Zipf 法則についての覚え書き

——云語人口密度分布の一特徴——

田　口　時　夫

(1969年9月　受付)

On Zipf's Law
—A Characterization of Distributions in Linguistics and Demography—

Tokio Taguchi

Continuous Zipf's distribution

$$f(x)=\frac{A}{x^k}\ ;\ 0<\kappa\leqq x\leqq\infty,\ 1<k\leqq 2,\ A>0$$

has not arithmetic mean and variance.

In this paper we introduce, at first, the notions of harmonic mean difference γgiven by

$$\gamma^{-1}=\int_\kappa^\infty\int_\kappa^\infty |x_i^{-1}-x_j^{-1}|\, f(x_i)\, f(x_j)\, dx_i dx_j,$$

harmonic concentration curve given by the following parametric represeintation

$$\mathcal{H}(x)=\int_\kappa^x f(x)\,dx$$

$$\mathfrak{H}(x)=H\int_\kappa^x x^{-1}f(x)\,dx,$$

where H expresses harmoniic mean, and harmonic concentration coefficient

$$J=\frac{H}{2\gamma}.$$

At the next, we apply the above notion over Zipf's distribution.
Consequently we see that the relationships between Zipf's distribution and harmonic notions is ultimately the same with the relationships between Paretoan distribution and concentration curve.

The process carried out in this paper will derives easily the dynamic notions and geometrical method.

The Institute of Mathematical Statistics.

1. はしがき

語彙分布や人口密度分布に適合性が高いといわれるものに Zipf 分布がある.
その本来の形式は密度関数を $f(x)$ で表わすと，離散型で

$$(1)\ f(x)=\frac{A}{x};\ x=1,2,\cdots,N,\ A>0$$

であるがその一般化としての連続型密度関数は

$$(1')\ f(x)=\frac{A}{x^k};\ 0<\kappa_1\leqq x\leqq\kappa_2\leqq\infty,\ k>0,\ A>0$$

で与えられる.

従って特に $k>1$ ならば Pareto 分布

$$(2)\ f(x)=\frac{\theta^\alpha}{\alpha x^{1+\alpha}};\ 0<\theta\leqq x\leqq\infty,\ \alpha>0$$

に含まれるのである.

然し後述4節の本質的差異に基づいてこゝでは主として (1′) に於て $1<k\leqq 2$ の場合を Zipf 分布*, $2<k\leqq 3$ の場合を Pareto 分布** と区別することにする.

その場合 Zipf 分布 (1′) に於て特に $\kappa_2=\infty$ の場合を除いては既にしばしば集中曲線法による解析が有効であることを指摘した [1] [6] [7]. 従ってこゝでの本質的な問題は (1′) に於て $1<k\leqq 2$, $\kappa_2=\infty$ で定義した Zipf 分布の場合であるが, この場合の最大の数量的特徴は分散のみならず平均が存在しないことである. 従って此の際集中曲線法を直接適用する訳にはいかない. 勿論モメントを基本的統計量とする教理統計学的手法についても同様のことが云える.

此の制約の下で他の基礎的統計量を導入し解析的方法を確立することが此の稿の目的である

2. 調和平均と調和平均差及び調和集中指数

N 個の測定値をもとにした時調和平均 H は云うまでもなく

$$\frac{1}{H}=\sum_{i=1}^{N}\frac{1}{x_i} \tag{3}$$

で与えられる. 度数 N_i を考慮すれば

$$\frac{1}{H}=\sum_{a=1}^{p}\frac{N_h}{N}\frac{1}{x_h} \tag{3′}$$

である.

こゝで通常の平均差

$$\Delta=\frac{1}{N^2}\sum_{i,j=1;2\cdots,N}\sum|x_i-x_j| \quad 又は \quad \sum_{h,k=1,2,\cdots,p}\sum\frac{N_hN_k}{N^2}|x_h-x_k| \tag{4}$$

に対して

(定義 1)
$$\frac{1}{\nabla}=\frac{1}{N^2}\sum_{i,j=1,2,\cdots,N}\sum\left|\frac{1}{x_i}-\frac{1}{x_j}\right| \quad または \sum_{h,k=1,2,\cdots,p}\sum\frac{N_hN_k}{N^2}\times\left|\frac{1}{x_h}-\frac{1}{x_k}\right| \tag{5}$$

によって与えられる ∇ を調和平均差と定義しよう.

