

クロス集計と独立性の検定

高木廣文
東邦大学医学部看護学科
国際保健看護学研究室

●クロス集計と独立性の検定

喫煙の状況を調べる場合、単に喫煙者数や割合を求めるだけでなく、男女差や、飲酒との関係なども知りたいと考えるのではないだろうか。

・ 質的データ間の関係を調べるには？

⇒ 「クロス集計 cross tabulation」

●関係をどのように調べるのか？

表 A薬とB薬の症状改善について

	著効	有効	不変	悪化	計
A薬	8	10	6	1	25
B薬	3	9	10	4	26
計	11	19	16	5	51

Q. どのようにすれば2つの薬剤の有効性に相違があるといえるのか？

クロス表の記号：

表 変数AとBのクロス表

A	B					計	
	B ₁	B ₂	...	B _j	...		B _l
A ₁	n ₁₁	n ₁₂	...	n _{1j}	...	n _{1l}	n _{1.}
⋮	⋮	⋮	...	⋮	...	⋮	⋮
A _i	n _{i1}	n _{i2}	...	n _{ij}	...	n _{il}	n _{i.}
⋮	⋮	⋮	...	⋮	...	⋮	⋮
A _k	n _{k1}	n _{k2}	...	n _{kj}	...	n _{kl}	n _{k.}
計	n _{.1}	n _{.2}	...	n _{.j}	...	n _{.l}	N

クロス表に関する用語：

2変数A, Bに関してクロス表を作成：

- ・ 変数Aはk個のカテゴリ,
- ・ 変数Bはl個のカテゴリ：
「k × lのクロス表」
(k × l) 個のマス目 - 「セル cell」

● 変数AについてカテゴリA_i, かつBについてB_jという特性をもつものの度数を n_{ij} のように表す。

クロス表に関する用語の続き

- ・ カテゴリA_iの度数の合計： $n_{i.}$
- ・ カテゴリB_jの度数の合計： $n_{.j}$

$$n_{i.} = n_{i1} + n_{i2} + \dots + n_{il}$$

$$n_{.j} = n_{1j} + n_{2j} + \dots + n_{kj}$$

- ・ 全体の標本数を Nとする。

セル度数の期待値

- 2つの変数A, Bが独立な場合:
- ・ AのカテゴリA_iの割合: $n_{i\cdot} / N$
 - ・ BのカテゴリB_jの割合: $n_{\cdot j} / N$

対応するセルの期待値: e_{ij}

$$e_{ij} = \frac{n_{i\cdot}}{N} \times \frac{n_{\cdot j}}{N} \times N = \frac{n_{i\cdot} \times n_{\cdot j}}{N}$$

クロス表の記号:

表 変数AとBのクロス表

A	B				計		
	B ₁	B ₂	...	B _j		...	B _l
A ₁	n ₁₁	n ₁₂	...	n _{1j}	...	n _{1l}	n _{1.}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A _i	n _{i1}	n _{i2}	...	n _{ij}	...	n _{il}	n _{i.}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A _k	n _{k1}	n _{k2}	...	n _{kj}	...	n _{kl}	n _{k.}
計	n _{.1}	n _{.2}	...	n _{.j}	...	n _{.l}	N

カイ 2 乗値統計量

$$\left[\frac{(\text{期待値} - \text{実測値})^2}{\text{期待値}} \right] \text{の全セルの合計}$$

検定のためのカイ 2 乗値の計算

2変数A, Bの関連の有無を検定するための統計量: χ_0^2

$$\chi_0^2 = N \left(\frac{n_{11}^2}{n_{1\cdot} \times n_{\cdot 1}} + \frac{n_{12}^2}{n_{1\cdot} \times n_{\cdot 2}} + \dots + \frac{n_{kl}^2}{n_{k\cdot} \times n_{\cdot l}} - 1 \right)$$

$$= N \left(\sum_{i=1}^k \sum_{j=1}^l \frac{n_{ij}^2}{n_{i\cdot} \times n_{\cdot j}} - 1 \right)$$

