DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

Nr. 1/99

Disparitätsmessung aus klassierten Daten mittels Schätzung von entropiemaximalen Dichtefunktionen

von

André Lucas

März 1999

DISKUSSIONSBEITRÄGE ZUR STATISTIK UND ÖKONOMETRIE

SEMINAR FÜR WIRTSCHAFTS- UND SOZIALSTATISTIK
UNIVERSITÄT ZU KÖLN

Albertus-Magnus-Platz, D-50923 Köln, Deutschland

DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

Nr. 1/99

Disparitätsmessung aus klassierten Daten mittels Schätzung von entropiemaximalen Dichtefunktionen

von

André Lucas*

März 1999

Zusammenfassung: Standardmethoden zur Schätzung von Disparitätsmaßen aus klassierten Daten basieren entweder auf der Bestimmung von Schranken, die den wahren Wert des jeweiligen Disparitätsmaßes einschließen (nichtparametrischer Ansatz) oder aber auf Annahmen bezüglich der den Daten zugrunde liegenden Verteilung, deren Parameter geschätzt werden müssen (parametrischer Ansatz). Die Parameterschätzung kann je nach angenommener Verteilung numerisch aufwendig sein und es ist nicht in jedem Fall gesichert, dass diese Verteilung eine gute Anpassung an die Daten darstellt. Die Bestimmung der Schranken ist hingegen nur dann sinnvoll, wenn diese nahe genug beieinander liegen (dies ist zumeist nur bei Vorliegen einer größeren Anzahl von Klassen der Fall). In diesem Beitrag wird die Schätzung von Disparitätsmaßen mittels Bestimmung von entropiemaximalen Dichtefunktionen dargestellt. Dabei wird in jeder Klasse die Entropie der geschätzten Dichtefunktion maximiert. Die durchgeführte Simulationsstudie bestätigt eine verbesserte Schätzung bei einem akzeptablen numerischen Aufwand auch bei einer kleinen Klassenanzahl.

Title: Inequality Measurement from Grouped Data by Maximum–Entropy Density Estimation

Abstract: Standard methods of inequality measurement from grouped data are usually based on the determination of bounds which enclose the true value of the inequality measure (nonparametric method), or on assumptions on the data generating distribution, for which it is necessary to estimate the parameters (parametric method). Estimating parameters for some distributions can cause great numerical efforts and it is not for sure, that the chosen distribution fits the data adequatly. On the other hand, the determination of bounds is only useful and practical if these bounds are close enough (this is usually the case only, if the data is divided in many groups). The aim of this paper is to present a method for estimating inequality measures, based on the estimation of distribution with the maximum–entropy method. With given data, the entropy of the estimated distribution in each group is maximized. Simulations confirm that this method is increasing the accuracy of the estimation, even with little number of groups and with an acceptable numerical effort.

Keywords: Maximum–entropy; density estimation; inequality measures; grouped data

JEL classification: C13, D31, D63

Correspondence to: André Lucas, Seminar für Wirtschafts- und Sozialstatistik, Universität zu Köln, 50923 Köln, Germany, Tel: +49-221-4704129, Fax: +49-221-4705074, email: lucas@wiso.uni-koeln.de

*Seminar für Wirtschafts- und Sozialstatistik, Universität zu Köln, 50923 Köln, Germany, Tel.: +49-221-470 4129, Fax: +49-221-470 5074, e-mail: lucas@wiso.uni-koeln.de

1 Einleitung

Dieser Beitrag beschäftigt sich mit der Bestimmung von relativer Konzentration (Disparität) für den Fall, daß keine Einzeldaten, sondern bereits in Klassen eingeteilte Daten vorliegen. Speziell untersucht wird hier die Eignung des 'Maximum-Entropy Density Estimation' (ME) Konzepts für die Bestimmung von Schätzwerten für ausgewählte Disparitätsmaße.

Zunächst kann man sich Gründe für das Vorliegen von klassierten Daten überlegen, die es notwendig erscheinen lassen, nicht nur Probleme zu untersuchen bei denen Einzelwerte vorliegen:

- Die Beschaffung der Primärdaten bzw. Einzeldaten ist zu kostenintensiv bzw. der aus diesen Daten gewonnene Informationszugewinn rechtfertigt nicht die für die Beschaffung notwendigen Kosten, so daß wenn möglich auf Sekundär–Material zurückgegriffen wird.
- Bei Untersuchungen z.B. der Einkommens- oder Vermögensverteilung vergangener Zeiten, in denen es noch keine Datenverarbeitung gab bzw. die Rohdaten nach der Erhebung nicht einzeln archiviert wurden oder werden konnten, besteht überhaupt nicht erst die Möglichkeit, Berechnungen mit Einzeldaten durchzuführen, da die Daten nicht nachträglich erhoben werden können.
- Aufgrund der Vorteile von standardisierten Befragungen[1] durch Marktforschungsinstitute (Interview über Fragebögen, in denen mehrere Antwortmöglichkeiten bereits vorgegeben sind, bzw. in denen sich der oder die Befragte in bestimmte Merkmalsausprägungs–Klassen einordnen muß), liegen Informationen aus Marktuntersuchungen oftmals in klassierter Form vor.

Problematisch ist nun aber die Schätzung der unbekannten Verteilungsfunktion mithilfe derer man den Wert des jeweiligen Disparitätsmaßes bestimmen möchte. Durch die Komprimierung von zumeist Tausenden von Einzeldaten auf einige wenige Kenndaten mittels Klassierung entstehen logischerweise Informationsverluste. Innerhalb der Klassen ist die Art der Verteilung unbekannt. Neben den absoluten oder relativen Klassenhäufigkeiten sind zusätzlich, wenn überhaupt, zumeist nur der Mittelwert, die bedingten Mittelwerte in den einzelnen Klassen oder die Varianz bekannt. Dieser Informationsverlust macht es auch unmöglich, die Ungleichheit in einer Klasse genau zu bestimmen.

Dieser Beitrag ist wie folgt aufgebaut: In Abschnitt 2 wird kurz das ME–Konzept sowie die Herleitung von entropiemaximalen Dichtefunktionen dargelegt und erläutert, wie sich diese Methode für Dichteschätzungen bei Vorliegen von klassierten Daten und bestimmten Informationen über die wahre Verteilung nutzen läßt. In Abschnitt 3 werden die Formeln für vier Disparitätsmaße (Gini–Koeffizient, Theil–Maß, Pietra–Maß, Logarithmische Varianz) und für bestimmte Informationssituationen hergeleitet, mithilfe derer man die Werte der Disparitätsmaße direkt berechnen kann. In Abschnitt 4 werden dann in gebotener Kürze andere Verfahren vorgestellt, die zur Bestimmung der Werte von Disparitätsmaßen aus klassierten Daten eingesetzt werden können. Abschnitt 5 beinhaltet abschließend eine vergleichende Simulationsstudie, wobei Einkommensdaten des SOEP (**S**ozio–**Oe**konomisches **P**anel) benutzt werden.

1) Vorteile: Vollständigkeit, Vergleichbarkeit, leichtere Quantifizierung der Ergebnisse.

2 Das Prinzip der Entropie in der Statistik

Der in der Statistik benutzte Begriff der 'Entropie' wird aus der Informationstheorie übernommen. Hier ist die Entropie ein Maß für den erwarteten Informationsgehalt einer Nachricht.

Um den Begriff ´Informationsgehalt´ zu verdeutlichen, kann man sich z.B. ein unfaires Münzwurfspiel mit folgenden Wahrscheinlichkeiten vorstellen:

$$P(\text{"Kopf"}) = 0.1 \qquad P(\text{"Zahl"}) = 0.9 \quad .$$

Erhält man mittels einer Nachricht Informationen über den Ausgang des Spieles, so kann man sich überlegen, daß sich bei Eintritt des Ereignisses ´Zahl´ die Überraschung hierüber in Grenzen halten wird, da man aufgrund der hohen Eintrittswahrscheinlichkeit davon ausgehen konnte, daß das Ereignis ´Zahl´ eintritt. Diese Nachricht beinhaltet also wenig neue Information. Erfährt man hingegen mittels dieser Nachricht, daß das Ereignis ´Kopf´ eingetreten ist, so enthält diese Nachricht sehr viel Information, da aufgrund der geringen Eintrittswahrscheinlichkeit nicht verstärkt mit dem Eintritt dieses Ereignisses zu rechnen war.

Je unwahrscheinlicher das Eintreten eines Ereignisses A vor Eintreffen der Nachricht über das Eintreten dieses Ereignisses war ($p = P(A) \to 0$), desto höher ist der Informationsgehalt dieser Nachricht. Zur Bestimmung des Informationsgehaltes einer Nachricht wählt man dementsprechend eine in Abhängigkeit von p streng monoton fallende Funktion. Als vorteilhaft erweist sich die negative Logarithmus-Funktion

$$h(p) = -\log(p),$$

da sie unter anderem Additivität bei stochastisch unabhängigen Ereignissen gewährleistet (eine axiomatische Begründung für diese Wahl findet sich in THEIL (1967), S. 6 ff.) [2].

Möchte man vor dem Eintreffen der Nachricht den zu erwartenden Informationsgehalt einer Nachricht ermitteln, muß man den Erwartungswert des Informationsgehaltes aller möglichen Ereignisse berechnen, über deren Eintreten man mittels einer Nachricht Kenntnis erlangen kann.

Gegeben seien nun sich gegenseitig ausschließende Ereignisse $A_1, A_2, \ldots, A_n$ mit ihren zugehörigen Eintrittswahrscheinlichkeiten $p_1, p_2, \ldots, p_n$, wobei gilt: $h(p_i) = -\log(p_i)$. Unbekannt sei, welches Ereignis eintritt. Es ist nun möglich, den *Erwartungswert des Informationsgehalts einer Nachricht* zu berechnen, die darüber Auskunft gibt, welches Ereignis A_i eingetreten ist (vgl. THEIL (1967), S. 24):

$$H = E(h(p_i)) = \sum_{i=1}^{n} p_i \cdot h(p_i) = -\sum_{i=1}^{n} p_i \cdot \log(p_i) \quad \text{mit} \quad 0 \le H \le \log(n).$$

Der *minimale* Wert von H wird erreicht für $p_i = 1$ und $p_j = 0$ für $j = 1, \ldots, n$ und $j \neq i$ (vollkommene Gewissheit, minimale Information), der *maximale* Wert von H für $p_i = \frac{1}{n}$

2) Seien A_1 und A_2 zwei stochastisch unabhängige Ereignisse mit den zugehörigen Eintrittswahrscheinlichkeiten p_1 und p_2, so ist $p_1 p_2$ die Wahrscheinlichkeit für das gemeinsame Auftreten von A_1 und A_2. Dann gilt:
$h(p_1 p_2)$ = Informationgehalt, wenn beide Ereignisse eintreten
$= -\log(p_1 p_2) = -\log(p_1) - \log(p_2) = h(p_1) + h(p_2)$.

für $i = 1, \ldots, n$ (maximale Ungewissheit, maximale Information). Diese Größe wurde 1948 von Claude E. Shannon als Entropie in die Informationstheorie eingeführt (vgl. SHANNON (1948), S. 19). Analog hierzu definiert man das Entropie–Maß für den stetigen Fall mittels der Dichtefunktion einer Zufallsvariablen X:

$$H = -E(\log f(x)) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

Der Erwartungswert H existiert jedoch nur dann, wenn das Integral absolut konvergiert.

2.1 Entropiemaximale Dichtefunktionen

Mithilfe dieses Entropie–Maßes ist es nun möglich, die Dichte einer Zufallsvariablen zu schätzen. Man kann eine Familie von entropiemaximalen Dichtefunktionen finden, bei der dann u.a. über die Wahl der Werte mehrerer Parameter, eine Anpassung der Dichten an die jeweiligen vorhandenen Informationen erfolgt.

