統計学2及び演習

分割表における独立性検定とその応用例

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21日

Contents

分割表

- 属性Aの事象$A_1, A_2, \ldots, A_R$
- 属性Bの事象$B_1, B_2, \ldots, B_C$

属性AとBについて集計した二次元度数分布表を
特に分割表 (contingency table) とよぶ

- x_{ij}: A_iかつB_jである観測度数
- $n = \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij}$:総観測度数
- $x_{i\cdot} = \sum_{j=1}^{C} x_{ij} \quad (i = 1, \ldots, R)$
- $x_{\cdot j} = \sum_{i=1}^{R} x_{ij} \quad (j = 1, \ldots, C)$

	B_1	B_2	$\cdots$	B_C	計
A_1	x_{11}	x_{12}	$\cdots$	x_{1C}	$x_{1\cdot}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2C}	$x_{2\cdot}$
$\vdots$					$\vdots$
A_R	x_{R1}	x_{R2}	$\cdots$	x_{RC}	$x_{R\cdot}$
計	$x_{\cdot 1}$	$x_{\cdot 2}$	$\cdots$	$x_{\cdot C}$	n

対応する同時確率

- $\Pr(A_i \cap B_j) = \pi_{ij}$: A_iかつB_jである確率
- $\sum_{i=1}^{R}\sum_{j=1}^{C}\pi_{ij} = 1$
- $\Pr(A_i) = \pi_{i.} = \sum_{j=1}^{C}\pi_{ij} \quad (i = 1, \ldots, R)$
- $\Pr(B_j) = \pi_{.j} = \sum_{i=1}^{R}\pi_{ij} \quad (j = 1, \ldots, C)$

観測度数

	B_1	B_2	$\cdots$	B_C	計
A_1	x_{11}	x_{12}	$\cdots$	x_{1C}	$x_{1.}$
A_2	x_{21}	x_{22}	$\cdots$	x_{2C}	$x_{2.}$
$\vdots$					$\vdots$
A_R	x_{R1}	x_{R2}	$\cdots$	x_{RC}	$x_{R.}$
計	$x_{.1}$	$x_{.2}$	$\cdots$	$x_{.C}$	n

対応する同時確率

	B_1	B_2	$\cdots$	B_C	計
A_1	π_{11}	π_{12}	$\cdots$	π_{1C}	$\pi_{1.}$
A_2	π_{21}	π_{22}	$\cdots$	π_{2C}	$\pi_{2.}$
$\vdots$					$\vdots$
A_R	π_{R1}	π_{R2}	$\cdots$	π_{RC}	$\pi_{R.}$
計	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.C}$	1

仮定する分布

- 観測度数

$$\boldsymbol{x} = (x_{11}, x_{12}, \dots, x_{1C}, \dots, x_{R1}, x_{R2}, \dots, x_{RC})$$

  - パラメータ$(n, \boldsymbol{\pi})$の多項分布からの実現値と考える

$$\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \dots, \pi_{1C}, \dots, \pi_{R1}, \pi_{R2}, \dots, \pi_{RC})$$

- 多項分布

$$\Pr(X_{11} = x_{11}, \dots, X_{RC} = x_{RC}) = \frac{n!}{\prod_{i=1}^{R}\prod_{j=1}^{C} x_{ij}!} \prod_{i=1}^{R}\prod_{j=1}^{C} \pi_{ij}^{x_{ij}}$$

期待値 $E[X_{ij}] = n\pi_{ij}$

共分散 $\mathrm{Cov}[X_{ij}, X_{st}] = -n\pi_{ij}\pi_{st} \ (i \neq s \text{ or } j \neq t)$

分散 $V[X_{ij}] = n\pi_{ij}(1 - \pi_{ij})$

独立性の検定

- 帰無仮説と対立仮説

  H_0: AとBは互いに統計的独立

  $\Leftrightarrow H_0: \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad (i = 1, \dots, R; j = 1, \dots, C)$

