Peter von der Lippe

Deskriptive Statistik

Formeln, Aufgaben, Klausurtraining

Ursprünglich verlegt bei Oldenbourg, hier in überarbeiteter Form als download zur Verfügung gestellt

Oldenbourg

von der Lippe: Deskriptive Statistik
Inhalt von Teil I (Formelteil)

Vorwort (zugleich eine Art Einführung)

Die Entstehungsgeschichte dieses Buches ist ähnlich der des Buches "Induktive Statistik" in der gleichen Reihe. Der Formel- und Aufgabenteil ist auch hier aus einer früheren Veröffentlichung hervorgegangen, wurde aber noch einmal überarbeitet. Ferner ist ein Teil "Klausuraufgaben" dem bisher im Oldenbourg Verlag erschienenen Buch "Klausurtraining in Statistik", 1.-4. Auflage entnommen worden, einem Buch, das somit in dem vorliegenden Buch sowie in dem Buch "Induktive Statistik" aufgegangen ist. Schließlich enthält dieses Buch als vierten Teil auch einige effektiv in letzter Zeit von uns an der Universität-Gesamthochschule Essen gestellte Klausuraufgaben.

Die Erfahrung hat gezeigt, dass es für das Erlernen der Statistik von großer Wichtigkeit ist, sich selbständig mit Kenntnis des Vorlesungsstoffs an das Lösen von Aufgaben zu machen. Dabei besteht auch ein Unterschied zwischen Übungsaufgaben, die sich jeweils auf einen Ausschnitt des (gerade gelernten) Stoffes beziehen und auch oft den Charakter von (in der Vorlesung benutzten) Demonstrationsbeispielen haben einerseits und Klausuraufgaben andererseits.

Das Buch ist gedacht als Begleitlektüre zu Vorlesungen und Übungen, wie sie üblicherweise unter dem Titel "Deskriptive Statistik" oder "Statistik I" an den meisten Hochschulen für Wirtschaftswissenschaftler angeboten werden. **Wenn** entsprechende Veranstaltungen besucht werden, sollte das Buch ausreichend sein zur Klausurvorbereitung. Dazu sind jedoch noch einige (etwas persönliche) Anmerkungen zum Was und Wie des Statistikstudiums angebracht.

Es wird nicht selten versucht, die Statistik als bloße Anwendung der Wahrscheinlichkeitsrechnung darzustellen oder die Unterscheidung zwischen Deskription und Induktion aufzulösen. Von dem, was man unter "Deskriptive Statistik" verstehen kann, bleiben dann allenfalls Gegenstände, wie sie hier in Kap. 3-5 sowie 7 und 8 (oder in Teilen dieser Kapitel) behandelt werden, übrig und sie werden quasi als Einführungen in bzw. Vorbemerkungen zu Darstellungen der entsprechenden Konzepte der Wahrscheinlichkeitsrechnung betrachtet. Ein solches Verständnis von Statistik wird m. E. weder der Leistungsfähigkeit der Statistik noch den Bedürfnissen der Nutzer von Statistik(en) in der Praxis (insbesondere auch der Wirtschaftspraxis) gerecht. Es mag auch mitverantwortlich sein für den Eindruck mancher Studenten, aber auch mancher Professoren der Wirtschaftswissenschaft, die Statistik sei eine mathematische Hexerei, die sich immer mehr in den Elfenbeinturm zurückzieht, und sie sei deswegen

eigentlich entbehrlich bzw. man könne sie sich ohne Mitwirkung von Statistikern von Fall zu Fall selbst aneignen. Nach unserem Verständnis ist aber Statistik nicht nur ein Teil der Mathematik und sie bietet viele Methoden zur Erkenntnisgewinnung aufgrund zahlenmäßiger Informationen, die nicht notwendig immer auf Wahrscheinlichkeitsüberlegungen beruhen. Gerade für Anwender aus der Wirtschaft sind "rein" beschreibende Methoden mindestens genau so wichtig wie stochastisch fundierte Methoden, und man kann sie nicht richtig verstehen und interpretieren, wenn man sie nur als Rechenaufgaben auffaßt. Man sollte also "Deskriptive Statistik" (und auch die hieran – was v.a. Kap. 10 und 12 zeigt – angrenzende "Wirtschaftsstatistik") als selbständige Gegenstände betrachten, die es wert sind, sich mit ihnen zu beschäftigen.

Mehr Daten, mehr Rechenfähigkeiten und auch mehr Zwang, etwas empirisch "belegen" zu wollen, führt nicht nur zu mehr Anwendung der Statistik, sondern auch zu mehr Fehlanwendung. Dabei kann mit der Art, wie man Statistik lernt, schon der Grundstein für Fehlanwendung gelegt werden. Statistik kann man weder durch bloßes Hören von Vorlesungen (oder gar Auswendiglernen von Begriffen) lernen, noch durch (Nach-) Rechnen von Aufgaben, die einem vorgerechnet werden. Man kann nicht mit ihr umgehen, wenn man nur im Abstrakten bleibt oder nur lernt, Zahlen in Formeln einzusetzen. Sowohl Vorlesungen (wofür der Formelteil quasi ein Notizgerüst liefert) als auch Übungen (also Aufgaben) sind notwendig und der Reiz (aber leider auch die Schwierigkeit für viele) besteht darin, beides zu verbinden, Methoden und ihre (Rechen-) Ergebnisse. Die Fähigkeit, Methoden und Anwendungen zu verbinden, eine Anwendbarkeit zu erkennen und ein Ergebnis zu interpretieren, verlangt Kenntnisse **und** Übung, Verstehen und auch Phantasie. So etwas zu erlernen kann einem niemand abnehmen; man kann nur versuchen, es zu erleichtern.

Welche Art von Übungsaufgabe man als hilfreich empfindet, um für Statistik motiviert zu werden oder vielleicht auch die angesprochenen Fähigkeiten zu erwerben, ist sicher zum großen Teil Geschmackssache. Für viele sind dafür tatsächliche Anwendungen mit großen und evtl. auch unhandlichen Datensätzen aus Betrieben besonders motivierend. Wir glauben jedoch, dass es ein Schritt weiter ist, angeregt zu werden, sich selbst "Aufgaben" auszudenken. Wer Anwendungen anderer studiert, wird daraus viel lernen, wer aber Spaß daran findet, auch eigene Anwendungen zu konstruieren, könnte einen Schritt mehr Souveränität (und damit auch Motivation) gewinnen. Auf längere Sicht wird man nur das wirklich können, was einem auch Freude macht. Solche Überlegungen stecken auch hinter der Art der Aufgaben, die hier zusammengestellt sind.

Im Unterschied zum Buch "Induktive Statistik" kann hier auch auf einen Begleittext verwiesen werden, der die mit der Formelsammlung präsentierten Stichworte durch Erläuterungen verbindet:

P. v.d. Lippe: Deskriptive Statistik, Reihe UTB
(Uni-Taschenbücher) Bd. 1632, Stuttgart, Jena, 1993.

Die Nummerierung von Formeln und Definitionen im vorliegenden Buch nimmt darauf Bezug.

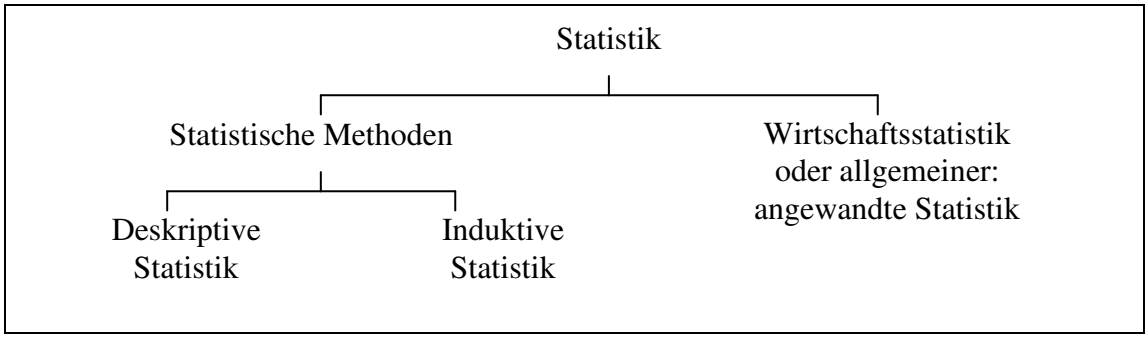
Bei der Vorbereitung des vorliegenden Buches wurde ich erneut von Herrn Dipl. Volkswirt Andreas Kladroba sehr tatkräftig unterstützt. Nach Einstellung von Herrn Dipl. Volkswirt Michael Westermann konnte die nicht unerhebliche Arbeit an der Überarbeitung und Neugestaltung der Texte auch etwas geteilt werden. Beide Mitarbeiter, Herr Kladroba und Herr Westermann, haben nicht nur mit viel Engagement die Veröffentlichung vorbereitet, sondern auch inhaltlich viel beigetragen aufgrund ihrer großen Erfahrungen mit Übungen und Tutorien sowie Klausuren. Ich danken ihnen sehr herzlich dafür. Ich danke auch Frau stud. rer. pol. Karla Behal und Frau stud. rer. pol. Alexandra Werner für die Arbeiten am PC, die sich wieder, wie beim Buch "Induktive Statistik" als aufwendiger und schwieriger herausstellten als wir zunächst dachten.

Essen, den 11.02.99

Kapitel 1: Gegenstand und Grundbegriffe der Statistik

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

Übersicht 1.1 Aufbau des Faches Statistik



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).

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.

Def. 1.3: diskret und stetig

Eine metrisch skalierte 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.

Def. 1.4: Messung

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.

Skala (Name, Typ)	definiert ist zusätzlich	zulässige Transformationen	anschauliches Beispiel	Mittelwert
Nominalskala	Äquivalenzrelation (=, ≠)	ein-eindeutige Transformation	Postleitzahlen Steuerklasse	Modus
Ordinalskala	Ordnungrelation (>, <)	streng monoton steigend	Windstärke (Beaufort)	Median
Intervall-Skala	Maßeinheit und Nullpunkt*	linear $y_v = a + bx_v$	Temperatur in Grad Celsius	$\bar{x}$
Ratio- bzw. Verhältnisskala	natürl. Nullpunkt (Maßeinheit noch wilkürlich)	proportional $y_v = bx_v$ $(a = 0)$**	Temperatur in Kelvin, Körpergröße	$\bar{x}_G$ $\bar{x}_H$
Absolutskala	auch natürliche Maßeinheit	identisch $y_v = x_v$ $(b = 1)$	Häufigkeit	

* beides (Nullpunkt und Maßeinheit) noch willkürlich.

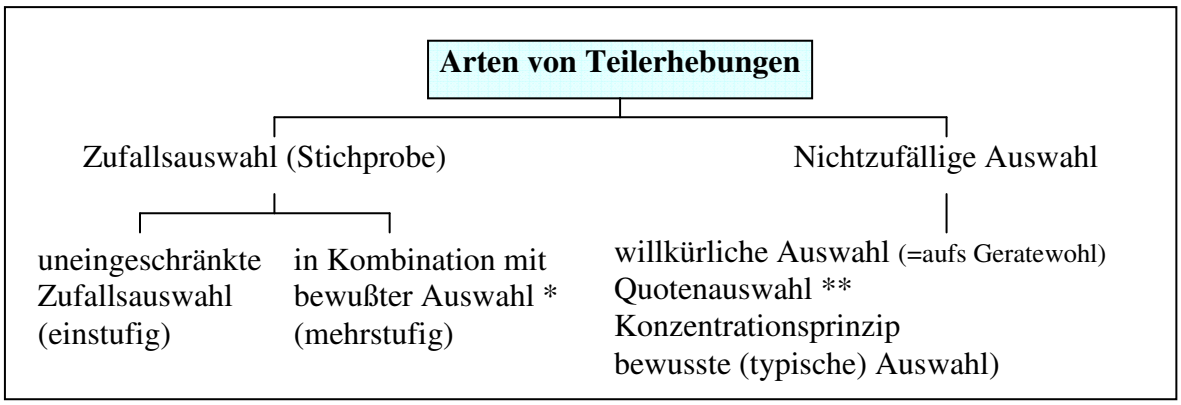
** d.h. der Nullpunkt ist nicht mehr willkürlich (er kann nicht durch $a \neq 0$ verschoben werden), wohl aber die Maßeinheit (weshalb $b \neq 1$ sein kann). Man kann sinnvoll Verhältnisse x_1/x_2 (Proportionen, engl. "ratios") bilden (denn $y_1/y_2 = x_1/x_2$).

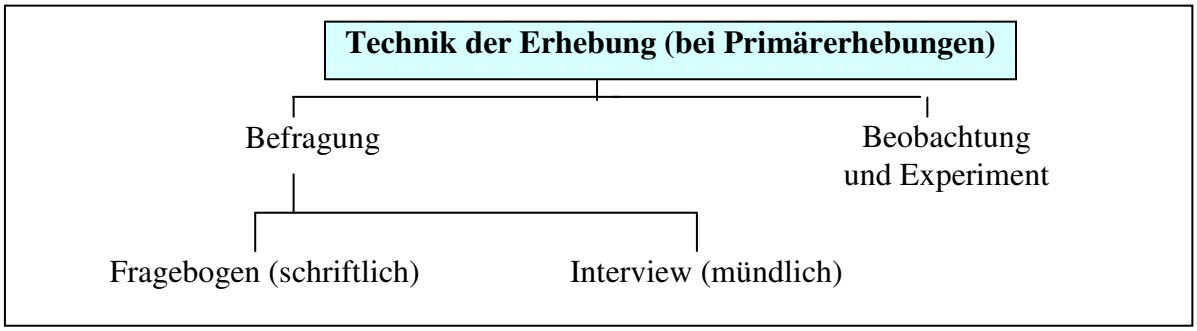
Kapitel 2: Daten, Maßzahlen und Axiomatik

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 zugrunde zu legenden Daten bilden einen Datensatz.

Übersicht 2.2: Methoden der Datengewinnung





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

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

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,

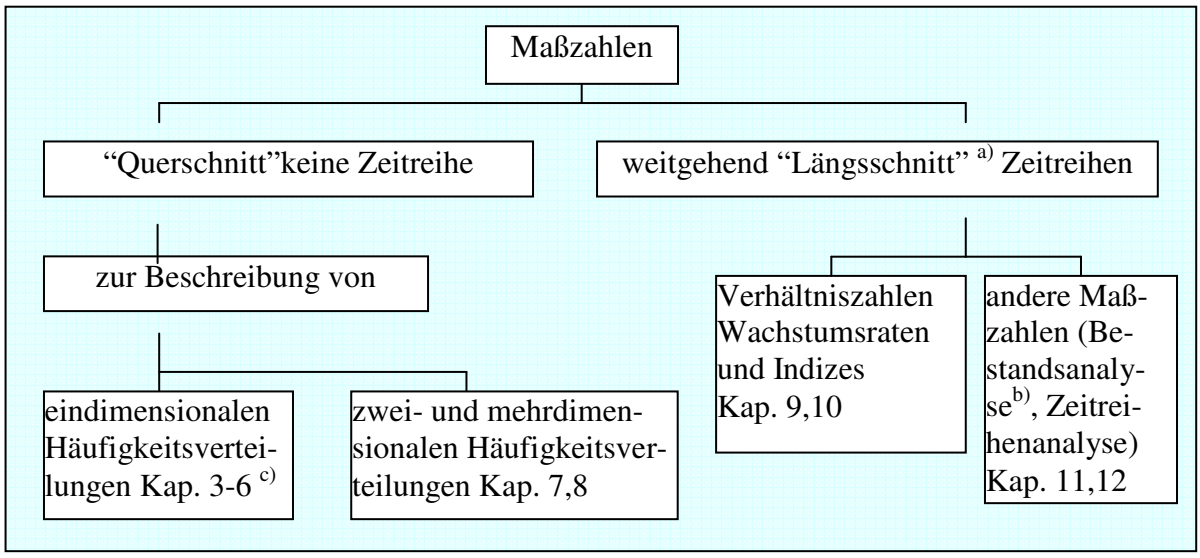
(2.1) $f: R^n \rightarrow R$, $M = f(x_1, x_2,...,x_n)$

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

Übersicht 2.3: Arten von Maßzahlen



a) Viele, aber nicht alle Methoden sind auf Zeitreihen (nicht verwechseln mit "Längsschnittsdaten") bezogen. Bestimmte Verhältniszahlen, wie Gliederungs- und Beziehungszahlen beziehen sich auf Querschnittsdaten.

b) Kennzahlen der Bestandsanalyse wie z.B. Durchschnittsbestand, Umschlagshäufigkeit, mittlere Verweildauer dienen der Beschreibung von Abläufen, die zu Bestandsänderungen führen.

c) Die Berechnung vieler der 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.

