



高知工科大学

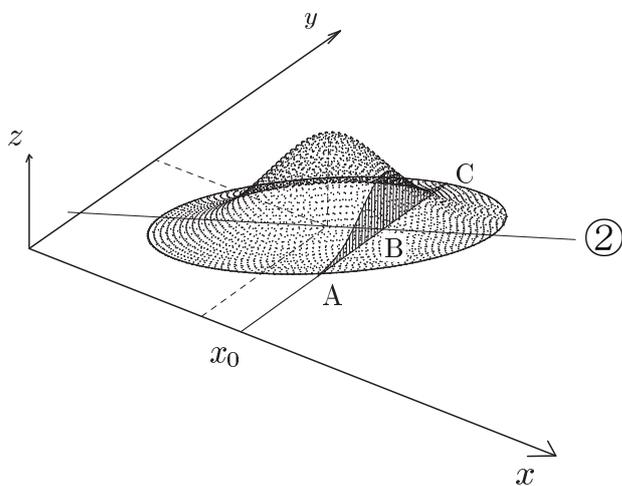
Kochi University of Technology

基礎数学ワークブック

(2005年度版)

番外編

「確率分布」



内容

- ◎ 確率分布
- ◎ 統計的推測
- ◎ ポアソン過程・ブラウン運動

井上 昌昭 著

< 離散型確率分布 1 >

確率変数 X のとる値が

$$x_1, x_2, \dots, x_n, \dots$$

のように定まっています。各値をとる確率が

$$P(X = x_k) = p_k \quad (k = 1, 2, \dots)$$

で与えられているとき、 X を離散型確率変数といい、その分布

$P(X = x_k) = p_k \quad (k = 1, 2, \dots)$ を離散型確率分布という。ここで

$p_k \geq 0, \sum_{k=1}^{\infty} p_k = 1$ である。このとき任意の関数 $f(x)$ に対し

$$E[f(X)] = \sum_{k=1}^{\infty} f(x_k)P(X = x_k) = \sum_{k=1}^{\infty} f(x_k)p_k$$

と定める。 X の平均と分散は

$$E[X] = \sum_{k=1}^{\infty} x_k p_k = m \quad : \text{平均}$$

$$V[X] = E[(X - m)^2] = \sum_{k=1}^{\infty} (x_k - m)^2 p_k \quad : \text{分散}$$

となる。

例 1 (二項分布) サイコロ投げやコイン投げをくり返し行うように、同じ試行をくり返して行うことを、「ベルヌーイ試行」という。成功確率 p の試行をくり返して n 回行う。

(これを成功確率 p のベルヌーイ試行という。) 各回は互いに独立である。成功した回数を X とすると

$$P(X = k) = {}_n C_k p^k (1-p)^{n-k} \quad (k = 0, 1, \dots, n)$$

となる。この分布を二項分布 $B(n, p)$ という。

平均と分散は

$$E[X] = np, \quad V(X) = np(1-p)$$

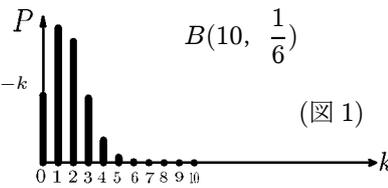
である。

< 離散型確率分布 2 >

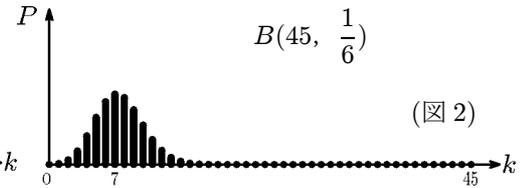
$p = \frac{1}{6}$ の場合二項分布

$$P = P(X = k) = {}_n C_k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}$$

の値を棒グラフにした



(図 1)



(図 2)

ものが図 1($n = 10$) と図 2($n = 45$) である。 n が大きくなると平均 np , 分散 $np(1-p)$ の正規分布に近づく。

例 2 < 幾何分布 >

成功確率 $p(0 < p < 1)$ のベルヌーイ試行で、初めて成功するまでの間に何回失敗したかを数え、その失敗の回数を X とする。 $X = k$ ということは、最初から連続 k 回失敗し、 $k + 1$ 回目に初めて成功した場合であるから、その確率は

$$P(X = k) = p(1-p)^k \quad (k = 0, 1, 2, \dots)$$

となる。この分布を幾何分布 $G(p)$ という。

平均と分散は

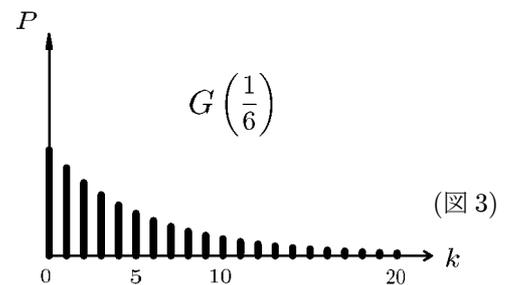
$$E[X] = \sum_{k=0}^{\infty} kp(1-p)^k = \frac{1-p}{p} \quad : \text{平均}$$

$$V[X] = \sum_{k=1}^{\infty} \left(k - \frac{1-p}{p}\right)^2 p(1-p)^k = \frac{1-p}{p^2} \quad : \text{分散}$$

である。 $p = \frac{1}{6}$ の場合の幾何分布

$$P(X = k) = \frac{1}{6} \left(\frac{5}{6}\right)^k \quad (k = 0, 1, 2, \dots)$$

を棒グラフにしたものが図 3 である。



(図 3)

< パスカル分布 >

例3 < パスカル分布=負の二項分布 >

成功確率 p ($0 < p < 1$) のベルヌーイ試行で, r 回成功するまでの失敗回数を X とすると

$$P(X = k) = {}_{r+k-1}C_k p^r (1-p)^k \quad (k = 0, 1, 2, \dots)$$

となる。この分布をパスカル分布または負の二項分布 $NB(r, p)$ という。

負の二項分布 (*negative binomial distribution*) と呼ばれるのは,

$f(x) = (1-x)^{-r}$ のマクローリン展開

$$\begin{aligned} (1-x)^{-r} &= 1 + \frac{r}{1!}x + \frac{(r+1)r}{2!}x^2 + \frac{(r+2)(r+1)r}{3!}x^3 + \dots \\ &= \sum_{k=0}^{\infty} {}_{r+k-1}C_k x^k \quad (\text{負の二項展開}) \end{aligned}$$

が負の二項展開と呼ばれるからである。この平均と分散は

$$E[X] = \sum_{k=0}^{\infty} k {}_{r+k-1}C_k p^r (1-p)^k = \frac{r(1-p)}{p} \quad : \text{平均}$$

$$V[X] = \sum_{k=0}^{\infty} \left(k - \frac{r(1-p)}{p} \right)^2 {}_{r+k-1}C_k p^r (1-p)^k = \frac{r(1-p)}{p^2} \quad : \text{分散}$$

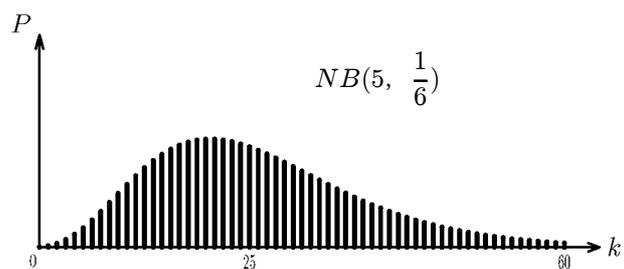
となる。

右図は $p = \frac{1}{6}$, $r = 5$ の

場合のパスカル分布

$$P = P(x = k) = {}_{k+4}C_k \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^k$$

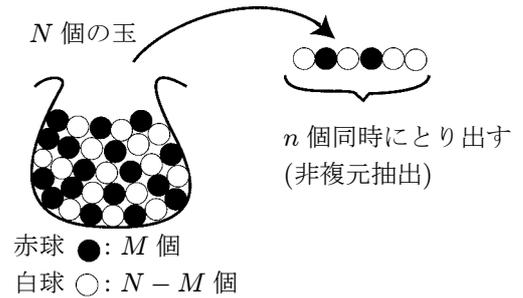
を棒グラフとしたものである。



< 超幾何分布 >

例4 < 超幾何分布 >

壺の中に N 個の玉が入っていて、そのうち M 個が赤球、 $N - M$ 個が白玉である。この壺から 1 度に n 個の玉をとり出す。このとき、とり出した玉は壺にもどさない (非復元抽出)。このとり出した n 個のうち赤球の数を X とする。このとき X の確率は



$$P(X = k) = \frac{M C_k \times (N - M) C_{n - k}}{N C_n} \quad (k = 0, 1, 2, \dots, n)$$

となる。この分布を超幾何分布 $H(N, n, p)$ (ただし $p = \frac{M}{N}$)

という。平均と分散は

$$E[X] = np, \quad V(X) = E[(X - np)^2] = n \left(\frac{N - n}{N - 1} \right) p(1 - p) = v$$

である。図1は $N = 300, n = 30, p = 0.4$

の場合の超幾何分布であり、図2は

$n = 30, p = 0.4$ の場合の二項分布である。

一般に N が十分大きいときは超幾何分布

は二項分布で近似できる。

定理 1

$$\lim_{N \rightarrow \infty} \frac{pN C_k \times (1-p)N C_{n-k}}{N C_n} = n C_k p^k (1-p)^{n-k}$$

図3の棒グラフは超幾何分布 $H(300, 30, 0.4)$

であり、曲線は正規分布曲線 $y = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-np)^2}{2v}}$

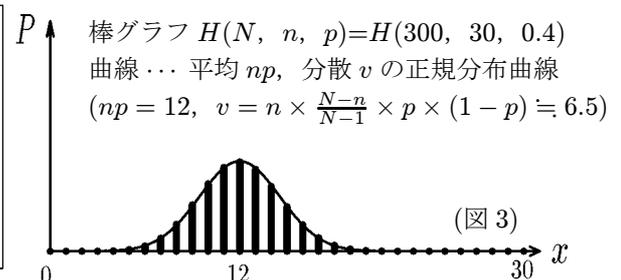
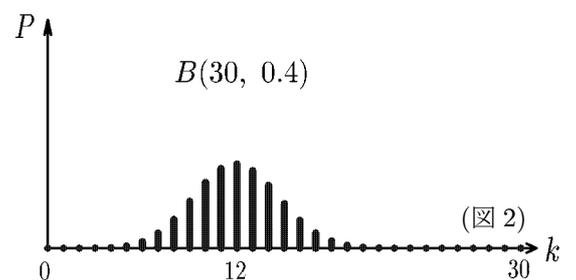
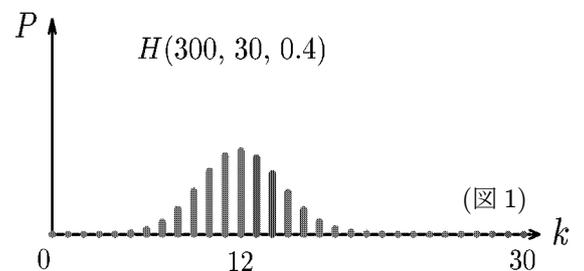
である。ただし、 $p = \frac{M}{N} (= 0.4)$, $v = n \times \frac{N-n}{N-1} p(1-p) = 30 \times \frac{270}{299} \times 0.4 \times 0.6 (= 6.5)$ である。

定理 2

$\frac{M}{N} = p, \frac{n}{N} = q$ が一定という条件で

$N \rightarrow \infty$ とするとき

$$\lim_{N \rightarrow \infty} \left\{ P(a < X < b) - \int_a^b \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-np)^2}{2v}} dx \right\} = 0$$

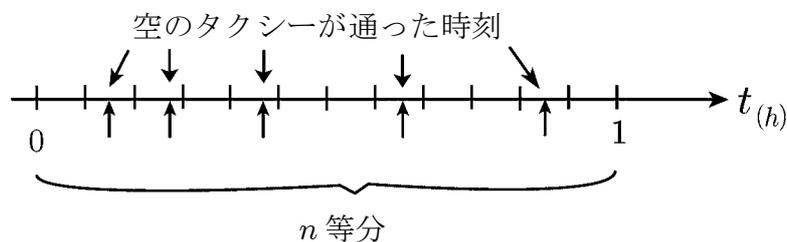


< ポアソン分布 1 >

例5 <ポアソン分布>

ある通りで空のタクシーが通る回数を調べたら、平均すると1時間に λ 回であった。空のタクシーがいつ通るかはまったく偶然であるが、微小時間に2台以上通ることはほとんどないとする。このとき1時間に通る空のタクシーの台数を X として、確率 $P(X = k)$ を求めたい。

1時間を n 等分して、微小時間に分ける。



n を大きくすれば各時間帯は2台以上通らない。すなわち1台通るか通らないかどちらかである。 $\frac{1}{n}$ 時間に空のタクシーが通る回数は平均 $\frac{\lambda}{n}$ 回であるから、この時間帯に空のタクシー1台が通る確率は $\frac{\lambda}{n}$ と考えてよい。各時間帯で空のタクシーが通るかどうかは無関係だから、独立に起こる。従って X は成功確率 $\frac{\lambda}{n}$ のベルヌーイ試行を n 回くり返したときの成功回数と同じであるから、二項分布 $B\left(n, \frac{\lambda}{n}\right)$ に従う。よって確率は

$$\begin{aligned}
 P(X = k) &= {}_n C_k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
 &= \frac{n(n-1)\cdots(n-k-1)}{k!} \times \frac{\lambda^k}{n^k} \times \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
 &= \frac{\lambda^k}{k!} \times \frac{n}{n} \times \frac{n-1}{n} \times \cdots \times \frac{n-k+1}{n} \times \left(1 - \frac{\lambda}{n}\right)^n \times \left(1 - \frac{\lambda}{n}\right)^{-k} \\
 &= \frac{\lambda^k}{k!} \times 1 \times \left(1 - \frac{\lambda}{n}\right) \times \cdots \times \left(1 - \frac{\lambda}{n}\right) \times \left\{\left(1 - \frac{\lambda}{n}\right)^{\frac{n}{\lambda}}\right\}^{-\lambda} \times \left(1 - \frac{\lambda}{n}\right)^{-k}
 \end{aligned}$$

< ポアソン分布 2 >

ここで

$$\lim_{n \rightarrow \infty} \left(1 + \frac{-\lambda}{n}\right)^{-\frac{n}{\lambda}} = \lim_{x \rightarrow -0} (1+x)^{\frac{1}{x}} = e \quad (\text{自然対数の底})$$

だから

$$\lim_{n \rightarrow \infty} P(X = k) = \frac{\lambda^k}{k!} \times 1 \times 1 \times \cdots \times 1 \times e^{-\lambda} \times 1 = e^{-\lambda} \frac{\lambda^k}{k!}$$

が成り立つ。

一般に定数 $\lambda > 0$ に対して,

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k = 0, 1, 2, \dots)$$

である確率分布をポアソン分布 $P(\lambda)$ という。この平均

と分散は

$$E[X] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \lambda, \quad V(X) = \sum_{k=0}^{\infty} (k - \lambda)^2 e^{-\lambda} \frac{\lambda^k}{k!} = \lambda$$

である。

(注1) 例の条件で「微小時間に空のタクシーが2台以上通ることはない」とした。このようにポアソン分布は「まれに起こる現象」の確率を表す。

(注2) 例の極限の結果をまとめると

$$\lim_{n \rightarrow \infty} {}_n C_k \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k = 0, 1, 2, \dots)$$

となる。すなわち二項分布の極限がポアソン分布である。このことを

「二項分布のポアソン近似」

とか

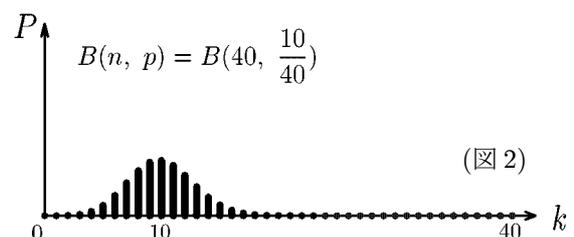
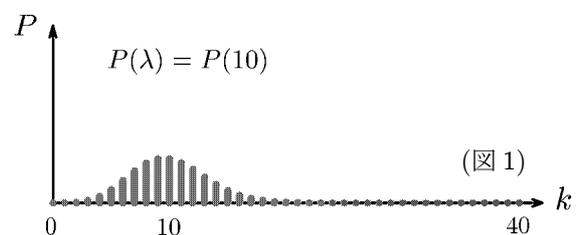
「ポアソンの少数の法則」

などと言う。

図1は $\lambda = 10$ の場合のポアソン分布

であり、図2は $n = 40$, $p = \frac{10}{40} = \frac{1}{4}$ の

場合の二項分布である。



< 多項分布 >

例 6 < 多項分布 >

二項分布を多次元に一般化したのが多項分布である。二項分布のコイン投げをサイコロ投げに変えたと考えれば良い。いま k 個の面をもつ仮想のサイコロを考える。第 i 番目の面が出る確率を p_i とする。 $p_i > 0$, $p_1 + p_2 + \cdots + p_k = 1$ である。このサイコロを N 回投げたときに i 番目の面が出た回数を X_i とおく。このとき

$$P(X_1 = n_1, X_2 = n_2, \cdots, X_k = n_k) = \frac{N!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

となる。ただし $n_1 + n_2 + \cdots + n_k = N$ である。この分布を**多項分布** $M(N, (p_i))$ という。多項分布と呼ばれるのは、多項展開式

$$(p_1 + p_2 + \cdots + p_k)^N = \sum_{n_1 + n_2 + \cdots + n_k = N} \frac{N!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

の各項を確率としているからである。

(X_1, X_2, \cdots, X_k) の分布は k 次元分布であり、各 i に対し X_i の分布は 1 次元分布である。この 1 次元分布を多次元分布の**周辺分布** という。 X_i の分布は二項分布 $B(N, p_i)$ であるから、その平均と分散は

$$E[X_i] = Np_i, \quad V[X_i] = E[(X_i - E[X_i])^2] = Np_i(1 - p_i)$$

である。また $X_i + X_j$ は二項分布 $B(N, p_i + p_j)$ に従うから、

$$E[X_i + X_j] = N(p_i + p_j), \quad V[X_i + X_j] = N(p_i + p_j)(1 - p_i - p_j)$$

である。さらに共分散 $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$ は

$$Cov(X_i, X_j) = \frac{1}{2} \{V(X_i + X_j) - V(X_i) - V(X_j)\} = -Np_i p_j$$

となる。これによって分散共分散行列 $(Cov(X_i, X_j))$ が求められる。

ただし $Cov(X_i, X_i) = V(X_i)$ である。

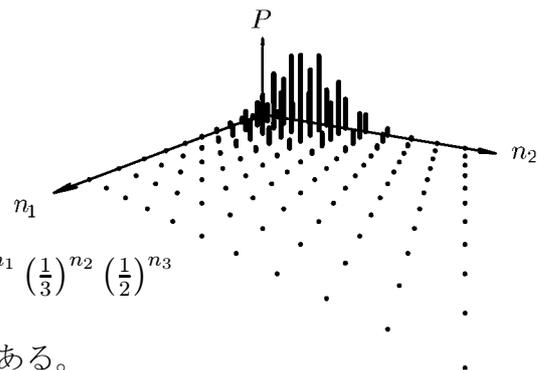
右図は $k = 3$, $p_1 = \frac{1}{6}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{2}$,

