t zu interpretieren.

5. Weitere Hilfsmittel zur Analyse von Zeitreihen sind die ***Autokovarianz- und Autokorrelationsfunktion*** (Korrelation von y_t mit y_{t-d}, wobei d = 0, 1, 2,... der Lag ist). Sie sind Ausgangspunkt weiterer Analyseverfahren (Spektralanalyse, Box Jenkins Verfahren), auf die jedoch hier nicht eingegangen werden soll.

Beispiel 11.10:

Man bilde für die Polynome

a)$y_t = a + bt = 2 + 3t$ und

b)$y_t = a + bt + ct^2 = 2 + 3t + 0{,}5t^2$

die Differenzen erster und zweiter Ordnung und stelle die Gleichungen dar als linear rekursive Funktionen.

Lösung 11.11:

a) Polynom vom Grade 1 (lineare-, Geradenfunktion)

t	y_t	δy_t	$\delta^2 y_t$
0	2		
1	5	3	
2	8	3	0
3	11	3	0
4	14	3	0

b) Polynom vom Grade 2 (Parabel)

t	y_t	δy_t	$\delta^2 y_t$	$\delta^3 y_t$
0	2			
1	5,5	3,5		
2	10	4,5	1	0
3	15,5	5,5	1	0
4	22	6,5	1	0

Man erkennt, dass bei einem Polynom vom Grade p die p-ten Differenzen konstant (und somit die p+1-ten Differenzen Null) sind.

a) $y_t = 2 + 3t$ ist äquivalent mit $y_t = 2y_{t-1} - y_{t-2}$ mit den Anfangswerten $y_0 = 2$ und $y_1 = 2 + 3 = 5$, so dass man erhält

$y_2 = 2y_1 - y_o = 2 \cdot 5 - 2 = 8$

$y_3 = 2y_2 - y_1 = 2 \cdot 8 - 5 = 11$

$y_4 = 2y_3 - y_2 = 2 \cdot 11 - 8 = 14 \cdot$

b) $y_t = 2 + 3t + ½t^2$ ist äquivalent mit $y_t = 3y_{t-1} - 3y_{t-2} + y_{t-3}$ mit den Anfangswerten $y_0 = 2$ und $y_1 = 2 + 3 + 0{,}5 = 5{,}5$ und $y_2 = 2 + 6 + 2 = 10$, so dass man erhält

$y_3 = 3y_2 - 3y_1 + y_0 = 3\cdot10 - 3\cdot5{,}5 + 2 = 15{,}5$

$y_4 = 3y_3 - 3y_2 + y_1 = 3\cdot15{,}5 - 3\cdot10 + 5{,}5 = 22$ usw.

c) Fourieranalytische Methoden

Eine kurze heuristische Einführung in die fourieranalytischen Methoden der Zeitreihenanalyse, auf die hier nicht weiter eingegangen werden kann, sollte zeigen, dass auch recht komplizierte Kurvenverläufe als Addition trigonometrischer Polynome "entstehen" können, so dass "umgekehrt" die Analyse einer empirischen Zeitreihe auch darin bestehen kann, zu zeigen, aus welchen Schwingungen sie sich zusammensetzt.

In der Funktion

$$(11.21) \quad y(t) = A\cdot\sin(Bt + C)$$

ist A die Amplitude (Maximalausschlag der Sinus-Schwingung), B die Kreisfrequenz (Anzahl der Schwingungen im Intervall $[0,2\pi]$; das übliche Symbol ist ω statt B) und C die Phase (Verschiebung gegenüber dem Ursprung).

Mit $a = A\cdot\sin(C)$ und $b = A\cdot\cos(C)$ läßt sich Gl. 11.21 umformen zu

$$(11.22) \quad y(t) = a\cdot\cos(Bt) + b\cdot\sin(Bt) \qquad \text{mit } 0 < B \le \pi$$

mit der Amplitude $A = \sqrt{a^2 + b^2}$.

Umfasst z.B. die gesamte Länge einer Zeitreihe T = 120 Monate, so erstreckt sich eine Sinusschwingung über die volle Länge von 120 Monaten, wenn $B = 2\pi/120$ ist. Man erhält zwei Schwingungen über den gesamten Bereich, wenn die Frequenz 2B ist, weil dann die Periodenlänge (Wellenlänge) $2\pi/2B = (2\pi)120/4\pi = 60$ Monate beträgt. Entsprechend bedeutet eine Frequenz von kB (mit k = 1, 2,..., 60), dass k Schwingungen ausgeführt werden, die jeweils die Periodenlänge von 120/k Monaten (allgemein T/k Einheitsintervalle) haben. Statt $\sin(\lambda_k Bt)$ mit $B = 2\pi/T$ kann man auch schreiben $\sin(\lambda_k 2\pi t)$ mit $\lambda_k = k/T$ als Frequenz (im Unterschied zur Kreisfrequenz B), d.h. λ_k ist die Anzahl der Schwingungen zwischen t = 0 und t = 1.

Eine Überlagerung derartiger Schwingungen, d.h. ein Ausdruck der Gestalt

(11.23) $y(t) = \Sigma[a_k \cos(kBt) + b_k \sin(kBt)]$ mit k = 1, 2, 3,...
und a_k, b_k = Fourierkoeffizienten

kann eine unregelmäßig erscheinende Kurve darstellen. Das bedeutet auch, dass sich eine unregelmäßige Kurve y(t), bzw. deren diskrete Werte y_t als Summe von Schwingungen verschiedener Frequenzen und Amplituden zerlegen lässt. Eine Darstellung der relativen Bedeutung von Schwingungen verschiedener Frequenz ist das Periodogramm.

Im folgenden Beispiel wurde quasi Gl. 11.23 von rechts nach links betrachtet, d.h. es wurden zwei Schwingungen g(t) und h(t) konstruiert und die sich daraus ergebende Gestalt der "Zeitreihe" y(t) hinsichtlich der "Bedeutung" (gemessen an den Amplituden) bestimmter Frequenzen interpretiert. In der Praxis ist die Blickrichtung natürlich genau umgekehrt, d.h. bei gegebener Zeitreihe y(t) (bzw. y_t) wird von der linken auf die rechte Seite der Gl. 11.23 geschlossen, d.h. auf die der Zeitreihe zugrundeliegenden Schwingen. Wie dies geschieht, kann hier nicht dargestellt werden. Das gilt auch für die Interpretation des Periodogramms, das grob gesprochen angibt, in welchem Ausmaß eine Zeitreihe von kürzer- und längerfristigen Vorgängen geprägt ist.

Beispiel 11.13:

Bestimmen Sie die Gestalt der Überlagerung $y_1(t)$ der Funktionen

$g_1(t) = 20 \sin(2\pi \cdot 0{,}35t)$ und
$h_1(t) = 5 \sin(2\pi \cdot 0{,}07t)$
$y_1(t) = g_1(t) + \ln(t)$ für t = 0 bis t = 30

Man berechne ferner das Periodogramm für $f_1(t)$ und für $f_2(t)$ mit

$y_2(t) = 20 \sin(2\pi \cdot 0{,}07t) + 5 \sin(2\pi \cdot 0{,}35t)$

$y_1(t) = h_2(t) + g_2(t)$.

Lösung 11.13:

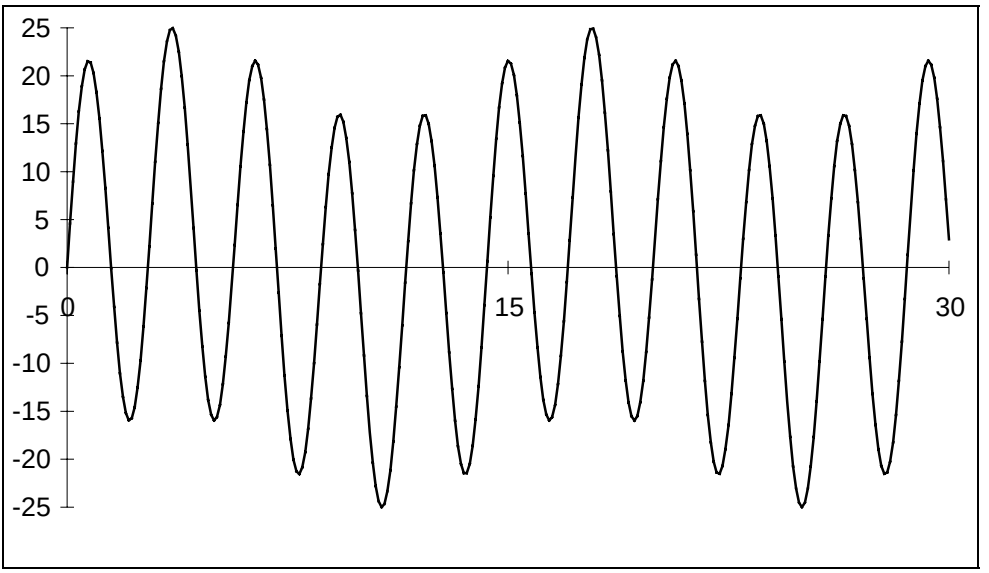
Der obere Teil der Abb. 11.3 und 11.4 zeigt die Funktionen $y_1(t)$ und $y_2(t)$ für t = 0 bis t = 30, d.h. unterschiedliche Überlagerungen von zwei Sinusschwingungen. Wie man sieht ist $y_1(t) = g_1(t) + h_1(t)$ als Zeitreihe von ganz anderer Gestalt als $y_2(t) = g_2(t) + h_2(t)$.

Die Periodogramme[1] zeigen, dass $y_1(t)$ stark von einer kurzen Welle (hohe Frequenz 0,35) geprägt ist, während es bei $y_2(t)$ genau umgekehrt ist. Das Beispiel wurde auch so konstruiert, dass in

- $y_1(t)$ die hochfrequente Reihe ($g_1(t)$) die dominierende Amplitude besitzt und in
- $y_2(t)$ "spiegelbildlich" die niedrigfrequente (langwellige) Reihe ($h_2(t)$) die dominierende Amplitude besitzt.

Das Ergebnis ist natürlich nicht überraschend, weil das Periodogramm jeweils genau das zu Tage gefördert hat, was in die fiktiven Zeitreihen "hineingelegt" wurde. Aber man erkennt an dieser extrem vereinfachten Betrachtung nicht nur wie unterschiedlich die konstruierten Zeitreihen im Zeitbereich (oberer Teil der Abbildungen) aussehen können, sondern auch dass sich dies bei der spektralen Darstellung im Frequenzbereich (unterer Teil der Abb. 11.3 und 11.4) widerspiegelt.

Abb. 11.3: Zeitreihe und Periodogramm von $y_1(t)$



[1] Es wurde berechnet von Herrn Thomas Lungwitz aufgrund von diskreten Zeitreihen (t=1,...,300) von $y_1(t)$ und $y_2(t)$ mit dem Programm RATS im Rahmen seiner Diplom-Arbeit im Fach Statistik. Empirische Zeitreihen haben i.d.R. eine diskrete Zeitvariable t. Zehn gleiche Abschnitte wie in Abb. 11.3 und 11.4 so dass T = 300 ist (statt 30 Werte wie in den Abbildungen) wurden gewählt, damit die Schätzung des Periodogramms zuverlässig ist.

