UNIVERSITÄT
DUISBURG
ESSEN

WIRTSCHAFTSWISSENSCHAFTEN
BWL VWL RECHTSWISSENSCHAFTEN
INFORMATIK WIRTSCHAFTSINFORMATIK
MEDIZIN-MANAGEMENT

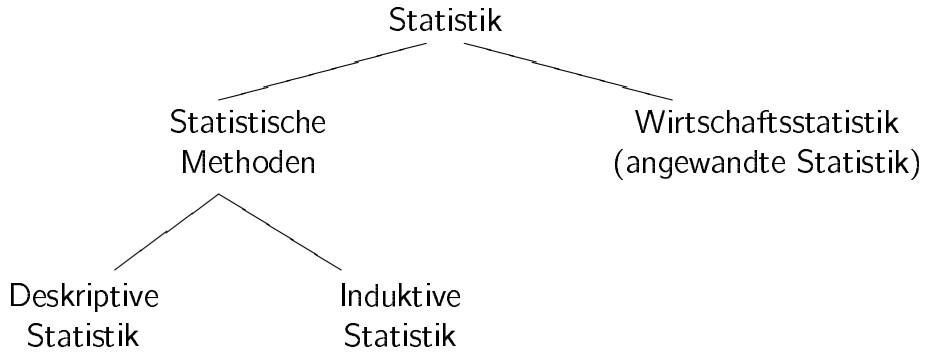
Folien zur Übung Deskriptive Statistik

Michael Westermann

zuletzt geändert: 28. Juli 2004

Gegenstand der Statistik

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



Wer liefert die Informationen?

Statistische Einheit

Unter *statistischen Einheiten* (Elemente, Merkmalsträger) versteht man die Träger von Informationen bzw. Eigenschaften, welche im Rahmen einer empirischen Untersuchung von Interesse sind.

Statistische Masse

Eine *statistische Masse* (Kollektiv, Population) ist eine sachlich, räumlich und zeitlich sinnvoll abgegrenzte Gesamtheit von statistischen Einheiten.

Grundgesamtheit	Teilgesamtheit
Bestandsmasse (zeit**punkt**bezogen)	Bewegungsmasse (zeit**raum**bezogen)
Bestand 0 + Bewegung = Bestand 1	

Was wird untersucht?

Merkmal/Realisationen

Ein *Merkmal* ist eine Eigenschaft einer statistischen Einheit, die bei einer statistischen Untersuchung von Interesse ist. Die unterschiedlichen Erscheinungsformen eines Merkmals heißen *Merkmalsausprägung* oder *Realisationen*.

Quantitativ
zähl– bzw. messbar
Intensiv
lediglich Durchschnittswerte sinnvoll interpretierbar
Diskret
Anzahl der Realisationen ist mit den natürlichen Zahlen abzählbar

Qualitativ
durch die Art unterscheidbar
Extensiv
auch Summen sind sinnvoll interpretierbar
Stetig
überabzählbar unendlich viele Realisationen

Wie können die Informationen verarbeitet werden?

Skalenarten

Unter einer *Skala* wird die Zahlenmenge definiert, die zur Bezeichnung von Merkmalsausprägungen verwendet werden kann.

- *Nominal*: keine Rangfolge, lediglich Unterscheidbarkeit der Realisationen.
- *Ordinal*: Rangfolge, die Unterschiede der Ausprägungen lassen sich jedoch nicht (sinnvoll) interpretieren.

Metrische Skalen:

- *Intervall*: Rangfolge und Abstände zwischen je zwei Ausprägungen sind definiert.
- *Ratio*: Rangfolge, Abstände und Verhältnis zweier Merkmalsausprägungen sind definiert ($\rightarrow$ natürlicher Nullpunkt).
- *Absolut*: zusätzlich ist eine natürliche Zähleinheit vorgegeben.

Informationsgehalt ↓

Daten, Datensatz und Datengewinnung

Daten

Statistische *Daten* sind der Ausgangspunkt weitergehender statistischer Auswertungen. Es sind Zahlenangaben über Merkmalsausprägungen, die an Einheiten beobachtet bzw. gemessen worden sind.

Datensatz

Alle sachlich zusammengehörigen und einer statistischen Auswertung zugrundeliegenden Daten bilden einen *Datensatz*.

Methoden der Datengewinnung

1. *Datenerhebung*: Systematische Gewinnung statistischer Daten durch Befragen/Messen/Experiment.
2. *Datenaufbereitung*: Urliste $\rightarrow$ geordnete Statistische Reihe $\rightarrow$ Häufigkeitsverteilung $\rightarrow \ldots \rightarrow$ graphische Darstellung
3. *Datenauswertung*: adäquate Anwendung statistischer Methoden

Maßzahlen

Maßzahlen dienen der zusammenfassenden Beschreibung von Daten. Ziel ist dabei die informationsverdichtende Beschreibung von Daten.
Eine Maßzahl ist eine Funktion f, die den Realisationen $x_1, x_2, \ldots, x_n$ des Merkmals X eine Zahl M zuordnet.

$$M = f(x_1, x_2, \ldots, x_n)$$

Die Funktion g heisst gewogene Maßzahl, wenn sie den Realisationen $x_1, x_2, \ldots, x_n$ in Verbindung mit den entsprechenden Gewichten $g_1, g_2, \ldots, g_n$ eine Zahl G zuordnet.

$$G = g\big[(x_1, g_1), (x_2, g_2), \ldots, (x_n, g_n)\big]$$

Beispiel: Arithmetisches Mittel: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i n_i$
Maßzahl $G = \bar{x}$, Gewichte $g_1 = n_1, g_2 = n_2, \ldots$, Funktion $g = \frac{1}{n} \sum_{i=1}^{n}$

Arbeitstabelle — unklassierte Daten

Datensatz vom Umfang n. Merkmal X mit insgesamt m verschiedenen Merkmalsausprägungen $x_1, x_2, \ldots, x_m$ und den entsprechenden absoluten Häufigkeiten $n_1, n_2, \ldots, n_m$.

	(1)	(2)	(3)	(4)	(5)
i	x_i	n_i	h_i	N_i	H_i
1	x_1	n_1	h_1	N_1	H_1
2	x_2	n_2	h_2	$N_2 = N_1 + n_2$	$H_2 = H_1 + n_2$
3	x_3	n_3	h_3	$N_3 = N_2 + n_3$	$H_3 = H_2 + n_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	x_m	n_m	h_m	n	1
Σ		n	1		

Aus den Spalten (1), (2) bzw. (3) wird das Stabdiagramm und aus den Spalten (1), (4) bzw. (5) die empirische Verteilungsfunktion erstellt.

Arbeitstabelle — klassierte Daten

Aufteilung der Merkmalsausprägungen in p Klassen mit den entsprechenden absoluten Klassenhäufigkeiten $n_1, n_2, \ldots, n_p$.

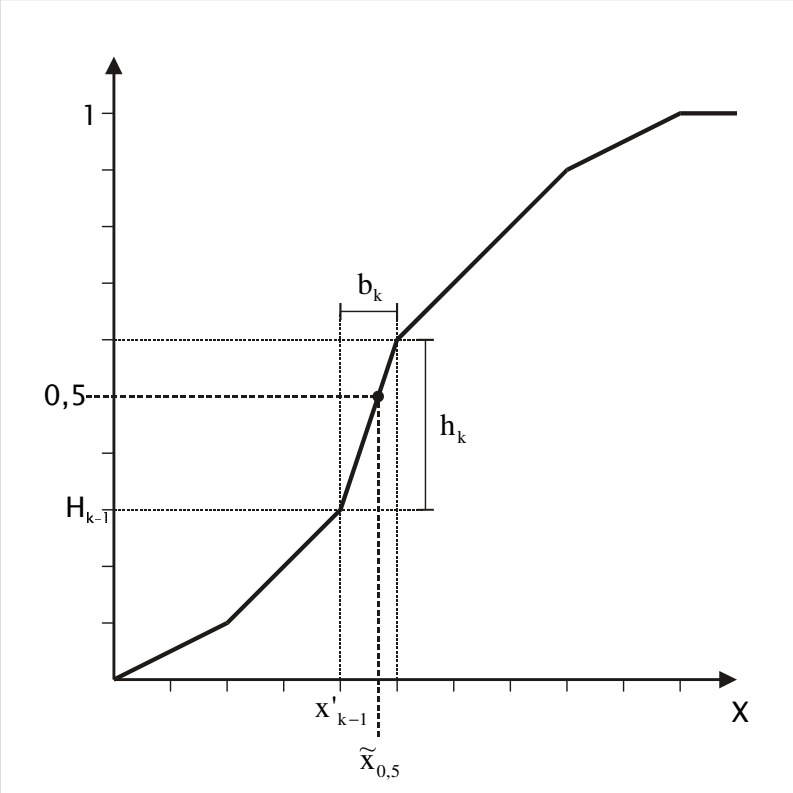
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
k	$(x'_{k-1}; x'_k]$	n_k	h_k	N_k	H_k	b_k	n^*_k	h^*_k
1	$(0; x'_1]$	n_1	h_1	N_1	H_1	b_1	n^*_1	h^*_1
2	$(x'_1; x'_2]$	n_2	h_2	N_2	H_2	b_2	n^*_2	h^*_2
3	$(x'_2; x'_3]$	n_3	h_3	N_3	H_3	b_3	n^*_3	h^*_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	$(x'_{p-1}; x'_p]$	n_m	h_m	n	1	b_k	n^*_k	h^*_k
Σ		n	1					

