STATISTIK I

BESCHREIBENDE STATISTIK

von

Alex Keel

15. Auflage

2000

Verlag Wilhelm Surbir St. Gallen

Vorlesung an der Universität St. Gallen - Hochschule für Wirtschafts-, Rechts- und Sozialwissenschaften (HSG)

Prof. Dr. Alex Keel, Bodanstrasse 4, CH-9000 St. Gallen, Tel. 071/224 24 31

Verlag Wilhelm Surbir, Dufourstrasse 48, CH-9000 St. Gallen, Tel. 071/224 23 01 und 071/298 36 16, Fax 071/224 26 46

Vorwort

Der vorliegende Band Statistik I *Beschreibende Statistik* ist der erste einer dreiteiligen Serie von Begleittexten zur Einführungsvorlesung in die Statistik an der Universität St. Gallen. Diese ist in drei Bereiche, nämlich beschreibende Statistik, Wahrscheinlichkeit und induktive Statistik gegliedert. Die beschreibende Statistik wurde vornehmlich aus zwei Gründen an den Anfang gestellt. Zunächst wird inhaltlich eine Auswahl jener Bereiche abgedeckt, welche der Statistik traditionellerweise zugeordnet werden. Zweitens erlaubt dieser Aufbau, formale Konzepte der eher mathematisch orientierten Wahrscheinlichkeitstheorie intuitiv vorzubereiten.

Der Text wendet sich primär an Studierende der Wirtschafts- und Sozialwissenschaften, welche erstmals mit der Statistik als Wissenschaft konfrontiert werden. Der Konzeption liegt die Vorstellung zugrunde, eine anwendungsbezogene Einführung in die statistische Methodenlehre bereitzustellen. Aus diesem Grund mag der eher formal interessierte Leser die mathematisch konzise Stringenz vermissen. Im Rahmen einer ersten Begegnung mit einer Wissenschaft, über deren konkreten Inhalt oft nur vage Vorstellungen existieren, wurde zur Vermittlung der ersten Einblicke in die statistische Denk- und Arbeitsweise der anschaulichen Darstellung jedoch höhere Priorität eingeräumt. Auf dieser Grundlage wurde versucht, auch an Studierende ohne spezielle mathematische Neigungen heranzukommen.

Im Text sind verschiedentlich erläuternde Beispiele eingefügt, es fehlen hingegen eigentliche Übungsbeispiele. Diese sind in einem separaten Skriptum mit Kurzlösungen zusammengefasst.

Mir bleibt die angenehme Aufgabe des Dankens. Meine Mitarbeiter Stefan Ott, Dipl.-Math.oec., Reto Leibundgut, lic.oec. sowie Anton Vucurovic, lic.oec. haben mich bei der Neufassung des Skriptums tatkräftig unterstützt. Ihre stete Bereitschaft zur aufbauenden Diskussion hat die Motivation zur Neukonzeption nachhaltig beeinflusst. Frau M.-C. Baumann hat einmal mehr die Koordination sowie die formelle Gestaltung des Skriptums geprägt. Herzlichen Dank.

INHALTSVERZEICHNIS

1. TEIL - BESCHREIBENDE STATISTIK

1. SAMMELN UND ORDNEN

Das zentrale Anliegen der beschreibenden Statistik besteht in der Aufarbeitung einer im allgemeinen umfangreichen Menge von Daten. Solche Datenmengen stammen etwa von Befragungen (Volks- und Betriebszählungen), von Experimenten oder von Beobachtungen der Wirtschaft (Börsenkurse) oder der Natur (Niederschlagsmengen). All diesen Beispielen sollte gemeinsam sein, dass *vor* der Datenerhebung bei der Formulierung des Untersuchungszieles klar festgelegt wird, *was* untersucht werden soll, *wer* - zumindest potentiell - in die Untersuchung einbezogen werden soll und schliesslich *wie* die Beobachtungen zu erfolgen haben.

1.1 Wer - Grundgesamtheit

Die Grundgesamtheit Ω (Kollektiv, Population) umfasst die Menge der möglichen Beobachtungsobjekte, über welche man Aussagen im Hinblick auf das Untersuchungsziel machen möchte. Zur eindeutigen Festlegung der Grundgesamtheit gehört neben der sachlichen meist noch eine zeitliche und lokale Abgrenzung (Berufstätige Bevölkerung einer bestimmten Gemeinde zu einem bestimmten Zeitpunkt).

Die Elemente von Ω werden als *statistische Einheiten* bezeichnet.

Die Schwierigkeiten bei der Abgrenzung von Ω werden in aller Regel unterschätzt. Im Rahmen einer Untersuchung über die Arbeitslosigkeit muss beispielsweise festgelegt werden, ob eine nur temporär arbeitslose Person zu Ω gehört, ob Hausfrau ein Beruf ist, wieviele wöchentlichen Arbeitsstunden notwendig sind, um als berufstätig zu gelten, ob Gelegenheitsarbeiter zu Ω gehören, wenn sie am Tage x zufällig beschäftigt sind.

Grundgesamtheiten sind entweder empirisch vorgegeben (wie im obigen Beispiel) oder bilden eine Fiktion (Gesamtheit der Menschen, welche eine bestimmte Krankheit hatten, haben oder haben werden).

1.2 Was - Merkmale und deren Skalierung

Durch das Untersuchungsziel wird vorgegeben, welche statistischen Merkmale bei den Untersuchungsobjekten (statistischen Einheiten) festzustellen sind. Die Feststellung der Merkmale erfolgt über deren "Messung". Jedes Merkmal besitzt einen wohldefinierten Ausprägungsbereich, für welchen zumindest prinzipiell beliebige Aufteilungen (Partitionen) möglich sind. Eine bestimmte Aufteilung des gesamten Ausprägungsbereichs, verbunden mit einer Vorschrift, wie die konkreten Ausprägungen der Merkmale den Elementen der Partition zuzuordnen sind, nennt man *Skalierung*. Die Partition selber heisst *Skala*. Messen bedeutet dann, die Beobachtungen entsprechend ihrer Ausprägung den Elementen der (Skalierungs)- Partition zuzuordnen.

Je nach den Eigenschaften der Elemente der Partition spricht man von Nominal-, Ordinal-, Intervall- und Verhältnisskala.

Nominalskala
Der Ausprägungsbereich wird in genau k voneinander verschiedene Kategorien aufgeteilt. Zwischen den Kategorien ist lediglich die Relation gleich (ungleich) definiert. Die Zuordnung zu den Kategorien dient im wesentlichen der Unterscheidung der Merkmalsausprägungen und erfolgt im allgemeinen auf rein qualitativer Basis.

Beispiele: Geschlecht, Zivilstand, Religionszugehörigkeit, Postleitzahl.

Solange die Zuordnung zu den Kategorien nicht gestört wird, können Nominaldaten beliebig transformiert werden; sie sind nur bis auf eineindeutige Transformationen bestimmt.

Ordinalskala
Bei ordinal skalierten Merkmalen ist zusätzlich zur Gleichheits- eine *Ordnungs*relation zwischen den Merkmalsausprägungen definiert.

Beispiele: Schulnoten, Intelligenzquotienten

Beachte: Besitzt A die Note 6, B die Note 5, C die Note 3 und D die Note 2, so kann daraus nicht geschlossen werden, dass A im selben Masse besser ist als B wie C besser ist als D. Es wird lediglich eine Ordnung, hingegen kein Abstand festgelegt.

Ordinaldaten sind nur bis auf ordnungserhaltende Transformationen eindeutig bestimmt.

Intervallskala
Bei der Intervallskala ist zusätzlich zur Ordnung auch der *Abstand* zwischen den Merkmalsausprägungen definiert. Er wird in einem prinzipiell frei wählbaren Einheitsmass gemessen. Für die Temperatur kann als Einheitsmass eine bestimmte Volumenänderung von Quecksilber verwendet werden. Änderungen der Temperatur werden dann in Vielfachen dieser Volumenänderung gemessen. Neben dem Einheitsmass ist auch der Nullpunkt frei wählbar.

Beispiele: Temperatur (sofern nicht in Grad Kelvin gemessen), Jahrgang

Beachte: Die Aussage "40°C ist doppelt so warm wie 20°C" ist nicht sinnvoll; sie hängt neben der Masseinheit auch von der Wahl des Nullpunktes ab. Hingegen ist die Aussage "Die Temperaturänderung von 40°C auf 60°C ist doppelt so gross wie jene von 10°C auf 20°C" sinnvoll. Für Intervalldaten selber ist die Differenz und der Quotient von Differenzen erklärt.

Intervalldaten sind bis auf lineare Transformationen der Form $f: x \to ax + b$ eindeutig bestimmt.

Verhältnisskala
Bei Verhältnisdaten ist zusätzlich zur Differenz auch der Quotient von zwei Messwerten sinnvoll definiert. Der Nullpunkt wird als fest vorausgesetzt. Das Einheitsmass ist hingegen frei wählbar.

Beispiele: Einkommen, Einwohnerzahl, Gewicht, Energieverbrauch

Verhältnisdaten sind nur bis auf eine von Null verschiedene Multiplikation, d.h. eine lineare Transformation der Form $f: x \to ax,\ a \neq 0$, eindeutig definiert.

Daten, welche mindestens auf dem Intervallniveau messbar sind, werden auch als *Kardinal*daten bezeichnet. Ihre Skalierung heisst metrische Skalierung. Innerhalb dieser Kategorie unterscheidet man ferner zwischen *diskreten* und *stetigen* Merkmalen. Diskrete Merkmale besitzen eine abzählbare Ausprägungsmannigfaltigkeit (Einwohnerzahl, Gewinn, Ausspielungen beim Lotto, Kinderzahl einer Familie). Stetige Merkmale sind in ihrer Ausprägungsmannigfaltigkeit nicht mehr abzählbar (Gewicht, Zeit, Länge). Messtechnische Restriktionen sind allerdings dafür verantwortlich, dass stetige Merkmale nur in der Modellvorstellung existieren. Die praktische Messgenauigkeit ist stets beschränkt, sodass in der konkreten Realisierung auch stetige Merkmale als diskret in Erscheinung treten.

Die statistische Aufbereitung beobachteter Daten hängt offensichtlich von ihrem Messniveau ab. Selbst wenn allen Beobachtungen Zahlen zugeordnet werden, so sind die mit diesen Zahlen erlaubten Operationen entscheidend vom Messniveau abhängig. Für ordinale Zahlen ist beispielsweise keine Addition definiert.

1.3 Wie - Erhebung

Im Rahmen der deskriptiven Statistik ist es unerheblich, ob die beobachteten Daten aus einer Vollerhebung stammen, oder ob es sich lediglich um eine Teilgesamtheit in Form einer Stichprobe handelt. Normalerweise liegen vor allem aus finanziellen und zeitlichen Gründen nur Stichprobenwerte zur Bearbeitung vor. Im Hinblick auf eine einheitliche Notation werden die verfügbaren Daten als empirischer Befund in Form einer *Stichprobe* betrachtet. Sie werden in der Reihenfolge ihrer Erhebung als Protokoll oder Urliste mit $x_1, x_2, \ldots, x_n$ bezeichnet.

Beispiel A
Qualitatives Merkmal „Zivilstand" mit den Ausprägungen

L ledig
H verheiratet
W verwitwet
G geschieden

50 Beobachtungen mit der Urliste
LWLLH HHGHL GHHLH GWLHL LHLHH
LWGLH WLGLL HLHGG LHHLW HLHHG

Beispiel B
Ordinales Merkmal „Zwischenprüfungsnoten" mit den Ausprägungen

S sehr gut (1)
G gut (2)
N nicht bestanden (3)

25 Beobachtungen mit der Urliste
GGNSG SGGGS GGSGG GSSNG GSSGG

Beispiel C
Kardinales Merkmal „Einkommen"

50 Beobachtungen mit der Urliste

2905	2669	3000	2798	3020	2827	2691	2200	2562	2467
2791	2969	2344	2302	2572	2614	2978	2858	2867	2353
2873	2502	2600	2770	2606	3007	2678	2504	2996	2375
2989	2306	2708	2286	2200	2507	3506	3229	3506	3252
2269	3200	2706	3582	2712	3609	2876	2826	2335	2598

1.4 Aufbereitung - Klassenbildung

Zur übersichtlichen Darstellung werden die Stichprobenwerte zunächst geordnet und merkmalsmässig gleiche Ausprägungen zusammengefasst. Die Ordnung erfolgt bei

- Nominaldaten willkürlich
- Ordinaldaten nach der Ausprägungsintensität
- Intervall- und Verhältnisdaten der Grösse nach.

Die Belegungshäufigkeit der einzelnen Ausprägungen bringt man vorerst in einer sogenannten *Strichliste* zum Ausdruck.

Beispiel A (Zivilstand)

L ||||| ||||| ||||| |||
H ||||| ||||| ||||| ||||
W |||||
G ||||| |||

Beispiel B (Noten)

S ||||| |||
G ||||| ||||| |||||
N ||

Beispiel C (Einkommen)

2200 ||
2269 |
2286 |
:
3506 ||
3582 |
3609 |

Enthält eine Stichprobe wie in Beispiel C sehr viele zahlenmässig verschiedene Werte und sind die einzelnen Ausprägungen nur schwach besetzt, so drängt sich eine *Klassenbildung* auf.

Bei der Klassenbildung wird ein geeignet gewähltes Intervall I, welches alle Beobachtungen überdeckt, in k Teilintervalle $I_j\,(j=1,2\ldots,k)$ unterteilt.

Sei

$x_{\min}$ = kleinster Wert in der Stichprobe
$x_{\max}$ = grösster Wert in der Stichprobe

R = Messbereich = $[x_{\min}, x_{\max}]$

$|R|$ = Spannweite (Range) = $x_{\max} - x_{\min}$

Das Überdeckungsintervall I ist etwas grösser als der Messbereich zu wählen.

Die Klassenbildung erfolgt nicht nach eindeutig festgelegten Regeln. Die dabei verfügbaren Freiräume sind sehr oft dafür verantwortlich, dass man der Statistik nachsagt, sie vermöge mit richtigen Zahlen Falsches und mit falschen Zahlen Richtiges zu "beweisen".

Bezüglich der *Anzahl* Klassen k existieren Faustregeln, z.B.

$$\begin{aligned} &5 \le k \approx \sqrt{n} \le 25 \\ &5 \le k \approx 5\cdot\log(n) \le 25 \qquad (1.1) \\ &5 \le k \approx 1+3.3\cdot\log(n) \le 25 \end{aligned}$$

Mit den Bezeichnungen

e_{j-1} = untere Grenze der j-ten Klasse
e_j = obere Grenze der j-ten Klasse

gilt

$$e_j - e_{j-1} = a_j = \text{Länge der } j\text{-ten Klasse}$$

$$\frac{(e_j + e_{j-1})}{2} = x_j = j\text{-te Klassenmitte}$$

Legt man das Überdeckungsintervall I mit der Länge $|I|$ symmetrisch über den Messbereich R, so gilt

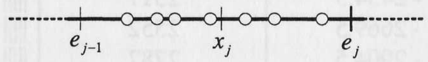
$$\begin{aligned} e_0 &= x_{\min} - \frac{(|I|-|R|)}{2} \\ e_k &= x_{\max} + \frac{(|I|-|R|)}{2} \end{aligned} \qquad (1.2)$$

woraus die übrigen Klassengrenzen

$$e_j = e_0 + \sum_{i=1}^{j} a_i \qquad (1.3)$$

folgen. Bei *konstanten* Klassenbreiten gilt $a_j = a = \dfrac{|I|}{k}$.

Die Klassenmitten $x_j\,(j=1,2,\ldots,k)$ repräsentieren die Stichprobenwerte innerhalb der j-ten Klasse. Die ursprünglichen Stichprobenwerte treten nach erfolgter Klassenbildung nicht mehr in Erscheinung; sie werden als in der Klassenmitte konzentriert betrachtet.



Mit abnehmender Anzahl Klassen vereinfacht sich zwar die Stichprobe, jedoch geht durch die Abbildung der Stichprobenwerte auf die Repräsentanten Information verloren. Andererseits lassen zu viele Klassen allfällig nichtrelevante Zufallsschwankungen zu stark hervortreten.

Aus praktischen Gründen sollten die Klassengrenzen so gelegt werden, dass sie nicht auf, sondern zwischen die beobachteten Messwerte fallen. Andernfalls ist eine Vorschrift nötig, die im Falle des Zusammenfallens festlegt, wie die Beobachtung den Nachbarklassen zuzuordnen ist (konsequent der linken, der rechten oder je hälftig den beiden Nachbarklassen). Aus diesem Grund ist das Überdeckungsintervall etwas grösser als der Messbereich zu wählen.

Klassenbildung für das Beispiel C (Einkommen)

$n = 50$

Spannweite: $|R| = 3609 - 2200 = 1409$

Überdeckungsintervall: $|I| = 1410$

Anzahl Klassen: $k = 6$ (freie Wahl)

Klassenbreite: $a = \frac{|I|}{k} = \frac{1410}{6} = 235$

Klassengrenzen:

$$e_0 = 2200 - \frac{(1410 - 1409)}{2} = 2199.5$$

$$e_j = e_0 + 235 \cdot j, \quad j = 1, 2, \ldots, 6$$

Klassenmitten: $x_j = e_{j-1} + \frac{235}{2}$

Tabellarische Zusammenstellung

Klassengrenzen	Klassenmitten	Häufigkeiten
2199.5 - 2434.5	2317	||||| |||||
2434.5 - 2669.5	2552	||||| |||||	
2669.5 - 2904.5	2787	||||| ||||| ||||
2904.5 - 3139.5	3022	||||| |||
3139.5 - 3374.5	3257	|||
3374.5 - 3609.5	3492	||||

Im Hinblick auf eine einheitliche Notation wird das zu beschreibende Datenmaterial künftig als (klassierte) Stichprobe $x_1, x_2, \ldots, x_k$ mit den Besetzungen $n_1, n_2 \ldots, n_k$ der jeweiligen Klassen bezeichnet. Falls das Datenmaterial trotzdem in Urform (d.h. nicht klassiert) vorliegt, gilt $n_1 = n_2 = \ldots = n_k = 1$ und $k = n$. Ferner wird für Daten auf mindestens ordinalem Niveau vorausgesetzt, dass die Ausprägungen $x_1, x_2, \ldots, x_k$ in natürlicher Reihenfolge geordnet sind. In Kurzform bezeichnen wir eine Stichprobe einfach mit den Klassenrepräsentanten $\{x_j, n_j\}$.

2. HÄUFIGKEITEN

2.1 Absolute und relative Häufigkeit

Für die Stichprobe $\{x_j, n_j\}$ bezeichnen:

$n_X(x_j) = n_j$ die *absolute* Häufigkeit und

$f_X(x_j) = f_j = \dfrac{n_j}{n}$ die *relative* Häufigkeit

der j-ten Merkmalsausprägung. Es gelten die Beziehungen

$$0 \le n_j \le n \quad \text{resp.} \quad 0 \le f_j \le 1$$
$$\sum n_j = n \quad \text{resp.} \quad \sum f_j = 1 \tag{2.1}$$

Beispiel B (Noten)

x_j	$n_X(x_j)$	$f_X(x_j)$
S	8	0.32
G	15	0.60
N	2	0.08
	$25 = n$	1.00

2.2 Häufigkeitsfunktion

Für Daten, welche mindestens auf dem Intervallniveau gemessen wurden, heisst

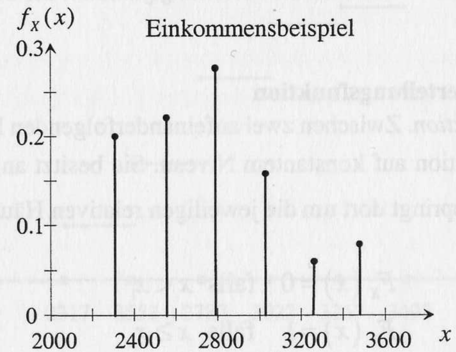
$$f_X(x) = \begin{cases} f_X(x_j) & \text{falls } x = x_j \quad j = 1, 2, \ldots, k \\ 0 & \text{sonst} \end{cases} \tag{2.2}$$

Häufigkeitsfunktion der Stichprobe.

Beispiel C (Einkommen)

x_j	$n_X(x_j)$	$f_X(x_j)$
2317	10	0.20
2552	11	0.22
2787	14	0.28
3022	8	0.16
3257	3	0.06
3492	4	0.08
$\sum$	50	1.00

Häufigkeitsfunktion $f_X(x)$



2.3 Summenhäufigkeit - Verteilungsfunktion

Während die relative Häufigkeit den Anteil der Stichprobenwerte mit einer bestimmten Merkmalsausprägung beschreibt, gibt die Summenhäufgkeit Auskunft auf die Frage nach dem Anteil der Stichprobenwerte, welche *kleiner oder gleich* einer bestimmten Ausprägung sind. Bezeichnet $\{x_j, f_j\}$ eine Stichprobe mit den relativen Häufigkeiten $f_1, f_2, \ldots, f_k$, so heisst

$$F_X(x_j) = \sum_{x_i \le x_j} f_X(x_i) \tag{2.3}$$

relative Summenhäufigkeit der j-ten Merkmalsausprägung. Die Summation erstreckt sich über die relativen Häufigkeiten jener Merkmalsausprägungen, welche den Wert x_j nicht übersteigen.

Analog zur Häufigkeitsfunktion definiert man die sogenannte Summenhäufigkeitsfunktion

$$F_X(x) = \sum_{x_i \le x} f_X(x_i) \tag{2.4}$$

$F_X(x)$ beschreibt den Anteil der Stichprobenwerte, deren Merkmalsausprägungen das Argument x nicht übersteigen. F ist, wie die Häufigkeitsfunktion f, nur sinnvoll definiert, wenn die Daten mindestens auf Intervallniveau messbar sind. Die Summenhäufigkeitsfunktion F wird auch als *Verteilungsfunktion* der Stichprobe bezeichnet.

Eigenschaften der Verteilungsfunktion

F ist eine *Treppenfunktion*. Zwischen zwei aufeinanderfolgenden Merkmalsausprägungen verläuft die Funktion auf konstantem Niveau. Sie besitzt an den Stellen x_j Unstetigkeitsstellen; sie springt dort um die jeweiligen relativen Häufigkeiten.

$$\begin{aligned} F_X(x) &= 0 \quad \text{falls } x < x_1 \\ F_X(x) &= 1 \quad \text{falls } x \ge x_k \end{aligned} \tag{2.5}$$

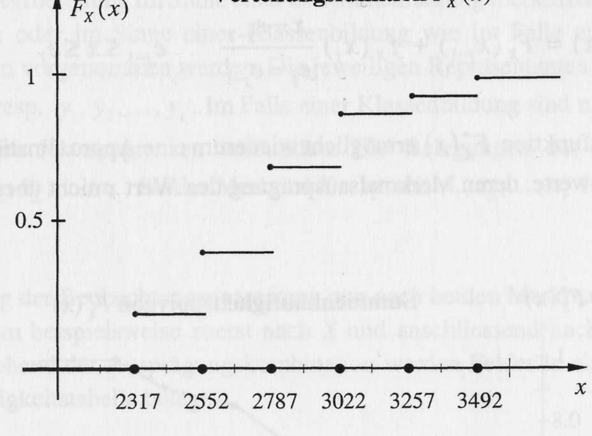
$$\begin{aligned} F_X(x_1) &= f_X(x_1) \\ F_X(x_j) &= F_X(x_{j-1}) + f_X(x_j) \end{aligned}$$

Die letzte Beziehung zeigt, wie sich Verteilungs- und Häufigkeitsfunktion auseinander ableiten lassen.

Beispiel C (Einkommen)

x_j	$n_X(x_j)$	$f_X(x_j)$	$F_X(x_j)$
2317	10	0.20	0.20
2552	11	0.22	0.42
2787	14	0.28	0.70
3022	8	0.16	0.86
3257	3	0.06	0.92
3492	4	0.08	1.00

Verteilungsfunktion $F_X(x)$



2.4 Summenhäufigkeitspolygon

Der Verteilungsfunktion F liegt ein Modell zugrunde, bei dem mit der Klassenbildung die Werte innerhalb einer Klasse auf die jeweilige Klassenmitte projiziert werden. Als Alternative dazu bietet sich ein Modell an, bei dem man die anteiligen Betrachtungen auf die jeweiligen Klassen*grenzen* bezieht und auf eine Projektion verzichtet. Für jede (obere) Klassengrenze e_j gilt stets, dass der Anteil der Stichprobenwerte, welche diese Klassengrenze nicht übersteigen, mit dem Wert der Verteilungsfunktion F an der dazugehörigen Klassenmitte x_j übereinstimmt. Ordnet man nach dieser Vorschrift allen oberen Klassengrenzen als Funktionswerte die Ausprägungen der Verteilungsfunktion F der jeweiligen Klassenmitte zu, und verbindet man die derart konstruierten Punkte geradlinig miteinander (lineare Interpolation), so entsteht ein Polygonzug, den man als Graphen der modifizierten Verteilungsfunktion F^* bezeichnet. Die Modellvoraussetzungen sind umso eher erfüllt, je mehr Werte eine Klasse enthält und je "gleichmässiger" sich die Beobachtungen innerhalb der Klasse aufteilen. In diesem Fall weist die ursprüngliche Verteilungsfunktion F entsprechend häufigere Sprungstellen mit jeweils kleineren Sprunghöhen auf.

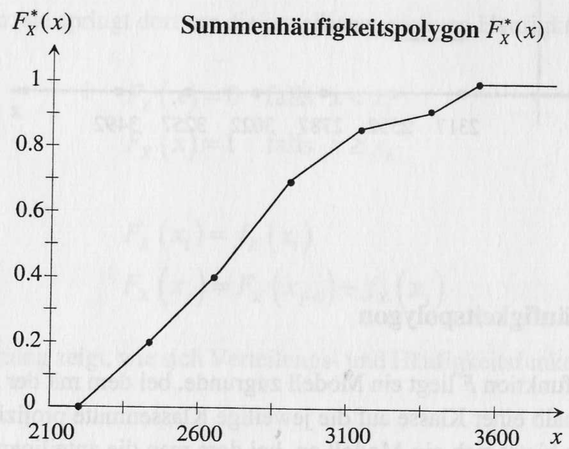
Für die funktionale Beziehung zwischen F^* und F gilt

$$F_X^*(e_j) = F_X(x_j) \qquad j=1,2,\dots,k$$

$$F_X^*(x) = F_X(x_{j-1}) + f_X(x_j)\,\frac{x-e_{j-1}}{e_j-e_{j-1}} \qquad e_{j-1} \le x \le e_j \tag{2.6}$$

Die Verteilungsfunktion $F_X^*(x)$ ermöglicht wiederum eine Approximation des Anteils der Stichprobenwerte, deren Merkmalsausprägung den Wert x nicht übersteigen.

Summenhäufigkeitspolygon $F_X^*(x)$



2.5 Mehrdimensionale Häufigkeitsverteilung

In der Praxis werden bei den statistischen Einheiten in aller Regel mehrere Merkmale gleichzeitig beobachtet. Fragebogen enthalten meistens zuviele Fragen, als dass sie vom Befragten noch sorgfältig genug beantwortet werden könnten. Die mehrdimensionale Beobachtung ist insbesondere dann nötig, wenn neben den Einzelbeschreibungen Interaktionen zwischen den Merkmalen von Interesse sind. Bei Betriebsvergleichen interessieren etwa neben der Betriebsgrösse die Finanzierung, der Jahresumsatz, der Gewinn, der Cash Flow etc, resp. deren gegenseitige Abhängigkeiten.