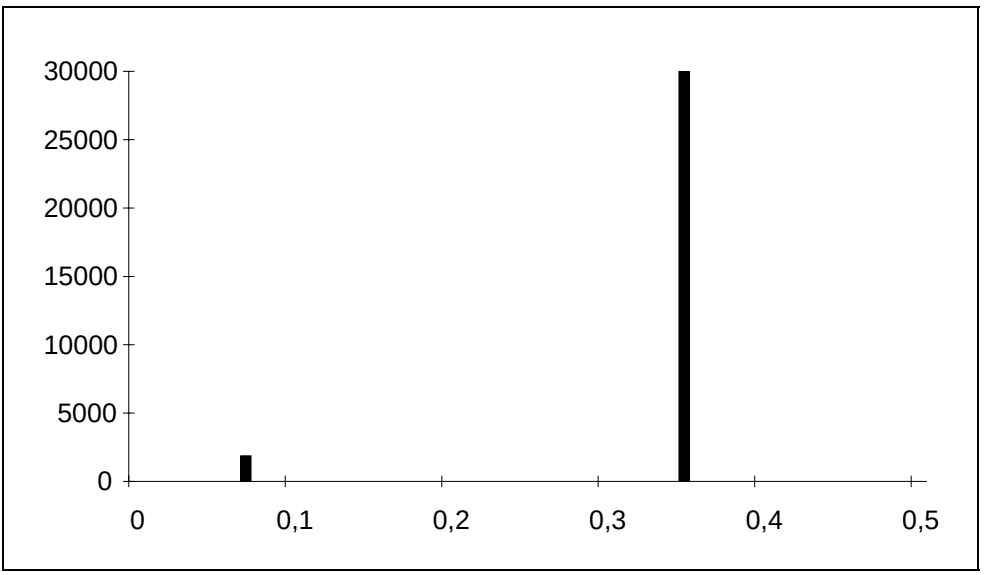
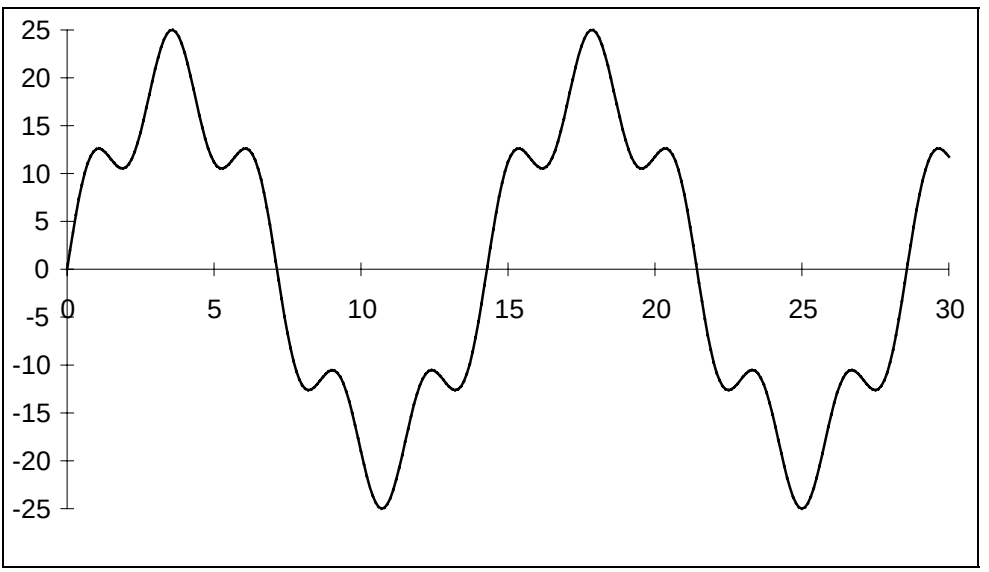
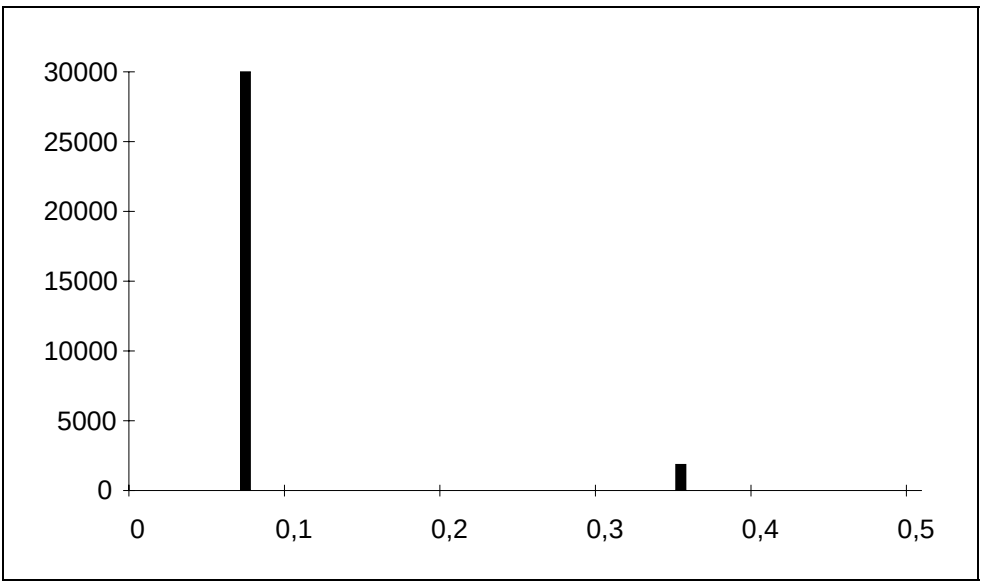


Abb. 11.4: Zeitreihe und Periodogramm von $y_2(t)$




30000
25000
20000
15000
10000
5000
0
0
0,1
0,2
0,3
0,4
0,5

Kapitel 12: Bestandsanalyse und Tafelrechnung

1. Bestands- und Bewegungsmassen

a) Definitionen

Gegenstand der Bestandsanalyse ist die (graphische) Darstellung und Beschreibung (durch Kennzahlen) von Bestandsveränderungen durch laufend auftretende Zu- und Abgänge. Es ist durch geeignete Kennzahlen diese Dynamik des Bestands und das Ein- und Austrittsverhalten der Einheiten darzustellen.

Def. 12.1: Bestandsmasse, Bewegungsmasse, Verweildauer

a) Eine statistische Masse, deren Einheiten (i=1,2,...,n) jeweils gemeinsam zu einem bestimmten Zeit**punkt** t_j in einem Bestand (über eine nicht näher bestimmte Zeit) verweilen, heisst **Bestandsmasse** (engl. **stock**). Der Umfang der Bestandsmasse zum Zeitpunkt t_j heißt Bestand $B(t_j) = B_j$. Er ist zu jedem Zeitpunkt $t = t_j$ durch die Bestandsfunktion B(t) gegeben. Die Zeit kann als diskrete ($t = t_0, t_1,...,t_j,...,t_m$) oder stetige Variable betrachtet werden.

b) Eine statistische Masse, deren Einheiten dadurch charakterisiert sind, dass sie zu einem bestimmten Zeitpunkt ihren Zustand ändern (was ein "Ereignis" darstellt) heißt **Bewegungsmasse** (Ereignismasse, Stromgröße, engl. **flow**). Der Umfang einer Bewegungsmasse ist die Anzahl derartiger Ereignisse in einem gegebenen Zeit**raum**

(Zeitintervall). Zustandsänderung kann insbesondere bedeuten: Zugang zu oder Abgang von einer Bestandsmasse.

c) Jede Einheit einer Bewegungsmasse (i=1,2,...,n) ist durch Zugangszeit (t_{zi}) und Abgangszeit (t_{ai}) gekennzeichnet. Der Zeitraum zwischen Zu- und Abgangszeit $d_i = t_{ai} - t_{zi}$ heißt **Verweildauer** (Verbleibdauer).

Bemerkungen zu Def. 12.1:

1. Einheiten einer Bestandsmasse haben eine Verweildauer $d_i>0$, d.h. sie befinden sich über einen nicht näher definierten Mindestzeit**raum** (Strecke) "im Bestand", während eine Bewegungsmasse aus Ereignissen besteht, die quasi in einem Zeit**punkt** passieren.
   Man spricht deshalb auch von Strecken- und Punktmasse, was jedoch insofern etwas verwirrend sein mag, weil erstere zu einem Zeitpunkt, zweite zu einem Zeitraum festgestellt wird (Nr. 2).

2. Umfang und Struktur einer Bestandsmasse wird zu einem Stichtag t_j erfaßt, während eine Bewegungsmasse für ein Zeitintervall $[t_0,t_j]$ definiert ist und i.d.R. sekundärstatistisch durch laufende Registrierung festgestellt wird.

3. Jeder Bestandsmasse (z.B. der Wohnbevölkerung) sind zwei Bewegungsmassen zugeordnet, die Zugangs- (Geburten und Einwanderungen) und die Abgangsmasse (Todesfälle und Auswanderungen).

   Das heißt jedoch nicht, dass jede Bewegungsmasse bestandsverändernd wirken muss: wegen Mehrfachbeschäftigung kann z.B. der Bestand an beschäftigten Personen nicht einfach mit der Zahl begonnener und beendeter Beschäftigungsverhältnisse fortgeschrieben werden.

4. Beispiele:

Bestandsmasse (stock)	Bewegungsmasse (flow)
Wohnbevölkerung	Geburten, Todesfälle, Wanderungen
Bruttoanlagevermögen	Investitionen, Verschrottungen
Kontostand	Gutschrift, Lastschrift
Vermögen	Einkommen
Auftragsbestand	Auftragseingang, Umsatz

Methoden der Erhebung von Bestands- und Bewegungsmassen:

Es gibt mehrere Möglichkeiten, Bestands- und Bewegungsmassen zu erfassen. Hiervon hängt es ab, wie informativ die Daten sind und welche Methoden zur Auswertung der Daten angewendet werden können.

1. Feststellung der Bewegungen (Bewegungsmassen)

a) durch individualisierte Erhebung aller Verläufe, d.h. für jede Einheit werden Zugangs- und Abgangszeit festgestellt (= Längsschnitts- oder Verlaufsanalyse);

b) laufende Registrierung aller Bestandsveränderungen und Auswertung der über ein Beobachtungsintervall (von t_0 bis t_j) kumulierten Zugänge (Z_{0j}) und Abgänge (A_{0j}), d.h. der Bruttoströme.

c) Feststellung der Bestandsveränderungen (d.h. der Salden- oder Nettoströme Z_{0j}-A_{0j}).

Im Vergleich zu den Bruttoströmen stellen Nettoströme eine erhebliche Verringerung des Informationsgehalts dar: es ist ein Unterschied, ob z.B. der Arbeitslosenbestand um 100.000 Personen wächst, weil 100.000 Personen arbeitslos geworden sind (Z_{0j} = 100.000) und niemand aus der Arbeitslosigkeit ausgeschieden ist (also A_{0j} = 0), oder ob es einen größeren Umschlag gab, also z.B. gilt Z_{0j} = 500.000 und A_{0j} = 400.000. Der Unterschied betrifft die mittlere Verweildauer und die Umschlagshäufigkeit der Einheiten.

2. Feststellung der Bestände (Bestandsmassen)

a) durch periodische Inventuren (Zählen oder Messen)

b) durch Fortschreibung für das Intervall $[t_0,t_j]$:

(12.1) $B_j = B_0 + Z_{0j} - A_{0j} \quad (j = 0,1,...m)$

In Gl. 12.1 ist B_0 der Anfangsbestand, B_j der Bestand zum Zeitpunkt t_j, Z_{0j} die Anzahl der Zugänge und A_{0j} die Anzahl der Abgänge im Beobachtungsintervall $[t_0,t_j]$.

c) Bei Kenntnis sämtlicher individueller Verläufe (wie in 1a), also bei Längsschnittdaten, ist, wie noch gezeigt wird, der Bestand zu jedem beliebigen Zeitpunkt bekannt.

Unter Querschnittsanalysen versteht man die Kombination 1b + 2a und unter Längsschnittsanalysen die Kombination 1a + 2c.

Die durch Gl. 12.1 definierte "Fortschreibung" einer Bestandsmasse erlaubt auch eine "**Bilanzdarstellung**" von Bestandsmassen (Anfangs- [=B_0] und Endbestand [=B_j]) :

(12.1a) $N_{0j} = B_0 + Z_{0j} = B_j + A_{0j}$ (N_{0j} = Bilanzsumme)

Aktiv	Passiv
B_0	B_j
Z_{0j}	A_{0j}

Die "Bilanzsumme" N_{0j} ist die Anzahl der Einheiten, die im Intervall $[t_0,t_j]$ jemals der Bestandsmasse angehörten (wobei eine Einheit auch mehrfach gezählt werden kann). Es ist eine Anzahl von Bewegungen (Zu- und Abgängen), von Ein- und Austritts**fällen**, nicht die Anzahl n der daran beteiligten **Personen** ($n \le N_{0j}$), weil eine Einheit (z.B. eine Person) mehrmals ein- und austreten kann.
Die Gleichungen 12.1 und 12.1a können für beliebige Intervalle aufgestellt werden, z.B. Gl. 12.1a auch für das gesamte Beobachtungsintervall von t_0 bis t_m

(12.1b) $N_{0m} = B_0 + Z_{0m} = B_m + A_{0m}$.

Aus der obigen Bilanzdarstellung kann man auch eine **kombinierte Fluß- und Bestandsgrößendarstellung** herleiten. Unterscheidet man die beiden Sektoren "Außenwelt"(AW) und "Bestand" oder "System"(BS) so ergeben sich folgende "Lieferbe-ziehungen":

	nach		
von	AW	BS	Σ
AW	**	Z_{0m}	**
BS	A_{0m}	R_{0m}	B_0
Σ		B_m	

Die Summe der drei Größen A_{0m}, Z_{0m} und R_{0m} ist N_{0m}.

Die mit ** bezeichneten Felder sind meist nicht von Interesse. In der Summenzeile, bzw. -spalte erscheinen End-, bzw. Anfangsbestand. R_{0m} sind die Einheiten des Anfangsbestan-
des die während des ganzen Intervalls im System geblieben sind. Die Abgänge A_{0m} kann man als "Lieferung" des Systems an die Außenwelt auffassen.

b) Beckersches Diagramm, Bestandsfunktion und Zeitmengenfläche

1. Beckersches Diagramm:

Eine graphische Darstellung der individuellen Verläufe ist das Beckersche Diagramm (Abb. 12.1 für das Beispiel 12.1). Die 45^o - Linie (Zugangsachse) z ist Konsequenz dessen, dass Zugangs- und Kalenderzeit (Abszisse t) synchron sind. Kreise bezeichnen Zugänge (auf der Achse z) und Kästchen Abgänge. Ihre horizontale Verbindung ist die Verweillinie der Einheit i, deren Länge die Verweildauer d_i ist. In Abb. 12.1 wurde davon ausgegangen, dass jeweils nur eine Einheit zu- oder abgeht.

2. Bestandsfunktion:

Es ist leicht zu sehen, wie aus dem Beckerschen Diagramm (oberer Teil von Abb. 12.1) die Bestandsfunktion B(t) (t stetig), bzw. B_j (Bestände zu den diskreten Zeitpunkten t_j) herzuleiten ist. Mit jedem Zugang (Abgang) einer Einheit erhöht (verringert) sich die Bestandsfunktion um 1.

3. Zeitmengenfläche:

Die schraffierte Fläche unter der Bestandsfunktion heißt Zeitmengenfläche F, oder genauer F_{0m} wenn die Fläche "über" dem Intervall $[t_0,t_m]$ betrachtet wird.

