Friedrich-Alexander-Universität Erlangen-Nürnberg

Wirtschafts-und Sozialwissenschaftliche Fakultät

Diskussionspapier
27 / 1999

Rangordnungsstatistiken als Verteilungsmaßzahlen für ordinalskalierte Merkmale

I. Streuungsmessung

Ingo Klein

Lehrstuhl für Statistik und Ökonometrie
Lehrstuhl für Statistik und empirische Wirtschaftsforschung
Lange Gasse 20 · D-90403 Nürnberg

RANGORDNUNGSSTATISTIKEN ALS VERTEILUNGSMASSZAHLEN FÜR ORDINALSKALIERTE MERKMALE

I. Streuungsmessung

Ingo Klein

Lehrstuhl für Statistik und Ökonometrie

Universität Erlangen–Nürnberg

Lange Gasse 20

D–90403 Nürnberg

Germany

E-mail: ingo.klein@wiso.uni-erlangen.de

Abstract

Well–known measures of dispersion für ordinal variables are considered. It will be shown that these measures can be represented as linearer rank statistics with special concave scores. Two further measures are discussed. One that can be derived from a special ordering of dispersion and one with convex scores. The first and the measure of Gini will be represented in an equivalent way with a very simple interpretation. This interpretation is based on the amount of frequency manipulations that is necessary to get minimal dispersion. These measures are related to the L_1– and L_2–distance. The problem of normalization such that the meaures have the value 1 as upper bound is raised.

1 Einleitung

Streuungsmaáe werden in der Literatur zumeist für quantitative Merkmale angegeben. Dies zeigt sich bereits daran, daá diese Maßzahlen von Differenzen von Quantilen, Differenzen von Merkmalsausprägungen und Mittelwerten und Differenzen von Teilmittelwerten ausgehen. Die gängigen Charakterisierungen von Streuungsmaßen setzen sogar stetige Verteilungsfunktionen voraus (siehe z.B. die šbersichten in Oja (1981) und in Handl (1986)). Streuungsmaße für nominalskalierte Merkmale werden z.B. in der Monographie von Uschner (1987) intensiv diskutiert und Streuungsmaße für ordinalskalierte Merkmale werden vor allem von Vogel & Dobbener (1982) und Vogel (1991) behandelt. Insbesondere in Vogel (1991) sind eine ganze Reihe von Streuungsmaßen zusammengetragen worden.

Wir wollen im folgenden versuchen, für die von Vogel genannten Vorschläge für Streuungsmaße einen einheitlichen Zugang anzubieten, der insbesondere die axiomatischen Anforderungen an Streuungsmaße berücksichtigt und auch eigene Streuungsmaße vorschlagen, die zum Teil in abgewandelter Form bereits aus der nichtparametrischen Statistik bekannt sind und den besonderen Anforderungen der deskriptiven Statistik Rechnung tragen. [1]

2 Formaler Rahmen

Sei E eine endliche Grundgesamtheit des Umfangs n und U ein ordinalskaliertes Merkmal mit abzählbar vielen geordneten Ausprägungen $u_i < u_{i+1}$ für $i \in \mathbb{N}$. n_i bezeichne die absolute Häufigkeit mit der die i–te Merkmalsausprägung in der Grundgesamtheit E vorkommt für $i \in \mathbb{N}$. Es sei $n = \sum_i n_i$. Dann sind $f_i = n_i/n$ die relativen Häufigkeiten, $R_i = \sum_{j \le i} n_i$ die kumulierten absoluten und $F_i = R_i/n$ die kumulierten relativen Häufigkeiten für $i \in \mathbb{N}$. Ist die Anzahl der möglichen Merkmalsausprägungen endlich ($= k$), so sind $R_k = n$ und $F_k = 1$.

In der nichtparametrischen Statistik wird

$$\sum_{i=1}^{k} c_i a^*(R_i) = \sum_{i=1}^{k} c_i a(F_i)$$

[1] Für Korrekturen und Hinweise danke ich Herrn Diplom–Mathematiker Hans Kiesl, Lehrstuhl für Statistik, Universität Bamberg.

mit $a(F_i) = a^*(R_i)$, $i = 1, 2, \ldots, k$ als lineare Rangordnungsstatistik bezeichnet. $c_1, \ldots, c_k$ heißen Regressionskoeffizienten und $a^*(.)$ bzw. $a(.)$ Scoresfunktion (vgl. z.B. Hajek & Sidak (1967), S. 57ff.).

3 Streuung für ordinalskalierte Merkmale

Die elementaren Anforderungen an ein Streuungsmaß betreffen zunächst die Extrempunkte, wann keine und wann eine maximale Streuung vorliegen soll.

