分割表の計算

2010 年 2 月 3 日版

目次

分割表における独立性の検定は、よく用いられるが各マスの度数が 5 以下の場合では、カイ 2 乗統計量をもちいることは近似式であり、ずれが大きいといわれる。ここでは近似をおこなわない計算、すなわち超幾何分布の確率計算とカイ 2 乗値との比較をふまえながら議論をしてみる。

これが注目されるようになった、階乗という困難な数式の処理が可能になってきたという理由である。

R.A.Fisher の直接計算とカイ 2 乗分布による近似計算を調べよう。

1 直接計算確率法

2×2 の分割表とは、つぎに与えられるような形:

度数	B_1	B_2	計 (周辺)
A_1	n_{11}	n_{11}	$n_{1.}$
A_2	n_{21}	n_{21}	$n_{2.}$
計 (周辺)	$n_{.1}$	$n_{.2}$	n

という行列形である。たとえば予防注射の効用を調べるとき、属性を A,B の二つでそれぞれがまた 2 つに分ける。A_1:予防注射を受けた, A_2:受けない、またその効果について、B_1:非罹患,B_2:罹患という形に分類する。4 個の列と行の合計（周辺度数）$n_{1.}$、$n_{2.}$、$n_{.1}$、$n_{.2}$ を与えて、条件：

$$n_{1.} = n_{11} + n_{12}, \qquad n_{2.} = n_{21} + n_{22}$$
$$n_{.1} = n_{11} + n_{21}, \qquad n_{.2} = n_{21} + n_{22}$$

を満たすような 4 つの数 $(n_{11}, n_{12}, n_{21}, n_{22})$ を求める。ただしセルの総合計を $n = n_{1.} + n_{2.} = n_{.1} + n_{.2}$ とおいた。もしこの属性が独立で、各々の項目に含む確率が属性 A については $p_1, p_2(= 1 - p_1)$ として、属性 B では $q_1, q_2(= 1 - q_1)$ とし、属性の間には、独立と仮定する。するとこれを満たすような場合の数は、超幾何分布に従う。これからの目的は、周辺度数 $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$ を与えたとき、各セルの数が $(n_{11}, n_{12}, n_{21}, n_{22})$ となる事象の条件付き確率（簡略化して表現して）$p(n_{ij} \,|\, n_{.i}, n_{j.})$, $(i, j = 1, 2)$ を求める。一般に r 行 s 列のデータ $i = 1, 2, \cdots, r, j = 1, 2, \cdots, s$ であっても同様の計算である。条件付き確率に対して $p(n_{ij} \,|\, n_{.i}, n_{j.}) = \dfrac{p(n_{ij})}{p(n_{.i}, n_{j.})}$ を求める。分子には同時確率であるから $p(n_{ij}) = \dfrac{n!}{\Pi_{i,j} n_{ij}!} \, (p_1^{n_{1\cdot}} p_2^{n_{2\cdot}}) \, (q_1^{n_{\cdot 1}} q_2^{n_{\cdot 2}})$ つまり

$$p(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}! n_{12}! n_{21}! n_{22}!} \, p_1^{n_{1\cdot}} p_2^{n_{2\cdot}} q_1^{n_{\cdot 1}} q_2^{n_{\cdot 2}}$$

もう一つの確率 $p(n_{i.}, n_{.j})$ は周辺密度として計算する。条件を満たすものにわたって和をとるから、$p(n_{i.}, n_{.j}) = \sum_{*} p(n_{ij})$, ここで $*$ は条件 $(i, j) : \sum_i n_{ij} = n_{.j}, \sum_j n_{ij} = n_{i.}$ となるすべての n_{ij} につい

ての和を表すとする。実際

$$\begin{aligned} p(n_{i.}, n_{j.}) &= \sum_{*} p(n_{ij}) = \sum_{*} \frac{n!}{\Pi_{i,j} n_{ij}!} \left(p_1^{n_{1\cdot}} p_2^{n_{2\cdot}}\right) \left(q_1^{n_{\cdot 1}} q_2^{n_{\cdot 2}}\right) \\ &= n! \sum_{*} \frac{1}{\Pi_{i,j} n_{ij}!} \left(p_1^{n_{1\cdot}} p_2^{n_{2\cdot}}\right) \left(q_1^{n_{\cdot 1}} q_2^{n_{\cdot 2}}\right) = \frac{(n!)^2}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \left(p_1^{n_{1\cdot}} p_2^{n_{2\cdot}}\right) \left(q_1^{n_{\cdot 1}} q_2^{n_{\cdot 2}}\right) \end{aligned}$$

