

# 名義尺度のデータの分析 (クロス表作成・カイ二乗検定)

2011/06/25

心理データ解析演習

M1 熊木 悠人(くまき ゆうと)

# 今日の予定

- 名義尺度のデータの扱い
- $2 \times 3$ のクロス表作成
- 連関係数
- $\chi^2$ 検定
- 残差分析
  
- (おまけ)  $2 \times 2$ のクロス表
- $\chi^2$ 検定が使えないとき  
Fisherの直接法による検定

# 名義尺度とは？

## 変数の分類

### □ 比率尺度

(0の点が一義的に決まっている、 $a \div b = c \div d$ )

### □ 間隔尺度

(データの変域によらず測定値の間隔が一定、 $a - b = c - d$ )

### □ 順位尺度

(測定値は大小のみを表す、 $a > b$  )

### □ 名義尺度

(測定値間に大小関係はない、 $a = b$  )

# 名義尺度とは？

## □ 例えば・・

- ・ 性別
- ・ 血液型
- ・ 出身地
- ・ 職種
- ・ 支持政党
- ・ 「Yes」 or 「No」
- ・ 「病気」または「健康」

などなど、名義尺度でしか測れない変数はたくさんある。

# 名義尺度による研究

Ex. 喫煙者は健常者と比べて肺がんになる率  
が高いか？

独立変数⇒喫煙者/非喫煙者

従属変数⇒肺がん患者/健常者

従属変数も、数値で表したり、順序をつけた  
りできない。

⇒名義尺度の変数とみなして分析。

# クロス集計

- 独立変数ごとに度数、比率（%）をクロス表に集計

表1. 肺がん患者と健常者における喫煙者の人数

	喫煙者	非喫煙者	計
肺がん患者	52	8	60
健常者	48	42	90
計	100	50	150

図解: 赤い楕円で60と100を囲い、赤い矢印が「周辺度数」を指す。青い楕円で150を囲い、青い矢印が「総度数」を指す。

喫煙者のほうが肺がん患者が多い？

⇒ 連関係数を見る

2 × 2 のクロス表 ⇒  $\phi$  係数

変数のカテゴリー数が 3 以上 ⇒ クラメールの連関係数

# φ係数（参考）

		カテゴリー		計
		1	2	
カテゴリー	1	$n_{11}$	$n_{12}$	$n_{1\cdot}$
	2	$n_{21}$	$n_{22}$	$n_{2\cdot}$
計		$n_{\cdot 1}$	$n_{\cdot 2}$	$N$

連関の強さを求める。  
具体的には以下の式

$$\phi = \frac{|n_{11}n_{22} - n_{12}n_{21}|}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$$

φ係数がとりうる値の範囲は $0 \leq \phi \leq 1$   
1に近づくほど連関が強いと判断される。

ちなみに、先程のデータでは

# Φ係数（参考）

## 注意すべき点

- 「各セルの度数が等しい  $\Rightarrow \phi=0$ 」  
だが、「 $\phi=0 \Rightarrow$  各セルの度数が等しい」ではない。
- 例えば、以下のクロス表のような値であれば、各セルの度数は全て異なるが、 $\phi=0$ となる。（連関はない）

	カテゴリー		計
	1	2	
カテゴリー 1	40	20	60
カテゴリー 2	60	30	90
計	100	50	150



# カイ二乗検定

- $\chi^2$ 分布に基づいて考えだされた統計的検定の総称  
1条件で従属変数のカテゴリーが複数ある場合、  
2×2のクロス表、  
条件数が3以上の場合 など、様々な $\chi^2$ 検定がある。
- 条件が複数の場合、条件間に対応のないケースで用いる。
- 帰無仮説  
「各条件によって従属変数の各カテゴリーの度数の比率に差はない。」

