関係を調べるための方法

高木廣文
東邦大学看護学部
国際保健看護学研究室

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

関係 relationship
- 相関 correlation：量
- 関連 association：質

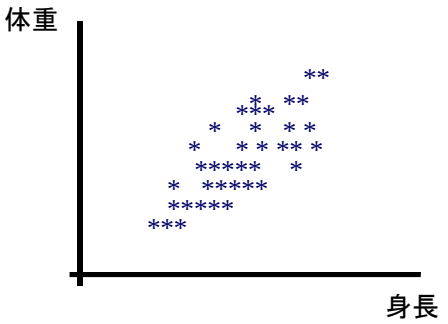
相関関係について

● ２つの量的な変数間の関係の程度を示す。

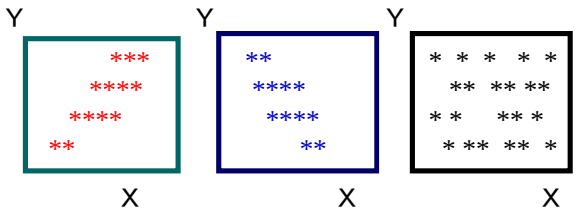
●相関図（散布図）
correlation diagram, scatter plots
２変数間の相関関係を図示。

●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） positive cor.
2) 負の相関（逆相関） negative cor.
3) 相関なし（無相関） null cor.

回帰 regression

回帰直線regression line：
Yの予測値＝[係数]×[Xの値]＋[定数]

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

回帰直線のパラメータ推定

回帰係数の推定：

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

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：

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

9

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

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

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

11

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。

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

12

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	60
165	50
180	65
168	70
162	55

平均値： 169　　60
分　散： 37.6　　50

共分散と相関係数の計算

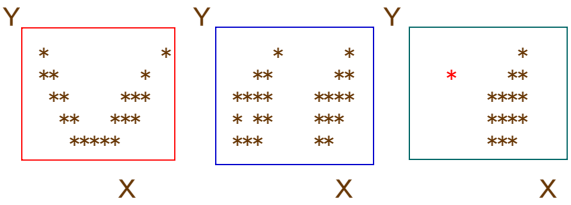
共分散＝[(170−169) × (60−60)＋(165−169) × (50−60)
＋(180−169) × (65−60)＋(168−169) × (70−60)
＋(162−169) × (55−60)**]**／5

＝ {1×0+(−4) × (−10)+11×5
＋(−1) × (10)＋(−7) × (−5)} ／5

＝120/5＝24

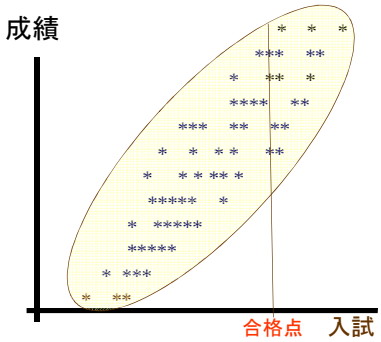
$$\text{単相関係数}=\frac{24}{\sqrt{37.6\times 50}}=0.554$$

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 truncated data



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	15	5	20
プラセボ	6	14	20
計	21	19	40

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：$\chi_Y^2 = 0$

四分表のカイ2乗検定の方法

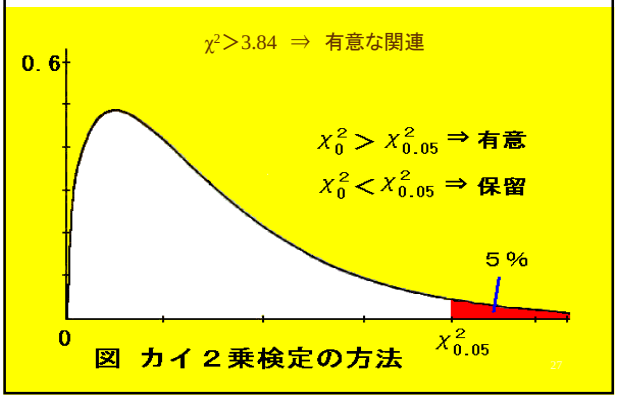


図　カイ２乗検定の方法

例題の計算：

$$\chi_Y^2 = \frac{40 \times (|15 \times 14 - 5 \times 6| - 40/2)^2}{20 \times 20 \times 21 \times 19} = 6.416$$

$$\chi_Y^2 > 3.84 \quad (p = 0.011)$$

【結論】

Ａ薬による自覚症状改善の程度は，プラセボより有意に効果がある。

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。

四分表の検定はどの方法で？

４分表の検定には，Fisher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などが必要。
ホームページでも計算できる。

Fisherの直接確率の計算

直接確率 p_F ＝0.010

有意確率は５％より小さい。

【結論】
Ａ薬には有意な効果がある。

31