Die folgenden Betrachtungen beschränken sich auf zwei Dimensionen. Die dazugehörigen Merkmale werden mit X und Y bezeichnet. Die Stichprobe besteht dann aus den n Wertepaaren $(x_1, y_1),(x_2, y_2),\dots,(x_n, y_n)$. Für beide Merkmale X und Y kann getrennt eine Aufbereitung im Sinne einer Zusammenfassung merkmalsmässig gleicher Ausprägungen oder im Sinne einer Klassenbildung wie im Falle eindimensionaler Beobachtungen vorgenommen werden. Die jeweiligen Repräsentanten seien $x_1, x_2,\dots,x_r$ resp. $y_1, y_2,\dots,y_c$. Im Falle einer Klassenbildung sind mit den Klassenmitten implizite Klassengrenzen verbunden. Die Belegungen der Repräsentanten werden mit $n_1, n_2,\dots,n_r$ für das Merkmal X resp. mit $n_1, n_2,\dots,n_c$ für das Merkmal Y bezeichnet.

Die Zuordnung der Beobachtungspaare kann nun nach beiden Merkmalen gleichzeitig erfolgen, indem beispielsweise zuerst nach X und anschliessend nach Y klassifiziert wird. Entsprechend der Ausprägungskombination werden Felder in einer zweidimensionalen Häufigkeitstabelle belegt.

Die Anzahl Beobachtungen in der Klasse (x_i, y_j) sei n_{ij}. Dann heissen

$$n_{ij} = \text{absolute Häufigkeit der Klasse } (ij)$$

$$\left.\begin{aligned} n_{i.} &= \sum_j n_{ij} \\ n_{.j} &= \sum_i n_{ij} \end{aligned}\right\} \text{absolute Häufigkeit der Randklassen} \tag{2.7}$$

$$\frac{n_{ij}}{n} = f_{XY}(x_i, y_j) \text{ relative Häufigkeit der Klasse } (ij)$$

$$\left.\begin{aligned} \frac{n_{i.}}{n} &= f_X(x_i) \\ \frac{n_{.j}}{n} &= f_Y(y_j) \end{aligned}\right\} \text{relative Randhäufigkeiten} \tag{2.8}$$

Zur besseren Übersicht werden die Häufigkeiten meist in einer zweidimensionalen Tabelle dargestellt.

X \ Y	y_1 y_2 ... y_j ... y_c	n_X
x_1	n_{11} n_{12} ... n_{1j} ... n_{1c}	$n_{1.}$
x_2	n_{21} n_{22} ... n_{2j} ... n_{2c}	$n_{2.}$
.	. . . . . .	.
x_i	n_{i1} n_{i2} ... n_{ij} ... n_{ic}	$n_{i.}$
.	. . . . . .	.
x_r	n_{r1} n_{r2} ... n_{rj} ... n_{rc}	$n_{r.}$
n_Y	$n_{.1}$ $n_{.2}$ ... $n_{.j}$ ... $n_{.c}$	n

Im folgenden wird das Modell der Häufigkeitsfunktion auf den Fall zweidimensionaler Merkmale übertragen. Analog zum eindimensionalen Fall werden für beide Merkmale mindestens intervallskalierte Daten vorausgesetzt. Dann heisst die Funktion

$$f_{XY}(x, y) = \begin{cases} f_{XY}(x_i, y_j) & \text{falls } x = x_i \text{ und } y = y_j \\ 0 & \text{sonst} \end{cases} \tag{2.9}$$

gemeinsame Häufigkeitsfunktion der Merkmale *X* und *Y*. Sie gibt den Anteil der Stichprobenwerte an, deren *X*-Merkmal in der Ausprägung x_i und *Y*-Merkmal in der Ausprägung y_j vorliegt.

Die Funktion

$$F_{XY}(x, y) = \sum_{x_i \le x} \sum_{y_j \le y} f_{XY}(x_i, y_j) \tag{2.10}$$

heisst *gemeinsame* Verteilungsfunktion der Merkmale *X* und *Y* und gibt den Anteil der Stichprobenwerte an, bei denen das *X*-Merkmal den Wert *x* und gleichzeitig das *Y*-Merkmal den Wert *y* nicht übersteigt. Die Summation über die gemeinsamen relativen Häufigkeiten läuft über den durch die obigen Bedingungen ebenfalls gemeinsam abgesteckten Bereich.

Beispiel

100 Unternehmen wurden bezüglich der Merkmale Jahresumsatz in 100'000 Fr. (*X*) und Gewinn in 1'000 Fr. (*Y*) befragt.

X \ *Y*	10 (0,20]	30 (20,40]	60 (40,80]	100 (80,120]	
10 (0,20]	1	3	3	7	14
30 (20,40]	9	16	14	11	50
50 (40,60]	11	6	2	1	20
80 (60,100]	10	5	0	1	16
	31	30	19	20	100

$f_{XY}(50,60) = 2/100 = 0.02$
$f_{XY}(50,58) = 0$
$F_{XY}(50,58) = (1+3+9+16+11+6)/100 = 0.46$

Die Funktionen

$$f_X(x) = \begin{cases} f_X(x_i) & \text{falls } x = x_i \\ 0 & \text{sonst} \end{cases} \tag{2.11}$$

$$f_Y(y) = \begin{cases} f_Y(x_y) & \text{falls } y = y_j \\ 0 & \text{sonst} \end{cases} \tag{2.12}$$

heissen *Randhäufigkeitsfunktionen* von X respektive Y.

Für das obige Beispiel erhält man

x	10	30	50	80
$f_X(x)$	0.14	0.50	0.20	0.16

Bedingte Verteilung

Aus der gemeinsamen Verteilung lassen sich neben den Randverteilungen auch *innerhalb* jeder Merkmalsausprägung einer Variablen die Verteilung der anderen Variablen beschreiben. Für die Unternehmungen mit einem vorgegebenen Umsatz von 30 besitzt das Merkmal Gewinn (Y) die Verteilung

$y \mid x = 30$	10	30	60	100
$f_{Y\mid X}\left(y \mid x = 30\right)$	0.18	0.32	0.28	0.22

Die gemeinsamen Häufigkeiten werden an den jeweiligen Randhäufigkeiten gemessen.

$\frac{n_{ij}}{n_{.j}} = f_{X\mid Y}\left(x_i \mid y_j\right)$ *bedingte relative Häufigkeit von X*, falls Y bereits in der Ausprägung y_j vorliegt.

$\frac{n_{ij}}{n_{i.}} = f_{Y\mid X}\left(y_j \mid x_i\right)$ *bedingte relative Häufigkeit von Y*, falls X bereits in der Ausprägung x_i vorliegt.

Die bedingten relativen Häufigkeiten sind aus den gemeinsamen und den Randhäufigkeiten abzuleiten.

$$f_{X\mid Y}(x_i \mid y_j) = \frac{n_{ij}}{n_{.j}} = \frac{\left(\frac{n_{ij}}{n}\right)}{\left(\frac{n_{.j}}{n}\right)} = \frac{f_{XY}(x_i, y_j)}{f_Y(y_j)} \tag{2.13}$$

Eine analoge Beziehung findet man für die bedingte Verteilung von Y bei gegebenem X. Löst man nach den gemeinsamen Häufigkeiten auf, so erhält man

$$f_{XY}(x_i, y_j) = f_{X\mid Y}(x_i \mid y_j)\ f_Y(y_j) = f_{Y\mid X}(y_j \mid x_i)\ f_X(x_i) \tag{2.14}$$

Im Falle zweidimensionaler Betrachtungen sind bedingte Verteilungen stets eindimensional. Bei höherdimensionalen Betrachtungen können bedingte Verteilungen, abhängig von den gestellten Bedingungen, allerdings auch mehrdimensional werden.

Der Vergleich bedingter Verteilungen kann sehr aufschlussreich sein. Unterschiedliche Randhäufigkeiten werden bei diesen Betrachtungen ausgeschaltet.

y	10	30	60	100
$f(y \mid x{=}10)$	1/14	3/14	3/14	7/14
$f(y \mid x{=}30)$	9/50	16/50	14/50	11/50
$f(y \mid x{=}50)$	11/20	6/20	2/20	1/20
$f(y \mid x{=}80)$	10/16	5/16	0	1/16

3. GRAPHISCHE DARSTELLUNG

Graphische Darstellungen schliessen an die oben skizzierte Primäraufbereitung an. Hauptsächlich dienen sie der Visualisierung der im Beobachtungsmaterial enthaltenen Grundstrukturen im Sinne etwa der Darstellung oder des Vergleichs von Häufigkeiten oder der Entwicklung von Grössen im Zeitablauf. In den Anfängen wurde die Photographie mit dem Slogan gepriesen, dass ein Bild mehr als tausend Worte zum Ausdruck bringe. Im übertragenen Sinn gilt dies auch für graphische Darstellungen in der Statistik. Der optische Eindruck gestattet sehr rasch einen ersten *Überblick* über die Daten. Daneben kann eine geschickte Darstellung auch schon Ausgangspunkt zu weitergehenden Analysen oder zur Formulierung von Hypothesen sein. Unter dem Begriff 'Graphische Methoden' wird sogar eine ganze Technik verstanden, welche sich vornehmlich graphischer Darstellungen bedient. Insbesondere hat dieser Ansatz bei der Analyse multivariater Daten Eingang gefunden, nicht zuletzt um den unbedachten, unreflektierten oder gar blinden Einsatz multivariater Methoden zu verhindern. An dieser Stelle sei gleichzeitig auch auf die Gefahren im Missbrauch graphischer Darstellungen hingewiesen. Ein Grossteil der Skepsis gegenüber statistischen Aussagen ist diesem Umstand zuzuschreiben. Man bedenke, dass auch ein Bild mehr zu lügen imstande ist als tausend Zahlen! Irreführende und manipulierte graphische Darstellungen sind in der Praxis leider häufiger als das anzustrebende Gegenteil.

Die manuelle Erstellung von Graphiken ist meist recht mühsam. Glücklicherweise verfügen jedoch die meisten Softwarepakete über ein breites Angebot der gängigsten Darstellungen. Neben der einfachen Erfassung, Organisation und Speicherung der Daten leisten solche Pakete eine nicht zu unterschätzende zusätzliche Hilfe bei der ersten Aufbereitung.

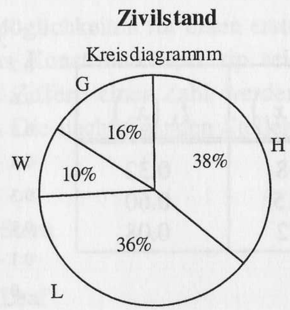
Das Spektrum der Darstellungsmöglichkeiten ist sehr breit. Neben dem Hinweis auf die zahlreich verfügbare Spezialliteratur beschränken sich die folgenden Ausführungen auf eine Auswahl von graphischen Techniken.

3.1 Kreisdiagramm

Häufigkeiten werden durch Kreissektoren dargestellt. Die Zentriwinkel α werden proportional zu den Häufigkeiten gewählt. Ihrer speziellen Form wegen sind Kreisdiagramme auch unter dem Begriff von 'Kuchendiagrammen' bekannt. Kreisdiagramme eignen sich insbesondere zur Darstellung von *Anteilen*. Ihr Einsatz ist auf jedem Messniveau möglich.

Beispiel A (Zivilstand)

x_j	$f_X(x_j)$	α_j°
L	0.36	129.6
H	0.38	136.8
W	0.10	36.0
G	0.16	57.6

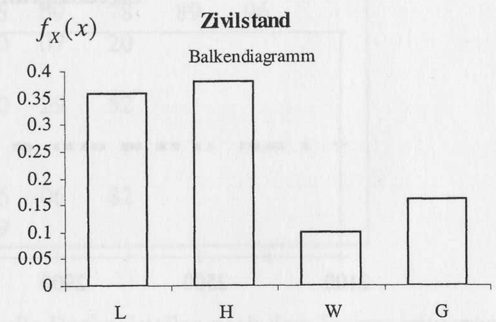


3.2 Balken- und Stabdiagramm

Die Häufigkeiten der Merkmalsausprägungen werden durch vertikale oder horizontale Balken respektive Stäbe dargestellt. Die Länge der jeweiligen Balken und Stäbe wird proportional zu den dazugehörigen relativen Häufigkeiten gewählt. Balkendiagramme werden vornehmlich bei Nominal- und Ordinaldaten angewendet, wobei die Ausprägungen in abnehmender Reihenfolge der Häufigkeiten aufgeführt werden. Die Balkenbreite wird konstant gehalten, um eine allfällige Assoziation zu einem Abstand zu vermeiden.

Beispiel A (Zivilstand)

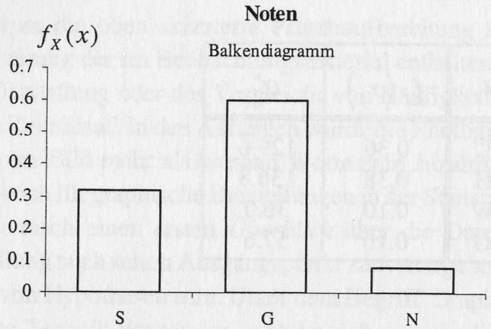
x_j	$n_X(x_j)$	$f_X(x_j)$
H	19	0.38
L	18	0.36
G	8	0.16
W	5	0.10



Bei Ordinaldaten kann mit der Reihenfolge die Stärke der Ausprägungsintensität zum Ausdruck gebracht werden. Für reine Ordinaldaten ist ebenfalls kein Abstand definiert, sodass man, wie im Falle von Balkendiagrammen, Stäbe von konstanter Breite zur Darstellung der relativen Häufigkeiten wählt.

Beispiel B (Noten)

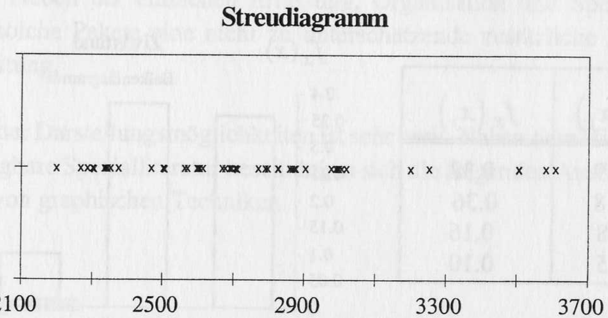
x_j	$n_X(x_j)$	$f_X(x_j)$
S	8	0.32
G	15	0.60
N	2	0.08



3.3 Streudiagramm

Für Intervall- und Verhältnisdaten sind Abstände zwischen den Merkmalsausprägungen definiert. Die einzelnen Beobachtungen lassen sich somit als Punkte auf der reellen Achse darstellen. Zweckmässigerweise erfolgt diese Darstellung vor einer eventuell notwendigen Klassenbildung.

Beispiel C (Einkommen)



Streudiagramme sind sehr einfach zu konstruieren. Sie bringen den Messbereich, Extremwerte sowie allfällig vorhandene Konzentrationen augenfällig zum Ausdruck. Der Stichprobenumfang sollte nicht allzu gross sein.

3.4 Stem-and-Leaf Diagramm

Stem-and-leaf Diagramme bieten effiziente Möglichkeiten für einen ersten Überblick über die Verteilung *numerischer* Daten. Das Konstruktionsprinzip sei anhand des Einkommensbeispiels dargestellt. Die ersten Ziffern einer Zahl werden primär zur Ordnung des Beobachtungsmaterials benutzt. Die nachfolgenden Stellen bringen die Häufigkeiten zum Ausdruck.

29 Stem

2905 —> 29 | 05

05 Leaf

Für das gesamte Einkommensbeispiel erhält man nach diesem Muster das folgende Stem-and-Leaf Diagramm

Tiefe		Stem	Leaf					
(50)	4	22	00	00	69	86		
(46)	10	23	02	06	35	44	53	75
(40)	11	24	67					
(39)	17	25	02	04	07	62	72	98
(33)	23	26	00	06	14	69	78	91
27	(29)	27	06	08	12	70	91	98
21	(35)	28	26	27	58	67	73	76
15	(40)	29	05	69	78	89	96	
10	(43)	30	00	07	20			
		31						
7	(46)	32	00	29	52			
		33						
		34						
4	(49)	35	06	06	82			
1	(50)	36	09					

Es brauchen nicht notwendigerweise alle Dezimalstellen nach dem Stamm mitgenommen zu werden. Zur besseren Übersicht erweist es sich oft als zweckmässig, nur eine Stelle nach dem Stamm in das Diagramm zu übernehmen.

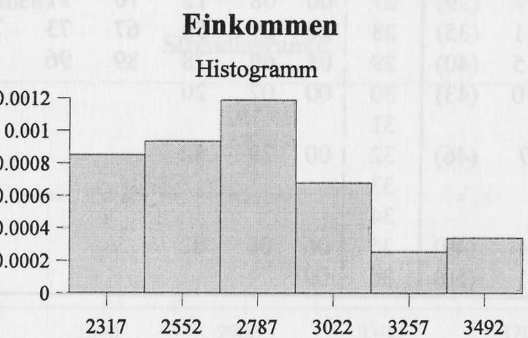
Stem-and-Leaf Diagramme liefern vor allem Informationen über die Symmetrie einer Verteilung, über allfällige Ausreisser, über Häufungspunkte und über eventuell nicht besetzte Bereiche. Die Konstruktion bietet keine Schwierigkeiten, vor allem, wenn das Sortierproblem computermässig gelöst werden kann.

Eine weitere Information, welche sofort aus stem-and-leaf Diagrammen abgeleitet werden kann, ist die sogenannte *Tiefe* einer Beobachtung. Sie gibt an, wie weit eine Beobachtung von den Extrempunkten in Richtung des Zentrums zu liegen kommt. Zu diesem Zweck werden die Stichprobenwerte über die Stämme von oben nach unten und umgekehrt ausgezählt. Der kleinere der beiden Ränge wird als Tiefe bezeichnet.

3.5 Histogramm

Die Idee von Stabdiagrammen lässt sich auf die graphische Darstellung von relativen Häufigkeiten für Intervall- und Verhältnisdaten übertragen. Zusätzlich wird die Tatsache berücksichtigt, dass für Daten auf diesem Niveau Abstände zwischen den Merkmalsausprägungen definiert sind. Analog wie beim Stabdiagramm wählt man zur Darstellung der relativen Häufigkeit *Rechtecke*, wobei allerdings nicht die Höhe, sondern deren *Fläche* proportional zur entsprechenden relativen Häufigkeit gewählt wird.

Beispiel C (Einkommen)

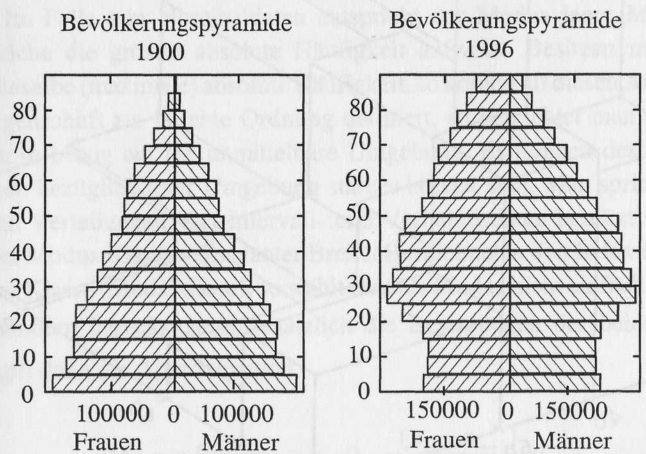


Das Konstruktionsprinzip für Histogramme macht die Darstellung der relativen Häufigkeiten insbesondere unabhängig von den Klassenbreiten. Im Falle einer Zusammenfassung von zwei Nachbarklassen zu einer einzigen Klasse würde bei 'höhentreuer' Darstellung relativer Häufigkeiten der optische Eindruck durch die addierte Klassenhöhe und die vereinigte Klassenbreite überzeichnet.

Bei konstanten Klassenbreiten sind neben den Flächen auch die Rechteckshöhen proportional zu den dazugehörigen relativen Häufigkeiten.

3.6 Spezielle Graphiken

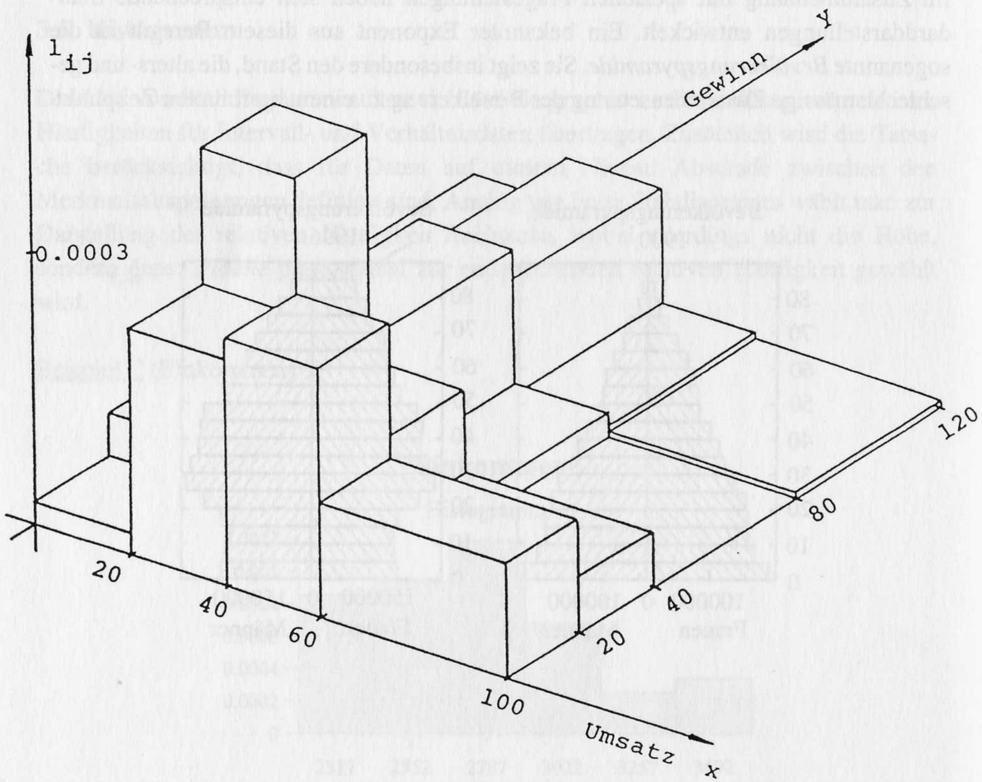
Im Zusammenhang mit speziellen Fragestellungen haben sich entsprechende Standarddarstellungen entwickelt. Ein bekannter Exponent aus diesem Bereich ist die sogenannte *Bevölkerungspyramide*. Sie zeigt insbesondere den Stand, die alters- und geschlechtsmässige Zusammensetzung der Bevölkerung zu einem bestimmten Zeitpunkt.



3.7 Verbundene Häufigkeitsverteilung

Im Falle verbundener Merkmale stellen sich keine prinzipiell neuen Probleme. Für zwei Merkmale wird unter Berücksichtigung des Messniveaus eine entsprechend zweidimensionale Darstellung im Raum gewählt.

Für das Umsatz-Gewinnbeispiel ergibt sich folgende Graphik der gemeinsamen relativen Häufigkeiten in Histogrammform.



4. STICHPROBENPARAMETER

Ein weiterer Schritt in der Aufbereitung des Beobachtungsmaterials nach tabellarischen und graphischen Darstellungen besteht in der Charakterisierung der Stichprobe durch wenige, aber möglichst aussagefähige Parameter. Man nennt solche Kenngrössen statistische Masszahlen oder Stichprobenparameter. Das Beobachtungsmaterial wird auf wenige Parameter verdichtet. Die nachfolgende Auswahl an Stichprobenparametern ist keineswegs abschliessend. Die Ordnung entspricht den Einsatzmöglichkeiten aufgrund des Messniveaus der Daten.

4.1 Lageparameter

4.1.1 Modus

Der Modus ist die Merkmalsausprägung einer Stichprobe mit der grössten relativen Häufigkeit. Im Falle von Nominaldaten entspricht der Modus jener Merkmalsausprägung, welche die grösste absolute Häufigkeit aufweist. Besitzen mehrere Ausprägungen dieselbe (maximale) absolute Häufigkeit, so kommt all diesen Ausprägungen die Modaleigenschaft zu. Ist eine Ordnung definiert, so betrachtet man die relativen Häufigkeiten in bezug auf die unmittelbare Umgebung. Es können deshalb mehrere Ausprägungen bezüglich ihrer Umgebung stärker besetzt sein. Man spricht dann von multimodalen Verteilungen. Bei Intervall- und Verhältnisdaten benützt man zur Bestimmung des Modus Klassen konstanter Breite. Bezeichnet m den Index der am stärksten besetzten Klasse (Modalklasse), so wählt man als Modus nicht einfach die Klassenmitte x_m, sondern berücksichtigt zusätzlich die Besetzungen der beiden Nachbarklassen. Es gilt dann die Approximation

$$Mo = e_{m-1} + (x_m - x_{m-1}) \frac{(n_m - n_{m-1})}{2n_m - n_{m-1} - n_{m+1}} \tag{4.1}$$

Im Einkommensbeispiel führt die Approximation zum Modus

$$Mo = 2669.5 + 235 \frac{(14-11)}{28-11-8} = 2747.8 \tag{4.2}$$

Der Modus ist für alle Messniveaus anwendbar. Er ist robust gegenüber Ausreissern, reagiert auf Klassenbildung und liefert vor allem Informationen über Häufungspunkte.

4.1.2 Quantile

Die Quantile $q_{ir}\ (i = 1, 2, \ldots, r-1)$ sind jene (r-1) Werte, welche eine geordnete Stichprobe so in r Abschnitte aufteilen, dass jeder Abschnitt denselben Anteil der Stichprobenwerte enthält. Gebräuchliche Werte für r sind:

$r = 2$	→	Median
$r = 4$	→	Quartile
$r = 10$	→	Dezile
$r = 100$	→	Perzentile

Der Median teilt die geordnete Stichprobe in zwei gleich grosse Hälften. Aufeinanderfolgende Quartile legen Intervalle fest, welche je 25% der Stichprobenwerte enthalten.

Für *Urdaten* erhält man die Quantile durch Abzählen in der aufsteigend geordneten Stichprobe $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Das i-te r-Quantil q_{ir} wird nach folgender Regel bestimmt:

$$q_{ir} = x_{(j+1)} \qquad \text{falls } n\frac{i}{r} \text{ nicht ganzzahlig}$$

$$q_{ir} = \frac{\left(x_{(j)} + x_{(j+1)}\right)}{2} \qquad \text{falls } n\frac{i}{r} \text{ ganzzahlig} \tag{4.3}$$

mit

$$j = \text{int}\left(n\frac{i}{r}\right) \qquad \left[\text{ganzzahliger Teil von } n\frac{i}{r}\right] \tag{4.4}$$

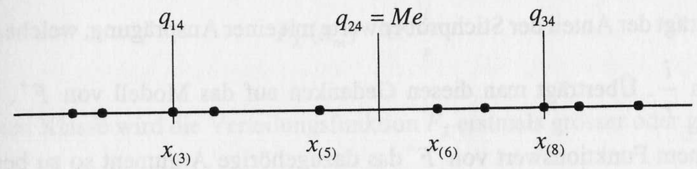
Das nach dieser Regel berechnete i-te r-Quantil besagt:

- der Anteil der Stichprobenwerte kleiner (<) als q_{ir} beträgt höchstens $\frac{i}{r}$ und
- der Anteil der Stichprobenwerte grösser (>) als q_{ir} beträgt höchstens $1-\frac{i}{r}$.

