〈一般研究課題〉　ベキ乗分布に基づく環境情報の統計学的
分析とその応用に関する研究

研 究 者 名　　愛知県立大学　　シイ シイ

ベキ乗分布に基づく環境情報の統計学的分析とその応用に関する研究

シイ シイ
（愛知県立大学）

A statistical analysis of environmental information based on fractional power distributions and its applications

Si Si
(Aichi Prefectural University)

1. はじめに

統計資料を整理したとき、ガウス分布ではなくて、いわゆるベキ乗分布と呼ばれる分布のグラフが現れることが度々あって、しかもそれが大事な意味を持つことが予想されるようになってきた。地震の頻度とか、ネットワークのノードにおける連結の数、あるいは、会話や文章に登場する単語の頻度、また一定地域での交通事故の数など、ベキ乗分布になる場合が多い。

ベキ乗分布とは、大きな数値が観測される頻度がゆっくり0に近づくような分布である。ガウス分布の場合は x が大きくなるとき急激に 0に近づく。よく例えられるベル型で、ずっと大きな値は殆ど無視できる。　これとは違って x が大きいとき密度関数は

$$\frac{c}{x^{\alpha+1}}$$

（c は正定数）で近似される場合である。この α はベキ乗分布の指数と呼ばれる。指数は、必ず0と2の間にある。

ベキ乗分布が得られた場合、それを統計的にどのように理解し、またそれに対応すればよいのかを考えてみたい。目標は

1-1. 統計的分析

ベキ乗分布の統計的な特性を知り、それが実際得られたデータのどの性質に対応するかを知りたい。例えば

i) 分布は自己相似である。これはデータのスケールに依存しない性質である。データが日本円で記録されようと、米ドルで記録されようと、関係ない。同じタイプの分布になり指数は変らない。

ii) 指数によって分類され、それによってタイプに分類できる。K を大きな値として、測定値 x が Kより大きくなる確率が比較できる。それにより指数の持つ意味を考えることができる。例えば、 α=1を境にした変化も考えられる。

1-2. 環境･情報源の理論的特性と応用

ベキ乗分布の発生する原因を知りたい。これには次のような方法が考えられる。

i) 類型を見る。同じタイプのベキ乗分布について、発生源のランダム性に共通点があるかを検討する。

ii) ベキ乗分布の吸引域を考える。一般に、加工していないデータは、グラフに書いても滑らかな曲線にはならない。何回もデータをとり、スムースなグラフにしたい。ただし、”ならし”は変数の1次変換によるもので、測定ごとの分布の算術平均をとることではない。

iii) 安定過程に埋め込む。ベキ乗分布の指数 α に対し、同じ指数の安定過程(時系列）に埋め込むことができる場合は、原因究明に大変有利である。加法性(時間的に独立増分）があるかどかをこれは別に詳しく述べる。

安定過程として扱うことの利点は、確率過程としての情報を得たことになる、いわば無限次元の情報が与えられる。ベキ乗とはいえ、一つの分布を知ることは1次元的な情報を得たに過ぎない。後に詳しく述べるように、これは Lévy 分解により、多くの素な過程であるポアソン過程の加重和としてあらわされ、その理論的な特性は実例であるランダム現象の解明に役立つ。

2. 安定分布

任意の n について、同じ分布 F に従う独立な確率変数 $X_1, X_2, ,\ldots, X_n$ の和 S_n が、ある一つの確率変数 X があって、$c_n X$ と同じ分布に従うとき, 分布 Fを安定分布 という。この定数 c_nは $c_n = n^{1/\alpha}$, と表され、$0 < \alpha \leq 2$, である。 α を安定分布の指数という。特別な値 $\alpha = 2$のときはガウス分布であるが、ここでは扱わない。

特に，指数$0 < \alpha < 2$, の安定分布で原点0について対称な場合を考えよう．そのとき分布(分布関数を $F(x)$ とする）の特性関数 $\varphi(t)$ (密度関数のフーリエ変換)は正の実数で、

log $\varphi(t)$も実数となり：

$$\varphi(t)=\int e^{itx}\ dF(x)$$

は次のように書ける。

$$\log \varphi(t) = \quad - \ c|t|^{\alpha} \quad , c \geq 0,$$

分布が対称だから平均 $m=0$ である. スケールを a 倍すれば $t \to at$ で定数 c が a 倍されるのみである. これは自己相似であることを意味しており, 分布のタイプを決めるのに役立つ.
特別な指数のとき, すなわち $\alpha = 1, 1/2, 3/2$ のときはその分布の様子が知られている。

