7. 関係を調べるための方法

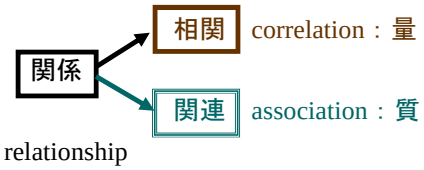
高木廣文
東邦大学看護学部

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

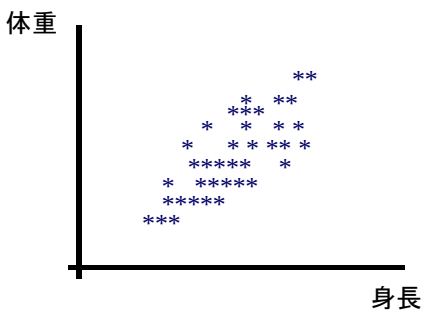
相関関係について

・２つの量的な変数間の関係の程度を示す。

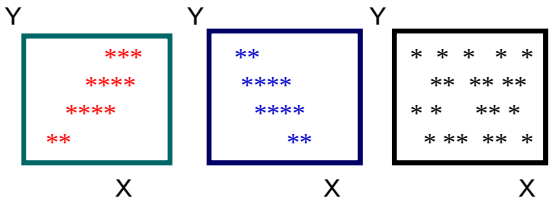
・相関図（散布図）
correlation diagram, scatter plots
２変数間の相関関係を図示。

・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

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

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：

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

10

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

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

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

12

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。

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

13

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

15

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

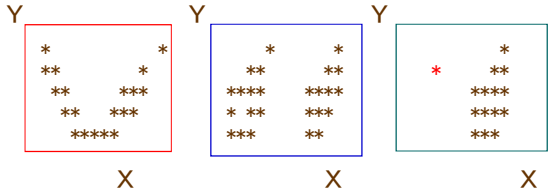
＝｛1×5.2+(−4)×(−2.8)+11×12.2
+(−1)×(−4.8)+(−7)×(−9.8)｝／5

＝224/5＝44.8

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

16

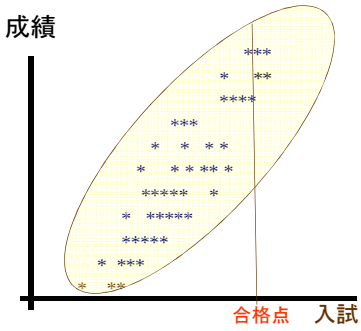
問題のある相関図



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）

17

打ち切りデータ



18

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。

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 」

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

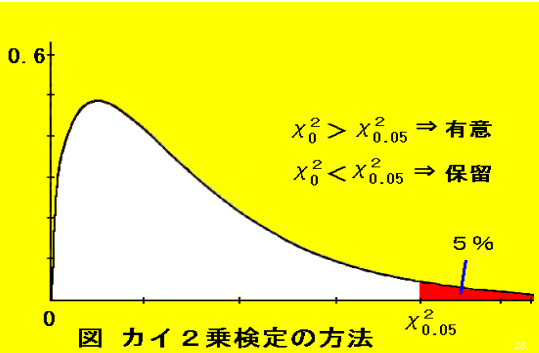


図　カイ２乗検定の方法

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。

８．交絡とバイアス

高木廣文
東邦大学看護学部
国際保健看護学研究室

交絡と効果の修飾について

ある要因と疾患の発生等の関係を調べる場合，他の要因の存在の影響により，正しくその要因のリスク等を評価できないことがある。

(1) 交絡 confounding
(2) 効果の修飾 effect modification

交絡とは

・特定の要因のある結果への影響を調べる場合，第３の要因が両者に影響を及ぼし，期待される要因の推定値が得られない場合。

・第３の要因：
交絡因子 confounding factor

交絡因子の条件

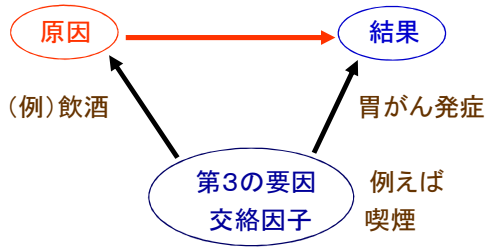
1) 交絡因子は問題となっている結果（疾患の発生，死亡等）のリスク因子である。
2) 問題となっている要因（原因）と関係がある。
3) 問題となっている要因と結果の因果連鎖の中間変数ではない。