$N = 10$ のとき確率

$$P(X_1 = n_1, X_2 = n_2, X_3 = n_3) = \frac{10!}{n_1! n_2! n_3!} \left(\frac{1}{6}\right)^{n_1} \left(\frac{1}{3}\right)^{n_2} \left(\frac{1}{2}\right)^{n_3}$$

を (n_1, n_2) 平面上の棒の高さで表現したものである。

ただし $n_3 = 10 - n_1 - n_2$ である。



< 連続型確率分布 >

確率変数 X に対し, 非負値関数テスト $p(x)$ が存在し

$$P(a < X < b) = \int_a^b p(x) dx \quad (a < b)$$

を満たすとき, X は連続型の確率変数といい, $p(x)$ を X の**確率密度関数**という。

(注) $p(x)$ が確率密度関数であれば

$$p(x) \geq 0 \quad , \quad \int_{-\infty}^{\infty} p(x) dx = 1$$

を満たす。

$p(x)$ を確率密度関数とする確率変数の平均と分散は

$$E[X] = \int_{-\infty}^{\infty} xp(x) dx = m \quad : \text{平均}$$

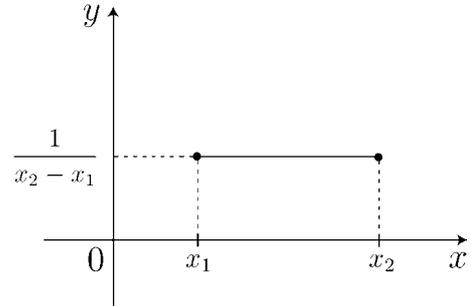
$$V[X] = E[(X - m)^2] = \int_{-\infty}^{\infty} (x - m)^2 p(x) dx \quad : \text{分散}$$

となる。

例 (一様分布)

定数 x_1, x_2 ($x_1 < x_2$) に対し

$$p(x) = \begin{cases} \frac{1}{x_2 - x_1} & : x_1 \leq x \leq x_2 \\ 0 & : \text{その他} \end{cases}$$



を確率密度関数とする確率変数 X

の分布を**一様分布**という。平均と分散は

$$E[X] = \frac{x_1 + x_2}{2} \quad , \quad V[X] = \frac{(x_2 - x_1)^2}{12} \quad \dots (*)$$

となる。

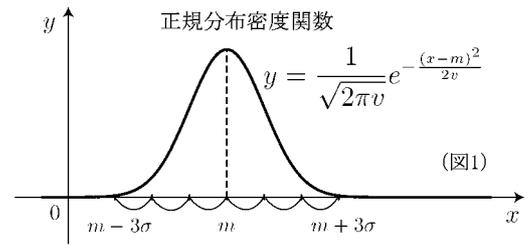
問 (*) 式を証明せよ。

< 正規分布 1 >

定数 $m, v (v > 0)$ に対し, 関数

$$p(x) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} \quad (x \in \mathbf{R})$$

を確率密度関数にもつ確率変数 X の分布を



正規分布 (normal distribution) といい, $N(m, v)$ で表す。

X の平均と分散は

$$E[X] = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} dx = m \quad : \text{平均}$$

$$V[X] = \int_{-\infty}^{\infty} \frac{(x-m)^2}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} dx = v \quad : \text{分散}$$

である。図1は $y = p(x)$ のグラフである。ここで標準偏差を $\sigma = \sqrt{v}$ とすると

$$P(m - k\sigma \leq X \leq m + k\sigma) = \int_{m-k\sigma}^{m+k\sigma} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} dx = \int_{-k}^k \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

より正規分布表で積分値を求めると

$$P(m - 2\sigma \leq X \leq m + 2\sigma) = 0.9544 \quad , \quad P(m - 3\sigma \leq X \leq m + 3\sigma) = 0.9973$$

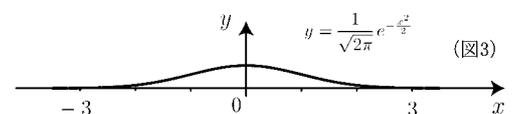
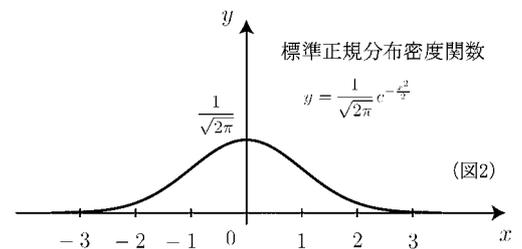
であるから X が平均から 2σ の範囲にある確率は 95.44 % であり, 平均から 3σ の範囲にある確率は 99.73 % である。

$m = 0, v = 1$ のときの分布 $N(0, 1)$ を **標準正規分布** という。図2は, その密度関数のグラフ $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ である。ただし図2は y 軸方向を拡大している。(注: $\frac{1}{\sqrt{2\pi}} \doteq 0.4$)

x 軸方向と y 軸方向を同じ長さにするると,

$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ のグラフは図3のようになる。

実際の標準正規分布曲線は図3のようになるが, 平たくなりすぎるので, 図2のような曲線として描いてあることが多い。



< 正規分布 2 >

定理 3 X が正規分布 $N(m, v)$ に従う確率変数とする。

a, b ($a \neq 0$) に対し, 確率変数

$$Y = aX + b$$

は正規分布 $N(am + b, av^2)$ に従う。特に

$$X^* = \frac{X - m}{\sqrt{v}}$$

は標準正規分布 $N(0, 1)$ に従う。

(注) 一般に $E[X] = m, V[X] = v$ のとき $Y = aX + b$ の平均と分散は

$$E[Y] = am + b, \quad V[Y] = a^2v$$

である。次の定理 4, 5 は正規分布特有の性質である。

定理 4 確率変数 X_1 は正規分布 $N(m_1, v_1)$ に従い,

確率変数 X_2 は正規分布 $N(m_2, v_2)$ に従う。

X_1 と X_2 が独立ならば, 和 $X_1 + X_2$ は

正規分布 $N(m_1 + m_2, v_1 + v_2)$ に従う。

定理 5 確率変数 X_1, X_2, \dots, X_n は独立で標準正規分布 $N(0, 1)$ に従う。

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{と} \quad \text{するとき,}$$

$\{X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_{n-1} - \bar{X}\}$ と \bar{X} は独立である。

系 1 確率変数 Y_1, Y_2, \dots, Y_n は独立で同じ正規分布 $N(m, v)$ に従う。

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{と} \quad \text{するとき,}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \quad \text{と} \quad \bar{Y} \quad \text{は独立である。}$$

系 2 確率変数 X_1, X_2, \dots, X_n は独立で同じ正規分布 $N(m, v)$ に従うならば

$$\chi^2 = \frac{1}{v} \sum_{i=1}^n (X_i - \bar{X})^2$$

は自由度 $n - 1$ の χ^2 分布に従う。

(注) χ^2 分布は 13 ページ参照。

< ガンマ関数とベータ関数 >

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad (\alpha > 0)$$

をガンマ関数という。部分積分より

$$\begin{aligned} \Gamma(\alpha) &= [-x^{\alpha-1} e^{-x}]_0^{\infty} - \int_0^{\infty} (\alpha-1)x^{\alpha-2}(-e^{-x})dx \\ &= (\alpha-1) \int_0^{\infty} x^{\alpha-2} e^{-x} dx \\ &= (\alpha-1)\Gamma(\alpha-1) \end{aligned}$$

より

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \quad (\alpha > 1)$$

が成り立つ。また

$$\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$$

であるから、自然数 $n (\geq 1)$ に対し

$$\Gamma(n) = (n-1)!$$

である。また $\lambda > 0$ に対し

$$\int_0^{\infty} x^{\alpha-1} e^{-\lambda x} dx = \frac{\Gamma(\alpha)}{\lambda^{\alpha}}$$

が成り立つ。

$\alpha > 0, \beta > 0$ に対し、関数

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

をベータ関数という。次式が成り立つ。

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = B(\beta, \alpha)$$

< ガンマ分布・指数分布 >

< ガンマ分布 >

定数 $\alpha, \beta > 0$ に対し

$$p(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} \quad (x > 0)$$

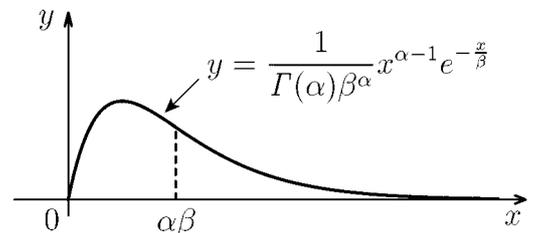
を密度とする確率分布を**ガンマ分布** $\text{Gamma}(\alpha, \beta)$ という。

α は形の母数, β は尺度母数といわれる。特に $\text{Gamma}(\alpha, 1)$ は形の母数 α の標準ガンマ分布という。 $\text{Gamma}(\alpha, \beta)$ の平均と分散は

$$\int_0^\infty x \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} dx = \alpha\beta \text{ (平均)}, \quad \int_0^\infty (x - \alpha\beta)^2 \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} dx = \alpha\beta^2 \text{ (分散)}$$

となる。

定理6 X_1, X_2 がそれぞれ $\text{Gamma}(\alpha_1, \beta)$, $\text{Gamma}(\alpha_2, \beta)$ に従う確率変数で X_1 と X_2 が独立ならば, 和 $X_1 + X_2$ は $\text{Gamma}(\alpha_1 + \alpha_2, \beta)$ に従う。



< 指数分布 > $\alpha = 1$ のガンマ分布を指数分布という。その密度は

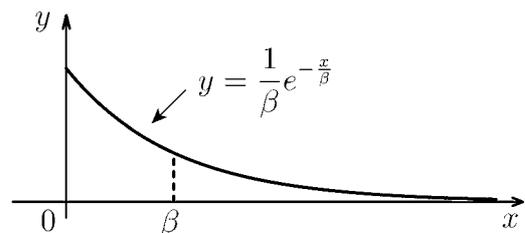
$$p(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \quad (x > 0)$$

である。平均は β , 分散は β^2 である。

例 ポアソン分布の例のタクシーの場合, 空のタクシーが平均 1 時間に λ 台通るとき, 1 台の空タクシーが通りすぎた後で, 次のタクシーが通るまでの時間を ξ とすると

$$P(\xi < s) = \int_0^s \lambda e^{-\lambda x} dx \quad \left(\beta = \frac{1}{\lambda} \text{ の指数分布} \right)$$

となる。この理由はポアソン過程の項で詳しく説明する。



< χ^2 分布, ベータ分布 >< χ^2 分布 >

X_1, X_2, \dots, X_n は独立で正規分布 $N(m, \sigma^2)$ に従うとする。このとき

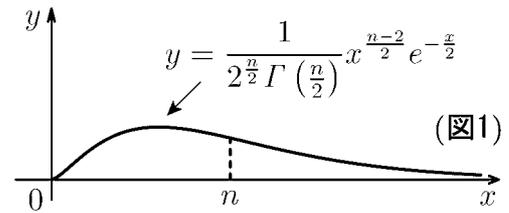
$X = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - m)^2$ は $\alpha = \frac{n}{2}, \beta = 2$ のガンマ分布に従う。すなわち

$$P(a < X < b) = \int_a^b \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n-2}{2}} e^{-\frac{x}{2}} dx \quad (a < b)$$

となる。この分布を**自由度 n の χ^2 分布** (カイ 2 乗分布) という。平均は

$E[X] = n$ であり, 分散は $V(X) = 2n$ である。

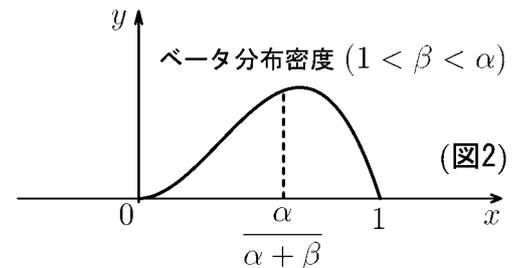
図 1 は $n = 5$ の場合の密度関数の図である。



< ベータ分布 >

正定数 α, β に対し, 関数

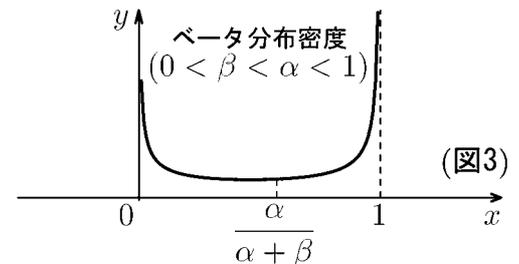
$$p(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & : 0 < x < 1 \\ 0 & : \text{その他} \end{cases}$$



を密度とする分布を形状母数 (α, β) の **ベータ**

分布 という。ただし $B(\alpha, \beta)$ はベータ 関数

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$



である。ベータ分布の平均は $\frac{\alpha}{\alpha + \beta}$ であり, 分散は $\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$ である。

図 2 は $\alpha = 3, \beta = 2$ の場合のグラフであり, 図 3 は $\alpha = 0.4, \beta = 0.3$ の場合の

グラフである。 $\alpha = \beta = 1$ の場合は一様分布になる。

定理 7 正定数 α, β, λ に対し, X を $\text{Gamma}(\alpha, \lambda)$ に従う確率変数, Y を $\text{Gamma}(\beta, \lambda)$

に従う確率変数で, X と Y は独立する。このとき $Z = \frac{X}{X+Y}$ の分布は形状母数

(α, β) のベータ分布である。

< t 分布, F 分布 >

< t 分布 > 正数 n に対し, 関数

$$t_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad (x \in \mathbf{R})$$

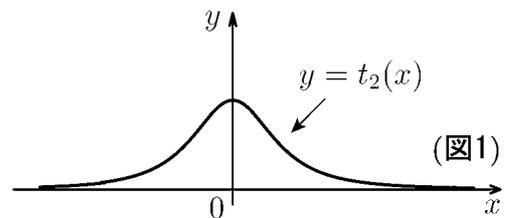
を密度とする分布を, **自由度 n の t 分布** という。 $n=1$ のときは *Cauchy* 分布 (コーシー分布) といい, 平均は存在しない。 $n > 1$ のとき平均は 0 である。 $n \leq 2$ のとき分散は存在しない。 $n > 2$ のとき分散は $\frac{n}{n-2}$ である。

定理 8 $\lim_{n \rightarrow \infty} t_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

定理 9 X が正規分布 $N(0, 1)$ に従い, Y が自由度 n の χ^2 分布に従う確率変数で, X と Y が独立ならば

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

は自由度 n の t 分布に従う。



系 X_1, X_2, \dots, X_n は独立で正規分布 $N(m, \sigma^2)$ に従うとき,

$$\frac{\sqrt{n}(\bar{X} - m)}{\sqrt{\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2}}$$

は自由度 $n-1$ の t 分布に従う。ただし $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ である。

< F 分布 > 正の整数 m, n に対して, 関数

$$F(x) = \frac{n^{\frac{n}{2}} m^{\frac{m}{2}} x^{\frac{n}{2}-1}}{B\left(\frac{n}{2}, \frac{m}{2}\right) (nx + m)^{\frac{n+m}{2}}} \quad (x > 0)$$

を密度とする分布を, **自由度 (n, m) の F 分布** という。 $n > 2$ のとき平均は $\frac{n}{n-2}$,

$n > 4$ のとき分散は $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ である。

定理 10 X と Y が独立で, それぞれ自由度 n, m の χ^2 分布に従うとき,

$$Z = \frac{\frac{X}{n}}{\frac{Y}{m}}$$

は自由度 (n, m) の F 分布に従う。

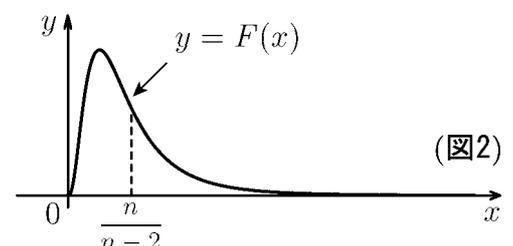


図 2 は $n=8, m=10$ の場合の $y = F(x)$ のグラフである。

< 2次元正規分布 1 >

定理11 X, Y は標準正規分布 $N(0, 1)$ に従い, X と Y が独立ならば

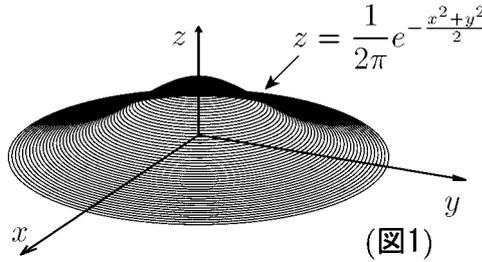
$$P((X, Y) \in A \times B) = \iint_{A \times B} \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} dx dy \quad (A, B \in \mathbf{R})$$

となる。このとき「 (X, Y) は **2次元標準正規分布** に従う」という。

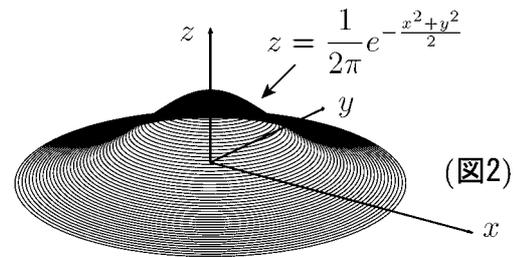
(注) 図1と図2は
この密度関数

$$z = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

が表す曲面である。



(図1)



(図2)

定理12 (U, V) を2次元標準正規分布に従うとする。定数 m_1, m_2, a, b, c, d ($ad - bc \neq 0$) に対して

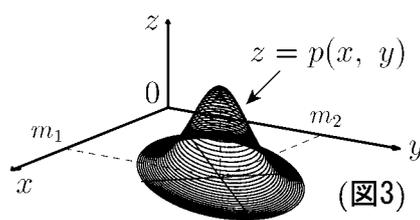
$$X = aU + cV + m_1, \quad Y = bU + dV + m_2$$

とおくと, (X, Y) の分布 $P((X, Y) \in A \times B) = \iint_{A \times B} p(x, y) dx dy$ の密度関数 $p(x, y)$ は

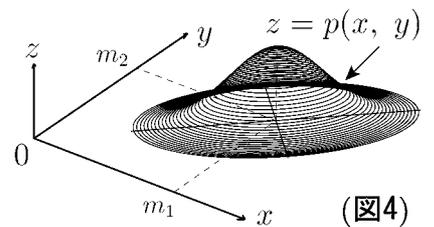
$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{x-m_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-m_1}{\sigma_1}\right)\left(\frac{y-m_2}{\sigma_2}\right) + \left(\frac{y-m_2}{\sigma_2}\right)^2 \right\}}$$

となる。ここで $\sigma_1 = \sqrt{a^2 + c^2}$, $\sigma_2 = \sqrt{b^2 + d^2}$, $\rho = \frac{ab + cd}{\sigma_1\sigma_2}$ である。

(注1) 図3と図4が
 $m_1 = 2, m_2 = 2,$
 $a = 0.4, b = 0.4,$
 $c = -0.2, d = 0.2$
の場合の $z = p(x, y)$
の曲面である。



(図3)



(図4)