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

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) $$M^* = M_u^* + (M - M_u)\frac{M_o^* - M_u^*}{M_o - M_u}$$

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

(2.3a) Normierung von M^* auf den Wertebereich $0 \le M^* \le 1$

(2.3b) Normierung von M^* auf den Wertebereich $-1 \le M^* \le +1$

Kapitel 3: Eindimensionale Häufigkeitsverteilungen

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,

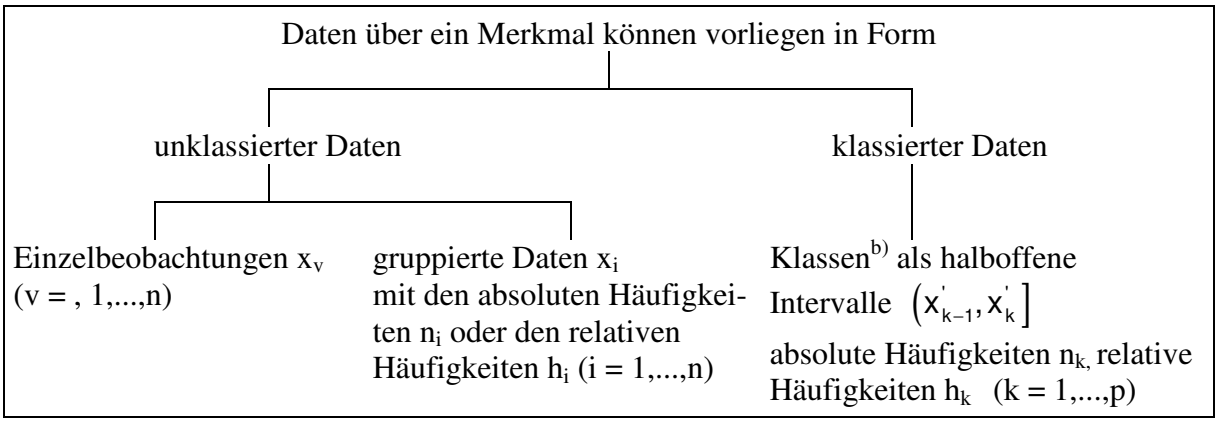
(3.1) $n_i = n(x_i)$ absolute Häufigkeit $(i = 1,2,...,m)$
und mit $n = \sum n_i$ (Gesamthäufigkeit, Umfang der Beobachtungsgesamtheit) der Quotient

(3.2) $h_i = h(x_i) = n_i/n$ relative Häufigkeit
der i-ten Ausprägung des Merkmals X. Es gilt $0 \le h_i \le 1$ und (wegen $n = \sum n_i$) $\sum h_i = 1$.

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. Eine Häufigkeitsverteilung ist also eine Zuordnung von Häufigkeiten (h_i oder n_i) zu Merkmalsausprägungen x_i. Graphische Darstellung durch ein Histogramm (Balken-, Block-, Stabdiagramm).

Übersicht 3.1:



[a)] In späteren Abschnitten (insbes. im Kap. 8 und 10) wird gelegentlich auch x_i anstelle von x_v verwendet.

[b)] 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. 3.3: Summenhäufigkeit, Verteilungsfunktion

Die Summe N_i der absoluten Häufigkeiten n_j ($j = 1,2, \ldots ,i$) aller Merkmalsausprägung 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). Entsprechend heißt

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

relative kumulierte Häufigkeit (relative Summenhäufigkeit).
Die Funktion

$$(3.5) \quad H(x) = \begin{cases} 0 & \text{für} \quad x < x_1 \\ H_j & \text{für} \quad x_j \le x < x_{j+1} \\ 1 & \text{für} \quad x \ge x_{m`} \end{cases}$$

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

Def. 3.4: Resthäufigkeit

Die Summe N_i^- der absoluten Häufigkeiten n_j ($j = i+1, i+2, \ldots, m$) aller Merkmalsaus-

prägungen, die größer als x_i sind, $N_i^- = N^-(x_i) = n(x > x_i) = \sum_{j=i+1}^{m} n_j = n - N_i$

heißt absolute Resthäufigkeit. Entsprechend: $H_i^- = 1 - H_i$ (relative) Resthäufigkeit und $H^-(x) = 1 - H(x)$ relative Resthäufigkeitsfunktion.

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.

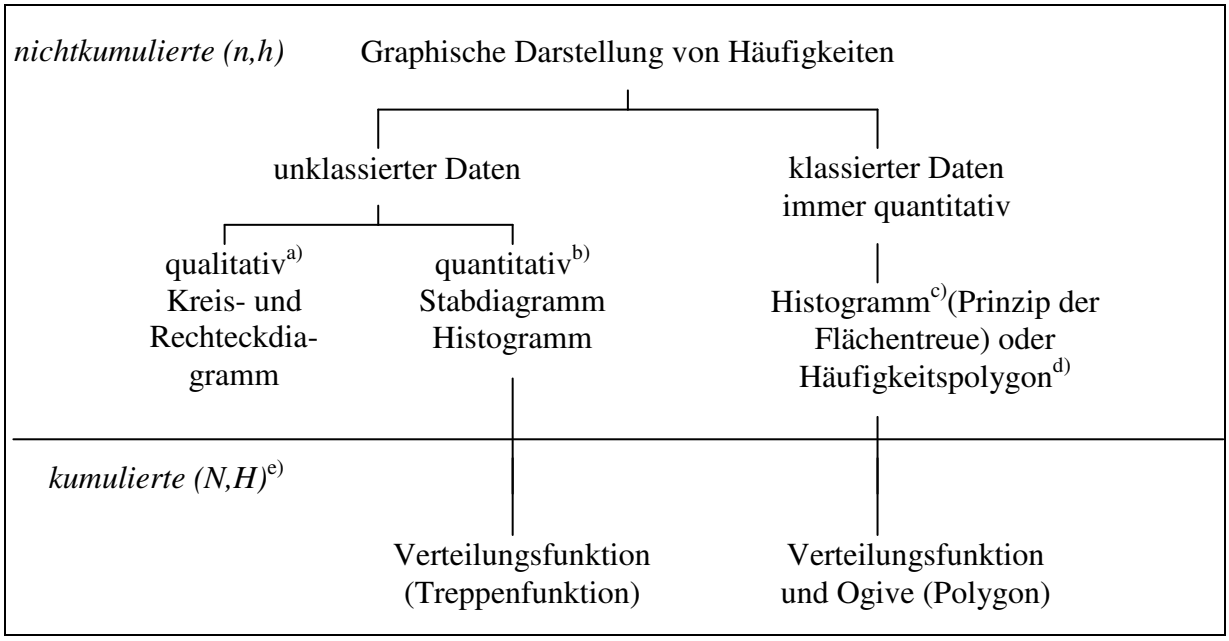
b) Die Differenz $b_k = x'_k - x'_{k-1}$ heißt *Klassenbreite* und die Größe $m_k = \frac{1}{2}\left(x'_{k-1} + x'_k\right)$ heißt *Klassenmitte* der k-ten Klasse.

c) Die Anzahl n_k der Einheiten, die in die k-te Klasse "fallen" $n_k = n(x'_{k-1} < x \le x'_k)$ ist die absolute *Klassenhäufigkeit* und der Anteil $h_k = n_k/n$ heißt relative Klassenhäufigkeit.

d) Der Quotient $h_k^* = h_k/b_k$ (Häufigkeit je Klassenbreite) ist die *Häufigkeitsdichte* oder einfach die *Dichte*.

Graphische Darstellung von Häufigkeitsverteilungen und Summenhäufigkeiten

Übersicht 3.2: Graphiken



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 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).

Kapitel 4: Mittelwerte und andere Lagemaße

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 $x_{(1)} \le M \le x_{(n)}$ bzw. bei Merkmalsausprägungen $x_1 \le M \le x_m$.
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 $\mathbf{x} \ge \mathbf{y}$ 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).

Def. 4.2: Arithmetisches Mittel

(4.4) $$\bar{x} = \frac{1}{n}\sum_{v=1}^{n} x_v$$ Berechnung aus Einzelbeobachtungen,

ungewogenes arithmetisches Mittel

oder

(4.5) $$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i = \sum_{i=1}^{m} x_i h_i$$ Berechnung aus Merkmalsausprägungen

gewogenes arithmetisches Mittel

Satz 4.1: Schwerpunkteigenschaft des arithmetischen Mittels

$$\sum_{v=1}^{n}(x_v - \bar{x}) = 0 \qquad \text{bzw.} \quad \sum_{i=1}^{m}(x_i - \bar{x})h_i = 0 .$$

Satz 4.2: Minimumeigenschaft

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

Satz 4.3: Lineartransformation des arithmetischen Mittels

Das arithmetische Mittel erfüllt das Mittelwertaxiom M3 für lineare Transformationen.
Aus $x_v^* = a + b \cdot x_v$ folgt

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

Arithmetisches Mittel bei klassierten 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) $$\bar{x} = \sum_{k=1}^{p} \bar{x}_k \, h_k$$

Andernfalls verwendet man die *Klassenmitten* m_k und erhält den *geschätzten* Gesamtmittelwert m (als Schätzung von $\bar{x}$) mit:

(4.10) $$m = \sum_{k=1}^{p} m_k h_k$$

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

Def. 4.3: Geometrisches Mittel

Die Maßzahl

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

(das Produktzeichen Π bedeutet $\Pi x_v = x_1 x_2 ... x_n$), bzw.

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

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

(4.13) $$\log \bar{x}_G = \frac{1}{n} \sum_{v=1}^{n} \log x_v$$

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.

Def. 4.4: harmonisches Mittel

Die Maßzahl

(4.14) $$\bar{x}_H = \frac{n}{\sum_v 1/x_v}$$ (bei der Berechnung aus Einzelbeobachtungen)

(4.15) $\bar{x}_H = \dfrac{n}{\sum_i n_i/x_i} = \dfrac{n}{\sum_i h_i/x_i}$ (bei gruppierten Daten, [Häufigkeitsverteilung])

heißt harmonisches Mittel ($x \neq 0$).
Es gilt: Der reziproke Wert von $\bar{x}_H$ ist das arithmetische Mittel der reziproken Werte (also der Werte $1/x_v$).

Def. 4.5: quadratisches- und antiharmonisches Mittel

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

(4.18) $\bar{x}_Q = +\sqrt{\dfrac{1}{n}\sum x_v^2}$

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

(4.19) $\bar{x}_Q = \sqrt{\sum x_i^2 h_i}$ berechnet.

b) Die Maßzahl

(4.20) $\bar{x}_A = \bar{x}_Q^2 / \bar{x}$

heißt antiharmonisches Mittel.

Def. 4.6: Potenzmittel

(4.21) $\bar{x}_{P,r} = \left[\dfrac{1}{n}\left(x_1^r + x_2^r + \ldots + x_n^r\right)\right]^{\frac{1}{r}} = \left(\dfrac{1}{n}\sum_{v=1}^{n} x_v^r\right)^{\frac{1}{r}}$ (ungewogene Berechnung), bzw.

(4.22) $\bar{x}_{P,r} = \left(x_1^r h_1 + x_2^r h_2 + \ldots + x_m^r h_m\right)^{1/r} = \left(\sum_{v=1}^{m} x_i^r h_i\right)^{1/r}$ (gewogene Berechnung)

Spezialfälle:

$r = -1$	harmonisches Mittel
$r \to 0$	geometrisches Mittel
$r = 1$	arithmetisches Mittel
$r = 2$	quadratisches Mittel

Ungleichung von Cauchy

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

Def. 4.7: Median

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

a) bei Einzelbeobachtungen die Maßzahl

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

Der Median ist der Wert, der in einer der Größe nach geordneten Reihe $x_{(1)} \leq x_{(2)} \leq \ldots$ $x_{(n)}$ in der Mitte, d.h. an der $\dfrac{1}{2}(n+1)$ - 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, Merkmalsausprägungen) gilt entsprechend für den Median

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

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

Interpolation des Medians

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

Dabei gilt: k = Medianklasse
b_k = Breite der Medianklasse
x'_{k-1} = Obergrenze der k-1-ten Klasse (= Untergrenze der k-ten Klasse)

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

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

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

Quantile bei klassierten Daten

$$(4.26a)\qquad \tilde{x}_p=x'_{k-1}+b_k\left(p-H_{k-1}\right)/h_k$$ für das interpolierte p-Quantil.

"Mittelpunkt" des Streubereichs (midrange)

$$(4.28)\qquad \tilde{x}_M=\tfrac{1}{2}\left(x_{min}+x_{max}\right)$$

Def. 4.9: Modus

Existiert bei einer diskreten Variable (einem diskreten Merkmal) X mit den Merkmalsausprägungen x_i genau ein Merkmalswert x_{i*} dergestalt, dass

$$(4.29)\qquad h(x = x_{i*}) = \max_i h(x_i),$$

so ist dieser Wert der Modus $D=\bar{x}_{mod}$ (oder der Modalwert, der dichteste oder häufigste Wert), also $D = x_{i*}$.
Der Modus ist derjenige Merkmalswert, der in einer Häufigkeitsverteilung am häufigsten (absolute oder relative Häufigkeit) vorkommt.

Kapitel 5: Streuung, Schiefe, Wölbung

Konstruktionsprinzipien für Streuungsmaße

1. Mittelwert aus Abständen (Abweichungen) der einzelnen Beobachtungen von einem Lageparameter (vgl. Übersicht 5.1)
2. Abstand zweier Ordnungsstatistiken untereinander (z.B. Spannweite)
3. Mittlerer Abstand der Merkmalswerte untereinander (z.B. Ginis Maß)

Übersicht 5.1: Einige Streuungsmaße nach dem Konstruktionsprinzip Nr. 1

Abweichung vom	Mittel der Abweichung	(absolutes) Streuungsmaß
arithmet. Mittel	Quadratisches Mittel	Standardabweichung
arithmet. Mittel*)	Arithmetisches Mittel	Varianz
Median**)	Arithmetisches Mittel	durchschn. Abweichung
Median**)	Median	Medianabweichung

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

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, dann soll für die Streuung S_y der mit $y_v = a+bx_v$ transformierten Variablen X gelten: $S_y = \|b\|S_x$, wobei $\|b\| > 0$. Für ein absolutes Streuungsmaß mit der quadrierten Maßeinheit der Merkmalswerte soll dann gelten $S_y = b^2 S_x$.

Def. 5.1: Relative Streuung

Die Maße der relativen Streuung (S_r) sind definiert als Quotienten eines absoluten Streuungsmaßes S und eines Mittelwertes M (wenn $M \neq 0$), $S_r = S/M$ sofern S die Maßeinheit der Merkmalswerte hat.

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

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

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

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

b) Die positive Quadratwurzel aus der Varianz heißt Standardabweichung s

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

Varianz des lineartransformierten Merkmals X

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

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

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

Verschiebungssatz

$$(5.5) \qquad s^2 = \frac{1}{n}\sum_{v=1}^{n}x_v^2 - \bar{x}^2 \qquad \text{(bei Einzelbeobachtungen) bzw.}$$

$$(5.6) \qquad s^2 = \frac{1}{n}\sum_{i=1}^{m}x_i^2 n_i - \bar{x}^2 = \sum_{i=1}^{m}x_i^2 h_i - \bar{x}^2 \qquad \text{(bei einer Häufigkeitsverteilung)}$$

Steinerscher Verschiebungssatz

$$(5.7) \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 - (\bar{x} - c)^2$$

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_c^2 bezeichnen kann. Zwischen s^2 (oder s_x^2) und s_c^2 besteht nach Gl. 5.7 die folgende Beziehung:

$$(5.7a) \qquad s_x^2 = s_c^2 - (\bar{x} - c)^2$$

Mit c = 0 erhält man Gl. 5.5 und 5.6 als Spezialfall.

