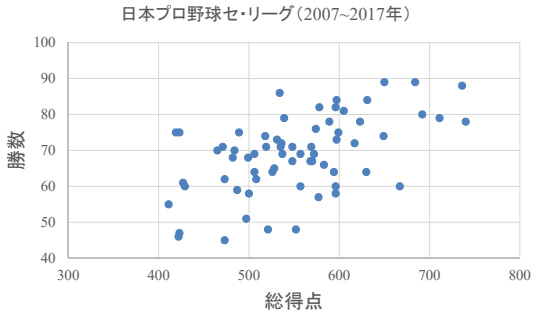
統計学入門 補助資料
～共分散と相関係数～

2022年度1学期:　月曜2限
担当教員:　石垣　司

2変数標本の可視化と定量化

- **散布図** 2変数の関係性を可視化
  - 観測値 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$



x_i: 総得点
y_i: 勝数

日本野球機構HPより
・2007~2015年144試合
・2016年以降143試合

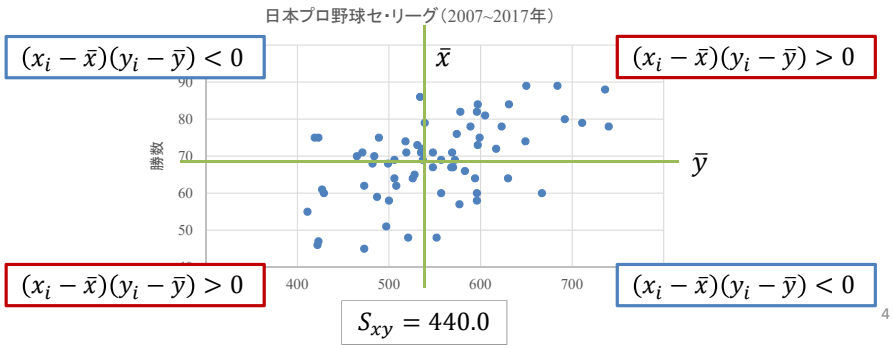
  - 総得点が多ければ勝数も多いという傾向
- 散布図の傾向を<u>共分散</u>や<u>相関係数</u>として定量化

標本共分散

- 2変数間の関連性の強さを表す代表値
  - 散布図の直線的な傾向（線形関係）の指標
- **標本共分散** $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - 値が大（正）⇒ x が大なら y も大の傾向
  - 値が小（負）⇒ x が大なら y が小の傾向
  - 値がゼロ付近 ⇒ x と y の関係性は小さい
- 問題
  - $S_{xy} = \frac{1}{n-1}(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y})$ を確かめなさい

標本共分散のイメージ

- 標本共分散 $S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
  - S_{xy} の値が正 ⇒ 右上がりの傾向
  - S_{xy} の値が負 ⇒ 右下がりの傾向
    - $(x_i - \bar{x})(y_i - \bar{y})$は各標本$(x_i, y_i)$に関して正or負の値をもつ
    - S_{xy}はその総和



標本相関係数

- 変数の単位に依存しない線形関係の代表値
  - 標本共分散を正規化
    - 標本共分散の値は標本の単位・スケールに依存するため異なるスケールの変数の比較に不向き
  - **標本相関係数**（ピアソンの積率相関係数）

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

    - S_{xx}は x の標本分散、S_{yy}は y の標本分散
  - 標本相関係数の値の範囲

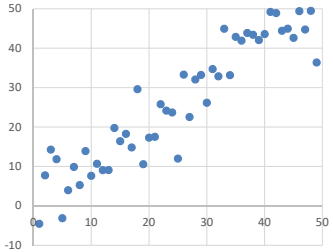
$$-1 \le r_{xy} \le 1$$ check!