# カイ二乗検定

□ おおざっぱに言えば、

$$\chi^2 = \sum \frac{(\text{観測度数} - \text{期待度数})^2}{\text{期待度数}}$$

観測度数と期待度数との差を計算しているものである。

ちなみに、期待度数は

$$E_{ij} = \frac{n_i n_j}{N}$$

で求められる。

	カテゴリー		計
	1	2	
カテゴリー 1	$n_{11}$	$n_{12}$	$n_{1\cdot}$
カテゴリー 2	$n_{21}$	$n_{22}$	$n_{2\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$N$

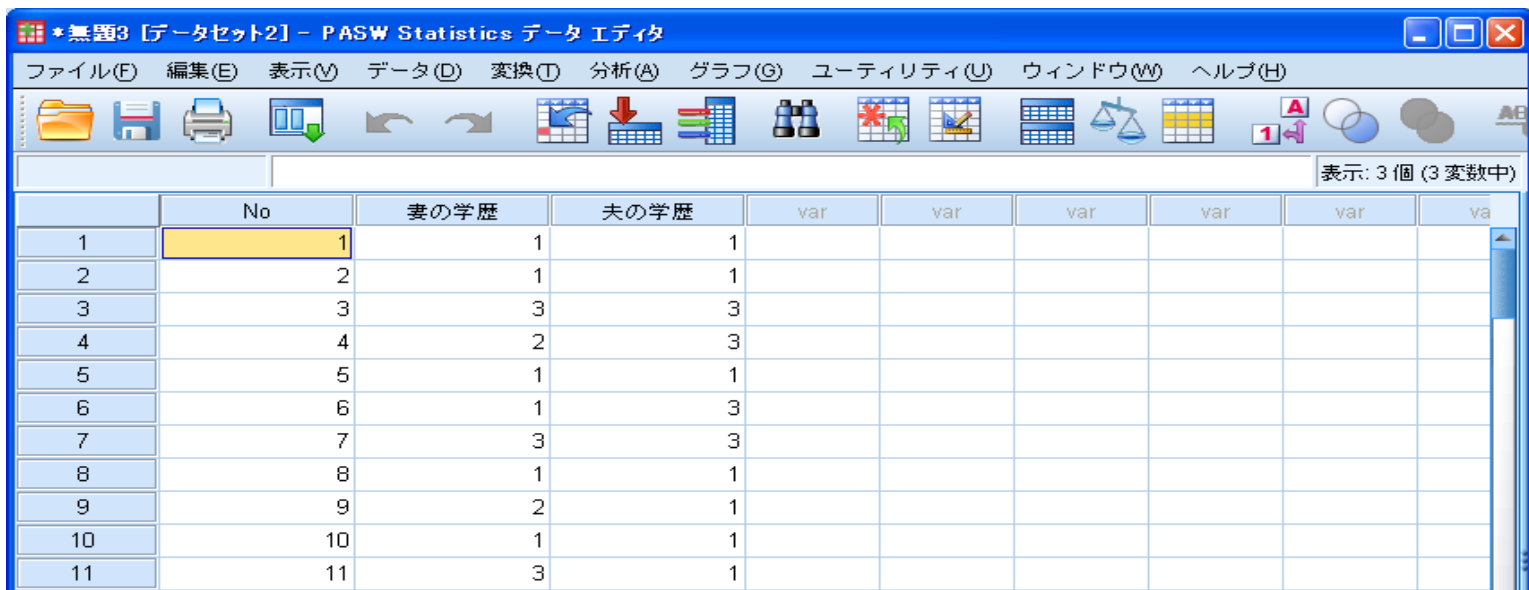
# カイ二乗検定

## カイ二乗検定を行うときの注意

- 条件間に対応のないケースのみ使用可能  
⇒対応がある場合、マクネマー検定、コクランのQ検定など、他の検定を使う。
- 観測度数が少ない ( $N < 20$ ) や、期待度数が5未満のセルがある場合、  
 $\chi^2$ 検定は行うべきでない。
- $\chi^2$ 検定では、何らかの連関があることは示せても、どのような連関があるかまでは示せない。  
⇒残差分析

