

4 Konzentrations- und Armutsmessung

4.0 Vorbemerkungen

Konzentration: Ausmaß der Ballung von großen Anteilen an der gesamten Merkmalssumme auf wenige Einheiten.

Literatur

- Fahrmeier, L. & Künstler, R. & Pigeot, I. & Tutz, G. (7. Auflage, 2010): Statistik: Der Weg zur Datenanalyse. Springer Verlag, Berlin, Heidelberg, New York.
- Toutenburg, H., Heumann, C (7. Auflage 2009): Deskriptive Statistik. Springer Verlag, Berlin, Heidelberg, New York.
- Wagschal, Uwe (1999): Statistik für Politikwissenschaftler. Oldenbourg Wissenschaftsverlag, Oldenbourg.

Bem. 4.1. *Durchgängige Annahmen in Kapitel 4.1 und 4.2*

- X sei ein verhältnisskaliertes Merkmal (mit Urliste x_1, \dots, x_n) und der Größe nach geordneten Ausprägungen a_1, \dots, a_k und h_1, \dots, h_k . Ferner seien f_1, \dots, f_k die zugehörigen absoluten bzw. relativen Häufigkeiten; die empirische Verteilungsfunktion werde mit $F_X(\cdot)$ oder $F(\cdot)$ bezeichnet.
- Zudem sei $x_i \geq 0$, für alle $i = 1, \dots, n$ und $\sum_{i=1}^n x_i > 0$

-
- Betrachtet werden die der Größe nach geordneten Daten:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

(Achtung: Die Klammern im Index werden in der Literatur weggelassen (z.B. in (Fahrmeir et al. (2010))); dann muss man vor dem Anwenden der dortigen Formeln die Daten ordnen. Allerdings wird auch hier angenommen, dass $a_1 < a_2 < \dots < a_k$ gilt, dass also diese Ausprägungen bereits geordnet sind.)

Warum überhaupt Konzentrationsmessung als eigenes Gebiet?

Bem. 4.2. [Zum Ungenügen der Varianz und des Variationskoeffizienten als (alleiniges) Maß für (relative) Konzentration]

Es gibt eine Unmenge an Konzentrations- und Ungleichheitsmaßen, z.B. Rinne (2003, Taschenbuch der Statistik. Frankfurt am Main) führt über 50 verschiedene Maße auf, hier natürlich nur Beschränkung auf ein paar wesentliche.

Im Folgenden wird öfter, um die Beschreibung nicht zu komplex werden zu lassen, von „reich“ und „arm“ gesprochen, auch wenn andere Merkmale als Einkommen und Vermögen betrachtet werden. Reich steht dann einfach für mit relativ großem Anteil an der Gesamtsumme (z.B. Umsatz, Stimmanteile).

4.1 Relative Konzentration

4.1.1 Die Lorenzkurve

Definition 4.3.

Gegeben sei die geordnete Urliste $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ eines verhältnisskalierten Merkmals X , das den Annahmen aus Bemerkung ?? genügt.

Die stückweise lineare Kurve durch die Punkte $(0, 0)$, (u_1, v_1) , (u_2, v_2) , \dots , $(u_n, v_n) = (1, 1)$, wobei für jedes $j = 1, \dots, n$

$$u_j := \frac{j}{n}$$

und

$$v_j := \frac{\sum_{i=1}^j x_{(i)}}{n} = \frac{\sum_{i=1}^j x(i)}{\sum_{i=1}^n x_i}$$

heißt Lorenzkurve von X .

- u_j ist der Anteil der j kleinsten Merkmalsträger
- v_j der anteilige Beitrag dieser Einheiten zur Gesamtsumme $\sum_{i=1}^n x_i = \sum_{i=1}^n x(i)$
(„kumulierte relative Merkmalssumme“)

Bem. 4.4.

- a) Insbesondere bei größeren Datensätzen vereinfacht sich die Berechnung wesentlich, wenn man die relativen/absoluten Häufigkeiten f_1, \dots, f_k bzw. h_1, \dots, h_k der der Größe nach geordneten Merkmalsausprägungen $a_1 < a_2 < \dots < a_k$ benutzt. Dann ist für $j = 1, \dots, k$

$$u_j = \sum_{l=1}^j \frac{h_l}{n} = \sum_{l=1}^j f_l = F(a_j) \quad (4.1)$$

und

$$v_j = \frac{\sum_{l=1}^j h_l \cdot a_l}{\sum_{l=1}^k h_l \cdot a_l} = \frac{\sum_{l=1}^j f_l \cdot a_l}{\sum_{l=1}^k f_l \cdot a_l} \quad (4.2)$$

