

SNS の Zipf 則と安定分布¹

Stable Distribution and Zipf's Law in SNS Language Data

山内康英/Yasuhide YAMANOUCHI¹ ・ 小松正/Tadashi KOMATSU² ・ 永井拓史/Takuji NAGAI³¹多摩大学情報社会学研究所 教授 ・ ²多摩大学情報社会学研究所 客員教授、³tn-works 代表

[Abstract]

Zipf's law is the remarkable feature that the rank-ordered distribution of word frequencies follows an approximate power law. The theoretical basis of this pattern, however, has not been fully grounded. Authors' group considered that the stable distribution is the key-factor creating Zipf's law in SNS language data and provided the explanation probabilistically. If SNS contains stable distributions, it is possible to derive Zipf's laws from them using the tutorial by Adamic (2000). According to Hida (2011) the stable distributions composed from compound-Poisson distributions in Lévy-Khinchin formula should be widely observed among power-laws and long-tailed social phenomena. This paper provides Hida's conjecture a good evidence by using Japanese Twitter's data and the computer simulation. This computer simulation indicates a new method to create fat-tailed distributions flexibly.

[キーワード]

安定分布、Zipf 則、SNS、複合ポアソン分布、Adamic の原理、ベキ法則

1. はじめに

Zipf 則は、単語の出現順位（ランク） x と出現回数（頻度） y の関係として、言語データに一般的にあらわれる次のようなベキ法則である。このパタンが出現する理論的な根拠は、十分に解明されていない。

$$y \sim cx^{-\alpha} \quad \alpha \text{ と } c \text{ は定数}$$

本研究の目的は、①SNS の言語データの単語の出現頻度を安定分布と考え、②安定分布を構成するパラメータを時系列的に測定して、社会状況との関係性を示すとともに、③SNS の言語データに安定分布のあらわれる理由を確率論的に説明することである。SNS の言語データの単語の出現頻度が安定分布であれば、本稿が示すように Adamic の原理から、その確率分布をランク図に変換して Zipf 則（ベキ法則）を導くことができる。

Corominas-Murata, Hanel, Thurner (2016)は、個別の著作の言語データにあらわれる Zipf 則の説明を、(1) 情報理論とエントロピー²、(2) 優先的選択 (preferential attachment)³、(3) 自己組織的臨界性 (self-organized criticality)⁴の 3 つの理論に分類し、これとは異なる新しいモデルとして、(4) 標本空間縮退 (sample-space collapse) 理論⁵を提唱した。本研究グループは、言語データとして個別の書籍ではなく、

SNS のログを対象として、3・11 のような大きな社会事象が日本語 Twitter の言語データのベキ法則にどのような影響を与えるのかを検討している。⁶ 個別の著作の Zipf 則では、個別の作品がジャンルや著者の性向によって、どのようなパラメータの差違を示すのかに注目するのに対して、SNS の言語データ分析では、SNS を取り巻く社会事象とパラメータの時間変化に注目することになる。本稿で分析の対象とした SNS の言語データは、日本語 Twitter の 2011 年 3 月 11 日午前 10 時から 16 時の全 tweet である。

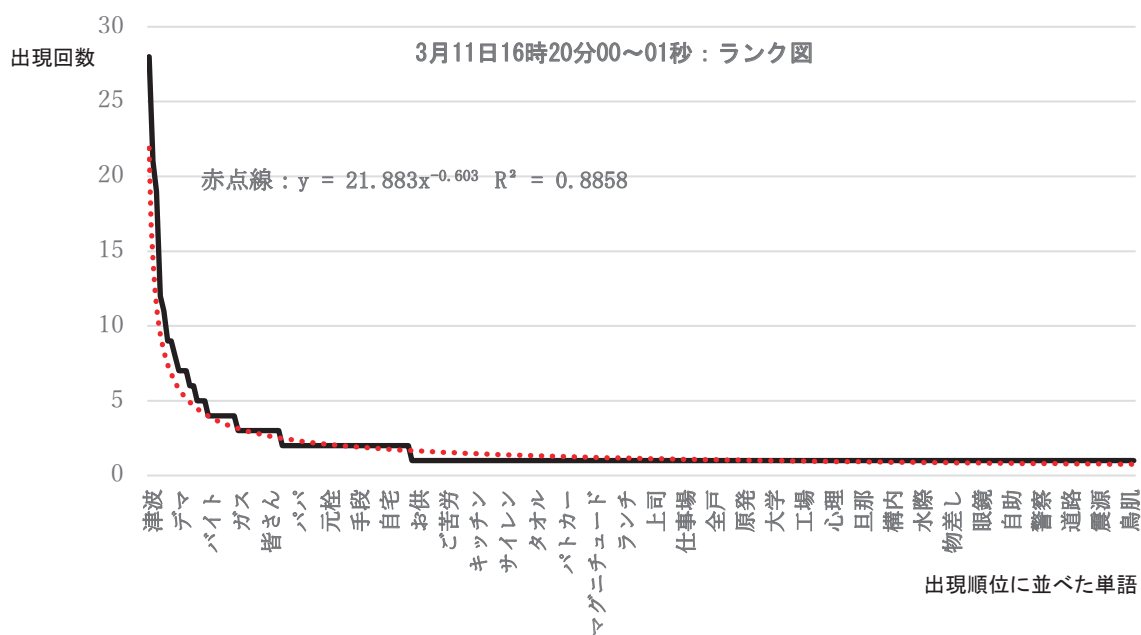
ベキ分布の背後に安定分布があるという本稿の仮説は、2011 年の飛田の著書に基づく。本稿では、これを飛田の予想 (Hida's conjecture) として、実データとシミュレーションをつかって検証した。⁷ 安定分布は、一種のフラクタル (自己相似的) な性質をもつ無限分解可能分布の族で、中心極限定理の自然な拡張として導くことができる。極限分布 (limiting distribution) は、1930 年代に P. Lévy や A. Khinchin が集中的に研究した領域である。拡張された極限分布の概念によれば、中心極限定理の導くガウス分布 (正規分布) は、安定指数 $\alpha=2$ 、歪度母数 $\beta=0$ の安定分布として位置付けられる。⁸

最近のパーソナル・コンピュータの演算能力の向上と解析ソフトの普及によって、安定分布の特性関数のパラメータを直接計算できるようになった。本研究グループも、日本語 Twitter の言語データの安定分布を構成するパラメータの時系列的な測定を課題としてきたが、第 3 節「図 5」が示すように、本稿は安定指数と歪度母数の連続的な測定に初めて成功したものである。1990 年代以降、数理経済学の分野でも Black-Scholes 方程式に対する再検討のなかで、ガウス分布の拡張として安定分布に対する関心が高まった。⁹ 本稿の第 3 節の測定結果 (「表 2」) によれば、SNS の言語データにあらわれる安定分布は、歪度母数が常に 1 の近傍で、安定指数の値が 1 以下の極端に小さな値をとることがあり、書籍にあらわれるベキ分布や金融市場にあらわれる安定分布とは性質が異なっている。

以上のように本稿は、SNS の Zipf 則を構成する確率的なメカニズム (the underlying probability mechanism) を安定分布から説明することを目的としており、本稿の説明の図式は既存のモデルとは異なっている。このために本稿では、第 2 節 (「確率論的説明: 複合ポアソン分布とランダムな個数の確率変数の和」) で言語データにあらわれる安定分布を確率論的に根拠付けた。つぎに第 3 節 (「SNS のデータに表れた安定分布の指数の測定と時間変化」) で、SNS の言語データをコンピュータで解析して安定分布のパラメータを推定した。第 4 節 (「シミュレーションを用いた検証と本研究で利用したソフトウェア」) では、安定分布の構成原理に基づいたシミュレーション・プログラムを作成して、SNS の実証データと比較した。本稿の結論は、SNS の Zipf 則の背後にある安定分布を裏付けている。SNS の言語データは、個別の書籍や金融市場のデータとは確率変数の構成のメカニズムを異にしており、Corominas-Murata 論文が分類した 4 つのモデルとの関係性は今後の検討課題になっている。