I. $\alpha = 1$ の場合
密度関数は

$$\frac{1}{\pi} \frac{c}{c^2+(x-m)^2}, c > 0$$

で コーシー分布である。
II. $\alpha = 1/2$ の場合密度関数は　$x > 0$ のみで、

$$\frac{a}{\sqrt{2\pi}} x^{-3/2} e^{-\frac{a^2}{2x}}, a > 0$$

[註]　この分布はブラウン運動 $B(t)$ について
の逆関数の分布として現れる。

$$\max_{s \leq t} B(s)$$

III.　$\alpha = 3/2$ の場合
星の重力場に関する Holtsmark 分布としても知られている (W. Feller　vol.II, Chap.6)。

吸引域 (domain of attraction)

安定分布については、それらの 吸引域(domain of attraction) が知られている。

分布関数 $F(x)$ が指数 α ($0 < \alpha < 2$) の安定分布の吸引域に属するための必要十分条件は
任意の $k > 0$ に対して

$$\frac{1-F(x)+F(-x)}{1-F(kx)+F(-kx)} \to k^{\alpha}$$

となることである.
これから資料によって上の式を確かめればよい。そのとき、指数 α もわかる。

複合ポアソン過程

ポアソン過程 $P(t)$, $t \geq 0$ は非負整数値をとる単調非減少な加法過程で、定常増分をもつ。　その分布は, $t > s > 0$ で, k を非負整数 として

$$\Pr(P(t) - P(s) = k) = \frac{(\lambda(t-s))^k e^{\lambda(t-s)}}{k!}$$

できまる。ここで λ は強度(intensity) と呼ばれる正定数である。

複合ポアソン過程は、いろいろな高さのジャンプを持つ、独立なポアソン過程を組あわせたものである。組みあわせ方を規定して、たとえば自己相似になるようにすれば安定過程がえられる。ポアソン(分布）過程の特性と自己相似性を組み合わせて、安定過程の intrinsic な性質を探ることができる。

加法過程のLévy分解

安定過程を含むより一般な Lévy 過程、すなわち確率連続で，時間的に一様な独立増分をもつ加法過程 $L(t),\ t\geq 0$ は次のように分解される。

$$L(t)=m(t)+\sum B(t)+\int_{R-0} uP_{du}(t)-\frac{tu}{(1+u^2)}dn(u)\ 。$$

ここで，$B(t)$はブラウン運動，P_{du}はポアソン過程による確率測度$dn(u)$はLévy 測度で

$$\int \frac{u^2}{1+u^2}dn\ (u)\langle\infty$$

である。文献 [1] 参照。

直感的にいえば，$L(t)$は定数を除き，ブラウン運動(これは見本関数が連続)と種々のジャンプを持つポアソン過程(収束化定数を引いて)の和として表される。後者の見本関数は第一種の不連続性を持つ。

指数 α の対称な安定過程の場合は

$$dn(u)=c|u|^{-(1+\alpha)},$$

である。ただしcは定数である。

3. ポアソン・ノイズによるベキ乗分布の数理

記号$P_u(t)$ はジャンプの高さがu のポアソン過程を表す。それは $uP(t)$ と同等である。時間の dilation : $t\rightarrow\ e^at$を考えると、$P_u(e^at)$ は $e^aP_u(t)$ と同等になり、強度の dilation となる。

