Excel による独立性・適合度のカイ二乗検定

樋口さぶろお http://hig3.net

龍谷大学理工学部数理情報学科

生活の中の統計技術　L13(2019-01-21 Mon)

最終更新: Time-stamp: "2019-01-21 Mon 09:01 JST hig"

今日の目標

- Excel のピボットテーブルを使ってクロス集計表を作れる
- Excel を使って独立性のカイ二乗検定を実行できる

L12-Q1

Quiz 解答:ベイズ推定

当外 \ 色	赤	白	計
あたり	2	18	20
はずれ	56	24	80
計	58	42	100

1. $\frac{20}{100}$
2. $\frac{42}{100}$
3. $\frac{2}{58}$

L12-Q2

Quiz 解答:ベイズ推定

E\ D	感染	非感染	計
陽性	$\frac{60}{100} \cdot \frac{1}{100}$	$\frac{10}{100} \cdot \frac{99}{100}$	$\frac{1050}{10000}$
陰性	$\frac{40}{100} \cdot \frac{1}{100}$	$\frac{90}{100} \cdot \frac{99}{100}$	$\frac{8950}{10000}$
合計	$\frac{1}{100}$	$\frac{99}{100}$	1

1. $1/100 = 0.01.$
2. $\dfrac{\frac{60}{10000}}{\frac{1050}{10000}} = 0.0571.$
3. $\dfrac{\frac{8910}{10000}}{\frac{8950}{10000}} = 0.9955.$

L12-Q3

Quiz 解答:ベイズの公式

1. $$P(Y = y|X = 1) = \begin{cases} 0.95 & (y = 10) \\ 0.05 & (y = 20) \end{cases}$$
$$P(Y = y|X = 2) = \begin{cases} 0.125 & (y = 10) \\ 0.875 & (y = 20) \end{cases}$$

2

$y\backslash x$	1	2
10	0.19	0.10
20	0.01	0.70

$$P(X=1|Y=10) = \frac{P(Y=10|X=1)P(X=1)}{\sum_x P(Y=10|X=x)P(X=x)} = \frac{0.95\times 0.2}{0.95\times 0.2 + 0.125\times 0.8} = \frac{19}{29}.$$

3

$y\backslash x$	1	2
10	0.76	0.025
20	0.04	0.175

$$P(X=2|Y=20) = \frac{P(Y=20|X=2)P(X=2)}{\sum_x P(Y=20|X=x)P(X=x)} = \frac{0.875\times 0.2}{0.05\times 0.8 + 0.875\times 0.2} = \frac{35}{43}.$$

L12-Q4

Quiz 解答:母平均値の区間推定 (母分散未知)

標本サイズは $n = 6$.
$T = \frac{\bar{X}-\mu}{\sqrt{s^2/n}}$ は自由度 $n-1 = 5$ の t 分布に従う. 表より
$t(n-1;\alpha/2) = t(5;0.005) = 4.032$. よって, 信頼係数 $1-\alpha = 0.99$ の信頼区間は,

$$204 - 4.032 \times \sqrt{2/6} < \mu < 204 + 4.032 \times \sqrt{2/6}.$$

L12-Q5

Quiz 解答:標本サイズと信頼区間

標本サイズ $N = 20$ のときの信頼区間は, $170 - 4 < \mu < 170 + 4$ で信頼区間の長さは 8cm.

1. 信頼区間の長さを $1/2$ にすればいいから,
   $N_1 = 20 \times (1/(1/2))^2 = 80$.