となる。なぜならば、多項式の展開 $(x_1 + x_2)^{n_{\cdot 1}} (x_1 + x_2)^{n_{\cdot 2}} = (x_1 + x_2)^{n_{\cdot 1} + n_{\cdot 2}} = (x_1 + x_2)^n$ において係数 $x_1^{n_{1\cdot}} x_2^{n_{2\cdot}}$ を比較すると $\sum_{*} \frac{n_{\cdot 1}!}{n_{11}!\, n_{21}!} \frac{n_{\cdot 2}!}{n_{12}!\, n_{22}!} = \frac{n!}{n_{1\cdot}!\, n_{2\cdot}!}$ であるから, $\sum_{*} \frac{1}{n_{11}!\, n_{21}!\, n_{12}!\, n_{22}!} = \frac{n!}{n_{1\cdot}!\, n_{2\cdot}!\, n_{\cdot 1}!\, n_{\cdot 2}!}$ を得る。したがって

$$\begin{aligned} p(n_{ij} \,|\, n_{i.}, n_{\cdot j}) &= \frac{(\Pi_i n_{i.}!)(\Pi_j n_{\cdot j}!)}{n!\, \Pi_{i,j} n_{ij}!}, \qquad \forall i, j \\ p(n_{11}, n_{12}, n_{21}, n_{22} \,|\, n_{1.}, n_{2.}, n_{\cdot 1}, n_{\cdot 2}) &= \frac{1}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \left(\frac{n_{1.}!\, n_{2.}!\, n_{\cdot 1}!\, n_{\cdot 2}!}{n!}\right) \\ &= (n_{ij}\text{によらない定数}) \frac{1}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!} \end{aligned}$$

上で述べた条件を満たすすべての $(n_{11}, n_{12}, n_{21}, n_{22})$ を加え合わせれば、1 となるものが（n_{ij} によらない定数) である。

具体例の数値で考えてみると、(宇野利雄、統計演習：２１１ページ)

	非罹患	罹患	計
予防	n_{11}	n_{12}	11
非予防	n_{21}	n_{22}	8
計	12	7	19

となるような n_{ij} は、どのくらいの確率で起こるか？を計算していたのである。

フィシャーによる考え方は観測データよりも偏ったものが起こる確率がある水準、たとえば 0.05 以下であれば独立であるとした仮説を棄却しようとするものである。偏るということはセルの最小値が k であれば、$k-1, k-2, \cdots, 0$ を総計した場合の確率を求める。

たとえば、上記の設定で、つぎのデータ $n_{11} = 9, n_{12} = 2, n_{21} = 3, n_{22} = 5$ であれば、偏る場合は、

$$\begin{array}{|cc|} \hline 9 & 2 \\ 3 & 5 \\ \hline \end{array} \quad \begin{array}{|cc|} \hline 10 & 1 \\ 2 & 6 \\ \hline \end{array} \quad \begin{array}{|cc|} \hline 11 & 0 \\ 1 & 7 \\ \hline \end{array}$$

であるから、

$$P\text{ 値} = \frac{11!8!12!7!}{19!} \left(\frac{1}{9!2!3!5!} + \frac{1}{10!1!2!6!} + \frac{1}{11!0!1!7!}\right) \fallingdotseq 0.067$$

しかし実際の計算実行は見かけほど簡単ではない。階乗の計算はきわめて厄介なものである。計算機の進歩で比較簡単に行えるになってきている。つぎはこの数式処理システムを述べよう。

2 数式処理での計算法 (Mathematica)

例題１ 宇野、統計演習：２１１ページ
つぎの資料から、予防注射の効能を調べよ。

	非罹病者	罹病者	計
予防注射を受けた者	9	2	11
受けなかった者	3	5	8
計	12	7	19

(解法) まず手入力で電卓流の計算を行う。

数式処理による直接的数値計算

```
In[1]:=a1 = 1/(9! * 2! * 3! * 5!)
```
Out[1]=$\frac{1}{522547200}$
```
In[2]:=a2 = 1/(10! * 1! * 2! * 6!)
```
Out[2]=$\frac{1}{5225472000}$
```
In[3]:=a3 = 1/(11! * 0! * 1! * 7!)
```
Out[3]=$\frac{1}{201180672000}$
```
In[4]:=(a1 + a2 + a3) * (11! * 8! * 12! * 7!)/19!
```
Out[4]=$\frac{283}{4199}$
```
In[5]:=N[%]
```
Out[5]=0.067397