Graphische Motivation für eine Stichprobe vom Umfang n=10.

$$n = 10; \quad q_{14} = x_{(3)}; \quad q_{24} = \frac{\left(x_{(5)} + x_{(6)}\right)}{2}; \quad q_{34} = x_{(8)} \tag{4.5}$$

Quartile



<u>Beispiel C</u> (Einkommen)

2200	2200	2269	2286	2302	2308	2335	2344	2353	2375
2467	2502	2507	2534	2562	2572	2598	2600	2606	2614
2669	2678	2691	2706	2708	2712	2770	2791	2798	2826
2827	2858	2867	2873	2876	2905	2969	2978	2989	2996
3000	3007	3020	3200	3229	3252	3506	3506	3582	3609

Median:

$$Me = q_{12} = \frac{\left(x_{(25)} + x_{(26)}\right)}{2} = \frac{(2708 + 2712)}{2} = 2710 \tag{4.6}$$

Quartile:

$$q_{14} = x_{(13)} = 2507$$

$$q_{24} = q_{12} = 2710 = \frac{(2708 + 2712)}{2} \tag{4.7}$$

$$q_{34} = x_{(38)} = 2978$$

Dezile:

$$q_{1,10} = \frac{\left(x_{(5)} + x_{(6)}\right)}{2} = \frac{(2302 + 2308)}{2} = 2305 \tag{4.8}$$

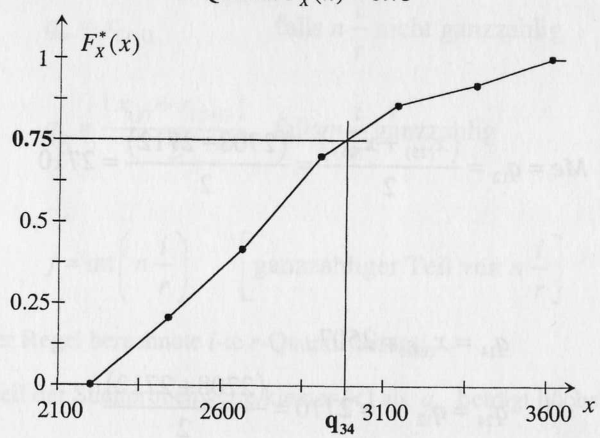
Liegt das Datenmaterial *nach erfolgter Klassenbildung* vor, bestimmt man das *i*-te *r*-Quantil q_{ir} aus der Verteilungsfunktion F^* (Summenhäufigkeitspolygon). Definitionsgemäss beträgt der Anteil der Stichprobenwerte mit einer Ausprägung, welche q_{ir} nicht übersteigen $\frac{i}{r}$. Überträgt man diesen Gedanken auf das Modell von F^*, so ist zu vorgegebenem Funktionswert von F^* das dazugehörige Argument so zu bestimmen, dass gilt

$$F_X^*(x) = F_X(x_{j-1}) + f_X(x_j)\frac{x - e_{j-1}}{e_j - e_{j-1}} = \frac{i}{r} \tag{4.9}$$

Die Lösung nach x liefert das *i*-te *r*-Quantil q_{ir}.

Quartile

3. Quartil: $F_X^*(x) = 0.75$



Die Verteilungsfunktion F^* ist intervallweise für jede Klasse definiert. Zunächst ist somit der Index jener Klasse zu bestimmen, innerhalb welcher sich das *i*-te *r*-Quantil befindet. Bezeichnet *m* den Index dieser Klasse, so muss gelten

$$\begin{aligned} F_X(x_{m-1}) &< \frac{i}{r} \\ F_X(x_m) &\geq \frac{i}{r} \end{aligned} \tag{4.10}$$

In der *m*-ten Klasse wird die Verteilungsfunktion F_X erstmals grösser oder gleich $\frac{i}{r}$.

Dann folgt für das *i*-te *r*-Quantil nach dem Strahlensatz

$$q_{ir} = e_{m-1} + a_m \frac{\left(\frac{i}{r} - F_X(x_{m-1})\right)}{f_X(x_m)} \tag{4.11}$$

<u>Beispiel C</u> (Einkommen)

j	e_{j-1}	e_j	$f_X(x_j)$	$F_X(x_j)$
1	2199.5	2434.5	0.20	0.20
2	2434.5	2669.5	0.22	0.42
3	2669.5	2904.5	0.28	0.70
4	2904.5	3139.5	0.16	0.86
5	3139.5	3374.5	0.06	0.92
6	3374.5	3609.5	0.08	1.00

$$\begin{aligned} q_{14} &= 2434.5 + 235\frac{(0.25-0.20)}{0.22} = 2487.91 \\ q_{6,10} &= 2669.5 + 235\frac{(0.60-0.42)}{0.28} = 2820.57 \end{aligned} \tag{4.12}$$

Eigenschaften der Quantile

- Quantile verlangen mindestens *Ordinaldaten*
- Quantile reagieren nicht auf *Ausreisser*
- Für den Median gilt:

$$\sum n_j \,|\, x_j - a \,| \;\rightarrow\; \text{min,} \quad \text{falls} \quad a = q_{12} \tag{4.13}$$

- Werden die Stichprobenwerte streng monoton *transformiert*, so ändert zwar der Zahlenwert, nicht aber die relative Position der Quantile. Der Median wird z.B. wieder in den Median transformiert.

- Bei der *Vereinigung* von zwei Stichproben zu einer einzigen, lassen sich die Quantile der vereinigten Stichprobe nicht aus den Quantilen der Teilstichproben ableiten.

- Mit den Quantilen können durch Durchschnittsbildung neue Lageparameter gebildet werden. Unter dem *Quantilsmittel* versteht man beispielsweise den Ausdruck

$$\frac{\left(q_{1r}+q_{r-1,r}\right)}{2} \tag{4.14}$$

  Gebräuchlich sind vor allem Quartils- und Dezilmittel.

- Der *Abstand* aufeinanderfolgender Quantile ermöglicht Rückschlüsse auf die Streuung der Stichprobenwerte in den dazugehörigen Intervallen. Je näher die Quantile beieinander liegen, desto enger sind die Stichprobenwerte innerhalb dieser Grenzen zusammengepackt.

4.1.3 Das arithmetische Mittel

Das arithmetische Mittel einer Stichprobe $\left\{x_j, n_j\right\}$ ist definiert als

$$\bar{x}=\frac{1}{n}\sum_{j=1}^{k} n_j x_j=\sum_{j=1}^{k} f_X(x_j)x_j \tag{4.15}$$

Beispiel C (Einkommen)

$\bar{x}$ = 1/50·(10·2317 + 11·2552 + ... + 4·3492) = 2763.50

Teilt man die gesamte Lohnsumme *zu gleichen Teilen* unter die 50 Personen auf, so erhält jeder Fr. 2763.50.

Eigenschaften des arithmetischen Mittels

- Das arithmetische Mittel setzt mindestens Intervalldaten voraus.
- Das arithmetische Mittel ist eine gewichtete Summe der Merkmalsausprägungen. Die Gewichte entsprechen den relativen Häufigkeiten.

- $n\bar{x}=\sum n_j x_j$ (Ersetzungssatz)
- $\sum n_j (x_j-a)^2 = f_X(a) \rightarrow$ min, falls $a=\bar{x}$
- Bei einer *linearen Transformation* der Stichprobenwerte erfährt das arithmetische Mittel dieselbe Transformation.

$$y_j=\frac{x_j-c_2}{c_1} \qquad x_j=c_1 y_j+c_2$$

$$\begin{aligned}\bar{y}&=\frac{1}{n}\sum_{j=1}^{k} n_j y_j=\frac{1}{n}\sum_{j=1}^{k} n_j\left(\frac{x_j-c_2}{c_1}\right)\\&=\frac{1}{c_1}\left[\frac{1}{n}\sum_{j=1}^{k} n_j x_j-c_2\right]\\&=\frac{\bar{x}-c_2}{c_1}\end{aligned} \tag{4.16}$$

- Werden k Teilstichproben mit den arithmetischen Mitteln $\left\{\bar{x}_i\right\}$ $(i=1,2,\ldots,k)$ zu einer Gesamtstichprobe mit dem arithmetischen Mittel $\bar{x}$ vereinigt, so gilt:

$$\bar{x}=\frac{\left(n_1\bar{x}_1+n_2\bar{x}_2+\ldots+n_k\bar{x}_k\right)}{n} \tag{4.17}$$

  wobei $n=\sum n_j$.

- $\bar{x}$ ist einfach zu berechnen und intuitiv verständlich. Sämtliche ursprünglichen Stichprobenwerte gehen mit demselben Gewicht in die Berechnung ein, worauf die Anfälligkeit gegenüber Ausreissern zurückzuführen ist.

4.1.4 Das geometrische Mittel

Das geometrische Mittel ist die n-te Wurzel aus dem Produkt der Stichprobenwerte. Typisches Beispiel eines geometrischen Mittels ist die mittlere Zuwachsrate eines Kapitals im Zeitablauf bei unterschiedlichen Zuwachsraten (Zinseszinsrechnung).

Das Kapital K_0 entwickle sich mit den Zuwachsraten $r_1, r_2 \ldots, r_n$. Das Endkapital nach n Jahren beträgt

$$K_n = K_0 \cdot r_1 \cdot r_2 \ldots r_n \tag{4.18}$$

Um gleichzeitig vom selben Anfangskapital zum selben Endkapital bei *konstanten Zuwachsraten* zu kommen, muss gelten

$$K_n = K_0 \cdot r \cdot r \ldots r = K_0 \cdot r^n \tag{4.19}$$

woraus für r folgt

$$r = (r_1 \cdot r_2 \cdot \ldots \cdot r_n)^{1/n} \tag{4.20}$$

Allgemein gilt für das geometrische Mittel der Stichprobe $\{x_j, n_j\}$ $(x_j > 0)$

$$g = \left(x_1^{n_1} \cdot x_2^{n_2} \cdot x_k^{n_k}\right)^{1/n} \tag{4.21}$$

- Das geometrische Mittel ist nur für *positive Verhältnisdaten* definiert.
- Durch das geometrische Mittel werden relative Änderungen der Merkmalsausprägungen gemittelt. Die Unterschiede zwischen den Stichprobenwerten werden nicht durch die Differenz (wie beim arithmetischen Mittel), sondern durch ihr Verhältnis ausgedrückt.
- Der Logarithmus des geometrischen Mittels ist das arithmetische Mittel der Logarithmen der Stichprobenwerte.

Beispiel

Die Bevölkerung einer Stadt habe sich gemäss folgender Tabelle entwickelt.

Zeit	Bestand	Änderungsrate gegenüber Vorjahr
0	76396	
1	74882	0.9802
2	73875	0.9865
3	73061	0.9890

Den unterschiedlichen Veränderungsraten entpricht eine konstante Rate von $(0.9802 \cdot 0.9865 \cdot 0.9890)^{1/3} = 0.98523$, welche in der gleichen Zeit vom selben Anfangs- in denselben Endzustand führt.

4.1.5 Das harmonische Mittel

Kauft jemand beim Tanken konstant 30 Liter Benzin, so berechnet sich der Durchschnittspreis als arithmetisches Mittel der jeweils bezahlten Literpreise. Anders verhält es sich bei einer Strategie, wonach jeweils für einen konstanten Betrag B Benzin eingekauft wird. Für den mittleren Literpreis gilt dann

$$\overline{p} = \frac{\text{totale Beträge}}{\text{totale Menge}} = \frac{nB}{\frac{B}{p_1} + \frac{B}{p_2} + \ldots + \frac{B}{p_n}} = \frac{n}{\frac{1}{p_1} + \frac{1}{p_2} + \ldots + \frac{1}{p_n}} \tag{4.22}$$

$\overline{p}$ heisst harmonisches Mittel der n Beobachtungen $p_1, p_2, \ldots, p_n$.

Allgemein definiert für eine Stichprobe $\{x_j, n_j\}$ $(x_j > 0)$

$$h = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_k}{x_k}} = \frac{1}{\frac{f_X(x_1)}{x_1} + \frac{f_X(x_2)}{x_2} + \ldots + \frac{f_X(x_k)}{x_k}} \tag{4.23}$$

das harmonische Mittel.

Das harmonische Mittel kommt vor allem für Daten mit "gebrochener Dimension" der Merkmale (Fr./Liter, Fr./Stück, km/h) in Frage. Bei Konstanz der Nennerdimension berechnet sich das Mittel gewöhnlich nach dem arithmetischen und bei Konstanz der Zählerdimension nach dem harmonischen Modell.

4.1.6 Lageregeln einiger Mittelwerte

Ist für eine Stichprobe $\{x_j, n_j\}$ sowohl das arithmetische, das geometrische und das harmonische Mittel definiert, so stehen sie in der Relation

harmonisches $\le$ geometrisches $\le$ arithmetisches Mittel

zueinander. Sie stimmen nur dann überein, wenn alle Stichprobenwerte dieselbe Merkmalsausprägung aufweisen.

Aus dem Zusammenspiel von Modus Mo, Median $Me = q_{12}$ und arithmetischem Mittel $\overline{x}$ kann auf die *Form* der Verteilung geschlossen werden.

Ist die Verteilung *symmetrisch*, so stimmen die drei obigen Lageparameter (*Mo*, *Me* und $\overline{x}$) überein. Umgekehrt kann bei Gleichheit der Parameter jedoch nur geschlossen werden, dass diese Konstellation auf eine symmetrische Verteilung hindeute. Der Schluss ist hingegen nicht zwingend.

Im Falle *rechtsschiefer* Verteilungen gilt für die Parameter die Relation

Modus ≤ Median ≤ arithmetisches Mittel.

Auch hier gilt die obige Bemerkung im Falle der Umkehrung. Aus der Reihenfolge der Parameter kann wiederum nicht eindeutig auf die Form der Verteilung geschlossen werden.

Für *linksschiefe* Verteilungen gilt schliesslich die Reihenfolge

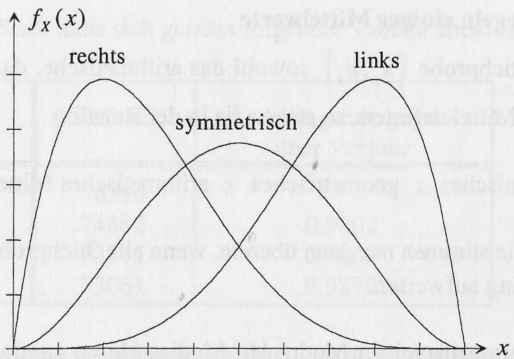
arithmetisches Mittel ≤ Median ≤ Modus.

Ferner gilt

$$g \cong \overline{x}\left(1-\frac{1}{2}\frac{s^2}{\overline{x}^2}\right) \tag{4.24}$$

wobei $s^2 = \dfrac{1}{n-1}\sum(x_i - \overline{x})^2$

Schiefe



4.2 Streuungsparameter

Lageparameter liefern Informationen über die Lage des zu beschreibenden Beobachtungsmaterials als Ganzes. Sie erlauben jedoch keine Aussagen über die gegenseitige Lage der Einzelwerte. Rein intuitiv wird man die beiden Stichproben

Stichprobe 1: 1, 3, 4, 4, 4, 6, 6
Stichprobe 2: -100, -50, -20, 4, 50, 70, 74

als verschieden betrachten, obschon sie bei gleichem Stichprobenumfang dasselbe arithmetische Mittel und denselben Median aufweisen. Der offensichtliche Unterschied zwischen den beiden Stichproben besteht darin, dass die *Abstände* der Einzelwerte wesentlich verschieden sind. Die 'Streuung' ist im ersten Fall bedeutend kleiner als im zweiten. Für diese Art der Charakterisierung sind neue Masszahlen, sogenannte Streuungsparameter notwendig. Auch hier sind, abhängig vom Messniveau, verschiedene Konzepte möglich. Wir beschränken uns auf Streuungsmasse für Daten, welche mindestens auf dem Intervallniveau messbar sind. Konzeptionell konzentrieren wir uns auf Masszahlen, welche entweder

- auf dem Abstand geeigneter *Ranggrössen* oder
- den *Abständen von einem Lageparameter*

aufbauen. Eine zentrale Stellung besitzt die Forderung der sogenannten Translationsinvarianz, welche besagt, dass Streuungsparameter auf eine reine Translation $f: x \mapsto x-a, a \in \mathbb{R}$, ("Parallelverschiebung") der Daten nicht reagieren sollen, zumal die gegenseitige Lage der Punkte nicht verändert wird.

4.2.1 Spannweite

Die Spannweite R ist der Abstand zwischen dem grössten und dem kleinsten Stichprobenwert.

$$R = x_{\max} - x_{\min}$$

Die Spannweite ist insofern eine recht 'grobe' Masszahl, als sie sich nur auf die beiden Extremwerte einer Stichprobe abstützt. Daraus resultiert eine meist unerwünschte Anfälligkeit gegenüber Ausreissern. Dieser Umstand zeichnet dafür verantwortlich, dass die Spannweite nur selten als Streuungsmass Anwendung findet. Im Regelfall werden die Extremwerte nur als Informationsparameter aufgeführt.

Praktische Relevanz besitzt die Spannweite etwa bei Börsenkursen einzelner Titel eines Handelstages oder bei der Aufzeichnung von Tagestemperaturen.

4.2.2 Quantilsabstand

Die Anfälligkeit der Spannweite gegenüber Ausreissern wird weitgehend ausgeschaltet, falls anstelle von Extremwerten Quantile miteinander verglichen werden.

$$d_r = \left(q_{r-1,r} - q_{1,r}\right) \tag{4.25}$$

heisst Quantilsabstand. Er bezeichnet die Länge jenes Intervalls, welches $100(1\text{-}2/r)\%$ der 'mittleren' Stichprobenwerte überdeckt. Im Gegensatz zur Spannweite ist der Quantilsabstand robust gegenüber Ausreissern.

Eine wichtige Anwendung von Streuungsmassen, basierend auf dem Konzept von Quartilen, sind die sogenannten *Box-Plots*. Die gegenseitige Lage der Extremwerte und Quartile, ergänzt durch "Ausreissergrenzen" G_U und G_O, werden in einer gemeinsamen Graphik zusammengefasst. Für die untere und obere Ausreissergrenze gilt:

$$G_U = q_{14} - 1.5\, d_4 \qquad G_O = q_{34} + 1.5\, d_4 \tag{4.26}$$

Stichprobenwerte, welche vom 1. Quartil um mehr als des 1.5-fache des Quartilsabstandes d_4 nach unten abweichen, werden als Ausreisser bezeichnet und verlangen eine eingehendere Beurteilung. Dasselbe gilt für Stichprobenwerte oberhalb von G_O. Stichprobenwerte ausserhalb der Ausreissergrenzen werden in Softwarepaketen speziell markiert.

Die wesentlichen Bausteine von Box-Plots sind die Quartile, welche ihrerseits gegenüber Ausreissern ein robustes Verhalten zeigen. Diese Eigenschaft kann entsprechend auf Box-Plots übertragen werden. Ausreisser gehen jedoch nicht verloren; sie werden sogar noch speziell markiert. Zur Beliebtheit von Box-Plots trägt weiter bei, dass in derselben Graphik sowohl Lageparameter (Quartile) als auch Streuungsparameter (Abstände der Quartile) zum Ausdruck kommen.

Besonders anschaulich zeigen sich die Eigenschaften von Box-Plots am Beispiel mittlerer Tagestemperaturen.

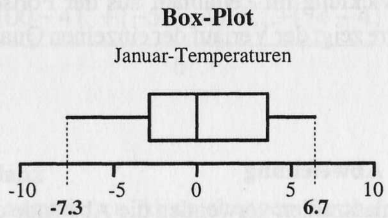
Die Originaldaten (Januartemperaturen)

6.7, 2.5, 1.1, 1.1, -1, -0.9, -1.4, -0.5, -0.3, 0.4, 0.1, -1.4, -2.6, -4.3, -0.1, -0.9, -2.2, -7.3, -3.2, 2, -2, -1.5, 3.9, 2.8, 3.1, 6.2, 4.6, 4.9, 2.6, 0.7, 0.4

werden zunächst geordnet

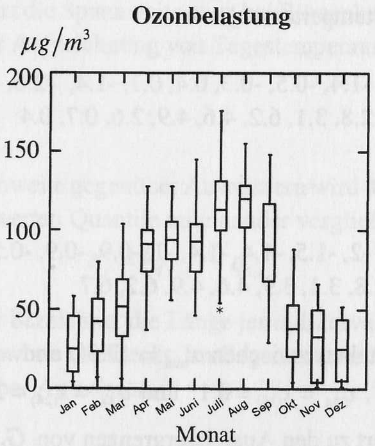
-7.3, -4.3, -3.2, -2.6, -2.2, -2, -1.5, -1.4, -1.4, -1, -0.9, -0.9, -0.5, -0.3, -0.1, 0.1, 0.4, 0.4, 0.7, 1.1, 1.1, 2, 2.5, 2.6, 2.8, 3.1, 3.9, 4.6, 4.9, 6.2, 6.7

Die Temperaturen schwanken zwischen $x_{\min} = -7.3°$ und $x_{\max} = 6.7°$. Die Quartile betragen $q_{14} = x_{(8)} = -1.4°$, $q_{24} = x_{(16)} = 0.1°$ und $q_{34} = x_{(24)} = 2.6°$. Der Quartilsabstand von $2.6°-(-1.4°) = 4°$ führt zu den Ausreissergrenzen von $G_U = -7.4°$ und $G_O = 8.6°$. In einer Graphik zusammengefasst erhält man schliesslich die Darstellung



Die Box enthält die mittleren 50% der Stichprobenwerte. Unterhalb und oberhalb der Box liegen somit ebenfalls je 25% der kleinsten resp. grössten Stichprobenwerte. SYSTAT markiert Stichprobenwerte ausserhalb der Ausreissergrenze mit speziellen Symbolen (*(outside value), ∘(far outside value)). In diesen Fällen werden anstelle von $x_{\min}$ oder $x_{\max}$ die Extremwerte innerhalb der Ausreissergrenzen als Endpunkte der Box-Plots verwendet.

Box-Plots können auch zu Vergleichszwecken herangezogen werden. Die nachfolgende Graphik enthält monatliche Box-Plots der Ozonbelastung im Verlaufe des Jahres.



Aus der Graphik erkennt man neben den monatlichen Informationen (Extremwerte, Niveau, Streuung) die Entwicklung im Zeitablauf aus der Fortschreibung ausgewählter Kenngrössen. Insbesondere zeigt der Verlauf der einzelnen Quartile eine offensichtliche Temperaturabhängigkeit.

4.2.3 Mittlere absolute Abweichung

Die künftigen Streuungsmasszahlen verwenden die Abstände der Stichprobenwerte zu einem Lageparameter. Die Abstände werden normalerweise durch das Quadrat der Differenzen oder durch die absoluten Abweichungen gemessen.

$$m_a = \frac{1}{n}\sum_{j=1}^{k} n_j \left| x_j - a \right| \tag{4.27}$$

heisst *mittlere absolute Abweichung von a*.

- Obschon für einen gegebenen Datensatz die Summe der absoluten Abweichungen vom Median minimal wird, werden die Abweichungen gewöhnlich gegenüber dem arithmetischen Mittel $\overline{x}$ bestimmt.

- Die absoluten Abweichungen sowohl vom Median als auch vom arithmetischen Mittel erfüllen die Bedingung der Translationsinvarianz.

4.2.4 Stichprobenvarianz

Die Stichprobenvarianz, bekannt auch unter dem Begriff der mittleren quadratischen Abweichung, ist definiert als

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j\left(x_j - \overline{x}\right)^2 = \frac{n}{n-1}\sum_{j=1}^{k} f_X(x_j)\left(x_j - \overline{x}\right)^2 \tag{4.28}$$

Die positive Quadratwurzel s der Stichprobenvarianz heisst *Standardabweichung*.

Für die eingangs erwähnten Stichproben 1 und 2 (Seite 37) manifestiert sich die unterschiedliche Streuung (bei gleichem arithmetischem Mittel von 4) durch die Masszahlen

$$s_1^2 = \frac{\left[(1-4)^2 + (3-4)^2 + \ldots + (6-4)^2\right]}{6} = 3 \tag{4.29}$$

und

$$s_2^2 = \frac{\left[(-100-4)^2 + (-50-4)^2 + \ldots + (74-4)^2\right]}{6} = 4280 \tag{4.30}$$

Eigenschaften der Varianz

- Die Division der summierten Abweichungen durch (n-1) anstatt durch n, wie intuitiv zu erwarten wäre, wird im Rahmen der induktiven Statistik im Zusammenhang mit den Eigenschaften von Schätzfunktionen begründet.

- $s^2 = 0 \quad \Leftrightarrow \quad x_1 = x_2 = \ldots = x_n$

- s^2 reagiert auf Ausreisser

- $\sum n_j\left(x_j - a\right)^2 \;\rightarrow\; \min \quad \text{falls} \quad a = \overline{x}$

- Lineare Transformation

$$y_j = \frac{x_j - c_2}{c_1} \qquad x_j = c_1 y_j + c_2 \tag{4.31}$$

$$s_y^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j\left(y_j - \overline{y}\right)^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j\left(\frac{x_j - c_2}{c_1} - \frac{\overline{x} - c_2}{c_1}\right)^2 = \frac{1}{c_1^2} s_x^2 \qquad (4.32)$$

- Die konkrete Berechnung der Stichprobenvarianz nach der obigen Definition macht es notwendig, zuerst das arithmetische Mittel zu berechnen, um nachher in einem zweiten Durchgang die Summe der Abweichungsquadrate zu bilden. Durch eine einfache Umformung lässt sich diese 'Doppelspurigkeit' vermeiden.

$$s^2 = \frac{1}{n-1}\left[\sum_{j=1}^{k} n_j x_j^2 - n\overline{x}^2\right] = \frac{1}{n-1}\left[\sum_{j=1}^{k} n_j x_j^2 - \frac{\left(\sum n_j x_j\right)^2}{n}\right] \qquad (4.33)$$

Die meisten Taschenrechner verfügen über feste Programme zur Berechnung des arithmetischen Mittels und der Varianz. Praktisch alle Modelle arbeiten nach der obigen Formel, die allerdings anfällig ist auf Rundungsfehler. Rechner mit einer Kapazität von 12 Dezimalstellen liefern für eine Stichprobe mit den Werten 99'999, 100'000, 100'001 anstelle einer Varianz von 1 eine solche von 0!

Varianzzerlegung

Wir betrachten die Varianz einer aus *r* Teilstichproben zusammengesetzten Datenmenge. x_{ij} bezeichne die *i*-te Merkmalsausprägung der *j*-ten Teilstichprobe. Zur einfacheren Notation sei vorausgesetzt, dass jede Merkmalsausprägung innerhalb der Teilstichproben nur einfach besetzt ist. Jede Teilstichprobe besitzt ein eigenes arithmetisches Mittel $\overline{x}_j$ und eine eigene Varianz s_{wj}^2. Aus den individuellen Informationen ist die Varianz der vereinigten Stichprobe gesucht. (Die Berechnung des arithmetischen Mittels der Gesamtstichprobe wurde früher gezeigt.)

