～数理モデルに関する科研費研究会（福井高専)～

べき指数の推定の仕方

梅野　善雄

元一関工業高等専門学校

2024 年 3 月 29 日

目　次

べき分布の定義

- べき分布の確率密度関数

$$f(x) = \frac{ab^a}{x^{a+1}} = \frac{\frac{a}{b}}{\left(\frac{x}{b}\right)^{a+1}} \quad (a > 0, x \geqq b)$$

  パレート分布と呼ばれ，Pareto(a, b) などで表される．
- 両対数グラフでは，右下がりの直線になる．

$$\log f(x) = \log C - (a+1) \log x \quad (C = ab^a)$$

- 累積分布関数は

$$P(X \leqq x) = \int_b^x \frac{ab^a}{t^{a+1}}\, dt = 1 - \frac{b^a}{x^a}$$

- 相補累積分布関数はべき関数になる．

$$P(X > x) = \frac{b^a}{x^a}$$

べき分布の平均と分散

- べき分布の平均

$$E(X) = \int_b^\infty x \cdot \frac{ab^a}{x^{a+1}}\,dx = \int_b^\infty \frac{ab^a}{x^a}\,dx$$
$$= \begin{cases} \infty & (0 < a \leqq 1) \\ \dfrac{ab}{a-1} & (a > 1) \end{cases}$$

- べき分布の分散

$$V(X) = \begin{cases} \infty & (0 < a \leqq 2) \\ \dfrac{ab^2}{(a-1)^2(a-2)} & (a > 2) \end{cases}$$

- べき分布の n 次モーメント

$$E(X^n) = \begin{cases} \infty & (0 < a \leqq n) \\ \dfrac{ab^n}{a-n} & (a > n) \end{cases}$$

べき分布 Pareto(a, b) の性質

- 確率密度関数に自己相似性がある．これは，べき関数に限る．

$$f(cx) = \frac{ab^a}{(cx)^{a+1}} = \frac{1}{c^{a+1}} f(x) \propto f(x)$$

- cX $(c > 0)$ は Pareto(a, bc) にしたがう．

$$P(cX \leqq x) = P\left(X \leqq \frac{x}{c}\right)$$

$$= 1 - \frac{b^a}{(x/c)^a} = 1 - \frac{(bc)^a}{x^a}$$

- X^n $(n > 0)$ は Pareto$\left(\frac{a}{n}, b^n\right)$ に従う．

$$P(X^n \leqq x) = P(X \leqq x^{\frac{1}{n}})$$

$$= 1 - \frac{b^a}{\left(x^{\frac{1}{n}}\right)^a} = 1 - \frac{(b^n)^{\frac{a}{n}}}{x^{\frac{a}{n}}}$$

- べき分布は，定数倍してもべき乗してもべき分布に従う．

べき指数 a の違いによるグラフ ($b = 1$ とする)

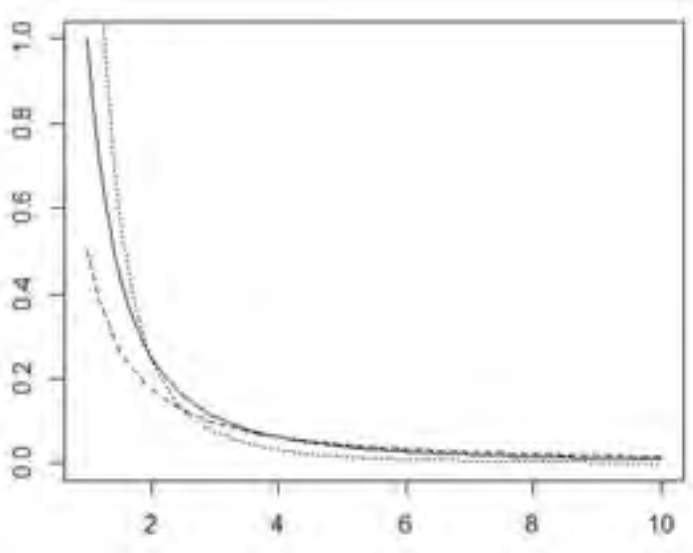
- 下図では，破線は $a = 0.5$，実線は $a = 1$，点線は $a = 2$ とする．

