Arcsine 変換に基づく 傾向性の Cochran-Armitage 型検定

2011SE009 荒木倖平

2011SE192 中津琢也

指導教員:白石高章

1 はじめに

統計学の基礎となる確率や事象を学んできた. それをもとに, 統計学の分布論や 2 項分布に関係した 1,2 標本モデルの統計解析法, 分散安定化変換などを学んだ.

Cochran-Armitage 検定は主に医学分野で使用されている (丹後, 小西 [1]). 統計学が実際にどのように用いられているかが明確になるにつれて Cochran-Armitage 検定に興味を持ち, この検定手法について研究することにした. そこで本論文では, 比率モデルにおける Arcsine 変換に基づく Cochran-Armitage 型検定について考察する.

2 Cochran-Armitage 型検定 I

ある定量的あるいは順序尺度変数 X の a 個の値を

$$x_1, \cdots, x_a$$

とする.

x_iにおける n_i個の観測値 $Y_{ij}(j=1,\cdots,n_i)$ は独立で, いずれも期待値 $p(x_i)$ をとるとすると

$$E(Y_{ij}) = p(x_i)$$

となり,$p(x)$ は $X = x_i$で $p(x_i)$ を広義の単調関数とする.

このとき,

$$p(x_1) \le \cdots \le p(x_a) \text{ または } p(x_1) \ge \cdots \ge p(x_a)$$

となる.

この関係の存在を示すために x_iに評点 $s(x_i)$ を与え, この評点上への Y の単回帰直線の傾きが正または負の値をとることを示せばよい.

ここに,

$$s(x_1) < \cdots < s(x_a)$$

とする. 次に仮説検定を考える. 帰無仮説を

$$H_0 : p_1 = \cdots = p_a$$

対立仮説を

$$H_1 : p_1 \le \cdots \le p_a \text{ または } p_1 \ge \cdots \ge p_a$$

であり, 少なくとも 1 つの不等号は等号が成り立たないものとする.ただし $p_i = p(x_i)$ とする.

(命題 1) 確率変数 Y の評点 s 上への単回帰係数の推定値を$\hat{b}$とすると

$$\hat{b} = \frac{\displaystyle\sum_{i=1}^{a} n_i(s_i - \bar{s})(\hat{p}_i - \bar{\hat{p}})}{\displaystyle\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}$$

となる.ただし $s_i = s(x_i)$ とする.

(証明) 白石 [2] より,

$$y_1, \cdots, y_{n1}, y_{n_1+1}, \cdots, y_{n_1+n_2}, \cdots, y_n$$

$$x_1, \cdots, x_{n1}, x_{n_1+1}, \cdots, x_{n_1+n_2}, \cdots, x_n$$

$$b' = \frac{\displaystyle\sum_{i=1}^{a}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{a}(x_i - \bar{x})^2}$$

がいえる.ただし,

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

とする.また,

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{a} n_i s_i$$

$$\bar{\hat{p}} = \frac{1}{n}\sum_{i=1}^{a} \hat{p}_i n_i = \frac{1}{n}\sum_{i=1}^{a} Y_i$$

である. b'の式で s_iを x_i, $\hat{p}_i$を y_iに対応させると

$$\begin{aligned}\sum_{i=1}^{a}(x_i - \bar{x})^2 &= (s_1 - \bar{s})^2 + (s_1 - \bar{s})^2 + \cdots + (s_1 - \bar{s})^2 \\ &\quad + (s_2 - \bar{s})^2 + (s_2 - \bar{s})^2 + \cdots + (s_2 - \bar{s})^2 \\ &\quad + \cdots + (s_a - \bar{s})^2 + (s_a - \bar{s})^2 + \cdots + (s_a - \bar{s})^2 \\ &= \textstyle\sum_{i=1}^{a} n_i(s_i - \bar{s})^2\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{a}(x_i - \bar{x})(y_i - \bar{y}) &= (s_1 - \bar{s})(Y_{11} - \bar{\hat{p}}) \\ &\quad + \cdots + (s_1 - \bar{s})(Y_{1n_1} - \bar{\hat{p}}) \\ &\quad + \cdots + (s_a - \bar{s})(Y_{a1} - \bar{\hat{p}}) + \cdots + (s_a - \bar{s})(Y_{an_a} - \bar{\hat{p}}) \\ &= \sum_{i=1}^{a}(s_i - \bar{s})(Y_i - n_i\bar{\hat{p}}) \\ &= \sum_{i=1}^{a} n_i(s_i - \bar{s})(\hat{p}_i - \bar{\hat{p}})\end{aligned}$$

