7. 関係を調べるための方法

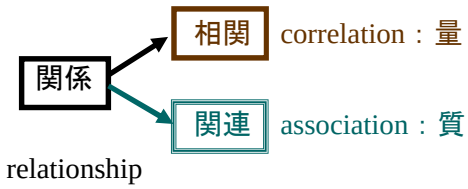
高木廣文
東邦大学看護学部

関係を調べるための方法

・身長が大きいと体重も重い。
・塩分の摂りすぎは高血圧症のもと。
・タバコの吸いすぎは肺ガンの原因。

物事には何らかの関係があるものが多い。
「関係」を統計学的に調べるには？

関係の分類



relationship

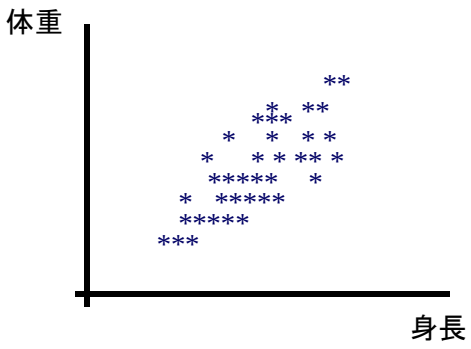
相関関係について

●２つの量的な変数間の関係の程度を示す。

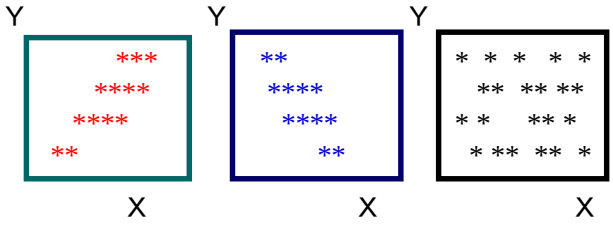
●相関図（散布図）
correlation diagram, scatter plots
２変数間の相関関係を図示。

●相関係数 correlation coefficient
２変数間の相関関係の大きさを数値化。

相関図の例



いろいろな相関関係



1) 正の相関（順相関）　2) 負の相関（逆相関）　3) 相関なし（無相関）

回帰 regression

回帰直線regression line:

Yの予測値＝[係数]×[Xの値]＋[定数]

$$\hat{y} = b\,x + a$$

最小2乗法によるパラメータ推定

最小2乗法 least squares method

$$L = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - bx_i - a\right)^2$$

Lを最小にするように，b，a を定める。

回帰直線のパラメータ推定

回帰係数の推定:

$$b = \frac{X\text{とY}の共分散}{X\text{の分散}}$$

定数項の推定:

$$a = \bar{y} - b \times \bar{x}$$

共分散 covariance

２変数の変動関係の指標

・２変数Ｘ，Ｙの偏差の積の平均値：

$$\text{共分散 } S_{xy} = \frac{1}{n}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

相関係数 correlation coefficient

量的な2変数間の関係の強さを示す指標。
通常は，2変数間の直線的な関係の程度を表すために用いられる。

Pearsonの積率相関係数，単相関係数
Pearson’s product moment cor. coef.
simple correlation coefficient

単相関係数の定義

$$r_{xy} = \frac{S_{xy}}{S_x \times S_y} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

$$-1 \le r_{xy} \le 1$$

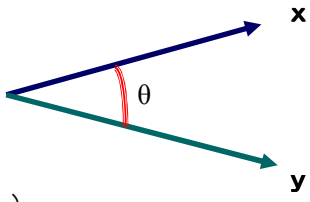
決定係数 coefficient of determination

・相関係数の２乗
一方の変動により，他方の変動を
どれだけ説明することができるか。

$$0 \le r_{xy}^{\ 2} \le 1$$

相関係数の幾何学的意味

2つのベクトルの余弦と相関係数



$$\cos\theta = \frac{(\mathbf{x},\ \mathbf{y})}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \text{XとYの相関係数 } r_{XY}$$

