統計学2及び演習

適合度検定とその例 (1)

創域理工学部
Faculty of Science and Technology

東京理科大学
創域理工学部情報計算科学科
安藤宗司

2023年6月7日

Contents

- さいころ投げ
- 多項分布
- 適合度検定
- ピアソンのカイ二乗統計量

さいころ投げ

- さいころの各目（1から6）が出る確率は1/6かどうか
- 実際にさいころをn回投げて，確かめる実験を考える

- 各目が出た回数

$$x_i \ (i = 1, 2, 3, 4, 5, 6)$$

- 各目が出た割合

$$p_i = \frac{x_i}{n} \ \ (i = 1, 2, 3, 4, 5, 6)$$

モチベーション

- さいころを120回繰り返しふった結果

出目	1	2	3	4	5	6	計
観測度数	16	28	16	15	25	20	120

  - この結果から正しいサイコロ（さいころの各目が出る確率は1/6かどうか）かを判定したい
  - さいころの各目が出る確率が1/6 のとき，期待度数はどの出目に対しても20になる
- 得られた観測度数と期待度数を比較し，仮定したさいころの分布がデータに適合しているかを検定したい

記号の定義

- 1回の試行でC個の排反な事象 $E_1, E_2, \dots, E_C$
- 事象E_i $(i = 1,2, \dots, C)$が生じる確率 $P(E_i) = \pi_i$
- 試行をn回繰り返し，事象E_iが生じた回数 X_i

排反な事象	E_1	E_2	$\cdots$	E_C	計
観測度数	X_1	X_2	$\cdots$	X_C	n
確率	π_1	π_2	$\cdots$	π_C	1
期待度数	$n\pi_1$	$n\pi_2$	$\cdots$	$n\pi_C$	

適合度検定

- 母集団をC個の排反な事象 $E_1, E_2, \ldots, E_C$に分割
- それらの事象が起こった観測度数と仮定する分布のもとでそれらの事象が起きる期待度数を比較
- 仮定した分布が適合しているかどうかを検定すること

多項分布

- 二項分布の一般化
- 多次元ベルヌーイ試行
  - 事象 i $(i = 1, \ldots, C)$ が起こる確率 π_i $(\pi_1 + \cdots + \pi_C = 1)$
- 試行回数 n
- 事象 i が起こった回数 X_i
- 確率変数ベクトル $\boldsymbol{X} = (X_1, \ldots, X_C)$

確率関数

$$\Pr(X_1 = x_1, \ldots, X_C = x_C) = \frac{n!}{x_1! \cdots x_C!} \pi_1^{x_1} \cdots \pi_C^{x_C} \qquad n = x_1 + \cdots + x_C$$

多項分布の性質

- 期待値 $E[X_i] = n\pi_i$

- 分散 $V[X_i] = n\pi_i(1 - \pi_i)$

- 共分散 $\mathrm{Cov}\left[X_i, X_j\right] = -n\pi_i\pi_j \ \ (i \neq j)$

$E[X_i]$ と $V[X_i]$ の導出

$S = \{ s \mid s \neq i, s = 1, \ldots C \}$

$X_i \; (i = 1, \ldots, C)$ の周辺確率関数

$$\Pr(X_i = x_i) = \sum_{s \in S} \sum_{x_s = 1}^{n - x_i} \frac{n!}{x_1! \cdots x_C!} \pi_1^{x_1} \cdots \pi_C^{x_C}$$

$$= \frac{n! \, (1 - \pi_i)^{n - x_i}}{(n - x_i)! \, x_i!} \pi_i^{x_i} \sum_{s \in S} \sum_{x_s = 1}^{n - x_i} \frac{(n - x_i)!}{\prod_{s \in S} x_s!} \prod_{s \in S} \left(\frac{\pi_s}{1 - \pi_i} \right)^{x_s}$$