同様に加算平均 μ を用いた集中係数 (Gini 指数) G は (4) に対して

$$G=\frac{\Delta}{2\mu} \tag{6}$$

で与えられるが, これに対応して

(定義 2) (7) $J^{-1}=\dfrac{2\nabla}{H}$ を充す J を調和集中指数とする.

今もし, 正領域 $0<\kappa_1\leqq x\leqq\infty$ で定義される連続な密度関数 $f(x)$ を対象にするならば (3′) (5) は勿論

$$H^{-1}=\int_{\kappa}^{\infty}x^{-1}f(x)\,dx \tag{8}$$

$$\nabla^{-1}=\int_{\kappa}^{\infty}\int_{\kappa}^{\infty}|x0^{-1}{}_i-x^{-1}{}_j|f(x_i)f(x_j)\,dx_idx_j \tag{9}$$

---

* Zipf 分布は [2] [3] [4] 等に見られる如く $1<k<1+\alpha$ を充していると仮定出来る。

** 事実 Pareto は $\alpha=1.5$ の場合即 $k=2.5$ を一つの基準状態と考えていた事が示されている。[4] Pareto 分布に於ては $\alpha>1$ とすることは例えば [3] に明快に示されている。

で表現される.

3. 調和集中曲線

連続な密度関数 $f(x)$ に対して定義される集中曲線 $\Lambda(X, Y)=0$ は

$$(10)\quad \begin{cases} X(x)=\int_{-\infty}^{x} f(x)\,dx \\ Y(x)=\dfrac{1}{\mu}\int_{-\infty}^{x} xf(x)\,dx \end{cases}$$

で助変数表示された.

我々はこれに対して

定義 3. 正領域 $0<\kappa\leqq x\leqq\infty$ で定義される連続な密度関数 $f(x)$ の調和集中曲線 $\Psi(\mathcal{H}, \mathfrak{Y})=0$ は

$$(11)\quad \begin{cases} \mathcal{H}(x)=\int_{\kappa}^{x} f(x)\,dx \\ \mathfrak{Y}(x)=H\int_{\kappa}^{x} x^{-1}f(x)\,dx \end{cases}$$

で助変数表示される曲線である.

を与えるならば $\Psi(\mathcal{H}, \mathfrak{Y})=0$ が $\Lambda(X, Y)=0$ に可成り類似した性質を示すことが認められる.

例えば

定理 1 (12) $0\leqq\mathcal{H}\leqq 1$ かつ $0\leqq\mathfrak{Y}\leqq 1$

$$(13)\quad \frac{d\mathfrak{Y}}{d\mathcal{H}}=\frac{H}{x}$$

$$(14)\quad \frac{d^2\mathfrak{Y}}{d\mathcal{H}^2}=-\frac{H}{x^2 f(x)}$$

等が得られる. (拙稿 [6] 111頁参照)

従って例えば正領域 $0<\kappa\leqq x\leqq\infty$ で定義される連続な密度関数の調和集中曲線は第 1 図に見られるごとく $(0, 0)$ および $(1, 1)$ を通る単調増大で上に凸な曲線であることが分る.

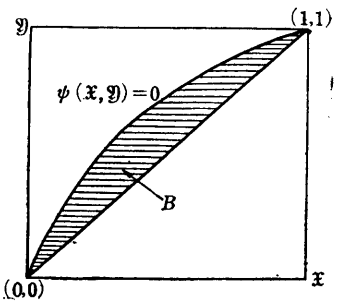
更に直線

$$\mathfrak{Y}=\mathcal{H}$$

は集中曲線の場合に準じて此の場合にも均等線と称することが出来るが, この時曲線

$$\Psi(\mathcal{H}, \mathfrak{Y})=0$$

と均等線とで囲まれる領域の面積を B とすれば, Gini 係数と集中曲線の対応する面積 A との関係に同等な関係が得られる.



