データサイエンスのための統計学

多次元の確率分布

by 尚晋　(名古屋大学経済学研究科助教)
on 2025 年 6 月 10 日

目次

* 復習
* 同時確率分布と周辺確率分布
* 条件付確率分布と独立な確率変数
* 独立な確率変数の和

目次

* 復習
* 同時確率分布と周辺確率分布
* 条件付確率分布と独立な確率変数
* 独立な確率変数の和

» 前回の復習　各確率分布の特徴と応用

二項分布：二値性　復元抽出する場合、製品の不良率、政権への支持率、テレビの視聴率、アンケートや投票結果、治療効果の有無

ポアソン分布：発生確率低い　航空機事故件数、火災件数、大量生産の不良品個数

指数分布：待ち時間分布　故障までの待ち時間、災害までの年数、生存関数

正規分布　生物・人体測定値、標本分布論の測定誤差

対数正規分布 (高い方には明確な限度がない)　所得, 貯蓄額

超幾何分布　非復元抽出する場合、捕獲再捕獲法

パレート分布　高額所得者の所得分布

» 前回の復習

計算例

株価の例：確率変数 X はある株式の利回り (%) で，平均 3, 分散 10 の正規分布 $N(3, 10)$ に従うものとする．株式投資が損となる確率を求む．

» 前回の復習

計算例

株価の例：確率変数 X はある株式の利回り (%) で，平均 3, 分散 10 の正規分布 $N(3, 10)$ に従うものとする．株式投資が損となる確率を求む．

解答：

$\sigma = \sqrt{10} \approx 3.162$ であるから、条件 $X < 0$ は

$$\begin{aligned} X &< 0 \\ \frac{X-3}{\sqrt{10}} &< \frac{0-3}{\sqrt{10}} \\ Z &< -0.949 \end{aligned}$$

と書き換えられる、ただし $Z \sim N(0,1)$

$$\begin{aligned} P(X<0) &= P(Z<-0.949) \qquad (\because \Phi(-0.949) = 1 - \Phi(0.949)) \\ &= P(Z>0.949) \end{aligned}$$

正規分布表によると、Z の値が 0.949 と対応する上側確率は 0.171 である。従って、株式投資が損となる確率は 17.1%.

目次

* 復習
* 同時確率分布と周辺確率分布
* 条件付確率分布と独立な確率変数
* 独立な確率変数の和

» 同時確率分布と周辺確率分布

離散型確率変数の同時確率分布　X, Y は離散型とする.$X = x$ であり同時に $Y = y$ である確率

$$P(X = x, Y = y) = f(x, y) \tag{1}$$

を, 2 次元確率変数 (X, Y) の同時確率分布という.

離散型確率変数の周辺確率分布　同時確率分布から、X, Y 単独の確率分布は

$$f_X(x) = P(X = x_i) = \sum_y f(x, y) \tag{2a}$$

$$f_Y(y) = P(Y = y_j) = \sum_x f(x, y) \tag{2b}$$

で求められる. それぞれ X, Y の**周辺確率分布**という.

- 同時確率関数 $f(x, y)$ は $f(x, y) \geq 0$, かつ, $\sum_{i=1}^{n} \sum_{j=1}^{m} f(x, y) = 1$ を満たす.

周辺確率関数 $f_X(x), f_Y(y)$ も同じ.

» 同時確率分布と周辺確率分布　例

例 1：次の同時分布の、X と Y それぞれの周辺確率分布 $f_x(x), f_y(y)$ は？

X \ Y	$Y=8$	$Y=9$	$f_x(x)$
$X=1$	0.1	0.1	0.2
$X=2$	0.2	0.3	0.5
$X=3$	0.1	0.2	0.3
$f_y(y)$	0.4	0.6	

X の周辺確率分布 $f_x(x)$ は定義通りに求める：

$$f_x(x) = \begin{cases} 0.2 & \text{for } x=1 \\ 0.5 & \text{for } x=2 \\ 0.3 & \text{for } x=3 \end{cases}$$

同じで、Y の周辺確率分布 $f_y(y)$ も定義通りに求める：

$$f_y(y) = \begin{cases} 0.4 & \text{for } y=8 \\ 0.6 & \text{for } y=9 \end{cases}$$