Die Parameter (und damit die Dichte) sollen so angepasst werden, daß der Erwartungswert des Informationsgehaltes (die Entropie) der vorhandenen Nachrichten (also die Informationen über die wahre Verteilung: Z.B. Momente und Häufigkeiten) maximiert wird.

Die Entropie bezüglich einer Dichte $f(x)$ ist nun genau dann maximal, wenn die Dichte $f(x)$ die Form

$$f(x) = e^{\theta_0 + \theta_1 h_1(x) + \ldots + \theta_n h_n(x)}$$

aufweist und wenn die Parameterwerte $\theta_0, \ldots, \theta_n$ existieren (s.a. S. 5), so daß die folgenden zwei Bedingungen eingehalten werden:

1) Seien $h_1(x), \ldots, h_n(x)$ integrierbare Funktionen auf $[a, b]$, so daß folgende Bedingungen für gegebene Konstanten $g_1, \ldots, g_n$ (z.B. Momente, Häufigkeiten) erfüllt sind (*Erhaltungsbedingungen*):

$$\int_a^b h_i(x) f(x) dx = g_i \qquad i = 1, \ldots, n.$$

2) Es muß zudem gewährleistet sein, daß

$$\begin{aligned} f(x) &> 0 \quad \text{für } x \in [a, b] \\ f(x) &= 0 \quad \text{sonst.} \end{aligned}$$

Ein Beweis hierfür findet sich u.a. in KAGAN/LINNIK/RAO (1973), S. 408-409 und RAO (1973), S. 59.

2.2 Entropiemaximale Dichtefunktionen bei klassierten Daten

Da dieser Beitrag klassierte Daten behandelt, liegen uns also keine Informationen über einzelne Datenpunkte vor, sondern über absolute oder relative Häufigkeiten innerhalb der Klassen sowie gegebenenfalls Informationen über Momente der zugrunde liegenden Verteilung.

Diese vorgegebenen Informationen über die Verteilung gehen bei der Entropie–Maximierung in die oben angeführten Erhaltungsbedingungen ein.

Im Rahmen dieser Arbeit sollen dabei drei verschiedene Informationssituationen untersucht werden (K = Anzahl vorhandener Klassen):

1) • Keine Momente bekannt
   • Relative Häufigkeit p_i im Intervall $[a_i, b_i]$ bekannt für alle $i = 1, \ldots, K$

2) • Bedingter Mittelwert μ_i im Intervall $[a_i, b_i]$ bekannt für alle $i = 1, \ldots, K$
   • Relative Häufigkeit p_i im Intervall $[a_i, b_i]$ bekannt für alle $i = 1, \ldots, K$

3) • Unbedingter Mittelwert μ bekannt
   • Relative Häufigkeit p_i im Intervall $[a_i, b_i]$ bekannt für alle $i = 1, \ldots, K$

Aufgrund der Art der vorliegenden Informationen bei klassierten Daten ist es sinnvoll, abschnittweise entropiemaximale Dichtefunktionen unter Beachtung der Erhaltungsbedingungen zu bestimmen (korrespondierend mit den Intervallen $[a_i, b_i]$) und die zugehörigen Verteilungsfunktionen dann auf den Wertebereich $[0, p_i]$ normiert zusammenzusetzen. Bei GOLANI/PHILLIPS (1990) findet sich ein ähnlicher Ansatz, jedoch wurde dort u.a. die notwendige Bedingung der Häufigkeitserhaltung in jedem Intervall bei Fall 3 nicht eingehalten.

zu Fall 1: Neben den relativen Häufigkeiten sind keine weiteren Informationen über die Verteilung vorhanden. Die allgemeine Form der ME–Dichte für die einzelnen Intervalle $[a_i, b_i]$ lautet[3]:

$$f_i(x) = e^{\theta_0^{(i)}} \quad .$$

Unter Beachtung der Erhaltungsbedingungen (siehe Anhang) erhält man für die Parameter

$$\theta_0^{(i)} = \ln \frac{p_i}{b_i - a_i} \quad \text{für} \quad i = 1, \ldots, K$$

und somit als ME–Dichtefunktion für das Intervall $[a_i, b_i]$:

$$f_i(x) = \frac{p_i}{b_i - a_i} \quad .$$

Wie man sieht, entspricht dieses Resultat der Annahme einer Rechteck–Verteilung innerhalb jeder Klasse.

Zu Fall 2: Zusätzlich zu den relativen Häufigkeiten sind noch die bedingten Mittelwerte, also die Mittelwerte für die einzelnen Klassen gegeben. Bei der Bestimmung der abschnittweisen entropiemaximalen Dichtefunktionen sind diese zusätzlichen Informationen in Form weiterer Erhaltungsbedingungen (siehe Anhang) zu berücksichtigen. Die allgemeine Form der ME–Dichte für die einzelnen Intervalle $[a_i, b_i]$ lautet:

$$f_i(x) = e^{\theta_0^{(i)} + \theta_1^{(i)} x} \quad .$$

Wiederum unter Beachtung der Erhaltungsbedingungen erhält man für die Parameter:

$$\theta_0^{(i)} = \ln \frac{\theta_1^{(i)} p_i}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}}$$

3) Wenn die Funktion $h_1(x)$ der konstanten Funktion 1 entspricht, werden die Parameter $\theta_0^{(i)}$ und $\theta_1^{(i)}$ additiv zu $\theta_0^{(i)}$ zusammengefaßt.

und

$$\frac{[(b_i - \frac{1}{\theta_1^{(i)}})e^{\theta_1^{(i)} b_i} - (a_i - \frac{1}{\theta_1^{(i)}})e^{\theta_1^{(i)} a_i}]}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}} = \mu_i \quad .$$

Wie man erkennt, ist es nicht möglich die $\theta_1^{(i)}$ direkt zu bestimmen. Vielmehr müssen die Werte für diese Parameter numerisch z.B. über ein Nullstellenberechnungsverfahren bestimmt werden. Dabei kann man zeigen, daß der linke Teil der Gleichung für $\theta_1^{(i)} \to 0$ gegen $(a_i + b_i)/2$ konvergiert und monoton steigt. Somit läßt sich die Menge, in der sich die gesuchte Nullstelle befindet genauer eingrenzen:

$$\begin{aligned} \frac{a_i + b_i}{2} > \mu_i \quad &\Longrightarrow \quad \theta_1^{(i)} \in \mathbb{R}^- \\ \frac{a_i + b_i}{2} < \mu_i \quad &\Longrightarrow \quad \theta_1^{(i)} \in \mathbb{R}^+. \end{aligned}$$

Wenn $(a_i + b_i)/2 = \mu_i$ gilt, existiert keine Nullstelle. Nach Einsetzen der Parameter in die allgemeine Form erhält man für die entropiemaximale Dichtefunktion für das Intervall $[a_i, b_i]$:

$$f_i(x) = e^{\theta_0^{(i)}} \cdot e^{\theta_1^{(i)} x} = p_i \cdot \frac{\theta_1^{(i)} e^{\theta_1^{(i)} x}}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}}.$$

Zu Fall 3: Anstelle der bedingten Mittelwerte ist in diesem Fall nur der unbedingte Mittelwert bekannt. Die allgemeine Form der ME–Dichte für die einzelnen Intervalle $[a_i, b_i]$ lautet:

$$f_i(x) = e^{\theta_0^{(i)} + \theta_1 x} \quad .$$

Als Parameterwerte erhält man:

$$\theta_0^{(i)} = \ln \frac{\theta_1 p_i}{e^{\theta_1 b_i} - e^{\theta_1 a_i}}$$

und

$$\sum_i p_i \frac{[(b_i - \frac{1}{\theta_1})e^{\theta_1 b_i} - (a_i - \frac{1}{\theta_1})e^{\theta_1 a_i}]}{e^{\theta_1 b_i} - e^{\theta_1 a_i}} = \mu.$$

Auch hier muß θ_1 numerisch ermittelt werden. Der linke Teil der Gleichung konvergiert für $\theta_1 \to 0$ gegen $\frac{1}{2}\sum_i (a_i + b_i)$ und somit kann man analog zu Fall 2 die Menge eingrenzen, in der sich die Nullstelle befindet:

$$\begin{aligned} \frac{1}{2}\sum_i (a_i + b_i) > \mu \quad &\Longrightarrow \quad \theta_1 \in \mathbb{R}^- \\ \frac{1}{2}\sum_i (a_i + b_i) < \mu \quad &\Longrightarrow \quad \theta_1 \in \mathbb{R}^+. \end{aligned}$$

Nach Einsetzen der Parameter in die allgemeine ME–Dichte erhält man dann für das Intervall $[a_i, b_i]$:

$$f_i(x) = e^{\theta_0^{(i)}} \cdot e^{\theta_1 x} = p_i \cdot \frac{\theta_1 e^{\theta_1 x}}{e^{\theta_1 b_i} - e^{\theta_1 a_i}}.$$

3 Berechnung der Werte von Disparitätsmaßen aus entropiemaximalen Dichtefunktionen

Im folgenden sollen nun die Formeln zur Berechnung der Disparitätsmaße ermittelt werden. Dabei müssen folgende Werte bekannt sein:

- Klassengrenzen a_i, b_i mit $b_i > a_i$ für K endliche Intervalle
- relative Häufigkeiten p_i der K Klassen oder absolute Häufigkeiten n_i für K Klassen, da $p_i = n_i / \sum_{i=1}^{K} n_i$
- Parameterwerte der ME–Dichtefunktionen

Es wird zudem von ausschließlich positiven Merkmalsausprägungen ($x > 0$) ausgegangen.

Da für die Berechnung mancher Disparitätsmaße der Wert des unbedingten Mittelwertes μ einer Verteilung einbezogen werden muß, dieser aber nur in Fall 2 und 3 bekannt ist, bedarf es einer anderweitigen Bestimmung des unbedingten Mittelwertes.

Zu Fall 1: Der unbedingte Mittelwert wird näherungsweise durch $\frac{1}{2} \sum_{i=1}^{K} p_i(a_i + b_i)$ bestimmt, d.h. man unterstellt, daß die Klassenmitten die bedingten Mittelwerte repräsentieren, die mit p_i gewichtet in μ eingehen.

Zu Fall 2: Da hier die bedingten Mittelwerte gegeben sind, kann man $\mu = \sum_{i=1}^{K} p_i \mu_i$ setzen.

3.1 Gini–Koeffizient

Der Gini–Koeffizient ist definiert als die zweifache Fläche zwischen der Lorenzkurve und der Winkelhalbierenden (= Gerade der vollkommenen Gleichverteilung) im Einheitsquadrat und berechnet sich im stetigen Fall gemäß

$$G = \int_0^1 (2y - 1) G^*(y) dy.$$

Dabei ist $G^*(y) = \frac{G(y)}{\mu}$ die normierte, inverse Verteilungsfunktion (d.h. bei Vorliegen einer streng monotonen Verteilungsfunktion: Die normierte Umkehrfunktion der Verteilungsfunktion F). Dabei gilt

$$\int_0^1 G(y) dy = \mu$$

(vgl. Piesch (1975), S. 17 und S. 28 - 30).

Das Transferprinzip, das Populationsprinzip und die Skaleninvarianz werden vom Gini–Koeffizienten erfüllt. Sein Wertebereich liegt im Intervall [0,1] und ist somit normiert (vgl. Cowell (1995), S. 66).