Karl Pearson 1857-1936

相関係数の性質

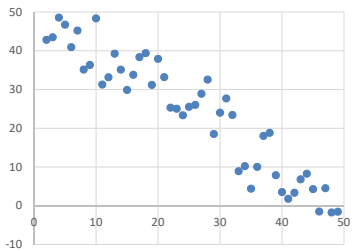
- 右上がりの関係 ⇒ 値が 1 に近い（正の相関）
- 右下がりの関係 ⇒ 値が -1 に近い（負の相関）
- 直線的な関係性が見出せない ⇒ 値が 0 に近い（無相関）

正の相関



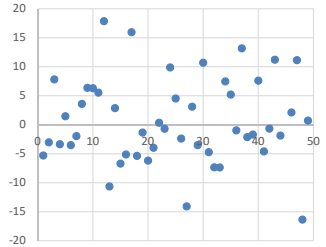
$r_{xy} = 0.92$

負の相関



$r_{xy} = -0.93$

無相関



$r_{xy} = -0.01$

相関係数の値の意味付け

- 社会調査における目安

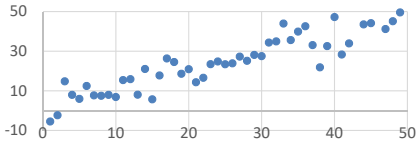
出典：岩永雅也、大塚雄作、高橋一男（編集）社会調査の基礎 (放送大学教材) 2003

  - 相関係数±0.7〜1.0 ⇒ かなり強い相関関係がある
  - 相関係数±0.5〜0.7 ⇒ 強い相関関係がある
  - 相関係数±0.4〜0.5 ⇒ 中程度の相関がある
  - 相関係数±0.3〜0.4 ⇒ ある程度の相関がある
  - 相関係数±0.2〜0.3 ⇒ 弱い相関関係がある
  - 相関係数±0.0〜0.2 ⇒ ほとんど相関関係がない

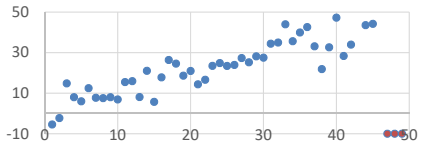
- 相関係数の利用は便利・簡便だが誤用も多いので，性質，特徴を正確に理解する必要あり

注意1〜外れ値に弱い

- 外れ値の影響で相関係数の値は大きく変化
  - 外れ値で相関係数の値が小さくなる例：

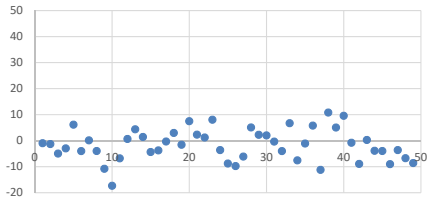


$r_{xy} = 0.92$

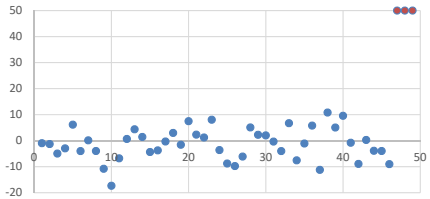


$r_{xy} = 0.49$

  - 外れ値で相関係数の値が大きくなる例：



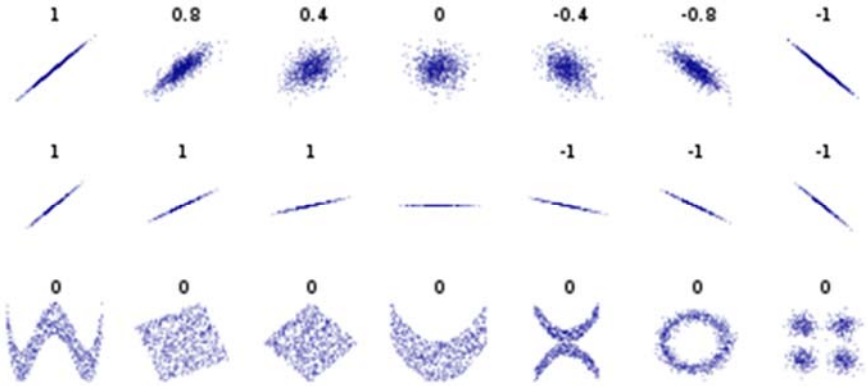
$r_{xy} = 0.00$



$r_{xy} = 0.41$

注意2～非線形関係

- 非線形関係(直線でない関係)の代表値には不適



"Pearson correlation coefficient" in Wikipedia. (public domain)

