2 つのカテゴリ変数の間の関係

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 L11(2018-12-17 Mon)

最終更新: Time-stamp: "2018-12-17 Mon 13:53 JST hig"

今日の目標

- 「独立である」と「関係がある」の関係を説明できる
- 独立性の指標のピアソンの χ^2 を説明できる
- よい/わるい測定の意味を説明できる

ここまで来たよ

カテゴリ変数

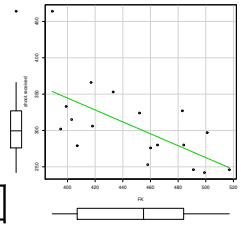
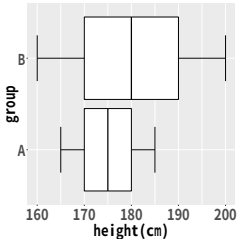
今回の対象=質的変数
その中でも, 名義変数=カテゴリ (カル) 変数
順序や距離がなくぜんぶが対等. 例: 血液型, 性別, 携帯電話番号, チーム
A 型, B 型などがカテゴリ

2 カテゴリなら, 0,1 のように番号を振って量的と思える
3 カテゴリ以上なら, 順序や間隔によるので離散型には帰着できない.

なぜなら

ここまで来たよ

関係ある/関係ないの判定

- x:量的, y 量的. x と y に相関があるか?
  - 散布図 
  - 相関係数 r, 回帰係数, 決定係数 $|r|^2$
- x:カテゴリ (そろばん ABC), y 量的. y は x によって違うの?
  - 箱ひげ図 

  - 2 群の平均値の差の検定, 分散分析, 級間平方和/級内平方和
- x:カテゴリ (男女), y カテゴリ: 禁煙の有無. x と y は関係あるの?
  - 二元分割表

	A 型	A 型以外
女子	1	2
男子	4	5

  - ピアソンの χ^2

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.

性別と血液型は関係ある?

‘関係ある’ 度を考えたい. 将来的には検定に使いたい.
「性別と血液型は関係ある」 の否定は,

- 性別と血液型は関係ない
- 性別と血液型は独立
- こんな感じになってる

	A 型	A 型以外
女子	$n_{11} = 1$	$n_{12} = 2$
男子	$n_{21} = 3$	$n_{22} = 6$

女 A:女 X=男 A:男=A:X=1:2
女 A:男 A=女 X:男 X=女:男=1:3
男 A の数 n_{21} は次のようにして計算できる $12 \times \frac{1}{3} \times \frac{3}{4}$.

L11-Q1

Quiz(無相関な二元分割表)

次の二元分割表を, 右利きと早生まれが独立である形に完成させよう. ただし, 標本サイズは $N = 28$ である.

	右利き	右利きでない
早生まれ	2	5
早生まれでない		

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 $(r-1)(c-1)$ のカイ二乗分布にしたがう.

L11-Q2

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.

L11-Q3

Quiz(ピアソンの χ^2)

次の 4 つの二元分割表について, ピアソンの χ^2 の大きさの順序は?

	A	B
X	40	0
Y	0	60

	A	B
X	0	40
Y	60	0

	A	B
X	50	10
Y	10	30

	A	B
X	16	24
Y	24	36

ここまで来たよ

クラメールの連関係数 V

クラメールの連関係数 V

χ^2:ピアソンの χ^2, N:サンプルサイズ.

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

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

クラメールの連関係数 V の性質

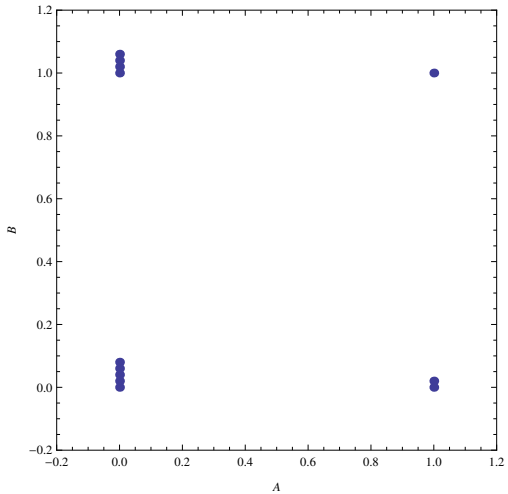
χ^2 を, 相関係数 r みたいに $0 \leq V \leq 1$ を満たすように変換したもの

- $V = 0$ 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?

2×2 のときの r と連関係数 V の関係

$$|r| = V$$

ここまで来たよ

混同行列

本当の性質と, 不正確な測定 (検査) についての, 2×2 の二元分割表のこと.

- 病気である/病気でない
- 検査で陽性になった/検査で陰性になった

の二元分割表

Confusion matrix 混同行列

	検査で陽性	検査で陰性
病気である	True Positive 真陽性	False Negative 偽陰性
病気でない	False Positive 偽陽性	True Positive 真陰性

「関係ある」ほど, よい測定

独立 $\rightarrow$ 何の意味もない測定

2 つの測定 (検査) のよさを比較する, χ^2 以外の指標

- Precision=適合率=精度=TP/(TP+FP)
- Recall=検出率=感度=TP/(TP+FN)
- Specificity=特異度=TN/(TN+FP)

ここまで来たよ

シンプソンのパラドクス

全体の比率は, 各組の比率からわかるか?

- A 組を A1 組と A2 組に分割.
- B 組を B1 組と B2 組に分割.

同じ試験を実施.

- A1 組と B1 組の合格率を比較すると A1 組が上.
- A2 組と B2 組の合格率を比較すると A2 組が上.

このとき, A 組全体の合格率は B 組全体より上?

	A1	B1	A2	B2	A	B
合格	2	30	25	1	27	31
不合格	1	20	25	2	26	22

こういう例の作り方

お知らせ

- 次回 2019-01-06 月 2 は (たぶん)5-203 で
- 図書館ミニ講義「確率を学ぶ〜年末ジャンボ宝くじが当たる確率は !? 〜」by 樋口
  - 2018-12-20 木 12:45-13:15
  - 生協コンビニ地下スチューデントコモンズ (瀬田) ミーティングスペース
- レポート 1(長くない)
  - Manaba で振り返りの作文的なもの https://manaba.ryukoku.ac.jp
  - 2018-12-17 月夜 まで
- 期末試験計画
  - 30 ピーナッツ/科目 100 ピーナッツ
  - 60 分
  - 2019-01-28 月