自由度 $\nu = (k-1) \times (l-1)$ のカイ2乗分布

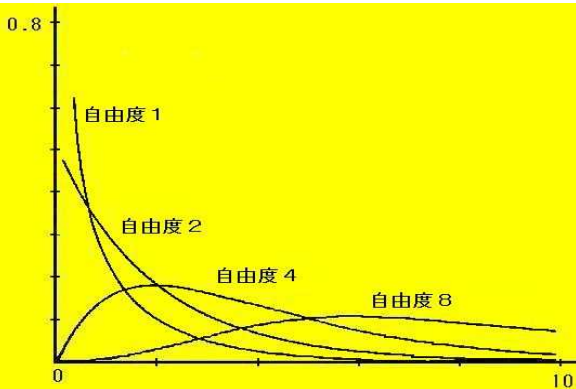


図 自由度による χ^2 分布の形状

クロス表の自由度について

表 A薬とB薬の症状改善について

	著効	有効	不変	悪化	計
A薬	*	*	*	*	25
B薬	*	*	*	*	26
計	11	19	16	5	51

Q. 最大でいくつ自由にセルの値を決められるのか?

検定の方法

χ_0^2 と自由度 ν のカイ2乗分布の上側5% (または1%) 点の値 $\chi_{0.05}^2(\nu)$ と比較:

$\Rightarrow \chi_0^2 > \chi_{0.05}^2(\nu) \rightarrow$ 有意

$\chi_0^2 < \chi_{0.05}^2(\nu) \rightarrow$ 態度保留

【注意】

実際には、 χ_0^2 と自由度 ν のカイ2乗分布からp値(有意確率)を求めて検定する。

クロス表の検定での注意事項

- ・カテゴリ数が大きくなるにつれ、セルの個数も急激に増加。
 - ・セルの度数が極小になることがある。
 - ・ χ_0^2 のカイ2乗分布への適合が悪化。
- ・正確には、各セルの期待値が問題: 期待値が5以下の場合、カイ2乗分布による近似はよくないといわれている。
 \Rightarrow 可能ならば隣合うカテゴリ同士を併合して、検定を行う。

クロス表の検定の例題:

前述のA薬とB薬の症状改善についてのクロス表の検定を行う。

$$\begin{aligned} \chi_0^2 &= 51 \times \left[\frac{8^2}{25 \times 11} + \frac{10^2}{25 \times 19} + \frac{6^2}{25 \times 16} + \frac{1^2}{25 \times 5} \right. \\ &\quad \left. + \frac{3^2}{26 \times 11} + \frac{9^2}{26 \times 19} + \frac{10^2}{26 \times 16} + \frac{4^2}{26 \times 5} - 1 \right] \\ &= 5.108 \end{aligned}$$

例題の続き:

χ_0^2 の自由度: $(2-1) \times (4-1) = 3$

\Rightarrow 自由度3のカイ2乗分布の上側5%点の値は7.815

$\Rightarrow \chi_0^2 = 5.108 < 7.815 \rightarrow$ 判断保留
($p = 0.164$)

\Rightarrow A薬とB薬の効果に差があるとはいえない。

●カテゴリ併合の例

期待値が5未満のセルがあるので、カテゴリの併合を行う。

表 カテゴリの併合

	著効	有効	その他	計
A薬	8	10	7	25
B薬	3	9	14	26
計	11	19	21	51

不変と悪化を併合し「その他」とした。

併合の続き:

検定のための統計量:

$\chi_0^2 = 4.641$

自由度: $(3-1) \times (2-1) = 2$

\Rightarrow 自由度2のカイ2乗分布の

上側5%点 $5.991 > \chi_0^2 = 4.641$

\Rightarrow 保留(有意でない)

- ・併合は無意味か?