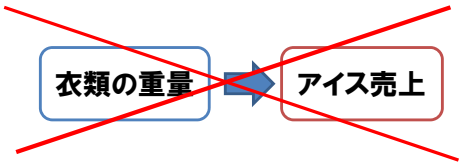
注意3～因果関係

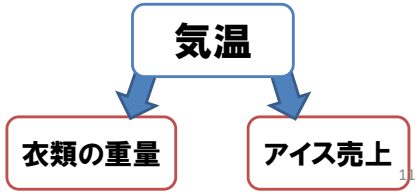
- 因果関係　原因があって結果が生じる関係
  - プロ野球の例
    - 勝数が大きいなら、総得点数も大きい傾向？
    - 総得点数が大きいなら、勝数も大きい傾向？
- 相関係数だけでは因果関係の向きは分からない
  - 野球というスポーツの特性上, 後者と判断
  - 国のGDPと学力は相関あり。因果の向きは？

変数 X（原因）→ 変数 Y（結果）　　変数 X（結果）← 変数 Y（原因）

注意4～疑似相関

- **疑似相関**　関係のない変数間に因果関係を推測
  - 変数 X :身に付けている衣類の重量
  - 変数 Y :アイスクリームの売上
    ⇒　変数 X が下がれば, 変数 Y は増大
- 本当の因果関係は変数 Z（気温）
  - 変数 Z が上がれば, 変数 X は下がる
  - 変数 Z が上がれば, 変数 Y は上がる



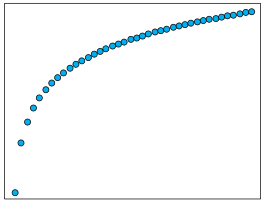


相関関係と因果関係

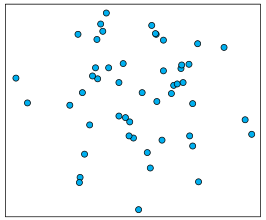
- 社会科学では疑似相関の要因を排除し, 因果関係を推測することが重要
  - 風が吹けば桶屋がもうかる、は本当か？
- 因果関係をとらえるためには・・・
  - 科学的知見や社会常識(多くの人の納得)の付与
  - 計量経済学の主眼
  - 論争例:(米国)白人は黒人よりIQが高い？
    AR Jensen　(1969) "How Much Can We Boost IQ and Scholastic Achievement?", Harvard Educational Review 39: 1–123 の結果と誤用に基づく社会論争。特に, 非白人系児童への知能テストの実施と特殊学級への配置の問題
- データのみから因果関係を理解することは, 特殊な場合を除き不可能

スピアマンの順位相関係数

- 順位尺度に基づく相関係数
  - 順位に意味のある場合に使用(相対評価のテストなど)
  - 単調増加(減少)の関係の尺度



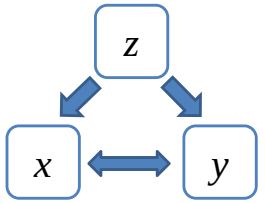
ピアソンの相関係数　$r_{xy} = 0.88$
スピアマンの相関係数　$r_{xy} = 1.00$



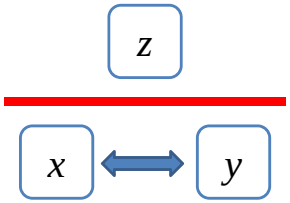
ピアソンの相関係数　$r_{xy} = -0.04$
スピアマンの相関係数　$r_{xy} = -0.03$

