高専数学教育研究会（盛岡：アイーナ）

平均も分散も存在しない確率分布

梅野　善雄

元一関工業高等専門学校

2022 年 11 月 19 日

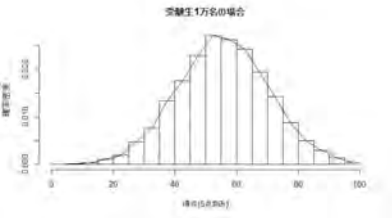
目　次

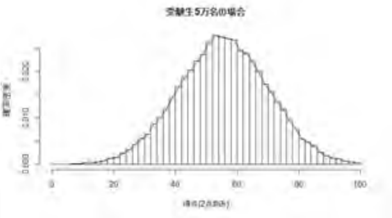
確率分布

- 統計データがどのように分布しているのかは，幾つかの階級に分けてヒストグラムで描画すると分かりやすい．

階級	$a_0 \sim a_1$	$a_1 \sim a_2$	$\cdots$	$a_{n-1} \sim a_n$	計
階級値	x_1	x_2	$\cdots$	x_n	
度数	N_1	N_2	$\cdots$	N_n	N
相対度数	p_1	p_2	$\cdots$	p_n	1

- 連続型変量で階級幅をいくらでも狭くできると仮定すると，縦軸を(相対度数)/(階級幅) にとったヒストグラムは確率密度曲線に近づいていく．





累積分布関数と相補累積分布関数

- 連続型確率変数 X の確率密度関数を $f(x)$ とする．累積分布関数は

$$F(x) = P(X \leqq x) = \int_{-\infty}^{x} f(x)\,dx, \qquad F'(x) = f(x)$$

- $[0, 1]$ での一様分布の累積分布関数は，

$$F(x) = \int_0^x 1\,dx = x \quad \text{よって} \quad P(U \leqq x) = x$$

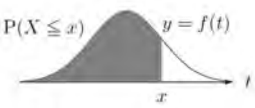
- ある確率分布の累積分布関数を $F(x)$ とし，その逆関数を $F^{-1}(x)$ とする．$[0, 1]$ での一様分布に従う確率変数を U とすると，

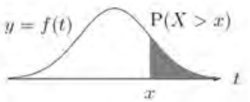
$$P(F^{-1}(U) \leqq x) = P(U \leqq F(x)) = F(x)$$

$F^{-1}(U)$ は，累積分布関数が $F(x)$ であるような確率分布に従う．

- 相補累積分布関数

$$1 - F(x) = P(X > x) = \int_{x}^{\infty} f(x)\,dx$$





確率分布の平均と分散

- 度数分布表で与えられたときの平均は，

$$
\begin{aligned}
E(X) &= \frac{1}{N}(x_1 N_1 + x_2 N_2 + \cdots + x_n N_n) \\
&= x_1 \frac{N_1}{N} + x_2 \frac{N_2}{N} + \cdots + \frac{N_n}{N} \\
&= x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_{i=1}^{n} x_i p_i
\end{aligned}
$$

- 確率密度関数を $f(x)$ とすると，細かく分割したときの相対度数に相当するのは $f(x)\,dx$ になるので，

$$
E(X) = \int_{-\infty}^{\infty} x\,f(x)\,dx
$$

- 分散は，平均 μ との差の 2 乗の平均なので，

$$
\begin{aligned}
V(X) &= E\left((X-\mu)^2\right) \\
&= E(X^2) - E(X)^2
\end{aligned}
$$

確率変数の標準化と，歪度・尖度

- 平均 μ，分散が σ^2 の確率分布に従う確率変数を X とすると，

$$Z = \frac{X-\mu}{\sigma}$$

とおくと，Z は平均が 0，分散が 1 に標準化される．

- 確率分布の左右対称性の指標として歪度があり，次の式で定義される．

$$(\text{歪度}) = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right)$$

- 正規分布と比較する指標として尖度があり，次の式で定義される．

$$(\text{尖度}) = E\left(\left(\frac{X-\mu}{\sigma}\right)^4\right) - 3$$

(注) 正規分布 $N(\mu, \sigma^2)$ の尖度は 3 になる．

確率分布のモーメント母関数

- 確率分布の平均・分散・歪度・尖度は，$E(X^n)$ $(n = 1, 2, 3, 4)$ が分かれば計算できる．n 次モーメントという．
- x^n $(n = 1, 2, \ldots)$ をすべて含む関数として e^x がある．

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

次のものを考える．モーメント母関数といい，$M_X(t)$ で表す．

$$\begin{aligned} E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots \end{aligned}$$

- モーメント母関数が分かれば，n 次モーメントは $M_X(t)$ の n 次導関数の $t = 0$ のときの値として求められる．

$$E(X^n) = M_X^{(n)}(0)$$

モーメント母関数と確率密度関数

- 幾つかの確率分布のモーメント母関数

  $[a, b]$ での一様分布では，$M_X(t) = \dfrac{e^{bt} - e^{at}}{(b-a)t}$

  正規分布 $N(\mu, \sigma^2)$ では，$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$

  自由度 n の χ^2 分布では，$M_X(t) = \dfrac{1}{(1-2t)^{\frac{n}{2}}} \quad \left(t < \frac{1}{2}\right)$

- モーメント母関数 $M_X(t)$ が分かれば，確率密度関数も分かる．

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-tx} M_X(t)\, dt \quad \text{(反転公式)}$$

- モーメント母関数が存在しない確率分布もある．
  t 分布やコーシー分布では，任意の t についてモーメント母関数は存在しない．

特性関数

- モーメント母関数は e^{tX} の平均であるが，虚数単位 i を付加した $E(e^{itX})$ を確率分布の特性関数といい，$\varphi_X(t)$ で表す．

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx$$

- モーメント母関数は確率分布によっては存在しない場合があるが，特性関数は任意の確率分布に対して必ず存在する．
- 特性関数が必ず存在するのは，次の式が成り立つことによる．

$$\left|\int_{-\infty}^{\infty} e^{itx} f_X(x)\,dx\right| \leqq \int_{-\infty}^{\infty} \left|e^{itx} f_X(x)\right| dx = \int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

- 特性関数が分かれば，確率密度関数は次式により定まる．

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t)\,dt$$

ただし，必ずしも初等関数で表せるとは限らない．

特性関数の性質

- 特性関数は，確率分布に対して一意的に定まり，特性関数が分かれば確率計算を次のように行うことができる．

$$P(a \leqq X \leqq b) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t)\, dt$$

- 確率変数 X の特性関数を $\varphi_X(t)$ とすると，$a + bX$ の特性関数は

$$\begin{aligned} \varphi_{a+bX}(t) &= E(e^{it(a+bX)}) \\ &= e^{iat} E(e^{i(bt)X}) \\ &= e^{iat} \varphi_X(bt) \end{aligned}$$

- 同一の確率分布に従う，互いに独立な確率変数を X, Y とするとき，$X + Y$ の特性関数は

$$\begin{aligned} \varphi_{X+Y}(t) &= E(e^{it(X+Y)}) \\ &= E(e^{itX}) E(e^{itY}) \\ &= \varphi_X(t) \varphi_Y(t) \end{aligned}$$