Beispiel 12.1:

Wegen des zur Nachsaison unsicheren Wetters ist das Badevergnügen oft von nur kurzer Dauer. Andererseits ergreifen jedoch die Urlauber wegen der ihnen entstandenen Kosten der weiten Reise jede sich bietende Gelegenheit, den Strand aufzusuchen. Am Strand von Katapulco gab es mithin an einem Vormittag (9 - 13 Uhr) ein ständiges Kommen und Gehen von (zwecks Rechenvereinfachung) nur fünf

Urlaubern A,...,E. Für die Zeiten galt:

	Zeitpunkt des Zugangs t_{zi}	Zeitpunkt des Abgangs t_{ai}
A	09^{30}	10^{00}
B	09^{45}	10^{45}
C	10^{30}	12^{30}
D	10^{45}	11^{15}
E	11^{45}	12^{45}

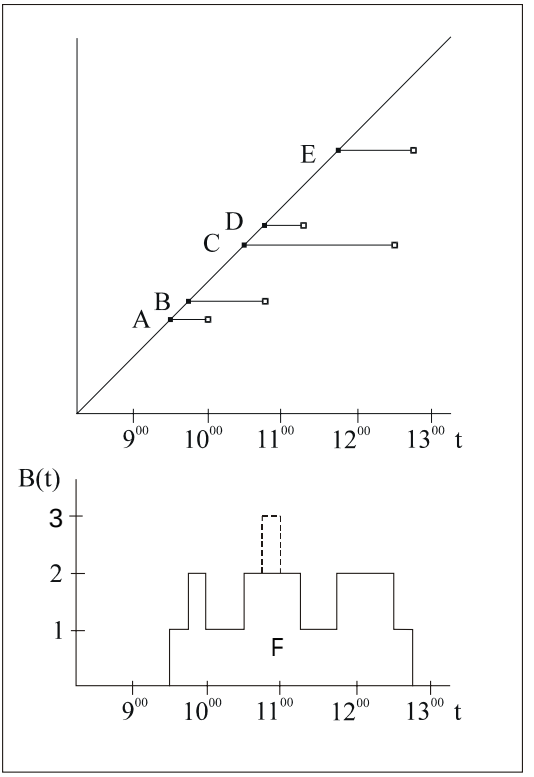
Man zeichne das Beckersche Diagramm und bestimme die Verweildauerverteilung.
Modifikation: B kommt um 09^{45} und geht nicht um 10^{45}, sondern um 11^{00}.

Lösung 12.1:

Man erhält für die Verweildauerverteilung:

	Zeitpunkt des Zugangs	Zeitpunkt des Abgangs	Verweildauer d_i (in Stunden)	Verweildauer bei der
A	09^{30}	10^{00}	0,5	0,5
B	09^{45}	10^{45}	1,0	1,25
C	10^{30}	12^{30}	2,0	2,0
D	10^{45}	11^{15}	0,5	0,5
E	11^{45}	12^{45}	1,0	1,0
		Verweilsumme	$\Sigma d_i = 5{,}0$	$\Sigma d_i = 5{,}25$

Abb. 12.1: Beckersches Diagramm und Bestandsfunktion Beispiel 12.1



Def. 12.2: Zeitmengenfläche, -bestand

Die Zeitmengenfläche F_{0j} des Intervalls $[t_0,t_j]$ ist die Fläche "unter" der Bestandsfunktion B(t) im Intervall $[t_0,t_j]$.

Bemerkungen zu Def. 12.2:

1. Im Beispiel 12.1 ist die Zeitmengenfläche $F_{0m} = 5$, das ist die schraffierte Fläche in Abb. 12.1, bzw. 5,25 im modifizierten Bsp. 12.1.

2. Die Zeitmengenfläche ist eine gewogene Summe der Anzahl der Einheiten, gewogen mit der Zeit, die sie in diesem Intervall im Bestand verbringen. Bei einer geschlossenen Masse (vgl. Def. 12.3) ist die Zeitmengenfläche gleich der Verweilsumme Σd_i.

3. Die Maßeinheit der Zeitmengenfläche ist je nach zugrundeliegendem Beobachtungsintervall (z.B. Stunden, Jahre) "Personenstunden" oder "-jahre". Sie hat eine Anzahl- (vertikal) und eine Zeitdimension (horizontal) und erlaubt deshalb die Herleitung des Durchschnittsbe-

stands $\bar{B}$ einerseits und der durchschnittlichen Verweildauer $\bar{d}$ und der Umschlagshäufigkeit U andererseits (vgl. Übers. 12.1).

Satz 12.1:

Das Beckersche Diagramm ist umfassender (informativer) als

a) die zeitliche Verteilung der Zu- und Abgänge, bzw. die im Intervall von t_0 bis t_j kumulierten Zu- und Abgänge (Z_{0j}, A_{0j})
b) die Bestandsfunktion B_j (Bestand zu jedem Zeitpunkt $t = t_j$)
c) die Verteilung der Verweildauer d_i, also die Häufigkeitsverteilung der Variable Verweildauer. Denn:

Man kann vom Beckerschen Diagramm eindeutig auf a), b) oder c) schließen, nicht aber umgekehrt.

Das liegt daran, dass das Beckersche Diagramm individualisierte Verläufe enthält, bei den genannten Funktionen dagegen jeweils Individuen zusammengefaßt werden (vgl. hierzu Beispiel 12.2).

Beispiel 12.2:

Gegeben seien die folgenden vier Fälle von Bestandsveränderungen (jeweils eine geschlossene Masse [Def. 12.3]):

	Fall 1		Fall 2		Fall 3		Fall 4	
Einheit	t_{zi}	t_{ai}	t_{zi}	t_{ai}	t_{zi}	t_{ai}	t_{zi}	t_{ai}
A	1	4	1	3	1	3	1	2
B	2	3	2	4	2	4	2	5
C	3	4	3	4	3	4	4	6
D	4	6	4	6	4	7	5	6
E	5	7	5	7	5	6	5	7

Wie unterscheiden sich diese Fälle hinsichtlich

- der zeitlichen Verteilung der Zu- und Abgänge,
- der Bestandsfunktion,
- der Verweildauerverteilung und
- hinsichtlich des Beckerschen Diagramms?

Lösung 12.2:

Die zeitliche Verteilung (Zeitreihe) der Zu- und Abgänge ist in den Fällen 1 - 3 gleich: es geht jeweils eine Einheit zu den Zeitpunkten 1,2,3,4 und 5 zu. Es gibt drei Zeitpunkte (3,6 und 7), an denen jeweils eine Einheit und einen Zeitpunkt (4) an dem jeweils zwei Einheiten abgehen. Welche Einheiten jeweils abgehen ist jedoch verschieden. Konsequenz: gleiche Bestandsfunktion, aber unterschiedliche Beckersche Diagramme und folglich reicht es nicht aus, Zeitreihen über Zu- und Abgänge zu haben. Solche Daten erlauben keinen eindeutigen Schluß auf die Verweildauerverteilung. Denn es unterscheiden sich die Verteilungen der Verweildauer d_i. Man erhält die folgenden Verteilungen in den Fällen Nr.1, 3 und 4:

d_i	n_i
1	2
2	2
3	1

während man im Fall 2, wie man leicht sieht, eine andere Verteilung der Verweildauer erhält. Es kann sogar sein, dass sowohl die Zeitreihen der Zu- und Abgänge, als auch die Verweildauerverteilungen gleich sind (Fälle 1 und 3) und trotzdem hat das Beckersche Diagramm eine andere Gestalt: auch wenn es z.B. jeweils nur eine Einheit ist, die eine Verweildauer von nur einer Periode hat, so unterscheiden sich doch die Zu- und Abgangszeitpunkte (im Fall 1 ist dies die Einheit B, die zum Zeitpunkt 2 zu- und zum Zeitpunkt 3 abgeht, im Fall 3 ist dies die Einheit C, die in 3 zu- und in 4 abgeht).

Zusammenfassung:

Abkürzungen: Zeitreihen der Zu- und Abgänge (ZZA), Verweildauerverteilung (VV), Beckersches Diagramm (BD).

Vergleich der Fälle	gleich ist	ungleich ist*)
1 mit 2	ZZA	VV, BD
1 mit 4 und 3 mit	VV	ZZA, BD
1 mit 3	ZZA und VV	nur BD

*) ungleich ist bei **jedem** Vergleich das Beckersche Diagramm (BD).

c) Offene und geschlossene Massen

Def. 12.3: offene-, geschlossene Masse

Eine Bestandsmasse heißt geschlossen bezüglich des Zeitintervalls $[t_0,t_m]$, wenn keine ihrer Einheiten vor t_0 zugegangen ist und nach t_m abgeht (endgültig aus dem Bestand ausscheidet). Eine Masse, die nicht beidseitig geschlossen ist, heißt offene Masse. Man kann auch halbseitig und beidseitig offene Massen unterscheiden.

Im Beispiel 12.1 ist die Bestandsmasse der Badegäste geschlossen bezüglich des Intervalls 9 bis 13 Uhr, oder auch von 9^{10} bis 12^{50}, sie ist dagegen offen bezüglich des Intervalls 9^{50} bis 12^{10} (Anfangsbestand $B_0 = 2$, Endbestand $B_m = 2$)

Satz 12.2:

Bei einer geschlossenen Masse gilt:

1. Es gibt keinen Anfangs- und keinen Endbestand ($B_0 = B_m = 0$);
2. Zugänge = Abgänge (im Bsp. 12.1: $Z_{0m} = A_{0m} = 5$), denn alle Zugänge, die nur nach t_0 stattfanden sind auch vor t_m wieder abgegangen;
3. die Zeitmengenfläche (F_{0m}) ist gleich der Verweilsumme (Σd_i) (im Beispiel ist $F_{0m} = \Sigma d_i = 5$), d.h. es gilt:

(12.2a) $Z_{0m} = A_{0m} = N_{0m} = N$
(12.2b) $F_{0m} = \Sigma d_i$.

Zu weiteren Besonderheiten einer geschlossenen Masse vgl. Bem. 3 zu Def. 12.5.

2. Kennzahlen der Dynamik eines Bestands: Durchschnittsbestand, durchschnittliche Verweildauer und Umschlagshäufigkeit

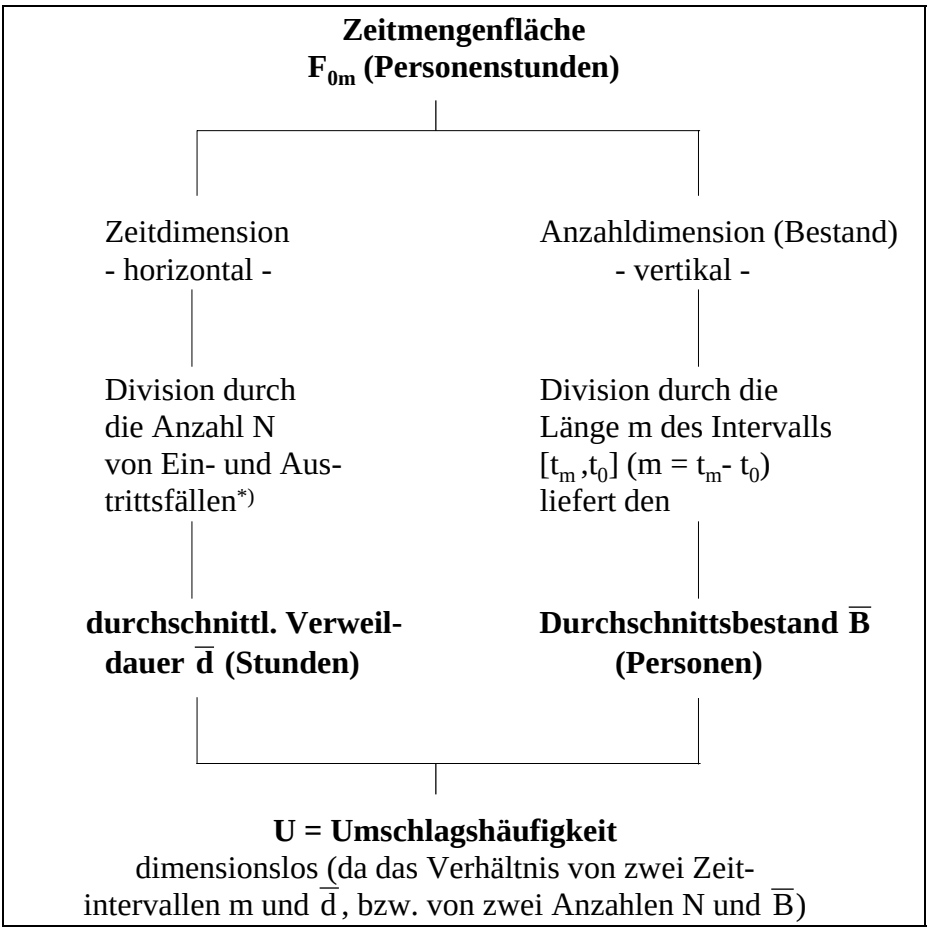
a) Einführende Übersicht

Gegenstand der Bestandsanalyse ist, wie gesagt, die Beschreibung der Dynamik eines Bestandes, d.h. der Bestandsveränderung aufgrund laufender Zu- und Abgänge durch geeignete Kennzahlen, wie

- Durchschnittsbestand ($\bar{B}$),
- durchschnittliche Verweildauer ($\bar{d}$) und
- Umschlagshäufigkeit (U).

Übersicht 12.1 zeigt den Zusammenhang zwischen diesen im folgenden dargestellten Kennzahlen zur Beschreibung der Bestandsentwicklung. Die Zeitmengenfläche F_{0m} (d.h. die Fläche unter der Bestandsfunktion) ist die Grundlage für alle weiteren Berechnungen. Sie hat eine Zeit- und eine Anzahldimension; ihre Maßeinheit ist deshalb z.B. bei Stunden als Einheitsintervall "Personenstunden".