Aus den Spalten (1), (7) bzw. (8) wird das Histogramm und aus den Spalten (1), (4) bzw. (5) die empirische Verteilungsfunktion erstellt.
Dabei ist das Prinzip der Flächentreue zu berücksichtigen, d.h. zur Darstellung der Daten im Histogramm müssen die absoluten bzw. relativen Häufigkeits**dichten** bestimmt werden [Spalten (6), (7) und (8)].

$$n^*_k = \frac{n_k}{b_k} \qquad \text{bzw.} \qquad h^*_k = \frac{h_k}{b_k}$$

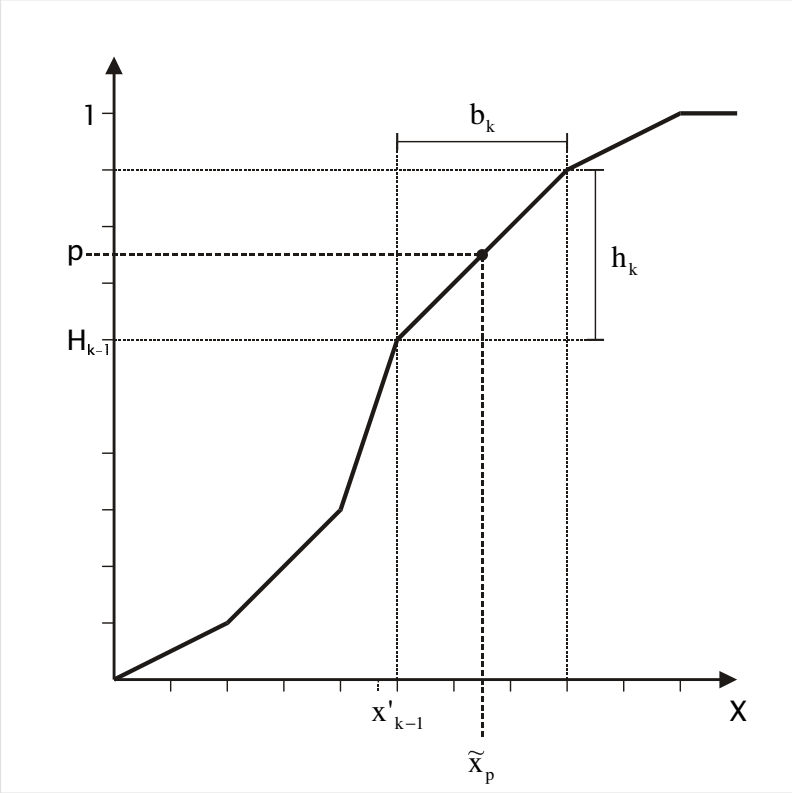
Median und Quantile

Bestimmung des Medians (p = 0,5)



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

Bestimmung von Quantilen



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

Geometrisches Mittel

$$\bar{w} = \left(\prod_{i=1}^{T} w_i \right)^{\frac{1}{T}} = \left(\frac{y_T}{y_0} \right)^{\frac{1}{T}}$$

In der folgenden Tabelle ist die Verschuldung der öffentlichen Haushalte in der Bundesrepublik Deutschland dargestellt:

Jahr	Verschuldung y_t (in Mio. DM)	t	w_t
1982	614820	0	
1983	671708	1	$w_1 = 671708/614820 = 1,09$
1984	717522	2	$w_2 = 717522/671708 = 1,07$
1985	760182	3	$w_3 = 760182/717522 = 1,06$
1986	800967	4	$w_4 = 800967/760182 = 1,05$

Berechnen Sie den durchschnittlichen Wachstumsfaktor bzw. die durchschnittliche Wachstumsrate der Verschuldung des Bundeshaushalts im betrachteten Zeitraum!
Zunächst werden die Wachstumsfaktoren ermittelt (letzte Spalte in der Tabelle):

$$w_t = \frac{y_t}{y_{t-1}}$$

Wird der Ausgangswert y_0 mit diesen (unterschiedlichen) Wachstumsfaktoren multipliziert, gelangt man zum Endwert y_4, also

$$614820 \cdot 1,09 \cdot 1,07 \cdot 1,06 \cdot 1,05 = 800967$$

Die Frage lautet: Welcher durchschnittliche (gleichbleibende) Wachstumsfaktor führt zum gleichen Ergebnis ($\rightarrow$ geometrisches Mittel)?

$$614820 \cdot \bar{w}^4 = 800967 \qquad \Rightarrow \bar{w} = \left(\frac{800967}{614820} \right)^{1/4} = 1,07$$

$$\Rightarrow \bar{w} = (1,09 \cdot 1,07 \cdot 1,06 \cdot 1,05)^{1/4} = 1,07$$

Wäre die Verschuldung jeweils um den (konstanten) Faktor 1,07 gewachsen, so wäre man am Ende des Zeitraums ebenfalls zu einer Verschuldung von 800967 gelangt.

Harmonisches Mittel

→ Aufgabe 4.12

Der Rocker R kam leider nie in den Genuss, eine Statistikvorlesung zu hören. Es gelingt ihm deshalb nicht, das folgende Problem zu lösen:

R möchte auf der Hin– und Rückfahrt zu seiner 4km entfernten Stammkneipe eine Durchschnittsgeschwindigkeit von 60km/h fahren. Dabei traut er sich auf dem Rückweg nur eine Geschwindigkeit von 30km/h zu. Muss er deshalb auf dem Hinweg 90km/h fahren?

Lösung über das Harmonische Mittel!

$$\bar{x}_h = \frac{8km}{\frac{4km}{90km/h} + \frac{4km}{30km/h}} = 45km/h$$

Seine Annahme ist somit falsch. Bei einer Geschwindigkeit von 90km/h auf der Hinfahrt erreicht er lediglich eine Durchschnittsgeschwindigkeit von 45km/h.

Anmerkung: Die gewünschte Durchschnittsgeschwindigkeit ist gar nicht mehr zu erreichen, da er für 8km bei einer Durchschnittsgeschwindigkeit von 60km/h genau 8 Minuten für die Gesamtstrecke benötigen würde. Diese 8 Minuten hat er jedoch schon für die Rückfahrt eingeplant (4km/30km/h = 8min).