正規分布 $N(\mu, \sigma^2)$

- Z が $N(0,1)$ に従うとき，$\mu + \sigma Z$ は $N(\mu, \sigma^2)$ に従う．
- X が $N(\mu, \sigma^2)$ に従うとき，定数 a, b に対して
  $a + bX = (a + b\mu) + (b\sigma)Z$ は $N(a + b\mu, b^2\sigma^2)$ に従う．
- $N(\mu, \sigma^2)$ に従う互いに独立な n 個の確率変数の平均は，
  $N\left(\mu, \dfrac{\sigma^2}{n}\right)$ に従うので，$\mu + \dfrac{\sigma}{\sqrt{n}}Z$ と $\overline{X}$ は同じ確率分布に従う．

$$\frac{1}{n}(X_1 + X_2 + \cdots + X_n) \overset{d}{=} \mu + \frac{\sigma}{\sqrt{n}}Z$$

$$\therefore \quad X_1 + X_2 + \cdots + X_n \overset{d}{=} n\mu + \sqrt{n}\sigma Z$$

- X が $N(\mu, \sigma^2)$ に従えば，標準化した Z は $\sigma Z = X - \mu$ であり

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} (n - \sqrt{n})\mu + \sqrt{n}X$$

同じ分布に従う確率変数の和が，同じ分布に従う 1 個の確率変数の 1 次式の分布と一致する．$\Rightarrow$ このような分布を「安定分布」という．

中心極限定理

定理

平均 μ，分散 σ^2 であるような確率分布において，その確率分布に従う互いに独立な確率変数を $X_1, X_2, \ldots, X_n$ とする．その平均と標準化を

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n), \quad Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

とすると，次の式が成り立つ．

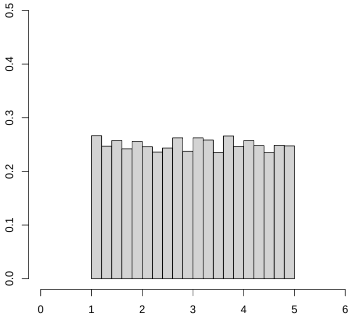
$$\lim_{n\to\infty} P(Z_n \leqq \alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} e^{-\frac{x^2}{2}}\, dx$$

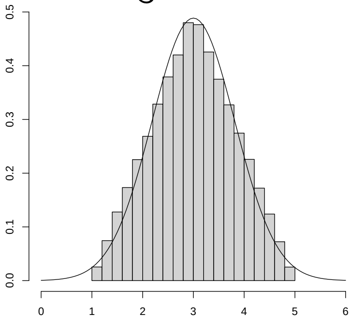
平均と分散が存在するような確率分布であれば，平均をとる個数が多くなるにつれ，

- 平均の標準化は標準正規分布 $N(0,1)$ に近づく．
- $\overline{X}$ は正規分布 $N\left(\mu, \dfrac{\sigma^2}{n}\right)$ に近づく．

具体例 (一様分布)

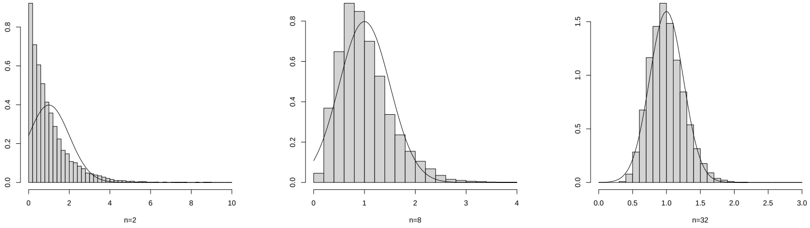
- 区間 $[a, b]$ における一様分布で n 個の平均と分散は
$$E(\overline{X}) = \frac{a+b}{2}, \quad V(\overline{X}) = \frac{(b-a)^2}{12n}$$
- 区間 $[1, 5]$ で考えて，$n = 2$ のとき 1 万個の乱数で確認すると，
$$E(\overline{X}) = 3, \quad V(\overline{X}) = \frac{2}{3}$$





具体例 (χ^2 分布)

- χ^2 分布では $E(X) = n, V(X) = 2n$ である．
- 自由度 1 の χ^2 分布では $E(X) = 1, V(X) = 2$ となるので，n 個の平均は $E(\overline{X}) = 1, V(\overline{X}) = \dfrac{2}{n}$ である．
- 自由度 1 で $n = 2, 8, 16$ の場合の平均のヒストグラム．曲線は $N\left(1, \dfrac{2}{n}\right)$ の確率密度曲線である．



$N(0,1)$ から得られる確率分布

Z_i, W を $N(0,1)$ に従う確率変数とする．

- $\mu + \sigma Z$ は $N(\mu, \sigma^2)$ に従う．
- $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ は，自由度 n の χ^2 分布に従う．
- $\dfrac{Z}{W}$ は，どのような分布に従うか？ $\dfrac{Z}{W} = \dfrac{Z}{\sqrt{W^2/1}}$ とすると，
  Z は $N(0,1)$，W^2 は自由度 1 の χ^2 分布に従うので
  $\dfrac{Z}{W}$ は自由度 1 の t 分布に従う．t 分布の確率密度関数は，

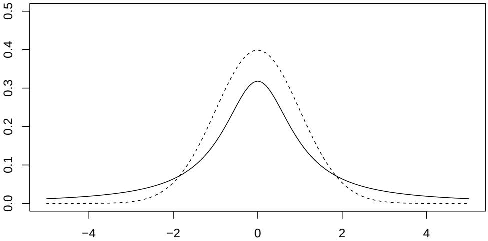
$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad (n \geqq 1)$$

  $n = 1$ のときは，次の式になり，コーシー分布になる．

$$f(x) = \frac{\Gamma(1)}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}\right)}\left(1+x^2\right)^{-1} = \frac{1}{\pi}\cdot\frac{1}{1+x^2}$$

コーシー分布の平均・分散

$N(0,1)$(破線) と，コーシー分布 (実線) の確率密度曲線．



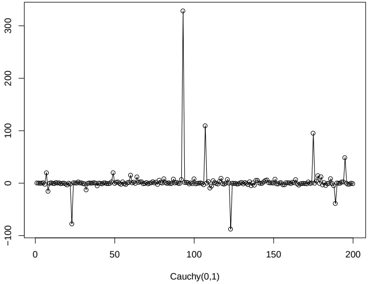
コーシー分布の平均は，

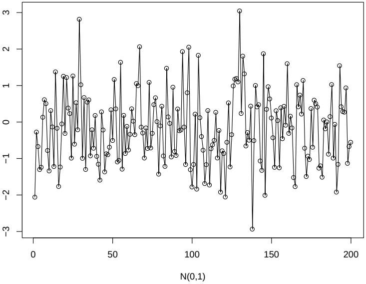
$$\begin{aligned} E(X) &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x}{x^2+1}\,dx \\ &= \frac{1}{2\pi}\left[\log(x^2+1)\right]_{-\infty}^{\infty} \\ &= \frac{1}{2\pi}\lim_{\substack{a\to-\infty \\ b\to\infty}}\left[\log(x^2+1)\right]_a^b \end{aligned}$$

