Formelsammlung Statistik

Andrey Behrens

August 2009

Das ist eine Formelsammlung für Statistik. Die Formelsammlung enthält alle Formeln aus dem Skript des Wintersemesters 2009/2010. Außerdem ein paar Sachen die mir sinnvoll erschienen und für die Klausur notwendig sein könnten, sowie Formblätter zum schnellen Ausfüllen während der Klausur.

Teil I

Begriffe

Statistische Masse	Umfang der Einheiten einer statistischen Untersuchung
Statistische Einheit	Untersuchungsobjekt einer statistischen Untersuchung. Träger der interessanten Informationen.
Merkmal	Zu betrachtendes Attribut einer Einheit. Etwa Einkommen, Altern, ...
Merkmalstypen	

diskrete	Merksmaltypen bestehen aus einer überschaubare, endliche Menge (etwa Geschlecht),
stetige	Merksmaltypen können in einem bestimmten Bereich jeden reelen Wert annehmen,
quasi-stetige	Merksmaletypen sind eigentlich diskret, enthalten aber sehr grosse Menge von möglichen Merkmalen

Gruppierung	Sortierung, gleiche Merkmalsausprägung
Klassifizierung	benachbarte Ausprägungen werden zu einer Klasse zusammengefasst. Übliche Schreibweise $[200; 400)$ mit der Bedeutung $200 \leq x < 400$.
Skalenniveau	

nominal	qualitativ (also keine Zahlen), etwa Geschlecht oder Studiengang. Darstellung als gruppierter Wert.
ordinal	Merkmalsausprägung mit objektiver Rangordnung, Abstände sind aber nicht bezifferbar (etwa Noten). Darstellung als gruppierter Wert.
metrisch	*Interval quantitativ*: reele Zahlen, natürliche Rangfolge, eindeutige Abstände, etwa Sparsumme, *Verhältnis quantitativ*: reele Zahlen, natürliche Rangfolge, eindeutige Abstände, absoluter Bezugspunkt (etwa Nullpunkt). Beispiel: Alter. Darstellung als klassierter Wert.

Teil II

Univariante Datenanalyse

1 Beispiele

Gruppiert: Für nominale und ordinale Werte

x_i	h_i	H_i	f_i	F_i	$\triangle x_i$	f_i^*	h_i^*
280	1	1	0,1	0,1	-	-	-
340	2	3	0,2	0,3	-	-	-
560	1	4	0,1	0,4	-	-	-
600	1	5	0,1	0,5	-	-	-
650	3	8	0,3	0,8	-	-	-
740	1	9	0,1	0,9	-	-	-
1180	1	10	0,1	1,0	-	-	-

Klassiert: Für metrische Werte

x_i	h_i	H_i	f_i	F_i	$\triangle x_i$	f_i^*	h_i^*
[200;400)	21	21	0,21	0,21	200	0,00105	0,1050
[400;700)	56	77	0,56	0,77	300	0,00187	0,1867
[700;1000)	19	96	0,19	0,96	300	0,00063	0,0633
[1000;1500)	2	98	0,02	0,98	500	0,00004	0,0040
[1500;2000)	2	100	0,02	1,00	500	0,00004	0,0040

2 Häufigkeiten

Name	Math		Formel	TR
abs. Häufigkeit	h_i	hi	-	-
abs. Summenhäufigkeit	H_i	shi	$h_1 + ... + h_i = \sum_{j=1}^{i} h_j$	cusum(hi)
relative Häufigkeit	f_i	fi	$\frac{h_i}{N}$ mit $\sum_{i=1}^{k} f_i$	$relhfg(hi)$
abs. Summenhäufigkeit	F_i	sfi	$f_1 + ... + f_i = \sum_{j=1}^{i} f_j$	$cumsum(relhfg(hi))$
Stat Masse	N	n	$\sum_{i=1}^{k} h_i$	$sum(hi)$
abs Häufigkeitsdichte	h_i^*	his	$\frac{h_i}{\Delta x_i}$	his
rel Häufigkeitsdichte	f_i^*	fis	$\frac{f_i}{\Delta x_i}$	fis