(注2) 変換

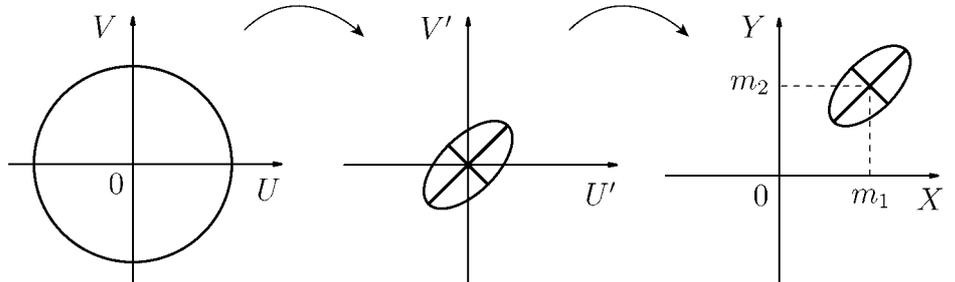
$(U, V) \rightarrow (X, Y)$
を1次変換(回転,
拡大, 縮小)と平行
移動に分けると右
図のようになる。

(1次変換)

$$\begin{pmatrix} U' \\ V' \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$$

(平行移動)

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} U' \\ V' \end{pmatrix} + \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$



(注3) 定理12の分
布を一般の**2次元正
規分布**という。

< 2次元正規分布 2 >

定理13 (X, Y) は一般の2次元正規分布(前ページ定理12)に従うとする。すなわち

$$P((X, Y) \in A \times B) = \iint_{A \times B} p(x, y) dx dy,$$

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{Q(x, y)}{2}}$$

$$Q(x, y) = \frac{1}{1-\rho^2} \left\{ \left(\frac{x-m_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x-m_1}{\sigma_1} \right) \left(\frac{y-m_2}{\sigma_2} \right) + \left(\frac{y-m_2}{\sigma_2} \right)^2 \right\}$$

とする。ただし $\sigma_1 > 0$, $\sigma_2 > 0$, $0 < \rho < 1$ である。このとき次式が成立する。

$$E[X] = \iint_{\mathbf{R}^2} xp(x, y) dx dy = m_1 \text{ (} X \text{ の平均)}, \quad E[Y] = \iint_{\mathbf{R}^2} yp(x, y) dx dy = m_2 \text{ (} Y \text{ の平均)}$$

$$V(X) = E[(X - m_1)^2] = \iint_{\mathbf{R}^2} (x - m_1)^2 p(x, y) dx dy = \sigma_1^2 \text{ (} X \text{ の分散)}$$

$$V(Y) = E[(Y - m_2)^2] = \iint_{\mathbf{R}^2} (y - m_2)^2 p(x, y) dx dy = \sigma_2^2 \text{ (} Y \text{ の分散)}$$

$$C_{ov}(X, Y) = E[(X - m_1)(Y - m_2)] = \iint_{\mathbf{R}^2} (x - m_1)(y - m_2)p(x, y) dx dy = \rho\sigma_1\sigma_2 \text{ (共分散)}$$

また X の密度関数を $p_X(x)$, Y の密度関数を $p_Y(y)$ と書くと

$$p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}} : \text{平均 } m_1, \text{ 分散 } \sigma_1^2 \text{ の1次元正規分布密度}$$

$$p_Y(y) = \int_{-\infty}^{\infty} p(x, y) dx = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-m_2)^2}{2\sigma_2^2}} : \text{平均 } m_2, \text{ 分散 } \sigma_2^2 \text{ の1次元正規分布密度}$$

となる。

(注) $p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$ となるのは任意の実数 a, b ($a < b$) に対し

$$\int_a^b p_X(x) dx = P(a < X < b) = P((X, Y) \in (a, b) \times \mathbf{R}) = \iint_{(a, b) \times \mathbf{R}} p(x, y) dx dy = \int_a^b \left\{ \int_{\mathbf{R}} p(x, y) dy \right\} dx$$

が成立するからである。この (X, Y) の分布に対して、 X だけの分布(または Y だけの分布)を**周辺分布**という。

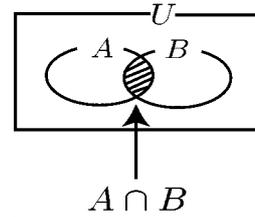
< 条件付確率 1 >

事象 A が起こったとき、事象 B の起こる確率を

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

と定め、 A が起こったとき B の起こる条件付確率

という。この定義より $P(A \cap B) = P(B | A) \times P(A)$ となる。



問1 A と B が独立事象であるとき $P(B | A) = P(B)$ であることを示せ。

例 全部で 100 本のくじの中に当たりが 10 本ある。

最初に A 君が引き、次に B 君が引いた。

A 君が当たりを引く事象を A ,

B 君が当たりを引く事象を B

とする。

$$P(A) = \frac{10}{100} = \frac{1}{10}, \quad P(A \cap B) = \frac{10 \times 9}{100 \times 99} = \frac{1}{10} \times \frac{1}{11} = \frac{1}{110}$$

より A 君が当たった後で B 君の当たる確率は

$$P(B | A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{110}}{\frac{1}{10}} = \frac{1}{11}$$

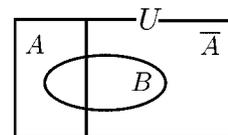
(別解) A 君が当たったとき、残りくじは 99 本で、当たりくじは 9 本残っているから

$$P(B | A) = \frac{9}{99} = \frac{1}{11}$$

問2 上の例で最初に A 君がはずれを引くという事象を \bar{A} とする。

$P(B | \bar{A})$ を求めよ。

問3 上の例で 2 人目の B 君が当たる確率を求めよ。



(ヒント) $P(B) = P(B \cap A) + P(B \cap \bar{A}) = P(B | A)P(A) + P(B | \bar{A})P(\bar{A})$

< 条件付確率 2 >

1. X, Y が離散型確率変数のとき,

「 $X = x$ が起こったとき, $Y = y$ の起こる条件付確率」を

$$P(Y = y | X = x) = \begin{cases} \frac{P(Y = y, X = x)}{P(X = x)} & : P(X = x) > 0 \\ 0 & : P(X = x) = 0 \end{cases}$$

と定める。

2. X, Y が連続型確率変数で

$$X \text{ の密度が } p_X(x) \left(\Leftrightarrow P(X \in A) = \int_A p_X(x) dx \right)$$

$$Y \text{ の密度が } p_Y(y) \left(\Leftrightarrow P(Y \in B) = \int_B p_Y(y) dy \right)$$

$$X \text{ と } Y \text{ の同時分布の密度が } p(x, y) \left(\Leftrightarrow P((X, Y) \in A \times B) = \iint_{A \times B} p(x, y) dx dy \right)$$

であるとき,

「 $X = x$ が起こったとき, $Y = y$ の起こる条件付確率密度」を

$$p(Y = y | X = x) = \begin{cases} \frac{p(x, y)}{p_X(x)} & : p_X(x) > 0 \\ 0 & : p_X(x) = 0 \end{cases}$$

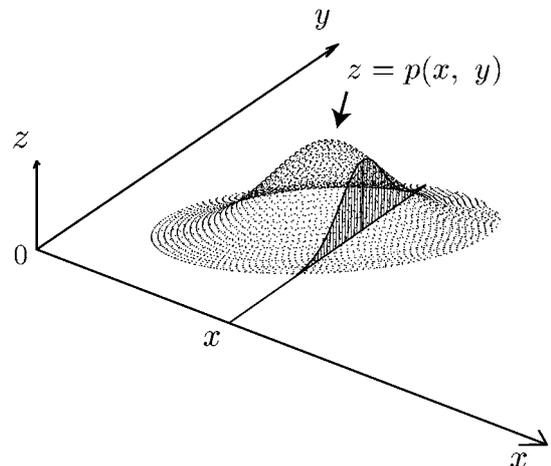
と定める。このように定めると

「 $X = x$ が起こったとき, Y が B に含まれる条件付確率」は

$$\begin{aligned} P(Y \in B | X = x) &= \int_B p(Y = y | X = x) dy \\ &= \int_B \frac{p(x, y)}{p_X(x)} dy = \frac{\int_B p(x, y) dy}{\int_{-\infty}^{\infty} p(x, y) dy} \end{aligned}$$

となる。

$$(注) \quad p_X(x) = \int_{-\infty}^{\infty} p(x, y) dy$$



< 2次元正規分布の周辺分布 >

(X, Y) を一般の 2次元正規分布に従う確率変数とする (p15 定理 2)。

確率密度関数 $p(x, y)$ は

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{Q(x, y)}{2}}$$

$$Q(x, y) = \frac{1}{1-\rho^2} \left\{ \left(\frac{x-m_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x-m_1}{\sigma_1} \right) \left(\frac{y-m_2}{\sigma_2} \right) + \left(\frac{y-m_2}{\sigma_2} \right)^2 \right\}$$

$$= \frac{1}{(1-\rho^2)\sigma_2^2} \left\{ (y-m_2)^2 - 2\rho\frac{\sigma_2}{\sigma_1}(x-m_1)(y-m_2) + \left(\frac{\sigma_2}{\sigma_1} \right)^2 (x-m_1)^2 \right\}$$

$$= \frac{1}{(1-\rho^2)\sigma_2^2} \left[\left\{ (y-m_2) - \rho\frac{\sigma_2}{\sigma_1}(x-m_1) \right\}^2 + (1-\rho^2) \left(\frac{\sigma_2}{\sigma_1} \right)^2 (x-m_1)^2 \right]$$

$$= \frac{1}{(1-\rho^2)\sigma_2^2} \left\{ (y-m_2) - \rho\frac{\sigma_2}{\sigma_1}(x-m_1) \right\}^2 + \frac{(x-m_1)^2}{\sigma_1^2}$$

となる。よって $X = x$ が起こったとき $Y = y$ の起こる条件付確率密度は

$$p(Y = y | X = x) = \frac{p(x, y)}{p_X(x)} = \frac{\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{Q(x, y)}{2}}}{\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-m_1)^2}{2\sigma_1^2}}}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2\sigma_2^2(1-\rho^2)} \{y-m_2-\rho\frac{\sigma_2}{\sigma_1}(x-m_1)\}^2}$$

である。これは平均 $m_2 + \rho\frac{\sigma_2}{\sigma_1}(x-m_1)$, 分散 $\sigma_2^2(1-\rho^2)$ の

1次元正規分布密度である。従って $X = x$ が起こったとき, Y の平均は

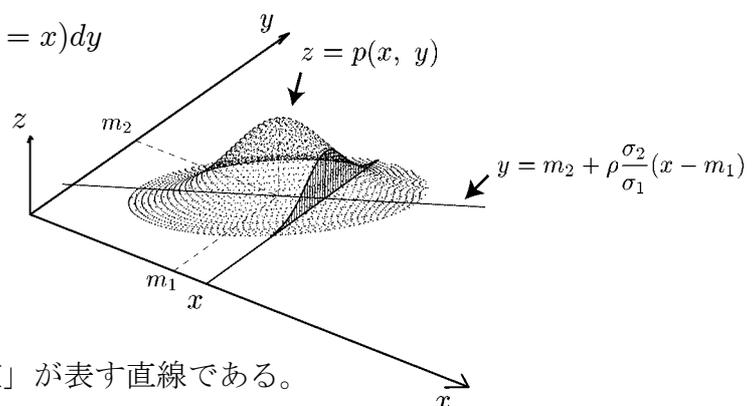
$$E[Y | X = x] = \int_{-\infty}^{\infty} yp(Y = y | X = x)dy$$

$$= m_2 + \rho\frac{\sigma_2}{\sigma_1}(x-m_1)$$

となる。この直線

$$y = m_2 + \rho\frac{\sigma_2}{\sigma_1}(x-m_1)$$

は「 $X = x$ のときの Y の条件付平均値」が表す直線である。



< 2次元正規分布に従うデータと回帰直線 1 >

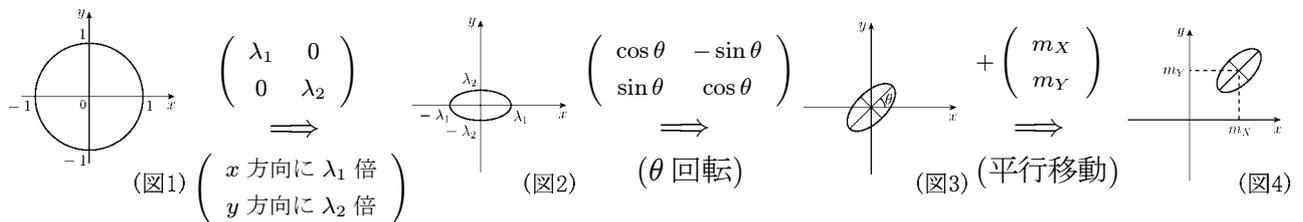
$(U_1, V_1), (U_2, V_2), \dots, (U_n, V_n)$ は独立な 2次元標準正規分布に従う確率変数列とし、2次元データ

$$(*) \quad \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} U_i \\ V_i \end{pmatrix} + \begin{pmatrix} m_X \\ m_Y \end{pmatrix} \quad (i = 1, 2, \dots, n)$$

を考える。この一次変換が、定数 $\lambda_1 > \lambda_2 > 0$, $0 < \theta < \frac{\pi}{2}$ に対して

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 \cos \theta & -\lambda_2 \sin \theta \\ \lambda_1 \sin \theta & \lambda_2 \cos \theta \end{pmatrix}$$

と表されている場合、この変換 (*) は図 1 → 図 2 → 図 3 → 図 4 のようになる。



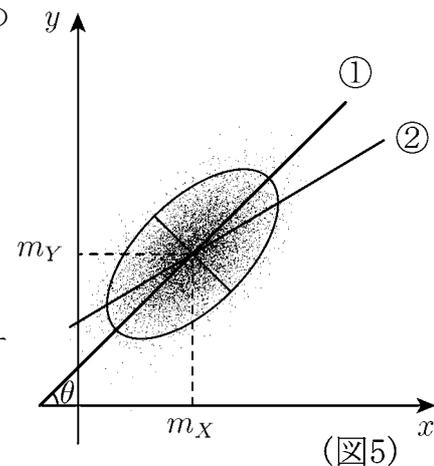
このとき 2次元データ (X_i, Y_i) ($i = 1, 2, \dots, n$) の散布図は図 5 のような (m_X, m_Y) を中心として、中心軸が直線

$$y = (\tan \theta)(x - m_X) + m_Y \cdots \textcircled{1}$$

である楕円の形になる。

図 5 の直線②は前ページで導いた 2次元正規分布の場合の「 $X = x$ のときの Y の条件付平均値」が表す直線

$$y = \rho \frac{\sigma_2}{\sigma_1} (x - m_X) + m_Y \cdots \textcircled{2}$$



である。ただし $\sigma_1 = \sqrt{a^2 + c^2}$, $\sigma_2 = \sqrt{b^2 + d^2}$, $\rho = \frac{ab + cd}{\sigma_1 \sigma_2}$ である。このとき、次が成り立つ。

① n が十分大きいとき、直線①は 2次元データ (X_i, Y_i) ($i = 1, 2, \dots, n$) の直交回帰直線とほぼ一致する。

② n が十分大きいとき、直線②は 2次元データ (X_i, Y_i) ($i = 1, 2, \dots, n$) の回帰直線とほぼ一致する。

< 2次元正規分布に従うデータと回帰直線 2 >

前ページの性質 [2] を示す。

各 (X_i, Y_i) は 2次元正規分布に従い、その密度関数は

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-m_X}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-m_X}{\sigma_1}\right)\left(\frac{y-m_Y}{\sigma_2}\right) + \left(\frac{y-m_Y}{\sigma_2}\right)^2\right\}}$$

とする。これより P.16 定理 3 から

$$E[X_i] = m_X, \quad E[Y_i] = m_Y, \quad V(X_i) = E[(X_i - m_X)^2] = \sigma_1^2$$

$$V(Y_i) = E[(Y_i - m_Y)^2] = \sigma_2^2, \quad Cov(X_i, Y_i) = E[(X_i - m_X)(Y_i - m_Y)] = \rho\sigma_1\sigma_2$$

となる。一方、2次元データ (X_i, Y_i) ($1 \leq i \leq n$) の統計量は

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_{xx} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$S_{yy} = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad S_{xy} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

である。それらは確率変数であるから、その平均をとると

$$E[\bar{X}] = m_X, \quad E[\bar{Y}] = m_Y, \quad E[S_{xx}] = \frac{n-1}{n}\sigma_1^2$$

$$E[S_{yy}] = \frac{n-1}{n}\sigma_2^2, \quad E[S_{xy}] = \frac{n-1}{n}\rho\sigma_1\sigma_2$$

となる。(証明は不偏分散の項です。) 大数の法則より、 $n \rightarrow \infty$ のとき平均に近づくので、 n が十分大きければ

$$\bar{X} \doteq m_X, \quad \bar{Y} \doteq m_Y, \quad S_{xx} \doteq \sigma_1^2, \quad S_{yy} \doteq \sigma_2^2, \quad S_{xy} \doteq \rho\sigma_1\sigma_2$$

とみなせる。

一方、データ (X_i, Y_i) の回帰直線の方程式は

$$y = \frac{S_{xy}}{S_{xx}}(x - \bar{X}) + \bar{Y} \quad (\text{データの回帰直線})$$

であるが、傾きは $\frac{S_{xy}}{S_{xx}} \doteq \frac{\rho\sigma_1\sigma_2}{\sigma_1^2} = \rho\frac{\sigma_2}{\sigma_1}$ であり $\bar{X} \doteq m_X, \bar{Y} \doteq m_Y$ より

この直線は

$$y = \rho\frac{\sigma_2}{\sigma_1}(x - m_X) + m_Y \quad (\text{前ページ直線②})$$

で近似できるので、性質 [2] が示された。

< 2次元正規分布に従うデータと回帰直線 3 >

20 ページの性質 **1** を示す。

2次元データ (X_i, Y_i) ($1 \leq i \leq n$) の直交回帰直線の方程式は

$$(1) \quad y = m(x - \bar{X}) + \bar{Y} \quad \dots\dots \quad (\text{直交回帰直線})$$

$$m = \frac{S_{yy} - S_{xy} + \sqrt{(S_{xx} - S_{xy})^2 + 4(S_{xy})^2}}{2S_{xy}}$$

である。大数の法則より $\bar{X} \doteq m_X$, $\bar{Y} \doteq m_Y$, $S_{xx} \doteq \sigma_1^2$, $S_{yy} \doteq \sigma_2^2$, $S_{xy} \doteq \rho\sigma_1\sigma_2$ であるから直線 (1) は

$$(1)' \quad y = m'(x - m_X) + m_Y$$

$$m' = \frac{\sigma_1^2 - \sigma_2^2 + \sqrt{(\sigma_1^2 - \sigma_2^2)^2 + 4(\rho\sigma_1\sigma_2)^2}}{2\rho\sigma_1\sigma_2}$$

で近似できる。ここで 20 ページの場合 $\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} \lambda_1 \cos \theta & -\lambda_2 \sin \theta \\ \lambda_1 \sin \theta & \lambda_2 \cos \theta \end{pmatrix}$ であるから

$$\sigma_1^2 = a^2 + c^2 = \lambda_1^2 \cos^2 \theta + \lambda_2^2 \sin^2 \theta, \quad \sigma_2^2 = b^2 + d^2 = \lambda_1^2 \sin^2 \theta + \lambda_2^2 \cos^2 \theta$$

$$\rho\sigma_1\sigma_2 = ab + cd = \lambda_1^2 \cos \theta \sin \theta - \lambda_2^2 \sin \theta \cos \theta$$