Gini's Dispersionsmaß (Gini's mittlere Differenz)

Berechnung

$$S_g = \frac{2}{n^2 - n} \sum_{v<w} |x_v - x_w| = \frac{2(a_1 + a_2 + \cdots + a_n)}{n^2 - n} = \frac{2A}{n^2 - n}$$

Herleitung

	(1)	(2)	(3)		(n)	
	x_1	x_2	x_3	...	x_n	Σ
x_1	0	$\lvert x_1 - x_2 \rvert$	$\lvert x_1 - x_3 \rvert$	...	$\lvert x_1 - x_n \rvert$	a_1
x_2		0	$\lvert x_2 - x_3 \rvert$	...	$\lvert x_2 - x_n \rvert$	a_2
x_3			0	...	$\lvert x_3 - x_n \rvert$	a_3
$\vdots$				$\ddots$	$\vdots$	$\vdots$
x_n					0	a_n
					Σ	**A**

- zeilenweise Summation der Spalten (1) bis (n) ergibt die Elemente a_1 bis a_n.
- aufaddieren der Teilsummen a_1 bis a_n ergibt **A**
- Insgesamt besteht die Tabelle aus n^2 Feldern. Davon haben genau n Elemente den Wert 0 (Alle Elemente auf der Hauptdiagonalen). Somit existieren $n^2 - n$ von Null verschiedene absolute Abweichungen.
- Berücksichtigt man die Symmetrie der absoluten Abweichungen ($|x_1 - x_2| = |x_2 - x_1|$), so wird klar, dass die Summe **aller** absoluten Abweichungen genau dem Zweifachen der errechneten Summe **A** entspricht.
- Gini's Maß ist das arithmetische Mittel aller absoluten Abweichungen, also

$$S_g = \frac{2A}{n^2 - n}.$$

Gini's Dispersionsmaß (Gini's mittlere Differenz)

Beispiel

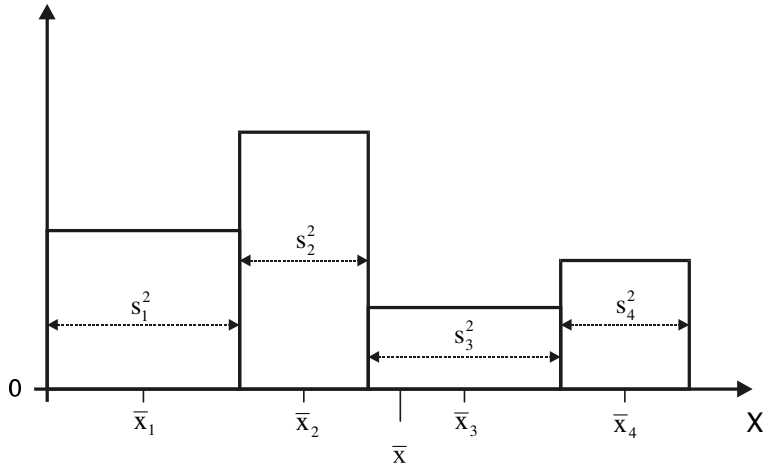
Man berechne Gini's mittlere Differenz für die folgende Zahlenfolge: $41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51$

	41	42	43	44	45	46	47	48	49	50	51	Σ
41	0	1	2	3	4	5	6	7	8	9	10	55
42		0	1	2	3	4	5	6	7	8	9	45
43			0	1	2	3	4	5	6	7	8	36
44				0	1	2	3	4	5	6	7	28
45					0	1	2	3	4	5	6	21
46						0	1	2	3	4	5	15
47							0	1	2	3	4	10
48								0	1	2	3	6
49									0	1	2	3
50										0	1	1
51											0	0
											Σ	**220**

$$S_g = \frac{2 \cdot A}{n^2 - n} = \frac{2 \cdot 220}{11^2 - 11} = \frac{440}{110} = 4$$

Varianz bei klassierten Daten

Merkmal X aufgeteilt in insgesamt p Klassen. Vorliegen der Klassenmittelwerte $\bar{x}_k$ und der Klassenvarianzen s_k^2



$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{k=1}^{p}(\bar{x}_k - \bar{x})^2 n_k + \frac{1}{n}\sum_{k=1}^{p} s_k^2 n_k \\
&= \sum_{k=1}^{p}(\bar{x}_k - \bar{x})^2 h_k \quad + \sum_{k=1}^{p} s_k^2 h_k \\
&= s_{ext}^2 \qquad\qquad\qquad + s_{int}^2
\end{aligned}
$$

Bei unbekannten Klassenmittelwerten sind die Klassen**mitten** zu verwenden. Die externe Varianz wird dann zu

$$
s_{ext}^2 = \sum_{k=1}^{p}(m_k - \hat{\bar{x}})^2 h_k, \qquad \text{mit } \hat{\bar{x}} = \sum_{k=1}^{p} m_k h_k.
$$

Herfindahlindex und Rosenbluthindex

Beispiel

Für fünf Tageszeitungen wurden folgende Marktanteile ermittelt:

Zeitung	A	B	C	D	E
Marktanteil	0,15	0,10	0,25	0,20	0,30

Berechnen Sie den Herfindahlindex und den Rosenbluthindex

Arbeitstabelle:

i	q_i	q_i^2	iq_i
1	0,3	0,0900	0,3
2	0,25	0,0625	0,5
3	0,20	0,0400	0,6
4	0,15	0,0225	0,6
5	0,1	0,0100	0,5
Σ	1,0	0,2250	2,5

Herfindahlindex

$$K_H = \sum_{i=1}^{n} q_i^2 = 0,225$$

Rosenbluthindex

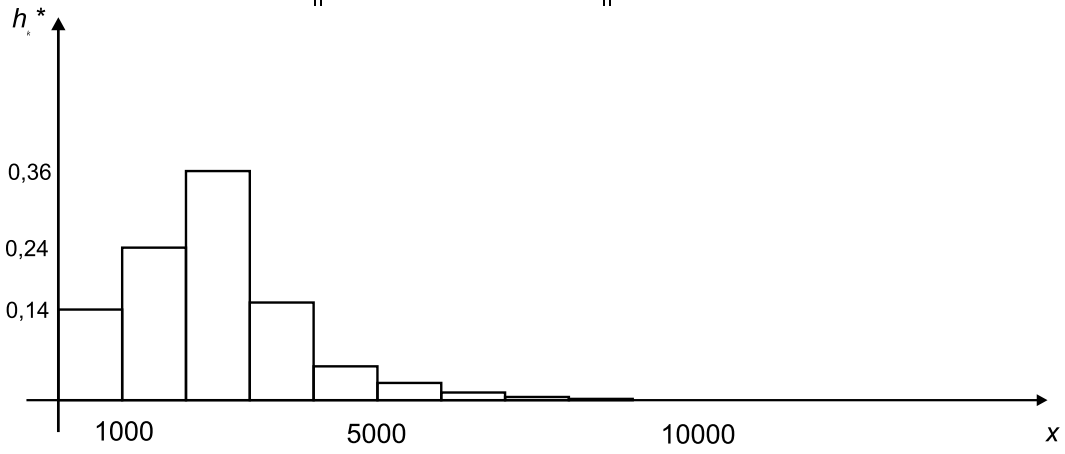
$$K_R = \frac{1}{\left(2\sum_{i=1}^{n} iq_i\right) - 1} = \frac{1}{(2 \cdot 2,5) - 1} = 0,25$$

Von der Häufigkeitsverteilung zur Lorenzkurve

Beispiel

Dem *SozioOEkonomischen Panel* (SOEP) des *Deutschen Instituts für Wirtschaftsforschung* (DIW) kann man für das Jahr 1996 folgende monatliche Nettopersoneneinkommen entnehmen:

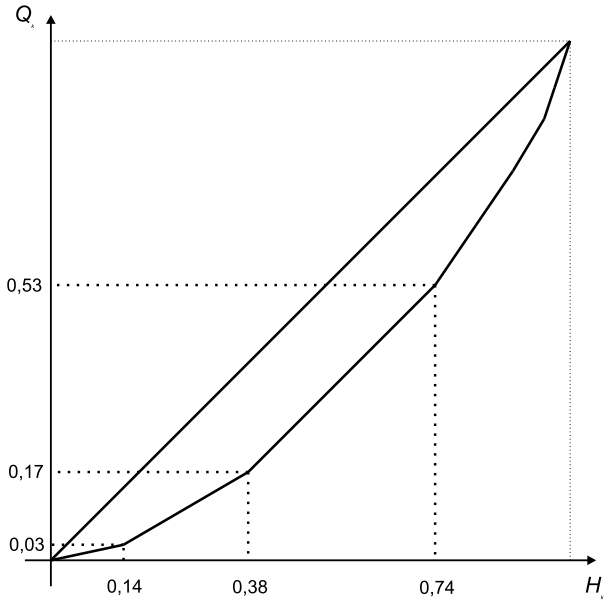
Einkommen(in DM) $[x'_{k-1}; x'_k)$	n_k	h_k	H_k	m_k	b_k	$m_k h_k$	h^*_k
[0; 999)	1060	0,14	0,14	499,5	999	70,9	0,14
[1000; 1999)	1785	0,24	0,38	1499,5	999	358,2	0,24
[2000; 2999)	2682	0,36	0,74	2499,5	999	897,1	0,36
[3000; 3999)	1145	0,15	0,89	3499,5	999	536,2	0,15
[4000; 4999)	397	0,05	0,95	4499,5	999	239,0	0,05
[5000; 5999)	204	0,03	0,97	5499,5	999	150,1	0,03
[6000; 6999)	90	0,01	0,99	6499,5	999	78,3	0,01
[7000; 7999)	40	0,01	0,99	7499,5	999	40,1	0,01
[8000; 8999)	33	0,00	1,00	8499,5	999	37,5	0,00
[9000; 9999)	15	0,00	1,00	9499,5	999	19,1	0,00
[10000; 30000)	22	0,00	1,00	20000	20000	58,9	0,00
	7473	1		2485,3			