Übers. 12.1: Kennzahlen zur Beschreibung der Bestandsentwicklung



*) Die Anzahl $N = N_{0m}$ ist eine Anzahl von Fällen, nicht notwendig gleich der Anzahl n von Personen (Einheiten) die ein- und ausgetreten (zu- und abgegangen) sind.

Für die Berechnung des Durchschnittsbestands ($\bar{B}$) reichen (anders als für $\bar{d}$ und U) Inventuren und Fortschreibungen aus. Es ist auch gleichgültig, ob die Masse offen oder geschlossen ist. Die korrekte Bestimmung der durchschnittlichen Verweildauer ist aber bei solchen Querschnittsdaten nicht möglich. Bei offenen Massen sind ferner in höherem Maße Schätzungen vorzunehmen als bei geschlossenen Massen (vgl. Übers. 12.2).
Es ist deshalb im folgenden zu unterscheiden ob Längsschnitts- oder Querschnittsdaten vorliegen und dabei jeweils, ob die Bestandsmasse im Beobachtungsintervall geschlossen oder offen ist.

b) Kennzahlen bei Kenntnis der individuellen Verläufe (Längsschnittsdaten)

1. Bei einer geschlossenen Masse

Bei Kenntnis aller individueller Zu- und Abgangszeiten ist zu allen Zeitpunkten die Bestandsfunktion bekannt. Sie ist eine stetige Funktion B(t), weil Zu- und Abgänge zu beliebigen Zeiten stattfinden können. Aus dem Konzept der Bestandsfunktion folgt unmittelbar die Definiton des Durchschnittsbestands.

Def. 12.4: Durchschnittsbestand

Zwei Bestandsfunktionen $B_1(t)$ und $B_2(t)$ sind im Intervall $[t_0,t_m]$ der Länge $m = t_m - t_0$ "äquivalent", wenn sie die gleiche Zeitmengenfläche F_{0m} haben. Der Durchschnittsbestand ist dann als derjenige konstante Bestand definiert, der äquivalent der beobachteten Bestandsfunktion ist. Folglich gilt:

$$(12.3)\quad \bar{B} = \frac{F_{0m}}{m}$$

Abb. 12.2 veranschaulicht den Gedanken. Die Fläche (schraffiert) unter der tatsächlichen Bestandsfunktion (linker Teil von Abb. 12.2) ist gleich der Fläche unter der Konstanten $\bar{B}$ (rechter Teil), da gem. Gl. 12.3 gilt:

$m\bar{B} = F_{0m}$.

Das im folgenden behandelte Konzept der durchschnittlichen Verweildauer ist am einfachsten im Falle einer geschlossenen Masse einzuführen.

Def. 12.5: durchschnittliche Verweildauer, Umschlagshäufigkeit

a) Die durchschnittliche Verweildauer $\bar{d}$ ist das arithmetische Mittel der Verweildauerverteilung:

$$(12.4)\quad \bar{d} = \frac{\Sigma d_i}{N}.$$

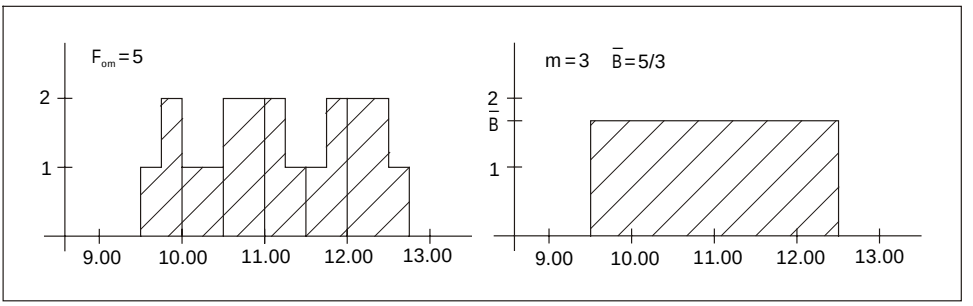
Nach Satz 12.2 gilt bei einer geschlossenen Masse $\Sigma d_i = F_{0m}$, so dass bei geschlossenen Massen Gl. 12.4 äquivalent ist mit

$$(12.4a)\quad \bar{d} = \frac{F_{0m}}{N}\ .$$

b) Die Umschlagshäufigkeit U ist gegeben durch

$$(12.5)\quad U = \frac{m}{\bar{d}}.$$

Abb. 12.2: Durchschnittsbestand



Bemerkungen zu Def. 12.5

1. Die Umschlagshäufigkeit U bringt zum Ausdruck, ob die durchschnittliche Verweildauer $\bar{d}$ größer (dann ist U < 1) oder kleiner (U > 1) als das Beobachtungsintervall m ist. Ist $\bar{d}$ < m, so können die Einheiten im Durchschnitt nicht die ganze Beobachtungszeit über im Bestand sein; der Bestand muss folglich mindestens einmal "umgeschlagen" sein. Bei gegebener Länge des Beobachtungsintervalls ist die Umschlagshäufigkeit indirekt proportional zur mittleren Verweildauer: je kürzer die Einheiten im Durchschnitt im Bestand verweilen, desto häufiger muss ein (konstanter) Bestand umgeschlagen sein.

2. Aus Gl. 12.3 und 12.4 folgt unmittelbar, dass auch gilt :

$$(12.5a)\quad U = \frac{N}{\bar{B}}$$

d.h. dass U das Verhältnis zwischen der Anzahl der Bewegungen N und dem Durchschnittsbestand $\bar{B}$ ist. Ist $Z_{0m} > \bar{B}$, also die Anzahl der Zugänge größer als der Durchschnittsbestand, so muss der Bestand mehrmals erneuert worden sein, also U > 1 sein.

3. Bei einer geschlossenen Masse ist stets $U \geq 1$, weil $\bar{d} \leq m$.

4. Im Falle einer Längsschnittsanalyse ist nicht nur die durchschnittliche Verweildauer, sondern die gesamte Verteilung der Verweildauer bekannt. Bei Querschnittsdaten kann es sein, dass die Verweildauerverteilung schwer zu schätzen ist, nicht aber deren Mittel, die durchschnittliche Verweildauer.

5. Im Beispiel 12.1 gilt:
   - durchschnittliche Verweildauer $\bar{d} = (1/N)\Sigma d_i = (1/n)\Sigma d_i = 5/5 = 1$ also eine Stunde;
   - Durchschnittsbestand (für die Zeit von 9 bis 13 Uhr) $\bar{B} = F_{0m}/(t_m - t_0) = 5/4 = 1{,}25$ Personen;
   - Umschlagshäufigkeit $U = m/\bar{d} = N/\bar{B} = 4$ (Bestand schlägt im Beobachtungsintervall viermal um). Zur Interpretation vgl. auch das Beispiel 12.3.

Beispiel 12.3:

Ein Lager werde zur Zeit $t_0 = 0$ mit vier Waren (A,...,D) gefüllt und der Lagerbestand soll während der ganzen Beobachtungszeit (von $t_0 = 0$ bis $t_m = 8$) konstant 4 betragen:

a) alle vier Waren haben die gleiche Verweildauer von 4 Perioden
b) zwei der vier Waren (A,B) haben eine Verweildauer von 2 Perioden und zwei Waren (C,D) von 4 Perioden.

Man bestimme die den beiden Teilen zugrundeliegenden Verteilungen der Verweildauer sowie die Umschlagshäufigkeit des Lagers in beiden Fällen.

Lösung 12.3:

Bezeichnet man die Waren von Typ A mit A_1, A_2,... wenn sie jeweils durch eine Ware des gleichen Typs A ersetzt werden (entsprechend B_1,B_2 usw.). Man erhält nun im Falle

a) Der Bestand besteht von t=0 bis t<4 aus A_1,B_1,C_1 und D_1 und von t=4 bis t<8 aus A_2, B_2,C_2 und D_2. Die Anzahl N der Ein- bzw. Austritte (Auslagerungen) ist 8, denn 4 Einheiten werden zur Zeit t=0 eingelagert (und bei t=4 ausgelagert) und 4 Einheiten werden zur Zeit t=4 ein- und zur Zeit t=8 ausgelagert. Der

Durchschnittsbestand ist $\bar{B} = 4$ und die Verweildauerverteilung hat die folgende Gestalt:

Verweildauer	$d_1 = 2$	$d_2 = 4$
Anzahl der Fälle	0	8

so dass die durchschnittliche Verweildauer $\bar{d} = 4$ und die Umschlagshäufigkeit 2 ist (denn das Lager wurde zweimal, zur Zeit t=0 und zur Zeit t=4 vollständig neu gefüllt).

b) Der Bestand besteht nun zwischen t=0 und t<2 aus A_1,B_1,C_1,D_1 und zwischen t=2 und t<4 aus A_2,B_2,C_1 und D_1, dann (t=4 bis t<6) aus A_3,B_3,C_2,D_2 und schließlich (t=6 bis t=8) aus A_4,B_4,C_2,D_2. Die Anzahl der Bewegungen (Ein- und Auslagerungen) ist 12 (nämlich $A_1,...,A_4,...,D_1,D_2$, so dass man die folgende Verweildauerverteilung erhält:

Verweildauer	$d_1 = 2$	$d_2 = 4$
Anzahl der Fälle	8	4

und die mittlere Verweildauer ist somit $2/3 \cdot 2 + 1/3 \cdot 4 = 32/12 = 8/3 = 2{,}667$ (und nicht, wie man meinen könnte, drei als [ungewogenes] arithmetisches Mittel aus 2 und 4; sie ist vielmehr das **harmonische** Mittel aus 2 und 4) und die Umschlagshäufigkeit ist 3, denn das Lager wurde im Mittel dreimal "umgeschlagen", zweimal bei den Waren C und D und viermal bei den Waren A und B.

2. Bei einer offenen Masse

Für die Definition des Durchschnittsbestands gilt weiterhin Gl. 12.3, die auch zur Berechnung von $\bar{B}$ herangezogen werden kann. Es ist aber nicht mehr von $\Sigma d_i = F_{0m}$ auszugehen. Vielmehr ist zur Berechnung der Verweilsumme Σd_i bezogen auf die $N = N_{0m}$ Einheiten, die im Beobachtungsintervall zu irgend einer Zeit dem Bestand angehörten, F_{0m} zu korrigieren um die Zeiten, welche die

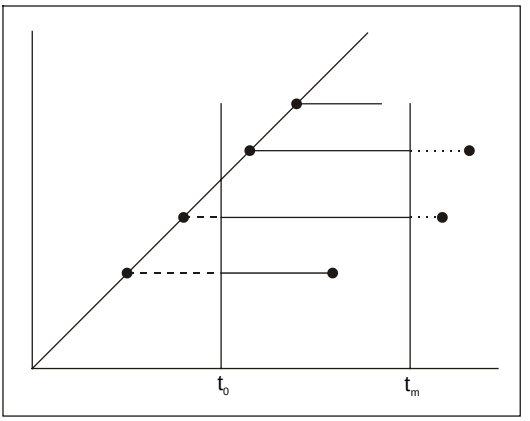
- B_0 Einheiten des Anfangsbestands vor t_0 bereits dem Bestand angehört hatten und die Zeiten welche die
- B_m Einheiten des Endbestands nach dem Ende des Beobachtungsintervalls, also nach t_m dem Bestand noch angehören werden.

In diesem Sinne spricht man von **Aufbauzeiten** und **Abbauzeiten** und es ist davon auszugehen, dass für die Zeiten vor t_0 und nach t_m keine Längsschnittsdaten vorliegen, so dass die Auf- und Abbauzeiten nur geschätzt werden können (zur Veranschaulichung dieser Zeiten vgl. Abb. 12.3). Die Verweilsumme ist unter diesen Voraussetzungen zu schätzen mit

$$(12.6) \quad G_{0m} = B_0\bar{d}_0 + F_{0m} + B_m\bar{d}_m$$

(B_0,B_m = Anfangs-, Endbestand; $\bar{d}_0$ ist die mittlere **Aufbau**zeit [d.h. die mittlere Verweildauer der B_0 Einheiten **vor** t_0] und $\bar{d}_m$ die mittlere **Abbau**zeit [mittlere restliche Verbleibdauer **nach** t_m im Bestand]).