Stichprobe 1	Stichprobe 2	....	Stichprobe *r*
x_{11} x_{21} . . $x_{n1,1}$	x_{12} x_{22} . . $x_{n2,2}$		x_{1r} x_{2r} . . $x_{nr,r}$
$\overline{x}_1$; s_{w1}^2	$\overline{x}_2$; s_{w2}^2	....	$\overline{x}_r$; s_{wr}^2

Bezeichnet s^2 die Gesamtvarianz der vereinigten Stichprobe mit dem arithmetischen Mittel $\overline{x}$, so gilt

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{x}\right)^2 \\ &= \frac{1}{n-1}\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left[\left(x_{ij} - \overline{x}_j\right) + \left(\overline{x}_j - \overline{x}\right)\right]^2 \\ &= \sum_{j=1}^{r}\frac{n_j - 1}{n-1}\,\frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{x}_j\right)^2 + \frac{1}{n-1}\sum_{j=1}^{r} n_j\left(\overline{x}_j - \overline{x}\right)^2 \\ &= \sum_{j=1}^{r}\frac{n_j - 1}{n-1} s_{wj}^2 + s_b^2 \end{aligned} \qquad (4.34)$$

Die Gesamtvarianz lässt sich in 2 Bestandteile zerlegen

- $$s_{wj}^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{x}_j\right)^2 \qquad (4.35)$$

  beschreibt die Varianz innerhalb (within) der *j*-ten Teilstichprobe und

- $$s_b^2 = \frac{1}{n-1}\sum_{j=1}^{r} n_j\left(\overline{x}_j - \overline{x}\right)^2 \qquad (4.36)$$

  die Varianz zwischen (between) den Stichprobenmittelwerten $\overline{x}_j$.

Mit

$$s_w^2 = \sum_{j=1}^{r} \frac{n_j - 1}{n-1} s_{wj}^2 \tag{4.37}$$

gilt somit der Zerlegungssatz

$$s^2 = s_w^2 + s_b^2 \tag{4.38}$$

Spezialfälle:

- Zwei Teilstichproben $r = 2$

  Aus

$$\overline{x} = \frac{n_1 \overline{x}_1 + n_2 \overline{x}_2}{n} \tag{4.39}$$

  folgt

$$s^2 = \frac{1}{n_1 + n_2 - 1}\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \frac{n_1 n_2}{n_1 + n_2}(\overline{x}_1 - \overline{x}_2)^2\right] \tag{4.40}$$

- Die Werte innerhalb der einzelnen Teilstichproben sind identisch.

  $\left(x_{1j} = x_{2j} = \ldots = x_{n_j j} \qquad j = 1, 2, \ldots, r\right)$

$$s_{wj}^2 = 0 \qquad (j = 1, 2, \ldots, r)$$
$$s^2 = s_b^2 \tag{4.41}$$

- Die Mittelwerte aller r Teilstichproben sind identisch $\overline{x}_1 = \overline{x}_2 = \ldots = \overline{x}_r$.

$$s_b^2 = 0$$
$$s^2 = s_w^2 \tag{4.42}$$

Die obige Zerlegung zeigt, wie die Gesamtvarianz einer vereinigten Stichprobe additiv aufgespaltet werden kann in Bestandteile, welche innerhalb der Teilstichprobe zustandekommen und in Bestandteile, welche die Varianz der Stichprobenmittel zum Gegenstand haben. Dasselbe Prozedere, allerdings in umgekehrter Richtung, erscheint im analytischen Teil der Statistik bei der Varianzanalyse.

Eine weitere Anwendung eröffnet sich bei der rekursiven Berechnung der Stichprobenvarianz.

Berücksichtigt man, dass eine zusätzliche Beobachtung von n_j Werten je mit derAusprägung x_j ein arithmetisches Mittel von x_j und eine Varianz von Null (Klassenbildung!) aufweist, so folgt insbesondere aus dem Spezialfall für $r = 2$

$$s_j^2 = \frac{1}{N_j - 1}\left[(N_{j-1} - 1)s_{j-1}^2 + \frac{N_{j-1} n_j}{N_j}(\overline{x}_{j-1} - x_j)^2\right] \tag{4.43}$$

mit

$$s_1^2 = 0 \quad \text{und} \quad \overline{x}_1 = x_1$$
$$N_j = \sum_{i=1}^{j} n_i \tag{4.44}$$
$$\overline{x}_j = \frac{1}{N_j}\left(N_{j-1}\, \overline{x}_{j-1} + n_j x_j\right)$$

s_j^2 ist die Varianz der ersten j Klassenmitten mit den dazugehörigen absoluten Häufigkeiten n_j.

Die rekursive Berechnung der Stichprobenvarianz vermeidet das Problem von Rundungsfehlern. Zudem lässt sich die Entwicklung der Varianz unter Beizug neuer Beobachtungen anschaulich verfolgen.

4.2.5 Variationskoeffizient

Empirische Beobachtungen zeigen, dass zahlenmässig grosse Stichprobenwerte normalerweise stärker streuen als kleine. Prozentual gleiche Abweichungen wirken sich in der Varianz bei grossen Stichprobenwerten in einer entsprechend grösseren Varianz aus. Um Varianzen verschiedener Merkmale unabhängig von der Grössenordnung der Zahlen miteinander vergleichbar zu machen, sind relative Streuungsmasse notwendig. Das bekannteste ist der Variationskoeffizient v, für den gilt

$$v = \frac{\text{Standardabweichung}}{\text{arithmetisches Mittel}} = \frac{s}{\overline{x}} \tag{4.45}$$

v ist nur für strikt positive Verhältnisdaten sinnvoll definiert und misst die Standardabweichung in Einheiten des arithmetischen Mittels. Kann eine Merkmalsausprägung sowohl positive als auch negative Werte annehmen, so ist ein arithmetisches Mittel von 0 möglich, womit der Variationskoeffizient in diesem Fall beliebig grosse Werte annehmen könnte. Dies ist unerwünscht, da ein arithmetisches Mittel von 0 sachlich keine spezielle Rechtfertigung besitzt.

Beispiel C (Einkommen)

$$v = \frac{s}{\bar{x}} = \frac{\sqrt{128'026}}{2'758.6} = 0.13 \qquad (4.46)$$

Die Standardabweichung beträgt 13% des arithmetischen Mittels.

- Der Variationskoeffizient eignet sich vor allem zu Varianzvergleichen, wenn die beobachteten Merkmale in unterschiedlichen Einheiten gemessen wurden, z.B. Preise eines Gutes in verschiedenen Währungen.
- Der Variationskoeffizient ist invariant gegenüber einer Multiplikation der Daten mit einer Konstanten; er reagiert hingegen auf allgemeine lineare Transformationen.

4.3 Formparameter

Die beiden konstruierten Stichproben 1 und 2 zeigen, dass das arithmetische Mittel und die Varianz auch gemeinsam eine Verteilung nur unvollständig zu beschreiben vermögen.

Stichprobe 1

x_j	n_j
10	2
20	3
30	5
40	15
50	5

$\bar{x} = 36$

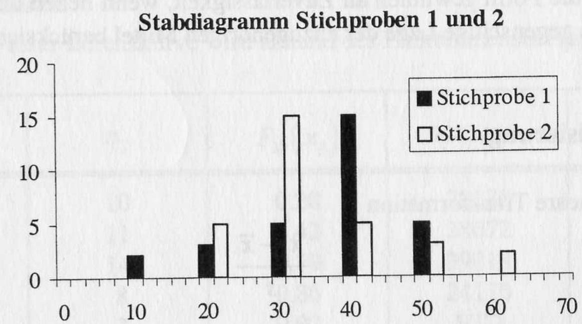
$s^2 = 121.38$

Stichprobe 2

x_j	n_j
22	5
32	15
42	5
52	3
62	2

$\bar{x} = 36$

$s^2 = 121.38$



Der offensichtliche Unterschied besteht in der "Form" der beiden Verteilungen, welche ja hinsichtlich der Varianz und des arithmetischen Mittels übereinstimmen. Gesucht sind somit weitere Parameter, welche diese Formunterschiede zum Ausdruck bringen.

4.3.1 Schiefe

$$m_3 = \frac{\frac{1}{n-1}\sum_{j=1}^{k} n_j \left(x_j - \bar{x}\right)^3}{s^3} \qquad (4.47)$$

heisst *drittes zentrales Moment*.

m_3 ist eine dimensionslose Grösse. Sie wird bei symmetrischen Verteilungen null, bei linksschiefen Verteilungen negativ und bei rechtsschiefen Verteilungen positiv. Die Aussagefähigkeit ist jedoch wegen der sensitiven Abhängigkeit gegenüber Ausreissern beschränkt. Die Schlüsse von den Parametern auf die Form der Verteilung haben deshalb mit der nötigen Vorsicht zu erfolgen.

Für die beiden obigen Beispiele gilt:

Stichprobe 1: $m_3 = -0.87$
Stichprobe 2: $m_3 = 0.87$

Aussagen über die Form gewinnen an Zuverlässigkeit, wenn neben den Schiefeparametern auch die gegenseitige Lage der dazugehörigen Mittel berücksichtigt werden.

4.4 Standardisierung

Die spezielle lineare Transformation

$$z_j = \frac{x_j - \bar{x}}{s} \tag{4.48}$$

heisst Standardisierung. z_j misst die Abweichungen der Werte x_j von $\bar{x}$ in Einheiten der Standardabweichung.

Das arithmetische Mittel der standardisierten Beobachtungen z wird 0, die Varianz besitzt den Wert 1.

$$\bar{z} = \frac{1}{n}\sum_{j=1}^{k} n_j z_j = \frac{1}{s}\frac{1}{n}\sum_{j=1}^{k} n_j\left(x_j - \bar{x}\right) = \frac{1}{ns}\left[\sum_{j=1}^{k} n_j x_j - n\bar{x}\right] = 0$$

$$s_z^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j\left(z_j - \bar{z}\right)^2 = \frac{1}{n-1}\sum_{j=1}^{k} n_j z_j^2 \tag{4.49}$$

$$= \frac{1}{s^2}\frac{1}{n-1}\sum_{j=1}^{k} n_j\left(x_j - \bar{x}\right)^2 = \frac{1}{s^2}s^2 = 1$$

4.5 Konzentrationsmasse

Konzentrationsmasse bilden eine Ergänzung zu den Streuungsparametern. Sie gestatten vor allem Aussagen über den Beitrag einzelner Merkmalsausprägungen zur Summe aller Merkmalsausprägungen (Beitrag einzelner Einkommensstufen zum gesamten Einkommen, Anteil von Landwirtschaftsbetrieben bezüglich Flächenbeitrag an der gesamten landwirtschaftlichen Fläche). Die Konzentration wird als absolut bezeichnet, wenn die gesamte Merkmalssumme auf ein einziges Element konzentriert ist. Umgekehrt spricht man von fehlender Konzentration, wenn die gesamte Merkmalssumme zu gleichen Teilen auf alle Merkmalsträger aufgeteilt ist.

4.5.1 Lorenzkurve

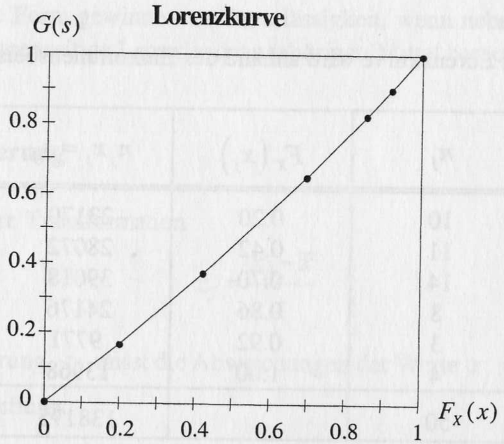
Die Grundidee einer Lorenzkurve wird anhand des Einkommensbeispiels dargestellt.

x_j	n_j	$F_X(x_j)$	$n_j x_j = s_j$	$G(s_j)$
2317	10	0.20	23170	0.1677
2552	11	0.42	28072	0.3708
2787	14	0.70	39018	0.6532
3022	8	0.86	24176	0.8282
3257	3	0.92	9771	0.8989
3492	4	1.00	13968	1.0000
	50		138175	

Die Kolonne $G(s_j)$ ist ähnlich zustandegekommen wie jene der Verteilungsfunktion F. Anstelle der summierten relativen Häufigkeiten, werden bei der Funktion G die aufsummierten Merkmalsbeiträge an der gesamten Merkmalssumme gemessen.

$$G(s_j) = \frac{\sum_{i=1}^{j} n_i x_i}{\sum_{i=1}^{k} n_i x_i}, \qquad j \le k \tag{4.50}$$

Für die Merkmalsausprägung 2552 gilt, dass 42% der untersten Lohnempfänger 37.08% der gesamten Lohnsumme auf sich vereinigen. Verbindet man in einem Koordinatensystem die Punkte mit den Koordinaten $\left[F_X(x_j), G(s_j)\right]$ geradlinig miteinander, so entsteht ein Polygonzug, den man als Lorenzkurve bezeichnet.



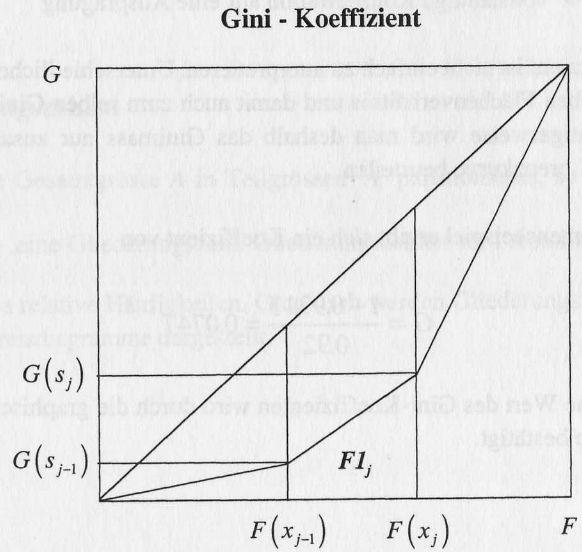
Eigenschaften der Lorenzkurve

- Die Lorenzkurve zeigt einen monoton wachsenden, konvexen Verlauf.
- Bei gleichmässiger Aufteilung der gesamten Merkmalssumme auf alle Merkmalsträger stimmt die Lorenzkurve mit der Winkelhalbierenden überein.
- Mit zunehmender Konzentration auf einzelne Merkmalsträger verläuft die Kurve flacher am Anfang und steiler am Ende.
- Die Interpretation der Lorenzkurve ist nicht beschränkt auf die Knickstellen, sondern erfolgt analog in den linearen Zwischenbereichen des Polygonzuges.
- Nach erfolgter Klassenbildung verläuft die Kurve generell weniger flach als im Falle der ursprünglichen Einzelbeobachtungen. Die Projektion der Beobachtungen innerhalb der Klassen bewirkt eine künstliche Konzentration auf die Klassenmitten.
- Die Lorenzkurve vermittelt einen *optischen* Eindruck der Konzentration.

4.5.2 Gini-Koeffizient

Die Lorenzkurve gestattet, ein Konzentrationsmass abzuleiten. Aus der Tatsache, dass die Konzentration tendenziell zunimmt, je stärker der Polygonzug von der Winkelhalbierenden nach unten abweicht, hat Gini vorgeschlagen, die Konzentration durch das Verhältnis der Fläche zwischen der Winkelhalbierenden und der Lorenzkurve zur maximal möglichen Fläche bei vollständiger Konzentration zu messen.

Die effektive und die maximal mögliche Fläche lassen sich aus der nachfolgenden Graphik ableiten



$$F1_j = \left[F_X(x_j) - F_X(x_{j-1})\right]\frac{G(s_{j-1}) + G(s_j)}{2} = f_X(x_j)\frac{G(s_{j-1}) + G(s_j)}{2}$$

$$F1 = 0.5 - \sum_{j=1}^{k} F1_j = 0.5\left\{1 - \sum_{j=1}^{k} f_X(x_j)\left[G(s_{j-1}) + G(s_j)\right]\right\} \tag{4.51}$$

$$F1_{\max} = \frac{\{1 - f_X(x_k)\}}{2} = \frac{F_X(x_{k-1})}{2}$$

Für den *Gini-Koeffizienten G* folgt aus diesen Beziehungen

$$G = \frac{F1}{F1_{\max}} = \frac{1 - \sum f_X\left(x_j\right)\left[G\left(s_{j-1}\right) + G\left(s_j\right)\right]}{F_X\left(x_{k-1}\right)} \tag{4.52}$$

- $0 \le G \le 1$
- $G = 0$ —> gleichmässige Verteilung über alle Ausprägungen
- $G = 1$ —> vollständige Konzentration auf eine Ausprägung

Der Gini-Koeffizient ist nicht einfach zu interpretieren. Unterschiedliche Lorenzkurven können zum selben Flächenverhältnis und damit auch zum selben Gini-Koeffizienten führen. Vorsichtigerweise wird man deshalb das Ginimass nur zusammen mit der dazugehörigen Lorenzkurve beurteilen.

Für das Einkommensbeispiel ergibt sich ein Koeffizient von

$$G = \frac{1 - 0.9313}{0.92} = 0.0747 \tag{4.53}$$

Der relativ kleine Wert des Gini-Koeffizienten wird durch die graphische Darstellung der Lorenzkurve bestätigt.

5. STATISTISCHE MASSZAHLEN

In der wirtschaftlichen Praxis ist man oft daran interessiert, Aspekte empirischer Sachverhalte (Sozialprodukt, Unternehmensgewinn, Preisentwicklung) in Form statistischer Masszahlen numerisch zum Ausdruck zu bringen. Dabei wird man zweckmässigerweise so vorgehen, dass zwischen den Zahlen und den durch sie abgebildeten Sachverhalten eine möglichst enge Verknüpfung besteht. Eng bedeutet dabei, dass gleichen Sachverhalten gleiche Zahlen und gleichen Zahlen gleiche Sachverhalte zugeordnet sein sollen. Schwierigkeiten bietet dabei vor allem die zweite Forderung, zumal keine eindeutigen Kriterien bekannt sind, an denen die Äquivalenz von Sachverhalten beurteilt werden könnte.

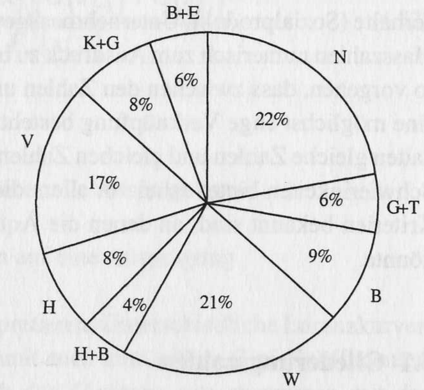
5.1 Gliederungszahlen

Lässt sich eine Gesamtgrösse A in Teilgrössen A_i partitionieren, so nennt man den Quotienten $\frac{A_i}{A}$ eine Gliederungszahl. Gliederungszahlen sind Anteile. Typische Beispiele sind etwa relative Häufigkeiten. Graphisch werden Gliederungszahlen vorzugsweise durch Kreisdiagramme dargestellt.

Beispiel 1

Ausgaben eines Durchschnittshaushalts

Nahrungsmittel	20
Getränke und Tabak	5
Bekleidung	8
Wohnungsmiete	19
Heizung, Beleuchtung	4
Haushalteinrichtung	7
Verkehr	15
Körper-/Gesundheit	7
Bildung und Erholung	15



Beispiel 2
Mittelung von Gliederungszahlen

Die Stimmbeteiligung betrug in 3 Stadtteilen 35%, 41% und 49%. Die Stimmberechtigten wohnen zu 20%, 30% und 50% in den entsprechenden Gebieten.

Für die gesamte Stadt mit B Stimmberechtigten ergibt sich eine Stimmbeteiligung von

$$b = \frac{B\cdot 0.20\cdot 0.35 + B\cdot 0.30\cdot 0.41 + B\cdot 0.50\cdot 0.49}{B} \tag{5.1}$$
$$= 0.20\cdot 0.35 + 0.30\cdot 0.41 + 0.50\cdot 0.49 = 0.438$$

b ist das arithmetische Mittel der Stimmbeteiligungen. Das Gewichtungsmuster entspricht den Bevölkerungsanteilen.

5.2 Beziehungszahlen

Während Gliederungszahlen Teile am Ganzen messen, werden bei Beziehungszahlen verschiedenartige Grössen zueinander in Beziehung gesetzt. Beispiele: Bevölkerungsdichte (Einwohner pro Flächeneinheit), Kapitalintensität (Anlagevermögen pro Beschäftigte), 'Wohlstand' (Bruttosozialprodukt pro Kopf).

Typische Exponenten von Beziehungszahlen im Bereich der Betriebswirtschaftslehre sind Unternehmungskennzahlen, welche vor allem Kontroll-, Planungs- und Steuerungszwecken dienen. Dazu werden ausgewählte Buchhaltungspositionen zueinander in Beziehung gesetzt. Gewöhnlich werden sie in 5 Gruppen eingeteilt.

- Liquidität (flüssige Mittel/kurzfristiges Fremdkapital)
- Sicherheit (Eigenkapital/Gesamtkapital)
- Rentabilität (Gewinn/investiertes Kapital = RoI)
- Cash-Flow (Effektivschulden/Cash-Flow)
- Umsatz (Auftragsbestand · 360/Umsatz des letzten Jahres)

Kennzahlen sind in erster Linie Orientierungsgrössen, deren isolierte Betrachtung nur beschränkte Einblicke herzugeben vermag. Sachadäquate Informationen sind eher aus dem Zusammenspiel verschiedener Kennzahlen mit zusätzlichen Bewertungen und Interpretationen zu erwarten.

Mittelung von Beziehungszahlen
Die Bevölkerungsdichte in 3 Kantonen sei 209, 25 und 158 Einwohner pro km^2. Die Einwohnerzahl (in 1'000) beträgt 421.7, 180.1 und 65.6. Dies ergibt eine Bevölkerungsdichte d für alle 3 Kantone zusammen von

$$d = \frac{\text{gesamte Einwohnerzahl}}{\text{gesamte Fläche}}$$

$$= \frac{421.7 + 180.1 + 65.6}{\frac{421.7}{209} + \frac{180.1}{25} + \frac{65.6}{158}} \tag{5.2}$$

$$= \frac{1}{\frac{0.63}{209} + \frac{0.27}{25} + \frac{0.10}{158}}$$

$$= 69.2$$

Das Resultat entspricht dem harmonischen Mittel der Beziehungszahlen 209, 25 und 158 Einwohner pro km^2. Das Gewichtungsschema leitet sich aus den Bevölkerungsanteilen ab.

5.3 Mess- und Indexzahlen

Wirtschaftlich interessante Fragen tauchen oft mit dem Verlauf einer sich im Zeitablauf entwickelnden Grösse auf. Man versucht die Entwicklung dadurch zu beschreiben, dass sämtliche Beobachtungen mit einer speziell ausgewählten Beobachtung verglichen werden. Dieses Referenzgrössenprinzip verlangt nicht unbedingt, dass die Bezugsgrösse selber eine Beobachtung sein muss; es kann beispielsweise auch ein Mittel eines Ausschnittes oder ein beliebig anderer Wert gewählt werden.

Messzahlen werden in erster Linie zur Beschreibung von Zeitreihen eingesetzt. Das Merkmal X wird im Zeitablauf beobachtet und das allgemeine Element mit x_t und die Referenzgrösse mit x_0 bezeichnet. Beispiele derartiger Zeitreihen wären etwa das Bruttosozialprodukt, der Unternehmensgewinn oder die Bevölkerung einer Agglomeration.

$$I_{0,t} = \frac{x_t}{x_0} \tag{5.3}$$

Der Quotient heisst Messzahl von x_t bezüglich der Basis x_0. Üblicherweise multipliziert man Messzahlen mit 100 und erhält aus dem Produkt prozentuale Aussagen. Speziell gilt $I_{0,0} = 100\%$.

$I_{0,t2} - I_{0,t1}$ wird als Änderung in *Prozentpunkten* bezeichnet. Änderungen in Prozentpunkten sind *keine* prozentualen Änderungen von x selber, sondern relative Änderungen von x im Intervall $[t_1, t_2]$.

$$I_{0,t_2} - I_{0,t_1} = \frac{x_{t_2}}{x_0} - \frac{x_{t_1}}{x_0} = \frac{1}{x_0}\left(x_{t_2} - x_{t_1}\right) \tag{5.4}$$

Bei der Konstruktion von Messzahlen ist die Wahl der Basis von entscheidender Bedeutung. Sie lässt sich nicht nach festen Regeln bestimmen; gefordert wird nur, eine Periode mit 'normaler' Entwicklung zu wählen. Insbesondere sind Perioden mit aussergewöhnlichen Einflüssen zu vermeiden.

Umbasierung

Bezeichnet t_0 eine alte Basis, welche durch eine neue Basis t_1 zu ersetzen ist, so bestimmen sich die neuen Messzahlen aus den alten durch die Beziehung

$$I_{t_1,t} = \frac{x_t}{x_{t_1}} = \frac{\frac{x_t}{x_{t_0}}}{\frac{x_{t_1}}{x_{t_0}}} = \frac{I_{t_0,t}}{I_{t_0,t_1}} \tag{5.5}$$

Für $t_0 = 0$ gilt speziell

$$I_{0,t} = I_{0,t_1} \cdot I_{t_1,t} \tag{5.6}$$

Diese Operation heisst *Verkettung* und zeigt die multiplikative Verknüpfung von Messzahlen über zwei Teilperioden zu einer Messzahl über die gesamte Periode.

Messzahlen sind keineswegs auf die Beschreibung von Zeitreihen beschränkt. Es können auch andere gleichartige Grössen auf einen aus diesen ausgewählten Repräsentanten (als Basis) bezogen werden. Beschreibt X etwa die Verkehrsintensität (Anzahl Fahrzeuge pro Zeiteinheit) und sind $x_1, x_2, \ldots$ Beobachtungen auf verschiedenen Autobahnabschnitten, so beschreiben Messzahlen die prozentualen Abweichungen der Verkehrsintensität bezüglich einer Referenzstrecke.

Viele, praktisch bedeutsame Fragestellungen führen zur Aufgabe, die Entwicklung nicht nur einer, sondern mehrerer, sachlich jedoch zusammengehöriger Zeitreihen durch eine einzige, gemeinsame Messzahl zu beschreiben. Konkret geht es etwa um Fragen der Form:

- Um welchen Prozentsatz haben sich die "Preise" verschiedener Güter in den letzten 2 Jahren verändert?

- Welche prozentuale Veränderung haben die Mieten im Durchschnitt während des letzten Jahres erfahren?

Wenn der Verlauf mehrerer Zeitreihen in einer einzigen Zahl zum Ausdruck kommen soll, so wird verständlich, dass der Aussagegehalt wesentlich von der Komplexität und dem Zusammenwirken der zu beschreibenden Zeitreihen abhängt. Von einer nach sinnvollen Kriterien konstruierten derartigen Globalgrösse erwartet man Antworten der Form, dass sich der momentane Durchschnittswert um *x%* gegenüber einem Durchschnittswert der Basisperiode verändert hat. Der Aussagegehalt entspricht jenem bei Messzahlen mit dem Unterschied, dass bei letzteren jeweils nur eine Messreihe im Zeitablauf beobachtet wird. Bei *Indexzahlen* geht es also darum, den Verlauf *mehrerer, sachlich zusammengehörender Zeitreihen durch eine alle Reihen umfassende Masszahl zu charakterisieren.*

Gegeben sei eine Gruppe von k ökonomischen Variablen (z.B. Preise verschiedener Güter), wobei jede im Zeitablauf beobachtet wird. Für die Zeitreihe der j-ten Variablen $(j=1,\dots,k)$ lässt sich gesondert eine Messzahl $I_{0,t}^{j}$ bestimmen. Die Basisperiode stimme bei allen k Variablen (nach einer allfällig notwendigen Umbasierung) überein.