より傾き m' は

$$m' = \frac{(\lambda_1^2 - \lambda_2^2)(\sin^2 \theta - \cos^2 \theta) + \sqrt{(\lambda_1^2 - \lambda_2^2)^2}}{2(\lambda_1^2 - \lambda_2^2) \sin \theta \cos \theta} = \frac{\sin^2 \theta - \cos^2 \theta + 1}{2 \sin \theta \cos \theta} = \frac{\sin \theta}{\cos \theta} = \tan \theta$$

であるから、直線 (1)' は 20 ページの直線①に等しい。

よって性質 **1** が示された。

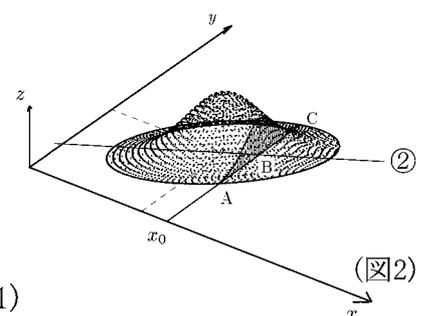
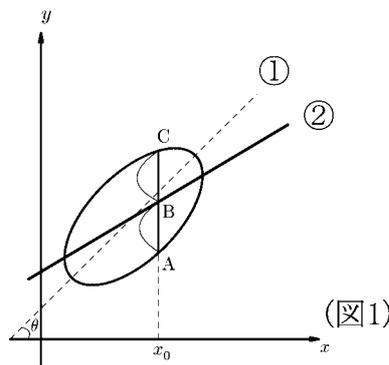
(注 1) 20 ページの直線②の傾きは ($\lambda_1 > \lambda_2 > 0$, $\sin \theta > 0$, $\cos \theta > 0$ より)

$$\rho \frac{\sigma_2}{\sigma_1} = \frac{ab + cd}{a^2 + c^2} = \frac{\lambda_1^2 \sin \theta \cos \theta - \lambda_2^2 \sin \theta \cos \theta}{\lambda_1^2 \cos^2 \theta + \lambda_2^2 \sin^2 \theta} = \frac{\sin \theta}{\cos \theta + \frac{\lambda_2^2}{(\lambda_1^2 - \lambda_2^2) \cos \theta}} < \tan \theta$$

となり、直線①の傾き $\tan \theta$ より小さい。

(注 2) 直線②の傾きは、楕円の中心軸である直線①の傾きより小さい(図 1)

図 2 の 2次元正規分布の曲面に対し、平面 $x = x_0$ の断面は、直線②との交点 B を境にして面積が等しい 2つの部分に分かれる。



< 母集団と標本 >

ある集団を調査したり、ある物質の性質を調べる実験をする場合、あらゆる可能性を全て調査(実験)をする事を**全数調査**と呼ぶことにする。この全数調査をするかわりに、いくつかの標本を抽出し、その標本を調べることによって全体を推測することを**統計的推測**という。推測すべき全体を**母集団**と呼ぶ。

例 1 10000 人の市民の内閣支持率を調べるために 100 人を選んで調査した。このとき 100 人の選び方に 2 通りある。

- ①< 復元抽出 > 10000 人の中から(乱数表などを使って)無作為に 1 人選び、支持するかしないかを聞き、記録する。これを 100 回くり返す。ただし同じ人が 2 回以上選ばれる可能性がある。
- ②< 非復元抽出 > 10000 人の中から無作為に 1 人選び、支持するかしないかを聞き、記録する。次に残りの 9999 人の中から無作為に 1 人選び、同じ事をする。このようにして、一度選んだ人は除外し、残りの中から順に 100 人選ぶ。

この例の場合、無作為に選ぶ事が確率を必要とする理由である。

①の場合 X_k を

$$X_k = \begin{cases} 1: k \text{ 番目の人が支持すれば} 1 \\ 0: \quad \quad \quad \text{支持しなければ} 0 \end{cases} \quad (k = 1, 2, \dots, 100)$$

とすると、 X_k は確率が

$$P(X_k = 1) = \frac{\text{支持する人数}}{10000}, \quad P(X_k = 0) = \frac{\text{支持しない人数}}{10000} \quad (k = 1, 2, \dots, 100)$$

である確率変数と考えられ、しかも X_1, X_2, \dots, X_{100} は独立である。この場合、和 $X_1 + X_2 + \dots + X_{100}$ は成功確率 $p = \frac{\text{支持する人数}}{10000}$ の二項分布 $B(100, p)$ に従っていると考えられる。

②の場合 X_k を①と同様に定めると、和 $X_1 + X_2 + \dots + X_{100}$ は 10000 人から 100 人を非復元抽出で選んだときの超幾何分布 $H(10000, 100, p)$ ($p = \frac{\text{支持する人数}}{10000}$) に従う確率変数と考えられる。

< 復元抽出による標本調査 1 >

例 2 全国の高校 3 年生の男子の身長を調べたい。全国の高校 3 年生の男子全体は N 人いるとして、その身長データを $\{y_1, y_2, \dots, y_N\}$ とする。これを母集団とし、その平均と分散を、それぞれ

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i \quad (\text{母平均}) \quad , \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 \quad (\text{母分散})$$

母平均、母分散という。この母集団全てを調査するのは大変なので、その中から n 人だけ調べて、全体を推測したい。 n 人を無作為に復元抽出し、その身長データを X_1, X_2, \dots, X_n とする。 X_k は k 番目に選んだ男子の身長である。このとき X_k は

$$P(X_k = y_i) = \frac{1}{N} \quad \begin{pmatrix} i = 1, 2, \dots, N \\ k = 1, 2, \dots, n \end{pmatrix}$$

である確率変数とみなせる。この標本 X_1, X_2, \dots, X_n の平均と分散を

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \quad (\text{標本平均}) \quad , \quad S^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2 \quad (\text{標本分散})$$

標本平均、標本分散という。このとき次の定理が成り立つ。

定理 14 (1) 各 $k = 1, 2, \dots, n$ に対し、確率変数 X_k の平均と分散は母平均と母分散に一致する。すなわち

$$E[X_k] = \mu \quad , \quad E[(X_k - \mu)^2] = \sigma^2$$

$$(2) \quad E[\bar{X}] = \mu \quad , \quad E\left[\sum_{k=1}^n (X_k - \bar{X})^2\right] = (n-1)\sigma^2$$

(注 1) この定理より標本分散の平均値は $E[S^2] = \frac{n-1}{n}\sigma^2$ となり母分散 σ^2 と異なる。そのため標本分散のかわりに

$$U = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 \quad (\text{不偏分散})$$

を用いる。この平均は $E[U] = \sigma^2$ となる。 U を**不偏分散**という。

(注 2) 統計では未知数の数を**自由度 (degree of freedom)**という。

例えば n 個の未知数 $\{x_1, x_2, \dots, x_n\}$ の自由度は n だが、

$$\text{その平均 } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ を引いた値} \\ \{x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}\}$$

の自由度は $n-1$ である。

問 $y_k = x_k - \bar{x}$ ($\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$) ($k = 1, 2, \dots, n$) とおくと、 y_n の値は

$\{y_1, y_2, \dots, y_{n-1}\}$ で表されることを示せ。

< 復元抽出による標本調査 2 >

前のページの定理 14 を証明する。

< (1) の証明 >

$$E[X_k] = \sum_{i=1}^N y_i P(X_k = y_i) = \frac{1}{N} \sum_{i=1}^N y_i = \mu \quad (\text{母平均})$$

$$E[(X_k - \mu)^2] = \sum_{i=1}^N (y_i - \mu)^2 P(X_k = y_i) = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 = \sigma^2 \quad (\text{母分散})$$

< (2) の証明 >

$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{k=1}^n X_k\right] = \frac{1}{n} \sum_{k=1}^n E[X_k] = \frac{1}{n} \sum_{k=1}^n \mu = \mu$$

$$\begin{aligned} E\left[\sum_{k=1}^n (X_k - \bar{X})^2\right] &= \sum_{k=1}^n E[(X_k - \mu + \mu - \bar{X})^2] \\ &= \sum_{k=1}^n \{E[(X_k - \mu)^2] + 2E[(X_k - \mu)(\mu - \bar{X})] + E[(\mu - \bar{X})^2]\} \end{aligned}$$

ここで

$$\begin{aligned} E[(X_k - \mu)(\mu - \bar{X})] &= E\left[(X_k - \mu) \left\{\frac{1}{n} \sum_{i=1}^n (\mu - X_i)\right\}\right] \\ \frac{1}{n} \sum_{i=1}^n E[(X_k - \mu)(\mu - X_i)] &= \frac{1}{n} \left\{-E[(X_k - \mu)^2] + \sum_{i \neq k} E[(X_k - \mu)]E[(\mu - X_i)]\right\} \\ &= -\frac{1}{n} E[(X_k - \mu)^2] = -\frac{\sigma^2}{n} \end{aligned}$$

また

$$\begin{aligned} E[(\mu - \bar{X})^2] &= E\left[\left\{\frac{1}{n} \sum_{i=1}^n (\mu - X_i)\right\}^2\right] \\ &= \frac{1}{n^2} \left\{\sum_{i=1}^n E[(\mu - X_i)^2] + \sum_{l \neq i} E[(\mu - X_i)(\mu - X_l)]\right\} = \frac{1}{n^2} \left(\sum_{i=1}^n \sigma^2\right) = \frac{\sigma^2}{n} \end{aligned}$$

となる。よって

$$\begin{aligned} E\left[\sum_{k=1}^n (X_k - \bar{X})^2\right] &= \sum_{k=1}^n \left\{\sigma^2 + 2\left(-\frac{\sigma^2}{n}\right) + \frac{\sigma^2}{n}\right\} = n\sigma^2 \left(1 - \frac{1}{n}\right) \\ &= (n-1)\sigma^2 \end{aligned}$$

(証明終)

< 非復元抽出による標本調査 1 >

例 3 2 ページ例 2 で、 n 人の標本を非復元抽出した場合を考える。すなわち N 人の身長データ $\{y_1, y_2, \dots, y_N\}$ の中から順に n 個のデータ $\{X_1, X_2, \dots, X_n\}$ を (無作為に) 非復元抽出する。このとき N 個のデータ $\{y_1, y_2, \dots, y_N\}$ の中の任意の n 個のデータ $\{y_{i_1}, y_{i_2}, \dots, y_{i_n}\}$ を選ぶ確率は

$$P(X_1 = y_{i_1}, X_2 = y_{i_2}, \dots, X_n = y_{i_n}) = \frac{1}{N(N-1)(N-2)\cdots(N-n+1)}$$

である。このとき、次の定理が成り立つ。

定理 15

$$(1) P(X_k = y_i) = \frac{1}{N} \quad (k = 1, 2, \dots, n, \quad y_i \in \{y_1, y_2, \dots, y_N\})$$

$$(2) P(X_k = y_i, X_\ell = y_j) = \frac{1}{N(N-1)} \quad (k \neq \ell, \quad i \neq j)$$

$$(3) E[X_k] = \mu, \quad E[(X_k - \mu)^2] = \sigma^2 \quad (k = 1, 2, \dots, n)$$

$$(4) E[\bar{X}] = \mu$$

$$(5) E[(\bar{X} - \mu)^2] = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

$$(6) E\left[\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2\right] = \frac{N}{N-1} \sigma^2$$

ただし $\mu = \frac{1}{N} \sum_{i=1}^N y_i$, $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$, $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ である。

系 $V = \frac{N-1}{N(n-1)} \sum_{k=1}^n (X_k - \bar{X})^2$ とおくと $E[V] = \sigma^2$ が成り立つ。

この $V = \frac{N-1}{N(n-1)} \sum_{k=1}^n (X_k - \bar{X})^2$ を非復元抽出の場合の不偏分散という。

定理 16

$$\begin{aligned} & E[(V - \sigma^2)^2] \\ &= \left(\frac{N-1}{N}\right)^2 \frac{\mu_4}{n} - \frac{n-3}{n(n-1)} \sigma^4 - \frac{N-1}{N^2 n(n-1)} \left\{ n^2 - 6n + 7 + \frac{4(n-2)(n-3)}{N-2} + \frac{6(n-2)(n-3)}{(N-2)(N-3)} \right\} \mu_4 \\ &+ \frac{1}{Nn(n-1)} \left\{ n^2 - 8n + 9 + \frac{(n-2)(n-3)(5N-9)}{(N-2)(N-3)} \right\} \sigma^4 \end{aligned}$$

ただし $\mu_4 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^4$ である。

< 非復元抽出による標本調査 2 >

前ページの定理 15 を証明する。

(1) の証明

$$\begin{aligned} P(X_k = y_i) &= \frac{N-1 P_{n-1}}{N P_n} \\ &= \frac{(N-1)(N-2)\cdots(N-1-(n-1)+1)}{N(N-1)(N-2)\cdots(N-n+1)} = \frac{1}{N} \end{aligned}$$

(2) の証明

$k \neq \ell, i \neq j$ に対し

$$\begin{aligned} P(X_k = y_i, X_\ell = y_j) &= \frac{N-2 P_{n-2}}{N P_n} \\ &= \frac{(N-2)(N-3)\cdots(N-2-(n-2)+1)}{N(N-1)(N-2)(N-3)\cdots(N-n+1)} = \frac{1}{N(N-1)} \end{aligned}$$

(3) の証明

$$\begin{aligned} E[X_k] &= \sum_{i=1}^N y_i P(X_k = y_i) = \frac{1}{N} \sum_{i=1}^N y_i = \mu \\ E[(X_k - \mu)^2] &= \sum_{i=1}^N (y_i - \mu)^2 P(X_k = y_i) = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2 = \sigma^2 \end{aligned}$$

(4) の証明

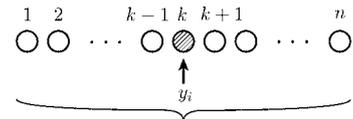
$$E[\bar{X}] = E\left[\frac{1}{n} \sum_{k=1}^n X_k\right] = \frac{1}{n} \sum_{k=1}^n E[X_k] = \frac{1}{n} \sum_{k=1}^n \mu = \mu$$

(5) の証明

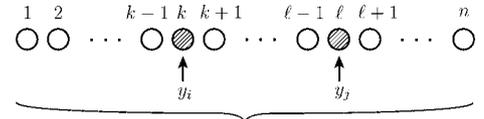
$$\begin{aligned} E[(\bar{X} - \mu)^2] &= E\left[\left(\frac{1}{n} \sum_{k=1}^n X_k - \mu\right)^2\right] = \frac{1}{n^2} E\left[\left\{\sum_{k=1}^n (X_k - \mu)\right\}^2\right] \\ &= \frac{1}{n^2} E\left[\sum_{k=1}^n (X_k - \mu)^2 + \sum_{k \neq \ell} (X_k - \mu)(X_\ell - \mu)\right] \end{aligned}$$

ここで $k \neq \ell$ に対し

$$\begin{aligned} E[(X_k - \mu)(X_\ell - \mu)] &= \sum_{i \neq j} (y_i - \mu)(y_j - \mu) P(X_k = y_i, X_\ell = y_j) \\ &= \frac{1}{N(N-1)} \sum_{i \neq j} (y_i - \mu)(y_j - \mu) = \frac{1}{N(N-1)} \left[\underbrace{\left\{\sum_{i=1}^N (y_i - \mu)\right\}^2}_{=0} - \sum_{i=1}^N (y_i - \mu)^2 \right] = -\frac{\sigma^2}{N-1} \end{aligned}$$



y_i 以外の $N-1$ 個の $\{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_N\}$ から $n-1$ 個選んで 1 列に並べる順列は ${}_{N-1}P_{n-1}$ 通り



y_i, y_j 以外の $N-2$ 個の $\{y\}$ から $n-2$ 個選んで 1 列に並べる順列は ${}_{N-2}P_{n-2}$ 通り

< 非復元抽出による標本調査 3 >

(5) の証明の続き 前ページより

$$\begin{aligned}
 E[(\bar{X} - \mu)^2] &= \frac{1}{n^2} \left\{ \sum_{k=1}^n E[(X_k - \mu)^2] + \sum_{1 \leq k \neq \ell \leq n} E[(X_k - \mu)(X_\ell - \mu)] \right\} \\
 &= \frac{1}{n^2} \left\{ n\sigma^2 + n(n-1) \times \left(-\frac{\sigma^2}{N-1} \right) \right\} = \frac{\sigma^2}{n(N-1)} \{ (N-1) - (n-1) \} = \frac{N-n}{(N-1)n} \sigma^2
 \end{aligned}$$

(6) の証明

$$\begin{aligned}
 E \left[\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 \right] &= \frac{1}{n-1} E \left[\sum_{k=1}^n X_k^2 - 2 \sum_{k=1}^n X_k \bar{X} + n\bar{X}^2 \right] \\
 &= \frac{1}{n-1} E \left[\sum_{k=1}^n X_k^2 - n\bar{X}^2 \right] = \frac{1}{n-1} E \left[\sum_{k=1}^n \{ (X_k - \mu) + \mu \}^2 - n \{ (\bar{X} - \mu) + \mu \}^2 \right] \\
 &= \frac{1}{n-1} \left\{ \sum_{k=1}^n E[(X_k - \mu)^2 - 2(X_k - \mu)\mu + \mu^2] - nE[(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\mu + \mu^2] \right\} \\
 &= \frac{1}{n-1} \left\{ \sum_{k=1}^n E[(X_k - \mu)^2] + n\mu^2 - n(E[(\bar{X} - \mu)^2] + \mu^2) \right\} \\
 &= \frac{1}{n-1} \left\{ \sum_{k=1}^n \sigma^2 - n \times \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \right\} = \frac{\sigma^2}{n-1} \left\{ n - \frac{N-n}{N-1} \right\} = \frac{N}{N-1} \sigma^2
 \end{aligned}$$

(証明終)

例題 (不偏分散の応用)

かんづめ 50 個の中から 5 個を無作為に抽出し、重さを測ったら

98, 101, 100, 98, 103 (単位はグラム)

であった。次の各場合に 50 個全体の平均と分散を推測せよ。

(1) 復元抽出の場合

(2) 非復元抽出の場合

(解) (1) 定理 14 より $\bar{X} = \frac{1}{5} \{98 + 101 + 100 + 98 + 103\} = 100$ (標本平均)

$$\begin{aligned}
 U &= \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 \\
 &= \frac{1}{5-1} \{ (98-100)^2 + (101-100)^2 + (100-100)^2 + (98-100)^2 + (103-100)^2 \} \\
 &= \frac{9}{2} = 4.5 \quad (\text{不偏分散})
 \end{aligned}$$

(2) 定理 15 より $\bar{X} = \frac{1}{5} \{98 + 101 + 100 + 98 + 103\} = 100$ (標本平均)

$$V = \frac{N-1}{N(n-1)} \sum_{k=1}^5 (X_k - \bar{X})^2 = \frac{49}{50 \times 4} \times 18 = 4.41 \quad (\text{非復元抽出のときの不偏分散})$$