3.1.1 Fall 1: Relative Häufigkeiten bekannt

Man geht zunächst aus von der entropiemaximalen Verteilungsfunktion, wenn keinerlei Informationen über Momente vorhanden sind

$$y = F(x) = p + p_j \cdot \frac{x - a_j}{b_j - a_j} \quad \text{für} \quad \begin{array}{l} x \in (a_j, b_j] \\ j = 1, \ldots, K. \end{array}$$

Dabei soll im folgenden gelten

$$p = \begin{cases} 0 & \text{für} \quad j = 1 \\ \sum_{i=1}^{j-1} p_i & \text{für} \quad j = 2, \ldots, K. \end{cases}$$

Man kann nun die zugehörige inverse Verteilungsfunktion bestimmen:

$$x = G(y) = a_j + (y - p)\frac{b_j - a_j}{p_j}.$$

Setzt man diese dann in die Bestimmungsgleichung für den Gini–Koeffizienten ein, erhält man

$$R = \int_0^1 (2y - 1)G^*(y)dy = \frac{1}{\mu}\sum_{j=1}^{K} \int_{F(a_j)}^{F(b_j)} (2y - 1)(a_j + (y - p)\frac{b_j - a_j}{p_j})dy.$$

Durch Integration erhält man

$$R = \frac{1}{\mu}\sum_{j=1}^{K} \left[\left(a_j - \frac{b_j - a_j}{p_j} p \right)(y^2 - y) + \frac{b_j - a_j}{p_j} \left(\frac{2}{3}y^3 - \frac{1}{2}y^2 \right) \right]_{F(a_j)}^{F(b_j)}.$$

Setzt man nun noch die Integrationsgrenzen in die Berechnungsformel ein, erhält man abschließend

$$R = \frac{1}{\mu}\sum_{j=1}^{K} \left(\left(a_j - \frac{b_j - a_j}{p_j} p \right) \left(F^2(b_j) - F(b_j) - F^2(a_j) + F(a_j) \right) \right.$$
$$\left. + \frac{b_j - a_j}{p_j} \left(\frac{2}{3}(F^3(b_j) - F^3(a_j)) - \frac{1}{2}(F^2(b_j) - F^2(a_j)) \right) \right).$$

3.1.2 Fall 2: Relative Häufigkeiten und bedingte Mittelwerte bekannt

Wir bestimmen zunächst aus

$$y = F(x) = p + p_j \frac{e^{\theta_1^{(j)} x} - e^{\theta_1^{(j)} a_j}}{e^{\theta_1^{(j)} b_j} - e^{\theta_1^{(j)} a_j}}$$

die Umkehrfunktion, wobei wir

$$c_j = \frac{p_j}{e^{\theta_1^{(j)} b_j} - e^{\theta_1^{(j)} a_j}}$$

setzen:

$$x = G(y) = \frac{1}{\theta_1^{(j)}} \ln\left(\frac{y-p}{c_j} + e^{\theta_1^{(j)} a_j}\right).$$

Man kann nun die normierte inverse Verteilungsfunktion in die Berechnungsformel für den Gini–Koeffizienten einsetzen:

$$R = \frac{1}{\mu} \sum_{j=1}^{K} \frac{1}{\theta_1^{(j)}} \int_{F(a_j)}^{F(b_j)} (2y-1) \ln\left(\frac{y-p}{c_j} + e^{\theta_1^{(j)} a_j}\right) dy.$$

Produktintegration mit $u' = 2y-1, u = y^2 - y, v = \ln(\frac{y-p}{c_j} + e^{\theta_1^{(j)} a_j}), v' = \frac{1}{c_j} \cdot 1/(\frac{y-p}{c_j} + e^{\theta_1^{(j)} a_j})$ führt zu:

$$R = \frac{1}{\mu} \sum_{j=1}^{K} \frac{1}{\theta_1^{(j)}} \left(\left[(y^2-y) \ln\left(\frac{y-p}{c_j} + e^{\theta_1^{(j)} a_j}\right) \right]_{F(a_j)}^{F(b_j)} \right.$$
$$\left. -\frac{1}{c_j} \left(\int_{F(a_j)}^{F(b_j)} \frac{y^2}{\frac{y}{c_j} - \frac{p}{c_j} + e^{\theta_1^{(j)} a_j}} dy - \int_{F(a_j)}^{F(b_j)} \frac{y}{\frac{y}{c_j} - \frac{p}{c_j} + e^{\theta_1^{(j)} a_j}} dy \right) \right).$$

Setzt man

$$A_j = \frac{1}{c_j} \qquad \text{und} \qquad B_j = e^{\theta_1^{(j)} a_j} - \frac{p}{c_j}$$

und integriert die verbleibenden Integrale, so erhält man

$$R = \frac{1}{\mu} \sum_{j=1}^{K} \frac{1}{\theta_1^{(j)}} \left[(y^2-y) \ln(A_j y + B_j) - A_j \left(\frac{1}{A_j^3} \left[\frac{1}{2} (A_j y + B_j)^2 \right.\right.\right.$$
$$\left.\left.\left. -2B_j (A_j y + B_j) + B_j^2 \ln|A_j y + B_j| \right] - \frac{y}{A_j} + \frac{B_j}{A_j^2} \ln|A_j y + B_j| \right) \right]_{F(a_j)}^{F(b_j)}.$$

Nach weiterem Zusammenfassen erhält man dann

$$R = \frac{1}{\mu} \sum_{j=1}^{K} \frac{1}{\theta_1^{(j)}} \left[y^2 \left(\ln(A_j y + B_j) - \frac{1}{2} \right) \right.$$
$$\left. -y \left(\ln(A_j y + B_j) - \frac{B_j}{A_j} - 1 \right) - \left(\frac{B_j}{A_j} + \frac{B_j^2}{A_j^2} \right) \ln|A_j y + B_j| + \frac{3}{2} \frac{B_j^2}{A_j^2} \right]_{F(a_j)}^{F(b_j)}.$$

Setzt man dann noch die Integrationsgrenzen ein erhält man abschließend

$$R = \frac{1}{\mu} \sum_{j=1}^{K} \frac{1}{\theta_1^{(j)}} \left(F^2(b_j) \left(\ln(A_j F(b_j) + B_j) - \frac{1}{2} \right) - F^2(a_j) \left(\ln(A_j F(a_j) + B_j) - \frac{1}{2} \right) \right.$$
$$-F(b_j) \left(\ln(A_j F(b_j) + B_j) - \frac{B_j}{A_j} - 1 \right) + F(a_j) \left(\ln(A_j F(a_j) + B_j) - \frac{B_j}{A_j} - 1 \right)$$
$$\left. -\left(\frac{B_j}{A_j} + \frac{B_j^2}{A_j^2} \right) \left(\ln|A_j F(b_j) + B_j| - \ln|A_j F(a_j) + B_j| \right) \right).$$

3.1.3 Fall 3: Relative Häufigkeiten und unbedingter Mittelwert bekannt

Wir bestimmen diesmal aus

$$y = F(x) = p + p_j \cdot \frac{e^{\theta_1 x} - e^{\theta_1 a_j}}{e^{\theta_1 b_j} - e^{\theta_1 a_j}}$$

die Umkehrfunktion, wobei wir nun

$$c_j = \frac{p_j}{e^{\theta_1 b_j} - e^{\theta_1 a_j}}$$

setzen. Man erhält

$$x = G(y) = \frac{1}{\theta_1} \ln\left(\frac{y-p}{c_j} + e^{\theta_1 a_j}\right).$$

Wie man erkennt, unterscheiden sich Fall 2 und Fall 3 nur durch die Anzahl der Parameter. Man berechnet

$$R = \frac{1}{\theta_1 \mu} \sum_{j=1}^{K} \int_{F(a_j)}^{F(b_j)} (2y-1) \ln\left(\frac{y-p}{c_j} + e^{\theta_1 a_j}\right) dy$$

analog wie in Fall 2 und erhält als Wert für den Gini–Koeffizienten

$$\begin{aligned} R = \frac{1}{\theta_1 \mu} \sum_{j=1}^{K} \Bigg(& F^2(b_j)\left(\ln(A_j F(b_j) + B_j) - \frac{1}{2}\right) - F^2(a_j)\left(\ln(A_j F(a_j) + B_j) - \frac{1}{2}\right) \\ & - F(b_j)\left(\ln(A_j F(b_j) + B_j) - \frac{B_j}{A_j} - 1\right) + F(a_j)\left(\ln(A_j F(a_j) + B_j) - \frac{B_j}{A_j} - 1\right) \\ & - \left(\frac{B_j}{A_j} + \frac{B_j^2}{A_j^2}\right)\left(\ln|A_j F(b_j) + B_j| - \ln|A_j F(a_j) + B_j|\right)\Bigg). \end{aligned}$$

Dabei definieren wir hier aber

$$A_j = \frac{1}{c_j} \qquad B_j = e^{\theta_1 a_j} - \frac{p}{c_j} \quad .$$

3.2 Theil–Maß

Das Theil–Maß ist ein Spezialfall der Verallgemeinerten Entropie–Maße welche aus der Informationstheorie heraus entwickelt wurden. Wenn man die Bestimmungsgleichung der Entropie $H = -\sum_{i=1}^{n} p_i \ln(p_i)$ mit

- n = Anzahl Merkmalsträger
- $p_i = \frac{x_i}{n\overline{x}}$ = Anteil der i–ten Merkmalsausprägung x_i des i–ten Merkmalsträgers an der gesammten Merkmalssumme $\overline{x}$

neu interpretiert und die Entropie dieser Verteilung von der maximal möglichen Entropie

($p_i = \frac{1}{n}$; alle Merkmalsträger besitzen die gleiche Merkmalsausprägung $\Longrightarrow$ vollkommene Gleichheit) abzieht, erhält man für den diskreten Fall (vgl. COWELL (1977), S. 56-57):

$$\begin{aligned} T &= \sum_{i=1}^{n} \frac{1}{n} h\Big(\frac{1}{n}\Big) - \sum_{i=1}^{n} p_i h(p_i) = \sum_{i=1}^{n} p_i \Big(h\Big(\frac{1}{n}\Big) - h(p_i)\Big) = \sum_{i=1}^{n} p_i \Big(\ln(p_i) - \ln\Big(\frac{1}{n}\Big)\Big) \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{\overline{x}} \ln\Big(\frac{x_i}{\overline{x}}\Big). \end{aligned}$$

Den stetigen Fall kann man analog interpretieren und erhält

$$T = \int_0^\infty \Big(\frac{x}{\mu}\Big) \ln\Big(\frac{x}{\mu}\Big) f(x) dx.$$

Das Theil–Maß ist ebenfalls skaleninvariant und erfüllt das Transfer- und Populationsprinzip, wobei sein Wertebereich jedoch im offenen Intervall $[0, \infty)$ liegt.
Im folgenden wird definiert:

$$c_{1,i} = \frac{p_i}{b_i - a_i} \qquad c_{2,i} = \frac{p_i}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}} \qquad c_{3,i} = \frac{p_i}{e^{\theta_1 b_i} - e^{\theta_1 a_i}}.$$

3.2.1 Fall 1: Relative Häufigkeiten bekannt

Klassenweise integrierend erhält man durch Produktintegration von

$$T = \frac{1}{\mu} \sum_{i=1}^{K} c_{1,i} \int_{a_i}^{b_i} x \ln\Big(\frac{x}{\mu}\Big) dx$$

mit

$$u' = x; \quad u = \frac{1}{2}x^2; \quad v = \ln\Big(\frac{x}{\mu}\Big); \quad v' = \frac{1}{x}$$

folgenden Term mit einfach zu berechnendem Integral

$$\frac{1}{\mu} \sum_{i=1}^{K} c_{1,i} \left(\Big[\frac{1}{2}x^2 \ln\Big(\frac{x}{\mu}\Big)\Big]_{a_i}^{b_i} - \int_{a_i}^{b_i} \frac{1}{2} x dx \right).$$