第 1 図 調和集中曲線

つまり

$$2B=\int_0^1 \mathfrak{Y}\,d\mathcal{H}-\int_0^1 \mathcal{H}\,d\mathfrak{Y}$$

$$2\frac{B}{H}=\int_{\kappa}^{\infty} dF(x)\int_{\kappa}^{x} y^{-1}dF(y)-\int_{\kappa}^{\infty} x^{-1}dF(x)\int_{\kappa}^{x} dF(y)=\int_{\kappa}^{\infty}\int_{\kappa}^{8}(y^{-1}-x^{-1})\,dF(x)\,dF(y)$$

扨 $\int_{\kappa}^{\infty}\int_{\kappa}^{\infty}(y^{-1}-x^{-1})\,dF(x)\,dF(y)=0$ であるから,

$$2\frac{B}{H}=\frac{1}{2}\left[\int_{\kappa}^{\infty}\int_{\kappa}^{x}(y^{-1}-x^{-1})\,dF(x)\,dF(y)-\int_{\kappa}^{\infty}\int_{x}^{\infty}(y^{-1}-x^{-1})\,dF(x)\,dF(y)\right]$$

$$=\frac{1}{2}\int_{\kappa}^{\infty}\int_{\kappa}^{\infty}|y^{-1}-x^{-1}|\,dF(x)\,dF(y)=\frac{\nabla^{-1}}{2}$$

つまり

定理 2. (15) $B=\dfrac{H}{4\nabla}=\dfrac{J}{2}$

が成立する.

従って，正値変量 x の密度関数 $f(x)$ に関しては常に $0\leqq G\leqq 1$ に対応して

系 1. (16) $0\leqq J\leqq 1$

が成立するのである.

また筆者は嚢に [6] 集中曲線の幾何学的性質から N の測定値の平均差 Δ は，測定値を単調な n クラスに分けた時各クラスの個数を N_h 平均を M_h また平均差を Δ_h とすれば，

$$(17)\quad \Delta=\sum_{h=1}^{n}p_h{}^2\Delta_h+\sum_{\substack{h,k=1,2\cdots n\\ h\neq k}}\sum p_hp_k|\mu_h-\mu_k|$$

こゝで $p_h=\dfrac{N_h}{N}$ (拙稿 [6] 122 頁で単調としなかったことは誤りである)

が成立し，従って第一項を級内平均差，第二項を級間平均差とすれば

平均差は分散同様，級内及び級間平均差に分解されることを示した.

今調和平均差について考えると (15) 式の成立は (17) の導出と全く同様の演算過程を与えうることを示している. 従って

定理 3. 測定値を単調なクラスに分けるならば

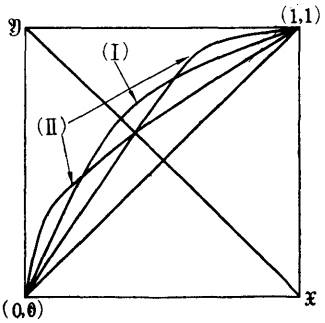
$$(18)\quad \nabla^{-1}=\sum_{h=1}^{n}p_h{}^2\nabla_h{}^{-1}+\sum_{\substack{h,k=1,2\cdots,k\\ h\neq k}}\sum p_hp_k|H_h{}^{-1}-H_k{}^{-1}|$$

の成立を知ることが出来る.

この場合に於ても第一項を級内調和平均差 ∇_{with} の逆数，第2項を級間調和平均差 $\nabla_{bet,}$ の逆数と定義すれば

$$(18')\quad \nabla^{-1}=\nabla_{with,}{}^{-1}+\nabla_{bet,}{}^{-1}$$

と表現されることになる.



第2図 自己及び相互対称調和集中曲線

従ってこの場合も集中曲係の場合 (拙稿 [6] 113頁参照) と同様均等線と異なる他の対角線 $\mathcal{H}+\mathcal{Y}=1$ に対する対称性について分布関数を特徴づけることが出来る.