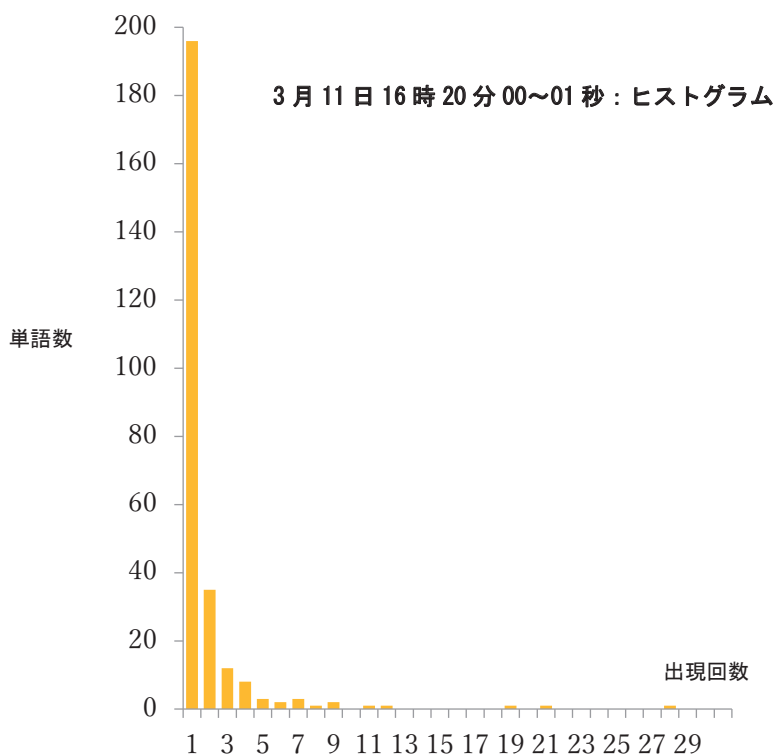
SNS の言語データにあらわれる Zipf 則

「図 1」の黒の実線は、本稿が分析の対象とする 2011 年 3 月 11 日の Twitter のデータのなかで、16 時 20 分 00 秒から 1 秒間の投稿にあらわれた 267 語の名詞を、横軸にその単語の出現回数の順番 (ランク) に並べ、縦軸にその単語の出現回数を取り、折れ線でプロットしたものである。この曲線は、赤点線のように $y=cx^{-a}$ ($c=21.883$, $a=0.603$) で近似することができて、SNS の言語データにあらわれる Zipf 則と長く伸びたロングテイルの特徴を良く示している。¹⁰



【図1：SNSの言語データにあらわれる Zipf 則】

「図2」に「図1」と同じ1秒間のデータをランク図ではなく確率分布のグラフで示した。「図2」は x 軸に級数として出現回数 $\{1, 2, 3, \dots, 26, 27, 28\}$ 、 y 軸に度数として各級数に対応する単語数 $\{196, 35, 12, \dots, 0, 0, 1\}$ を取ったヒストグラム（関数で近似すれば確率密度関数）である。出現回数のもっとも多い単語が確率密度関数では右端（tail）に、ランク図では左端（head）にあらわれるので、両者は横軸の尺度が左右入れ替わることに注意する。



【図2：SNSの言語データ：ヒストグラム】

この Zipf 則（と確率密度関数）は、SNS の言語データの任意の時刻を開始時間とする 1 秒間のデータで検証することができる。3 月 11 日 14 時 46 分の東北大震災の発災の前後で、Zipf 則のパラメータが変化する。¹¹ SNS の Zipf 則の背後にある確率論的なメカニズムや、そのメカニズムの構成する確率分布は同一なので、社会的な事件を受けて確率分布のパラメータが時間的に変化した、と考えることができる。ランク図としての Zipf 則のベキ分布と、確率分布としての安定分布が同一であれば、安定分布のパラメータを調べることによって、ベキ分布を用いるよりも正確に時間変化の様子を観測することが可能になるであろう。

ベキ分布と安定分布に関する飛田の予想

飛田は、2011 年の著書で『レヴィ過程、特に複合ポアソン過程の中で、時間の変形に関して一般の自己相似性をもつものは特に興味がある』¹²として、ベキ分布と安定分布の関係について以下のように述べている。¹³

『統計的なデータの処理で、いわゆるベキ分布（long tail または fat tail をもつ分布としてガウス分布と区別される）が、しばしば現れる。（ベキ分布が）ガウス分布のように素直な分布（特殊な個性的分布ではなく）であれば、「ならし」の累積を想定するのは自然であろう。¹⁴ 背後にレヴィ過程、このときは安定過程があることを仮定してもよいであろう。（中略）吸引域にあることが推測される。

1) そこで試してみたいことは、安定分布からの標本ではなかろうかということ¹⁵である。安定分布なら指数 α （ガウス分布でないため $0 < \alpha < 2$ である）が決まり、その密度関数は $|x|$ が大きいとき $\frac{1}{|x|^{\alpha+1}}$ に比例する。安定分布の特性量である α を決めるために、データに最小 2 乗法を用いるのは適切ではない。安定分布の密度関数は（中略） $x=0$ の近くの様子はわかっていないこと、また密度関数の様子にもよるからである。

2) （中略） $|x|$ が大きいときは漸近的に

$$p_N(x) \sim c|x|^{-(1+N)}$$

が知られている。たとえば Feller(1971)参照。』

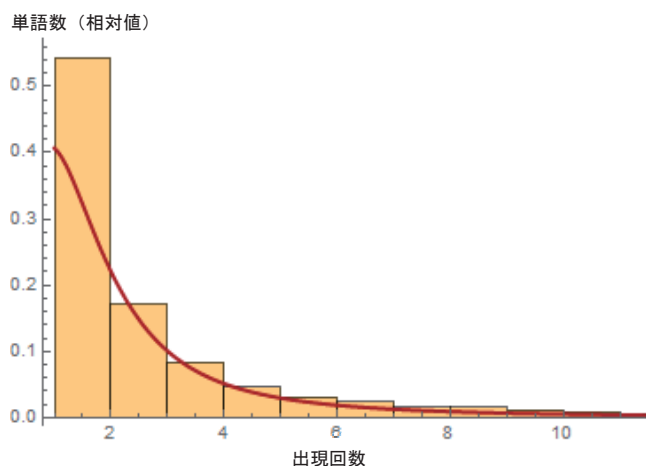
飛田の予想が述べるところの、SNS の言語データの単語の出現頻度の背後にある確率的なメカニズムは、次の通りである。日本語 Twitter のデータは、きわめて多数の投稿者が、それぞれの文脈で投稿する tweet の時系列の累積から出来ている。言語データの語彙頻度の特殊性から、すべての単語の相対的な出現頻度はきわめて低い。¹⁶ 個別の名詞の出現は、少数の法則からポアソン分布に従う。¹⁷ 他方、データベースに累積する個々の単語は、多数の相互に無関係なテキストの入力結果からなる独立な確率変数の合成積として、このポアソン分布を作り出している。 x_i を特定の単語の出現頻度のデータで独立同分布のある確率分布に従う確率変数とし、 n を強度 λ の単語のポアソン分布に従う変数とすれば、言語データにあらわれる個々の単語の分布は、 $P(X, \lambda) = \sum_{i=1}^n x_i$ であり、これは複合ポアソン分布に

