

クロス集計と独立性の検定

青木繁伸

1 目的

クロス集計表を作り、独立性の検定を行う。

2 使用法

```
from xtabs import xtabs
xtabs(df, x, y, row_percent=True, col_percent=True, verbose=True)
```

2.1 引数

<code>df</code>	データフレーム
<code>x</code>	クロス集計表の表側にする変数名
<code>y</code>	クロス集計表の表頭にする変数名
<code>row_percent</code>	行方向の % を付加する
<code>col_percent</code>	列方向の % を付加する
<code>verbose</code>	必要最小限のプリント出力をする (デフォルトは <code>True</code>)

2.2 戻り値の名前

<code>"table"</code>	分割表
<code>"colsums"</code>	列和
<code>"rowsums"</code>	行和
<code>"n"</code>	サンプルサイズ
<code>"chisq"</code>	χ^2 値
<code>"df"</code>	自由度
<code>"p value"</code>	p 値
<code>"chisq(Yates)"</code>	イエーツの補正による χ^2 値
<code>"p value(Yates)"</code>	イエーツの補正による p 値

3 使用例

```
import pandas as pd

df = pd.DataFrame({
    "a" : ["B", "D", "C", "E", "E", "A", "C", "E", "C", "C", "E", "C",
          "D", "C", "A", "E", "B", "A", "B", "E", "E", "D", "D", "E", "D",
```

```

"D", "C", "C", "B", "A", "E", "E", "D", "D", "A", "C", "D", "B",
"B", "B", "A", "C", "C", "B", "A", "A", "B", "C", "B", "E"],
"b" : [1, 2, 3, 1, 2, 1, 1, 3, 3, 2, 2, 1, 2, 1, 3, 2, 3, 3, 3, 2,
3, 2, 3, 1, 2, 1, 2, 2, 2, 1, 1, 3, 2, 3, 1, 2, 3, 3, 3, 1, 1,
2, 2, 2, 1, 1, 3, 1, 2, 2],
"c" : ["bar", "foo", "bar", "baz", "bar", "baz", "baz", "bar", "bar",
"foo", "baz", "foo", "foo", "baz", "baz", "foo", "bar", "baz",
"bar", "bar", "bar", "foo", "foo", "foo", "bar", "baz", "foo",
"foo", "foo", "baz", "bar", "baz", "baz", "baz", "bar", "bar",
"baz", "baz", "baz", "bar", "foo", "bar", "foo", "foo", "baz",
"foo", "foo", "foo", "foo", "baz"],
"d" : [2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1, 1,
2, 2, 1, 1, 1, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2,
1, 2, 1, 2, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2, 1, 1,
2, 1, 2],
"e" : [1, 2, 2, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1,
2, 1, 1, 2, 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 2, 2,
1, 2, 1, 1, 1, 1, 2, 1, 2, 2, 2, 1, 2, 1, 2, 1,
2, 2, 2]})
df.index = range(50)

import sys
sys.path.append("statlib")
from xtabs import xtabs

a = xtabs(df, "b", "c")

```

	[[c]]			
[[b]]	(1)	(2)	(3)	Sum
(bar)	4	5	6	15
row %	26.7	33.3	40.0	100.0
col %	25.0	26.3	40.0	30.0
(baz)	7	3	7	17
row %	41.2	17.6	41.2	100.0
col %	43.8	15.8	46.7	34.0
(foo)	5	11	2	18
row %	27.8	61.1	11.1	100.0
col %	31.2	57.9	13.3	36.0
Sum	16	19	15	50
row %	32.0	38.0	30.0	100.0
col %	100.0	100.0	100.0	100.0

Pearson's Chi-squared test
Chi squared = 8.4988, df = 4, p value = 0.0749
Chi-squared approximation may be incorrect

```
a = xtabs(df, "a", "c")
```

	[[c]]					
[[a]]	(A)	(B)	(C)	(D)	(E)	Sum
(bar)	1	4	4	1	5	15
row %	6.7	26.7	26.7	6.7	33.3	100.0

col %	12.5	40.0	33.3	11.1	45.5	30.0
(baz)	5	2	2	4	4	17
row %	29.4	11.8	11.8	23.5	23.5	100.0
col %	62.5	20.0	16.7	44.4	36.4	34.0
(foo)	2	4	6	4	2	18
row %	11.1	22.2	33.3	22.2	11.1	100.0
col %	25.0	40.0	50.0	44.4	18.2	36.0
Sum	8	10	12	9	11	50
row %	16.0	20.0	24.0	18.0	22.0	100.0
col %	100.0	100.0	100.0	100.0	100.0	100.0

Pearson's Chi-squared test

Chi squared = 9.1105, df = 8, p value = 0.3331

Chi-squared approximation may be incorrect

```
a = xtabs(df, "d", "e")
```

	[[e]]		
[[d]]	(1)	(2)	Sum
(1)	13	12	25
row %	52.0	48.0	100.0
col %	52.0	48.0	50.0
(2)	12	13	25
row %	48.0	52.0	100.0
col %	48.0	52.0	50.0
Sum	25	25	50
row %	50.0	50.0	100.0
col %	100.0	100.0	100.0

Pearson's Chi-squared test

Chi squared = 0.0800, df = 1, p value = 0.7773

Pearson's Chi-squared test with Yates' continuity correction

Chi squared = 0.0000, df = 1, p value = 1.0

```
a = xtabs(df, "a", "b", row_percent=False, col_percent=False)
```

	[[b]]					
[[a]]	(A)	(B)	(C)	(D)	(E)	Sum
(1)	6	2	4	1	3	16
(2)	0	3	6	5	5	19
(3)	2	5	2	3	3	15
Sum	8	10	12	9	11	50

Pearson's Chi-squared test

Chi squared = 13.1089, df = 8, p value = 0.1082

Chi-squared approximation may be incorrect