Streuungszerlegung

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

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

(5.9) $s_{ext}^2 = \sum_{k=1}^{r} h_k (\bar{x}_k - \bar{x})^2 \quad \text{mit } h_k = n_k / n$

ein gewogenes Mittel der quadrierten Abstände zwischen den r Mittelwerten der Teilgesamtheiten (x_k; das müssen nicht Mittelwerte von Klassen, also von Teilgesamtheiten im Sinne einer klassierten Verteilung sein) und dem Gesamtmittelwert x. Die interne Varianz ist demgegenüber das gewogene Mittel der Varianz s_k^2 innerhalb der Teilgesamtheiten

(5.10) $s_{int}^2 = \sum_{k=1}^{r} h_k - s_k^2$

mit den relativen Häufigkeiten h_k als Gewichte.

Varianz bei klassierten Daten

(5.11) $s^2 = \sum_{k=1}^{r} h_k (\bar{x}_k - \bar{x})^2 + \sum_{k=1}^{r} h_k s_k^2 = s_{ext}^2 + s_{int}^2$

wobei s_k^2 die Varianz innerhalb der k-ten Klasse ist. Gl. 5.11 ist also ein Spezialfall von Gl. 5.8-5.10 (Teilgesamtheiten als aufeinanderfolgende Größenklassen). Bei unbekannten Klassenmittelwerten gilt als Näherung für s_{ext}^2 :

(5.11a) $s_m^2 = \sum (m_k - m)^2 h_k$

(mit m [m_k] als wahren oder geschätzten Gesamtmittelwert [Klassenmittelwert])

Sheppard-Korrektur

Sie geht davon aus, dass die Varianz durch s_m^2 häufig systematisch überschätzt wird. Deshalb ist eine bessere Approximation für s^2 gegeben, wenn man SK = $b^2/12$ (SK = Sheppard-Korrektur) von s_m^2 subtrahiert.

Def. 5.3: Durchschnittliche und Medianabweichung

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

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

und a_1, a_2,...,a_m seien die entsprechenden absoluten Abweichungen der Merkmalsausprägungen x_1, x_2,..., x_m .

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

Dann ist das arithmetische Mittel der absoluten Abweichungen vom Median

(5.25) $d_x = \frac{1}{n}\sum_{v=1}^{n} a_v$ bei Einzelwerten

(5.26) $d_x = \sum_{i=1}^{m} a_i h_i$ 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) $$m_x = \begin{cases} a_{(n+1)/2} & \text{, falls n ungerade ist} \\ \frac{1}{2}\left[a_{(n/2)} + a_{(n/2+1)}\right] & \text{, falls n gerade ist.} \end{cases}$$

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.

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) $$d_x^* = \begin{cases} \frac{1}{n}\sum_{v=1}^{n} |x_v - \bar{x}| & \text{bei Einzelwerten} \\ \sum_{i=1}^{n} |x_i - \bar{x}| h_i & \text{bei Häufigkeitsverteilungen} \end{cases}$$

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

Def. 5.4: Spannweite, Quartilsabstand, Quantilsabstände

a) Die Differenz zwischen dem (der) größten und kleinsten Beobachtungswert (Merkmalsausprägung) heißt *Spannweite R* (range, Wertebereich, Variationsbreite). Sie ist bei Einzelwerten durch

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

und bei Häufigkeitsverteilungen durch die Differenz zwischen kleinster und größter Merkmalsausprägung gegeben (die Berechnung von R ist jedoch vorwiegend bei Einzelwerten 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* $Q_{0,25}$ (Semiquartilsabstand) ist durch Gl. 5.33 gegeben:

(5.32) $Q_{0,25} = Q_3 - Q_1$ und (5.33) $\overline{Q}_{0,25} = \frac{1}{2} Q_{0,25} = \frac{1}{2}\left[(Q_3 - Q_2) + (Q_2 - Q_1)\right]$

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) $Q_p = \tilde{x}_{1-p} - \tilde{x}_p$ mit $0 < p < 0{,}5$.

Analog zu Gl. 5.33 heißt dann die Maßzahl $\overline{Q} = ½ Q_p$ *mittlerer Quantilsabstand* (Semiquantilsabstand).

Größenbeziehung zwischen d_x, s und R: $d_x \le s \le R$.

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

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

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

Seltener ist das Maß

(5.40a) $$S_G^* = \frac{1}{n^2} \sum \sum |x_v - x_w| \qquad v, w = 1, \ldots, n$$

Variationskoeffizient

(5.51) $$V = \frac{s}{\overline{x}}$$ (Standardabweichung/arithmetisches Mittel)

Def. 5.8: Quartilsdispersionskoeffizient

Setzt man den mittleren Quartilsabstand $\overline{Q}_{0,25} = ½ (Q_3 - Q_1)$ als Maß der absoluten Streuung ins Verhältnis zum Wert $½ (Q_1 + Q_3)$, den man als eine Art Mittelwert interpretieren kann (analog zu Gl. 4.28), so erhält man QD, den *Quartilsdispersionskoeffizient*

(5.52) $$QD = (Q_3 - Q_1) / (Q_3 + Q_1).$$

Der Quartilsdispersionskoeffizient kann auch mit dem Median ($\tilde{x}_{0,5} = Q_2$) berechnet werden, man erhält dann

(5.52a) $$QD^* = (Q_3 - Q_1) / Q_2 .$$

Auf der Basis des Medians lassen sich auch andere Maße der relativen Streuung konstruieren, etwa

(5.52b) $$RD = d_x / Q_2 = d_x / \tilde{x}_{0,5},$$

eine relativierte durchschnittliche Abweichung.

Def. 5.9: Momente

a) Mit der beliebigen reellen Konstanten a ist der folgende Ausdruck definiert als das k-te Moment um a:
   - bei Einzelwerten (ungewogene Berechnung)

(5.53) $m_{k(a)} = \frac{1}{n}\sum_{v=1}^{n}(x_v - a)^k$ [k-tes Moment um a]

- bei Häufigkeitsverteilungen (gewogene Berechnung)

(5.54) $m_{k(a)} = \frac{1}{n}\sum_{j=1}^{m}(x_j - a)^k n_j = \sum(x_j - a)^k h_j$

b) Spezialfälle: Anfangsmomente (oder Momente um Null) und zentrale Momente sind Spezialfälle des Moments um a (Übers. 5.2).

c) Von geringerer Bedeutung sind absolute Momente: analog Gl. 5.53 ist das k-te absolute Moment um a definiert als

(5.53a) $m^*_{k(a)} = n^{-1}\sum| x_v - a |^k$ [k-tes absolutes Moment um a].

Bei einer geraden Zahl sind die absoluten Momente gleich den "gewöhnlichen Momenten" [= Momente im Sinne von a) bzw. b)]

d) Für mehrdimensionale Verteilungen sind Produktmomente definiert (vgl. Kap. 7).

Bei proportionaler Transformation $y_v = bx_v$ gilt für zentrale Momente und Anfangsmomente: $z_{k(y)} = b^k z_{k(x)}$.

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

(5.57) $h(\tilde{x}_{0,5} - c) = h(\tilde{x}_{0,5} + c) \qquad c > 0.$

Dabei ist $h(\tilde{x}_{0,5} - c)$ die relative Häufigkeit der Merkmalsausprägung $\tilde{x}_{0,5} - c$ und $h(\tilde{x}_{0,5} + c)$ ist entsprechend definiert. Eine Verteilung ist schief oder asymmetrisch, wenn Gl. 5.57 nicht gilt. Diese Definition ist jedoch nicht generell brauchbar für die Konstruktion von Schiefemaßen.

Fechnersche Lageregel

(5.59) linkssteil: $\bar{x}_{mod} < \tilde{x}_{0,5} < \bar{x}$ rechtssteil: $\bar{x} < \tilde{x}_{0,5} < \bar{x}_{mod}$

($\bar{x}_{mod}$ = Modus, $\tilde{x}_{0,5}$ = Median)

Def. 5.12: Schiefemaße

a) Die von Bowley und Fisher eingeführte *Momentschiefe* (Momentkoeffizient der Schiefe) lautet:

(5.60) $SK_M = \frac{z_3}{s^3}$ (z_3 ist das dritte zentrale Moment)

linkssteil: $SK_M > 0$

rechtssteil: $SK_M < 0$

symmetrisch: $SK_M = 0,$ da $z_3 = 0.$

b) Als Quantilskoeffizient der Schiefe wird bezeichnet:

(5.61) $$SK_{Q,p} = \frac{(\tilde{x}_{1-p} - Q_2) - (Q_2 - \tilde{x}_p)}{\tilde{x}_{1-p} - \tilde{x}_p} \qquad p < ½$$

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) $$SK_Q = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$, mit $-1 \le SK_Q \le +1$.

c) Auf der Fechnerschen Lageregel beruhen die folgenden, von (Yule und Pearson) vorgeschlagenen Schiefemaße (Pearsonsche Schiefemaße):

(5.63) $$SK_{p1} = \frac{\bar{x} - \bar{x}_{mod}}{s} \qquad (\bar{x}_{mod} = \text{Modus})$$

(5.64) $$SK_{p2} = \frac{3(\bar{x} - \tilde{x}_{0,5})}{s} \qquad (\tilde{x}_{0,5} = \text{Median})$$

Der zweite Koeffizient hat Vorteile, weil der Modus evtl. schwer zu bestimmen ist.

Def. 5.13: Symmetrisierende Potenztransformation

Die Variable X wird in die Variable Y nach Maßgabe einer Potenztransformation transformiert, wenn gilt

(5.66) $$y_v = (x_v + c)^p \quad \text{für } p \neq 0 \qquad \text{und} \qquad y_v = \ln(x_v + c)^p \quad \text{für } p = 0.$$

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) $$W_M = \frac{z_4}{s^4} - 3.$$

b) Weniger bekannt sind Wölbungsmaße auf der Basis von Quantilen, etwa ein Quantilskoeffizient W_Q der Wölbung:

(5.68) $$W_Q = 1 - \frac{\tilde{x}_{1-p} - \tilde{x}_p}{\tilde{x}_{1-q} - \tilde{x}_q}$$, mit $0 < q < p < ½$.

- $W_M = 0$ 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).

Übersicht 5.2: Momente

Moment um a
a ist eine beliebige reelle Konstante

(5.53) $m_{k(a)} = \frac{1}{n}\sum_{v=1}^{n}(x_v - a)^k$ [k-tes Moment um a]

(ungewogene Berechnung, die gewogene Berechnung erfolgt analog, vgl. Def. 5.9)

Spezialfälle

Anfangsmoment $a = 0$

k-tes Anfangsmoment

(5.54) $m_k = \frac{1}{n}\sum x_v^k$

(5.54a) $m_k = \sum x_i^k h_i$

zentrales Moment $a = \bar{x}$

k-tes zentrales Moment

(5.55) $z_k = \frac{1}{n}\sum (x_v - \bar{x})^k$

(5.55a) $z_k = \sum (x_i - \bar{x})^k h_i$

Spezialfälle:

$m_0 = 1$

$m_1 = \bar{x}$

$z_1 = 0$ (Schwerpunkteigenschaft!)

$z_2 = s^2$ (Varianz)

Zusammenhänge zwischen Anfangs- und zentralen Momenten:

$z_2 = m_2 - m_1^2$ (Verschiebungssatz für die Varianz) Analog folgt:

$z_3 = m_3 - 3m_1m_2 + 2(m_1)^3$

$z_4 = m_4 - 4m_1m_3 + 6(m_1)^2 m_2 - 3(m_1)^4$

Kapitel 6: Konzentrations- und Disparitätsmessung

Def. 6.1: Anteile, Merkmalsanteile

1. die Anzahl n (absolute Konzentration) der Merkmalsträger, bzw. die Anteile h_i an der Gesamtheit der Merkmalsträger (relative Konzentration, Disparität)
2. die Merkmalsanteile q_i, d.h. die Anteile an dem Merkmalsbetrag (Summe der Merkmalsbeträge der zu verteilenden Größe).

Übersicht 6.1

Darstellung	(absolute) Konzentration	(relative) Konzentration = Disparität
a) graphisch	Konzentrationskurve	Lorenzkurve
b) Maße		
summarisch	Rosenbluth-Index	Gini-Koeffizient
	Herfindahl-Index	Variationskoeffizient
diskret	concentration ratios	Maximaler Nivellierungssatz

Def. 6.2: Disparitäts- und Gleichheitsmaß

Ist D ein Disparitätsmaß, so ist $G=1-D$ ein Gleichheitsmaß

Lorenzkurve und Gini-Koeffizient bei Einzelbeobachtungen

Def. 6.7: Lorenzkurve, Gini-Koeffizient bei Einzelbeobachtungen

a)Lorenzkurve

Wird der Merkmalsanteil des i-ten Merkmalsträgers bei einer Ordnung nach zunehmender Größe

(6.18) $$q_i = x_{(i)} / \sum x_{(j)} = x_{(i)} / \sum x \qquad i, j=1,2,\ldots,n$$

genannt, dann ist

(6.19) $$Q_i = \sum_{j=1}^{i} q_j$$

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 im H-Q-Kordinatensystem heißt Lorenzkurve.

Für die H_i gilt im Fall von Einzelbeobachtungen: $H_i = i/n$.

b) Gini-Koeffizient

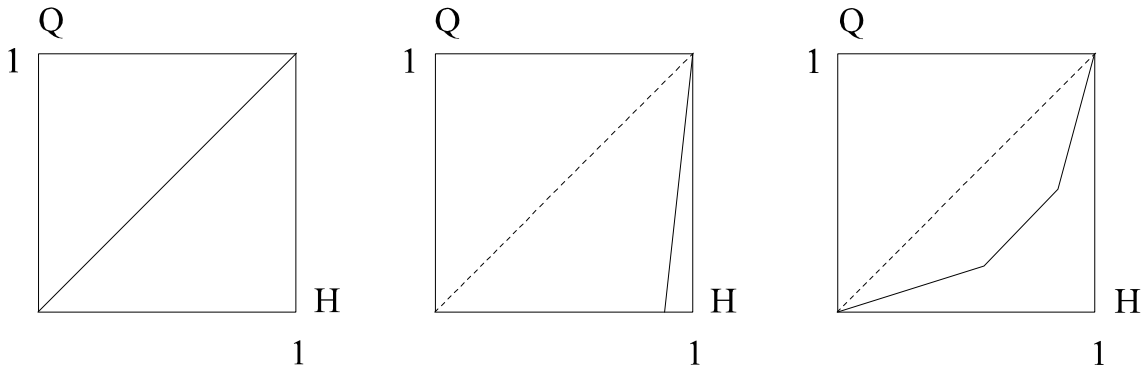
Die Größe

$$(6.20) \qquad D_G = \sum_{i=1}^{n} \frac{2i-n-1}{n} q_i \qquad 0 \le D_G \le 1-1/n$$

heißt Disparitätskoeffizient von Gini (oder einfach Gini-Koeffizient). Zur entsprechenden Formel bei gruppierten und klassierten Daten vgl. Gl. 6.28 und 6.29.

Zusammenhang von q_j und h_j: (6.21) $\frac{q_j}{h_j} = \frac{x_j}{\bar{x}}$.

Abb. 6.6: Lorenzkurve und extreme Fälle von Disparität



Fläche zwischen der Lorenzkurve und der Gleichverteilungsgeraden, F

$$(6.22) \qquad F = \sum_{i=1}^{n} \frac{2i-n-1}{2n} q_i .$$

Daraus folgt, dass der Gini-Koeffizient das Verhältnis zwischen F und der Dreiecksfläche unterhalb der Gleichverteilungsgeraden ist (diese Dreiecksfläche beträgt $1/2$), so dass $D_G = 2F$.

Steigung der Lorenzkurve

$$(6.23) \qquad q_i / h_i = \frac{x_{(i)} / n\bar{x}}{1/n} = \frac{x_i}{\bar{x}}$$

(Steigung der Lorenzkurve, vgl. Gl. 6.21. Danach ist nicht nur die Lorenzkurve sondern auch die Steigung der Lorenzkurve monoton steigend)

Beziehung zwischen H und Q

$H_i \ge Q_i$ ($H_i = Q_i$ gilt außer bei egalitärer Verteilung nur für i=0 und i=n), denn

$$H_i = \frac{i}{n} \ge \frac{i\bar{x}_i}{n\bar{x}} = Q_i \qquad \text{(Lorenzkurve ist monoton steigend)}$$

wobei $\bar{x}_i$ der mittlere Merkmalsbetrag der ersten i Merkmalsträger ist, der wegen der Reihenfolge der Merkmalsträger notwendig stets kleiner ist als der Mittelwert $\bar{x}$, der sich auf alle n Merkmalsträger bezieht.