Abb. 12.3: Auf- und Abbauzeiten



Aufbauzeit ---- Abbauzeit ·····

Man beachte, dass sich die Summanden in Gl. 12.6 aus sehr unterschiedlichen Verweilsummen zusammensetzen. Mit den Größen der oben behandelten kombinierten Fluß- und Bestandsgrößendarstellung erhält man die folgenden Verweilsummen:

Verweilsummen

Gruppe	vor t_0	im Intervall	nach t_m
A_{0m}	V_1	V_3	
R_{0m}	V_2	V_4	V_6
Z_{0m}		V_5	V_7

$N_{0m} = A_{0m} + R_{0m} + Z_{0m}$

Es ist offensichtlich, dass sich die Verweilsumme G_{0m} in Gl. 12.6 wie folgt zusammensetzt:

$$\begin{aligned} B_0\bar{d}_0 &= V_1 + V_2 \\ F_{0m} &= V_3 + V_4 + V_5 \\ B_m d_m &= V_6 + V_7 \,. \end{aligned} \qquad (12.7)$$

Zur Berechnung der durchschnittlichen Verweildauer ist G_{0m} an die Stelle von Σd_i in Gl. 12.4 einzusetzen, so dass für die durchschnittliche Verweildauer aller N_{0m} Einheiten, die jemals (zu irgendeiner Zeit) im Intervall $[t_0,t_m]$ zum Bestand gehört haben gilt

$$\bar{d}_N = \frac{G_{0m}}{N_{0m}} \,. \qquad (12.8)$$

Für die Umschlagshäufigkeit U gilt weiter Gl. 12.5.

c) Kennzahlen bei Querschnittsdaten

1. Übersicht

Es ist davon auszugehen, dass für die Praxis allein der Fall einer offenen Masse relevant ist.
Ein Problem, das bei dieser Art von Daten zusätzlich auftritt und bisher nicht zu behandeln war, ist die Schätzung der Zeitmengenfläche, weil nicht individuelle Zu- und Abgänge beobachtet werden, d.h. diese nicht als Einzelfälle mit genauem Ein- und Austrittszeitpunkt festgehalten werden, sondern nur summarisch. Dies gilt selbst dann, wenn die Zu- und Abgänge im Beobachtungsintervall stattfinden. Übersicht 12.2 zeigt, wie sich die Berechnung von Kennzahlen in Abhängigkeit der Daten verkompliziert:

Übersicht 12.2: Schätzprobleme bei den Kennzahlen

Daten/Masse	Zeitmengenfläche F_{0m}	Verweilsumme Σd_i
1. Längsschnitt a) geschlossen b) offen	kein Problem, da Bestandsfunktion B(t) für jedes t bekannt	a) identisch mit F_{0m} b) als G_{0m} aus F_{0m} zu schätzen mit Gl. 12.6
2. Querschnitt (nur offene Masse)	zu schätzen mit Gl. 12.10, da B(t) nur zu bestimmten Zeitpunkten bekannt ist	Übergang von F_{0m} zu G_{0m} wie im Fall 1b, aber Annahmen über Auf- u. Abbauzeiten nötig

Der Übergang zu den aus F_{0m} bzw. Σd_i abgeleiteten Maßzahlen $\bar{B}$ und $\bar{d}$, sowie U ergibt sich analog zu Gl. 12.3 bis 12.5.

2. Zeitmengenfläche und Durchschnittsbestand:

Wenn die Bestandsänderungen ausschließlich genau zu den Beobachtungszeitpunkten t_j (j=1,2,...,m) stattfinden, dann ist die Zeitmengenfläche gegeben durch:

$$F_{0m} = \Sigma B_{j-1}(t_j - t_{j-1}) \quad (12.9)$$

denn der Bestand ist dann von t_{j-1}, als die letzte Bestandsänderung stattfand, bis t_j konstant B_{j-1}. Sind die Beobachtungszeitpunkte t_j (mit j = 1,2,... m) äquidistant, so dass $t_j - t_{j-1}=1$ (für alle j) und $t_m - t_0 = m$, so folgt aus Gl. 12.9

(12.9a) $F = F_{0m} = \Sigma B_{j-1} = B_0 + B_1 +...+B_{m-1}.$

Wegen $\Sigma(t_j - t_{j-1}) = m$ (bei j = 1,2,...,m und $t_m - t_0 = m$) ist erkennbar, dass $\bar{B} = F_{0m}/m$ ein mit Zeitintervallen gewogenes Mittel von Beständen ist.
Gl. 12.9a geht davon aus, dass Bestandsänderungen (Zu- und Abgänge) jeweils am **Ende** einer Einheitsperiode stattfinden. Entsprechend wäre bei Bestandsänderungen jeweils am **Anfang** einer Einheits- (Beobachtungs-) Periode die Zeitmengenfläche mit

(12.9b) $F^* = \Sigma B_j = B_1 + B_2+...+ B_{m-1} + B_m$

zu schätzen. Es liegt nahe, einen Mittelwert von F und F^* zu berechnen, so dass F_{0m} zu schätzen ist mit

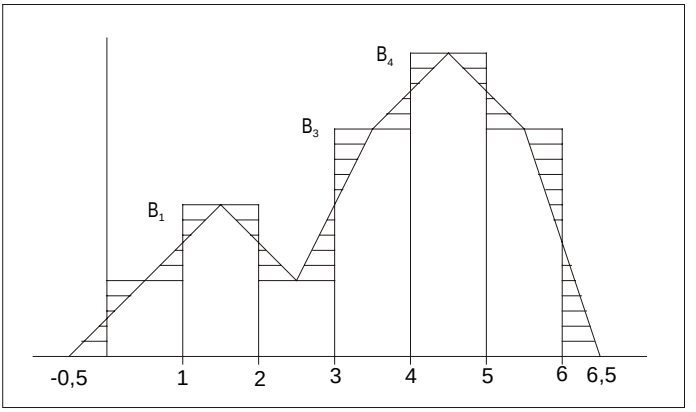
(12.10) $F_{0m} = ½B_0 + B_1 +...+ B_{m-1} + ½B_m ,$

woraus $\bar{B}$ mit Gl. 12.3 zu errechnen ist.

Bemerkungen zu Gleichung 12.10:

1. Gl. 12.10 ergibt sich bei linearer Approximation der Bestandsfunktion (die dann ein Polygonzug ist); sie gilt bei offener **und** geschlossener Masse (vgl. Abb. 12.4). Sie beruht auf der Annahme, dass im Intervall von t_{j-1} bis t_j jeweils im Durchschnitt ½ $(B_{j-1} + B_j)$ Einheiten im Bestand sind.

Abb. 12.4: Bestandsfunktion als Polygonzug



2. Gl. 12.10 wird auch als **chronologisches Mittel** (einer Folge von Bestandszahlen) bezeichnet. Die Formel ist auch bekannt aus der Behandlung zentrierter gleitender Mittelwerte.

3. Eine exakte Berechnung der Zeitmengenfläche durch Gl. 12.9a bis 12.10 liegt nur dann vor, wenn Zu- und Abgänge von jeweils einer oder mehrerer Einheiten stets alle genau am Anfang, am Ende oder in der Mitte der Beobachtungsintervalle stattfinden. In allen anderen Fällen kann Gl. 12.10 nur eine Näherung sein. Dies gilt insbesondere dann, wenn für jedes Intervall nur die **kumulierten** (d.h. die Summe aller im Intervall zu beliebigen Zeitpunkten insgesamt erfolgten) Zu- und Abgänge bekannt sind.

4. Man beachte, dass Gl. 12.10 nicht identisch ist mit der häufig als Schätzung des Durchschnittsbestands verwendeten Mittelung von Anfangs und Endbestand ($½B_0 + ½B_m$), die als Näherung noch wesentlich gröber ist und nur vertretbar ist, wenn die Bestandsfunktion im Intervall von t_0 bis t_m linear verläuft.

3. Durchschnittliche Verweildauer und Umschlagshäufigkeit:

Der Durchschnittsbestand ergibt sich einfach mit Gl. 12.10 in Verbindung mit Gl. 12.3. Die Schätzung der durchschnittlichen Verweildauer ist jedoch schwieriger, weil die Verweilsumme Σd_i aus F_{0m} zu schätzen ist. Es ist dabei analog zu Gl. 12.6 vorzugehen, d.h. G_{0m} ist die geeignete Schätzung von Σd_i. Anders als bei der Betrachtung von Gl. 12.6 ist jedoch jetzt davon auszugehen, dass keine Längsschnittsdaten vorliegen und über die durchschnittliche Aufbau- und Abbauzeit nichts bekannt ist.

Es sind nun Annahmen über die mittlere Aufbauzeit ($\bar{d}_0$) und Abbauzeit ($\bar{d}_m$) nötig. Üblich ist die Annahme

$$\bar{d}_0 = \delta\bar{d} \text{ und } \bar{d}_m = (1-\delta)\bar{d} \quad \text{mit } 0 < \delta < 1 ,$$

d.h. die B_0 Einheiten des Anfangsbestands haben im Durchschnitt bereits ein δ-tel ihrer gesamten Verweildauer vor t_0 zurückgelegt und die B_m Einheiten des Endbestands werden nach t_m noch ein (1-δ)-tel ihrer gesamten Verweildauer im Bestand bleiben.
Eingesetzt in (12.6) liefert das

$$G_{0m} = \bar{d}N_{0m} = B_0\delta\bar{d} + F_{0m} + B_m(1-\delta)\bar{d} \text{ und schließlich}$$

$$(12.11)\quad \bar{d} = \bar{d}_N = \frac{F_{0m}}{dZ_{0m} + (1-d)A_{0m}}$$

Welchen Wert kann man nun für δ in Gl. 12.11 ansetzen?

1. Wie man leicht sieht, bedeutet δ = 0

   $$(12.11a)\quad \bar{d}_1 = m\bar{B}/A_{0m}$$

   und bei δ = 1 gilt

   $$(12.11b)\quad \bar{d}_2 = m\bar{B}/Z_{0m} \quad .$$

   Der "Grenzwert" δ = 0 bedeutet, dass die Aufbauzeit der Zugänge vor t_0 genau $\bar{d}$ und folglich die Verweildauer im Beobachtungsintervall 0 beträgt. Es sind also Z_{0m} Einheiten zugegangen, und zwar ausschließlich vor t_0. Entsprechend widersprüchlich ist die Annahme δ = 1.

2. Sehr verbreitet ist nun die Annahme δ = ½, was zu der sehr bekannten Formel

   $$(12.12)\quad \bar{d} = \frac{2m\bar{B}}{Z_{0m}+A_{0m}}$$

   führt. Sie gilt auch bei einer geschlossenen Masse, da dort stets $N_{0m} = Z_{0m} = A_{0m} = (Z_{0m}+A_{0m})/2$ ist. Gilt $Z_{0m} = A_{0m}$, was bei einer offenen Masse meist dann anzunehmen ist, wenn das Intervall von t_0 bis t_m genügend lang ist und sich der Bestand nicht wesentlich verändert, so erhält man unabhängig von der Wahl von δ aus Gl. 12.11 stets Gl. 12.12.

   Als Faustregel gilt, dass die Länge m des Intervalls etwa (mindestens) die vierfache durchschnittliche Verweildauer $\bar{d}$ sein sollte, um $\bar{d}$ mit Gl.12.12 zuverlässig schätzen zu können.

   Zur Anwendung der Gl. 12.12 bei der Schätzung der durchschnittlichen Verweildauer vgl. auch Beispiel 12.4.

3. Man kann davon ausgehen, dass sich ein größerer Bestand langsamer auf- und abbaut als ein kleinerer Bestand und so z.B. annehmen, dass die (auf die Anzahl der Einheiten bezogene) durchschnittliche Aufbauzeit des Anfangsbestands sich zu der durchschnittlichen Abbauzeit des Endbestands wie Anfangsbestand zu Endbestand verhält, also $\overline{d}_0 / \overline{d}_m = B_0 / B_m$. Das führt dann zu der folgenden häufig verwendeten Annahme über δ:

$$(12.13) \qquad \delta = \frac{B_0}{B_0+B_m}$$

Gleichbleibender Bestand $B_0 = B_m$ bedeutet danach $\delta = ½$ und zunehmender Bestand ($B_0 < B_m$) kürzere Aufbauzeiten, also $\delta < ½$. Bei gegebener Fläche F_{0m} ist der Nenner $\delta(Z_{0m}-A_{0m})+A_{0m}$ in Gl. 12.11 bei zunehmendem Bestand wegen $Z_{0m} > A_{0m}$ (sonst könnte der Bestand ja auch nicht zunehmen) größer als bei abneh-mendem Bestand $Z_{0m} < A_{0m}$, so dass in diesem Fall auch die Verweildauer sinkt.

Beispiel 12.4:

Für die Bundesrepublik galten für die Zeit von 1982 bis 1987 die folgenden Zahlen der Arbeitslosenstatistik:

Bestand, Zugang und Abgang an Arbeitslosen

Jahr	Bestand	Zugang	Abgang
1982	1.833.244	3.706.655	3.187.165
1983	2.258.235	3.704.185	3.578.551
1984	2.265.559	3.672.791	3.696.594
1985	2.304.014	3.750.240	3.728.294
1986	2.228.004	3.637.266	3.766.214
1987	2.228.788	3.726.460	3.636.411

Quelle: J. Kühl, 15 Jahre Massenarbeitslosigkeit, Aspekte einer Halbzeitbilanz, in: Aus Politik und Zeitgeschichte, Beilage zur Wochenzeitschrift "Das Parlament" vom 16.9.1988.

Berechnen Sie den Durchschnittsbestand und schätzen Sie die durchschnittliche Verweildauer (in der Arbeitslosigkeit) sowie die Umschlagshäufigkeit des Arbeitslosenbestandes!

(Anmerkung zu den Daten: Es wird häufig vergessen, dass ein hoher Bestand an Arbeitslosen von über 2,2 Millionen [ab 1983] auch einhergeht mit einer erheblichen Bewegung auf dem Arbeitsmarkt. In diesen Jahren wurden nämlich jeweils über 3½ Millionen Zu- und Abgänge zur, bzw. aus der Arbeitslosigkeit gezählt. Ist die Anzahl der Bewegungen groß gegenüber dem Bestand [so bei Arbeitslosigkeit oder Krankenhausbelegung], so ist die durchschnittliche Verweildauer kurz. Die

Umkehrung gilt bei der Bevölkerung: die Anzahl der jährlich beobachteten Bewegungen (Geburten, Todesfälle) ist gering, verglichen mit dem Bestand, also ist die Verweildauer (Lebensdauer) relativ groß.

Lösung 12.4:

Die Anwendung von Gl. 12.10 (für F_{0m}) und Schätzung von $\bar{B}$ nach Gl. 12.3/12.12 führt zu: $F_{0m} = 11.086.828$, $\bar{B} = 2.217.365{,}6$ (man beachte, dass $m = 5$ ist, $B_0 = B_{1982}$ und $B_m = B_{1987,}$ so dass $½(B_0+B_m) = 2.031.016 < \bar{B}$, wobei $\bar{B}$ aufgrund von mehr als nur zwei Bestandszahlen geschätzt wurde); $Z_{0m} = 22197597$ und $A_{0m} = 21593229$. Man erhält daraus mit Gl. 12.12 für die durchschnittliche Verweildauer $\bar{d} = 0{,}50635$. Man erkennt übrigens auch, dass die Wahl von δ keine große Rolle spielt, denn:
für $\delta = 0$ erhält man $\bar{d}_0 = F_{0m}/A_{0m} = 0{,}51344$
und für $\delta = 1$ erhält man $\bar{d}_u = F_{0m}/ Z_{0m} = 0{,}49946$.