ジャンプ u のポアソン過程$P_u(t)$ を時間微分したポアソン・ノイズ$P_u(t)$にうつる。その見本関数は次の形をしている。

$$\Sigma_s\ u\ \delta_s(t),\quad 0\leq t\ \langle\infty$$

その特性汎関数$C_u(\xi)$は$P_u(t)$をテスト関数でsmear して平均値を$E(\exp\{i\int \dot{P}_u(t)\xi(t)\}dt)$計算するが、それは

$$C_u(\xi)=exp\ \psi(\xi)$$

と書けて

$$\psi(\xi)=\lambda\int_0^\infty(e^{i\xi(t)u}-1)dt$$

である。 これをψ関数というが、この関数がポアソン・ノイズの分布を決めているのである。

ジャンプ u をパラメータとするポアソン・ノイズをくみあわせる； 詳しくは u についてウエイト $f(u)$ をつけて積分する：

$$Y(t) = \int P_u(t) f(u) du.$$

これもテスト関数 ξ(t)により smear して、ψ関数を求めると

$$\lambda \iint (e^{i\xi(t)u} - 1) f(u) du dt$$

となる。

ここで$Y(t)$が自己相似 (self-similar) になるための $f(u)$ に対する条件を求めてみると、乗法についての定数を除き

$$f(u) = u^{-1-\alpha}$$

がえられる。

ここまでは形式的な計算によるが、$f(u)$ がこのように決まったとき、ψ-関数を表わす積分の収束を確かめなければならない。

i) $0 < \alpha < 1$のときは積分の収束に問題はない。

ii) $1 \leq \alpha < 2$のとき、適当な補正が必要である。実際

$$e^{iu\xi(t)} - 1 \;\rightarrow\; e^{iu\xi(t)} - 1 - iu\xi(t)$$

とすればよい。補正した項 $iu\xi(t)$ は定数を減ずることに対応し、扱いは全く容易である。i) のときに補正しても無害である。

以上まとめて, 定数を除き

$$\psi(\xi) = \lambda \iint e^{iu\xi(t) - 1 - iu\xi(t)} |u|^{-1-\alpha} du dt$$

すなわち安定分布に到達する。

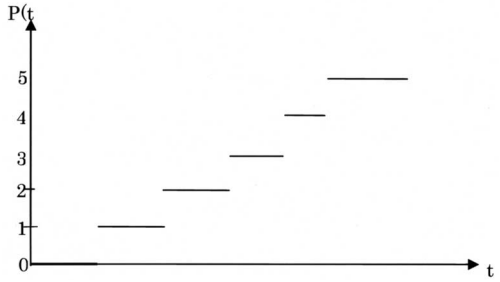
4. ベキ乗分布の特徴づけ

対象となるデータが安定過程のある時点での統計であることがわかった場合は、安定過程およびその構成要素（成分）であるポアソン過程の理論的な特性をみて、問題の現象と照らし合わせることができる。それによって、もとのランダム現象が解明されるであろう。

理論面から主張できることを列挙してみよう。詳細は Si Si [5] 参照。

a) $P_u(t)$ はすべての u について同型だから $P_1(t) = P(t)$ を扱う。

見本関数のグラフは下図のようになる。



各ステップでの滞在時間は指数分布にしたがう。$P(t)=k$ である時間区間を X_k とすれば

$$P(X_k > x) = e^{-\lambda x},\ x \geq o$$

で、λ はポアソン過程の強度と一致する。ついでながら、この分布の平均値は λ^{-1} である。
$X_k, k \geqq 0$, は i.i.d. である。双対性がなりたつ：

$$\sum_{0}^{n} X_k < t \ \leftrightarrow\ P(t) > n.$$

b) 指数分布の性質として lack of memory がある。各 X_k を代表して X とかき 、それを待ち時間として理解したとき, 条件つき確率について $x, h > 0$ のとき

$$P(X > x + h / X > x) = P(X > h).$$

すなわち、X だけ待った効果はない。これを、$P(t)$ が値 k をとったとき、状態 k が、あとどれだけの時間持続するかは、それ以前の持続時間に無関係である。

c) 指数分布について　lack of memory とは相対的に未来を無視してよいともいうべき性質がある。いま、$P(t)=k$ を知り、その後どれだけ状態 k が続いたかを知ったとする。t 以前の状態は知らないとしよう。わかっていることは時刻 t までに k 回ジャンプしたことである。
「ジャンプした時点は、$[0,t]$ 上を独立に、しかも一様に分布する k 個の時点を大きさの順に並べたもの」ということができる。t から先の未来のことには無関係である。

d) データから推定される安定過程は $P_u(t)$ の組み合わせである。実際には近似的に有限個のを選ぶことになろう。それらは独立である。得られた資料に a), b) c) の性質を独立にあてはめて、推定、特徴づけなどを行えばよいことになる。

