数理統計学まとめ（その 6）：第 4 章 母集団と標本

1 標本抽出

調査対象の集団を**母集団** という.（日本人全体，兵庫県在住の成人男子，神戸市の女子学生全体など）

調べ方

- 全部調べる（**悉皆調査**）：金も時間もかかる．(例：国勢調査）
- 標本の小規模な集団を抜き出して調べる（**標本調査**）：全体の動向を反映しているかが問題.（例：アンケート調査）

標本抽出（標本を母集団から取り出すこと.）

- **復元抽出**： 取り出した標本を元に戻してから次の標本を取り出す.
- **非復元抽出**：標本を元に戻さない.
- **有意抽出法**：調査者が自分の経験や知識によって標本を選ぶ．統計的推測はできない.
- **無作為抽出法**：標本を偶然性に基づいて取り出す.

統計的推測　無作為抽出法で取り出された標本 $X_1, \ldots, X_n$ はそれぞれ母集団の分布に従う確率変数で，互いに独立と考える．標本の値を使って母集団の分布を推測することを統計的推測と言う.

2 標本分布

標本のデータの関数 $T(X_1, \ldots, X_n)$ を（標本）**統計量**という．これは母集団の特性値（平均，分散など．**母数**と呼ぶ.）を推測するために使う．標本統計量 $T(X_1, \ldots, X_n)$ の分布を標本分布と言う．標本の総数 n を標本の**大きさ**という.

- 母平均 μ に対して 標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 母分散 σ^2 に対して標本分散 $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

3　標本平均の分布

標本平均 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ の標本分布については次が言える．

定理 3.1　（教科書 p.83 定理 4.1）
母平均 μ 母分散 σ^2 の母集団から無作為に抽出された大きさ n の標本を $(X_1, \ldots, X_n)$ とする．このとき標本平均 $\bar{X}$ の分布について次が成立．

$$E(\bar{X}) = \mu, \qquad (\bar{X} \text{ の平均})$$
$$V(\bar{X}) = \frac{\sigma^2}{n}. \qquad (\bar{X} \text{ の分散})$$

証明には次の定理が基本的

定理 3.2　（教科書 p.73 定理 3.5 と定理 3.7）
X, Y を確率変数とする．

(1) X, Y が独立ならば，その共分散 $\mathrm{Cov}(X, Y)$ は 0.

(2) $V(X+Y) = V(X) + V(Y) - 2\mathrm{Cov}(X, Y)$ でしたがって X, Y が独立ならば $V(X+Y) = V(X) + V(Y)$.

証明　(1) 第 3 章定理 5.1 教科書 p.73 定理 3.5 として前回紹介した．証明しておく．
X, Y の平均をそれぞれ μ_X, μ_Y と書く．

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

X, Y が独立なので，$E(XY)$ は

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(X = x_i)P(Y = y_j) = \left(\sum_{i=1}^{n} x_i P(X = x_i)\right)\left(\sum_{j=1}^{m} y_j P(Y = y_j)\right)$$

に等しく，これは $E(X)E(Y) = \mu_X\mu_Y$ と等しい．$\therefore \mathrm{Cov}(X, Y) = 0$.
(2) については，

$$\begin{aligned} V(X+Y) &= E\left(\{(X+Y) - (\mu_X + \mu_Y)\}^2\right) \\ &= E\left((X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2\right) \\ &= V(X) + 2\mathrm{Cov}(X, Y) + V(Y). \end{aligned}$$

定理 3.1 の証明をしよう．

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{n\mu}{n} = \mu.$$

だから $\bar{X}$ の平均は μ になる．一方，$X_1, \ldots, X_n$ が独立なので，定理 3.2 により，

$$\begin{aligned} V(\bar{X}) &= E\left(\left\{\frac{1}{n}\sum_{j=1}^{n}(X_j - \mu)\right\}^2\right) \\ &= \frac{1}{n^2}\sum_{j=1}^{n} V(X_j) + \frac{2}{n^2}\sum_{1\le i<j\le n} \mathrm{Cov}(X_i, X_j) \\ &= \frac{1}{n}\sigma^2. \end{aligned}$$

定理 3.3 （**大数の法則**, 教科書 p.84 定理 4.2）
母集団の平均を μ 分散を σ^2 とする．このときこの母集団からとった標本 $X_1, \ldots, X_n$ について，任意の小さな $\varepsilon > 0$ に対して

$$P(|\bar{X} - \mu| > \varepsilon)$$

は $n \to \infty$ のとき 0 に近づく．

定理 3.4 （教科書 p.84, 定理 4.3）
母集団が正規分布に従えば，無作為標本の標本平均 $\bar{X}$ も正規分布に従う

定理 3.5 （**中心極限定理**, 教科書 p.85 定理 4.4）
母集団の平均を μ，分散を σ^2 とする．どちらも有限な値とする．このときこの母集団からとった標本 $X_1, \ldots, X_n$ について，

$$Z = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

の分布は $n \to \infty$ のとき，標準正規分布に限りなく近づく．

例 3.1 （教科書 p.85 例題 4.5）確率変数 X が 2 項分布 $B_N(20, 0.4)$ に従うとき，

$$P(X \le 10) = 0.8725$$

を得るが，この確率を正規分布による近似から求めよ．

解答　まず正規分布は連続な分布であり，Z が標準正規分布のとき $P(Z \leq a)$ は密度関数 $\dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}$ と x 軸と $x=a$ とで囲まれた面積を表す．従って，考えている 2 項分布でもそのヒストグラムで $x = 10$ を含む階級 $(9.5 \sim 10.5)$ 以下に対応する面積を対応させる．

従って考えるのは $P(X \leq 10.5)$ である．

$B_N(20, 0.4)$ は確率 0.4 の (0-1) 分布に従う独立な確率変数の 20 個の和 R の分布として現れる．確率 p の (0-1) 分布の平均は p 分散は $p(1-p)$ (標準偏差は $\sigma = \sqrt{p(1-p)}$) だから，中心極限定理により

$$Z = \sqrt{20}\left(\frac{R}{20} - 0.4\right)\frac{1}{\sqrt{0.24}} = \frac{R-8}{\sqrt{4.8}}$$

の分布は標準正規分布で近似できる．したがって

$$R \leq 10.5 \Leftrightarrow Z \leq \frac{2.5}{\sqrt{4.8}}$$

右辺は小数第 2 位まで求めると 1.14.（第 3 位まで求めると 1.137）標準正規分布表で対応する値は

$$P(0 \leq Z \leq 1.14) = 0.3729$$

比例配分で第 3 位までの値に対応させると

$$P(0 \leq Z \leq 1.137) = 0.3708 + (0.3729 - 0.3708) \times 0.7 = 0.3723$$

したがって求める確率はこれに $P(Z \leq 0) = 0.5$ を加えて $P(R \leq 10.5)$ は 0.8729 (比例配分で求めると 0.8723) と近似できている．