Weitere Interpretationen von Gl. 12.11 und 12.12

- Es ist offensichtlich, dass die Gl. 12.11 und 12.12 nicht zu verwenden sind, wenn es keine Zu- und Abgänge gibt, wenn also $Z_{0m} = A_{0m} = 0$.
- Von der gerade durchgeführten Betrachtung, ob $Z_{0m} > A_{0m}$ oder umgekehrt $Z_{0m} > A_{0m}$, ist zu unterscheiden, wie sich bei gegebenem Z_{0m} eine Zunahme von A_{0m} (bzw. bei gegebenem A_{0m} eine Zunahme von Z_{0m}) auf die durchschnittliche Ver-weildauer auswirkt. In beiden Fällen ist wegen des Nenners der Gl. 12.12 mit einer Abnahme der Verweildauer zu rechnen.
- Man sollte sich klarmachen, dass Gl. 12.12 in der Regel nur eine grobe Schätzung der mittleren Verweildauer darstellen kann. Das folgende Beispiel mit drei bzw. vier Personen (A,...,D) möge das zeigen.

Beispiel/Lösung 12.5:

Gegeben seien die folgenden Angaben über die Verweildauer von vier Einheiten:

Verweildauer

Person	vor t_0	im Intervall	nach t_m
A	2	2	0
B	0	3	3
C	0	5	0
D	3	3	0

Rechnet man nur mit den Personen A,B und C, so erhält man die durchschnittliche Verweildauer $\bar{d} = (4+6+5)/3 = 5$ und nach Gl. 12.11 wegen $F_{0m}= 2+3+5=10$ sowie $Z_{0m}= A_{0m} = 2$ (Zugänge: B,C; Abgänge: A,C)und $\delta = ½$ ebenfalls $\bar{d} = 5$. Die Berechnung für

die vier Personen A,..., D führt dagegen zu einer tatsächlichen Verweilsumme von Σd_i =21 also $\bar{d}$ = 21/4 = 5,25, aber nach Gl. 12.11 wegen F_{0m} = 2+3+5+3=13 sowie Z_{0m} = 2 und A_{0m}= 3 sowie δ = ½ zu $\bar{d}$= 26/5 = 5,2.

4. Weitere Kennzahlen bei Querschnittsdaten: Verhältniszahlen

Es ist davon auszugehen, dass Zugänge (Abgänge) nicht individuell zum Zeitpunkt ihres Zugangs (Abgangs), sondern nur summarisch erfaßt werden, so dass Z_{0m} (A_{0m}) die Gesamtzahlen der Zu-, bzw. Abgänge sind, die irgendwann im Intervall $[t_0,t_m]$ stattfanden. Entsprechend sei dann $\bar{B}$ der Durchschnittsbestand im Intervall bzw. der Bestand in der **Mitte** dieses Intervalls.

Def. 12.6: Zugangs- und Abgangsraten

Die mit

$$(12.14)\quad z_{0m} = \frac{Z_{0m}}{\bar{B}} \quad \text{und} \quad a_{0m} = \frac{A_{0m}}{\bar{B}}$$

definierten Größen werden auch Raten genannt. Die (rohe)[1] Geburtenrate ist vom Typ einer Zugangsrate (z_{0m}) und die (rohe) Todesrate eine Abgangsrate (a_{0m}). Raten sind Beziehungszahlen und somit spezielle Verhältniszahlen (vgl. Kap. 9).

Der Ausdruck Rate ist in diesem Fall durchaus sinnvoll. Würde man im Nenner nicht den Durchschnitts- sondern den Anfangsbestand ansetzen, so wäre offensichtlich die Differenz der so definierten Raten $z^* = Z_{0m}/B_0$ und $a^* = A_{0m}/B_0$ die Wachstumsrate des Bestands. Hier zeigt sich übrigens auch die Problematik der Längsschnittsinterpretation von Querschnittsdaten: Wenn beispielsweise die Scheidungsquote (Scheidungen eines Jahres/Bestand an Ehen in der Mitte des Jahres) σ = 0,007 ist, dann heißt das nicht, dass nur 0,7% aller Ehen vor dem Scheidungsrichter enden, denn in σ gehen ja nur die Scheidungen eines Jahres ein und die durchschnittliche Verweildauer in der Ehe ist jedoch länger als ein Jahr. Damit ist die Lösung des folgenden Beispiels klar.

Beispiel 12.6:

(querschnitts- und längsschnittsanalytische Scheidungsquote)

---

[1] Der Zusatz "roh" (oder "allgemein") ist üblich, weil es auch alters- oder geschlechtsspezifische Raten gibt. So ist z.B. die altersspezifische Sterberate (Todesrate) der 60-jährigen definiert als die Anzahl der im Alter von 60 Gestorbenen, dividiert durch die Wohnbevölkerung (also die Lebenden) im Alter von 60 Jahren. Die altersspezifischen Sterberaten sind die Grundlage für die Berechnung der Sterbewahrscheinlichkeiten, die ihrerseits Basis für die Konstruktion der Sterbetafel sind.

In den letzten Jahren betrug die sog. "Scheidungsquote" in der Regel meist etwa um 20 je 10.000 Einwohner und 70 je 10.000 bestehende Ehen.
Bedeutet dies, dass nur jeder 500te Bundesbürger geschieden wird und dass nur etwa jede 143te Ehe (also noch nicht einmal 1%, genau genommen nur 0,7% [70/10000]) vor dem Scheidungsrichter endet?

Lösung 12.6:

Die Lösung ergibt sich aus dem Text vor dem Beispiel.

3. Stationäre Bevölkerung und Tafelrechnung

a) Stationäre Bevölkerung

Das Modell der stationären Bevölkerung ist eine stark vereinfachte Darstellung einer Entwicklung, bei der abstrahiert wird von Wachstum und Strukturveränderung. Es erlaubt, eine offene Masse wie eine geschlossene Masse zu behandeln und von der Querschnitts- auf die Längsschnittsbetrachtung zu schließen. Es liegt der Berechnung einer Sterbetafel zugrunde, weshalb man auch von "Sterbetafelbevölkerung" spricht.
In einer stationären Bevölkerung wird jeder Geburtsjahrgang (jede "Kohorte") durch eine

a) gleich große und
b) gleich strukturierte

Kohorte ersetzt. Wegen a) wird im Modell abstrahiert von einer Dynamik (Wachstum) und wegen b) von einer Strukturveränderung (z.B. in Form einer Veränderung der Sterbewahrscheinlichkeiten und damit der Abgangsordnung). Es sind deshalb zunächst die Begriffe Kohorte und Abgangsordnung zu definieren.

Def. 12.7: Kohorte, Abgangsordnung, stationäre Bevölkerung

a) Eine Zugangskohorte oder einfach Kohorte ist die Gesamtheit der gleichzeitig (zum gleichen Zeitpunkt t_j, bzw. im gleichen Intervall geringer Länge $[t_{j-1}, t_j]$) zugehenden Einheiten. Ihr Umfang, d.h. die Zahl

   der zugehenden Einheiten ist l_0.

b) Die Abgangsordnung l_x (wobei $x = 0,1,...,w$ das Alter, d.h. die Anzahl der vollendeten Jahre ist) ist die Anzahl der Überlebenden des Alters x. Es ist der Restbestand einer Geburtskohorte des Umfangs l_0 nach Vollendung von x Jahren in Abhängigkeit von x.

c) Bei einer stationären Bevölkerung (Sterbetafelbevölkerung) wird jede Kohorte (jeder Geburtsjahrgang) in jedem aufeinanderfolgenden Intervall (in allen folgenden Jahren) durch eine gleich große Kohorte (so dass $Z_{j-1,j} = l_0$ für alle j) mit gleicher Abgangsordnung (d.h. gleicher "Struktur"; l_x ist nicht von j sondern nur von x abhängig) ersetzt.

Folgerungen aus Def. 12.7:

1) Die Stärke der Geburtsjahrgänge und die Abgangsordnung l_x ist nicht abhängig vom Geburtsjahr (von der absoluten-, objektiven Zeit). Die Abgangsordnung, d.h. die Folge $l_0,l_1,...l_w$, und alle im folgenden eingeführten Tafelfunktionen q_x, T_x und e_x hängen bei einer stationären Bevölkerung allein vom Alter x ab (von der relativen Zeit). Das ist der Grund, weshalb hier (und nur hier) vom Querschnitt auf den Längsschnitt geschlossen werden kann.

2) Die Abgangsordnung (auch "Absterbeordnung", "Überlebensordung" oder "Absterbefunktion") ist eine monoton (nicht streng-monoton) abnehmende Funktion des Alters: ist das Alter z größer als das Alter x, so ist $l_z \leq l_x$.

Beispiel 12.7:

Trotz intensivster ärztlicher Bemühungen haben sich die fünf ersten Exemplare einer neu gezüchteten Pferderasse als nicht sonderlich überlebensfähig herausgestellt. Das Pferd Egon konnte noch nicht einmal seinen ersten Geburtstag feiern. Das Alter (in vollendeten Jahren), das die fünf Pferde erreichten betrug leider nur:

Pferd	Alter
Egon (E)	0
Doris (D)	1
Boris (B),Clara	2
Augustus (A)	3

Man bestimme aus diesen Daten die Abgangsordnung l_x und die Altersverteilung der Kohorte (des Geburtsjahrgangs).
Das Beispiel soll im fol-

genden erweitert werden.

Lösung 12.7:

Abgangsordnung

Alter x	Pferde	l_x
0	A,B,C,D,E	5
1	A,B,C,D	4
2	A,B,C	3
3	A	1
4		0

Verteilung des Sterbealters

Von den $l_0 = 5$ Pferden erreichen d_x das Alter (Anzahl der vollendeten Jahre) von x Jahren:

x	d_x	(Pferd)
0	1	Egon
1	1	Doris
2	2	Boris,
3	1	Augustus

Fortführung des Beispiels 12.7:
Man stelle sich nun vor, Diplom-Kaufmann K aus E betreibe seine Pferdezucht weiter und jedes Jahr werden 5 Pferde A,...,E geboren, die in der gleichen Weise leider nur sehr kurz leben und spätestens (Pferd Augustus) kurz vor Erreichen des 4-ten Geburtstags sterben. Man erhält nun beginnend mit dem Jahr 1991 die folgende "Pferdebevölkerung" (die Geburtsjahrgänge [Kohorten, engl. vintages] werden mit 1,2,... bezeichnet).

Jahr	Pferde im Alter von ...Jahren			
		1	2	3
91	A_1,B_1,C_1,D_1,E_1			
92	A_2,B_2,C_2,D_2,E_2	A_1,B_1,C_1,D_1		
93	A_3,B_3,C_3,D_3,E_3	A_2,B_2,C_2,D_2	A_1,B_1,C_1	
94	A_4,B_4,C_4,D_4,E_4	A_3,B_3,C_3,D_3	A_2,B_2,C_2	A_1

Wie man sieht ist ab 1994 die Altersverteilung der Pferde und der Umfang der Pferdebevölkerung konstant. Es gilt für die Altersverteilung der $T^*,_0 = \Sigma l_x = 13$ Pferde :

Altersverteilung des Bestands

Alter x	0	1	2	3	Summe
Pferde	$l_0=5$	$l_1=4$	$l_2=3$	$l_3=1$	$T_0^*=\Sigma\, l_x=13$

Wie man sieht gilt: Die Anzahl der Pferde des Alters x ist stets genau l_x. Bei einer stationären Bevölkerung hängt also der Bestand mit der Absterbeordnung eng zusammen (vgl. Satz 12.3). Es ist nun auch von Interesse, die Altersverteilung des Bestands mit derjenigen der Kohorte (der jedes Jahr stattfindenden Geburten und Todesfälle) zu vergleichen:

Altersverteilung der Kohorte

Alter x	0	1	2	3	Summe
Pferde	d_0=1	d_1=1	d_2=2	d_3=1	l_0= Σd_x=5

Es ist auch offensichtlich, dass das Durchschnittsalter der Lebenden (des Bestands) kleiner ist als das durchschnittliche Sterbealter (Durchschnittsalter der Gestorbenen). Schließlich besteht zwischen den Größen l_x und d_x folgender Zusammenhang:

Absterbeordnung l_x und Anzahl der Gestorbenen d_x

Alter x	Anzahl der Überlebenden l_x	im Beispiel 12.7
0	$l_0 = d_0 + d_1 + d_2 + \ldots + d_w$	$d_0 + d_1 + d_2 + d_3 = 5$
1	$l_1 = d_1 + d_2 + \ldots + d_w$	$d_1 + d_2 + d_3 = 4$
2	$l_2 = d_2 + \ldots + d_w$	$d_2 + d_3 = 3$
...	... ...	...
w-2	$l_{w-2} = d_{w-1} + d_w$	$d_2 + d_3 = 3$
w-1	$l_{w-1} = d_w$	$d_3 = 1$

Es gilt also:

(12.15) $l_x = \Sigma\, d_y$ (mit $y \le x$).

Satz 12.3:

1. Die stationäre Bevölkerung hat einen konstanten Umfang von $P_t = T_0^*$ (P = population) zu allen Zeitpunkten t, wobei

   (12.16) $T_0^* = \Sigma\, l_x \quad (x = 0,1,...,w)$.

