カテゴリ変数と独立性の検定

樋口さぶろお

龍谷大学理工学部数理情報学科

確率統計☆演習 II L04(2015-05-01 Fri)

最終更新: Time-stamp: "2015-05-01 Fri 22:28 JST hig"

今日の目標

- 2 変数のカテゴリ変数の標本が与えられたとき, Excel のピボットテーブルを使ってクロス集計表が作れる
- 2 変数のカテゴリ変数の標本が与えられたとき, χ^2 が計算でき, 独立性の検定が行える.

http://hig3.net

L03-S1

Quiz 解答:条件付き分布

1. $$P(X=x)=\begin{cases}\frac{7}{12} & (x=2)\\ \frac{5}{12} & (x=3)\\ 0 & (\text{他})\end{cases},\quad P(Y=y)=\begin{cases}\frac{3}{12} & (y=3)\\ \frac{9}{12} & (y=7)\\ 0 & (\text{他})\end{cases}$$

2. $$P(X=x|Y=3)=\begin{cases}\frac{2}{3} & (x=2)\\ \frac{1}{3} & (x=3)\\ 0 & (\text{他})\end{cases}$$

   $$P(Y=y|X=3)=\begin{cases}\frac{1}{5} & (y=3)\\ \frac{4}{5} & (y=7)\\ 0 & (\text{他})\end{cases}$$

L03-S2

Quiz 解答:ベイズの公式

1.

$y\backslash x$	1	2
10	21/40	4/40
20	9/40	6/40

2.

$$P(X=x|Y=10)=\begin{cases}\frac{21}{25} & (x=1)\\ \frac{4}{25} & (x=2)\\ 0 & (他)\end{cases}$$

ここまで来たよ

1. 略解: 条件付き確率とベイズの公式

2. カテゴリ変数と独立性の検定
   - カテゴリ変数
   - 質的変数が 2 つ
   - 独立性の検定

カテゴリ変数

きょうは寄り道だけど実用的に重要な回
確率統計☆演習 II の主な対象=量的変数

- 離散型 表 2 項分布, ポアソン分布, $\cdots$, x は整数
- 連続型 確率密度関数 正規分布, χ^2 分布,$\cdots$, x は実数

今日の対象=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
A 型, B 型などがカテゴリ

2 カテゴリなら, 0,1 に置きかえて離散型と思える
$\rightarrow$ 比率

確率統計☆演習 I(2014)L11

3 カテゴリ以上なら, [ ] 離散型には帰着できない.

ここまで来たよ

1. 略解: 条件付き確率とベイズの公式

2. カテゴリ変数と独立性の検定
   - カテゴリ変数
   - 質的変数が 2 つ
   - 独立性の検定

2 つのカテゴリカル変数

未知の母分布

	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

分割表, クロス集計表

Excel

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

$f_{ij}, 1 \leq i \leq c, 1 \leq j \leq r.$ 行数 r, 列数 $c2.$

性別と血液型は関係ある?

関係ある の否定は,

- 関係ない
- 性別と血液型は確率変数として独立である
- P(血液型=A 型, 性別=男)=P(性別=男)× P(血液型=A 型) $= p_i \times q_j$.

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

	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	$12 \times \frac{9}{12} \times \frac{5}{12} = 3.75$	$12 \times \frac{9}{12} \times \frac{7}{12} = 5.25$	9
計	5	7	12

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

(ずれ)^2 = (度数 − 期待度数)^2

「独立でない度」:ピアソンの χ^2(カイ 2 乗)

$p_i (i = 1, \ldots, c)$, q_j $(j = 1, \ldots, r)$ を, 標本から推定した周辺分布としたとき,

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

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 2 乗) の性質

- $0 \leq \chi^2$.
- 大きいほど '独立でなさそう'
- 実は, 自由度 $(r-1)(c-1)$ の χ^2 分布にしたがう.

Example

Excel で分割表を作って χ^2 を求めようピボットテーブルという Excel の機能を使うのが便利
RaMMoodle `https://el.math.ryukoku.ac.jp/moodle` のデータをクロス集計表にして, 独立性の検定をして, 課題にアップロード.
標本のデータ部分を選択して, 挿入 > ピボットテーブル.

ここまで来たよ

1. 略解: 条件付き確率とベイズの公式

2. カテゴリ変数と独立性の検定
   - カテゴリ変数
   - 質的変数が 2 つ
   - 独立性の検定

統計的仮説検定

ぎりぎりのデータから Yes/No のいちおうの結論を出す, 科学業界で合意された方法が

確率統計☆演習 I(2014)L12

検定 (test)=統計的仮説検定 (statistical hypothesis test)

真の母平均値は 55g と異なる, を証明したい.

