学術フロンティア講義 数理工学のすすめ

「計算代数統計」（導入編の答え）

2018 年 5 月 10 日 (木)

清 智也 sei@mist.i.u-tokyo.ac.jp

先週配布した資料の「導入編」では，ある講義における S1 タームと S2 タームの教員の選択結果に関連性が見られるか，という問題を扱った。しかし，その答えをまだ与えていなかった。一言で言えば，答えは

関連性があるとは言えない

となる。実際, 統計ソフトウェア R で以下のように入力すると p 値 (後述) が 0.3245 となり，十分小さいとは言えない。

```
> x = matrix(c(
  3, 0, 0, 0, 1, 1,
  0, 0, 2, 0, 1, 2,
  0, 1, 1, 2, 0, 0,
  0, 2, 0, 1, 1, 1,
  1, 2, 0, 1, 0, 1,
  1, 0, 1, 1, 2, 0), 6, 6, byrow=TRUE)

> fisher.test(x)

Fisher's Exact Test for Count Data

data:  x
p-value = 0.3245
alternative hypothesis: two.sided
```

以下，この出力結果の意味を説明する。一般的な仮説検定や p 値については統計学の入門書[1]を参照されたい。

話を簡単にするため，まず表 1 のような 2×2 の分割表で考える。各セルの数字は度数を表す。たとえば S1 タームに教員 A，S2 タームに教員 a を選んだ学生が 9 人であったことを表す。

表 1: 分割表の例

	a	b
A	9	5
B	5	10

[1] たとえば竹村彰通，「統計」，共立出版など。

ここで次のような**モデル**を考える。すなわち，各学生は独立に，教員の組 (i, j) を確率 p_{ij} で選ぶというモデルである。これも表の形で

p_{11}	p_{12}
p_{21}	p_{22}

と書ける。ここで p_{ij} は $p_{ij} \geq 0$ かつ $p_{11} + p_{12} + p_{21} + p_{22} = 1$ を満たす実数である。すると，行と列に関連性がないことを表すモデル（**独立モデル**と呼ばれる）は

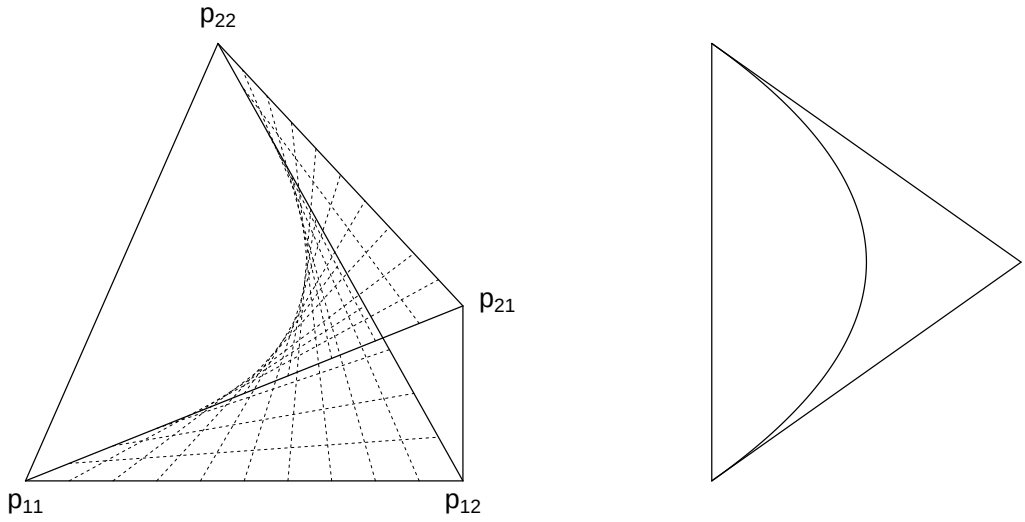
$$p_{ij} = \alpha_i \beta_j$$

と表される。ここで $\alpha_i \geq 0$, $\alpha_1 + \alpha_2 = 1$, $\beta_j \geq 0$, $\beta_1 + \beta_2 = 1$ とする。

集合

$$\{(p_{11}, p_{12}, p_{21}, p_{22}) \mid p_{ij} \geq 0,\ p_{11} + p_{12} + p_{21} + p_{22} = 1\}$$

は確率単体と呼ばれる。確率単体は 4 次元空間の集合であるが，たとえば p_{12}, p_{21}, p_{22} の動く範囲を考えれば 3 次元空間内の四面体と見なせる。確率単体の中で独立モデルは図 1 のような線織面となる。



