ベキ乗則を導く確率モデルと映画の統計データへの適用

山本 健 （中央大学理工学部）

1 はじめに

自然科学や社会科学において，様々な量がベキ分布にしたがうことが知られている [1]．ベキ分布とは，確率密度関数（の裾）がベキ乗関数で表わされるような統計分布である．ベキ分布は裾での減衰が指数関数よりも緩慢であることから，正規分布や指数分布に比べてサイズの大きな事象が生じやすいという特徴がある．フラクタル [2] や臨界現象 [3] がベキ分布の研究の端緒である．フラクタル図形を直径 ε の円板で被覆するのに必要な最小個数 $N(\varepsilon)$ は，ε が小さいところでベキ乗の依存性 $N(\varepsilon) \propto \varepsilon^{-D}$ をもつ．ここに現れるベキ指数 D はフラクタル図形のフラクタル次元とよばれる．また，パーコレーション問題において占有率 p を大きくしていき臨界確率 p_c に近づけると，クラスターの平均サイズ S は $S \propto (p_c - p)^{-\gamma}$ というふるまいを示す（2 次元のパーコレーションでは $\gamma = 43/18$ である）．ベキ指数 γ は臨界指数とよばれるものの 1 つである．これらの例では，ベキ指数の値が図形や現象の特徴を定める．より複雑な現象も，ベキ分布を用いて分析が試みられている．たとえば，複雑ネットワークや経済物理学などの研究においてベキ分布は重要な役割をもつ．

ベキ分布の典型例が，1 つの店舗で販売される各々の商品の売上の分布である．売上額が s よりも大きな商品の割合が $P_>(s) \propto s^{-\beta}$ で表される．このようにサイズが s よりも大きな要素の割合を s の関数とみなした $P_>(s)$ を，本稿では**累積分布**とよぶ．累積分布 $P_>(s)$ は確率密度関数を s より大きい範囲で積分したものである．ベキ指数 β は正であり，たとえばアマゾン ($Amazon.com$) の売上データでは $\beta = 1.1$ から 1.2 程度である [4]．売上のベキ分布は，よく売れる少数の人気商品とほとんど売れない大多数のマイナーな商品の売上の間に大きな格差があることを意味する．（一方で，大多数のマイナーな商品の売上を足し上げると，店舗の総売上のうち無視できない割合を占めるという “ロングテール現象”[5] もベキ分布の帰結である．）

ベキ分布の中でも，累積分布のベキ指数が $\beta = 1$ である場合を特に**ジップの法則**とよぶ．本来，ジップの法則とは，多数の要素（データ）の集合において要素のサイズが順位に反比例するという経験則であり [6], $\beta = 1$ はこれと等価な表現である．元々は文章中の単語の出現回数に関する統計的な関係として見出されたが，言語の統計以外にも地震のエネルギー ($\beta = 0.95$) [7] や科学論文の被引用数 ($\beta = 1.09$) [8] などがほぼジップの法則にしたがう．上述したアマゾンの売上データもほぼジップの法則にしたがうといえる．

本研究では商品が売れていく様子を単純化した確率過程を提案し，定常なベキ分布が導かれることを紹介する．アメリカの映画の興行収入のデータを例にとり，提案したモデルが実際の分布をよく説明することを示す．

2 モデル

本研究では，ある 1 種類の商品の売上の時間変化に対するモデルとして，次の確率過程を考える：

$$x_{t+1} = \mu_t x_t, \tag{1a}$$

$$S_{t+1} = S_t + x_t. \tag{1b}$$

$t = 1, 2, \ldots$ は離散的な時刻を表す．初期条件は $x_1 = 1$, $S_1 = 0$ とする．確率変数 x_t は t 期の売上を表す．式 (1a) は，ある期間の売上が大きいほど次の期間でも売上が大きくなりやすいという正のフィードバックの効果を単純化したものである．さらに議論の単純化のため成長率 μ_t は各時刻 t で独立かつ