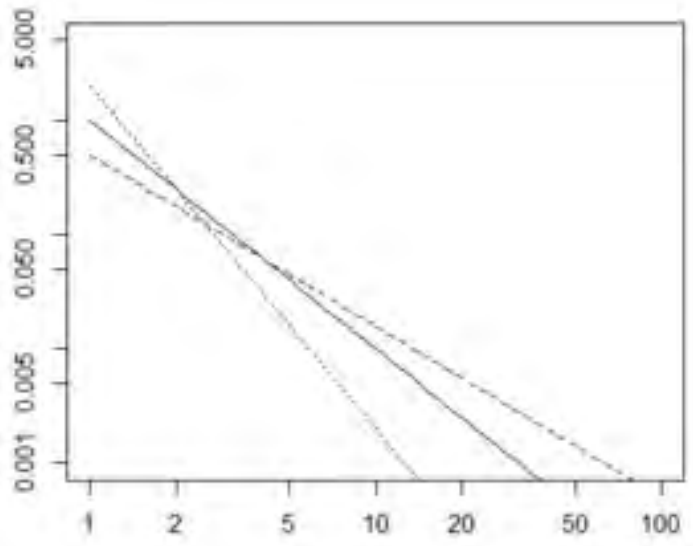
  - 確率密度曲線

    $$f(x) = \frac{a}{x^{a+1}}$$

  - 両対数グラフ

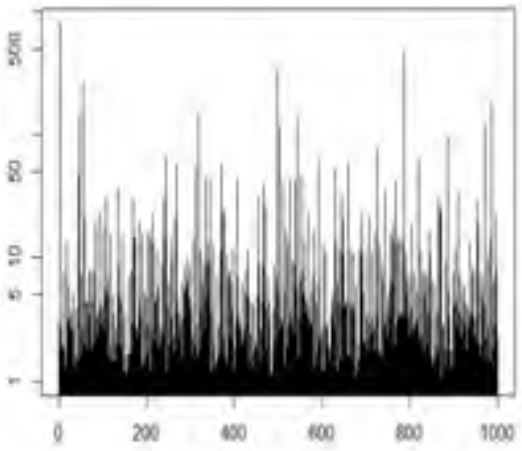
    $$\log f(x) = \log a - (a+1) \log x$$





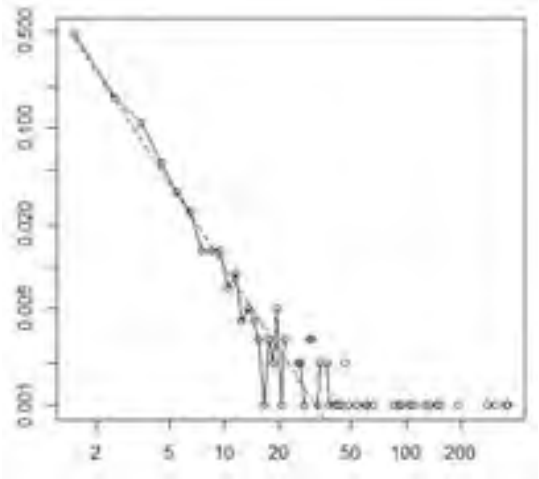
乱数データからのべき指数 α の推定

(1) Pareto(1,1) の乱数の生データ



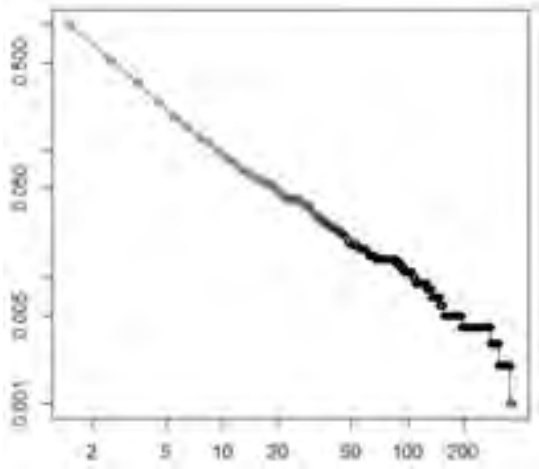
横軸は生成順.
縦軸は対数軸.
4 桁以上の場合もある

(2) ヒストグラム (折れ線) の両対数グラフ



大きい値は生成数が少ないので，折れ線表示では上下動が大きい.
回帰直線の傾きは
-1.07.

