Prof. Dr. Hans Peter Litz **Sommersemester 2005**

Statistische Formelsammlung

Teil I

Deskriptive Statistik

Unter Mitarbeit von: Kamatchy Selvarajah

A Voraussetzungen

1 Das Rechnen mit Summenzeichen: Σ

1.1 Definition

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + ... + X_{N-1} + X_N$$

Anmerkungen: X_i = Variable

i = Laufindex über die Objekte $(i = 1...N)$

N = Anzahl der Objekte

1.2 Rechenregeln

$$\sum_{i=1}^{N} c \cdot X_i = c \cdot \sum_{i=1}^{N} X_i$$

$$\sum_{i=1}^{N} (X_i + Y_i) = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i$$

$$\text{aber: } \sum_{i=1}^{k} X_i \cdot f_i = X_1 f_1 + ... + X_k f_k$$

$$\sum_{i=1}^{N} a_i = N \cdot a$$

$$\sum_{i=1}^{N} X_i = \sum_{i=1}^{n} X_i + \sum_{i=n+1}^{N} X_i$$

$$\sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij} = \sum_{i=1}^{N} (X_{i1} + X_{i2} + ... X_{iM})$$

$$= (X_{11} + X_{12} + ... + X_{1M}) + (X_{21} + X_{22} + ... + X_{2M}) + ... + (N_{N1} + X_{N2} + ... + X_{NM})$$

$$\sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij} = \sum_{j=1}^{M} \sum_{i=1}^{N} X_{ij}$$

Anmerkungen: X_i, Y_i = Variablen $(i = 1...N)$

a_i , c = Konstante

f_i = Häufigkeit eines Variablenwertes $(i = 1...k)$

k = Anzahl unterschiedlicher Variablenwerte

X_{ij} = doppelt indizierte Variablen

i = Laufindex über die Objekte $(i = 1...N)$

j = Laufindex über die Variablen $(j = 1...M)$

2 Die Bestimmung des Skalenniveaus

	Skalentyp	**empirische Relation**	**zulässige Transformationen**	**Beispiele**
nicht metrische Skalen	Nominal-Skala	Äquivalenz-beziehungen: gleich – ungleich	eindeutige	Autokennzeichen, Berufe, Familien-stand, Geschlecht
	Ordinal-(Rang-)Skala	zusätzlich: Rangbeziehungen	monotone	Richterskala, MOHSsche - Härte-Skala, Ranglisten, Soziale Schichtung
metrische Skalen	Intervall-Skala	zusätzlich: Differenzen-beziehungen	lineare: $y = a + bx$	Kalenderzeit, Tempe-raturen in Celsius und Fahrenheit
	Verhältnis-Skala	zusätzlich: Proportional-beziehungen, absoluter Nullpunkt	proportionale: $y = bx$	Länge, Gewicht, Zeitintervalle, Währungen
	Absolut-Skala	zusätzlich: natürliche Intervalle	identische: $y = x$	Zählvorgänge bei Gegenständen, Häufigkeiten

B Statistische Maßzahlen eindimensionaler Häufigkeitsverteilungen

1 Vorarbeiten

1.1 Die Konstruktion von statistischen Tabellen

Allgemeine Form einer Häufigkeitstabelle

Merkmal X_i	**abs. Häufigkeit** f_i
x_1	f_1
x_2	f_2
.	.
.	.
.	.
.	.
x_{k-1}	f_{k-1}
x_k	f_k
insgesamt	N

Allgemeine Form einer auf- bzw. ab(wärts)kumulierten Häufigkeitsverteilung

Auf(wärts)kumulation		**Ab(wärts)kumulation**	
bis unter X_i	**kum. Hf.** $f_i^{\uparrow}$	X_i **und mehr**	**kum. Hf.** $f_i^{\downarrow}$
x_1	0	x_1	N
x_2	f_1	x_2	$N - f_1$
x_3	$f_1 + f_2$	x_3	$N - f_1 - f_2$
x_4	$f_1 + f_2 + f_3$	x_4	$N - f_1 - f_2 - f_3$
.	.	.	.
.	.	.	.
.	.	.	.
x_{k-1}	$N - f_k - f_{k-1}$	x_{k-1}	$f_k + f_{k-1}$
x_k	$N - f_k$	x_k	f_k
x_{k+1}	N	x_{k+1}	0

1.2 Die Zeichnung statistischer Graphiken

Bei unterschiedlicher Klassenbreite mit **Häufigkeitsdichten** f_i^d oder mit **modifizierten Häufigkeitsdichten** $\tilde{f}_i$ arbeiten!

$f_i^d = \frac{f_i}{c_i}$ [c_i (Klassenbreite) = X_o (Klassenobergrenze) - X_u (Klassenuntergrenze)]

$\tilde{f}_i = \frac{f_i}{c_i} \cdot \tilde{c}$ [$\tilde{c}$ = Standardklassenbreite]

1.3 Die statistische Arbeitstabelle

Arbeitstabelle zur Ermittlung der Häufigkeitsdichten

Merkmalswerte von ... bis unter ...	abs. Hf. f_i	Klassenmitte m_i	Klassenbreite c_i	modifizierte Häufigkeitsdichte $\tilde{f}_i$
$x_1 - x_2$				
$x_2 - x_3$				
.				
.				
.				
$x_{k-1} - x_k$				

Arbeitstabelle für die Berechnung statistischer Maßzahlen aus einer klassierten Häufigkeitsverteilung

m_i	f_i	$m_i \cdot f_i$	m_i^2	$m_i^2 \cdot f_i$	$\left\|m_i - \overline{X}\right\|$	$\left\|m_i - \overline{X}\right\|^2 \cdot f_i$
.						
.						
.						
.						
Σ	N	$\Sigma\, m_i f_i$	-	$\Sigma m_i^2 \cdot f_i$	-	$\Sigma\left\|m_i - \overline{X}\right\|^2 \cdot f_i$

2. Maßzahlen der zentralen Tendenz (Mittelwerte)

2.1 Verwendbarkeit der Mittelwerte bei gegebenem Skalenniveau

	Mittelwerte		
Skalenniveau	Modus	Median	arith.Mittel
nominal	(x)	-	-
ordinal	x	x	-
metrisch	x	x	x

2.2 Lagetypische Mittelwerte

2.2.1 Der Modus (häufigster Wert): Mod

bei gruppierten Daten:

$$\text{Mod} = X_i \text{ bei } f_i = \max$$

bei klassierten Daten:

feinberechneter Modus mit modifizierten Häufigkeitsdichten:

$$\text{Mod} = L_{\text{Mod}} + c_{\text{Mod}} \cdot \left[\frac{\tilde{f}_{\text{Mod}} - \tilde{f}_{\text{Mod}-1}}{2\tilde{f}_{\text{Mod}} - (\tilde{f}_{\text{Mod}-1} + \tilde{f}_{\text{Mod}+1})} \right]$$