< 母平均の区間推定 1 >

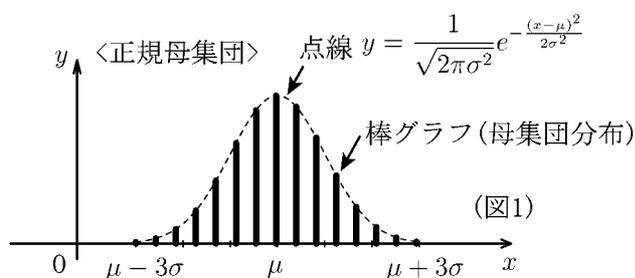
前ページの例題のように母集団の平均や分散を1つの数値で推定する方法を**点推定**という。この点推定の場合、本当の母集団の平均値(また分散値)と推定値との間には誤差がある。

母集団の分布型がわかっている場合や、標本数が多い場合には、もう少し精密な推定ができる。本当の母集団の値と推定値との間の誤差の評価ができる。例えば「95%の確率で平均値は98.4と102.6との間にある」というように、求めるべき母集団の値が含まれる区間を示す幅をつけた推定法を**区間推定**という。いくつかの場合に母平均の区間推定をする。

まず母集団が正規分布で近似できる場合を考える。例えば、全国の満18歳男子の身長の測定値のつくる集団とか、工業製品の寸法を多数計測したときの測定値の作る集団などである。このように、過去の調査結果などから母集団分布(図1の棒グラフ)が正規分布曲線

$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ (図1の点線) で近似できる場合、この集団を**正規母集団** $N(\mu, \sigma^2)$ という。

この場合、母平均は μ 、母分散は σ^2 である。



X_1, X_2, \dots, X_n を正規母集団 $N(\mu, \sigma^2)$ からの標本とすると、標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ は正規分布 $N(\mu, \frac{\sigma^2}{n})$ に従うので、その標準化 $\bar{X}^* = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ は標準正規分布に従うから

$$P\left(|\bar{X} - \mu| < \frac{\sigma}{\sqrt{n}}\alpha\right) = P\left(|\bar{X}^*| < \alpha\right) = \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = f(\alpha)$$

が成り立つ。よって

$$\text{確率 } p = f(\alpha) \text{ で } \boxed{\bar{X} - \frac{\sigma\alpha}{\sqrt{n}} < \mu < \bar{X} + \frac{\sigma\alpha}{\sqrt{n}}} \cdots (*)$$

であると推定できる。(*) を確率 $p = f(\alpha)$ の信頼区間という。

< 母平均の区間推定 2 >

例 ある大都市の満 17 歳の男子の平均身長 μ (cm) を求めるのに、50 名の標本調査を無作為復元抽出により行った。そのデータを X_1, X_2, \dots, X_{50} としたとき、平均と標準偏差は

$$\bar{X} = \frac{1}{50} \sum_{i=1}^{50} X_i = 169.5(\text{cm}) \quad : \quad \text{標本平均}$$

$$s = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (X_i - 169.5)^2} = 5.6 \quad : \quad \text{標本標準偏差}$$

であった。過去のデータからみて、この都市における 17 歳男子の身長分布は正規分布 $N(\mu, \sigma^2)$ で近似できる。また標本の大きさ (= 50) が大きいので、 σ^2 を $s^2 = 5.6^2$ で代用する。正規分布の性質より、平均 $\bar{X} = \frac{1}{50} \sum_{i=1}^{50} X_i$ は正規分布 $N(\mu, \frac{5.6^2}{50})$ に従うので、前ページより、 $n = 50$, $\alpha = 1.96$ のとき

$$P\left(|\bar{X} - \mu| \leq \frac{\sigma}{\sqrt{50}} \times 1.96\right) = \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.95 (= 95\%)$$

であるから、 $\bar{X} = 169.5$, $\sigma = 5.6$ を代入すると、確率 $p = 0.95$ で

$$\begin{aligned} |169.5 - \mu| \leq \frac{5.6}{\sqrt{50}} \times 1.96 &\Leftrightarrow 169.5 - \frac{5.6}{\sqrt{50}} \times 1.96 \leq \mu \leq 169.5 + \frac{5.6}{\sqrt{50}} \times 1.96 \\ &\Leftrightarrow \boxed{167.95 \leq \mu \leq 171.05} \quad \dots (*) \end{aligned}$$

と推定できる。(*) が確率 95% の信頼区間である。

問 付録の正規分布表を見て、この例の場合に、確率 99% の信頼区間を求めよ。

(注) 95% の確率で区間推定を行うとき、95% を**信頼度**という。

一般に正規母集団 $N(\mu, \sigma^2)$ からの標本数 n が大きいとき

信頼度 $f(\alpha) = \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ の信頼区間は

$$\boxed{\bar{X} - \alpha \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + \alpha \times \frac{s}{\sqrt{n}}} \quad \dots (*)'$$

である。ここで $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$ は標本標準偏差である。

< 母平均の区間推定 3 >

正規母集団からの標本数 n が小さいときは、7 ページ (*) 式や 8 ページ (*)' 式は適用できない。この場合は t -分布を使う。

正規母集団 $N(\mu, \sigma^2)$ から無作為復元抽出により n 個の標本 X_1, X_2, \dots, X_n を得た。その標本平均 \bar{X} と不偏分散 U を

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad U = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

とする。各 X_i は $N(\mu, \sigma^2)$ に従う確率変数であり、 $\{X_1, X_2, \dots, X_n\}$ は独立だから、正規分布の性質によって \bar{X} と U は独立になり、

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{U}{n}}} \quad (T\text{-統計量})$$

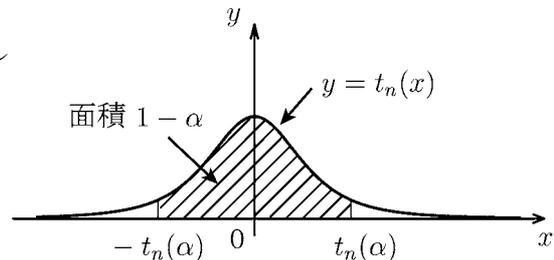
は自由度 $n-1$ の t -分布に従う。すなわち

$$P(a < T < b) = \int_a^b t_{n-1}(x) dx,$$

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

が成り立つ。ここで $0 < \alpha < 1$ なる α に対し

$$\int_{-k}^k t_n(x) dx = 1 - \alpha$$



をみたす正数 k を $t_n(\alpha)$ とおく。

T の密度関数が $t_{n-1}(x)$ だから

$$P\left(\left|\frac{\bar{X} - \mu}{\sqrt{\frac{U}{n}}}\right| \leq t_{n-1}(\alpha)\right) = 1 - \alpha$$

が成り立つ。よって

$$\text{確率 } 1 - \alpha \text{ で } \boxed{\bar{X} - t_{n-1}(\alpha)\sqrt{\frac{U}{n}} \leq \mu \leq \bar{X} + t_{n-1}(\alpha)\sqrt{\frac{U}{n}}} \quad (**)$$

が成り立つ。(**) が信頼度 $1 - \alpha$ の信頼区間である。

< 母平均の区間推定 4 >

例題 ある都市の満 17 歳の平均身長 μ (cm) を求めたい。母集団は正規母集団とする。標本として 10 人を無作為復元抽出し、その身長を X_1, X_2, \dots, X_{10} とする。今、平均と不偏分散が

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = 169.3 \quad , \quad U = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2 = 6.43$$

であった。 μ の 95% 信頼区間を求めよ。

(解) 標本数 $n = 10$ が小さいので、前ページの(**)式を使う。確率 $1 - \alpha = 0.95$ の信頼区間を求める。 $\alpha = 0.05$ のとき t 分布表より $t_{n-1}(\alpha) = t_9(0.05) = 2.262$ だから、前ページ(**)式より

$$169.3 - 2.262 \times \sqrt{\frac{6.43}{10}} \leq \mu \leq 169.3 + 2.262 \times \sqrt{\frac{6.43}{10}}$$

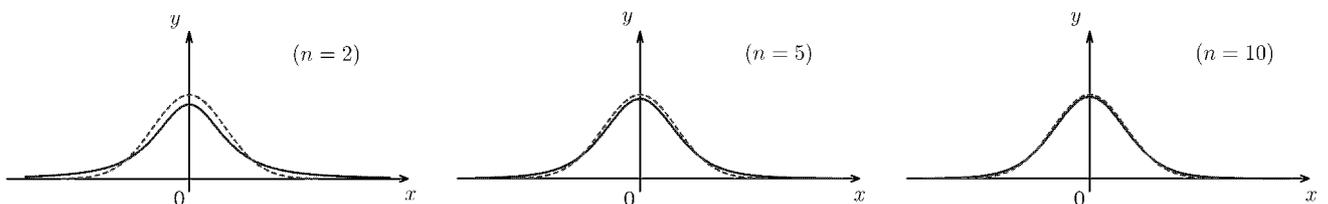
(答) $167.486 \leq \mu \leq 171.114$

問 例題と同じく 10 人の身長を測ったら、その値 X_1, X_2, \dots, X_{10} の平均と不偏分散は

$$\bar{X} = \frac{1}{10} \sum_{i=1}^{10} X_i = 169 \quad , \quad U = \frac{1}{9} \sum_{i=1}^{10} (X_i - \bar{X})^2 = 6.25$$

であった。 t 分布表を用いて、 μ の 99.5% 信頼区間を求めよ。

(注) 標本数 n の大小により、信頼区間の式がちがうが、 n が大きくなると $t_n(x)$ は標準正規分布 $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ に近づくので、 n が大きいときも(**)式を使って良い。 $t_n(x)$ のグラフ(実線)と $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ のグラフ(点線)を比べてほしい。



< 母平均の区間推定 5 >

◎ 母集団の分布型がわからない場合、標本数が小さいと区間推定した場合に信頼性が全くないので、この場合は区間推定できない。

◎ 母集団の分布がわからなくても、標本数が多ければ区間推定できる。

(1) [復元抽出した場合] 中心極限定理より、標本平均 \bar{X} の分布は正規分布に近づくので、30 ページ (*)' が適用できる。

(2) [非復元抽出した場合] 母平均 μ , 母分散 σ^2 , 大きさ N の母集団から n 個の標本 X_1, X_2, \dots, X_n を非復元抽出した場合、標本平均 \bar{X} の確率変数としての平均と分散は定理 15 より

$$E[\bar{X}] = \mu, \quad E[(\bar{X} - \mu)^2] = \frac{N-n}{N-1} \times \frac{\sigma^2}{n}$$

である。さらに N, n が共に大きいとき

(☆) \bar{X} は近似的に正規分布 $N\left(\mu, \frac{N-n}{N-1} \times \frac{\sigma^2}{n}\right)$ に従う

ことが知られている。従って \bar{X} の標準化

$$(\bar{X} - \mu) / \sqrt{\frac{N-n}{N-1} \times \frac{\sigma^2}{n}}$$

は近似的に標準正規分布に従うから

$$P\left(\left|(\bar{X} - \mu) / \sqrt{\frac{N-n}{N-1} \times \frac{\sigma^2}{n}}\right| \leq \alpha\right) \doteq \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx (= f(\alpha))$$

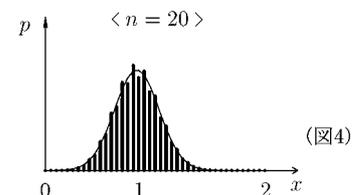
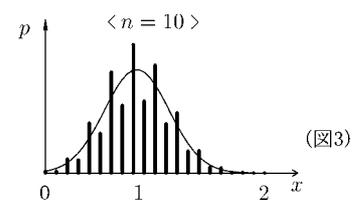
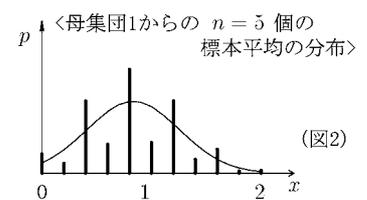
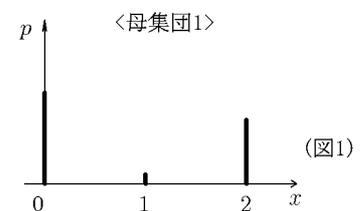
が成り立つ。よって信頼度 $f(\alpha)$ の信頼区間は

$$\bar{X} - \alpha \sqrt{\frac{N-n}{N-1} \times \frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + \alpha \sqrt{\frac{N-n}{N-1} \times \frac{\sigma^2}{n}} \quad (***)$$

なお母分散 σ^2 が未知のときは、 N, n が共に大きくなる

と、定理 16 より不偏分散 $V = \frac{N-1}{N(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2$

が σ^2 に近づくので、 σ^2 のかわりに V で代用する。



(注) 右の図は(☆)の例を示している。母集団が図1のように3つの値をもつ集合

$\{0, 0, 0, \dots, 1, 1, \dots, 2, 2, \dots, 2, 2\}$ (0が100個, 1が10個, 2が70個)

の場合、この母集団から無作為に非復元抽出した標本 X_1, X_2, \dots, X_n の平均

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ の分布が図2($n=5$), 図3($n=10$), 図4($n=20$)である。 $n=20$ ぐらい

になると、 \bar{X} の分布はほぼ正規分布といってもよいであろう。

< 母比率の区間推定 1 >

ある集団の内閣支持率を調べる場合、母集団は支持する人と支持しない人に分かれる。この比率 $\frac{\text{支持する人の人数}}{\text{母集団の全人数}} = p$ を **母比率** と呼ぶ。この母比率を求めたい。この母集団から n 人を復元抽出し、

$$X_k = \begin{cases} 1 & : k \text{ 番目に選んだ人が支持者の場合} \\ 0 & : \text{そうでないとき} \end{cases}$$

とおくと、和 $\sum_{k=1}^n X_k$ は 2 項分布 $B(n, p)$ に従う確率変数である。従って平均 $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ は $E[\bar{X}] = p$, $E[(\bar{X} - p)^2] = \frac{p(1-p)}{n}$ であるので、その標準化 $\bar{X}^* = (\bar{X} - p) / \sqrt{p(1-p)/n}$ は近似的に標準正規分布に従う。すなわち

$$P\left(\left|\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq \alpha\right) \doteq \int_{-\alpha}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = f(\alpha)$$

である。従って信頼度 $f(\alpha)$ の信頼区間は

$$\boxed{\bar{X} - \alpha \sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + \alpha \sqrt{\frac{p(1-p)}{n}}} \quad (***)$$

である。この式 (***) は p に関する 2 次不等式であり、これを解くと

$$\boxed{\frac{b - \sqrt{D}}{2a} \leq p \leq \frac{b + \sqrt{D}}{2a}} \quad (***)'$$

$$\left(a = \frac{1}{n} + \frac{1}{\alpha^2}, \quad b = \frac{1}{n} + \frac{2\bar{X}}{\alpha^2}, \quad c = \frac{\bar{X}^2}{\alpha^2}, \quad D = b^2 - 4ac\right)$$

である。

(注) n が十分大きい場合は (***)' のかわりに (***) 式の $p(1-p)$ を $\bar{X}(1-\bar{X})$ で代用した式

$$\bar{X} - \alpha \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + \alpha \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$$

を用いることもある。

< 母比率の区間推定 2 >

大きさ N の母集団の母比率 p を非復元抽出による n 個の標本 X_1, X_2, \dots, X_n によって推定したい。母平均は $\mu = p$, 母分散は $\sigma^2 = p(1-p)$ より 11 ページ (***) 式から、信頼度

$f(\alpha) = 2 \int_0^\alpha \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ の信頼区間は

$$\boxed{\bar{X} - \alpha \sqrt{\frac{N-n}{N-1} \times \frac{p(1-p)}{n}} \leq p \leq \bar{X} + \alpha \sqrt{\frac{N-n}{N-1} \times \frac{p(1-p)}{n}} \quad (*^5)}$$

となる。この p に関する 2 次不等式を解くと

$$\boxed{\frac{b - \sqrt{D}}{2a} \leq p \leq \frac{b + \sqrt{D}}{2a}} \quad (*^5)'$$

$$\left(a = \frac{N-n}{(N-1)n} + \frac{1}{\alpha^2}, \quad b = \frac{N-n}{(N-1)n} + \frac{2\bar{X}}{\alpha^2}, \quad c = \frac{\bar{X}^2}{\alpha^2}, \quad D = b^2 - 4ac \right)$$

である。

例題 10 万人の有権者のいる都市で、内閣支持率 p を調べるのに 2000 人を無作為非復元抽出し、支持者の数を調べたら、2000 人中 800 人が支持者であった。母比率 p の 95% 信頼区間を求めよ。また信頼区間の幅を求めよ。

(解) $\bar{X} = \frac{800}{2000} = 0.4$, $N = 100000$, $n = 2000$, $\alpha = 1.96$ とする。(*⁵)' より

$$\frac{b - \sqrt{D}}{2a} = 0.378952, \quad \frac{b + \sqrt{D}}{2a} = 0.421424 \quad \text{となるから}$$

$$\underline{\text{信頼区間は } 0.378952 \leq p \leq 0.421424}$$

$$\underline{\text{信頼区間の幅は } 0.421424 - 0.378952 = 0.042472}$$

問 10 万人の有権者のいる都市で、内閣支持率 p を調べるのに 3000 人を無作為非復元抽出し、支持者の数を調べたら 3000 人中 1500 人が支持者であった。母比率 p の 95% 信頼区間を求めよ。また信頼区間の幅を求めよ。

< 必要な標本の大きさ >

例題 10万人の有権者のいる都市で、内閣支持率 p を調べるのに標本調査を行った。信頼区間の幅が 0.04 以内になるように、信頼度 95% で p の区間推定をしたい。抽出する有権者の数を何名以上にすれば良いか？

(解) 前ページの非復元抽出の場合の信頼区間の式を使いたいが、 \bar{X} が不明なので(*5)式を使う。

$$\bar{X} - \alpha \sqrt{\frac{N-n}{N-1} \times \frac{p(1-p)}{n}} \leq p \leq \bar{X} + \alpha \sqrt{\frac{N-n}{N-1} \times \frac{p(1-p)}{n}}$$

より

$$|\bar{X} - p| \leq \alpha \sqrt{\frac{N-n}{N-1} \times \frac{p(1-p)}{n}} \quad \dots \quad (1)$$

となる。信頼区間の幅が 0.04 以下になるためには (1) 式の右辺が 0.02 以下になれば良い。すなわち

$$\alpha \sqrt{\frac{N-n}{N-1} \times \frac{p(1-p)}{n}} \leq 0.02 \quad \dots \quad (2)$$

であれば良い。ここで $\alpha = 1.96$ である。一方

$$p(1-p) = -p^2 + p = -\left(p - \frac{1}{2}\right)^2 + \frac{1}{4} \leq \frac{1}{4}$$

だから

$$\alpha \sqrt{\frac{N-n}{N-1} \times \frac{1}{4n}} \leq 0.02 \quad \dots \quad (3)$$

であれば良い。 $\alpha = 1.96$ を代入して両辺を 2 乗すると

$$1.96^2 \times \frac{N-n}{N-1} \times \frac{1}{4n} \leq (0.02)^2$$

$$N-n \leq 4 \left(\frac{0.02}{1.96}\right)^2 (N-1)n$$

$$n \geq \frac{N}{4 \times \left(\frac{0.02}{1.96}\right)^2 (N-1) + 1} = \frac{100000}{4 \times \left(\frac{2}{196}\right)^2 \times 99999 + 1} = 2344.73$$