$$= \binom{n}{x_i} \pi_i^{x_i} (1 - \pi_i)^{n - x_i}$$

X_i の周辺確率関数は，二項分布の確率関数に一致することから

$$E[X_i] = n\pi_i, V[X_i] = n\pi_i(1 - \pi_i)$$

$\mathrm{Cov}\left[X_i, X_j\right]$の導出

X_i と X_j $(i \neq j)$ の同時確率関数

$T = \{\, t \mid t \neq i, t \neq j, t = 1, \dots C\}$

$$\Pr(X_i = x_i, X_j = x_j) = \sum_{t \in T} \sum_{x_t = 1}^{n - x_i - x_j} \frac{n!}{x_1! \cdots x_C!} \pi_1^{x_1} \cdots \pi_C^{x_C}$$

$$= \frac{n! \left(1 - \pi_i - \pi_j\right)^{n - x_i - x_j}}{\left(n - x_i - x_j\right)! \, x_i! \, x_j!} \pi_i^{x_i} \pi_j^{x_j} \sum_{t \in T} \sum_{x_t = 1}^{n - x_i - x_j} \frac{\left(n - x_i - x_j\right)!}{\prod_{t \in T} x_t!} \prod_{t \in T} \left(\frac{\pi_t}{1 - \pi_i - \pi_j}\right)^{x_t}$$

$$= \frac{n!}{\left(n - x_i - x_j\right)! \, x_i! \, x_j!} \pi_i^{x_i} \pi_j^{x_j} \left(1 - \pi_i - \pi_j\right)^{n - x_i - x_j}$$

$\mathrm{Cov}[X_i, X_j]$の導出

$\Sigma\Sigma$については，$1 \leq x_i \leq n, 1 \leq x_j \leq n, x_i + x_j \leq n$ を満たすx_i, x_jについて和を取る

$$E[X_i X_j] = \sum\sum x_i x_j \frac{n!}{(n - x_i - x_j)!\, x_i!\, x_j!} \pi_i^{x_i} \pi_j^{x_j} (1 - \pi_i - \pi_j)^{n - x_i - x_j}$$

$$= n(n-1)\pi_i \pi_j \sum\sum \left[\frac{(n-2)!}{\{(n-2) - (x_i - 1) - (x_j - 1)\}!\,(x_i - 1)!\,(x_j - 1)!} \times \pi_i^{x_i - 1} \pi_j^{x_j - 1} (1 - \pi_i - \pi_j)^{(n-2) - (x_i - 1) - (x_j - 1)} \right]$$

$y_i = x_i - 1$

$y_j = x_j - 1$

$$= n(n-1)\pi_i \pi_j \sum\sum \left[\frac{(n-2)!}{\{(n-2) - y_i - y_j!\}\, y_i!\, y_j!} \pi_i^{y_i} \pi_j^{y_j} (1 - \pi_i - \pi_j)^{(n-2) - y_i - y_j} \right]$$

$$= n(n-1)\pi_i \pi_j$$

$\mathrm{Cov}[X_i, X_j]$の導出

$$\mathrm{Cov}[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])]$$

$$= E[X_i X_j - X_i E[X_j] - E[X_i] X_j + E[X_i] E[X_j]]$$

$$= E[X_i X_j] - E[X_i] E[X_j] - E[X_i] E[X_j] + E[X_i] E[X_j]$$

$$= E[X_i X_j] - E[X_i] E[X_j]$$

$$= n(n-1)\pi_i \pi_j - n\pi_i \, n\pi_j$$

$$= -n\pi_i \pi_j$$

仮説の設定と検定統計量

帰無仮説と対立仮説

$H_0 : \pi_i = p_i \ \ (i = 1,2,\ldots,C)$　ただし，$p_i \ \ (i = 1,2,\ldots,C)$は既知

$H_1 : H_0$ではない $\Leftrightarrow H_1 :\ ^{\exists}t \ (t = 1,2,\ldots,C)$に対して$\pi_t \neq p_t$

検定統計量

- Pearson（ピアソン）のカイ二乗統計量

帰無仮説のもとで

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - np_i)^2}{np_i} \approx \chi^2_{(C-1)}\text{分布}$$

n: 大きいとき

棄却域

- 帰無仮説が正しいとき

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - np_i)^2}{np_i} \quad \text{は小さくなる}$$

- 帰無仮説が正しくないとき

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - np_i)^2}{np_i} \quad \text{は大きくなる}$$