Gini-Koeffizient bei einer linearen Transformation

Eine Lineartransformation der Merkmalswerte mit $y_i = a + bx_i$ wirkt auf die Merkmalsanteile wie folgt

$$q_i^* = a/n\bar{y} + b(\bar{x}/\bar{y})q_i,$$

so dass man für das Disparitätsmaß von Gini (vgl. Gl. 6.20) erhält

(6.26) $D_G^* = b(\bar{x}/\bar{y})D_G$, da $\sum(2i-n-1) = 2\sum i - n^2 - n$, für $i = 1,2,\ldots,n.$

Bei proportionaler Transformation (Axiom K1) gilt a=0 und wegen $y = bx$ auch $D_G^* = D_G$. Bei einer Niveauänderung im Sinne des Axioms K3 gilt b=1 und folglich

(6.26a) $D_G^* = \dfrac{\bar{x}}{a+\bar{x}} D_G$, so dass bei a > 0 gilt $D_G^* < D_G$.

Ginis "mittlere Differenz" und Ginis Disparitätsmaß

(6.27) $S_G^* = \sum_i \sum_k |x_i - x_k| / n^2$ i,k = 1,2,...,n (siehe Gl. 5.40a)

Zusammenhang zwischen D_G und S_G^*: $D_G = S_G^* / 2\bar{x}$.

Daraus folgt übrigens auch, dass man D_G darstellen kann als

(6.27a) $D_G = \left(\sum_i \sum_k |q_i - q_k| \right) \Big/ 2n$ $i,k = 1,2,\ldots,n.$

<u>Lorenzkurve und Gini-Koeffizient bei gruppierten und klassierten Daten</u>

Def. 6.8: Lorenzkurve bei gruppierten und klassierten Daten

Die lineare Verbindung der Punkte $P_i(H_i, Q_i)$ (i=0,1,...,m) mit $P_0(0,0)$ und $P_m(1,1)$ heißt Lorenzkurve.

Gini-Koeffizient bei gruppierten Daten*)

(6.28) $D_G = 1 - \sum h_i (Q_i + Q_{i-1})$ oder (6.29) $D_G = \sum q_i (H_i + H_{i-1}) - 1$

i = 1,2,...,m ($Q_0 = H_0 = 0$)

*) Wenn bei klassierten Daten die Disparität innerhalb der Klasse berücksichtigt werden soll, vgl. P. von der Lippe, Deskriptive Statistik, UTB Nr. 1632, Gl. 6.29.

Gini-Koeffizient bei zwei Klassen (Spezialfall)

Die Lorenzkurve hat dann nur drei Punkte: $P_0(0,0), P_1(h,q)$ und $P_2(1,1)$. Ginis Dispersions-

maß ist dann der senkrechte Abstand zwischen dem Punkt $P_l(h,q)$ und der Gleichverteilungsgeraden, also die Strecke $D_G = h - q$.

Normiertes Quadrat des Variationskoeffizienten als Disparitätsmaß

(6.30) $NV = V^2/n$ normiertes Quadrat des Variationskoeffizienten V

Axiome für Disparitäts- und Konzentrationsmaße

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. Die ersten drei Axiome (K1 bis K3) gelten für Konzentrations- und Disparitätsmaße in der gleichen Weise, bei den zweiten drei Axiomen (K4 bis K6) unterscheidet sich das Verhalten von Konzentrationsmaßen einerseits und Disparitätsmaßen andererseits.

K1	***Unabhängigkeit von der Maßeinheit:*** Ein Konzentrations- oder Disparitätsmaß 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.
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 $C(y) = \begin{cases} < C(x) & \text{, wenn } a > 0 \text{ (abnehmende Konzentration)} \\ > C(x) & \text{, wenn } a < 0 \text{ (zunehmende Konzentration)} \end{cases}$
K4	***Proportionalitätsprobe, Anzahleffekt:*** 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: $D^* = D$ (Disparität bleibt unverändert) und für das "neue" Konzentrationsmaß K^* im Vergleich zum "alten" $K^* = K/k$ (Fall der Dekonzentration). Entsprechend soll im "umgekehrten" Fall einer Fusion von k gleich großen Einheiten zu einer Einheit gelten $D^* = D$ und $K^* = kK$.
K5	***Ergänzungsprobe, Nullergänzung, Disparitätseffekt:*** Fügt man einer Verteilung m Einheiten, deren Merkmalsbeträge jeweils Null sind ("Nullträger") hinzu, so soll gelten $K^* = K$ und $D^* > D$.
K6	***Wertebereiche:*** Als Wertebereiche sollen $1/n \le K \le 1$ (für Konzentrationsmaße) und $0 \le D \le 1 - 1/n$ (für Disparitätsmaße) gelten.

Kapitel 7: Zweidimensionale Häufigkeitsverteilungen

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 $x_1,x_2,...,x_m$ 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 Merkmalsausprägungen 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 und j=1,...,k).

Für die relativen Häufigkeiten gilt analog zur eindimensionalen Häufigkeitsverteilung

(7.2) $h_{ij} = n_{ij}/n$ mit $n = \sum_i \sum_j n_{ij} = \sum_{i,j} n_{ij}$.

c) bei klassierten Daten gilt b) analog.

Def. 7.2: Zweidimensionale Häufigkeitsverteilung (joint distribution)

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_{1j}	...	h_{1k}	Randverteilung von X
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}	
	Randverteilung von y						

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 der Merkmale X und Y (klassierte Daten) bezeichnen.

Verteilungen		
eine zweidimensionale gemeinsame Verteilung h_{ij} (auch kumulierte Verteilung H_{ij}) von x_i,y_j	eindimensionale Verteilungen	
	zwei Randverteilungen Def. (7.3)	m+k bedingte Verteilungen Def. (7.4)
Beschreibende Kennzahlen		
Kovarianz Def.(7.7) Korrelationskoeffizient Def.(7.8)	Mittelwerte $\bar{x}, \bar{y}$ und Varianzen der Randverteilungen	Bedingte Mittelwerte, Regressionslinie Def.(7.6)

Def. 7.3: Randverteilungen (marginal distribution)

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

$$(7.4) \qquad h_{i\cdot} = \sum_{j=1}^{k} h_{ij} = h\left(X = x_i\right).$$

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:

$$(7.5) \qquad h_{\cdot j} = \sum_{i=1}^{m} h_{ij} = h\left(Y = y_j\right).$$

Die Randverteilungen ausgedrückt in absoluten Häufigkeiten $n_x(x)$ mit den über k Spalten summierten absoluten Häufigkeiten einer Zeile

$$(7.4a) n_{i.} = n_{i1} + n_{i2} + ... + n_{ik}$$

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						Summe
	y_1	y_2	...	y_j	...	y_k	$h_x(x)$
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}	$h_{m.}$
$h_y(y)$	$h_{.1}$	$h_{.2}$	...	$h_{.j}$	...	$h_{.k}$	1

Die Summenspalte $h_x(x)$ ist die Randverteilung von X und die Summenzeile $h_y(y)$ ist die Randverteilung von Y.

Def. 7.4: bedingte Verteilung (conditional distribution)

Die durch Gl. 7.6 definierten bedingten relativen Häufigkeiten $h_{i|j}$ stellen die bedingte Häufigkeitsfunktion (-verteilung) von X, gegeben $Y = y_j$ dar

(7.6) $$h_{i|j} = \frac{h_{ij}}{h_{\cdot j}} = \frac{n_{ij}}{n_{\cdot j}} = h\left(x \mid Y = y_j\right).$$

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) $$h_{j|i} = \frac{h_{ij}}{h_{i\cdot}} = \frac{n_{ij}}{n_{i\cdot}} = h\left(y \mid X = x_i\right).$$

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, so sind X und Y unabhängig (gleichzeitig gilt: Gleichheit der m bedingten Verteilungen $h_{j|i}$ des Merkmals Y also Unabhängigkeit von X und Y, [Unabhängigkeit ist eine symmetrische Relation]).
2. Im Falle der Unabhängigkeit ergeben sich die absoluten, bzw. relativen gemeinsamen Häufigkeiten aus den entsprechenden Häufigkeiten der Randverteilungen gem.

(7.8) $$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$ bzw. (7.8a) $$h_{ij} = h_{i\cdot} h_{\cdot j}.$$

Unabhängigkeit impliziert Unkorreliertheit aber nicht umgekehrt, d.h. Unkorreliertheit kann bestehen, obgleich die Variablen X und Y nicht unabhängig sind.

Mittelwert und Varianz der Randverteilungen

Mittelwert $\bar{x}$ der Randverteilung $h_x(x)$

(7.9) $$\bar{x} = \sum_i x_i h_{i\cdot} = \sum_i \sum_j x_i h_{ij}$$

und die Varianz

(7.10) $$s_x^2 = \sum_i x_i^2 h_{i\cdot} - \bar{x}^2.$$

Die entsprechenden Parameter der Randverteilung $h_y(y)$ sind analog definiert.

Parameter der bedingten Verteilungen

a) Die wichtigsten Parameter der bedingten (Häufigkeits-) Verteilungen sind die bedingten Mittelwerte

(7.11) $$\bar{x} \mid y = \bar{x}\left(y_j\right) = \sum_{i=1}^{m} x_i h_{i|j}$$

(7.12) $$\bar{y} \mid x = \bar{y}\left(x_i\right) = \sum_{j=1}^{k} y_j h_{j|i}$$

b) Seltener ist die Berechnung der bedingten Varianzen (notwendig zur Berechnung des Korrelationsverhältnisses)

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.

Def. 7.7: Kovarianz

Die Kovarianz ist als beschreibende Kennzahl einer zweidimensionalen Verteilung definiert als

(7.13) $$s_{xy} = \frac{1}{n}\sum_{v=1}^{n}(x_v - \bar{x})(y_v - \bar{y})$$ bei n Einzelbeobachtungen

bzw. bei gruppierten Daten

(7.14) $$s_{xy} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})n_{ij}$$ mit absoluten Häufigkeiten

(7.14a) $$s_{xy} = \sum_{i=1}^{m}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})h_{ij}$$ mit relativen Häufigkeiten

Kovarianz bei Lineartransformation

(7.15) $s_{x^*y^*} = bds_{xy}$,wenn x*=a+bx und y*=c+dy

Verschiebungssatz für die Kovarianz

Auch für die Kovarianz gilt der Verschiebungssatz:

(7.13a) $$s_{xy} = \frac{1}{n}\sum_{v} x_v y_v - \bar{x}\,\bar{y}$$

bzw. bei gruppierten Daten

(7.14a) $$s_{xy} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k} x_i y_j n_{ij} - \bar{x}\,\bar{y}$$ mit absoluten Häufigkeiten

(7.14b) $$s_{xy} = \sum_{i=1}^{m}\sum_{j=1}^{k} x_i y_j h_{ij} - \bar{x}\,\bar{y}$$ mit relativen Häufigkeiten

oder: $s_{xy} = \overline{xy} - \bar{x}\,\bar{y}$

Hierin ist $\overline{xy}$ der Mittelwert des Produkts der x und y Werte und $\bar{x}\,\bar{y}$ ist das Produkt der Mittelwerte.
Die damit gegebene Beziehung zwischen dem Anfangsproduktmoment $\overline{xy}$ und dem zentralen Produktmoment s_{xy} führt auch wegen der Schwerpunkteigenschaft des arithmetischen Mittels zu folgenden Darstellungen der Kovarianz:

(7.17) $$s_{xy} = \frac{1}{n}\sum_{v}(x_v - \bar{x})y_v = \frac{1}{n}\sum_{v}(y_v - \bar{y})x_v\ .$$

Satz 7.2:

Verschwindet eine der Varianzen (etwa $s_x^2 = 0$), so ist auch die Kovarianz null. 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 mit einer Konstanten k ist stets Null, also $s_{xk} = 0$ oder $s_{yk} = 0$

Satz 7.3: Schwarz'sche Ungleichung

(7.18) $0 \le (s_{xy})^2 \le s_x^2 s_y^2$

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) $r_{xy} = s_{xy}/s_x s_y$

(7.20a) $-1 \le r_{xy} \le +1$ (wegen 7.18).

Somit ist r_{xy} die auf den Wertebereich von -1 bis +1 normierte Kovarianz s_{xy} (während s_{xy} nicht beschränkt ist).

Def. 7.9: Scheinkorrelation, spurious correlation

Sind zwei Variablen X und Y hoch miteinander korreliert, weil sie gemeinsam abhängig sind von einer dritten Variablen Z, so spricht man von Scheinkorrelation.

Kapitel 8: Regressionsanalyse

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 eine 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 $X_1, X_2, X_3, ..., X_p$ gibt.

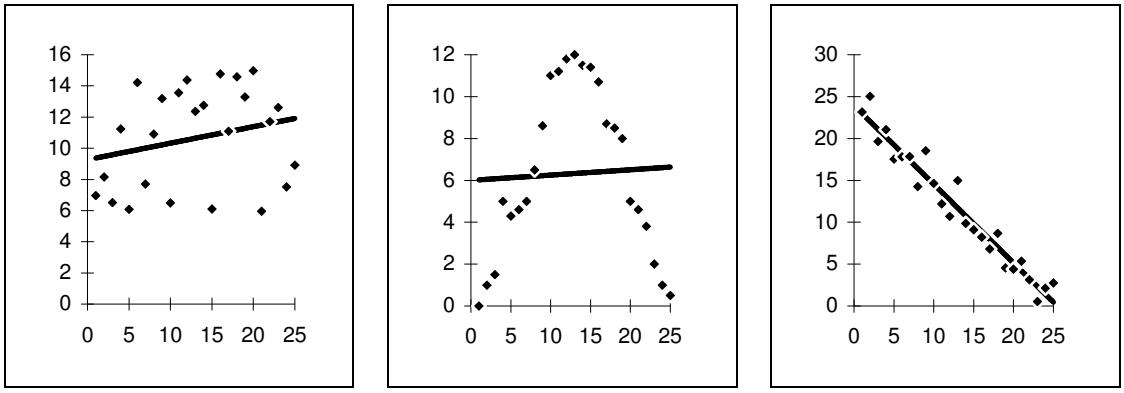
c) lineare und nichtlineare Regression:

Eine Regressionsfunktion ist linear (in den Variablen und in den Parametern), wenn gilt:

$\hat{y}_v = a + bx$ [a und b heißen Regressionskoeffizienten] (einfache lineare Regression) oder

$\hat{y}_v = b_0 + b_1 x_{1v} + b_2 x_{2v} + \ldots + b_p x_{pv}$ (multiple lineare Regression, p Regressoren), andernfalls ist sie nichtlinear.

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 einen relativ 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.

Def. 8.2: Regressionsgerade

a) Die lineare Regressionsfunktion (Regressionsgerade) zur Bestimmung von Y (abhängige Variable) durch X (unabhängige Variable) lautet:

(8.1) $\hat{y}_v = a + bx$

dabei ist $\hat{y}_v$ der Regresswert für die v-te Beobachtung (Einheit) mit v = 1,2,...,n und für die einzelne Beobachtung (x_v,y_v) gilt:

(8.1a) $y_v = \hat{y}_v + u_v = a + bx_v + u_v$,

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) $\hat{x}_v = c + dy_v$ zu berechnen, wobei für x_v gilt:

(8.2a) $x_v = c + dy_v + v_v$.

Die Störgröße V ist jeweils der waagrechte Abstand zwischen einem Beobachtungspunkt

x_v und $\hat{x}_v$ im x,y-Koordinatensystem.

Schätzung der Koeffizienten bei der linearen, einfachen Regression

Nach der "Methode der kleinsten Quadrate" erhält man die Normalgleichungen:

(8.4a)	$an + b\sum x_v = \sum y_v$	1. Normalgleichung
(8.4b)	$a\sum x_v + b\sum x_v^2 = \sum x_v y_v$	2. Normalgleichung

Wird dieses Normalgleichungssystem nach a und b aufgelöst so erhält man als Schätzwerte zur Bestimmung der Regressionskoeffizienten a und b:

(8.5a) $$a = \frac{\sum x_v^2 \sum y_v - \sum x_v \sum x_v y_v}{n\sum x_v^2 - \left(\sum x_v\right)^2}$$

(8.6a) $$b = \frac{\sum (x_v - \bar{x})(y_v - \bar{y})}{\sum (x_v - \bar{x})^2} = \frac{s_{xy}}{s_x^2}\,.$$

Wie man leicht sieht, gilt aufgrund der ersten Normalgleichung:

(8.6b) $$a = \bar{y} - b\bar{x}$$

Man erhält die entsprechenden Formeln zur Bestimmung von c und d indem man in den Normalgleichungen bzw. in den Formeln für a und b x und y vertauscht.