Von der Häufigkeitsverteilung zur Lorenzkurve

Zur graphischen Darstellung der Disparität wird nun die (geschätzte) Merkmalssumme ($\rightsquigarrow$ **Klassenmitten**) der jeweiligen Einkommensklassen herangezogen, so dass den Merkmalsträgern der auf sie entfallende Anteil an der Merkmalssumme gegenübergestellt werden kann. Die lineare Verbindung der $(H_k; Q_k)$–Kombinationen ergibt dann die Lorenzkurve.[1]

Einkommen(in DM) $[x'_{k-1}; x'_k)$	n_k	h_k	$\mathbf{H_k}$	$m_k n_k$	$q_k = \frac{m_k n_k}{\sum m_k n_k}$	$\mathbf{Q_k}$
[0; 999)	1060	0,14	**0,14**	529470,0	0,03	**0,03**
[1000; 1999)	1785	0,24	**0,38**	2676607,5	0,14	**0,17**
[2000; 2999)	2682	0,36	**0,74**	6703659,0	0,36	**0,53**
[3000; 3999)	1145	0,15	**0,89**	4006927,5	0,22	**0,75**
[4000; 4999)	397	0,05	**0,95**	1786301,5	0,10	**0,85**
[5000; 5999)	204	0,03	**0,97**	1121898,0	0,06	**0,91**
[6000; 6999)	90	0,01	**0,99**	584955,0	0,03	**0,94**
[7000; 7999)	40	0,01	**0,99**	299980,0	0,02	**0,95**
[8000; 8999)	33	0,00	**1,00**	280483,5	0,02	**0,97**
[9000; 9999)	15	0,00	**1,00**	142492,5	0,01	**0,98**
[10000; 30000)	22	0,00	**1,00**	440000,0	0,02	**1,00**
	7473	1		18572774,5	1	



[1] Zur exakten Darstellung müssten die Werte genauer gerundet werden, da bei der gewählten Darstellung die letzten drei H_k–Werte nicht mehr zu unterscheiden sind.

Lorenzkurve und Gini–Koeffizient

Beispiel

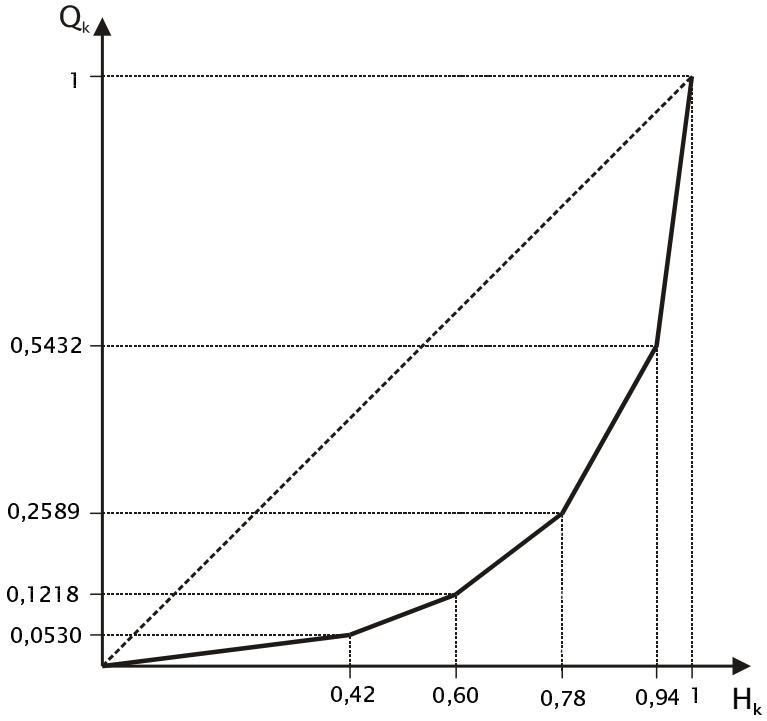
In einer Region wurde bei 50 landwirtschaftlichen Betrieben die Größe der landwirtschaftlichen Nutzfläche ermittelt. Zeichnen Sie die Lorenzkurve und ermitteln Sie den Gini–Koeffizienten.

Fläche (in ha) $[x'_{k-1}; x'_k)$	n_k	h_k	H_k	m_k	$m_k n_k$	$q_k = \frac{m_k n_k}{\sum m_k n_k}$	Q_k	$q_i(H_i + H_{i-1})$
$[0; 5)$	21	0,42	0,42	2,5	52,5	0,0533	0,0533	0,0224
$[5; 10)$	9	0,18	0,60	7,5	67,5	0,0685	0,1218	0,0699
$[10; 20)$	9	0,18	0,78	15	135,0	0,1371	0,2589	0,1891
$[20; 50)$	8	0,16	0,94	35	280,0	0,2843	0,5432	0,4889
$[50; 250)$	3	0,06	1,00	150	450,0	0,4569	1,00	0,8863
$\sum$	$n = 50$	1			985,0	1		1,6566

Gini–Koeffizient

$$D_G = \sum_{i=1}^{n} q_i \left(H_i + H_{i-1}\right) - 1 = 1,6566 - 1 = 0,6566$$

Lorenzkurve



Interpretation:

- Auf die 42% kleinsten Betriebe entfallen 5,3% der landwirtschaftlichen Nutzfläche.
- Auf die 60% kleinsten Betriebe entfallen 12,18% der landwirtschaftlichen Nutzfläche.
- Auf die 78% kleinsten Betriebe entfallen 25,89% der landwirtschaftlichen Nutzfläche.
- Auf die 94% kleinsten Betriebe entfallen 54,32% der landwirtschaftlichen Nutzfläche.
- Auf die 6% größten Betriebe entfallen 45,68% der landwirtschaftlichen Nutzfläche.

Bivariate Häufigkeitstabelle

Zwei Merkmale X und Y mit m bzw. k verschiedenen Merkmalsausprägungen.

	y_1	y_2	y_3	$\dots$	y_k	$n_{i\cdot}$
x_1	n_{11}	n_{12}	n_{13}	$\dots$	n_{1k}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	n_{23}	$\dots$	n_{2k}	$n_{2\cdot}$
x_3	n_{31}	n_{32}	n_{33}	$\dots$	n_{3k}	$n_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	n_{m1}	n_{m2}	n_{m3}	$\dots$	n_{mk}	$n_{m\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$\dots$	$n_{\cdot k}$	n

gemeinsame absolute bzw. relative Häufigkeiten

$$n_{ij} = n(X = x_i \text{ und } Y = y_j) \quad \text{mit} \quad n_{ij} \geq 0 \qquad \text{und} \quad \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} = n$$

$$h_{ij} = \frac{n_{ij}}{n} \qquad \text{mit} \quad 0 \leq h_{ij} \leq 1 \quad \text{und} \quad \sum_{i=1}^{m}\sum_{j=1}^{k} h_{ij} = 1$$

Randverteilungen von X bzw. Y

$$n_{i\cdot} = n(X = x_i) = \sum_{j=1}^{k} n_{ij} \quad \text{bzw.} \quad h_{i\cdot} = h(X = x_i) = \sum_{j=1}^{k} h_{ij}$$

$$n_{\cdot j} = n(Y = Y_j) = \sum_{i=1}^{m} n_{ij} \quad \text{bzw.} \quad h_{\cdot j} = h(Y = y_j) = \sum_{i=1}^{m} h_{ij}$$