   T_0^* ist die Summe der von der Kohorte von l_0 Einheiten (von l_0 Neugeborenen) insgesamt zu durchlebenden Jahre (eine Verweilsumme mit der Maßeinheit "Personenjahre" bzw. im Beispiel "Pferdejahre").

2. Ihre Altersverteilung ist gegeben durch die Abgangsordnung: dem Alter x ist die absolute Häufigkeit l_x zugeordnet.

3. Das durchschnittliche Sterbealter $\bar{x}_D = \Sigma x d_x / \Sigma d_x$ der Geburtskohorte (der gleichzeitig Geborenen) des Umfangs l_0 beträgt $\bar{x}_D = e_0^* - 1$ mit

   (12.17) $$e_0^* = \frac{T_0^*}{l_0}$$

(e_0^* ist die durchschnittliche Verweildauer, die Lebenserwartung eines [einer] Neugeborenen). Folglich gilt bei einer stationären Bevölkerung, dass der Bestand (Umfang) zu jedem Zeitpunkt konstant ist und zwar $T_0^* = e_0^* l_0$, also:

> Bei einer stationären Bevölkerung ist der (konstante) Bevölkerungsstand gleich T_0^* und damit gleich dem Produkt der Zahl der Zugänge und der mittleren Verweildauer.

4. Das mittlere Alter der Lebenden (Mittelwert der Altersverteilung gem. Nr. 2) ist mit $\Sigma x l_x / \Sigma l_x$ in der Regel verschieden von der Lebenserwartung $e^*{}_0$ (gem. Gl. 12.17). Im Beispiel 12.7 erhält man $\Sigma x l_x / \Sigma l_x = 1$ und $\Sigma x d_x / \Sigma d_x = 8/5 = 1{,}6$.

5. Zwischen der Altersverteilung des Bestands und derjenigen der Kohorte (der konstanten Zu- und Abgänge), also zwischen l_x und d_x gilt der Zusammenhang von Gl. 12.15.

Man beachte: bei der Definition des durchschnittlichen Sterbealters $\overline{x}_D$ (Nr. 3) und des durchschnittlichen Alters der Lebenden $\overline{x}_L$ (Nr.4) ist angenommen, dass jede der d_X Einheiten, die das Alter x (Anzahl der vollendeten Jahre) erreicht hat, praktisch gleich nach ihrem Geburtstag stirbt und bei e_0^* , dass sie praktisch kurz vor ihrem x+1ten Geburtstag stirbt. Später (vgl. Satz 12.5) wird die Definition des Sterbealters dahingehend modifiziert, dass die Sterbezeitpunkte als gleichmäßig über ein Altersjahr (Intervall [x, x+1]) angenommen werden.

Beispiel 12.8:

Es sei angenommen, dass jedes Jahr $l_0 = 6$ Zugänge (Einheiten A,..., F) zu verzeichnen sind, die wie folgt abgehen: $d_0 = 3$ Einheiten (D,E,F) im Alter $x = 0$ (also vor dem ersten Geburtstag), $d_1 = 2$ Einheiten (B, C) im Alter $x = 1$ und $d_2 = 1$ Einheit (A) im Alter $x = 2$. Man bestimme die Altersverteilung des (beginnend mit der dritten Kohorte) konstanten Bestands und der Abgänge.

Lösung 12.8:

Altersverteilung (j = Subskript der Kohorte):

des Bestands am Beginn des dritten Jahres:

Alter x	0	1	2
Lebende	$A_3,...,F_3$	A_2,B_2,C_2	A_1
Anzahl	$l_0=6$	$l_1=3$	$l_2=1$
Anteil	60%	30%	10%

der Abgänge am Ende des dritten Jahres:

Alter x	0	1	2
Gestorbene	D_3,E_3,F_3	B_2,C_2	A_1
Anzahl	$d_0=3$	$d_1=2$	$d_2=1$
Anteil	50%	33,3%	16,7%

Man verifiziert leicht, dass das Durchschnittsalter der Gestorbenen (durchschnittliches Sterbealter) $\overline{x}_D$ - was die Regel ist - größer ist als das Durchschnittsalter der Lebenden (des Bestands) $\overline{x}_L$ (es gilt $\overline{x}_D = 2/3$ und $\overline{x}_L = 0,5$) und dass gilt :

(12.15) $l_x = \Sigma d_y$ (mit $y \le x$):

$$\left.\begin{array}{rrrrrrrrr} l_0 = & d_0 & + & d_1 & + & d_2 & = & 3 + 2 + 1 & = 6 \\ l_1 = & & & d_1 & + & d_2 & = & 2 + 1 & = 3 \\ l_2 = & & & & & d_2 & = & 1 & = 1 \end{array}\right\} \text{ Bestand der Bevölkerung } \Sigma l_x = T_0^* = 10$$

Beispiel 12.9:

Es sei angenommen, dass jedes Jahr $l_0 = 4$ Zugänge (Einheiten A bis D) zu verzeichnen sind, die alle im Alter von x = 2 (nach Vollendung von 2 Jahren) abgehen (rechteckige Abgangsordnung). Man bestimme die Altersverteilung des konstanten Bestands und der Abgänge.

Lösung 12.9:

Altersverteilung (j = Subskript der Kohorte) am Beginn/Ende des dritten Jahres:

Beginn

Alter x	0	1	2
Lebende	$A_3 - D_3$	$A_2 - D_2$	$A_1 - D_2$
Anzahl	$l_0 = 4$	$l_1 = 4$	$l_2 = 4$

Ende

Alter x	2
Gestorbene	$A_1 - D_1$
Anzahl	$d_2 = 4$

Es gilt $d_0=d_1=0$ und $d_2=4$ so dass $l_0=l_1=l_2=4$.

Auch jetzt gilt, dass das Durchschnittsalter des Bestands $\overline{x}_L = 1$ kleiner ist als das durchschnittliche Sterbealter $\overline{x}_D = 2$ bzw. die Lebenserwartung eines Nulljährigen $e_0^* = 3$ (vgl. Satz 12.5).

b) Sterbetafel

Die folgenden Ausführungen setzten den ansonsten in der Deskriptiven Statistik noch nicht eingeführten Begriff der Wahrscheinlichkeit voraus.
Die einjährige Sterbewahrscheinlichkeit der x-jährigen Männer bzw. Frauen wird bestimmt aus der Anzahl P^*_{xj} der Personen (Männer bzw. Frauen) der Bevölkerung, die zum Zeitpunkt t_j zur Wohnbevölkerung gehören (Datenbasis für die Sterbetafel; in der Praxis verwendet man nicht ein, sondern meist drei benachbarte Jahre um die Wirkung von Zufallseinflüssen auf die Sterbewahrscheinlichkeiten zu verringern) und der Anzahl D_{xj} der Abgänge (Todesfälle) in der Altersgruppe x in einem Jahresintervall um t_j. Die alters-spezifischen Sterbeziffern lauten somit: $m_x = D_{xj} / P^*_{xj}$. Aus ihnen werden nach gewissen Korrekturen und Glättungen (z.B. Anpassung von Regressionsfunktionen, gleitende Durchschnitte, Spline-Funktionen usw.) die einjährigen Sterbewahrscheinlichkeiten q_x be-stimmt. Unter der Annahme, dass sich die Todesfälle im Altersintervall (x, x+1) gleichmäßig verteilen, so dass P^*_{xj} durch $P_{xj} = P^*_{xj} + ½D_{xj}$ zu ersetzen ist, erhält man $q_x = D_{xj} / P_{xj} = m_x / (1 + ½m_x)$.

Def. 12.8: Tafelfunktionen q, p, l, L

a) Die einjährige Sterbewahrscheinlichkeit q_x der x-jährigen ist die (bedingte) Wahrscheinlichkeit dafür, dass eine Person, die das Alter von x erreicht hat, das Alter von x+1 nicht mehr erreichen wird (mit x = 0,1,...,w für das Alter in vollendeten Jahren).

b) Die einjährige Überlebenswahrscheinlichkeit p_x ist demzufolge $p_x = 1 - q_x$, denn eine x-jährige Person kann in ihrem nächsten Lebensjahr entweder sterben oder dieses überleben. Auch die einjährige Überlebenswahrscheinlichkeit p_x ist eine bedingte Wahrscheinlichkeit, das Alter x zu überleben (x+1 zu erleben), **wenn** (bedingt dadurch, dass) man bis x gelebt hat.

c) Sämtliche Sterbetafelfunktionen sind allein Funktionen des Alters x und sie sind mit der Folge der Sterbewahrscheinlichkeiten q_x und dem willkürlich gewählten Anfangsbestand (Geburten) l_0 eindeutig gegeben:

- die Absterbeordnung l_x bezeichnet die Anzahl der Personen, die mindestens das Alter x erreichen. Sie ist ausgehend von einem fiktiven Anfangsbestand von $l_0 = 100.000$ Personen rekursiv zu berechnen mit

$$(12.18)\quad l_{x+1} = l_x p_x = l_x(1 - q_x).$$

- Entsprechend ist die Anzahl d_x der im Altersintervall (x, x+1) gestorbenen Personen gegeben mit

$$(12.19)\quad d_x = l_x q_x = l_x - l_{x+1} \geq 0$$

da die Abgangsordnung l_x monoton fallend ist ($l_{x+i} \leq l_x$).

d) Mit L_x wird die Anzahl der von allen überlebenden x-jährigen Personen bis zum Alter x+1 durchlebten Jahre (die Anzahl der im Intervall (x,x+1) verlebten Personenjahre) bezeichnet.

(12.20) $L_x = ½(l_x + l_{x+1})$.

Zur Begründung für diese Modifikation der Abgangsordnung vgl. Bem. Nr. 2 zu dieser Definition.

e) Für einige Zwecke ist es auch sinnvoll, mehrjährige unbedingte Überlebenswahrscheinlichkeiten vom Alter 0 bis zu einem bestimmten Alter x als Produkt einjähriger Überlebenswahrscheinlichkeiten zu definieren mit

(12.21) $p^*_{0x} = p_0 p_1 p_2 \cdots p_{x-1}$,

so dass gilt:

(12.22) $l_x = l_0 p^*_{0x} = l_0 p_0 p_1 \cdots p_{x-1}$,

d.h. die Folge der einjährigen Sterbe- und damit auch Überlebenswahrscheinlichkeiten bestimmt zusammen mit der willkürlichen Anzahl l_0 der Geburten (der fiktiven Kohorte) eindeutig die gesamte Absterbeordnung und damit auch die gesamte Sterbetafel.

Bemerkungen zu Def. 12.8:

1. Alle Sterbetafelfunktionen q_x, p_x, l_x, d_x und L_x sowie die noch zu definierenden Funktionen T_x und e_x (Def. 12.8) sind bei gegebenem (fiktiven) Anfangsbestand l_0 der angenommenen Geburtskohorte Funktionen des Alters x und allein abhängig von q_x.

2. Der Berechnung von L_x als ungewogenes Mittel von l_x und l_{x+1} liegt der Gedanke zugrunde, dass die l_{x+1} Personen, die auch ihren x+1-ten Geburtstag erleben, zu L_x ein volles Lebensjahr beitragen, dagegen diejenigen, die zwar das Alter (Anzahl der vollendeten Jahre) x, nicht aber das Alter x+1 Jahre erreichen, im Mittel nur ein halbes Jahr beitragen, so dass gilt:
$L_x = 1 \cdot l_{x+1} + ½ \cdot d_x = l_{x+1} + ½(l_x - l_{x+1}) = ½(l_x + l_{x+1})$.

3. Berechnet man die mit der Sterbetafel gegebene Lebenserwartung einer x-jährigen Person (vgl. Def. 12.8) auf der Basis der Folge der l_x-Werte (in Def. 12.8 dann e^*_x statt e_x genannt) statt der L_x - Werte, so ist die Lebenserwartung jeweils (für alle x) um ein halbes Jahr größer: $e^*_x = e_x + ½$.

Beispiel 12.10:

Gegeben sei die monoton steigende Folge von Sterbewahrscheinlichkeiten q_x und Überlebenswahrscheinlichkeiten $p_x = 1 - q_x$:

x	0	1	2	3
q_x	1/5	1/4	2/3	1
p_x	4/5	3/4	1/3	0

Bestimmen Sie ausgehend von einer fiktiven Kohorte von $l_0 = 5$ Personen (A,B,C,D und E) die Absterbeordnung l_x und alle weiteren bisher besprochenen Tafelfunktionen (d_x,l_x), sowie die noch zu definierenden Funktionen (vgl. Def. 12.9) T_x und e_x (bzw. T^*_x und e^*_x)!