なる。したがって言語データ全体にあらわれる分布は、それぞれが個別の重みを持つきわめて多数の複合ポアソン分布の確率変数の和、すなわち畳み込み（重畳：convolution）になる。このような複合ポアソン分布の列の極限は、無限分解可能な分布を構成する。¹⁸ 本研究の分析結果が示すように、この無限分解可能な分布はサンプリングの時間幅ごとに、いたるところで分布収束している。¹⁹ このような SNS の言語データの構成の直接の結果として、単語の出現頻度は全体として安定分布を作り出すのである。すべての単語は複合ポアソン分布に従うため、これを Lévy - Khinchin 表現で見た場合、ガウス分布の構成要素が無く、したがってこの安定分布の指数 α は必ず 2 よりも小さい。次節では以上の議論を数学的に構成して安定分布を導出する。

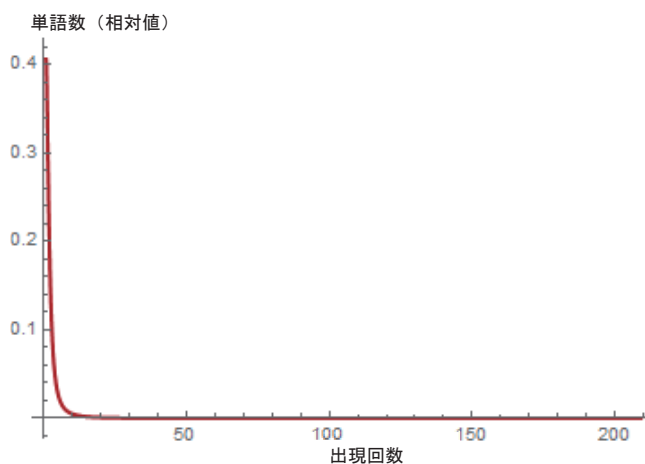
清水（1976）にならって安定分布 ($\alpha \neq 1$) を、つぎの特性関数 $\psi(t)$ によって定義する。²⁰ 安定分布の特性関数は安定指数 α ($0 < \alpha \leq 2$)，歪度母数 β ($-1 \leq \beta \leq 1$)，位置母数 μ （任意の実数），尺度母数 σ （任意の正の実数）を持つ。

$$\psi(t) = \exp\{i\mu t - \sigma|t|^\alpha (1 + i\beta \text{sign}(t) \tan \frac{\pi}{2}\alpha)\} \quad (1)$$

「図 3」「図 4」は、3 月 11 日の発災前の SNS 言語データの状況を示すために、10 時 00 分から 2 分間のデータを Mathematica で解析して、安定分布のパラメータを推定し、そのパラメータを使って確率密度関数を描出したものである。「図 3」は、「図 4」の左端（head）を拡大したもので、実データとの関係を示すためにヒストグラムを描き込んだ。データの大きさは 5357 語、安定分布の 4 つの指標は $\{\alpha, \beta, \mu, \sigma\} = \{1.08947, 1., 5.99234, 0.683748\}$ である。「図 4」はこのデータの全体を示している。歪度母数 $\beta = 1$ は、この分布が最大限左に偏っていることに対応している。「表 1」に 3 月 11 日 10 時 00 分から 2 分間のデータにあらわれた名詞を記載した。左 2 列は最も多く出現した名詞を 20 単語、右 3 列は 2 回および 1 回出現した名詞のなかから 10 単語および 20 単語を選んだ。1 回出現した名詞の数は 2905 単語、2 回出現した名詞の数は 925 単語、3 回 449、4 回 249 である。



【図 3：3 月 11 日 10 時 00 分_2 分データ_head】



【図 4：3 月 11 日 10 時 00 分_2 分データ_全体】

【表 1 : 3 月 11 日 10 時 00 分_2 分データにあらわれた名詞】

自分	208	自動	59	主役	2	友垣	1	古事記	1
定期	195	休み	58	主題歌	2	友引	1	台場	1
花粉	157	お腹	57	乗換	2	友情	1	台所	1
学校	80	電車	54	予備	2	双方	1	右上	1
天気	78	言葉	52	事業主	2	双葉	1	右目	1
感じ	77	ご飯	47	二世	2	取り込み	1	右翼	1
楽しみ	70	友達	47	二次元	2	取手	1	右肩	1
先生	67	コーヒー	46	互換	2	受け	1	右脳	1
気持ち	65	世界	46	人工	2	受取	1	右腕	1
会社	64	人間	45	人手	2	受話器	1	右足	1

2. 確率論的説明：複合ポアソン分布とランダムな個数の確率変数の和²¹

言語データにあらわれる確率分布の構成は、 X_i 個の確率変数から n 個をとって確率変数の和を作るとき n が特定の分布、つまり個別に特定の強度を持つポアソン分布であることを示している。²² この構成は確率密度と確率変数の和を組み合わせた次の混合分布 (mixed distribution) を作り出す。 X_i の分布を $H(x)$ 、混合分布 $Y(x, \lambda)$ の特性関数をそれぞれ $\xi(t)$ 、 $\eta(t)$ とすると、ポアソン分布の確率密度関数 $e^{-\lambda} \frac{\lambda^n}{n!}$ と自然対数の底 e の公式 $\sum_{n=0}^{\infty} \frac{a^n}{n!} = e^a$ を用いて、

$$\eta(t) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} \xi^n(t) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda \xi(t))^n}{n!} = e^{-\lambda} \cdot e^{\lambda \xi(t)} = \exp\{\lambda(\xi(t) - 1)\} \quad (2)$$

ここで $\xi^n(t)$ は X_i の n 個の畳み込み (確率変数の和) をあらわしている。特性関数と確率分布の定義式 $\xi(t) = \int_{-\infty}^{\infty} e^{itx} dH(x)$ 、 $1 = \int_{-\infty}^{\infty} dH(x)$ を使うと、 $H(x)$ を要素とする複合ポアソン分布の特性関数を定めることができる。²³

$$\eta(t) = \exp\left\{\lambda \int_{-\infty}^{\infty} (e^{itx} - 1) dH(x)\right\} \quad (3)$$

(3) の複合ポアソン分布の特性関数は、この分布が無限分解可能な分布であることを示している。なぜならば、 $\eta_n(t) = \exp\left\{\frac{\lambda}{n} \int_{-\infty}^{\infty} (e^{itx} - 1) dH(x)\right\}$ とおくと、 $(\eta_n(t))^n = \eta(t)$ を満足する。言い換えると分布 $Y(x, \lambda)$ を持つ確率変数は、すべての n に対して、ある分布の n 個の連続的な畳み込み、すなわち標本の和=データの蓄積によってあらわすことができるからである。²⁴