L_{mod} = der untere Klassenrand der modalen Klasse

c_{mod} = die Klassenbreite der modalen Klasse

$\tilde{f}_{Mod}$ = die modifizierte Häufigkeitsdichte der modalen Klasse

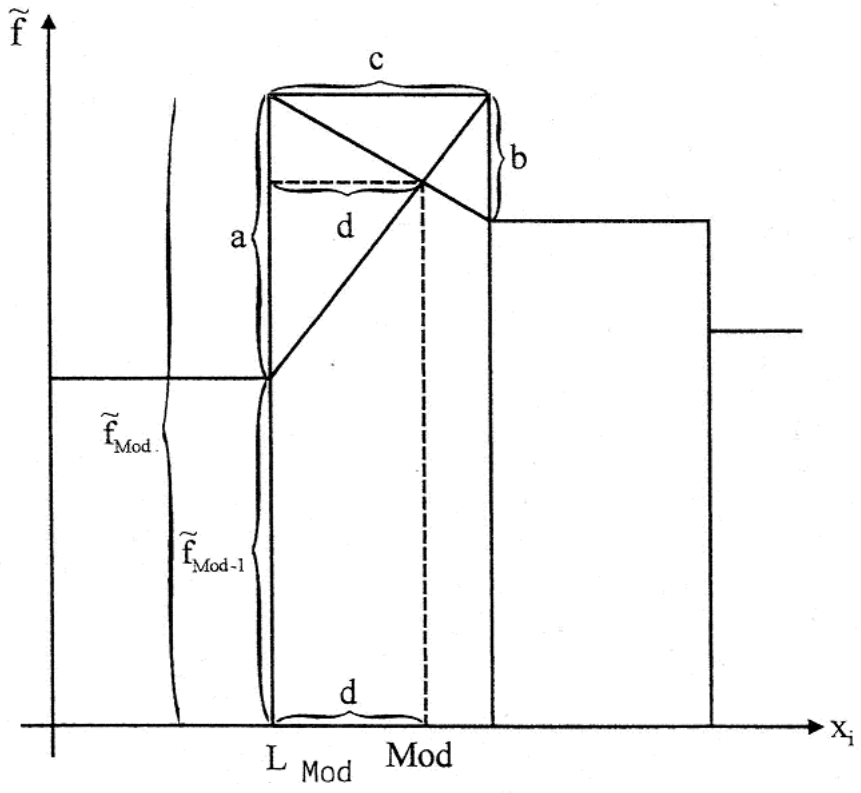
$\tilde{f}_{Mod\ -1}$ = ie modifizierte Häufigkeitsdichte der, der modalen Klasse vorausgehenden Klasse

$\tilde{f}_{Mod\ +1}$ = die modifizierte Häufigkeitsdichte der, der modalen Klasse nachfolgenden Klasse

Anmerkung: - sind nur die Häufigkeitsdichten gegeben, werden die $\tilde{f}_i$ durch f_i^d ersetzt

- sind die Klassen gleich breit, werden die $\tilde{f}_i$ durch f_i ersetzt

Graphische Bestimmung des feinberechneten Modus



$$\frac{d}{a} = \frac{c-d}{b} \quad \text{und} \quad \text{Mod} = L_{\text{Mod}} + d$$

$$d = c \cdot \frac{a}{a+b} \quad \text{mit}$$

$$a = \tilde{f}_{\text{Mod}} - \tilde{f}_{\text{Mod}-1} \quad \text{und}$$

$$b = \tilde{f}_{\text{Mod}} - \tilde{f}_{\text{Mod}+1} \ .$$

2.2.2 Der Median (Zentralwert): Med

bei geordneten Urlistenwerten:

$$X_1 \le X_2 \le ..... \le X_{\frac{N}{2}} \le X_{\frac{N}{2}+1} \le .....X_N$$

- **für gerade N:**

$$X_{\frac{N}{2}} = \text{Wert des } \frac{N}{2}\text{- ten Elementes}$$

$$\text{Med} = \frac{1}{2}(X_{\frac{N}{2}} + X_{\frac{N}{2}+1})$$

- **für ungerade N:**

$$\text{Med} = X_{\frac{N+1}{2}}$$

feinberechneter Median bei klassierten Daten $X_i (i = 1...k)$

$$\text{Med} = L_{\text{Med}} + c_{\text{Med}} \cdot \left(\frac{\frac{N}{2} - (\Sigma f)_{L_{\text{Med}}}}{f_{\text{Med}}} \right)$$

Hinweis: mit aufkumulierten Häufigkeiten arbeiten

L_{Med} = unterer Klassenrand der medianen Klasse (d.h. der Klasse, die den Median enthält)

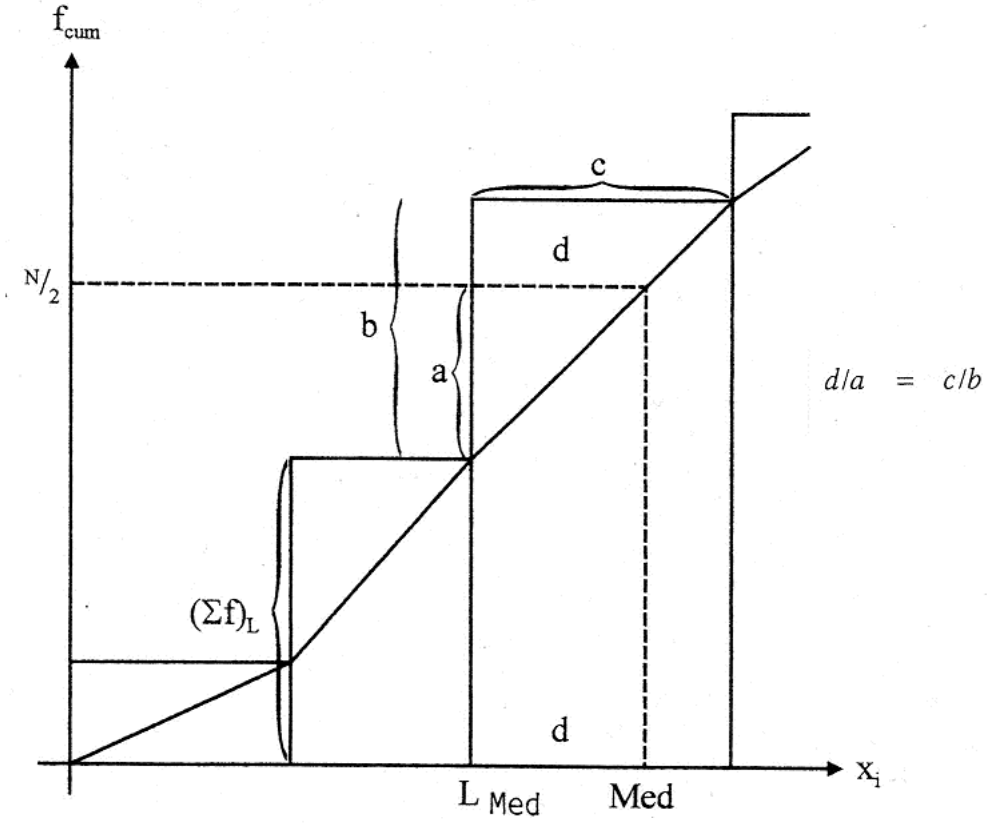
c_{Med} = Klassenbreite der medianen Klasse

N = Anzahl der Fälle

$(\Sigma f)_{L_{\text{Med}}}$ = Summe der Häufigkeiten in allen Klassen, die kleiner als die mediane Klasse sind

f_{Med} = absolute Häufigkeit der medianen Klasse

Graphische Bestimmung des feinberechneten Medians



$$\text{Med} = L_{Med} + c \cdot \frac{a}{b} .$$

$$a = \frac{N}{2} - (\Sigma f)_L \text{ und } b = f_{\text{Med}}$$

2.3 Rechnerische Mittelwerte

2.3.1 Das Arithmetische Mittel: $\overline{X}$

bei Urlistenwerte:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

N = Anzahl der Fälle

Anmerkung: wichtige Eigenschaft von $\overline{X}$: $\Sigma(X_i - \overline{X}) = 0$

bei gruppierten Daten:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{g} X_i \cdot f_i \qquad f_i = \text{Häufigkeit}; \quad N = \sum_{i=1}^{g} f_i$$

bei klassierten Daten $X_i = m_i (i = 1...k)$

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} m_i \cdot f_i$$ m_i = Klassenmitten (i = 1,2,.....,k) , $m_i = \frac{X_o + X_u}{2}$,

X_o = Klassenobergrenze, X_u = Klassenuntergrenze

Anmerkung: Die beiden ersten Berechnungsformeln führen zu identischen und exakten Werten; Das $\overline{X}$ aus klassierten Werten weicht i.d.R. davon ab. Es ist der ungenauere Wert und deshalb nur zu verwenden, wenn die Urlistenwerte oder die gruppierten Daten nicht verfügbar sind.