（参考）ベクトルの知識

・**ベクトルの掛け算**：

$\mathbf{x}=(x_1, x_2, \ldots, x_N)$，　$\mathbf{y}=(y_1, y_2, \ldots, y_N)$

$\mathbf{xy}' = x_1y_1 + x_2y_2 + \cdots + x_Ny_N$

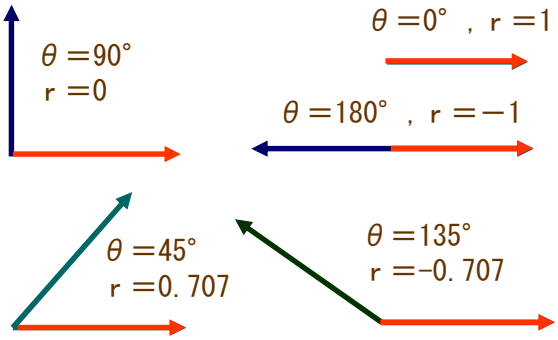
対応する要素の積の和。

「'」はベクトルの縦，横を転置する意味

・**ベクトルのノルム**（要素の２乗の和）：

$\mathbf{xx}' = x_1x_1 + x_2x_2 + \cdots + x_Nx_N = \|\mathbf{x}\|^2$

ベクトル表示による相関関係



相関係数の大きさと関係の強さ

絶対値	相関関係の強さ
0. 0～0. 2	ほとんど相関関係がない
0. 2～0. 4	やや（弱い）相関関係がある
0. 4～0. 7	かなり（強い）相関関係がある
0. 7～1. 0	強い相関関係がある

単相関係数の計算例：

●５人の身長と体重のデータ

身長	体重
170	65
165	57
180	72
168	55
162	50
平均値： 169	59. 8
分　散： 37. 6	60. 6

共分散と相関係数の計算

共分散＝[(170−169)×(65−59.8)+(165−169)×(57−59.8)
+(180−169)×(72−59.8)+(168−169)×(55−59.8)
+(162−169)×(50−59.8)]／5

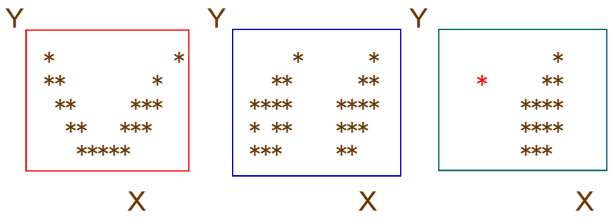
＝{1×5.2+(−4)×(−2.8)+11×12.2
+(−1)×(−4.8)+(−7)×(−9.8)}／5

＝224/5＝44.8

$$\text{単相関係数}=\frac{44.8}{\sqrt{37.6\times 60.6}}=0.939$$

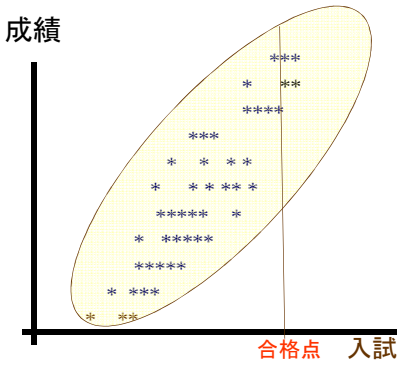
19

問題のある相関図



1) 曲線相関　2) 層化が必要　3) 外れ値
4) 打ち切りデータ（入試のデータなど）

打ち切りデータ



順位相関係数 rank cor. coef.

- データは量的でも，その数値自体の信頼性が低い場合。
- データがもともと大小関係のみが与えられた順序尺度である場合。

データの順位に基づく相関係数を求める。

「順位相関係数」

よく用いられる順位相関係数

(1)スピアマンの順位相関係数
Spearman's rank correlation coefficient
順位データにピアソンの積率相関係数の計算方法をそのまま用いたもの。

(2)ケンドールの順位相関係数
Kendall's rank correlation coefficient
データの大小関係のみから計算する。

スピアマンの順位相関係数

n標本中のケース i の２変数のデータ (x_i, y_i) が順位で与えられている場合，
スピアマンの順位相関係数 r_s ：

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}$$

タイのある場合のスピアマンの順位相関係数の補正（1）

変数 x に k 個の同順位のデータの組。
各組には t_{x_k} 個のデータ。
変数 y には m 個の同順位のデータの組。
各組には t_{y_m} 個のデータ。

修正項：

$$T_x = \sum_{i=1}^{k} t_{x_i}(t_{x_i}{}^2 - 1)$$

$$T_y = \sum_{i=1}^{m} t_{y_i}(t_{y_i}{}^2 - 1)$$

タイのある場合のスピアマンの順位相関係数の補正（2）

修正したスピアマンの順位相関係数：

$$r_s = \frac{n(n^2-1) - (T_x + T_y)/2 - 6\sum_i (x_i - y_i)^2}{\sqrt{[n(n^2-1) - T_x/2] \times [n(n^2-1) - T_y/2]}}$$