単語の出現頻度の配列と確率分布の収束

単語の出現頻度の 2 重系列からなる配列 (array) を用いて、データの確率変数全体の和の確率分布の極限を考える。SNS の言語データにあらわれる単語は一つ一つが異なるポアソン分布の強度を持ち、データを取る時間幅に相応した出現回数 $\{1, 2, \dots, n\}$ を持つ。各出現回数は、一定範囲の強度を持つ単語を含んでいるので、これを X_{nk_n} の系列として、つぎの確率変数列の配列 (4) を構成する。²⁵

$$\begin{array}{cccc}
X_{11}, & X_{12}, & X_{13}, \dots, & X_{1k_1} \\
X_{21}, & X_{22}, & X_{23}, \dots, & X_{2k_2} \\
X_{31}, & X_{32}, & X_{33}, \dots, & X_{3k_3} \\
\dots & \dots & \dots & \dots \\
X_{n1}, & X_{n2}, & X_{n3}, \dots, & X_{nk_n}
\end{array} \quad (4)$$

nk_n の添字 k_n の値は、その系列の確率変数の数を示す。 k_n には大小があるので配列の全体の形を凹凸を含む三角形にすることができる。²⁶ 配列 $\{X_{nj}\}$ が、つぎの条件を満たすとき、極限において微少 (infinitesimal) であり、漸近的に無視可能な成分を持つ。

$$\lim_{n \rightarrow \infty} \max_{1 \leq j \leq k_n} \Pr\{|X_{nj}| > \varepsilon\} = 0$$

配列 $\{X_{nj}\}$ が微少であるための必要十分条件は、 $\{X_{nj}\}$ の特性関数を a_{nj} とするとき、 t の任意有限区間で一様に (5) が成り立つことである。²⁷

$$\lim_{n \rightarrow \infty} \max_{1 \leq j \leq k_n} |a_{nj}(t) - 1| = 0 \quad (5)$$

$\{a_{nj}\}_{j=1, \dots, k_n; n=1, 2, \dots}$ は標本 $\{X_{nj}\}$ の特性関数の配列で (6) および (7) の条件を満たすとする。

$$\lim_{n \rightarrow \infty} \max_{1 \leq j \leq k_n} |1 - a_{nj}| = 0 \quad (6)$$

$$\sum_{j=1}^{k_n} |1 - a_{nj}| \leq C < \infty \quad (C \text{ は正の定数}) \quad (7)$$

このとき次式が成り立つ。

$$\lim_{n \rightarrow \infty} \left| \prod_{j=1}^{k_n} a_{nj} - \exp \left\{ \sum_{j=1}^{k_n} (a_{nj} - 1) \right\} \right| = 0 \quad (8)$$

証明 $\delta_n = \max_j |1 - a_{nj}|$ とおく。 $a_{nj} \neq 0$ なので $\log a_{nj}$ をとれば、 $\log a + (1 - a)$ を $a = 1$ の近傍で展開した

$\log a + (1 - a) = -\frac{1}{2}(a - 1)^2 + \frac{1}{3}(a - 1)^3 + O[a - 1]^4$ を用いて、²⁸

$$\left| \log \left(\prod_{j=1}^{k_n} a_{nj} \right) / \exp \left\{ \sum_{j=1}^{k_n} (a_{nj} - 1) \right\} \right| \leq \sum_{j=1}^{k_n} |\log a_{nj} + (1 - a_{nj})| \leq \frac{1}{2} |1 - a_{nj}|^2 + \frac{1}{3} |1 - a_{nj}|^3 + \dots$$

右辺から $|1 - a_{nj}|^2$ を括りだして (6) (7) で置き換えれば、

$$\leq \delta_n \sum_{j=1}^{k_n} |1 - a_{nj}| \left(\frac{1}{2} + \frac{1}{3} |1 - a_{nj}| + \dots \right) \leq \delta_n C = 0 \quad (n \rightarrow \infty)$$

各出現回数に収まるポアソン分布の強度を持つ単語の数は有限なので、最初の配列は (6) (7) を満たす。(8) と (3) の式のかたちを比較すれば、すべての確率変数の特性関数が収束して、その極限は複合ポアソン分布の特性関数に等しく、したがって最初の配列 (4)、すなわち言語データにあらわれる単語の出現頻度の確率分布は $n \rightarrow \infty$ で無限分解可能である。

安定過程にある無限分解可能分布

すべての範囲で標本の確率変数 X に吸引域 (domain of attraction) があるとき、この標本は安定過程にある。²⁹ 吸引域 Z_n を持つとは、 X が独立同分布 (i.i.d.) の確率変数の系列であって、正の実数列 $A_n (\lim A_n = \infty)$ と B_n に対して以下を満たすことである。

$$\frac{X_1 + X_2 + \cdots + X_n}{A_n} - B_n \xrightarrow{d} Z_n \quad (9)$$

$\xrightarrow{d}$ は分布収束をあらわす記号である。「図 2」の 1 秒間のヒストグラムは、この言語データがあらゆる範囲で分布収束していること、したがって吸引域にあることを示している。無限分解可能な分布が安定過程にあるとき、無限分解可能分布の Lévy - Khinchin 表現と (9) から、安定分布に固有な Lévy - Khinchin 表現の Lévy 測度が決まり、安定分布の特性関数の式 (1) と安定指数の値 $0 < \alpha \leq 2$ が定まる。

³⁰ ここで Lévy 測度とは複合ポアソン分布の要素となる $H(x)$ に積分関数の収束条件を加味したものである。³¹

Adamic の原理：ランク図と確率密度関数の相互交換性について

「図 1」のランク図は、横軸 (X 軸) で r 番目の事象が縦軸 (Y 軸) で n 回の頻度をもつことを示している。これは Y 軸で n 以上の頻度をもつ事象が X 軸で r 個あることに等しい。つまり $P[Y \geq y] \sim X^{-b}$ である。このように Y と X の関係でランク図が一種の累積確率分布になっていることに注意して、 $y = x^{-b} (b > 0)$ つまり $y^{-1/b} = x$ 【ランク図】ならば、 X 軸と Y 軸を交換して $y = x^{-1/b}$ 【累積分布関数】、したがって確率密度はこれを 1 回微分した $y = \frac{1}{b} x^{-(1+1/b)}$ 【確率密度関数】になる。この関係は、Zipf 則

のランク図を構成する過程が必ずベキ分布の確率密度関数をもつことを示しており、逆も成立する。³²

「図 4」が示すように、SNS の言語データの安定分布は単峰が左端で消失しており、その密度関数は、形として x の大きな部分のベキ分布と連続してつながっている。³³ ベキ分布の確率密度関数をもてば Zipf 則のランク図を構成するから、以上によって SNS の言語データの単語の出現頻度に Zipf 則のあらわれる理由を説明することが出来た。言語データのランク図が Zipf 則を作り出す理由は、 β が常に 1 の近傍にあって歪度が最大である要因が大きい。

3. SNS のデータに表れた安定分布の指数の測定と時間変化

以上の SNS 言語データの構成のメカニズムは、安定分布を作り出す混合分布の要素となる $H(x)$ の確率分布の変化が、安定分布のパラメータに直接影響することを示している。2011 年 3 月 11 日 14 時 46 分 18 秒の東日本大震災の発災と同時に、Twitter の言語データの安定指数に大きな変化が生じた。

安定指数 α は通常、平時の 1 前後で安定しているが、大震災と津波が Twitter の話題を占有するに従って α の値は振動し、もっとも小さなきには平時の 10 分の 1 以下に低下した。この様子を「図 5」と「表 2」に示した。 α の最大値は 10 時 52 分の 1.19338、最小値は 15 時 34 分の 0.089433 である。「表 2」では、この 2 つの時間を黄色で、発災の時間を赤色で示した。1 冊の作品の言語データが、このように小さな安定指数を示すことは考えられない。