(3) 相補累積分布としてまとめ直したグラフ



x 以上の値としてまとめると乱れが少ない.
(3) の傾きを $-\alpha$ とすると，(2) の傾きは $-(\alpha+1)$ になる.

確率分布の母数を標本から推定する

- ある確率分布に従うデータ $(x_1, x_2, \ldots, x_n)$ が得られたとき，その確率分布の母平均と母分散を推定したいとする．
- 母平均 μ は標本平均 $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ により，母分散 σ^2 は不偏分散 $u^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ により推定できる．点推定である．
- 母数 θ が不明な確率分布の確率密度関数を $f(x:\theta)$ とする．
- 独立に n 個の標本 $x_1, x_2, \ldots, x_n$ を抽出すると，そのようなデータが得られる確率は次式と考えられる．

$$\begin{aligned} F(\theta) &= f(x_1:\theta) f(x_2:\theta) \cdots f(x_n:\theta) \\ &= \prod_{i=1}^{n} f(x_i:\theta) \end{aligned}$$

- $F(\theta)$ を尤度関数という．$F(\theta)$ が最大になるような θ を求める．

べき指数を最尤推定法で求める

- 積は扱いにくいので対数をとった式が最大になるような θ を求める．

$$L(\theta) = \log F(\theta) = \sum_{i=1}^{n} \log f(x_i : \theta)$$

- べき分布 Pareto$(a, 1)$ に従うと思われるデータが得られて，べき指数 a が不明なときは $f(x : a) = \dfrac{a}{x^{a+1}}$ であるので，

$$\begin{aligned} L(a) &= \sum_{i=1}^{n} (\log a - (a+1) \log x_i) \\ &= n \log a - (a+1) \sum_{i=1}^{n} \log x_i \end{aligned}$$

- a で微分すると，$\dfrac{dL}{da} = \dfrac{n}{a} - \sum_{i=1}^{n} \log x_i$
- $\dfrac{dL}{da} = 0$ となる a は，$a = n \left(\sum_{i=1}^{n} \log x_i \right)^{-1}$

べき指数の確率密度関数

- 前に利用した乱数 1000 個で計算すると，$a \approx 1.0035$ が得られる．
- $\beta = \sum_{i=1}^{n} \log x_i$ とおくと，$F(a) = \prod_{i=1}^{n} \frac{a}{x_i^{a+1}}$ より，

$$\begin{aligned} L(a) &= \log F(a) \\ &= \log a^n - (a+1)\beta \\ &= \log a^n + \log e^{-(a+1)\beta} \end{aligned}$$

- これより，$F(a)$ は次式で表わされる．

$$F(a) = a^n e^{-(a+1)\beta}$$

- データ $(x_1, x_2, \ldots, x_n)$ が与えられたとき，a を確率変数とみると，その確率密度関数は次式と考えられる．

$$\frac{F(a)}{\int_0^\infty F(a)\, da}$$

$\int_0^\infty F(a)\,da$ を求める

- $\int_0^\infty F(a)\,da = \int_0^\infty a^n e^{-(a+1)\beta}\,da$ は，べき関数と指数関数の積の積分で，ガンマ関数 $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,dt$ と同じタイプ．
- $a\beta = t$ とおくと，$da = \dfrac{1}{\beta}\,dt$ であり

$$\begin{aligned}\int_0^\infty a^n e^{-(a+1)\beta}\,da &= \int_0^\infty \left(\frac{t}{\beta}\right)^n e^{-t}e^{-\beta}\frac{1}{\beta}\,dt \\ &= \frac{e^{-\beta}}{\beta^{n+1}}\int_0^\infty t^{(n+1)-1}e^{-t}\,dt \\ &= e^{-\beta}\beta^{-(n+1)}\Gamma(n+1)\end{aligned}$$