5. 検証

ここでは，実際に文章から得たデータがベキ乗分布といえるのかを検証する。

[手順1.] まずサンプルとなる文章100ページ間から任意の単語の数をプログラミングにかけることで数える。これを接続詞・前置詞・副詞・動詞の四種類の品詞に分けて行い，グラフする．(本実験では接続詞30個，前置詞49個，副詞118個，動詞378個をサンプルとして検索した。)

[手順2.] 検索した単語のそれぞれの個数の和をとり，その値で各単語の個数の商をとり，縦軸をその値，横軸を検索した単語としてグラフにする。

[手順3.] 手順2で得たグラフの縦軸と横軸の値のlogをとり，それをグラフにする．(ただし横軸のlogをとる際は文章中の単語の存在数が多い順に1,2,3...と置き換えて考える。)

[手順4.] 手順3で得たデータによりベキ乗分布の式を求める。

以下に式を実際に求める.まず，手順1、2、3で得たデータ図1，図2として載せておく。

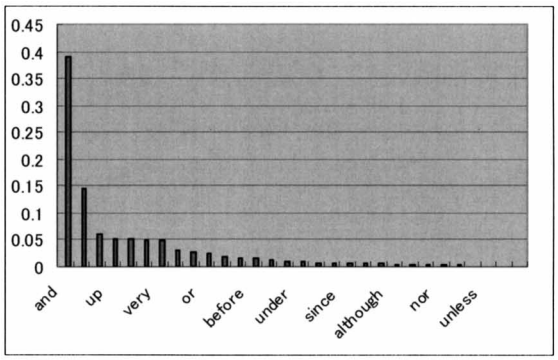


図1　接続詞に従うベキ乗分布

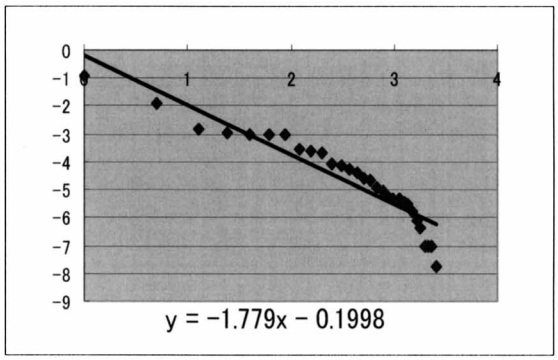


図2：図1の両対数グラフ化

図2のグラフは直線$y = -1.779\,x^{-0.1998}$にほぼ一致することが分かる。この式を用いて図1のグラフの式は

$$y = e^{-0.1998}\,x^{-1.779}$$