この極限値は，存在しない．したがって，分散も存在しない．

コーシー分布に従う乱数

コーシー分布に従う乱数を 200 個生成し，$N(0, 1)$ に従う正規乱数と比較する．下図は，生成順の値である．

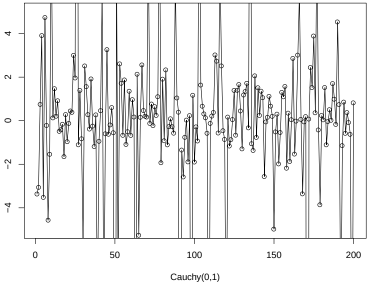


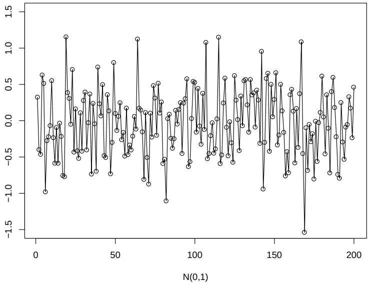


- コーシー分布では，ときおり，とんでもない値が出現する．
- $N(0, 1)$ では，$[-3, 3]$ の範囲内にあり，多くは $[-1, 1]$ 内にある．

コーシー分布に従う乱数 5 個の平均

コーシー分布に従う乱数 5 個の平均を，$N(0,1)$ に従う乱数 5 個の平均と比較する．生成順に 200 個を示す．





- コーシー分布に平均は，激しく上下し，どのような値になりそうか予想できない．平均が ± 1000 を超える場合もある．
  平均を取る個数をいくら増やしても，この状況は変わらない．
- $N(0,1)$ では 0 の近くに分布し，概ね $[-1,1]$ の範囲内にある．
  分散は $\dfrac{1}{5} = 0.2$ なので，標準偏差は $\sqrt{0.2} = 0.447$ である．

乱数平均の具体的な生成値

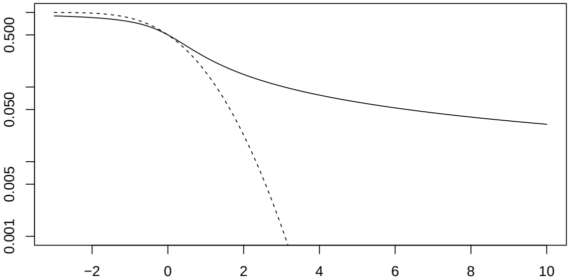
下の表は，5 個の乱数平均が 100 を超えた場合の，具体的な生成値である．

$Mean$	$rand1$	$rand2$	$rand3$	$rand4$	$rand5$
108.56	6.62	-1.18	536.27	0.839	0.239
202.08	-0.607	1013.5	-1.05	-0.580	-0.867
430.81	2156.6	-4.27	0.232	0.729	0.716
1186.5	-0.724	5945.3	-11.9	0.035	-0.197

- 乱数 5 個の平均を 200 個生成したので，乱数生成を 1000 回繰り返したことになる．
- 大部分は 0 の近くの値が生成されているが，1 個だけ桁の異なる値が紛れ込んだために，平均が大きくなっていることが分かる．

コーシー分布の相補累積分布曲線

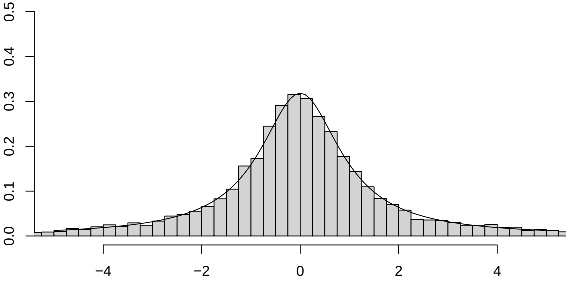
下図は，相補累積分布曲線 $P(X > x)$ である．縦軸は対数軸とした．破線は $N(0, 1)$ のものである．



- $N(0, 1)$ の相補累積分布は，$x \geqq 1$ を超えると急激に減少している．
- コーシー分布ではなかなか減少せず，むしろ減少の度合いが緩んでいる．
- $P(X \geqq 10) = 0.032$, $P(X \geqq 100) = 0.0032$ などとなっている．

5 個の乱数平均 1 万個のヒストグラム

5 個の乱数平均を 1 万回計算して，そのヒストグラムを描画すると，下図のようになる．



- 曲線は，コーシー分布の確率密度関数である．
- つまり，5 個の乱数平均のヒストグラムは，もとの確率密度曲線と一致する．
- コーシー分布では，このことが任意個数の平均で成立する．

一般化されたコーシー分布の確率密度関数

- ある確率分布の確率密度関数と累積分布関数を $f(x), F(x)$，確率変数を X とする．このとき，$a + bX$ $(b > 0)$ の累積分布関数を $G(x)$ とすると，$a + bX$ の確率密度関数 $g(x)$ は，

$$\begin{aligned}
G(x) &= P(a + bX \leqq x) = P\left(X \leqq \frac{x-a}{b}\right) \\
&= F\left(\frac{x-a}{b}\right) \\
\therefore \quad g(x) &= \frac{1}{b}F'\left(\frac{x-a}{b}\right) = \frac{1}{b}f\left(\frac{x-a}{b}\right)
\end{aligned}$$

- X がコーシー分布に従うとき，$a + bX$ $(b > 0)$ の確率密度関数は

$$\frac{1}{b} \cdot \frac{1}{\pi} \frac{1}{\left(\frac{x-a}{b}\right)^2 + 1} = \frac{1}{\pi} \cdot \frac{b}{(x-a)^2 + b^2}$$

一般化されたコーシー分布といい，$\mathrm{Cauchy}(a, b)$ で表す．

コーシー分布の特性関数

- Cauchy(0,1) の特性関数は，複素積分の留数を利用して計算される．

$$\varphi_X(t) = E(e^{itX}) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{e^{itx}}{1+x^2}\,dx = e^{-|t|}$$

- Cauchy(0,1) に従う互いに独立な確率変数を X, Y とすると，

$$\begin{aligned}\varphi_{X+Y}(t) &= E(e^{it(X+Y)}) = E(e^{itX})E(e^{itY}) \\ &= \varphi_X(t)\varphi_Y(t)\end{aligned}$$

- Cauchy(0,1) に従う確率変数の平均 $\overline{X}$ の特性関数は，

$$\begin{aligned}\varphi_{\overline{X}}(t) &= \varphi_{X_1+X_2+\cdots+X_n}\left(\frac{t}{n}\right) \\ &= \varphi_{X_1}\left(\frac{t}{n}\right)\varphi_{X_2}\left(\frac{t}{n}\right)\cdots\varphi_{X_n}\left(\frac{t}{n}\right) \\ &= e^{-|\frac{t}{n}|}e^{-|\frac{t}{n}|}\cdots e^{-|\frac{t}{n}|} \\ &= \left(e^{-|\frac{t}{n}|}\right)^n = e^{-|t|}\end{aligned}$$

レビ分布

- $N(0,1)$ に従う確率変数を Z とするとき，$U = \dfrac{1}{Z^2}$ が従う確率分布をレビ分布という．
- $U = \dfrac{1}{Z^2}$ の累積分布関数を $F(x)$ とすると，