(答) 2345 人以上を抽出すれば良い。

問 20万人の有権者のいる都市で、内閣支持率 p を調べるのに標本調査を行った。信頼区間の幅が 0.04 以内になるように、信頼度 95% で p の区間指定をしたい。抽出する有権者の数を何名以上にすればよいか？

< ポアソン過程 >

5,6 ページのタクシーの例で、1 時間に平均 λ 回空のタクシーが通る場合、1 時間に通る空のタクシーの台数を X とすると、

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k = 0, 1, 2, \dots)$$

であった。 $t > 0$ に対し、 t 時間に通るタクシーの台数を $X(t)$ とすると、(平均 λt 回通るから)

$$(1) \quad P(X(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad (k = 0, 1, 2, \dots)$$

となる。また「時刻 t までの通過台数 m 」と「時刻 t から後の通過台数」とは独立だから、 $s > 0$ に対して

$$(2) \quad P(X(t+s) - X(t) = k \mid X(t) = m) = P(X(t+s) - X(t) = k) = e^{-\lambda s} \frac{(\lambda s)^k}{k!}$$

となる。(1), (2) を満たす確率過程 $\{X(t) : t > 0\}$ をポアソン過程という。

$\{X(t)\}$ のグラフは右図のように階段形のグラフになる。

今時刻 t までに m 台通過したとき、 $m+1$ 台目が通るまでの待ち時間を ξ とすると、 $s > 0$ に対し

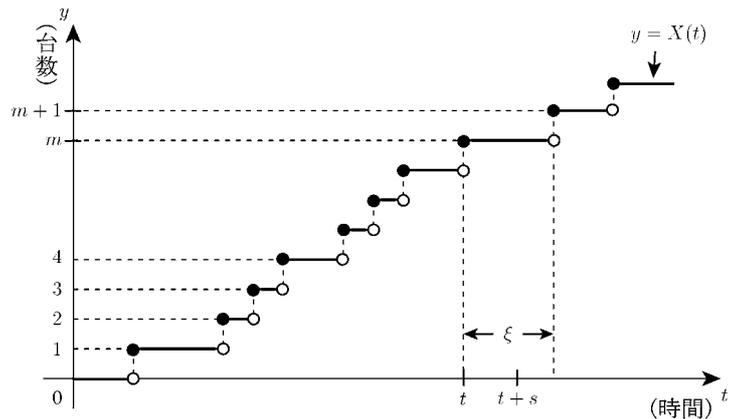
$$\begin{aligned} & P(\xi > s \mid X(t) = m) \\ &= P(X(t+s) = m \mid X(t) = m) = P(X(t+s) - X(t) = 0 \mid X(t) = m) \\ &= e^{-\lambda s} \frac{(\lambda s)^0}{0!} = e^{-\lambda s} = \int_s^{\infty} \lambda e^{-\lambda x} dx \end{aligned}$$

より、 ξ は密度 $\lambda e^{-\lambda x}$ の指数分布に従う。この待ち時間の平均は

$$(3) \quad \boxed{E[\xi] = \frac{1}{\lambda}}$$

である。

問 (3) 式を証明せよ。



< ブラウン運動 >

1827年植物学者の Brown は水に浮かぶ花粉の粒子が不思議な動きをしているのを発見した。花粉の粒子は突然にある一定距離移動し、止まる。これを繰り返し、ジグザグな動きをする。花粉以外でも、水に浮かぶ微粒子ならば同様の動きをすることがわかった。この運動する粒子をブラウン粒子と呼び、その運動を「ブラウン運動」と呼ぶ。

1905年アインシュタイン (Einstein) はブラウン粒子の時刻 t 、位置 x に存在する確率密度 $u(t, x)$ が熱方程式

$$\frac{\partial}{\partial t} u(t, x) = D \frac{\partial^2}{\partial x^2} u(t, x) \quad (t \geq 0, x \in \mathbf{R})$$

を満たすことを示した。ここで D は拡散係数と呼ばれる正の定数である。なお空間が 3 次元の場合は $\frac{\partial u}{\partial t} = D \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} \right)$ となる。

1923年ウィナー (Wiener) はブラウン運動の数学的モデルを次の $\boxed{1}, \boxed{2}, \boxed{3}$ をみたす確率過程 $\{B(t) : t \geq 0\}$ として提案した。

$\boxed{1}$ $\{B(t) : t \geq 0\}$ は t の関数として連続である。

$\boxed{2}$ [独立増分性]

$\left(\begin{array}{l} \text{任意の時間分点 } 0 \leq t_0 < t_1 < t_2 < \cdots < t_n \text{ に対して、} \\ B(t_1) - B(t_0), B(t_2) - B(t_1), \cdots, B(t_n) - B(t_{n-1}) \text{ は独立} \end{array} \right)$

$\boxed{3}$ [Gauss 分布 (正規分布)]

任意の $0 \leq s < t$ に対して、

$$P(B(t) \in A \mid B(s) = a) = \int_A \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{(x-a)^2}{2(t-s)}} dx \quad (a \in \mathbf{R}, A \subset \mathbf{R})$$

が成り立つ。

$\boxed{1}, \boxed{2}, \boxed{3}$ を満たす確率過程 $\{B(t) : t \geq 0\}$ を「ブラウン運動」または「Wiener 過程」と言う。なお空間が 3 次元の場合は $\boxed{3}$ のかわりに

$$\boxed{3'} \quad P(B(t) \in A \mid B(s) = a) = \iiint_A (2\pi(t-s))^{-\frac{3}{2}} e^{-\frac{|x-a|^2}{2(t-s)}} dx_1 dx_2 dx_3$$

$$(x = (x_1, x_2, x_3) \in \mathbf{R}^3, a \in \mathbf{R}^3, A \subset \mathbf{R}^3)$$

を用いる。

(注) $\boxed{1}, \boxed{2}, \boxed{3}$ を満たす $\{B(t)\}$ の存在確率密度 $u(t, x) \left(P(B(t) \in A) = \int_A u(t, x) dx \right)$ は $D = \frac{1}{2}$ の場合の熱方程式の解である。

< Wiener 過程 >

ウィナー (Wiener) はブラウン運動 $\{B(t) : 0 \leq t \leq \pi\}$ を次のように構成した。

< Wiener によるブラウン運動の構成 >

$X_0, X_1, X_2, \dots, X_n, \dots$ は独立同分布な確率変数列で、全て標準正規分布 $N(0, 1)$ に従う。すなわち $P(X_n \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ ($A \subset \mathbf{R}$) が成り立つ。このとき

$$(*) \quad B(t) = \frac{t}{\sqrt{\pi}} X_0 + \sum_{n=1}^{\infty} \sqrt{\frac{2}{\pi}} X_n \frac{\sin(nt)}{n} \quad (0 \leq t \leq \pi)$$

と置くと、確率過程 $\{B(t) : 0 \leq t \leq \pi\}$ は前ページの性質 [1](#), [2](#), [3](#) を満たす。Wiener はブラウン運動 $B(t)$ をフーリエ級数展開 (*) によって構成した。

(注) ブラウン運動 $\{B(t)\}$ の時間微分 $\dot{B}(t)$ を「ホワイト・ノイズ」と呼ぶのは (*) 式を形式的に時間微分すると

$$(*)' \quad \dot{B}(t) = \frac{1}{\sqrt{\pi}} X_0 + \sum_{n=1}^{\infty} \sqrt{\frac{2}{\pi}} X_n \cos(nt)$$

となるからである。周期 $\frac{2\pi}{n}$ (= 周波数 $\frac{n}{2\pi}$) の波 $\cos(nt)$ の強度 $\sqrt{\frac{2}{\pi}} |X_n|$ の平均は $E \left[\sqrt{\frac{2}{\pi}} |X_n| \right] = \frac{2}{\pi}$ で n に無関係に一定である。全ての周波数の強度が一定なスペクトルをもつ可視光が白色になるから「ホワイト」と呼ばれる。また「ノイズ」(雑音) の由来は、ブラウン運動 $\{B(t)\}$ がジグザグな動きであり、その時間微分は各瞬間ごとにまったくでたらめな値をとるランダムな過程であり、このような波形を持つ音は純粋な雑音として聞こえるからである。ただし (*)' 式の右辺は収束しないので、数学者はこの式 (*)' を使わない。

Wiener のブラウン運動は P. Levi 等によってその数学的性質が詳しく研究された。さらに 1951 年にドンスカー (Donsker) はランダム・ウォークからブラウン運動を構成した。次ページからその話をする。

< 乱歩 >

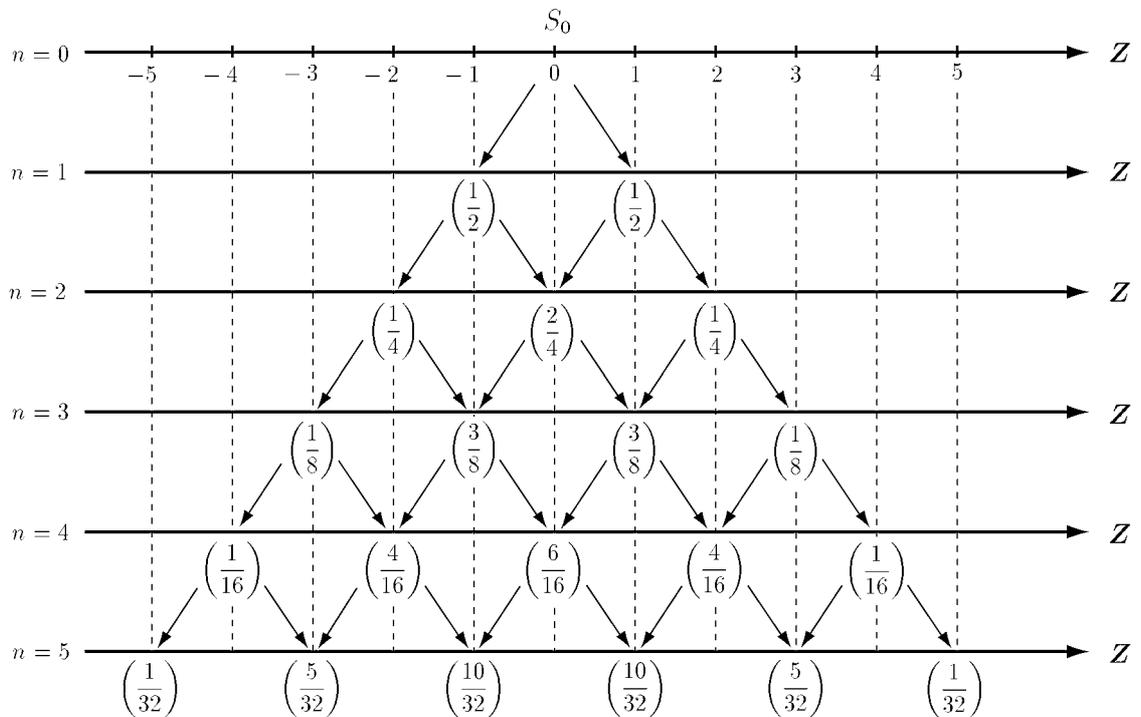
1次元乱歩 (random walk) というのは1次元の格子点上を動く粒子の運動である。コイン投げを考える。表が出れば一步進む。裏が出れば一步後退する。このコイン投げを無限回繰り返す。 $k \geq 1$ に対し

$$X_k = \begin{cases} 1 & : k \text{ 回目に表が出たとき} \\ -1 & : k \text{ 回目に裏が出たとき} \end{cases}$$

として、

$$S_n = X_1 + X_2 + \cdots + X_n, \quad S_0 = 0$$

とおく。 $\{S_n\}$ を0から出発し、整数 Z 上を動く乱歩 (ランダム・ウォーク) という。 $n = 0$ から $n = 5$ までの $\{S_n\}$ の動く可能性のある道を図示すると下図のようになる。



Z 上の点 j の下の () 内は $P(S_n = j)$ の確率である。

$$P(S_{n+1} = j) = \frac{1}{2}P(S_n = j + 1) + \frac{1}{2}P(S_n = j - 1)$$

が成り立つ。

< 乱歩の極限 >

前ページの乱歩 $S_n = \sum_{k=1}^n X_k$ に対し、 X_1, X_2, \dots, X_n は独立で同分布な確率変数で

$$E[X_k] = 0, E[(X_k)^2] = 1 \text{ より、}$$

$$E[S_n] = 0, E[S_n^2] = n$$

が成り立つ。さらに次の定理が成り立つ。

定理 17 任意の正数 $C > 0$ に対し、次の極限式が成立する。

$$(1) \lim_{n \rightarrow \infty} P(|S_n| \leq C) = 0$$

$$(2) \lim_{n \rightarrow \infty} P(|S_n| \leq C \cdot n) = 1 \quad (\text{大数の法則})$$

$$(3) \lim_{n \rightarrow \infty} P(|S_n| \leq C\sqrt{n}) = \int_{-C}^C \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \quad (\text{中心極限定理})$$

(証明) (1) 略

(2) チェビシエフの不等式より

$$P(|S_n| \geq C \cdot n) \leq \frac{1}{C^2 n^2} E[S_n^2] = \frac{1}{C^2 n} \rightarrow 0 \quad (n \rightarrow \infty)$$

(3) S_n の標準化は $S_n^* = \frac{S_n}{\sqrt{n}}$ で、中心極限定理より

$$P(|S_n| \leq C\sqrt{n}) = P(|S_n^*| \leq C) \xrightarrow{(n \rightarrow \infty)} \int_{-C}^C \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

(証明終)

[系] $t > 0$ とする。極限式

$$\lim_{n \rightarrow \infty} P\left(a < \frac{S_{nt}}{\sqrt{n}} < b\right) = \int_a^b \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} dx$$

が任意の定数 $a, b (a < b)$ に対し成立する。

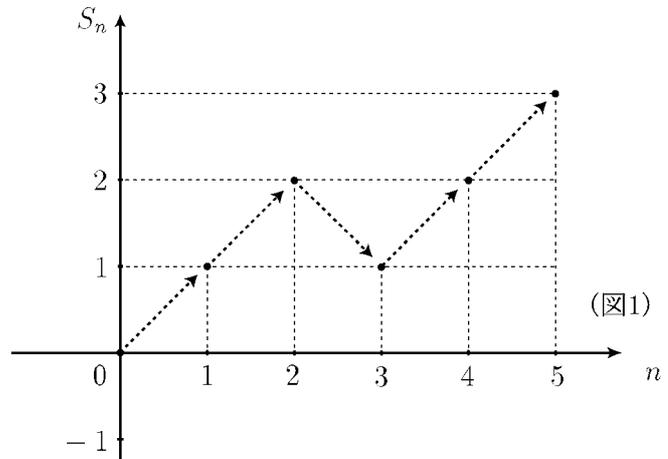
問 中心極限定理を用いて系を証明せよ。

< ランダムウォークによるブラウン運動の構成 >

前ページのランダムウォーク

$$S_n = \sum_{k=1}^n X_k, \quad S_0 = 0$$

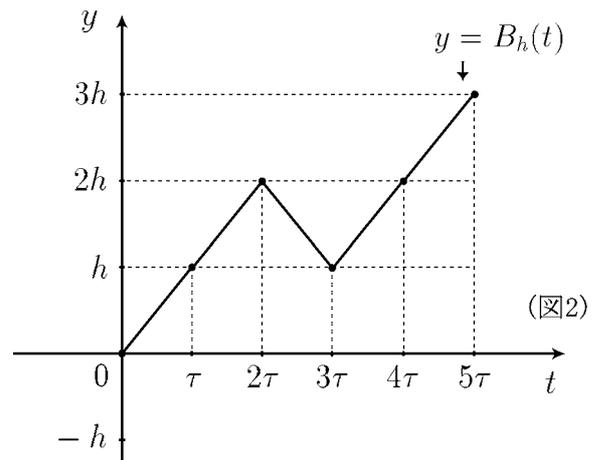
は \mathbf{Z} 上を動く。 $h > 0$ に対して $\tau = h^2$ とおき、 $\{B_h(t) : t > 0\}$ を次式で定義する。



$$B_h(t) = hS_{\lceil \frac{t}{\tau} \rceil} + \left(\frac{t}{\tau} - \left\lceil \frac{t}{\tau} \right\rceil \right) h \left\{ S_{\lceil \frac{t}{\tau} \rceil + 1} - S_{\lceil \frac{t}{\tau} \rceil} \right\}$$

ここで

$\lceil t \rceil = t$ を超えない最大整数 (ガウス記号) である。もし S_n が図1のような動きであれば $B_h(t)$ は図2のような折れ線である。Donsker (1951) は $h \rightarrow 0$ のとき、 $\{B_h(t) : t \geq 0\}$ の極限としてブラウン運動 $\{B(t) : t \geq 0\}$ を構成した。それは関数空間上の確率測度の収束であるが、厳密な定義は省略する。Donsker の定理から、時刻 $t > 0$ を止めたときの存在確率の収束



$$\lim_{h \rightarrow 0} P(a < B_h(t) < b) = \int_a^b \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} dx = P(a < B(t) < b)$$

がわかる。

問 $u(t, x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$ とおくと $\frac{\partial}{\partial t} u(t, x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(t, x)$ が成立することを証明せよ。

< 確率密度の収束 1 >

なめらかな関数 $f(x)$ に対し, 次の極限式が成り立つ。

$$(\textcircled{1}) \quad \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = f'(x) \quad \dots \quad 1 \text{ 階導関数}$$

$$(\textcircled{2}) \quad \lim_{h \rightarrow 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = f''(x) \quad \dots \quad 2 \text{ 階導関数}$$

問 ロピタルの定理を用いて $(\textcircled{2})$ を証明せよ。

この極限式より h が十分小さいとき, 次の近似式が成り立つ。

$$f'(x) \doteq \frac{f(x+h) - f(x)}{h}, \quad f''(x) \doteq \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

< 熱方程式の差分近似 >

$$(1) \quad \frac{\partial}{\partial t} u(t, x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} u(t, x) \quad \dots \quad \text{熱方程式 (拡散方程式)}$$

上記の近似式によって熱方程式 (1) を差分式 (2) になおす。時間の増分を $\Delta t = \tau$, 空間の増分を $\Delta x = h$ とおく。(1) の差分式は

$$(2) \quad \frac{u(t+\tau, x) - u(t, x)}{\tau} = \frac{u(t, x+h) - 2u(t, x) + u(t, x-h)}{2h^2}$$

となる。ここで $\tau = h^2$ のとき (2) より

$$u(t+\tau, x) = \frac{1}{2}u(t, x+h) + \frac{1}{2}u(t, x-h)$$

となり, $t = n\tau, x = jh$ のとき

$$(3) \quad u((n+1)\tau, jh) = \frac{1}{2}u(n\tau, (j+1)h) + \frac{1}{2}u(n\tau, (j-1)h)$$

が成り立つ。一方ランダムウォーク S_n の確率 $P(S_n = j) = P(S_n h = jh)$ は

$$(4) \quad P(S_{n+1}h = jh) = \frac{1}{2}P(S_n h = (j+1)h) + \frac{1}{2}P(S_n h = (j-1)h)$$

< 確率密度の収束 2 >

より, $P(S_n h = jh)$ と $u(n\tau, jh)h$ は同じ差分式をみたす。

そこで $n = 0$ のときの値 (初期値) を同じにする。

$$(5) \quad u(0, jh)h = \left\{ \begin{array}{l} 1 : j = 0 \text{ のとき} \\ 0 : \text{その他} \end{array} \right\} = P(S_0 = j) \quad (j \in \mathbf{Z})$$