$\left(\hat{p}_i = \frac{Y_i}{n_i} \Leftrightarrow Y_i = \hat{p}_i n_i \text{より}\right)$
以上より,

$$\hat{b} = \frac{\sum_{i=1}^{a} n_i(s_i - \bar{s})(\hat{p}_i - \bar{\hat{p}})}{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}$$

が成り立つ. □
また,

$$(\hat{b}\text{の分子}) = \sum_{i=1}^{a} n_i(s_i - \bar{s})(\hat{p}_i - \bar{\hat{p}}) = \sum_{i=1}^{a} n_i(s_i - \bar{s})\hat{p}_i$$

となる.ここで,

$$\hat{b} = \frac{\sum_{i=1}^{a} n_i(s_i - \bar{s})\hat{p}_i}{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}$$

となる. また, $\hat{b}$の分散は,

$$V(\hat{b}) = V\left(\frac{\sum_{i=1}^{a} n_i(s_i - \bar{s})\hat{p}_i}{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}\right)$$

$$V(\hat{p}_i) = V\left(\frac{Y_i}{n_i}\right) = \frac{n_i p_i(1-p_i)}{{n_i}^2}$$

の関係を使うと,

$$(\text{与式}) = \sum_{i=1}^{a} n_i(s_i - \bar{s})^2 p_i(1-p_i)$$

とできる.
よって,

$$V(\hat{b}) = \frac{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2 p_i(1-p_i)}{\left\{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2\right\}^2}$$

である.次に,
帰無仮説 $H_0 : p_1 = \cdots = p_a$**vs.**対立仮説 $H_1 : p_1 \le \cdots \le p_a$または $p_1 \ge \cdots \ge p_a$に対する検定を考える.このとき少なくとも 1 つの不等式は等号が成り立たないものとする.

H_0の下で, $p_1 = \cdots = p_a = p_0$であり, $\beta = 0$ かつ

$$V_0(\hat{b}) = \frac{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2 p_i(1-p_i)}{\{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2\}^2} = \frac{p_0(1-p_0)}{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}$$

となる.
また, T_bを

$$T_b = \frac{\hat{b} - E_0(\hat{b})}{\sqrt{V_0(\hat{b})}} = \frac{\hat{b}}{\sqrt{\frac{p_0(1-p_0)}{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}}}$$

とおく.また次の (条件 1) を仮定する.
(条件 1) $\quad 0 < \lim_{i\to\infty} \frac{n_i}{n} = \lambda_i < 1$
このとき, H_0のもとで, $n \to \infty$ として

$$T_b = \frac{\sum_{i=1}^{a} n_i(s_i - \bar{s})(\hat{p}_i - \bar{\hat{p}})}{\sqrt{p_0(1-p_0)\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}}$$

$$= \frac{\sum_{i=1}^{a} \frac{n_i}{n}(s_i - \bar{s})\sqrt{n}(\hat{p}_i - \bar{\hat{p}})}{\sqrt{p_0(1-p_0)\sum_{i=1}^{a} \frac{n_i}{n}(s_i - \bar{s})^2}}$$

$$\xrightarrow{L} \frac{\sum_{i=1}^{a} \lambda_i(s_i - \bar{s})\sqrt{\frac{p_i(1-p_i)}{\lambda_i}}Z_i}{\sqrt{p_0(1-p_0)\sum_{i=1}^{a} \lambda_i(s_i - \bar{s})^2}}$$

$$\sim N(0.1)$$

となる.ただし $\xrightarrow{L}$ は分布収束を表す.
H_0のもとで $p_0 = \overline{p}$より,

$$\hat{V}_0 = \frac{\bar{\hat{p}}(1-\bar{\hat{p}})}{\sum_{i=1}^{a} n_i(s_i - \bar{s})^2}$$