Parameter der Randverteilungen

Mittelwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_{i\cdot} = \sum_{i=1}^{m} x_i h_{i\cdot}$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{k} y_j n_{\cdot j} = \sum_{j=1}^{k} y_j h_{\cdot j}$$

Varianz

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{m} x_i^2 n_{i\cdot} - \bar{x}^2 = \sum_{i=1}^{m} x_i^2 h_{i\cdot} - \bar{x}^2$$

$$s_y^2 = \frac{1}{n}\sum_{j=1}^{k} y_j^2 n_{\cdot j} - \bar{y}^2 = \sum_{j=1}^{k} y_j^2 h_{\cdot j} - \bar{y}^2$$

Bedingte Verteilungen

	y_1	y_2	y_3	$\dots$	y_k	$n_{i\cdot}$
x_1	n_{11}	n_{12}	n_{13}	$\dots$	n_{1k}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	n_{23}	$\dots$	n_{2k}	$n_{2\cdot}$
x_3	n_{31}	n_{32}	n_{33}	$\dots$	n_{3k}	$n_{3\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_m	n_{m1}	n_{m2}	n_{m3}	$\dots$	n_{mk}	$n_{m\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$\dots$	$n_{\cdot k}$	n

Bedingte Verteilungen

$$h_{i|j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{h_{ij}}{h_{\cdot j}} = h(x|Y = y_j)$$

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

Bedingte Mittelwerte

$$\bar{x}|y = \sum_{i=1}^{m} x_i h_{i|j}$$

$$\bar{y}|x = \sum_{j=1}^{k} y_j h_{j|i}$$

Empirische Unabhängigkeit

Bei Unabhängigkeit der Merkmale gilt:

$$n_{ij} = \frac{n_{i.} n_{.j}}{n} \quad \text{bzw.} \quad h_{ij} = h_{i.} h_{.j}$$

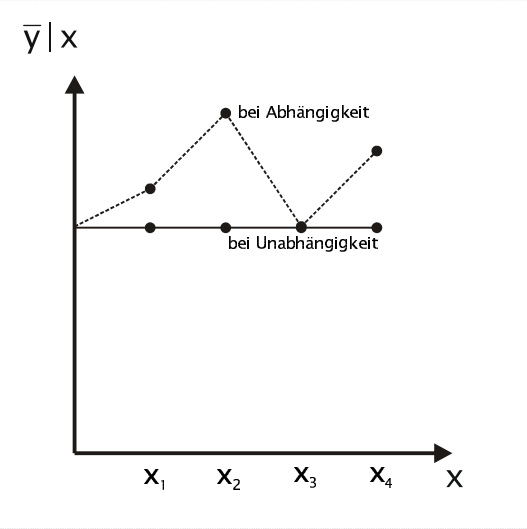
$$n_{i|j} = n_{i.} \quad \text{bzw.} \quad h_{i|j} = h_{i.} \Rightarrow \bar{x}|y = \bar{x}$$

$$n_{j|i} = n_{.j} \quad \text{bzw.} \quad h_{j|i} = h_{.j} \Rightarrow \bar{y}|x = \bar{y}$$

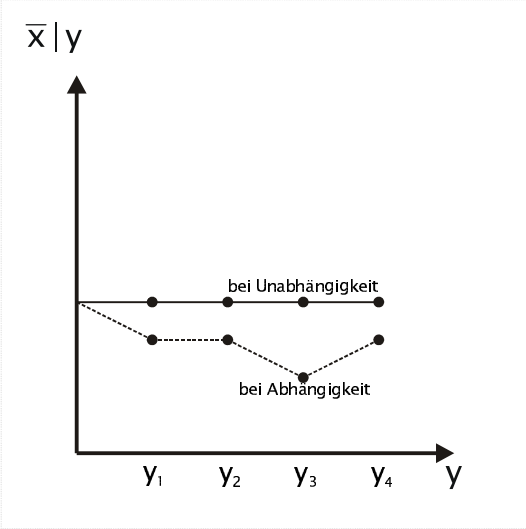
Empirische Regressionslinie

Lineare Verbindung aller bedingten Mittelwerte eines Merkmals

emp. Regressionslinie für Y



emp. Regressionslinie für X



Kovarianz und Korrelation

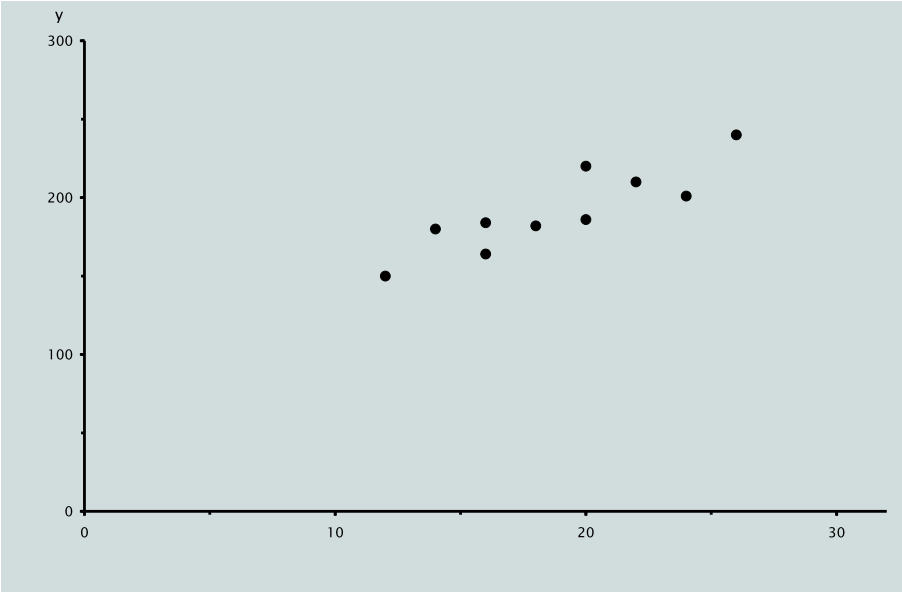
Beispiel

Der Geschäftsführer einer Großbäckerei wünscht Informationen über die Entwicklung des monatlichen Umsatzes und ob seine Marketingaktivitäten erfolgreich sind. Eine Überprüfung von zehn zufällig ausgewählten Monatsumsätzen und den dazugehörigen Marketingausgaben ergibt folgende Zahlen:

Umsatz y	201	184	220	240	180	164	186	150	182	210
Marketingausgaben x	24	16	20	26	14	16	20	12	18	22

Quelle: Schira (2003) — Statistische Methoden der VWL und BWL, S. 113.

Einen ersten Eindruck über den möglichen Zusammenhang der beiden Größen erhält man, indem man ein Streudiagramm anfertigt, welches für den vorliegenden Fall wie folgt aussieht:



Beispiel — Fortsetzung

Um den Zusammenhang zwischen Marketingausgaben und Umsatz numerisch beziffern zu können, fertigt man eine Arbeitstabelle an. Mit dieser können dann die Kovarianz sowie der Korrelationskoeffizient der beiden Größen ermittelt werden:

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	24	201	576	4824	40401
2	16	184	256	2944	33856
3	20	220	400	4400	48400
4	26	240	676	6240	57600
5	14	180	196	2520	32400
6	16	164	256	2624	26896
7	20	186	400	3720	34596
8	12	150	144	1800	22500
9	18	182	324	3276	33124
10	22	210	484	4620	44100
$\sum$	188	1917	3712	36968	373873

$$\bar{x} = \frac{188}{10} = 18,8$$

$$\bar{y} = \frac{1917}{10} = 191,7$$

$$s_x^2 = \tfrac{1}{10}\sum x^2 - \bar{x}^2 = 371,2 - (18,8)^2 = 17,76$$

$$s_y^2 = \tfrac{1}{10}\sum y^2 - \bar{y}^2 = 37387,3 - (191,7)^2 = 638,41$$

$$s_{xy} = \tfrac{1}{10}\sum xy - \bar{x}\cdot\bar{y} = 3696,8 - 18,8\cdot 191,7 = 92,84$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{92,84}{\sqrt{17,76\cdot 638,41}} = 0,872$$