因果関係のパスモデル

（例）飲酒　　胃がん発症

・原因と結果を矢印で結び，因果関係があることを示す（パス図，パスダイアグラム）

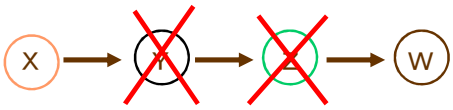
交絡のある因果関係



因果連鎖モデル

・一つの結果が，次の原因になり，鎖のように繋がっている関係

交絡因子はなぜ因果連鎖の中間変数ではないのか？



・変数ＹやＺがなければ，変数Ｘの影響は結果変数Wに及ばない！

効果の修飾とは何か？

・ある要因の結果への影響を示す効果の指標（リスク差やリスク比等）の数値を変化させるような第３の要因がある場合，「効果の修飾」があるという。

⇒第３の要因＝効果の修飾因子
effect-modifier

・交絡　⇒ 除去すべきもの
・効果の修飾 ⇒ 詳細な影響評価

交絡因子を除く方法は？

・交絡因子を除き，効果の修飾を見つける ための方法は？

・層別解析：
Mantel-Haeszel検定／推定
マッチング matching

・多変量解析：
条件付きロジスティックモデル
比例ハザードモデル

層別解析

・性，年齢，喫煙習慣などの交絡因子の影響を除き，評価したい要因の影響を正しく推定するための方法

交絡因子の各カテゴリ（層）ごとに集計し，各層内では交絡因子が均一になるようにする。

（例）性別，年齢別，喫煙習慣別など

表．Avadex と肺癌

層		投与	非投与	計
系Ｘ：雄	肺癌	4	5	9
	なし	12	74	86
	小計	16	79	95
系Ｘ：雌	肺癌	2	3	5
	なし	14	84	98
	小計	16	87	103
系Ｙ：雄	肺癌	4	10	14
	なし	14	80	94
	小計	18	90	108
系Ｙ：雌	肺癌	1	3	4
	なし	14	79	93
	小計	15	82	97

表. 層がない場合のAvadex と肺癌

	投与	非投与	計
肺癌	１１	２１	３２
なし	５４	３１７	３７１
計	６５	３３８	４０３

・リスク比＝(11/65) ／(21/338)＝2.724
95%信頼区間(1.315, 5.497)
・オッズ比＝11×317／(21×54) ＝3.075
95%信頼区間(1.304, 7.162)

Mantel-Haenszel 推定量について

・層別四分表について，各層に共通するオッズ比（共通オッズ比）を推定するための近似的な方法

・全体で層の個数はＫ個
・第ｋ層のオッズ比ψ_kは層によらず一定
（均一性の仮定）
・共通オッズ比ψを推定する

表. 第k層の四分表

	要因 あり	要因 なし	計
患者	a_k	b_k	m_k
対照	c_k	d_k	n_k
計	x_k	y_k	N_k

（ｋ＝1, 2, . . . , Ｋ）

共通オッズ比の推定

・Mantel-Haenszel 推定量 ψ_{MH}

$$\psi_{MH} = \frac{\sum_{k=1}^{K}(a_k d_k / N_k)}{\sum_{k=1}^{K}(b_k c_k / N_k)} = \frac{\sum_{k=1}^{K} R_k}{\sum_{k=1}^{K} S_k} = \frac{R_+}{S_+}$$