Klammert man nun noch $\frac{1}{2}x^2$ aus und setzt die Integrationsgrenzen ein, so erhält man als Lösung

$$\begin{aligned} T &= \frac{1}{\mu} \sum_{i=1}^{K} c_{1,i} \cdot \left[\frac{1}{2}x^2 \left(\ln\Big(\frac{x}{\mu}\Big) - \frac{1}{2} \right) \right]_{a_i}^{b_i} \\ &= \frac{1}{\mu} \sum_{i=1}^{K} c_{1,i} \cdot \left(\frac{1}{2} b_i^2 \left(\ln\Big(\frac{b_i}{\mu}\Big) - \frac{1}{2} \right) - \frac{1}{2} a_i^2 \left(\ln\Big(\frac{a_i}{\mu}\Big) - \frac{1}{2} \right) \right). \end{aligned}$$

3.2.2 Fall 2: Relative Häufigkeiten und bedingte Mittelwerte bekannt

Wir müssen

$$T = \frac{1}{\mu}\sum_{i=1}^{K} c_{2,i} \int_{a_i}^{b_i} \left(x \ln\left(\frac{x}{\mu}\right)\right) \theta_1^{(i)} e^{\theta_1^{(i)} x} dx$$

berechnen und Integration ergibt dann

$$\begin{aligned}
T &= \frac{1}{\mu}\sum_{i=1}^{K} c_{2,i} \left(\left[x \ln\left(\frac{x}{\mu}\right) e^{\theta_1^{(i)} x} \right]_{a_i}^{b_i} - \int_{a_i}^{b_i} \left(1 + \ln\left(\frac{x}{\mu}\right)\right) e^{\theta_1^{(i)} x} dx \right) \\
&= \frac{1}{\mu}\sum_{i=1}^{K} c_{2,i} \left(\left[x \ln\left(\frac{x}{\mu}\right) e^{\theta_1^{(i)} x} \right]_{a_i}^{b_i} - \left(\left[\frac{1}{\theta_1^{(i)}} e^{\theta_1^{(i)} x} \left(1 + \ln\left(\frac{x}{\mu}\right)\right) \right]_{a_i}^{b_i} - \frac{1}{\theta_1^{(i)}} \int_{a_i}^{b_i} \frac{e^{\theta_1^{(i)} x}}{x} dx \right) \right) \\
&= \frac{1}{\mu}\sum_{i=1}^{K} c_{2,i} \left[\ln\left(\frac{x}{\mu}\right) e^{\theta_1^{(i)} x} \left(x - \frac{1}{\theta_1^{(i)}} \right) - \frac{1}{\theta_1^{(i)}} e^{\theta_1^{(i)} x} + \frac{1}{\theta_1^{(i)}} \left(\ln x + \sum_{r=1}^{\infty} \frac{(\theta_1^{(i)} x)^r}{r \cdot r!} \right) \right]_{a_i}^{b_i}.
\end{aligned}$$

Mittels des Quotientenkriteriums kann man die Konvergenz der unendlichen Reihe zeigen und nach Einsetzen der Integrationsgrenzen und Zusammenfassung erhält man schließlich

$$\begin{aligned}
T &= \frac{1}{\mu}\sum_{i=1}^{K} c_{2,i} \left(\left(b_i - \frac{1}{\theta_1^{(i)}} \right) \ln\left(\frac{b_i}{\mu}\right) e^{\theta_1^{(i)} b_i} - \left(a_i - \frac{1}{\theta_1^{(i)}} \right) \ln\left(\frac{a_i}{\mu}\right) e^{\theta_1^{(i)} a_i} \right. \\
&\quad \left. + \frac{1}{\theta_1^{(i)}} \left(e^{\theta_1^{(i)} a_i} - e^{\theta_1^{(i)} b_i} + \ln\left(\frac{b_i}{a_i}\right) + \sum_{r=1}^{\infty} \frac{(\theta_1^{(i)})^r (b_i^r - a_i^r)}{r \cdot r!} \right) \right).
\end{aligned}$$

3.2.3 Fall 3: Relative Häufigkeiten und unbedingter Mittelwert bekannt

Wie man schon bei der Berechnung für den Gini–Koeffizienten gesehen hat, bedarf es für den Fall 3 nur einer Veränderung der Parameter von Fall 2. Wir erhalten

$$\begin{aligned}
T &= \frac{1}{\mu}\sum_{i=1}^{K} c_{3,i} \left(\left(b_i - \frac{1}{\theta_1} \right) \ln\left(\frac{b_i}{\mu}\right) e^{\theta_1 b_i} - \left(a_i - \frac{1}{\theta_1} \right) \ln\left(\frac{a_i}{\mu}\right) e^{\theta_1 a_i} \right. \\
&\quad \left. + \frac{1}{\theta_1} \left(e^{\theta_1 a_i} - e^{\theta_1 b_i} + \ln\left(\frac{b_i}{a_i}\right) + \sum_{r=1}^{\infty} \frac{(\theta_1)^r (b_i^r - a_i^r)}{r \cdot r!} \right) \right).
\end{aligned}$$

3.3 Pietra Maß / Maß von Kuznet

Das Pietra–Maß, auch bekannt als das Maß von Kuznets (vgl. LÜTHI (1981), S. 30)

$$P = \int \frac{|x - \mu|}{2\mu} f(x) dx$$

läßt sich anschaulich als maximaler Abstand der Lorenzkurve von der Gleichverteilungsgeraden (Winkelhalbierende) im Einheitsquadrat deuten.

Das Maß repräsentiert die normierte, mittlere Abweichung. Dabei wird der absolute Abstand der einzelnen Ausprägungen zum Mittelwert der Verteilung gemessen. Mit 2 multipliziert erhält man die bekannte relative, mittlere Abweichung.

Die Tatsache, daß das Pietra–Maß sich durch nur einen Punkt der Lorenzkurve darstellen läßt, deutet darauf hin, daß dieses Maß wesentliche Informationen einer Verteilung unberücksichtigt läßt. So werden Umverteilungen zwischen Merkmalsträgern, deren Merkmalsausprägungen alle oberhalb bzw. alle unterhalb des Mittelwertes liegen, überhaupt nicht erfaßt, d.h. für beliebige Lorenzkurven erhält man, unter der einzigen Bedingung, daß die Lorenzkurve durch den oben angegebenen Punkt hindurchläuft, den gleichen Wert für das Pietra–Maß (vgl. COWELL (1977), S. 72 und LÜTHI (1981), S. 32).

Es ist offensichtlich, daß das Transfer–Prinzip bei Umverteilungen zwischen Klassen oberhalb bzw. unterhalb des Mittelwertes nicht erfüllt wird. Das Pietra–Maß erfüllt dieses Prinzip nur bei Umverteilungen über den Mittelwert hinweg. Nur hier zeigt das Pietra–Maß maximale Veränderung der Ungleichheit an und nur in diesem Fall kann man bei diesem Maß von Sensitivität sprechen (vgl. SCHMID (1991), S. 159/163 und LÜTHI (1981), S. 91). Das Pietra Maß ist aber skaleninvariant, was für die Berechnung von Vorteil sein wird, wie wir später sehen werden.

3.3.1 Fall 1: Relative Häufigkeiten bekannt

Beachtet man die notwendige Fallunterscheidung wegen des Auftretens des Absolutbetrages, so erhält man (dabei soll im folgenden gelten: $\sum_{i=1}^{j-1} \ldots = 0$ für $j = 1$)

$$\begin{aligned}
\frac{1}{2\mu}\sum_{i=1}^{K} c_{1,i}\int_{a_i}^{b_i}|x-\mu|dx \;=\; & \frac{1}{2\mu}\left(\sum_{i=1}^{j-1} c_{1,i}\left(-\left[\frac{1}{2}x^2-\mu x\right]_{a_i}^{b_i}\right)\right. \\
& +c_{1,j}\left(-\left[\frac{1}{2}x^2-\mu x\right]_{a_j}^{\mu}+\left[\frac{1}{2}x^2-\mu x\right]_{\mu}^{b_j}\right) \\
& \left.+\sum_{i=j+1}^{K} c_{1,i}\left[\frac{1}{2}x^2-\mu x\right]_{a_i}^{b_i}\right) \qquad \text{für } \mu\in(a_j,b_j]
\end{aligned}$$

Einsetzen der Integrationsgrenzen und Zusammenfassen ergibt dann

$$\begin{aligned}
P \;=\; & \frac{1}{2\mu}\left(\sum_{\substack{i=1\\ i\neq j}}^{K}\operatorname{sgn}(b_i-\mu)c_{1,i}\Big(\frac{1}{2}(b_i^2-a_i^2)-\mu(b_i-a_i)\Big)\right. \\
& \left.+c_{1,j}\Big(\frac{1}{2}(b_j^2+a_j^2+2\mu^2)-\mu(b_j+a_j)\Big)\right) \qquad \text{für } \mu\in(a_j,b_j]
\end{aligned}$$

Dabei ist $\operatorname{sgn}(b_i-\mu)=1$ für $(b_i>\mu)\wedge(\mu\notin(a_i,b_i])$ und $\operatorname{sgn}(b_i-\mu)=-1$ für $(b_i<\mu)\wedge(\mu\notin(a_i,b_i])$.

3.3.2 Fall 2: Relative Häufigkeiten und bedingte Mittelwerte bekannt

Wir berechnen

$$
\begin{aligned}
P = \frac{1}{2\mu}\sum_{i=1}^{K} c_{2,i}\int_{a_i}^{b_i} |x-\mu|\theta_1^{(i)} e^{\theta_1^{(i)}x}dx \\
&= \frac{1}{2\mu}\Big(-\sum_{i=1}^{j-1} c_{2,i}\int_{a_i}^{b_i}(x-\mu)\theta_1^{(i)}e^{\theta_1^{(i)}x}dx \\
&\quad +c_{2,j}\Big(-\int_{a_j}^{\mu}(x-\mu)\theta_1^{(j)}e^{\theta_1^{(j)}x}dx + \int_{\mu}^{b_j}(x-\mu)\theta_1^{(j)}e^{\theta_1^{(j)}x}dx\Big) \\
&\quad +\sum_{i=j+1}^{K} c_{2,i}\int_{a_i}^{b_i}(x-\mu)\theta_1^{(i)}e^{\theta_1^{(i)}x}dx\Big) \qquad \text{für } \mu\in(a_j,b_j]
\end{aligned}
$$

Unter Beachtung von

$$
\theta_1^{(i)}\int_{a_i}^{b_i}(x-\mu)e^{\theta_1^{(i)}x}dx = \Big[e^{\theta_1^{(i)}x}(x-\mu-\frac{1}{\theta_1^{(i)}})\Big]_{a_i}^{b_i}
$$

erhält man als Ergebnis

$$
\begin{aligned}
P &= \frac{1}{2\mu}\Big(\sum_{\substack{i=1\\ i\neq j}}^{K}\operatorname{sgn}(b_i-\mu)c_{2,i}\Big(e^{\theta_1^{(i)}b_i}(b_i-\mu-\frac{1}{\theta_1^{(i)}}) - e^{\theta_1^{(i)}a_i}(a_i-\mu-\frac{1}{\theta_1^{(i)}})\Big) \\
&\quad +c_{2,j}\Big(e^{\theta_1^{(j)}b_j}(b_j-\mu-\frac{1}{\theta_1^{(j)}}) + e^{\theta_1^{(j)}a_j}(a_j-\mu-\frac{1}{\theta_1^{(j)}}) + \frac{2}{\theta_1^{(j)}}e^{\theta_1^{(j)}\mu}\Big)\Big).
\end{aligned}
$$

3.3.3 Fall 3: Relative Häufigkeiten und unbedingter Mittelwert bekannt

Man muß auch hier nur wieder die Parameter austauschen und es ergibt sich

$$
\begin{aligned}
P &= \frac{1}{2\mu}\Big(\sum_{\substack{i=1\\ i\neq j}}^{K}\operatorname{sgn}(b_i-\mu)c_{3,i}\Big(e^{\theta_1 b_i}(b_i-\mu-\frac{1}{\theta_1}) - e^{\theta_1 a_i}(a_i-\mu-\frac{1}{\theta_1})\Big) \\
&\quad +c_{3,j}\Big(e^{\theta_1 b_j}(b_j-\mu-\frac{1}{\theta_1}) + e^{\theta_1 a_j}(a_j-\mu-\frac{1}{\theta_1}) + \frac{2}{\theta_1}e^{\theta_1\mu}\Big)\Big).
\end{aligned}
$$

3.4 Logarithmische Varianz

Die stetige Form der Logarithmischen Varianz lautet

$$
LV = \int_0^{\infty}\Big(\ln\Big(\frac{x}{\mu}\Big)\Big)^2 f(x)dx.
$$

Durch die konkave Funktionsform des Logarithmus werden absolut gleiche Transfers im Bereich niedrigerer Ausprägungen stärker gewichtet als bei höheren Ausprägungen, d.h. die Logarithmische Varianz reagiert besonders empfindlich auf Veränderungen im untersten Bereich einer Verteilung. Das Transferprinzip ist nicht erfüllt, da das Maß steigen kann, obwohl von höheren Merkmalswerten zu relativ kleineren Merkmalswerten umverteilt wird. Alle anderen Prinzipien, mit Außnahme der Normierung (Wertebereich $[0, \infty)$), werden erfüllt. Problematisch ist das starke Anwachsen der Logarithmischen Varianz bei Berücksichtigung von Ausprägungen nahe Null.