2.3.2 Das Harmonische Mittel: $\overline{X}_h$

$$\overline{X}_h = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + ... \frac{1}{X_N}}$$

2.3.3 Das Quadratische Mittel: $\overline{X}_q$

$$\overline{X}_q = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2}$$

bei gruppierten Daten:

$$\overline{X}_q = \sqrt{\frac{1}{N}\sum_{i=1}^{g} X_i^2 \cdot f_i}$$

bei klassierten Daten:

$$\overline{X}_q = \sqrt{\frac{1}{N}\sum_{i=1}^{k} m_i^2 \cdot f_i}$$

2.3.4 Das Geometrische Mittel: $\overline{X}_g$

$\overline{X}_g = \sqrt[N]{X_1.X_2.....X_N}$ → in logarithmischer Schreibweise: $\log \overline{X}_g = \frac{1}{N}\sum_{i=1}^{N} \log X_i$

bei gruppierten Daten:

$\overline{X}_g = \sqrt[N]{X_1^{f_1}.X_2^{f_2}.....X_g^{f_g}}$ → in logarithmischer Schreibweise: $\log \overline{X}_g = \frac{1}{N}\sum_{i=1}^{g} f_i \log X_i$

Berechnung der durchschnittlichen Zuwächse:

Der Wachstumskoeffizienten (Wachstumsfaktor) y_i ist das Verhältnis aufeinander folgender Jahreswerte Y_i:

$$y_i = \frac{Y_i}{Y_{i-1}} = 1 + w_1$$

Die Wachstumsrate $w_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}}$

Der durchschnittliche Wachstumskoeffizient (Wachstumsfaktor): $\overline{Y}$ $(\equiv \overline{X}_g)$

$$\overline{y} = \sqrt[N]{y_1.y_2.....y_N} = \sqrt[N]{Y_N / Y_0}$$

Die durchschnittliche Wachstumsrate: $\overline{w}$

$$\overline{w} = \overline{y} - 1 = \sqrt[N]{Y_N / Y0} - 1$$ vgl.:

Zinseszinsformel: $Y_N = (1 + \overline{w})^N \cdot Y_o$

3 Maßzahlen der Streuung (Dispersionsparameter)

3.1 Lagetypische Streuungsmaße

3.1.1 Die Spannweite: S

bei gruppierten Daten:

$$S = X_g - X_1$$

3.1.2 Der (Inter-) Quartilsabstand: QA

$$QA = Q_{III} - Q_I$$

Q_I = erstes Quartil , Q_{III} = drittes Quartil

Anmerkung: In diesem Bereich liegen die mittleren 50 % der Werte

Bestimmung der Quartile bei geordneten Urlistenwerten $X_i (i = 1...N)$

Segmentierung der Verteilung in 4 gleich umfangreiche Blöcke

Das 1. Quartil	Das 2. Quartil	Das 3. Quartil
$X_{\frac{N}{4}} \le Q_I < X_{\frac{N}{4}+1}$	$X_{\frac{N}{2}} \le Q_{II} < X_{\frac{N}{2}+1}$	$X_{\frac{3N}{4}} \le Q_{III} < X_{\frac{3N}{4}+1}$

Anmerkung: Diese Indizierung gilt bei durch 4 teilbarem N. Ist N nicht durch 4 teilbar, ergeben sich die Indexwerte aus den abgerundeten Quotienten im Index.

Feinberechnung der Quartile bei klassierten Daten:

$$Q_I = L_{Q_I} + c_{Q_I} \cdot \left(\frac{\frac{N}{4} - (\Sigma f)_{L_{Q_I}}}{f_{Q_I}} \right) \qquad Med = Q_{II} = L_{Med} + c_{Med} \cdot \left(\frac{\frac{N}{2} - (\Sigma f)_{L_{Med}}}{f_{Med}} \right)$$

$$Q_{III} = L_{Q_{III}} + c_{Q_{III}} \cdot \left(\frac{\frac{3N}{4} - (\Sigma f)_{L_{Q_{III}}}}{f_{Q_{III}}} \right)$$

Anmerkung: (Symbole analog 2.2.2)

3.1.3 Der Semiquartilsabstand: (SQA)

$$SQA = \frac{Q_{III} - Q_I}{2}$$

Anmerkung: SQA drückt den durchschnittlichen Abstand des Medians von den Quartilen aus.

3.1.4 Die Kelly- Range

Segmentierung der Verteilung in 10 gleich umfangreiche Blöcke analog 3.1.2

$\rightarrow$ Dezentile $D_1 ... D_9$

$$\text{Kelly} - \text{Range} = D_9 - D_1$$

Anmerkung: In diesem Bereich liegen die mittleren 80 % der Werte

3.2 Rechnerische Streuungsmaße

3.2.1 Die Mittlere Absolute Abweichung: MA

bei Urlistenwerten $X_i (i = 1...N)$

$$MA = \frac{1}{N} \sum_{i=1}^{N} \left| X_i - \overline{X} \right|$$

Anmerkung: ($\overline{X}$ vgl. 2.3.1)

bei gruppierten Daten:

$$MA = \frac{1}{N} \sum_{i=1}^{g} \left| X_i - \overline{X} \right| \cdot f_i$$

bei klassierten Daten:

$$MA = \frac{1}{N} \sum_{i=1}^{k} \left| m_i - \overline{X} \right| \cdot f_i$$

Anmerkung: m_i vgl. 2.3.1

3.2.2 Die Mittlere Quadratische Abweichung (Varianz); $VAR(X)$

bei Urlistenwerten:

$$VAR(X) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$$

Rechenformel : $VAR(X) = \frac{1}{N}\sum_{i=1}^{N} X_i^2 \; - \overline{X}^2$

bei gruppierten Daten:

$$VAR(X) = \frac{1}{N}\sum_{i=1}^{g}\left(X_i - \overline{X}\right)^2 f_i$$

Rechenformel : $VAR(X) = \frac{1}{N}\sum_{i=1}^{g} X_i^2 \cdot f_i \; - \overline{X}^2$

bei klassierten Daten:

$$VAR(X) = \frac{1}{N}\sum_{i=1}^{k}\left(m_i - \overline{X}\right)^2 f_i$$

Rechenformel : $VAR(X) = \frac{1}{N}\sum_{i=1}^{k} m_i^2 \cdot f_i \; - \overline{X}^2$

3.2.3 Die Standardabweichung: s_X

$$s_X = \sqrt{VAR(X)}$$

4 Relative Streuung und Schiefe der Verteilung

4.1 Die Relative Streuung / der Variationskoeffizient: V

$$V = \frac{s}{\overline{X}}$$

Anmerkung: zum Vergleich von Streuungen unterschiedlicher Verteilungen in zeitlicher oder regionaler Hinsicht; oft in Prozent. Drückt den Betrag der Standardabweichung in Prozent des Betrags des Mittelwertes aus.

