

統計学入門

～共分散と相関係数～

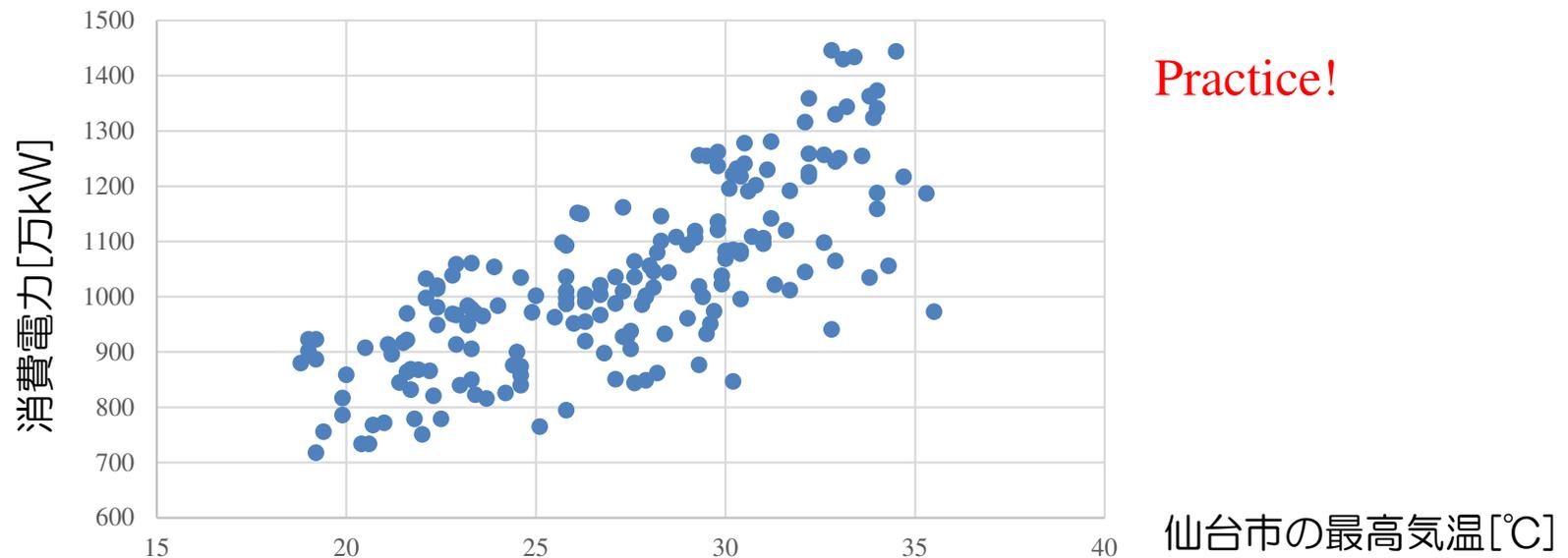
2025年度1学期: 月曜2限

担当教員: 石垣 司

2変数標本の可視化と定量化

散布図：2変数の分布を可視化した図

- 例：変数 x (最高気温)と y (消費電力)のプロット
- 観測値 $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$



散布図の傾向を共分散や相関係数として定量化

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

- 値が正 $\Rightarrow x$ が大なら y も大の傾向
- 値が負 $\Rightarrow x$ が大なら y が小の傾向
- 値がゼロ付近 $\Rightarrow x$ と y の関係性は小さい