Korrelationskoeffizienten r_{xy}

(8.7) $$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{s_{xy}}{s_x s_y} = \begin{cases} \sqrt{b \cdot d} & \text{wenn } b, d > 0 \\ -\sqrt{b \cdot d} & \text{wenn } b, d < 0 \end{cases}$$

Varianzzerlegung

(8.8)	$\frac{1}{n}\sum (y_i - \bar{y})^2$	=	$\frac{1}{n}\sum (\hat{y}_i - \bar{y})^2$	+	$\frac{1}{n}\sum (y_i - \hat{y}_i)^2$
	totale Varianz	=	**erklärte Varianz**	+	**Residualvarianz**
	s_y^2	=	$s_{\hat{y}}^2$	+	s_u^2

Bestimmtheitsmaß B_{yx} und Unbestimmtheitsmaß U_{yx}

$$B_{xy} = \frac{\text{erklärte Varianz}}{\text{totale Varianz}} = \frac{S_{\hat{y}}^{\,2}}{S_y^{\,2}} \qquad 0 \le B_{yx} \le 1$$

$$U_{yx} = \frac{\text{Residualvarianz}}{\text{totale Varianz}} = \frac{S_u^{\,2}}{S_y^{\,2}} = 1 - B_{yx} \quad 0 \le U_{xy} \le 1$$

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 $\left(B_{yx} = r_{xy}^2\right)$.

(8.9) $$B_{yx} = \frac{s_{xy}^2}{s_x^2 s_y^2} = b \cdot d = r_{xy}^2$$

Die mit x erklärte Varianz ist $s_{\hat{x}}^2 = d^2 s_y^2$ so dass $B_{xy} = d^2 s_y{}^2 / s_x{}^2 = s_{xy}{}^2 / s_x{}^2 s_y{}^2 = b \cdot d = r_{xy}{}^2 = B_{yx}$.

Man kann zeigen:

1. Für den Winkel α zwischen den Regressionsgeraden gilt:

(8.10) $$\tan(\alpha) = \frac{s_{xy}\left(1 - r^2\right)}{r^2\left(s_x^2 + s_y^2\right)}$$

2. Die Steigung der Regressionsgeraden $\hat{x}$ im x,y-Koordinatensystem ist betragsmäßig stets größer ist als die Steigung b der

Eigenschaften der KQ-Schätzung

(8.11) $\Sigma u_v = \bar{u} = 0$, die geschätzte Regressionsgerade verläuft durch den Schwerpunkt

(8.12) $\Sigma y_v = \Sigma \hat{y}_v$ und somit $\bar{y} = \bar{\hat{y}}$.

(8.13) $\Sigma x_v u_v = 0$ und $s_{ux} = r_{ux} = 0$.

(8.14) $\Sigma y_v u_v = r_{\hat{y}u} = 0$.

(8.15) $\Sigma y_v u_v = \Sigma u_v^2$ und $s_u^2 = (\Sigma u_v^2)/n = s_{uy}$. Hieraus folgt $r_{yu} = s_u/s_y$ und damit auch

(8.16) $(r_{yu})^2 = s_u^2 / s_y^2 = 1 - (r_{xy})^2 = U_{xy}$

(8.17) $r_{xy} = r_{\hat{y}y}$.

Kapitel 9: Verhältniszahlen, Wachstumsraten und Aggregation

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 . Mit dem Umfang n_i der i-ten Teilgesamtheit und n der Gesamtheit bzw. den Merkmalssummen S_i und S ist eine Gliederungszahl

(9.1) $$G_i = \frac{n_i}{n} \quad \text{oder} \quad G_i = \frac{S_i}{S}$$

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.

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.

Eigenschaften von Gliederungszahlen

Es ergibt sich als unmittelbare Folgerung aus Gl. 9.1:

(9.2) $0 \le G_i \le 1$ und (9.3) $\Sigma G_i = 1 \quad (i = 1,2,...,m)$.

Eigenschaften von Beziehungszahlen

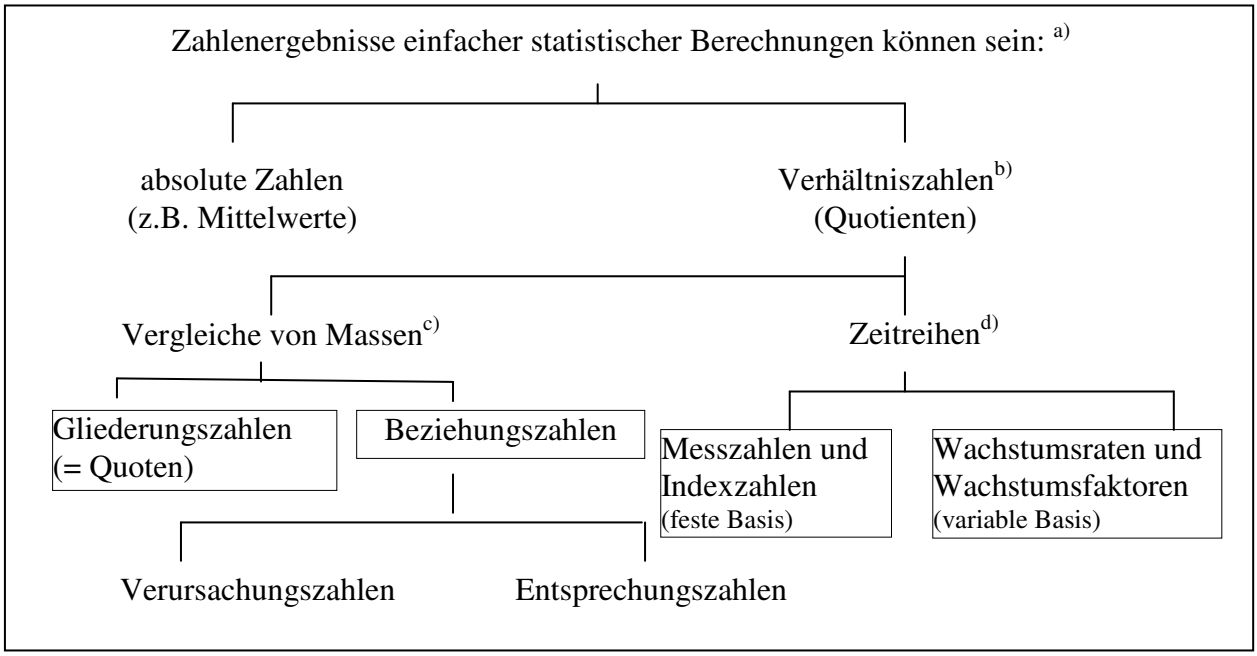
1. Dimension: Anders als Gliederungs- und Messzahlen haben Beziehungszahlen meist eine Maßeinheit.
2. Umkehrbarkeit: Beziehungszahlen sind grundsätzlich umkehrbar.
3. Zusammenhang mit Mittelwerten: Beziehungszahlen sind nichts anderes als Mittelwerte wenn eine Merkmalssumme (Zähler) zu einer entsprechend abgegrenzten Personengesamtheit (Nenner) ins Verhältnis gesetzt wird.

Alle Verhältniszahlen sind ferner Mittelwerte durch Aggregation, 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).

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).

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.

Messzahlen

Die Messzahl m_{0t} (d.h. zur Basis 0, Berichtszeit t) einer Variablen Y ist nach Def. 9.1 die Größe:

(9.6) $\quad m_{0t} = y_t/y_0,$

bei der diskreten Zeitvariable t = 0,1,2,...,T bzw. die mit 100 multiplizierten Größen

(9.6a) $\quad m^*_{0t} = 100m_{0t} = 100y_t/y_0.$

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.

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	$m(\alpha)_{0t} = \alpha y_t/\alpha y_0 = m_{0t} = y_t/y_0$ Unabhängigkeit von der Maßeinheit der Messwerte
Zeitumkehrbarkeit (Reversibilität)	$m_{t0} = m^{-1}_{ot}$ Vertauschung von Basis- und Berichtsperiode ($m_{t0}m_{t0} = 1$)
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 $m^W_{0t} = m^P_{0t} \cdot m^{Q\,*)}_{0t}$

*) eine Wertmesszahl ist das Produkt aus Preis- und Mengenmesszahl.

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_{st} 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: $m_{st} = m_{0t}/m_{0s}$
Verkettung: $m_{0t} = m_{0s} \cdot m_{st}$

b) Messzahlen m^*_{0t}, m^*_{st} (mit 100 multiplizierte Messzahlen):

Umbasierung: $m^*_{st} = \left(m^*_{0t} / m^*_{0s}\right) \cdot 100$

Verkettung: $m^*_{0t} = \left(m^*_{0s} \cdot m^*_{st}\right) / 100$

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)	$r_t = (y_t - y_{t-1})/y_{t-1} = w_t - 1$	(r_t: Wachstumsrate)
(9.8)	$w_t = y_t/y_{t-1} = r_t + 1$	(w_t: Wachstumsfaktor)

b) Für ein Wachstum mit konstanter Wachstumsrate [z.B. Verzinsung mit Zinseszins] r (r_t = r für alle t) gilt:

(9.9) $y_t = y_o \cdot w^t = y_o \cdot (1+r)^t$ (Wachstum mit konstanter Rate r).

Bei variierenden Wachstumsraten r_t lautet die Wachstumsgleichung:

(9.10) $y_T = y_0(1+r_1)(1+r_2)...(1+r_T) = y_0 \prod_{t=1}^{T}(1+r_t) = y_0 \prod_{t=1}^{T} w_t$.

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) $r = (w_1 w_2 ... w_T)^{1/T} - 1 = \left[\prod_{t=1}^{T} w_t\right]^{1/T} - 1$.

Mittlere Wachstumsrate

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*.

(9.11) $r = (y_t/y_0)^{1/t} - 1$ (mittlere Wachstumsrate),

bzw. in Prozent:

(9.11a) $r = [(y_t/y_0)^{1/t} - 1] \cdot 100$.

Def. 9.4: Wachstumsrate bei stetiger Zeit

a) Die Wachstumsrate r(t) einer stetigen Funktion y = y(t) ist

(9.14) $r(t) = \frac{y'(t)}{y(t)} = \frac{dy/dt}{y} = \frac{d\ln(y)}{dt}$.

b) Bei konstanter Wachstumsrate $r(t) = \alpha$ (für jeden Wert von t) ist die stetige Zeitreihe y(t) gegeben mit

(9.15) $y(t) = y(0)e^{\alpha t} = y(0)\exp(\alpha t)$.

Beziehung zwischen den Wachstumsraten α (stetige Zeit) und r (diskrete Zeit)

(9.18) $e^{\alpha} = w = 1+r$, so dass gilt (9.19) $\alpha = \ln(1+r)$.

Man erhält somit im Zusammenhang mit der Reihenentwicklung von e^{α} und ln(1+r) die folgenden Umrechnungen

(9.20) $$r = e^{\alpha} - 1 = \alpha + \frac{\alpha^2}{2!} + \frac{\alpha^3}{3!} + \frac{\alpha^4}{4!} + \ldots$$

für die Umrechnung von α in r (so dass $\alpha < r$) und

(9.21) $$\alpha = \ln(1+r) = r - \frac{r^2}{2!} + \frac{r^3}{3!} + \frac{r^4}{4!} + \ldots$$

für die Umrechnung von r nach α.

Wie man sieht gilt nur bei kleinen Wachstumsraten $r \approx \alpha$.

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 = x/y	$w_z = w_x/w_y$	$r_z(t) = r_x(t) - r_y(t)$
Kehrwert z = 1/y	$w_z = 1/w_y$	$r_z(t) = -r_y(t)$

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=1,2,...,J)

(9.22) $Q = \Sigma Q_j g_{yj}$.

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) $$Q_A^* = \sum Q_{Aj} g_j^* \quad \text{mit} \quad Q_B^* = \sum Q_{Bj} g_j^* ,$$

d.h. man vergleicht Beziehungszahlen, die unter Zugrundelegung der gleichen Gewichte (Standardgewichte) g_j^* berechnet sind. Die Größen Q^* heißen dann standardisierte Beziehungszahlen.

Übersicht 9.5.: Wachstumsraten ausgewählter Funktionen

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).

	Funktion y(t)	Ableitung	Wachstumsrate
1 1a 1b 1c 1d	$(a+bt)^{\alpha}$ (2) a=1: Gerade y = a+bt a=-1: 1/(a+bt) a=1/2: $\sqrt{a+bt}$ Potenzfunktion bt^{α}	$\alpha b(a+bt)^{\alpha-1}$ b $-b(a+bt)^{-2}$ $1/2b(a+bt)^{-1/2}$ $b\alpha t^{\alpha-1}$	$ab(a+bt)^{-1}$ $b/(a+bt) = r_G$ $-b/(a+bt) = -r_G$ $b/2(a+bt) = 1/2r_G$ α/t (hyperbolisch)
2	Parabel(3) $a+bt+ct^2$	$b+2ct$	$(b+2ct)/(a+bt+ct^2)$
3 3a 3b	$a\cdot\exp(bt^{\alpha})$ $\alpha=1$: ae^{bt} oder: ar^t mit $r=e^b$ a = -1: $ae^{b/t}$	$y\cdot\alpha bt^{\alpha-1}$ yb $y\cdot\ln(r)$ yb/t^2	$\alpha bt^{\alpha-1}$ b b = ln(r) $-b/t^2$
4	$k+be^{ct}$ oder $y=k+br^t$ mit $r=e^c$ (k: Sättigungsniveau)	cbe^{ct} strebt gegen 0 wenn r < 1, c=lnr < 0	$(-c)[(k-y)/y] = -cR$ speziell: c = -1, dann r(t) = R;
5	k + b/(c+t) (Hyperbel)	$-b/(c+t)^2$	$-b/[k(c+t)^2+b(c+t)]$
6	k(t+a)/(t+b) (b>a)	$k(b-a)/(t+b)^2$	$(b-a)/(t+a)(t+b)$
7	$\exp(K+br^t)$ mit $r=e^c$ oder: $\ln(y) = K+be^{ct}$ $k=e^K$ Sättigungsniveau	$ya/(b+t)^2$	$a/(b+t)^2$
8	$k/(1+e^{a-bt})$ a,b,k>0 k: Sättigungsniveau	$by(k-y)/k$	$b+\beta y$ ($\beta = -b/k$) (4)
9	$\ln(y)_K = K - a/(b+t)$ $k=e^K$ Sättigungsniveau	$ya/(b+t)^2$	$a/(b+t)^2$

(1) $r(t) = y'/y$

(2) In dieser allgemeinen Form liegt eine Polynomfunktion vom Grade α vor, wobei $\alpha \in Z^+$ ist. Bei $\alpha=1/2$ liegt eine Wurzelfunktion und bei ganzzahligem α und a > 0 eine Potenzfunktion vor (Fall 1d).

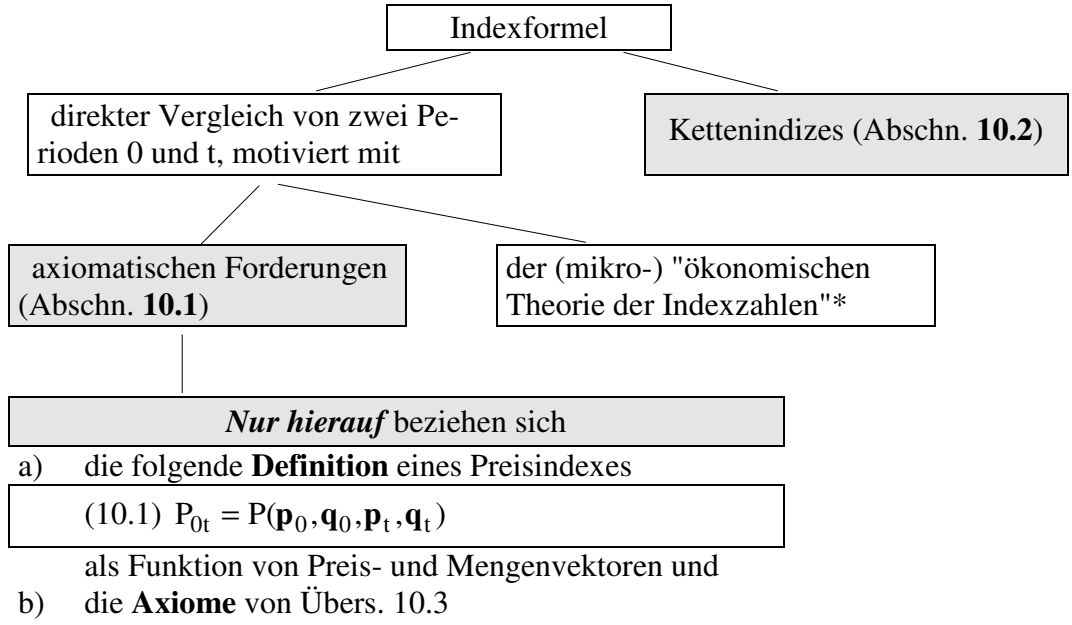
(3) Kann entsprechend verallgemeinert werden wie Funktion Nr. 1.