4.2 Schiefemaße

4.2.1 Graphische Darstellung symmetrischer und schiefer Verteilungen

symmetrische Verteilung	rechtsschiefe Verteilung	linksschiefe Verteilung
$\overline{X}$=Med=Mod	$Mod < Med < \overline{X}$	$\overline{X} < Med < Mod$

Anmerkung: die Verteilungen weisen gleiche arithmetischen Mittel und gleiche Standardabweichungen auf, sind aber von unterschiedlicher Gestalt.

4.2.2 Das Schiefemaß nach Pearson: PSM

$$PSM = \frac{\overline{X} - Mod}{s}$$

Anmerkung: PSM = 0 → Verteilung: symmetrisch
PSM > 0 → Verteilung: rechtsschief
PSM < 0 → Verteilung: linksschief

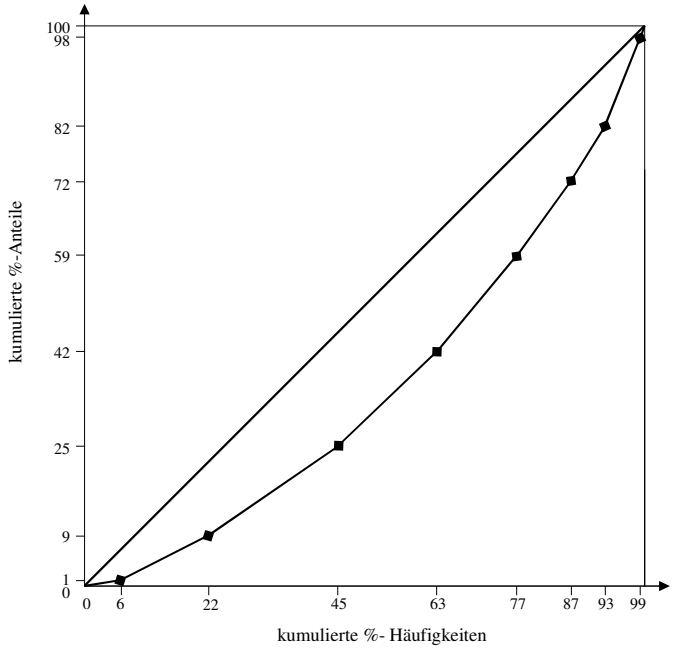
5 Konzentrationsmaße

5.1 Relative Konzentration

geordnete X_i bei aufsteigenden Merkmalsausprägungen $X_1 \leq X_2 \leq \ldots X_N$
(bzw. X_g oder)

5.1.1 Die Lorenzkurve

Abb. 6.2: Lorenz-Kurve der Einkommenskonzentration



Anmerkung: Die Lorenzkurve beschreibt, welcher Anteil des Merkmalsgesamtbetrages $(h_i^{\%\uparrow})$ auf einen vorgegebenen Anteil $(f_i^{\%\uparrow})$ der, der Größe nach geordneten Merkmalsträger entfällt. Je größer die Fläche K desto größer die Konzentration. Bei Gleichverteilung fällt die Kurve mit der Diagonalen zusammen.

Arbeitstabelle zur Lorenzkurve und zum Gini-Koeffizienten für klassierte Daten

Klasse	f_i	$f_i^{\%}$	$m_i f_i$	$h_i^{\%} = \frac{m_i f_i}{\Sigma m_i f_i} 100$	$h_i^{\%\uparrow}$	$f_i^{\%\uparrow}$	$(h_i^{\%\uparrow} + h_{i-1}^{\%\uparrow}) f_i^{\%}$
1	2	3	4	5	6	7	8
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Σ							$\Sigma \rightarrow$ Gini-Koeff.

Anmerkungen zum Berechnen und Zeichnen der Lorenzkurve:

Bei gruppierten Daten werden für Spalte 4 die Summen der Merkmalsbeträge der einzelnen Klassen herangezogen.

$h_i^{\%\uparrow}$ = Ordinatenwerte der Lorenzkurve; $f_i^{\%\uparrow}$ = Abszissenwerte der Lorenzkurve

5.1.2 Der Gini-Koeffizient: G

$$G = 1 - \sum_{i=1}^{k} f_i' \left(h_i'^{\uparrow} + h_{i-1}'^{\uparrow} \right) = 1 - \frac{\sum f_i^{\%} (h_i^{\%\uparrow} + h_{i-1}^{\%\uparrow})}{10000}$$

Anmerkung: Da die Summe aus Spalte 8 der Arbeitstabelle auf den Prozentwerten der Spalten 6 und 7 beruht, muss sie 2 mal um den Faktor 100 gekürzt werden (f' bzw. h' sind die relativen, $f^{\%}$ bzw., $h^{\%}$ die prozentualen Häufigkeiten).

5.2 Absolute Konzentration

geordnete X_i bei abfallenden Merkmalsausprägungen $X_1 \geq X_2 \geq X_3 \ ... \geq X_N$

5.2.1 Die Konzentrationsrate: C

für die r größten Merkmalsträger des geordneten Merkmals $X_i \ (i = 1...N)$

mit $X_1 \geq X_2 \geq X_3 \geq ... \geq X_r \geq ... \geq X_N$

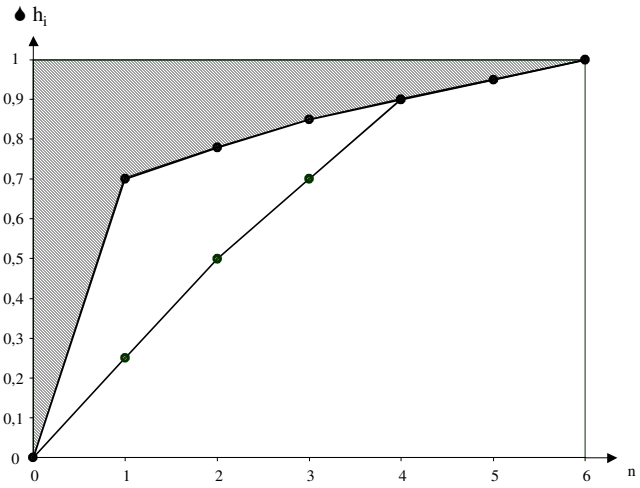
$C = \sum_{i=1}^{r} h_i$, mit:

$h_i =$ Merkmalsanteil des Objekts:

$$h_i = \frac{X_i}{\sum_{i=1}^{N} X_i}$$

5.2.2 Die Konzentrationskurve

Abb. 6.4: Konzentrationskurve



5.2.3 Der Rosenbluth- Index: C_R

$$C_R = \frac{1}{2\sum_{i=1}^{N} i \cdot h_i - 1}$$

Eigenschaft: $\frac{1}{N} \leq C_R \leq 1$, $C_R = 1$ bei maximaler Konzentration; $C_R = \frac{1}{N}$ bei Gleichverteilung

5.2.4 Der Herfindahl-Index: C_H

$$C_H = \sum_{i=1}^{N} h_i^2$$

Eigenschaft, $C_H = 1$ bei maximaler Konzentration; $C_H = \frac{1}{N}$ bei Gleichverteilung