» 同時確率密度関数と周辺確率密度関数

連続型確率変数の同時確率密度関数 連続型の場合，$a < X < b$ かつ $c < Y < d$ となる確率 $P(a < X < b, c < Y < d)$ が同時確率密度関数 $f(x, y)$ を用いて

$$P(a < X < b, c < Y < d) = \int_c^d \int_a^b f(x, y)\,dx\,dy \tag{3}$$

で与えられる．

連続型確率変数の周辺確率密度関数 連続型の場合も，X, Y の周辺確率密度関数は，同時確率密度関数から

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \tag{4a}$$

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx \tag{4b}$$

で与えられる．

■ 同時確率密度関数 $f(x, y)$ は $f(x, y) \geq 0$，かつ，$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) = 1$ を満たす．周辺確率密度関数 $f_x(x), f_y(y)$ も同じ．

» 同時確率密度関数と周辺確率密度関数

Y の値にかかわらず, X が $a < X < b$ となる確率 $P(a < X < b)$ は同時確率密度関数および周辺確率密度関数を用いて

$$P(a < X < b) = \int_{-\infty}^{\infty} \int_{a}^{b} f(x, y)\, dx\, dy \tag{5a}$$

$$= \int_{a}^{b} f_x(x)\, dx \tag{5b}$$

で与えられる.

» 共分散

共分散 X, Y の共分散 covariance は

$$Cov(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} \quad (\mu_x = E(X), \mu_y = E(Y)) \tag{6}$$

で定義される。なお、共分散 $Cov(X, Y)$ は同時確率分布で

$$Cov(X, Y) = \begin{cases} = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - \mu_x)(y_j - \mu_y) f(x_i, y_j) & \text{離散型} \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dxdy & \text{連続型} \end{cases}$$

と表される。共分散は X と Y が, それぞれの平均 μ_x, μ_y からたがいに関連しながら, ばらつく程度を表す.

■ $Cov(X, Y) > 0$ なら, X, Y は大小が同傾向, < 0 なら反対傾向の関係を表す (即ち、関係の方向).

» 共分散

性質

* $V(X+Y) = V(X) + V(Y) + 2Cov(X, Y)$。
例えば、株式投資で、A 石油の株式と B 石油の株式というように, 同一業種の株式に同時に投資することは, 一般に勧められない. なぜなら, エネルギー危機など共通の経済的要因によって A も B も同傾向に連動するから, $Cov(X, Y) > 0$ となり, 単独の分散（ばらつき）の和 $V(X) + V(Y)$ 以上にばらつくからである.
* $V(aX+bY) = a^2 V(X) + 2abCov(X, Y) + b^2 V(Y)$
* $Cov(X, Y) = E(XY) - E(X)E(Y)$

» 共分散　資産運用への応用：ポートフォリオ選択

異なる複数の資産（投資対象）を組み合わせて形成される資産総額をポートフォリオ(portfolio)という。
今、二つの資産 A_X, A_Y でポートフォリオ A_W を作る問題を考える。

- A_X の保有比率を r、A_Y の保有比率を $(1-r)$ とおく。各資産の収益率を確率変数 (X, Y) で表せば、このポートフォリオのリターン（収益）は

$$W = rX + (1-r)Y \tag{7}$$

  となる、資産保有者が、比率 r を自由に決める。
- 個別資産のリターン X, Y の期待値・分散・共分散を
  $\mu_X = E(X), \mu_Y = E(Y), \sigma_X = V(X), \sigma_Y = V(Y), \sigma_{XY} = Cov(X, Y)$ とおく。
- W の期待値 $E(W)$ を期待リターン、分散 $V(W)$ をリスクと呼ぶ。
- ポートフォリオ A_W の期待リターンは $\mu_W = r\mu_X + (1-r)\mu_Y$ となる。
- ポートフォリオ A_W のリスクは
  $\sigma_W^2 = V(rX + (1-r)Y) = r^2\sigma_X^2 + (1-r)^2\sigma_Y^2 + 2r(1-r)\sigma_{XY}$

» 共分散 資産運用への応用：ポートフォリオ選択 (続き)