- 棄却域の設定

$$W = \left\{ (x_1, x_2, \dots, x_C) \,\middle|\, \chi^2 > \chi^2_{(C-1)}(\alpha) \right\}$$

$\chi^2_{(C-1)}(\alpha)$: 自由度$C-1$のカイ二乗分布の上側100α%点

適合度検定の検定方式

■ 検定方式

$\chi^2 \in \left(\chi^2_{(C-1)}(\alpha), \infty\right)$ のとき，帰無仮説を棄却する

$\chi^2 \notin \left(\chi^2_{(C-1)}(\alpha), \infty\right)$ のとき，帰無仮説を採択する

この検定方式をPearsonのカイ二乗検定という

適用例

- さいころを120回繰り返しふった結果

出目	1	2	3	4	5	6	計
観測度数	16	28	16	15	25	20	120
期待度数	20	20	20	20	20	20	120

$$\chi^2 = \sum_{i=1}^{6} \frac{(X_i - np_i)^2}{np_i} = \frac{(16-20)^2}{20} + \frac{(28-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(20-20)^2}{20}$$

$$= \frac{16+64+16+25+25+0}{20} = \frac{146}{20} = 7.3 < \chi^2_{(5)}(0.05) = 11.07$$

この結果から，さいころが正しくないと断定することはできない

Pearsonのカイ二乗統計量

□ 定理

帰無仮説のもとで

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - np_i)^2}{np_i} \underset{n \to \infty}{\approx} \chi^2_{(C-1)}\text{分布}$$

定理の証明 (1)

- 多項分布の確率関数

$$\Pr(X_1 = x_1, \dots, X_C = x_C) = \frac{n!}{\prod_{i=1}^{C} x_i!} \prod_{i=1}^{C} \pi_i^{x_i}$$

$$n = x_1 + \cdots + x_C$$

$$x_i \geq 0, \mathrm{integer}$$

- 記号の定義

$$Y_i = \frac{X_i - np_i}{\sqrt{np_i}} \quad (i = 1,2, \dots, C)$$

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_C \end{bmatrix}$$

$C \times 1$ベクトル

定理の証明 (2)

- $\boldsymbol{Y}$の積率母関数

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_C \end{bmatrix} \quad C \times 1 \text{ベクトル}$$

$$g_n(\boldsymbol{\theta}) = E\left[e^{\boldsymbol{\theta}^\top \boldsymbol{Y}}\right] = E\left[e^{\theta_1 Y_1 + \cdots + \theta_C Y_C}\right]$$

$$= E\left[\exp\left(\sum_{i=1}^{C} \theta_i \frac{X_i - np_i}{\sqrt{np_i}}\right)\right] = \exp\left(-\sum_{i=1}^{C} \theta_i \sqrt{np_i}\right) E\left[\exp\left(\sum_{i=1}^{C} \frac{\theta_i}{\sqrt{np_i}} X_i\right)\right]$$

帰無仮説
のもとで

$$= \xi \sum_{\substack{n = x_1 + \cdots + x_C \\ x_i \geq 0, \text{integer}}} \exp\left(\sum_{i=1}^{C} \frac{\theta_i}{\sqrt{np_i}} x_i\right) \frac{n!}{\prod_{t=1}^{C} x_t!} \prod_{t=1}^{C} p_t^{x_t}$$

$$= \xi \sum \frac{n!}{\prod_{t=1}^{C} x_t!} \prod_{t=1}^{C} \left(p_t \exp\left(\frac{\theta_t}{\sqrt{np_t}}\right)\right)^{x_t}$$

定理の証明 (3)

- $\boldsymbol{Y}$の積率母関数

$$g_n(\boldsymbol{\theta}) = \xi \sum \frac{n!}{\prod_{t=1}^{C} x_t!} \prod_{t=1}^{C} \left(p_t \exp\left(\frac{\theta_t}{\sqrt{np_t}} \right) \right)^{x_t}$$

多項定理

$$= \xi \left(\sum_{i=1}^{C} p_i \exp\left(\frac{\theta_i}{\sqrt{np_i}} \right) \right)^n$$