Eine Indexzahl ist dann ein gewichtetes Mittel der k Messzahlen $I_{0,t}^{j}$.

Wird speziell ein *arithmetisches Mittel* unterstellt, so folgt als Indexzahl nach diesem Modell

$$M_{0,t}=\sum_{j=1}^{k} g_j\, I_{0,t}^{j} \qquad 0\le g_j \le 1; \qquad \sum g_j = 1 \tag{5.7}$$

In Abhängigkeit von der Art der Zeitreihen unterscheidet man bei ökonomischen Betrachtungen im wesentlichen 3 Arten von Indexzahlen.

Ist die Zeitreihe eine Folge von *Preisen* $p_{j,t}$, so wird der Index als *Preisindex* $P_{0,t}$ bezeichnet.

$M_{0,t}$ heisst *Mengenindex* $Q_{0,t}$, falls die Zeitreihe aus Mengen besteht.

Betrachtet man schliesslich Umsätze $x_{jt}=p_{jt}\cdot q_{jt}$ im Zeitablauf, so wird der dazugehörige Index als *Umsatzindex* $U_{0,t}$ bezeichnet.

Zur einfacheren Bezeichnung werden Preis- und Mengenreihen auch als Vektoren dargestellt. $\mathbf{p}(t)=(p_{1t},p_{2t},\dots,p_{kt})$ steht für das Preissystem zum Zeitpunkt t. Entsprechend bezeichnet $\mathbf{q}(t)=(q_{1t},q_{2t},\dots,q_{kt})$ das Mengensystem für denselben Zeitpunkt. Analog verfährt man bei Umsätzen.

Bei gegebenen Variablen und bekannter Basisperiode sind die Indizes nur noch vom Gewichtungssystem g_j abhängig. Aus der ökonomischen Praxis hat sich der Gedanke durchgesetzt, in den Gewichten die wertmässige Bedeutung der Variablen zum Ausdruck zu bringen. Bekannt sind die beiden Gewichtungsmodelle von *Laspeyres* und *Paasche.*

Fasst man die Variablen als Güter mit einem bestimmbaren Preis auf, so schlägt *Laspeyres* als Gewichtungsfaktor für die j-te Variable ihren *wertmässigen Anteil in der Basisperiode* vor.

Paasche verwendet in seinem Konzept ebenfalls den wertmässigen Anteil des j-ten Gutes am Wert aller Güter. Im Unterschied zu Laspeyres betrachtet Paasche jedoch die Anteile der *Vergleichsperiode*. Bei Paasche variiert somit das Gewichtungsschema im Zeitablauf, während es bei Laspeyres dank der Orientierung an der Basisperiode konstant bleibt.

5.3.1 Indizes nach Laspeyres

Indizes nach *Laspeyres* sind *arithmetisch* gemittelte Preisverhältnisse. Das Gewichtungsschema des Basiszeitpunktes

$$g_j=\frac{p_{j0}q_{j0}}{\sum_{i=1}^{k} p_{i0}q_{i0}} \tag{5.8}$$

führt zum *Preisindex*

$$ {}_L P_{0t} = \sum_{j=1}^{k} \frac{p_{jt}}{p_{j0}} \frac{p_{j0} q_{j0}}{\sum_{i=1}^{k} p_{i0} q_{i0}} = \frac{\sum_{j=1}^{k} p_{jt} q_{j0}}{\sum_{i=1}^{k} p_{i0} q_{i0}} = \frac{\mathbf{p}(t)\mathbf{q}(0)}{\mathbf{p}(0)\mathbf{q}(0)} \tag{5.9} $$

$\mathbf{p}(t)$ resp. $\mathbf{q}(t)$ symbolisieren Preis- und Mengenvektoren des Warenkorbes zum Zeitpunkt t.

Zähler und Nenner enthalten denselben Mengenvektor $\mathbf{q}(0)$ der Basisperiode. Bei konstanten Basismengen reagiert der Index nur auf Preisänderungen. Der Preisindex von Laspeyres beschreibt die prozentuale Veränderung der Gesamtausgaben für einen festen 'Warenkorb der Basisperiode' zu Preisen der Basis- respektive der Vergleichsperiode.

Mengenindizes nach Laspeyres sind arithmetisch gemittelte Mengenänderungsverhältnisse der Vergleichs- und Basisperiode unter konstanten Preisstrukturen der Basisperiode.

$$ {}_L Q_{0t} = \sum_{j=1}^{k} \frac{q_{jt}}{q_{j0}} \frac{p_{j0} q_{j0}}{\sum_{i=1}^{k} p_{i0} q_{i0}} = \frac{\sum_{j=1}^{k} p_{j0} q_{jt}}{\sum_{i=1}^{k} p_{i0} q_{i0}} = \frac{\mathbf{p}(0)\mathbf{q}(t)}{\mathbf{p}(0)\mathbf{q}(0)} \tag{5.10} $$

Das Verhalten des Indexes wird bei festen Basispreisen nur durch Mengenänderungen bestimmt.

Indizes nach Laspeyres beruhen auf konstanten Strukturen der *Basis*periode. Es liegt nahe, im Gegensatz dazu, Konzepte mit konstanten Strukturen der Vergleichsperiode zu betrachten. Die daraus resultierenden Indizes werden als sogenannte *Paasche*-Indizes bezeichnet.

5.3.2 Indizes nach Paasche

Zu Indizes nach Paasche gelangt man auch dadurch, dass man in den entsprechenden Indizes von Laspeyres die konstanten Strukturen der Basisperiode konsequent durch jene der Vergleichsperiode ersetzt.

Der Preisindex nach Paasche

$$ {}_P P_{0t} = \frac{\sum_{j=1}^{k} p_{jt} q_{jt}}{\sum_{i=1}^{k} p_{i0} q_{it}} = \frac{\mathbf{p}(t)\mathbf{q}(t)}{\mathbf{p}(0)\mathbf{q}(t)} \tag{5.11} $$

misst das Verhältnis der gesamten Aufwendungen für einen 'Warenkorb' der Vergleichsperiode zu Basis- respektive zu Vergleichsperiodenpreisen. Ein Index von 1.2 besagt, dass der Warenkorb der Vergleichsperiode in der Vergleichsperiode 20% mehr kostet, als derselbe Warenkorb in der Basisperiode gekostet hätte. Der Warenkorb der Vergleichsperiode ist im Konzept von Paasche für die Basisperiode oft als Fiktion zu betrachten, zumal der Markt in aller Regel so schnell reagiert, dass die Produkte des Vergleichskorbes in der zurückliegenden Basisperiode gar nicht verfügbar waren.

Für den *Mengenindex* nach Paasche folgt analog

$$ {}_P Q_{0t} = \frac{\sum_{j=1}^{k} p_{jt} q_{jt}}{\sum_{i=1}^{k} p_{it} q_{i0}} = \frac{\mathbf{p}(t)\mathbf{q}(t)}{\mathbf{p}(t)\mathbf{q}(0)} \tag{5.12} $$

Ein Index von 1.2 bedeutet, dass der durch Mengenänderungen verursachte Wert des Vergleichsperiodenkorbes zu Preisen der Vergleichsperiode 20% höher ist als in der Basisperiode.

Der formalen Struktur nach sind die Indizes nach Paasche *harmonische* Mittel von Preis- respektive Mengenverhältnissen. Das Gewichtungskonzept bezieht sich auf den anteiligen Wert des *j*-ten Gutes am Gesamtwert in der *Vergleichsperiode*.

5.3.3 Vergleich von Laspeyres- und Paascheindizes

Im Rahmen der Gegenüberstellung beschränken wir uns auf Preisindizes. Die konstanten Mengenstrukturen verursachen unterschiedliche Konsequenzen. Laspeyres-Indizes sind sehr einfach zu berechnen. Für jeden Zeitpunkt muss lediglich der aktuelle Preisvektor erhoben werden. Der Warenkorb bei Paasche wird stets den momentanen Marktverhältnissen angepasst, womit für jeden Berechnungszeitpunkt Preise und Mengen neu zu erheben sind.

Bei aktiven Märkten mit starken Angebots- und Nachfrageschwankungen wird der

Basiskorb von Laspeyres bereits in kurzen Zeitintervallen obsolet und muss den neuen Marktverhältnissen angepasst werden. Analoge Effekte werden durch rasche technologische Entwicklungen ausgelöst. Das Laspeyreskonzept verlangt somit wie jenes von Paasche nach einer periodischen Aktualisierung, allerdings in längeren Intervallen. Ein ähnliches Problem stellt sich aber auch bei Paasche. Zwar widerspiegelt der Warenkorb den momentanen Markt. Doch muss dieses für die Basisperiode fiktive Warenbündel auch mit Preisen gewichtet werden. Neue Produkte besitzen jedoch keine Vergangenheitspreise, womit auch dem Paasche-System Limiten gesetzt sind.

Während Indizes von Laspeyres zu verschiedenen Zeitpunkten dank der einheitlichen Basis untereinander vergleichbar sind, trifft dies für Indizes nach Paasche nur bedingt zu. Unterschiedliche Preisindizes von Paasche zu zwei verschiedenen Zeitpunkten bedeuten keineswegs, dass sich die Preise im Mittel gemäss der Indexentwicklung verändert haben. Der Unterschied kann ebensogut selbst bei konstanten Preisen von einer mengenmässigen Umlagerung des Warenkorbs herrühren. Weder beim Laspeyres- noch beim Paaschekonzept werden ökonomisch begründbare Beziehungen zwischen Preisen und effektiv nachgefragten Mengen berücksichtigt. Ökonomisch handelnde Haushalte werden, wenn immer möglich, teure Güter durch billige ersetzen. Dadurch ändert sich die faktische Zusammensetzung des Güterkorbes, welche bei Laspeyres unberücksichtigt bleibt. Die budgetwirksame Preisentwicklung wird bei Laspeyres in solchen Situationen überschätzt. Und dies umso mehr, je weiter Basis- und Vergleichsperiode auseinanderliegen.

Umgekehrt werden bei Paasche Güter mit stark steigenden Preisen zu wenig stark gewichtet. Der Mengenkorb der Vergleichsperiode hat durch entsprechende Mengenanpassungen bereits auf Preissteigerungen reagiert.

Bei rationalem Konsumverhalten wird die Nachfrage nach Gütern mit steigenden Preisen und damit auch steigenden Preisverhältnissen zurückgehen. Das Mengengerüst von Laspeyres bleibt jedoch konstant, woraus für solche Situationen die Indizes von Paasche und Laspeyres auseinanderdriften, d.h.

$$ {}_{P}P_{0,t} \le {}_{L}P_{0,t} \tag{5.13}$$

Die Diskrepanz der beiden Indizes wird in effizienten Märkten umso grösser, je weiter Basis- und Vergleichszeitpunkt auseinanderliegen.

In formaler Hinsicht erfüllen weder Paasche noch Laspeyres die Verknüpfungseigenschaft, welche insbesondere bei Basisänderungen relevant werden.

$$ {}_{L}P_{0,t} \ne {}_{L}P_{0,t^*} \cdot {}_{L}P_{t^*,t}$$

$$\left(\frac{\mathbf{p}(t)\cdot\mathbf{q}(0)}{\mathbf{p}(0)\cdot\mathbf{q}(0)} \ne \frac{\mathbf{p}(t^*)\cdot\mathbf{q}(0)}{\mathbf{p}(0)\cdot\mathbf{q}(0)} \cdot \frac{\mathbf{p}(t)\cdot\mathbf{q}(t^*)}{\mathbf{p}(t^*)\mathbf{q}(t^*)}\right) \tag{5.14}$$

$$ {}_{P}P_{0,t} \ne {}_{P}P_{0,t^*} \cdot {}_{P}P_{t^*,t}$$

$$\left(\frac{\mathbf{p}(t)\cdot\mathbf{q}(t)}{\mathbf{p}(0)\cdot\mathbf{q}(t)} \ne \frac{\mathbf{p}(t^*)\cdot\mathbf{q}(t^*)}{\mathbf{p}(0)\cdot\mathbf{q}(t^*)} \cdot \frac{\mathbf{p}(t)\cdot\mathbf{q}(t)}{\mathbf{p}(t^*)\mathbf{q}(t)}\right) \tag{5.15}$$

Letztlich sind es vor allem erhebungstechnische Gründe und Interpretationsvorteile, welche die dominante Stellung des Laspeyreskonzeptes ausmachen.

5.3.4 Teilindizes

Die Berechnung eines Gesamtindexes aus Teilindizes sei am Beispiel eines Preisindexes von Laspeyres dargestellt. Die Gruppe aller *k* Variablen wird in zwei Teilgruppen von *m* respektive *k-m* Variablen aufgeteilt. Für jede Teilgruppe lässt sich gesondert ein Index bestimmen.

$$ {}_{L}P^1_{0,t} = \frac{\mathbf{p}_1(t)\mathbf{q}_1(0)}{\mathbf{p}_1(0)\mathbf{q}_1(0)}; \quad {}_{L}P^2_{0,t} = \frac{\mathbf{p}_2(t)\mathbf{q}_2(0)}{\mathbf{p}_2(0)\mathbf{q}_2(0)} \tag{5.16}$$

Die Vektoren $\mathbf{p}_i(\tau)$ und $\mathbf{q}_i(\tau)(i=1,2)$ bezeichnen Preise und Mengen für die *i*-te Gruppe zum Zeitpunkt $\tau(\tau=0,t)$.

$$ {}_{L}P_{0,t} = \frac{\mathbf{p}_1(0)\mathbf{q}_1(0)\dfrac{\mathbf{p}_1(t)\mathbf{q}_1(0)}{\mathbf{p}_1(0)\mathbf{q}_1(0)} + \mathbf{p}_2(0)\mathbf{q}_2(0)\dfrac{\mathbf{p}_2(t)\mathbf{q}_2(0)}{\mathbf{p}_2(0)\mathbf{q}_2(0)}}{\mathbf{p}(0)\mathbf{q}(0)} \tag{5.17}$$

$$= g_{1L}P^1_{0,t} + g_{2L}P^2_{0,t}$$

Der Gesamtindex ist eine gewichtete Summe der Teilindizes beider Gruppen. Die Gewichte enthalten den wertmässigen Anteil der Variablen innerhalb der Gruppen am Gesamtwert aller Variablen. Sie bringen die wertmässige Bedeutung der Gruppen zum Ausdruck.

5.3.5 Der schweizerische Landesindex der Konsumentenpreise

Der Landesindex der Konsumentenpreise gehört zur Kategorie der wichtigen volkswirtschaftlichen Kenngrössen und dient dazu, die *Bewegung der Detailhandelspreise von Waren und Dienstleistungen im Zeitablauf* wiederzugeben. In der Schweiz wird der Index im Prinzip nach der Methode von Laspeyres berechnet.

Verbrauchsgrundlage

Das feste Verbraucherschema wird periodisch bei den sogenannten Indexhaushalten erfragt. Aus den Haushaltrechnungen werden die jeweils aktuellen Verbrauchsgewohnheiten (Warenkorb) eines 'Durchschnittshaushaltes' abgeleitet. Auf der obersten Stufe erfolgt die Aufteilung der Ausgaben auf die Bedarfsgruppen. Die entsprechenden Quoten sind in nachfolgender Tabelle zusammengestellt.

Bedarfsgruppen	Quoten %		
	1966	1977	1982
Nahrungsmittel	31	20	21
Getränke und Tabakwaren	5	5	5
Bekleidung	13	8	7
Wohnungsmiete	17	19	18
Heizung und Beleuchtung	6	4	5
Haushalteinrichtung und Unterhalt	7	7	6
Verkehr	9	15	14
Körper- und Gesundheitspflege	7	7	8
Bildung und Erholung	5	15	16

Der Durchschnittshaushalt verwendet im Jahre 1982 18% seines verfügbaren Einkommens für die Wohnungsmiete.

Die Bedarfsgruppen werden weiter bis zur Stufe einzelner Indexpositionen (Milch, Fleisch, Bücher, Ferien, Schmerzmittel) gegliedert.

Preiserhebung

Die Preise werden bei den Verkaufsstellen der Indexartikel unter Berücksichtigung regionaler (Stadt, Land) und angebotsspezifischer (Absatzkanäle) Differenzierungen erfasst. Die Preise werden monatlich (Nahrungsmittel, Heizung und Beleuchtung), vierteljährlich (Getränke und Tabakwaren, Bekleidung, Haushalteinrichtung und Haushaltunterhalt, Verkehr, Körper- und Gesundheitspflege, Bildung und Erholung) oder halbjährlich (Mieten) neu erhoben.

Berechnungsmethode

Aus den einzelnen Preismeldungen wird zunächst eine durchschnittliche Preisveränderung pro Absatzkanal innerhalb einer Gemeinde ermittelt. Über alle Gemeinden wird daraus ein Index pro Absatzkanal und daraus durch Mittelung über alle Absatzkanäle eine Veränderung pro Indexposition aufgerechnet. Aus diesen Teilindizes pro Warengruppe wird schliesslich ein Gesamtindex gebildet.

Allgemeine Bemerkungen

In der Schweiz wird der Landesindex der Konsumentenpreise nach der Methode der *Basispreisrelation* bestimmt. Für jede einzelne Preismeldung wird zunächst die Veränderung p_{jt} / p_{j0} gegenüber der Basisperiode berechnet. Diese Preisverhältnisse werden anschliessend arithmetisch zur Preisänderung des jeweiligen Produktes gemittelt. Es werden also nicht mittlere Preise der Vergleichs- und Basisperiode zueinander in Beziehung gesetzt, wie es die reine Theorie vorsehen würde. Begründet wird dieses Vorgehen dadurch, dass Indexpositionen nur selten so eindeutig sind, dass ein wohldefiniertes Produkt in seiner Preisentwicklung beobachtet werden kann. Die meisten Positionen beschreiben sogenannt *inhomogene* Güter wie zum Beispiel Arzneimittel. Auf dem Markt existiert jedoch kein eindeutiges Produkt "Arzneimittel", sodass nach schweizerischer Praxis jeder Preismelder jenes Produkt auswählt, das seiner Meinung nach typisch ist für die Position Arzneimittel. Für das einmal gewählte Produkt werden dann Preisveränderungen ermittelt, welche im weiteren Vorgehen mit den Entwicklungen der Preisverhältnisse der übrigen unter der Rubrik Arzneimittel gewählten Positionen gemittelt werden. Mit diesem Vorgehen erhofft man sich neben den methodischen Vorteilen, möglichst nahe an die effektiven Marktverhältnisse heranzukommen.

Schwierigkeiten ähnlicher Art bieten Artikel mit saisonalen Schwankungen (Gemüse, Früchte, Reisen), da sie eine unterjährige Anpassung des Warenkorbes notwendig machen.

5.3.6 Kaufkraftvergleiche

Indexzahlen sind keineswegs auf die Bearbeitung von Preisreihen im Zeitablauf beschränkt. Bezeichnen p_{jA} die Inlandpreise von Gütern des Warenkorbes q_{jA} sowie p_{jB} und q_{jB} dieselben Angaben für das Land B, so sind Indizes von Laspeyres und von Paasche sogenannte Verbraucherparitäten.

$$ {}_{L}P_{AB} = \frac{\sum_{j=1}^{k} p_{jB} q_{jA}}{\sum_{j=1}^{k} p_{jA} q_{jA}} = \frac{\mathbf{p}(B)\mathbf{q}(A)}{\mathbf{p}(A)\mathbf{q}(A)} \tag{5.18} $$

Der Laspeyresindex ${}_{L}P_{AB}$ vergleicht die Preise eines Warenkorbes des Landes A (in Währung A) mit den Kosten desselben Korbes im Land B (in Währung B). Ein Index von 5.25 bedeutet, dass für den Warenkorb des Landes A im Lande B 5.25 mal soviele Geldeinheiten des Landes B aufzuwenden waren wie im Lande A. Bei einem freien Güterverkehr zwischen diesen beiden Ländern müsste sich auch der Wechselkurs beider Währungen etwa bei diesem Verhältnis einspielen.

Analoge Überlegungen lassen sich mit dem Paaschekonzept anstellen. Es setzt die Kosten des Warenkorbes B in Währung von B mit den Kosten desselben B-Korbes, jedoch zu Preisen des Landes A zueinander in Beziehung.

5.3.7 Der Index nach Fisher und seine Axiome

Fisher hat vorgeschlagen, die Indizes von Laspeyres und Paasche geometrisch zu mitteln. Sein Preisindex beträgt

$$ {}_{F}P_{0t} = \sqrt{{}_{L}P_{0t} \cdot {}_{P}P_{0t}} \tag{5.19} $$

Als geometrisches Mittel besitzt der Index von Fisher keine zu den Indizes von Laspeyres und Paasche analoge Interpretation als Kostenverhältnis eines konstanten Warenkorbes.

Fisher hat mit seinem Konzept nicht nach einem Kompromiss zwischen den Indizes von Laspeyres und Paasche gesucht, obschon sein Index stets zwischen diese beiden Indizes zu liegen kommt. Fisher war viel eher bemüht, einen Index mit formal möglichst optimalen Eigenschaften zu finden. Er gilt in dieser Sicht als Begründer einer axiomatisch-statistischen Indextheorie. Indizes werden unter diesem Gesichtspunkt primär nicht am inhaltlichen Gehalt sondern an den formalen Eigenschaften gemessen, die man an einen

Index stellt. Zu diesem Zwecke definierte Fisher im Jahre 1922 ein System von Anforderungen, denen seiner Ansicht nach ein Index genügen sollte. Zur Darstellung des Fisher'schen Testsystems erweist es sich als vernünftig, einen Index als Funktion P der Argumente $\mathbf{p}(0), \mathbf{q}(0), \mathbf{p}(t)$ und $\mathbf{q}(t)$ darzustellen. Ein Index

$$ P\left[\mathbf{p}(0), \mathbf{q}(0), \mathbf{p}(t), \mathbf{q}(t)\right] \tag{5.20} $$

ist dann eine Abbildung des Preis-Mengenraumes auf die reelle Achse.

Proportionalitätstest

Verändern sich sämtliche Preise der Basisperiode um das π-fache, so soll der Index den Wert π annehmen.

$$ P\left[\mathbf{p}(0), \mathbf{q}(0), \pi\mathbf{p}(0), \mathbf{q}(t)\right] = \pi \tag{5.21} $$

Zirkularitätstest

Ein Index, der die Basisperiode 0 mit der Vergleichsperiode t verknüpft, lässt sich als Produkt eines Indexes mit der Basis 0 und der Zwischenvergleichsperiode s sowie einem Index mit der Basis s und der Vergleichsperiode t darstellen.

$$ \begin{aligned} & P\left[\mathbf{p}(0), \mathbf{q}(0), \mathbf{p}(t), \mathbf{q}(t)\right] \\ & = P\left[\mathbf{p}(0), \mathbf{q}(0), \mathbf{p}(s), \mathbf{q}(s)\right] \cdot P\left[\mathbf{p}(s), \mathbf{q}(s), \mathbf{p}(t), \mathbf{q}(t)\right] \end{aligned} \tag{5.22} $$

Mit t=0 erhält man den sogenannten *Zeitumkehrtest*, welcher besagt, dass beim Austausch von Basis- und Vergleichsperiode der Index in seinen reziproken Wert übergehen soll.

$$ P\left[\mathbf{p}(0), \mathbf{q}(0), \mathbf{p}(t), \mathbf{q}(t)\right] \cdot P\left[\mathbf{p}(t), \mathbf{q}(t), \mathbf{p}(0), \mathbf{q}(0)\right] = 1 \tag{5.23} $$

Bestimmtheitstest

Dieser Test verlangt, dass der Index gegen eine positive reelle Zahl konvergieren soll, wenn ein beliebiges skalares Element (bei Fisher allerdings nur die Preise) gegen Null strebt.

Kommensurabilitätstest
Eine Änderung der Masseinheiten der Güter verursacht keine Indexänderung. Werden die Mengen zum Beispiel anstatt in kg in Tonnen gemessen und die Preise entsprechend angepasst, soll der Index keine Änderung erfahren.

Faktorumkehrtest
Werden die Argumente $\mathbf{p}(0)$ mit $\mathbf{q}(0)$ und $\mathbf{p}(t)$ mit $\mathbf{q}(t)$ vertauscht, so kann der Index $P\left[\mathbf{q}(0),\mathbf{p}(0),\mathbf{q}(t),\mathbf{p}(t)\right]$ als Mengenindex aufgefasst werden. Der Faktorumkehrtest verlangt, dass das Produkt eines Preisindexes mit einem daraus nach obiger Vorschrift abgeleiteten Mengenindex einen Wertindex ergibt. Der Wertindex ist das Verhältnis der Werte der Warenkörbe der Vergleichs- und der Basisperiode zu jeweiligen Preisen.

Die nachfolgende Tabelle zeigt, inwiefern die Indizes von Laspeyres, Paasche und Fisher die obigen Tests erfüllen.

	Laspeyres	Paasche	Fisher
Proportionalität	+	+	+
Zirkularität	-	-	-
Zeitumkehr	-	-	+
Bestimmtheit	+	+	+
Kommensurabilität	+	+	+
Faktorumkehr	-	-	+

Der Index von Fisher vermag offensichtlich mehr Tests als jene von Laspeyres und Paasche zu erfüllen. Fisher hat ihn deshalb als 'idealen' Index bezeichnet. Die Tabelle zeigt ebenfalls, dass selbst der Fisher'sche Index nicht alle Tests gleichzeitig zu befriedigen vermag. Eichhorn und Voeller (Eichhorn, W., Voeller, J.; 1976; Theory of the Price Index) haben bewiesen, dass es keinen Index gibt, der sämtliche Tests von Fisher erfüllt. Sie haben deshalb ein modifiziertes Axiomensystem für Indexzahlen aufgestellt.

Monotonie
Ein Preisindex ist eine streng monoton fallende Funktion in den Preisen der Basisperiode und eine streng monoton steigende Funktion in den Preisen der Vergleichsperiode.

Lineare Homogenität
Ändern sämtliche Preise der Vergleichsperiode um einen konstanten positiven Faktor π, so erfährt der Preisindex dieselbe Veränderung um den Faktor π.

$$P\left[\mathbf{p}(0),\mathbf{q}(0),\pi\mathbf{p}(t),\mathbf{q}(t)\right]=\pi\, P\left[\mathbf{p}(0),\mathbf{q}(0),\mathbf{p}(t),\mathbf{q}(t)\right] \tag{5.24}$$

Identität
Bei übereinstimmenden Preisen der Basis- und Vergleichsperiode nimmt der Preisindex den Wert 1 an.

$$P\left[\mathbf{p}(0),\mathbf{q}(0),\mathbf{p}(0),\mathbf{q}(t)\right]=1 \tag{5.25}$$

Dimensionalität
Eine Änderung der Masseinheit der Währung verursacht keine Preisänderung.

$$P\left[\pi\mathbf{p}(0),\mathbf{q}(0),\pi\mathbf{p}(t),\mathbf{q}(t)\right]=P\left[\mathbf{p}(0),\mathbf{q}(0),\mathbf{p}(t),\mathbf{q}(t)\right] \tag{5.26}$$

Kommensurabilität
Eine Änderung der Masseinheiten der Güter verursacht keine Preisänderung.

