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。

従来 epitab 系コマンドの仕様は [ST] マニュアル中に記載されていましたが、Stata14 からは [R] マニュアルの方に移設されました。

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

ここでは表 4 のようなリスクデータが与えられたときに、曝露の効果をどう判定するかについて考えてみることにします。

表 4　リスクデータ

	リスク因子		
	曝露	非曝露	計
症例	$s\ (=c_{11})$	(c_{12})	m_1
非症例	(c_{21})	(c_{22})	m_2
計	n_1	n_2	n

今、曝露群の人数 n_1 と非曝露群の人数 n_2 を所与とするなら、$n = n_1 + n_2$ も所与となります。次にこの n 人中で疾病に罹患した人数 m_1 を固定して考えるなら、非症例の人数 m_2 も固定値 $n - m_1$ となります。このように周辺度数 m_1, m_2, n_1, n_2, n が固定された場合、各セルの値は c_{11} の値によってすべて規定されることになります。例えば c_{11} の値を s と書くことにしたとき、c_{12} の値は $m_1 - s$ となります。曝露群と非曝露群とで発症の確率に差がないとしたとき、c_{11} の値が s であるテーブルが生成される確率 P_s は

$$P_s = \frac{{}_{n_1}C_s \cdot {}_{n_2}C_{m_1-s}}{{}_{n}C_{m_1}} \tag{1}$$

で与えられることになります。このような確率関数で表現される離散型分布を**超幾何分布** (hypergeometric distribution) と呼びます。

s について特定の値が与えられたとき、(1) 式に基づき確率を計算することによってそのような分割表の得られやすさを算出し、

H_0 : 曝露群と非曝露群とで発症の確率は等しい

という帰無仮説を検定するのが Fisher **の正確検定** (Fisher's exact test) です。

(1) csi コマンド実行例

今、次のような 2×2 分割表が与えられたとします。

表 5　サンプルデータ

	リスク因子		
	曝露	非曝露	計
症例	7	2	9
非症例	8	15	23
計	15	17	32

データ件数が合計 32 件と少ないので Fisher の正確検定を選択することにします。

- Statistics ▷ Epidemiology and related ▷ Tables for epidemiologists
  ▷ Cohort study risk-ratio etc. calculator と操作

次のようにデータを直接入力します[*1]。なお、Fisher's exact p という項目への ✓ を忘れないようにしてください。

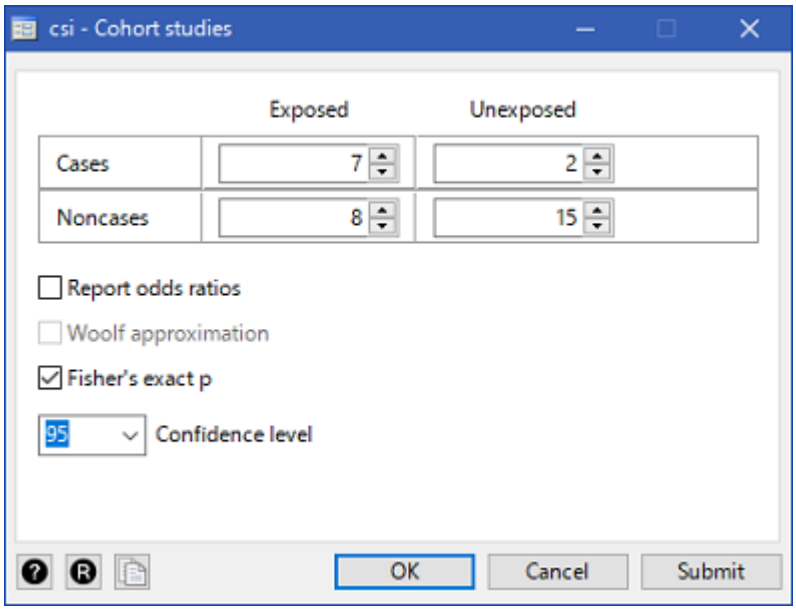

図 1 `csi` ダイアログ

```
. csi 7 2 8 15, exact

                   |  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       |
-------------------+-----------------------+-----------
                                1-sided Fisher's exact P = 0.0353
                                2-sided Fisher's exact P = 0.0491
```

[*1] `csi` は immediate タイプのコマンドなのでデータセットは必要としません。

`csi` コマンドの出力中には曝露群のリスク R_1 が 0.47、非曝露群のリスク R_0 が 0.12、従ってリスク比 RR は $R_1/R_0 = 3.97$ とレポートされています。両群で発症確率に差がないとしたときの正確検定の結果は出力末尾に示されており、片側検定の場合の p 値が 0.0353、両側検定の場合の p 値が 0.0491 と示されています。

(2) 検定結果の検証

評価版では割愛しています。

3. カイ 2 乗検定

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。

■