統計学2及び演習

適合度検定とその例 (2)

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14日

Contents

- ポアソン分布
- 交通事故の死亡者数
- 適合度検定
  - 帰無仮説が未知の場合
- 検定統計量
  - Pearson（ピアソン）のカイ二乗統計量
  - 尤度比カイ二乗統計量

ポアソン分布

- ある時間間隔で発生する事象の回数を表す離散確率分布
- ある時間内での生起回数の確率
  - 指数分布は生起間隔の確率
- 具体例
  - 1時間にある交差点を通過する車の数
  - 1日に受け取る電子メールの件数
- 確率変数 X: 0 以上の整数を値にとる

確率関数

$$\Pr(X = x) = \frac{\exp(-\lambda)\,\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots$$

ポアソン分布の性質

- 期待値 $E[X] = \lambda$

平均と分散が1つの母数 λ で表現される

- 分散 $V[X] = \lambda$
- 二項分布の母数について$n \to \infty,\ \pi \to 0$ を適用した分布
  - $n\pi \to \lambda$ となる極限
- λ が十分に大きいときは正規分布で近似できる

$E[X]$の導出

$$E[X] = \sum_{x=0}^{\infty} x \frac{\exp(-\lambda)\,\lambda^x}{x!} = \lambda\exp(-\lambda) \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$y = x - 1$

$$= \lambda\exp(-\lambda) \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda$$

$$\left(\because \exp(\lambda) = \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right)$$

指数関数の定義

$V[X]$の導出

$$E[X(X-1)] = \sum_{x=0}^{\infty} x(x-1)\frac{\exp(-\lambda)\,\lambda^x}{x!} = \lambda^2\exp(-\lambda)\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!}$$

$$y = x - 2$$

$$= \lambda^2\exp(-\lambda)\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda^2$$

$$E[X^2] = E[X(X-1)] + E[X] = \lambda^2 + \lambda$$

$$V[X] = E[X^2] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

二項分布の極限としてのポアソン ン分布

- $X \sim \mathrm{Bin}(n,\pi)$とし，$\lim_{n\to\infty} n\pi = \lambda$であるとする。
このとき，Xの従う確率分布は 母数λ のポアソン分布に収束する。

（証明）

n が十分大きいとき，$n\pi = \lambda$ であることから，X の確率関数は

$$\Pr(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \approx \frac{n!}{x!\,(n-x)!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{n(n-1)\cdots\big(n-(x-1)\big)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= 1\times\left(1-\frac{1}{n}\right)\times\cdots\times\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n$$

証明（続き）

ここで，x を固定したまま n を無限大にすると

$$\lim_{n\to\infty}\left(1\pm\frac{x}{n}\right)^{n}=\exp(\pm x)$$

であることから，次式を得る。

$$\begin{aligned}\lim_{n\to\infty}\Pr(X=x)&\approx\lim_{n\to\infty}1\times\left(1-\frac{1}{n}\right)\times\cdots\times\left(1-\frac{x-1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-x}\frac{\lambda^{x}}{x!}\left(1-\frac{\lambda}{n}\right)^{n}\\&\approx\lim_{n\to\infty}\frac{\lambda^{x}}{x!}\left(1-\frac{\lambda}{n}\right)^{n}\\&=\frac{\lambda^{x}}{x!}\exp(-\lambda)\end{aligned}$$

