mwp-085

疫学系テーブルの分析　【 評価版 】

疫学の分野では分割表（クロス表）に基づく分析が統計的推論の基盤となります。Stata にはそれを支援するための機能が Epitab 系コマンドとして一式用意されていますが、本 whitepaper ではそれらのコマンドを使用して行く上で前提となる基本的事項について、情報を整理しておきます。

1. 分割表

Epitab 系コマンドとしては cs, ir, cc 等、10 種類ほどのコマンドが用意されているわけですが、いずれの場合においても分析対象となるのは分割表 (contingency table) あるいはクロス表 (cross tabulation) と呼ばれるテーブルです。前提となる研究スタイルによって異なった形式のテーブルが用いられるわけですが、基本となるのは次の 3 種類です。

(1) リスクデータ分析用の分割表 [コホート研究]
(2) 罹患率データ分析用の分割表 [コホート研究]
(3) 症例対照データ分析用の分割表 [症例対照研究]

なお以下においては、最も基本となる 2×2 分割表を前提に説明を行って行きます。

(1) リスクデータ

この場合、分析対象となる分割表は表 1 の形式となります。

表 1　リスクデータ分析用の分割表

	リスク因子		
	曝露	非曝露	計
症例	a	b	$a+b$
非症例	c	d	$c+d$
計	n_1	n_0	n

特定のリスク因子に関してあらかじめ曝露群 (exposed group) と非曝露群 (unexposed group) を設定した上で、症例 (cases) の発生を追跡し計測する形となるので、研究スタイルとしてはコホート研究 (cohort study) に区分されます。セルに含まれる数値は人数を表します。従って、$n(=a+b+c+d)$ 人全員が同一の期間観察されること、すなわち途中打切り (censoring) が発生しないことが前提となる点に注意してください。

Stata ではこの形式のデータを累積罹患データ (cumulative incidence data) と呼んでいますが、この形式のデータの場合、効果の判定に用いられる指標は**リスク**です。すなわち曝露群の場合には $\frac{a}{n_1}(=\frac{a}{a+c})$、非曝露群の場合には $\frac{b}{n_0}(=\frac{b}{b+d})$ という値（確率）がリスクとなるわけで、これらの値に基づき曝露の効果が評価されます。表 1 の形式のデータを分析する場合には `cs/csi` コマンドが用いられるわけですが、その用法については *mwp-012* をご参照ください。

(2) 罹患率データ

評価版では割愛しています。

(3) 症例対照データ

評価版では割愛しています。

2. Fisher の正確検定

評価版では割愛しています。

3. カイ 2 乗検定

Fisher の正確検定は超幾何分布に基づく確率を計算し検定を行うので、厳密な結果を得ることはできますが、n の値（表 4 参照）の増大に伴い、演算の負荷は膨大なものとなります。このため、その適用は n の値が比較的小さいケースに限られます。これに対し、`epitab` 系コマンドにおいてデフォルトで実行される χ^2 検定は近似計算に基づくものだけに負荷は軽く、n の値が大きなケースに対しても問題なく適用できます。

今、表 7 のような観測データが与えられたとします。

表 7 観測データ

	曝露	非曝露	計
症例	d_{11}	d_{12}	m_1
非症例	d_{21}	d_{22}	m_2
計	n_1	n_2	n

ただし

$$m_1 = d_{11} + d_{12} \qquad n_1 = d_{11} + d_{21}$$
$$m_2 = d_{21} + d_{22} \qquad n_2 = d_{12} + d_{22}$$
$$n = d_{11} + d_{12} + d_{21} + d_{22}$$

また、曝露群と非曝露群の間で発症の確率に差はないとしたときに期待される分割表を表 8 のように書くものとします。

表 8 期待値データ

	曝露	非曝露	計
症例	e_{11}	e_{12}	m_1
非症例	e_{21}	e_{22}	m_2
計	n_1	n_2	n

ただし、各セルの値は周辺度数より

$$\begin{aligned} e_{11} &= n_1 \cdot \tfrac{m_1}{n} \qquad & e_{12} &= n_2 \cdot \tfrac{m_1}{n} \\ e_{21} &= n_1 \cdot \tfrac{m_2}{n} \qquad & e_{22} &= n_2 \cdot \tfrac{m_2}{n} \end{aligned} \tag{1}$$

のように算出できる点に注意してください。このとき次のように定義される統計量を T とします。

$$T = \frac{(d_{11} - e_{11})^2}{e_{11}} + \frac{(d_{12} - e_{12})^2}{e_{12}} + \frac{(d_{21} - e_{21})^2}{e_{21}} + \frac{(d_{22} - e_{22})^2}{e_{22}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(d_{ij} - e_{ij})^2}{e_{ij}} \tag{2}$$

帰無仮説を

$$H_0:\ \text{曝露群と非曝露群とで発症の確率は等しい}$$

としたとき、統計量 T は近似的に自由度 1 の χ^2 分布に従うことが知られています。この性質を利用して検定を行うのが χ^2 検定です。

(1) csi コマンド実行例

最初にセクション 2 と同じ 2×2 分割表を用いて χ^2 検定を実行してみます。

```
. csi 7 2 8 15
```

```
. csi 7 2 8 15

                |  Exposed   Unexposed  |      Total
----------------+-----------------------+-----------
          Cases |        7           2  |          9
       Noncases |        8          15  |         23
----------------+-----------------------+-----------
          Total |       15          17  |         32
                |                       |
           Risk |  .4666667   .1176471  |     .28125
                |                       |
                |    Point estimate     |  [95% conf. interval]
----------------+-----------------------+-----------------------
Risk difference |        .3490196       |    .0537288    .6443104
     Risk ratio |        3.966667       |    .9686604    16.24351
Attr. frac. ex. |        .7478992       |   -.0323535    .9384369
Attr. frac. pop |        .5816993       |
----------------+-----------------------+-----------------------
                          chi2(1) =     4.80  Pr>chi2 = 0.0284
```

検定統計量が $\chi^2(1) = 4.80$ と示されていますが、( ) 内の 1 という数字は自由度を意味しています。また検定結果の p 値は 0.0284 とレポートされています。

(2) 検定結果の検証

評価版では割愛しています。

4. 層化データ

評価版では割愛しています。

5. 回帰モデル

評価版では割愛しています。

補足１ – 超幾何分布の確率関数値の算出

評価版では割愛しています。

補足 2 – グラフ作成コマンド操作

評価版では割愛しています。

■