p値の変化: $p = 0.164 \rightarrow 0.098$

四分表の検定

- 慢性前立腺患者に、無作為にA薬とプラセボを各15例ずつ投与し、自覚症状について効果を調べたところ、下の表に示したような結果が得られた。このデータから、A薬に効果があるといえるだろうか。

	自覚症状		計
	改善	不変	
A薬	11	4	15
プラセボ	6	9	15
計	17	13	30

四分表の一般的な表現

A	B		計
	O	x	
O	a	b	n_1
x	c	d	n_2
計	m_1	m_2	N

$$n_1 = a + b, n_2 = c + d, m_1 = a + c, m_2 = b + d$$

四分表の期待値

- ・ 変数AがO, かつ変数BがOのセル:
観察値: a
期待値: $e(a)$ AとBが独立ならば

$$e(a) = \frac{n_1}{N} \times \frac{m_1}{N} \times N = \frac{n_1 \times m_1}{N}$$

期待値とカイ2乗値

- 各セルの期待値と観察値から得られる統計量
- カイ2乗値: χ_0^2

$$\chi_0^2 = \frac{[\text{期待値} - \text{観察値}]^2}{\text{期待値}} \text{の合計}$$

四分表の独立性の検定

- 2変数の両方が2カテゴリからなる場合:
「4分表」 four-fold table
「独立性の検定」のための統計量:

$$\chi_0^2 = \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

- ・ カイ2乗分布への当てはまりが悪い。

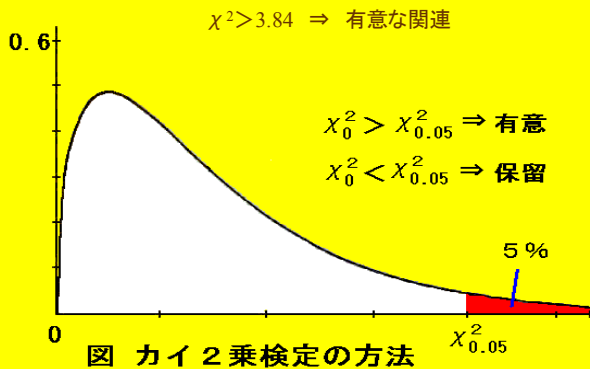
イエーツの連続性の補正

- 四分表の検定の「連続性の補正」:
カイ2乗分布への当てはまりがよくなる

$$\chi_Y^2 = \frac{N(|ad - bc| - N/2)^2}{n_1 n_2 m_1 m_2}$$

- ・ 連続性の補正: $|ad - bc| > N/4$ の場合
- ・ そうでない場合には: $\chi_Y^2 = 0$

カイ2乗検定の方法



例題の計算:

$$\chi^2 = \frac{30 \times (|11 \times 9 - 4 \times 6| - 30/2)^2}{15 \times 15 \times 17 \times 13}$$

$$= 2.172$$

$$\Rightarrow \chi^2 < 3.84 \quad (p = 0.141)$$

【結論】

A薬による自覚症状改善の程度が、プラセボより効果があるとはいえない。

Fisherの直接確率(正確な確率)

- 2 × 2クロス表の検定のための有意確率を直接求める方法。
⇒ t検定やカイ2乗検定などとは異なり、特定の統計量の分布を考えなくてよい。
- 直接確率は、偶然に観測された4つのセルの度数が a, b, c, d となる確率、およびそれ以上に特定方向に偏った状態の確率を求めるための方法。

正確な確率の計算:

- 計の欄の数値を固定した場合:
変数Aが○, 変数Bも○のセルの度数が a となる場合を考える:
(1) 全標本数 N のうち, n_1 が変数Aが○となる場合の数:

$${}_N C_{n_1} = \frac{N!}{n_1! \times (N - n_1)!} = \frac{N!}{n_1! \times n_2!}$$

正確な確率の計算(その2):

- (2) n_1 のうち a が変数Bの○に, b が × となる組み合わせの数:
⇒ Bの○: m_1 人中 a 人
Bの×: m_2 人中 b 人
となる場合の数の積:

$${}_m C_a \times {}_m C_b = \frac{m_1! \times m_2!}{a! \times b! \times c! \times d!}$$