Neben den Indizes von Laspeyres, Paasche und Fisher erfüllen noch weitere Indizes sämtliche Axiome von Eichhorn. Axiomensysteme vermögen die Menge möglicher Indizes normalerweise nicht auf ein einziges Element zu reduzieren, gewöhnlich wird nur eine ganze Klasse von potentiell möglichen Indizes festgelegt. Es kann anderseits auch vorkommen, dass im Falle inkonsistenter Axiome die Klasse möglicher Indizes leer wird; die von Fisher als wünschbar bezeichneten Tests sind ein Beispiel dafür.

5.3.8 Ökonomische Indizes

Aus ökonomischer Sicht wird an der statistischen Indextheorie bemängelt, dass Preise und Mengen als voneinander isolierte, unabhängige Variablen betrachtet werden. Demgegenüber unterstellt die ökonomische Theorie etwa des Nachfrageverhaltens wechselseitige Beziehungen zwischen den Preisen eines Gutes und seiner nachgefragten Menge. Das Konzept eines festen Warenkorbes bei variablen Preisen vermag demnach unter ökonomischen Gesichtspunkten nicht mehr zu befriedigen. Als Alternative dazu wurden ökonomisch fundierte Konzepte vorgeschlagen.

In einer Basisperiode herrsche ein Preissystem $\mathbf{p}(0)$. Bei einem Einkommen $I(0)$ wird ein nutzenmaximierendes Wirtschaftssubjekt seinen Warenkorb $\mathbf{q}(0)$ derart zusammenstellen, dass seine Bedürfnisse optimal befriedigt werden.

In der Vergleichsperiode wird jenes Einkommen $I(t)$ gesucht, das bei einem Preissystem $\mathbf{p}(t)$ denselben Grad an Bedürfnisbefriedigung sicherstellt. Der Warenkorb wird wieder nach denselben Optimalitätskriterien zusammengestellt.

Das Verhältnis der beiden minimal notwendigen Einkommen $I(0)$ und $I(t)$ wird als ökonomischer Index ${}_0P$ bezeichnet.

Während bei statistischen Preisindizes der feste Warenkorb das konstante Element darstellt, ist es bei ökonomischen Indizes der Grad der Bedürfnisbefriedigung, was allerdings noch einer expliziteren Formulierung bedarf. Ebenso sind zusätzliche Annahmen über das Konsumverhalten notwendig.

In Abhängigkeit von seiner Zusammensetzung bringt ein Warenkorb $\mathbf{q}(t) = (q_{1t} \dots q_{kt})$ dem Käufer einen Nutzen $U(q)$. Funktional werde diese Abhängigkeit durch eine Beziehung vom *Cobb-Douglas-Typ* beschrieben.

$$U(\mathbf{q}(t)) = \prod_{j=1}^{k} (q_{jt})^{g_j} \qquad 0 \le g_j \le 1; \ \sum g_j = 1 \tag{5.27}$$

Das Nutzenniveau U wird als konstantes Element vorgegeben.

Beim Preissystem $\mathbf{p}(\tau) = (p_{1\tau} \dots p_{k\tau})$ ist das minimal notwendige Einkommen $I(\tau)$ gesucht, welches einen vorgegebenen Nutzen U gerade noch garantiert.

Das Minimumproblem lautet

$$I(\tau) = \sum_{j=1}^{k} p_{j\tau} q_{j\tau} \rightarrow \text{Min} \tag{5.28}$$

unter der Nebenbedingung

$$\prod_{j=1}^{k} (q_{j\tau})^{g_j} = U \tag{5.29}$$

Die partielle Ableitung der Prinzipalfunktion

$$\Phi(q(\tau), \lambda) = \sum_{j=1}^{k} p_{j\tau} q_{j\tau} - \lambda \left(\prod_{j=1}^{k} q_{j\tau}^{g_j} - U \right) \tag{5.30}$$

nach $q_{j\tau}$ führt zunächst zu

$$p_{j\tau} q_{j\tau} = g_j \ I(\tau) \tag{5.31}$$

Ersetzt man $q_{j\tau}$ in der Nebenbedingung, so folgt

$$I(\tau) = \frac{U}{\prod_{j=1}^{k} \left(\frac{g_j}{p_{j\tau}} \right)^{g_j}} \tag{5.32}$$

Setzt man für τ die entsprechenden Zeitpunkte der Basis- und Vergleichsperiode ein, so erhält man den *ökonomischen Preisindex*

$${}_0P_{0,t} = \frac{I(t)}{I(0)} = \prod_{j=1}^{k} \left(\frac{p_{jt}}{p_{j0}} \right)^{g_j} \tag{5.33}$$

Unter dem obigen Nutzenoptimierungsmodell vom Cobb-Douglas-Typ ist der ökonomische Preisindex ein geometrisches Mittel der Preisverhältnisse der einzelnen Güter.

Das ökonomische Indexkonzept verlangt ein konstantes Gewichtungsmuster g_j über die Zeit. Dabei ist zu beachten, dass aus der Beziehung

$$p_{j\tau}q_{j\tau} = g_j\ I(\tau) \qquad (5.34)$$

g_j als wertmässiger Anteil des j-ten Gutes am Gesamtwert des Warenkorbes $q(\tau)$

folgt. Die Aufteilung der Gesamtausgaben nach den Gewichten g_j muss also ebenfalls im Zeitablauf konstant bleiben.

Aus der Theorie von Cobb-Douglas-Funktionen ist bekannt, dass g_j ebenfalls der Elastizität des Index bezüglich des Preisverhältnisses des j-ten Gutes entspricht. Auch hier wird Zeitinvarianz unterstellt.

Mögliche Gewichtsmuster lassen sich analog zu den Konzepten von Laspeyres und Paasche bilden.

Wählt man jenes von Laspeyres

$$g_j = \frac{p_{j0}q_{j0}}{\sum p_{i0}q_{i0}} \qquad (5.35)$$

so ist der ökonomische Preisindex nie grösser als jener von Laspeyres (als arithmetisches Mittel).

Wählt man die Gewichte nach Paasche

$$g_j = \frac{p_{jt}q_{jt}}{\sum p_{it}q_{it}} \qquad (5.36)$$

so ist der Preisindex nach Paasche (als harmonisches Mittel) nie grösser als der ökonomische Index.

Wenn die Gewichte nach Laspeyres und Paasche nicht zu stark auseinanderdriften, liegt der ökonomische Index stets zwischen jenen von Paasche und Laspeyres. Diese "Kompromisskonsequenz" wird oft als Begründung für die Wahl des geometrischen Mittels bei der Indexbildung herangezogen.

Beispiel:

Neben $g_1 = 0.2$ und $g_2 = 0.8$ gelten weiter

	$t = 0$	$t = 1$
p_1	1	1
p_2	4	8
I	50	87.06
q_1	10	17.41
q_2	10	8.71

Aus den Zwischenresultaten:

$$U(0) = U(1) = 10^{0.2}\,10^{0.8} = 10 \qquad (5.37)$$

$$I(0) = \frac{10}{\left(\frac{0.2}{1}\right)^{0.2}\left(\frac{0.8}{4}\right)^{0.8}} = 50$$

$$I(1) = \frac{10}{\left(\frac{0.2}{1}\right)^{0.2}\left(\frac{0.8}{8}\right)^{0.8}} = 87.06 \qquad (5.38)$$

folgt zunächst der neue Warenkorb zum Zeitpunkt 1

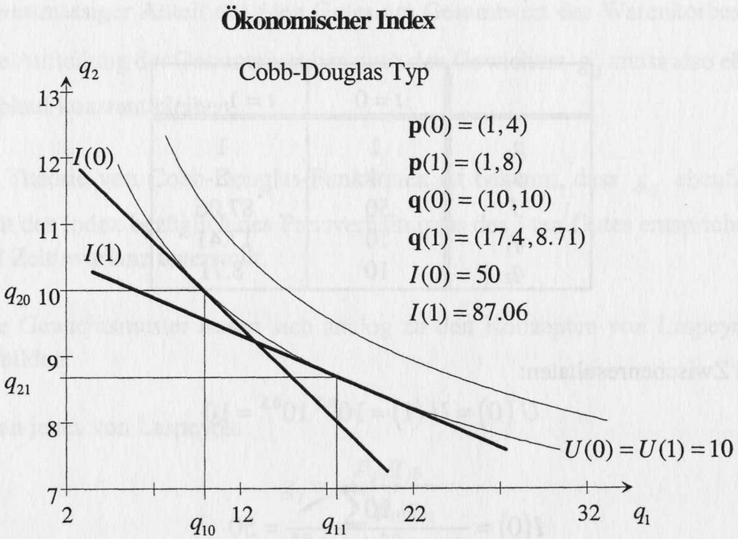
$$q_{11} = \frac{g_1 I(1)}{p_{11}} = \frac{0.2 \cdot 87.06}{1} = 17.41$$

$$q_{21} = \frac{g_2 I(1)}{p_{21}} = \frac{0.8 \cdot 87.06}{8} = 8.71 \qquad (5.39)$$

und daraus der ökonomische Index

$${}_0P_{0t} = \frac{87.06}{50} = \left(\frac{1}{1}\right)^{0.2} \cdot \left(\frac{8}{4}\right)^{0.8} = 1.74 \qquad (5.40)$$

Graphische Veranschaulichung



Die Steigung der Budgetgeraden wird durch die relativen Preise $-p_{1t} / p_{2t}$ bestimmt. Während beim Preissystem $\mathbf{p}(0)$=(1,4) zum Zeitpunkt 0 ein Einkommen von 50 notwendig ist, um einen Nutzen von 10 zu realisieren, beläuft sich das minimal notwendige Einkommen $I(1)$ = 87.06 zum Zeitpunkt 1 mit dem Preissystem $\mathbf{p}(1)$=(1,8). Das Verhältnis der minimal notwendigen Einkommen entspricht dem ökonomischen Index in der Höhe von 1.74.

Beim System von Laspeyres dreht sich die Budgetgerade um den festen Punkt (q_{10}, q_{20}) und führt bei Änderungen der relativen Preise stets auf ein höheres Nutzenniveau.

Ändert sich das Gewichtungsschema $g_j\,(j = 1, \ldots, k)$ in den beiden Vergleichszeitpunkten nicht, so liegt der ökonomische Index stets zwischen dem Paasche- und dem Laspeyresindex.

5.3.9 Dow Jones Indizes

Die Dow Jones Indizes (Dow Jones & Company) gehören zu den am weitesten verbreiteten Kenngrössen zur Bewertung von Niveau und Trends auf den Aktienmärkten der Vereinigten Staaten. Die Indexbildung geht zurück auf Charles Henry Dow, Gründungsherausgeber des Wall Street Journal und Mitbegründer von Dow Jones & Company, sowie auf Edward D. Jones.

In seiner ursprünglichen, aus dem Jahr 1886 stammenden Version, umfasste der Dow Jones Industrial Average (DJIA) zwölf repräsentative Werte und wurde als arithmetisches Mittel der entsprechenden Aktienkurse gebildet. Noch vor dem DJ Industrial Average wurde im Jahre 1884 der damals auf elf Aktienkursen basierende Dow Jones Railroad Average ins Leben gerufen. Der jüngste Index der Dow Jones Familie ist der seit dem Jahre 1929 gebildete Dow Jones Utility Index (DJUA), welcher aufgrund von Werten des Dienstleistungssektors gebildet wird.

Mittlerweile wurde das in die Indizes eingehende Aktionsspektrum erweitert, der DJIA wird auf der Basis von 30 Werten gebildet, der Dow Jones Transportation Average (DJTI), der aus dem Dow Jones Railroad Average hervorgegangen ist, umfasst heute 20 Werte, der DJUA basiert auf 15 Gesellschaften.

In die Bildung des DJIA gehen folgende Werte ein:

AlliedSignal Inc.	International Business Machines Corp.
Aluminum Co. of America	International Paper Co.
American Express Co.	J.P. Morgan & Co.
AT&T Corp.	Johnson & Johnson
Boeing Co.	McDonald's Corp.
Caterpillar Inc.	Merck & Co.
Chevron Corp.	Minnesota Mining & Manufacturing Co.
Coca-Cola Co.	Philip Morris Cos.
DuPont Co.	Procter & Gamble Co.
Eastman Kodak Co.	Sears, Roebuck & Co.
Exxon Corp.	Travelers Group Inc.
General Electric Co.	Union Carbide Corp.
General Motors Corp.	United Technologies Corp.
Goodyear Tire & Rubber Co.	Walt Disney Co.
Hewlett-Packard Co.	Wal-Mart Stores Inc.

Ursprünglich erfolgte die Indexbildung als arithmetisches Mittel der Notierungen, d.h.

$$DJ = \frac{1}{n}\sum p_i \qquad (5.41)$$

p_i: Kurs der i-ten Aktie.

n wird als Dow Jones Divisor bezeichnet; er entsprach also der Anzahl der enthaltenen Aktienkurse. Dieser Ansatz erwies sich im Falle von Mutationen unter den Titeln oder bei Splits als unzweckmässig. Bei einem 3-zu-1-Split werden beispielsweise für jede alte Aktie drei neue ausgegeben, was im allgemeinen etwa zur Drittelung des Kurses führt. Betrachtet man nun einen Index, welcher aufgrund der drei Aktien A, B und C im Wert von \$5, \$10 bzw. \$15 gebildet wird, so sollte dieser durch den Split von C in seinem Wert nicht verändert werden. Vor dem Split gilt

$$DJ = \frac{5+10+15}{3} = 10 \qquad (5.42)$$

Ohne Modifikation des Divisors gilt nach dem Split

$$DJ = \frac{5+10+5}{3} = 6.67 \qquad (5.43)$$

was offensichtlich unerwünscht ist; der "richtige" Divisor lautet 2.

Zur Wahrung der Kontinuität der Indizes im Falle von Aktiensplits wird eine Anpassung des Divisors gemäss folgender Formel vorgenommen.

- Notierungen von A, B, C $\quad p_1, p_2, p_3$
- Splitfaktor $\quad c$
- „neuer“ Divisor $\quad ND$

Dann soll gelten

$$\frac{p_1+p_2+p_3}{3} = \frac{p_1+p_2+\frac{p_3}{c}}{ND}$$

$$ND\left(p_1+p_2+p_3\right) = 3p_1+3p_2+\frac{3}{c}p_3 \qquad (5.44)$$

$$ND = \frac{3p_1+3p_2+\frac{3}{c}p_3}{p_1+p_2+p_3}$$

$$\left(\text{im Beispiel}: \ ND = \frac{15+30+15}{5+10+15} = \frac{60}{30} = 2\right)$$

Allgemein gilt

- Notierungen $\quad p_1, \dots, p_n$
- “alter” Divisor $\quad AD$
- Splitfaktor $\quad c$ für Aktie j
- “neuer” Divisor $\quad ND$

$$\frac{\sum p_i}{AD} = \frac{\sum_{i\neq j} p_i + \frac{p_j}{c}}{ND} \qquad (5.45)$$

$$ND = \frac{AD\left(\sum_{i\neq j} p_i + \frac{p_j}{c}\right)}{\sum p_i}$$

Auch im Falle der Aufnahme neuer Aktien in den Index sowie beim Austausch von Werten ist eine Anpassung des Divisors erforderlich. Würde beispielsweise Philip Morris (\$39.25) durch Microsoft (\$117.38) ersetzt, so hätte dies bei einem aktuellen Divisor von 0.24275214 ohne Anpassung des Divisors eine Erhöhung des Index um

$$\frac{117.38-39.25}{0.24275214} = 321.85 \qquad (5.46)$$

Punkte zur Folge.

Daher ist in diesem Fall der Divisor wie folgt anzupassen:

Austausch von p_j durch p_{j_0}

$$\frac{\sum p_i}{AD} = \frac{\sum_{i \neq j} p_i + p_{j_0}}{ND}$$

$$ND = \frac{\sum_{i \neq j} p_i + p_{j_0}}{\sum p_i} AD \tag{5.47}$$

Man beachte, dass bei einem Divisor D die Veränderung eines Kurses um \$1 generell eine Veränderung des Index um $\frac{1}{D}$ Punkte impliziert.

Eine offensichtliche Schwäche des Index liegt in der Preisgewichtung. Einerseits besitzen hoch notierte Aktien einen starken Einfluss auf den Index und andererseits wird im Falle eines Splits das Gewicht der gesplitteten Aktie entsprechend reduziert und das aller übrigen leicht angehoben.

Für die praktische Interpretation bedeutet dies, dass der DJ nicht als Benchmark für ein Portefeuille verwendet werden kann; er signalisiert bestenfalls die Richtung der Börsenentwicklung.

5.3.10 Swiss-Market-Index (SMI)

Der SMI gehört zur Indexfamilie der Schweizer Börse. Seine Funktionspalette ist vielfältig und reicht vom Börsenbarometer über ein Performancemass bis zur Basis derivativer Produkte. Der SMI wurde am 30. Juni 1988 bei einem Indexstand von 1500 eingeführt. Die Berechnung erfolgt "real time", d.h. sobald mit einem sog. SMI-Titel ein Abschluss erfolgt, wird der Index neu berechnet.

Der SMI baut methodisch, wie alle übrigen Indizes der Swiss-Index-Familie, auf der Methode von Laspeyres auf. Er umfasst bis zu 25 der wichtigsten und liquidesten Titel (sog. Blue Chips) von hochkapitalisierten Schweizer Unternehmen, welche ca. 80% der Gesamtkapitalisierung an der Schweizer Börse repräsentieren. Die Gewichtung der einzelnen Titel erfolgt gemäss ihrem Anteil an der Gesamtkapitalisierung. Für den Initialindex gilt

$$I_{88} = \frac{\sum p_{i,88}\, x_{i,88}}{\sum p_{i,88}\, x_{i,88}} 1500 = k_{88} \sum p_{i,88}\, x_{i,88} \tag{5.48}$$

k_{88} wird als sog. Kapitalfaktor bezeichnet.

Für den aktuellen Indexstand I_s gilt

$$I_s = k_t \sum_{i=1}^{M} p_{i,s}\, x_{i,t} \tag{5.49}$$

mit

t	=	aktueller Tag
s	=	aktueller Zeitpunkt am Tag t
k_t	=	Kapitalfaktor am Tag t
M	=	Anzahl Indextitel
$p_{i,s}$	=	letzter bezahlter Kurs vom Valor i
$x_{i,t}$	=	Anzahl Titel für Valor i am Tag t

Der SMI beschreibt die wertmässige Entwicklung eines wohldefinierten Portefeuilles, welches durch den sog. Kapitalfaktor in einem gewissen Sinne normiert wird. Der Titelkorb bleibt jedoch im Zeitablauf nicht konstant; er wird vielmehr durch exogene Ursachen wie Kapitalerhöhungen, Aktiensplits oder Austausch von Titeln laufend verändert.

Derartige Ereignisse verlangen operative Anpassungen, welche sowohl den Titelkorb (x) als auch den Kapitalfaktor betreffen.

Allfällige Änderungen der Titelzahl (x) erfolgen einmal pro Jahr per 1. Juli. Der Kapitalfaktor wird laufend angepasst, sobald eine Veränderung in der Kapitalisierung stattfindet, welche nicht durch Marktschwankungen verursacht wird. Die Anpassung erfolgt unmittelbar vor dem Eintritt des Kapitalereignisses. Der Kapitalfaktor wird dabei so angepasst, dass das Produkt aus der neuen Kapitalisierung mit dem neuen Kapitalfaktor denselben Indexstand ergibt wie das Produkt aus der Kapitalisierung vor dem Ereignis mit dem alten Kapitalfaktor.

Mit dem eben beschriebenen Anpassungsmechanismus bei Änderungen in der Anzahl

Titel wird im Laspeyres-Index eine systemfremde Verkettung vollzogen. Dies hat Konsequenzen bezüglich der Interpretation des Index. Der SMI beschreibt somit nicht die Wertentwicklung eines konstanten Portfolios sondern jene eines den Marktgegebenheiten angepassten Titelkorbes. Die klassische Laspeyres-Interpretation des Initial-SMI wird mit der Zeit mehr und mehr verwässert.

Der Hauptunterschied zum Dow Jones liegt darin, dass in den SMI der kapitalisierte Wert eines Valors und nicht nur sein Kurs eingeht. Ein Split verändert z.B. den SMI nicht, da mit der Mengen- automatisch eine Preisanpassung erfolgt, welche das Produkt der beiden Faktoren konstant lässt. Aus diesem Grund eignet sich der SMI sehr viel besser als Benchmark für Referenzportfolios.

6. BESTANDS- UND BEWEGUNGSMASSEN

6.1 Grundbegriffe

Statistische Massen können danach charakterisiert werden, ob ihre Erhebung zu einem bestimmten *Zeitpunkt* oder über ein bestimmtes *Zeitintervall* sinnvoll ist. Im ersten Fall spricht man von Bestands- und im zweiten Fall von Bewegungsmassen. Bestandsmassen sind oft in natürlicher Art und Weise von Bewegungsmassen eingeschlossen.

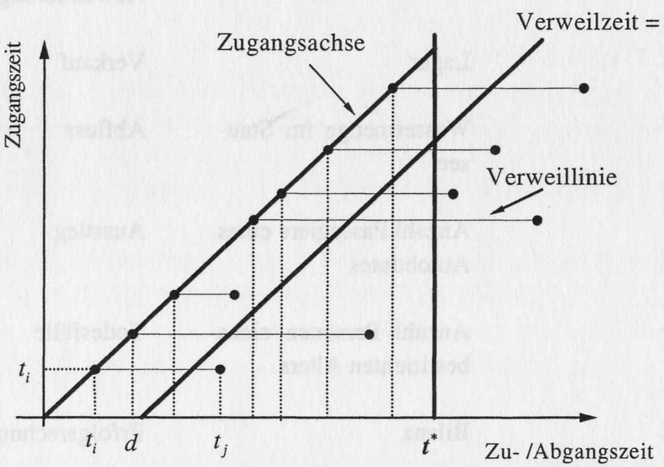
Bewegungsmasse	Bestandsmasse	Bewegungsmasse
Zugang	Bestand	Abgang
Geburt, Zuwanderung	Bevölkerung einer Stadt	Todesfälle, Abwanderungen
Einkauf	Lager	Verkauf
Zufluss	Wassermenge im Stausee	Abfluss
Zustieg	Anzahl Passagiere eines Autobusses	Ausstieg
	Anzahl Personen eines bestimmten Alters	Todesfälle
	Bilanz	Erfolgsrechnung

Der Beobachtungszeitraum wird durch das Intervall $[t_0, t_n]$ abgegrenzt. $T = t_n - t_0$ bezeichnet die Länge des Beobachtungszeitraumes. Mutationen finden im Bestand zu den Zeitpunkten $t_1, t_2, \ldots, t_n$ statt, wobei gilt $t_1 < t_2 < \ldots < t_n$.

B_i ist der Bestand zum Zeitpunkt t_i. A_i und Z_i symbolisieren die gesamten Ab- und Zugänge bis und mit dem Zeitpunkt *i*. Bezeichnet t_i den Zugangs- und t_j den Abgangszeitpunkt eines bestimmten Elementes zur Masse, so heisst $t_j - t_i$ seine Verweildauer.

Für das Kurrikulum eines konkreten Individuums im Bestand benötigt man lediglich den Zugangs- und Ausscheidezeitpunkt. Zur Darstellung der Bestände und Bewegungen dient die *Becker'sche* Graphik. In einem rechtwinkligen Koordinatensystem sind Ordinate und Abszisse Zeitachsen. Die Zugangszeit t_i eines Elementes wird durch einen Punkt mit den Koordinaten (t_i, t_i) markiert. Somit liegen sämtliche Zugangszeitpunkte auf der ersten Winkelhalbierenden (Zugangsachse). Dem Ausscheidezeitpunkt t_j wird ein Punkt mit den Koordinaten (t_j, t_i) zugeordnet. Die zur Abszisse parallele Verbindungsgerade von Zugangs- und Ausscheidepunkt symbolisiert die Verweildauer des Elementes im Bestand.

Becker'sche Graphik



Auf der Zugangsachse ist die Verweildauer null. Parallelen zur Zugangsachse markieren Punkte übereinstimmender Verweildauer. Der Bestand zum Zeitpunkt t^* entspricht der Anzahl Schnittpunkte einer Senkrechten zur Abszisse über t^* mit den Verweilstrecken.

Normalerweise reicht eine globale Betrachtung des Bestandes aus. Anstelle einer individuellen Betrachtung genügt es meist, neben dem Bestand, die Zu- und Abgänge zu kennen. Als Mutationsregel wird unterstellt, dass stets das älteste Element im Bestand ausscheidet (fifo: first in first out).

6.2 Bestandsfortschreibung - Zeitmengenfläche

Der Bestand zum Zeitpunkt t ist

$$B(t) = B(t_0) + Z(t_0, t) - A(t_0, t) \tag{6.1}$$

Dabei bedeuten:

$B(t_0)$: Anfangsbestand

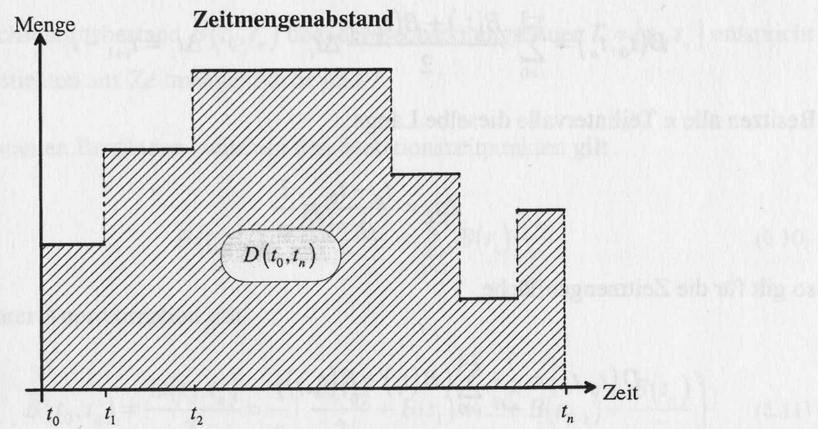
$Z(t_0, t)$: Summe der Zugänge zwischen t_0 und t

$A(t_0, t)$: Summe der Abgänge zwischen t_0 und t

In den folgenden Ausführungen betrachten wir nur noch den Saldo $S(t_0, t)$ der Zu- und Abgangsbewegungen und schreiben

$$B(t) = B(t_0) + S(t_0, t) \tag{6.2}$$

Ferner wird zunächst unterstellt, dass der Bestand zwischen zwei Mutationszeitpunkten konstant bleibt.



Die Fläche unter der Bestandsfunktion B wird als Zeitmengenfläche $D(t_0, t)$ bezeichnet, wobei zwischen den Zeitpunkten t_0 und t_n gilt

$$D(t_0, t_n) = \sum_{i=0}^{n-1} B(t_i)\,\Delta t_i \qquad \Delta t_i = t_{i+1} - t_i \tag{6.3}$$

In der Dimension von D (Zeit · Menge) kommt ihre Doppelfunktion zum Ausdruck. Die Zeitmengenfläche heisst

$$\begin{array}{ll} \text{beidseitig offen, falls} & B(t_0)>0 \text{ und } B(t_n)>0 \\ \text{beidseitig geschlossen, falls} & B(t_0)=0 \text{ und } B(t_n)=0 \end{array} \tag{6.4}$$

$$\text{einseitig offen, falls} \begin{cases} B(t_0)>0 \quad \text{und } B(t_n)=0 \\ \quad\text{oder} \\ B(t_0)=0 \quad \text{und } B(t_n)>0 \end{cases} \tag{6.5}$$

Zwei Bestandsfunktionen heissen über demselben Zeitintervall äquivalent, falls ihre Zeitmengenbestände übereinstimmen. Es werden nicht identische Funktionsverläufe, sondern nur gleiche Flächen unter der Bestandsfunktion verlangt.

