分割表の解析：カイ2乗検定とフィッシャー検定

渡邊直樹

2021年11月10日

たとえば，顧客名簿からランダムに100人を抽出し，電話を用いて，あるキャンペーンの存在を知っていたか，キャンペーン期間中に来店したかを訊ねたとする．このとき，質問項目の同時集計を次のような表にした．この表からキャンペーンの周知が来店者数の増加と関連があるかを統計的に検討することができる．このようなデータの集計方法を**クロス集計**（cross tabulation）といい，クロス集計された表を**分割表**（contingency table）という．

表 1: 2×2 分割表

	来店した	来店しなかった
知っていた	38	20
知らなかった	17	25

分割表は上記の例のように2行2列のよりも大きなものでもよい．たとえば，項目Aをm個（$a_1, \ldots, a_m$）の基準で，項目Bをn個（$b_1, \ldots, b_n$）の基準で分類することもでき，この場合の分割表を $m \times n$ 分割表という．表1の分割表は 2×2 型である．

	b_1	$\ldots$	b_n
a_1	x_{11}	$\ldots$	x_{1n}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_m	x_{m1}	$\ldots$	x_{mn}

一般的な $m \times n$ 分割表の解析にはカイ 2 乗検定が適用されるが，カイ 2 乗分布は連続変数に対して定義された分布であり，上記の表に見られる個数は離散変数であるため，それらの個数の生起確率を近似的に求めている．近似であるがゆえに，後述する期待度数が極端に小さい場合には，検定がなされないソフトウェアも多い．これに対して，2×2 分割表では，**超幾何分布**（hypergeometric distribution）によって正確な確率を計算できることが知られている．

フィッシャーの正確検定

フィッシャーの正確検定（Fisher's exact test）で実行されている計算を簡潔に説明しよう．次の例は奥村晴彦氏の「R で楽しむ統計」（共立出版，2016）の 76 ページに記載されているものである．そこでは，ある事案に対する賛否が性別ごとに集計されており，男女間に意見の相違はあるかが問われている．世論調査などの結果を参照すると，賛否の件数を見る限り，男女間での意見の相違はありそうに見える．

表 2: 奥村（2016）の例

	賛成	反対
男性	3	2
女性	1	4

まず，10 件のサンプルうち 4 件が賛成，6 件が反対であることを確認しよう．（分割表における件数を縦計している．）男性の意見は 5 件ある．（分割表における件数を横計している．）そこで，10 人の中から 5 人をランダムに選んだとき，そのうち 3 人が賛成，2 人が反対である確率は

$$\frac{{}_4C_3 \cdot {}_6C_2}{{}_{10}C_5}$$

なので，R の choose() 関数を用いて計算すると，0.2380952 であることがわかる．R のコマンドは，EZR ではなく，R のコンソールで次のように入力する．

```
> choose(4,3)*choose(6,2)/choose(10,5)
```

賛成4件，反対6件を維持しつつ，10人中5人が男性である（縦計の件数と横計の件数が表2のものと同じ）分割表は他にも次のものがある．

	賛成	反対	賛成	反対	賛成	反対	賛成	反対
男性	4	1	2	3	1	4	0	5
女性	0	5	2	3	3	2	4	1

これらの分割表に対しても，それぞれ，${}_4C_0 \cdot {}_6C_1/{}_{10}C_5 = 0.02380952$，${}_4C_2 \cdot {}_6C_3/{}_{10}C_5 = 0.4761905$，${}_4C_1 \cdot {}_6C_4/{}_{10}C_5 = 0.2380952$，${}_4C_0 \cdot {}_6C_5/{}_{10}C_5 = 0.02380952$ となる．以上の準備の後，10人の中から5人をランダムに選んだときに,「3人が賛成，2人が反対である事象」となる生起確率にそれよりも生起確率が低い事象の生起確率を足し合わせると，0.5238になる．これがフィッシャーのフィッシャーの正確検定における帰無仮説のp値である．ここでの帰無仮説は男女間で意見の相違がないことであり，p値が0.05以下であれば，多くの場合，帰無仮説を棄却する．上述の奥村晴彦氏の例では，男女間での意見の相違はありそうに見えるが，その差は統計的には有意ではないと判断できる．

この手順を実行するRの関数がfisher.test()である．Rのコンソールのプロンプトから使用する場合のコマンドとその出力結果は次のとおりである．

```
> fisher.test(matrix(c(3,1,2,4), nrow=2))

Fisher's Exact Test for Count Data

data:  matrix(c(3, 1, 2, 4), nrow = 2)
p-value = 0.5238
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.218046 390.562917
sample estimates:
odds ratio
  4.918388
```

EZRでは，件数を直接ポップアップ画面から入力し，次のような手順を踏む．

1. EZR のリボンにある「統計量」ボタンをクリックし,「分割表」>「2 元表の入力と分析...」を選択する.
2. 「2 元表を入力」というポップアップが出てくるので,「表」のページで変数の名前（オプショナル）を付け，行数と列数をスライダーで選択し，同じポップアップ上にある「数を入力」と書かれた欄に件数を入力する．デフォルトでは 2×2 型になっている.
3. 「同じポップアップの「統計量」のページで「フィッシャーの正確検定」のチェックボックスにチェックを入れる．デフォルトでは「独立性のカイ 2 乗検定」になっている.（カイ 2 乗検定はデータに期待度数が 5 以下の件数があると実行されない．奥村 (2016) の例で試してみるとよい.）

カイ 2 乗検定

表 2 の奥村（2016）の例において，男女での意見の相違がまったくない（独立）であれば，縦計と横計により，男性の比率 0.5，賛成比率 0.4，反対比率 0.6 より，次のような件数が観察されることが期待（予想）される.