正確な確率の計算(その3):

- (3) (1) と (2) から, 4つのセルの度数が偶然に a, b, c, d となる確率 p_0 :

$$p_0 = \frac{{}_m C_a \times {}_m C_b}{{}_N C_{n_1}} = \frac{n_1! \times n_2! \times m_1! \times m_2!}{N! \times a! \times b! \times c! \times d!}$$

正確な確率の計算（その4）：

検定のためには、表の状態を更に偏らせた場合の確率が必要：

表 AとBの四分表

		B		
A		O	×	計
O	a + i	b - i	n_1	
×	c - i	d + i	n_2	
計	m_1	m_2	N	

正確な確率の計算（その5）：

- $a d > b c$ とする：
偏った状態とは、表に示したように、積の大きい方はより大きく、小さい方はより小さくなるように、各セルに1ずつ度数を増減すること。
- ・ b と c のうち小さい方の値を k 。
各セルの度数は0以上：
 i のとり得る値は、 $0, 1, \dots, k$
すべての4分表の確率の計算が必要

正確な確率の計算（その6）：

- $i = 1$ の場合の確率： p_1

$$p_1 = \frac{b \times c}{(a+1) \times (d+1)} \times p_0$$

- $i = 2$ の場合の確率： p_2

$$p_2 = \frac{(b-1) \times (c-1)}{(a+2) \times (d+2)} \times p_1$$

正確な確率の計算（その7）：

- 一般に $i + 1$ の場合の確率： p_{i+1}

$$p_{i+1} = \frac{(b-i) \times (c-i)}{(a+i+1) \times (d+i+1)} \times p_i$$

正確な確率の計算（その8）：

- すべての確率の合計を計算： p_F

$$p_F = p_0 + p_1 + p_2 + \dots + p_k$$

【注意】

確率 p_F は、特定方向のみの偏りを問題としており、片側検定である。
通常、この確率を2倍した値を検定のための有意確率として用いる場合が多い。

四分表の検定はどの方法で？

【問題点】

4分表の検定には、フィッシャーの直接確率を用いるか、カイ2乗検定を用いるか？

A. できる限りフィッシャーの直接確率

標本数が大きい場合、階乗計算は困難。
適当な統計パッケージHALBAU/HALWINなどが必要（ちょっと宣伝）。
ホームページでも計算できる。

例題の計算：

- ①「計」の欄の数値が一定の場合に、表のようなクロス表が偶然できる確率を計算：

$$p_0 = \frac{15! \cdot 15! \cdot 17! \cdot 13!}{30! \cdot 11! \cdot 4! \cdot 6! \cdot 9!} = 0.057046$$

例題の計算（その2）：

	自覚症状		計
	改善	不変	
A 薬	15	0	15
プラセボ	2	13	15
計	17	13	30

- ・ A薬で「不変」のセルが3, 2, 1となる3つのクロス表の確率もすべて計算する。

例題の計算（その3）：

- ② A薬で不変のセルの度数が3の場合：

$$p_1 = \frac{4 \times 6}{12 \times 10} \times p_0 = 0.011409$$

- ③ 度数が2の場合：

$$p_2 = \frac{3 \times 5}{13 \times 11} \times p_1 = 0.001197$$

例題の計算（その4）：

- ④ 度数が1の場合と0の場合：

$$p_3 = \frac{2 \times 4}{14 \times 12} \times p_2 = 0.00005699$$

および

$$p_4 = \frac{1 \times 3}{15 \times 13} \times p_3 = 0.000000088$$

例題の計算（その5）：

すべての確率の和 p_F ：

$$p_F = p_0 + p_1 + p_2 + p_3 + p_4 = 0.0697$$

$$\Rightarrow 2p_F = 0.139$$

有意確率は13.9% \Rightarrow 有意でない。

【結論】

A薬の効果があるとはいえない。