基礎勉強会第7回(多分)P47∼P54

新B4 福田真悟

2020.3.3

3　多次元の確率分布

3.1　2次元の確率分布

3.1.1　同時確率分布

2つの確率変数X, Yを定義する.この2つの変数が離散型で$X = x_i, Y = y_j$となる確率は，

$$P(X = x_i, Y = y_j) = f(x_i, y_j) \tag{1}$$

となる．これを同時確率分布という．連続型のときは，確率を範囲で考えるので$(x, y), (x + \Delta x, y), (x, y + \Delta y), (x + \Delta x, y + \Delta y)$の4点で囲まれた長方形に入る確率を考えるので，

$$f(x, y) = \frac{P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)}{\Delta x \Delta y} \tag{2}$$

$\Delta x \to 0, \Delta y \to 0$と極限をとったときの$f(x, y)$を同時確率密度関数という． 2変数における累積分布関数は，連続型と離散型でそれぞれ，

$$F(x, y) = P(X \leq x, Y \leq y) = \begin{cases} \displaystyle\sum_{u \leq x} \sum_{v \leq y} f(u, v) & \text{(離散型)} \\ \displaystyle\int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv & \text{(連続型)} \end{cases} \tag{3}$$

となる．これを同時分布関数または，同時累積分布関数という．

3.1.2　共分散と相関係数

2つの確率分布が存在する場合に知りたい情報として，その2つの情報の関係性がある．そのときに関係性を表す指標として，共分散と相関係数が用いられる．まず共分散は，

$$\mathrm{cov}(X, Y) = \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] \tag{4}$$

で書かれる．また，2変数のときの期待値演算子Eは，

$$E[\varphi(X, Y)] = \begin{cases} \displaystyle\sum_{x} \sum_{y} \varphi(x, y) f(x, y) & \text{(離散型)} \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x, y) f(x, y) dx dy & \text{(連続型)} \end{cases} \tag{5}$$

となる．よって，式(4)は，

$$\sigma_{XY} = \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y) & (\text{離散型}) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y) dxdy & (\text{連続型}) \end{cases} \tag{6}$$

となる．　相関係数は，共分散をX, Yの個々のばらつき(分散)に依存しない(スケール不変性)で2変数の関係性のみを評価するために基準化を行ったものである．式は，

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{7}$$

となる．個々の確率変数のスケールを$X = aU + b, Y = cV + d$と変えたとき，共分散と各標準偏差は，

$$\sigma_{UV} = ac\sigma_{XY} \tag{8}$$

$$\sigma_U = a\sigma_X \tag{9}$$

$$\sigma_V = c\sigma_Y \tag{10}$$

となる．よってスケールをU, Vに変えたときの相関係数は，

$$\rho = \frac{ac\sigma_{XY}}{a\sigma_X c\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{11}$$

となり，スケールを変えても相関係数は変化しないことがわかる．具体的にイメージするために確率変数X, Yの条件のもと，試行をN回繰り返したときのi回目の試行の結果を(x_i, y_i)とする．ここでのNは十分に大きい整数とする．X, Yの結果とそれぞれの平均値との差のベクトル$\mathbf{X} = (x_1 - \mu_X, x_2 - \mu_X, \cdots, x_N - \mu_X), \mathbf{Y} = (y_1 - \mu_Y, y_2 - \mu_Y, \cdots, y_N - \mu_Y)$としたとき，$\sigma_{XY}$は，大数の法則弱より，内積$\mathbf{X} \cdot \mathbf{Y}$を回数$N$で除したものとなる．$\sigma_X^2, \sigma_Y^2$は，それぞれ$|\mathbf{X}|^\mathbf{2}, |\mathbf{Y}|^\mathbf{2}$を回数$N$で除したものとなる．このとき，相対係数$\rho$は，

$$\rho = \frac{X \cdot Y}{|X||Y|} \tag{12}$$

となる．ベクトルで考えたとき，二つのベクトル$\mathbf{X}, \mathbf{Y}$のなす角度をθとおいたとき，相対係数は，内積の公式から，

$$\rho = \cos\theta \tag{13}$$

となることがわかる．このため範囲が$-1 \leq \rho \leq 1$となる．また，このことから相関係数が示しているのはベクトルのなす角度であり，相関係数はあくまで線形の関係の$\mathbf{Y} = A\mathbf{X} + B$があるかどうかを示していることがわかる．また，相対係数の結果を比較を行うとき$\cos\theta$のため，大小の比較しか行えないことがわかる．

3.1.3　周辺確率分布，条件付確率分布および独立

- 周辺確率

XとYの個々の確率分布を周辺確率分布という．それぞれ離散型と連続型は，

$$g(x) = \begin{cases} \sum_y f(x,y) & (\text{離散型}) \\ \int_{-\infty}^{\infty} f(x,y) dy & (\text{連続型}) \end{cases} \tag{14}$$

$$h(y) = \begin{cases} \sum_{x} f(x,y) & \text{(離散型)} \\ \int_{-\infty}^{\infty} f(x,y)dx & \text{(連続型)} \end{cases} \tag{15}$$

となる.