  H_1: H_0ではない

- 帰無仮説の別表現

$$H_0: \pi_{ij} = g_{ij}(\pi_{1\cdot}, \dots, \pi_{R-1\cdot}; \pi_{\cdot 1}, \dots, \pi_{\cdot C-1}) \quad (i = 1, \dots, R; j = 1, \dots, C)$$

$$\pi_{R\cdot} = 1 - (\pi_{1\cdot} + \cdots + \pi_{R-1\cdot}) \qquad \pi_{\cdot C} = 1 - (\pi_{\cdot 1} + \cdots + \pi_{\cdot C-1})$$

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

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(X_{ij} - n\hat{\pi}_{ij}\right)^2}{n\hat{\pi}_{ij}}$$

$$= \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(X_{ij} - \frac{X_{i\cdot}X_{\cdot j}}{n}\right)^2}{\frac{X_{i\cdot}X_{\cdot j}}{n}} \approx \chi^2_{(RC-1-s)}\text{分布}$$

n: 大きいとき

帰無仮説のもとでの
π_{ij}の最尤推定量

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n}$$

$(i = 1, \dots, R; j = 1, \dots, C)$

$H_0: \pi_{ij} = g_{ij}(\pi_{1\cdot}, \dots, \pi_{R-1\cdot}; \pi_{\cdot 1}, \dots, \pi_{\cdot C-1})$ $(i = 1, \dots, R; j = 1, \dots, C)$ であることから

$$s = R - 1 + C - 1 \qquad RC - 1 - s = RC - 1 - (R - 1 + C - 1) = (R - 1)(C - 1)$$

棄却域と検定方式

棄却域

$$W = \{ \boldsymbol{x} \mid \chi^2 > \chi^2_{((R-1)(C-1))}(\alpha) \}$$

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

検定方式

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

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

帰無仮説のもとでの$\boldsymbol{\pi_{ij}}$の最尤推定量 (1)

$$H_0: \pi_{ij} = \pi_{i.}\pi_{.j} \quad (i = 1, \ldots, R; j = 1, \ldots, C)$$

$$L\left(\{\pi_{ij}\} \mid \{x_{ij}\}\right) = \frac{n!}{\prod_{i=1}^{R}\prod_{j=1}^{C} x_{ij}!} \prod_{i=1}^{R}\prod_{j=1}^{C} \pi_{ij}^{x_{ij}}$$

$$L\left(\{\pi_{i.}\}, \{\pi_{.j}\} \mid \{x_{ij}\}\right) = \frac{n!}{\prod_{i=1}^{R}\prod_{j=1}^{C} x_{ij}!} \prod_{i=1}^{R}\prod_{j=1}^{C} \left(\pi_{i.}\pi_{.j}\right)^{x_{ij}}$$

$$\log L\left(\{\pi_{i.}\}, \{\pi_{.j}\} \mid \{x_{ij}\}\right) = \text{Const} + \sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij} \left(\log \pi_{i.} + \log \pi_{.j}\right)$$

$$\text{Const} = \log \frac{n!}{\prod_{i=1}^{R}\prod_{j=1}^{C} x_{ij}!}$$

帰無仮説のもとでの$\boldsymbol{\pi_{ij}}$の最尤推定量 (2)

ラグランジュの未定乗数法より

$$\log L = \text{Const} + \sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij}\left(\log \pi_{i\cdot} + \log \pi_{\cdot j}\right) - \phi_1\left(\sum_{i=1}^{R} \pi_{i\cdot} - 1\right) - \phi_2\left(\sum_{j=1}^{C} \pi_{\cdot j} - 1\right)$$

① $\dfrac{\partial \log L}{\partial \pi_{k\cdot}} = \dfrac{x_{k\cdot}}{\pi_{k\cdot}} - \phi_1 \quad (\equiv 0)$
$(k = 1, \ldots, R)$

② $\dfrac{\partial \log L}{\partial \pi_{\cdot l}} = \dfrac{x_{\cdot l}}{\pi_{\cdot l}} - \phi_2 \quad (\equiv 0)$
$(l = 1, \ldots, C)$