即特に連続型分布関数についてその調和集中曲線を

(19) $\mathcal{Y}=\phi(\mathcal{H})$ と

陽関数で表示すると次式で与える対称条件

$$(20)\quad \frac{\phi''(\mathcal{H})}{\phi''(1-\mathcal{H})}=\phi'^{3}(\mathcal{H})$$

(拙稿 [6] 114頁参照)

に定理1を適用して

定理 4. 調和集中曲線が均等線外の対角線に対して自己対称となる分布関数は

$$(21)\quad \frac{f(x')}{f(x)}=\left(\frac{H}{x'}\right)^{3}\qquad xx'=H^2$$

を充している.

同様にして二つの集中曲線が相互に上記の対角線に対して対称である条件は

定理 5. 二つの相互対象は集中曲線をもつ連続な密度関数 $f_1(x)$, $f_2(x)$ は

$$(22)\quad \frac{H_1 f_2(x')}{H_2 f_1(x)}=\left(\frac{H_1}{x'}\right)^3 \qquad xx'=H_1H_2$$

をみたす. (拙稿 [6] 114頁参照)
ことである.

4. 特異分布について

歪み型分布の条件 [2], [3] [4] は通常

$$(23)\quad -\frac{\log f(x)}{\log x}\to k>0 \qquad (x\to\infty)$$

によって与てられるが, これはたしかに Pearson 型を初めとするいわゆる一般の分布関数の充す条件

$$(24)\quad -\frac{\log f(x)}{\log x}\to\infty \qquad (x\to\infty)$$

とは区別されるものである。

またこの条件は集中曲線の最大曲率の位置と大きさによってグラフ的に意味づけられることは筆者が示した (拙稿 [6] 124, 5 頁参照)

然しこの条件は既に序文及び [8] で触れたように本質的に次のように細分されるべきものである. 即 (23) (24) に基づいて第1表を得ることが出来る.

第1表

0	条 件	特徴 (特に連続分布について)	代表的分布
(i)	$-\infty<k\leqq 1$	存在しない	有界領域になる
(ii)	$1<k\leqq 2$	平均なし 分散なし	Zipf 分布
(iii)	$2<k\leqq 3$	平均あり 分散なし	Pareto
(iv)	$3<k<\infty$	平均あり 分散あり	正規分布

特に (iv) の条件を充すものは従来の数理統計学的解析法の適用可能な本来のジャンルを含んでいる. 又 (iii) については集中曲線に対する曲線解析の本来の適用の場であり, かつ実際に有効である事は周知の事実である。

従って 残された (ii) の条件こそが従来の方法に対して問題を提起するものであり, Zipf 分布の意義と, 2, 3 節に与えた諸定義を斎す源泉である. つまり分散及び平均が存在しなくても H 及び ∇ が従って調和集中曲線が存在しうるからである.

5. Zipf 分布の性質

一般に Zipf 分布は有界な不連続分布として, 又一般化された Zipf 分布 (1) は $0<\kappa\leqq x\leqq\infty$ であるが, 不連続型として説明されている [2] [3] [4]。

しかし, 1, 2, 3 節で述べた方法は, 連続型の場合は勿論, 不連続型の場合にも折線グラフを考えれば本質的に同様の結果を与える (拙稿 [6] 110 頁 (5) 式参照) から, 以下では (1) は連続な密度分布と仮定する. 又特に有界領域とする必要はない (有界の場合には次式の upper tremcation として拙稿 [6] 115 頁3節に述べたことと類似の関係が得られる) から特に問題を次の密度関数に限定しても問題の本質を失わない. 即

$$(25)\quad f(x)=\frac{A}{x^k}\ ;\ 0<\kappa\leqq x\leqq\infty,\quad 1<k\leqq 2,\ A>0.$$

この場合調和集中曲線 $\mathscr{V}(\mathscr{H},\mathfrak{P})=0$ は (11) により

$$(26)\quad \begin{cases} \mathcal{H}(x)=\int_{\kappa}^{x}\frac{A}{x^{k}}dx \\ \mathcal{Y}(x)=H\int_{\kappa}^{x}\frac{A}{x^{k+1}}dx \end{cases}$$