2.1 Funktion der relatitiven Summenhäufigkeit/Verteilungsfunktion

2.1.1 Bei gruppierte Daten

$$F(x) = \begin{cases} 0 & x < x_1 \\ F_i & x_i \leq x < x_{i+1} \\ 1 & x \geq x_k \end{cases}$$

Als Rechenbeispiel:

F(500)=0,30 -> Es wird nicht gerechnet, sondern aus dem Diagramm abgelesen, da es sich um gruppierte Werte handelt!

Als grafische Lösung (Treppendiagramm, keine Zwischenwerte!) siehe Abbildung **??**

2.1.2 Bei klassierten Daten

$$F(x) = \begin{cases} 0 & x < x_1^u \\ F(x_i^u) + \frac{f_i}{\Delta x_i} * (x - x_i^u) & x_i^u \leq x < x_i^o \\ 1 & x \geq x_k^o \end{cases}$$

als Rechenbeispiel:

1. Klasse aus Diagramm ablesen (H_i), untere und obere Grenzen der Klasse herauslesen.
2. In Formel einsetzen: $F(500) = 0.21 + \frac{0{,}56}{300}(500 - 400) = 0,397 = 39,7\%$

als grafische Lösung siehe Funktionsdiagramm **??**

2.2 Darstellung der relativen Häufigkeiten

gruppiert Stabdiagramm siehe Abbildung **??**

klassiert Histogramm, siehe Abbildung **??**

3 Statistische Maßzahlen

Name	Math	TR	nominal	ordinal	metrisch	Vor- und Nachteile
Modal	x_D	xd	ja	ja	ja	Ist die Merkmalsausprägung, die am häufigsten vorkommt. Es kann mehrere Modalwerte geben.
Median	x_z	xz	?	ja	ja	Mitte aller Merkmalsträger, bzw. welcher Merkmalswert wird von der Hälfte aller Merkmalsträger nicht überschritten. Vorteil: Robust gegen Ausreißer.
Quantil	x_p	xp	?	?	?	ein Teil aller Merkmalsträger (etwa 0,25x oder 0,75x) bzw. welcher Merkmalswert wird von einem Teil aller Merkmalsträger nicht überschritten. Dabe ist das $x_p = x_{0.5} = x_z$
Arith. Mittelw.	$\bar{x}$	xs	nein	nein	ja	Der Durchschnitt oder Mittelwert aller Merkmale
Geom Mittelw.	x_G	xg	?	?	ja	Mittelwert für Produkte, etwa bei Verhältnissen oder Wachstumswerten. Nur für Zahlen>0 sinnvoll.

Tabelle 3.1: Überblick Lageparameter

Math	TR	Formel		Erläuterung
x_D	xd	Gruppen	da x_i wo f_i am größten ist	Ist die Merkmalsausprägung, die am häufigsten vorkommt. Es kann mehrere Modalwerte geben.
		Klassen	Mitte der modalen Klasse $x_D = \frac{x_i^u + x_i^o}{2} = x_i^{'}$	
x_z	xz			Median bzw. Zentralwert ist der Wert, der in der Mitte der Variantsreihe liegt. Ist N gerade, wird der Mittelwert der zwei mittelsten Werte ermittelt. Beispiel: Zuerst Klasse bestimmen und dann $400 + \frac{0.5-0.21}{0.56} * 300 = 555.36$
		Gruppe	$x_z = 0.5N$	
		Klasse	$x_z = x_i^u + \frac{0.5-F(x_i^u)}{f_i} * \Delta x_i$	
x_p	xp			Wird eine Variationsreihe in gleich große Teile zerlegt, entstehen Quantile. Typisch sind 0,25, 05, 0,75. Der Quantilabstand ist $Q = x_{0.75} - x_{0.25}$. Das 0,5-Quantil ist gleich dem Median.
		Gruppe	$x_p = p \cdot N$	
		Klasse	$x_p = x_i^u + \frac{p-F(x_i^u)}{f_i} * \Delta x_i$	
$\bar{x}$	xa			Arithmetischer Mittelwert bzw. Durchschnitt. Durchschnitte werden mit dieser Formel addiert: $\bar{x} = \frac{\sum_{m=1}^{k} N_m * \bar{x_m}}{\sum_{m=1}^{k} N_m}$
		Gruppe	$\bar{x} = \sum_{i=1}^{k} x_i f_i$	
		Klasse	$\bar{x} = \sum_{i=1}^{k} x_i^{'} f_i$	
x_G	xg			Der Geometrische Mittelwert wird bei der Mittelung von Wachstumsraten oder multiplikativ verknüpften Daten angewendet.
		$x_G = \sqrt[N]{\prod_{i=1}^{k} x}$		