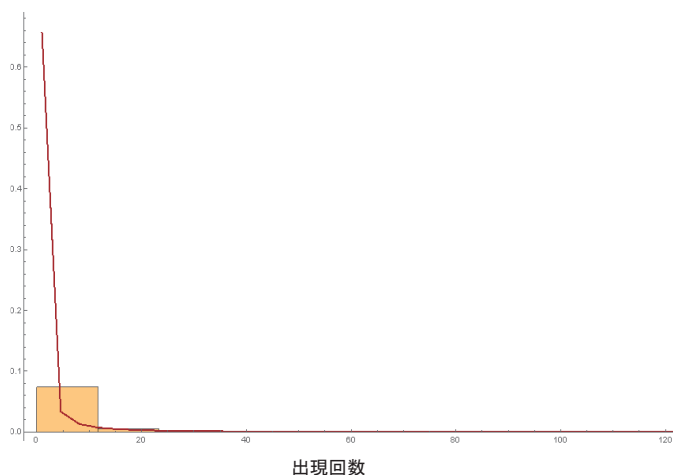
【図 5：安定分布のパラメータの時間変化】

【表 2：3 月 11 日 10 時 58 分から 16 時 04 分の 2 分データで計測したパラメータの推移】

時刻	10時48分	10時50分	10時52分	10時54分	10時56分	10時58分	14時30分	14時32分	14時34分
α	1.08696	1.08943	1.19338	1.08888	1.08704	1.08928	1.01787	1.01787	1.01787
β	1	1	1	1	1	1	1	1	1
時刻	14時36分	14時38分	14時40分	14時42分	14時44分	14時46分	14時48分	14時50分	14時52分
α	1.01016	1.08706	1.01787	1.01787	1.01787	1.01787	0.959312	0.802252	0.786335
β	1	1	1	1	1	1	1	1	1
時刻	14時54分	14時56分	14時58分	15時00分	15時02分	15時04分	15時06分	15時08分	15時10分
α	1.01787	0.525855	0.786335	0.135611	0.496137	0.162305	0.15211	0.095572	0.132695
β	1	1	1	0.763742	1	0.999998	1	1	0.999991
時刻	15時12分	15時14分	15時16分	15時18分	15時20分	15時22分	15時24分	15時26分	15時28分
α	0.200563	0.865712	0.722238	0.162806	0.143558	0.812476	0.802536	0.132009	0.118253
β	0.999963	1	1	0.999372	0.999789	1	1	0.999999	0.967173
時刻	15時30分	15時32分	15時34分	15時36分	15時38分	15時40分	15時42分	15時44分	15時46分
α	0.153917	0.100113	0.089433	0.462056	0.433477	0.868265	0.150689	0.109361	0.1122
β	1	0.962716	1	1	1	1	0.999987	1	1
時刻	15時48分	15時50分	15時52分	15時54分	15時56分	15時58分	16時00分	16時02分	16時04分
α	0.711686	0.702045	0.109404	0.121793	0.111944	0.140519	0.133219	0.791007	0.156628
β	1	1	1	0.784185	0.999998	0.999999	0.996241	1	0.999998

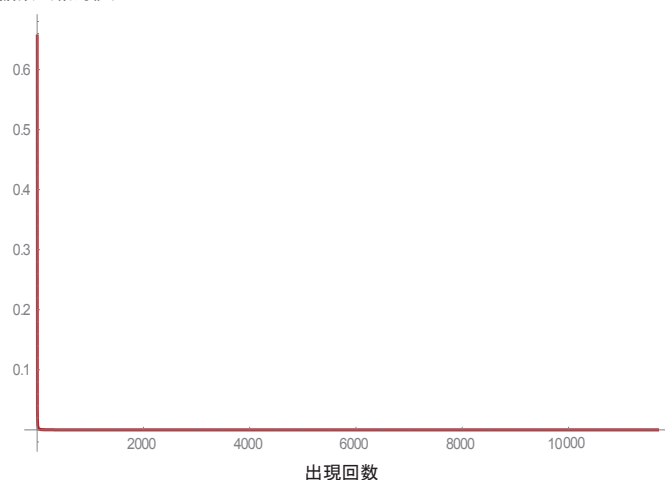
15 時 02～04 分のデータの安定分布のグラフと 2 分間のデータにあらわれた名詞を「図 6」「図 7」および「表 3」に記載した。データの大きさは 6125 語である。地震や震度といった最頻出語の数が莫大であるため、このデータにあてはめた安定分布のグラフは右端が大きく伸び、「図 3」に比べて L 字型が強調されている。

単語数（相対値）



【図 6：3 月 11 日 15 時 02 分_2 分データ_head】

単語数（相対値）



【図 7：3 月 11 日 15 時 02 分_2 分データ_全体】

【表 3：15 時 02 分～04 分のデータにあらわれた名詞と出現回数】

地震	11676	火事	1016	火災	629	建物	380	八雲	1
震度	6841	気持ち	967	電気	585	フィギュア	363	公国	1
揺れ	2954	皆さん	965	東北	548	ニュース	320	公文	1
津波	2702	実家	940	食器	487	マンション	320	公民館	1
台場	2582	電車	896	家族	478	沿岸	316	公爵	1
部屋	1633	被害	780	宮城	440	パソコン	313	共食い	1
余震	1586	自分	774	北部	404	状況	308	兵隊	1
テレビ	1253	本棚	744	感じ	401	マグニチュ	307	典型	1
ビル	1208	情報	706	友達	400	地方	307	内耳	1
震源	1056	会社	691	大震災	381	パニック	299	内閣	1

4. シミュレーションを用いた検証と本研究で利用したソフトウェア

本研究グループは、SNS の言語データの背後にある確率的なメカニズムをシミュレーションによって検証するために、複合ポアソン分布の確率変数の和を用いて安定分布を構成するプログラムを作成し、このプログラムによって SNS の言語データの示す確率分布の形が再現出来ることを確認した。このため変数に以下の前提をおいて仮想的なデータを作成した。

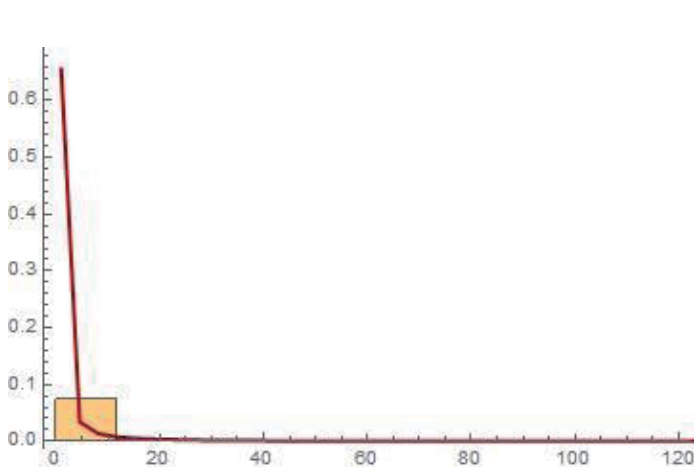
(i) x を乱数発生の時間幅、 y を生起回数、 z を事象が 1 回起こる小区間の個数、 v を重み付けの変数とする。³⁴

```

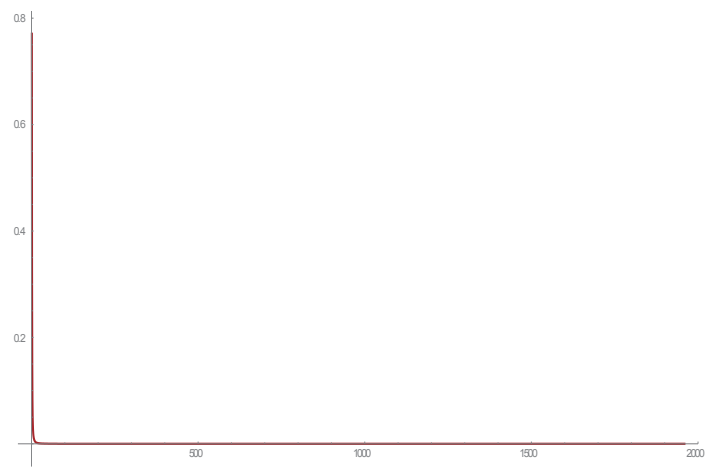
z = 600; y = 200;
x = 400; v = 1000;
a = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 200; v = 500;
b = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 100; v = 200;
c = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 50; v = 100;
d = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 40; v = 20;
e = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 20; v = 10;
f = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 10; v = 5;
g = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 5; v = 4;
h = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 1; v = 3;
i = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
x = 0.1; v = 2;
j = Replace[Flatten[Table[Total[Table[Length[Select[RandomReal[x, y], LessThan[1/z]], z]], v]],
  _?PossibleZeroQ -> Nothing, {1}];
data = Flatten[{a, b, c, d, e, f, g, h, i, j}];
estimate = EstimatedDistribution[data, StableDistribution[ $\alpha$ ,  $\beta$ ,  $\mu$ ,  $\sigma$ ]];
pdf = PDF[estimate, n];
hist = Histogram[data, Automatic, "PDF"];
Show[hist, Plot[pdf, {n, 1, Max[data] + 1}, PlotRange -> All, PlotStyle -> {Darker[Red], Thick}]]
estimate
Length[data]
data
Plot[pdf, {n, 1, Max[data] + 1}, PlotRange -> All, PlotStyle -> {Darker[Red], Thick}]

