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 1.1 und 1.2*

- X sei ein verhältnisskaliertes Merkmal (mit Urliste $x_1, \ldots, x_n$) und der Größe nach geordneten Ausprägungen $a_1, \ldots, < a_k$. und $h_1, \ldots, h_k$. Ferner seien $f_1, \ldots, f_k$ die zugehörigen absoluten bzw. relativen Häufigkeiten; die empirische Verteilungsfunktion werde mit $F_X(\cdot)$ oder $F(\cdot)$ bezeichnet.

- Zudem $x_i \geq 0$, für alle $i = 1, \ldots, n$ und $\sum_{i=1}^{n} x_i > 0$ (d.h mindestens ein Wert ist von Null verschieden)

- Betrachtet werden die der Größe nach geordneten Daten:

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

(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 $a_1 < a_2 < \ldots < a_k$ 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[1]) 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 groöem Anteil an der Gesamtsumme (z.B. Umsatz, Stimmanteile).

[1]Rinne, H (2003): Taschenbuch der Statistik. Frankfurt am Main.

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

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

und

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

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 $f_1, \ldots, f_k$ bzw. $h_1, \ldots, h_k$ der der Größe nach geordneten Merkmalsausprägungen $a_1 < a_2 <, \ldots, < a_k$ benutzt. Dann ist für $j = 1, \ldots, k$

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

und

$$v_j = \frac{\sum\limits_{l=1}^{j} h_l \cdot a_l}{\sum\limits_{l=1}^{k} h_l \cdot a_l} = \frac{\sum\limits_{l=1}^{j} f_l \cdot a_l}{\sum\limits_{l=1}^{k} f_l \cdot a_l} \tag{4.2}$$

b) Ist bei klassierten Daten mit den Klassen $[c_0, c_1), [c_1, c_2), \ldots, [c_{k-1}, c_k]$ 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 $m_l = \frac{c_{l-1}+c_l}{2}$ fallen. Damit erhält man mit f_ℓ und h_ℓ, $\ell = 1, \ldots, k$, als relative bzw. absolute Klassenhäufigkeiten und $a_\ell = m_\ell$, $\ell = 1, \ldots, k$:

$$v_j = \frac{\sum_{l=1}^{j} h_l \cdot m_l}{\sum_{l=1}^{k} h_l \cdot m_l} = \frac{\sum_{\ell=1}^{j} f_\ell \cdot m_\ell}{\sum_{\ell=1}^{k} f_\ell \cdot m_\ell} \quad . \tag{4.3}$$

Während normalerweise bei Lorenzkurven nur die Punkte $(0,0), (u_1, v_1), \ldots$ 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 Vermägensverteilung

Klasse	1	2	3
$(c_{\ell-1}, c_\ell)$	[0;10)	[10;227.5]	[227.5; ∞)
	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 $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ eines verhältnisskalierten Merkmals X, das den Annahmen aus Bemerkung 4.1 genügt.

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

heißt <u>Gini-Koeffizient</u> von X und

$$G^* := \frac{n}{n-1} \cdot G$$

<u>normierter Gini-Koeffizient</u> (Lorenz-Münzner-Koeffizient) von X.

Bem. 4.8.

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

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

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

- Betrachtet man die geordneten Ausprägungen $a_1 < a_2 < \ldots < a_k$ mit den Häufigkeiten $h_1, h_2, \ldots, h_k$, so gilt mit (4.2) und (**??**):

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

- Es gilt bei minimaler Konzentration $G=0$ und bei maximaler Konzentration $G=\dfrac{n-1}{n}$;
- 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 $G^*\approx G$.)

Bsp. 4.9. [Konzentrationsmessung] *Fortsetzung von Bsp. 4.6*

Man veranschauliche und berechne in der Situation von Bsp. 4.6 den Ginikoeffizienten.

4.1.3 Quantilsbezogene relative Konzentrationsmessung

Oft stehen die Daten in einer anderen Form zur Verfügung:
Gegeben sind dann N Quantile (typischerweise Quartile, Quintile oder Dezile) und die Anteile des Merkmals, die auf das l-te jeweilige Quantil $l = 1, ..., z$ entfallen.

Wie kann man dann immer noch die Lorenzkurve berechnen?

Bsp. 4.10. *[Zur internationalen Einkommensverteilung]*

Bem. 4.11. *(„Quantilsbezogene Darstellung")*

Gegeben sei eine „Einteilung der Abszisse"

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

und die zugehörigen Quantile $x_{\alpha_l}, l = 1, \ldots, q$ (, 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 $l = 1, \ldots, q$

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

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

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

Umgekehrt werden Daten dieser Form $(\alpha_l, z_l)_{l=1,\ldots,q}$ 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

$$u_l^* = \alpha_l$$

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

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

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

mit

$$f_l^* := \alpha_l - \alpha_{l-1}, l = 1, \ldots, q$$

als *induzierter Gini-Koeffizient*.

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

4.1.4 Einige weitere quantilsbasierte Maße

Insbesondere basierend auf der „natürlichen, äquidistanten Quantilseinteilung“ lassen sich weitere relative Konzentrationsmaße definieren:

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

$$z_l \quad \text{Anteil im l-ten Quantil, } \ell = 1, \ldots, q$$

und mit

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

$$RHI = \sum_{\ell=1}^{q} |\alpha - z_\ell|_+ \left(= \sum_{\ell=1}^{q} |z_\ell - \alpha|_+ = \frac{1}{2} \sum_{\ell=1}^{q} |z_\ell - \alpha| \right) \tag{4.4}$$

Es gilt ferner

$$RHI = \max_\ell (\ell \cdot \alpha - v_\ell); \tag{4.5}$$

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

Bem. 4.12. *Quantilverhältnisse*

Ein weiteres anschauliches Maß sind Quantilsverhältnisse, 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. 4.13.

Dezilverhältnisse 90:10 des Einkommens von Vollbeschäftigten im internationalen Vergleich $\rightarrow$ Wagschal (1999, S.138)

Norwegen	Italien
Schweden	Neuseeland
Dänemark	Japan
Belgien	Frankreich
Finnland	Großbritannien
Deutschland	Österreich
Niederlande	Kanada
Schweiz	Portugal
Australien	USA

4.2 Absolute Konzentration

4.2.1 Vorbemerkungen

4.2.2 Einige Maßzahlen der absoluten Konzentration

Definition 4.14.

Sei $0 \leq x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ 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. 4.15.

Zweitstimmenanteile polit. Parteien bei Bundestagswahlen 1949-2009

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

Definition 4.16.

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. 4.19. [Herfindahl- und Rae-Index des deutschen Parteienwesens]

Bsp. 4.20. [„Durchschnittliche Fraktionalisierung von Parteiensystemen“]

(Wagschal S. 145 (1999), Riedl (2011))

Land	durchschnittl. Rae-Index 1945-93
USA	0.53
Österreich	0.60
Großbritannien	0.62
Deutschland	0.64
Schweiz	0.71
Italien	0.75
Frankreich	0.79
Niederlande	0.79
Finnland	0.82