問題

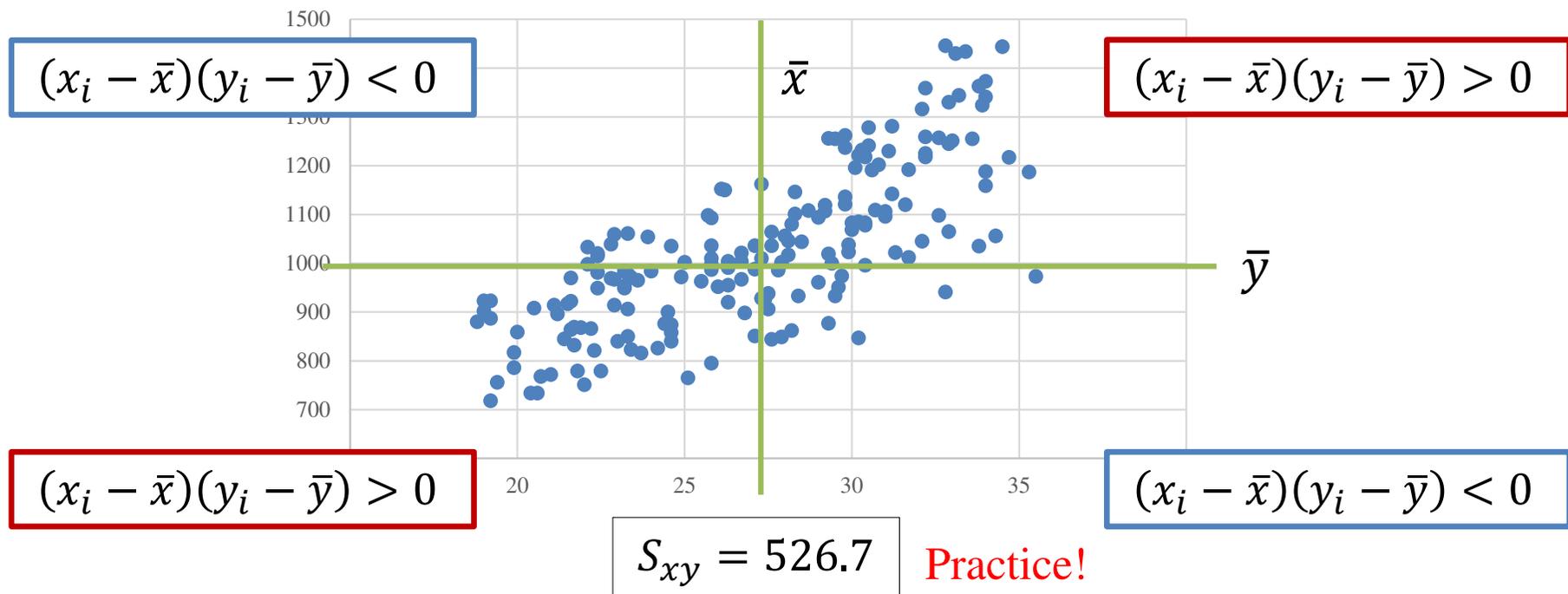
- $S_{xy} = \frac{1}{n-1} (\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y})$ を確かめなさい

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 \Rightarrow 右上がりの傾向
- S_{xy} の値が負 \Rightarrow 右下がりの傾向

$(x_i - \bar{x})(y_i - \bar{y})$ は各標本 (x_i, y_i) に関して正か負の値をもち、 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

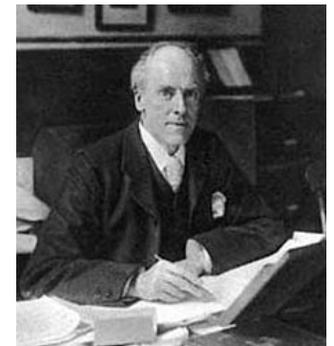
– 標本相関係数(ピアソンの積率相関係数)