とおくと,

$$\frac{\hat{b}}{\sqrt{\hat{V}_0}} \xrightarrow{L} N(0,1)$$

$$\frac{\hat{b}^2}{\hat{V}_0} \xrightarrow{L} \chi_1^2$$

水準αの検定は,

$$\frac{\hat{b}^2}{\hat{V}_0} > \chi_1^2(\alpha) \Rightarrow H_0\text{を棄却}$$

ただし$\chi_1^2(\alpha)$ は自由度 1 のχ^2分布の上側 100α%点である.

3 Cochran-Armitage 型検定 II

次に $\arcsin\sqrt{\hat{p}_i}$を用いた検定方式について考える.

(補題 2) (条件 1) が満たされているとすると, $n \to \infty$ として

$$\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i}) \xrightarrow{L} N(0, \frac{1}{4\lambda_i})$$

が成り立つ.
(証明) 白石 [2] の命題 7.8 より

$$T_n = \hat{p}_i, \theta = p_i$$

とすると,
$$\sqrt{n}(\hat{p}_i - p_i) \xrightarrow{L} N(0, h(p_i))$$
とおける.
$$g(x) = \arcsin\sqrt{x}$$
とすると,
$$g(\hat{p}_i) = \arcsin\sqrt{\hat{p}_i},\ g(p_i) = \arcsin\sqrt{p_i}$$
となる.
$$\{g'(p_i)\}^2 = \{(\arcsin\sqrt{p_i})'\}^2 = \frac{1}{4p_i(1-p_i)}$$
であるので,
$$\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})$$
$$\xrightarrow{L} N(0, \{g'(p_i)\}^2 * h(p_i)) = N(0, \frac{1}{4\lambda_i})$$
となる. □
$$b_s \equiv \frac{\sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})}{\sqrt{\sum_{i=1}^{a} n_i^2(s_i-\bar{s})^2\frac{n}{4n_i}}}$$
とする.
$$(b_s\text{の分子}) = \sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})$$
このとき, (条件 1) を用いて, 分子を n で割ると
$$(\text{与式}) = \sum_{i=1}^{a} \frac{n_i}{n}(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})$$
$$\xrightarrow{L} \sum_{i=1}^{a} \lambda_i(s_i-\bar{s})\frac{Z_i}{2\sqrt{\lambda_i}}$$
また, このときの分散は,
$$V\left(\sum_{i=1}^{a} \lambda_i(s_i-\bar{s})\tfrac{Z_i}{2\sqrt{\lambda_i}}\right) = \sum_{i=1}^{a} V\left(\lambda_i(s_i-\bar{s})\tfrac{Z_i}{2\sqrt{\lambda_i}}\right) = \sum_{i=1}^{a} \lambda_i^2(s_i-\bar{s})^2\tfrac{1}{4\lambda_i}V(Z_i)$$
となる.

(補題 3) (条件 1) が満たされているとすると, $n \to \infty$ として
$$b_s \equiv \frac{\sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})}{\sqrt{\sum_{i=1}^{a} n_i^2(s_i-\bar{s})^2\frac{n}{4n_i}}}$$
$$\xrightarrow{L} N(0,1)$$
が成り立つ.