Regressionsanalyse

Es soll eine Gerade durch die Punktewolke der gemeinsamen Beobachtungen angepasst werden (Vgl. vorheriges Beispiel). Es wird ein linearer funktionaler Zusammenhang zwischen X und Y angenommen, also

$$y = f(x) = y = a + bx.$$

Da nicht davon ausgegangen werden kann, dass **alle** gemeinsamen Beobachtungen exakt auf der Geraden liegen, wird der funktionale Zusammenhang um eine Störgröße (*Residuum*) ergänzt:

$$y_i = a + bx_i + u_i.$$

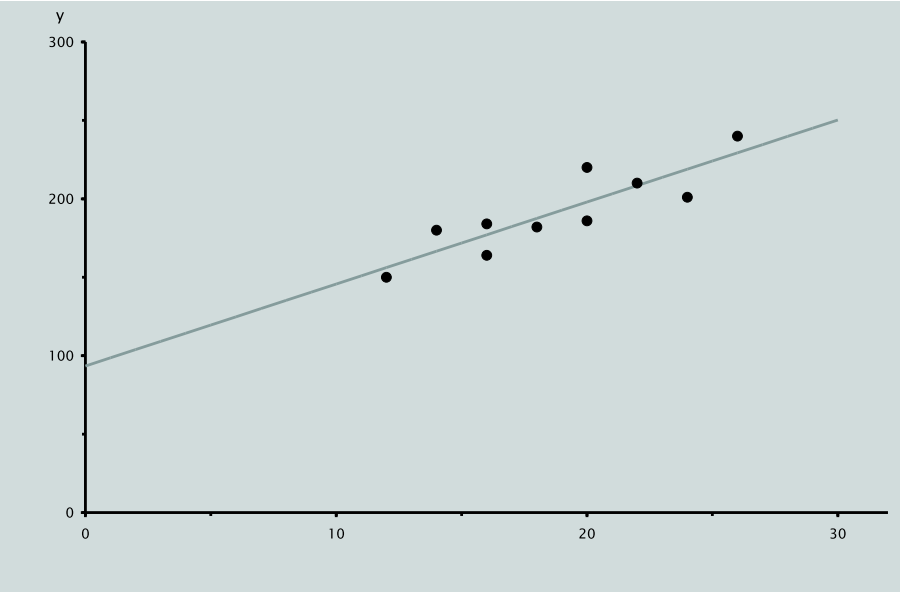
Die *Methode der Kleinsten Quadrate*[2] bestimmt die Parameter a und b so, dass die Quadratsumme der Residuen minimiert wird. Man gelangt dann zu den Bestimmungsgleichungen

$$b = \frac{s_{xy}}{s_x^2}$$
$$a = \bar{y} - b\bar{x}$$

[2] In englischer Sprache: OLS — *Ordinary Least Squares*.

Beispiel — Fortsetzung

Für das vorherige Beispiel erhält man demnach folgende geschätzte Parameter der Regressionsgeraden:



$$b = \frac{s_{xy}}{s_x^2} = \frac{92,84}{17,76} = 5,23$$

$$a = \bar{y} - b\bar{x} = 191,7 - 5,23 \cdot 18,8 = 93,42$$

$$\Rightarrow \quad \hat{y} = 93,42 + 5,23x$$

Einleitung und Wertindex

Ein Index ist eine Maßzahl der aggregierten Veränderung, d.h. eine Maßzahl für den Vergleich von Merkmalsgesamtheiten. Gegeben sind zwei Zeitpunkte $(0, t)$, die Güter $i = 1, \ldots, n$, sowie die Preise (p_{i0}, p_{it}) und die Mengen (q_{i0}, q_{it}) der n Güter zu den zwei verschiedenen Zeitpunkten. Der Wert eines Gutes i des Aggregats wird (zum Zeitpunkt 0) bestimmt durch

$$\text{Wert}_{i0} = \text{Preis}_{i0} \cdot \text{Menge}_{i0}$$

und kann durch eine reine Preisänderung, eine reine Mengenänderung oder durch eine gemeinsame Änderung beider Größen beeinflußt werden.
Der Wertindex als Maßzahl der tatsächlichen Ausgabenveränderung ist definiert als:

$$W_{0t} = \frac{\sum_{i=1}^{n} p_{it} q_{it}}{\sum_{i=1}^{n} p_{i0} q_{i0}} = \frac{(2)}{(1)}$$

Indizes nach Laspeyres und Paasche

Ein Index kann dargestellt werden als ein gewogener Meßzahlenmittelwert.

$$I_{0t} = \sum_{i=1}^{n} m_{0t}^{(i)} g_i$$

Je nach Wahl der zu mittelnden Größe (Preis– oder Mengenmeßzahl) und der Art der Gewichte (aus der Basisperiode oder aus der Berichtsperiode) gelangt man zu Preis– bzw. Mengenindizes nach Laspeyres oder Paasche. So läßt sich z.B. der Preisindex nach Laspeyres als gewogenes arithmetisches Mittel der Preismeßzahlen ermitteln. Die Gewichte berechnen sich dabei als Ausgabenanteil des jeweiligen Gutes in der Basisperiode:

$$P_{0t}^{L} = \sum_{i=1}^{n} \frac{p_{it}}{p_{i0}} \cdot \frac{p_{i0} q_{i0}}{\sum p_{i0} q_{i0}}$$

Indizes nach Laspeyres und Paasche (Aggregatformel)

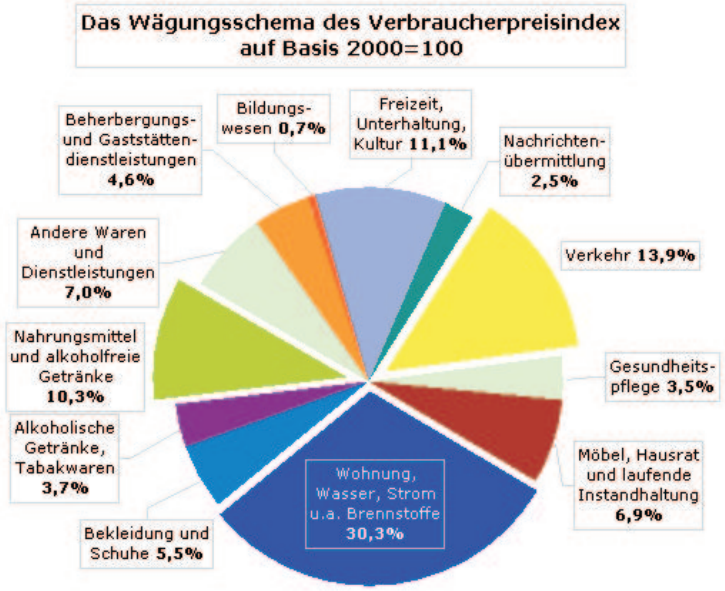
	Preismeßzahlen	$p \leftrightarrow q$	Mengenmeßzahlen
Laspeyres (Basisperiode)	$P_{0t}^{L} = \frac{\sum_{i=1}^{n} p_{it} \cdot q_{i0}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i0}} = \frac{(4)}{(1)}$		$Q_{0t}^{L} = \frac{\sum_{i=1}^{n} q_{it} \cdot p_{i0}}{\sum_{i=1}^{n} q_{i0} \cdot p_{i0}} = \frac{(3)}{(1)}$
Paasche (Berichtsperiode)	$P_{0t}^{P} = \frac{\sum_{i=1}^{n} p_{it} \cdot q_{it}}{\sum_{i=1}^{n} p_{i0} \cdot q_{it}} = \frac{(2)}{(3)}$		$Q_{0t}^{P} = \frac{\sum_{i=1}^{n} q_{it} \cdot p_{it}}{\sum_{i=1}^{n} q_{i0} \cdot p_{it}} = \frac{(2)}{(4)}$
Fisher (geometrisches Mittel)	$P_{0t}^{F} = \sqrt{P_{0t}^{L} \cdot P_{0t}^{P}}$		$Q_{0t}^{F} = \sqrt{Q_{0t}^{L} \cdot Q_{0t}^{P}}$

Arbeitstabelle

	$t=0$		$t=1$		(1)	(2)	(3)	(4)
i	p_{i0}	q_{i0}	p_{it}	q_{it}	$p_{i0}q_{i0}$	$p_{it}q_{it}$	$p_{i0}q_{it}$	$p_{it}q_{i0}$
1	p_{10}	q_{10}	p_{1t}	q_{1t}	$p_{10}q_{10}$	$p_{1t}q_{1t}$	$p_{10}q_{1t}$	$p_{1t}q_{10}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	p_{n0}	q_{n0}	p_{nt}	q_{nt}	$p_{n0}q_{n0}$	$p_{nt}q_{nt}$	$p_{n0}q_{nt}$	$p_{nt}q_{n0}$
Σ					(1)	(2)	(3)	(4)

Was ist der Verbraucherpreisindex?

