Deskriptive Statistik
Zusammenfassung und Formelsammlung

Verfasser: Christoph Molnar

Quellen: Vorlesungsfolien von Prof. Küchenhoff, Statistik-Bibel, Wikipedia

Einführung

Bezeichnungen

$\Omega = \{\omega_1, \ldots, \omega_n\}$	Grundgesamtheit
$\omega_1 \ldots \omega_n$	statistische Einheiten, Merkmalsträger
$\mathbb{S} = \{a_1, \ldots, a_k\}$	Merkmalsraum
n	Anzahl der untersuchten Einheiten
X	Merkmal
$x_1, \ldots, x_i, \ldots x_n, \ i = 1, \ldots, n$	beobachtete Werte / Ausprägungen
$\{x_1, \ldots, x_i, \ldots x_n\}\ i = 1, \ldots, n$	Urliste, Rohdaten
$x_{(j)}$	x_i mit dem j-kleinsten Wert (bei nominal beliebig)
$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$	geordnete Urliste

Unterscheidung

- Deskriptive Statistik: beschreibend
- Explorative Statistik: Suche nach Strukturen
- Induktive Statistik: Schlüsse von Daten auf Grundgesamtheit und allgemeine Phänomene

Merkmalstypen

- diskret: endlich oder abzählbar unendlich viele Ausprägungen
- stetig: alle Werte eines Intervalls sind mögliche Ausprägungen
- quasi-stetig: diskretes Merkmal, viele Ausprägungen, wie stetig behandelt
- dichotom: es gibt nur zwei Ausprägungen
- nominalskaliert: Ausprägungen sind Namen, keine Ordnung möglich
- ordinalskaliert: Ordnung möglich, Abstände nicht interpretierbar
- intervallskaliert: Ausprägungen sind Zahlen, Abstände interpretierbar
- verrhältnisskaliert: sinnvoller, absoluter Nullpunkt
- absolutskaliert: keine Transformationen möglich, Anzahlen

	sinnvoll interpretierbare Berechnungen				
Skalenart	auszählen	ordnen	Differenzen bilden	Quotienten bilden	Transformationen
nominal	ja	nein	nein	nein	alle
ordinal	ja	ja	nein	nein	alle monotonen
intervall	ja	ja	ja	nein	Y = aX+b
verhältnis	ja	ja	ja	ja	Y = aX
absolut	ja	ja	ja	ja	keine

Erhebungsarten

- Vollerhebung: Alle statistischen Einheiten der Grundgesamtheit werden untersucht
- Stichprobe = Teilerhebung
- Zufallsstichprobe: zufällige Ziehung aus der Grundgesamtheit
- Bewusste Auswahlverfahren (= "Expertenauswahl")
- Quotenauswahl
- Querschnittsdaten: einmalige Erhebung zu bestimmtem Zeitpunkt
- Zeitreihe
- Longitudinal-, Längsschnitts-, oder Paneldaten

Univariate Deskription

Allgemeines

$h_1, \ldots, h_k$ mit $h_j = h(a_j)$	absolute Häufigkeit der Ausprägung a_j
$f_1, \ldots, f_k$ mit $f_j = f(a_j) = \frac{h_j}{n}$	relative Häufigkeit
$F(x) = f(a_1) + \cdots + f(a_j) = \sum_{i:a_i \leq x} f_i$	Empirische Verteilungsfunktion, kummulierte Wahrscheinlichkeit

Lagemaßzahlen

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$	arithmetisches Mittel
$\bar{x}_\lambda = \sum_{i=1}^{n} \lambda_i x_i$ mit $\sum_{i=1}^{n} \lambda_i = 1$ und $0 \leq \lambda_i \leq 1$	gewichtetes arithmetische Mittel
x_{mod}	Modus (häufigster Wert)
x_{med}	Median (teilt die geordnete Urliste in der Hälfte)
x_p	Quantil
$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$ (x_i *proz. Wachstumsrate*)	geometrisches Mittel (Wachstums- oder Zinsfaktor)
$\bar{x}_H := \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i}}$	harmonisches Mittel (z.B. mittlere Geschwindigkeit)
$\bar{x}_\alpha = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} x_{(i)}$	getrimmtes Mittel