(証明) 初めに b_sの分母, 分子を n で割ると
$$(b_s\text{の分子}) = \sum_{i=1}^{a} \frac{n_i}{n}(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})$$
$$\xrightarrow{L} \sum_{i=1}^{a} \lambda_i(s_i-\bar{s})\frac{Z_i}{2\sqrt{\lambda_i}}$$
$$(b_s\text{の分母}) = \sqrt{\sum_{i=1}^{a} \frac{n_i^2}{n^2}(s_i-\bar{s})^2\frac{n}{4n_i}}$$
$$\xrightarrow{P} \sqrt{\sum_{i=1}^{a} \lambda_i^2(s_i-\bar{s})^2\frac{1}{4\lambda_i}}$$
スラツキーの定理より,
$$b_s \xrightarrow{L} \frac{\sum_{i=1}^{a} \lambda_i(s_i-\bar{s})\frac{Z_i}{2\sqrt{\lambda_i}}}{\sqrt{\sum_{i=1}^{a} \lambda_i^2(s_i-\bar{s})^2\frac{1}{4\lambda_i}}}$$
このとき,
$$A = \frac{\sum_{i=1}^{a} \lambda_i(s_i-\bar{s})\frac{Z_i}{2\sqrt{\lambda_i}}}{\sqrt{\sum_{i=1}^{a} \lambda_i^2(s_i-\bar{s})^2\frac{1}{4\lambda_i}}}$$
とおくと.
$$b_s \sim N(A \cdot 0, A^2 \cdot 1) = N(0,1)$$
よって示せた. □
ここで,
$$\hat{b}_s = \frac{\sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \overline{\arcsin\sqrt{P}})}{\sqrt{\sum_{i=1}^{a} n_i^2(s_i-\bar{s})^2\frac{n}{4n_i}}}$$
とする.ただし,
$$\overline{\arcsin\sqrt{P}} = \sum_{i=1}^{a} \frac{n_i}{n}\arcsin\sqrt{\hat{p}_i}$$
とする.

(定理 4) (条件 1) が満たされているとすると, $n \to \infty$ として, H_0の下で
$$\hat{b}_s = \frac{\sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \overline{\arcsin\sqrt{P}})}{\sqrt{\sum_{i=1}^{a} n_i^2(s_i-\bar{s})^2\frac{n}{4n_i}}}$$
$$\xrightarrow{L} N(0,1)$$
が成り立つ.

(証明)
$$(\hat{b}_s - b_s\text{の分子}) = \sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \overline{\arcsin\sqrt{P}})$$
$$-\sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}(\arcsin\sqrt{\hat{p}_i} - \arcsin\sqrt{p_i})$$
$$= \sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}\arcsin\sqrt{\hat{p}_i} - \sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}\overline{\arcsin\sqrt{P}}$$
$$-\sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}\arcsin\sqrt{\hat{p}_i} + \sum_{i=1}^{a} n_i(s_i-\bar{s})\sqrt{n}\arcsin\sqrt{p_1}$$

より,
(与式) $= \sum_{i=1}^{a} n_i(s_i - \bar{s})\sqrt{n}\arcsin\sqrt{\hat{p}_i} - \sum_{i=1}^{a} n_i(s_i - \bar{s})\sqrt{n}\arcsin\sqrt{\bar{p}_i} = 0$

$$\Rightarrow \hat{b_s} - b_s = \frac{0}{\sqrt{\sum_{i=1}^{a} n_i^2(s_i - \bar{s})^2 \frac{n}{4n_i}}}$$

よって,

$$\hat{b}_s - b_s \xrightarrow{P} 0 \qquad \square$$

以上より,

$$\hat{b}_s^2 \xrightarrow{L} \chi_1^2$$

となる. 水準αの検定は,

$$\hat{b}_s^2 > \chi_1^2(\alpha) \Rightarrow H_0 \text{を棄却}$$

4　C 言語によるプログラム解説

Cochran-Armitage 検定による検定結果と Arcsine 変換を用いた検定結果を出力するプログラムを C 言語により作成した. 以下のプログラムは今回作成したプログラムの main プログラムである. ただし $s_i = i$ としている.

```
int main(void){
input();
teigi();
keisan();
hantei();
keisan2();
hantei2();
return 0;
}
```

1.input 関数の中で, 標本の個数 a, 順序尺度変数 ($x_1, \cdots, x_a$), 有意水準 α(ALPHA), 観測値の個数 ($n_1, \cdots, n_a$) を入力する.

