2 つの質的変数の相関

樋口さぶろお

龍谷大学理工学部数理情報学科

使える統計!　L14(2014-01-15 Wed)

今日の目標

1. 2×2 クロス集計表からピアソンの χ^2 が計算できる
2. 2×2 クロス集計表からクラメールの連関係数 V が計算できる

http://hig3.net

L13-S1

Quiz 解答:区間推定

これはサイズ 10 の標本.
標本平均値は

$$\frac{1}{10}[0+0+0+0+0+0+10+10+30+100] = 15(\text{円})$$

. よって, 母平均値は 15 円と推定される.
標本 (不偏) 分散は

$$\frac{1}{10-1}[(0-15)^2\times 6+(10-15)^2\times 2+(30-15)^2+(100-15)^2] = 930.6(\text{円}^2)$$

. よって, 母分散は 930.6 円と推定される.
母平均値 μ の信頼係数 95%の信頼区間は,

$$15-1.96\times\sqrt{\frac{930.6}{10}} < \mu < 15+1.96\times\sqrt{\frac{930.6}{10}}$$

すなわち,

$$-3.9 < \mu < 33.9$$

L13-S2

Quiz 解答:区間推定

母平均値 μ の推定値は, 標本平均値で与えられ,

$\overline{x} = \frac{1}{5}[10 + 20 + 30 + 30 + 110] = 40$(分)

母分散の推定値は, 標本 (不偏) 分散で与えられ,

$s^2 = \frac{1}{5-1}[(10-40)^2 + (20-40)^2 + (30-40)^2 + (30-40)^2 + (110-40)^2] = 1600$(分^2)

よって, 母平均値 μ の信頼係数 99% の信頼区間は,

$$40 - 2.58 \times \sqrt{\frac{1600}{5}} < \mu < 40 + 2.58 \times \sqrt{\frac{1600}{5}}$$

すなわち,

$$-6.1 < \mu < 86.1.$$

L13-S7

Quiz 解答:母比率の区間推定

1. 母比率 p の推定値は, $\frac{35}{50} = 0.7$.
2. 分散は $\frac{1}{50} \cdot 0.7 \times (1 - 0.7) = \frac{1}{50} \times 0.21 = 0.0042$ と見積もられる.
   母比率 p の信頼係数 95%の信頼区間は,

$$\begin{aligned} 0.7 - 1.96 \times \sqrt{0.0042} < & p < 0.7 + 1.96 \times \sqrt{0.0042} \\ 0.7 - 0.13 < & p < 0.7 + 0.13 \\ 0.57 < & p < 0.83 \end{aligned}$$

   信頼係数 95%で当選確実ってことですね.
3. 母比率 p の信頼係数 95%の信頼区間は,

$$\begin{aligned} 0.7 - 2.58 \times \sqrt{0.0042} < & p < 0.7 + 2.58 \times \sqrt{0.0042} \\ 0.7 - 0.17 < & p < 0.7 + 0.17 \\ 0.53 < & p < 0.87 \end{aligned}$$

   信頼係数 99%でも当選確実ってことですね.

大注意 $p = 35/50 = 7/10$ だからといって, $n = 10$ としてはいけない. n は標本サイズだから 50. これが大きいほど, 信頼区間は短くなり, 推定は正確になる.

性別と血液型って無関係?

データの個数 $N = 12$.

質的変数が 1 つ! 度数 (人)

	A 型	A 型以外
	3	9

母比率 $= \frac{3}{12} = 0.25$.

質的変数が 2 つ! クロス集計表

	A 型	A 型以外
女子	1	2
男子	4	5

性別と血液型って ‘無関係’? ‘関係ある’?

関係ある

	A 型	A 型以外
女子	1	2
男子	1	8

関係ない

	A 型	A 型以外
女子	1	2
男子	3	6

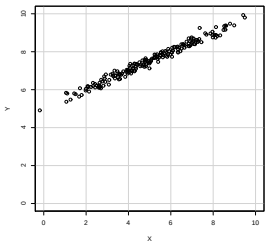
A 型の母比率は同じ

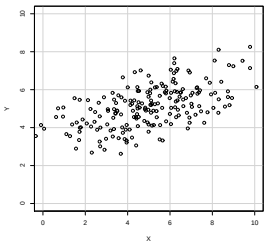
女子の母比率は同じ

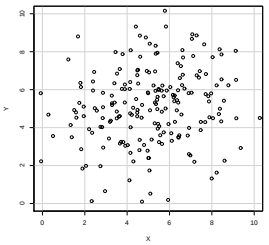
関係の程度を表す数値が欲しい!