- bei Häufigkeitsdaten ist $\bar{x} = \sum_{j=1}^{k} a_j f_j$
- Modus: häufigster Wert, nicht immer eindeutig, alle Skalenniveaus
- Quantil: p der Daten sind kleiner als x_p,(1-p) größer
  $x_p = \begin{cases} x_{[np]+1} & \text{falls } np \text{ nicht ganzzahlig} \\ \in [x_{(np)}, x_{(np+q)}] & \text{falls } np \text{ ganzzahlig} \end{cases}$
  Der Median ist ein Spezialfall des Quantils mit p = 0,5
  $x_{0,25}$ unteres Quartil, $x_{0,75}$ oberes Quartil
- Es gilt im Allgemeinen: $\bar{x_H} \leq \bar{x_G} \leq \bar{x}$

Symmetrie und Schiefe

$g_p = \frac{(x_{1-p}-x_{med})-(x_{med}-x_p)}{x_{1-p}-x_p}$	Quantilskoeffizient
$g_m = \frac{m_3}{\tilde{s}^3} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{\tilde{s}^3}$	Momentkoeffizient der Schiefe

Verteilung ist:
symmetrisch, falls $\bar{x} \approx x_{med} \approx x_{mod}$, $g_p = 0$, $g_m = 0$
linkssteil (rechtsschief), falls $\bar{x} > x_{med} > x_{mod}$, $g_p > 0$, $g_m > 0$
rechtssteil (linksschief), falls $\bar{x} < x_{med} < x_{mod}$, $g_p < 0$, $g_m < 0$

Streuungsparameter

$q = x_{max} - x_{min}$	Spannweite(Range)
$q = x_{0,75} - x_{0,25}$	Quartilsabstand
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	Varianz bei Stichprobe
$\tilde{S^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{1=1}^{n} x_i^2 - \bar{x}^2$	Varianz bei Vollerhebung
$S = \sqrt{S^2}$	Standardabweichung
$v = \frac{\tilde{S}}{\bar{x}}$	Variationskoeffizient
$\frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$	Mittlere Absolute Abweichung (MAD)

Schichtenbildung

$n_j, j = 1, \ldots, n$	absolute Häufigkeit der Klasse j
$\bar{x} = \frac{1}{n}(n_1\bar{x}_1 + \cdots + n_r\bar{x}_r) = \frac{1}{n}\sum_{k=1}^{r} n_j\bar{x}_j$	arithmetisches Mittel bei Schichtenbildung
$\tilde{s}^2 = \frac{1}{n}\sum_{j=1}^{r} n_j\tilde{s}_j^2 + \frac{1}{n}\sum_{j=1}^{r} n_j(\bar{x}_j - \bar{x})^2$	Varianz bei Schichtenbildung, Streuungszerlegung

- Schichten entstehen, wenn man die Daten zu Schichten zusammenfasst. Wenn man Käse, der im Supermarkt angeboten wird, untersucht, kann man die verschieden Einheiten zu Schichten zusammenfassen, z.B. normaler Käse, stinkender Käse und Analogkäse.
- Gesamtstreuung = Streuung innerhalb der Schichten + Streuung zwischen den Schichten

Klassierte Daten

$[c_{i-1}, c_i)$	Klasse von x-Werten c_{i-1} bis c_i
$d_i = c_{i-1} - c_i$	Klassenbreite
$x_{mod,grupp}$	Klassenmitte der Klasse mit der höchsten Beobachtungszahl
$x_{med,grupp} = c_{i-1} + \frac{d_i(0,5-F(c_{i-1}))}{f_i}$	Median bei gruppierten Daten, wobei $[c_{i-1}, c_i)$ Klasse in die der Median fällt
$x_{p,grupp} = c_{i-1} + \frac{d_i(p-F(c_{i-1}))}{f_i}$	x_p bei klassierten Daten mit $[c_{i-1}, c_i)$ Klasse, in die x_p fällt.
$m_j = \frac{c_{j-1}+c_j}{2}, j = 1, \ldots, k$	Klassenmitten
$\bar{x}_{grupp} = \sum_{i=1}^{k} f_j m_j$	arithmetisches Mittel bei klassierten Daten

- Gruppierung bei metrischen, stetigen oder quasi-stetigen Merkmalen
- Wenn ein Merkmal zu viele Ausprägungen hat, kann man es klassieren

Dichtekurven

$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$	Kerndichteschätzer
Kerne:	
$K(u) = \frac{3}{4}(1-u^2)$ für $-1 \leq u \leq 1$	Epanechnikov-Kern
$K(u) = \frac{15}{16}(1-u^2)^2$ für $-1 \leq u \leq 1$	Bisquare-Kern
$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$ für $u \in \mathbb{R}$	Gauß-Kern