2.teigi 関数の中で, $\hat{p}_i$(ph[i]), n(m), $\bar{\hat{p}}$(phb), $\bar{s}$(sb), s_i(s[i]) の値を求める.

3.keisan 関数の中で, $\hat{b}(bh)$, $\hat{V}_0(V0h)$, $\frac{\hat{b}^2}{\hat{V}_0}(vb)$, $\chi_1^2(\alpha)(X1)$ の値を求める.

4.hantei 関数の中で, keisan 関数により得た値を用い検定を行う.

5.keisan2 関数の中で, arcsin を用いた値を計算, 出力する.

6.hantei2 関数の中で, keisan2 関数により得た値を用い検定を行う.

4.1　検定内容

飲酒, 喫煙がそれぞれ呼吸器疾患に関係があるかどうかを調べるにあたり, 男女合わせて 775 人分のデータ (参考文献 [3] より) を使用した.
このとき, 男女ともに,『飲酒, 喫煙ともになし』,『飲酒なし, 喫煙はあり』,『飲酒, 喫煙ともにあり』の 3 標本についての検定を行った.

4.2　実行結果

今回はページの都合上, 有意水準 α=0.01 の男子の場合のみを記載する.
有意水準 0.01 での呼吸器疾患におけるタバコとお酒関係性（男）

```
順序尺度変数 a 個の値を入力 3
有意水準αを入力してください (0<α ≦ 1)0.01
x_1 における観測値の個数 n_1 を入力 141
x_2 における観測値の個数 n_2 を入力 133
x_3 における観測値の個数 n_3 を入力 185
Y_1 を入力 38
Y_2 を入力 102
Y_3 を入力 135
Y_1=38.000 ph[1]=0.270 m=459 phb=0.599 sb=2.096
Y_2=102.000 ph[2]=0.767 m=459 phb=0.599 sb=2.096
Y_3=135.000 ph[3]=0.730 m=459 phb=0.599 sb=2.096
bh1=70.638 bh2=321.782 bh=0.220 v0h1=0.240174
v0h2=321.782135 v0h=0.000746 vb=86502.912
帰無仮説 H0:p1=… =pa vs 対立仮説 H1:p1<=… <=pa
または p1>=… >=pa
T0>=X1 より H0 を棄却する
aph1=0.546 aph2=1.067 aph3=1.024 apb=0.890
bsh1=1571.490 bsh2=36924.500 bhs=8.178
帰無仮説 H0:p1=… =pa vs 対立仮説 H1:p1<=… <=pa
または p1>=… >=pa
T1>=X1 より H0 を棄却する
```

5　考察

有意水準 $\alpha = 0.01$ において帰無仮説 H_0 が棄却されることがわかった. このことから喫煙と飲酒が呼吸器疾患におおいに関係していることがわかった. また本稿では男性、女性による違いがあるのかを検証したがどちらも同様の結果になり男女差はないといえることがわかった.

6　おわりに

C 言語によってプログラムを作成し, Arcsine 変換を用いた検定方式でも同様の結果を導くことができた. また実際にプログラムを作成したことで, 今まで研究してきたことへの理解がより深まった.

参考文献

[1] 丹後俊郎, 小西貞則:『医学統計学の辞典』. 朝倉書店, 東京 2012 年

[2] 白石高章:『統計学の基礎』. 日本評論社, 東京,2012 年.

[3] Daniel,W.W.:『Biostatistics』.Wiley and Sons,1987 年