偏相関係数

- 3つの変数 (x, y, z) から1つの変数の影響を取り除いた, 残り2つの変数の相関係数
  - z が x と y の共通の説明要因であると, x と y の間に疑似相関が発生
- 回帰(←次回の授業で説明する)により変数 z の影響を x と y からそれぞれ除去し, その残差間の相関



相関係数



偏相関係数

偏相関係数～計算方法

- 標本偏相関係数の定義
  - 回帰により z の影響を除去した x の残差

$$u_i = x_i - (\hat{c} + \hat{d} z_i)$$

  - 回帰により z の影響を除去した y の残差

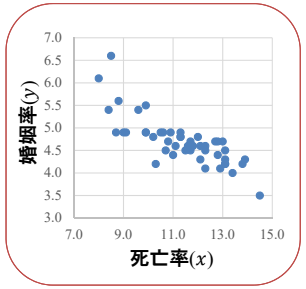
$$v_i = y_i - (\hat{e} + \hat{f} z_i)$$

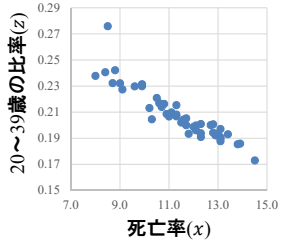
  - z の影響を除去した x と y の標本偏相関係数

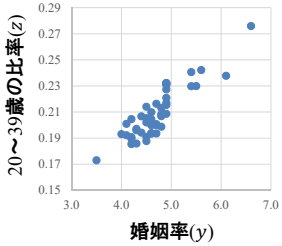
$$r_{xy,z} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2 \sum_{i=1}^{n} v_i^2}} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

偏相関係数～人口動態の相関係数

- 各都道府県の人口動態データ(2015年)
  - 1000人当たりの死亡率(x), 婚姻率(y)
    ⇒ 相関係数　$r_{xy} = -0.80$ (その解釈は？)
  - 20～39歳の比率(z)
    ⇒ 相関係数　$r_{xz} = -0.92, r_{yz} = 0.88$



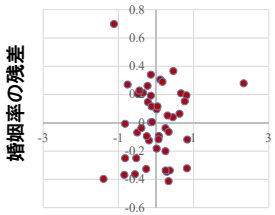




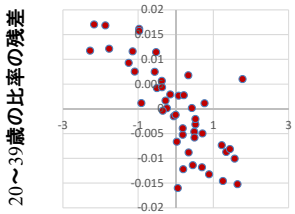
総務省統計局「2-15都道府県別出生・死亡数と婚姻・離婚件数(平成27年)」,「第2表都道府県、年齢(5歳階級)総人口(平成27年10月1日)」

偏相関係数～人口動態の偏相関係数

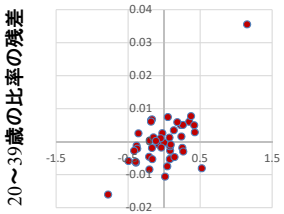
- **偏相関係数**
  - **偏相関係数** $r_{xy,z} = 0.09$
    - 20～39**歳の比率**(z)**が同じときの死亡率**(x), **婚姻率**(y)**の相関係数**
  - **偏相関係数** $r_{yz,x} = 0.62, r_{xz,y} = -0.77$



z の影響を除去した残差



y の影響を除去した残差



※ただし、右上(東京都)と左下(奈良県)のデータ2点を除去の場合 $r_{yz,x} = 0.24$

x の影響を除去した残差

演習問題

- **問題　次の文章内の結論と, その結論を導く過程には統計的な誤用が複数存在する。その誤用を2つ以上答えなさい**
  - **薬AとBの投入量と血糖値の低下値について適切な実験を行い, その相関係数を得た**
  - **薬A投入量と血糖値低下の相関係数**　0.11（**症例数**75**件**）
  - **薬B投入量と血糖値低下の相関係数**　0.14（**症例数**4**件**）
  - **結論：薬Bの方が治療効果が高い**