Lösung 12.10:

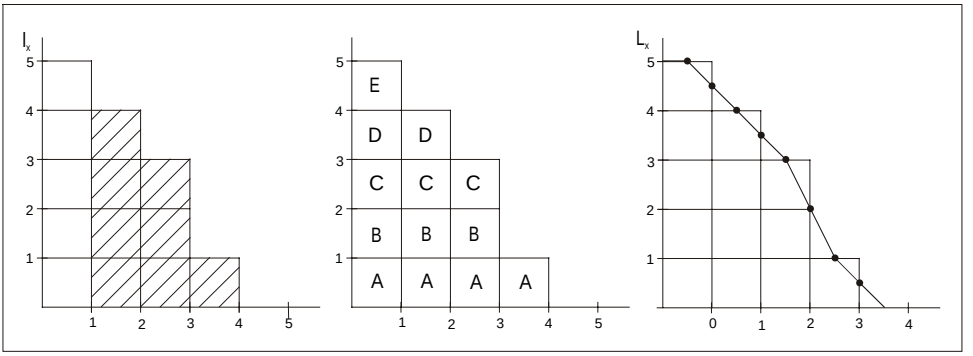
Wie man leicht sieht, handelt es sich hinsichtlich der Abgangsordnung um die gleichen Zahlen, wie in dem ausführlich behandelten Beispiel 12.7. Die folgende Tabelle[1)] enthält die Tafelfunktionen.

x	Personen	l_x	d_x	(Pers.)	L_x	T_x	e_x	T^*_x	e^*_x
0	A,B,C,D,E	5	1	E	4,5	10,5	2,1	13	2,6
1	A,B,C,D	4	1	D	3,5	6	1,5	8	2
2	A,B,C	3	2	B,C	2	2,5	0,833	4	1,333
3	A	1	1	A	0,5	0,5	0,5	1	1
4	-	0							

1) es gilt: $e^*_x = T^*_x / l_x$ und $e_x = T_x / l_x$

Zur graphischen Darstellung einiger Tafelfunktionen vgl. Abb. 12.5. In den Bemerkungen zur Def. 12.9 wird auf die Interpretation der Tafelfunktionen dieses Beispiels eingegangen.

Abb. 12.5: Zur Interpretation der Verweilsummen T_x und T_x^* (Bsp. 12.10)



Satz 12.4:

Es gilt:

(12.23) $\Sigma\, d_x = l_o \quad (x = 0,1,...,w)$

(was nur den trivialen Sachverhalt ausdrückt, dass die gesamte Kohorte vom Umfang l_0 bis Erreichung des Maximalalters w vollständig abstirbt).

Beweis:

Für Partialsummen, etwa für $\Sigma d_x = d_0 + d_1 + d_2$ (x = 0,1,2), läßt sich unter Verwendung von Gl. 12.21 leicht herleiten

$$\sum_{x=0}^{2} d_x = l_0(1 - p_0p_1p_2) = l_0\,(1 - p_{03}^*) = l_0 - l_3.$$

Dieser Zusammenhang gilt allgemein, so dass

$$\sum_{x=0}^{w} d_x = l_0 - l_{w+1} = l_0 \text{ gilt,}$$

da $p_w = 0$ und deshalb $l_{w+1} = 0$ (keiner erreicht das Alter w+1).

Def. 12.9: Tafelfunktionen T, e

a) Die Tafelfunktion T_x, die Zahl der von den Überlebenden des Alters x noch zu durchlebenden Jahre ist die Summe der Größen L_x, L_{x+1}, L_{x+2},...,L_w.

(12.24) $$T_x = \sum_{y=x}^{w} L_y \qquad x \le y \le w.$$

Häufig ist es auch sinnvoll, von der Größe

$$(12.25) \qquad T_x^* = \sum_{y=x}^{w} l_y = T_x + \tfrac{1}{2} l_x$$

auszugehen. Die Größen T_x und T_x^* sind Verweilsummen; Maßeinheit: "Personenjahre".

b) Dividiert man T_x^* durch die Anzahl der Überlebenden des Alters x, also durch l_x, so erhält man mit

$$(12.26) \qquad e_x = \frac{T_x}{l_x} = \frac{T_x^*}{l_x} - \tfrac{1}{2} = e_x^* - \tfrac{1}{2}$$

die (mittlere, durchschnittliche weitere) Lebenserwartung einer x-jährigen Person (spricht man von "der" Lebenserwartung, so ist e_0 gemeint).

Bemerkungen zu Def. 12.9:

1. Die Größen T_x^* sind die von unten (d.h. vom Maximalalter w an) aufkumulierten Werte l_x. Dies zeigt sich im Beispiel 12.10 etwa bei (w=3)
   $T_1^* = l_3 + l_2 + l_1 = 1 + 3 + 4 = 8$.
   Die Interpretation als die Gesamtzahl der von den $l_1 = 4$ Personen (bzw. Pferde) "noch zu durchlebenden Jahre" läßt sich wie folgt verdeutlichen: D lebt noch 1 Jahr, B und C leben noch jeweils 2 Jahre und A noch 3 Jahre (8 = 1+2+2+3). Entsprechend erhält man $T_0^* = l_4 + l_3 + l_2 + l_1 = 13$. Auch dies ist eine Verweilsumme. Es ist jeweils ein Jahr hinzuzurechnen, also A noch 4, B und C noch 3, D noch 2 und A ein Jahr, zusammen also 4 + 3 + 3 + 2 + 1 = 13 Jahre.

2. Diese Interpretation von T_x^* als Verweilsumme lässt sich auch graphisch veranschaulichen: T_x^* ist in Abb. 12.5 die schraffierte Fläche unter der Abgangsordnung (als Treppenfunktion) von rechts (Alter w) nach links (Alter x) gesehen.

3. Die Verweilsumme T_x ist die Fläche unter der Abgangsordnung als Polygonzug (Abb. 12.5 rechts) von rechts (Alter w) nach links (Alter x) gesehen. Auch L_x ist eine Verweilsumme und Fläche unter der Abgangsordnung.

4. Wenn $l_1 = 4$ Personen zusammen noch $T_1 = 6$ Jahre leben, dann liegt es nahe, die durchschnittliche weitere Lebenserwartung jeder einzel-

nen Person als $e_1 = T_1/l_1 = 6/4 = 1{,}5$ Jahre anzusetzen.

5. Berücksichtigt man, dass die Personen nicht jeweils kurz vor Erreichen ihres x+1 Geburtstags, sondern über das (Alters-) Jahr gleichmäßig verteilt sterben (vgl. Bem. 2 zu Def. 12.8), so erhält man T_x, also beispielsweise $T_1 = L_3 + L_2 + L_1 = 0{,}5 + 2 + 3{,}5 = 6$ Personenjahre (bzw. Pferdejahre). D lebt noch 0,5 Jahre (statt 1 Jahr), B und C leben noch jeweils 1,5 Jahre (statt 2 Jahre) und A noch 2,5 Jahre (statt 3 Jahre).

6. Man kann T_x auch in Abhängigkeit von den Größen l_x darstellen. Danach gilt:

$$(12.27)\quad T_x = \sum_{y=x}^{w} l_y - \tfrac{1}{2} l_x = T_x^* - \tfrac{1}{2} l_x \,.$$

   Daraus folgt, dass - wie erwähnt (Bem. 3 zu Def. 12.8) - die Verwendung von L_x anstelle von l_x zu einer um ein halbes Jahr kürzeren Lebenserwartung bei allen Altersklassen x = 0,1,...,w führt.

7. Einem verbreiteten Mißverständnis zufolge, nimmt die Lebenserwartung in dem Maße ab, in dem man älter wird, so dass gilt $e_{x+d} = e_x - d$. Man kann jedoch leicht zeigen, dass dies nicht der Fall ist, es sei denn, es gelte:

$$q_x = \begin{cases} 1 & \text{für } x = w \\ 0 & \text{sonst} \end{cases}$$

   was eine rechteckige Absterbeordnung impliziert. Die Lebenserwartung e_x ist auch (anders als die Abgangsordnung l_x) nicht notwendig monoton fallend, sie kann sogar vorübergehend ansteigen. Ein Ansteigen der Lebenserwartung (so dass $e_{x+1} > e_x$) ist möglich, wenn

   $$(12.28)\quad q_x e_{x+1}^* > 1.$$

   Ansteigende Lebenserwartungen kommen in jungen Jahren vor. Hat man ein kritisches Alter, im Bereich der Säuglingssterblichkeit oder im Bereich von 18 bis 19 Jahren (Verkehrstote!) überschritten, so kann die Lebenserwartung (meist nur) für ein Jahr ansteigen.

8. Der Zusammenhang

   $$(12.29)\quad e_{x+1}^* - e_x^* = e_{x+1}^* q_x - 1$$

   ergibt sich leicht aus der folgenden Überlegung: Man erhält mit
   $T_x^* = T_{x+1}^* + l_x$ sowie $l_{x+1} = l_x p_x$
   $e_x^* = 1 + e_{x+1}^* p_x$. Ferner gilt wegen $e_x^* = e_x - \tfrac{1}{2}$ auch

   $$(12.30)\quad p_x = \frac{e_x - \tfrac{1}{2}}{e_{x+1} + \tfrac{1}{2}}$$

9. Zwischen der mehrjährigen Überlebenswahrscheinlichkeit p_{xy}^* und der Lebenserwartung besteht folgender Zusammenhang: Aus
   $T_x^* = l_x (1 + p_x + p_x p_{x+1} + p_x p_{x+1} p_{x+2} + ... + p_x p_{x+1} \cdots p_{w-1})$ folgt

(12.31) $e_x^* = e_x + ½ = 1 + \sum p_{x,y}^*$ mit y = x+1,x+2,...w

Die Lebenserwartung wird also durch die Summe von Überlebenswahrscheinlichkeiten von x bis zum Maximalalter w bestimmt.

10. Nach Satz 12.1 gilt $T_0 = l_0 e_0$, d.h. bei einer stationären Bevölkerung ist der Bevölkerungsstand zur Zeit t, der P_t genannt sei, für alle t gleich T_0 und gleich dem Pro-dukt aus den Zugängen und der durchschnittlichen Verweildauer. Er setzt sich zusammen aus den Geburten (Zugängen) der Vorperioden Z_{t-x} "gewogen" mit den Überlebenswahrscheinlichkeiten

$$P_t = \sum_{x=0}^{x=w} Z_{t-x}\, p_{0x}^* ,$$

denn von der Kohorte Z_{t-x} ist noch ein Anteil von p_{0x}^* nach x Jahren (zur Zeit t) am Leben. Da alle Kohorten vom gleichen Umfang l_0 sind und die Überlebenswahr-scheinlichkeiten von den Zeitpunkten t unabhängig sind, ist der Umfang der Bevölkerung durch die Summe der Überlebenden l_x (bzw. L_x) also durch T_x^* (bzw. T_x) gegeben. Er hängt somit allein ab von e_0 und von der Stärke der Kohorten (Geburten) Z_t. Da für alle t gilt $Z_t = l_0$, ist die Geburtenrate $b_t = Z_t / P_t = l_0 / T_0 = (e_0)^{-1}$ und da die stationäre Bevölkerung nicht wächst, ist die konstante rohe Geburtenrate zugleich die (konstante rohe) Todesrate.

Bei einer stationären Bevölkerung ist die rohe Geburtenrate b_t und die rohe Todesrate gleich dem reziproken Wert der Lebenserwartung e_0.

Anders als die empirische Geburtenrate einer (nichtstationären) Bevölkerung ist die Geburtenrate einer stationären Bevölkerung nicht abhängig von Schwankungen des Altersaufbaus, die es ja bei einer stationären Bevölkerung ex definitione nicht gibt.

11. Der konstante Altersaufbau der stationären Bevölkerung (die Folge $l_0, l_1, l_2, ..., l_w$, also die Abgangsordnung) ist allein durch die Überlebenswahrscheinlichkeiten gegeben. Für das Durchschnittsalter der Lebenden $\bar{x}_L$ und das durchschnittliche Sterbealter (Durchschnittsalter der Gestorbenen) $\bar{x}_D$ gilt (Summen von x=0 bis x=w):

(12.33) $$\bar{x}_L = \frac{\Sigma(x+½)l_x}{\Sigma l_x}$$

(12.34) $$\bar{x}_D = \frac{\Sigma(x+½)d_x}{\Sigma d_x} = e_0 .$$

Man beachte, dass bei dieser Definition von $\bar{x}_L$ und $\bar{x}_D$, anders als bei der Betrachtung in Ziff. 3ff von Satz 12.3 und Bsp. 12.8f., die Korrektur um 1/2 für die gleichmäßige Verteilung der Sterbezeitpunkte auf das Intervall [x,x+1] vorgenommen wurde.

Satz 12.5:

Man kann die Lebenserwartung eines Nulljährigen e_0 (auch "mittlerer Sterbezeitpunkt" genannt) als durchschnittliches Sterbealter $\bar{x}_D$ interpretieren, d.h. es gilt Gl. 12.34

Beweis:

Es gilt

$$\begin{aligned} T_0^* &= l_0 + l_1 + ... + l_w \\ &= l_0 + (l_0-d_0) + (l_0-d_0-d_1) + ... + (l_0-d_0-...-d_{w-1}) \\ &= (w+1)l_0 - wd_0 - (w-1)d_1 - (w-2)d_2 - ...- 2d_{w-2} - 1d_{w-1}. \end{aligned}$$

Mit Satz 12.2 erhält man hieraus

$$\begin{aligned} T_0^* &= l_0 + 0\cdot d_0 + 1\cdot d_1 + 2\cdot d_2 + ... + w\cdot d_w = l_0 + \Sigma x d_x = l_0[1+(\bar{x}_D-½)] = \\ &= l_0(\bar{x}_D+½), \end{aligned}$$

da $\bar{x}_D = [\Sigma(x+½)d_x]/\Sigma d_x$ (mit x=0,1,...,w), so dass

$e_0^* = T_0^*/l_0 = \bar{x}_D + ½$ und wegen $e_0^* = e_0 + ½$ gilt $e_0 = \bar{x}_D$.

Satz 12.5 besagt auch: Die Fläche T_0 unter der Abgangsordnung (Polygonzug) ist das l_0-fache durchschnittliche Sterbealter. Man verifiziert diesen Zusammenhang leicht anhand des obigen Beispiels 12.10:

Person(en)	Sterbealter
E	0,5
D	1,5
C	2,5
B	2,5
A	3,5
Summe	10,5.

Somit ist: $\bar{x}_D = 10,5/5 = 2,1 = e_0$.