計量書誌学的分布における集中度：指標の感度と標本量依存性

芳鐘　冬樹（東京大学大学院教育学研究科, E-Mail: fuyuki@p.u-tokyo.ac.jp）
影浦　峡（国立情報学研究所, E-Mail: kyo@nii.ac.jp）

Abstract

Lotkaが観察した著者に対する論文の分布，Bradfordが観察した雑誌に対する論文の分布など，計量書誌学が扱う現象において偏った分布がしばしば観察される。そのような分布の「集中度」を測る指標が数多く提案されているにも拘わらず，集中度の概念そのものは必ずしも明らかではなく，指標の特徴を比較した研究はほとんど存在しない。本研究では,「相対的集中」および「絶対的集中」という2つの観点から，指標の特徴を分析・比較する。あわせて，計量書誌学的データに指標を適用する場合考慮しなければならない問題，指標の標本量依存性に関する考察も行う。

1　はじめに

「ごく一部の事象が頻繁に出現する一方で，非常にわずかしか出現しないような事象が多数存在する」あるいは「少数の者に大部分の富が集中する一方で，残りの富はその他の多数の者に広く分散する」偏った分布は，社会一般の様々な現象において観察される。例えば，Simon (1955) は，そのような偏った分布の例として，テキストに出現する語の頻度分布や所得分布等を挙げている。

計量書誌学が扱う範囲でも，この類の集中・分散はしばしば観察される。最もよく知られているのは，Lotka (1926) の法則・Bradford (1934) の法則に従う，著者の発表論文数や雑誌の掲載論文数の分布だろう。Lotkaの法則は，ある分野の論文が特定の著者に集中する様子を，Bradfordの法則は，ある主題の論文が特定の雑誌に集中する様子を示した計量書誌学分野の経験則である。

このような偏りのある分布の偏りの程度，つまり分布の集中度を測る指標は数多く提案されている。しかし，その一方で，集中度の概念そのものは必ずしも明らかになっていない。この状況に関して，Egghe & Rousseau (1991) は“集中という概念は，集中度を測るために用いる指標の値を通して定義されている”という循環論法に陥っていると指摘している。集中度の指標の特徴を整理・比較した研究がほとんどないのは，集中度の概念に関する共通の基盤が存在しないことに起因していると考えられる。

そこで本研究では,「相対的集中」および「絶対的集中」という2つの観点を提示し，それぞれの観点から指標の特徴を分析する。さらに，計量書誌学的データの特性を考慮に入れたとき，母集団推定において浮かび上がる指標の特徴，あるいは問題点に関する考察も行う。

2　相対的集中と絶対的集中

集中という概念は，相補う2つの観点から整理することができる。2つの観点とは，すなわち「絶対的にどれだけの数の事象に集中して出現しているか」という絶対的集中，そして「出現する事象の間で，頻度がどの程度，相対的に不平等であるか」という相対的集中である。前者は，言わば「出現事象への集中」，後者は「出現事象の中での集中」であり，2つの観点を合わせたものが「集中」であると考えることができる。

著者の発表論文数の分布を例にとると，絶対的集中は「ある分野の論文が，何人の著者に集中しているか」を表し，相対的集中は「それらの著者の中での論文数の偏り」を表す。集中度の記述には，この2つの観点が必要である。

相対的集中という観点を幾何的に表現したものがローレンツ曲線である。図1にローレンツ曲線を例示した。曲線は，事象を出現頻度の昇順に並べたときの累積相対頻度の推移を表している。すべての事象が等しい頻度で出現する一様分布の場合，ローレンツ曲線は対角線と重なり，分布の不平等が増すにつれて，曲線は対角線から逸れる。ローレンツ曲線では，横軸・縦軸ともに相対値を座標にとっているため,「何%の事象で何%の頻度をカバーしている」という相対的集中の有様が変わらなければ，事象数や頻度総数が変わっても，曲線の軌跡は変化しない。