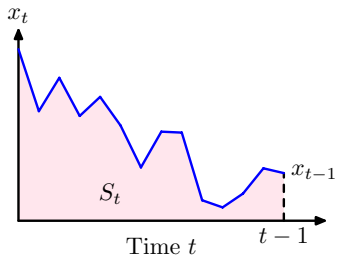

図 1 モデル (1) の模式図．x_t は式 (1a) にしたがって時間変化する．S_t は x_t のグラフと t 軸の間の面積に相当する．

同一の分布にしたがうものとする．一方，確率変数 S_t は各期間の売上 x_t を時刻 $t-1$ まで累計した量（$S_t = x_1 + \cdots + x_{t-1}$）である．充分に大きな時刻 t において，S_t は商品の総売上を表す．モデル (1) を模式的に表したのが図 1 である．各期間の売上 x_t の動きを t の関数としてグラフ化すると，総売上 S_t は x_t のグラフより下側の面積に相当する．

式 (1) の解析 [9] によると，t が大きいとき S_t は定常な分布をもち，サイズが大きいところでベキ的な裾 $P(S_t > s) \propto s^{-\beta}$ をもつ．指数 $\beta(>0)$ は条件

$$E(\mu_t^{\beta}) = 1 \tag{2}$$

により特徴づけられる．$E(\cdot)$ は期待値を表す．特に，成長率の期待値が 1 ($E(\mu_t) = 1$) である場合には $\beta = 1$ となり，ジップの法則が導かれる．

3　映画の興行収入との比較

シンプルな形の式 (1) が商品の売上を記述するモデルとして現実的なのかを，実際のデータとの比較によって検討する．モデルは実際の売上プロセスを非常に単純化しているので，売上データの詳細まで再現することは望めない．一方，売上がベキ分布になることやベキ指数の値は現象の詳細にはほとんど依存しないと考えられる（統計物理学の普遍性クラスという概念と関係する）．式 (2) から評価した指数 β が実際の値と近くなることが確認できれば，モデル (1) は実際の現象の重要なポイントを押さえていると期待される．

実際の売上データとして，アメリカの映画の興行収入を取り上げる．式 (2) を用いて指数を評価するには成長率 μ_t に関する情報，つまり期間ごとの売上 x_t の情報が必要であるのだが，映画の興行収入については必要なデータがオンラインで無料公開されている．今回利用したデータベースは *The Numbers* および *Box Office Mojo* である [10].

直感的には，モデルの S_t は 1 つの映画作品の総収入額に対応すると考えられるのだが，この見方には少し問題がある．一般に，異なる映画の興行収入は統計的に同等とはいえない．とりわけ，映画の経済的な規模が興行収入に関係し，費用のかかった映画ほど大きな興行収入を得やすい [11]．一方，モデル (1) は **1 つの商品**の売上が成長率 μ_t のゆらぎによって確率的な値をとることを表す．したがって，複数の映画の興行収入額からつくった分布はモデルにおける S_t の定常分布に対応しない．そこで，興行収入を制作費で割って正規化した**投資利益率** (return on investment, ROI) とよばれる量に注目する．ROI は映画の成功・失敗を測る指標である．ROI が 1 より大きければその映画は黒字（制作費を上回る収入が得

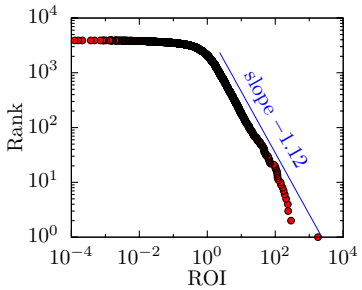

図 2 データベース *The Numbers* による ROI の累積分布．分布の裾はベキ指数 $\beta = 1.12$ のベキ分布にしたがう．

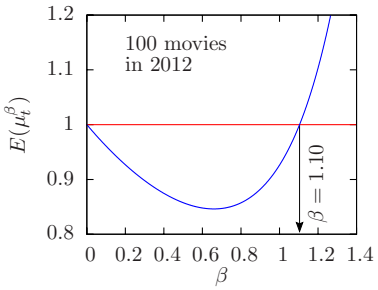

図 3 式 (2) を用いた指数 β の評価．$E(\mu_t^\beta)$ を β の関数として表示したグラフ．$E(\mu_t^\beta) = 1$ の解として $\beta = 1.10$ が得られる．

られた）であり，1 より小さければ赤字である．ROI では映画ごとの規模の違いの影響が小さくなっていると考えられる．

映画の ROI の分布とモデルの S_t の定常分布を比較する．図 2 はデータベース *The Numbers* のデータ（1915 年から 2014 年 4 月までの 3906 タイトル）による ROI の累積分布である．ROI が 1 より大きい裾の部分でベキ分布的な減衰がみられ，そのベキ指数は $\beta = 1.12 \pm 0.03$ と求められる．つまり，黒字だった映画の ROI はおおむねジップの法則にしたがう．ちなみに，興行収入額自体の分布では，ジップの法則はおろか明瞭なベキ分布すらみえない．(興行収入よりも ROI の方が統計的に単純な指標であるといえるかもしれない.)