でありよって図1はベキ乗分布を表わすことが分かる。

以下前置詞・副詞・動詞の集まりについても同様の検証を行った際，全てベキ乗分布であることが確認された．以下に図とその方程式を載せておく．

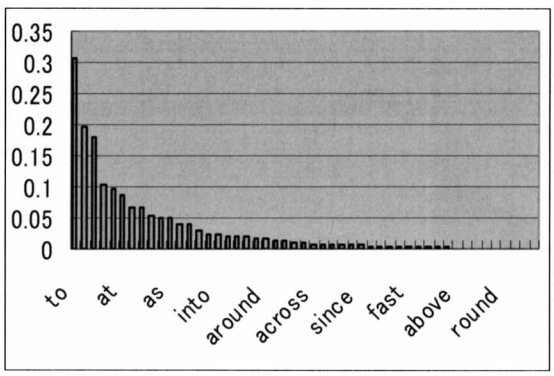


図3　前置詞に従うベキ乗分布

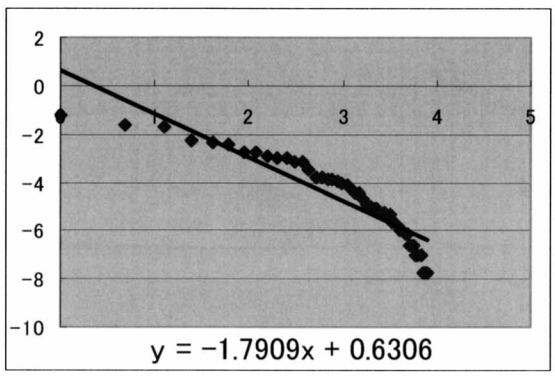


図4：図3の対数グラフ化

図3の方程式　：　$y = C_2\,x^{-1.7909}$

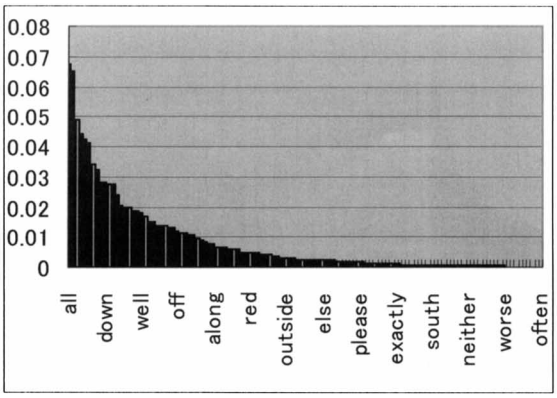


図5　副詞に従うベキ乗分布

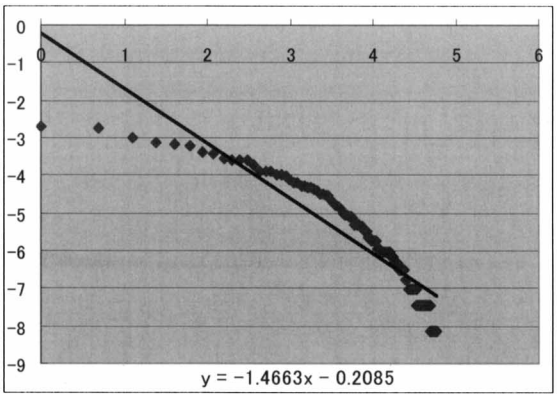


図6：図5の対数グラフ化

図5の方程式　：　$y = C_3\,x^{-1.4663}$

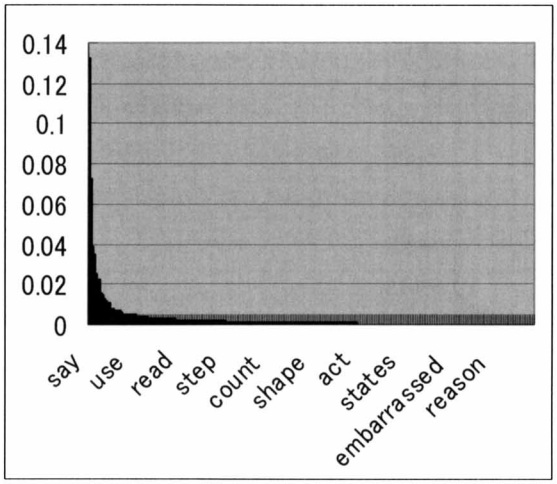


図7　動詞に従うベキ乗分布

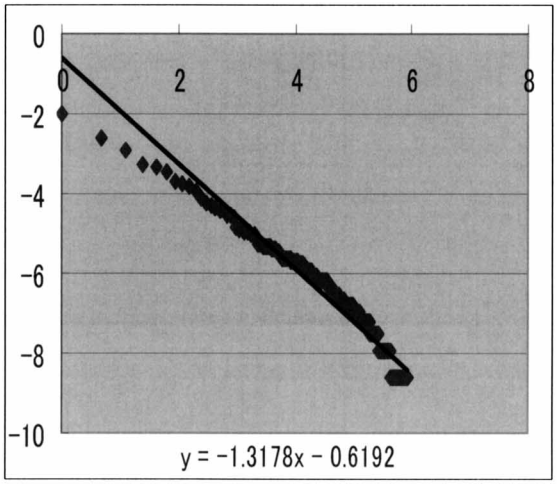


図8：図7の対数グラフ化

図7の方程式： $y = C_4 x^{-1.3178}$

参考文献［5］のHemingwayの作品を使用したが，次に参考文献［6］Sheldonの作品で同様の実験を行い、接続詞、前置詞、副詞。動詞の分布を調べ、結果は下記の通りです。