b) Ist bei klassierten Daten mit den Klassen $[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k]$ die Merkmalsverteilung in den Klassen nicht bekannt - und will man als Funktionswert dennoch eine einzelne Zahl -, so nimmt man wie beim arithmetischen Mittel als Approximation an, dass alle Ausprägungen in dieser Klasse auf die Klassenmitte $m_\ell = \frac{c_{\ell-1} + c_\ell}{2}$ fallen. Damit erhält man mit f_ℓ und h_ℓ , $\ell = 1, \dots, k$, als relative bzw. absolute Klassenhäufigkeiten und $a_\ell = m_\ell$, $\ell = 1, \dots, k$:

$$v_j = \frac{\sum_{\ell=1}^j h_\ell \cdot m_\ell}{\sum_{\ell=1}^k h_\ell \cdot m_\ell} = \frac{\sum_{\ell=1}^j f_\ell \cdot m_\ell}{\sum_{\ell=1}^k f_\ell \cdot m_\ell} . \quad (4.3)$$

Während normalerweise bei Lorenzkurven nur die Punkte $(0, 0), (u_1, v_1), \dots$ interpretierbar sind, interpretiert man bei klassierten Daten auch die linearen Zwischenstücke.

Bem. 4.5. [Zur Interpretation der Lorenzkurve]

Bsp. 4.6. Übungsaufgabe

Gegeben sei ein Land mit folgender klassierter Vermögensverteilung

Klasse	1	2	3
(c_{l-1}, c_l)	$[0;10)$	$[10;227.5)$	$[227.5; \infty)$
	arm	mittel	reich
f_l	$f_1 = 0.5$	$f_2 = 0.4$	$f_3 = 0.1$

mit $m_3 := 500$

Man bestimme die Lorenzkurve.

4.1.2 Der Gini-Koeffizient

Definition 4.7. [Gini-Koeffizient]

Gegeben sei die geordnete Urliste $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ eines verhältnisskalierten Merkmals X , das den Annahmen aus Bemerkung ?? genügt.

$$G := \frac{2 \cdot \sum_{i=1}^n i \cdot x_{(i)}}{n \sum_{i=1}^n x_i} - \frac{n+1}{n} \quad (4.4)$$

heißt *Gini-Koeffizient* von X und

$$G^* := \frac{n}{n-1} \cdot G \quad (4.5)$$

normierter Gini-Koeffizient (Lorenz-Münzner-Koeffizient) von X .

Bem. 4.8.

- Man kann durch die Herleitung über die Trapezformel zeigen (Toutenburg & Heumann (2009)), dass gilt:

$$\begin{aligned} G &= \frac{\text{Fläche zwischen Winkelhalbierender und Lorenzkurve}}{\text{Fläche zwischen Winkelhalbierender und Abszisse}} \\ &= 2 \cdot \text{Gesamtfläche} - \text{Fläche unter der Lorenzkurve} \\ &= 2 \cdot \text{Fläche zwischen Winkelhalbierender und Lorenzkurve} \end{aligned}$$

In der Literatur gibt es verschiedene äquivalente Formen, den Gini-Koeffizient zu berechnen.

-
- Betrachtet man wie in Bemerkung 4.3 die geordneten Ausprägungen $a_1 < a_2 < \dots < a_k$ mit den Häufigkeiten h_1, h_2, \dots, h_k , so gilt:

$$G = \frac{\sum_{l=1}^k (u_{l-1} + u_l) f_l \cdot a_l}{\sum_{l=1}^k f_l \cdot a_l} - 1 = \frac{\sum_{l=1}^k (u_{l-1} + u_l) h_l \cdot a_l}{\sum_{l=1}^k h_l \cdot a_l} - 1 = 1 - \sum_{l=1}^k f_l (v_{l-1} + v_l) \quad (4.6)$$

- Es gilt bei minimaler Konzentration $G = 0$ und bei maximaler Konzentration $G = \frac{n-1}{n}$;
- damit ist also $G^* = 0$ bei minimaler Konzentration und $G^* = 1$ bei maximaler Konzentration. (Ist n sehr groß, so ist $\frac{n-1}{n} \approx 1$, also $G^* \approx G$.)

Man veranschauliche und berechne in der Situation von Bsp. 4.5 den Ginikoeffizienten.