Der Verbraucherpreisindex misst die **durchschnittliche Preisveränderung aller Waren und Dienstleistungen, die von privaten Haushalten für Konsumzwecke gekauft werden.**[...] Die Ausgaben für diese Waren und Dienstleistungen **repräsentieren die durchschnittlichen Verbrauchsgewohnheiten** privater Haushalte, wie sie sich im sogenannten Warenkorb widerspiegeln. **Die Gewichtung dieser einzelnen Positionen des Warenkorbs ist im Wägungsschema festgelegt.**



Quelle: Statistisches Landesamt Baden-Württemberg — http://www.statistik-bw.de/Indexzahlen/vWkInfo.asp

Einleitung

Eine *Zeitreihe* ist eine zeitlich geordnete Folge von Beobachtungen.
Die *Zeitreihenanalyse* befaßt sich mit Methoden zur Beschreibung dieser Vorgänge. Zu den Aufgaben der Zeitreihenanalyse gehören u.a.

- die Prognose zukünftiger Werte,
- die Analyse und Erkennen von Mustern und deren Ursachen, sowie
- die Kontrolle von Prozessen.

Das Komponentenmodell

Es liegen die Beobachtungen $y_1, y_2, y_3, \ldots, y_T$ vor, wobei von jeder Beobachtung angenommen wird, dass sie sich aus einer systematischen und einer nichtsystematischen Komponente zusammensetzt. Die systematische Komponente besteht aus

Trend:	langfristiger Einfluß, welcher nicht periodisch ist
Konjunktur:	mehrjährige, wiederkehrende Schwankungen
Saison:	zeitlich (saisonal) bedingte Schwankungen

$$y_t = \underbrace{\boxed{\underbrace{m_t}_{\text{Trend}} + \underbrace{k_t}_{\text{Konjunktur}} + \underbrace{s_t}_{\text{Saison}}}}_{\text{systematischer Teil}} + \underbrace{r_t}_{\text{Rest}}$$

Beispiel

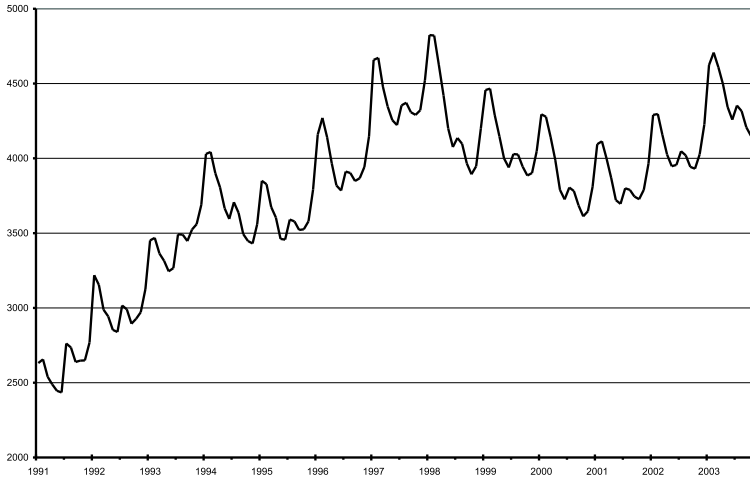
Jeweilige Monatsendstände der Zahl der reistrierten Arbeitslosen von 1991 bis 2003 (in 1000).

	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
1991	2.631	2.656	2.539	2.489	2.446	2.435	2.762	2.735	2.638	2.647	2.649	2.769
1992	3.219	3.154	2.988	2.943	2.854	2.839	3.016	2.990	2.894	2.928	2.971	3.126
1993	3.451	3.469	3.364	3.315	3.245	3.266	3.492	3.490	3.447	3.525	3.560	3.689
1994	4.029	4.042	3.900	3.807	3.665	3.595	3.707	3.636	3.493	3.447	3.430	3.560
1995	3.850	3.827	3.674	3.605	3.461	3.457	3.591	3.578	3.521	3.526	3.579	3.791
1996	4.159	4.270	4.141	3.967	3.818	3.785	3.912	3.902	3.848	3.867	3.942	4.148
1997	4.658	4.672	4.477	4.347	4.256	4.222	4.354	4.372	4.308	4.290	4.322	4.522
1998	4.824	4.821	4.625	4.421	4.200	4.076	4.137	4.097	3.966	3.894	3.947	4.198
1999	4.456	4.466	4.289	4.147	3.999	3.939	4.029	4.026	3.944	3.885	3.902	4.048
2000	4.295	4.278	4.142	3.988	3.789	3.726	3.805	3.782	3.685	3.612	3.646	3.810
2001	4.095	4.114	4.000	3.869	3.721	3.696	3.800	3.789	3.744	3.727	3.789	3.964
2002	4.290	4.297	4.157	4.025	3.948	3.956	4.048	4.020	3.943	3.931	4.027	4.227
2003	4.624	4.707	4.610	4.497	4.343	4.259	4.353	4.316	4.208	4.151	4.184	4.315

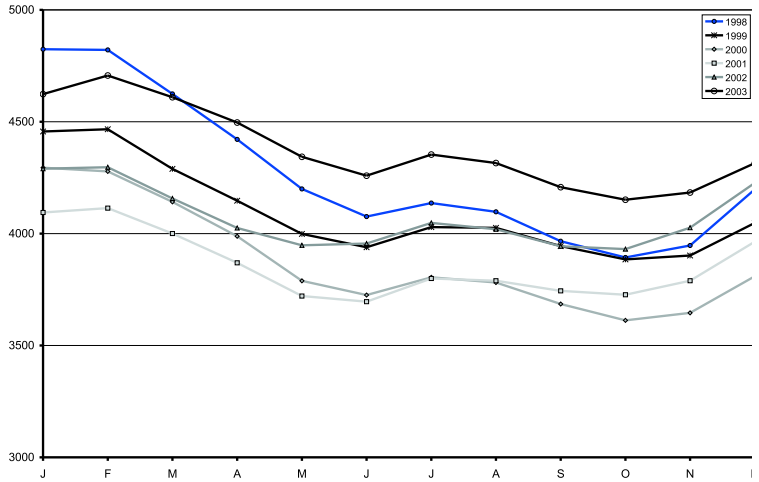
Quelle: Deutsche Bundesbank

Graphische Darstellung

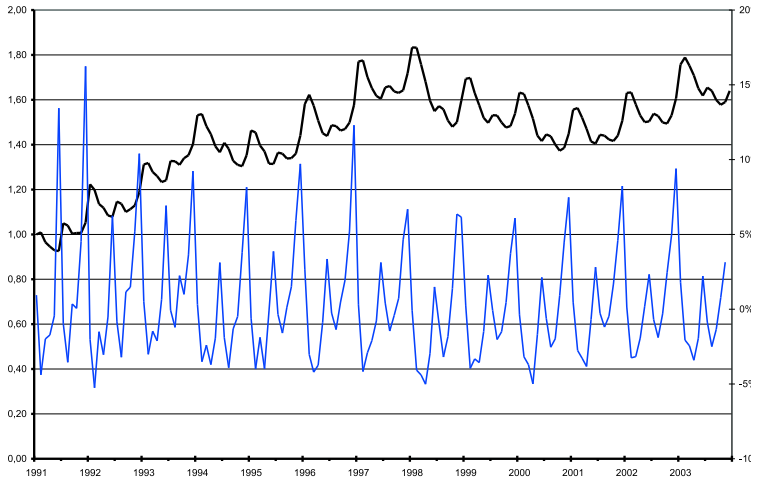
Jahresverlauf



Saisonaler Verlauf



Messziffern (linke Achse) und Wachstumsraten (rechte Achse)