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

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

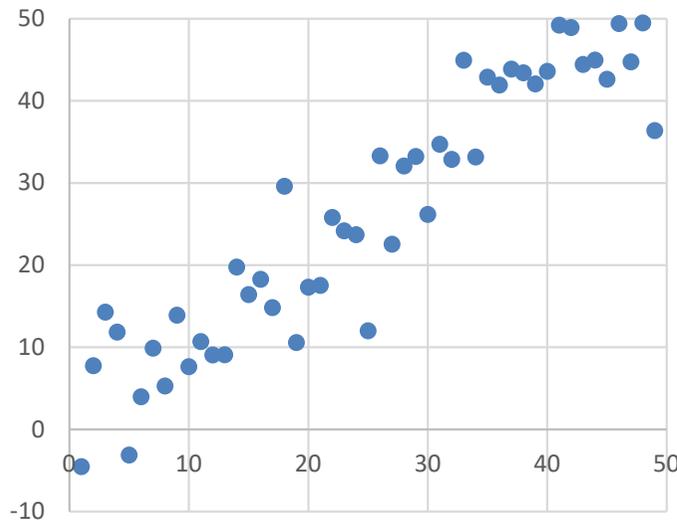
$$-1 \leq r_{xy} \leq 1 \quad \text{check!}$$



相関係数の性質

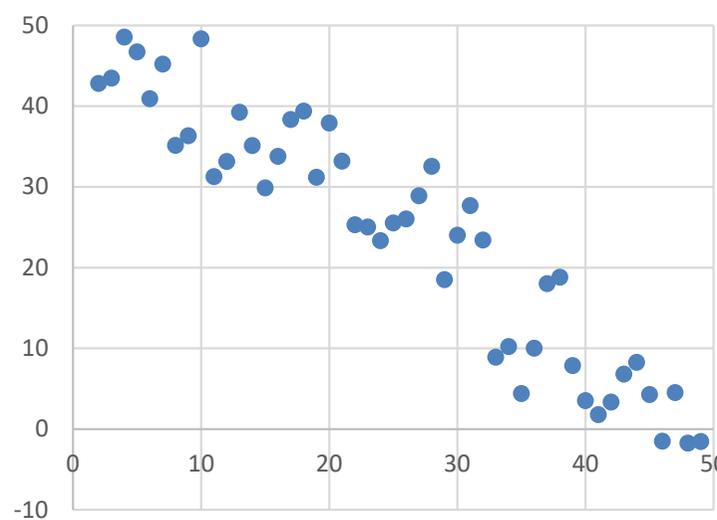
- 直線的な右上がりの関係 \Rightarrow 値が 1 に近い(正の相関)
- 直線的な右下がりの関係 \Rightarrow 値が -1 に近い(負の相関)
- 直線的な関係性が見出せない \Rightarrow 値が 0 に近い(無相関)