# データの読み込み

- Excelデータをダウンロードし、SPSSを立ち上げる。
- 【ファイル(F)】 → 【開く(O)】 → 【データ(A)】 から、さきほどダウンロードしたExcelデータを読み込む
  - 太郎丸(2005)のデータを一部改変

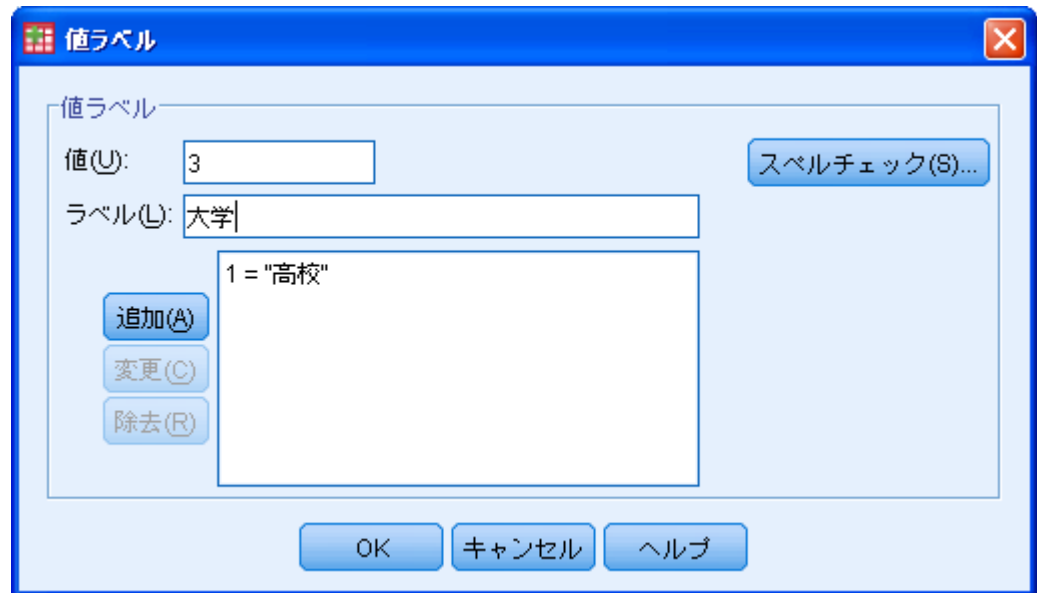


The screenshot shows the PASW Statistics Data Editor window. The title bar reads '\*無題3 [データセット2] - PASW Statistics データ エディタ'. The menu bar includes 'ファイル(F)', '編集(E)', '表示(V)', 'データ(D)', '変換(T)', '分析(A)', 'グラフ(G)', 'ユーティリティ(U)', 'ウィンドウ(W)', and 'ヘルプ(H)'. The toolbar contains various icons for file operations, data manipulation, and analysis. The data table has 11 rows and 10 columns. The first column is 'No', the second is '妻の学歴', the third is '夫の学歴', and the remaining seven are labeled 'var'. The first row is highlighted in yellow.

	No	妻の学歴	夫の学歴	var	var	var	var	var	va
1	1	1	1						
2	2	1	1						
3	3	3	3						
4	4	2	3						
5	5	1	1						
6	6	1	3						
7	7	3	3						
8	8	1	1						
9	9	2	1						
10	10	1	1						
11	11	3	1						