(4) Kennzeichnend für die logistische Funktion: r(t) = f[y(t)] (f: linear).

Kapitel 10: Indexzahlen

Indexzahlen sind Maßzahlen (beschreibende Kennzahlen) für den Vergleich einer Gesamtheit von Erscheinungen. Indizes sind Maße der aggregierten Veränderung, z.B. ein Preisindex ist i.d.R. ein summarisches (zusammengefasstes) Maß von Preisveränderungen (im zeitlichen Vergleich)[1], etwa ein Mittelwert von Preis- Messzahlen für i = 1,2,...,n Waren.

10.1 Prinzipien der Konstruktion von Indexformeln



* wird hier nicht behandelt

10.1 Direkte Indexformeln

a) Vorläufer (historische ungewogene Indizes)

Mit den ungewogenen arithmetischen Mitteln der einzelnen Preise $\overline{p}_{t=} \sum p_{it}/n$ und $\overline{p}_0$ entsprechend erhält man die Preisindexformel

$$(10.2) \qquad P_{0t}^{D} = \frac{\overline{p}_t}{\overline{p}_0} = \frac{\sum p_{it}}{\sum p_{i0}} \qquad \text{von (D =) } \textbf{Dutot}$$

eine Verhältniszahl (Messzahl) von Mittelwerten: sie ist nicht sinnvoll, weil sie nicht die Kommensurabilität, Axiom P5 in Übers. 10.3 erfüllt. Es gilt

> eine Messzahl von Mittelwerten erfüllt nicht Minimalforderungen an Indizes (z.B. Axiom P5), wohl aber ein Mittelwert von Messzahlen.

Ein ungewogenes arithmetisches Mittel von Preismesszahlen ist die Preisindexformel

$$(10.3) \qquad P_{0t}^{C} = \frac{1}{n} \sum \frac{p_{it}}{p_{i0}} \text{ von (C =) } \textbf{Carli}.$$

[1] zumindest direkte Indizes (vgl. Übers. 10.1) werden auch für den interregionalen (z.B. internationalen) Vergleich benutzt.

Ein ungewogenes geometrisches Mittel von Preismesszahlen ist die Formel von **Jevons**

$$P_{0t}^{J} = \left[\frac{p_{1t}}{p_{10}} \cdot \frac{p_{2t}}{p_{20}} \cdot \ \cdot \frac{p_{nt}}{p_{n0}} \right]^{\frac{1}{n}}.$$

b) Aktuelle Indexformeln (insbes. Laspeyres und Paasche)

Preisindizes nach Laspeyres und Paasche haben eine doppelte Interpretation, als

- gewogenes Mittel von Preismesszahlen (**Messzahlenmittelwertformel**) und als
- Verhältnis von Ausgaben- bzw. Einnahmenaggregate (**Aggregatformel**)[2].

Zur Vereinfachung ist im Folgenden das Subskript i (Warenart) weggelassen worden. Die Vektorschreibweise zeigt, dass die Indizes *lineare* Indizes sind (Preisindizes linear in den Preisen, Mengenindizes linear in den Mengen):

Formel von	**Messzahlenmittelwertformel**	**Aggregatformel**
Laspeyres (L)	(10.4) $P_{0t}^{L} = \sum \frac{p_t}{p_0} \frac{p_0 q_0}{\sum p_0 q_0}$ gewogenes **arithmetisches** Mittel Gewichte: Ausgabenanteile zur **Basis**zeit	(10.5) $P_{0t}^{L} = \frac{\sum p_t q_0}{\sum p_0 q_0} = \frac{\mathbf{p}_t' \mathbf{q}_0}{\mathbf{p}_0' \mathbf{q}_0}$ Zähler: fiktives Ausgabenaggregat Nenner: tatsächliche Ausgaben
Paasche (P)	(10.6) $P_{0t}^{P} = \sum \frac{p_t}{p_0} \frac{p_0 q_t}{\sum p_0 q_t}$ oder: gewogenes **harmonisches** Mittel, Gewichte: Ausgabenanteile zur **Berichts**zeit	(10.7) $P_{0t}^{P} = \frac{\sum p_t q_t}{\sum p_0 q_t} = \frac{\mathbf{p}_t' \mathbf{q}_t}{\mathbf{p}_0' \mathbf{q}_t}$ Zähler: tatsächliche Ausgaben, Nenner: fiktives Ausgabenaggregat

In manchen Lehrbüchern (nicht in der Praxis) spielt auch der Preis- (oder gar der Mengen-) index nach **Lowe** eine gewisse Rolle. Ein solcher Index erfüllt jedoch nicht die Axiome von Übers 10.3. Kein Index, der Durchschnittspreise verwendet, kann kommensurabel sein. Schon wegen der Unmöglichkeit, kg-, Liter-, Stück-Mengen usw. zu einer "Gesamtmenge" zu addieren, sind Durchschnittsmengen auch meist gar nicht definiert.

Wertindex (z.B. Lebenshaltungskostenindex im Unterschied zum Preisindex für die Lebenshaltung nach Laspeyres)

(10.8) $$W_{0t} = \frac{\sum p_{it} q_{it}}{\sum p_{i0} q_{i0}} \quad \text{oder einfach} \quad W_{0t} = \frac{\sum p_t q_0}{\sum p_0 q_0}$$

Mengenindizes gewinnt man aus Preisindizes durch Vertauschen von Mengen und Preisen (vgl. Übers. 10.2):

(10.9) $$Q_{0t}^{L} = \frac{\sum q_t p_0}{\sum q_0 p_0}$$ (**Laspeyres**) und (10.10) $$Q_{0t}^{P} = \frac{\sum q_t p_t}{\sum q_0 p_t}$$ (**Paasche**)

Wertindex als Indexprodukt (10.11) $W_{0t} = P^L Q^P = P^P Q^L$ (Produkttest)[3]

[2] Der berühmte "Idealindex" von I. Fisher oder auch Kettenindizes aller Art besitzen keine der beiden Interpretationen.

[3] Das Indexpaar Laspeyres-Paasche erfüllt den Produkttest, nicht jedoch die anspruchsvollere (und in ihrer Bedeutung meist völlig überschätzte) Faktorumkehrbarkeit.

10.2 Übersicht über die Indexformeln

Preise p, Mengen q, Subskripte t = Berichtszeit, 0 = Basiszeit, Summierung über alle n Waren

Wertindex $W_{0t} = \frac{\sum p_t q_0}{\sum p_0 q_0}$

Laspeyres Preisindex $P_{0t}^L = \frac{\sum p_t q_0}{\sum p_0 q_0}$

Verwendung für: spezielle Preisniveaus (z.B. Preisindizes für die Lebenshaltung)

Paasche Preisindex $P_{0t}^P = \frac{\sum p_t q_t}{\sum p_0 q_t}$

Verwendung: 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 $Q_{0t}^L = \frac{\sum q_t p_0}{\sum q_0 p_0}$

Paasche Mengenindex $Q_{0t}^P = \frac{\sum q_t p_t}{\sum q_0 p_t}$

Es gilt die grundlegende Formel als Basis für die Preisbereinigung:

(10.11) $W_{0t} = P_{0t}^L Q_{0t}^P = P_{0t}^P Q_{0t}^L$

Preisbereinigung (Deflationierung; auch Realwert- oder Volumenrechnung genannt).

aus einem	ist zu errechnen ein	Vorgehensweise
Wert $= \sum p_t q_t$ (einer ***nominalen*** Größe, zu ***jeweiligen*** Preisen)	***Volumen*** $= \sum p_0 q_t$ (eine ***reale*** Größe, zu ***konstanten*** Preisen des Basisjahres)	Division durch einen[1)] Paasche Preisindex $V_t = \frac{W_t}{P_{0t}^P}$
Wertindex $W_{0t} = \frac{\sum p_t q_0}{\sum p_0 q_0}$	Laspeyres-Mengenindex* $Q_{0t}^L = \frac{\sum q_t p_0}{\sum q_0 p_0}$	Division durch einen[1)] Paasche Preisindex $Q_{ot}^L = \frac{W_{0t}}{P_{0t}^P}$

1) sich auf das gleiche Aggregat beziehende
2) als Maß für die Veränderung von Volumen

Strukturelle Konsistenz (der Deflationierung)

Gilt für nominale Teilaggregate $W = W_1 + W_2 + ... + W_m$ und soll dann für die realen Teilaggreagte $V_j = W_j/P_j$ (j = 1,2,...,m) gelten

$$\frac{W}{P} = V = \frac{W_1}{P_1} + ... + \frac{W_m}{P_m} = V_1 + ... + V_m, \text{ dann (10.12)} \quad P^{-1} = \sum \frac{W_j}{W} P_j^{-1}$$

d.h. dann muss der Gesamt-Deflator P ein harmonisches Mittel der m Teil-Deflatoren sein (Gewichte $W_j/\Sigma W_j = W_j/W$), also ein direkter Paasche Preisindex. Deflationierung mit einem

anderen Index als P^P liefert strukturell inkonsistente Ergebnisse (Volumen addieren sich nicht in gleicher Weise wie Werte, das Ergebnis der Deflationierung ist abhängig vom Aggregationsgrad).

Additive Konsistenz (der Indexformel)

Wenn ein Gesamtaggregat zu Teilaggregaten j = 1, ... , m zerlegt werden kann, dann soll sich der Gesamtindex aus den Teilindizes in der gleichen Weise zusammensetzen, wie der Gesamtindex aus den Messzahlen. Im Falle von P^L gilt z.B. der folgende Zusammenhang

$$(10.13) \quad P^L_{0t} = \sum_j \frac{W_{j,o}}{\sum W_{j,o}} P^L_{j,0t}$$

d.h. der Gesamtindex P^L_{0t} ist das arithmetisches Mittel der m Teil-Indizes $P^L_{j,0t}$ mit den Wertanteilen zur Basiszeit als Gewichte. Für den Paasche Preisindex P^P gilt Gl. 10.12. Lineare (=additive) Indizes (vgl. Übers. 10.3) sind additiv konsistent. Die Umkehrung gilt nicht.

Formel von Ladislaus v. Bortkiewicz
(Größenrelation zwischen Laspeyres- und Paasche-Preisindex)

Die Kovarianz von Preis- (b_i) und Mengenmesszahlen (c_i) mit den Gewichten g_i (Ausgabenanteile zur Basiszeit) lautet: $C = \sum (b_i - P^L)(c_i - Q^L) g_i = Q^L (P^P - P^L)$.

Daraus folgt $W = P^L Q^L + C$, denn mit $g_i = \frac{p_{i0} q_{i0}}{\sum p_{i0} q_{i0}}$, $b_i = \frac{p_{it}}{p_{i0}}$, und $c_i = \frac{q_{it}}{q_{i0}}$ gilt

$$P^L_{0t} = \sum g_i b_i \quad \text{und} \quad Q^L_{0t} = \sum g_i c_i \quad \text{sowie} \quad W_{0t} = \sum b_i c_i g_i = P^L Q^P = P^P Q^L$$

Dann gilt für die Kovarianz

$$(10.14) \quad C = Q^L(P^P - P^L) = P^L(Q^P - Q^L) \text{ also}$$

wenn negative Kovarianz $C < 0$ dann $P^L > P^P$ und $Q^L > Q^P$
wenn positive Kovarianz $C > 0$ dann $P^L < P^P$ und $Q^L < Q^P$

c) Einige Axiome und ein Axiomensystem (von Eichhorn/Voeller)

Zu einigen fundamentalen Forderungen an sinnvolle Indexformeln (Index-"axiome") vgl. Übers. 10.3. Wichtige Axiome, die erst in neuerer Zeit mehr bachtet werden sind ferner die Aggregationseigenschaften, wie z.B. strukturelle - und additive Konsistenz (s.o.).

Eine große Rolle spielen jedoch auch immer noch Axiome (oder "Proben", "Tests"), die aus der Indexphilosophie von Irving Fisher stammen wie:

Zeitumkehrbarkeit (Z)

Vertauschung von Basis- und Berichtsperiode führt zum reziproken Preisindex

(10.14) $P_{0t} P_{t0} = 1$ (Zeitumkehrprobe). Nicht erfüllt vom Paar Laspeyres/Pasasche, denn

$$P^L_{0t} P^P_{t0} = P^P_{0t} P^L_{t0} = 1 \, .$$

Faktorumkehrprobe (F)

Die Wertsteigerung kann in das Produkt einer *nach der gleichen Indexformel* berechenten Preis- und Mengenkomponente zerlegt werden. Fisher's "Idealindex"

(10.15) $$P^F_{0t} = \sqrt{P^L_{0t} P^P_{0t}}$$

das geometrische Mittel aus der Laspeyres- und Paasche Preisindexformel (bei Mengenindizes analog Q^F als geometr. Mittel aus Q^L und Q^P) erfüllt F (und Z, nicht aber T), denn

(10.16) $$W_{0t} = P^F_{0t} Q^F_{0t}$$

Zirkularität (Verkettbarkeit, Transitivität, T)

Nach dieser Forderung (auch "Rundprobe" ["circular test"]) soll für *beliebige*, Einteilungen dens Intervalls [0,t) in [0,s) und [s,t) also für ***jedes*** s gelten:

(10.17) $$P_{0t} = P_{0s} P_{st} .$$

Die Einteilung in zwei Teilintervalle mit $0 < s < t$ ist nicht zwingend (es könnten auch drei oder mehr Teilintervalle sein, etwa $0 < r < s < t$, so dass gilt $P_{0t} = P_{0r} P_{rs} P_{st}$, oder auch $0 > s > t$). T wird oft dahingehend missverstanden, dass ein als Produkt definierter Index, wie der Kettenindex "verkettbar" sei. Dabei wird auch vergessen, dass bei Gl. 10.17 betont werden muss "für *jedes* s".

Wenn Identität gilt, dann folgt Z aus T (Umkehrung gilt nicht)..

Umbasierung und Verkettung (vgl. Kap. 9)

Von Zeitumkehrbarkeit und Verkettbarkeit als Axiome ist zu unterscheiden, dass entsprechende Berechnungen [als Hilfslösungen] vorgenommen werden:

	gegeben	gesucht	Lösung
Umbasierung (rescaling)	ein Index zur Basis 0 für die Perioden 0, ..., s,..., t	ein Index zur Basis s (meist: zur aktuelleren Basis s)	(10.18) $P_{st} = \frac{P_{0t}}{P_{0s}}$
Verkettung (splicing)	ein Index zur Basis 0 (berechnet mindestens bis zur Periode s) und einer zur Basis s	Bildung einer langen Reihe zur Basis 0 aus zwei oder mehreren sich überlappenden Indexreihen	(10.19) $P_{0t} = P_{0s} P_{st}$ (für $0 < s < t$)

Man sieht, dass die Rechenoperationen äquivalent sind und beide (Gl. 10.18 und19) auf dem simplen "Dreisatz": $P_{0t}/P_{0s} = P_{st}/P_{ss}$ (mit $P_{ss} = 1$) beruhen, der jedoch - genau genommen - nicht zutreffend ist, wenn Verkettbarkeit (wie z.B. bei P^L und P^P) nicht erfüllt ist

Additivität (= Linearität) der Indexfunktion

(als spezielle Form der Monotonie) bedeutet in der Notation der Übersicht 10.3:

Fall a) unterschiedliche Preise in der Berichtsperiode:

$$P(\mathbf{p}_0, \mathbf{p}^*_t) = P(\mathbf{p}_0, \mathbf{p}_t) + P(\mathbf{p}_0, \Delta\mathbf{p}^*_t)$$ wenn für $\mathbf{p}^*_t$, $\mathbf{p}_t$ und $\Delta\mathbf{p}^*_t$ gilt: $\mathbf{p}^*_t = \mathbf{p}_t + \Delta\mathbf{p}^*_t$

und entsprechend

Fall b) unterschiedliche Preise in der Basisperiode:

$$\left[P(\mathbf{p}^*_0, \mathbf{p}_t)\right]^{-1} = \left[P(\mathbf{p}_0, \mathbf{p}_t)\right]^{-1} + \left[P(\Delta\mathbf{p}^*_0, \mathbf{p}_t)\right]^{-1}$$ wenn entsprechend gilt: $\mathbf{p}^*_0 = \mathbf{p}_0 + \Delta\mathbf{p}^*_0$.