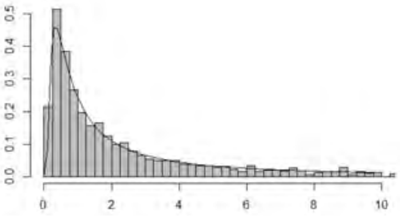
$$\begin{aligned}
F(x) &= P\left(\frac{1}{Z^2} \leqq x\right) = P\left(Z^2 \geqq \frac{1}{x}\right) \\
&= P\left(Z \leqq -\frac{1}{\sqrt{x}}\right) + P\left(Z \geqq \frac{1}{\sqrt{x}}\right) \\
&= 2P\left(Z \geqq \frac{1}{\sqrt{x}}\right) = 2\left\{1 - P\left(Z \leqq \frac{1}{\sqrt{x}}\right)\right\} \\
&= 2\left\{1 - \int_{-\infty}^{\frac{1}{\sqrt{x}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du\right\}
\end{aligned}$$

レビ分布の確率密度関数

- 累積分布関数 $F(x)$ を微分すると確率密度関数が得られる．

$$f(x) = -\frac{2}{\sqrt{2\pi}}e^{-\frac{1}{2x}} \cdot \left(-\frac{1}{2x^{\frac{3}{2}}}\right)$$
$$= \frac{1}{\sqrt{2\pi}}\frac{1}{x^{\frac{3}{2}}}e^{-\frac{1}{2x}} \quad (x > 0)$$

- 下図は，$N(0,1)$ に従う乱数 2000 個による $\dfrac{1}{Z^2}$ のヒストグラムと，上記の確率密度曲線である．



レビ分布の累積分布関数

- 累積分布関数は，下記の積分を計算することになる．

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \frac{1}{u^{\frac{3}{2}}} e^{-\frac{1}{2u}}\, du$$

- $t = \sqrt{\frac{1}{2u}}$ として置換積分を行うと，誤差関数で表すことができる。

$$dt = -\frac{1}{2\sqrt{2}} \frac{1}{u^{\frac{3}{2}}}\, du, \qquad \begin{array}{c|ccc} u & 0 & \to & x \\ \hline t & \infty & \to & \sqrt{\frac{1}{2x}} \end{array}$$

$$\begin{aligned} F(x) &= \int_{\infty}^{\sqrt{\frac{1}{2x}}} \frac{1}{\sqrt{2\pi}} (-2\sqrt{2} e^{-t^2}) dt \\ &= \frac{2}{\sqrt{\pi}} \int_{\sqrt{\frac{1}{2x}}}^{\infty} e^{-t^2}\, dt \\ &= 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt = 1 - \mathrm{erf}\,(x) \end{aligned}$$

レビ分布の平均と分散

- レビ分布の平均は次の積分を計算することになる．

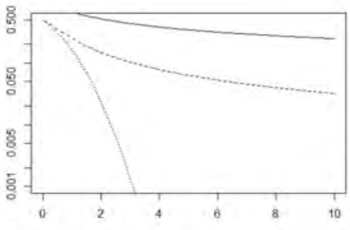
$$
\begin{aligned}
E(X) &= \frac{1}{\sqrt{2\pi}}\int_0^\infty x\cdot\frac{1}{x^{\frac{3}{2}}}e^{-\frac{1}{2x}}\,dx = \frac{1}{\sqrt{2\pi}}\int_0^\infty \frac{1}{\sqrt{x}}e^{-\frac{1}{2x}}\,dx \\
&> \frac{1}{\sqrt{2\pi}}\int_1^\infty \frac{1}{\sqrt{x}}e^{-\frac{1}{2x}}\,dx \quad (\text{積分範囲を縮小}) \\
&> \frac{1}{\sqrt{2\pi}}\int_1^\infty \frac{1}{\sqrt{x}}e^{-\frac{1}{2}}\,dx \quad \left(-\frac{1}{2x} > -\frac{1}{2}\right) \\
&= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}\left[2\sqrt{x}\right]_1^\infty = \infty
\end{aligned}
$$

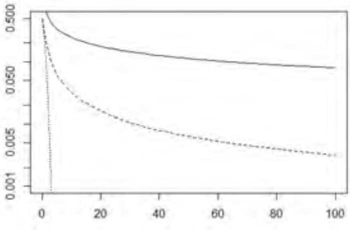
- 平均が存在しないので，分散も存在しない．
- 歪度も尖度もモーメント母関数も存在しない．
- 特性関数は，次の式で与えられる．

$$
\varphi_X(t) = E(e^{itX}) = e^{-\sqrt{|t|}(1+i\,\mathrm{sgn}\,t)}
$$

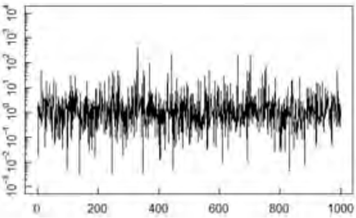
レビ分布の相補累積分布曲線と乱数

- 相補累積分布 $1 - F(x)$ を標準正規分布・コーシー分布と比較する.

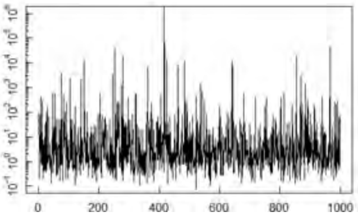




- 乱数を 1000 個発生させてコーシー分布と比較する.



コーシー分布の絶対値



レビ分布

レビ分布の性質

- U がレビ分布に従うとき，$a + bU$ $(b > 0)$ の確率密度関数は
$$\frac{1}{b} f\left(\frac{x-a}{b}\right) = \frac{1}{b}\frac{1}{\sqrt{2\pi}}\frac{1}{\left(\frac{x-a}{b}\right)^{\frac{3}{2}}} e^{-\frac{b}{2(x-a)}}$$
$$= \sqrt{\frac{b}{2\pi}}\frac{1}{(x-a)^{\frac{3}{2}}} e^{-\frac{b}{2(x-a)}} \qquad (x > a)$$
- この確率分布を Levy(a, b) で表す．特性関数は，
$$\varphi_{a+bX}(t) = e^{iat - \sqrt{b|t|}(1+i\,\mathrm{sgn}\,t)}$$
- Levy(a,b) に従う独立な確率変数 X, Y に対し $X + Y$ を考えると，
$$\varphi_{X+Y}(t) = E(e^{it(X+Y)}) = E(e^{itX})E(e^{itY})$$
$$= e^{-\sqrt{|t|}(1+i\,\mathrm{sgn}\,t)} e^{-\sqrt{|t|}(1+i\,\mathrm{sgn}\,t)}$$
$$= e^{2iat - 2\sqrt{b|t|}(1+i\,\mathrm{sgn}\,t)} = e^{i(2a)t - \sqrt{2^2 b|t|}(1+i\,\mathrm{sgn}\,t)}$$
- $X + Y$ は Levy$(2a, 2^2 b)$ に従うことを示す (再生性がある)