3.1 Minimale Streuung

Keine Streuung liegt offenbar dann vor, wenn eine der potentiellen Merkmalsausprägungen die gesamte Masse auf sich vereint. D.h. es gibt ein $j \in \mathbb{N}$ mit

$$F_i = 0 \text{ für } i = 1, 2, \ldots, j \text{ und } F_i = 1 \text{ für } i = j + 1, \ldots$$

3.2 Maximale Streuung bei fester Anzahl der Merkmalsausprägungen

Sei $k < \infty$ die Anzahl möglicher Merkmalsausprägungen, dann tritt der Fall der maximale Streuung auf, wenn nur die beiden extremen Merkmalsausprägungen u_1 und u_k die gesamte Häufigkeitsmasse auf sich vereinen und zwar möglichst im Sinne einer Gleichverteilung. D.h.

$$F_i = 1/2 \text{ oder } 1 - 2F_i = 0 \text{ oder } 1 - F_i = F_i \text{ für } i = 1, 2, \ldots, k - 1.$$

Dies ist jedoch nur möglich, wenn n gerade ist. Für ungerades n wird diejenige Verteilung F_i, $i = 1, 2, \ldots, k-1$ gesucht, die der genannten Gleichverteilung am ähnlichsten ist. Dies ist für

$$F_i = 1/2 \pm 1/(2n) \text{ für } i = 1, 2, \ldots, k - 1$$

der Fall, wenn einer der gängigen Abstandsmaße, wie sie im Abschnitt 4.2 behandelt werden, verwendet wird.

3.3 Funktionale Abhängigkeit

Bezeichne S im folgenden ein Verteilungsmaß. Ein solches Maß kann nun von den Vektoren der Merkmalsausprägungen $u = (u_1, u_2, \ldots)$, der Häufigkeiten $f = (f_1, f_2, \ldots)$ und n abhängen. Wir schreiben kurz $S(u, f, n)$. Es läßt sich nun nachweisen, daß S nur von den kumulierten absoluten oder relativen Häufigkeiten abhängen kann, wenn als geringe Forderung die sog. Vergleichsinvarianz postuliert wird. Diese spezielle Invarianz bedeutet, daß

$$S(u, f, n) = S(u', f', n') \iff S(g(u), f, n) = S(g(u'), f', n')$$

für alle zulässigen Skalentransformationen g gelten muß. Zulässige Skalentransformationen sind im Falle der Ordinalskala die streng monoton zunehmenden, stetigen Transformationen.

Klein ((1994), S. 179) hat gezeigt, wenn für eine beliebige Maßzahl die Vergleichsinvarianz gilt, und es mindestens eine streng monotone Transformation g gibt, so daß

$$S(u, f, n) = S(g(u), f, n)$$

ist, diese Maßzahl S nur von den kumulierten relativen Häufigkeiten $F = (F_1, F_2, \ldots)$ und n abhängen kann. Streuungsmaße sind aber zumindest invariant gegenüber Translationen, so daß die vorstehende Bedingung offensichtlich erfüllt ist. Wir schreiben deshalb im folgenden kurz $S(F, n)$. Da sich die Form der Verteilung nicht ändert, solange F unverändert bleibt, darf S nicht von n abhängen, so daß im folgenden lediglich $S(F)$ betrachtet wird.

3.4 Partialordnung der Streuung bei endlicher Anzahl von Ausprägungen

Ausgangspunkt für jede Messung ist die Definition einer geeigneten Ordnung von Verteilungen unterschiedlicher Streuung. Geht man von einer endlichen Anzahl k von Merkmalsausprägungen aus, so liegt es nahe, zwei k–dimensionale Vektoren kumulierter relativer Häufigkeiten F und F' entsprechend der folgenden Relation zu ordnen: F streut nicht mehr als F' (kurz: $F \preceq_S F$), wenn

$$|F_i - 1/2| \geq |F'_i - 1/2|$$

für $i = 1, 2, \ldots, k-1$ gilt. Dies bedeutet, daß F' von dem Extremum der maximalen Streuung "nicht weiter entfernt" ist als F.

Von Streuungsmaßen für ordinalskalierte Merkmale ist zu verlangen, daß sie diese sehr spezielle Ordnung erhalten. D.h. es gilt für ein Streuungsmaß S:

$$F \preceq_S F' \Longrightarrow S(F) \leq S(F').$$

3.5 Spiegelung der Verteilung

Bei Permutationen der Komponenten des k–dimensionalen Vektors der relativen Häufigkeiten $f = (f_1, \ldots, f_k)$ wird sich im Regelfall die Streuung des ordinalskalierten Merkmals verändern. Dies gilt allerdings nicht, wenn statt f die gespiegelte Verteilung $\overline{f} = (f_k, f_{k-1}, \ldots, f_1)$ betrachtet wird, zu der der Vektor

$$\overline{F} = (\overline{F}_1, \ldots, \overline{F}_{k-1}, 1) = (1 - F_{k-1}, \ldots, 1 - F_1, 1)$$

der kumulierten relativen Häufigkeiten gehört. Das Streuungsphänomen bleibt von dieser Spiegelung unberührt, da im wesentlichen die Form der Verteilung erhalten bleibt. Dies muß ein geeignetes Streuungsmaß berücksichtigen.

3.6 Ein allgemeines Konstruktionsprinzip

Es ist zunächst zu klären, was unter einem (additiven) Streuungsmaß für ordinalskalierte Merkmale zu verstehen ist. Wir beschränken uns zunächst auf eine feste Anzahl k von potentiellen Merkmalsausprägungen.

Definition 3.1 *Seien $\mathcal{F}_k$ die Menge aller Vektoren von kumulierten relativen Häufigkeiten für k Merkmalsausprägungen. $S : \mathcal{F}_k \to \mathbb{R}$ heißt Streuungsmaß, wenn*

1. *$S(F) = 0$, falls $F = (F_1, \ldots, F_k)$ mit $F_i \in \{0, 1\}$ für $i = 1, 2, \ldots, k-1$,*
2. *$F \preceq_S F' \Longrightarrow S(F) \leq S(F')$ für $F, F' \in \mathcal{F}_k$,*
3. *$SO(F) = SO(F_1, \ldots, F_{k-1}, 1) = SO(1 - F_{k-1}, \ldots, 1 - F_1, 1) = SO(\overline{F})$ für $F \in \mathcal{F}_k$.*

gelten.