	賛成	反対
男性	$10 \times 0.5 \times 0.4 = 2$	$10 \times 0.5 \times 0.6 = 3$
女性	$10 \times 0.5 \times 0.4 = 2$	$10 \times 0.5 \times 0.6 = 3$

平均して E 回起こる事象が O 回起こるとすると，確率変数 O は平均 E，分散 E のポアソン分布に従うが，この E が十分に大きい時には O を Z 変換した $\chi = (O - E)/\sqrt{E}$ は（中心極限定理より）標準正規分布で近似できる.（標準正規分布の平均は 0，分散は 1 である.）標準正規分布に従う互いに独立な確率変数 χ_i が n 個（$i = 1, \ldots, n$）あれば,

$$\sum_{i=1}^{n} \chi_i^2 = \chi_1^2 + \chi_2^2 + \cdots + \chi_n^2$$

が従う確率分布を自由度 n の**カイ 2 乗分布**（chi-squared distribution）という．これを奥村（2016）の例に適用すると，縦計と横計で行の合計と

列の合計が固定されているので，自由に動かせる変数の個数は 3 つ減り，自由度 1 のカイ 2 乗分布に（近似的に）従う．このことを利用した分割表における独立性の検定を**ピアソンのカイ 2 乗検定**（Pearson's chi-squared test）という．

ピアソンのカイ 2 乗検定を実行する R の関数が chisq.test() である．R のコンソールのプロンプトから使用する場合のコマンドとその出力結果は次のとおりである．

```
> chisq.test(matrix(c(3,1,2,4), nrow=2), correct=FALSE)

Pearson's Chi-squared test

data:  matrix(c(3, 1, 2, 4), nrow = 2)
X-squared = 1.6667, df = 1, p-value = 0.1967

 警告メッセージ:
 chisq.test(matrix(c(3, 1, 2, 4), nrow = 2), correct = FALSE) で:
   カイ自乗近似は不正確かもしれません
```

このように,「カイ自乗近似は不正確かもしれません」との出力がなされるのは，上述の導出過程で言及されたように，平均して生起する件数 E_i が小さいと，$\chi_i = (O_i - E_i)/\sqrt{E_i}$ が従う確率分布を標準正規分布で近似することが困難になるからである．この不正確さは，ピアソンのカイ 2 乗検定のための p 値（0.1967）が，フィッシャーの正確検定におけるそれ（0.5238）と比較して著しく低いことにより，帰無仮説を誤って棄却してしまう確率を高めてしまうことを示唆している．

そこで，

$$\sum_{i=1}^{n}(|O_i - E_i| - 0.5)^2/E_i$$

と修正することが提案され，この方法は**イェーツの連続性補正**（Yates' continuity correction）と呼ばれている．この補正を施したカイ 2 乗検定は**イェーツのカイ 2 乗検定**とも呼ばれる．

イェーツのカイ 2 乗検定を実行するには，R の関数 chisq.test() を使う．R のコンソールのプロンプトから使用する場合のコマンドとその出力結果は次のとおりである．

```
> chisq.test(matrix(c(3,1,2,4), nrow=2))

Pearson's Chi-squared test with Yates' continuity correction

data:  matrix(c(3, 1, 2, 4), nrow = 2)
X-squared = 0.41667, df = 1, p-value = 0.5186

 警告メッセージ:
 chisq.test(matrix(c(3, 1, 2, 4), nrow = 2)) で:
   カイ自乗近似は不正確かもしれません
```

不正確性に関する警告メッセージは出されているが，帰無仮説が成立する確率である p 値（0.5186）はフィッシャーの正確検定のためのもの（0.5238）に近づいている．

練習問題：表 1 の分割表をピアソンのカイ 2 乗検定，イェーツのカイ 2 乗検定，フィッシャーの正確検定で解析せよ．

Appendix: より複雑な課題

分割表の解析は現在でも頻繁になされているが，実務者が扱うデータの解析は分割表に類似してはいるがかなり様相の異なる分析を必要とするものも多い．分析のために取得するべきデータとその構造はそれに適用される統計手法とのセットで選択されなければならず，データセットを作成した後では適用できない手法もある．たとえば，次の問題を考えてみよう．

ある小売店が，契約者数と彼らの取引金額を伸ばすために，電話やダイレクトメールなどによるいくつかの施策を行っている．これまでに得られたデータは次のとおりである．同一の契約者が複数の案件について契約する場合は，単純化のため，ここでは考えないことにする．

各施策への対象者の割当てはランダムに行われているわけではなく，それは小売店と潜在的契約者の選択によるものなので，何らかの対応が必要である．同一人物が異なる施策の対象者であることもありうる．その人物が異なる案件に関して別々に約定に至っていることもあるだろう．このような困難のために，ランダムな割り当てではない部分をどのように考慮するかはここでは問わないことにする．さらに，契約者が購入した

表 3: 施策の結果

施策	対象人数	契約者数	約定率	購入金額（円）mean (std. dev.)	原価率 (%) mean (std. dev.)
A	100	20	20%	5641 (354.88)	66.1 (0.043)
B	80	24	30%	4899 (366.46)	59.2 (0.066)
C	90	18	20%	4193 (351,22)	63.7 (0.052)

商品は契約ごとに異なり，原価率も異なる．そこで，施策ごとの「利益の分布」を推定してみることになるだろう．

この課題に回答するには**マルコフチェーンモンテカルロ法**（MCMC）を用いた**ベイズ推定**（Bayesian estimation）が適用される．MCMC を実行するには，R や RStan の利用が推奨される．簡単なものであれば Excel でも実行できないわけではないが，数万回にも及ぶ乱数を発生させる必要があり，かなり時間がかかる．