正の相関



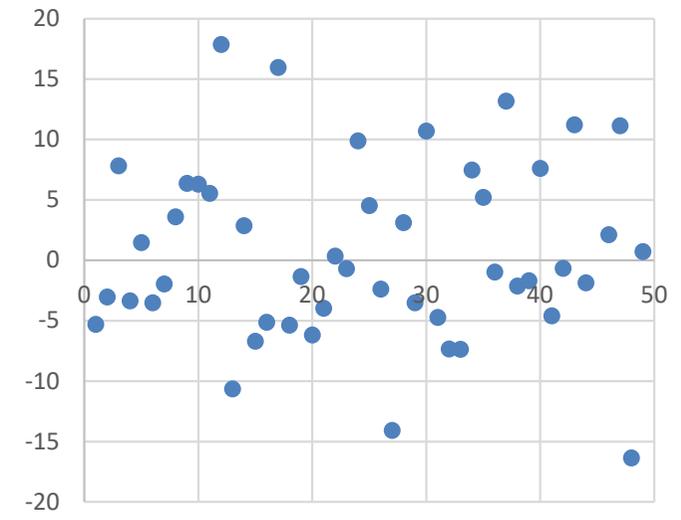
$$r_{xy} = 0.92$$

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

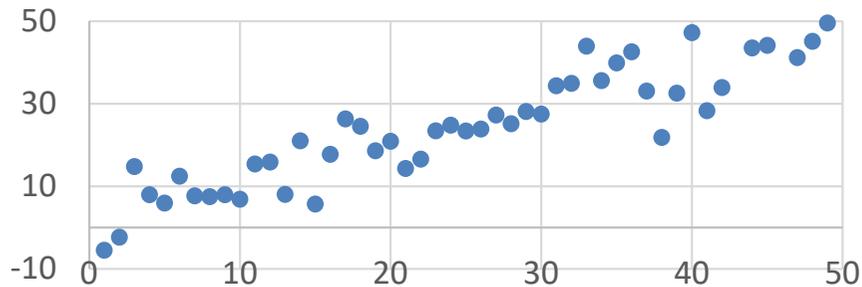
- 相関係数 $\pm 0.7 \sim 1.0$ \Rightarrow かなり強い相関関係がある
- 相関係数 $\pm 0.5 \sim 0.7$ \Rightarrow 強い相関関係がある
- 相関係数 $\pm 0.4 \sim 0.5$ \Rightarrow 中程度の相関がある
- 相関係数 $\pm 0.3 \sim 0.4$ \Rightarrow ある程度の相関がある
- 相関係数 $\pm 0.2 \sim 0.3$ \Rightarrow 弱い相関関係がある
- 相関係数 $\pm 0.0 \sim 0.2$ \Rightarrow ほとんど相関関係がない

相関係数の利用は便利・簡便。しかし、誤用も多いので、性質や特徴を正確に理解する必要あり

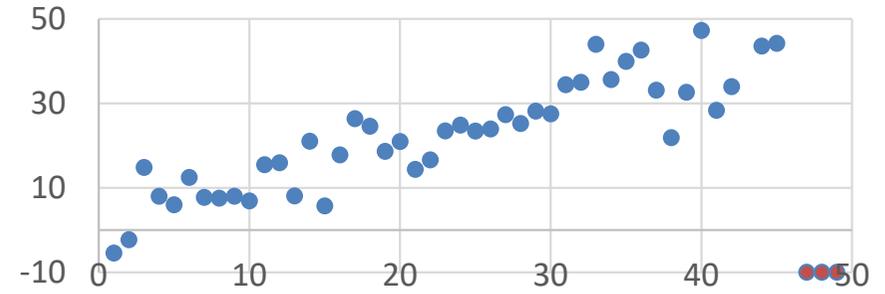
注意1 外れ値に弱い

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:

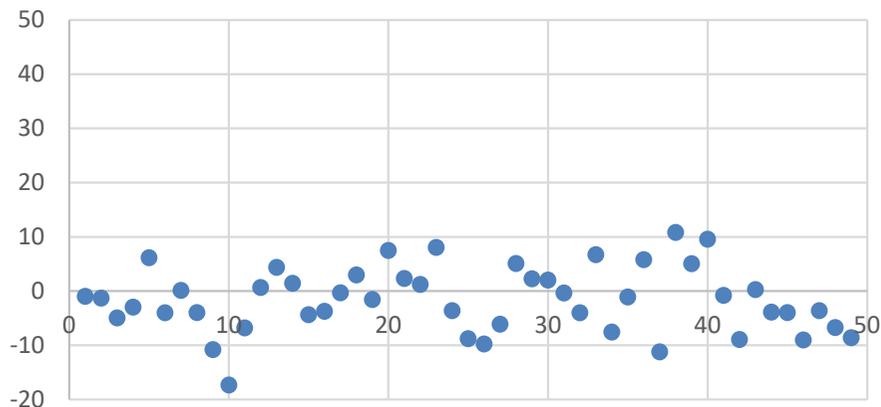


$$r_{xy} = 0.92$$

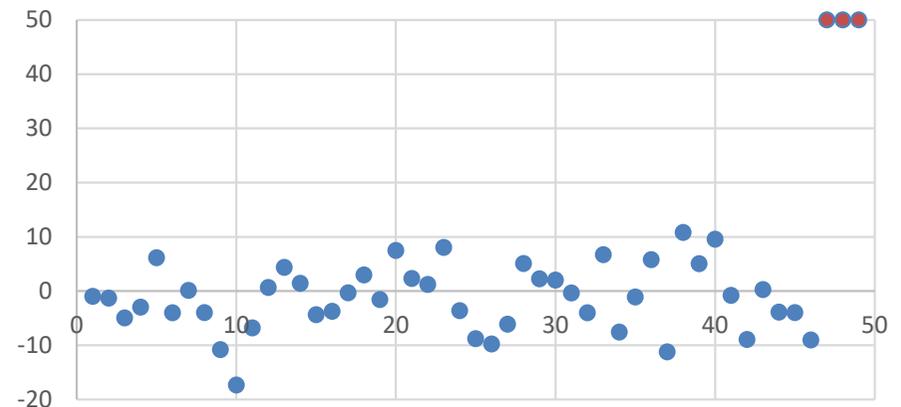


$$r_{xy} = 0.49$$

– 外れ値で相関係数の値が大きくなる例:



