9. Konzentrationsmaße

Ich stehe Statistiken etwas skeptisch gegenüber. Denn laut Statistik haben ein Millionär und ein armer Kerl jeder eine halbe Million.

Franklin D. Roosevelt

9.1. Begriff der Konzentration und Definitionen

Die statistische Analyse von Konzentrationen macht quantitative (zahlenmäßige) Aussagen von z.B. folgenden Sachverhalten:

- Verteilung der Haushalts-Einkommen in Deutschland
- Aufteilung der Kfz-Produktion auf die Autofirmen
- Häufigkeit, mit der in diesem Text bestimmte Worte vorkommen: ("und", "die", "Konzentrationsmerkmale" kommen z.B. sicher konzentriert vor ;-)

All dem ist gemeinsam, dass statistische Merkmale untersucht werden, bei denen eine Summenbildung $möglich$ und $sinnvoll$ ist. Man nennt solche Merkmale auch **Konzentrationsmerkmale** oder **extensive Merkmale** (im Gegensatz zu intensiven Merkmalen).

Neben den bisher schon verwendeten Anteilen an der Summe der Merkmals$träger$: Relative Häufigkeit f_i und relative Summenhäufigkeit F_i sind bei Konzentrationsmerkmalen auch Anteile an der Summe der $Merkmale$, der sogenannten **Merkmalssumme**, sinnvoll: Der Anteil p_i eines Merkmalsträgers oder einer Klasse an der Merkmalssumme sowie der kumulierte Anteil P_i.

9.1(b) Definitionen

- Die **Merkmalssumme**

$$M = \sum_{i=1}^{n} x_i$$

Frage: Kann man die Merkmalssumme auch durch das arithmetische Mittel des Merkmals ausdrücken? Wie sieht die Merkmalssumme für klassierte Daten aus?

- Der **Anteil** eines Merkmalsträgers an der Merkmalssumme,

$$p_i = \frac{x_i}{M}$$

Frage: Wie sieht der Merkmalssummen-Anteil für klassierte Daten aus?

- Der **Kumulierter Merkmalssummen-Anteil**

$$P_i = \sum_{i'=1}^{i} p_{i'}$$

Hinweis: Bei manchen Maßen dieses Kapitels müssen die Merkmalsträger nach steigendem Wert des Konzentrationsmerkmals geordnete sein. Also:

Die "Kleinsten" und "Ärmsten" kommen zuerst dran!

9.1(c) Verständnisfragen

1. Nennen Sie bei folgenden Sachverhalten jeweils die Merkmale, Merkmalsträger, Merkmalssummen M und Merkmalsträgersummen:

   - Energieverbrauch pro Kopf
   - Einwohnerzahlen und Flächen der Länder dieser Welt
   - Umsatz, Beschäftigtenzahl und Produktionsrate von Autofirmen (warum wäre das Merkmal "Gewinn" problematisch?)
   - Länge und mittleres Verkehrsaufkommen des Streckennetzes der Straßen, klasiert nach erster Ordnung (BAB) bis vierter Ordnung (Kreisstraßen)
   - Häufigkeit, mit der in diesem Text bestimmte Worte vorkommen

2. Warum müssen Konzentrationsmerkmale kardinalskaliert (sogar verhältnisskaliert) sowie nichtnegativ sein?

3. Sind folgende Merkmale Konzentrationsmerkmale oder nicht?

   - Zahl x_i der verkauften Autos der Marke i
   - Familienstand: ledig,verheiratet,geschieden, ...
   - Körpergröße X der Bürger in den USA
   - tägliche Niederschlagsmenge X

9.1(d) Absolute *vs.* relative Konzentration

Man unterscheidet zwei Arten von Konzentrationen:

- **Absoluten Konzentration**: Konzentration, die durch *Ausscheiden* von Merkmalsträgern entsteht. Diese ist hoch, wenn ein großer Anteil der Merkmalssumme auf eine kleine *Zahl* von Merkmalsträgern entfällt ("*wenige* Firmen machen einen Großteil des Umsatzes"),

- **Relative Konzentration**, auch **Disparität** genannt. Konzentration, die durch das Wachsen der Großen und Schrumpfen der Kleinen entsteht, also die *Ungleichheit* erhöht. Diese ist hoch, wenn ein großer Anteil der Merkmalssumme auf einen kleinen *Anteil* von Merkmalsträgern entfällt ("ein *geringer Prozentsatz* der Firmen macht einen Großteil des Umsatzes").

Ferner kann man Konzentration als *Zustand* und als *Prozess* auffassen.