```

【図 8 : シミュレーションで用いた Mathematica のプログラム】



【図 9 : シミュレーション結果_head】



【図 10 : シミュレーション結果_全体】

(ii) $\frac{y}{x} = \lambda$ はポアソン分布の強度になる。前頁の「図 8」は、このシミュレーションで用いた Mathematica のプログラムで、 $\{a, b, c, d, e, f, g, h, i, j\}$ は、数式 (4) の配列（この場合は 10 行）で確率変数 $X_m (m=10)$ のデータ系列を構成する。 $\{a, b, \dots\}$ で x と y を毎回指定することによって強度 $= \{0.5, 1, 2, 4, 5, 10, 20, 40, 100, 1000\}$ 、重み付け $= \{1000, 500, 200, 100, 20, 10, 5, 4, 3, 2\}$ の確率変数のデータ系列を生成する。ここで用いた強度と重み付けの数値は、強度のより小さなポアソン分布が全データのなかでより大きな割合を持つ、という関係をあらわす以上のものではない。

(iii) 「図 8」のプログラムの $data = Flatten[\{a, b, c, d, e, f, g, h, i, j\}]$ で、確率変数のデータ系列の和を取って全データを構成する。この Mathematica のコマンドは、数式では全ての複合ポアソン分布の畳み込み積分に相当する。これは、Lévy 表現³⁵を逆に使って、多数の複合ポアソン分布から安定分布を構成するプロセスを意味し、ガウス分布を含まないホワイトノイズ解析になっている。³⁶ このデータに対して $EstimatedDistribution[data, StableDistribution[\alpha, \beta, \mu, \sigma]]$ を用いて安定分布のパラメータを推定する。このコマンドは Mathematica の組み込み関数³⁷である。

シミュレーションの結果、 $\{\alpha, \beta\} = \{0.835538, 1.\}$ を得た。極端な L 字型を成す安定分布のグラフは、「図 6」「図 7」のグラフと同じ形を持つ。(ii) の変数は、「地震」や「津波」のような過度に強度の高い確率変数 ($\{\lambda = 1000, 100, \dots\}$) に小さな重み付け ($\{2, 3, \dots\}$) を振る極端な構成になっている。ポアソン分布の強度、配列の行の重み付け、配列の行数を変えることによって、ロングテイルをもつ単峰のグラフとパラメータを連続的に得ることができる。

言語データの解析手法と安定分布の解析ツール

本研究で扱う Twitter のデータは大量なので、言語データのパラメータの推計の前工程として、任意の開始時間と時間幅で tweet の時系列データを切り出し、茶釜を用いて形態素解析をする仕組みを作った。Twitter のデータ（投稿文）を形態素解析に掛ける際に、次の 2 つのルールを適用した。

【ルール 1】分析対象としない投稿文の除外ルール：①リツイートは除外する。具体的には、(ア) ツイート本文中に RT(Retweet)を含むツイートを除外する。(イ) ツイート本文中に QT(Quote Tweet)を含むツイートを除外する。ただし、除外するのは大文字の場合 (RT および QT) のみであり、小文字の場合 (rt および qt) は除外しない。②ウェブサイトのリンクが張られたツイートを除外する。

【ルール 2】名詞抽出ルール：①茶釜による品詞分類の結果から「名詞一般」を抽出したうえで、以下を除外する。(ア) 2 文字未満、(イ)「固有名詞」および「代名詞」、(ウ) 全てひらがなの名詞、(エ) 全て英字（全角）の名詞。

安定分布のパラメータを推計する後工程に研究者の利用できる汎用ソフトウェアとして、①R の安定分布解析コマンド³⁸、②Nolan の運営する Robust Analysis 社が提供する R および MATLAB のライブラリ³⁹、および、③Mathematica のコマンドなどがある。⁴⁰ McCulloch は、安定分布のパラメータを推定するために、データ・サンプルから 5 つの分位数を用いる手法を開発し、いくつかのライブラリはこの手法を採用している。McCulloch の手法は計算測度が速いが、 α が 0.6 以下のデータには対応できな

い。⁴¹ 本研究グループは、Mathematica を用いてデータを解析し⁴²、さらに R を用いてパラメータを推計したが、上記のライブラリとの関係からパラメータの範囲に制限があり、Mathematica の計算結果との突合は今後の課題になっている。

5. まとめにかえて

本稿では、飛田の予想から出発して、言語データにあらわれる Zipf 則に確率論的な根拠付けを与えるとともに、2011 年 3 月 11 日の Twitter のデータを用いて、安定分布を確率的な根拠とするベキ分布が広く社会現象に見られるという飛田の予想に例証を与えた。

Adamic が強調するように、順位-規模法則としての Zipf 則、確率密度分布としてのベキ分布、ランク図の累積分布関数であるところの Pareto 分布は相互に変換可能である。したがって、この 3 種類の統計手法の意味するところは、いずれも非 - 正規分布としての安定分布と密接に関係していることになる。「図 1」と「図 2」を比較すれば明らかなように、ランク図は確率密度関数にはない情報を有しており、人口に膾炙したロングテイルの言説⁴³が示すように、データ分析の実践的ツールとしての価値が高い。これはパレート図が日本の品質管理（QC 活動）のツールになってきたのと軌を一にしている。

⁴⁴ ランク図、パレート図、ロングテイルの確率論的＝統計学的な根拠付けとシミュレーションのアルゴリズムが、このように非 - 正規分布としての安定分布を通じて成されるとすれば、その意義は大きいであろう。

その際、安定分布の構成のメカニズムが、数式 (3) の $H(x)$ を要素とする複合ポアソン分布に根拠を置いている、という点は重要である。安定分布は、複合ポアソン分布の要素となるきわめて多数の独立な確率変数と混合することを通じて外部の確率事象、たとえば社会的事件の情報と結び付いているのである。どのような $H(x)$ が、言語データのもつ特異な歪度母数や安定指数と関係するのかという検討は、今後の課題である。

