

統計学第7回 「カテゴリ変数2つの解析(2)」

中澤 港

<http://phi.ypu.jp/stat.html>

[<minato@ypu.jp>](mailto:minato@ypu.jp)



2つのカテゴリ変数の関連の強さを みるには？

- ▶ カイ二乗検定やフィッシャーの直接確率検定では、関連の強さはわからない。
- ▶ 1つのカテゴリ変数をグループ分けを示す変数と考え、もう1つのカテゴリ値の取り方が、グループ間で何倍違うかを見る→リスク比, オッズ比(, 罹患率比)
- ▶ カテゴリの増減が一致している程度をみる→ユールのQ, ファイ係数(ρ), 四分相関係数
- ▶ 再測定したときの一致度をみる→ κ 係数

リスク・オッズ・罹患率

- ▶原因と結果にどれくらいの関連があるかを知りたいとき、原因が無いグループの結果に対して、原因があるグループの結果がどれくらい大きいかで評価するという考え方(疫学)
- ▶病気を例にとって説明すると……
- ▶病気のリスクとは、全体のうちでその病気を発症する人の割合
- ▶病気のオッズとは、発症した人の発症していない人に対する人数比
- ▶ちなみに、病気の罹患率とは、延べ観察時間あたりの病気の発生数をいう

リスク比

- ▶ リスク比は原因なし群のリスクに対する、原因あり群のリスクの比。Risk Ratio, Relative Risk などという。RR と略記する。
- ▶ 全体のうちどれくらい発症するかという情報が必要なので、曝露時点での全体がわからないデザイン(断面研究や症例対照研究)では計算できない。
- ▶ 原因の有無と発症が無関係なら1。

▶		発症あり	発症なし	
曝露あり	X	m1-X	m1	
曝露なし	Y	m2-Y	m2	
	n1	n2	N	

とすると、曝露あり群のリスクが $(X/m1)$ 、曝露なし群のリスクが $(Y/m2)$ となるので、

リスク比は、 $(X/m1)/(Y/m2) = (X*m2)/(Y*m1)$

リスク比の信頼区間

- ▶ リスク比 (RR) の分布は N が大きくなると正規分布に近づく。しかし右裾を引いているので、対数変換した方が近似が良い (RR の推定値が大きいときは立方根変換した方が近似がいいが面倒)。
- ▶ 対数変換では、95% 信頼区間の下限は、 $RR \cdot \exp(-1.96 \cdot \sqrt{1/X - 1/m_1 + 1/Y - 1/m_2})$ 、上限は、 $RR \cdot \exp(1.96 \cdot \sqrt{1/X - 1/m_1 + 1/Y - 1/m_2})$ となる。
- ▶ 一般に、95% 信頼区間が 1 をまたいでいるかどうかでリスク比の有意性を判断することが多い。しかし本質的には、信頼区間の幅そのものの情報をきちんと評価すべきである。

オッズ比

- ▶ オッズ比は原因なし群のオッズに対する、原因あり群のオッズの比。Odds Ratio。ORと略記する。
- ▶ オッズは比なので、断面研究でも症例対照研究でも計算でき、しかもそれらが数学的に一致する。
- ▶ 原因の有無と疾病の有無が無関係なら1

▶		疾病あり	疾病なし	
曝露あり	a	b	m1	
曝露なし	c	d	m2	
	n1	n2	N	

とすると、曝露あり群のオッズは a/b ，なし群は c/d となるので、オッズ比は、 $(a/b)/(c/d) = (a*d)/(b*c)$

- ▶ ただし、Rも含めて統計ソフトでは連続修正されたり最尤推定されることが多いので、この単純な定義通りにはならない

オッズ比の信頼区間

- ▶ リスク比の場合と同様，オッズ比 (OR) も右裾を引いた分布に従うので，対数変換または Cornfield の方法で正規分布に近づけ，正規近似を使って 95% 信頼区間を求める。
- ▶ 対数変換なら，95% 信頼区間の下限は $OR \cdot \exp(-1.96 \cdot \sqrt{1/a+1/b+1/c+1/d})$ 上限は $OR \cdot \exp(1.96 \cdot \sqrt{1/a+1/b+1/c+1/d})$ となる。
- ▶ 実は R には `vcd` というライブラリがあって，`install.packages("vcd")` などによって CRAN から `vcd` をインストール済みならば，`library(vcd)` して，`summary(oddsratio(クロス表, log=F))` で全部やってくれる (但し連続修正されている)。



関連性の指標の意味

- ▶ オッズ比はリスク比の良い近似となることが多い。稀な疾患の場合(「注目している事象が起こる確率がきわめて低い場合」), 患者対照研究によってオッズ比を求める方が効率が良い(テキストの例を参照)
- ▶ その他の関連性の指標としては, リスク差, 相対差, 曝露寄与率, 母集団寄与率, Yule の Q , ファイ係数 (ρ : 四分点相関係数ともいう), クラメールの V , 一致係数など。
- ▶ 同じ質問を繰り返して行った場合に同じ答えが返ってくるかどうかを見るには κ 係数を使う。

2 × 2クロス表からの関連性の指標

	発症	非発症	
曝露あり	X	m1-X	m1
曝露なし	Y	m2-Y	m2
	n1	n2	N

無修正のカイ二乗値を C と書くことにすると,

▶ リスク差 (RD) = $(X/m1) - (Y/m2)$

▶ 相対差 = $RD / (1 - (Y/m2))$

▶ 曝露寄与率 = $RD / (X/m1)$

▶ 母集団寄与率 = $\{(X+Y)/N - (Y/m2)\} / \{(X+Y)/N\}$

▶ Yule の Q = $(OR - 1) / (OR + 1)$

▶ ファイ係数 (ρ) = $\sqrt{RD * (X/n1 - (m1 - X)/n2)} = \sqrt{C/N}$

▶ クラメールの V = $\sqrt{C * N}$

▶ 一致係数 = $\sqrt{C / (C + N)}$

▶ R では下3つは, `library(vcd)` の `assoc.stats(クロス表)` で計算できる。



κ (カッパ) 係数

- ▶ 2回の繰り返し調査をしたときに、あるカテゴリ変数がどれくらい一致するかを示す指標。test-retest-reliability の尺度。
- ▶ 別々のカテゴリ変数間のクロス集計と見かけは同じだが、意味は違う。カイ二乗検定やフィッシャーの直接確率で帰無仮説とした、2つのカテゴリ変数が独立という仮定はそもそも無理なので、それらを求めても無意味。

▶	1回目○	1回目×	
2回目○	a	b	m1
2回目×	c	d	m2
	n1	n2	N

とすると、偶然の一致割合 $Pe=(n1m1/N+n2m2/N)/N$

実際の一致割合 $Po=(a+d)/N$ となるので、 $\kappa=(Po-Pe)/(1-Pe)$ という量を考えると、この統計量 κ は、完全一致で 1、偶然の一致と同じとき 0、偶然より一致度が低いとき負の値をとる。

- ▶ $V(\kappa)=Pe/\{N(1-Pe)\}$ から、 $\kappa/\sqrt{V(\kappa)} \sim N(0,1)$ を使って検定できる。R では `library(vcd)` に `Kappa(クロス表)` がある。

memo



memo