Bestimmung der Trendkomponente m_t

Methode der kleinsten Quadrate

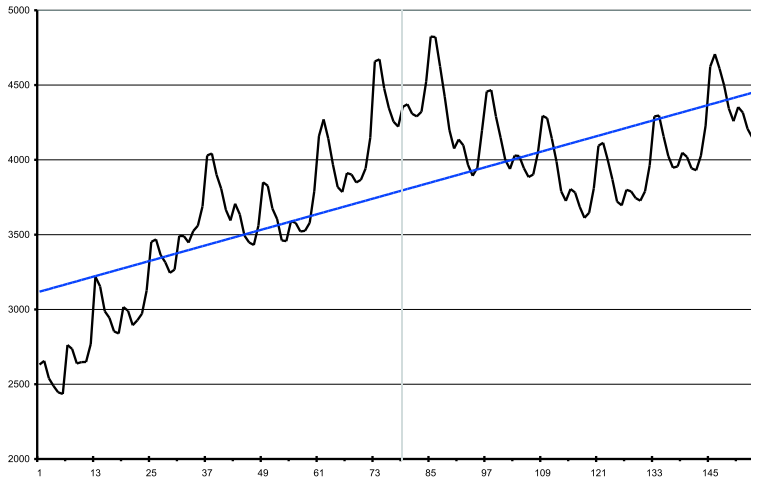
Gesucht: $\hat{y}_t = a + b \cdot t$

Lösung
$(t = 1, 2, 3, \ldots, T)$
$b = \frac{s_{yt}}{s_t^2} = \frac{\sum(y_t - \bar{y})(t - \bar{t})}{\sum(t - \bar{t})^2} = \frac{\overline{yt} - \bar{y} \cdot \bar{t}}{\overline{t^2} - \bar{t}^2}$
$a = \bar{y} - b \cdot \bar{t}$

Alternative
$(t_z = \ldots, -3, -2, -1, 0, 1, 2, 3, \ldots)$
$b = \frac{\sum y_{t_z} \cdot t_z}{\sum t_z^2}$
$a = \bar{y}$

Lösung:

$\hat{y}_t = 3110 + 8,69 \cdot t$ bzw. $\hat{y}_{t_z} = 3792 + 8,69 \cdot t_z$



Bestimmung der glatten Komponente ($m_t + k_t$)

Gleitende (zentrierte) Durchschnitte/Mittelwerte

ungerader Ordnung: $p = 2k + 1 \qquad \Rightarrow \qquad k = \frac{p-1}{2}$

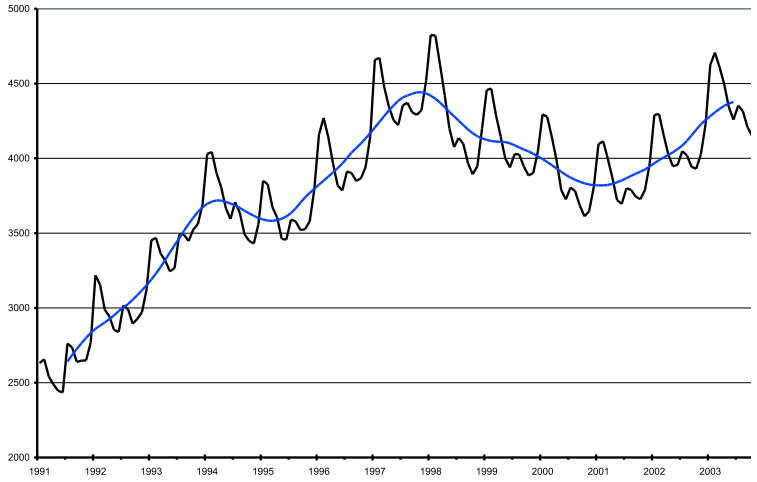
$$\tilde{y}_t = \frac{1}{p} \sum_{h=-k}^{+k} y_{t+h} = \frac{1}{p} \left(y_{t-k} + \cdots + y_t + \cdots + y_{t+k} \right)$$

gerader Ordnung: $p = 2k \qquad \Rightarrow \qquad k = \frac{p}{2}$

$$\tilde{y}_t = \frac{1}{p} \left[\left(\sum_{h=-(k-1)}^{+(k-1)} y_{t+h} \right) + \frac{y_{t-k} + y_{t+k}}{2} \right] = \frac{1}{p} \left(\frac{y_{t-k}}{2} + y_{t-(k-1)} + \cdots + y_t + \cdots + y_{t+(k-1)} + \frac{y_{t+k}}{2} \right)$$

Hier: gleitende Mittelwerte 12. Ordnung, da die Beobachtungen als Monatsdaten vorliegen, d.h. $p = 12 \Rightarrow k = 12/2 = 6$

$$\tilde{y}_t = \frac{1}{12} \left(\frac{y_{t-6}}{2} + y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3} + y_{t+4} + y_{t+5} + \frac{y_{t+6}}{2} \right)$$

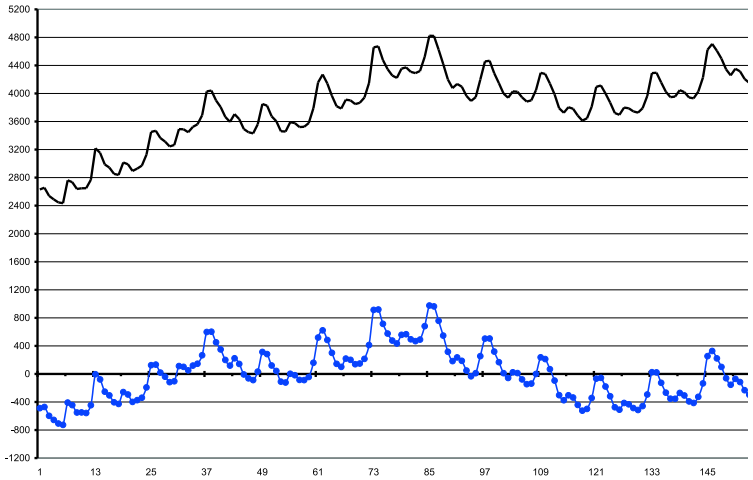


Trendbereinigung

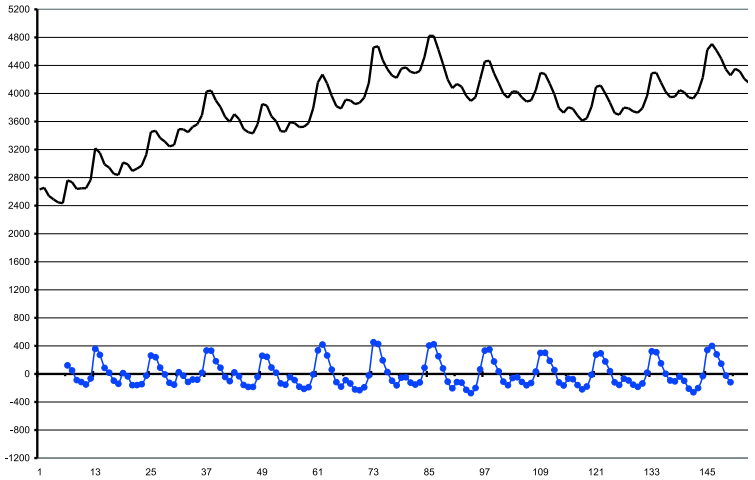
Die tatsächlichen Beobachtungen werden um die geschätzten Werte bzw. um die gleitenden Mittelwerte bereinigt und bewegen sich somit um die Abszisse.

$$d_t = y_t - \hat{y}_t \qquad \text{bzw.} \qquad d_t = y_t - \tilde{y}_t$$

OLS



gleitende Mittelwerte



UNIVERSITÄT DUISBURG ESSEN

WIRTSCHAFTSWISSENSCHAFTEN
BWL VWL RECHTSWISSENSCHAFTEN
INFORMATIK WIRTSCHAFTSINFORMATIK
MEDIZIN-MANAGEMENT

FB 5

Kontakt:	Michael Westermann
Anschrift:	Universität Duisburg–Essen — Campus Essen — Fachbereich Wirtschaftswissenschaften Universitätsstr. 12 45112 Essen
e–Post:	westermann@vwl.uni-essen.de
Homepage:	www.vwl.uni-essen.de/westermann/