ケンドールの順位相関係数τ

n 組のデータから，任意に２組のデータを取り出した場合：

- ２組でのＸとＹの順位の大小関係が一致する組の個数をＰ
- 一致しない組の個数をＱ

$$\tau = \frac{2(P-Q)}{n(n-1)}$$

タイのある場合のケンドールの順位相関係数の修正

順位に同順位（タイ）がある場合：
T_x, T_y を変数 x, y の各タイの組み合わせ数の総数。

$$\tau = \frac{P-Q}{\sqrt{[n(n^2-1) - T_x/2] \times [n(n^2-1) - T_y/2]}}$$

順位相関の計算例（1）

●５人の身長と体重のデータ

スピアマンの順位相関係数のための表

身長	体重	身長順位	体重順位	差	差の２乗
170	65	4	4	0	0
165	57	2	3	−1	1
180	72	5	5	0	0
168	55	3	2	1	1
162	50	1	1	0	0

順位相関の計算例(2)

順位の差の2乗和＝0＋1＋0＋1＋0＝2

スピアマンの順位相関係数

＝1－6×2／[5×(5×5－1)]

＝1－12／120

＝0．9

順位相関の計算例(3)

ケンドールの順位相関係数の計算のための
データの大きさの比較

番号	1	2	3	4	5
1	--				
2	大, 大	--			
3	小, 小	小, 小	--		
4	大, 大	小, 大	大, 大	--	
5	大, 大	大, 大	大, 大	大, 大	--
一　致	4	2	2	1	--
不一致	0	1	0	0	--

順位相関の計算例(4)

・一致数Ｐ＝４＋２＋２＋１＝９
・不一致数Ｑ＝０＋１＋０＋０＝１

ケンドールの順位相関
　＝２×（９－１）／[５×（５－１）]
　＝１６／２０
　＝０．８

●クロス集計と独立性の検定

喫煙の状況を調べる場合，単に喫煙者数や割合を求めるだけでなく，男女差や，飲酒との関係なども知りたいと考えるのではないだろうか。

・質的データ間の関係を調べるには？

⇒ 「クロス集計 cross tabulation 」

●関係をどのように調べるのか？

表　Ａ薬とＢ薬の症状改善について

	著効	有効	不変	悪化	計
Ａ薬	8	10	6	1	25
Ｂ薬	3	9	10	4	26
計	11	19	16	5	51

Ｑ．どのようにすれば２つの薬剤の有効性に相違があるといえるのか？

クロス表の記号：

表　変数ＡとＢのクロス表

A	B						計
	B_1	B_2	⋯	B_j	⋯	B_l	
A_1	n_{11}	n_{12}	⋯	n_{1j}	⋯	n_{1l}	$n_{1\cdot}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
A_i	n_{i1}	n_{i2}	⋯	n_{ij}	⋯	n_{il}	$n_{i\cdot}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
A_k	n_{k1}	n_{k2}	⋯	n_{kj}	⋯	n_{kl}	$n_{k\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	⋯	$n_{\cdot j}$	⋯	$n_{\cdot l}$	N

クロス表に関する用語：

２変数Ａ，Ｂに関してクロス表を作成：
・変数Ａはｋ個のカテゴリ，
・変数Ｂはｌ個のカテゴリ：
「ｋ×ｌのクロス表」
（ｋ×ｌ）個のマス目 －「セル cell 」

●変数ＡについてカテゴリA_1，かつＢについてB_1という特性をもつものの度数を n_{11} のように表す。

クロス表に関する用語の続き

・カテゴリA_iの度数の合計：$n_{i\bullet}$

・カテゴリB_jの度数の合計：$n_{\bullet j}$

$$n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{il}$$

$$n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{kj}$$

・全体の標本数を Nとする。

セル度数の期待値

２つの変数Ａ，Ｂが独立な場合：

・ＡのカテゴリA_iの割合：$n_{i\bullet}/N$

・ＢのカテゴリB_jの割合：$n_{\bullet j}/N$

対応するセルの期待値：e_{ij}

$$e_{ij} = \frac{n_{i\bullet}}{N} \times \frac{n_{\bullet j}}{N} \times N = \frac{n_{i\bullet} \times n_{\bullet j}}{N}$$