- Glättet/approximiert Histogramme
- Für eine Dichtefunktion muss gelten: $\int_{-\infty}^{\infty} f(x)dx = 1$
- Die Flächen unter der Kurve sollen den approximativen Wahrscheinlichkeiten entsprechen: $\int_a^b f(x)\,dx = \frac{1}{n}\#\{x_i | a \leq x_i \leq b\}$
- Der Median teilt Dichtekurve in zwei gleichgroße Teile

Multivariate Deskription

Allgemeines

$f_{ij} = \frac{h_{ij}}{n}$	Die relative Häufigkeit der Kombination (a_i, b_j)
$f_{i\cdot} = \sum_{j=1}^{m} = \frac{h_{i\cdot}}{n}$	relativen Randhäufigkeiten zu X
$f_{\cdot j} = \sum_{i=1}^{k} = \frac{h_{\cdot j}}{n}$	relativen Randhäufigkeiten zu Y
$f_Y(a_1\|b_i) = \frac{h_{i1}}{h_{i\cdot}}, \ldots, f_Y(b_m\|a_1) = \frac{h_{im}}{h_{i\cdot}}$	Bedingte Häufigkeitsverteilung von Y mit Bedingung $X = a_i$

Kontingenztafel

absolute Häufigkeiten:

	b_1	…	b_m	
a_1	h_{11}	…	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	…	h_{2m}	$h_{2\cdot}$
⋮	⋮		⋮	⋮
a_k	h_{k1}	…	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	…	$h_{\cdot m}$	n

relative Häufigkeiten:

	b_1	…	b_m	
a_1	f_{11}	…	f_{1m}	$f_{1\cdot}$
a_2	f_{21}	…	f_{2m}	$f_{2\cdot}$
⋮	⋮		⋮	⋮
a_k	f_{k1}	…	f_{km}	$f_{k\cdot}$
	$f_{\cdot 1}$	…	$f_{\cdot m}$	1

Spezialfall 2 x 2:

a	b	(a+b)
c	d	(c+d)
(a+c)	(b+d)	

Odd's ratio (relative Chancen, Kreuzproduktverhältnis)

- $\gamma(1,2|X=3,\, X=4) = \frac{\gamma(1,2|X=3)}{\gamma(1,2|X=4)} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}}$
- "Population 3 besitzt ein γ mal größere Chance auf 1 gegenüber 2 als Population 4"
- Nur bei diskreten Merkmalen mit wenig Ausprägungen, mind. nominalskaliert
- Auch bei höheren Skalen wird nur das Nominalniveau benutzt

χ^2- und Kontingenzkoeffizient

$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{h_{i\cdot}h_{\cdot j}}{n}\right)^2}{\frac{h_{i\cdot}h_{\cdot j}}{n}} \in [0, \infty)$	χ^2-Koeffizient
$\chi^2 = n\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(f_{ij} - f_{i\cdot}f_{\cdot j})^2}{f_{i\cdot}f_{\cdot j}}$	χ^2-Koeffizient mit relativen Häufigkeiten
$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$	χ^2-Koeffizient für 2 x 2 Kontingenztaffeln (Sonderfall)
$K = \sqrt{\frac{\chi^2}{n+\chi^2}}$ mit $K \in \left[0, \sqrt{\frac{M-1}{M}}\right]$	Kontingenzkoeffizient
$K^* = \frac{K}{\sqrt{\frac{M-1}{M}}}$ mit $K^* \in [0, 1]$	korrigierte Kontingenzkoeffizient
mit $M = min\{k, m\}$	

- Merkmale diskret, mit wenigen Ausprägungen
- Für alle Skalen geeignet, benutzt aber nur das Nominalniveau der Daten
- Invariant gegenüber Spalten-/Zeilentausch
- Idee: Wie sollten die Felderinhalte aussehen, wenn kein Zusammenhang zwischen den beiden Merkmalen besteht? χ^2 summiert die normierten Quadrate der Differenz zwischen Wert bei Unabhängigkeit und tatsächlichem Wert
- Es wird nur die Stärke des Zusammenhangs gemessen und nicht die Richtung
- χ^2 wächst mit dem Stichprobenumfang n

Kappa-Koeffizient

$\kappa = \frac{f_0 - f_e}{1 - f_e}$	Kappa Koeffizient
$f_0 = \frac{\sum_{i=1}^{l} h_{ii}}{n}$	Der Anteil der Übereinstimmung beider Beobachter
$f_e = \sum_{i=1}^{l} f_{i\cdot}f_{\cdot i} = \frac{\sum_{i=1}^{l} n_{i\cdot}n_{\cdot i}}{n^2}$	Zufällige Übereinstimmung, wenn kein Zusammenhang besteht