Wir betrachten hier ausschließlich additive Maßzahlen der Form

$$S(F) = \sum_{i=1}^{k-1} a(F_i)$$

mit einer sog. Scoresfunktion $a(.)$. $S(F)$ ist Spezialfall der eingangs eingeführten linearen Rangordnungsstatistik mit $n_i = 1$, $c_i = 1$, $a^*(i) = a(i/k)$, $i = 1, 2, \ldots, k-1$ und $n_k = n - \sum_{i=1}^{k-1} n_i$, $c_k = 0$, $a^*(k) = a(1)$.

Im weiteren sollen für die Scoresfunktion $a(.)$ die folgenden Annahmen getroffen werden:

(1) $a : [0,1] \to [0,\infty)$,

(2) a ist stetig im Intervall $[0,1]$,

(3) a ist symmetrisch, d.h. $a(p) = a(1-p)$ für $p \in [0,1]$,

(4) a ist streng monoton zunehmend auf dem Intervall $[0,1/2]$,

(5) $a(0) = 0$.

Insbesondere nimmt dann $a(.)$ an der Stelle $p = 1/2$ ein eindeutiges Maximum an. Dies bedeutet aber sofort, daß auch $S(F)$ maximal ist für $F_i = 1/2$, $i = 1, 2, \ldots, k-1$. Als Wertebereich gilt $0 \leq S(F) \leq (k-1)a(1/2)$.

Beispiele, wie $a(.)$ gewählt werden kann, sind u.a. Vogel (1991) zu entnehmen. Wir numerieren die verschiedenen Scoresfunktion entsprechend den bei Vogel genannten Streuungsmaßen:

- $a_1(p) = -p\mathrm{ld}p - (1-p)\mathrm{ld}(1-p)$ (siehe auch Vogel (1981))
- $a_2(p) = p(1-p)$
- $a_3(p) = 1 - (p^2 + (1-p)^2) = 2p(1-p)$
- $a_4(p) = 1 - p^p(1-p)^{1-p}$
- $a_6(p) = 1 - p^{-p}(1-p)^{-(1-p)} = 1 - 2^{a_1(p)}$

Als weitere Scoresfunktionen sollen im folgenden

- $a_7(p) = 1/2 - |p - 1/2|$
- $a_8(p) = \exp(-|p - 1/2|) - \exp(-1/2)$

betrachtet werden.

Alle diese Funktionen erfüllen offensichtlich die geforderten Eigenschaften (1) bis (5). [2]

Wenn $a(.)$ die Eigenschaften (1) bis (5) besitzt, dann kann gezeigt werden, daß die zugehörige additive Maßzahl S ein Streuungsmaß ist.

Theorem 3.1 *Sei $S : \mathcal{F}_k \to \mathbb{R}$ eine (additive) Maßzahl mit Scoresfunktion $a(.)$, die die Eigenschaften (1) bis (5) erfüllt, dann ist S ein Streuungsmaß für ordinalskalierte Merkmale mit k Merkmalsausprägungen.*

Beweis:

1. Sei $F_i = 1$ für $i = j + 1, \ldots, k$ und $F_i = 0$ für $i = 1, 2, \ldots, j$, dann ist $a(F_i) = 0$ für $i = 1, 2, \ldots, k - 1$ und damit $S(F) = 0$.

2. Zu zeigen bleibt für zwei Vektoren F und F' kumulierter relativer Häufigkeiten

$$F \preceq_S F' \Longrightarrow S(F) \leq S(F').$$

   Es ist

$$F \preceq_S F' \iff |F_i - 1/2| \geq |F_i' - 1/2| \text{ für } i = 1, 2, \ldots, k - 1,$$

   womit es reicht, nachzuweisen, daß

$$|F_i - 1/2| \geq |F_i' - 1/2| \iff a(F_i) \leq a(F_i')$$

   für $i = 1, 2, \ldots, k - 1$ gilt. Dazu nehmen wir eine Fallunterscheidung vor:

   2.1 Sei $F_i \leq 1/2$ und $F_i' \leq 1/2$, dann ist

$$\begin{aligned} |F_i - 1/2| \geq |F_i' - 1/2| &\iff 1/2 - F_i \geq 1/2 - F_i' \\ &\iff F_i \leq F_i' \\ &\iff a(F_i) \leq a(F_i'), \end{aligned}$$

   da $a(.)$ monoton wachsend auf $[0, 1/2]$ ist.

[2] Die Funktion $a_5(.)$ fehlt, da sie von Vogel (1991) für ein multiplikativ verknüpftes Streuungsmaß verwendet wird.

2.2 Sei $F_i > 1/2$ und $F_i' > 1/2$, dann ist

$$\begin{aligned} |F_i - 1/2| \geq |F_i' - 1/2| &\iff F_i - 1/2 \geq F_i' - 1/2 \\ &\iff F_i \geq F_i' \\ &\iff a(F_i) \leq a(F_i'), \end{aligned}$$

da $a(.)$ monoton fallend auf $[1/2, 1]$ ist.