- そこで，$F(a)$ を改めて次のように定める．

$$F(a) = \frac{a^n e^{-(a+1)\beta}}{e^{-\beta}\beta^{-(n+1)}\Gamma(n+1)}$$

べき指数 a の平均を求める

- a の平均 $E(a)$ を計算する．

$$E(a) = \int_0^\infty aF(a)\,da = \frac{\displaystyle\int_0^\infty a \cdot a^n e^{-(a+1)\beta}\,da}{e^{-\beta}\beta^{-(n+1)}\Gamma(n+1)}$$

- $\beta a = t$ と置換すると，$\Gamma(n+2) = (n+1)\Gamma(n+1)$ より

$$\begin{aligned}(top) &= \int_0^\infty \frac{t}{\beta}\left(\frac{t}{\beta}\right)^n e^{-t}e^{-\beta}\frac{1}{\beta}\,dt \\ &= e^{-\beta}\beta^{-(n+2)}\int_0^\infty t^{(n+2)-1}e^{-t}\,dt \\ &= e^{-\beta}\beta^{-(n+2)}\Gamma(n+2) \\ &= e^{-\beta}\beta^{-(n+2)}(n+1)\Gamma(n+1)\end{aligned}$$

- したがって，

$$E(a) = \frac{e^{-\beta}\beta^{-(n+2)}(n+1)\Gamma(n+1)}{e^{-\beta}\beta^{-(n+1)}\Gamma(n+1)} = \frac{n+1}{\beta}$$

べき指数の平均 $E(a)$

- 乱数 1000 個のデータでは，$\beta = 996.487$ となるので，

$$E(a) = \frac{1001}{996.487} \approx 1.0045$$

- データ数 n と平均 $E(a)$

n	10	50	100	200	500
β	9.44	49.7	109.5	207.5	469.0
$E(a)$	1.16	1.02	0.922	0.968	1.068

- $n = 50$, $\beta = 49.7$ のときの確率密度関数

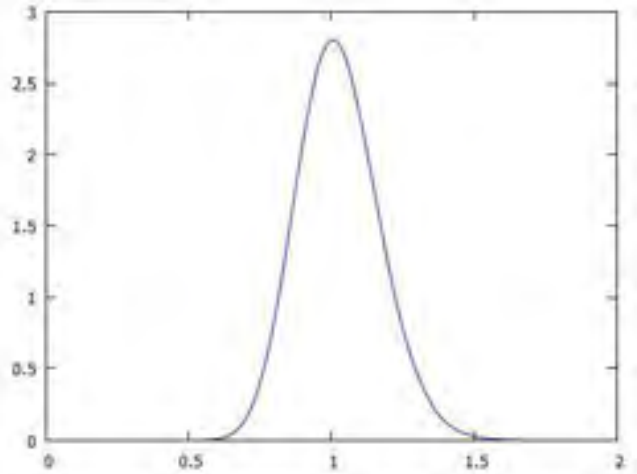
$$F(x) = \frac{x^{50} e^{-(x+1)\cdot 49.7}}{e^{-49.7} 49.7^{-51} \Gamma(51)}$$

$e^{-49.7} \approx 2.603 \times 10^{-22}$

$49.7^{-51} \approx 3.061 \times 10^{-87}$

$\Gamma(51) \approx 3.041 \times 10^{64}$

$e^{-2\cdot 49.7} \approx 6.778 \times 10^{-44}$



おわりに

- 関数 $F(x)$ は，データ $(x_1, x_2, \ldots, x_n)$ に依存する．
- $F(x)$ のグラフの頂点は，必ずしも $a \approx 1$ の箇所になるわけではない．データを何度か生成して $a \approx 1$ となるようなデータを利用しただけである．
- 最尤推定法の紹介や，ガンマ関数が現れる具体例として利用できるかもしれない．

【注１】本解説は，下記の「APPENDIX B」を参照しました．
M. E. J. Newman: Power laws, Pareto distributions and Zipf's law, 2006
APPENDIX B: Maximum likelihood estimate of exponents
[URL] https://arxiv.org/pdf/cond-mat/0412004.pdf

【注２】「べき分布」については、下記も参照してください．
[URL] https://yunavi.lsv.jp/powerlaw.html

【注３】「パーコレーション」は、下記の「＞理学部数学科＞複雑系」を参照下さい．
[URL] https://yunavi.lsv.jp/mathstudy.html