共通オッズ比の計算例-1

Avadexデータの共通オッズ比の計算を行う。

各層のオッズ比：
第1層：4×74／(5×12)＝ 4.933
第2層：2×84／(3×14)＝ 4.000
第3層：4×80／(10×14)＝ 2.286
第4層：1×79／(3×14)＝ 1.881

共通オッズ比の計算例-2

・R_+, S_+を求める。

R_+＝4×74/95＋2×84/103＋4×80/108
＋1×79/97＝8.524

S_+＝5×12/95＋3×14/103＋10×14/108
＋3×14/97＝2.769

共通オッズ比の推定値：

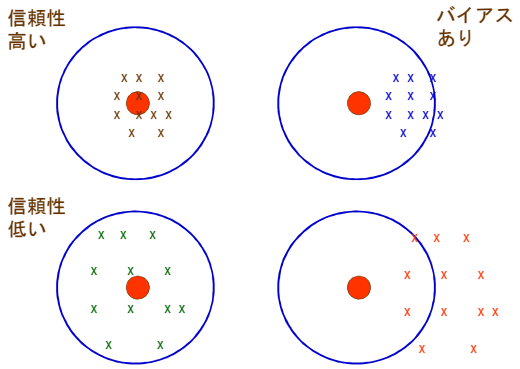
ψ_{MH}＝R_+／S_+＝8.524／2.769＝3.079

バイアスの問題

バイアス：
真実からの結果や影響の偏り，もしくはそのような偏りをもたらすプロセス

真実とは体系的に異なる結論を導くデータ収集，分析，解釈，出版，レビューにおけるあるゆる傾向

49



50

バイアスの分類

1. 情報バイアス，観察バイアス
2. 選択バイアス

そのほかにも多くのバイアスがある

51

1. 情報バイアス，観察バイアス

曝露や結果のデータの測定における欠陥。比較するグループ間の情報の質（正確性）が異なるという結果をもたらす。

52

なぜ情報バイアスが生じるのか？

1. リコール・バイアス（思いだしバイアス）
過去のイベントや経験の記憶を思い出すことの正確さや完全さでの差による，系統的エラー
（患者と健常者での病気に関する意識の差など）
2. 面接者によるバイアス
選択的にデータを集めてしまうという面接者の潜在意識や意識による系統的エラー
3. 観察者によるバイアス
観察者の変動（能力などのばらつき）による真実の値と実際に観察された値の間の系統的差

53

情報バイアスを避けるには

1. より客観的な情報を得る
2. 盲検（マスク化）の方法を用いる
3. より正確な情報を得るための方法について観察者の訓練をする

54

2. 選択バイアス

研究のために選ばれた人と選ばれなかった人の間での特性についての系統的差によるエラー

55

なぜ選択バイアスが生じるのか？

1. Bersonのバイアス
2. デザイン・バイアス
3. 健康労働者効果

56

1. Bersonのバイアス

選択バイアスの一つ。ケース・コントロール研究において病院のケースとコントロールが互いに系統的に異なってしまう。

研究上の曝露と疾病の組み合わせが，病院への管理のリスクを増加させる場合に，このバイアスが生じる。
病院のケースでの曝露率が病院のコントロールよりも系統的に大きくなり，これが順にオッズ比を系統的に歪めてしまう。

57

2. デザイン・バイアス

真の値と誤った研究デザインの結果として得られた値の間の差。

例には，制御ができていない研究がある。2つ以上のプロセスを分離できない（交絡のある）研究や，母集団の定義のあいまいな研究，不適当なコントロール群で実施された研究である。

58

3. 健康労働者効果

職業病の研究で初めに観察された現象。
労働者は通常，一般人口集団よりも，全ての死亡率で低値を示す。
これは，重大な病気や 慢性的な不具は，通常は雇用から除外されるからである。

59

選択バイアスを避けるには

1. 無作為標本抽出
2. 研究上の疾病以外には，ケースとコントロールを選ぶために，同一基準を用いる。

60