Tabelle 3.2: Lageparameter

Math	TR	Formel		Erläuterung
R	r	Gruppe	$R = x_{max} - x_{min}$	Die Spannweite ist die Differenz zwischen größtem und kleinstem Merkmalswert.
		Klasse	$R = x_k^o - x_1^u$	
Q	q	$Q = x_{0.75} - x_{0.25}$		Der Quantilsabstand ist der Abstand zwischen oberem und unterem Quantil.
s_x^2	s2x			Die Varianz ist die mittlere quadratische Abweichung aller Merkmalsausprägungen vom arith. Mittelwert. Alternative Zeichen der Varianz sind $s_x^2 = s^2 = \sigma^2$
		Gruppe	$s_x^2 = \left\{ \sum_{i=1}^{k} \left[(x_i)^2 \cdot f_i \right] \right\} - \overline{x}$	
		Klasse	$s_x^2 = \left\{ \sum_{i=1}^{k} \left[(x_i')^2 \cdot f_i \right] \right\} - \overline{x}^2$	
		$s^2 = \frac{N_1 s_1^2 + N_2 s_2^2}{N_1 + N_2} + \frac{N_1 (\bar{x_1} - \bar{x})^2 + N_2 (\bar{x_2} - \bar{x})^2}{N_1 + N_2}$		Varianz der Grundgesamtheit. Gleichungsbeispiel bei der Annahme, dass es zwei Teilmengen gibt und die jeweils die Varianzen und Mittelwerte bekannt sind.
s_x	sx	$s_x = \sqrt{s_x^2}$		Standardabweichung mittlere Abweichung vom Mittelwert. Nachteil: Bei großen Merkmalsmengen nimmt die Schwankungsbreite zu. Ein Vergleich zwischen Messreihen mit großen und kleinen Verteilungen ist daher ggf. nicht mehr sinnvoll. Statt dessen: Variationskoeffizient.
v	v	$v = \frac{s_x}{\bar{x}}$		Variationskoeffizient = Auf den Mittelwert bezogenes relatives Streuungsmaß, sofern nur positive Werte auftreten.

Tabelle 3.3: Streuungsparameter

Math	TR	Formel	Erläuterung
p_I	pi	$p_i = \frac{x_i' \cdot hi}{N \cdot \bar{x}}$	Konzentrationskoeffizient berechnet den Anteil eines Merkalswertes an der Merkmalssumme
P_i	spi	$P_i = \sum_{j=1}^{i} p_j$	Das Konzentrationsmaß beschreibt die relative Merkmalssumme. Die Lorenzkurve veranschaulicht das Konzentrationsmaß grafisch.
-	-	-	Die Fläche zwischen Gleichverteilung und Lorenzkurve wird als Lorenzfläche bezeichnet und ist ein weiteres Konzentrationsmaß. Je größer die Fläche, desto größer die Konzentration. Beispiel zur Lorenzkurve siehe **??**
G	g	$G = \frac{0.5 - A(L)}{0.5}$	Der Gini-Koeffizient misst die Höhe der relatitiven Konzentration über das Verhältnis der Lorenzfläche zur Fläche bei maximaler Konzentration mit 0.5
$A(L)$	al	$A(L) = \sum_{i=1}^{k} \frac{P_{i-1} + P_i}{2} \cdot f_i$ mit $P_0 = 0$	Fläche unter der Trapezkurve Hinweis: Sind großgeschriebene P's also spi.