接続詞に従うベキ乗分布の確率密度関数　：　$f(x)=C'_1 x^{-1.721}$

前置詞に従うベキ乗分布の確率密度関数　：　$f(x)= C'_2 x^{-1.9024}$

副詞に従うベキ乗分布の密度関数　：　$f(x)=C'_3 x^{-1.2906}$

動詞に従うベキ乗分布の密度関数方程式　：　$f(x)=C'_4 x^{-1.1555}$

Sheldonの作品についても，品詞や文章の内容にかかわらずベキ乗分布となることが分かった．

6. 発展

今回のような文章に登場する単語の頻度については，以前からベキ乗分布になる場合が多いと言われている．しかし，文章を構成する単語一つ一つの分布に目を向けた場合，それらはどのような分布になっているのであろうか。

ここでは接続詞，前置詞，副詞，動詞の実験を行った中から特にand，to，all，sayを例として挙げる．図9，図11，図13，図15は横軸を単語の出現行とし縦軸を各行数までに出現した単語の頻度の和とした．また，図10，図12，図14，図16は横軸を単語の出現間隔とし，縦軸を出現間隔の等しいものの数とした。

なお動詞sayに関してはプログラムの関係上saidについてのグラフとする。

まずは接続詞andの確率過程と出現行による間隔の分布について，これがどのような分布であるか考えてみる。

図9と図10よりandはポアソン過程に従うことは明らかである。単位行ごとに生起する事象の数がパラメータλを決める。

$$\lambda = 1.151$$

である．

以下と同様に次項に載せた図11-図16についても全てポアソン過程と指数分布である．つまり，ベキ乗分布の裏側にはポアソン過程と指数分布が潜んでいる}ことが明らかとなった．

