カイ 2 乗分布について

SHIMURA Masato
jcd02773@nifty.ne.jp

2016 年 9 月 8 日

目次

1 カイ 2 乗値による検定

鈴木義一郎「J 言語による統計分析」からカイ 2 乗 (χ^2, C^2) に関連する部分を取りだし、スクリプトを復元する

一般に n 人中 x 人がある属性を持ち、その属性が P で与えられているとするとすると、理論と実際のズレは次式で与えられる。

$$C^2 = \frac{(x-np)^2}{np} + \frac{(n-x-n(1-p))^2}{n(1-p)} = \left(\frac{x-np}{\sqrt{np(1-p)}}\right)^2$$

x を 2 項分布に従う変量と考えると次式は平均 0 標準偏差 1 の標準正規分布に従う

$$Z = \frac{x-np}{\sqrt{np(1-p)}}$$

従って次のような関係が成立している

$$P(C^2 \geq 4) = P(|Z| \geq 2) = 0.05$$

理論と実際のズレの値 C^2 が 4 を越えるようなら、実際の値は理論からずれていることとなる。

例題 あるサイコロを 60 回投げた結果が次のようであった。これは正しいサイコロと見做してよいだろうか

サイコロの目	1	2	3	4	5	6
度数	12	8	10	13	8	9

1. 期待値は 10
2. 理論と実際のズレの平方値を期待度数で割り、合計する

$$\sum \frac{2^2}{10} + \frac{\_2^2}{10} + \frac{0^2}{10} + \frac{3^2}{10} + \frac{\_2^2}{10} + \frac{\_1^2}{10} = 2.2$$

   $\_2^2$ は J 流の表記法で $\_2$ が一つの数であり、$(-2)^2$ のような括弧は不要
3. Script

```
mean=: +/ % #
chigf=: 3 : '(+/ *: y - e0 ) % e0=. mean y'
NB. Usage:   chigf 12 8 10 13 8 9
```

   - +/ $\sum$
   - % $\div$
   - # n
   - *: a^2

```
   chigf 12 8 10 13 8 9
2.2
```

4. 偏ったサイコロでは数値が大きくなるが例題はそれ程でもない。

1.1 傾向性の検定

例題 年齢別成人女子の不眠症数

年齢階級	18 – 24	25 – 34	35 – 44	45 – 54	55 – 64	65 – 74	70–
調査人数	534	746	784	705	443	299	70
不眠症	150	250	264	302	238	176	36

1. DATA

```
     X0,:Y0
534 746 784 705 443 299 70
150 250 264 302 238 176 36
```

2. 合計と比率を求める

```
    A, ({:% {.) A=.  DATA,.+/"1 DATA=. X0,:Y0
  534   746   784   705   443   299   70   3581
  150   250   264   302   238   176   36   1416
0.281 0.335 0.337 0.428 0.537 0.589 0.514 0.395
```

3. 数式とアルゴリズム

- 傾向性が全くないとする仮説

$$H_0 : \qquad p_{01} = p_{02} = \cdots = p_{0k}$$

- 増加傾向があるといったような傾向性があるという仮説

$$H_1 : \qquad p_{01} \leq p_{02} \leq \cdots \leq p_{0k}$$

- 観測された比率

$$P_i = \frac{x_i}{n_i} \qquad , (i = 1, 2, 3, \cdots k)$$

- カイ 2 乗値

$$C_0^2 = \sum n_i \frac{(P_i - p_{0i})^2}{p_{0i}(1 - p_i)} = \sum \left(\frac{x_i - n_i p_{0i}}{\sqrt{n_i p_{0i}(1 - p_{0i})}} \right)^2$$

これが近似的に自由度 k のカイ 2 乗分布に従う

- この値は理論比率 p_{0i} がわからないと計算できない。
仮説 H_0 の下では p_{0i}g が共通の値であることから次の値で代用する

$$\bar{p} = \frac{p_1 + p_2 + p_3 + \cdots + p_k}{k}$$

- そこで次式も自由度 $k - i$ のカイ 2 乗分布で近似できる

$$C^2 = \sum n_i \frac{(P_i - \bar{P})^2}{\bar{P}(1 - \bar{P})}$$

4. Script

```
chitr=: 4 : ' (+/ y * *: p-q) % q * -. q=. mean p=. x %y'
```

5. 計算

```
   Y0 chitr X0
158.702
```

6. カイ 2 乗値はかなり大きく傾向性がないとする H_0 仮説は棄却される

1.2 適合度検定

理論比率が等確率ではなく、適当に特定された理論比率からのズレを問題としなければならない場合。

一般に k 通りある理論比率 $(p_1, p_2.p_3 \cdots, p_k)$ に対する n 回の実験結果 $(x_1, x_2, x_3 \cdots x_k)$ のように与えられた場合の理論と実際のズレを求める式