Tabelle 3.4: Relative Konzentration

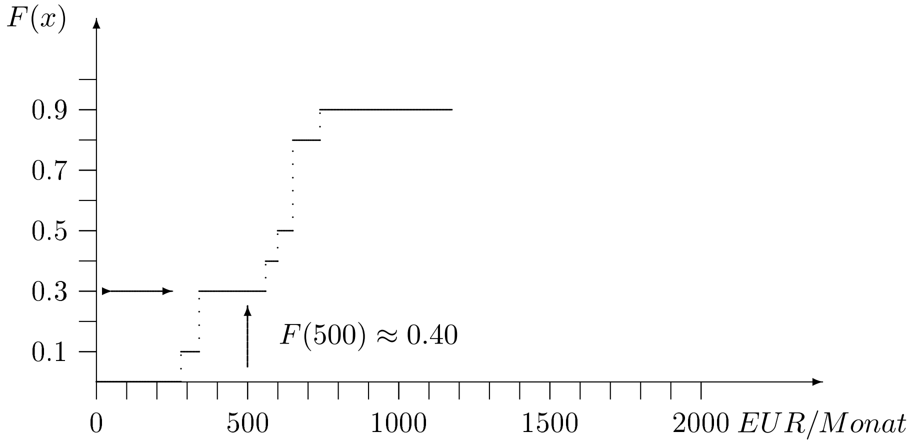

Abbildung 3.1: Funktion relativer Sumenhäufigkeit F(x) bei gruppierten Daten

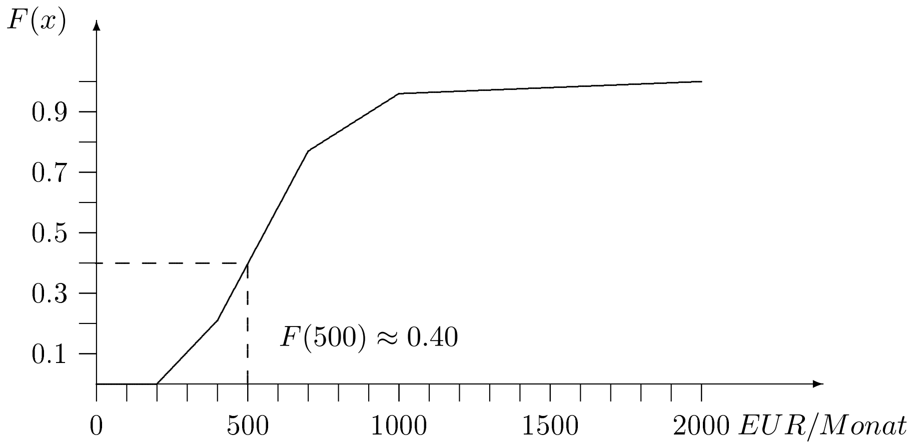

Abbildung 3.2: Funktion relativer Summenhäufigkeit bei klass. Daten

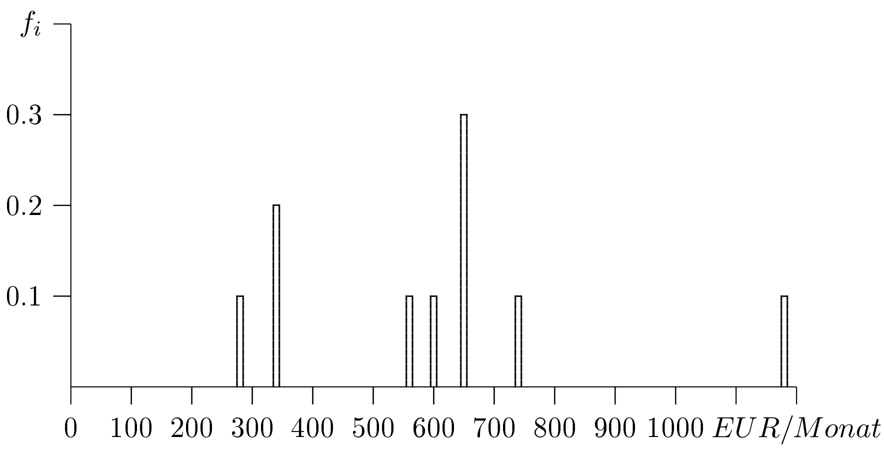

Abbildung 3.3: Relative Häufigkeit von Gruppen: Stabdiagramm