クロス表の記号：

表　変数ＡとＢのクロス表

A	B_1	B_2	⋯	B_j	⋯	B_l	計
A_1	n_{11}	n_{12}	⋯	n_{1j}	⋯	n_{1l}	$n_{1\bullet}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
A_i	n_{i1}	n_{i2}	⋯	n_{ij}	⋯	n_{il}	$n_{i\bullet}$
⋮	⋮	⋮	⋯	⋮	⋯	⋮	⋮
A_k	n_{k1}	n_{k2}	⋯	n_{kj}	⋯	n_{kl}	$n_{k\bullet}$
計	$n_{\bullet 1}$	$n_{\bullet 2}$	⋯	$n_{\bullet j}$	⋯	$n_{\bullet l}$	N

カイ２乗値統計量

$$\left[\frac{(\text{期待値} - \text{実測値})^2}{\text{期待値}} \right] \text{の全セルの合計}$$

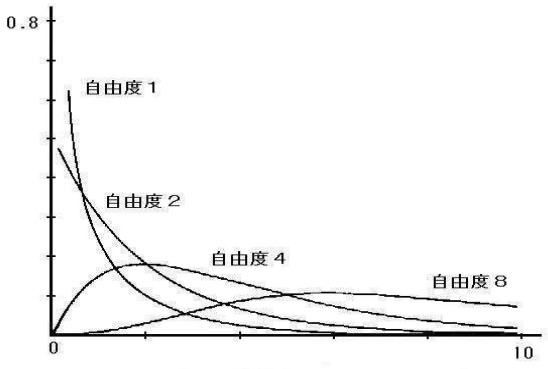
検定のためのカイ２乗値の計算

２変数Ａ，Ｂの関連の有無を検定するための統計量：χ_0^2

$$\chi_0^2 = N\left(\frac{n_{11}^2}{n_{1\bullet} \times n_{\bullet 1}} + \frac{n_{12}^2}{n_{1\bullet} \times n_{\bullet 2}} + \cdots + \frac{n_{kl}^2}{n_{k\bullet} \times n_{\bullet l}} - 1 \right)$$

$$= N\left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n_{i\bullet} \times n_{\bullet j}} - 1 \right)$$

自由度 $\nu = (k-1) \times (l-1)$ のカイ2乗分布



クロス表の自由度について

表　Ａ薬とＢ薬の症状改善について

	著効	有効	不変	悪化	計
Ａ薬	＊	＊	＊	＊	25
Ｂ薬	＊	＊	＊	＊	26
計	11	19	16	5	51

Q．最大でいくつ自由にセルの値を決められるのか？

検定の方法

χ_0^2と自由度νのカイ２乗分布の上側５％（または１％）点の値$\chi_{0.05}^2(\nu)$と比較：

⇒ $\chi_0^2 > \chi_{0.05}^2(\nu)$ → 有意

$\chi_0^2 < \chi_{0.05}^2(\nu)$ → 態度保留

【注意】

実際には，χ_0^2と自由度νのカイ２乗分布からｐ値（有意確率）を求めて検定する。

クロス表の検定での注意事項

・カテゴリ数が大きくなるにつれ，セルの個数も急激に増加。
  ・セルの度数が極小になることがある。
  ・χ_0^2のカイ２乗分布への適合が悪化。

・正確には，各セルの期待値が問題：
期待値が５以下の場合，カイ２乗分布による近似はよくないといわれている。
⇒可能ならば隣合うカテゴリ同士を併合して，検定を行う。

クロス表の検定の例題:

前述のＡ薬とＢ薬の症状改善についてのクロス表の検定を行う。

$$\chi_0^2 = 51 \times \left[\frac{8^2}{25\times11} + \frac{10^2}{25\times19} + \frac{6^2}{25\times16} + \frac{1^2}{25\times5} + \frac{3^2}{26\times11} + \frac{9^2}{26\times19} + \frac{10^2}{26\times16} + \frac{4^2}{26\times5} - 1 \right]$$

$$= 5.108$$

例題の続き:

χ_0^2の自由度：(2−1)×(4−1)＝3

⇒　自由度３のカイ２乗分布の上側５％点の値は7.815

⇒　$\chi_0^2 < 7.815$ → 判断保留

⇒　Ａ薬とＢ薬の効果に差があるとはいえない。

●カテゴリ併合の例

期待値が５未満のセルがあるので，カテゴリの併合を行う。

表　カテゴリの併合

	著効	有効	その他	計
Ａ薬	8	10	7	25
Ｂ薬	3	9	14	26
計	11	19	21	51

不変と悪化を併合し「その他」とした。

併合の続き：

検定のための統計量：

$\chi_0^2 = 4.641$

自由度：$(3-1) \times (2-1) = 2$

⇒　自由度２のカイ２乗分布の
　　上側５％点$5.991 > \chi_0^2 = 4.641$

⇒　保留（有意でない）

・併合は無意味か？

p値の変化：$p = 0.164 \rightarrow 0.098$

四分表の検定

●慢性前立腺患者に，無作為にA薬とプラセボを各15例ずつ投与し，自覚症状について効果を調べたところ，下の表に示したような結果が得られた。このデータから，A薬に効果があるといえるだろうか。

表　Ａ薬とプラセボの投与結果

	自覚症状		計
	改善	不変	
Ａ　薬	11	4	15
プラセボ	6	9	15
計	17	13	30

四分表の一般的な表現

表　ＡとＢの四分表

A	B ○	B ×	計
○	a	b	n_1
×	c	d	n_2
計	m_1	m_2	N

$n_1 = a+b$, $n_2 = c+d$, $m_1 = a+c$, $m_2 = b+d$

四分表の期待値

・変数Ａが○，かつ変数Ｂが○のセル：

観察値：ａ

期待値：$e(a)$（ＡとＢが独立ならば）

$$e(a) = \frac{n_1}{N} \times \frac{m_1}{N} \times N = \frac{n_1 \times m_1}{N}$$

期待値とカイ２乗値

各セルの期待値と観察値から得られる統計量
－ カイ２乗値：χ_0^2

$$\chi_0^2 = \frac{[\text{期待値} - \text{観察値}]^2}{\text{期待値}} \text{の合計}$$

四分表の独立性の検定

●２変数の両方が２カテゴリからなる場合：
「４分表」four-fold table
「独立性の検定」のための統計量：

$$\chi_0^2 = \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2}$$

・カイ２乗分布への当てはまりが悪い。

イエーツの連続性の補正

●四分表の検定の「連続性の補正」：
カイ２乗分布への当てはまりがよくなる

$$\chi_Y^2 = \frac{N(|ad - bc| - N/2)^2}{n_1 n_2 m_1 m_2}$$

・連続性の補正：$|ad - bc| > N/4$　の場合

・そうでない場合には：$\chi_Y^2 = 0$

カイ2乗検定の方法 ： $\chi^2 > 3.84$ ⇒有意な関連

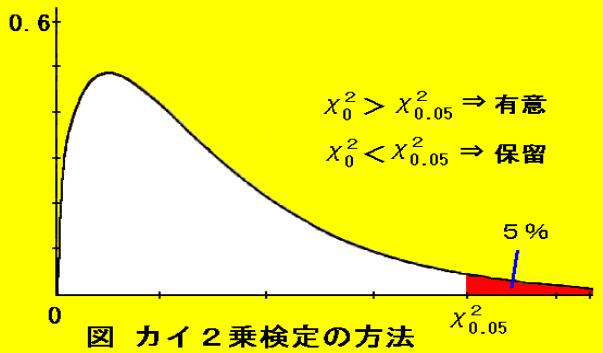


図　カイ２乗検定の方法

例題の計算:

$$\chi_Y^2 = \frac{30 \times (|11 \times 9 - 4 \times 6| - 30/2)^2}{15 \times 15 \times 17 \times 13}$$

$= 2.172$

⇒ $\chi_Y^2 < 3.84$　（p＝0.141）

【結論】
Ａ薬による自覚症状改善の程度が，プラセボより効果があるとはいえない。

Fisherの直接確率（正確な確率）

●２×２クロス表の検定のための有意確率を直接求める方法。
⇒　ｔ検定やカイ２乗検定などとは異なり，特定の統計量の分布を考えなくてよい。

●直接確率は，偶然に観測された４つのセルの度数が a，b，c，d となる確率，およびそれ以上に特定方向に偏った状態の確率を求めるための方法。

正確な確率の計算：

●計の欄の数値を固定した場合：
変数Ａが○，変数Ｂも○のセルの度数がａとなる場合を考える：
(1) 全標本数Nのうち，n_1 が変数Ａが○となる場合の数：

$$_N C_{n_1} = \frac{N!}{n_1! \times (N - n_1)!} = \frac{N!}{n_1! \times n_2!}$$

正確な確率の計算（その２）：

(2) n_1 のうち a が変数Ｂの○に，b が×となる組み合わせの数：
⇒　Ｂの○：m_1 人中 a 人
Ｂの×：m_2 人中 b 人
となる場合の数の積：

$$_{m_1} C_a \times _{m_2} C_b = \frac{m_1! \times m_2!}{a! \times b! \times c! \times d!}$$

