2026 年度　データサイエンス数学基礎/統計学　（担当：松澤　寛）　レジメ

No.10　母集団分布と標本分布

- 全国の中学 1 年生の身長の分布や平均，分散などを調べる際には，全員を調査するのではなく，その一部を抜き出して調べる標本調査と呼ばれる手法が用いられる．全国の中学 1 年生全体（これを母集団という）から 1 人を無作為に選ぶと，その人の身長は選ばれるまで分からないので，身長を確率変数とみなすことができる．さらに，100 人を無作為に選んだとき，それぞれの身長を順に $X_1, \cdots, X_{100}$ とおくと、100 個の確率変数を扱うことになる．このとき，ある生徒が選ばれたという事実が，その後に選ばれる生徒に影響を与えるかどうかは重要な問題である．一度選ばれた生徒を再び抽出の対象とする方法を**復元抽出**といい、一度選ばれた生徒を以後の抽出対象から除く方法を**非復元抽出**という．非復元抽出では，先に誰が選ばれたかによって後の抽出結果の確率が変化する．一方，母集団の大きさが標本数に比べて十分大きい場合には，その影響は非常に小さくなり，復元抽出とみなしてよいことが多い．このような考察のために，次節では事象や確率変数の独立性について学ぶ．

確率変数の独立性

- 確率変数 X_1, X_2 について，**X_1 に関する任意の事象と X_2 に関する任意の事象が独立**であるとき，確率変数 X_1, X_2 は**互いに独立**であるという．
- 具体的には X_1, X_2 が離散型確率変数の場合，任意の a, b

$$P(X_1 = a, X_2 = b) = P(X_1 = a)P(X_2 = b)$$

  が成り立つことである．
- 連続型確率変数の場合は，任意の a, b, c, d $(a < b, c < d)$ について

$$P(a \leq X_1 \leq b,\ c \leq X_2 \leq d) = P(a \leq X_1 \leq b)P(c \leq X_2 \leq d)$$

  が成り立つことである．
- $X_1, X_2, \cdots, X_n$ が**互いに独立**であるとは X_i と X_j $(i \neq j)$ が独立であることをいう．

注

- $P(X_1 = a, X_2 = b)$ は 事象 $X_1 = a$ と $X_2 = b$ が同時に起こる確率である．
- $P(a \leq X_1 \leq b, c \leq X_2 \leq d)$ は事象 $a \leq X_1 \leq b$ と $c \leq X_2 \leq d$ が同時に起こる確率である．

例 さいころを 2 回投げて，1 回目および 2 回目の目の数をそれぞれ X_1, X_2 とする．このとき X_1, X_2 は独立である．

共分散

X, Y を確率変数とし，$E[X] = \mu_X$, $E[Y] = \mu_Y$ とする．

$$E[(X - \mu_X)(Y - \mu_Y)]$$

を X, Y の**共分散**といい $\mathrm{Cov}(X, Y)$ と表す．

共分散の性質

(1) $\mathrm{Cov}(X,Y) = E[XY] - \mu_X\mu_Y$

(2) a, b を定数とするとき，$\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$

(3) $V[X+Y] = V[X] + \mathrm{Cov}(X,Y) + V[Y]$

(4) X と Y が独立ならば $E[XY] = E[X]E[Y] = \mu_X\mu_Y$ が成り立つ．特にこのとき $\mathrm{Cov}(X,Y) = 0$ が成り立つ．

(5) X と Y が独立ならば $V[X+Y] = V[X] + V[Y]$

証明

(1)

$$\begin{aligned}\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] &= E[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= E[XY] - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y\end{aligned}$$

(2) $E[aX] = a\mu_x, E[bY] = b\mu_y$ より，

$$\begin{aligned}\mathrm{Cov}(aX,bY) &= E[(aX - a\mu_x)(bY - b\mu_y)] \\ &= E[ab(X-a\mu_x)(Y-\mu_y)] \\ &= abE[(X-a\mu_x)(Y-\mu_y)] = ab\mathrm{Cov}(X,Y)\end{aligned}$$