Die Indizes von Laspeyres und Paasche sind additiv.

Übersicht 10.3: Axiomensystem von Eichhorn und Voeller

Notation:
Preis- und Mengenvektoren (jeweils n Komponenten [Waren]) $\mathbf{p}_0$, $\mathbf{q}_0$, $\mathbf{p}_t$, $\mathbf{q}_t$. Die Indexfunktion P: $R^{4n} \Rightarrow R$ sollte danach die folgenden Axiome erfüllen:

P1	***Monotonie:*** a) in Berichtspreisen $P(\mathbf{p}_0,\mathbf{p}_t^*) > P(\mathbf{p}_0.\mathbf{p}_t)$, wenn $p_{it}^* \geq p_{it}$ und für mindestens eine Ware i gilt: $p_{it}^* > p_{it}$ b) in Basispreisen $P(\mathbf{p}_0,\mathbf{p}_t) > P(\mathbf{p}_0^*.\mathbf{p}_t)$ wenn analog gilt: $p_{i0}^* \geq p_{i0}$ und $p_{i0}^* > p_{i0}$ für mindestens ein i (eine Ware)
P2	***Lineare Homogenität:***[a] $P(\mathbf{p}_0,\mu\mathbf{p}_t) = \mu P(\mathbf{p}_0,\mathbf{p}_t)$ mit $\mu \in R_+$ (nicht zu verwechseln mit Proportionalität: $P(\mathbf{p}_0,\mu\mathbf{p}_0) = \mu$, wobei $p_{it} = \mu p_{i0}$ für alle i)
P3	***Identität:***[b] $P(\mathbf{p}_0,\mathbf{p}_t) = 1$ wenn $p_{it} = p_{i0}$ für alle i also $\mathbf{p}_t = \mathbf{p}_0$
P4	***Dimensionalität:*** $P(\mu\mathbf{p}_0,\mu\mathbf{p}_t) = P(\mathbf{p}_0,\mathbf{p}_t)$ mit $\mu \in R_+$ (Unabhängigkeit von der Währungseinheit der Preise)
P5	***Kommensurabilität:*** $P(\Lambda\mathbf{p}_0, \Lambda\mathbf{p}_t, \Lambda^{-1}\mathbf{q}_0, \Lambda^{-1}\mathbf{q}_t) = P(\mathbf{p}_0,\mathbf{p}_t,\mathbf{q}_0,\mathbf{q}_t)$ mit Λ = diag(λ_1, ..., λ_n) und $\lvert\lambda_i\rvert > 0$ (Unabhängigkeit von der Mengeneinheit, auf die sich die Preisnotierung bezieht).

a) Unter Homogenität vom Grade -1 versteht man die Forderung $P(\mu\mathbf{p}_0, \mathbf{p}_t) = \mu^{-1}P(\mathbf{p}_0, \mathbf{p}_t)$. Sie ist erfüllt, wenn P2 und P4 gelten.

b) Axiome P2 und P3 stellen zusammen sicher, dass die sog. Proportionalitätsprobe erfüllt ist.

10.2 Kettenindizes

Die Standardkritik am direkten Preisindex nach Laspeyres P_{0t}^L (als Maß der Inflation) bzw. an Volumen, die man als Ergebnis einer Deflationierung mit einem (direkten) Paasche Preisindex erhält ist, dass in P_{0t}^L die Mengen, bzw. in P_{0t}^P die Preise (allgemein die Gewichte) für eine gewisse Zeit (in Interesse des reinen Preisvergleichs) konstant gehalten werden und dass das Wägungsschema veraltet. Es müsse stattdessen jeweils mit möglichst aktuellen Gewichten gerechnet werden. Nicht viel mehr als dies steckt hinter der in neuerer Zeit vehement wiederbelebten Forderung nach Kettenindizes[4].

Die Definition eines Kettenindexes umfasst stets zwei Elemente(Übers. 10.4),

- die **Kette** $\overline{P}_{0t}^C$ (C = chain), das konstantes Element: Zwei-Periodenvergleich (zwischen 0 und t) indirekt als Produkt $P_{0t} = P_{01}P_{12}....P_{t-1,t}$.(analog zur Verkettung)
- (das variable Element) das **Kettenglied** $P_t^C = P_{t-1,t}$ (link), das je nach verwendeter Indexformel unterschiedlich ist, z.B. nach Laspeyres, Paasche usw. (Übers. 10.4)

Befürworter von Kettenindizes vergleichen meist P_t^C (statt $\overline{P}_{0t}^C$) mit P_{0t}. Dem Vorteil, dass $\overline{P}_{0t}^{LC}$ (auch) von den aktuelleren Mengen q_{t-1} abhängt, nicht nur von den "veralteten" Mengen

[4] Sie sind (leider) in internationalen Empfehlungen für die Verwendung in der amtlichen Statistik vorgeschrieben worden.

10.4 Definition von Kettenindizes

Kettenindex

Definition der **Kettenglieder** (links)	Verkettung zur **Kette**
$P_t^C = P_{t-1,t}$ ein Index (genügt Axiomen)	$\overline{P}_{0t}^C = P_1^C P_2^C ... P_t^C$ kein Index*
Beispiele: Laspeyres $P_t^{LC} = \frac{\sum p_t q_{t-1}}{\sum p_{t-1} q_{t-1}}$, Paasche $P_t^{PC} = \frac{\sum p_t q_t}{\sum p_{t-1} q_t}$	Beispiele: $\overline{P}_{0t}^{LC} = P_1^{LC} P_2^{LC} ... P_t^{LC}$ $\overline{P}_{0t}^{PC} = P_1^{PC} P_2^{PC} ... P_t^{PC}$
ein Warenkorb q_{t-1} bzw. q_t **	**viele** Warenkörbe $q_0, q_1, ...$

* muss nicht Axiome (z.B. im Sinne von Übers. 10.3) erfüllen, selbst wenn das einzelne Kettenglied dies tut.

** bei unterjährigem Vergleich zum Vorjahr (ein Monat verglichen mit dem gleichem Monat im Vorjahr) jedoch schon beim einzelnen Kettenglied **zwei** Warenkörbe.

q_0 wie in P_{0t} stehen folgende **Nachteile** gegenüber:

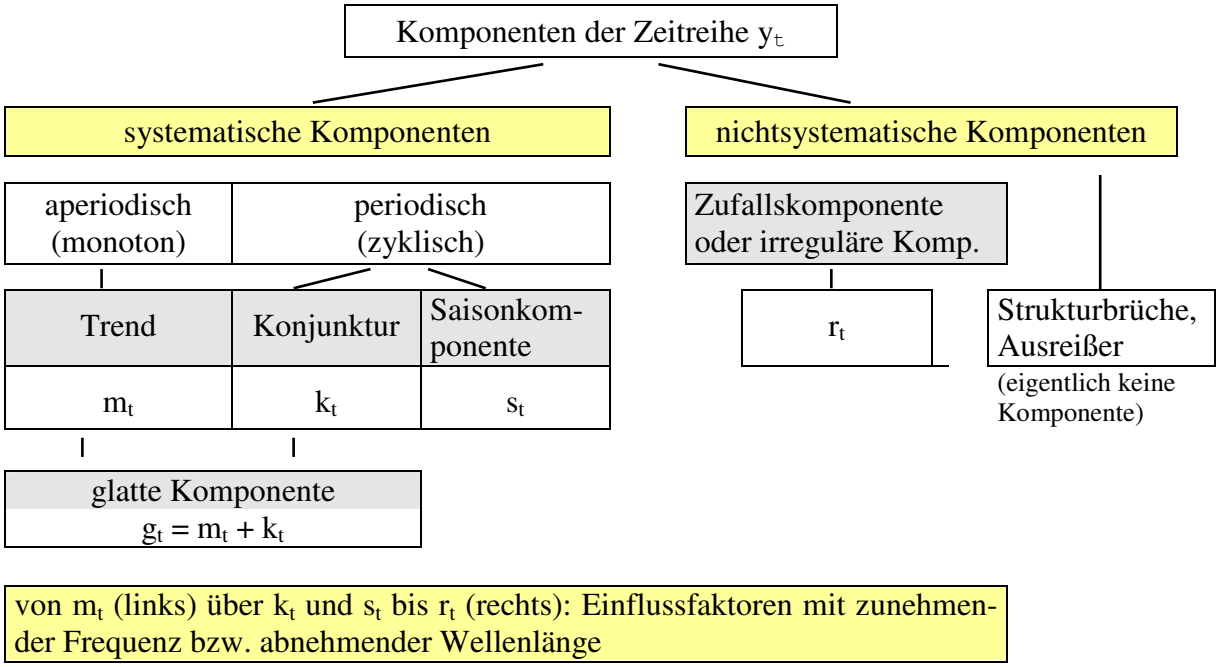
1. Kettenindizes erlauben keine **Interpretation** im Sinne des reinen Preisvergleichs, als Messzahlenmittelwert oder als Verhältnis von Aggregaten. Axiome sind auf sie nicht anwendbar: Trotz gleicher Preise in Periode 0 und 2 muss nicht gelten $P_{02} = 1$ (Identität verletzt, ebenso können Monotonie und andere Axiome von Übers. 10.3 verletzt sein).
2. Verkettung als Form der **zeitlichen Aggregation** ist **pfadabhängig**: ein Kettenindex ist kein Zwei-Perioden-**Vergleich**, sondern ein summarisches Maß für die Gestalt einer Zeitreihe (für einen **Verlauf**). Das Ergebnis für das Intervall von 0 bis t ist i.d.R. unterschiedlich, je nach dem, wie es in Teilintervalle zerlegt wird und wie sich Preise und Mengen in den Zwischenperioden 1,...,t-1 entwickeln. Bei zyklischer Bewegung der Preise (der Verlauf zwischen 0 und t wiederholt sich) kann die Kette für Periode 2t, 3t, ... im Wert ständig zunehmen (wenn der Index $P_{0t} > 1$ ist) oder abnehmen (wenn $P_{0t} < 1$), selbst dann wenn die Preise in 0, t, 2t,... alle gleich sind.
3. Ungünstige Aggregationseigenschaften: **additive** - und (bei Deflationierung) **strukturelle Konsistenz** nicht erfüllt. Volumen V_t nicht nur abhängig von q_t und p_0, auch von Preisen p_1, ..., p_t., so dass man kaum von "in *konstanten*" Preisen sprechen kann.
4. Erheblicher **Mehraufwand** für Datenbeschaffung (häufigere Feststellung der Warenkörbe).

Kapitel 11: Einführung in die Zeitreihenanalyse

1. Komponentenmodell

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".



Verknüpfung (Überlagerung) der Komponenten

Man unterscheidet zwei Grundmodelle:

1) additive Überlagerung: wenn bei steigendem bzw. fallendem Trend die zyklischen Einflüsse 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:

$$(11.1) \quad y_t = m_t + k_t + s_t + r_t \qquad \text{(additive Überlagerung)}$$

2) multiplikative Überlagerung: wenn z.B. die Schwankungen der zyklischen Komponenten, insbesondere der Saisonkomponente mit steigendem Niveau der Zeitreihe zunehmen:

$$(11.2) \quad y_t = m_t k_t s_t r_t \qquad \text{(multiplikative Überlagerung)}$$

2. Methoden zur Berechnung des Trends (der glatten Komponente)

1. Trendberechnung mit der Methode der kleinsten Quadrate

Bei einem Trend mit einer 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 Gesamtheit der übrigen Kom-

ponenten mit den Messwerten k_t, s_t und r_t ist das Residuum u_t.

Die Parameter a und b eines linearen Trends $y_t = m_t = a + bt$ (t = 1,2,...,T) werden mit den Normalgleichungen wie folgt

(11.3a)	$aT + b\sum t = \sum y_t$	1. Normalgleichung
(11.3b)	$a\sum t + b\sum t^2 = \sum ty_t$	2. Normalgleichung

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 und a und b direkt aus jeweils einer der beiden Normalgleichungen zu bestimmen ist: $a = (\Sigma y)/T$ und $b= (\Sigma ty)/T^2$.

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 p-gliedrige Durchschnitt lautet bei ungeradzahligem p = 2k+1

(11.4) $$\tilde{y}_t = \frac{1}{p}\sum_{h=-k}^{k} y_{t+h} \quad (p = 2k + 1, \text{ungeradzahlig}) \quad \text{oder}$$

$$\tilde{y}_t = (y_{t-k} + y_{t-k+1} + ... + y_t + ... + y_{t+k-1} + y_{t+k})/p.$$

b) Bei geradzahligem p = 2k wäre der Durchschnitt $\tilde{y}_t$ der Periode t - ½ und $\tilde{y}_{t+1}$ der Periode t + ½ zuzuordnen. Es liegt daher nahe einen ungewogenen Durchschnitt hieraus zu berechnen. Dieser der Periode t zugeordnete zentrierte gleitenden Durchschnitt lautet: ½

(11.5) $$\tilde{y}_t^z = \frac{1}{p}\left(\sum_{h=-(k-1)}^{k-1} y_{t+h} + \frac{y_{t-k} + y_{t+k}}{2}\right) \quad (p =2k, \text{geradzahlig}) \quad \text{etwa bei } p = 4$$

$$\tilde{y}_t^z = \frac{1}{4}\left(\tfrac{1}{2}y_{t-2} + y_{t-1} + y_t + y_{t+1} + \tfrac{1}{2}y_{t+2}\right)$$

Am Anfang und Ende fallen beim gleitenden Durchschnitt jeweils k Glieder weg. Der erste gleitende Durchschnitt fällt auf den k+1-ten Wert.

	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

3. Berechnung der Saisonkomponente

Konstante (starre) Saisonfigur (Saisonnormale) bei **additiver** Überlagerung

Ursprungswerte y und trendbereinigte Werte y* (bzw. bereinigt von glatter Komponente) für die Jahre j = 1, 2, ..., J und Unterzeitraum z = 1, 2, ..., Z (Z = Anzahl der Unterzeiträume, bei Quartalen Z = 4, bei Monatsdaten Z= 12)

(11.7) $$y^*_{jz} = y_{jz} - g_{jz}, \text{ etwa mit } g_{jz} = \tilde{y}_{jz}$$

nicht-normierte Saisonnormale

(11.8) $$S_z = \frac{1}{J}\sum_{j=1}^{J} y^*_{jz} \quad \text{mit dem Mittel} \quad \overline{S} = \frac{\sum_z S_z}{Z}$$

normierte (auf einen mittleren Wert 0) Saisonnormale (11.10) $S^*_z = S_z - \overline{S}$.
Bei multiplikativer Überlagerung Division statt Subtraktion

(11.7a) $y^*_{jz} = y_{jz} / g_{jz}$,

jedoch S_Z und mittlere Saisonnormale $\overline{S}$ auch als arithmet. Mittel und

(11.10a) $S^*_z = S_z / \overline{S}$

(Division durch mittlere nicht normierte Saison), normiert auf ein Mittel von 1.