3.4.1 Fall 1: Relative Häufigkeiten bekannt

Berechnung von

$$LV = \sum_{i=1}^{K} c_{1,i} \int_{a_i}^{b_i} \ln^2 \left(\frac{x}{\mu}\right) dx$$

ergibt zunächst

$$LV = \sum_{i=1}^{K} c_{1,i} \left[\left(\ln\left(\frac{x}{\mu}\right) - 2 \right) x \ln\left(\frac{x}{\mu}\right) + 2x \right]_{a_i}^{b_i}$$

und abschließend

$$LV = \sum_{i=1}^{K} c_{1,i} \left(\left(\ln\left(\frac{b_i}{\mu}\right) - 2 \right) b_i \ln\left(\frac{b_i}{\mu}\right) + 2b_i - \left(\ln\left(\frac{a_i}{\mu}\right) - 2 \right) a_i \ln\left(\frac{a_i}{\mu}\right) - 2a_i \right).$$

3.4.2 Fall 2: Relative Häufigkeiten und bedingte Mittelwerte bekannt

Man berechnet aus

$$LV = \sum_{i=1}^{K} c_{2,i} \int_{a_i}^{b_i} \theta_1^{(i)} e^{\theta_1^{(i)} x} \ln^2 \left(\frac{x}{\mu}\right) dx$$

im ersten Schritt

$$LV = \sum_{i=1}^{K} c_{2,i} \left(\left[e^{\theta_1^{(i)} x} \ln^2 \left(\frac{x}{\mu}\right) \right]_{a_i}^{b_i} - 2 \int_{a_i}^{b_i} \frac{e^{\theta_1^{(i)} x}}{x} \ln\left(\frac{x}{\mu}\right) dx \right)$$

und man erhält über

$$\int_{a_i}^{b_i} \frac{e^{\theta_1^{(i)} x}}{x} \ln\left(\frac{x}{\mu}\right) dx = \left[\left(\ln(x) + \sum_{r=1}^{\infty} \frac{(\theta_1^{(i)} x)^r}{r \cdot r!} \right) \ln\left(\frac{x}{\mu}\right) - \frac{1}{2} \ln^2(x) - \sum_{r=1}^{\infty} \frac{(\theta_1^{(i)} x)^r}{r \cdot r!} \right]_{a_i}^{b_i}$$

das Zwischenergebnis

$$LV = \sum_{i=1}^{K} c_{2,i} \left[e^{\theta_1^{(i)} x} \ln^2 \left(\frac{x}{\mu}\right) - 2 \left(\sum_{r=1}^{\infty} \frac{(r \ln(\frac{x}{\mu}) - 1)(\theta_1^{(i)} x)^r}{r^2 \cdot r!} + \ln(x) \ln\left(\frac{\sqrt{x}}{\mu}\right) \right) \right]_{a_i}^{b_i}.$$

Mittels des Quotientenkriteriums kann man wiederum die Konvergenz der unendlichen Reihe zeigen und Einsetzen der Integrationsgrenzen und Umformen ergibt schließlich

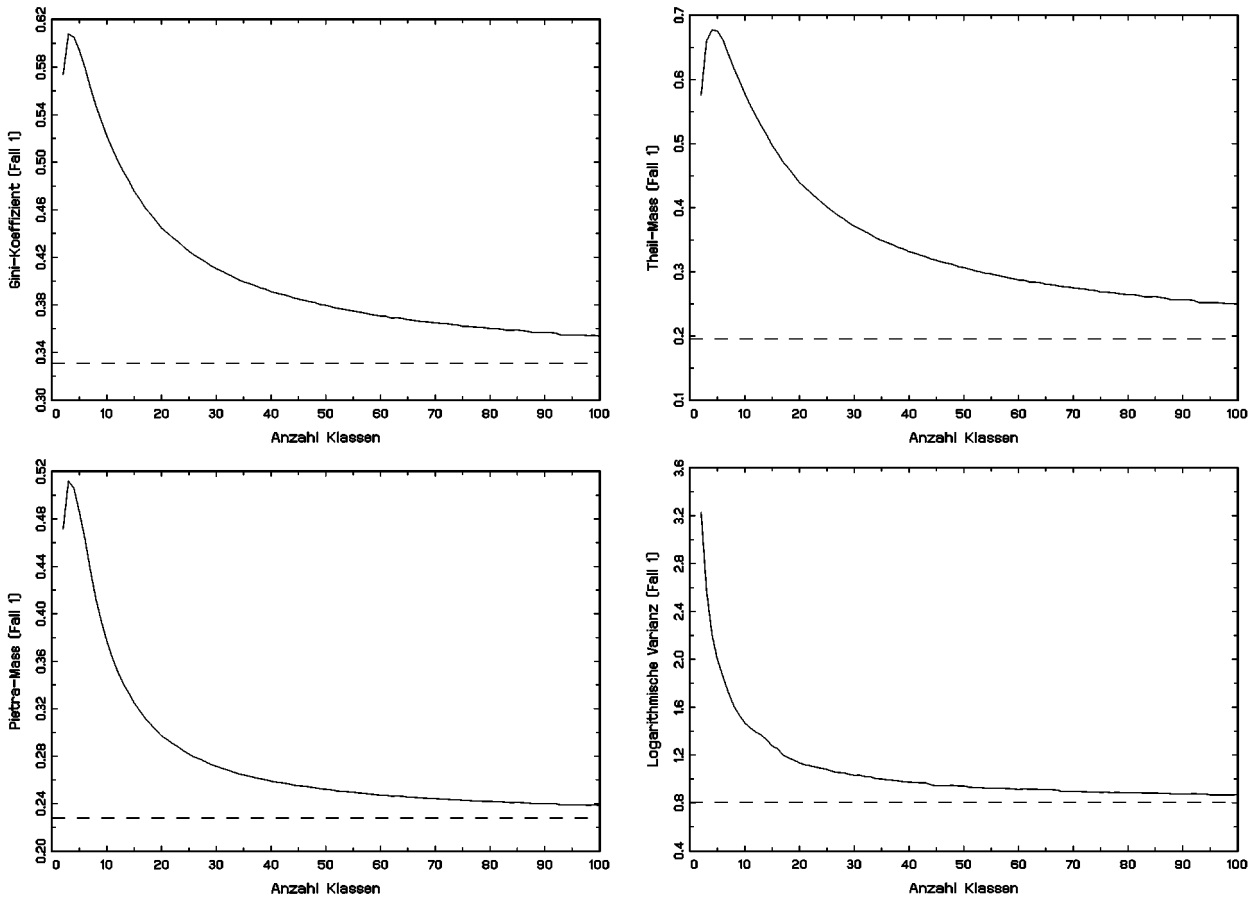
$$\begin{aligned} LV &= \sum_{i=1}^{K} c_{2,i} \Bigg(2 \sum_{r=1}^{\infty} \frac{(\theta_1^{(i)})^r \left(a_i^r (r \ln(\frac{a_i}{\mu}) - 1) - b_i^r (r \ln(\frac{b_i}{\mu}) - 1) \right)}{r^2 \cdot r!} \\ &\quad + e^{\theta_1^{(i)} b_i} \ln^2 \left(\frac{b_i}{\mu} \right) - e^{\theta_1^{(i)} a_i} \ln^2 \left(\frac{a_i}{\mu} \right) + 2 \ln(a_i) \ln \left(\frac{\sqrt{a_i}}{\mu} \right) - 2 \ln(b_i) \ln \left(\frac{\sqrt{b_i}}{\mu} \right) \Bigg). \end{aligned}$$

3.4.3 Fall 3: Relative Häufigkeiten und unbedingter Mittelwert bekannt

Nach Austausch der Parameter erhalten wir als Ergebnis

$$\begin{aligned} LV &= \sum_{i=1}^{K} c_{3,i} \Bigg(2 \sum_{r=1}^{\infty} \frac{(\theta_1)^r \left(a_i^r (r \ln(\frac{a_i}{\mu}) - 1) - b_i^r (r \ln(\frac{b_i}{\mu}) - 1) \right)}{r^2 \cdot r!} \\ &\quad + e^{\theta_1 b_i} \ln^2 \left(\frac{b_i}{\mu} \right) - e^{\theta_1 a_i} \ln^2 \left(\frac{a_i}{\mu} \right) + 2 \ln(a_i) \ln \left(\frac{\sqrt{a_i}}{\mu} \right) - 2 \ln(b_i) \ln \left(\frac{\sqrt{b_i}}{\mu} \right) \Bigg). \end{aligned}$$

Abb. 1: Schätzwerte der Disparitätsmaße bei Anwendung der ME–Methode auf klassierte Daten für Fall 1 (Anzahl Klassen von 2 bis 100) für das Erhebungsjahr 1991