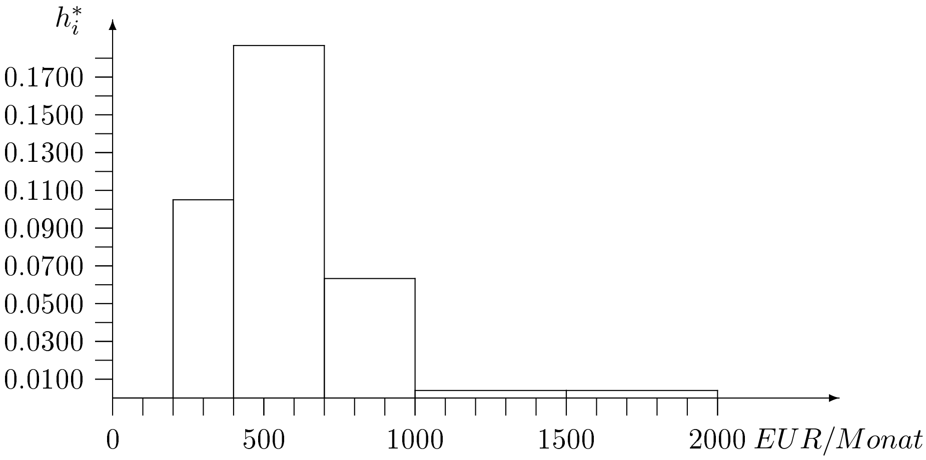

Abbildung 3.4: Relative Häufigkeit von Klassen: Histogramm

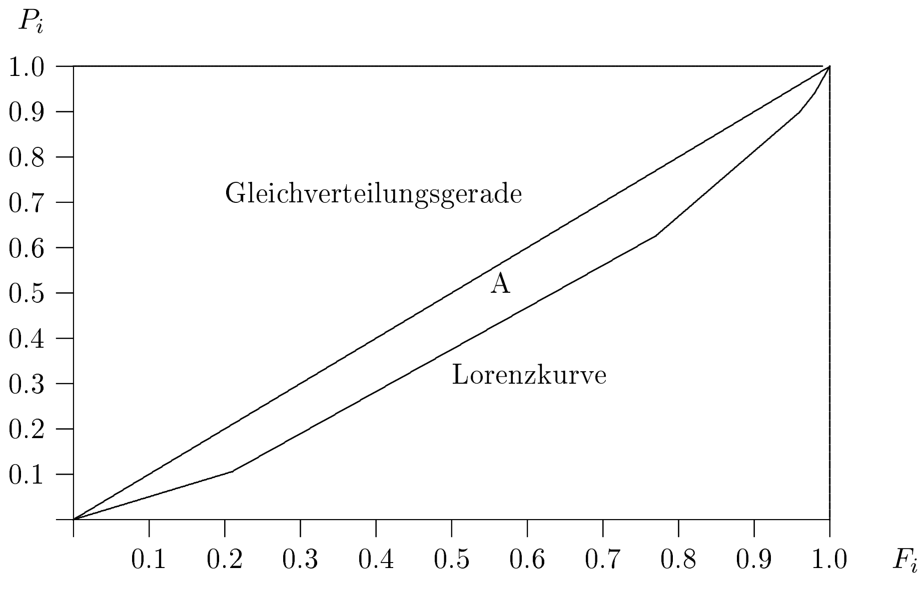

Abbildung 3.5: Lorenzkurve

4 Quellen

(1) Statistikscript Prof. Dr. Müller, HS Wismar

(2) Taschenbuch der Wirtschaftsmathematik, Wolfgang Eichholz und Eberhard Vilkner