一方，絶対的集中は出現事象数によって測ることができる。ローレンツ曲線の横軸は事象の累積相対度数を表しているため，出現事象数は，横軸の絶対的スケールに対応する。

本研究では，ローレンツ曲線による幾何的表現を利用して，2つの観点から見た集中度の指標の特徴を明らかにする。

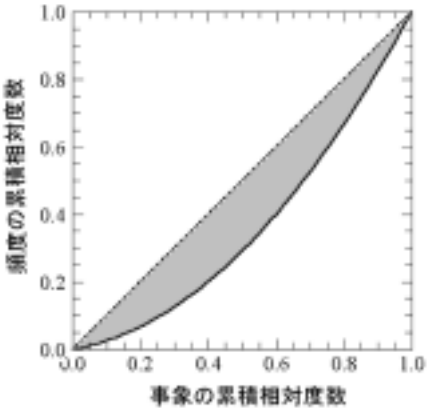


図 1: ローレンツ曲線

3　集中度の指標

今回の研究では以下の 6 つの指標を分析の対象とし，次節以降それぞれの特徴を明らかにする。

変動係数：　$C_V = \frac{\sigma}{\mu}$
対数標準偏差：　$\sigma_L = \sum_{i=1}^{V} \left(\log \frac{f_i}{\mu}\right)^2$

ジニ係数：　$G = \sum_{i=1}^{V} \sum_{j=1}^{V} \frac{|f_i - f_j|}{2\mu V^2}$

Theil (1967) の指標：　$Th = \frac{1}{V} \sum_{i=1}^{V} \frac{f_i}{\mu} \log \frac{f_i}{\mu}$

Herfindahl (1950) の指標：　$HH = \sum_{i=1}^{V} p_i^2$

Pratt (1977) の指標：　$C = \frac{2(\frac{V+1}{2} - q)}{V-1}$

ただし，$q = \sum_{i=1}^{V} ip_i$

ただし $p_1, p_2, ..., p_V$ は降順に並ぶものとする。

ここで，V は事象数，f_i は事象 $x_i (1 \leq i \leq V)$ の頻度を，p_i は事象 x_i の相対頻度を意味する。また，μ は頻度の平均を表す。これらのうち，G，Th，HH はもともとは経済学分野で提案されたものであり，C は計量書誌学分野独自の指標である。

4　指標の特徴

2 節で導入した相対的集中と絶対的集中という観点から，3 節で紹介した指標の特徴を分析する。

4.1　相対的集中に対する感度

ローレンツ曲線を描いたとき，対角線からの逸れ（曲線と対角線に囲まれた図形の面積）が大きければ大きいほど，低頻度の事象の相対出現頻度がより低く，高頻度の事象の相対出現頻度がより高くなるため，分布の不平等，つまり相対的集中の程度はより高いと言える。

より出現頻度が低い事象から，より出現頻度が高い事象に頻度を移動させたとき，ローレンツ曲線の逸れは大きくなる。今回取り上げた指標すべて，移動の原則[1]を満たし，ローレンツ曲線の逸れがより大きい分布により高い値を与えるため，基本的には相対的集中に対して感度を持つと言える。それぞれの指標が持つ感度の性質を把握するため，次に述べる簡単なシミュレーションを行った。

図 2 は，頂点 A（A1，A2）を通るローレンツ曲線（三角形）を例示したものである。このシミュレーションでは，単純化のため，曲線を三角形の 2 辺で代用する。曲線（三角形の 2 辺 BA，AC）は，事象を頻度の昇順に並べたときの累積相対頻度の推移を表している。