一方、*Mathematica* (マニュアル 178 頁) にしたがい、プログラムを入力して実行すると次の結果が得られた。

分割表におけるカイ 2 乗近似値計算プログラム

```
ClearAll["Global*"]
In[1]:=Needs["HypothesisTesting`"];
In[2]:=chiSquareContingencyTable[x_]:=
    Module[{l, m, t, chi, df}, {l, m} = Dimensions[x];
    t = N[Outer[Times, Total[Transpose[x]], Total[x]]/
    Total[Flatten[x]]];
    chi = ∑_{i=1}^{l} ∑_{j=1}^{m} (x[[i,j]]-t[[i,j]])^2/t[[i,j]];
    df = (l - 1) * (m - 1);
    {chi, df, ChiSquarePValue[chi, df]}]
In[3]:= chiSquareContingencyTable[{{9, 2}, {3, 5}}]
```
Out[4]={3.9095, 1, OneSidedPValue → 0.048014}

これを直接的計算結果と比較すると、多少のずれが当然出ている。(解法終り)

さらに *Mathematica* マニュアル 180 頁には、*data* として、分割表を行列入力して計算するための下記のプログラムがある。2 行 2 列のデータから、片側と両側の P 値 (PValue) を出力する。

```
twowayTableFishersTest[data_]:=Module[{e, f, g, h, m, m1, m2, one, two, s},
    {e, f} = Total[data]; {g, h} = Total[data, {2}];
    m = Table[{{x, g - x}, {y = (e - x), h - y}}, {x, 0, g}];
    fishersExactTest[{{a_, b_}, {c_, d_}}]:=Block[{g = a + c, h = b + d, e = a + b, f = c + d, n = e + f},
        (e! * f! * g! * h!)/(n! * a! * b! * c! * d!)//N];
    m1[{{a_, b_}, {c_, d_}}]:=Block[{}, {a * d - b * c, fishersExactTest[{{a, b}, {c, d}}]}];
    m2 = Map[m1, m]; s = m1[data][[1]];
    one = Rule["One SidedPValue", Total[If[s > 0, Select[m2, #[[1]] ≥ s&], Select[m2, #[[1]] ≤ s&]][[2]]];
    two = Rule["TwoSidedPvalue", Total[Select[Map[fishersExactTest, m],# ≤ fishersExactTest[data]&]]];
    {one, two}
]
```

例題 2 つぎはマニュアルにある例題を確認のために実行した。

```
twowayTableFishersTest[{{10, 2}, {9, 10}}]
```
{One SidedPValue → 0.0499879, TwoSidedPvalue → 0.0651757}

例題 3（宇野 211 頁）前の例題を再度行った結果である。

twowayTableFishersTest[{{9, 2}, {3, 5}}]

{One SidedPValue → 0.067397, TwoSidedPvalue → 0.0739462}

3 Mathematica の基本操作 (2 段組み)

..... **行列の入力** .....

a = {{10, 20}, {30, 40}}

{{10, 20}, {30, 40}}

MatrixForm[a]

$$\begin{pmatrix} 10 & 20 \\ 30 & 40 \end{pmatrix}$$

a//MatrixForm

$$\begin{pmatrix} 10 & 20 \\ 30 & 40 \end{pmatrix}$$

a[[1]]　% リスト（行列）の第 1 の成分 (列)

{10, 20}

a[[2]]　% リストの第 2 の成分 (列)

{30, 40}

..... **行列の演算**.....

a = {{a11, a12}, {a21, a22}}

{{a11, a12}, {a21, a22}}

Total[a]

{a11 + a21, a12 + a22} %　リストで表現していることに注意。Total はベクトルの和を求めている。

b = {{b11, b12}, {b21, b22}}

{{b11, b12}, {b21, b22}}

Dimensions[b]

{2, 2}

行列の和

a + b

{{a11 + b11, a12 + b12}, {a21 + b21, a22 + b22}}

MatrixForm[a + b]

$$\begin{pmatrix} \text{a11} + \text{b11} & \text{a12} + \text{b12} \\ \text{a21} + \text{b21} & \text{a22} + \text{b22} \end{pmatrix}$$

行列の積

MatrixForm[a.b]

$$\begin{pmatrix} \text{a11b11} + \text{a12b21} & \text{a11b12} + \text{a12b22} \\ \text{a21b11} + \text{a22b21} & \text{a21b12} + \text{a22b22} \end{pmatrix}$$

第 1 行を抽出する（リストであるから、列は別のやり方で）

a[[1]]

{a11, a12}

a[[2, 1]]　% 2 行 1 列の成分要素を取り出す

a21

a//MatrixForm

$$\begin{pmatrix} \text{a11} & \text{a12} \\ \text{a21} & \text{a22} \end{pmatrix}$$

Total[a, {2}]　% 1 行 2 列の新ベクトル作ること。

{a11 + a12, a21 + a22}

.. **階層（リストレベル）を同一レベルに平坦化** ..

a

{{a11, a12}, {a21, a22}}

Flatten[a]

{a11, a12, a21, a22}

..... **Table は行列の生成** .....