# データの読み込み

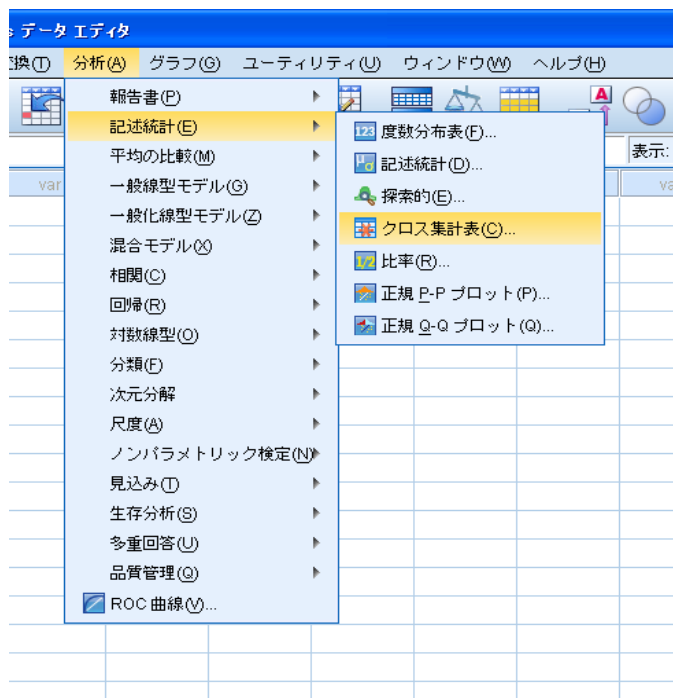
- 値ラベルの入力
- 変数ビューを開き、  
「妻の学歴」の「値」に、1=高校、2=短大、3=大学  
「夫の学歴」に、1=高校、3=大学  
を入れる。



# クロス集計してみる

先程ダウンロードしたデータを用いて、

- 分析(A)→記述統計(E)→クロス表集計(C)を選択
- 行に「夫の学歴」、列に「妻の学歴」を入れる。
- セル表示の設定で、観測度数にチェック



# クロス集計してみる

以下のように出力される。

処理したケースの要約

	ケース					
	有効数		欠損		合計	
	N	パーセント	N	パーセント	N	パーセント
夫の学歴 * 妻の学歴	148	100.0%	0	.0%	148	100.0%

夫の学歴と妻の学歴のクロス表

度数

		妻の学歴			合計
		高校	短大	大学	
夫の学歴	高校	72	10	4	86
	短大	26	20	16	62
合計		98	30	20	148

# クロス表を集計してみる

- 先程のクロス表から仮説を考える
- 1. 妻は自分と同程度の学歴の夫を選ぶ傾向にある。
- 2. 妻は自分よりも高い学歴の夫を選ぶ傾向にある。
- 3. 妻の学歴と夫の学歴には連関はない。  
(統計的に独立である)

など... これらの仮説について検討する。



# カイ二乗検定

## □ ちなみに...

期待度数はSPSSでオプションで出力できる。

## □ 「クロス表集計」 → 「セル」 → 「度数」の中の「期待」にチェック

処理したケースの要約

	ケース					
	有効数		欠損		合計	
	N	パーセント	N	パーセント	N	パーセント
夫の学歴 * 妻の学歴	148	100.0%	0	.0%	148	100.0%