Ist der Bestand zu den Zeitpunkten $t_i = (i = 0, 1, \dots, n)$ bekannt und unterstellt man zwischen den Zeitpunkten einen linear-gleichmässigen Zu- oder Abgang, so modifiziert sich die Zeitmengenfläche zu folgender Beziehung

$$D(t_0,t_n) = \sum_{i=0}^{n-1} \frac{B(t_i)+B(t_{i+1})}{2}\,\Delta t_i \qquad \Delta t_i = t_{i+1}-t_i \tag{6.6}$$

Besitzen alle n Teilintervalle dieselbe Länge

$$\Delta t_i = \frac{t_n - t_0}{n} = \frac{T}{n} \tag{6.7}$$

so gilt für die Zeitmengenfläche

$$\begin{aligned} D(t_0,t_n) &= \frac{T}{2n}\sum_{i=0}^{n-1}\left(B(t_i)+B(t_{i+1})\right) \\ &= \frac{T}{n}\left(\frac{B(t_0)}{2}+B(t_1)+\ldots+B(t_{n-1})+\frac{B(t_n)}{2}\right) \end{aligned} \tag{6.8}$$

<u>Beispiel:</u> Der Versichertenbestand beträgt jeweils am Monatsende

t	$B(t)$	t	$B(t)$
Dez	1211	Juli	1160
Jan	1428	Aug	1190
Feb	1398	Sept	1228
März	1370	Okt	1278
April	1406	Nov	1255
Mai	1380	Dez	1280
Juni	1290		

$$D(\text{Jan, Dez}) = \frac{1\text{ Jahr}}{12}\left(\frac{1211}{2}+1428+\ldots+1255+\frac{1280}{2}\right) = 1302 \text{ Personenjahre} \tag{6.8}$$

6.3 Masszahlen von Bestandsmassen

6.3.1 Durchschnittsbestand

Der Durchschnittsbestand $\overline{B}(t_0,t_n)$ über der Beobachtungsdauer $T = [t_0, t_n]$ entspricht dem Quotienten aus Zeitmengenfläche und T.

Bei konstanten Beständen zwischen den Mutationszeitpunkten gilt

$$\overline{B}(t_0,t_n) = \frac{D(t_0,t_n)}{T} = \sum_{i=0}^{n-1} B(t_i)\frac{\Delta t_i}{T} \tag{6.10}$$

Bei linearer Approximation gilt

$$\overline{B}(t_0,t_n) = \frac{D(t_0,t_n)}{T} = \frac{1}{n}\left(\frac{B(t_0)}{2}+B(t_1)+\ldots+B(t_{n-1})+\frac{B(t_n)}{2}\right) \tag{6.11}$$

In beiden Fällen ist der Durchschnittsbestand ein arithmetisches Mittel. Geometrisch entspricht der Durchschnittsbestand der Höhe eines zur Zeitmengenfläche $D(t_0,t_n)$ flächengleichen Rechtecks über dem Zeitintervall $[t_0,t_n]$.

Beispiel
Der Durchschnittsbestand der Versicherten pro Jahr beträgt

$$\overline{B}(t_0,t_n) = \frac{D(\text{Jahr})}{1\,\text{Jahr}} = \frac{1302\ \text{Personenjahre}}{1\ \text{Jahr}} = 1302\ \text{Personen} \qquad (6.12)$$

d.h. ein konstanter Bestand von 1'302 Versicherten führt über ein Jahr betrachtet zur selben Zeitmengenfläche wie der variable Bestand aus der Tabelle S 85.

6.3.2 Mittlere Verweildauer

Als mittlere Verweildauer eines Elementes in einem Bestand definiert man den Quotienten zwischen dem Zeitmengenbestand und jener Anzahl Elemente, welche im Verlaufe des Beobachtungsintervalls zu diesem Zeitmengenbestand Beiträge geleistet haben.

$$\text{Anzahl Elemente} = B(t_0) + Z(t_0,t_n) = B(t_n) + A(t_0,t_n)$$

Die gesamten Verweildauern über dem Intervall $[t_0,t_n]$ sind im Zeitmengenbestand $D(t_0,t_n)$ enthalten. Die Elemente, welche zum Zeitpunkt t_0 bereits dem Bestand angehören, besitzen ebenso eine Vorgeschichte wie die Elemente des Endbestandes $B(t_n)$, welche noch eine Restlaufzeit vor sich haben. Bezeichnen v_0 und v_n ihre Verweildauern, so folgt für die relevante Zeitmengenfläche

$$D^*(t_0,t_n) = B(t_0)v_0 + D(t_0,t_n) + B(t_n)v_n \qquad (6.13)$$

woraus die mittlere Verweildauer $\overline{V}^*(t_0,t_n)$ folgt

$$\overline{V}^*(t_0,t_n) = \frac{B(t_0)v_0 + D(t_0,t_n) + B(t_n)v_n}{B(t_0) + Z(t_0,t_n)} \qquad (6.14)$$

Trifft man für v_0 und v_n die Annahmen

$$\begin{aligned} v_0 &= \lambda\,\overline{V}^*(t_0,t_n) \\ v_n &= (1-\lambda)\,\overline{V}^*(t_0,t_n) \end{aligned} \qquad 0 \le \lambda \le 1 \qquad (6.15)$$

d.h. die Elemente des Anfangsbestandes haben den Anteil λ der mittleren Verweildauer bereits hinter sich und jene des Endbestandes noch $(1-\lambda)\,\overline{V}^*$ vor sich, so findet man nach einigen Umformungen

$$\overline{V}^*(t_0,t_n) = \frac{D(t_0,t_n)}{\lambda\, Z(t_0,t_n) + (1-\lambda)\, A(t_0,t_n)} \qquad 0 \le \lambda \le 1 \qquad (6.16)$$

Mangels eines besseren Grundes wählt man für λ meist den Wert ½. Die obige Formel für die mittlere Verweildauer gilt für alle Formen von Zeitmengenflächen (beidseitig offen oder geschlossen und halboffen).

6.3.3 Umschlagshäufigkeit

Die Umschlagshäufigkeit im Beobachtungsintervall (t_0,t_n) ist definiert als

$$\overline{U}(t_0,t_n) = \frac{t_n - t_0}{\overline{V}^*(t_0,t_n)} = \frac{T}{\overline{V}^*(t_0,t_n)} \qquad (6.17)$$

Beispiel
Wir betrachten ein Zugsabteil mit 4 Sitzplätzen. Gegeben sind die Zu- und Abgänge $Z(0,t)$ bzw. $A(0,t)$. Ferner sei $B(0) = 0$.

t	0	1	2	3	4	5	6	7	8	9	10
$Z(0,t)$	0	2	3	3	3	6	6	8	8	9	9
$A(0,t)$	0	0	0	1	2	2	3	5	5	5	9

Daraus folgen die Bestände

$B(t)$	0	2	3	2	1	4	3	3	3	4	0

und die Zeitmengenfläche

$$D(0,10) = 2 + 3 + 2 + \ldots + 4 = 25$$

Die mittleren Zu- und Abgangsmengen sind pro Zeiteinheit

$$\overline{Z}(0{,}10)=\frac{Z(0{,}10)}{10}=0.9$$
$$\overline{A}(0{,}10)=\frac{A(0{,}10)}{10}=0.9 \tag{6.18}$$

Der Durchschnittsbestand $\overline{B}$ während der gesamten Periode beträgt

$$\overline{B}(0{,}10)=\frac{D(0{,}10)}{10}=\frac{25}{10}=2.5 \tag{6.19}$$

Die mittlere Verweildauer ist

$$\overline{V}(0{,}10)=\frac{D(0{,}10)}{Z(0{,}10)}=\frac{25}{9}=2.78 \tag{6.20}$$

und für die Umschlagshäufigkeit $\overline{U}$ folgt

$$\overline{U}(10)=\frac{T}{\overline{V}(0{,}10)}=\frac{10}{2.78}=3.6 \tag{6.21}$$

Die Zeitmengenfläche ist beidseitig abgeschlossen.

6.4 Abgangsmodelle

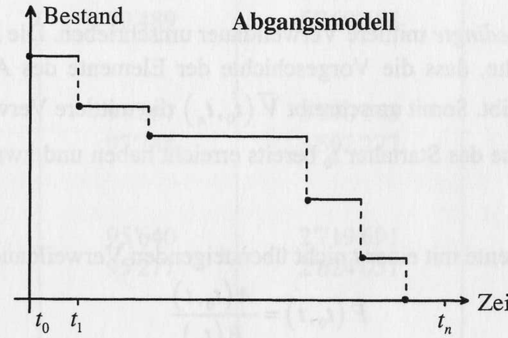
Fortschreibungsmodelle, welche die Bedingungen

1. der Zugang aller Elemente erfolgt zu einem festen Zeitpunkt t_0,
2. es fallen keine Neuzugänge an,
3. zum Zeitpunkt t_n hat sich der Bestand erschöpft,

erfüllen, heissen Abgangsmodelle. $B(t)$ ist mangels Neuzugängen eine monoton fallende Treppenfunktion mit

$$B(t)=B(t_0)-A(t_0,t)$$
$$B(t_n)=0 \tag{6.22}$$

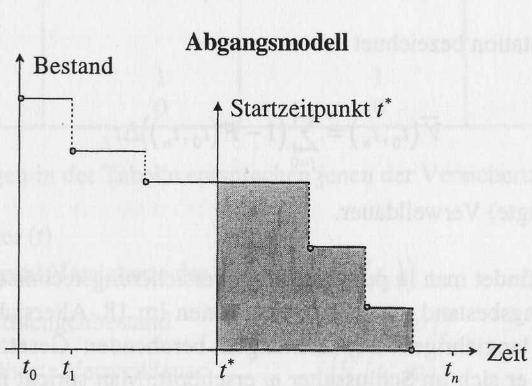
Typische Beispiele für Abgangsmodelle sind auslaufende Lager, Geburtenjahrgänge oder Kommutationsbestände in versicherungstechnischen Grundlagen.



Die Zeitmengenfläche ist analog zu den allgemeinen Bestandsmodellen

$$D(t_0,t_n)=\sum_{j=0}^{n-1}B(t_j)\Delta t_j \qquad \text{mit } \Delta t_j=t_{j+1}-t_j \tag{6.23}$$

Der Zeitpunkt t_0 kann beliebig verschoben werden und erzeugt stets wieder einen neuen Ausgangszeitpunkt für ein Abgangsmodell.



Zumal keine Neuzugänge anfallen, liefern nur die Elemente des Anfangsbestandes $B(t_0)$ Beiträge zur Zeitmengenfläche.

Definiert man als mittlere Verweildauer

$$\bar{V}(t_0,t_n)=\frac{D(t_0,t_n)}{B(t_0)} \tag{6.24}$$

so wird damit eine *bedingte* mittlere Verweildauer umschrieben. Die Bedingung resultiert aus der Tatsache, dass die Vorgeschichte der Elemente des Anfangsbestandes unberücksichtigt bleibt. Somit umschreibt $\bar{V}(t_0,t_n)$ die mittlere Verweildauer derjenigen Elemente, welche das Startalter t_0 bereits erreicht haben und zwar ausgehend von diesem Zeitpunkt.

Der Anteil der Elemente mit einer t nicht übersteigenden Verweildauer beträgt

$$F(t_0,t)=\frac{A(t_0,t)}{B(t_0)} \tag{6.25}$$

woraus für den Bestand zum Zeitpunkt t

$$B(t)=B(t_0)-A(t_0,t)=B(t_0)\big(1-F(t_0,t)\big) \tag{6.26}$$

und die Zeitmengenfläche

$$D(t_0,t_n)=\sum_{j=0}^{n-1}B(t_j)\,\Delta t_j=B(t_0)\sum_{j=0}^{n-1}\big(1-F(t_0,t_j)\big)\Delta t_j \tag{6.27}$$

folgt. In dieser Notation bezeichnet

$$\bar{V}(t_0,t_n)=\sum_{j=0}^{n-1}\big(1-F(t_0,t_n)\big)\Delta t_j \tag{6.28}$$

die mittlere (bedingte) Verweildauer.

Abgangsmodelle findet man in praktisch allen versicherungstechnischen Grundlagen. Ein fiktiver Anfangsbestand von 100'000 Personen im 18. Altersjahr baut sich nach bestimmten, auf langjährigen Beobachtungen beruhenden Gesetzmässigkeiten im Zeitablauf ab, bis er sich im Schlussalter ω erschöpft. Man spricht in diesem Zusammenhang kurz von der Absterbeordnung, welche periodisch wieder aktualisiert werden muss.

x	l_x	D_x	e_x
18	100'000	5'868'365	58.68
19	99'900	5'768'365	57.74
20	99'797	5'668'465	56.80
21	99'692	5'568'668	55.86
22	99'588	5'468'975	54.92
23	99'489	5'369'388	53.97
40	97'887	3'690'163	37.70
41	97'744	3'592'277	36.75
50	95'640	2'719'691	28.44
51	95'277	2'624'051	27.54
62	87'257	1'608'020	18.43
63	86'030	1'520'763	17.68
64	84'707	1'443'733	16.94
65	83'280	1'350'026	16.21
80	44'522	326'726	7.34
90	11'019	45'925	4.17
108	1	1	1
109	0	0	

Die Bezeichnungen in der Tabelle entsprechen jenen der Versicherungsmathematik.

x	Alter (t)	
l_x	Anzahl Versicherte des Alters x	$(B(t_x))$
D_x	Zeitmengenbestand	$(D(t_x,\omega))$
e_x	bedingte Verweildauer	$(\bar{V}(t_x,\omega))$
Δt	1 Jahr	

Von 100'000 18-jährigen Versicherten erreichen 83'280 das 65. Altersjahr.

$$D_{50} = \sum_{x=50}^{108} l_x = 2'719'691 \qquad (6.29)$$

bedeutet, dass die 95'640 Versicherten des Alters 50 noch insgesamt 2'719'691 Personenjahre erleben.

Die mittlere Verweildauer

$$e_x = \frac{D_x}{l_x} \qquad (6.30)$$

ist altersabhängig und wird als (bedingte) Lebenserwartung eines x-jährigen bezeichnet. Versicherte, welche das 65. Altersjahr erreicht haben, besitzen noch eine (bedingte) Lebenserwartung von 16.21 Jahren. Das mittlere Sterbealter der Personen, welche das 80. Altersjahr erreicht haben, beträgt 87.34 Jahre.

7. DESKRIPTIVE ZEITREIHENANALYSE

Unter einer Zeitreihe versteht man eine im Zeitablauf erhobene und danach geordnete diskrete Folge von Beobachtungen $y_1, y_2, \ldots, y_t$. Meistens ist der zeitliche Abstand aufeinanderfolgender Beobachtungen bereits konstant oder kann durch eine einfache Transformation entsprechend angepasst werden.

Beispiel: Tägliche Energienachfrage in MWh

	Mo	Di	Mi	Do	Fr	Sa	So
1. Woche		372	370	373	344	236	225
2. Woche	367	391	373	376	351	257	258
3. Woche	392	393	391	398	369	286	268
4. Woche	424	426	422	424	394	295	286
5. Woche	428	542	427	373			

Die deskriptiv-statistische Zeitreihenanalyse beschäftigt sich im wesentlichen mit zwei Hauptaufgaben. In einer ersten Phase steht die Modellierung der Zeitreihe zur Diskussion. Insbesondere stellt sich die Frage nach (theoretischen) Modellen, respektive Modellgleichungen, welche das konkret vorhandene Datenmaterial adäquat zu beschreiben vermögen. In zweiter Linie wird versucht, durch Glättungsmechanismen zeitliche Strukturen der Daten offenzulegen.

7.1 Das klassische Zeitreihenmodell

Ökonomische Prozesse erzeugen häufig Zeitreihen, welche neben einer längerfristigen Entwicklung ein in bestimmten Intervallen wiederkehrendes Muster zeigen. Diese empirisch gestützten Erkenntnisse motivieren die Vorstellung, dass eine Zeitreihe aus dem Zusammenwirken unterschiedlicher Faktoren zustandegekommen ist. Im Bemühen der formalen Beschreibung solcher Zeitreihen wurden Modelle entwickelt, welche derartigen faktischen Gegebenheiten Rechnung tragen.

Aus klassischer Sicht geht man von folgenden 4 Komponenten aus:

1. *Trend* (T), der eine langfristige systematische Veränderung des mittleren Niveaus beschreibt.

2. *Konjunktur* (*K*), mit der mittel- bis längerfristige Schwankungen berücksichtigt werden, wobei allerdings kein unbedingt sich wiederholendes Muster vorausgesetzt wird.

3. *Saison* (*S*), in der jene kurzfristigen Schwankungen zusammengefasst werden, welche sich periodisch nach demselben Muster wiederholen.

4. *Restkomponente* (*U*), als Auffangbecken der nicht systematisch wirksamen Faktoren wie etwa unvorhergesehene Ereignisse oder Messfehler. Da in *U* nur nicht systematische Abweichungen eingehen, gleichen sich deren Realisationen über die ganze Zeitreihe betrachtet etwa aus.

Zunächst wird in einer ersten Grobspezifikation die formale Verknüpfung der Komponenten festgelegt. Man unterscheidet im wesentlichen zwei Modelle.

Additives Modell

$$Y_t = T_t + K_t + S_t + U_t \qquad (7.1)$$

Trend und Konjunktur werden oft zur sogenannt glatten Komponente G_t zusammengefasst. Bekannt ist auch die Vereinigung von Konjunktur und Saison zur zyklischen Komponente Z_t

$$Y_t = G_t + S_t + U_t \quad \text{resp.}$$
$$Y_t = T_t + Z_t + U_t \qquad (7.2)$$

Multiplikatives Modell

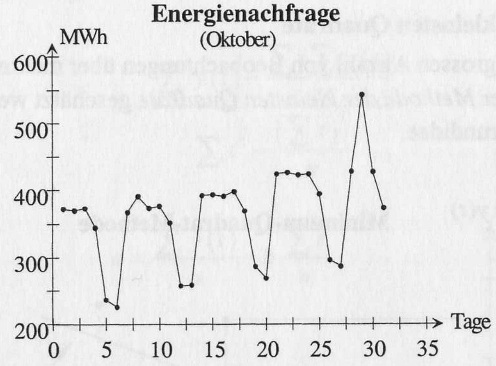
Häufig stellt man fest, dass zyklische Ausschläge im Zeitablauf (mit wachsendem Trend) grösser werden. Ein derartiges Verhalten wird durch ein *multiplikatives* Modell zum Ausdruck gebracht

$$Y_t = G_t \cdot S_t \cdot U_t \qquad (7.3)$$

welches allerdings durch Logarithmieren auf ein additives zurückgeführt werden kann. Denkbar sind auch Mischformen. In der Folge werden nur additive Modelle betrachtet.

Ein zentraler Aspekt der Zeitreihenanalyse besteht darin, Einflussfaktoren spezifischer Komponenten auf die Zeitreihe herauszuarbeiten. Entsprechend dem unterstellten Modell werden die einzelnen Komponenten aus der Zeitreihe gesondert isoliert.

Trägt man die Wertepaare (t, y_t) als Punkte in einem kartesischen Koordinatensystem ein, so erhält man zunächst ein sogenanntes Zeitreihendiagramm resp. Zeitreihenpolygon.



Aus dem Streudiagramm einer Zeitreihe können bereits auf visueller Basis Rückschlüsse auf die grobe Struktur der Komponenten gezogen werden. Auffallend sind vor allem die tendenzielle Entwicklung in langer Frist, sowie zeitlich wiederkehrende Datenmuster.

In der nächsten, über den optischen Eindruck hinausgehenden Phase, sind die berücksichtigten Faktoren näher zu spezifizieren, sei es im Sinne eines funktionalen Ansatzes oder einfach durch eine Vorschrift, welche für konkrete Zeitpunkte Angaben über das Ausmass der jeweiligen Wirkungen gestattet.

7.1.1 Trend

Die folgenden Funktionen sind neben einer Vielzahl möglicher Alternativen geeignet, die längerfristige Entwicklung einer Zeitreihe zu beschreiben.

$$T(t) = a + bt \qquad \text{(linear)}$$

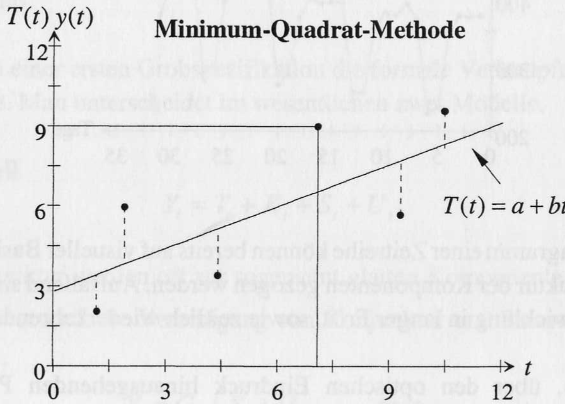
$$T(t) = a\exp(bt) \qquad \text{(exponentiell)}$$

$$T(t) = \frac{a}{1 + b\exp(-ct)} \qquad \text{(Sättigungskurve)}$$

Nach erfolgter Wahl eines Ansatzes sind allfällig unbekannte Parameter aus den vorliegenden Beobachtungen zu schätzen. Beispielhaft sei ein mögliches Verfahren für den linearen Ansatz dargestellt.

7.1.2 Methode der kleinsten Quadrate

Bei einer genügend grossen Anzahl von Beobachtungen über mehrere Zyklen kann der Trend T mit Hilfe der *Methode der kleinsten Quadrate* geschätzt werden. Die folgende Graphik zeigt die Grundidee.



Aus der Punktewolke $\left(t, y(t)\right)$ sind für die lineare Trendfunktion

$$T(t) = a + bt \tag{7.4}$$

die unbekannten Parameter a und b zu bestimmen. Nach dem Minimum-Quadrat-Prinzip werden a und b so gewählt, dass

$$\sum \left[y(t) - T(t)\right]^2 = \sum \left[y(t) - a - bt\right]^2 = f(a,b) \quad \to \min \tag{7.5}$$

Notwendigerweise müssen am Minimum die partiellen Ableitungen von f nach a und b verschwinden

$$\frac{\partial f(a,b)}{\partial a} = 0\,; \quad \frac{\partial f(a,b)}{\partial b} = 0 \tag{7.6}$$

Daraus folgt das Normalgleichungssystem

$$\begin{aligned} n\,a + b\sum t &= \sum y(t) \\ a\sum t + b\sum t^2 &= \sum t y(t) \end{aligned} \tag{7.7}$$

welches nach a und b aufgelöst

$$b = \frac{\sum t y(t) - \dfrac{\sum t \sum y(t)}{n}}{\sum t^2 - \dfrac{\left(\sum t\right)^2}{n}} \qquad a = \frac{\sum y(t)}{n} - b\frac{\sum t}{n} \tag{7.9}$$

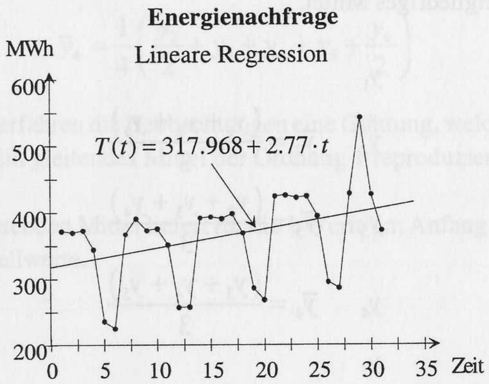
ergibt.

Das Energiebeispiel liefert folgende Regressionsgrössen

$\sum t\, y(t) = 186'566$ $\quad \sum t = 496$ $\quad \sum y(t) = 11'231$

$\sum t^2 = 10'416$ $\quad n = 31$

$b = 2.770$ $\quad a = 317.968$



Unter der Annahme eines konstanten Trends erlaubt die Trendfunktion $T(t)$ prognostische Aussagen über die Zukunft.

$$y(t)-T(t)=y(t)-a-b(t) \tag{7.10}$$

wird als *trendbereinigte* Reihe bezeichnet.

Nach analogem Muster können mit dem Minimum-Quadrat-Prinzip die Parameter für die übrigen Ansätze bestimmt werden. Die Eigenschaften von Minimum-Quadrat-Schätzungen werden im Rahmen der induktiven Statistik diskutiert.

Bei langen Zeitreihen erweisen sich mathematisch funktionale Trendansätze oft als zu abstrakt und zwar sowohl über den verfügbaren Datenbereich als auch im Zusammenhang mit Prognosebetrachtungen.

Man hat deshalb nach Ersatzmethoden gesucht, welche einen empirischen Befund weniger stark formalisieren und trotzdem den charakteristischen Verlauf zum Ausdruck bringen. Ein derartiges Verfahren bilden gleitende Mittel.

7.1.3 Glatte Komponente als gleitende Mittel

Einfache gleitende Durchschnitte einer Zeitreihe sind arithmetische Mittel aufeinanderfolgender Zeitreihenwerte. Den konkreten Beobachtungen y_t werden sukzessive arithmetische Mittel einer zu y_t symmetrischen Umgebung gegenübergestellt. Das Prinzip zeigt ein dreigliedriges Mittel.

$$\begin{array}{ll} y_1 & \\ y_2 & \bar{y}_2=\dfrac{(y_1+y_2+y_3)}{3} \\ y_3 & \bar{y}_3=\dfrac{(y_2+y_3+y_4)}{3} \\ y_4 & \bar{y}_4=\dfrac{(y_3+y_4+y_5)}{3} \\ y_5 & \\ . & \end{array} \tag{7.11}$$

Beim *gleitenden Mittel der Ordnung* $\lambda = 2k+1$ werden mit einer Beobachtung y_t k vorangehende und k nachfolgende Beobachtungen zu einer Gruppe zusammengefasst. Das daraus gebildete arithmetische Mittel wird als Modellwert dem Zeitpunkt t zugeordnet.

$$\bar{y}_t=\frac{1}{2k+1}\sum_{i=-k}^{k} y_{t+i} \tag{7.12}$$

y_t wird in eine zum Zeitpunkt t symmetrische Geschichte eingebettet. Voraussetzung für diese zeitliche Eingliederung ist allerdings eine ungerade Anzahl $2k+1$ von Zeitreihenwerten.