資産	A_X	A_Y	ポートフォリオ A_W(A_X の比率は r)
リターン	X	Y	$W = rX + (1-r)Y$
期待リターン	μ_X	μ_Y	$\mu_W = r\mu_X + (1-r)\mu_Y$
リスク	σ_X	σ_Y	$\sigma_W^2 = r^2\sigma_X^2 + (1-r^2)\sigma_Y^2 + 2r(1-r)\sigma_{XY}$

次の基準に基づくポートフォリオ選択を平均・分散基準という。保有比率 r の異なる二つのポートフォリオ A_{W1} (リターン = W1)、A_{W2} (リターン = W2) について

1. 高リターン選好なら：$\sigma_{W1}^2 = \sigma_{W2}^2, \mu_{W1} > \mu_{W2}$　→W1 を採用。
2. 低リスク選好なら：$\mu_{W1} = \mu_{W2}, \sigma_{W1}^2 < \sigma_{W2}^2$　→W1 を採用。

質問：もしポートフォリオリスクを最小化にしたいなら、比率 r をいくらにするか？
(Hint: このアプローチは最小分散ポートフォリオと呼ぶ。)

» 共分散 資産運用への応用：ポートフォリオ選択 (続き)

資産	A_X	A_Y	ポートフォリオ A_W(A_X の比率は r)
リターン	X	Y	$W = rX + (1-r)Y$
期待リターン	μ_X	μ_Y	$\mu_W = r\mu_X + (1-r)\mu_Y$
リスク	σ_X	σ_Y	$\sigma_W^2 = r^2\sigma_X^2 + (1-r^2)\sigma_Y^2 + 2r(1-r)\sigma_{XY}$

次の基準に基づくポートフォリオ選択を平均・分散基準という。保有比率 r の異なる二つのポートフォリオ A_{W1} (リターン = W1)、A_{W2} (リターン = W2) について

1. 高リターン選好なら：$\sigma_{W1}^2 = \sigma_{W2}^2, \mu_{W1} > \mu_{W2}$　→W1 を採用。
2. 低リスク選好なら：$\mu_{W1} = \mu_{W2}, \sigma_{W1}^2 < \sigma_{W2}^2$　→W1 を採用。

質問：もしポートフォリオリスクを最小化にしたいなら、比率 r をいくらにするか？
(Hint: このアプローチは最小分散ポートフォリオと呼ぶ。)

$$\frac{d\sigma_W^2}{dr} = 0 \text{と置く} \Longrightarrow r^* = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

この r が最小分散ポートフォリオにおける資産 A_X の最適比率となる.

» 相関係数

相関係数 X, Y の関係の強さの程度を判断する基準である. 共分散 $Cov(X, Y)$ の値を標準偏差で割って, 確率変数 X, Y の**相関係数**を

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{V(X)} \cdot \sqrt{V(Y)}} \tag{8}$$

で定義される。

- $\rho > 0$ なら X, Y は同じ大小の向きに変化する傾向があり, $\rho < 0$ なら逆である.
- もっとも極端な場合は, $\rho = \pm 1$ であり, このときは X, Y の間には厳密に次の 1 次式の関係が成り立つ:

$$Y = aX + b \qquad \text{ただし } \rho = 1 \text{ なら、} a > 0、\rho = -1 \text{ なら、} a < 0$$

- 逆に $\rho = 0$(つまり $Cov(X, Y = 0)$ の場合は X, Y はどちらの関係をもつともいえない. この場合, X, Y は無相関という. 無相関とは,「関連がない」.

» 計算例

同時確率分布

$$f(x,y)\begin{cases} = 6(x-y), & 0 \le y < x \le 1 \\ = 0, & \text{それ以外} \end{cases}$$

を持つ X, Y の周辺確率分布、X の期待値、分散、および X, Y の相関係数を求めよう. ここ $(E(Y) = 1/4, V(Y) = 3/80, E(XY) = 1/5)$ はすでに算出したとする.