ポアソン分布の確率関数に一致

交通事故死亡者数

- ある地域の100日間の交通事故死亡者数

人数 (i)	0	1	2	3	4	5以上	計
度数	18	27	35	17	3	0	100

- 仮説
  - 1日辺りの死亡者数は平均2のポアソン分布に従うかどうか
  - 帰無仮説と対立仮説

$$H_0: \Pr(X = i) = p_i = \frac{\exp(-2)\, 2^i}{i!} \qquad i = 0, 1, 2, \ldots$$

$H_1: H_0$ではない $\Leftrightarrow H_1: {}^\exists t\ (t = 0, 1, 2, \ldots)$に対して

$$\Pr(X = t) \neq \frac{\exp(-2)\, 2^t}{t!}$$

交通事故死亡者数

- 帰無仮説のもとでの期待度数

$$n \times p_i = 100 \times \frac{\exp(-2)\, 2^i}{i!}$$

- Pearson（ピアソン）のカイ二乗統計量

$$\chi^2 = \sum_{i=0}^{5} \frac{(X_i - np_i)^2}{np_i}$$

$$= \frac{\left(18 - 100 \times \frac{\exp(-2)\, 2^0}{0!}\right)^2}{100 \times \frac{\exp(-2)\, 2^0}{0!}} + \frac{\left(27 - 100 \times \frac{\exp(-2)\, 2^1}{1!}\right)^2}{100 \times \frac{\exp(-2)\, 2^1}{1!}} + \cdots + \frac{\left(0 - 100 \times \frac{\exp(-2)\, 2^5}{5!}\right)^2}{100 \times \frac{\exp(-2)\, 2^5}{5!}}$$

$$\fallingdotseq 11.49 > \chi^2_{(5)}(0.05) = 11.07$$

この結果から，1日辺りの死亡者数は平均2のポアソン分布に従わないと判断する

適合度検定

- 仮定する分布
  - 多項分布

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

- 帰無仮説と対立仮説

$\theta_1, \dots, \theta_s$は未知

$$H_0 : \pi_i = g_i(\theta_1, \dots, \theta_s) \ \ (i = 1, 2, \dots, C)$$

関数形g_iは既知　　　ただし，$s < C - 1, \theta_1, \dots, \theta_s$は線形独立

$H_1 : H_0$ではない $\Leftrightarrow H_1 : {}^{\exists}t \ (t = 1, 2, \dots, C)$に対して$\pi_t \neq g_t(\theta_1, \dots, \theta_s)$

検定統計量 (1)

Pearson（ピアソン）のカイ二乗統計量

帰無仮説のもとで

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} \approx \chi^2_{(C-1-s)}\text{分布}$$

n: 大きいとき

前回の講義と自由度が異なることに注意

ただし， $\hat{\pi}_i = g_i(\hat{\theta}_1, \dots, \hat{\theta}_s)$は帰無仮説のもとでの$\hat{\pi}_i$の最尤推定量

$$\max_{\{\theta_i\}} \frac{n!}{x_1! \cdots x_C!} \left(g_1(\theta_1, \dots, \theta_s)\right)^{x_1} \cdots \left(g_C(\theta_1, \dots, \theta_s)\right)^{x_C}$$

$$= \frac{n!}{x_1! \cdots x_C!} \left(g_1(\hat{\theta}_1, \dots, \hat{\theta}_s)\right)^{x_1} \cdots \left(g_C(\hat{\theta}_1, \dots, \hat{\theta}_s)\right)^{x_C}$$

検定統計量 (2)

尤度比カイ二乗統計量

帰無仮説のもとで

$$G^2 = 2\sum_{i=1}^{C} X_i \log \frac{X_i}{n\hat{\pi}_i} \approx \chi^2_{(C-1-s)}\text{分布}$$

n: 大きいとき

ただし，$\hat{\pi}_i = g_i(\hat{\theta}_1, \dots, \hat{\theta}_s)$は帰無仮説のもとでの$\hat{\pi}_i$の最尤推定量

$$\max_{\{\theta_i\}} \frac{n!}{x_1! \cdots x_C!}\left(g_1(\theta_1, \dots, \theta_s)\right)^{x_1} \cdots \left(g_C(\theta_1, \dots, \theta_s)\right)^{x_C}$$
$$= \frac{n!}{x_1! \cdots x_C!}\left(g_1(\hat{\theta}_1, \dots, \hat{\theta}_s)\right)^{x_1} \cdots \left(g_C(\hat{\theta}_1, \dots, \hat{\theta}_s)\right)^{x_C}$$