ここで，頂点 A が直線 DE に沿って D から E に移動すると仮定する。直線 DE が BC に並行であることから，A

[1] Dalton (1920) の提唱による。出現頻度が低い事象から高い事象に頻度を移動させたとき，集中度の値は増加しなければならないという原則。

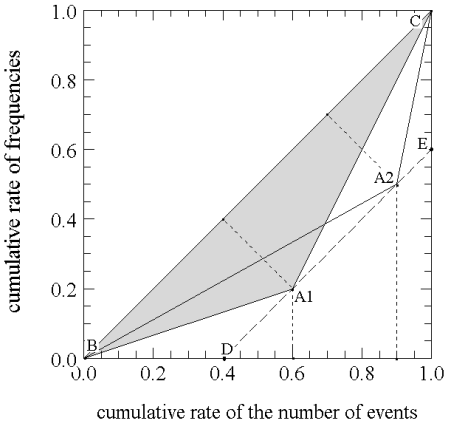


図 2: ローレンツ曲線の推移

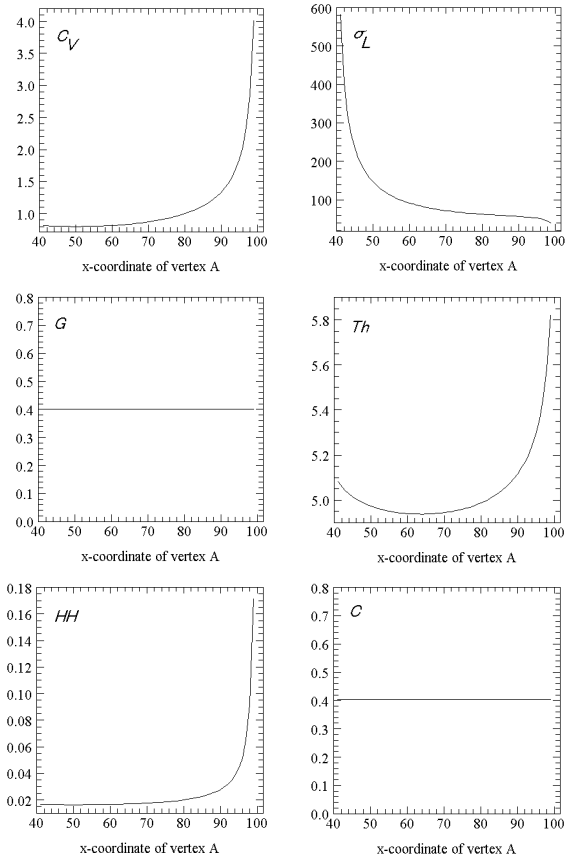


図 3: 頂点 A の座標の変化に伴う指標の値の変化

が DE 上にある限り三角形 ABC の面積は変化しない。例えば，三角形 A1BC と A2BC は同じ面積を持つ。つまり，この条件下では，「逸れの大きさ」は一定値を保つことが分かる。このときの各指標の振る舞い（頂点 A の位置と指標の値の関係）を示したものが図 3 である[2]。横軸には，頂点 A の位置を示すため図 2 における頂点 A の x 座標をとり，縦軸には指標の値をとった。

頂点 A が D から E へ移動するということは，ローレンツ曲線の逸れの頂点（三角形でなく曲線で考えれば，傾きが大きく変わる座標，つまり不平等が顕著に現れる部位）が，低頻度の事象から高頻度の事象へ移動するということである。このときの指標の値の挙動は，4 つの類型に大別できる。タイプ 1 は逸れの頂点の座標に関わらず一定値を保つ指標であり，G と C がこれにあたる。これらは総体として平均からどれだけ逸れているかのみを評価する指標であり，逸れの大きさが同じであれば，どの層に不平等の歪みが見られるかは問題にしない指標であると考えられる。

タイプ 2 は，頂点 A が E の近傍に寄るにつれて，値が急激に増加する指標である。C_V と HH がこれに属する。これらの指標は，平均出現頻度との差の 2 乗，あるいは相対出現頻度の 2 乗をもとに算出されるため，2 乗の寄与により，最頻事象群の影響を非常に強く受ける。これらは，最頻事象群に対する感度が非常に高い指標である。総体としての平均からの逸れが同程度であっても，不平等の歪みが

[2] 事象数 100，頻度総数 10,000 という条件のもと，シミュレーションを実行した。事象数が一定であるため，絶対的集中の関与は無視できる。

高頻度の事象側に現れる分布の方が相対的な集中度はより高いと，これらの指標は評価する。

3つめのタイプは，D の近傍において値の急激な変化が観察される指標である。σ_L がタイプ 3 に属す。関数 $y = \log x$ は，$x \to 0$ のとき $y \to -\infty$ であるため，出現頻度が 0 に近い事象が存在するとき $(\log x_i/\mu)^2$ の総和である対数標準偏差の値は非常に大きくなる。この指標は，タイプ 2 とは逆に，最も頻度が低い事象群に対して非常に高い感度を持つことが分かる。