レビ分布も安定分布

- 同様にすると，Levy(a, b) に従う独立な確率変数を X_i $(1 \leqq i \leqq n)$ とすると $X_1 + X_2 + \cdots + X_n$ は Levy$(na, n^2 b)$ に従う．
- X が Levy(a, b) に従えば，Levy(0,1) に従う確率変数 U を用いて $X = a + bU$ と表せる．このとき，$c + dX = (ad + c) + bdU$ は Levy$(ad + c, bd)$ に従う．特に，dX は Levy(ad, bd) に従う．
- $X_1 + X_2 + \cdots + X_n$ は Levy$(na, n^2 b)$ に従うので，$\overline{X}$ は Levy(a, nb) に従う．
- したがって，X が Levy(a, b) に従うとき，$a(n - n^2) + n^2 X$ は

$$an^2 + a(n - n^2) + bn^2 U = an + bn^2 U$$

  となるので，Levy$(na, n^2 b)$ に従う．
- 以上より次の関係が成り立つので，レビ分布も安定分布である．

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} a(n - n^2) + n^2 X$$

べき分布の定義

- 確率密度関数 $f(x)$ がべき関数で表される確率分布を考える．

$$f(x) = \frac{C}{x^a} \quad (a > 0, x \geqq b, C > 0)$$

- 定数 C は $\int_b^\infty f(x)\,dx = 1$ より $C = (a-1)b^{a-1}$ となる．
形を整えるために，$a - 1$ を a で置き換えると

$$f(x) = \frac{ab^a}{x^{a+1}} = \frac{\frac{a}{b}}{\left(\frac{x}{b}\right)^{a+1}}$$

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

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

べき分布の累積分布関数

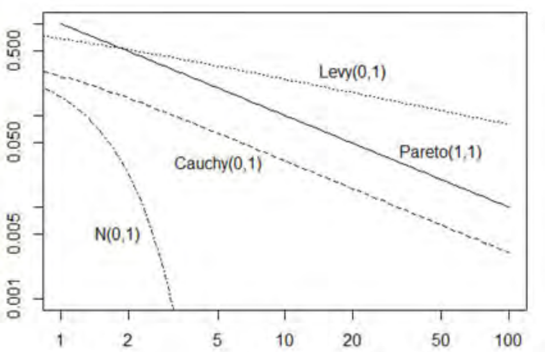
- 累積分布関数は

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

- したがって，相補累積分布関数もべき関数になる．

$$1 - F(x) = \frac{b^a}{x^a}$$

- 相補累積分布曲線を比較する．



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

$$V(X) = \frac{ab^2}{(a-1)^2(a-2)} \quad (a > 2)$$

- べき分布の n 次モーメント

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

べき分布のモーメント母関数

- べき分布のモーメント母関数は，次式で計算される．

$$M_X(t) = E(e^{tX}) = ab^a \int_b^\infty x^{-a-1} e^{tx}\,dx$$

- 右辺はガンマ関数 $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\,dt$ と類似している．
- $t > 0$ のときは発散し，$t = 0$ のときは 1 である．
- $t < 0$ のとき，ガンマ関数に近づけるため $tx = -u$ とおくと，

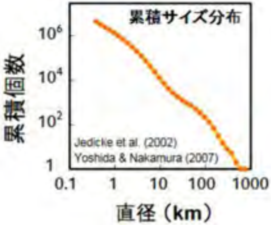
$$\begin{aligned} M_X(t) &= -\frac{ab^a}{t} \int_{-bt}^\infty \left(-\frac{u}{t}\right)^{-a-1} e^{-u}\,du \\ &= -\frac{ab^a}{t} \left(-\frac{1}{t}\right)^{-a-1} \int_{-bt}^\infty u^{-a-1} e^{-u}\,du \\ &= a(-bt)^a \int_{-bt}^\infty u^{-a-1} e^{-u}\,du \end{aligned}$$

- 第 2 種不完全ガンマ関数を用いると，次のように表される．

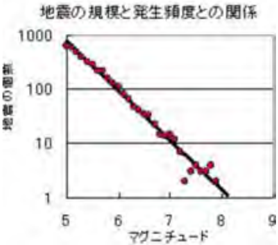
$$M_X(t) = a(-bt)^a \Gamma(-a, -bt)$$

べき分布に従う現象例

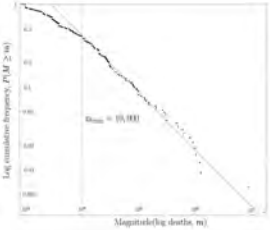
小惑星の直径と個数



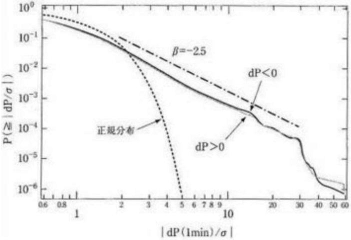
地震の規模と発生頻度



戦死者の数と戦争の数

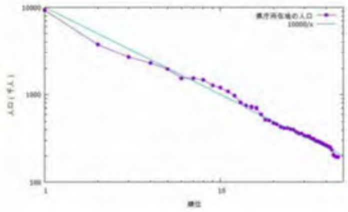


為替レートの変動

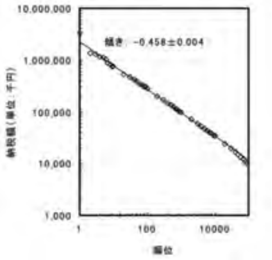


べき分布に従う現象例 (2)〜ジップの法則〜

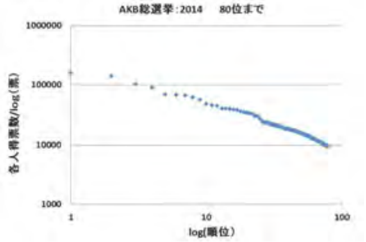
県庁所在地の人口と順位 (2003)



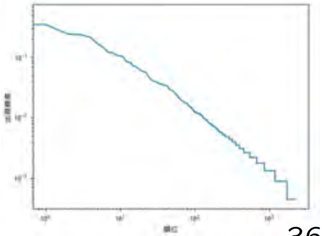
高額納税者と納税額



AKB の総選挙



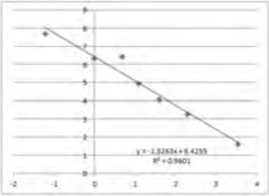
latex.ltx の制御綴り



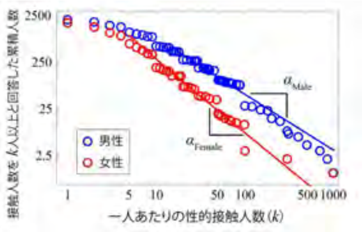
べき分布に従う現象例 (3)

ステンレス上の水滴

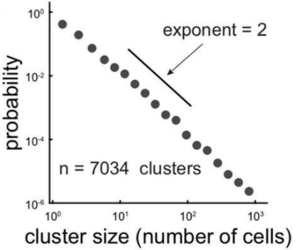
大きさと個数分布



性接触の人数



枯れ草菌のクラスターサイズ



べき分布 Pareto(a, b) の性質

- cX $(c > 0)$ の累積分布関数を $G(x)$ とすると，

$$\begin{aligned} G(x) &= P(cX \leqq x) = P\left(X \leqq \frac{x}{c}\right) \\ &= F\left(\frac{x}{c}\right) = 1 - \frac{b^a}{\left(\frac{x}{c}\right)^a} = 1 - \frac{(bc)^a}{x^a} \end{aligned}$$

  したがって，cX は，Pareto(a, bc) に従う．
- X^n $(n > 0)$ の累積分布関数を $G(x)$ とすると，

$$\begin{aligned} G(x) &= P(X^n \leqq x) = P(X \leqq x^{\frac{1}{n}}) \\ &= 1 - \frac{b^a}{\left(x^{\frac{1}{n}}\right)^a} = 1 - \frac{(b^n)^{\frac{a}{n}}}{x^{\frac{a}{n}}} \end{aligned}$$