2.3 Sei $F_i \leq 1/2$ und $F_i' > 1/2$, dann ist

$$\begin{aligned} |F_i - 1/2| \geq |F_i' - 1/2| &\iff 1/2 - F_i \geq F_i' - 1/2 \\ &\iff F_i \leq 1 - F_i' \\ &\iff a(F_i) \leq a(1 - F_i') = a(F_i'), \end{aligned}$$

da $1 - F_i' < 1/2$, $a(.)$ monoton wachsend auf $[0, 1/2]$ und a symmetrisch auf $[0, 1]$ sind.

2.4 Sei $F_i > 1/2$ und $F_i' \leq 1/2$, dann ist

$$\begin{aligned} |F_i - 1/2| \geq |F_i' - 1/2| &\iff F_i - 1/2 \geq 1/2 - F_i' \\ &\iff F_i \geq 1 - F_i' \\ &\iff a(F_i) \leq a(1 - F_i') = a(F_i'), \end{aligned}$$

da $1 - F_i' > 1/2$, $a(.)$ monoton fallend auf $[1/2, 1]$ und a symmetrisch auf $[0, 1]$ sind.

3. Sei $\overline{F}$ die Spiegelung von F. Dann ist

$$\begin{aligned} SO(\overline{F}) &= \sum_{i=1}^{k-1} a(\overline{F_i}) \\ &= \sum_{i=1}^{k-1} a(1 - F_{k-i}) \\ &= \sum_{i=1}^{k-1} a(F_{k-i}) \\ &= \sum_{i=1}^{k-1} a(F_i) = SO(F), \end{aligned}$$

wegen $a(p) = a(1 - p)$ für $p \in [0, 1]$. $\square$

Der vorstehende Satz läßt eine große Freiheit der Konstruktion von Streuungsmaßen für ordinalskalierte Merkmale, die sich in der Vielzahl der von Vogel gemachten Vorschläge niederschlägt. Vogel ((1991), S. 309) bemerkt, daß es offensichtlich nicht möglich sei, Streuung unabhängig von der verwendeten Maßzahl zu definieren, so daß jedes Maß eine eigene Gewichtung der Abweichungen von den genannten Extremsituationen vornimmt.

3.7 Einige ausgewählte Streuungsmaße

Wenn für ordinalskalierte Merkmale ein Streuungsmaß nur von den kumulierten Häufigkeiten abhängen kann, dann bietet es sich an, die oben angegebene Charakterisierung der Situationen minimaler und maximaler Streuung unmittelbar für ein Streuungsmaß zu verwenden.

Geht man von $|F_i - 1/2|$ aus, so definiert

$$S_7(F) = \sum_{i=1}^{k-1} a_7(F_i) = \sum_{i=1}^{k-1} 1/2 - |F_i - 1/2|$$

ein Streuungsmaß, das unmittelbar von der Partialordnung $\preceq_S$ ausgeht und als Maximum den Wert $(k-1)/2$ annimmt. [3]

Dieses Streuungsmaß ist eng verwandt zu linearen Rangordnungsstatistiken, die für Tests auf Skalenalternativen verwendet werden. So basiert laut Hajek & Sidak ((1967), S. 95) ein für die Dichte

$$f_X(x) = 1/2(1 + |x|)^{-1} \text{ für } x \in \mathbb{R},$$

asymptotisch optimaler Test (sog. Ansari–Bradley–Test) auf der Statistik

$$\sum_{i=1}^{n} c_i 2|2R_i - 1| = 4 \sum_{i=1}^{n} c_i |R_i - 1/2|$$

wobei R_i, $i = 1, 2, \ldots, n$ die Rangstatistiken einer Zufallsstichprobe $X_1, \ldots, X_n$ sind.

Eine weitere naheliegende Formulierung eines Streuungsmaßes ist

$$S_2(F) = \sum_{i=1}^{k-1} a_2(F_i) = \sum_{i=1}^{k-1} F_i(1 - F_i)$$

[3] Dieses Streuungsmaß wurde in etwas abgewandelter Form auch von Vogel (1994) betrachtet.

mit $0 \leq S_2 \leq (k-1)/4$. Von diesem Streuungsmaß wird noch nachgewiesen, daß es für bestimmte quantitative Merkmale mit dem Streuungsmaß von Gini übereinstimmt (siehe auch die Literaturhinweise in Vogel (1991), S. 304). Es besitzt ohnehin sehr viel Žhnlichkeit mit dem Streuungsmaß für qualitative Merkmale

$$\sum_{i=1}^{k} f_i(1-f_i),$$

das ebenfalls Gini zugesprochen wird und das im Mittelpunkt der Arbeit von Uschner (1987) steht.

Vogel & Dobbener (1981) haben ein Streuungsmaß für ordinalskalierte Merkmale auf der Basis der Entropie vorgeschlagen. Dieses lautet

$$S_1(F) = \sum_{i=1}^{k-1} a_1(F_i) = \sum_{i=1}^{k-1} \left(-F_i \mathrm{ld} F_i - (1-F_i)\mathrm{ld}(1-F_i)\right).$$

Der Wertebereich ist $0 \leq S_1 \leq k-1$.

Die von Vogel (1991) implizit verwendeten Scoresfunktionen sind sämtlich streng konkav. Dies erklärt auch, daß jeweils auf sehr ähnliche Weise die Streuung in den betrachteten Beispielen quantifiziert wird. Trotzdem sind die Streuungsmaße nicht konsistent. Konsistenz bedeutet dabei, daß durch die Maßzahlen jeweils dieselbe Ordnung von Grundgesamtheiten induziert wird.