2. 信頼区間の長さを $1/4$ にすればいいから,
   $N_1 = 20 \times (1/(1/4))^2 = 320.$

L12-Q6

Quiz 解答:母比率の区間推定

標本サイズは $n = 10$. 標本比率は $\frac{1}{5}$

1. 標本比率 r は, 母平均値 p, 母分散 $\frac{1}{n}p(1-p)$ の正規分布に近似的に従うので,

$$\frac{1}{5} - 1.96 \times \sqrt{\frac{1}{10}\frac{1}{5}(1-\frac{1}{5})} < p < \frac{1}{5} + 1.96 \times \sqrt{\frac{1}{10}\frac{1}{5}(1-\frac{1}{5})}$$

2. $p = \frac{m}{120}$ より $m = 120p$. よって,

$$120\times\left(\frac{1}{5} - 1.96 \times \sqrt{\frac{1}{10}\frac{1}{5}(1-\frac{1}{5})}\right) < m < 120\times\left(\frac{1}{5} + 1.96 \times \sqrt{\frac{1}{10}\frac{1}{5}(1-\right.$$

L12-Q7

Quiz 解答:標本サイズと信頼区間　　0.55 ± 0.14 であるが, 0.14 が 0.05 未満になればよい. 信頼区間の長さはサンプルサイズの平方根に反比例するので,

$\frac{\sqrt{n}}{\sqrt{50}} = \frac{0.05}{0.14}$.

よって, $n = 392$.

ここまで来たよ

12 ベイズ推定・標本サイズの決定

13 Excel による独立性・適合度のカイ二乗検定
- 独立性の指標 ピアソンの χ^2 とカイ二乗検定
- 適合度の指標 ピアソンの χ^2 とカイ二乗検定

2 つのカテゴリ変数

未知の母分布

$Y \setminus X$	A 型	A 型以外
女子	P(血液型=A 型, 性別=女)	P(血液型=A 型以外, 性別=女)
男子	P(血液型=A 型, 性別=男)	P(血液型=A 型以外, 性別=女)

標本

出席番号	血液型	性別
1	A 型以外	男
2	A 型以外	女
⋮	⋮	⋮
12	A 型	女

標本サイズ $N = 12$

分割表, クロス集計表

ピボット →

	A 型	A 型以外
女子	$n_{11} = 1$	$n_{12} = 2$
男子	$n_{21} = 4$	$n_{22} = 5$

度数 n_{ij}, $1 \leq i \leq c, 1 \leq j \leq r$. 行数 r, 列数 c.

標本の周辺分布

母分布の周辺分布を, 標本の周辺分布で推定

$y \setminus x$	A 型	A 型以外	計
女子	1	2	3
男子	4	5	9
計	5	7	12

- P(性別=女) は $p_1 = \frac{3}{12}$ くらい
- P(血液型=A 型) は $q_1 = \frac{5}{12}$ くらい

期待度数

もし, 性別と血液型が無関係 (=独立) なら. A 型の女子は

$$\text{期待度数} = N \times p_1 \times q_1 = 12 \times \frac{3}{12} \times \frac{5}{12} = 1.25$$

人くらいのはず

「独立でない度」:ピアソンの χ^2

期待度数

	A 型	A 型以外	計
女子	Np_1q_1	Np_1q_2	Np_1
男子	Np_2q_1	Np_2q_2	Np_2
計	Nq_1	Nq_2	N

$(\text{ずれ})^2 = \sum(\text{度数} - \text{期待度数})^2$

「独立でない度」:ピアソンの χ^2(カイ二乗)

p_i $(i = 1, \ldots, r)$, q_j $(j = 1, \ldots, c)$: 標本から推定した周辺分布.

$$\chi^2 = \frac{(\text{度数} - \text{期待度数})^2}{\text{期待度数}} \text{の合計} = \sum_{1 \le i \le r, 1 \le j \le c} \frac{(n_{ij} - Np_iq_j)^2}{Np_iq_j}$$

いまの場合

$$\chi^2 = \frac{(1-1.25)^2}{1.25} + \frac{(2-1.75)^2}{1.75} + \frac{(4-3.75)^2}{3.75} + \frac{(5-5.25)^2}{5.25} = 0.11685.$$

ピアソンの χ^2(カイ二乗) の性質

- $0 \leq \chi^2$.
- 大きいほど ‘独立でなさそう’=関係ありそう
- 実は, 自由度 $n = (r-1)(c-1)$ のカイ二乗分布にしたがう.