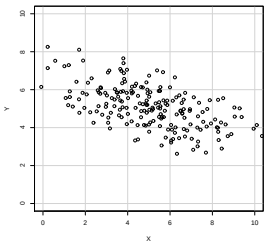
前にも似たことやってた: 身長と体重って無関係?

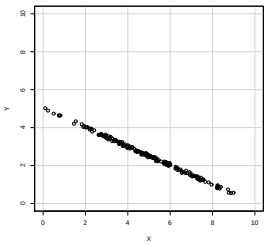
量的変数が 2 個! X:身長, Y:体重
散布図


強い正の相関
$r = 0.99$


弱い正の相関
$r = 0.55$


無相関
$r = 0$


弱い負の相関
$r = -0.55$


強い負の相関
$r = -0.99$

$$\text{相関係数 } r = \frac{X, Y \text{ の共分散 } C_{XY}}{(X \text{ の標準偏差} \sigma_X) \times (Y \text{ の標準偏差} \sigma_Y)}$$

- $-1 \leq r \leq +1$. 絶対値 $|r|$ が大きいほど '関係が強い'
- $r = \pm 1$: データ点が一直線上に載っちゃう

性別-血液型でも r みたいなのあればいいのに〜

もし無関係だったら?: 期待度数

まず合計欄を追加. 周辺分布

	A 型	A 型以外	計
女子	1	2	3
男子	4	5	9
計	5	7	12

- 全体の女子の母比率は $\frac{3}{12}$
- 全体の A 型の母比率は $\frac{5}{12}$

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

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

人くらいのはず

ピアソンの χ^2

期待度数一覧

	A 型	A 型以外	計
女子	$12 \times \frac{3}{12} \times \frac{5}{12} = 1.25$	$12 \times \frac{3}{12} \times \frac{7}{12} = 1.75$	3
男子	$12 \times \frac{9}{12} \times \frac{5}{12} = 3.75$	$12 \times \frac{7}{12} \times \frac{4}{12} = 5.25$	9
計	5	7	12

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

	A 型	A 型以外
女子	$(1 - 1.25)^2$	$(2 - 1.75)^2$
男子	$(4 - 3.75)^2$	$(5 - 5.25)^2$

ピアソンの χ^2(カイ 2 乗)

$$\chi^2 = \frac{(\text{度数} - \text{期待度数})^2}{\text{期待度数}} \text{の合計}$$

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 2 乗) の性質

- $0 \leq \chi^2$.
- 大きいほど ‘独立ではなさそう’
- データの個数 n が大きいほど大きくなる.

クラメールの連関係数 V

クラメールの連関係数 V

$$V = \sqrt{\frac{\chi^2}{n}}$$

例 $V = \sqrt{\frac{0.11685}{12}} = 0.0987$

クラメールの連関係数 V の性質

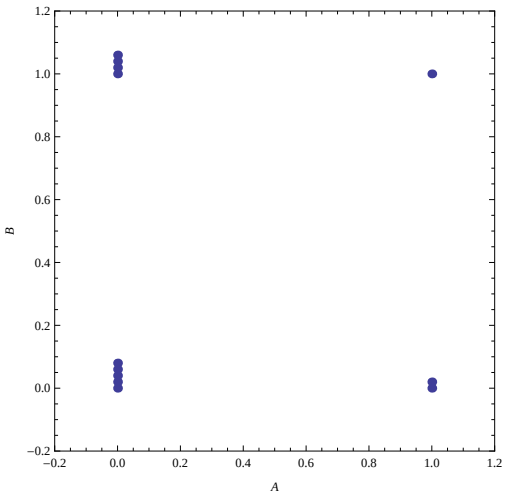
$0 \leq V \leq 1.$

- $V = 0$ 関係なし
- $V = 1$ 関係あり

相関係数との関係:ダミー変数

- 女子 $A = 1$, 男子 $A = 0$.
- A 型 $B = 1$, A 型以外 $B = 0$.

というように量的変数にしちゃえば? …ダミー変数



	A 型	A 型以外
女子	1	2
男子	4	5

$\rightsquigarrow$ 相関係数 r が求まる. 意味あるの?

- 0 と 100 じゃいけないの?
- 0 と 1 を逆にしたら?

r と連関係数 V の関係

$$|r| = V$$

r は (符号くらいしか) 変化しない. 意味ある.

2×2 よりサイズが大きいとき

- χ^2: 同じ定義. いくらでも大きくなる.
- V: 定義をちょっと変更すると, いつでも $0 \leq V \leq 1$.
- r: うまく定義できない. だって…

L14-Q1

Quiz(ピアソンの χ^2 とクラーメルの連関係数 V)

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. クラーメルの連関係数 V を求めよう.