4.1.3 Quantilsbezogene relative Konzentrationsmessung

Oft stehen die Daten in einer anderen Form zur Verfügung:

Gegeben sind dann bestimmte Quantile (typischerweise Quartile, Quintile oder Dezile) und die Anteile des Merkmals, die auf das jeweilige Quantil entfallen.

Wie kann man dann immer noch die Lorenzkurve berechnen?

Bem. 4.9. („Quantilsbezogene Darstellung“)

Gegeben sei eine „Einteilung der Abszisse“

$$0 =: \alpha_0 < \alpha_1 < \dots < \alpha_l < \dots < \alpha_{q-1} < 1 =: \alpha_q$$

und die zugehörigen Quantile $x_{\alpha_l}, l = 1, \dots, q$ (, so, dass $\alpha_l \cdot n$ ganzzahlig ist. Ferner seien die Ausprägungen von X echt größer als 0 und die Quantile eindeutig).

Dann ergibt für $l = 1, \dots, q$

$$z_l^* := v_{\alpha_l \cdot n} - v_{(\alpha_{l-1} \cdot n) + 1}$$

denjenigen Anteil der Merkmalssumme, der auf Beobachtungen mit einer Merkmalsausprägung in $(x_{\alpha_{l-1}}, x_{\alpha_l}]$ entfällt.

Wählt man insbesondere die $\alpha_l, l = 0, \dots, q$, äquidistant, also $\alpha_l = l \cdot \frac{1}{q}$, so erhält man sozusagen *quantilsbezogene* Daten; $z_l^*, l = 1, \dots, q$, sei dann als *l-ter Quantilsanteil* bezeichnet.

Umgekehrt werden Daten dieser Form $(\alpha_l, z_l^*)_{l=1, \dots, q}$ mit z_l^* wie oben häufig verwendet, um Konzentrationsverhältnisse ohne Angabe der konkreten Merkmalsausprägungen zu charakterisieren. Die Kurve durch die Punkte (u_l^*, v_l^*) mit

$$u_l^* = \alpha_l \tag{4.7}$$

$$v_l^* = \sum_{r \leq l} z_r^* \tag{4.8}$$

für $l = 0, \dots, q$ werde als *induzierte Lorenzkurve* bezeichnet, und

$$G^* = 1 - \sum_{l=1}^q f_l^* (v_{l-1}^* + v_l^*) \tag{4.9}$$

mit

$$f_l^* := \alpha_l - \alpha_{l-1}, \quad l = 1, \dots, q \tag{4.10}$$

als *induzierter Gini-Koeffizient*.

Die induzierte Lorenzkurve ist wieder eine Lorenzkurve (zum Beispiel für das fiktive Merkmal X^* mit Ausprägungen $x_l^* = z_l^*/f_l^*$, $l = 1, \dots, q$, und Häufigkeitsverteilung f_l^* , $l = 1, \dots, q$). Ihr Graph verläuft nicht unterhalb der Graphes der Lorenzkurve des Ausgangsmerkmals X . ; beide Graphen schneiden sich in den Punkten $(u_l^*, v_l^*)_{l=1, \dots, q}$. Ist G der Ginikoeffizient von X , so gilt $G \geq G^*$.

		1. Bereich	2. Bereich	...	q -ter Bereich
$\alpha_l = u_l^*$	kumul. Häufigkeiten	α_1	α_2	...	1
f_l^*	Einzelhäufigkeiten	α_1	$\alpha_2 - \alpha_1$...	$1 - \alpha_{q-1}$
Z_l^*	Einzelanteile	z_1	z_2	...	z_q
v_l^*	Gesamtanteile	z_1	$z_1 + z_2$...	$z_1 + z_2 + \dots + z_q = 1$

Bsp. 4.10. *[Verteilung des Privatvermögens in Deutschland gemäß dem Entwurf des 4. Armuts- und Reichtumsberichts]*

Anteile	in der unteren Hälfte	in den 6. bis 9. Dezil	im obersten Dezil
1998	4 %	51 %	45 %
2003	3 %	48 %	49 %
2008	1 %	46 %	53 %

Bestimmen Sie die Lorenzkurven und den Gini-Koeffizient der Verteilung,

4.1.4 Einige weitere quantilsbasierte Maße

Insbesondere basierend auf der „natürlichen, äquidistanten Quantileinteilung“ $\alpha, 2 \cdot \alpha, 3 \cdot \alpha, \dots$ lassen sich direkt weitere relative Konzentrationsmaße definieren, die natürlich entsprechend verallgemeinert werden können.:

Robin-Hood-Index (maximaler Nivellierungssatz, Schutzkoeffizient, (z.B. Wagschal (1999, S.135ff))

- Wie viel müsste den Reichen weggenommen werden, um zu einer Konzentration von 0 zu kommen?
- Ermittle für jedes Quantil mit einem Anteil von höchstens α den Abstand seines Anteils zu α !
- Aufaddieren dieser Abstände liefert den Robin-Hood-Index. Dieser Anteil müsste verteilt werden, um zu einer gleichen Verteilung zu kommen!