$$C^2 = \sum_{i=1}^{k} n_i \frac{(x_i - np_i)^2}{np_i}$$

もう一つのカイ 2 乗値であり、自由度 k のカイ 2 乗分布に従う。

例題　ある病院で生まれた新生児 900 人のうち男児は 480 人、女児は 420 人であった。

1. Script

```
testgf=: 4 : ' +/ (*: y -t)%t=. x * +/ y'
P0=: (1.06 % 2.06),1%2.06
```

2. 計算

P0 は男女出生性比率 1.06:1

```
   P0
0.514563 0.485437


  P0 testgf 480 420
1.26943
```

3. 値は小さい

1.3　分割票の独立性の検定

	分類	基準 (I)	周辺度数
分類基準 (II)	$\frac{a \cdot b}{n}$	$\frac{a \cdot (n-b)}{n}$	a
	$\frac{(n-a) \cdot b}{n}$	$\frac{(n-a)(n-b)}{n}$	$n-a$
周辺度数	b	$n-b$	n

データが次のようであれば

e	f
g	h

周辺度数は次のような定まる

$$\begin{aligned} a &= e + f \\ b &= e + g \\ n &= e + f + g + h \end{aligned}$$

データから期待度数は次のように計算できる

$$e_0 = \frac{(e+f)(e+g)}{e+f+g+h} \quad f_0 = \frac{(e+f)(f+h)}{e+f+g+h}$$

$$g_0 = \frac{(g+h)(e+g)}{e+f+g+h} \quad h_0 = \frac{(g+h)(f+h)}{e+f+g+h}$$

カイ 2 乗値は次により求まる。

$$C^2 = \frac{(e-e_0)^2}{e_0} + \frac{(f-f_0)^2}{f_0} + \frac{(g-g_0)^2}{g_0} + \frac{(h-h_0)^2}{h_0} = \frac{(e+f+g+h)(eh-fg)^2}{(e+f)(e+g)(f+h)(g+h)}$$

例題　新薬と旧薬の薬比効果

	治癒	無効果	計
新薬	304	104	408
旧薬	353	166	519
計	657	270	927

新薬の期待度数は $\frac{408}{927} \times 657 = 289$ となる。

1. Script

```
chic=: 3 : '(+/ ,y)*(*: det y) % */(+/"1 y),+/ y '
det=: -/ . *  NB. 行列式 eh-fg
```

2. 計算

```
   D0
304 104
353 166
   chic D0
4.66718
```

3. 結果は 4 より若干大きいので 2 の要因間には関連性が認められ、新薬の効果はある (若干)
4. 汎用の分割表の独立性の検定 Script

```
NB. general chic
testc=: 3 : 0
p=.,(+/ "1 y) */ +/ y
+/ (*: p-(*+/),y)%p*+/ ,y
)
```

5. 計算

```
   testc D0
4.66718
```

2　カイ２乗分布

2.1　確率密度関数（CDF)

$$\begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{\frac{x}{2}} & (x\ 0) \\ 0 & (x \leq 0) \end{cases}$$

jsoftware.com の Essay に載っているカイ２乗分布の CDF

```
NB. *********C square CDF*************
NB. from Jsoftware.com Essay
gamma   =: ! & <:
ig0     =: 4 : '(1 H. (1+x) % x&((* ^) * (^ -)~)) y'
incgam  =: ig0 % gamma@[  NB. incomplete gamma
chisqcdf=: incgam&-:
```

横山氏によるスクリプト (部分)

```
Gamma=: !@<:
chisq=:(([]^(-:@<:@<:@[))*(^@(-@-:@ ])))%((2: ^ -:@[)*(Gamma@-:@[))
```

2.2　累積密度関数

$$F(x;k) = \frac{\gamma(\frac{k}{2}, \frac{x}{2})}{\Gamma(\frac{k}{2})}$$

3　References

鈴木義一郎「J 言語による統計分析」森北出版　1996