- Der Zähler entspricht Differenz aus der beobachteten Übereinstimmung und der unter Zufälligkeit zu erwartenden Übereinstimmung
- Nur die Einträge auf der Diagonalen werden gewertet
- Probleme: Es kann sein, dass ein Beobachter generell immer höher/besser bewertet (falsche Kalibrierung) oder Beobachter schätzen Subjekte falsch ein

Bravais-Pearson-Korrelationskoeffizient

$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	Kovarianz
$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \in [-1, 1]$	Bravais-Pearson-Korrelationskoeffizient
$r_{XY} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}\right)}}$	Bravais-Pearson in besser
$r_{XY\|T} = \frac{r_{XY} - r_{XT}r_{YT}}{\sqrt{(1 - r_{XT}^2)(1 - r_{YT}^2)}}$	Bravais-Pearson mit Trendbereinigung der Variable T

- Misst den linearen Zusammenhang, nichtlineare Zusammenhänge werden unter Umständen nicht gemessen

- Beide Variablen müssen mindestens intervallskaliert oder dichotom sein
- Unverändert bei linearen Transformationen und es gilt: $r_{XY} = r_{YX}$
- empfindlich gegenüber Ausreißern
- "schwache" Korrelation" $|r| < 0,5$
  "mittlere Korrelation" $0,5 \leq |r| < 0,8$
  "starke Korrelation" $0,8 \leq |r|$
- $r > 0$ positve Korrelation,
  $r < 0$ negative Korrelation,
  $r = 0$ kein linearer Zusammenhang

Spearmans (Rang-)Korrelationskoeffizient

$r_{SP} = \frac{\sum(rg(x_i)-\bar{rg}_X)(rg(y_i)-\bar{rg}_Y)}{\sqrt{\sum(rg(x_i)-\bar{rg}_X)^2 \sum(rg(y_i)-\bar{rg}_Y)^2}} \in [-1,1]$	Spearmans Korrelationskoeffizient
mit $\bar{rg}_x = \frac{n+1}{2} = \bar{rg}_y$	
$r_{SP} = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$	vereinfachter Spearman ohne Bindungen
mit $d_i = rg(x_i) - rg(y_i)$	

- X,Y müssen mindestens ordinal skaliert sein
- Bei Bindungen werden mittlere Ränge gebildet
- Wie Bravais-Pearson, nur mit Rängen anstatt mit der Urliste, daher weniger ausreißerempfindlich
- Unverändert bei streng monotonen Transformationen und $r_{SP}(X,Y) = r_{SP}(Y,X)$
- Misst den monotonen Zusammenhang
- $r_{SP} > 0$ gleichsinniger monotoner Zusammenhang
  $r_{SP} < 0$ gegensinniger monotoner Zusammenhang
  $r_{SP} \approx 0$ kein monotoner Zusammenhang

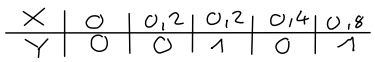
Kendall´s Tau

$\tau_a = \frac{N_C - N_D}{\frac{n(n-1)}{2}} \in [-1,1]$	Kendalls Tau ohne Bindungen (Ties)
$\tau = \frac{N_D - N_C}{\sqrt{(N_C+N_D+T_X)(N_C+N_D+T_Y)}}$	Kendalls Tau mit Bindungen (Ties)

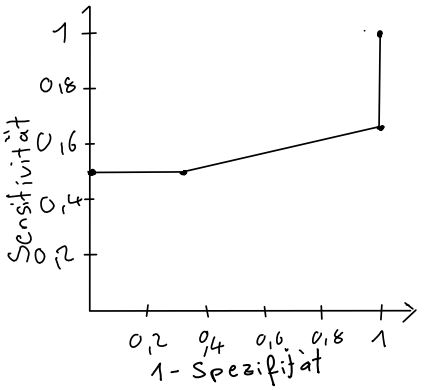
- N_C Anzahl der konkordanten Paare
  $(x_i < x_j$ und $y_i < y_j$ oder $x_i > x_j$ und $y_i > y_j)$
- N_D Anzahl der diskordanten Paare
  $(x_i > x_j$ und $y_i < y_j$ oder $x_i > x_j$ und $y_i < y_j)$
- T_X Anzahl der Paare, bei denen $x_i = x_j$
- T_Y Anzahl der Paare, bei denen $y_i = y_j$
- Falls bei einem Paar $x_i = x_j \wedge y_i = y_j$, dann wird dieses Paar nicht berücksichtigt
- Im Gegensatz zu Spearman werden alle Wertepaare untereinander verglichen und nicht nur die zwei Werte eines Paares
- In der Regel kleiner als Spearmans Korrelationskoeffizient