Fragen:

- Die deutschen Autohersteller sind im Wesentlichen VW, Daimler, BMW und Porsche. Angenommen, die beiden Kleinsten (BMW und Porsche) fusionieren. Wie ändern sich qualitativ die beiden Konzentrationsarten?
- Wie sieht es aus, wenn die beiden Größten fusionierten?
- Erläutern Sie Konzentration als Zustand und als Prozess am Beispiel der Einkommensunterschiede.

9.2. Maßzahlen der absoluten Konzentration

Die einfachste Maßzahl ist der

Herfindahl-Index: $K_H = \sum_{i=1}^{N} p_i^2$

- Im Monopolmarkt $(p_1, \cdots, p_{n-1} = 0, p_n = 1)$ gilt $H = 1$
- Bei völliger Gleichverteilung der Anteile $(p_i = 1/n)$ gilt $H = 1/n$.
- Firmen mit sehr geringen Umsatzanteilen beeinflussen den Index kaum, auch wenn es sehr viele sind.

Der Herfindahl-Index hängt mit dem Variationskoeffizient zusammen:

$$K_H = \frac{1}{n}(V^2 + 1) = \frac{1}{n}\left(\frac{s^2}{\bar{x}^2} + 1\right)$$

Aufgabe: Leiten Sie diesen Zusammenhang her!

9.2. Maßzahlen der absoluten Konzentration II

Mathematisch und auch intuitiv bessere Eigenschaften als der Herfindahl-Index hat der

Exponentialindex $K_E = \prod_{i=1}^{n} p_i^{p_i} = p_1^{p_1} p_2^{p_2} \cdots p_n^{p_n}.$

Er wird direkt aus dem sehr mächtigen Konzept des **Shannon'schen Informationsgehalts** (negative Entropie) einer Verteilung hergeleitet: $H = -\sum_{i=1}^{n} p_i \log_2 p_i$ gibt die Information in Bits an, die im Mittel nötig ist, um eine Firma aus der Verteilung herauszupicken: $H = 0$ beim absoluten Monopol, $H = 1$ (Bit) bei zwei Firmen mit je 50% Marktanteil, $H = 2$ bei vier Firmen mit je 25% Marktanteil, etc. *H gibt die mittlere Zahl der nötigen Ja-Nein-Fragen bei optimaler Fragestrategie an.*

Um das Ganze nun auf den Wertebereich von 0 (keinerlei Konzentration, $H \to \infty$) bis 1 (Monopol, $H = 0$) zu bringen, kommt H in den Exponenten:

$$\begin{aligned} K_E &= 2^{-H} = 2^{\sum_{i=1}^{n} p_i \log_2 p_i} \\ &= \prod_{i=1}^{n} 2^{p_i \log_2 p_i} = \prod_{i=1}^{n} (2^{\log_2 p_i})^{p_i} = \prod_{i=1}^{n} p_i^{p_i}. \end{aligned}$$

Aufgabe: Bei einem Fragespiel, bei dem nur Ja-Nein Fragen gestellt werden dürfen, soll von Leuten aus dem Publikum das Lieblings-Musikstück ermittelt werden. Der Interviewer mit der geschicktesten Fragestrategie brauchte im Mittel 5.5 Ja-Nein-Fragen dafür. Wie hoch ist der Exponentialindex der Konzentration der Lieblingsstücke mindestens?

9.2. Maßzahlen der absoluten Konzentration III

Beispiel: Verteilung des Gesamtumsatzes auf verschiedene Firmen

Firma	v_1	v_2	v_3	v_4	v_5	v_6	v_7
1	2	20	2	10	30	20	20
2	1	10	2	10	10	15	10
3			1	10	10	10	1
4			1		10	5	1
5							1
6							1

Wie groß ist bei den sieben Verteilungen $v_1, \cdots, v_7$ jeweils der Herfindahl-Index und des Exponentialindex? Diskutieren Sie die Eigenschaften dieser Indices!

9.3. Analyse der relativen Konzentration

Das wichtigste Analysemittel ist die **Lorenzkurve**, bei der der kumulierte Anteil P_i der $Merkmalssumme$ als Funktion der relative Summenhäufigkeit F_i, d.h., des kumulierten Anteils der $Merkmalsträgersumme$ aufgetragen wird. Also konkret:

Gegeben ist eine nach Größe geordnete Urliste eines Konzentrationsmerkmals X, also $0 \leq x_1 \leq \cdots \leq x_n$: Dann ist die **Lorenzkurve** durch Verbinden der $(n+1)$ Punkte (F_i, P_i), $i = 0, \cdots, n$, gegeben, wobei

$$F_i = \frac{i}{n},$$

$$P_i = \sum_{j=1}^{i} p_j = \frac{\sum_{j=1}^{i} x_j}{M}, \quad P_0 := 0.$$

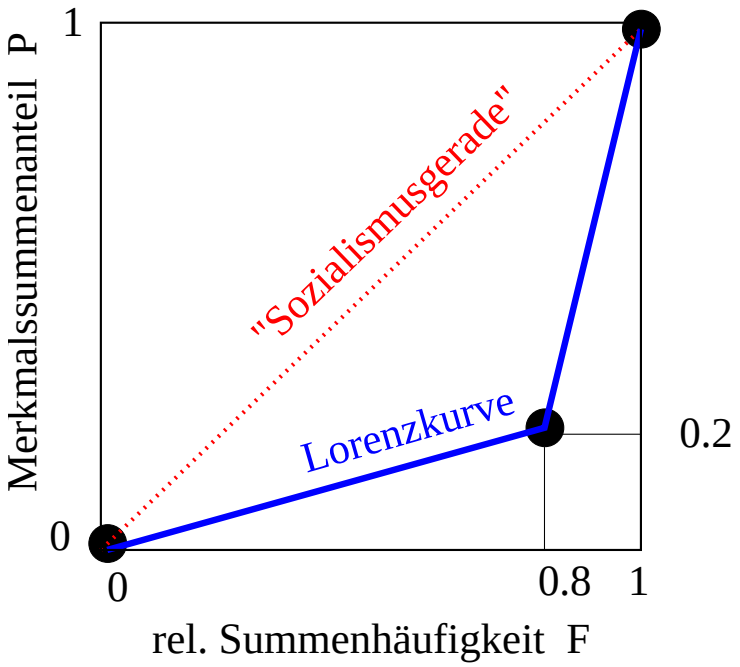
Wichtig bei der Erstellung der Lorenzkurve:

Die Kleinsten und Ärmsten kommen zuerst!

9.3(b) Lorenzkurve für klassierte Daten

Für klassierte Daten muss man einfach die relative Summenhäufigkeit und den kumulierten Anteil der Merkmalssumme durch die Klassen ausdrücken. Eingesetzt ergibt sich

$$F_k = \sum_{j=1}^{k} f_j = \sum_{j=1}^{k} \frac{h_j}{n}, \qquad P_k = \sum_{j=1}^{k} p_j = \frac{\sum_{j=1}^{k} x_j^* h_j}{M}$$



Beispiel:
Die "20−80−Regel"

(20% der Leute haben
80% des Vermögens,
für die restlichen 80%
verbleiben nur noch
20%

Arbeitstabelle:

Klasse	f_k	F_k	p_k	P_k
Arme				
Reiche				

9.3(c) Fragen zur Lorenzkurve

Verständnisfragen

- Warum können beide Koordinaten der Lorenzkurve nur Werte von 0 bis 1 annehmen?
- Warum liegt die Lorenzkurve nie oberhalb der Diagonalen (also nie $P_i > F_i$)?
- Wie würde die Lorenzkurve aussehen, wenn (i) alle das gleiche besitzen, (ii) einer alles besitzt? [1]

Übungsaufgaben

- Zu 10 großen Firmen, die einen Markt gleichmäßig unter sich aufteilten, kommen nun 10 weitere hinzu, die zunächst noch keinen Umsatz machen. Wie ändert sich der Herfindahl-Index und die Lorenzkurve der Umsatzaufteilung?
- Berechnen Sie für einige der Verteilungen aus Kap. 9.2 die Lorenzkurve
- Nach der Weltbank waren im Jahre 1999 die reichsten 20% der Weltbevölkerung im Mittel 12 mal so reich wie die ärmsten 20% (gemessen in Kaufkraft-Einheiten) sowie die "Mittelklasse" (alle anderen). im Mittel viermal so reich. Zeichnen Sie die "Lorenzkurve

[1] Ein Artikel in der ZEIT bringt die Lorenzkurven-Analyse von Vermögensverteilungen mit dem Titel "im weiten Bogen um die Gerechtigkeit" auf den Punkt ...

9.3(d) Gini-Koeffizient

Der **Gini-Koeffizient**

$$
\begin{aligned}
G &= \frac{\text{Fläche A zwischen Diagonalen und Lorenzkurve}}{\text{Fläche unter der Diagonalen}} \\
&= 1 - \sum_{i=1}^{n} (P_i + P_{i-1}) \frac{1}{n} \quad \text{(Urliste)} \\
&= 1 - \sum_{k=1}^{K} (P_k + P_{k-1}) f_k \quad \text{(klassierte Daten)}
\end{aligned}
$$

ist ein quantitatives Maß für die relative Konzentration.