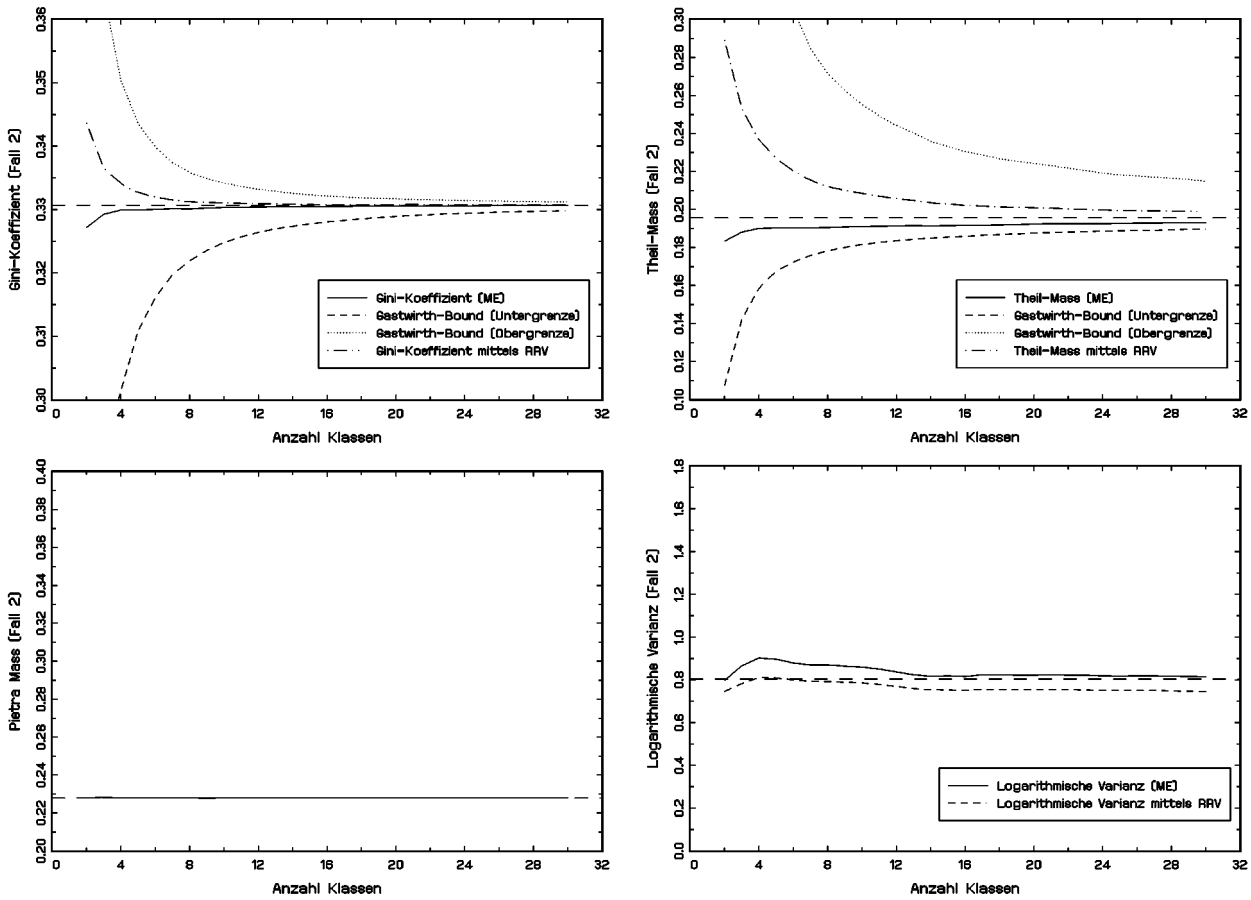
4 Lösungsansätze in der Literatur

Für das Problem der unbekannten Verteilung innerhalb einer Klasse gibt es in der Literatur verschiedene Lösungsversuche.

Neben Ansätzen wie dem "**frequency curve approach**", bei dem einfach die Klassenmittelpunkte eines Histogramms miteinander verbunden werden, oder wie der Annäherung an die Verteilung durch Polynome (z.B. mit Hilfe **kubischer Splines**), ist eine weit verbreitete Methode, gewisse Annahmen in bezug auf die Zugehörigkeit der Verteilung eines Merkmals zu einer **Verteilungsfamilie** (z.B. Log–Normalverteilung, Singh–Maddala–Verteilung) zu treffen. Dabei ist eine Verteilungsannahme über alle Klassen hinweg oder aber für jede Klasse einzeln vorzunehmen. Die für die Anwendung dieser Verteilungen notwendigen Parameter kann man dann beispielsweise über das Maximum–Likelihood Verfahren bestimmen, was allerdings u.U. über ein aufwendigeres numerisches Verfahren geschehen muß. Dieser Ansatz soll hier nicht weiter verfolgt werden. Ein einfacherer parametrischer Ansatz kann durch die Annahme einer Mittelwert erhaltenden Rechteck–Rechteck–Verteilung realisiert werden. Die Bestimmungsgleichungen für die hier betrachteten Disparitätsmaße finden sich bei SCHADER/SCHMID (1988), S.453/454.

Es besteht auch die Möglichkeit auf jegliche Verteilungsannahme zu verzichten und mittels eines nichtparametrischen Ansatzes die Werte der Disparitätsmaße zu ermitteln. GASTWIRTH (1972) hat gezeigt, daß es für bestimmte Maße möglich ist eine Abschätzung deren

Abb. 2: Schätzwerte der Disparitätsmaße bei Anwendung der ME–Methode auf klassierte Daten für Fall 2 (Anzahl Klassen von 2 bis 30) für das Erhebungsjahr 1991



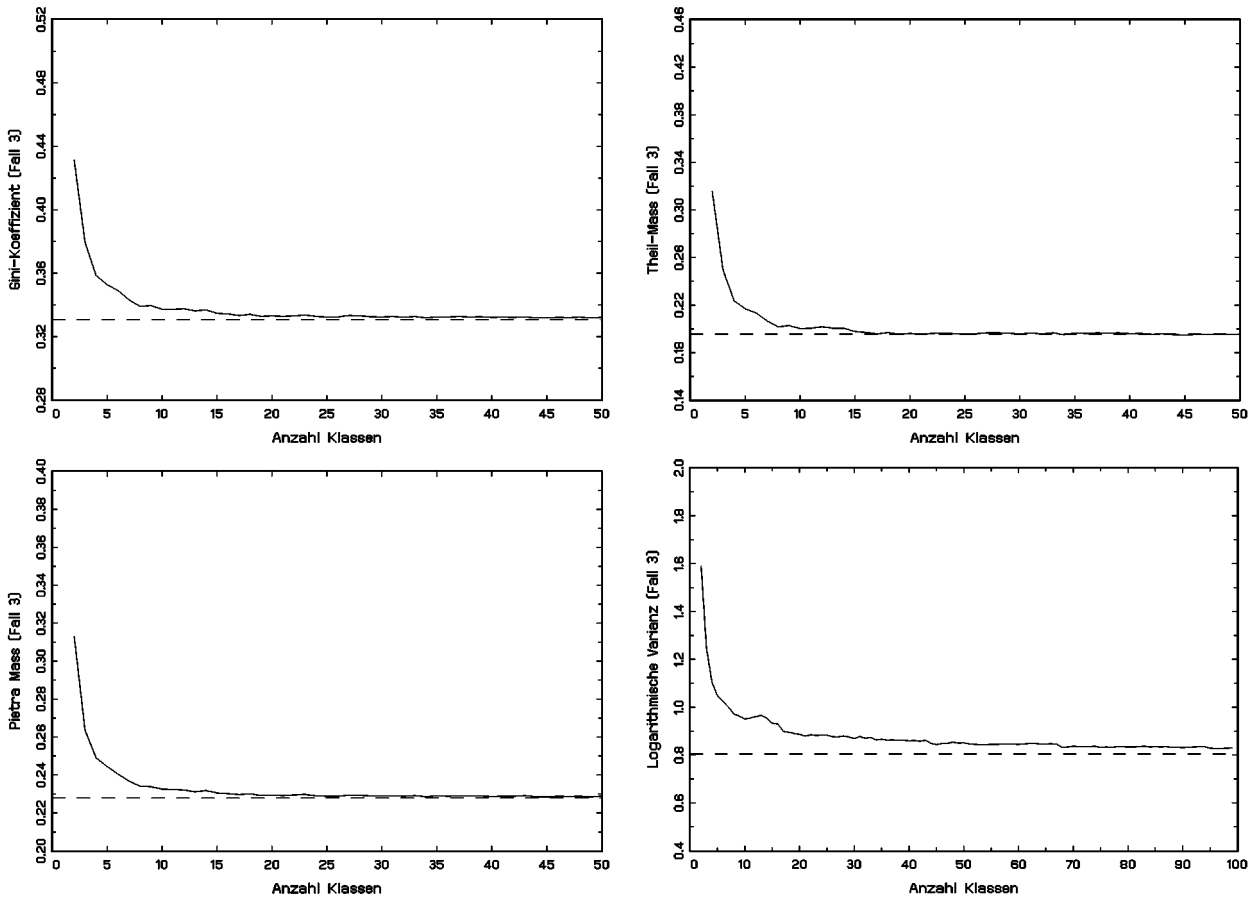
Werte nach unten und nach oben vorzunehmen (sog. **Gastwirth–Bounds**). Die Ergebnisse dieses Verfahrens können jedoch nur dann sinnvoll eingesetzt werden, wenn die Grenzen nahe genug beieinander liegen (ein methodischer und empirischer Vergleich der parametrischen Ansätze und der Gastwirth–Bounds findet sich in SCHADER/SCHMID (1988)).

Für den weiteren Vergleich mit dem ME–Verfahren, werden als ebenfalls einfach anzuwendende Verfahren die Gastwirth–Bounds sowie der parametrische Ansatz mittels einer Mittelwert erhaltenden Rechteck–Rechteck–Verteilung in der folgenden Simulationsstudie für Fall 2 mitberücksichtigt.

5 Simulationsstudie

Für die empirische Untersuchung des in dieser Arbeit vorgestellten Verfahrens zur Schätzung von Disparitätsmaßen wurden Daten des SOEP (**S**ozio–**Oe**konomisches **P**anel) zu Einkommensverhältnissen von zufällig ausgewählten, in West–Deutschland wohnhaften Einzelpersonen verwendet, die vom Deutschen Institut für Wirtschaftsforschung (DIW) kontinuierlich ermittelt werden, wobei für diese Arbeit Daten aus dem Erhebungsjahr 1991 herangezogen wurden. Dabei gingen solche Personen in die hier verwendete Stichprobe ein, die in zwei aufeinanderfolgenden beliebigen Jahren ein positives Einkommen aufwiesen.

Abb. 3: Schätzwerte der Disparitätsmaße bei Anwendung der ME–Methode auf klassierte Daten für Fall 3 (Anzahl Klassen von 2 bis 50 bzw. 100) für das Erhebungsjahr 1991



Neben den hier nicht relevanten weitergehenden Erhebungen (Berufsgruppe, höchster Bildungsabschluß,...) wurden folgende Daten erhoben, die für die weitere Untersuchung Verwendung fanden: Anzahl der Monate mit Arbeitseinkommen, durchschnittliches Brutto–Monats–Arbeitseinkommen (in den Monaten der Erwerbstätigkeit), 13. und 14. Monatsgehalt (Weihnachts- und Urlaubsgeld) und sonstige Zahlungen. Diese Einzelposten wurden für diese Arbeit zu einem Jahreseinkommen addiert und auf 12 Monate verteilt, um somit ein durchschnittliches Gesamt–Brutto–Monatseinkommen (Einkommen vor Steuern) für alle Monate zu erhalten (unabhängig davon, ob alle Personen in einem Jahr durchgängig erwerbstätig waren oder nicht).

Tabelle 1: Kenndaten der benutzten Einkommensdaten von 1991

Anzahl Daten	Minimum	Maximum	Arithmetisches Durchschnittseinkommen
5188	23,33 DM	40666,67 DM	3365,73 DM

Die monatlichen Durchschnittseinkommen wurden klassiert, nachdem die "wahren" Werte der verschiedenen Disparitätsmaße aus diesen Einzeldaten berechnet wurden. Anschließend wurden die Werte der Klassengrenzen so skaliert, daß die Obergrenzen der letzten Klasse bei jeder Klasseneinteilung den Wert 1 aufwiesen.

Da alle hier verwendeten Disparitätsmaße skaleninvariant sind, gibt es durch diese Transformation keine Veränderung der Werte der Disparitätsmaße, erleichtert aber die Berechnung der Werte für diese Simulationsstudie.

Um die Auswirkung der Anzahl der vorhandenen Klassen zu erkennen, wurde die Klasseneinteilung mittels Quantile der Ordnung 2 bis 100 vorgenommen, d.h. für alle drei Fälle wurden die Daten gemäß obigem Verfahren durchgehend aufgeteilt in 2 bis 100 Klassen (Fall 1 und 3) bzw. in 2 bis 30 Klassen (Fall 2). Für jede dieser verschiedenen Klassenzahlen wurden dann die Schätzungen der Disparitätsmaße errechnet.

Bei Fall 2 wurde die maximale Klassenzahl auf 30 begrenzt, da schon hier eine gute Konvergenz aller Maße hin zu ihrem wahren Wert sichtbar wurde. Zudem treten bei großen Klassenzahlen vermehrt Probleme bei der numerischen Berechnung der unendlichen Reihen von T und LV auf.