ROC-Kurve

Beispiel:



- X ist metrisch, Y ist dichotom
- Meistens für medizinische Tests verwendet.
  Y = 0 gesund, Y=1 krank, X ist das Ergebnis des Testes
- X-Achse: 1-Spezifität (= falsch als positiv eingestuft)
  $f\left(\hat{Y}=1|Y=0\right)=f\left(x\geq c|Y=0\right)=S_0(c)$
- Y-Achse: Sensitivität (= richtig als positiv eingestuft)
  $f\left(\hat{Y}=0|Y=1\right)=f\left(x\geq c|Y=1\right)=S_1(c)$



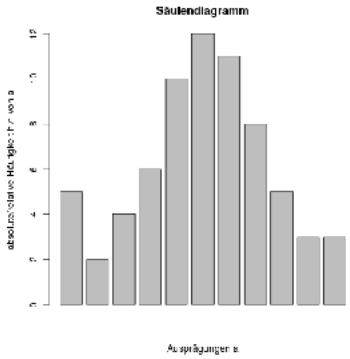
- Die ROC-Kurve besteht aus den Punkten $S_0(c), S_1(c)$
- Für c setzt man einzeln alle x-werte ein
  und errechnet die Sensivität und 1-Spezifität, um so die Punkte für die Kurve zu bekommen.
- AUC: Fläche unter der Kurve
- GINI-Koeffizient: Normierte Fläche zwischen Winkelhalbierender und ROC-Kurve
  $GINI=\frac{N_C-N_D}{N}$ mit $N=Anzahl\,der\,Paare$ und N_C, N_D wie bei Kendall's Tau

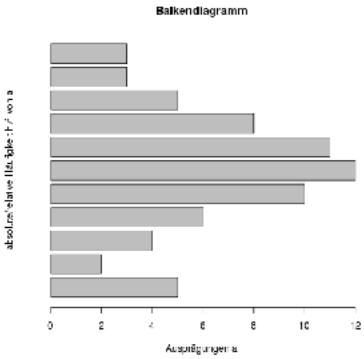
Lineare Regression

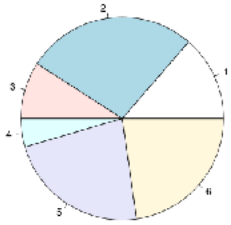
$\hat{y}=\alpha+\beta\hat{x}$	lineare Regression
$\hat{\beta}=\frac{S_{XY}}{S_X^2}$	Steigung der Regressionsgeraden
$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$	Y-Achsenabschnitt der Regressionsgeraden
$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}y_i-\bar{y})^2}$	Bestimmtheitsmaß R^2
$R^2=r_{XY}^2\in[0,1]$	Bestimmtheitsmaß
$X=\gamma+\delta Y \quad \delta=\frac{S_{XY}}{S_Y^2}$	Umkehrregression
$\beta\cdot\delta=r^2$	

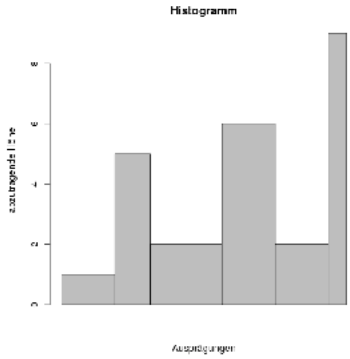
- Lineares Modell: $Y=\alpha+\beta X+\epsilon$
- α und β minimieren die Funktion $Q(\alpha,\beta)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{1}{n}\sum_{i=}^{n}(y_i-(\alpha+\beta x_i))^2$
- Die Regressionsgerade geht durch $(\bar{x},\bar{y})$
- R^2 gibt an, welcher Anteil der Streuung durch die Gerade erklärt werden kann
  Je größer R^2 desto besser ist das lineare Modell.
- Scheinkorrelation: hohe Korrelation zw. 2 Merkmalen, die aber inhaltlich unsinnig ist. Nicht-gemessene 3. Variable, die Einfluss auf die beiden anderen hat.
- Verdeckte Korrelation: eine Korrelation wird verdeckt durch nicht Berücksichtigen einer weiteren Variablen bzw. man hätte die Population in mehr Teilpopulationen teilen müssen.

Diagramme









Histogramm Howto:
Häufigkeiten entsprechen den Flächen der Rechtecke
Aufteilung in Klassen (falls noch nicht geschehen)
Bestimmung der relativen Häufigkeiten $f_j = \frac{n_j}{n}$
Abzutragende Höhen $h_j = \frac{f_j}{d_j}$ mit d_j Breite der Klasse j