夫の学歴と妻の学歴のクロス表

期待度数

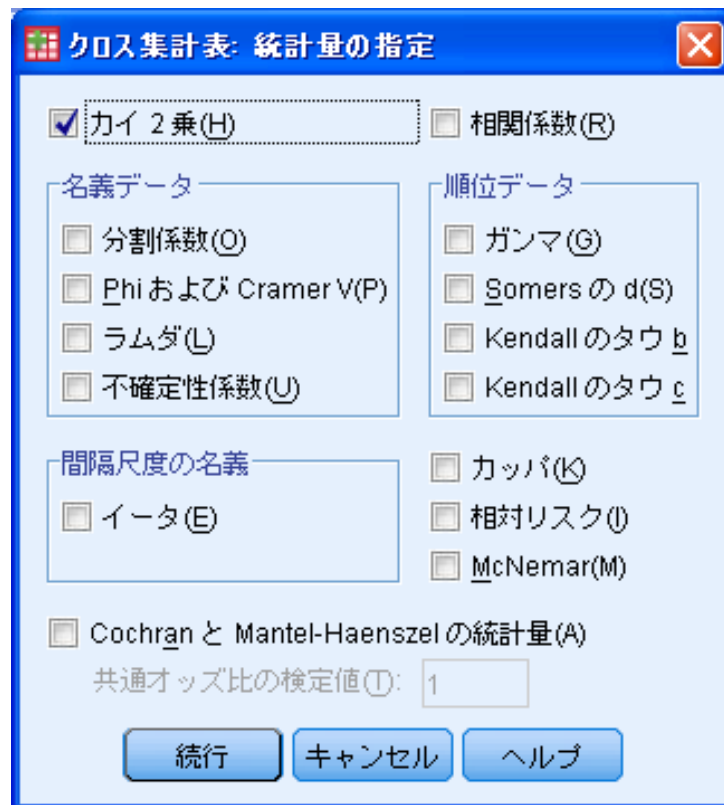
		妻の学歴			合計
		高校	短大	大学	
夫の学歴	高校	56.9	17.4	11.6	86.0
	短大	41.1	12.6	8.4	62.0
合計		98.0	30.0	20.0	148.0

# カイ二乗検定を試してみる

先程と同じように

- 分析(A)→記述統計(E)→クロス表集計(C)を選択
- 行に「夫の学歴」、列に「妻の学歴」を入れる。

- 「統計量の決定」で「カイ二乗」にチェックを入れる。



# カイ二乗検定を試してみる

## 結果の出力

処理したケースの要約

	ケース					
	有効数		欠損		合計	
	N	パーセント	N	パーセント	N	パーセント
夫の学歴*妻の学歴	148	100.0%	0	.0%	148	100.0%

夫の学歴と妻の学歴のクロス表

度数

		妻の学歴			合計
		高校	大学	短大	
夫の学歴	高校	72	4	10	86
	大学	26	16	20	62
合計		98	20	30	148

カイ2乗検定

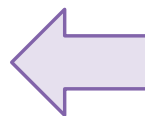
	値	自由度	漸近有意確率 (両側)
Pearsonのカイ2乗	28.996 <sup>a</sup>	2	.000
尤度比	29.663	2	.000
有効なケースの数	148		