In der Tabelle 2 sind die Ergebnisse für die verschiedenen Fälle dargestellt. Zum einen die absoluten Werte der Schätzung, sowie die prozentuale Abweichung (PA) der Schätzwerte von den 'wahren' Werten gemäß

$$PA = \frac{\text{Schätzwert} - \text{Wahrer Wert}}{\text{Wahrer Wert}} \cdot 100.$$

In den Abbildungen 1 bis 4 sind die Ergebnisse grafisch dargestellt und man erkennt (nicht überraschend) die Abnahme der Fehlschätzung bei zunehmender Klassenzahl bei allen Verfahren. Die Abbildungen 1 bis 3 stellen die absoluten Abweichungen der jeweiligen geschätzten Werte von den aus den Einzeldaten berechneten 'wahren' Werten dar (gestrichelte Parallele zur Abszisse).

Zu Fall 1: Wie oben bereits erwähnt, entspricht dieser Fall der Annahme einer Rechteckverteilung innerhalb der Klassen. Es ist deutlich zu erkennen, daß bei Fehlen von Informationen über Momente der wahren Verteilung keine allzu guten Ergebnisse für kleine Klassenzahlen erzielt werden können. Für große Klassenzahlen erhält man für den Gini–Koeffizienten, das Pietra–Maß und die Logarithmische Varianz akzeptable Ergebnisse. Große Klassenzahlen kommen in der Praxis aber nicht so häufig vor. Die maximale prozentuale Abweichung (siehe Abb. 4) ist bei kleinen Klassenzahlen bei G und bei P deutlich geringer als bei den anderen beiden Maßen.

Zu Fall 2: Für G und T wurden hier vergleichend neben den ME–Schätzungen noch die Gastwirth–Bounds berechnet, für G, T und LV ebenso die jeweiligen Schätzungen der Disparitätsmaße unter Zugrundelegung einer mittelwerterhaltenden Rechteck–Rechteck Verteilung.

Schon bei kleinen Klassenzahlen erkennt man bei G, T und besonders bei P eine gute Annäherung an den wahren Wert bei Anwendung der ME–Methode. Beim Pietra–Maß ist schon ab 2 Klassen fast keine Abweichung vom wahren Wert mehr festzustellen.

Bei größeren Klassenzahlen reduziert sich dann der Unterschied zwischen den einzelenen Verfahren, sodaß in diesem Fall bei größeren Klassenzahlen die Verfahrenswahl eine untergeordnete Rolle spielt.

Zu Fall 3: Schon bei kleineren Klassenzahlen ist zu erkennen, daß die wahren Werte durch die ME–Schätzwerte bei G, T und P gut angenähert werden. Bei größeren Klassenzahlen gibt es dann keine nennenswerte Abweichung von den wahren Werten mehr. Die Logarithmische Varianz fällt hier etwas negativ aus dem Rahmen: Der Schätzwert weicht sogar bei 100 Klassen noch über 2% vom wahren Wert ab. Dies liegt vermutlich an der recht hohen Sensitivität dieses Maßes im unteren Wertebereich. Durch die wenigen Informationen über

Tabelle 2: Ergebnisse der ME–Schätzungen für Fall 1, 2 und 3

Fall 1								
	Gini		Theil		Pietra		Log. Var.	
w. Wert	0.33070		0.19569		0.22788		0.80360	
Anzahl Klassen	Wert	PA	Wert	PA	Wert	PA	Wert	PA
2	0.57365	(73.463)	0.57583	(194.254)	0.47154	(106.924)	3.22703	(301.570)
3	0.60790	(83.820)	0.65961	(237.066)	0.51204	(124.697)	2.58286	(221.410)
4	0.60515	(82.988)	0.67785	(246.387)	0.50564	(121.888)	2.20779	(174.736)
5	0.59429	(79.704)	0.67574	(245.309)	0.48645	(113.467)	2.00202	(149.130)
6	0.57956	(75.250)	0.66187	(238.221)	0.46277	(103.076)	1.85438	(130.758)
7	0.56300	(70.243)	0.63973	(226.907)	0.43687	(91.710)	1.72358	(114.482)
8	0.54762	(65.592)	0.61733	(215.461)	0.41289	(81.187)	1.60874	(100.191)
9	0.53475	(61.700)	0.59838	(205.777)	0.39343	(72.648)	1.53302	(90.768)
10	0.52199	(57.842)	0.57782	(195.271)	0.37675	(65.328)	1.46656	(82.498)
15	0.47572	(43.851)	0.49737	(154.160)	0.32480	(42.531)	1.27571	(58.749)
20	0.44452	(34.416)	0.43913	(124.399)	0.29703	(30.345)	1.13388	(41.100)
35	0.39944	(20.785)	0.34920	(78.444)	0.26404	(15.868)	0.99900	(24.315)
55	0.37486	(13.352)	0.29677	(51.652)	0.24946	(9.470)	0.92336	(14.903)
75	0.36188	(9.427)	0.26849	(37.201)	0.24256	(6.442)	0.88664	(10.333)
95	0.35463	(7.235)	0.25178	(28.662)	0.23900	(4.880)	0.86748	(7.949)

Fall 2								
	Gini		Theil		Pietra		Log. Var.	
w. Wert	0.33070		0.19569		0.22788		0.80360	
Anzahl Klassen	Wert	PA	Wert	PA	Wert	PA	Wert	PA
2	0.32719	(-1.063)	0.18337	(-6.296)	0.22795	(0.031)	0.79787	(-0.713)
3	0.32921	(-0.452)	0.18806	(-3.900)	0.22825	(0.162)	0.86517	(7.661)
4	0.32990	(-0.243)	0.19011	(-2.852)	0.22794	(0.026)	0.90218	(12.267)
5	0.32994	(-0.231)	0.19032	(-2.745)	0.22801	(0.057)	0.89595	(11.492)
6	0.33004	(-0.201)	0.19029	(-2.760)	0.22791	(0.013)	0.87844	(9.313)
7	0.33013	(-0.174)	0.19026	(-2.776)	0.22789	(0.004)	0.87002	(8.265)
8	0.33011	(-0.180)	0.19039	(-2.709)	0.22791	(0.013)	0.86895	(8.132)
9	0.33023	(-0.143)	0.19078	(-2.510)	0.22781	(-0.031)	0.86372	(7.481)
10	0.33034	(-0.110)	0.19108	(-2.357)	0.22790	(0.009)	0.85955	(6.962)
15	0.33045	(-0.077)	0.19145	(-2.167)	0.22790	(0.009)	0.81817	(1.813)
20	0.33057	(-0.041)	0.19231	(-1.728)	0.22787	(-0.005)	0.82172	(2.254)
25	0.33061	(-0.028)	0.19262	(-1.570)	0.22788	(-0.000)	0.81801	(1.793)
30	0.33064	(-0.019)	0.19304	(-1.355)	0.22789	(0.004)	0.81479	(1.392)

Fall 3								
	Gini		Theil		Pietra		Log. Var.	
w. Wert	0.33070		0.19569		0.22788		0.80360	
Anzahl Klassen	Wert	PA	Wert	PA	Wert	PA	Wert	PA
2	0.43121	(30.391)	0.31569	(61.320)	0.31299	(37.348)	1.58896	(97.729)
3	0.37942	(14.731)	0.24954	(27.517)	0.26363	(15.688)	1.24743	(55.230)
4	0.35848	(8.399)	0.22368	(14.302)	0.24905	(9.290)	1.10368	(37.341)
5	0.35279	(6.678)	0.21687	(10.822)	0.24447	(7.280)	1.04908	(30.547)
6	0.34888	(5.496)	0.21324	(8.967)	0.24042	(5.503)	1.02385	(27.407)
7	0.34304	(3.730)	0.20659	(5.569)	0.23671	(3.875)	1.00120	(24.589)
8	0.33903	(2.518)	0.20173	(3.086)	0.23403	(2.699)	0.97217	(20.976)
9	0.33957	(2.681)	0.20279	(3.627)	0.23382	(2.606)	0.96186	(19.693)
10	0.33726	(1.982)	0.20025	(2.329)	0.23260	(2.071)	0.95056	(18.287)
15	0.33461	(1.181)	0.19776	(1.057)	0.23056	(1.176)	0.93220	(16.003)
20	0.33295	(0.679)	0.19599	(0.153)	0.22944	(0.684)	0.88639	(10.302)
35	0.33238	(0.507)	0.19638	(0.352)	0.22897	(0.478)	0.86600	(7.765)
55	0.33190	(0.362)	0.19555	(-0.072)	0.22857	(0.303)	0.84386	(5.010)
75	0.33154	(0.253)	0.19536	(-0.169)	0.22838	(0.219)	0.83253	(3.600)
95	0.33182	(0.337)	0.19586	(0.086)	0.22857	(0.303)	0.82764	(2.991)

diesen unteren Bereich bei Fall 3, kann dort eine ungenaue Schätzung der Dichte das Ergebnis stark beeinflussen.

Vom Autor wurden die gleichen Berechnungen für Einkommensdaten der Jahre 1983–1990 durchgeführt, die hier aus Platzgründen nicht dargestellt werden. Die hierbei erzielten Resultate stimmten grundsätzlich mit denen für das Jahr 1991 erzielten überein.

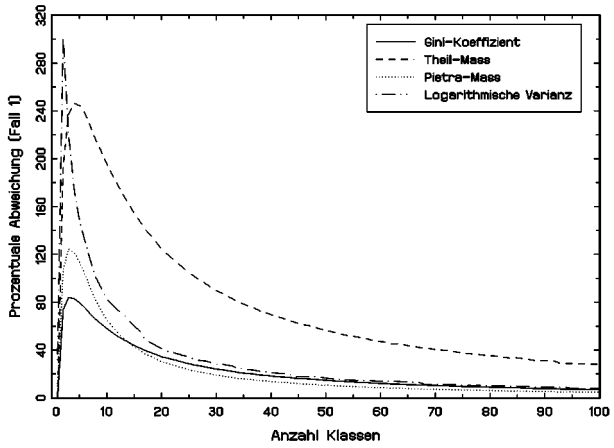
6 Fazit

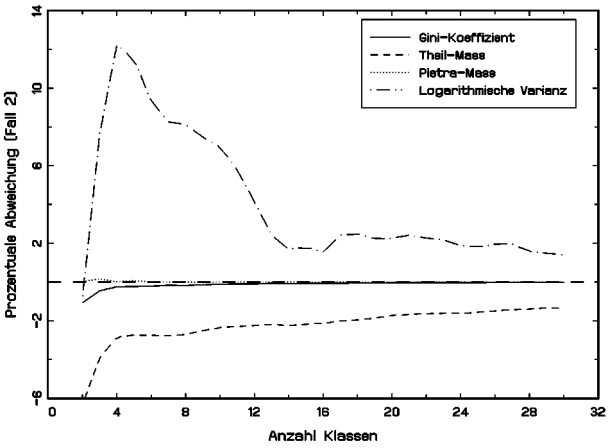
Das Konzept der entropiemaximalen Dichteschätzung scheint für den Fall der Bestimmung von Disparitätsmaßen aus klassierten Daten gut geeignet, sofern Informationen über Momente der wahren Verteilung bekannt sind. Auch bei nicht allzu großen Klassenzahlen können gute Annäherungen an die 'wahren' Disparitätswerte erzielt werden. Dabei erfordert die Berechnung keinen zu großen Aufwand. Eine Bestimmung von Parametern bei einer ML–Schätzung mittels eines nichtlinearen Optimierungsverfahrens (bei Anpassung einer 'komplexeren' Verteilung an die vorhandenen Daten) entfällt hier.