- Leiten Sie die math. Ausdrücke für den Gini-Koeffizienten her
- Warum ist immer $0 \leq G \leq 1$?
- Was bedeutet $G = 0$ und $G \approx 1$?
- Berechnen Sie den Gini-Koeffizienten für einige der Umsatzverteilungen von Kap. 9.2. Warum ist sofort klar, dass G für die ersten drei Verteilungen $\{2, 1\}$, $\{20, 10\}$ und $\{20, 20, 10, 10\}$ derselbe ist?
- Firmengrößen gehorchen häufig annähernd einer "abgeschnittenen Pareto-Verteilung": Bei z.B. 10 000 Firmen hat die größte einen Umsatzanteil von 20%, die Nummern 2-10 zusammen 20%, sowie die Firmen 11-100, 101-1000 und 1001-10000 zusammen jeweils weitere 20%. Wie groß ist G?
- Ermitteln Sie den "Ginikoeffizienten des Weltreichtums" (vgl. letzte Frage zur Lorenzkurve).

9.3(e) Kontinuierliche Daten

Bei einer sehr großen Zahl n der zur Merkmalssumme M beitragenden statistischen Einheiten, wie die Beiträge

- der Einkommen zum Gesamteinkommen
- der Umsätze zum Gesamtumsatz eines Sektors,
- der einzelnen Dateien an der gesamten Speicherbelegung eines Computers,

ist es sinnvoll, den kumulierten Anteil F der Merkmalsträger sowie den kumulierten Anteil P an der Merkmalssumme als stetige Größe anzusehen. Dies entspricht dem Grenzfall, dass man bei klassierten Daten die Klassenbreite $d_k = \Delta x_k = x_k^o - x_k^u$ gegen Null und die Klassenzahl gegen Unendlich gehen lässt.

Mit der bereits bei der Analyse klassierter Daten vorgestellten Dichte $f(x_k^*) = f_k^D = f_k/\Delta x_k$ erhält man für die Merkmalssumme

$$M = n\bar{x} = n\sum_k x_k^* f_k = n\sum_k x_k^* f(x_k^*)\Delta x_k$$

und damit nach der Grenzwertbildung $\sum_k \Delta x_k = \int dx$, $x_k^* \to x$ ein Integral:

$$M = n\int_0^\infty x f(x)\ \mathrm{d}x.$$

9.3(e) Kontinuierliche Daten II

Der kumulierte Anteil an der Merkmalssumme ergibt sich analog, wenn man bis zur Klasse k summiert bzw. bis zum Wert $x_k^* = x$ integriert (man beachte, dass die Variable x', über die integriert wird, nicht dieselbe wie die der Integrationsgrenze x sein darf):

$$P(x) = \frac{n}{M} \int_0^x x' f(x') \ \mathrm{d}x'.$$

Der kumulierte Anteil der Merkmalsträger ergibt sich einfach als Umkehrung der Definition $f(x) = dF/dx$ der Dichtefunktion:

$$F(x) = \int_0^x f(x') \ \mathrm{d}x'.$$

Schließlich ergibt sich mit $P_k + P_{k-1} \to 2P(x)$ der Gini-Koefizient:

$$G = 1 - 2 \int_0^\infty P(x) f(x) \ \mathrm{d}x.$$

Verständnisfrage: Warum ist hier das Prinzip "die Kleinsten und Ärmsten zuerst" automatisch erfüllt?

Aufgabe: Die Einkommen amerikanischer Bürger gehorchen in guter Näherung der Dichtefunktion

$$f(x) = \lambda e^{-\lambda x}$$

mit $1/\lambda = 20\ 000$ \$. Berechnen Sie Lorenzkurve und Gini-Koeffizienten.

9.3(f) Diskussion des Gini-Koeffizienten

In einigen Fällen verhält sich G nicht, wie man es erwarten sollte:

(a) Den Markt für Betriebssysteme in Europa bestreiten zu 80% Firma M und je 10% die Firmen A und UL. Verpackte Lebensmittel werden zu 80% von 10 multinationalen Firmen, zu 10% von 5 weiteren Firmen und zu 10% von 15 kleineren Nischenfirmen geliefert. Auf welchem Markt ist das Konzentrationsmaß "Gini-Koeffizient" höher? Was ist somit die offensichtliche Schwäche des Gini-Koeffizienten?

(b) Einkommensverteilungen sind üblicherweise nur durch die Einkommenssteuerstatistik erfassbar. Setzt man nun die Mindeststeuergrenze nach unten, ändert sich die Lorenzkurve nach links-unten und G wird größer, obwohl sich am Sachverhalt nichts ändert.

(c) Unterschiedliche Arten von Ungleichheiten können denselben Gini-Koeffizient haben z.B.
   (i) 10% der Leute besitzen 50% des Vermögens ("Nur Reiche, keine Armen"),
   (ii) 50% der Leute besitzen nur 10% des Vermögens ("nur Arme, keine Reichen")

Merke:

Konzentrationsindices sind i.A. nur zum *Vergleich* ähnlicher Sachverhalte geeignet. Ein absolutes Maß können sie nicht darstellen!