Arbeitstabelle zur Berechnung von C_R und C_H

Objekt	X_i	h_i	Rang (i)	$i \cdot h_i$	h_i^2
. . .					
Σ			-		

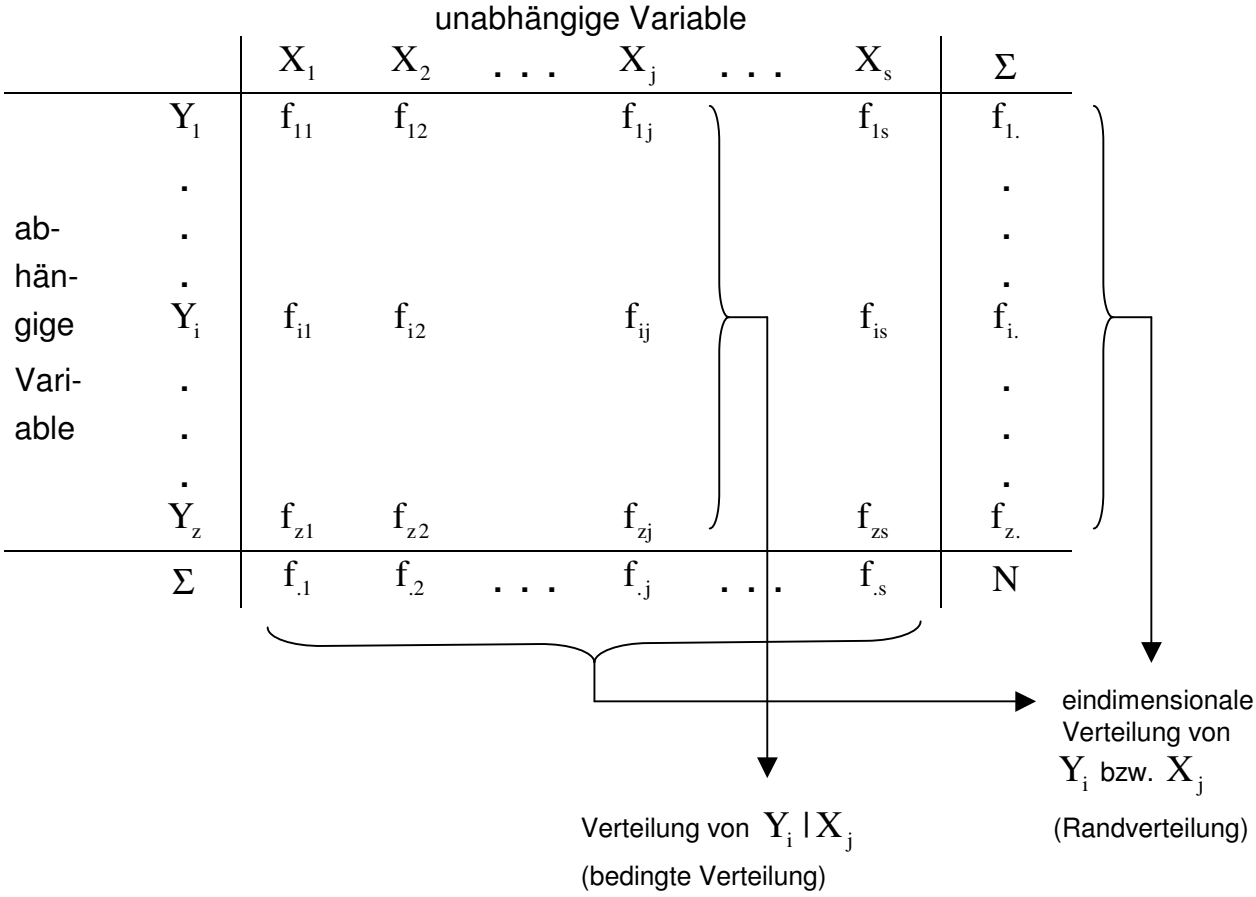
Anmerkung: Wenn $\overline{X}$ und s (Standardabweichung) gegeben, lässt sich C_H über den Variationskoeffizienten V mit der Formel

$C_H = \frac{V^2 + 1}{N}$ berechnen.

C Statistische Maßzahlen zweidimensionaler Häufigkeitsverteilungen

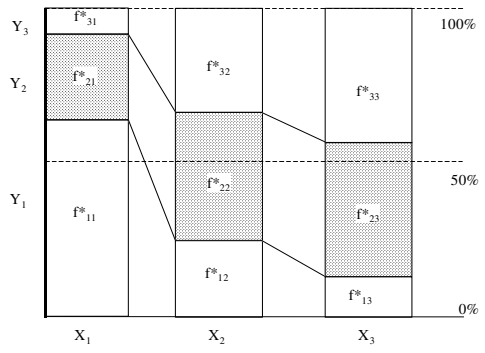
5 Vorarbeiten

6.1 Formaler Aufbau einer zweidimensionalen Häufigkeitstabelle



6.2 Graphische Darstellung bedingter relativer/prozentualer Häufigkeiten

Abb. 7.1: Graphische Darstellung bedingter relativer Häufigkeiten



7 Sätze zur Statistischen Unabhängigkeit

Satz 1: zwei Variablen $Y_i \ (i=1...2)$ und $X_j \ (j=1...s)$ sind statistisch unabhängig, wenn:

$$f'(Y_i \mid X_1) = f'(Y_i \mid X_2) = ... f'(Y_i \mid X_s)$$

$$\frac{f_{i1}}{f_{\cdot 1}} = \frac{f_{i2}}{f_{\cdot 2}} = \frac{f_{is}}{f_{\cdot s}}$$

Satz 2: Zwei Variablen Y_i und X_i sind statistisch unabhängig, wenn:

$$f'(Y_i \mid X_j) = f'(Y_i)$$

$$\frac{f_{ij}}{f_{\cdot j}} = \frac{f_{i\cdot}}{N}$$

Satz 3: Die relative Häufigkeit des gemeinsamen Auftretens zweier Variablen Y_i und X_j ist:

$$f'(Y_i \ , X_j) = f'(Y_i \mid X_j) \cdot f'(X_j)$$

$$\frac{f_{ij}}{N} = \frac{f_{ij}}{f_{\cdot j}} \cdot \frac{f_{\cdot j}}{N}$$

Satz 4: Die relative Häufigkeit des gemeinsamen Auftretens zweier Variablen Y_i und X_j bei Unabhängigkeit ist:

$$f'(Y_i \ , X_j) = f'(Y_i) \cdot f'(X_j)$$

$$\frac{f_{ij}}{N} = \frac{f_{i\cdot}}{N} \cdot \frac{f_{\cdot j}}{N}$$

unter der obigen Bedingung gilt deshalb:

$$f_{ij} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{N}$$

8 Zusammenhangsmaße für nominalskalierte Daten

8.1 Maßzahlen auf Basis von Chi-Quadrat (χ^2)

8.1.1 Definition von χ^2

$$\chi^2 = \sum_{b,e=1}^{z\cdot s} \frac{(f_b - f_e)^2}{f_e}$$

Anmerkungen:

z = Anzahl der Zeilen

s = Anzahl der Spalten

f_b = absolute Häufigkeiten der Kontingenztabelle $(b = 1...z \cdot s)$

f_e = absolute Häufigkeiten der Indifferenztabelle $(e = 1...z \cdot s)$

8.1.2 Häufigkeiten der Indifferenztabelle

Eine beliebige Zelle f_{ij} in der Indifferenztabelle wird wie folgt ermittelt:

$f_{ij} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{N}$ N = Anzahl der Fälle; i : Zeilenindex; j : Spaltenindex

Hinweis: - χ^2 niemals auf der Basis von prozentualen Häufigkeiten berechnen!

- Die Anzahl der Freiheitsgrade (FG) einer Indifferenztabelle beträgt: $FG = (z-1)\cdot(s-1)$
- χ^2 variiert direkt mit N, daher ist es sinnvoll eine der folgenden, normierten Maßzahlen zu verwenden.