正確な確率の計算（その３）：

(3) (1) と (2) から，４つのセルの度数が偶然に a, b, c, d となる確率 p_0：

$$p_0 = \frac{{}_{m_1}C_a \times {}_{m_2}C_b}{{}_{N}C_{n_1}} = \frac{n_1! \times n_2! \times m_1! \times m_2!}{N! \times a! \times b! \times c! \times d!}$$

正確な確率の計算（その４）：

検定のためには，表の状態を更に偏らせた場合の確率が必要：

表　ＡとＢの四分表

	B		
A	○	×	計
○	a＋i	b－i	n_1
×	c－i	d＋i	n_2
計	m_1	m_2	N

正確な確率の計算（その５）：

● $ad > bc$ とする：
偏った状態とは，表に示したように，積の大きい方はより大きく，小さい方はより小さくなるように，各セルに１ずつ度数を増減すること。

・ｂとｃのうち小さい方の値をｋ。
各セルの度数は０以上：
ｉのとり得る値は，０，１，．．．，ｋ
すべての４分表の確率の計算が必要

正確な確率の計算（その６）：

● ｉ＝１の場合の確率：p_1

$$p_1 = \frac{b \times c}{(a+1) \times (d+1)} \times p_0$$

● ｉ＝２の場合の確率：p_2

$$p_2 = \frac{(b-1) \times (c-1)}{(a+2) \times (d+2)} \times p_1$$

正確な確率の計算（その７）：

●一般にｉ＋１の場合の確率：p_{i+1}

$$p_{i+1} = \frac{(b-i) \times (c-i)}{(a+i+1) \times (d+i+1)} \times p_i$$

正確な確率の計算（その８）：

●すべての確率の合計を計算：p_F

$$p_F = p_0 + p_1 + p_2 + \cdots + p_k$$

【注意】
確率 p_F は，特定方向のみの偏りを問題としており，片側検定である。
通常，この確率を２倍した値を検定のための有意確率として用いる場合が多い。

四分表の検定はどの方法で？

【問題点】
４分表の検定には，フィッシャーの直接確率を用いるか，カイ２乗検定を用いるか？

できる限りフィッシャーの直接確率

標本数が大きい場合，階乗計算は困難。
適当な統計パッケージHALBAU/HALWINなどが必要（ちょっと宣伝）。
ホームページでも計算できる。

例題の計算：

①「計」の欄の数値が一定の場合に，表のようなクロス表が偶然できる確率を計算：

$$p_0 = \frac{15!\cdot 15!\cdot 17!\cdot 13!}{30!\cdot 11!\cdot 4!\cdot 6!\cdot 9!} = 0.057046$$

例題の計算（その２）：

表　最も偏りの大きい場合

	自覚症状		計
	改善	不変	
A　薬	15	0	15
プラセボ	2	13	15
計	17	13	30

・Ａ薬で「不変」のセルが３，２，１となる３つのクロス表の確率もすべて計算する。

例題の計算（その３）：

②Ａ薬で不変のセルの度数が３の場合：

$$p_1 = \frac{4\times 6}{12\times 10}\times p_0 = 0.011409$$

③度数が２の場合：

$$p_2 = \frac{3\times 5}{13\times 11}\times p_1 = 0.001197$$

例題の計算（その４）：

④度数が１の場合と０の場合：

$$p_3 = \frac{2\times 4}{14\times 12}\times p_2 = 0.00005699$$

および

$$p_4 = \frac{1\times 3}{15\times 13}\times p_3 = 0.000000088$$

例題の計算（その５）：

すべての確率の和p_F：

$$p_F = p_0+p_1+p_2+p_3+p_4 = 0.0697$$

$$\Rightarrow \quad 2p_F = 0.139$$

有意確率は13.9%⇒有意でない。

【結論】
Ａ薬の効果があるとはいえない。