多項定理

$$(p_1 + p_2 + \cdots + p_C)^n$$

$$= \sum \frac{n!}{\prod_{i=1}^{C} x_i!} \prod_{i=1}^{C} p_i^{x_i}$$

$n = x_1 + \cdots + x_C$
$x_i \geq 0, \text{integer}$

定理の証明 (4)

$$g_n(\boldsymbol{\theta}) = \xi\left(\sum_{i=1}^{C} p_i \exp\left(\frac{\theta_i}{\sqrt{np_i}}\right)\right)^n$$

$$\xi = \exp\left(-\sum_{i=1}^{C} \theta_i\sqrt{np_i}\right)$$

$$\Leftrightarrow \log g_n(\boldsymbol{\theta}) = -\sum_{i=1}^{C} \theta_i\sqrt{np_i} + n\log\left(\sum_{i=1}^{C} p_i \exp\left(\frac{\theta_i}{\sqrt{np_i}}\right)\right)$$

$$\Leftrightarrow \log g_n(\boldsymbol{\theta}) = -\sum_{i=1}^{C} \theta_i\sqrt{np_i} + n\log\left(\sum_{i=1}^{C} p_i\left(1 + \frac{\theta_i}{\sqrt{np_i}} + \frac{1}{2!}\left(\frac{\theta_i}{\sqrt{np_i}}\right)^2 + O(n^{-\frac{3}{2}})\right)\right)$$

$$\Leftrightarrow \log g_n(\boldsymbol{\theta}) = -\sum_{i=1}^{C} \theta_i\sqrt{np_i} + n\log\left(1 + \frac{1}{\sqrt{n}}\sum_{i=1}^{C} \theta_i\sqrt{p_i} + \frac{1}{2n}\sum_{i=1}^{C} \theta_i^2 + O(n^{-\frac{3}{2}})\right)$$

$$\exp\left(\frac{\theta_i}{\sqrt{np_i}}\right) = 1 + \frac{\theta_i}{\sqrt{np_i}} + \frac{1}{2!}\left(\frac{\theta_i}{\sqrt{np_i}}\right)^2 + O(n^{-\frac{3}{2}})$$

$$p_1 + p_2 + \cdots + p_C = 1$$

定理の証明 (5)

$$\log(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots$$

$$\log g_n(\boldsymbol{\theta}) = -\sum_{i=1}^{C} \theta_i \sqrt{np_i} + n \log\left(1 + \frac{1}{\sqrt{n}} \sum_{i=1}^{C} \theta_i \sqrt{p_i} + \frac{1}{2n} \sum_{i=1}^{C} \theta_i^2 + O(n^{-\frac{3}{2}})\right)$$

$$= -\sum_{i=1}^{C} \theta_i \sqrt{np_i} + n \left[\left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{C} \theta_i \sqrt{p_i} + \frac{1}{2n} \sum_{i=1}^{C} \theta_i^2 + O(n^{-\frac{3}{2}})\right\} - \frac{1}{2}\left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{C} \theta_i \sqrt{p_i} + \frac{1}{2n} \sum_{i=1}^{C} \theta_i^2 + O\left(n^{-\frac{3}{2}}\right)\right\}^2 + \frac{1}{3}\left\{\frac{1}{\sqrt{n}} \sum_{i=1}^{C} \theta_i \sqrt{p_i} + \frac{1}{2n} \sum_{i=1}^{C} \theta_i^2 + O\left(n^{-\frac{3}{2}}\right)\right\}^3 - \cdots\right]$$

定理の証明 (6)

$$\log g_n(\boldsymbol{\theta}) = \frac{1}{2}\left\{\sum_{i=1}^{C}\theta_i^2 - \left(\sum_{i=1}^{C}\theta_i\sqrt{p_i}\right)^2\right\} + O(n^{-\frac{1}{2}})$$

$$\left(\sum_{i=1}^{C}\theta_i\sqrt{p_i}\right)^2 = \sum_{i=1}^{C}\theta_i^2 p_i + 2\sum_{i\neq j}\theta_i\theta_j\sqrt{p_i p_j}$$

$$= \frac{1}{2}\left\{\sum_{i=1}^{C}(1-p_i)\theta_i^2 - \sum_{i\neq j}\theta_i\theta_j\sqrt{p_i p_j}\right\} + O(n^{-\frac{1}{2}})$$

$$= \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Lambda}\boldsymbol{\theta} + O(n^{-\frac{1}{2}})$$