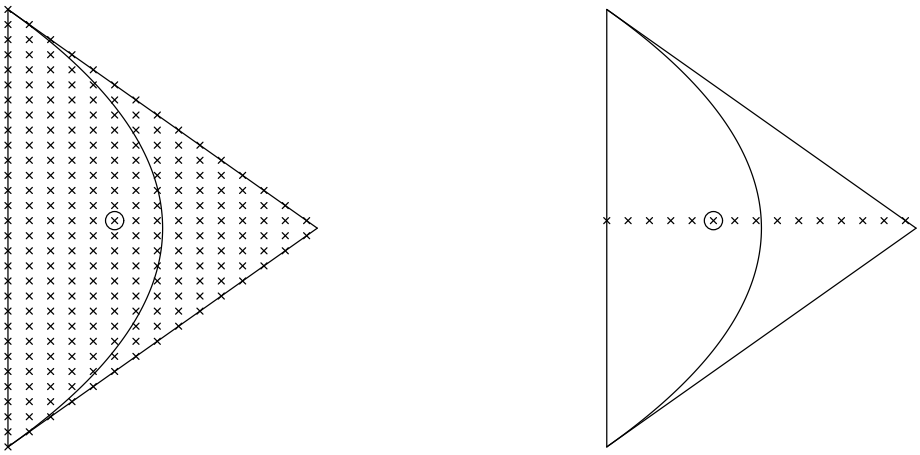
(a) 立体図　　(b) $p_{12} = p_{21}$ における断面

図 1: 確率単体と独立モデル。

さて，表 1 のデータから相対度数を求めれば

9/29	5/29
5/29	10/29

となり，これも確率単体の中の 1 点である。この点は図 2 (a) のように独立モデルから少し離れた位置にある。というより，総度数 29 で独立モデルに属すような観測値はほとんど存在しない（例外はどの点か？）。



(a) 総度数が 29 となる観測値（断面図）　　(b) 行和と列和を固定したファイバー

図 2: 可能な観測値の集合。表 1 に対応する観測値は丸で囲った。

したがって，関連性の有無を独立モデルに属すかどうかで判断しようとすれば，どんな観測値も独立でない（関連性がある）という判断になってしまう。しかしこれは実用的でない。そこで仮説検定の概念が必要となる。

仮説検定では，検定統計量 $T(\boldsymbol{x})$ という関数を用意し，事象 $\{\boldsymbol{x} \mid T(\boldsymbol{x}) \geq T(\boldsymbol{x}_0)\}$ の確率を計算する。ここで $\boldsymbol{x}_0$ は実際の観測値である。また確率を計算するときは帰無仮説（いまの場合は独立モデル）に基づいて行う。こうして得られる確率が **p 値**である。検定統計量の選び方はいろいろあるが，先のプログラムの中で採用されているのは $T(\boldsymbol{x}) = -2\log p(\boldsymbol{x})$ である[2]。ここで $p(\boldsymbol{x})$ は観測値 $\boldsymbol{x}$ が得られる確率である。

独立モデルに基づいて確率を計算すると述べたが，独立モデルに属す確率分布は無数にある。そこで前回説明した超幾何分布を考えることになる。復習すると，まずモデル p_{ij} のもとで観測値 x_{ij} が得られる同時確率は多項分布

$$\frac{N!}{x_{11}!x_{12}!x_{21}!x_{22}!}p_{11}^{x_{11}}p_{12}^{x_{12}}p_{21}^{x_{21}}p_{22}^{x_{22}}$$

となり（$N = \sum_i \sum_j x_{ij}$），独立モデルの場合は

$$\frac{N!}{x_{11}!x_{12}!x_{21}!x_{22}!}\alpha_1^{r_1}\alpha_2^{r_2}\beta_1^{c_1}\beta_2^{c_2}$$

となる（$r_i = \sum_j x_{ij}$, $c_j = \sum_i x_{ij}$）。そして，行和と列和を固定したもとでの条件付き確率は超幾何分布

$$p(\boldsymbol{x}) = \frac{r_1!r_2!c_1!c_2!}{N!x_{11}!x_{12}!x_{21}!x_{22}!}$$

[2]前回配布した「理論編」例 3 の尤度比検定統計量はこの検定統計量とは異なる。ただしスターリングの公式 $\log N! \simeq N\log N$ を用いれば両者は漸近的に同じ式になることが示される。

となり，α_i, β_j の値によらない。

図 2 (b) には行和と列和を固定したときの可能な観測値を示した。これは独立モデルに対する**ファイバー**と呼ばれる。ファイバーの各要素が得られる確率は左から順番に

x_{12}	0	1	2	3	4	5	6	7	8	9
$p(\boldsymbol{x})$	0.000	0.000	0.000	0.002	0.018	<u>0.078</u>	0.194	0.285	0.249	0.129

x_{12}	10	11	12	13	14
$p(\boldsymbol{x})$	0.039	0.006	0.001	0.000	0.000

となる。ただし観測値に対応する確率に下線を引いた。そして p 値は

$$(p\text{ 値}) = 0.000 \times 5 + 0.001 + 0.002 + 0.006 + 0.018 + 0.039 + 0.078 = 0.144$$

と計算される。実際に R で実行すると以下のようになる。

```
> x = matrix(c(9, 5, 5, 10), 2, 2, byrow=TRUE)

> fisher.test(x)

Fisher's Exact Test for Count Data

data:  x
p-value = 0.1431
alternative hypothesis: true odds ratio is not equal to 1
```

以上の議論は 6×6 分割表であっても全く同様であり，これで冒頭の R プログラムの出力結果の意味が説明された。ただし内部の計算はネットワークアルゴリズムと呼ばれる手法に基づいている。興味がある者は文献

> C. R. Mehta and N. R. Patel (1982). A network algorithm for performing Fisher's exact test in r × c contingency tables, *J. Amer. Statist. Assoc.*, **78**, 427–434.
> (学内からダウンロード可)

を参照されたい。

おまけ問題の答え　与えられた表と同じ行和・列和を持つ分割表を数えるのは簡単な問題ではないらしい。LattE[3]というソフトウェアを使って計算すると，全部で

$$16,415,075,151$$

通りあるようである。Mac Pro (Mid 2010, 2.66GHz 6-core) で 7984 秒かかった。

[3] `https://www.math.ucdavis.edu/~latte/`