③ $\dfrac{\partial \log L}{\partial \phi_1} = \sum_{i=1}^{R} \pi_{i\cdot} - 1 \quad (\equiv 0)$

④ $\dfrac{\partial \log L}{\partial \phi_2} = \sum_{j=1}^{C} \pi_{\cdot j} - 1 \quad (\equiv 0)$

帰無仮説のもとでのπ_{ij}の最尤推定量 (3)

$① \Leftrightarrow x_{k\cdot} - \phi_1 \pi_{k\cdot} = 0 \quad (k = 1, \ldots, R)$

$② \Leftrightarrow x_{\cdot l} - \phi_2 \pi_{\cdot l} = 0 \quad (l = 1, \ldots, C)$

和をとると

$$\sum_{k=1}^{R} (x_{k\cdot} - \phi_1 \pi_{k\cdot}) = 0 \Leftrightarrow \phi_1 = n$$

$$\sum_{l=1}^{C} (x_{\cdot l} - \phi_2 \pi_{\cdot l}) = 0 \Leftrightarrow \phi_2 = n$$

これらの結果を①, ②式に代入すると

$$\hat{\pi}_{k\cdot} = \frac{X_{k\cdot}}{n} \quad (k = 1, \ldots, R) \qquad \hat{\pi}_{\cdot l} = \frac{X_{\cdot l}}{n} \quad (l = 1, \ldots, C)$$

したがって

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{X_{i\cdot}}{n}\frac{X_{\cdot j}}{n} \qquad (i = 1, \ldots, R; j = 1, \ldots, C)$$

2×2分割表

- 属性Aの事象A_1, A_2
- 属性Bの事象B_1, B_2

	B_1	B_2	計
A_1	x_{11}	x_{12}	$x_{1.}$
A_2	x_{21}	x_{22}	$x_{2.}$
計	$x_{.1}$	$x_{.2}$	n

- Pearson（ピアソン）のカイ二乗統計量

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(X_{ij} - \frac{X_{i.}X_{.j}}{n}\right)^2}{\frac{X_{i.}X_{.j}}{n}} = \frac{n(X_{11}X_{22} - X_{12}X_{21})^2}{X_{1.}X_{2.}X_{.1}X_{.2}} \approx \chi^2_{(1)}\text{分布}$$

n: 大きいとき

イエーツの補正

- χ^2分布への近似をよくするための補正

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(\left|X_{ij} - \frac{X_{i\cdot}X_{\cdot j}}{n}\right| - \frac{1}{2}\right)^2}{\frac{X_{i\cdot}X_{\cdot j}}{n}} = \frac{n\left(|X_{11}X_{22} - X_{12}X_{21}| - \frac{n}{2}\right)^2}{X_{1\cdot}X_{2\cdot}X_{\cdot 1}X_{\cdot 2}} \approx \chi^2_{(1)}\text{分布}$$

  - セル観測度数が少ない（例えば5未満）ときに有効
  - 総観測度数が少ないときは，イエーツの補正でも近似の精度は十分ではないため，フィッシャーの正確検定を用いる

応用例

- 新規治療と標準治療の有効性の比較

	有効	無効	計
標準治療	63	20	83
新規治療	22	6	28
計	85	26	111

$$\chi^2 = \frac{111(63 \times 6 - 20 \times 22)^2}{83 \times 28 \times 85 \times 26} = 0.083 < \chi^2_{(1)}(0.05) = 3.84$$

$$\chi^2 = \frac{111\left(|63 \times 6 - 20 \times 22| - \frac{111}{2}\right)^2}{83 \times 28 \times 85 \times 26} = 0.0009 < \chi^2_{(1)}(0.05) = 3.84$$

問

- 次式が成り立つことを示せ

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{\left(X_{ij} - \frac{X_{i\cdot}X_{\cdot j}}{n}\right)^2}{\frac{X_{i\cdot}X_{\cdot j}}{n}} = \frac{n(X_{11}X_{22} - X_{12}X_{21})^2}{X_{1\cdot}X_{2\cdot}X_{\cdot 1}X_{\cdot 2}}$$