タイプ 4 の指標は，逸れの頂点 A が低頻度の事象側から高頻度の事象側に移動したとき，値の変化が凹型の軌跡を描くものである。Th がこれに属する。この指標は，高頻度と低頻度，両側の事象に対する感度が，中程度の順位の事象に対する感度よりも高いと考えられる。

シミュレーションの結果から，相対的集中に対する感度の性質は指標によって異なることが確認できる。総体としての平均からの「逸れ」の大きさが同じ分布を比べたとき，指標が高い感度を持つ部位において不平等が大きい分布の方が，指標の値は大きくなり，相対的な集中度は高く評価される。

4.2 絶対的集中に対する感度

絶対的集中の程度は，出現する事象の数で測ることができる。当然，その数が少なければ少ないほど，分散の規模は小さく，絶対的集中の程度は高いと言える。したがって，何%の事象で何%の頻度をカバーするという相対的集中の様子を保ったまま，出現事象数を変化させた場合の指標の振る舞いを見ることで，それぞれの指標の絶対的集中に対する感度を調べることができる。

図 4 は，2つの分布例（一様分布 A と一様でない分布 B[3]）について，事象数を定数倍していった[4]ときの指標の値の変化を表したものである。このシミュレーションでは，分布の相対的集中（ローレンツ曲線の軌跡）の様子は保たれるため，絶対的集中（ローレンツ曲線の横軸の絶対的なスケール）に対する指標の感度を観察できる。

σ_L，HH，C は，事象数のスケールが大きくなるにつれて，値が単調減少あるいは単調増加していることが分かる。それらのうち，HH は一様分布の場合も事象数の変化に応じて値が変化しているが，残りの 2 つは変化しない。つまり，HH，C は一様分布の場合には絶対的集中に対して感度を持たない。

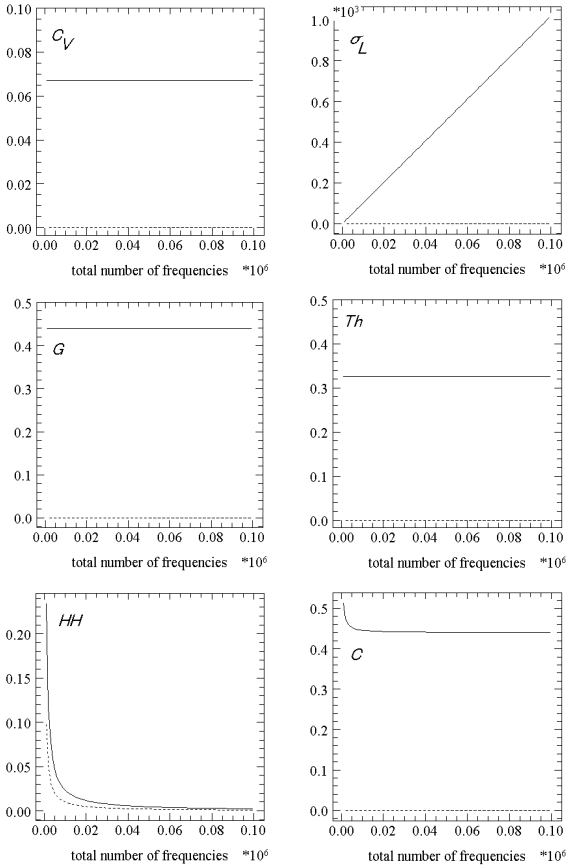
C_V，G，Th は式の中に事象数 V のスケールを標準化する項 $1/V$ を含むため，事象数を等倍しても，その変化は相殺され指標の値は変わらない。つまり，これらの指標は絶対的集中に対して感度を持たない。

一方，σ_L が事象数に比例して増加するのは，単に，事象数のスケールを標準化していないことに起因する。絶対的な分散の規模が大きくなる，すなわち集中度が低くなるにつれて，値が増加してしまうこの指標は，異なる事象数の分布の比較には適していないと言える。