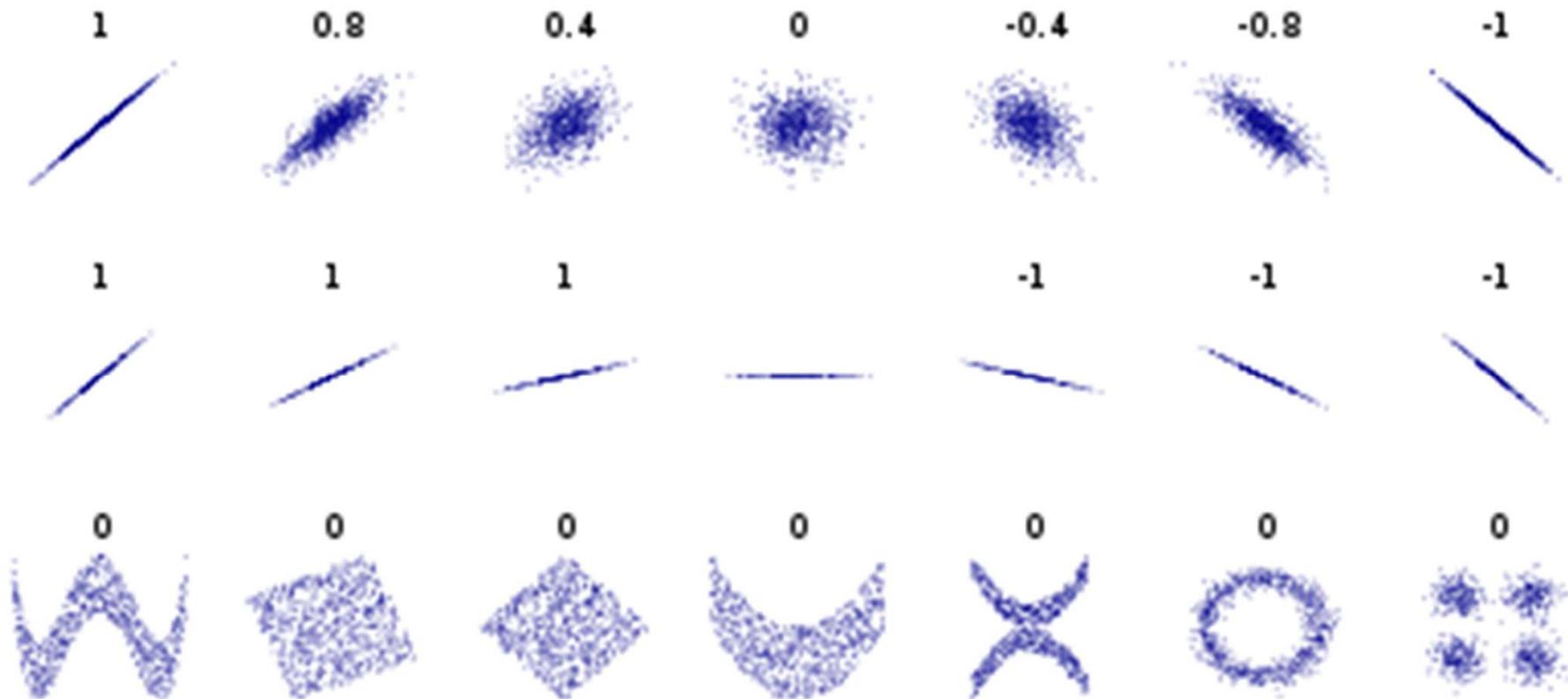
$$r_{xy} = 0.00$$



$$r_{xy} = 0.41$$

注意2 非線形関係

非線形関係(直線でない関係)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

– 最高気温と消費電力の例

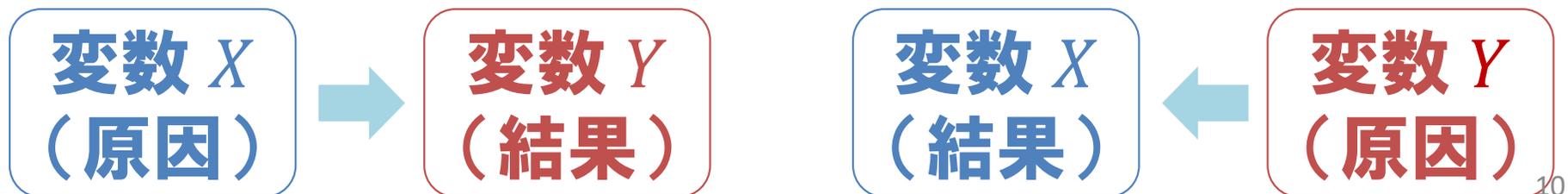
最高気温が大ならば、消費電力も大きい傾向？

消費電力が大ならば、最高気温も大きい傾向？

相関係数だけでは因果関係の向きは分からない

– 我々は、“夏の最高気温が高い時には冷房を利用するため消費電力が大きくなる”，というメカニズムを知っているために前者と判断できる

– 例：国のGDPと学力は相関あり。因果の向きは？



相関関係と因果関係

社会科学やビジネス環境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・ビジネス上の知見や社会的合意の付与
- 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**

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

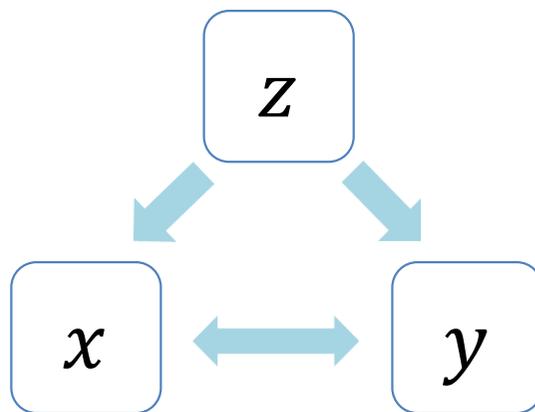


偏相関係数

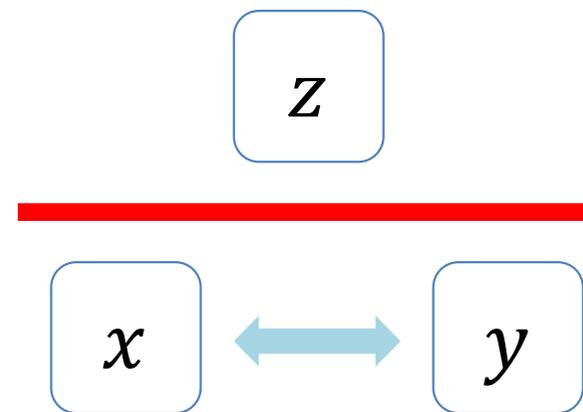
3つの変数 (x, y, z) から1つの変数の影響を取り除いた、残り2つの変数の相関係数