(3) $E[X+Y] = \mu_x + \mu_y$ より，

$$\begin{aligned}V[X+Y] &= E[(X+Y-\mu_x-\mu_y)^2] \\ &= E[(\{X-\mu_x\} + \{Y-\mu_y\})^2] \\ &= E[\{X-\mu_x\}^2 + 2\{X-\mu_x\}\{Y-\mu_y\} + \{Y-\mu_y\}^2] \\ &= E[\{X-\mu_x\}^2) + 2E(\{X-\mu_x\}\{Y-\mu_y\}) + E(\{Y-\mu_y\}^2] \\ &= V[X] + 2\mathrm{Cov}(X,Y) + V[Y]\end{aligned}$$

(4) 離散型確率変数の場合を示す．X, Y が独立ならば,

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

が成り立つので

$$\begin{aligned}E[XY] &= \sum_{i=1}^{n}\sum j = 1^m x_i y_j P(X = x_i, Y = y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(X = x_i)P(Y = y_j) \\ &= \left(\sum_{i=1}^{n} x_i P(X = x_i)\right)\left(\sum_{j=1}^{m} y_j P(Y = y_j)\right) = E[X]Y[Y]\end{aligned}$$

(5) (4) より $\mathrm{Cov}(X,Y) = 0$ であるから (3) より成り立つ．

確率変数 $X_1, X_2, \cdots, X_n$ が互いに独立で, 同じ分布に従うとする．このときこの分布を**母集団分布**という．全国中学校 1 年生の身長やある製品の耐久時間なども母集団分布として考える．このとき n 個の確率変数 $X_1, \cdots, X_n$ を**標本**といい，$X_1, X_2, \cdots, X_n$ の関数のことを**推定量**という．例

- 標本平均 $\overline{X} = \dfrac{1}{n}\sum_i X_i$
- 不偏分散 $V = \dfrac{1}{n-1}\sum_i (X_i - \overline{X})^2$

標本平均の性質

$X_1, X_2, \cdots, X_n$ を互いに独立で平均 μ，分散 σ^2 である同じ確率分布に従う確率変数であるとする（つまり $\mu = E(X_i), \sigma^2 = V(X_i)$）．このとき

$$\textbf{標本平均 } \overline{X} = \frac{1}{n}\sum_i X_i$$

は

$$E[\overline{X}] = \mu, V[\overline{X}] = \frac{\sigma^2}{n}$$

を満たす．

証明 平均の性質から

$$E[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_i \mu = \mu$$

を得る．また，$X_1, \cdots, X_n$ は互いに独立だから，分散の性質から

$$V\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} V[X_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2.$$

である．したがって

$$V[\overline{X}] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} V\left[\sum_{i=1}^{n} X_i\right] = \frac{\sigma^2}{n}.$$

を得る．□

相関係数

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V[X]V[Y]}}$$

を X と Y の**相関係数**という．

注

- a, b を定数とするとき

$$\rho(aX, bY) := \frac{\mathrm{Cov}(aX, bY)}{\sqrt{V[aX][V[bY]}} = \frac{ab\mathrm{Cov}(X,Y)}{\sqrt{a^2V[X]b^2V[Y]}} = \rho(X,Y)$$

- $-1 \leq \rho(X,Y) \leq 1$ が成り立つ．実際，t を実数とするとき $u(t) := V[tX + Y]$ について，

$$\begin{aligned} 0 \leq u(t) &= V[tX+Y] \\ &= V[tX] + V[Y] + 2\mathrm{Cov}(tX, Y) \\ &= V[X]t^2 + 2\mathrm{Cov}(X,Y)t + V[Y] \end{aligned}$$

よって, 判別式より $\mathrm{Cov}(X,Y)^2 - V(X)V(Y) \leq 0$ 即ち, $|\rho(X,Y)| \leq 1$ を得る．

注 等号成立 $(|\rho(X,Y)| = 1) \Leftrightarrow u(t) = V[tX + Y] = 0$ を満たす t_0 が存在
即ち，定数 a, b を用いて $Y = aX + b$ と表されること