次に，映画の興行収入の推移がモデル (1) にしたがうと仮定し，式 (2) を用いて指数 β を見積もってみる．日ごとの興行収入のデータでは曜日の周期性（週末に客数が伸びる）が強く現れるので，週ごとのデータを用いた．成長率 μ_t は，映画が公開されて t 週目と $t+1$ 週目の興行収入の比 $(\mu_t = x_{t+1}/x_t)$ として求める．データベース *Box Office Mojo* において，2012 年に公開された映画の中でアメリカ国内の興行収入額の上位 100 作品について各週の μ_t を算出した．こうして得られた成長率を μ_t の標本値とし，β の関数として $E(\mu_t^\beta)$ を表したのが図 3 である．$E(\mu_t^\beta) = 1$ となるのは $\beta = 1.10$ であり，S_t の定常分布の裾は $P(S_t > s) \propto \beta^{-1.10}$ という形であると見積もられる．この結果は図 2 の指数 $\beta = 1.12$ ととても近い．よって，モデル (1) は映画が興行収入をあげる過程の特徴をよくとらえていると結論できる．同様の計算により，2012 年の興行収入額上位 200 作品のデータからは $\beta = 1.09$, 2010 年の上位 100 作品のデータからは $\beta = 1.08$ が得られた．したがって，指数 β の見積もりは順位や年度にはほとんど依存しないと考えられる．

4　まとめと展望

本研究では，商品の売上を模擬する確率モデル (1) を提案した．確率変数 S_t は定常な分布をもち，その裾は式 (2) で特徴づけられる β を指数とするベキ分布にしたがう．次に，この理論的な結果を映画の ROI のデータと比較した．実際の ROI の累積分布は裾が $\beta = 1.12$ 程度のベキ分布であったのに対して，モデル (1) を仮定して評価したベキ指数は $\beta = 1.10$ であった．式 (2) から見積もった指数が実際の値に近かったことから，提案したモデル (1) が映画の集客の特徴をうまくとらえていると期待できる．

式 (1) は商品の売上のベキ分布を導く確率過程として，簡素な形をしている．本モデル化では，口コミ・広告宣伝・消費者の嗜好・流行などの効果は個別には考慮されていない．これらはすべて確率変数 μ_t の中に押し込められている．このように単純化したモデルであっても，ベキ指数のような普遍性をもつ特徴はよく再現されている．

なお，制作費のデータは興行収入額に比べて不正確な推定値である（正確な制作費は公開されていない）．したがって ROI にも不正確さが含まれることに注意する．映画以外の売上データでも本稿と同様の分析をおこない，モデル (1) の妥当性を検証することが必要である．

謝辞　本研究は科研費（若手研究 (B), 課題番号 25870743）の補助を受けたものである．

参考文献

[1] M. Buchanan, Ubiquity: Why Catastrophes Happen, Three Rivers Press, 2001; マーク・ブキャナン（水谷淳訳），歴史は「べき乗則」で動く，早川書房，2009.

[2] 松下貢，フラクタルの物理 (I), (II), 裳華房, 2002, 2004.

[3] H. E. Stanley, Introduction to Phase Transitions and Critical Phenomena, Oxford University Press, 1987; ユージン・スタンリー（松野孝一郎訳），相転移と臨界現象，東京図書，1987.

[4] T. Fenner, M. Levene, and G. Loizou, Physica A 389, 2416 (2010).

[5] C. Anderson, The Long Tail, Hyperion, 2008; クリス・アンダーソン（篠森ゆりこ訳），ロングテール，早川書房，2009.

[6] G. K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949.

[7] P. Bak, K. Christensen, L. Danon, and T. Scalon, Phys. Rev. Lett. 88, 178501 (2002).

[8] A. M. Petersen, H. E. Stanley, and S. Succi, Sci. Rep. 1, 181 (2011).

[9] K. Yamamoto, Phys. Rev. E 89, 042115 (2014).

[10] Box Office Mojo. `http://www.boxofficemojo.com/`
The Numbers. `http://www.the-numbers.com/`

[11] R. J. Pan and S. Sinha, New J. Phys. 12, 115004 (2010).