$\boldsymbol{\Lambda}^\top = \boldsymbol{\Lambda}$ (対称行列)

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1-p_1 & -\sqrt{p_1p_2} & \cdots & -\sqrt{p_1p_C} \\ -\sqrt{p_1p_2} & 1-p_2 & & -\sqrt{p_2p_C} \\ \vdots & & \ddots & \vdots \\ -\sqrt{p_1p_C} & \cdots & \sqrt{p_{C-1}p_C} & 1-p_C \end{pmatrix}$$

$C \times C$行列

定理の証明 (7)

帰無仮説のもとで

$$\lim_{n\to\infty} g_n(\boldsymbol{\theta}) = \exp\left(\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Lambda}\boldsymbol{\theta}\right)$$

となることから,

$$\underset{n\,\to\,\infty}{\boldsymbol{Y} \approx N(\boldsymbol{0}, \boldsymbol{\Lambda})}$$

$$\boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda} \ \text{(べき等行列)}, \qquad \boldsymbol{\Lambda}^\top = \boldsymbol{\Lambda} \ \text{(対称行列)}$$

$$\Lambda_{11}^2 = (1-p_1)^2 + p_1(p_2 + \cdots + p_c) = (1-p_1)^2 + p_1(1-p_1) = 1 - p_1 = \Lambda_{11}$$

$$\Lambda_{12}^2 = -\sqrt{p_1p_2}(1-p_1) - \sqrt{p_1p_2}(1-p_2) + \sqrt{p_1p_3}\sqrt{p_2p_3} + \cdots + \sqrt{p_1p_C}\sqrt{p_2p_C}$$

$$= -\sqrt{p_1p_2}(1-p_1) - \sqrt{p_1p_2}(1-p_2) + \sqrt{p_1p_2}(1-p_1-p_2) = -\sqrt{p_1p_2} = \Lambda_{12}$$

多変量正規分布の積率母関数

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$g_{\boldsymbol{X}}(\boldsymbol{\theta}) = E[\exp(\boldsymbol{\theta}^\top \boldsymbol{X})]$$

$$= \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}\boldsymbol{\theta}\right)$$

定理の証明 (8)

$\boldsymbol{\Lambda}$の固有値は0または1なので

$$\operatorname{rank} \boldsymbol{\Lambda} = \operatorname{tr} \boldsymbol{\Lambda} = \sum_{i=1}^{C} (1 - p_i) = C - 1$$

$\boldsymbol{\Lambda}$は対称行列であるので，適当な$C \times C$直交行列$\boldsymbol{U}$により対角化すると

$$\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & 0 \\ \vdots & & 1 & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{Y}$とすると

$$E[\boldsymbol{Z}] = E[\boldsymbol{U}\boldsymbol{Y}] = \boldsymbol{U}E[\boldsymbol{Y}] = \boldsymbol{0}$$

$$V[\boldsymbol{Z}] = V[\boldsymbol{U}\boldsymbol{Y}] = \boldsymbol{U}V[\boldsymbol{Y}]\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$$

定理の証明 (9)

帰無仮説のもとで, $\boldsymbol{Y} \underset{n \to \infty}{\approx} N(\mathbf{0}, \boldsymbol{\Lambda})$ となることから,

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{Y} \underset{n \to \infty}{\approx} N_C(\mathbf{0}, \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top) \qquad \boldsymbol{Z}^\top\boldsymbol{Z} \underset{n \to \infty}{\approx} \chi^2_{(C-1)}$$

$$\begin{aligned}
\boldsymbol{Z}^\top\boldsymbol{Z} &= (\boldsymbol{U}\boldsymbol{Y})^\top\boldsymbol{U}\boldsymbol{Y} \\
&= \boldsymbol{Y}^\top\boldsymbol{U}^\top\,\boldsymbol{U}\boldsymbol{Y} \\
&= \boldsymbol{Y}^\top\boldsymbol{Y} \\
&= \sum_{i=1}^{C} Y_i^2 = \sum_{i=1}^{C} \frac{(X_i - np_i)^2}{np_i}
\end{aligned}$$