a. 0セル(.0%)は期待度数が5未満です。最小期待度数は8.38です。

Pearsonのカイ二乗を見る。  
1%水準で有意

2×2のクロス表の場合、「連続修正」の値を見る。

ちなみに、期待度数が5未満のセルがある場合、カイ二乗検定は使うべきでない。



# カイ二乗検定を試してみる

□ 今の検定からわかったこと…

「条件によって、従属変数の各カテゴリーの度数に差がある。」

⇒これだけでは、

1. 夫と妻は同じくらいの学歴である傾向が強い
  2. 夫は妻より学歴が高い傾向が強い
- といった仮説は検討できていない。

あくまで、

3. 夫と妻の学歴は統計的に独立である。（関連はない）
- という仮説を棄却しただけである。

# 残差分析

## □ 残差とは？

セルの観測値と期待値の差

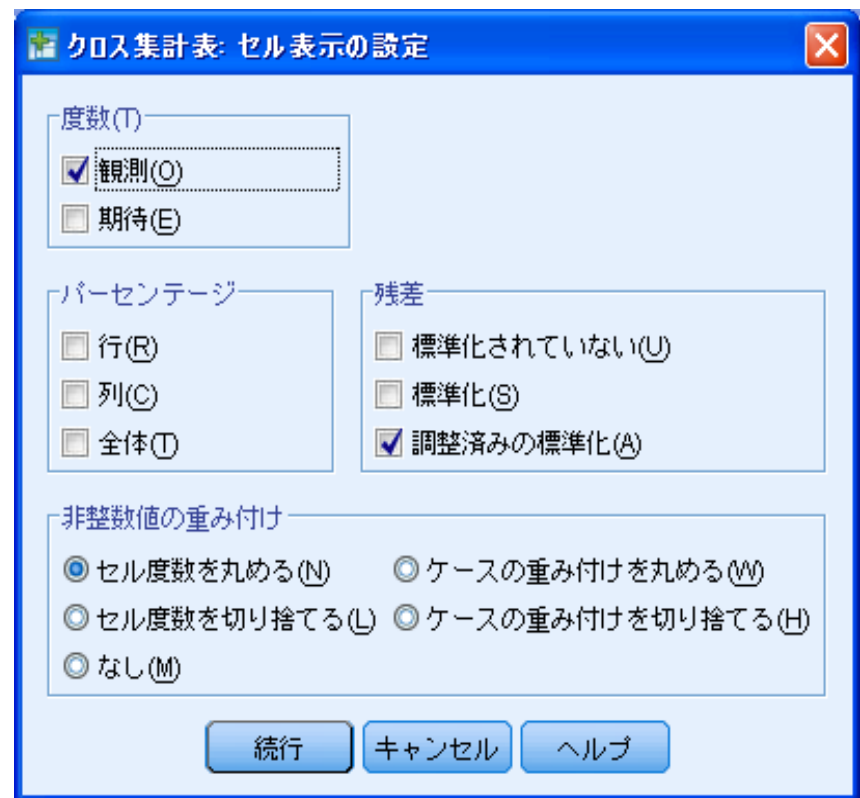
観測値が期待値よりも大きければ正、  
観測値が期待値よりも小さければ負、の値をとる。

→調整残差は正規分布に近似するので、検定可能  
検定が有意であれば、残差の符号を見ることで、  
判断する。

# 残差分析

- 先程と同じようにクロス表集計を開き、「セルの表示の決定」の中の「残差」の項目を見る。その中の「調整済みの標準化」にチェックを入れる。

調整残差は標準残差より正規分布に近似するため、通常、調整残差を用いる。



# 残差分析

## □ 結果の出力

処理したケースの要約

	ケース					
	有効数		欠損		合計	
	N	パーセント	N	パーセント	N	パーセント
夫の学歴*妻の学歴	148	100.0%	0	.0%	148	100.0%

夫の学歴と妻の学歴のクロス表

			妻の学歴			合計
			高校	大学	短大	
夫の学歴	高校	度数	72	4	10	86
		調整済み残差	5.3	-3.7	-3.1	
	大学	度数	26	16	20	62
		調整済み残差	-5.3	3.7	3.1	
合計		度数	98	20	30	148

全てのセルの残差が1%水準で有意。  
あとは、+か-かを見る。

標準正規分布を使った検定の限界値

	1%水準	5%水準
両側検定	2.58	1.96
片側検定	2.33	1.64

Habermanの残差検定

カイ2乗検定

	値	自由度	漸近有意確率 (両側)
Pearsonのカイ2乗	28.996 <sup>a</sup>	2	.000
尤度比	29.663	2	.000
有効なケースの数	148		

a. 0セル(.0%)は期待度数が5未満です。最小期待度数は8.38です。

# 残差分析

今回の残差分析の結果からわかること

- 妻が「高校」の場合、夫も「高校」となることが多い。
- 妻が「短大」の場合、夫は「大学」となることが多い。
- 妻が「大学」の場合、夫も「大学」となることが多い。

⇒ 「妻の学歴と夫の学歴は同程度である傾向が強い」  
という仮説 1 が確かめられる。



# 2×2のクロス表（おまけ）

- 先程の2×3のクロス表では、普通にカイ二乗の出力を見ればよかったが…

クロス表が2×2の場合、 $\frac{1}{E}$ のため、修正をかけた値を見る必要がある。

また、総度数が極端に少なかったり、期待度数の少ないセルが存在する場合には、 $\chi^2$ 検定を用いることは適切でないので、別の検定を用いなければならない。

⇒（おまけ）ではここを説明。

# 2×2のクロス表(おまけ)

先程と同様に

- Excelデータをダウンロードし、SPSSを立ち上げる。
- 【ファイル(F)】 → 【開く(O)】 → 【データ(A)】 から、さきほどダウンロードしたExcelデータを読み込む。
- 変数ビューの「値」を開き  
職種を「1=教師」「2=カウンセラー」  
評価を「1=反社会性重視」「2=非社会性重視」  
に設定。