» 計算例

同時確率分布

$$f(x,y)\begin{cases} = 6(x-y), & 0 \le y < x \le 1 \\ = 0, & \text{それ以外} \end{cases}$$

を持つ X, Y の周辺確率分布、X の期待値、分散、および X, Y の相関係数を求めよう. ここ $(E(Y) = 1/4, V(Y) = 3/80, E(XY) = 1/5)$ はすでに算出したとする.

$$f_x(x) = \int_0^x 6(x-y)dy = 3x^2$$

$$E(X) = \int_0^1 x \cdot 3x^2 dx = 3/4$$

$$Cov(X,Y) = 1/5 - (3/4)(1/4) = 1/80$$

$$\rho = \frac{1/80}{\sqrt{3/80} \cdot \sqrt{3/80}} = 1/3$$

目次

* 復習
* 同時確率分布と周辺確率分布
* 条件付確率分布と独立な確率変数
* 独立な確率変数の和

» 条件付確率分布

条件付確率関数 Y が y と与えられたときの X の条件付確率関数

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} \tag{9}$$

と定義される。同じで、X が x と与えられたときの Y の条件付確率関数 $f(y|x) = \frac{f(x,y)}{f_X(x)}$ と定義される。$|$ の後が条件，前が変数である．即ち、$f(x|y)$ は x の関数で，他方 y は指定された値に固定されている．

- x で和をとると，$\sum_x f(x|y) = \frac{\sum_x f(x,y)}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1$ が成立し，確かに確率分布の条件を満たしている．このように，条件付確率分布もひとつの確率分布である．
- その条件付期待値は $E(X|y) = \sum_x xf(x|y) = \mu_{x|y}$，

条件付分散は $V(X|y) = \sum_x (x - \mu_{x|y})^2 f(x|y)$ で計算できる。

- 連続型も同じの考え方で計算できる。

» 条件付確率分布

条件付期待値計算例

下記の X の条件付き確率分布を用いて, $E(X|3), E(X|5)$ を求めよう.

表; X の条件付き確率分布

x	1	2	3	4	5	6
$g(x\|3)$	0	0	1/7	2/7	2/7	2/7

x	1	2	3	4	5	6
$g(x\|5)$	0	0	0	0	1/3	2/3

条件付期待値は $E(X|y) = \sum_x x \cdot g(x|y) = \mu_{x|y}$ で計算できるより;

» 条件付確率分布

条件付期待値計算例

下記の X の条件付き確率分布を用いて, $E(X|3), E(X|5)$ を求めよう.

表; X の条件付き確率分布

x	1	2	3	4	5	6
$g(x\|3)$	0	0	1/7	2/7	2/7	2/7

x	1	2	3	4	5	6
$g(x\|5)$	0	0	0	0	1/3	2/3

条件付期待値は $E(X|y) = \sum_x x \cdot g(x|y) = \mu_{x|y}$ で計算できるより;

$$E(X|3) = 3 \cdot (1/7) + 4 \cdot (2/7) + 5 \cdot (2/7) + 6 \cdot (1/7) = 33/7$$

$$E(X|5) = 3 \cdot (1/3) + 6 \cdot (2/3) = 17/3$$

» 独立な確率変数

独立な確率変数 同時確率分布において、あらゆる x, y について、条件

$$f(x, y) = f_x(x) f_y(y) \tag{10}$$

が成り立つならば, X, Y はたがいに**独立**であるという.

- 独立のときは, X, Y の同時確率分布は X, Y それぞれの確率分布（周辺確率分布）を知るだけで求められる.
- 式10を式9に代入すると、

$$f(x|y) = f_x(x) \tag{11}$$

となって X の出方は y によらないことがわかる. 同様に, $f(y|x) = f_y(y)$.

- 一般には,

$$f(x, y) = f_x(x) f(y|x) \tag{12}$$

が成り立つ. 同様に, $f(x, y) = f_y(y) f(x|y)$.

» 独立な確率変数

独立性から導出できる性質

* X, Y が独立ならば

$$E(XY) = E(X)E(Y) \tag{13}$$

が成立する.

* X, Y が独立ならば

$$Cov(X, Y) = 0 \tag{14}$$

が成り立つ。即ち、独立なら無相関であるが、逆に、無相関でも独立とは限らない.