Teil III

Formblätter

x	x	Klasse oder Gruppe einer statistischen Zählung. Variable kann Zeichen haben wie 1, i, k die für das 1-te, i-te oder letzte Gruppe/Klasse stehen.
x_d	xd	Modalwert, der Wert mit der häufigsten Merkmalsausprägung
x_z	xz	Median, Mitte aller Merkmalsausprägungen, d.h. nach oben und unten gleich viele Merkmalsausprägungen
x_p	xp	Quantile überschreiten einen gewissen Anteil von Merkmalsausprägungen *nicht*
$x_i^{'}$		Klassenmitte der$i$-ten Klasse
x_i^u x_i^o		untere bzw. obere Grenze der i-ten Klasse
h	h	Anzahl von Einheiten innerhalb einer Gruppe oder Klasse. Tiefgestellte Zeichen gleiche Bedeutung wie bei x Die Summe aller h ist die statistische Masse
H_i	shi	absolute Summenhäufigkeit, wie h_i aber aufsteigend addiert. Der größte Wert=N
f_i	fi	relative Häufigkeit. Summe aller $f_i = 1$ Entspricht dem prozentualen Anteil an der statistischen Masse.
F_i	sfi	relative Summenhäufigkeit. Wie f_i aber aufsummiert. Der größte Wert = 1
Δx_i	dxi	Klassenbreite der i-ten Klasse
s_i	si	relative Summenhäufigkeit einer Klasse
N	n	Statistische Masse, also die Menge aller Merkmalsausprägungen.