8.1.3 Verkürztes Rechenverfahren für eine 2 x 2 Felder-Tabelle

	X_1	X_2	Σ
Y_1	a	b	a+b
Y_2	c	d	c+d
Σ	a+c	b+d	N

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

8.1.4 Der Phi-Koeffizient (φ)

$$\varphi = \sqrt{\frac{\chi^2}{N}}$$

Anmerkungen: 1) $0 \le \varphi = \le 1$ für 2 x 2 Tabelle

2) u.U. $1 < \varphi$ für größere Tabelle

8.1.5 Cramers V

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(z-1, s-1)}}$$

Anmerkung: $0 \le V \le 1$

8.1.6 Der Kontingenzkoeffizient von Pearson (C)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Anmerkung: $0 \le C \le C_{max} \le 1$

$C_{max} = \sqrt{\frac{z-1}{z}}$ für quadratische Tabellen

$C_{max} = \frac{1}{2}\left(\sqrt{\frac{z-1}{z}} + \sqrt{\frac{s-1}{s}}\right)$ für rechteckige Tabellen

Hinweis: für kleinere Tabellen $C_{max} < 1$

In diesem Fall empfiehlt sich die Verwendung des **korrigierten Kontingenzkoeffizienten**:

$$C_{korr} = \frac{C}{C_{max}}$$

Anmerkung: $0 \le C_{korr} \le 1$

8.2 Die Maße der prädiktiven Assoziation von Goodman und Kruskal (PRE- Maße „Lambda“)

Anmerkung: PRE = proportional reduction of error

8.2.1 λ_y - Asymmetrisch (y = abhängige Variable)

$$\lambda_{y=f(x)} = \frac{F_1 - F_2}{F_1}$$

$$0 \le \lambda_{y=f(x)} \le 1$$

$$F_1 = N - \max(f_{i\cdot}) \;;\; F_2 = \sum_{j=1}^{s}\left[f_{\cdot j} - \max(f_{ij})\right]$$

8.2.2 λ_x - Asymmetrisch (x = abhängige Variable)

$$\lambda_{x=f(y)} = \frac{F_1^* - F_2^*}{F_1^*}$$

$$F_1^* = N - \max(f_{\cdot j}) \;;\; F_2^* = \sum_{i=1}^{z}\left[f_{i\cdot} - \max(f_{ij})\right]$$

8.2.3 λ - Symmetrisch x ↔ y

$$\lambda_{sym} = \frac{F_1 + F_1^* - \left(F_2 + F_2^*\right)}{F_1 + F_1^*}$$

9 Zusammenhangsmaße für ordinal skalierte Daten

9.1 Konkordenzmaße (Maße des Paarvergleichs)

	X_1	X_2	X_3	X_4	X_5	X_6	Σ
Y_1				Ties			
		k P		in	d P		
Y_2				X			
Y_3	Ties in Y			Ties in X u. Y	Ties in Y		
Y_4				Ties			
		d P		in	k P		
Y_5				X			
Σ							

9.1.1 Kendalls Tau-a: τ_a

$$\tau_a = \frac{N_k - N_d}{N_p}$$

Anmerkung:

N_p = Anzahl aller möglichen Paare; $N_p = \frac{N(N-1)}{2}$

N_k = Anzahl der konkordanten Paare

N_d = Anzahl der diskordanten Paare

Hinweis: geeignet für Daten, die keine Ties enthalten, nur dann gilt der Wertebereich
$-1 \le \tau_a \le +1$

9.1.2 Kendalls Tau-b: τ_b

$$\tau_b = \frac{N_k - N_d}{\sqrt{(N_k + N_d + T_x)}\sqrt{(N_k + N_d + T_y)}}$$

Anmerkung: T_X = Anzahl der Ties in X
T_Y = Anzahl der Ties in Y

Hinweis:
- $1 \le \tau_b \le +1$
- nur bei quadratischen Tabellen max 1
- symmetrisch

9.1.3 Kendalls Tau-c: τ_c

$$\tau_c = \frac{N_k - N_d}{\frac{1}{2}N^2\left(\frac{m-1}{m}\right)}$$

Anmerkung: m = Minimum von z und s;

Hinweis:
- $-1 \le \tau_c \le +1$
- symmetrisch
- für rechteckigeTabellen:

9.1.4 Somers d - asymmetrisch $[y = f(x)]$

$$d_{y=f(x)} = \frac{N_k - N_d}{N_k + N_d + T_y}$$

Hinweis:
- Y = Zeilenvariable, X = Spaltenvariable
- für Tabellen beliebiger Größe
- $-1 \le d \le +1$

9.1.5 Somers d - asymmetrisch $[x = f(y)]$

$$d_{x=f(y)} = \frac{N_k - N_d}{N_k + N_d + T_x}$$

Hinweis:
- X = Zeilenvariable, Y = Spaltenvariable
- für Tabellen beliebiger Größe
- $-1 \le d \le +1$

9.1.6 Somers d -symmetrisch

$$d_{sym} = \frac{N_k - N_d}{N_k + N_d + \frac{1}{2}(T_y + T_x)}$$

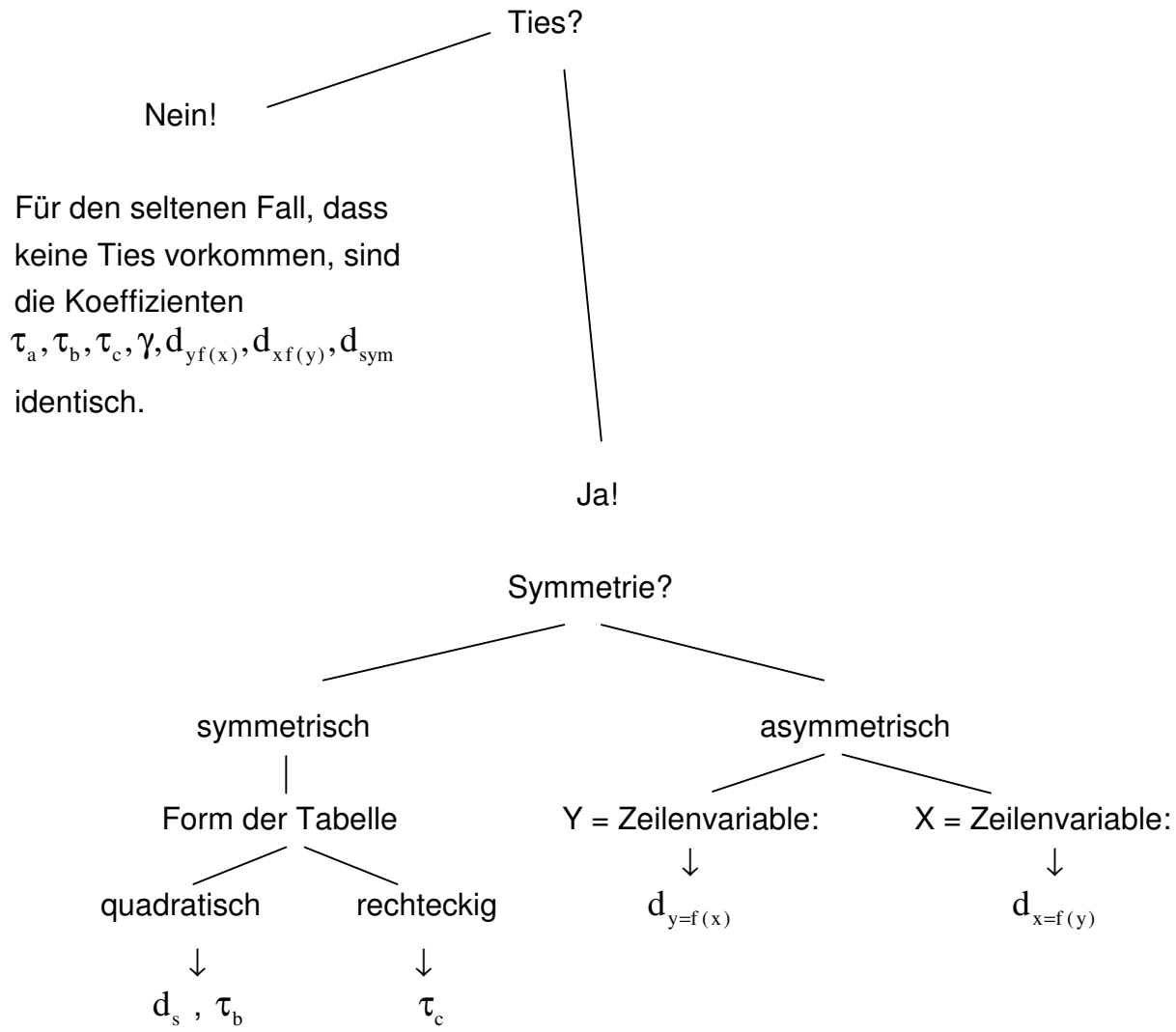
Hinweis:
- symmetrische Beziehung
- für Tabellen beliebiger Größe
- $-1 \le d_{sgm} \le +1$