とおくと

$$u(n\tau, jh)h = P(S_n h = jh) \quad (n = 0, 1, 2, \dots, j \in \mathbf{Z})$$

が成立する。(3), (5) の解 $u(n\tau, jh)$ に対し

$$u_h(t, x) = u\left(\left[\frac{t}{\tau}\right]\tau, \left[\frac{x}{h}\right]h\right) \quad (\tau = h^2)$$

とおくと $h \rightarrow 0$ のときの極限の関数

$$\lim_{h \rightarrow 0} u_h(t, x) = u(t, x) \quad (\text{差分解の収束})$$

は (1) の解で初期条件

$$(0) \quad \lim_{t \rightarrow 0} u(t, x) = \left\{ \begin{array}{l} +\infty : x = 0 \text{ のとき} \\ 0 : \text{その他} \end{array} \right., \quad \int_{-\infty}^{\infty} u(t, x) dx = 1$$

をみたす。この解 $u(t, x)$ は

$$u(t, x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}}$$

となる。ランダムウォーク $\{S_n\}$ に対し, 折れ線 $\{B_h(t)\}$ の存在確率は,

$t = n\tau$ のとき $B_h(t) = B_h(n\tau) = S_n h$ だから

$$P(a < B_h(t) < b) = P(a < S_n h < b)$$

$$= \sum_{\substack{j \\ a < jh < b}} P(S_n h = jh) = \sum_{\substack{j \\ a < jh < b}} u(n\tau, jh)h = \sum_{\substack{j \\ a < jh < b}} u_h(t, jh)h \doteq \int_a^b u_h(t, x) dx$$

$$\xrightarrow{(h \rightarrow 0)} \int_a^b u(t, x) dx = \int_a^b \frac{1}{\sqrt{2\pi t}} e^{-\frac{x^2}{2t}} dx \quad \text{より } \bullet \text{ が成り立つ。}$$

< ランダムウォークの族 >

$\tau/h^2 \leq 1$ をみたす正定数 τ , h に対し, $h\mathbf{Z} = \{hz : z \text{ は整数}\}$ 上を動くランダムウォーク $\{S_h^\tau(n)\}$ を次で定める。

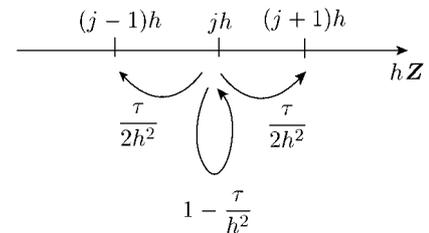
< $\{S_h^\tau(n)\}$ の定義 >

[I] (マルコフ性)

$$\begin{aligned} P\left(S_h^\tau(n+1) = x_{n+1} \mid S_h^\tau(n) = x_n, S_h^\tau(n-1) = x_{n-1}, \dots, S_h^\tau(1) = x_1\right) \\ = P\left(S_h^\tau(n+1) = x_{n+1} \mid S_h^\tau(n) = x_n\right) \end{aligned}$$

[II] (推移確率)

$$\begin{aligned} P\left(S_h^\tau(n+1) = jh \mid S_h^\tau(n) = jh\right) &= 1 - \frac{\tau}{h^2} \\ P\left(S_h^\tau(n+1) = (j+1)h \mid S_h^\tau(n) = jh\right) &= \frac{\tau}{2h^2} \\ P\left(S_h^\tau(n+1) = (j-1)h \mid S_h^\tau(n) = jh\right) &= \frac{\tau}{2h^2} \end{aligned}$$



[III] (初期値)

$$P(S_h^\tau(0) = 0) = 1, \quad P(S_h^\tau(0) = j) = 0 \quad (j \neq 0)$$

この定義より次式が成立する

$$\begin{aligned} (*1) \quad &P(S_h^\tau(n+1) = jh) \\ &= \left(1 - \frac{\tau}{h^2}\right) P\left(S_h^\tau(n) = jh\right) + \frac{\tau}{2h^2} P\left(S_h^\tau(n) = (j+1)h\right) + \frac{\tau}{2h^2} P\left(S_h^\tau(n) = (j-1)h\right) \end{aligned}$$

この式から $u_j^n = P(S_h^\tau(n) = jh)h^{-1}$ とおくと次式が成り立つ。

$$(*2) \quad \frac{u_j^{n+1} - u_j^n}{\tau} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2h^2}$$

問 (*1), (*2) を証明せよ。

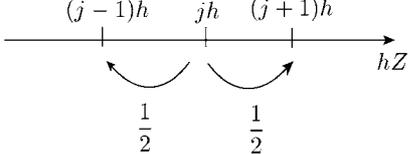
< ジャンプ型マルコフ過程 >

$\{X_1, X_2, \dots, X_n, \dots\}$ は独立で同分布な確率変数列で,
 $P(X_n = 1) = P(X_n = -1) = \frac{1}{2}$ であるとする。 $h > 0$ に対し

$$S_h(n) = \begin{cases} \left(\sum_{k=1}^n X_k\right)h & : n \geq 1 \\ 0 & : n = 0 \end{cases}$$

とおくと, $\{S_h(n) : n \geq 0\}$ は hZ 上を動くランダムウォークであり,
 推移確率は

$$P\left(S_h(n+1) = (j+1)h \mid S_h(n) = jh\right) = \frac{1}{2}$$

$$P\left(S_h(n+1) = (j-1)h \mid S_h(n) = jh\right) = \frac{1}{2}$$


となる。次の極限定理は「ポアソンの少数の法則」の変形である。

定理 18 前のページのランダムウォーク $\{S_h^\tau(n) : n = 0, 1, 2, \dots\}$ に対し,
 $\left\{S_h^\tau\left(\left[\frac{t}{\tau}\right]\right) : t \geq 0\right\}$ は連続時間の確率過程である。 $\tau \rightarrow 0$ のとき
 「 $\left\{S_h^\tau\left(\left[\frac{t}{\tau}\right]\right) : t \geq 0\right\}$ の分布」は「 $\{S_h(N_h(t)) : t \geq 0\}$ の分布」に収束する。
 ただし $N_h(t)$ は $\{S_h(n) : n \geq 0\}$ と独立な ($\lambda = h^{-2}$ の場合の) ポアソン過程
 であり,

$$P(N_h(t) = n) = e^{-\frac{t}{h^2}} \times \frac{\left(\frac{t}{h^2}\right)^n}{n!} \quad (n = 0, 1, 2, \dots)$$

 を満たす。

$\{S_h(n)\}$ と $\{N_h(t)\}$ の定義から次式が成り立つ。

$$(*3) \quad P\left(S_h(N_h(t)) = jh\right) = \sum_{n=0}^{\infty} P\left(S_h(n) = jh\right) e^{-\frac{t}{h^2}} \times \frac{\left(\frac{t}{h^2}\right)^n}{n!}$$

$$(*4) \quad P\left(S_h(n+1) = jh\right) = \frac{1}{2}P\left(S_h(n) = (j+1)h\right) + \frac{1}{2}P\left(S_h(n) = (j-1)h\right)$$

確率過程 $\{S_h(N_h(t)) : t \geq 0\}$ はジャンプ型のマルコフ過程である。

問 (*3), (*4) を証明せよ。

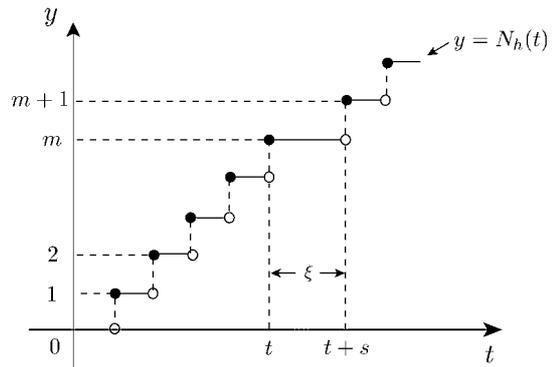
< ブラウン運動への収束 >

$u_h(t, jh) = P(S_h(N_h(t)) = jh)h^{-1}$ とおくと次式が成り立つ。

$$(*5) \quad \frac{d}{dt}u_h(t, jh) = \frac{u_h(t, (j+1)h) - 2u_h(t, jh) + u_h(t, (j-1)h)}{2h^2}$$

問 (*5) を証明せよ。

ポアソン過程 $\{N_h(t) : t \geq 0\}$ は右図のような階段型のグラフであり、 $N_h(t) = m$ であるとき、次にジャンプするまでの時間 ξ は密度 $\frac{1}{h^2}e^{-\frac{x}{h^2}}$ の指数分布に従う。このとき 37 ページより待ち時間 ξ の平均は



$$(*6) \quad E[\xi] = h^2$$

となる。

定理 19 $\{B(t) : t \geq 0\}$ を 0 から出発する ($B(0) = 0$) 1 次元ブラウン運動とする。

$h \rightarrow 0$ のとき

「 $\{S_h(N_h(t)) : t \geq 0\}$ の分布」は「 $\{B(t) : t \geq 0\}$ の分布」に収束する。

定理 20 $\tau > 0$, $h > 0$, $\tau h^{-2} \leq 1$ なる定数 τ, h に対し, $\{S_h^\tau(n)\}$

を 45 ページで定義したランダムウォークとする。また $\{B(t) : t \geq 0\}$

を 0 から出発する ($B(0) = 0$) 1 次元ブラウン運動とする。

$\frac{\tau}{h^2} \leq 1$ の関係を保ったまま, $\tau \rightarrow 0, h \rightarrow 0$ のとき

「 $\left\{S_h^\tau\left(\left[\frac{t}{\tau}\right]\right) : t \geq 0\right\}$ の分布」は「 $\{B(t) : t \geq 0\}$ の分布」に収束する。

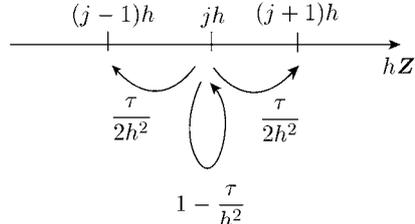
定理 18~20 の証明は関数空間における確率測度の収束を示さねばならないので省略する。

< 乱歩・ポアソン過程・ブラウン運動 >

ランダムウォーク, ポアソン過程, ブラウン運動についての関係をまとめる。
 ブラウン運動の離散モデルであるランダムウォーク (乱歩) の時間幅を $\Delta t = \tau$,
 空間幅を $\Delta x = h$ ($\tau \leq h^2$) とする。

< ① ランダムウォーク $\{S_h^\tau(n) : n = 0, 1, 2, \dots\}$ >

確率密度 $u_j^n = P(S_h^\tau(n) = jh)h^{-1}$ の満たす式は

$$\frac{u_j^{n+1} - u_j^n}{\tau} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2h^2} \quad (\text{差分方程式})$$


このランダムウォークに対し連続時間にした $\left\{S_h^\tau\left(\left[\frac{t}{\tau}\right]\right) : t \geq 0\right\}$ は
 $\tau \rightarrow 0$ のとき次のジャンプ型マルコフ過程に収束する。

< ② ジャンプ型マルコフ過程 $\{S_h(N_h(t)) : t \geq 0\}$ >

$\{S_h(n) : n = 0, 1, 2, \dots\}$ は $h\mathbf{Z}$ 上を動く乱歩で, 左右に 1 歩動く
 確率はそれぞれ $\frac{1}{2}$ である。また $\{N_h(t) : t \geq 0\}$ は $\{S_h(n)\}$ と独立な

ポアソン過程で, $P(N_h(t) = n) = e^{-\frac{t}{h^2}} \times \frac{(\frac{t}{h^2})^n}{n!}$ ($n = 0, 1, 2, \dots$) である。

確率密度 $u(t, jh) = P(S_h(N_h(t)) = jh)h^{-1}$ の満たす式は

$$\frac{d}{dt}u(t, jh) = \frac{u(t, (j+1)h) - 2u(t, jh) + u(t, (j-1)h)}{h^2} \quad (\text{微分差分方程式})$$

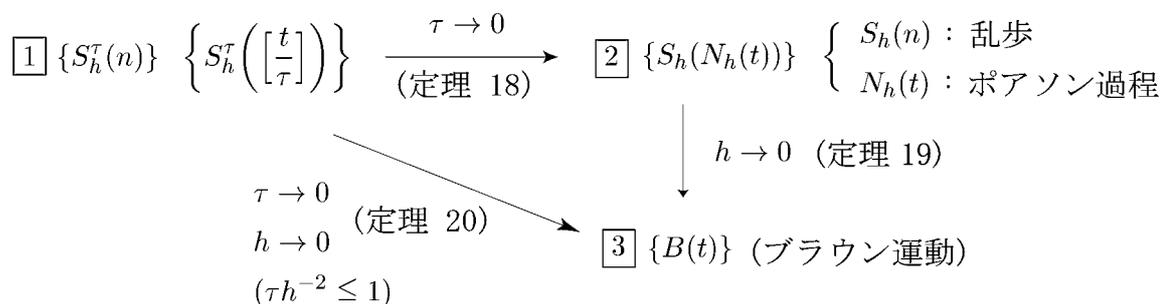
となる。ポアソン過程 $\{N_h(t)\}$ はジャンプする時刻を決める確率過程である。
 このジャンプ型マルコフ過程は $h \rightarrow 0$ のときブラウン運動に収束する。

< ③ ブラウン運動 $\{B(t) : t \geq 0\}$ >

確率密度関数 $u(t, x)$ $\left(\Leftrightarrow P(B(t) \in A) = \int_A u(t, x) dx\right)$ の満たす式は

$$\frac{\partial}{\partial t}u(t, x) = \frac{1}{2} \frac{\partial^2}{\partial x^2}u(t, x) \quad (\text{熱方程式, 拡散方程式})$$

以上まとめると下の図式になる。定理 18 は「少数の法則」, 定理 19, 20 は
 「中心極限定理」である。



< 独立確率変数の和の分布 1 >

定理 1 X と Y は整数値をとる確率変数

$$P(X = k) = p_k, \quad P(Y = k) = q_k \quad (k = 0, \pm 1, \pm 2, \dots)$$

X と Y が独立ならば

$$P(X + Y = k) = \sum_{j=-\infty}^{\infty} p_j \times q_{k-j} \quad \left(= \sum_{j=-\infty}^{\infty} p_{k-j} \times q_j \right)$$

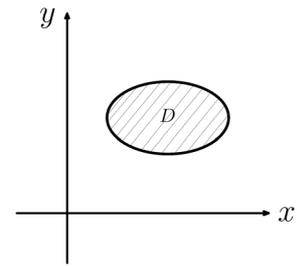
$$\begin{aligned} \text{(証明)} \quad P(X + Y = k) &= \sum_{j=-\infty}^{\infty} P(X = j, Y = k - j) \\ &= \sum_{j=-\infty}^{\infty} P(X = j)P(Y = k - j) = \sum_{j=-\infty}^{\infty} p_j \times q_{k-j} \quad (\text{証明終}) \end{aligned}$$

定理 2 X と Y は実数値連続密度関数をもつ確率変数

$$P(X \in A) = \int_A p(x)dx, \quad P(Y \in B) = \int_B q(x)dx$$

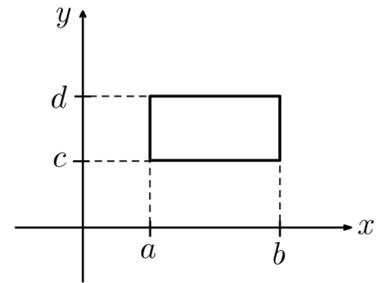
X と Y が独立ならば, 平面上の任意の領域 D に対し

$$P((X, Y) \in D) = \iint_D p(x)q(y)dxdy$$



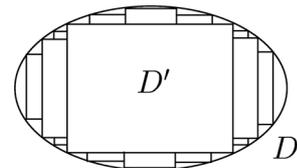
(証明) D が長方形領域 $[a, b] \times [c, d]$ のとき

$$\begin{aligned} P((X, Y) \in D) &= P(X \in [a, b], Y \in [c, d]) \\ &= P(X \in [a, b]) \times P(Y \in [c, d]) = \int_a^b p(x)dx \times \int_c^d q(y)dy \\ &= \int_a^b \left\{ \int_c^d p(x)q(y)dy \right\} dx = \iint_{[a, b] \times [c, d]} p(x)q(y)dy = \iint_D p(x)q(y)dxdy \end{aligned}$$



D が一般領域のときは D を長方形領域の集まり

D' で近似すれば良い。 (証明終)



< 独立確率変数の和の分布 2 >

定理 3 X と Y は実数値連続密度関数をもつ確率変数で, X の密度関数は $p(x)$, Y の密度関数は $q(x)$ とする。

X と Y が独立ならば, $X + Y$ の密度関数は

$$(p * q)(x) = \int_{-\infty}^{\infty} p(x-y)q(y)dy = \int_{-\infty}^{\infty} p(y)q(x-y)dy$$

である。この関数 $p * q(x)$ を p と q の「たたみこみ」(convolution) という。

(証明) 任意の実数 k に対し

$$D = \{(x, y) : x + y \leq k\}$$

とおくと, X と Y は独立だから

定理 2 より

$$P(X + Y \leq k) = P((X, Y) \in D)$$

$$= \iint_D p(x)q(y)dx dy = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{k-x} p(x)q(y)dy \right\} dx$$

$$= \int_{-\infty}^{\infty} p(x) \left\{ \int_{-\infty}^{k-x} q(y)dy \right\} dx \quad (y = u - x)$$

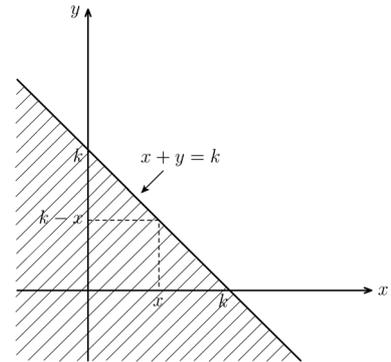
$$= \int_{-\infty}^{\infty} p(x) \left\{ \int_{-\infty}^k q(u-x)du \right\} dx$$

$$= \int_{-\infty}^k \left\{ \int_{-\infty}^{\infty} p(x)q(u-x)dx \right\} du = \int_{-\infty}^k (p * q)(u)du$$

$$= \int_{-\infty}^k (p * q)(x)dx \text{ より}$$

$$P(X + Y \in A) = \int_A (p * q)(x)dx$$

(証明終)



< 独立確率変数の和の分布 3 >

< 10 ページ 定理 4 >

X_1 は正規分布 $N(m_1, v_1)$ に従い, X_2 は正規分布 $N(m_2, v_2)$ に従う確率変数で, X_1 と X_2 は独立とする。このとき和 $X_1 + X_2$ は正規分布 $N(m_1 + m_2, v_1 + v_2)$ に従う。

この定理を証明するためには

$$p_1(x) = \frac{1}{\sqrt{2\pi v_1}} e^{-\frac{(x-m_1)^2}{2v_1}}, \quad p_2(x) = \frac{1}{\sqrt{2\pi v_2}} e^{-\frac{(x-m_2)^2}{2v_2}}$$

に対し

$$(p_1 * p_2)(x) = \int_{-\infty}^{\infty} p_1(x-y)p_2(y)dy = \frac{1}{\sqrt{2\pi(v_1+v_2)}} e^{-\frac{(x-m_1-m_2)^2}{2(v_1+v_2)}}$$