- 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

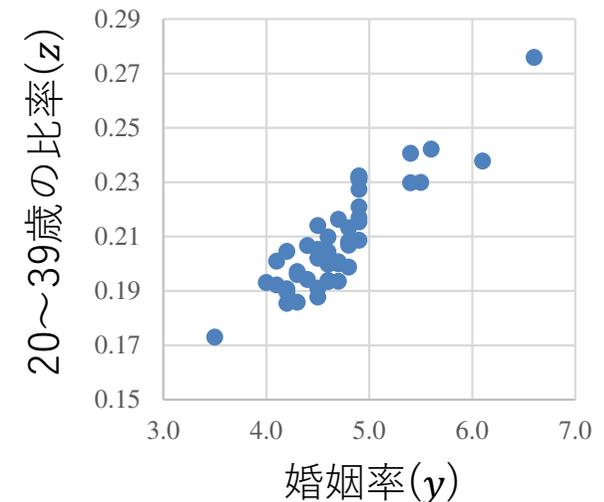
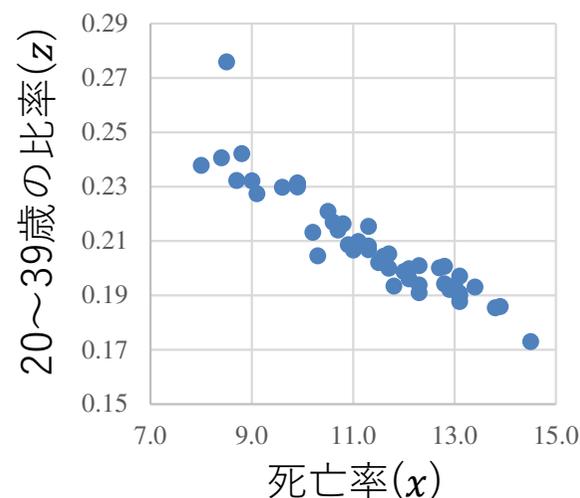
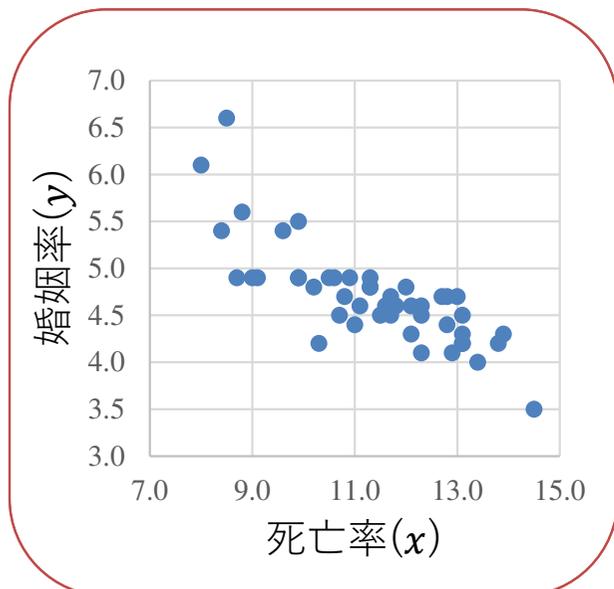


偏相関係数

偏相関係数～人口動態の相関係数

各都道府県の人口動態データ(2015年)

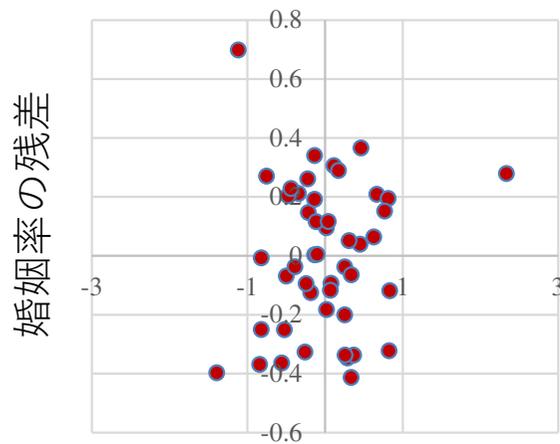
- 1000人当たりの死亡率(x), 婚姻率(y)
⇒ 相関係数 $r_{xy} = -0.80$ (その解釈は?)
- 20～39歳の比率(z)
⇒ 相関係数 $r_{xz} = -0.92, r_{yz} = 0.88$