HH，C は，事象数の増加に応じて単調減少するため，我々の感覚に適う形で絶対的集中の評価を値に反映させていると言える。ただし，C は，先ほど述べたように一様分布の絶対的集中に対する感度を持たない。C は，絶対的集中の程度を評価することよりも，一様分布の場合，事象数に関わらず常に値を最低値 0 にすることを優先していると言えよう。

[3] 分布例 A は各事象の頻度が 100 の一様分布，分布例 B の事象の頻度は，350, 258, 156, 104, 73, 44, 10。

[4] すべての m について，m 回出現する事象の数を定数倍する。



点線：分布例 A（一様分布），実線：分布例 B

図 4: 指標の絶対的集中に対する感度

4.3 まとめ

この節では，シミュレーションの解析に基づき，集中度の指標の性質を分析した。ここで，集中度に関する 2 つの観点，すなわち絶対的集中と相対的集中から見た指標の性質を 1 つの表にまとめてみる。表 1 がそれである。

	絶対的集中に対する感度	相対的集中に対する感度
C_V	–	高頻度
σ_L	×	低頻度
G	–	
Th	–	高・低
HH	○	高頻度
C	△	

表 1: 指標の性質

絶対的集中については，妥当な評価を値に反映させているかどうかを表に示した（○：妥当，△：妥当（一様分布の場合，感度なし），×：妥当でない，–：感度なし）。相対的集中については，どの層の事象に対して特に高い感度を持つかを示した。

5　指標の標本量依存性

集中度の指標のほとんどは，頻度総数に対して感度を持たない。3 節で紹介した指標も例外ではなく，各事象の出現頻度を等倍しても指標の値は変わらない。標本抽出調査に基づいて，母集団分布の特徴を推定しようとする場合，この性質（頻度総数を反映しないという性質）に注意を払う必要がある。

最も単純な例を考えてみよう。10 人の著者がそれぞれ 1 本ずつ論文を持っている標本と，同じく 10 人が，それぞれ 10 本ずつ持っている標本があったとする。標本量を増やしていったとき，最初の 10 人以外の著者が現れ，出現著者数が増加する確率に，どんな違いがあるだろうか。全員が既に 10 回現れている後者の分布よりも，まだ 1 回ずつしか現れていない前者の分布の方が，たまたまそれまで現れていなかった著者が標本に加わる可能性が高いと考えるのが自然だろう。したがって，前者の方が，母集団分布の絶対的集中度は低いと予想できる。また，標本に出現しないような低頻度の事象が存在するわけであるから，前者の母集団はより裾が広い分布になると予想できる。それゆえ，母集団分布の相対的集中度は，逆に前者の方が高くなるだろう。つまり，標本において絶対的集中や相対的集中の様子が同じであっても，頻度総数が違えば，予測される母集団の特徴は異なったものになると言える。

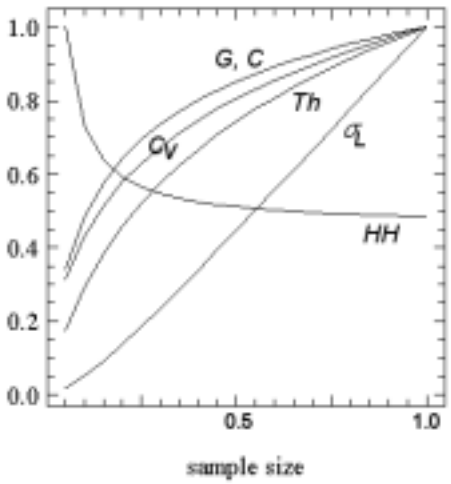


図 5: 指標の標本量依存性

頻度総数を反映しない指標は，標本そのものの集中度を測っているにすぎないため，統計的に不十分な標本では，標本量によって値が変化してしまう。指標の標本量依存性を確認するため，学会発表データベース[5]から抽出した情報処理学会の発表論文データを標本として[6]，標本量の変化に伴う指標の値の変化を観察した[7]。