  したがって，X^n は Pareto$\left(\frac{a}{n}, b^n\right)$ に従う．
- べき分布は，定数倍してもべき乗してもべき分布に従う．

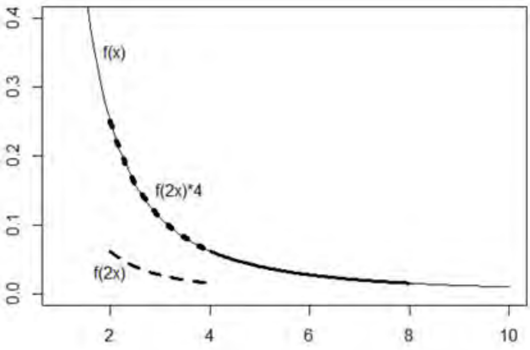
べき分布の自己相似性

- Pareto(a,b) の確率密度関数 $f(x)$ に対して

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

横軸方向の定数倍は，縦軸方向の定数倍と一致する．

- 下図では，Pareto(1,1) の確率密度関数を $f(x)$ とする．



自己相似性を持つ確率分布

定数 c に対して，$f(cx) = g(c)f(x)$ を満たす関数はどのような関数か？

- $x = 1$ とすると $f(c) = g(c)f(1)$ より，$f(cx) = \dfrac{f(c)}{f(1)}f(x)$
- 両辺を c で微分すると，$xf'(cx) = \dfrac{f'(c)}{f(1)}f(x)$ となる．

  $c = 1$ のときは，$xf'(x) = \dfrac{f'(1)}{f(1)}f(x)$ となる．
- $K = \dfrac{f'(1)}{f(1)}$ とおいて解くと，

$$\frac{f'(x)}{f(x)} = \frac{K}{x} \quad \text{積分して} \quad \int \frac{f'(x)}{f(x)}\,dx = K\int \frac{1}{x}\,dx$$

- これを解いて，

$$f(x) = Cx^K$$

- 自己相似性を持つ関数は，べき関数に限られる．

自己相似性による性質

- Pareto$(a, 1)$ で考えて $f(x) = \dfrac{a}{x^{a+1}}$ とする．$[1, c]$ での積分は

$$\int_1^c \frac{a}{x^{a+1}}\,dx = \left[-\frac{1}{x^a}\right]_1^c = 1 - \frac{1}{c^a} = p$$

$$\therefore \quad \frac{1}{c^a} = 1 - p$$

- 残りの $[c, \infty)$ で同じ割合 p になる部分は，全体では

$$(1-p)p = \frac{p}{c^a}$$

- $\dfrac{x}{c} = t$ と変数変換すると，t に関する $[1, c]$ での積分と一致する．

$$\int_1^c f(ct)\,c\,dt = \int_1^c \frac{a}{(ct)^{a+1}}\,c\,dt = \frac{1}{c^a}\int_1^c \frac{a}{t^{a+1}}\,dt = \frac{p}{c^a}$$

自己相似性による性質 (2)

- 確率変数 X は $\mathrm{Pareto}(a, b)$ に従うとする．
- 定数 c を $b < c$ として，
  $X > c$ という条件の下で $P(X > x)$ を考える．
- 条件付き確率であるから，

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

- これは，$X > c$ という条件のもとでは，X は $\mathrm{Pareto}(a, c)$ に従うことを示している．

べき分布と一様分布

- Pareto$(a, 1)$ に従う確率変数を X とするとき，$U = \dfrac{1}{X^a}$ はどのような分布に従うか？
- X の累積分布関数を $F(x)$ とすると，

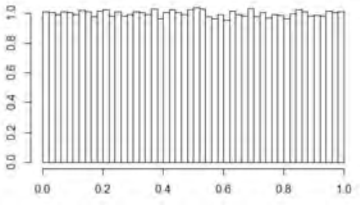
$$
\begin{aligned}
P(U \leqq x) &= P\left(\frac{1}{X^a} \leqq x\right) = P\left(X \geqq \frac{1}{x^{\frac{1}{a}}}\right) \\
&= 1 - F\left(\frac{1}{x^{\frac{1}{a}}}\right) = \frac{1}{\left(\frac{1}{x^{\frac{1}{a}}}\right)^a} = x
\end{aligned}
$$

- $U = \dfrac{1}{x^a}$ は $P(U \leqq x) = x$ を満たすので，区間 $[0, 1]$ における一様分布に従う．
- 逆に，U が区間 $[0, 1]$ での一様分布に従うとすると，$X = \dfrac{1}{U^{\frac{1}{a}}}$ は Pareto$(a, 1)$ に従う．

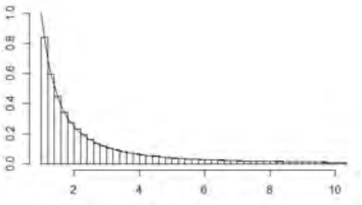
べき分布と一様分布 (2)

- Pareto(a, b) の場合も同様のことが成り立つ．
  X が Pareto(a, b) に従うとき $\left(\dfrac{b}{X}\right)^a$ は一様分布にしたがい，
  U が $[0, 1]$ での一様分布に従うとき $\dfrac{b}{U^{\frac{1}{a}}}$ は Pareto(a, b) に従う．
- Pareto$(1, 1)$ の場合に，乱数 10 万個を利用した確認．

$\dfrac{1}{X}$ のヒストグラム



$\dfrac{1}{U}$ のヒストグラム



べき分布と指数分布

- 確率密度関数が次式の確率分布を母数 λ の指数分布という．

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geqq 0) \\ 0 & (x < 0) \end{cases}$$

- 母数 λ の指数分布の累積分布関数は $F(x) = 1 - e^{-\lambda x}$
- X が $\mathrm{Pareto}(a, 1)$ に従うとき，$T = \log X$ は母数 a の指数分布に従う．

$$P(\log X \leqq t) = P(X \leqq e^t) = 1 - \frac{1}{(e^t)^a} = 1 - e^{-ta}$$

- T が母数 a の指数分布に従えば，$X = e^T$ は $\mathrm{Pareto}(a, 1)$ に従う．

$$\begin{aligned} P(e^T \leqq x) &= P(T \leqq \log x) = 1 - e^{-a \log x} \\ &= 1 - e^{\log x^{-a}} = 1 - x^{-a} \end{aligned}$$

- 確率変数 X が $\mathrm{Pareto}(a, 1)$ に従うのは，$\log X$ が母数 a の指数分布に従うときである．

べき分布とベータ分布

- ベータ関数：$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$

  ベータ分布：$f(x:a,b) = \dfrac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$

- $B(a,1) = \int_0^1 t^{a-1}\,dt = \dfrac{1}{a}$ より $f(x:a,1) = ax^{a-1}$

  このときの累積分布関数は　$F(x) = \int_0^x ax^{a-1}\,dx = x^a$

- X が Pareto$(a,1)$ に従えば，$V = \dfrac{1}{X}$ は $(a,1)$ のベータ分布

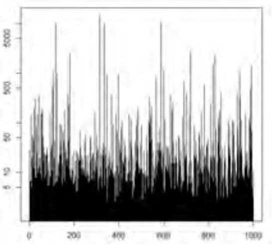
$$P\left(\frac{1}{X} \leqq v\right) = P\left(X \geqq \frac{1}{v}\right) = \frac{1}{(1/v)^a} = v^a$$