以下の図に関してはすべて拡大図のみとする．

「to」,「all」,「say」に関するパラメターλの値はそれぞれ1.121, 1.031, 1.101 である。

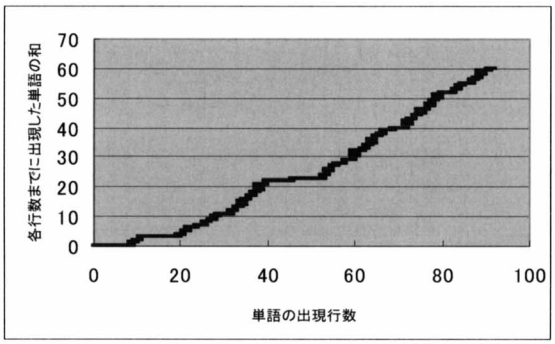


図9　andに従う確率過程

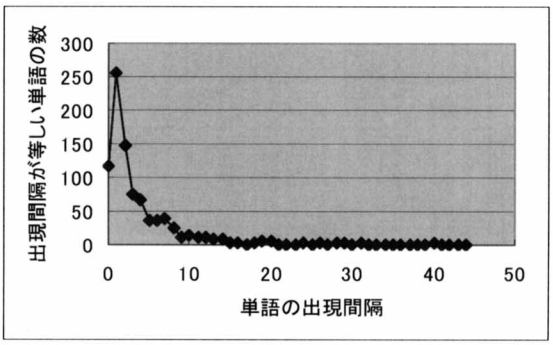


図10　andに従う確率分布

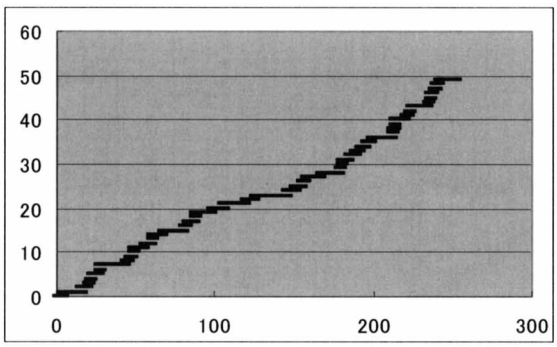


図11　toに従う確率過程

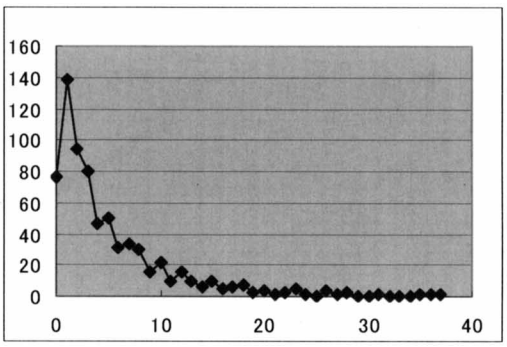


図12　toに従う確率分布

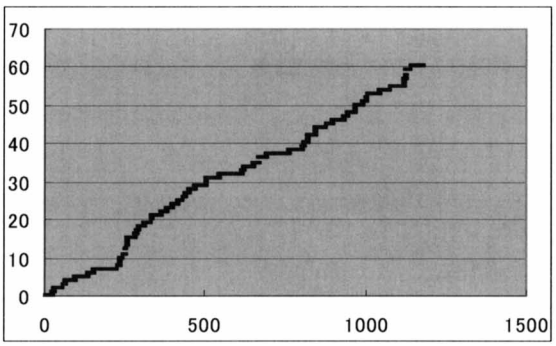


図13　allに従う確率過程

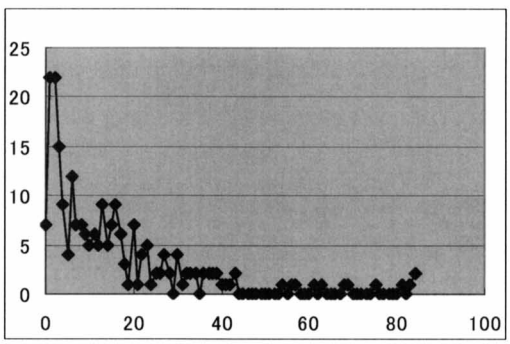


図14　allに従う確率分布

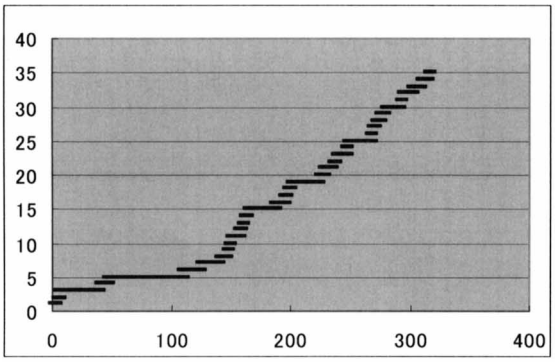


図15　saidに従う確率過程

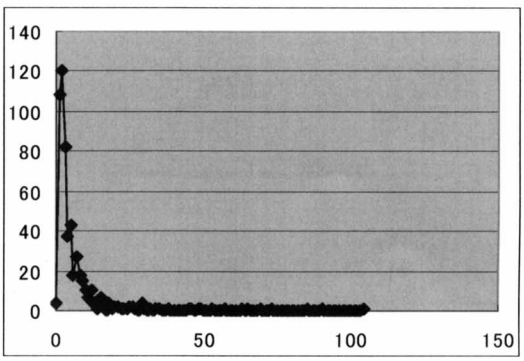


図16　saidに従う確率分布

7. 後書き

ベキ乗分布を安定過程に埋め込み、 Lévy 分解を適用して、ポアソン過程に分解する方法を文学的な作品に表れる単語の頻度の分布に適用し理論とよく一致することを確かめた。今回は英文について2人の著者について考察したが、和文の場合も同様な結果が得られることが予想される。今後の検討課題としたい。さらに進んだ研究として単語間のリンク数もまたベキ乗分布になることが予想される。これも確かめて総合的に理論化して、より一般の情報ネットワークに応用できるまで県研究を進めたい。

参考文献

[1] T. Hida, Stationary stochastic processes. Princeton Univ. Press. 1970.

[2] T. Hida and Si Si. Innovation approach to random fields. An application of white noise theory. World Sci. Pub. Co. 2004.

[3] T. Hida and Si Si, Lectures on white noise functionals. World Sci. Pub. Co. 2008.

[4] S. Kummon, Introduction to information sociology. NTT Pub. 2004.

[5] Si Si, Effective determination of Poisson noise. IDAQP Vol.6 (2003), 600-617.

[6] Si Si, An aspect of quadratic Hida distributions in the realization of a duality between Gaussian and Poisson noises, Infinite dimensional Analysis, Quantum Probability and Related Topics, Vol 11(2008).