棄却域

- 帰無仮説が正しいとき

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}, \qquad G^2 = 2\sum_{i=1}^{C} X_i \log\frac{X_i}{n\hat{\pi}_i}$$ は小さくなる

- 帰無仮説が正しくないとき

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}, \qquad G^2 = 2\sum_{i=1}^{C} X_i \log\frac{X_i}{n\hat{\pi}_i}$$ は大きくなる

- 棄却域の設定

$$W = \{ (x_1, x_2, \dots, x_C) \mid \chi^2 (\text{or } G^2) > \chi^2_{(C-1-s)}(\alpha) \}$$

$\chi^2_{(C-1-s)}(\alpha)$: 自由度$C-1-s$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

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

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

この検定方式をカイ二乗適合度検定という

交通事故死亡者数

- ある地域の100日間の交通事故死亡者数

人数 (i)	0	1	2	3	4	5以上	計
度数	18	27	35	17	3	0	100

- 仮説
  - 1日辺りの死亡者数はポアソン分布に従うかどうか
  - 帰無仮説と対立仮説

$$H_0: \Pr(X = i) = \frac{\exp(-\lambda)\,\lambda^i}{i!} \qquad i = 0, 1, 2, \ldots \qquad g_i(\lambda) = \frac{\exp(-\lambda)\,\lambda^i}{i!}$$

$H_1: H_0$ではない $\Leftrightarrow H_1: {}^{\exists}t\ (t = 0, 1, 2, \ldots)$に対して

$$\Pr(X = t) \neq \frac{\exp(-\lambda)\,\lambda^t}{t!}$$

帰無仮説のもとでの最尤推定量$\hat{\pi}_i$ (1)

$$L(\lambda) = \frac{n!}{x_1! \cdots x_C!} \pi_1^{x_1} \cdots \pi_C^{x_C}$$

$$= \frac{n!}{x_1! \cdots x_C!} \left(g_1(\theta_1, \dots, \theta_s)\right)^{x_1} \cdots \left(g_C(\theta_1, \dots, \theta_s)\right)^{x_C}$$

$$= \frac{n!}{x_1! \cdots x_C!} \left(\frac{\exp(-\lambda)\,\lambda^1}{1!}\right)^{x_1} \cdots \left(\frac{\exp(-\lambda)\,\lambda^C}{C!}\right)^{x_C}$$

$$\log L(\lambda) = \mathrm{Const} - \lambda \sum_{i=1}^{C} x_i + \sum_{i=1}^{C} x_i \times i \log \lambda$$

$$= \mathrm{Const} - \lambda n + \sum_{i=1}^{C} x_i \times i \log \lambda$$

$$\mathrm{Const} = \log \frac{n!}{x_1! \cdots x_C!} - \sum_{i=1}^{C} x_i \log i!$$

帰無仮説のもとでの最尤推定量$\hat{\pi}_i$ (2)

$$\log L(\lambda) = \mathrm{Const} - \lambda n + \sum_{i=1}^{C} x_i \times i \log \lambda$$

$$\frac{\partial \log L(\lambda)}{\partial \lambda} = -n + \sum_{i=1}^{C} \frac{i x_i}{\lambda} \quad (\equiv 0) \quad \Leftrightarrow \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{C} i X_i \qquad \hat{\pi}_i = \frac{\exp\left(-\hat{\lambda}\right) \hat{\lambda}^i}{i!}$$

交通事故死亡者数のデータに対する最尤推定値を求めると

$$\hat{\lambda} = \frac{1}{100} \sum_{i=0}^{5} i x_i = \frac{1}{100} (0 \times 18 + 1 \times 27 + 2 \times 35 + 3 \times 17 + 4 \times 3 + 5 \times 0) = \frac{160}{100} = 1.6$$