であり，かつこの曲線は (12) 式を充しているとを考えると，Pareto 分布

$$(27)\quad f(x)=\frac{A'}{x^{k}+1};\ 0<\kappa\leqq x\leqq\infty,\ 1<k\leqq 2,\ A>0$$

の (10) 式による集中曲線 $\Lambda(X,Y)=0$ と一定の関係をもつことが分る．即

$$(28)\quad \Psi(Y,X)=0$$

が成立する．つまり $\mathcal{H}$, $\mathcal{Y}$ 軸を X, Y 軸に重ねると，二つの曲線は均等線に対して対称な位置関係にある．

従って (6)，(7) の性質を併せ考えると

定理 5 Zipf 分布 (22) の調和集中曲線は

$$(29)\quad 1-\mathcal{Y}=(1-\mathcal{H})^{k/k-1}$$

である．又調和集中指数は

$$(30)\quad J=\frac{1}{2k-1}$$

である．

証明，拙稿 [6] 131 頁第3表 (G) 項でそれぞれ k を $k+1$, Y を $\mathcal{H}$, X を $\mathcal{Y}$ とすれば (25) 式が得られる

又拙稿 [6] 133 頁第6表 (G) で k を $k+1$ とした時の Gni 係数が J に等しいことから (30) を得る．

又同時に

定理 6. (1′) をみたす連続な密度関数の調和集中曲線は

(i) $k=0.5$, $\kappa_2<\infty$ のとき自己対称である．又領域を同じくする二つの調和集中曲線について

(ii) $k=0.5+p$, $k_2=0.5-p$ の二つの曲線は $\kappa_2<\infty$ のとき相互対称である．

証明 拙稿 [6] p 114 頁における集中曲線の対称性が $k-1$ としてそのまゝ調和曲線の対称性になる事を考えればよい．

更に集中曲線と同様調和集中曲線の最大曲率をもつ点の接線が $\mathcal{Y}$ 軸となす角を φ とすれば

$$(31)\quad \cos 2\varphi=1-\frac{2}{3}(k+1)=\frac{1-2k}{3}$$

(拙稿 [6] 127 頁 [III] 参照)

が成立つ事を容易に知る事が出来る．従って

定義 4. $$(32)\quad T=\cos 2\varphi$$

で与えられる T を調和集中曲線の歪度とすることが出来る．

この場合 $T=0$ ならば (1′) の曲線は自己対称であり，$T=\pm 1$ の時 (1′) の曲線は最大曲率をそれぞれ両端 (0, 0) 及び (1, 1) で有することになる．

更に

定義 5. 調和集中曲線の尖度 L は最大曲率で与えることが出来る．

この場合集中曲線おいて Pareto 分布に対し成立つ関係と同様

定理 7. Zipf 分布は常に

$$(33)\quad -\frac{1}{3}<T\leqq 1\quad \text{かつ}\quad T=1\text{において}\ L<\infty$$

である.

以上によって Pareto 分布が集中曲線によって明確にされたと全く同程度に Zipf 分布は調和集中係数等によって解明されることが分るであろう.

勿論従来通りの観点の下に

例えば Zipf 分布 (21) は対数変換

(34) $y_1=\log x$ よって指数分布に又

(35) $y_2=\frac{1}{x^2}$ によって Pareto 分布に変換されることから二, 三の帰結を得ることは出来るであろう. しかし此種の形式的変換によって得られる関係は例えば対数 moment 或は更に複雑な統計量を生む事になり, 検推定の方式に多少のヒントは与えうるが, 分布の構造的解明に直接つながるとは思われない.

こうした観点の下では更に Zipf 分布の具体的モデルビルディングに遡って検討することを必要とする.

6. Zipf 分布に関するモデル

Zipf 分布に関しては H. A. Simon を初めとする stochastic model, 或は特に云語問題に関して entropy $H=-f\log f dx$ による model binldug 又別に例えば重力と関係づけた力学的モデル [5] が存在する.

人口, 云語問題について門外漢の筆者が此等について解説を加える事は勿論適切でないが, stochastic modcl は状態を単線的に記述しているが構造観念に薄くいわば無重力的に都市や文章の構成を計量的に表現したものであり, entropy 概念はかゝる記述を脱却したやや具象性を帯びた観念として論理的には興味があるが説明原理としては, 猶試論の域を出ないといえよう.

