統計学第６回
「カテゴリ変数２つの解析（１）」

看護学部　中澤　港
〈minato@ypu.jp〉
http://phi.ypu.jp/statlib/l6-2003.pdf

- ２つの変数の母比率の比較
- ２つの変数の関係の分析
  - 独立かどうか？
  - 関連の程度はどうか？（※次回説明する）

2つの変数の母比率の比較

- 例：山口県立大のキャンパスを隔てるバイパスの交通状態の観察データ
- 進行方向，型，色を変数とする。
- 進行方向で2群に分けると，津和野方面と山口市街地方面で別々に型と色という変数ができる。津和野方面の型という変数での乗用車の割合と，山口市街地方面の型という変数での乗用車の割合を比較する。

観察データの符号化

	変数		
オブザーベーション	進行方向	車の種類	車の色
1	津和野	乗用車	白
2	市街地	乗用車	白
3	市街地	乗用車	銀
:	:	:	:

典型的なデータを1としてコード化

	変数		
Obs	Dest	Type	Color
1	1	1	1
2	2	1	1
3	2	1	3
:	:	:	:

津和野方面行きと市街地方面行きとで，乗用車割合は違うか？

- R では attach() 後に table(Dest,Type) で，津和野方面は，乗用車 57 台，それ以外 3 台，市街地方面は，乗用車 25 台，それ以外 4 台とわかる。
- 津和野方面行きの乗用車割合の母比率 p1 の推定値は，57/60（標本比率を使う）
- 市街地方面行きの乗用車割合の母比率 p2 の推定値は，25/29（標本比率を使う）
- p1 と p2 に差が無いという帰無仮説の下で Z を計算し，N(0,1) で検定する。

Zの計算式

$$Z=\frac{\hat{p}_1-\hat{p}_2-E\left(\hat{p}_1-\hat{p}_2\right)}{\sqrt{V\left(\hat{p}_1-\hat{p}_2\right)}}=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(1/n_1+1/n_2\right)}}\sim N(0,1)$$

ただし $n_1p_1>5$ かつ $n_2p_2>5$ でなければならない（そうでないときはフィッシャーの正確な確率を出す）。連続修正（次画面参照）したものが次の式

$$Z=\frac{\left|\hat{p}_1-\hat{p}_2\right|-\left(1/n_1+1/n_2\right)/2}{\sqrt{\hat{p}\hat{q}\left(1/n_1+1/n_2\right)}}\sim N(0,1)$$

連続性の補正（連続修正）

- 正規分布は連続分布なのに，カテゴリ変数の各カテゴリに該当する標本数の値は離散量なので，Z がとりうる値は飛び飛びになってしまい，そのままでは近似の精度がやや良くない。
- そこで，平均から (1/n1+1/n2)/2 を足したり引いたりするのが連続性の補正である。
- 式が若干異なるけれども，後で説明するカイ二乗検定での連続性の補正も同じ考え方に基づく。

母比率の差の信頼区間

- 母比率の差の 95% 信頼区間を求めるには，標本数が多ければ，差から分散の平方根の 1.96 倍を足したり引いたりしてやればよい。
  - 1.96 は標準正規分布の 97.5 %点（上側 2.5 %点）。R では qnorm(0.975,0,1)
  - 母比率の差の分散は

$$var\left(\hat{p}_1-\hat{p}_2\right)=\hat{p}_1\left(1-\hat{p}_1\right)/n_1+\hat{p}_2\left(1-\hat{p}_2\right)/n_2$$

- 通常は連続性の修正が必要なので，さらに下限から標本数の逆数の平均（ =(1/n1+1/n2)/2 ）を引き，上限には同じ値を足す。

２つのカテゴリ変数の関係

- 研究のデザインによってさまざまな分析
  - 症例対照研究 (Case Control Study) は基本的に一時点で症例群と対照群のデータを比較し，差があるかどうか検討する。肺がんについて，過去における喫煙率を症例群と対照群で比較するとたいていの場合有意な差が出る，ということから，肺がんのリスクファクターとしての喫煙を示したことは，症例対照研究の成果である。
  - 過去における喫煙率を症例群と対照群で比較する，ということは，統計的には，過去における喫煙という変数と患者かどうかという変数が独立かどうかを調べることに他ならない。
  - 独立でないならば，どの程度関連があるのかを調べることになる。喫煙がどの程度肺がん発症率を上げるのかを調べるには，コホート研究でリスク比やオッズ比を求める。オッズ比は症例対照研究でも計算できるが，リスク比はコホート研究やコホート内症例対照研究でないと求められない。

２つのカテゴリ変数の独立性

- まず，組み合わせの度数を調べた表を作成する。これをクロス集計表と呼ぶ。
- ２つの変数が，ともにカテゴリ数２個ずつ（つまり2値変数）のときのクロス集計表は，とくに２×２分割表とか，２×２クロス集計表と呼ばれ，統計的性質が良く調べられている。
- クロス集計表をもとにして独立性を調べるには，独立である場合に期待される各セルの度数と，実際の度数が適合しているかどうかをカイ二乗検定で調べるのが1つの方法。

独立性の検定の公式

- イェーツ (Yates) の連続性の補正を行ったカイ二乗値を計算し，それが自由度1のカイ二乗分布に従うと考えて計算する。
- 標本数が少ない場合は第1種の過誤が大きくなるので，フィッシャーの正確な確率検定 (Fisher's Exact Probability Test) をする。周辺度数が決まっている場合にありうるすべての組み合わせを考え，実際に得られている表が偶然の可能性の中で得られる確率がどれくらいあるかを求めるのだが，手計算では大変なので，普通はパッケージに計算させる。

イェーツの補正をしたカイ二乗値

$$\chi_c^2 = \frac{N\left(\left|ad-bc\right|-N/2\right)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$\chi_c^2 < qchisq(0.95,1)$ ならば，有意水準 5% で，２つのカテゴリ変数が独立であるという帰無仮説は棄却されない。つまり，独立でないとはいえないことになる。

フィッシャーの直接確率

- クロス集計表を

	あり	なし	
あり	a	m1-a	m1
なし	n1-a	m2-(n1-a)	m2
	n1	n2	N

とすれば,

- 周辺度数を固定したときにこの表が得られる確率は, choose(m1,a) * choose(m2,(n1-a)) / choose(N,n1)
- つまり, m1!m2!n1!n2!/{a!(m1-a)!(n1-a)!(m2-n1+a)!N!} となる。フィッシャーの正確な確率は, それより小さな確率が得られる表の確率も全部足したものになる(両側検定の場合)ので, 計算が大変。R では fisher.test() 関数を使えば一発