人数 (i)	0	1	2	3	4	5以上	計
度数	18	27	35	17	3	0	100

交通事故死亡者数

- 帰無仮説のもとでの期待度数

$$n\hat{\pi}_i = 100 \times \frac{\exp(-\hat{\lambda})\hat{\lambda}^i}{i!}$$

- Pearson（ピアソン）のカイ二乗統計量

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}$$

$$= \frac{\left(18 - 100 \times \frac{\exp(-1.6)1.6^0}{0!}\right)^2}{100 \times \frac{\exp(-1.6)1.6^0}{0!}} + \frac{\left(27 - 100 \times \frac{\exp(-1.6)1.6^1}{1!}\right)^2}{100 \times \frac{\exp(-1.6)\,1.6^1}{1!}} + \cdots + \frac{\left(0 - 100 \times \frac{\exp(-1.6)\,1.6^5}{5!}\right)^2}{100 \times \frac{\exp(-1.6)\,1.6^5}{5!}}$$

$$\fallingdotseq 8.01 > \chi^2_{(6-1-1)}(0.05) = \chi^2_{(4)}(0.05) = 9.49$$

この結果から，1日辺りの死亡者数はポアソン分布に従わないとはいえない

Pearsonと尤度比カイ二乗統計量の関係 (1)

- Pearson（ピアソン）のカイ二乗統計量

$$\chi^2 = \sum_{i=1}^{C} \frac{(X_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}$$

- 尤度比カイ二乗統計量

$$G^2 = 2\sum_{i=1}^{C} X_i \log \frac{X_i}{n\hat{\pi}_i}$$

$$G^2 = \chi^2 + O_p(n^{-\frac{1}{2}})$$

Pearsonと尤度比カイ二乗統計量の関係 (2)

$$G^2 = 2\sum_{i=1}^{C} X_i \log\frac{X_i}{n\hat{\pi}_i} = -2\sum_{i=1}^{C} X_i \log\frac{n\hat{\pi}_i}{X_i}$$

$$\log(1-x) = -\sum_{n=0}^{\infty}\frac{x^n}{n}$$

$$(-1 < x \leq 1)$$

$$\log\frac{n\hat{\pi}_i}{X_i} = \log\left(\frac{X_i - X_i + n\hat{\pi}_i}{X_i}\right) = \log\left(1 - \frac{X_i - n\hat{\pi}_i}{X_i}\right)$$

$$= -\frac{X_i - n\hat{\pi}_i}{X_i} - \frac{1}{2}\left(\frac{X_i - n\hat{\pi}_i}{X_i}\right)^2 - \frac{1}{3}\left(\frac{X_i - n\hat{\pi}_i}{X_i}\right)^3 + \cdots$$

$$= -\frac{1}{2}\left(\frac{X_i - n\hat{\pi}_i}{X_i}\right)^2 + O_p(n^{-\frac{1}{2}})$$

$$\frac{X_i}{n} = \hat{\pi}_i + O_p(n^{-\frac{1}{2}})$$

Pearsonと尤度比カイ二乗統計量の関係 (3)

$$G^2 = -2\sum_{i=1}^{C} X_i \log \frac{n\hat{\pi}_i}{X_i}$$

$$= -2\sum_{i=1}^{C} X_i \left[-\frac{1}{2}\left(\frac{X_i - n\hat{\pi}_i}{X_i}\right)^2 + O_p(n^{-\frac{1}{2}}) \right]$$

$$= \sum_{i=1}^{C} \left[\frac{(X_i - n\hat{\pi}_i)^2}{X_i} + O_p(n^{-\frac{1}{2}}) \right]$$

$$\frac{X_i}{n} = \hat{\pi}_i + O_p(n^{-\frac{1}{2}})$$

$$= \sum_{i=1}^{C} \left[\frac{(X_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} + O_p(n^{-\frac{1}{2}}) \right] = \chi^2 + O_p(n^{-\frac{1}{2}})$$