3 シグマ管理限界といった製造業の工程管理が正規分布を利用するのに対して、原則的に非-正規分布の分布族を必要とする社会科学系のデータ・サイエンスの領域として、保険統計、災害予測、確率論的リスク評価、オンライン・ビジネスおよび金融理論を挙げることができる。産業活動の保険数理、地震や津波などの被害予測、原子力発電所の確率論的リスク解析では、ビジネスの運営や政策策定の前提として、きわめて希に起こる事象や大きな外れ値を含むデータを整合的に取り扱う必要がある。またオンライン・ビジネスでは、イノベーションを引き起こす必要から、確率変数の収束として裾の厚い分布をビジネスモデルとして取り扱うことになる。このような領域では必然的に何らかの混合分布が登場する。

インターネットでは、きわめて多数の自律分散的な活動を、サーバやルータのデータで集約的に観測することができる。サーバやルータの指標は、これに接続する外部の多数の独立同分布の確率変数の和＝合成積が作り出す確率変数であって、その結果は何らかの混合分布になっており、合成積が作り出す確率変数が複合ポアソン分布であれば、その重み付け付きの畳み込みは全体として安定分布を構成する。このような訳でインターネットに多くの Zipf 則があらわれるのは自然である。⁴⁵

【註】

¹ 本稿は、多摩大学情報社会学研究所が株式会社構造計画研究所との 2019 年度の委託契約に基づいて実施した調査研究を利用したものである。研究の実施にあたって飛田武幸名誉教授（名古屋大学）、公文俊平教授（情報社会学研究所）、服部正太取締役社長（株式会社構造計画研究所）、SiSi 教授（ヤンゴン工科大学）からご指導を戴いた。

² Mandelbrot (1953)は Zipf(1949)のオリジナルの説明を再検討し Shannon の情報理論を用いて発展させた。

³ Yule の研究を発展させた Simon(1955)のモデルは、単語の既出現回数に比例して、その単語の出現頻度が高くなるという仮定を利用する。Barabási, Albert-László and Réka Albert (1999)は、同じメカニズムをネットワーク理論の観点から優先的選択理論として定式化した。藤田・クルーグマン・ベナブルズ (2000) は、第 12 章「都市規模に関する実証分析」で都市の規模・順位に関する Zipf 則を、Simon(1955)に遡って検討し、都市の規模のベキ分布をこれとは異なる経済学的前提にもとづいて導出しようとした。

⁴ Bak (1996).

⁵ Thurner, Stefan, Rudolf Hanel, Bo Liu, and Bernat Corominas-Murtra (2015).

⁶ 山内康英、小松正 (2016)。

⁷ 飛田武幸教授 (1927～2017 年)は確率論、関数解析学、ホワイトノイズ解析の研究者で、名古屋大学理学部長、確率過程論国際学会会長、名城大学理工学部長を歴任した。本研究グループは 2016 年まで飛田教授と共同研究を行った。この共同研究の趣旨について公文 (2006) 9 頁の以下の記述を参照。『さらに最近始まったばかりの研究として、ブラウン運動論の世界的権威である飛田武幸によるベキ法則の数理の研究がある。飛田はわたしたちの求めに応じて、ベキ法則を「数学的に安定的な法則」とみなす立場から、ベキ法則を安定分布へ、さらに安定過程へ、最終的には（複合）ポアソン過程に還元して、その数理の分析を試み始めている。』

⁸ 安定指数と歪度母数については数式 (1) を参照。安定分布の簡潔な説明として、フェラー (1969) 第 VI 章、飛田 (2011) 第 10 章がある。確率過程の時点 t における分布が、 $t \rightarrow \infty$ のとき初期条件に無関係な分布に収束するとき、その分布を極限分布という。中心極限定理の条件を満たせば、はじめの分布に関係なく、その確率過程はガウス分布に収束するという意味で、ガウス分布は極限分布である。

⁹ McCulloch (1996).

¹⁰ 「図1」は267語を取ったので横軸に示した単語は飛び飛びになる。したがって黒実線のプロットは表示されない語で変化している。このグラフで用いたデータの加工については、第4節第2項「言語データの解析手法と安定分布の解析ツール」の前工程を参照。Clauset, et.al. (2009)は、異なる研究領域から24件のデータを取り上げてベキ分布 (power law distribution) に対する厳密な検定手法を検討した。このなかで両対数グラフの傾きや最小2乗法を使ったベキ指数の特定が不正確であるため、Adamicの主張するようにパレート分布の累積確率分布を用いるべきだ、と主張している。かれらの結論によればベキ分布と呼ぶべき分布は明らかに存在する。

¹¹ この Twitter のデータに関するベキ分布のパラメータの変化については山内康英、小松正 (2016) を参照。

¹² 飛田 (2011)、223 頁。

¹³ 飛田 (2011)、228 頁。

¹⁴ 『ならし』の原語は smearing である。これは確率過程で独立同分布の確率変数が重なっていくにつれて、確率分布曲線が急速に滑らかになっていく様子を、塗装の塗り重ねに例えて表現している。ここで『ガウス分布のように素直な分布』とは、ロングテイルを持つ分布もやはり極限分布であることを意味する。

¹⁵ 飛田教授の予想を明確にするため引用に執筆者が下線を付した。

¹⁶ 『テキストにおける語の分布には非常にむらがある。高頻度語はごく少数で、比較的高頻度の語が

いくらかあり、ごく低頻度の語が大半なのである。』スタッブズ (2011)、53 頁。高頻度語は機能語（助動詞、代名詞、前置詞、接続詞など）であり、内容語（名詞、形容詞、副詞、本動詞）は低頻度である。本研究では形態素解析により名詞を選択したので、単語の出現頻度は低頻度である。

¹⁷ 単語の出現頻度がポアソン分布にしたがうことの検証については、山内康英、小松正 (2016) を参照。

¹⁸ 清水 (1976)、95 頁の定理 3.4。『(ただし) 連続分布は複合ポアソン分布ではありえないから、無限分解可能な分布の族は複合ポアソン分布の族より広い。』

¹⁹ 本稿では分布収束と弱収束を同意義に用いる。いたるところで分布収束するというのは吸引域にある、ということに等しい。

²⁰ 清水 (1976)、129 頁。パラメータのギリシャ文字アルファベットは Mathematica に合わせた。

²¹ 本稿の数学的な説明は清水 (1976)、飛田 (2011)、佐藤 (1990) に依拠した。

²² フェラーは、『確率論とその応用 I (下巻)』第 XII 章で確率母関数を使って複合ポアソン分布から無限分解可能な分布の導出を、また『確率論とその応用 II (上巻)』第 VI 章では特性関数を使って安定分布の概要を述べている。

²³ 清水 (1976)、90~92 頁。飛田 (2011)、109 頁。

²⁴ 確率変数の和は特性関数の積になる。

²⁵ 清水 (1976)、156 頁。

²⁶ この項の記述は清水 (1976)、156~158 頁に依る。フェラーはこの 2 重系列を 3 角形配列 (triangular array) と名付けた。フェラー (1969)、242 頁。

²⁷ この証明は、清水 (1976)、第 5 章補助定理 1、156~7 頁。

²⁸ 清水 (1976)、第 5 章補助定理 2、158 頁。出現しない単語は計測されないので、言語データの出現頻度では $P(X=0)=0$ である。これは確率密度関数の左端のかたちに関係する。