であることを示せば良い。積分の変数変換だけでできる問題である。

詳しい証明は練習問題とする。

< 12 ページ 定理 6 >

X_1, X_2 がそれぞれ $Gamma(\alpha_1, \beta), Gamma(\alpha_2, \beta)$ に従う確率変数で X_1 と X_2 が独立ならば, 和 $X_1 + X_2$ は $Gamma(\alpha_1 + \alpha_2, \beta)$ に従う。

この定理を証明するためには

$$p_\alpha(x) = \begin{cases} \frac{1}{\Gamma(\alpha) \beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}} & : x > 0 \\ 0 & : x \leq 0 \end{cases}$$

に対し

$$(p_{\alpha_1} * p_{\alpha_2})(x) = \int_{-\infty}^{\infty} p_{\alpha_1}(x-y)p_{\alpha_2}(y)dy = \int_0^x p_{\alpha_1}(x-y)p_{\alpha_2}(y)dy = p_{\alpha_1+\alpha_2}(x)$$

であることをベータ関数とガンマ関数の関係を用いて示せば良い。

詳しい証明は練習問題とする。

< 一様分布の独立和の標準化 1 >

例1
$$p(x) = \begin{cases} 1 & : 0 \leq x \leq 1 \\ 0 & : \text{それ以外} \end{cases}$$

を密度関数とする確率変数 X_1, X_2 が独立とする。

$X_1 + X_2$ の確率密度関数を $p_2(x)$ とすると

$$p_2 = (p * p)(x) = \int_{-\infty}^{\infty} p(x-y)p(y)dy = \int_0^1 p(x-y)p(y)dy = \int_0^1 p(x-y)dy$$

より

$$\textcircled{1} x < 0 \text{ のとき} \quad p_2(x) = \int_0^1 \underbrace{p(x-y)}_{\ominus} dy = 0$$

$$\textcircled{2} 0 \leq x \leq 1 \text{ のとき} \quad p_2(x) = \int_0^x \underbrace{p(x-y)}_{=1} dy + \int_x^1 \underbrace{p(x-y)}_{=0} dy = \int_0^x 1 dy = x$$

$$\textcircled{3} 1 \leq x \leq 2 \text{ のとき} \quad p_2(x) = \int_0^{x-1} \underbrace{p(x-y)}_{=0} dy + \int_{x-1}^1 p(x-y) dy = \int_{x-1}^1 1 dy = 2 - x$$

$$\textcircled{4} 2 < x \text{ のとき} \quad p_2(x) = \int_0^1 \underbrace{p(x-y)}_{1 \text{ より大}} dy = 0$$

従って、 $y = p_2(x)$ のグラフは右図のようになる。

一方

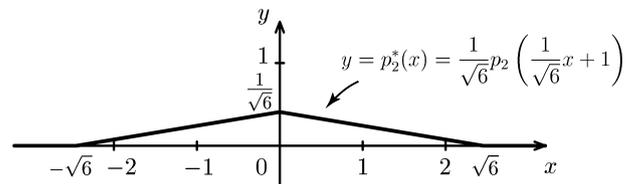
$$E[X_1] = E[X_2] = \int_0^1 xp(x)dx = \frac{1}{2}$$

$$V(X_1) = V(X_2) = \int_0^1 \left(x - \frac{1}{2}\right)^2 p(x)dx = \frac{1}{12}$$

であるから $E[X_1 + X_2] = 1$, $V(X_1 + X_2) = \frac{2}{12} = \frac{1}{6}$ で、和 $X_1 + X_2$ の標準化

$(X_1 + X_2)^* = \frac{X_1 + X_2 - 1}{\sqrt{\frac{1}{6}}} = \sqrt{6}(X_1 + X_2 - 1)$ の密度関数を $p_2^*(x)$ とすると、

$y = p_2^*(x)$ のグラフは右のとおり。



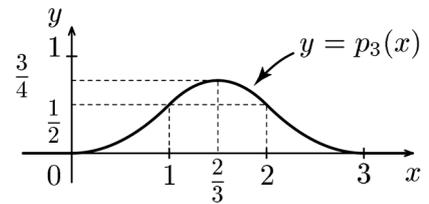
$$p_2^*(x) = \frac{1}{\sqrt{6}} p_2\left(\frac{1}{\sqrt{6}}x + 1\right) = \begin{cases} 0 & : x < -\sqrt{6} \\ \frac{x}{6} + \frac{1}{\sqrt{6}} & : -\sqrt{6} \leq x \leq 0 \\ \frac{1}{\sqrt{6}} - \frac{x}{6} & : 0 \leq x \leq \sqrt{6} \\ 0 & : \sqrt{6} < x \end{cases}$$

< 一様分布の独立和の標準化 2 >

例 2 X_1, X_2, X_3 は共に $p(x) = \begin{cases} 1 & : 0 \leq x \leq 1 \\ 0 & : \text{それ以外} \end{cases}$ を密度関数とする確率変数で、独立とする。 $X_1 + X_2 + X_3$ の確率密度関数を $p_3(x)$ とすると、例 1 より $p_3(x)$ は $p(x)$ と $p_2(x)$ とのたたみこみであるから

$$p_3(x) = (p * p_2)(x) = \int_{-\infty}^{\infty} p(x-y)p_2(y)dy = \int_0^2 p(x-y)p_2(y)dy$$

$$= \begin{cases} 0 & : x < 0 \\ \frac{1}{2}x^2 & : 0 \leq x \leq 1 \\ -(x - \frac{3}{2})^2 + \frac{3}{4} & : 1 \leq x \leq 2 \\ \frac{1}{2}(x-3)^2 & : 2 \leq x \leq 3 \\ 0 & : 3 < x \end{cases}$$



となる。一方

$$E[X_k] = \frac{1}{2}, \quad V(X_k) = \frac{1}{12} \quad (k = 1, 2, 3)$$

なので

$$E[X_1 + X_2 + X_3] = \frac{3}{2}, \quad V(X_1 + X_2 + X_3) = \frac{3}{12} = \frac{1}{4}$$

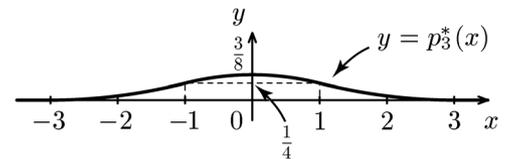
より、和 $X_1 + X_2 + X_3$ の標準化は

$$(X_1 + X_2 + X_3)^* = \frac{X_1 + X_2 + X_3 - \frac{3}{2}}{\sqrt{\frac{1}{4}}} = 2(X_1 + X_2 + X_3) - 3$$

である。この標準化された確率変数 $(X_1 + X_2 + X_3)^*$ の確率密度関数を $p_3^*(x)$ とすると

$$p_3^*(x) = p_3\left(\frac{x+3}{2}\right) \times \frac{1}{2}$$

$$= \begin{cases} 0 & : x < -3 \\ \frac{1}{16}(x+3)^2 & : -3 \leq x \leq -1 \\ -\frac{1}{8}x^2 + \frac{3}{8} & : -1 \leq x \leq 1 \\ \frac{1}{16}(x-3)^2 & : 1 \leq x \leq 3 \\ 0 & : 3 < x \end{cases}$$



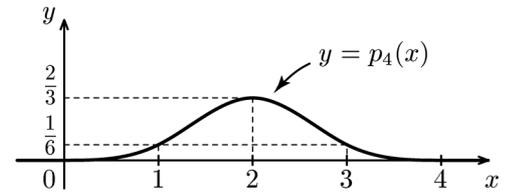
となる。

< 一様分布の独立和の標準化 3 >

例 3 X_1, X_2, X_3, X_4 は共に $p(x) = \begin{cases} 1 & : 0 \leq x \leq 1 \\ 0 & : \text{それ以外} \end{cases}$ を密度関数とする確率変数で、独立とする。 $X_1 + X_2 + X_3 + X_4$ の確率密度関数を $p_4(x)$ とすると、例 2 より $p_4(x)$ は $p(x)$ と $p_3(x)$ とのたたみこみであるから

$$p_4(x) = (p * p_3)(x) = \int_{-\infty}^{\infty} p(x-y)p_3(y)dy = \int_0^3 p(x-y)p_3(y)dy$$

$$= \begin{cases} 0 & : x < 0 \\ \frac{1}{6}x^3 & : 0 \leq x \leq 1 \\ \frac{1}{6} - \frac{1}{2}(x-1)(x^2 - 3x + 1) & : 1 \leq x \leq 2 \\ \frac{1}{6} + \frac{1}{2}(x-3)(x^2 - 5x + 5) & : 2 \leq x \leq 3 \\ -\frac{1}{6}(x-4)^3 & : 3 \leq x \leq 4 \\ 0 & : 4 < x \end{cases}$$



となる。一方

$$E[X_k] = \frac{1}{2}, \quad V(X_k) = \frac{1}{12} \quad (k = 1, 2, 3, 4)$$

であるから

$$E[X_1 + X_2 + X_3 + X_4] = 2, \quad V(X_1 + X_2 + X_3 + X_4) = \frac{1}{3}$$

より、和 $X_1 + X_2 + X_3 + X_4$ の標準化は

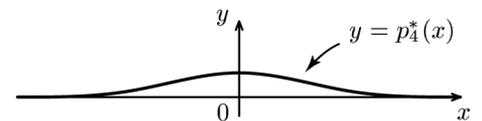
$$(X_1 + X_2 + X_3 + X_4)^* = \frac{X_1 + X_2 + X_3 + X_4 - 2}{\sqrt{\frac{1}{3}}}$$

である。この標準化された確率変数 $(X_1 + X_2 + X_3 + X_4)^*$ の確率密度関数を $p_4^*(x)$

とすると

$$p_4^*(x) = p_4\left(\frac{x}{\sqrt{3}} + 2\right) \frac{1}{\sqrt{3}}$$

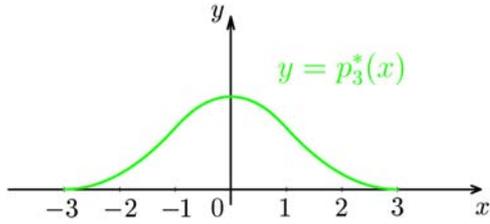
$$= \begin{cases} 0 & : x < -2\sqrt{3} \\ \frac{1}{54}(x+2\sqrt{3})^3 & : -2\sqrt{3} \leq x \leq -\sqrt{3} \\ -\frac{x^3}{18} - \frac{\sqrt{3}x^2}{9} + \frac{2\sqrt{3}}{9} & : -\sqrt{3} \leq x \leq 0 \\ \frac{x^3}{18} - \frac{\sqrt{3}x^2}{9} + \frac{2\sqrt{3}}{9} & : 0 \leq x \leq \sqrt{3} \\ -\frac{1}{54}(x-2\sqrt{3})^3 & : \sqrt{3} \leq x \leq 2\sqrt{3} \\ 0 & : 2\sqrt{3} < x \end{cases}$$



となる。

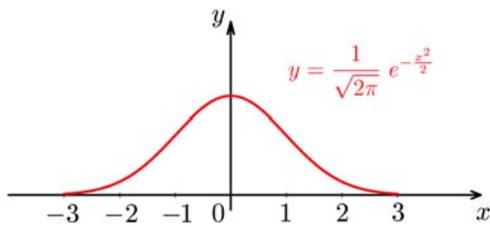
< 一様分布の独立和の標準化 4 >

$X_1, X_2, X_3, X_4 : [0, 1]$ 上に一様分布する独立確率変数 (*independent random variable distributed uniformly in $[0, 1]$*)

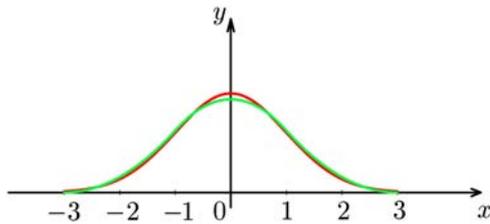


$$E[X_1 + X_2 + X_3] = \frac{3}{2}, \quad V[X_1 + X_2 + X_3] = \frac{1}{4}$$

$$P\left(\frac{X_1 + X_2 + X_3 - \frac{3}{2}}{\frac{1}{2}} \in A\right) = \int_A p_3^*(x) dx$$

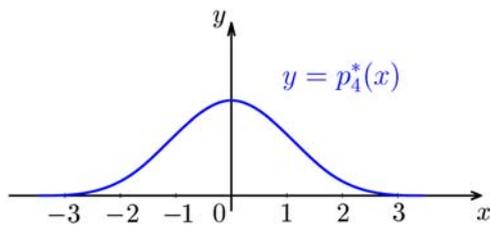


標準正規分布 (*normal distribution*) $N(0, 1)$



赤 : 標準正規分布 $N(0, 1)$
(*red*) (*normal distribution*)

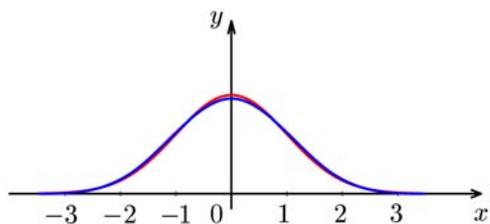
緑 : $y = p_3^*(x)$
(*green*)



$$E[X_1 + X_2 + X_3 + X_4] = 2$$

$$V[X_1 + X_2 + X_3 + X_4] = \frac{1}{3}$$

$$P\left(\frac{X_1 + X_2 + X_3 + X_4 - 2}{\sqrt{\frac{1}{3}}} \in A\right) = \int_A p_4^*(x) dx$$



赤 : 標準正規分布 $N(0, 1)$
(*red*) (*normal distribution*)

青 : $y = p_4^*(x)$
(*blue*)

< チェビシエフの不等式 >

確率変数 X は 2 次モーメント $E[X^2]$ が有界であるとする。

このとき任意の定数 $\varepsilon > 0$ に対し

$$P(|X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} E[|X|^2]$$

が成り立つ。

X が連続型の確率変数であるとき、すなわち

$$P(X \in A) = \int_A p(x) dx \quad (A \subset \mathbf{R})$$

をみたす確率密度 $p(x)$ を持つときに証明する。

(証明)

$$\begin{aligned} P(|X| \geq \varepsilon) &= \int_{|x| \geq \varepsilon} p(x) dx \\ &\leq \int_{|x| \geq \varepsilon} \frac{x^2}{\varepsilon^2} p(x) dx = \frac{1}{\varepsilon^2} \int_{|x| \geq \varepsilon} x^2 p(x) dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{\infty} x^2 p(x) dx = \frac{1}{\varepsilon^2} E[X^2] \end{aligned}$$

(証明終)

問 X が離散型の確率変数であるとき、すなわち

$$P(X = x_n) = p_n \quad (n = 1, 2, 3, \dots)$$

$$p_n \geq 0, \quad \sum_{n=1}^{\infty} p_n = 1, \quad \sum_{n=1}^{\infty} (x_n)^2 p_n < +\infty$$

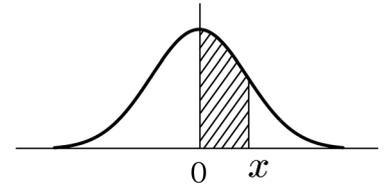
であるときチェビシエフの不等式 $P(|X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2} E[X^2]$

を証明せよ。

正規分布表

 $N(0, 1)$

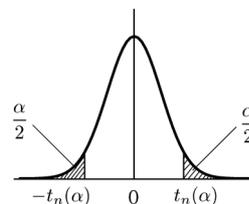
$$\Phi(x) = \int_0^x f(x)dx = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{x^2}{2}} dx$$



x	00	01	02	03	04	05	06	07	08	09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.49379	.49396	.49413	.49430	.49446	.49461	.49477	.49492	.49506	.49520
2.6	.49534	.49547	.49560	.49573	.49585	.49597	.49609	.49621	.49632	.49643
2.7	.49653	.49664	.49674	.49683	.49693	.49702	.49711	.49720	.49728	.49736
2.8	.49744	.49752	.49760	.49767	.49774	.49781	.49788	.49795	.49801	.49807
2.9	.49813	.49819	.49825	.49831	.49836	.49841	.49846	.49851	.49856	.49860
3.0	.49865	.49869	.49874	.49878	.49882	.49886	.49889	.49893	.49897	.49900

t 分布表

$$P(|T| \geq t_n(\alpha)) = \alpha$$



<i>n</i> \ <i>P</i>	0.50	0.25	0.10	0.05	0.025	0.02	0.01	0.005
1	1.000	2.414	6.314	12.706	25.452	31.821	63.657	127.32
2	0.816	1.604	2.920	4.303	6.205	6.965	9.925	14.089
3	0.765	1.423	2.353	3.182	4.177	4.541	5.841	7.453
4	0.741	1.344	2.132	2.776	3.495	3.747	4.604	5.598
5	0.727	1.301	2.015	2.571	3.163	3.365	4.032	4.773
6	0.718	1.273	1.943	2.447	2.969	3.143	3.707	4.317
7	0.711	1.254	1.895	2.365	2.841	2.998	3.499	4.029
8	0.706	1.240	1.860	2.306	2.752	2.896	3.355	3.833
9	0.703	1.230	1.833	2.262	2.685	2.821	3.250	3.690
10	0.700	1.221	1.812	2.228	2.634	2.764	3.169	3.581
11	0.697	1.215	1.796	2.201	2.593	2.718	3.106	3.497
12	0.695	1.209	1.782	2.179	2.560	2.681	3.055	3.428
13	0.694	1.204	1.771	2.160	2.533	2.650	3.012	3.373
14	0.692	1.200	1.761	2.145	2.510	2.624	2.977	3.326
15	0.691	1.197	1.753	2.131	2.490	2.602	2.947	3.286
16	0.690	1.194	1.746	2.120	2.473	2.583	2.921	3.252
17	0.689	1.191	1.740	2.110	2.458	2.567	2.898	3.223
18	0.688	1.189	1.734	2.101	2.445	2.552	2.878	3.197
19	0.688	1.187	1.729	2.093	2.433	2.539	2.861	3.174
20	0.687	1.185	1.725	2.086	2.423	2.528	2.845	3.153
21	0.686	1.183	1.721	2.080	2.414	2.518	2.831	3.135
22	0.686	1.182	1.717	2.074	2.406	2.508	2.819	3.119
23	0.685	1.180	1.714	2.069	2.398	2.500	2.807	3.104
24	0.685	1.179	1.711	2.064	2.391	2.492	2.797	3.091
25	0.684	1.178	1.708	2.060	2.385	2.485	2.787	3.078
26	0.684	1.177	1.706	2.056	2.379	2.479	2.779	3.067
27	0.684	1.176	1.703	2.052	2.373	2.473	2.771	3.057
28	0.683	1.175	1.701	2.048	2.369	2.467	2.763	3.047
29	0.683	1.174	1.699	2.045	2.364	2.462	2.756	3.038
30	0.683	1.173	1.697	2.042	2.360	2.457	2.750	3.030
40	0.681	1.167	1.684	2.021	2.329	2.423	2.704	2.971
60	0.679	1.162	1.671	2.000	2.299	2.390	2.660	2.915
120	0.677	1.156	1.658	1.980	2.270	2.358	2.617	2.860
∞	0.674	1.150	1.645	1.960	2.241	2.326	2.576	2.807