9.1.7 Gamma von Goodman und Kruskal (γ)

$$\gamma = \frac{N_k - N_d}{N_k + N_d}$$

Hinweis:
- für Tabellen beliebiger Größe
- ignoriert Ties, auch wenn diese gegen einen Zusammenhang sprechen.
- $-1 \le \gamma \le +1$

9.1.8 Entscheidungsbaum zur Anwendung der Formeln für ordinal skalierte Daten (Paarvergleich)



Hinweis: allg. gilt $\tau_a \leq \tau_b ... d_{sym} \leq \gamma$

9.2 Der Spearmansche Rangkorrelationskoeffizient: r_s

$$r_s = 1 - \frac{6\sum d_i^2}{N\left(N^2 - 1\right)}$$

Anmerkung: N - Anzahl der rangplazierten Untersuchungseinheiten,
d_i - Differenz zwischen den Rangplätzen, die die i-te Untersuchungseinheit bezüglich der Variablen X und Y aufweist $(X_i - Y_i)$,
Σd_i^2 - Summe der quadrierten Rangplatzdifferenzen $\left(\Sigma(X_i - Y_i)^2\right)$

Hinweis: - Voraussetzung: Nach zwei Merkmalen rangplatzierte Untersuchungseinheiten

- Vorsicht: - r_s produziert beim Vorliegen von Ties überhöhte Werte und überschätzt deshalb die Stärke des Zusammenhang
- Differenzbildung von Rangplätzen ist eigentlich nicht zulässig
- Eigenschaft: $-1 \le r_s \le +1$

10 Zusammenhangsmaße für metrisch skalierte Daten

10.1 Das einfache lineare Regressionsmodell

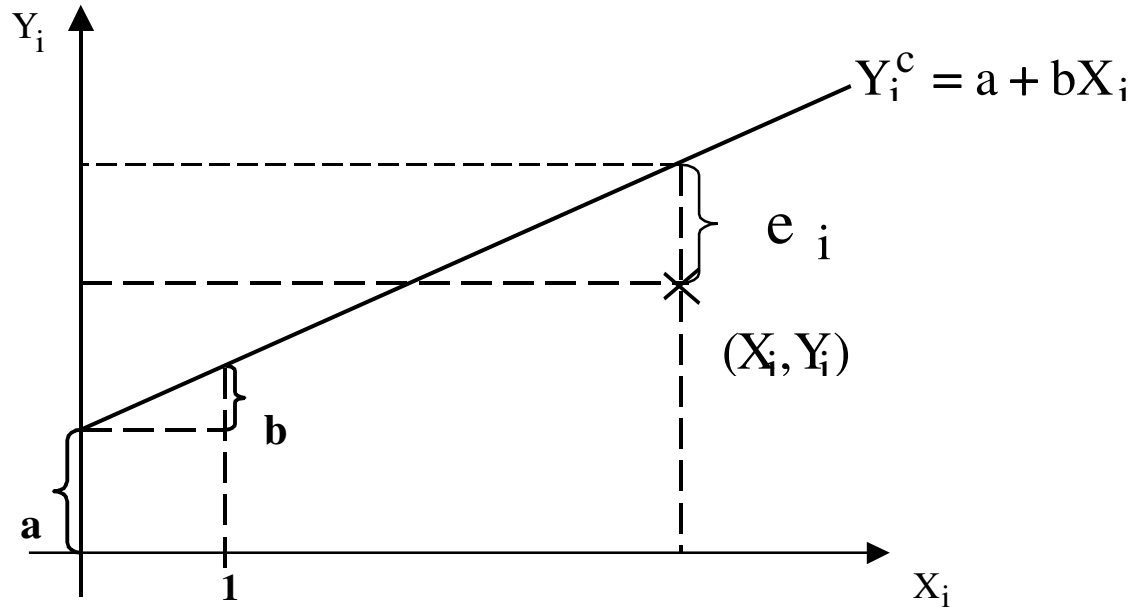
10.1.1 Die lineare Regressionsfunktion für die Wertepaare (Y_i , X_i)

$$Y_i = Y_i^c + e_i$$
$$Y_i^c = a + b\, X_i$$

Erläuterungen:

Y_i = beobachtete Werte der abhängigen Variablen
Y_i^c = Funktionswerte der abhängigen Variablen
X_i = beobachtete Werte der unabhängigen Variablen
e_i = $Y_i - Y_i^c$ = Residuen / Fehler (nicht aus dem Regressionsmodell erklärbare Anteile der beobachteten, abhängigen Variablen

10.1.2 Graphische Darstellung der Punktwolke und der Regressionsfunktion



Hinweis:
- a: Ordinatenachsenabschnitt
- b: Steigung der Gerade, die angibt, um wie viel Einheiten Y wächst, wenn X um eine Einheit wächst.

10.1.3 Die Methode der kleinsten Quadrate zur Bestimmung der Regressionsparameter

$$\Sigma e_i^2 = \Sigma\left(Y_i - Y_i^c\right)^2 = \Sigma\left(Y_i - a - b\,X_i\right)^2 = \min !$$

Es gilt: $\bar{e} = \frac{1}{N}\Sigma e_i = 0 \;;\; \mathrm{Var}(e) = \Sigma e_i^2 = \min !$

Aus $\Sigma e_i^2 = \min!$ resultieren die beiden Normalgleichungen und die Regressionsparameter nach 10.1.4:

$$1.\,\mathrm{NG}: -\Sigma Y_i + Na + b\Sigma X_i = 0$$
$$2.\,\mathrm{NG}: -\Sigma Y_i X_i + a\Sigma X_i + b\Sigma X_i^2 = 0$$

10.1.4 Die Parameter der Regressionsgeraden: $Y_i^c = a + b\,X_i$

Die Regessionskonstante: $a = \overline{Y} - b\overline{X}$

Der Regressionskoeffizient: $$b = \frac{\frac{1}{N}\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

Als Rechenformel: $$b = \frac{\frac{1}{N}\sum X_i Y_i - \overline{X} \cdot \overline{Y}}{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \overline{X}^2}$$

Bei gruppierten Daten: $$b = \frac{\frac{1}{N}\sum X_i Y_i f_i - \overline{X} \cdot \overline{Y}}{\frac{1}{N}\sum X_i^2 f_i - \overline{X}^2}$$