Table .1: Überblick Variablen

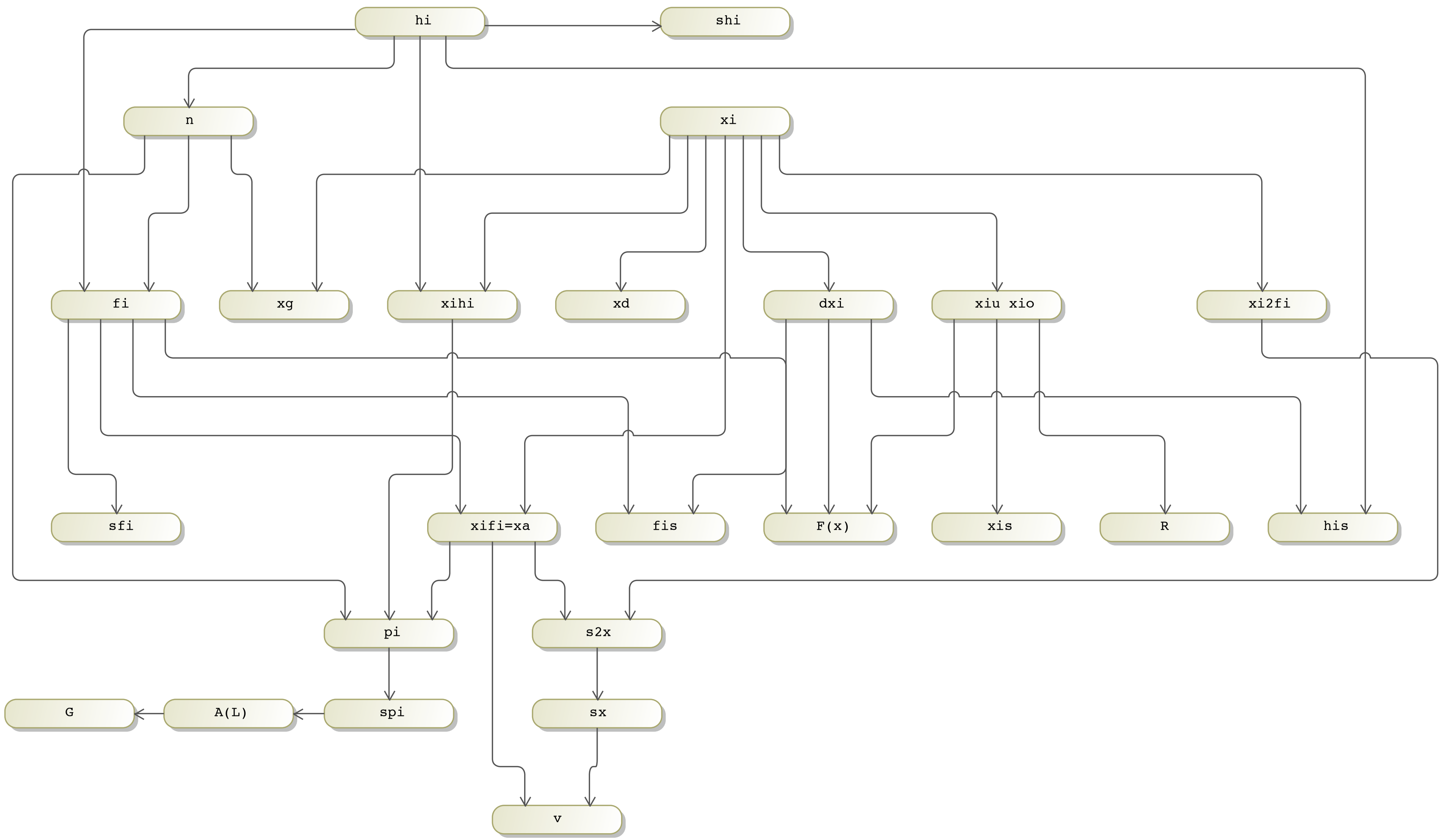

Figure .1: Zusammenhänge von Variablen

Gruppe	abs. Häufig	abs. Summen-häufig	rel. Häufig	rel. Summen-häufig				Konz- koeff.	Konz- maß	Fläche unter Lorenzkurve
x_i	h_i	H_i	f_i	F_i	$x_i f_i$	$x_i \cdot h_i$	$x_i^2 \cdot h_i$	p_i	P_i	$A(L)$
$\sum$	$N =$	-	$= 1$	-	$\bar{x}=$	$=$	$=$	-	-	$=$

Gruppe	abs. Häufig	abs. Summen-häufig	rel. Häufig	rel. Summen-häufig				Konz- koeff.	Konz- maß	Fläche unter Lorenzkurve
x_i	h_i	H_i	f_i	F_i	$x_i f_i$	$x_i \cdot h_i$	$x_i^2 \cdot h_i$	p_i	P_i	$A(L)$
$\sum$	$N =$	-	$= 1$	-	$\bar{x}=$	$=$	$=$	-	-	$=$

Gruppe	abs. Häufig	abs. Summen- häufig	rel. Häufig	rel. Summen- häufig				Konz- koeff.	Konz- maß	Fläche unter Lorenzkurve
x_i	h_i	H_i	f_i	F_i	$x_i f_i$	$x_i \cdot h_i$	$x_i^2 \cdot h_i$	p_i	P_i	$A(L)$
$\sum$	$N =$	-	$= 1$	-	$\bar{x}=$	$=$	$=$	-	-	$=$

Gruppe	abs. Häufig	abs. Summen- häufig	rel. Häufig	rel. Summen- häufig				Konz- koeff.	Konz- maß	Fläche unter Lorenzkurve
x_i	h_i	H_i	f_i	F_i	$x_i f_i$	$x_i \cdot h_i$	$x_i^2 \cdot h_i$	p_i	P_i	$A(L)$
$\sum$	$N =$	-	$= 1$	-	$\bar{x}=$	$=$	$=$	-	-	$=$