²⁹ Samorodnitsky は、安定分布の定義を 4 つに整理し互いに同値であることを示している。

Samorodnitsky & Taqqu (1986)、pp.1-10. 安定分布を吸引域と関係付ける定義は、Samorodnitsky が列挙した 3 番目のもので、Khinchine の研究に起源を持つ。Zolotarev (1986)、pp.3-5. 清水 (1976)、121 頁。

³⁰ $\alpha=2$ のときガウス分布となるように Lévy - Khinchin 表現は作られている。

³¹ 清水 (1976)、95~101 頁。飛田 (2011)、109~110 頁。

³² Zipf 則のランク図を構成する過程は原理的にベキ分布の確率密度関数をもつ。Adamic の "ranking tutorial" にしたがって、本稿ではこれを Adamic の原理と呼ぶ。Adamic (2000)。

³³ 註 13 の飛田 (2016) の引用の 2) を参照。

³⁴ (i) はポアソン過程の定義に他ならない。0 の近傍では確率密度関数が発散するためパラメータの推定に失敗する。 $_{?}PossibleZeroQ \rightarrow Nothing$ により確率変数からすべての 0 を取り除いた。言語データの出現回数は 1 以上の自然数であるため、これは標本空間の性質と矛盾しない。

³⁵ Lévy 表現は、積分方程式をガウス分布と複合ポアソン分布に分けて、特定の複素数値関数が無限分解可能な特性関数であるための必要十分条件を示すものである。清水 (1976)、99 頁、数式 (3.31) を参照。

³⁶ ポアソン・ノイズについて飛田 (2014)、35~44 頁。

³⁷ これに対して R ではさまざまなライブラリを組み合わせることができる。

³⁸ デフォルトでは McCulloch 法を初期値として最尤推定法を用いるが初期値を手動とする設定もある。

³⁹ Robust Analysis Inc., *User Manual for STABLE 5.3 R Version*, 18 July, 2017. <http://www.robustanalysis.com/>

⁴⁰ Koutrouvelis のアプローチは、計算機のマルチコア CPU に対応したライブラリで、並列計算が可能になる。Royuela-del-Val, Javier, Federico Simmross-Wattenberg, Carlos Alberola-López (2017)。

⁴¹ Wuertz (2017)。

⁴² 日本語版 Mathematica は 5 月にリリースされた Ver.12 を用いた。

⁴³ アンダーソン (2009)。

⁴⁴ 『パレート (Pareto) 図』というのは、一種の度数分布である。たとえば、不良、手直し、ロス、クレームなどの件数、損害金額やパーセントを、その原因別あるいは状況別にデータを取り (中略) 大ききの順に並べた図をいう。このように並べ、さらに実線のようにつぎつぎ加えて曲線 (累積曲線) を描いてみると、不良のうち最初の 2~3 の不良項目で不良の 70~80%以上を占める場合が多く、これを撃滅すれば不良の大半はなくなって、不良率が非常に減少することがわかる。一般に、社会におけ

る不良、ロス、事故、その他いろいろな問題は、項目や原因は非常にたくさんあっても、そのうち 2 ～ 3 の主な項目や原因が圧倒的に影響していることが多い。これをわれわれはパレートの原則 “vital few, trivial many、重要な問題は数が少なく、くだらない問題はたくさんある”、といている。』石川 (1989)、127 頁。著者は日本の品質管理の父と呼ばれている。

⁴⁵ Adamic, and Huberman (2002).

【引用文献】

アンダーソン、クリス (2009) 『ロングテール：「売れない商品」を宝の山に変える新戦略 (アップデート版)』 篠森ゆりこ訳、早川書房。

石川馨 (1989) 『品質管理入門 (第 3 版)』 日科技連。

公文俊平「情報社会学の諸側面」情報社会学会編『情報社会学会誌』Vol. 1 No. 1。

佐藤健一 (1990) 『加法過程』 紀伊國屋書店。

清水良一 (1976) 『中心極限定理』 教育出版。

スタッブズ、マイケル (2006) 『コーパス語彙意味論』 南出康世、石川慎一郎監訳、研究社。

飛田武幸 (2011) 『確率論の基礎と発展』 共立出版。

飛田武幸 (2012) 「時空と偶然」『数学セミナー』 日本評論社、5～8 号。

飛田武幸 (2014) 『ホワイトノイズ』 丸善出版。

フェラー、ウィリアム (1961) 『確率論とその応用 I (下巻)』 河田龍夫監訳、紀伊國屋書店。

フェラー、ウィリアム (1969) 『確率論とその応用 II (上巻)』 国沢清典監訳、紀伊國屋書店。

フェラー、ウィリアム (1970) 『確率論とその応用 II (下巻)』 国沢清典監訳、紀伊國屋書店。

藤田昌久、ポール・クルーグマン、アンソニー・ベナブルズ (2000) 『空間経済学：都市・地域・国際貿易の新しい分析』 東洋経済新報社。

山内康英、小松正 (2016) 「SNS 言語空間のベキ法則と安定分布」情報社会学会編『情報社会学会誌』、Vol. 11 No. 1。

Adamic, Lada A. (2000). “Zipf, Power-laws, and Pareto - a ranking tutorial.”
<https://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>

Adamic, L.A. and Bernardo A. Huberman (2002). “Zipf’s law and the Internet,” *Glottometrics*, 3.

Barabási, Albert-László and Réka Albert (1999). “Emergence of scaling in random network,” *Science*, Vol.286, 15 October.

Bak, Per (1996). *How Nature Works: the science of self-organized criticality*, Copernicus.

Clauset, Aaron, Cosma Rohilla Shalizi, and M.E.J. Newman, (2009). “Power-Law Distributions in Empirical Data,” *Society for Industrial and Applied Mathematics, SIAM Review*, Vol. 51, No. 4, December.

Corominas-Murtra, Berna, Rudolf Hanel, and Stefan Thurner (2015). “Understanding scaling through history-dependent processes with collapsing sample space,” *PNAS*, April 28.

Feller, William (1971). *An Introduction to Probability Theory and its Application*, Volume II, Second Edition, Wiley.

Mandelbrot, Benoît (1953). “An information theory of the statistical structure of language,” W. E. Jackson(ed.), *Communication Theory*, Academic Press.

McCulloch, J. Huston (1996). “Financial Applications of Stable Distributions,” G.S. Maddala and C.R.Mao, eds, *Handbook of Statistics 14, Statistical Methods in Finance*, Elsevier Science B.V.

Royuela-del-Val, Javier, Federico Simmross-Wattenberg, Carlos Alberola-López (2017). “libstable: Fast,

Parallel, and High-Precision Computation of α -Stable Distributions in R, C/C++, and MATLAB," *Journal of Statistical Software*, Volume 78, Issue 1, June.

Samorodnitsky, Gennady and Murad S. Taqqu, (1994). *Stable Non-Gaussian Random Processes, Stochastic Models with Infinite Variance*, Chapman & Hall.

Simon, Herbert A. (1955). "On a class of skew distribution functions," *Biometrika*, Volume 42, Issue 3-4, December.

Thurner, Stefan, Rudolf Hanel, Bo Liu, and Bernat Corominas-Murtra1 (2015). "Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation," *Journal of Royal Society Interface* 12, 30 May.

Wuertz, D., T. Setz, Y. Chalabi, M. Maechler. (2017). "Rmetrics -Markets and Basic Statistics," *Rmetrics Association*, Nov.17.

Zipf, George Kingsley, (1949). *Human Behavior and the Principle of Least Effort*, Addison-Wesley.

Zolotarev, V. M. (1986). *One-dimensional Stable Distributions*, American Mathematical Society.

(2019 年 7 月 27 日受理)