Die Inkonsistenz wird besonders deutlich, wenn das Streuungsmaß S_7 und die von Vogel (1991) betrachteten Streuungsmaße mit streng konkaver Scoresfunktion betrachtet werden. Es gibt Situationen, in denen S_7 dieselben Streuungswerte produziert, ohne daß dies für die Streuungsmaße S_1 bis S_6 gilt. Eine Umkehrung der Streuungsbewertung ist sogar möglich, wenn als Streuungsmaß

$$S_8(F) = \sum_{i=1}^{k-1} a_8(F_i) = \sum_{i=1}^{k-1} \left(\exp(-|F_i - 1/2|) - \exp(-1/2)\right)$$

betrachtet wird, dessen Scoresfunktion streng konvex auf $[0, 0.5]$ ist. Diese Problematik illustriert das folgende Beispiel:

Beispiel 3.1 *Vogel (1991) hat Vektoren absoluter Häufigkeiten für $k = 5$ Merkmalsausprägungen $n = 10$ Beobachtungen betrachtet und für einige die von ihm diskutierten*

Streuungsmaße berechnet. Viele dieser Häufigkeitsvektoren lassen sich bezüglich $\preceq_S$ *ordnen. So ist*

$$F = (0.8, 1, 1, 1, 1) \preceq_S F' = (0.8, 0.9, 1, 1, 1),$$

da $|F_i - 1/2| \geq |F'_i - 1/2|$ *für* $i = 1, 2, 3, 4$*, wie die nachstehende Tabelle zeigt:*

i	*1*	*2*	*3*	*4*
F_i	*0.8*	*1.0*	*1.0*	*1.0*
$\lvert F_i - 1/2\rvert$	*0.3*	*0.5*	*0.5*	*0.5*
F'_i	*0.8*	*0.9*	*1.0*	*1.0*
$\lvert F'_i - 1/2\rvert$	*0.3*	*0.4*	*0.5*	*0.5*

Betrachtet man jedoch zwei weitere Verteilungen F *bzw.* F' *mit den absoluten Häufigkeiten* $F = (0.8, 0.9, 0.9, 1, 1)$ *und* $F' = (0.8, 0.8, 1, 1, 1)$*, dann ist* $F_2 > F'_2$ *und* $F_3 < F'_3$*, wie die folgende Tabelle zeigt.*

i	*1*	*2*	*3*	*4*
F_i	*0.8*	*0.9*	*0.9*	*1.0*
$\lvert F_i - 1/2\rvert$	*0.3*	*0.4*	*0.4*	*0.5*
F'_i	*0.8*	*0.8*	*1.0*	*1.0*
$\lvert F' - 1/2\rvert$	*0.3*	*0.3*	*0.5*	*0.5*

Die Streuungsmaße verhalten sich nun in dieser inkompatiblen Situation gänzlich unterschiedlich. Dies liegt am Krümmungsverhalten der Scoresfunktion $a(.)$ *im Bereich* $[1/2, 1]$*. Ist* $a(.)$ *dort streng konkav, so ist* $a(0.8) - a(0.9) < a(0.9) - a(1.0)$*, so daß für* F *ein größerer Wert des zugehörigen Streuungsmaßes ausgewiesen wird als für* F'*. Dieses Verhalten zeigen sämtliche von Vogel (1991) betrachteten Maßzahlen. Ist hingegen* $a(.)$ *im Intervall* $[1/2, 1]$ *streng konvex, so besitzt* F *eine kleinere Streuung als* F'*. Dies betrifft das Streuungsmaß* S_8*. Im linearen Falle hingegen (z.B.* a_7*) werden die beiden Verteilungssituationen als streuungsäquivalent eingestuft. Es bedarf also einer weitergehenden Axiomatik des Krümmungsverhaltens der zu verwendenden Scoresfunktion, um die Menge der möglichen Streuungsmaße für ordinalskalierte Merkmale weiter einzuschränken.*

3.8 Weitere Streuungsmaße

Transformiert man ein Streuungsmaß mit einer streng monoton zunehmenden Transformation, die den Nullpunkt festhält, so ergibt sich offensichtlich ein neues Streuungsmaß, das wiederum eine andere Gewichtung des Abstandes zur Einpunktverteilung vornimmt. Vogel ((1991), S. 305) betrachtet

$$S_5(F) = 1 - \prod_{i=1}^{k-1} F_i^{F_i}(1 - F_i)^{1-F_i}.$$

Es ist aber

$$S_5(F) = 1 - 2^{-\sum_{i=1}^{k-1} a_1(F_i)} = 1 - 2^{-S_1(F)}$$

eine streng monoton zunehmende Transformation von $S_1(F)$.

Es ist mithin wichtig, die zu empfehlenden Streuungsmaße durch informelle Eigenschaften einzuschränken. Uschner ((1987), S. 45) nennt hier die Anschaulichkeit und leichte Berechenbarkeit. In diesem Sinne sollen im folgenden zwei der genannten Streuungsmaße eingehender diskutiert werden.