10.1.5 Arbeitstabelle zur Regression und Korrelation für gruppierte Wertepaare

i	Y_i	X_i	f_i	$Y_i X_i \cdot f_i$	$X_i^2 \cdot f_i$	$Y_i^2 \cdot f_i$
1 . . g						
Σ						

Hinweis: für Einzelwerte gilt $f_i = 1$

10.2 Das einfache lineare Korrelationsmodell

10.2.1 Die Kovarianz zweier Merkmale X_i und: $COV(X, Y)$

$$COV\ (X_i\ ,\ Y_i) = \frac{1}{N}\Sigma(X_i - \overline{X})(Y_i - \overline{Y})$$

10.2.2 Der Korrelationskoeffizient nach Bravais-Pearson r

(Maßzahl für die Stärke des Zusammenhangs)

Hinweis: $-1 \le r \le +1$

Als Rechenformel:
$$r = \frac{\frac{1}{N}\sum X_i Y_i - \overline{X}\cdot\overline{Y}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \overline{X}^2}\cdot\sqrt{\frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \overline{Y}^2}}$$

Bei gruppierten Daten:
$$r = \frac{\frac{1}{N}\sum X_i Y_i f_i - \overline{X}\cdot\overline{Y}}{\sqrt{\frac{1}{N}\sum X_i^2 f_i - \overline{Y}^2}\cdot\sqrt{\frac{1}{N}\sum_{i=1}^{N} Y_i^2 f_i - \overline{Y}^2}}$$

10.2.3 Beziehung zwischen Regressionskoeffizient und Korrelationskoeffizient

$$r = b\cdot\frac{s_X}{s_Y}$$

10.2.4 Der Determinationskoeffizient: r^2

$$\text{Def}: r^2 = \frac{\text{erklärte Varianz}}{\text{Gesamt varianz}} = \frac{\sum\left(Y_i^C - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

Hinweis: - Berechnung über b nach 10.1.4: $r^2 = b^2\left(\frac{Var(X)}{Var(Y)}\right)$

- Berechnung über r nach 10.2.1: $r^2 = r\cdot r$

10.2.5 Die Varianzzerlegung:

$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \frac{1}{N}\sum(Y_i^c - \overline{Y})^2 + \frac{1}{N}\sum(Y_i - Y_i^c)^2$$

Gesamt varianz von Y = erklärte Varianz + nichterklärte Varianz (Fehler Varianz)

- daraus folgt die Eigenschaft: $0 \le r^2 \le 1$

11 Indexzahlen

11.1 Die Messziffer: MZ

Preismesszahl $MZ_{p,t} = \frac{p_t}{p_0} \cdot 100$

Mengenmesszahl $MZ_{q,t} = \frac{q_t}{q_0} \cdot 100$

Erläuterungen:
p_{oi} = Preis des Gutes i im Basisjahr
p_{ti} = Preis des Gutes i im laufenden Jahr $t\ (t = 1...T)$
q_{oi} = Menge des Gutes i im Basisjahr
q_{ti} = Menge des Gutes i im laufenden Jahr $t\ (1 = 1...T)$

Hinweis: Angaben über 100 verweisen im Vergleich zur Basis auf eine Steigerung
Angaben unter 100 in Vergleich zur Basis auf eine Abnahme

11.2 Der Volumenindex: V

$$V = \frac{\sum_{i=1}^{N} q_{ti} \cdot p_{ti}}{\sum_{i=1}^{N} q_{0i} \cdot p_{0i}} \cdot 100 \quad (i = 1.....N)$$

Erläuterungen: vgl. 11.1
Hinweis: V erlaubt keinen Aufschluss über die spezifische Preis- und Mengenentwicklung!

11.3 Preisindizes nach Laspeyres und Paasche

(Erläuterungen vgl. 11.1)

11.3.1 Der Preisindex nach Laspeyres: P_L

$$P_L = \frac{\sum_{i=1}^{N} p_{ti} \cdot q_{0i}}{\sum_{i=1}^{N} p_{0i} \cdot q_{0i}} \cdot 100$$

11.3.2 Der Preisindex nach Paasche: P_P

$$P_P = \frac{\sum_{i=1}^{N} p_{ti} \cdot q_{ti}}{\sum_{i=1}^{N} p_{0i} \cdot q_{ti}} \cdot 100$$

11.4 Mengenindizes nach Laspeyres und Paasche

(Erläuterungen vgl. 11.1)

11.4.1 Der Mengenindex nach Laspeyres: Q_L

$$Q_L = \frac{\sum_{i=1}^{N} q_{ti} \cdot p_{0i}}{\sum_{i=1}^{N} q_{0i} \cdot p_{0i}} \cdot 100$$

11.4.2 Der Mengenindex nach Paasche: Q_P

$$Q_P = \frac{\sum_{i=1}^{N} q_{ti} \cdot p_{ti}}{\sum_{i=1}^{N} q_{0i} \cdot p_{ti}} \cdot 100$$

11.5 Zusammenhänge zwischen Volumen-, Preis- und Mengenindizes

$$V = \frac{P_L \cdot Q_P}{100} = \frac{P_P \cdot Q_L}{100}$$

11.6 Die Umbasierung bestehender Indizes

$$\text{Umbasierte} - \text{Indexzahl} = \frac{\text{jeweiliger} - \text{Indexwert}}{\text{Wert der Reihe in der Periode der neuen Basis}} \cdot 100$$

$$I_t^a = \frac{I_t^o}{I_a^o}$$

Erläuterungen: I_t^o $o \leftarrow$ Basisjahr o bzw. a
$t \leftarrow$ lfd. Jahr t bzw. a

Hinweis: Zweck der Umbasierung ist der Vergleich von Indizes verschiedener Basisperioden. Zum Vergleich wird einer der beiden Indizes auf die Basis des anderen umgerechnet.

11.7 Die Verkettung von Indizes unterschiedlicher Basisjahre

$$\text{Index} - \text{A} \oplus \text{B} = \frac{\text{jeweiliger Wert des Index A} \otimes \text{periodengleicher Wert des Index B}}{\text{periodengleicher Wert des Index A}}$$

$$I_{a+j}^o = \frac{I_{a+j}^a}{I_a^a} \cdot I_a^o$$

Erläuterungen: vgl. 11.6. Zu verketten sind die beiden Indizes I^o und. Der Index I^o wird auf der Basis des gemeinsamen Jahres a für die Jahre $a + j$ fortgeschrieben.

Hinweis: Zweck der Verkettung ist es, einen alten Index, der in dem Berichtsjahr nicht mehr berechnet wird, mit einem neuen Index zu verknüpfen, um gewisse Aufschlüsse über die laufende Entwicklung zu bekommen. Es sollten dabei zwei Voraussetzungen erfüllt sein:

1. Die Reihen sollen etwa den gleichen Inhalt haben.
2. Die Verkettung kann nur vorgenommen werden, wenn beide Reihen wenigstens für eine gemeinsame Periode verfügbare Werte aufweisen.

11.8 Kaufkraftparitäten und Terms of Trades

11.8.1 Laspeyres – Kaufkraftparität

$$K_{LS} = \frac{\sum_{i=1}^{N} p_{si} q_{0i}}{\sum_{i=1}^{N} p_{0i} q_{0i}} \cdot 100$$

11.8.2 Terms of Trade

$$T_L = P_L^E \Big/ P_L^I = \frac{\sum_{i=1}^{N} p_{ti}^E \cdot q_{0i}^E}{\sum_{i=1}^{N} p_{ti}^I \cdot q_{0i}^I} \Bigg/ \frac{\sum_{i=1}^{N} p_{0i}^E \cdot q_{0i}^E}{\sum_{i=1}^{N} p_{0i}^I \cdot q_{0i}^I}$$