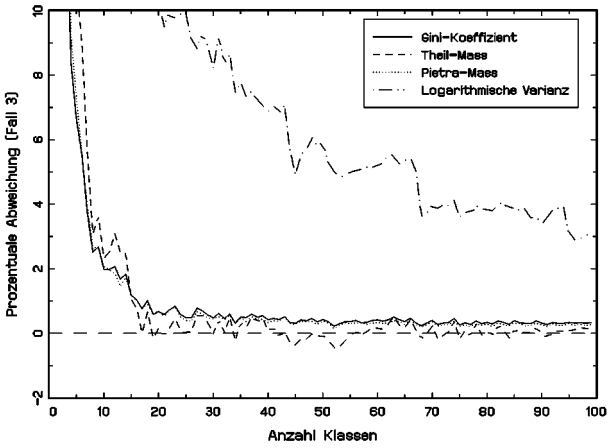
Vor allem für den in der Praxis häufig benutzten Gini–Koeffizienten, scheint sich die Anwendung des ME–Verfahrens zu lohnen. Bei größeren Klassenzahlen erzielen auch die anderen einfachen Verfahren gute Ergebnisse, wobei jedoch deren Anwendung auf Fall 2 (Kenntnis der bedingten Mittelwerte) beschränkt bleibt.

Stehen keine Informationen über Momente der wahren Verteilung zur Verfügung, so sollte auf einen parametrischen Ansatz zurückgegriffen werden (vgl. SCHADER/SCHMID (1988)).

Abb. 4: Prozentuale Abweichungen der geschätzten Disparitätsmaße von den 'wahren' Werten bei Anwendung der ME–Methode für die Fälle 1, 2 und 3







7 Anhang – Berechnung der ME–Dichtefunktionen

Fall 1:
Relative Häufigkeit p_i im Intervall $[a_i, b_i]$ bekannt.
Allgemeine Form der ME–Dichte:

$$f_i(x) = e^{\theta_0^{(i)}}$$

Erhaltungsbedingungen für $f_i(x)$:

$$1) \quad h_1(x) = 1 : \quad \int_{a_i}^{b_i} h_1(x) f_i(x) dx \quad = \quad \int_{a_i}^{b_i} f_i(x) dx = p_i$$

Bestimmung der Parameter:

$$\begin{aligned} \int_{a_i}^{b_i} f_i(x) dx = p_i \quad &\Longleftrightarrow \quad e^{\theta_0^{(i)}}(b_i - a_i) \quad = \quad p_i \\ &\Longleftrightarrow \quad \theta_0^{(i)} \qquad\qquad = \quad \ln \frac{p_i}{b_i - a_i} \end{aligned}$$

ME–Dichte:

$$f_i(x) = \frac{p_i}{b_i - a_i}$$

Fall 2:

a) Bedingter Mittelwert μ_i im Intervall $[a_i, b_i]$ bekannt

b) Relative Häufigkeit p_i im Intervall $[a_i, b_i]$ bekannt

Allgemeine Form der ME–Dichte:

$$f_i(x) = e^{\theta_0^{(i)} + \theta_1^{(i)} x}$$

Erhaltungsbedingungen für $f_i(x)$:

$$1) \quad h_1(x) = 1: \quad \int_{a_i}^{b_i} h_1(x) f_i(x) dx = \int_{a_i}^{b_i} f_i(x) dx = p_i$$

$$2) \quad h_2(x) = x: \quad \int_{a_i}^{b_i} h_2(x) f_i(x) dx = \int_{a_i}^{b_i} x f_i(x) dx = p_i \mu_i$$

Bestimmung der Parameter:
zu 1):

$$\begin{aligned}
& \int_{a_i}^{b_i} f_i(x) dx && = p_i \\
\iff & \int_{a_i}^{b_i} e^{\theta_0^{(i)} + \theta_1^{(i)} x} dx && = p_i \\
\iff & e^{\theta_0^{(i)}} \int_{a_i}^{b_i} e^{\theta_1^{(i)} x} dx && = p_i \\
\iff & \frac{e^{\theta_0^{(i)}}}{\theta_1^{(i)}} e^{\theta_1^{(i)} x} \Big|_{a_i}^{b_i} && = p_i \\
\iff & \frac{e^{\theta_0^{(i)}}}{\theta_1^{(i)}} (e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}) && = p_i \\
\iff & e^{\theta_0^{(i)}} && = \frac{\theta_1^{(i)} p_i}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}} \\
\iff & \theta_0^{(i)} && = \ln \frac{\theta_1^{(i)} p_i}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}}
\end{aligned}$$

zu 2):
Produktintegration (mit $v = x; v' = 1; u' = e^{\theta_1^{(i)} x}; u = (1/\theta_1^{(i)}) e^{\theta_1^{(i)} x}$):

$$\begin{aligned}
& \int_{a_i}^{b_i} x \frac{f_i(x)}{p_i} dx && = \mu_i \\
\iff & \int_{a_i}^{b_i} x e^{\theta_0^{(i)} + \theta_1^{(i)} x} dx && = p_i \mu_i \\
\iff & e^{\theta_0^{(i)}} \left(\left[\frac{x}{\theta_1^{(i)}} e^{\theta_1^{(i)} x} \right]_{a_i}^{b_i} - \frac{1}{\theta_1^{(i)}} \int_{a_i}^{b_i} e^{\theta_1^{(i)} x} dx \right) && = p_i \mu_i \\
\iff & e^{\theta_0^{(i)}} \left(\left[\frac{x}{\theta_1^{(i)}} e^{\theta_1^{(i)} x} - \frac{1}{\theta_1^{(i)2}} e^{\theta_1^{(i)} x} \right]_{a_i}^{b_i} \right) && = p_i \mu_i \\
\iff & e^{\theta_0^{(i)}} \left(\frac{b_i}{\theta_1^{(i)}} e^{\theta_1^{(i)} b_i} - \frac{1}{\theta_1^{(i)2}} e^{\theta_1^{(i)} b_i} - \frac{a_i}{\theta_1^{(i)}} e^{\theta_1^{(i)} a_i} + \frac{1}{\theta_1^{(i)2}} e^{\theta_1^{(i)} a_i} \right) && = p_i \mu_i
\end{aligned}$$

Einsetzen von $\theta_0^{(i)}$ in obige Gleichung:

$$\begin{aligned}
&\iff \frac{\theta_1^{(i)} p_i}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}} \left(\frac{b_i}{\theta_1^{(i)}} e^{\theta_1^{(i)} b_i} - \frac{1}{\theta_1^{(i)2}} e^{\theta_1^{(i)} b_i} - \frac{a_i}{\theta_1^{(i)}} e^{\theta_1^{(i)} a_i} + \frac{1}{\theta_1^{(i)2}} e^{\theta_1^{(i)} a_i} \right) &= p_i \mu_i \\
&\iff \frac{[(b_i - \frac{1}{\theta_1^{(i)}}) e^{\theta_1^{(i)} b_i} - (a_i - \frac{1}{\theta_1^{(i)}}) e^{\theta_1^{(i)} a_i}]}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}} &= \mu_i \quad (1)
\end{aligned}$$

ME–Dichte:

$$f_i(x) = e^{\theta_0^{(i)}} \cdot e^{\theta_1^{(i)} x} = p_i \cdot \frac{\theta_1^{(i)} e^{\theta_1^{(i)} x}}{e^{\theta_1^{(i)} b_i} - e^{\theta_1^{(i)} a_i}}$$

Dabei ist $\theta_1^{(i)}$ numerisch über eine Nullstellenberechnung aus der Gleichung (1) zu bestimmen.

Fall 3:

a) Unbedingter Mittelwert μ bekannt

b) Relative Häufigkeit p_i im Intervall $[a_i, b_i]$ bekannt für alle $i = 1, \ldots, K$

Allgemeine Form der ME–Dichte:

$$f_i(x) = e^{\theta_0^{(i)} + \theta_1 x}$$

Erhaltungsbedingungen für $f_i(x)$:

$$\begin{aligned}
&1) \quad h_1(x) = 1: \quad \int_{a_i}^{b_i} h_1(x) f_i(x) dx &= \int_{a_i}^{b_i} f_i(x) dx = p_i \\
&2) \quad h_2(x) = x: \quad \sum_i^K \int_{a_i}^{b_i} h_2(x) f_i(x) dx &= \sum_i^K \int_{a_i}^{b_i} x f_i(x) dx = \mu
\end{aligned}$$

Bestimmung der Parameter:
zu 1): siehe Fall 2:

$$\theta_0^{(i)} = \ln \frac{\theta_1 p_i}{e^{\theta_1 b_i} - e^{\theta_1 a_i}}$$

zu 2):

$$\begin{aligned}
&\sum_i \int_{a_i}^{b_i} x e^{\theta_0^{(i)} + \theta_1 x} dx &= \mu \\
\iff &\sum_i e^{\theta_0^{(i)}} \left(\frac{b_i}{\theta_1} e^{\theta_1 b_i} - \frac{1}{\theta_1^2} e^{\theta_1 b_i} - \frac{a_i}{\theta_1} e^{\theta_1 a_i} + \frac{1}{\theta_1^2} e^{\theta_1 a_i} \right) &= \mu \\
\iff &\sum_i p_i \frac{[(b_i - \frac{1}{\theta_1}) e^{\theta_1 b_i} - (a_i - \frac{1}{\theta_1}) e^{\theta_1 a_i}]}{e^{\theta_1 b_i} - e^{\theta_1 a_i}} &= \mu \quad (2)
\end{aligned}$$

ME–Dichte:

$$f_i(x) = e^{\theta_0^{(i)}} \cdot e^{\theta_1 x} = p_i \cdot \frac{\theta_1 e^{\theta_1 x}}{e^{\theta_1 b_i} - e^{\theta_1 a_i}}$$

Dabei ist θ_1 numerisch über eine Nullstellenberechnung aus der Gleichung (2) zu bestimmen.

Literatur

[1] **Cowell, F. A. (1995):**

Measuring Inequality, London School of Economics and Political Science, Second Edition.

[2] **Foster, James E. (1983):**

An axiomatic characterization of the Theil measure of income inequality, in: Journal of Economic Theory 31, S. 105 - 121.

[3] **Gastwirth, Joseph L. (1972):**

The estimation of the Lorenz Curve and Gini Index, in: Review of Economics and Statistics 54, S. 306 - 316.

[4] **Golani, B. / Phillips, F.Y. (1990):**

A maximum–entropy based heuristic for density estimation from data in grouped form, in: Decision Science 21, S. 862 - 881.

[5] **Kagan, A. M. / Linnik, Yu. V. / Rao, C. Radhakrishna (1973):**

Characterization Problems in Mathematical Statistics, New York.

[6] **Leipnik, Roy B. (1990):**

A maximum relative entropy principle for distribution of personal income with derivations of several known income distributions, in: Communication in Statistics 19, S. 1003 - 1036.

[7] **Lüthi, Ambros (1981):**

Messung wirtschaftlicher Ungleichheit, Dissertation, Lecture Notes in Economics and Mathematical Systems 189, Berlin.

[8] **Piesch, Walter (1975):**

Statistische Konzentrationsmaße - Formale Eigenschaften und verteilungstheoretische Zusammenhänge, Tübingen.

[9] **Rao, Calyampudi Radhakrishna (1973):**

Linear Statistical Inference and Its Applications, New York.

[10] **Schader, Martin / Schmid, Friedrich (1988):**

Zur Messung der relativen Konzentration aus gruppierten Daten - Ein Vergleich parametrischer und nichtparametrischer Ansätze, in: Jahrbücher für Nationalökonomie und Statistik, Vol. 204/5, S. 437 - 455.

[11] **Schmid, Friedrich (1991):**

Zur Sensitivität von Disparitätsmaßen, in: Allgemeines Statistisches Archiv 75, S. 155 - 167.

[12] **Shannon, Claude E. (1948):**

The mathematical theory of communication, in: Bell System Technical Journal, July and October 1948.

[13] **Theil, Henri (1967):**

Economics and Information Theory, Studies in Mathematical and Managerial Economics, Vol. 7, Amsterdam.

[14] **Zellner, A. / Highfield, R. (1988):**

Calculation of maximum entropy distributions and approximation of marginal posterior distributions, in: Journal of Econometrics 37, S. 195 - 209.