- 条件付確率

  条件付確率は,

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \tag{16}$$

  となる．周辺確率と合わせると,

$$g(x|y) = P(X = x|Y = y) = \frac{f(x,y)}{h(y)} \tag{17}$$

  となる．これは連続型でも拡張ができ，条件付確率密度関数という．また，条件付確率も総和は1となる条件を満足する．条件付確率の期待値，分散は，それぞれの確率密度関数を$f(x,y)$から$g(x|y)$などに変えることで導出できる.

- 独立

  独立のときは，条件付確率がそれぞれ,

$$g(x|y) = g(x), h(y|x) = h(y) \tag{18}$$

  となる．この式から,

$$f(x,y) = g(x)h(y) \tag{19}$$

  となる．上記の式とX, Yが独立であることは必要十分条件である.

- 独立の場合の相関係数

$$\begin{aligned} \mathrm{cov}(X,Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)g(x)h(y)dxdy \\ &= \int_{-\infty}^{\infty}(x-\mu_X)g(x)dx\int_{-\infty}^{\infty}(y-\mu_Y)h(y)dy \\ &= 0 \end{aligned} \tag{20}$$

  となるため，相関係数は0となる．相関係数が0のとき，X, Yが独立になるわけではない．反例として，$Y = X^2$の相関がある場合が挙げられる．独立の場合,

$$E[\varphi(X)\psi(Y)] = E[\varphi(X)]E[\psi(Y)] \tag{21}$$

  が成り立つ.

3.1.4 確率変数の和の分布と期待値と分散

- 確率変数の和の分布

  Zを2つの確率変数X, Yの和$Z = X + Y$としたときのZの確率密度関数は,

$$k(z) = \begin{cases} \sum_{x} f(x, z-x) & \text{(離散型)} \\ \int_{-\infty}^{\infty} f(x, z-x)dx & \text{(連続型)} \end{cases} \tag{22}$$

となる．特にX, Yが独立の場合は，

$$k(z) = \begin{cases} \sum_x g(x)h(z-x) & (\text{離散型}) \\ \int_{-\infty}^{\infty} g(x)h(z-x)dx & (\text{連続型}) \end{cases} \tag{23}$$

となり，これをたたみこみといい，$k = g * h$で表現される．

- 再生性

一般的に確率変数の和の分布を導出するには，式(22)を用いるが，一部の分布においては独立な確率変数の和は同一の分布形になるものがある．このような分布を再生的という．二項分布，ポアソン分布，負の二項分布，正規分布，ガンマ分布などがある．各分布に関しての再現性について下記に記載する．

まず二項分布について．$X \sim Bi(n_1, P), Y \sim Bi(n_2, P) \Rightarrow Z = X + Y \sim Bi(n_1 + n_2, P)$となる．$X, Y, Z$のそれぞれは，

$$g(x) = {}_{n_1}C_x P^x (1-P)^{n_1 - x} \tag{24}$$

$$h(z-x) = {}_{n_2}C_{z-x} P^{z-x} (1-P)^{n_2 - (z-x)} \tag{25}$$

$$k(z) = \sum_x {}_{n_1}C_x \cdot {}_{n_2}C_{z-x} P^z (1-P)^{n_1 + n_2 - z} \tag{26}$$

となる．ここで

$${}_{n_1+n_2}C_z = \sum_x {}_{n_1}C_x \cdot {}_{n_2}C_{z-x} \tag{27}$$

となるので，式(26)に代入すると，

$$k(z) = {}_{n_1+n_2}C_z P^z (1-P)^{n_1+n_2-z} = Bi(n_1 + n_2, P) \tag{28}$$

となる．また，負の二項分布も同様である．

次にポアソン分布について．$X \sim Po(\lambda_1), Y \sim Po(\lambda_2) \Rightarrow Z = X + Y \sim Po(\lambda_1 + \lambda_2)$となる．$X, Y, Z$のそれぞれは，

$$g(x) = \frac{\mathrm{e}^{-\lambda_1} \lambda_1^x}{x!} \tag{29}$$

$$h(z-x) = \frac{\mathrm{e}^{-\lambda_2} \lambda_2^{z-x}}{(z-x)!} \tag{30}$$

$$k(z) = \sum_x \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)} \lambda_1^x \lambda_2^{z-x}}{x!(z-x)!} \tag{31}$$