- V が $(a,1)$ のベータ分布に従えば，$X = \dfrac{1}{V}$ は Pareto$(a,1)$

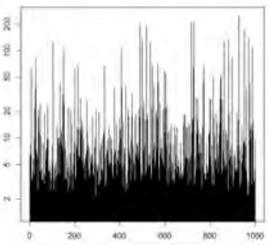
$$P\left(\frac{1}{V} \leqq x\right) = P\left(V \geqq \frac{1}{x}\right) = 1 - \left(\frac{1}{x}\right)^a = 1 - \frac{1}{x^a}$$

べき分布に従う乱数の 10 個平均

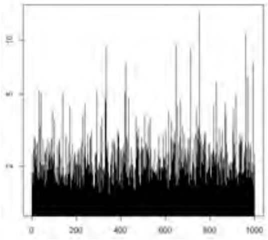
$a = 0.75,\ b = 1$



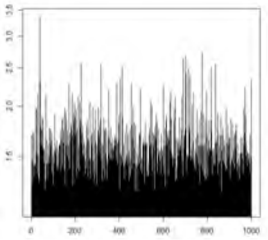
$a = 1,\ b = 1$



$a = 2,\ b = 1$

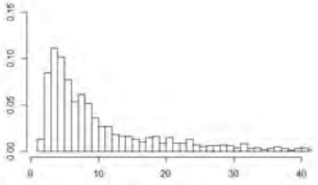


$a = 3,\ b = 1$

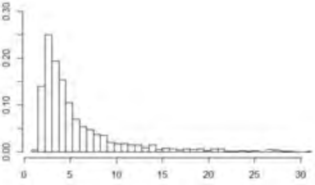


べき分布に従う乱数 10 個平均のヒストグラム

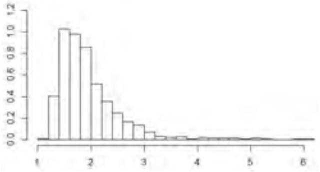
$a = 0.75,\ b = 1$



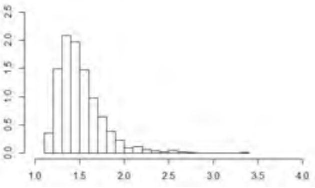
$a = 1,\ b = 1$



$a = 2,\ b = 1$



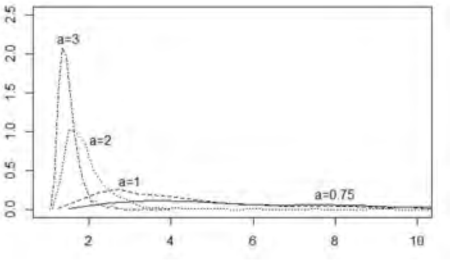
$a = 3,\ b = 1$



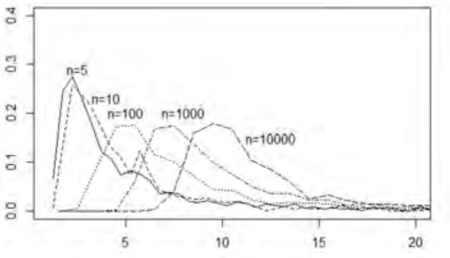
★平均の分布はべき分布ではない! $\Rightarrow$ べき分布は安定分布ではない.

べき分布に従う乱数平均の折れ線グラフ

- Pareto(α, 1) で指数 α を変えた乱数 10 個平均の場合

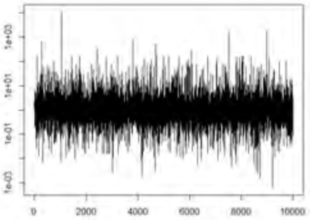


- Pareto(1, 1) で平均を取る個数を増やした場合

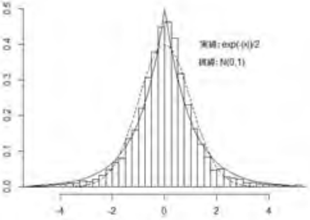


平均の比の分布

- Pare(1,1) に従う確率変数の 10 個平均の比 (生成順)



- 10 個平均の比の対数値のヒストグラム



株価変動でも，価格の比の対数値はラプラス分布に従う．

安定分布

- ある確率分布が安定分布であるのは，次の条件を満たすときである．同一の確率分布に従う互いに独立な確率変数を $X_1, X_2, \ldots, X_n$ とし，X も同じ分布に従うとき，ある定数 a_n, b_n $(b_n > 0)$ が存在して，次のことが成り立つ．

$$X_1 + X_2 + \cdots + X_n \overset{d}{=} a_n + b_n X$$

  a_n は centering parameter，b_n は norming parameter と呼ばれ，$b_n = n^{\frac{1}{\alpha}}$ $(0 < \alpha \leqq 2)$ の形である．α は指数と呼ばれる．
- 安定分布であるのは，特性関数が次の形のときに限られる．

$$\varphi_X(t) = \exp\left(itc - (d|t|)^\alpha \{1 + i\beta \operatorname{sgn}(t)\, u_\alpha(t)\}\right)$$

  α は指数 (index)，β は歪度 (skewness)，c は位置 (location)，d (> 0) はスケール (scale) を表すパラメータであり，$u_\alpha(t)$ は次式で定義される．

$$u_\alpha(t) = \begin{cases} \tan\dfrac{\pi\alpha}{2} & (\alpha \neq 1) \\ \dfrac{2}{\pi}\log|t| & (\alpha = 1) \end{cases}$$

べき分布と安定分布

- Pareto(α, 1) に従う n 個の互いに独立な確率変数の和を S_n，定数 a_n, b_n と指数 α を持つ安定分布の累積分布関数を $F_\alpha(x)$ とすると，

$$\lim_{n\to\infty} P\left(\frac{S_n - a_n}{b_n} < x\right) = F_\alpha(x)$$

- a_n は次式の通り．2 つ目の式は近似式，C はオイラー定数．

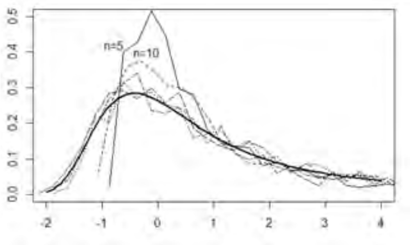
$$a_n = \begin{cases} 0 & (0 < \alpha < 1) \\ n\log n + n\left(1 - C - \log\dfrac{2}{\pi}\right) & (\alpha = 1) \\ \dfrac{n\alpha}{\alpha - 1} & (1 < \alpha < 2) \end{cases}$$

- $b_n = n^{\frac{1}{\alpha}} C_\alpha$ で，C_α は次式の通り．

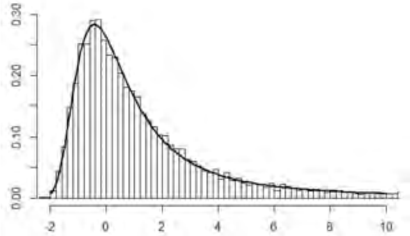
$$C_\alpha = \begin{cases} \left(\Gamma(1-\alpha)\cos\dfrac{\pi\alpha}{2}\right)^{\frac{1}{\alpha}} & (\alpha \neq 1) \\ \dfrac{\pi}{2} & (\alpha = 1) \end{cases}$$