# 2×2のクロス表(おまけ)

- 分析(A)→記述統計(E)→クロス表集計(C)を選択
- 行に「評価」、列に「職種」を入れる。
- セル表示の設定で、観測度数にチェック

[データセット1]

⇒クロス表を出力

処理したケースの要約

	ケース					
	有効数		欠損		合計	
	N	パーセント	N	パーセント	N	パーセント
職種 * 評価	23	100.0%	0	.0%	23	100.0%

職種と評価のクロス表

度数

		評価		合計
		反社会的行動重視	非社会的行動重視	
職種	教師	12	1	13
	カウンセラー	6	4	10
合計		18	5	23

# 2×2のクロス表（おまけ）

先程と同様に

- 「統計量の決定」で「カイ二乗」にチェックを入れて、 $\chi^2$ 検定を出力。

カイ2乗検定

	値	自由度	漸近有意確率 (両側)	正確有意確率 (両側)	正確有意確率 (片側)
Pearsonのカイ2乗	3.468 <sup>a</sup>	1	.063		
連続修正 <sup>b</sup>	1.829	1	.176		
尤度比	3.574	1	.059		
Fisherの直接法				.127	.089
有効なケースの数	23				

a. 2セル (50.0%)は期待度数が5未満です。最小期待度数は2.17です。

b. 2x2表に対してのみ計算

通常、クロス表が2×2であれば、この「連続修正」の値を用いる。（イエーツの連続性の修正）

# 2×2のクロス表（おまけ）

□ しかし、今回は…

カイ 2 乗検定

	値	自由度	漸近有意確率 (両側)	正確有意確率 (両側)	正確有意確率 (片側)
Pearsonのカイ 2 乗	3.468 <sup>a</sup>	1	.063		
連続修正 <sup>b</sup>	1.829	1	.176		
尤度比	3.574	1	.059		
Fisherの直接法				.127	.089
有効なケースの数	23				

a. 2セル (50.0%)は期待度数が 5 未満です。最小期待度数は 2.17 です。

b. 2x2 表に対してのみ計算

極端にケース数が少なかったり、期待度数が5未満のセルがあるときには、 $\chi^2$ 検定を行うのは適切でない。

⇒Fisherの直接法による検定を行う。(SPSSでは自動で出力)

# Fisherの直接法

- ノンパラメトリック検定
- 対応のない2条件間の比率の比較を行う。
- 周辺度数に10前後の小さな値がある、期待度数が0に近い数字がある時に用いる。

# まとめると・・・

- 独立変数または従属変数のカテゴリーが3以上  
⇒ $\chi^2$ 検定を行い、  
出力では「Pearsonのカイ二乗」の値を見る。
- $2 \times 2$ のクロス表  
⇒ $\chi^2$ 検定を行い、  
出力では「連続修正」の値を見る。
- 期待度数が5未満のセルが有る場合、  
⇒「Fisherの直接法」の値を見る。

# 参考文献

- 浅野 弘明(2010) 『実習で学ぶSPSSと統計学の基礎』  
プレデラス出版
- 太郎丸 博(2005) 『人文・社会科学のためのカテゴリカル・データ解析入門』 ナカニシヤ出版
- 森 敏昭 吉田寿夫(2009) 『心理学のためのデータ解析テクニカルブック』 北大路書房
- Alan Agresti著 渡邊裕之他訳(2003) 『カテゴリカルデータ解析入門』 サイエンティスト社