Um auch für den Fall einer geraden Anzahl von Zeitpunkten ein *gleitendes Mittel der Ordnung* $\lambda = 2k$ zu bilden, wird definitorisch festgelegt, dass man, ausgehend von einem Zeitpunkt t, das Mittel über dieselben $2k+1$ Beobachtungen wie bei der Ordnung $(2k+1)$ bildet, wobei die beiden Randwerte jedoch nur je mit dem Gewicht 0.5 berücksichtigt werden. Damit wird auch im Falle einer geraden Anzahl von Zeitreihenwerten eine Beobachtung in eine zum Zeitpunkt t symmetrische Geschichte eingeordnet.

$$\bar{y}_t=\frac{1}{2k}\left(\frac{y_{t-k}}{2}+\sum_{i=-k+1}^{k-1} y_{t+i}+\frac{y_{t+k}}{2}\right) \tag{7.13}$$

Ein gleitendes Mittel der Ordnung 4 für den Zeitpunkt $t = 4$ ergibt

$$\bar{y}_4=\frac{1}{4}\left(\frac{y_2}{2}+y_3+y_4+y_5+\frac{y_6}{2}\right) \tag{7.14}$$

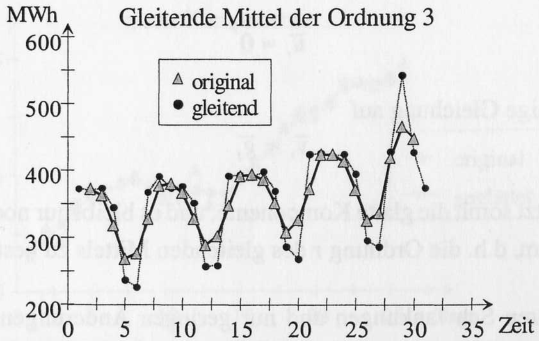
Durch die Mittelung erfahren die Beobachtungen eine Glättung, welche mit wachsender Ordnung zunimmt. Ein gleitendes Mittel der Ordnung 1 reproduziert die Zeitreihe.

Die Methode der gleitenden Mittel liefert für die k Werte am Anfang und am Ende einer Zeitreihe keine Modellwerte.

Beispiel: Tägliche Energienachfrage

y_t	$\bar{y}_t$ $\lambda = 3$	$\bar{y}_t$ $\lambda = 4$	$\bar{y}_t$ $\lambda = 7$
372			
370	371.7		
373	362.3	347.8	
344	317.7	312.6	326.7
236	268.3	293.8	329.4
225	276.0	298.9	329.9
367	327.7	321.9	330.3
391	377.0	357.9	331.3
373	380.0	374.8	334.3
376	366.7	356.0	339.0
351	328.0	324.9	342.6
257	288.7	312.5	342.9
258	302.3	319.9	345.4
392	347.7	341.8	348.6
393	392.0	376.0	351.1
391	394.0	390.6	355.3
398	386.0	374.4	356.7
369	351.0	345.6	361.3
286	307.7	333.5	366.0
268	326.0	343.9	370.4
424	372.7	368.0	374.1
426	424.0	404.5	377.7
422	424.0	420.3	379.0
424	413.3	400.1	381.6
394	371.0	366.8	382.1
295	325.0	350.3	398.7
286	336.3	369.3	399.4
428	418.7	404.3	392.1
542	465.7	431.6	
427	447.3		
373			

Energienachfrage



Bei der oben dargestellten arithmetischen Mittelung werden alle Beobachtungen mit demselben Gewicht berücksichtigt. Diese Annahme ist keineswegs zwingend. Es sind auch Modelle denkbar, bei denen Beobachtungen um so weniger gewichtet werden, je weiter sie vom zeitlichen Mittelpunkt entfernt sind. Voraussetzung ist nur, dass die Summe der nichtnegativen Gewichte 1 ergibt.

Anstelle des arithmetischen Mittels kann auch der Median als gleitendes Mittel verwendet werden. Das Mittelungsverfahren ändert im Prinzip nicht. Vor- und Nachteile der beiden Methoden lassen sich aus den Eigenschaften der involvierten Parameter ableiten.

7.1.4 Modell ohne Saisonkomponente

Im reduzierten Modell ohne Saisonkomponente

$$y_t = G_t + U_t \tag{7.15}$$

ist der Beitrag von G_t zu y_t zu schätzen.

Im wesentlichen geht es um die Elimination zufälliger Schwankungen, welche in U_t aufgefangen werden. Falls das Modell richtig spezifiziert ist, enthält G den systematischen und U den zufallsbedingten Beitrag zu y. Letztere beschreiben Abweichungen, welche sich im Mittel ausgleichen. Betrachtet man das einfache arithmetische Mittel von y über einen bestimmten Zeitraum, so gilt

$$\bar{y}_t = \bar{g}_t + \bar{u}_t \tag{7.16}$$

Unterstellt man wieder eine richtige Modellspezifikation, so gilt unter der Zufälligkeitshypothese über U

$$\bar{u}_t \approx 0 \tag{7.17}$$

womit sich die obige Gleichung auf

$$\bar{y}_t \approx \bar{g}_t \tag{7.18}$$

reduziert. $\bar{y}_t$ schätzt somit die glatte Komponente, und es bleibt nur noch, den optimalen Mittelungszeitraum, d.h. die Ordnung r des gleitenden Mittels zu bestimmen.

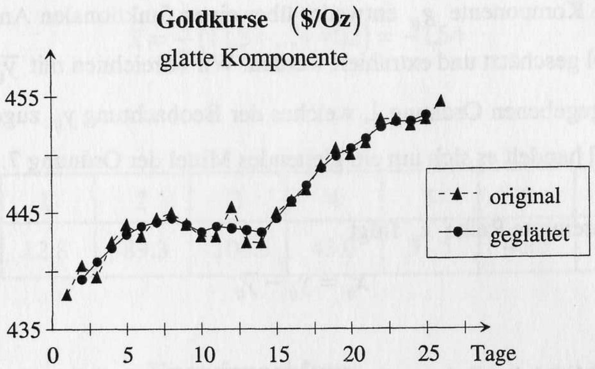
Bei stark irregulären Schwankungen und nur geringen Änderungen des Trends ist r möglichst gross zu wählen und umgekehrt. Gelingt es, ein derart optimales r zu finden, so schätzt

$$\bar{y}_t \cong \bar{g}_t \cong g_t \tag{7.19}$$

die glatte Komponente der Zeitreihe y_t.

<u>Beispiel</u>: Tägliche Goldkurse der Ordnung 3, gleitend gemittelt

Zeit	Kurs ($/Oz) y_t	gleit. Mittel $\bar{y}_t = g_t$ $\lambda = 3$	$u(t)$
1	438		
2	440.5	439.33	-1.17
3	439.5	440.83	1.33
4	442.5	442.00	-0.50
5	444	443.16	-0.83
6	443	443.83	0.83
.	...	.....	....
.	...	.....	....
.	...	.....	....
21	451	451.33	0.33
22	453	452.33	-0.66
23	453	452.83	-0.16
24	452.5	452.83	0.33
25	453	453.33	0.33
26	454.5		
			-0.33



Ein offensichtlicher Nachteil der oben beschriebenen Methode ist, dass wieder am Anfang und insbesondere auch am Ende einer Zeitreihe Schätzwerte für die glatte Komponente fehlen. Für solche Fälle hat man alternative Schätzfunktionen entwickelt, die im wesentlichen ebenfalls auf der Idee gleitender Mittelwerte aufbauen.

Eine allfällig vorhandene Konjunkturkomponente schätzt man als Differenz zwischen der glatten Komponente und dem Trend.

7.1.5 Modell mit konstanter Saisonstruktur

Wir gehen aus vom allgemeinen Modell

$$y_t = g_t + s_t + u_t \tag{7.20}$$

und bezeichnen die Zeitreihe zur Notationsvereinfachung mit

$$y_{ij} \qquad i = 1, 2, \dots k \;\; j = 1, 2, \dots, \lambda \tag{7.21}$$

y_{ij} ist die j-te Beobachtung innerhalb des i-ten von insgesamt k Zyklen. Unter einer konstanten Saisonstruktur versteht man ein periodisch wiederkehrendes Verlaufsmuster nach λ Zeiteinheiten.

Aus dem Modell

$$y_{ij} = g_{ij} + s_j + u_{ij} \tag{7.22}$$

kann die glatte Komponente g_{ij} entweder über einen funktionalen Ansatz oder über gleitende Mittel geschätzt und extrahiert werden. Wir bezeichnen mit $\overline{y}_{ij}$ das gleitende Mittel der vorgegebenen Ordnung λ, welches der Beobachtung y_{ij} zugeordnet ist. Im obigen Beispiel handelt es sich um ein gleitendes Mittel der Ordnung 7.

Für die trendbereinigte Reihe x_{ij} folgt

$$x_{ij} = y_{ij} - \overline{y}_{ij} \tag{7.23}$$

Ein gleitendes Mittel der Ordnung $\lambda = 7$ aus dem Energiebeispiel ergibt

i \ j	1	2	3	4	5	6	7
1	17.3	-93.4	-104.9	36.7	59.7	38.7	37.0
2	8.4	-85.9	-87.4	43.4	41.9	35.7	41.3
3	7.7	-80.0	-102.4	49.9	48.3	43.0	42.1
4	11.9	-103.7	-113.4	35.9			
$\overline{x}_j$	11.3	-90.8	-102.0	41.5	50.0	39.1	40.1

Durch die Mittelung der saisonbedingten Abweichungen über alle verfügbaren Perioden

$$\overline{x}_j = \frac{1}{k^*} \sum_i x_{ij} \qquad k^* = \text{Anzahl verfügbarer Zyklen} \tag{7.24}$$

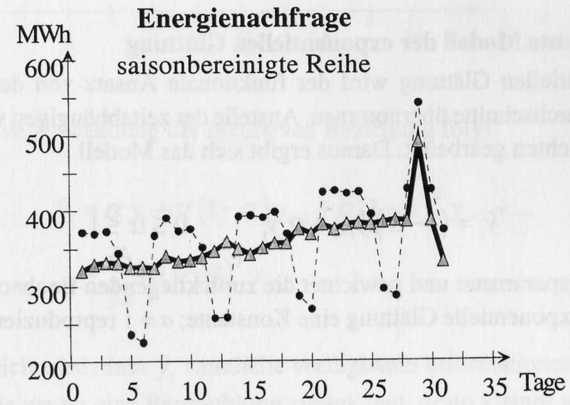
wird neben einer ersten Schätzung der saisonalen Beiträge die Restkomponente ausgeschaltet. Zur definitiven Schätzung der Saisonkomponenten werden die $\overline{x}_j$ so normiert, dass deren Summe Null ergibt. Dies wird durch Transformation

$$\hat{s}_j = \overline{x}_j - \frac{1}{\lambda} \sum_{j=1}^{\lambda} \overline{x}_j = \overline{x}_j - \overline{x} \tag{7.25}$$

erreicht und führt im obigen Beispiel zu folgenden Resultaten

$$\overline{x} = \frac{1}{7}(11.3 + \ldots + 40.1) = -1.54 \tag{7.26}$$

j	1	2	3	4	5	6	7
$\hat{s}_j$	12.8	-89.3	-100.5	43.0	51.5	40.6	41.6



7.2 Glättung

Bei der Glättung versucht man, 'Unebenheiten' aus den Ursprungsdaten, welche möglicherweise die zeitliche Struktur einer Datenreihe verwischen, herauszufiltern.

7.2.1 Fortlaufende Durchschnitte

Im Zusammenhang mit der Diskussion von Stichprobenparametern wurde das arithmetische Mittel als Repräsentant einer numerischen Datenmenge eingeführt. Bei einer Zeitreihe lässt sich die Idee der arithmetischen Mittelung übernehmen, indem als Modellwert für den Beobachtungszeitpunkt t das arithmetische Mittel aller n bisher verfügbaren Beobachtungen $y_t, y_{t-1}, \ldots, y_{t-n+1}$ gewählt wird.

$$\hat{y}_t = \frac{1}{n}\sum_{i=0}^{n-1} y_{t-i} = \frac{1}{n}\sum_{i=1}^{n-1} y_{t-i} + \frac{1}{n} y_t = \frac{n-1}{n}\hat{y}_{t-1} + \frac{1}{n} y_t \qquad (7.27)$$

Der Modellwert $\hat{y}_t$ ist eine Linearkombination des Modellwertes der Vorperiode sowie der aktuellen Beobachtung. Mit fortschreitender Zeit nimmt das Gewicht der aktuellen Beobachtung ab und jenes des letzten Modellwertes zu. Diese auf Stabilität hinauslaufende Eigenschaft wird oft als nachteilig empfunden. Einen Ausweg aus dem Dilemma bieten sowohl gleitende Mittel als auch die Methode der exponentiellen Glättung.

7.2.2 Das konstante Modell der exponentiellen Glättung

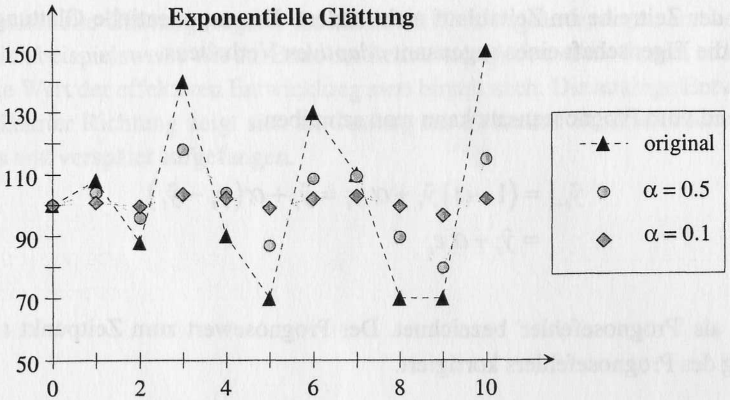
Bei der exponentiellen Glättung wird der funktionale Ansatz von der Methode der fortlaufenden Durchschnitte übernommen. Anstelle der zeitabhängigen wird jedoch mit konstanten Gewichten gearbeitet. Daraus ergibt sich das Modell

$$\hat{y}_t = (1-\alpha)\,\hat{y}_{t-1} + \alpha\, y_t \qquad 0 \le \alpha \le 1 \qquad (7.28)$$

α heisst Glättungsparameter und gewichtet die zurückliegenden Beobachtungen. Für $\alpha = 0$ liefert die exponentielle Glättung eine Konstante; $\alpha = 1$ reproduziert die Zeitreihe.

Beispiel:

t	0	1	2	3	4	5	6	7	8	9	10
$y(t)$	100	108	88	140	90	70	130	110	70	70	150



Durch sukzessive Anwendung der rekursiven Beziehung folgt

$$\begin{aligned}\hat{y}_t &= \alpha\, y_t + \alpha(1-\alpha)\, y_{t-1} + \alpha(1-\alpha)^2\, y_{t-2} + \ldots \\ &= \sum_{j=0}^{\infty} \alpha(1-\alpha)^j\, y_{t-j}\end{aligned} \qquad (7.29)$$

woraus ersichtlich wird, dass $\hat{y}_t$ sämtliche verfügbaren Informationen der Vergangenheit vereinigt. Je weiter eine Beobachtung zurückliegt, desto kleiner wird ihr Gewicht. Die Summe der Gewichte beträgt 1.

In praktischen Anwendungen können die Beobachtungen nicht beliebig weit in die Vergangenheit zurückverfolgt werden. In solchen Fällen wird die geglättete Zeitreihe durch die Vorschrift $\hat{y}_1 = y_1$ initialisiert. Die Summe der Gewichte erreicht den Wert 1 nicht mehr ganz.

Die exponentielle Glättung wird auch zur *kurzfristigen Prognose* verwendet. Zu diesem Zwecke wählt man den für die Periode t geglätteten Wert als Prognose für den Zeitpunkt $t+1$ und schreibt

$$\hat{y}_{t+1} = \alpha\; y_t + \alpha(1-\alpha)\; y_{t-1} + \ldots \qquad (7.30)$$

Das Modell zeigt deutlich, dass künftige Werte einzig aus dem bisherigen Verlauf der Zeitreihe geschätzt werden. Andere Einflussfaktoren finden keine Beachtung. Trotz der ausschliesslichen Abstützung auf die vergangenen Beobachtungen muss nicht im gleich strengen Mass wie bei den gleitenden Mitteln vorausgesetzt werden, dass sich die

Struktur der Zeitreihe im Zeitablauf nicht ändere. Die exponentielle Glättung besitzt nämlich die Eigenschaft eines sogenannt *adaptiven* Verhaltens.

Ausgehend vom Prognoseansatz kann man schreiben

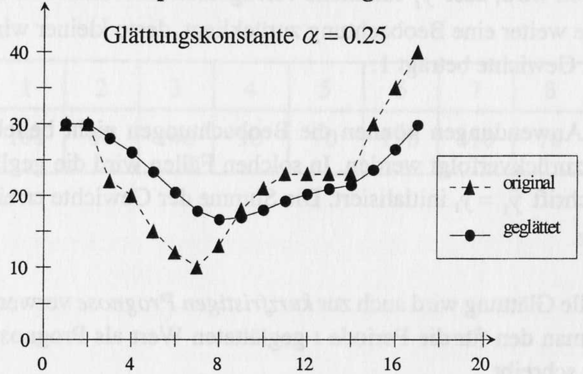
$$\begin{aligned} \hat{y}_{t+1} &= (1-\alpha)\,\hat{y}_t + \alpha\, y_t = \hat{y}_t + \alpha\,(y_t - \hat{y}_t) \\ &= \hat{y}_t + \alpha\, e_t \end{aligned} \tag{7.31}$$

e^t wird als 'Prognosefehler' bezeichnet. Der Prognosewert zum Zeitpunkt t wird in Richtung des Prognosefehlers korrigiert.

Das oben dargestellte Verfahren der exponentiellen Glättung ist vor allem dann angebracht, wenn die Zeitreihe weder einen Trend noch saisonale Einflüsse aufweist. Sind diese Voraussetzungen nicht erfüllt, so ist das Modell entsprechend zu erweitern.

Die nachfolgende Graphik zeigt einige Eigenschaften und Konsequenzen der exponentiellen Glättung.

Exponentielle Glättung



Die exponentielle Glättung reagiert vor allem bei Richtungsänderungen erst verzögert. Ändert sich beispielsweise wie am Ende der Zeitreihe angedeutet der Trend, so hinkt der geglättete Wert der effektiven Entwicklung *stets* hinten nach. Die analoge Entwicklung in umgekehrter Richtung zeigt sich am Anfang der Zeitreihe. Strukturbrüche werden ebenfalls erst verspätet aufgefangen.

LITERATURVERZEICHNIS

Bamberg G., Baur F.
Statistik, 4. A., Oldenbourg, München, Wien 1985

Bohley P.
Statistik, Einführendes Lehrbuch für Wirtschafts- und Sozialwissenschaftler, 2. A., Oldenbourg, München, 1987

Bleymüller J., Gehlert G., Gütlicher H.
Statistik für Wirtschaftswissenschaftler, 4. A., Vahlen, München 1985

Calot G.
Cours de Statistique déscriptive, 2. A., Dunod, Paris 1984

Esenwein-Rothe I.
Die Methoden der Wirtschaftsstatistik, Vandenhoeck & Ruprecht, Göttingen 1976

Hackl P., Katzenbeisser W., Panny W.
Statistik, 3. A., Oldenbourg, München, Wien 1981

Hartung J., Elpelt B., Klösner K.-H.
Statistik, 2. A., Oldenbourg, München, Wien 1984

Kreyszig E.
Statistische Methoden und ihre Anwendung, 7. A., Vandenhoeck & Ruprecht, Göttingen 1982

Rutsch M.
Statistik 1, Mit Daten umgehen, Birkhäuser Skripten, Basel, Boston, Stuttgart, 1986

Zwer R.
Internationale Wirtschafts- und Sozialstatistik, Oldenbourg, München, Wien 1981

INDEX

Verlag Wilhelm Surbir

Dufourstrasse 48 · CH-9000 St. Gallen
Tel. +4171-2242301/2242300/2983616 · Fax 2242646 · E-Mail: Norbert.Reetz@UNISG.CH

Lieferbare Titel

Allgoewer, Elisabeth, Dr.
Ökonomische Theoriebildung und Zeit. Eine methodenkritische Analyse anhand ausgewählter Arbeiten J.R. Hicks', 1992 (St. Galler Dissertation), Fr. 42.00

Bartmann, Hermann, Prof. Dr.
Allokationstheorie. Vorlesung, 2. Auflage 1993, Fr. 25.00

Bartmann, Hermann, Prof. Dr. und ***John***, Klaus-Dieter, Prof. Dr.
Grundkonzeptionen der Konjunktur- und Wachstumsanalyse. Beiträge zur Wirtschaftstheorie
Band 1, Klassik, Neoklassik, Keynes und Keynesianismus, 4. Auflage 1994, Fr. 20.00
Band 2, Monetaristisch-neoklassische Position und Supply-Side-Ökonomien, 4. Auflage 1994, Fr. 20.00
Band 3, Postkeynesianismus, 4. Auflage 1994, Fr. 20.00

Bartmann, Hermann, Prof. Dr., ***Busch***, Andreas A., Diplom-Volkswirt und ***Schwaab***, Jan A., Diplom-Volkswirt, Preis- und Wettbewerbstheorie. Vorlesung, 6. Auflage 1999, Fr. 45.00

Borchers, Henning, Dr.
Regulierte Strommärkte. Ein Beitrag zur (De-)Regulierungsdebatte in der Elektrizitätswirtschaft, 1994 (Mainzer Dissertation), Fr. 39.00

Brauchlin, Emil, Prof. Dr., ***Schips***, Bernd, Prof. Dr., ***Stier***, Winfried, Prof. Dr. und ***Studer***, Hans-Peter, Dr.
Statistische Methoden. Ihr sachgerechter Einsatz in der empirischen Wirtschafts- und Sozialforschung. Ein Kompendium, 3. Auflage 1987, Fr. 33.00

Einführung in die Wissenschaftstheorie für Nationalökonomen. Verfaßt von der Volkswirtschaftlichen Abteilung des Doktorandenseminars für Wissenschaftstheorie an der Hochschule St. Gallen
Band 1, hrsg. v. Prof. Dr. Walter Adolf ***Jöhr*** in Zusammenarbeit mit Dr. Gerhard ***Schwarz***, 1979, Fr. 33.00
Band 2, hrsg. v. Prof. Dr. Walter Adolf ***Jöhr*** und Prof. Dr. Bernd ***Schips*** in Zusammenarbeit mit Dr. Gerhard ***Schwarz***, 1980, Fr. 16.00

Föller, Alex, Dr.
Umwelthaftungsrecht und Schadensprävention. Eine ökonomische Analyse der Haftung für Umweltschäden unter Einbeziehung juristischer, ökologischer und versicherungstheoretischer Aspekte, 1994 (Mainzer Dissertation), Fr. 45.00

Frenkel, Michael, Prof. Dr.
Einführung in die Makroökonomik offener Volkswirtschaften, 2. Auflage 1993 (unveränderter Nachdruck 1995), Fr. 39.50

Gauglhofer, Margrit, Prof. Dr. und ***Müller***, Heinz, Prof. Dr.
Mathematik für Oekonomen
Band 1, 9. Auflage 1999, Fr. 36.00
Band 2, 9. Auflage 2000, Fr. 20.00

Guyer, Philipp, Dr.
Der „Non-Market-Clearing"-Ansatz der Ungleichgewichtstheorie und seine Anwendung auf das keynesianische makroökonomische Standardmodell, 1981 (St. Galler Dissertation), Fr. 37.00

John, Klaus-Dieter, Prof. Dr.
Verteilungskonflikte, Inflation und Beschäftigung. Ungleichgewichtsökonomische Ansätze und sozialwissenschaftliche Erweiterungen, 1982 (Mainzer Dissertation), Fr. 44.00

KANTIge Worte und Sprueche aus berufenem Munde, von den Traegern eben derselben autorisiert und zurecht gerueckt. Zu Nutzen und Frommen nachfolgender Schuelergenerationen gesammelt an der hochwohlloeblichen und ehrbaren mathematischen Abteilung der Kantonsschule St. Gallen von deren ehemaligen Zoeglingen Carola und Matthias ***Reetz***, 1986, Fr. 12.00

Keel, Alex, Prof. Dr.
Statistik
Band 1, Beschreibende Statistik, 14. Auflage 1999, Fr. 21.00
Band 2, Wahrscheinlichkeit, 13. Auflage 1999, Fr. 21.00
Band 3, Induktive Statistik, 14. Auflage 2000 , Fr. 23.00

Knecht, René, Dr.
Die Humankapitaltheorie als Ansatz zur Erklärung der personellen Arbeitseinkommensverteilung, 1988 (St. Galler Dissertation), Fr. 42.00

Koch, Christine, Dr.
Wachstum und Einkommensverteilung in postkeynesianischen Ansätzen, 1999 (Mainzer Dissertation), Fr. 48.00

Matthes, Rainer, Dr.
Zur ökonometrischen Spezifikation von Beschäftigungsfunktionen. Eine empirische Untersuchung für die BR Deutschland, 1991 (Mainzer Dissertation), Fr. 42.00

Minsch, Jürg, Dr.
Ursache und Verursacherprinzip im Umweltbereich. Zur theoretischen Fundierung einer verursacherorientierten Umweltpolitik, 1988 (St. Galler Dissertation), Fr. 45.00

Von Musen, Müttern und der Mathematik: Frauen(an)sichten. hrsg. von Annabeth ***Naef-Hinderling*** und Johanna ***Schönenberger-Deuel***, 1998, Fr. 20.00

Räth, Norbert, Dr.
Die Zwangsanleihe als finanzpolitisches Instrument, 1980 (Mainzer Dissertation), Fr. 39.50

Reetz, Gesine, Sozialarbeiterin (grad.)
Rückfallprognose in der Bewährungshilfe. Eine Untersuchung anhand von Erfahrungen mit Probanden der Reutlinger Bewährungshilfe 1960-1971, 1979, Fr. 10.00

Reetz, Norbert, Prof. Dr.

Konjunktur und Wachstum. Eine Einführung in die reale Theorie. Vorlesung, 5. Auflage 1987, Fr. 20.00

Produktionstheorie. Vorlesung, 2. Auflage 1989, Fr. 20.00

Grundzüge der makroökonomischen Theorie. Vorlesung, 5. Auflage 1990, Fr. 30.00

Grundzüge der mikroökonomischen Theorie. Vorlesung, 5. Auflage 1991, Fr. 30.00

Einführung in die mikroökonomische Theorie. Vorlesung, 9. Auflage 2000, Fr. 28.00

Anhang zu „Einführung in die mikroökonomische Theorie. Vorlesung (9. Auflage 2000)". Klausuren. Aufgaben und Lösungen, 2000, 20.00 Fr.

Reine Theorie der Außenwirtschaft. Vorlesung, 1995, Fr. 40.00

Schierjott, Alexander, Dr.
Mengenrationierung und Arbeitsmarkt. Theoretische Untersuchungen und empirische Ergebnisse für die Bundesrepublik Deutschland, 1984 (Mainzer Dissertation), Fr. 29.00

Schilling, Günter, Dr.
Rationale Erwartungen in makroökonomischen Modellen, 1987 (Mainzer Dissertation), Fr. 35.00

Schindler, Rosemarie, Dr.
Die Marktpolitik des Roheisenverbandes während der Weimarer Republik, 1978 (Tübinger Dissertation), Fr. 39.20

Schlotjunker, Stefan, Dr.
The Constructed Evolution of Technology. A Constructivist-Evolutionary Approach to Technological Change and its Empirical Evidence, 1994 (St. Galler Dissertation), Fr. 42.00

Schmidt, Joachim, Dr.
Regionales Konsumverhalten. Theoretische Überlegungen und empirische Ergebnisse für ausgewählte Bundesländer der Bundesrepublik Deutschland, 1987 (Mainzer Dissertation), Fr. 42.00

Schmidt, Norbert, Dr.
Investorenverhalten und konjunkturelle Stabilität, 1987 (Mainzer Dissertation), Fr. 42.00