Table[r[i, j], {i, 2}, {j, 2}]

{{r[1, 1], r[1, 2]}, {r[2, 1], r[2, 2]}}

$\sum_{i=1}^{2} r[i, j]$

r[1, j] + r[2, j]

SparseArray[{i_, j_}/;i>=j → p[i, j], {3, 3}]//MatrixForm

$$\begin{pmatrix} p[1,1] & 0 & 0 \\ p[2,1] & p[2,2] & 0 \\ p[3,1] & p[3,2] & p[3,3] \end{pmatrix}$$

Table[{x, 10 − x}, {x, 1, 9}]

{{1, 9}, {2, 8}, {3, 7}, {4, 6}, {5, 5}, {6, 4}, {7, 3}, {8, 2}, {9, 1}}

..... **Module の機能** .....

numTable[x_]:=Module[{f1, f2, f3},
{f1, f2, f3} = {x, x^2, Sqrt[x]}]

numTable[5]

$\{5, 25, \sqrt{5}\}$

Table[numTable[x], {x, 1, 5}]

$\{\{1, 1, 1\}, \{2, 4, \sqrt{2}\}, \{3, 9, \sqrt{3}\}, \{4, 16, 2\}, \{5, 25, \sqrt{5}\}\}$

Total[Table[numTable[x], {x, 1, 5}]]

$\{15, 55, 3 + \sqrt{2} + \sqrt{3} + \sqrt{5}\}$

命令 Outer は外積

Outer[Times, {1, 2, 3}, {c, d}]

$\{\{c, d\}, \{2c, 2d\}, \{3c, 3d\}\}$

Outer[Times, {{1, 2}, {3, 4}}, {{e, f}, {g, h}}]

$\{\{$
$\{\{e, f\}, \{g, h\}\},$
$\{\{2e, 2f\}, \{2g, 2h\}\}\},$
$\{\{\{3e, 3f\}, \{3g, 3h\}\},$
$\{\{4e, 4f\}, \{4g, 4h\}\}$
$\}\}$

数式から数値で計算

N[(1 + Sqrt[5])/2]

1.61803

新たに行列を入力

b = {{b1, b2}, {b3, b4}}

$\{\{\text{b1}, \text{b2}\}, \{\text{b3}, \text{b4}\}\}$

b//MatrixForm

$$\begin{pmatrix} \text{b1} & \text{b2} \\ \text{b3} & \text{b4} \end{pmatrix}$$

行列の積（ベクトルであれば内積）

a.b//MatrixForm

$$\begin{pmatrix} 10\text{b1} + 20\text{b3} & 10\text{b2} + 20\text{b4} \\ 30\text{b1} + 40\text{b3} & 30\text{b2} + 40\text{b4} \end{pmatrix}$$

a * b//MatrixForm

$$\begin{pmatrix} 10\text{b1} & 20\text{b2} \\ 30\text{b3} & 40\text{b4} \end{pmatrix}$$

即時実行命令

a21 = RandomInteger[{1, 6}, 3]

$\{3, 4, 6\}$

Table[a21, {3}]

$\{\{3, 4, 6\}, \{3, 4, 6\}, \{3, 4, 6\}\}$

遅延実行命令

a22:=RandomInteger[{1, 6}, 3]

Table[a22, {3}]

$\{\{4, 3, 6\}, \{4, 1, 3\}, \{2, 6, 5\}\}$

サイコロ投げの例

ClearAll

m = RandomInteger[{1, 6}, {5, 3}]

$\{\{3, 4, 6\}, \{4, 3, 6\}, \{2, 6, 1\}, \{5, 1, 1\}, \{4, 1, 2\}\}$

和を計算する

m/.{p_, q_, r_} → $p + q + r$

$\{13, 13, 9, 7, 7\}$

結果の集計する命令，｛値, 度数｝の形式

freq = Tally[m/.{p_, q_, r_} → $p + q + r$]