安定分布への収束状況

- Pareto(1,1) に従う確率変数の $n = 5, 10, 100, 1000, 10000$ の平均を，係数 a_n, b_n で規格化した場合の折れ線グラフ．



- 太線は，$\alpha = 1, \beta = 1, c = 0, d = 1$ の場合の確率密度曲線．
- 10000 個の平均 10000 個の場合で確認



安定分布のべき関数による近似

- 安定分布 $S(\alpha, \beta, c, d)$ は 4 つのパラメータで規定される．
- 安定分布の確率密度関数を初等関数で表現可能な確率分布は，正規分布・コーシー分布・レビ分布の 3 つに限られる．具体的には，$N(\mu, \sigma^2)$ は $S\left(2, 0, \mu, \dfrac{\sigma}{\sqrt{2}}\right)$，Cauchy$(a, b)$ は $S(1, 0, b, a)$，そして Levy(a, b) は $S\left(\dfrac{1}{2}, 1, a+b, b\right)$ に対応する．
- $\alpha = \dfrac{2}{3}, \dfrac{3}{2}$ のときは，ホイッタカー (Whittaker) 関数で表せる．
- X が $S(\alpha, \beta, c, d)$ に従うとし，$c_\alpha = \Gamma(\alpha) \sin \dfrac{\pi\alpha}{2}$ とすると，$x \to \infty$ のとき，安定分布の確率密度関数について次が成り立つ．

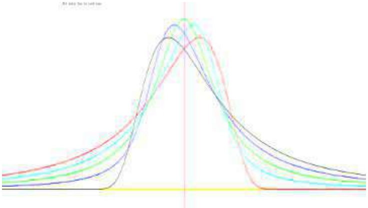
$$f(x) \approx \alpha \cdot \frac{d^\alpha c_\alpha (1+\beta)}{x^{\alpha+1}}, \qquad \frac{\alpha}{x^{\alpha+1}} \approx \frac{f(x)}{d^\alpha c_\alpha (1+\beta)}$$

安定分布の裾部分はべき関数で近似され，その逆もいえる．

安定分布のグラフ

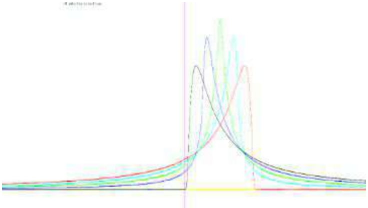
$S(1, \beta, 0, 1)$

$-1 \to 1$ step 0.5



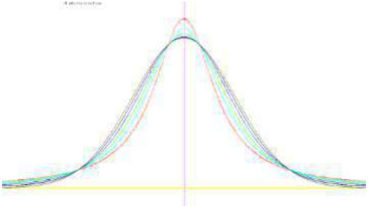
$S(0.5, \beta, 1, 1)$

$-1 \to 1$ step 0.5



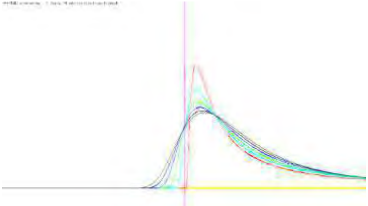
$S(\alpha, 0, 0, 1)$

$1 \to 2$ step 0.25



$S(\alpha, 1, 1, 1)$

$0.5 \to 1$ step 0.1



確率分布の裾の挙動分析

- 確率分布の裾部分を分析するために正則変動関数 $f(x)$ が用いられる．

$$f(x) = x^{\rho}\ell(x), \quad \lim_{x\to\infty}\frac{\ell(\lambda x)}{\ell(x)} = 1$$

- $\ell(x)$ としては，次のような関数がある．緩慢変動関数という．

$$\log x, \ (\log x)^2, \ e^{\sqrt{\log x}}, \ e^{\sqrt[3]{\log x}\cos\sqrt[3]{\log x}}$$

- $\ell(x)$ は次の性質を満たす．$\varepsilon > 0$ は任意．

$$\lim_{x\to\infty} x^{\varepsilon}\ell(x) = \infty, \quad \lim_{x\to\infty} x^{-\varepsilon}\ell(x) = 0, \quad \lim_{x\to\infty}\frac{\log \ell(x)}{\log x} = 0$$

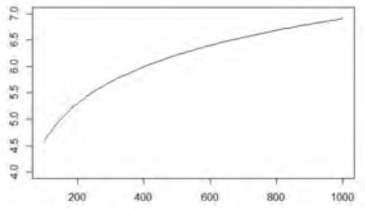
- 同一の確率分布 F に従う確率変数 X_i $(1 \leqq i \leqq n)$ に対して，その最大値を M_n とする．定数 a_n, b_n に対して

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G$$

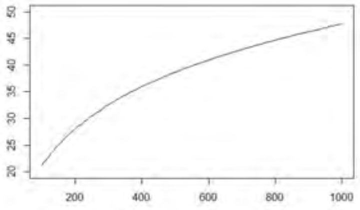
となるとき，G を極値分布といい，F は G に吸引されるという．

緩慢変動関数のグラフ

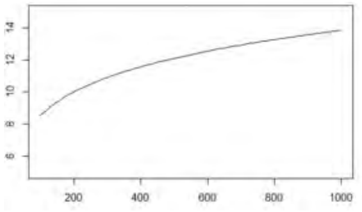
$\log x$



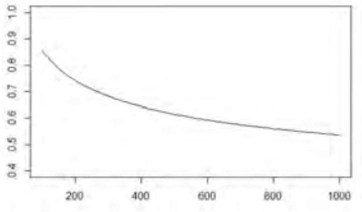
$(\log x)^2$



$e^{\sqrt{\log x}}$



$e^{\sqrt[3]{\log x}\cos\sqrt[3]{\log x}}$

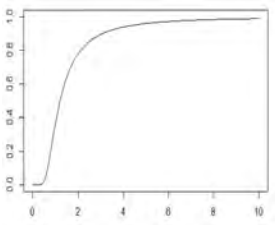


極値分布パターン

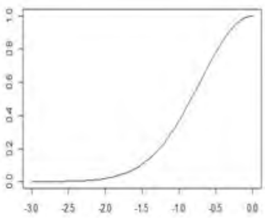
極値分布は次の 3 つに限られる．$\alpha > 0$, $F(x)$ は累積分布関数．

1. フレシェ (Fréchet) 分布 $F(x) = \begin{cases} 0 & (x \leqq 0) \\ e^{-x^{-\alpha}} & (x > 0) \end{cases}$
2. (負の) ワイブル (Weibull) 分布 $F(x) = \begin{cases} e^{-(-x)^{\alpha}} & (x \leqq 0) \\ 1 & (x > 0) \end{cases}$
3. ガンベル (Gumbel) 分布 $F(x) = e^{-e^{-x}}$
4. 下図は，$\alpha = 2$ の場合のグラフである．

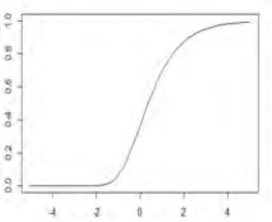
フレシェ分布



ワイブル分布



ガンベル分布



参考文献

- https://yunavi.lsv.jp/powerlaw.html

ご静聴ありがとうございました！