目次

* 復習
* 同時確率分布と周辺確率分布
* 条件付確率分布と独立な確率変数
* 独立な確率変数の和

» 独立な確率変数の和 (1)

統計学では, ランダムな誤差を含んだ観測値, 測定値, データなどの集計を行うが, そのなかでも和をとることは基本的な操作である. 確率変数の和 $X+Y$ については, その確率分布は意外と求めにくい. 独立性があるときは, 事はやや扱いやすくなる.

分散の加法性 X, Y が独立であるときには, 分散の加法性など

$$V(X \pm Y) = V(X) + V(Y) \tag{15}$$

が成立する。

n 個の場合 n 個の確率変数 $X_1, X_2, .., X_n$ に対しても同じく、$X_1, X_2, .., X_n$ が独立のときには、

$$V(X_1 + X_2 + .. + X_n) = V(X_1) + V(X_2) + ... + V(X_n) \tag{16}$$

も成立する。

» 独立な確率変数の和 (2)

同一分布 $X_1, X_2, .., X_n$ が同一の（共通の）確率分布に従うとし，これらの期待値，分散を μ, σ^2 とすれば，$E(X_1) = E(X_2) = ... = E(X_n) = \mu$, $V(X_1) = V(X_2) = ... = V(X_n) = \sigma^2$ であるから，

$$E(X_1 + X_2 + .. + X_n) = n\mu, \qquad V(X_1 + X_2 + .. + X_n) = n\sigma^2 \quad (17)$$

が成立する。したがつて，標準偏差は，$D(X_1 + X_2 + .. + X_n) = \sqrt{n}\sigma$ となる．

相加平均 $X_1, X_2, .., X_n$ を n で割った相加平均を $\bar{X} = \frac{(X_1+X_2+..+X_n)}{n}$ とおくと、

$$E(\bar{X}) = \mu, \qquad V(\bar{X}) = \frac{\sigma^2}{n} \quad (18)$$

を得る．この結果は重要である．相加平均 $\bar{X}$ は，期待値は n に無関係に常に μ に一致するが，分散は n に反比例して減少し 0 に収束する．

» 独立な確率変数の和 (3)

和の確率分布　離散型の確率変数 X, Y が独立であるとし，その確率分布を $g(x), h(y)$ としよう．和 $X+Y$ の確率分布 $k(z)$ は確率 $P(X+Y=z)$ を考えれば得られる．
関数 $k(z)$ は下記となる：

$$k(z) = \sum_x g(x)h(z-x) \tag{19}$$

関数 g, h から k を作る数学操作を g, h のたたみこみ convolution といい，$k = g * h$ と書く．g, h が密度関数のときも同様で，たたみこみは積分となる。

再生的
1. 二項分布　$Bi(n,p) * Bi(m,p) = Bi(n+m,p)$
2. ポアソン分布　$Po(\lambda 1) * Po(\lambda 2) = Po(\lambda 1 + \lambda 2)$
3. 正規分布　$N(\mu_1, \sigma_1^2) * N(\mu_2, \sigma_2^2) = N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

このように，たたみこみの結果として，ふたたび同一種類の確率分布が得られる場合，この確率分布は再生的という．

» 独立な確率変数の和 (4)

正規分布の再生性 正規分布については，再生性は大きなメリットがあり，統計的推測によく用いられているのであらかじめ知っておくとよい．

* $X_1, X_2, .., X_n$ が独立で，それぞれ正規分布 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2), .., N(\mu_n, \sigma_n^2)$ 従っているならば，
  1. $X_1 + X_2 + .. + X_n$ は $N(\mu_1 + \mu_2 + ... + \mu_n, \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2)$ 従い，
  2. $c_1X_1 + c_2X_2 + .. + c_nX_n$ は $N(c_1\mu_1 + c_2\mu_2 + ... + c_n\mu_n, c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + ... + c_n^2\sigma_n^2)$ 従う。
* とくに $X_1, X_2, .., X_n$ の確率分布がすべて正規分布 $N(\mu, \sigma^2)$ なら，
  1. $X_1 + X_2 + .. + X_n$ は $N(n\mu, n\sigma^2)$ に従い，
  2. $\bar{X} = \frac{(X_1+X_2+..+X_n)}{n}$ は $N(\mu, \frac{\sigma^2}{n})$ に従う。