L13-Q1

Quiz(ピアソンの χ^2 と独立性の検定)

日本人の高校生から標本を抽出し, 6 人を, 右利きかどうか, 早生まれかどうかで分類すると, 度数 (人数) は下の表のようになった.

	右利き	右利きでない
早生まれ	1	1
早生まれでない	3	1

1. ピアソンの χ^2 を求めよう.
2. 早生まれかどうかと右利きであるかどうかは独立か. 有意水準 $\alpha = 0.05$ で, 独立性のカイ二乗検定を行って判定しよう. 「○○○ (不等式) なので, 帰無仮説を棄却する/しない. X と Y には関係がある/あるとは言えない」の形で答えよう.

独立性のカイ二乗検定

独立性のカイ二乗検定

- 検定統計量=検査薬の発色レベルピアソンの χ^2.
- Excel での計算方法 `chisq.test(`度数の範囲`,` 期待度数の範囲`)` で計算される p 値を 有意水準 α と比較し, p 値 $<$ 有意水準 α なら発色, 「独立でない」「関係ある」と判定する.

ここまで来たよ

12 ベイズ推定・標本サイズの決定

13 Excel による独立性・適合度のカイ二乗検定
- 独立性の指標 ピアソンの χ^2 とカイ二乗検定
- 適合度の指標 ピアソンの χ^2 とカイ二乗検定

質的変数が 1 つのときの適合度

母分布

カテゴリの個数 $C = 4$.

カテゴリ	O 型	A 型	AB 型	B 型
確率 f_i	$f_1 = 0.12$	$f_2 = 0.51$	$f_3 = 0.17$	$f_4 = 0.20$

$$\sum_{i=1}^{C} f_i = 1.$$

標本

出席番号	血液型
1	B 型
2	O 型
⋮	⋮
12	A 型

ピボット →

度数分布表

カテゴリ	O 型	A 型	AB 型	B 型
度数 n_i	$n_1 = 2$	$n_2 = 3$	$n_3 = 6$	$n_4 = 1$

$$\sum_{i=1}^{C} n_i = N = 12.$$

適合度を表す量

期待度数=母分布の確率 × 標本サイズ

分布にあってない度:ピアソンの適合度基準 χ^2

$$\chi^2 = \frac{(\text{度数} - \text{期待度数})^2}{\text{期待度数}} \text{の合計} = \sum_{i=1}^{C} \frac{(n_i - Nf_i)^2}{Nf_i}$$

ピアソンの適合度基準 χ^2 の性質

- $0 \leq \chi^2$
- 大きいほど, 想定した母分布とちがいそう.
- 実は自由度 $C-1$ のカイ二乗分布にしたがう.

L13-Q2

Quiz(ピアソンの χ^2 と適合度の検定)

ある商品のサイコロは, 1 から 6 までの目が, 確率 $\frac{1}{6}$ ででるとされている. これが本当か確かめるために, 実際に $N = 60$ 回投げて試してみた. 度数 (人数) は下の表のようになった.

目	1	2	3	4	5	6
度数	14	8	6	12	11	9

1. ピアソンの適合度基準 χ^2 を求めよう.
2. この標本が, 想定される母分布に適合するかどうか, 有意水準 $\alpha = 0.05$ で, 適合度のカイ二乗検定を行って判定しよう.

適合度のカイ二乗検定

適合度のカイ二乗検定

- 検定統計量=検査薬の発色レベルピアソンの適合度基準 χ^2.
- Excel での計算方法 `chisq.test(`度数の範囲`,` 期待度数の範囲`)` で計算される p 値を 有意水準 α と比較し, p 値 $<$ 有意水準なら発色.
「分布にしたがっているとはいえない」

お知らせ

- 2019-01-22 火 1 は補講．期末試験シミュレーション問題演習．この日の出席や提出による加点はありません．
- 2019-01-28 月 2 期末試験
  - 30 ピーナッツ/科目 100 ピーナッツ
  - 60 分
  - 紙は何でも持込可．