図 5 は，元の標本から大きさの異なる部分標本を抽出し，元の標本の大きさの 20 分の 1，20 分の 2，... と標本量を大きくしていったときの各指標の値の変化をプロットしたものである。部分標本の抽出は標本量ごとに 1,000 回反復し，指標の値はその平均をとった。図の横軸は，元の標本の大きさに対する相対的な標本量を，縦軸は，観察された最大の値が 1 になるよう補正を施した指標の値を表す。

図から明らかなように，すべての指標が標本量に依存して系統的に変化する。指標の標本量依存性は，今回用いた標本だけに見られる問題ではない。一般に，Lotka 型・Bradford 型データ等，計量書誌学が扱うデータの多くは，1 回しか出現しない事象（著者・雑誌）が多数を占めるという特徴を持ち，そのような特徴を持つデータでは，集中度の指標を含むほとんどすべての統計量が標本量の変化に応じて系統的に変化してしまうことが分かっている (Tweedie & Baayen, 1998)。したがって，集中度の指標を使って，計量書誌学的標本に基づく比較分析を行う場合，指標の標本量依存性を考慮した統計的枠組みが必要となる。

6　おわりに

本研究では，相対的集中と絶対的集中という 2 つの観点から集中度の指標の特徴を整理した。また，計量書誌学的標本に指標を適用した場合，標本量に依存して値が変化してしまうという問題を確認した。

標本量に依存する指標を使う場合，標本量の変化に伴う動的な値の推移を追う必要がある。標本量依存性に関わる問題を回避するためのもうひとつの方法は，集中度の指標に対する母集団分布の不偏推定値を構成する統計量を使うことである。ただし，本研究で取り上げた指標に限れば，Herfindahl の HH（Simpson (1949) の D が不偏推定値を構成する）以外，不偏推定値は知られていない[8]。

参考文献

[1] **Bradford, S. C.** (1934) “Sources of information on specific subjects,” *Engineering*, Vol. 137, p. 85–86.

[2] **Dalton, H.** (1920) “The measurement of the inequality of incomes,” *The Economic Journal*, Vol. 30, p. 348–361.

[3] **Egghe, L. and Rousseau, R.** (1991) “Transfer principles and a classification of concentration measures,” *Journal of the American Society for Information Science*, Vol. 42, No. 7, p. 479–489.

[4] **Good, I. J.** (1953) “The population frequencies of species and the estimation of population parameters,” *Biometrika*, Vol. 40, No. 3–4, p. 237–264.

[5] **Herfindahl, O. C.** (1950) *Concentration in the Steel Industry*, Ph.D. dissertation, Columbia University, p. 15–24.

[6] **影浦峡** (2000)『計量情報学：図書館/言語研究への応用』丸善.

[7] **Lotka, A. J.** (1926) “The frequency distribution of scientific productivity,” *Journal of the Washington Academy of Sciences*, Vol. 16, No. 12, p. 317–323.

[8] **Theil, H.** (1967) *Economic and Information Theory*. Amsterdam, North-Holland Publishing Company, p. 91–96.

[9] **Pratt, A. D.** (1977) “A measure of class concentration in bibliometrics,” *Journal of the American Society for Information Science*, Vol. 28, No. 5, p. 285–292.

[10] **Simpson, E. H.** (1949) “Measurement of diversity,” *Nature*, Vol. 163, p. 688.

[11] **Simon, H. A.** (1955) “On a class of skew distribution functions,” *Biometrika*, Vol. 42, p. 425–440.

[12] **Tweedie, F. J. and Baayen, R. H.** (1998) “How variable may a constant be?: measures of lexical richness in perspective,” *Computers and the Humanities*, Vol. 32, p. 323–352.

[5] 国立情報学研究所が提供するデータベース。

[6] 標本の基本的数量は，
論文数 27,047，のべ著者数 79,372，異なり著者数 24,271

[7] 共著の場合，各々の著者がそれぞれ 1 本の論文を発表しているものとして指標の値を算出した。

[8] D が HH の不偏推定値となっていることは，Good (1953)，影浦 (2000) によって指摘・証明されている。