4 Streuungsmaße auf der Basis einer Austauschoperation

4.1 Häufigkeitsaustausch

Es sei ein Vektor $n = (n_1, \ldots, n_k)$ absoluter Häufigkeiten gegeben. Unter einer Häufigkeitsumschichtung verstehen wir eine Operation, die $n_i > 0$ für ein geeignetes $i \in \{1, 2, \ldots, k\}$ um Eins reduziert und die benachbarte Häufigkeit n_{i+1} oder n_{i-1} um Eins erhöht. Durch sukzessive Anwendung der Häufigkeitsumschichtung kann n z.B. derart verändert werden, daß der Fall einer Einpunktverteilung vorliegt.

Betrachtet man

$$U_{ij} = |i - j| f_j$$

dann mißt

$$nU_{ij} = |i - j| n_j$$

gerade die Anzahl der Häufigkeitsumschichtungen, die nötig sind, um n_j auf Null zu reduzieren und n_i um n_j zu steigern. Dann gibt

$$nU_i = n\sum_{j=1}^{k} U_{ij} = \sum_{j=1}^{k} |i-j|n_j$$

die Anzahl von Häufigkeitsumschichtungen an, die benötigt werden, um eine Einpunktverteilung zu erzeugen mit $F_j = 0$, $j = 1, 2, \ldots, i-1$ und $F_j = 1$, $j = i, i+1, \ldots, k$.

Als mögliche Streuungsmaße lassen sich die minimale Anzahl von Umschichtungen betrachten, die bis zu einer Einpunktverteilung benötigt werden:

$$\min_i U_i = \min_i \sum_{j=1}^{k} |i-j|f_j.$$

Eine Alternative bietet die durchschnittliche Anzahl von Umschichtungen bis zur Einpunktverteilung:

$$\sum_{i=1}^{k} U_i f_i = \sum_{i=1}^{k}\sum_{j=1}^{k} |i-j|f_j f_i.$$

Es läßt sich nun zeigen, daß diese Maßzahlen mit den bereits betrachteten Maßen S_7 und bis auf einen positiven Faktor auch mit S_2 (Gini–Maß) übereinstimmen. Wir betrachten zunächst S_7:

Theorem 4.1 *Sei $F = (F_1, \ldots, F_k)$ ein Vektor kumulierter relativer Häufigkeiten mit $F_k = 1$, dann gilt:*

$$S_7(F) = \sum_{i=1}^{k-1}(1/2 - |F_i - 1/2|) = \min_i \sum_{j=1}^{k} |i-j|f_j = \min_i U_i.$$

Beweis: Es ist

$$\begin{aligned}
U_i &= \sum_{j=1}^{k} |i-j|f_j = \sum_{j=1}^{i}(i-j)f_j + \sum_{j=i+1}^{k}(j-i)f_j \\
&= i\sum_{j=1}^{i} f_j - \sum_{j=1}^{i} j f_j + \sum_{j=i+1}^{k} j f_j - i\sum_{j=i+1}^{k} f_j \\
&= iF_i - (iF_i - \sum_{j=1}^{i-1} F_j) + (i(1-F_i) + \sum_{j=i}^{k-1}(1-F_j)) - i(1-F_i)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{i-1} F_j + \sum_{j=i}^{k}(1-F_j) \\
&= (k-1)/2 - (\sum_{j=1}^{i-1}(1/2-F_j) + \sum_{j=i}^{k-1}(F_j-1/2)),
\end{aligned}
$$

da

$$
\begin{aligned}
\sum_{j=1}^{i} j f_j &= f_1 + 2f_2 + \ldots + if_i \\
&= f_i + (f_i + f_{i-1}) + \ldots + (f_i + f_{i-1} + \ldots + f_1) \\
&= (F_i - F_{i-1}) + (F_i - F_{i-2}) + \ldots + F_i = iF_i - \sum_{j=1}^{i-1} F_j
\end{aligned}
$$

und

$$
\begin{aligned}
\sum_{j=i+1}^{k} j f_j &= (i+1)f_{i+1} + (i+2)f_{i+2} + \ldots + (i+k-i)f_k \\
&= i(f_{i+1} + f_{i+2} + \ldots + f_k) + f_{i+1} + 2f_{i+2} + \ldots + (k-i)f_k \\
&= i(1-F_i) + f_k + (f_k + f_{k-1}) + \ldots + (f_k + f_{k-1} + \ldots + f_{i+1}) \\
&= i(1-F_i) + (1-F_{k-1}) + \ldots + (1-F_i) \\
&= i(1-F_i) + \sum_{j=1}^{k-i}(1-F_{k-j}) \\
&= i(1-F_i) + \sum_{j=i}^{k-1}(1-F_j)
\end{aligned}
$$

gelten.

Dann wird U_i minimal, wenn $1/2 - F_j \geq 0$ für $j = 1, 2, \ldots, i-1$ und $F_j - 1/2 > 0$ für $j = i, \ldots, k-1$ sind, womit

$$\min_i U_i = (k-1)/2 - \sum_{j=1}^{k-1} |F_j - 1/2| = S_7$$

ist. □

Žhnliches gilt für das Gini–Maß S_2:

Theorem 4.2 *Sei $F = (F_1, \ldots, F_k)$ ein Vektor kumulierter relativer Häufigkeiten mit $F_k = 1$, dann gilt:*

$$S_2(F) = \sum_{i=1}^{k-1} F_i(1-F_i) = 1/2 \sum_{i=1}^{k}\sum_{j=1}^{k} |i-j| f_j f_i = 1/2 \sum_{i=1}^{k} U_i f_i.$$