$\{\{13, 2\}, \{9, 1\}, \{7, 2\}\}$

２番目の要素にのみを取り出す

freq[[All, 2]]

$\{2, 1, 2\}$

成分ごとの和を計算する

Total[freq]

$\{29, 5\}$

行列データの入力

md2 = {{a1, a2}, {b1, b2}}

$\{\{\text{a1}, \text{a2}\}, \{\text{b1}, \text{b2}\}\}$

md2//MatrixForm

$$\begin{pmatrix} \text{a1} & \text{a2} \\ \text{b1} & \text{b2} \end{pmatrix}$$

列の和

Total[md2]

$\{\text{a1} + \text{b1}, \text{a2} + \text{b2}\}$

行の和

Total[md2, {2}]

{a1 + a2, b1 + b2}

..... ３次元のデータの場合 .....

mdata = {{a1, a2, a3}, {b1, b2, b3}, {c1, c2, c3}}

{{a1, a2, a3}, {b1, b2, b3}, {c1, c2, c3}}

mdata//MatrixForm

$$\begin{pmatrix} a1 & a2 & a3 \\ b1 & b2 & b3 \\ c1 & c2 & c3 \end{pmatrix}$$

..... 列の和 .....

Total[mdata]

{a1 + b1 + c1, a2 + b2 + c2, a3 + b3 + c3}

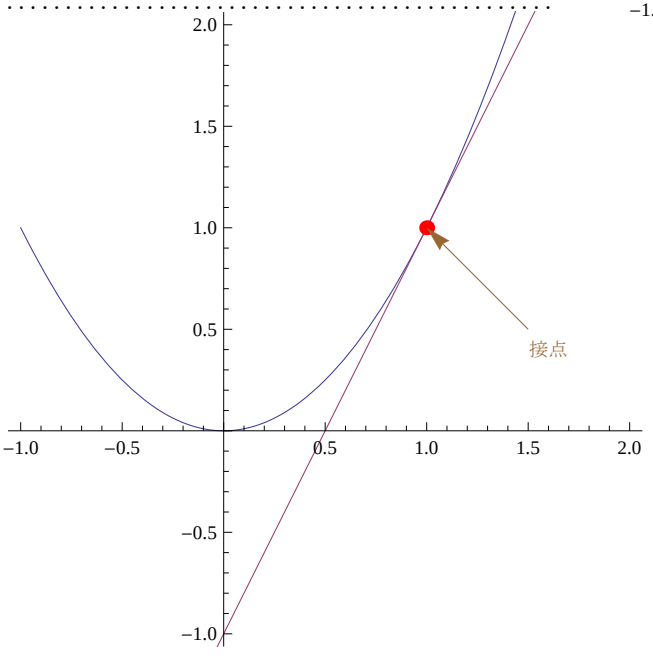
..... 行の和 .....

Total[mdata, {2}]

{a1 + a2 + a3, b1 + b2 + b3, c1 + c2 + c3}

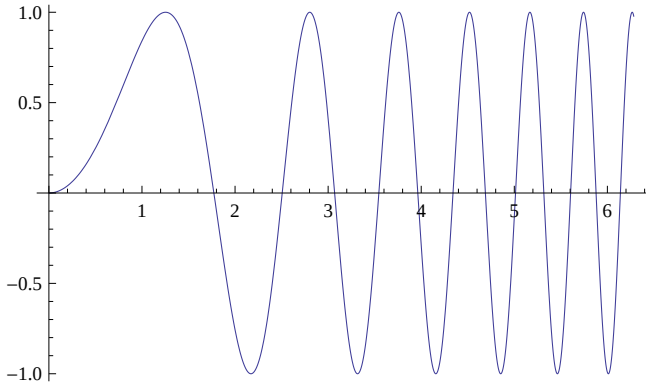
..... グラフに文字を入れるプログラム .....

```
Show[Plot[{x^2, 2x − 1}, {x, −1, 2}],
Graphics[{Red, PointSize[Large],
Point[{1, 1}], Brown, Arrow[{{1.5, 0.5}, {1, 1}}],
Text[" 接点 ", {1.6, 0.4}]}],
 AspectRatio → Automatic, PlotRange → {−1, 2}]
```



..... 基本数学アシスタントでパイを入力 .....

Plot[Sin[x^2], {x, 0, 2π}]



..... ２本のグラフを表示する .....

```
Plot[{Sin[x^2], Cos[x^2]}, {x, 0, 2π},
PlotStyle → {{Blue, Thick}, Dashed}]
```