Also mit der Notation

z_ℓ^* Anteil im ℓ -ten Quantil, $\ell = 1, \dots, q$

und mit

$$|a|_+ := \begin{cases} a & a > 0 \\ 0 & \text{sonst} \end{cases}$$

$$RHI = \sum_{\ell=1}^q |\alpha - z_\ell^*|_+ \left(= \sum_{\ell=1}^q |z_\ell^* - \alpha|_+ = \frac{1}{2} \sum_{\ell=1}^q |z_\ell^* - \alpha| \right) \quad (4.11)$$

Es gilt ferner mit v_ℓ^* gemäß (4.8)

$$RHI = \max_{\ell} (\ell \cdot \alpha - v_\ell^*); \quad (4.12)$$

Der RHI ist also der „maximale senkrechte Abstand“ zwischen der Winkelhalbierenden und der Lorenzkurve durch $(u_\ell, v_\ell)_\ell$.

Bem. 4.11. *Quantilverhältnisse*

Ein weiteres anschauliches Maß sind Quantilverhältnisse, etwa das

- z.B. *Dezilratio* $90 : 10 = \frac{x_{0.9}}{x_{0.1}}$, falls $x_{0.1} > 0$
- beim Einkommensvergleich: also um welchen Faktor ist der untere Wert der 10% Reichsten größer als der obere Wert der 10% Ärmsten
- Minimale Konzentration: alles in einem Punkt $x_{0.1} = x_{0.9}$
 \Rightarrow *Dezilratio* = 1,
- Umgekehrt Vorsicht: bei extremer Konzentration.
Die Maßzahl könnte für Entwicklungsländer eventuell problematisch sein,
- aber etwa für Einkommensverhältnisse in OECD-Ländern sehr anschauliches Maß.

Bsp. 4.12.

4.2 Absolute Konzentration

4.2.1 Vorbemerkungen: relative versus absoluter Konzentration

Vgl. Einleitung: Manche Autor(innen) verwenden für die „relative“ Konzentration den Begriff „Disparität“. Dann wird „absolute Konzentration“ meist schlicht als Konzentration bezeichnet.

4.2.2 Einige Maßzahlen der absoluten Konzentration

Definition 4.13.

Sei $0 \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ die geordnete Urliste eines verhältnisskalierten Merkmals mit $\sum_{i=1}^n x_i > 0$.

Mit

$$p(i) := \frac{x_{(i)}}{\sum_{j=1}^n x_j}$$

heißt

$$CR_g := \sum_{i=n-g+1}^n p(i)$$

Konzentrationsrate (vom Grade g).

Bsp. 4.14.

Zweitstimmenanteile polit. Parteien bei Bundestagswahlen 1949-2009

	1949	1953	1965	1972	1983	1994	2002	2005	2009	2013
CDU/CSU	31,0%	45,2%	47,6%	44,9%	48,8%	41,5%	38,5%	35,2%	33,8%	41,5%
SPD	29,2%	28,8%	39,3%	45,8%	38,2%	36,4%	38,5%	34,2%	23,0%	25,7%
FDP	11,9%	9,5%	9,5%	8,4%	7,0%	6,9%	7,4%	9,8%	14,6%	4,8%
Grüne	-	-	-	-	5,6%	7,3%	8,6%	8,1%	10,7%	8,4%
PDS/Die Linke	-	-	-	-	-	4,4%	4,0%	8,7%	11,9%	8,6%
Sonstige	27,9%	16,5%	3,6%	0,9%	0,4%	3,5%	3,0%	4,0%	7,0%	11,0%

Definition 4.15.

In der Situation von Def 1.12 heißt

$$H := \sum_{i=1}^n p_{(i)}^2$$

Herfindahl-Index. Die Größe $1 - H$ wird auch *Rae-Index* genannt.

In der Politikwissenschaft wird $\frac{1}{H}$ auch als *Anzahl der effektiven Parteien* bezeichnet.

Bsp. 4.18. [Herfindahl- und Rae-Index des deutschen Parteienwesens]

Bsp. 4.19. [„Durchschnittliche Fraktionalisierung von Parteiensystemen“]