4. Hinweise auf weiterführende Verfahren

4.1 Exponential Smoothing (exponentielles Glätten)

1. Prognose als gewogenes Mittel aus den letzten Werten

(11.11) $$y^p_{t+1} = \alpha y_t (1-\alpha) y^p_t \qquad \text{mit: } 0 < \alpha < 1.$$

2. Prognose als Mittel aller vergangener Beobachtungen

(11.12)

$$y^p_{t+1} = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \ldots + \alpha(1-\alpha)^n y_{t-n} + (1-\alpha)^{n+1} y^p_{t-n}$$

$$= \sum_{i=0}^{n} \alpha(1-\alpha)^i y_{t-i} + (1-\alpha)^{n+1} y^p_{t-n}.$$

3. Prognose als partielle (mit α gewogene) Korrektur einer Fehlschätzung F

(11.13) $$y^p_{t+1} = y^p_t + \alpha\left(y_t - y^p_t\right) = y^p_t + \alpha F.$$

4.2 Filter, Operatoren, Polynome

1. Ein "Filter" verwandelt eine Zeitreihe y_t (input) in eine transformierte Zeitreihe (output) z_t. Einfache lineare Filter sind z.B. gleitende Mittelwerte oder Differenzenbildung (Output $z_t = y_t - y_{t-1}$). Ein nichtlinearer Filter ist z.B. die Bildung von Wachstumsraten $r_t = (y_t - y_{t-1})/y_{t-1}$.
2. Operatoren: Verschiebungen der Variable t bewirkt der Backshift- oder Lag-Operator: $Ly_t = y_{t-1}$, $L^2 y_t = y_{t-2}$ usw.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 ∇) $\nabla y_t = y_t - y_{t-1}$. Hintereinanderausführen heißt Potenzieren des Operators $\nabla^2 y_t = \nabla y_{t+1} - \nabla y_t = y_{t+2} - 2y_{t+1} + y_t$. Man beachte, dass $\nabla^2 y_t$ nicht identisch ist mit $y_{t+2} - y_t$. Vor- und Rückwärtsdifferenzen für mehrere Perioden, etwa $\nabla_4 y_t = y_t - y_{t-4}$ oder $\nabla_{12} y_t = y_t - y_{t-12}$ beim Vorjahresvergleich mit Quartals- oder Monatsdaten sind "saisonale Differenzen".
3. Der Ausdruck $A_p(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_p t^p$ ist ein Polynom in t vom Grade p und $A_p(\ )$ heißt Polynomenoperator. Ein autoregressives Schema (eine linear-rekursive Funktion) ist $A_p(L)y_t = a(_0 + a_1 L + a_2 L^2 + \ldots + a_p L^p)\, y_t$. Lineare Filter kann man als Lagpolynome darstellen und Polynome in t als linear-rekursive Funktionen. Einem Polynom $y = A_p(t)$ ist eine linear rekursive Funktion $B_{p+1}(L)y$ äquivalent:

   p=1: das der Funktion $y_t = a_0 + a_1 t$ (Polynom vom Grade 1) äquivalente Lagpolynom ist $y_t = 2y_{t-1} - y_{t-2}$ mit den Anfangswerten $y_0 = a_0$ und $y_0 = a_0 + a_1$.

   p=2: der Funktion $y_t = a_0 + a_1 t + a_2 t^2$ 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 $y_2 = a_0 + 2a_1 + 4a_2$.

Kapitel 12: Bestandsanalyse und Tafelrechnung

12.1. Bestands- und Bewegungsmassen

Def. 12.1: (Bestandsmasse, Bewegungsmasse, Verweildauer)

a) Eine statistische Masse, deren Einheiten (i=1,2,...,n) jeweils gemeinsam zu einem bestimmten Zeitpunkt 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 Zeitraum (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**.

Methoden der Erhebung von Bestands- und Bewegungsmassen:

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_o bis t_j) kumulierten Zugänge (Z_{oj}) und Abgänge (A_{oj}), d.h. der Bruttoströme.

c) Feststellung der Bestandsveränderungen (d.h. der Salden- oder Nettoströme Z_{oj}-A_{oj}).

2. Feststellung der Bestände (Bestandsmassen)

a) durch periodische Inventuren (Zählen oder Messen)

b) durch **Fortschreibung** für das Intervall $[t_o,t_j]$: (12.1) $B_j = B_o + Z_{oj} - A_{oj}$ (j = 0, 1, ..., m)

In Gl. 12.1 ist B_o der Anfangsbestand, B_j der Bestand zum Zeitpunkt t_j, Z_{oj} die Anzahl der Zugänge und A_{oj} die Anzahl der Abgänge im Beobachtungsintervall $[t_o,t_j]$.

c) Bei Kenntnis sämtlicher individueller Verläufe (wie in 1a), also bei Längsschnittdaten, ist der Bestand zu jedem beliebigen Zeitpunkt bekannt.

> **Querschnitts**analysen sind die Kombination 1b + 2a
> **Längsschnitts**analysen die Kombination 1a + 2c.

Beckersches Diagramm, Bestandsfunktion und Zeitmengenfläche

1. Beckersches Diagramm: Eine graphische Darstellung der individuellen Verläufe ist das Beckersche Diagramm (Abb. 12.1 für Aufgabe 12.1 [siehe unten Aufgabenteil]).

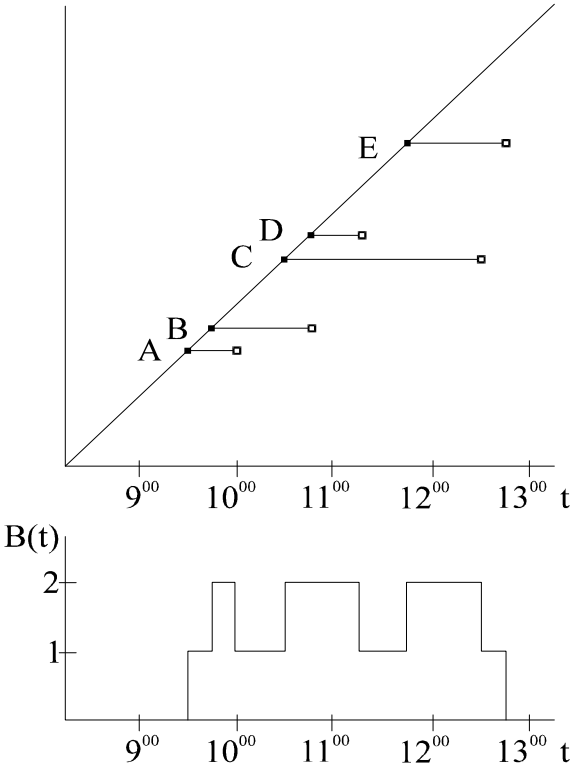
2. Bestandsfunktion: Es ist leicht zu sehen, wie aus dem Beckerschen Diagramm (oberer Teil von Abb. 12.1) die Bestandsfunktion B(t) (t stetig), bzw. B_j (Bestände zu den diskreten Zeitpunkten t_j) herzuleiten ist. Mit jedem Zugang (Abgang) einer Einheit erhöht (verringert) sich

die Bestandsfunktion um 1.

3. Zeitmengenfläche: Die schraffierte Fläche unter der Bestandsfunktion heißt Zeitmengenfläche F, oder genauer F_{om} wenn die Fläche "über" dem Intervall $[t_o,t_m]$ betrachtet wird.

	Zeitpunkt	
	Zugang	Abgang
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}

Abb. 2.1 Beckersches Diagramm und Bestandsfunktion (Bsp.)



Def. 12.2: (offene-, geschlossene Masse)

Eine Bestandsmasse heißt geschlossen bezüglich des Zeitintervalls $[t_o,t_m]$, wenn keine ihrer Einheiten vor t_o 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.

12.2. Kennzahlen der Dynamik eines Bestands: Durchschnittsbestand, durchschnittliche Verweildauer, Umschlagshäufigkeit

Berechnung der Kennzahlen (vgl. Übers. 12.1)

a) bei Kenntnis der individuellen Verläufe (Längsschnittsdaten)

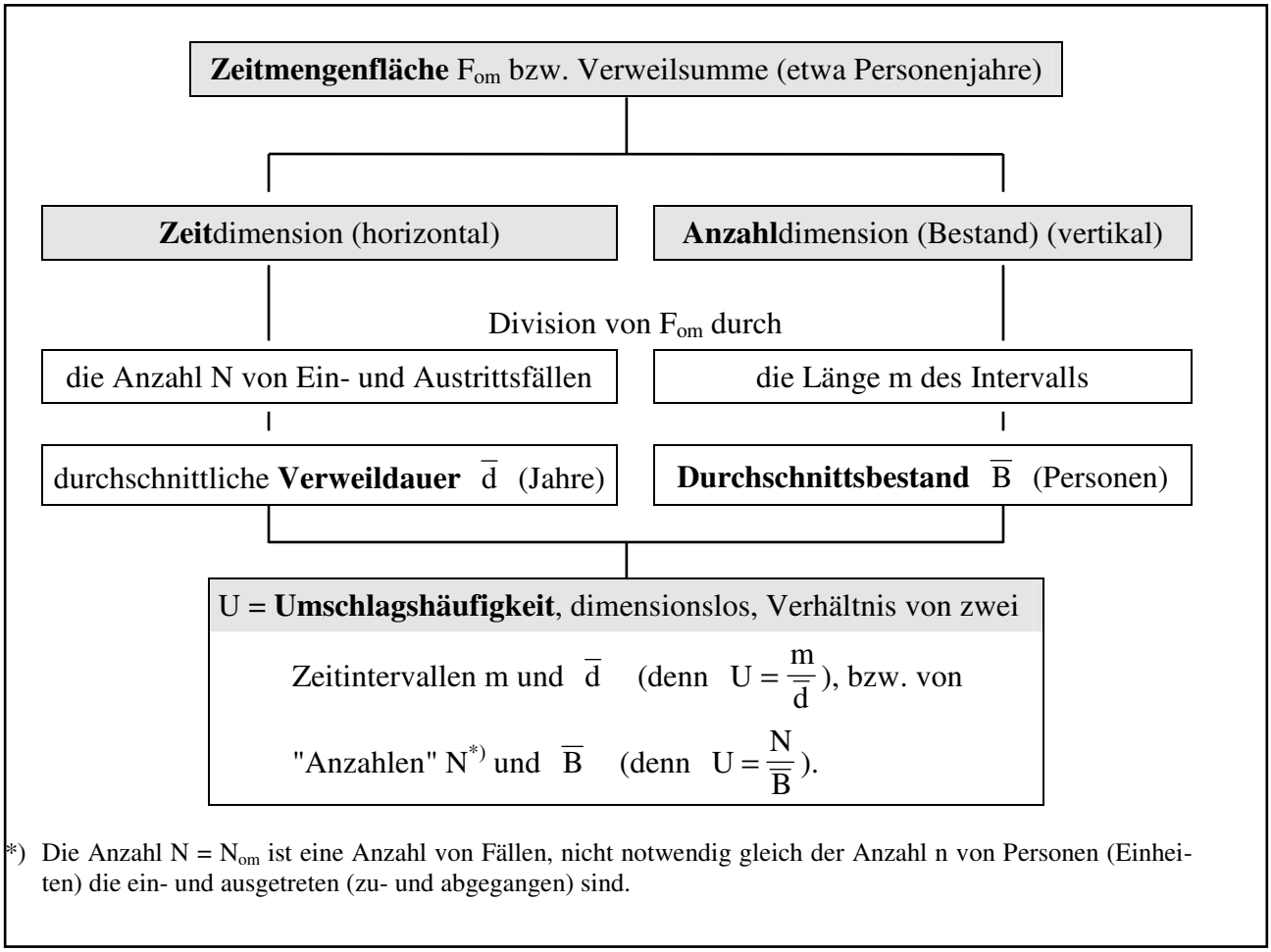
(12.3) $\overline{B} = \frac{F_{om}}{m}$ **(Durchschnittsbestand)**

(12.4) $\overline{d} = \frac{\sum d_i}{N_{0m}}$ (**durchschnittliche Verweildauer**).

Bei geschlossener Masse ist Zeitmengenfläche = Verweilsumme (und deshalb $\overline{d} = \frac{F_{om}}{N_{0m}}$)

(12.5) $U = \frac{m}{\overline{d}} = \frac{N_{0m}}{\overline{B}}$ (**Umschlagshäufigkeit**)

Übers. 12.1: Zusammenhänge zwischen Kennzahlen zur Beschreibung der Bestandsentwicklung



*) Die Anzahl $N = N_{om}$ ist eine Anzahl von Fällen, nicht notwendig gleich der Anzahl n von Personen (Einheiten) die ein- und ausgetreten (zu- und abgegangen) sind.

b) bei Querschnittsdaten

Zeitmengenfläche und Durchschnittsbestand

Finden die Bestandsänderungen ausschließlich genau zu den Beobachtungszeitpunkten t_j (j = 1, ..., m) statt, dann ist die Zeitmengenfläche $F_{om} = \sum_j B_{j-1}\left(t_j - t_{j-1}\right)$. 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_o = m, so gilt:

(12.10) $F_{om} = \frac{1}{2}B_o + B_1 + \ldots + B_{m-1} + \frac{1}{2}B_m$ (chronologisches Mittel)

woraus $\bar{B}$ wieder mit Gl. 12.3 zu errechnen ist.

Durchschnittliche Verweildauer und Umschlagshäufigkeit

Es ist nicht mehr von $\Sigma d_i = F_{om}$ auszugehen. Vielmehr ist F_{om} zu korrigieren um die Zeiten, welche

- die B_o Einheiten des Anfangsbestands vor to bereits dem Bestand angehört hatten (*Aufbauzeiten* d_o) und die Zeiten, welche die
- B_m Einheiten des Endbestands nach t_m dem Bestand noch angehören werden (*Abbauzeiten* d_m).

F_{om} ist also mit (geschätzten) durchschnittlichen Auf- und Abbauzeiten zu G_{om} (geschätzte Verweilsumme Σd_i) zu korrigieren

(12.6) $$G_{om} = B_o \bar{d}_o + F_{om} + B_m \bar{d}_m$$

(12.8) $$\bar{d} = \frac{G_{om}}{N_{om}}.$$

Es sind jetzt Annahmen über die mittlere Aufbau- und Abbauzeit nötig. üblich ist die Annahme $\bar{d}_o = \delta\bar{d}$ und $d_m = (1-\delta)\bar{d}$ mit $0 < \delta < 1$ liefert das

(12.11) $$\bar{d} = \frac{F_{om}}{\delta Z_{om} + (1-\delta)A_{om}}$$ und mit δ = ½ die bekannten Formel

(12.12) $$\bar{d} = \frac{2m\bar{B}}{Z_{om} + A_{om}}.$$

12.3. Stationäre Bevölkerung und Tafelrechnung

Def. 12.3: (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. Der Umfang dieser Masse, d.h. die Anzahl der zugehenden Einheiten ist l_o.

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_o nach Vollendung von x Jahren. l_x ist monoton fallend.

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 für die Zugänge Z gilt $Z_{j-1,j} = l_o$ für alle j) mit gleicher Abgangsordnung ersetzt (d.h. gleicher "Struktur"; l_x ist nicht von j sondern nur von x abhängig).

Def. 12.4 (Tafelfunktionen l, q, p, d, 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). Die einjährige Überlebenswahrscheinlichkeit p_x ist demzufolge $p_x = 1 - q_x$. Auch p_x ist eine bedingte Wahrscheinlichkeit.

b) 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_o eindeutig gegeben:

die Absterbeordnung l_x ist ausgehend von einem fiktiven Anfangsbestand von l_o = 100.000 Personen rekursiv zu berechnen mit

(12.18) $$l_{x+1} = l_x p_x = l_x(1-q_x).$$

Entsprechend ist die Anzahl $d_x \geq 0$ der im Altersintervall (x, x+1) gestorbenen Personen

(12.19) $$d_x = l_x q_x = l_x - l_{x+1}.$$

Wie man leicht sieht, ist $\sum_x d_x = l_o$.

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 [eine Zeitmengenfläche, bzw. lineare Interpolation der Abgangsordnung l_x]) bezeichnet.

$$L_x = \frac{1}{2}\left(l_x + l_{x+1}\right). \qquad (12.20)$$

Def. 12.5: (Tafelfunktionen T, T*, e, 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.

$$T_x = \sum_{y=x}^{w} L_y \qquad x \le y \le w. \qquad (12.24)$$

$$T^*_{*x} = \sum_{y=x}^{w} l_y = T_x + \tfrac{1}{2} l_x \qquad (12.25)$$

Die Größen T_x und T_x^* sind Verweilsummen; Maßeinheit: "Personenjahre".

b) Dividiert man T_x bzw. T_x^* durch die Anzahl der Überlebenden des Alters x, also durch l_x, so erhält man mit

$$e_x = \frac{T_x}{l_x} = \frac{T_x^*}{l_x} - \frac{1}{2} = e_x^* - \frac{1}{2} \qquad (12.26)$$

die (mittlere, durchschnittliche) weitere **Lebenserwartung** einer x-jährigen Person (spricht man von "der" Lebenserwartung, so ist e_0 gemeint). Die Größe e_0 ist zugleich das durchschnittliche Sterbealter der stationären Bevölkerung, eine Verweilsumme, und es gilt

Bestand (T_0) = Zugang (l_0) · durchschnittliche Verweildauer (e_0)

bei einer **stationären** Bevölkerung.

Ende des Formelteils