もし現実的な構造や趨勢を客観的に方程式系として捉えようとするならば, 事実 Zipf や Steward [5] が試みたように. 何等かの意味で力学的観点が要請されるのではなかろうか.

この意味で, Pareto 分布に代表される経済諸分布が集中曲線を通して, 競争原理や支配度概念に対応する Δ とその発展形式で理解されるごとく, ([1] [6] 及び [7] をみよ) 調和集中曲線を通して, 言語の使用率や, Stewart [5] の都市の人口引力を ∇ に対応させることによって統計的に Zipf 分布を理解しようとすることは可能でありかつ肯定され得る面をもつのではなかろうか.

例えば今仮に人口密度 x と都心かの距離 d との間に理想状態に於て

$$x=\frac{\lambda}{d} \tag{36}$$

が成立するとすれば $x_i^{-1}=x_j^{-1}=\frac{1}{\lambda}(d_i-d_j)$ 従って定義 1 (5) 式で支えた ∇ は

$$\nabla=\frac{\lambda}{E|d_i-d_j|} \tag{37}$$

であり, (平均的な意味で) 人口引力や potential の概念に接近した形をとる.

又言語使用率 u が同じく理想状態で例えば語の長さ x に対して

$$u=\frac{\kappa}{x} \tag{38}$$

であれば $x_i^{-1}-x_j^{-1}=\frac{1}{\kappa}(u_i-u_j)$ 従って ∇ は

$$\nabla=\frac{1}{\kappa E|u_i-u_j|} \tag{39}$$

となり，語に対する撰択或は需要の強さの表現を与え得よう．

勿論こうした単純な設定が屈折した諸状況の中で直接的な形での成立が許されるとは思はないが，前節までの所論に立脚して，唯次の事実だけは数量的な面で指摘することが出来る．

つまり都市えの人口集中を表現する為に用いられる，都市からの距離順に累積した地積，人口累計百分率の関係を示した集中図は，逆数変換した集中曲線を媒介にして考える時，座標軸を変換した調和集中曲線のグラフとも看做すことが出来ることである．又この時曲線と均等線との間の面積の倍数は調和集中指数 J に等しくなる．

問題は，更に Zipf 分布 (25) を想定する限りは，**経験的集中図の理解は通常の集中曲線の実限値としてではなく以上のような座標的変換つまり横軸を面積累計とするした調和集中曲線の実現と考えるのが妥当であるという点にある。**

何故ならば既述のように Zipf 分布の集中曲線は理論的に存在せず，強いて集中図を集中曲線の実現として理解する場合には (25) 式の upper truncation, つまり (1′) 型分布の集中曲線，集中指数の実現として受取るべきであり，その際には，常数 k のみならず上界 κ_2 の効果 (拙稿 [6] 各章参照) を充分加味せねばならないからである．

参 考 文 献

[1] P. E. Hart and S. J. Prais, "The analysis of buisiness concentration," J.R.S.S., Ser. A, 119(1156), 150-191.

[2] G. Heldon, "The relation between dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics" Biometrika, Vol. 45, 1958, 222-267.

[3] B. Mandelbrot, "A Note on a class of skew distribution function analysis and critique of a paper by H.A. Simon", Information and Control, Vol. 2, 1950, 90-99.

[4] H. A. Simon, "On a class of skew distribution functions," Biometrika, Vol. 42, 1955, 425-411.

[5] 館稔 」人口分析の方法」形成選書，古今最院 (第2刷)，昭和40年

[6] T. Taguchi, "Concentration-curve methods and structures of skew populations — a methodology for the analysis of economic data — " Annals of the Institute of statistical Mathematics, Vol. 20, No. 1, 1968, 107-141.

田口時夫「競争システムにおける統計 —イタリー学派の展望— 統計数理研究所彙報，第16巻，第1号，1968, 1-37.

田口時夫「パレート分布とパレート曲線の分析」統計曲線研究所彙報，第12巻，第1号 No.22 創立20周年記念号，1964, 293-314,

統計数理研究所