しか〜し, $\neq$ の証明はやりにくい 54g である, ことが証明できれば十分だけど, 有限個の標本からはとうてい無理.

こういうときの常套手段は［　　］. 否定した命題「55g である」を仮定して矛盾を導く.

注意

以下, 証明, 矛盾 は, 証明みたいなもの, 矛盾みたいなもの (統計的な, 確率 $\alpha = 0.05$ で間違っている) です. この回の授業のローカル用語.

α: 有意水準. どれだけの誤りを許すか. 大きいほど大胆/頼りない証明.

ふつうは 0.01 or 0.05.

帰無仮説と対立仮説

- H_0:帰無仮説 (null hypothesis) = 背理法の仮定 = 「真の母平均値 μ_1 は $\mu_0 = 55\mathrm{g}$ に等しい」「この標本はこの母分布から抽出したものである」
- H_1:対立仮説 (alternative hypothesis) = 示したい命題 = 「真の母平均値 μ_1 は $\mu_0 = 55\mathrm{g}$ でない」「この標本はこの母分布から抽出したものでない」

棄却・採択・有意

H_0 から 矛盾 が導かれるとき,

- H_0 を棄却 (reject) する
- (H_1 が採択 (accept) される)
- 差が有意である (significant)

などという. H_1 が 証明 されたということ.

- H_0 を棄却できない
- H_0 を採用 (accept) する
- μ_0 と μ_1 の差が有意でない (not significant) である

などという. このとき, H_1 でないことを 証明 できたわけではない

[ ] したケースに相当=標本は帰無仮説と 矛盾 しない. 結論なし.

答案や論文での検定の書き方

1. 有意水準を書く
2. (検定の名前があれば)「…検定」を行う, と書く
3. 帰無仮説を書く
4. 選択した検定統計量 Y と, それが (帰無仮説のもとで) 従う分布を書く
5. 標本に対する検定統計量の値 y_1 を書く.
6. Y が y_1 より極端な値となる確率を求める ($=p$). それが α より大きい/未満なら, 帰無仮説を採択する/棄却する (=有意でなかった/有意だった) と書く.

まあ最初のうちは, 参考書を見て, この状況ではこの検定, という解法パターン的対処でもやむをえないかも. ただし, 不適切な検定を無理に使わないようにしよう.

独立性の検定

1. 「有意水準 $\alpha = \ldots$ で,
2. 独立性の検定を行う」
3. 帰無仮説を, 'X,Y は独立である' とする.
4. このとき, ピアソンの適合度基準 χ^2 は自由度 $(c-1)(r-1)$ の χ^2 分布に従う.
5. いま, $\chi^2 = \ldots$ である.
6. $\chi^2_\alpha(k-1)$ との大小関係は, ... なので, 帰無仮説を採択する/棄却する. すなわち, X と Y には有意な関係がなかった/あった.

L04-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.

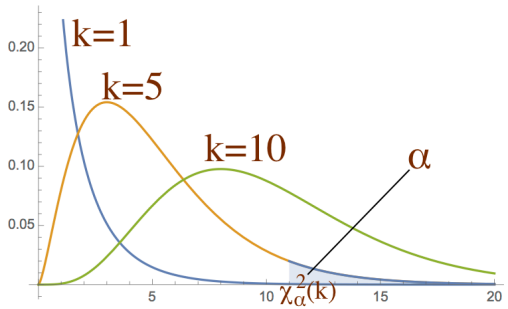
	右利き	右利きでない
早生まれ	1	1
早生まれでない	3	1

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 χ^2 検定を行って判定しよう.

χ^2 分布表

$\alpha = P(\chi^2 > \chi^2_\alpha(k))$.

$k\backslash\alpha$	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00003927	0.0001571	0.0009821	0.003932	0.01579	2.706	3.841	5.024	6.635	7.879
2	0.01003	0.02010	0.05064	0.1026	0.2107	4.605	5.991	7.378	9.210	10.60
3	0.07172	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.34	12.84
4	0.2070	0.2971	0.4844	0.7107	1.064	7.779	9.488	11.14	13.28	14.86
5	0.4117	0.5543	0.8312	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.6757	0.8721	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.9893	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2



Math ラウンジ=チューター

月火水木昼, 1-614

統計検定を取ろう!

http://www.toukei-kentei.jp/

1. 2 級 or 3 級をお奨めします
2. 2015-05-15 申込締切, 2015-06-21 検定実施

manaba 出席カード提出

https://attend.ryukoku.ac.jp