Beweis: Es ist

$$\begin{aligned}
1/2\sum_{i=1}^{k}\sum_{j=1}^{k}|i-j|f_jf_i &= \sum_{i=1}^{k}\sum_{j=1}^{i}(i-j)f_if_j \\
&= \sum_{i=1}^{k}\sum_{j=1}^{i}if_if_j-\sum_{i=1}^{k}\sum_{j=1}^{i}jf_if_j \\
&= \sum_{i=1}^{k}if_iF_i-\sum_{i=1}^{k}(iF_i-\sum_{j=1}^{i-1}F_j)f_i \\
&= \sum_{i=2}^{k}\sum_{j=1}^{i-1}F_jf_i \\
&= f_2F_1+f_3(F_1+F_2)+\ldots+f_k(F_1+F_2+\ldots+F_{k-1}) \\
&= F_1\sum_{i=2}^{k}f_i+F_2\sum_{i=3}^{k}f_i+\ldots+F_{k-1}f_k \\
&= \sum_{i=1}^{k-1}F_i(1-F_i)
\end{aligned}$$

wegen

$$\sum_{j=1}^{i}jf_j = iF_i-\sum_{j=1}^{i-1}F_j$$

(siehe Beweis des vorstehenden Satzes). □

4.2 Streuungsmaße auf der Basis von Abstandsmaßen

Vogel (1991) argumentiert, daß die betrachteten Maßzahlen von einem unterschiedlichen Abstand zur Einpunktverteilung ausgehen. Wir wollen dies am Beispiel der beiden Maßzahlen S_2 und S_7 präzisieren.

Bezeichnen $\iota_i = (F_1, \ldots, F_k)$ die Einpunktverteilung mit $F_j = 0$ für $j = 1, 2, \ldots, i-1$ und $F_i = 1$ für $j = i, \ldots, k$ bzw. $\mu = (F_1, \ldots, F_k)$ die extreme Zweipunktverteilung mit $F_1 = F_{k-1} = 1/2$.

Seien $F = (F_1, \ldots, F_k)$ und $F' = (F'_1, \ldots, F'_k)$ beliebige k–dimensionale Vektoren kumulierter relativer Häufigkeiten, dann definieren

$$d_7(F, F') = \sum_{i=1}^{k-1}|F_i - F'_i|,$$

$$d_2(F,F') \;=\; \sqrt{\sum_{i=1}^{k-1}(F_i - F_i')^2}$$

die bekannten L_1– bzw. L_2–Metriken.

Speziell sind

$$\begin{aligned} d_7(F,\mu) &= \sum_{i=1}^{k-1} |F_i - 1/2|, \\ d_7(F,\iota_i) &= \sum_{j=1}^{i-1} F_j + \sum_{j=i}^{k-1}(1-F_j) = U_i, \\ d_7(\mu,\iota_i) &= (k-1)/2, \end{aligned}$$

womit sich

$$\begin{aligned} S_7(F) &= (k-1)/2 - \sum_{i=1}^{k-1} |F_i - 1/2| = \min_i U_i = \min d_7(F,\iota_i) \\ &= d_7(\mu,\iota_i) - d_7(F,\mu) \end{aligned}$$

als Differenz des Abstandes zwischen Einpunkt– und extremer Zweipunktverteilung und zwischen F und der extremen Zweipunktverteilung darstellt und zugleich der minimale Abstand zwischen F und den möglichen Einpunktverteilungen ist. Eine unmittelbare Interpretation von S_7 als Abstand zur Einpunktverteilung ist nicht möglich.

Betrachtet man die Maßzahl S_2, so ist diese im Sinne von

$$S_2(F) = 1/2 \sum_{i=1}^{k} U_i f_i = 1/2 \sum_{i=1}^{k} d_7(F,\iota_i) f_i$$

als durchschnittlicher Abstand von F zu den denkbaren Einpunktverteilungen zu verstehen.

Für den L_2–Abstand gilt analog:

$$\begin{aligned} d_2(F,\mu) &= \sqrt{\sum_{i=1}^{k-1}(F_i - 1/2)^2} \\ d_2(F,\iota_i) &= \sqrt{\sum_{j=1}^{i-1} F_j^2 + \sum_{j=i}^{k-1}(1-F_j)^2} \\ d_2(\mu,\iota_i) &= \sqrt{(k-1)/4}, \end{aligned}$$

womit

$$\begin{aligned} S_2(F) &= \sum_{i=1}^{k-1} F_i(1-F_i) = (k-1)/4 - \sum_{i=1}^{k-1}(F_i - 1/2)^2 \\ &= d_2(\mu, \iota_i)^2 - d_2(F, \mu)^2 \end{aligned}$$

die Differenz der quadrierten Abstände zwischen einer Einpunktverteilung und der extremen Zweipunktverteilung einerseits und F und der extremen Zweipunktverteilung andererseits ist. Wiederum kann keine direkte Interpretation von S_2 als Abstandsmaß vorgenommen werden.

Für die anderen Streuungsmaße, die z.B. auf der Entropie basieren, ist selbst eine mittelbare Interpretation im Sinne eines Abstands zur Einpunktverteilung kaum möglich.

5 Streuungsvergleiche für unterschiedliche Anzahlen von Merkmalsausprägungen

5.1 Weglassen und Aufnehmen von Nullhäufigkeiten

Die vorgeschlagenen Maßzahlen hängen von der Anzahl k der möglichen Merkmalsausprägungen ab. Generell können aber gewisse Merkmalsausprägungen in einer konkreten Untersuchung nicht vorkommen, d.h. die korrespondierenden Häufigkeiten sind Null.