となる．ここで${}_zC_x = \frac{z!}{x!(z-x)!}$を用いると式(31)は，

$$k(z) = \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)}}{z!} \sum_x {}_zC_x \lambda_1^x \lambda_2^{z-x} \tag{32}$$

となり，二項定理から，

$$\sum_x {}_zC_x \lambda_1^x \lambda_2^{z-x} = (\lambda_1 + \lambda_2)^z \tag{33}$$

となるので代入すると，

$$k(z) = \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)} (\lambda_1 + \lambda_2)^z}{z!} = Po(\lambda_1 + \lambda_2) \tag{34}$$

となる．

次に正規分布について．$X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2) \Rightarrow Z = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$となる．$X, Y, Z$のそれぞれは，

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right] \tag{35}$$

$$g(z-x) = \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left[\frac{-((z-x)-\mu_2)^2}{2\sigma_2^2}\right] \tag{36}$$

$$k(z) = \int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[\frac{-(x-\mu_1)^2}{2\sigma_1^2} + \frac{-((z-x)-\mu_2)^2}{2\sigma_2^2}\right]dx \tag{37}$$

となる．ガウス積分などを用いて計算を行うと，

$$g(x) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\exp\left[\frac{-(z-(\mu_1+\mu_2))^2}{2(\sigma_1^2+\sigma_2^2)}\right] = N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2) \tag{38}$$

となる．

最後にガンマ分布について．$X \sim Ga(\alpha_1, \beta), Y \sim Ga(\alpha_2, \beta) \Rightarrow Z = X + Y \sim Ga(\alpha_1 + \alpha_2, \beta)$となる．$X, Y, Z$のそれぞれは，

$$g(x) = \frac{1}{\beta^{\alpha_1}\Gamma(\alpha_1)}x^{\alpha_1-1}\mathrm{e}^{-\frac{x}{\beta}} \tag{39}$$

$$h(z-x) = \frac{1}{\beta^{\alpha_2}\Gamma(\alpha_2)}(z-x)^{\alpha_2-1}\mathrm{e}^{-\frac{z-x}{\beta}} \tag{40}$$

$$k(z) = \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1)\Gamma(\alpha_2)}\mathrm{e}^{-\frac{z}{\beta}}\int_0^z x^{\alpha_1-1}(z-x)^{\alpha_2-1}dx \tag{41}$$

となる．積分区間については，ガンマ分布が$x < 0$の範囲では0となるので，[0,z]の区間になっている．ここでベータ関数$B(\alpha_1, \alpha_2)$を導入すると，

$$B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1}(1-t)^{\alpha_2-1}dt \tag{42}$$

となる．ここで$t = x/z$(zは定数，xは変数)で置換すると，

$$B(\alpha_1, \alpha_2) = \int_0^z z^{-(\alpha_1-1)}x^{\alpha_1-1}z^{-(\alpha_2-1)}(z-x)^{\alpha_2-1}z^{-1}dx \tag{43}$$

となる．まとめると，

$$B(\alpha_1, \alpha_2) = \frac{1}{z^{\alpha_1+\alpha_2-1}}\int_0^z x^{\alpha_1-1}(z-x)^{\alpha_2-1}dx \tag{44}$$

となる．さらにベータ関数とガンマ関数の間には下記の等式が成り立つ．

$$B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)} = \frac{1}{z^{\alpha_1+\alpha_2-1}}\int_0^z x^{\alpha_1-1}(z-x)^{\alpha_2-1}dx \tag{45}$$

となる．式(45)を式(41)に代入すると，

$$k(z) = \frac{1}{\beta^{\alpha_1+\alpha_2}\Gamma(\alpha_1+\alpha_2)}z^{\alpha_1+\alpha_2-1}\mathrm{e}^{-\frac{z}{\beta}} = Ga(\alpha_1+\alpha_2, \beta) \tag{46}$$

- 確率変数の和の期待値と分散

  確率変数の和の期待値は，

$$E(Z) = E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy = E(X)+E(Y) = \mu_X + \mu_Y \tag{47}$$

  となり，期待値もX, Yの和となる．一方で分散は，

$$\begin{aligned} V(X+Y) &= E[\{(X+Y)-(\mu_X+\mu_Y)\}^2] \\ &= E[(X-\mu_X)^2] + E[(Y-\mu_Y)^2] + 2E[(X-\mu_X)(Y-\mu_Y)] \\ &= V(X) + V(Y) + 2\mathrm{cov}(X,Y) \\ &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \end{aligned} \tag{48}$$

  となり，ただの和にはならないので注意が必要である．

参考文献

[1] 東京大学工学教程基礎系数学確率統計Ⅰ，縄田和満，平成25年10月10日