偏相関係数～人口動態の偏相関係数

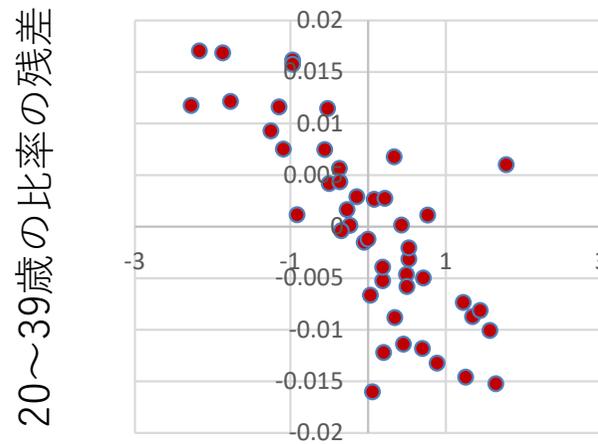
偏相関係数(計算方法は次回)

- 偏相関係数 $r_{xy,z} = 0.09$ **Practice!**
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 $r_{yz,x} = 0.62, r_{xz,y} = -0.77$



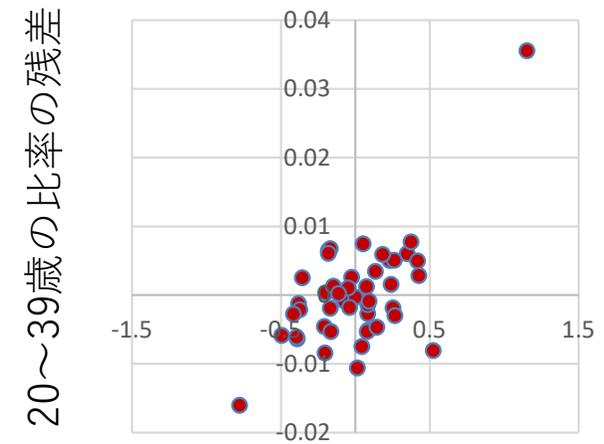
死亡率の残差

z の影響を除去した残差



死亡率の残差

y の影響を除去した残差



婚姻率の残差

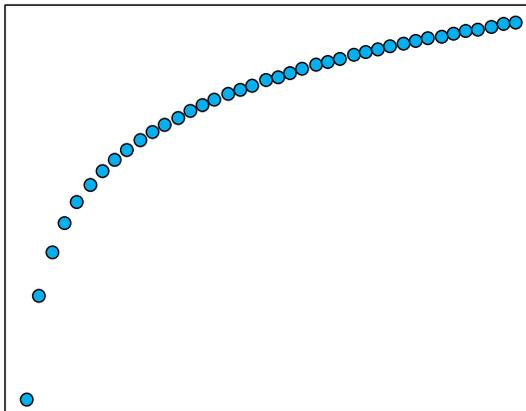
※ただし、右上(東京都)と左下(奈良県)のデータ2点を除去の場合 $r_{yz,x} = 0.24$

x の影響を除去した残差

補足：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

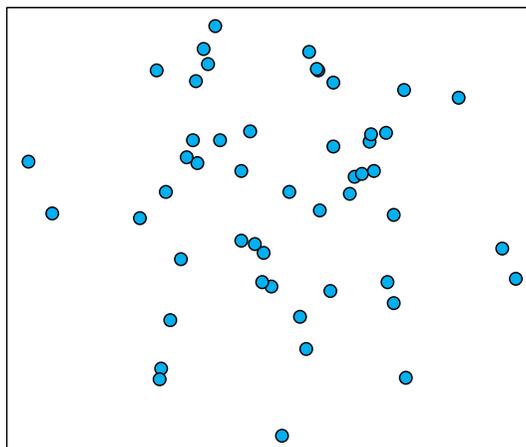


ピアソンの相関係数

$$r_{xy} = 0.88$$

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

$$r_{xy} = -0.03$$

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**