Der Wert eines Streuungsmaßes sollte von diesen Nullhäufigkeiten nicht beeinflußt werden, wenn sie an den Rändern auftreten. So besitzen die beiden Häufigkeitsvektoren

$$(0,0,7,1,0,3,0,0,0) \text{ und } (7,1,0,3)$$

dieselbe Streuung. Dieser Forderung entsprechen sämtliche hier betrachteten Streuungsmaße wegen $a(0) = a(1) = 0$.

Nullhäufigkeiten innerhalb eines Häufigkeitstupels haben durchaus einen Einfluß auf die Streuung. Läßt man sie weg, so verringert sich die Streuung, da die extremen Merkmalsausprägungen weniger "weit voneinander entfernt" liegen. Dem entsprechen die betrachteten Streuungsmaße, da eine Nullhäufigkeit $n_i = 0$ zu $F_i = F_{i-1} > 0$ und damit $a(F_i) > 0$ führt, womit $\sum_{j=1}^{k-1} a(F_j)$ gerade um $a(F_i)$ größer ist als im Falle eines Verzichts auf die i–te Ausprägung.

Weitergehende Aussagen über den Einfluß zusätzlicher Merkmalsausprägungen auf die Streuungsmaße sind kaum möglich. So argumentiert Uschner ((1987), S. 33) daß die Streuung eines nominalskalierten Merkmals steigen soll, wenn eine Merkmalsausprägung in dem Sinne gesplittet wird, daß ein Teil der ursprünglichen Häufigkeit auf die neue Ausprägung entfällt. Dies ist für ordinalskalierte Merkmale nicht sinnvoll zu fordern, da z.B. eine Reduktion der Häufigkeiten für die kleinste und größte Ausprägung zu einem Sinken der Streuung führen muß.

5.2 Normierung

Vogel (1991) berechnet für sämtliche von ihm betrachteten Streuungsmaße normierte Versionen, wobei das Streuungsmaß jeweils durch den maximalen Wert des Maßes dividiert wird. Der maximale Wert stellt sich stets als

$$\sum_{i=1}^{k-1} a(1/2) = (k-1)a(1/2).$$

ein, so daß die normierten Maße durch

$$S^*(F) = S(F)/(k-1)a(1/2))$$

gegeben sind.

Ein Problem ist jedoch, daß das zuvor geforderte Verhalten von Streuungsmaßen bezüglich der Aufnahme bzw. dem Weglassen von Merkmalsausprägungen verloren geht. Insbesondere Nullhäufigkeiten für die kleinsten und/oder größten Ausprägungen beeinflussen zwar nicht das Streuungsmaß, aber sehr wohl das Maximum und damit das normierte Streuungsmaß.

6 Zusammenfassung

Wir haben aus der Literatur bekannte Vorschläge für Streuungsmaße ordinalskalierter Merkmale als lineare Rangordnungsstatistiken mit speziellen Scores dargestellt und ausgehend von einer einfachen Streuungsordnung ein weiteres Streuungsmaß betrachtet. Anschließend wurden zwei Streuungsmaße herausgegriffen, für die eine einfache Interpretation als minimale bzw. durchschnittliche Anzahl von Häufigkeitsumschichtungen, die nötig

sind, um zu dem Extremum einer Einpunktverteilung zu gelangen. Das zweite Maß geht auf Gini (1955) zurück. Abschließend wurden die Bezüge dieser Maßzahlen zu bekannten Distanzmaßen aufgezeigt, das Problem der Normierung angesprochen.

7 Literatur

1. Bickel, P.J. & Lehmann, E.L. (1975). Descriptive statistics for nonparametric models. I. Introduction. II. Location. *Annals of Statistics* **3**, 1038–1069.

2. Bickel, P.J. & Lehmann, E.L. (1976). Descriptive statistics for nonparametric models. III. Dispersion. *Annals of Statistics* **4**, 1139–1158.

3. Dabrowska, D. (1985). Descriptive parameters of location, dispersion and stochastic dependence. *Statistics* **16**, 63–88.

4. Gini, C. (1955). *Memorie di metodologica statistica I. Variabilita e concentrazione.* Eredi Virgilio Veschi, Roma.

5. Hajek, J. & Sidak, Z. (1968). *Theory of rank tests.* Academic Press, New York.

6. Klein, I. (1994). *Mögliche Skalentypen, invariante Relationen und wissenschaftliche Gesetze.* Vandenhoeck & Ruprecht, Göttingen.

7. Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics* **8**, 154–168.

8. Uschner, H. (1987). *Streuungsmessung nominaler Merkmale mit Hilfe von Paarvergleichen.* Dissertation, Nürnberg.

9. Vogel, F. (1991). Streuungsmessung ordinalskalierter Merkmale. *Jahrbücher für Nationalökonomie und Statistik* **208**, 299–318.

10. Vogel, F. (1994). Ein einfaches und gut interpretierbares Streuungsmaß für norminale Merkmale. *Allgemeines Statistisches Archiv* **78**, 421–433.

11. Vogel, F. & Dobbener, R. (1982). Ein Streuungsmaß für komparative Merkmale. *Jahrbücher für Nationalökonomie und Statistik* **197**, 145-157.