https://bit.ly/2OWCVNI

KYOTO UNIVERSITY

統計的モデリング基礎②
～確率分布・2変量間の関係・相関係数～

鹿島久嗣
（情報学科 計算機科学コース）

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

参考書

実証分析のための計量経済学
–正しい手法と結果の読み方
／山本 勲（2015）

A5判/260頁
ISBN 978-4-502-16811-6

経済学対象にしているが、回帰分析を中心に、因果分析等についてもカバーしている

（かなり細かいところまで書いてある）参考書

現代数理統計学の基礎

[久保川 達也](#) 著・[新井 仁之](#)・[小林 俊行](#)・[斎藤 毅](#)・[吉田 朋広](#) 編

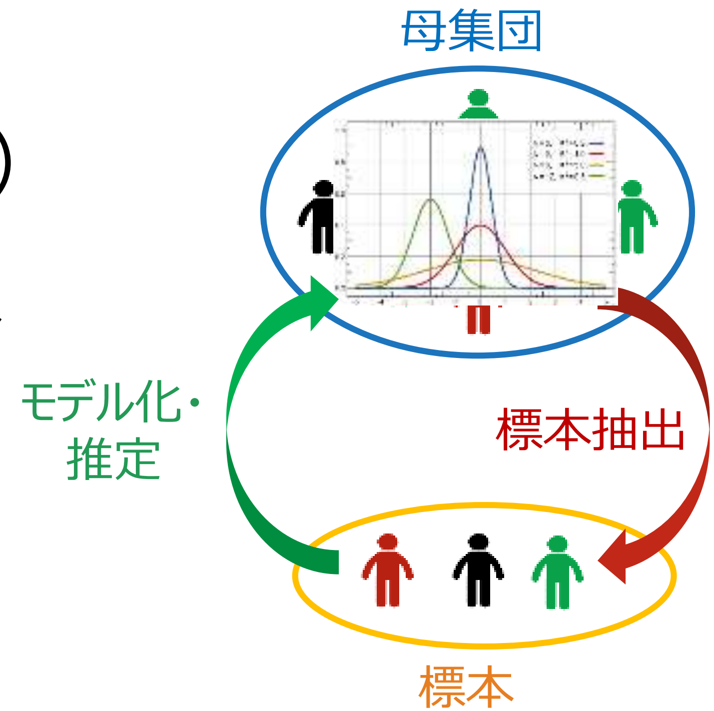
シリーズ名	共立講座 数学の魅力　全14巻+別巻1 【11】巻
ISBN	978-4-320-11166-0
判型	A5
ページ数	328ページ
発売日	2017年04月11日
本体価格	3,200円

理論的な背景含め基礎的事項がきちんと説明されている

統計的モデリングの考えかた

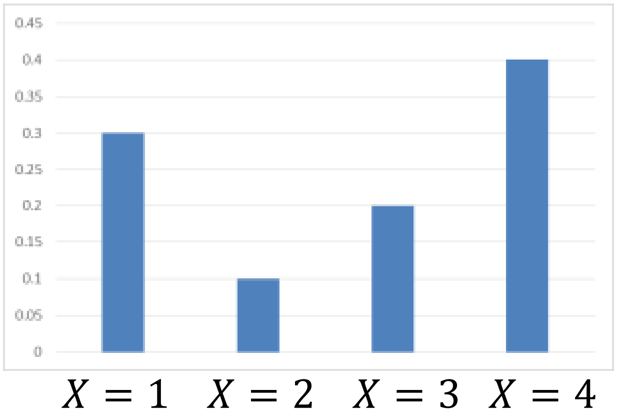
統計モデリングの考え方：部分から全体について知る

- 母集団：
  - 興味のある集合のすべての要素
  - 確率分布
    （分布のクラスやパラメータで指定される）
- 標本：母集団からの無作為抽出あるいは確率分布に従った抽出
  - 確率変数：確率的に値が決まる変数
- 標本から母集団について推測する
  （標本抽出の逆）



離散型確率変数の代表的な確率分布：離散分布、ベルヌーイ分布と2項分布

- 離散分布 $P(X=k)=f(k)$（但し $\sum_{k\in\mathcal{X}} f(k)=1,\ f(k)\geq 0$）
- ベルヌーイ分布：$\mathcal{X}=\{0,1\}$上の離散分布
- 二項分布
  - ベルヌーイ試行： 1が出る確率pの
    ベルヌーイ分布からn回 独立に抽出する
  - 二項分布：ベルヌーイ試行において1がk回出る確率を与える



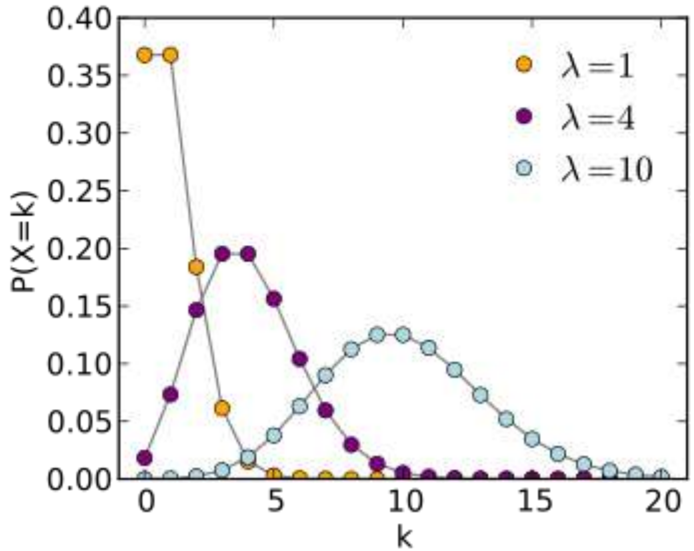
$$P(X=k\mid p)=\binom{n}{k}p^k(1-p)^{n-k}$$

n回の試行中のどこでk回の
1が現れるかの場合の数

- モデルパラメータpによって
  分布の形が一意に決定される

離散型確率変数の代表的な確率分布：ポアソン分布（2項分布の極限）、その他

- ポアソン分布：$P(X=k \mid \lambda)=\dfrac{\lambda^k}{k!}\exp(-\lambda)$
  - 比較的稀な事象が何回起こるかを表現
    - 1分あたりのWebサーバアクセス数
    - ロットあたりの不良品数
  - パラメータ$\lambda>0$
    - 2項分布のパラメータ（n, p）がない
    - 2項分布で$np=\lambda$として、$n\to\infty, p\to 0$とするとポアソン分布になる



https://en.wikipedia.org/wiki/Poisson_distribution#/media/File:Poisson_pmf.svg

- ほか、離散型の確率分布には幾何分布、負の2項分布などがある

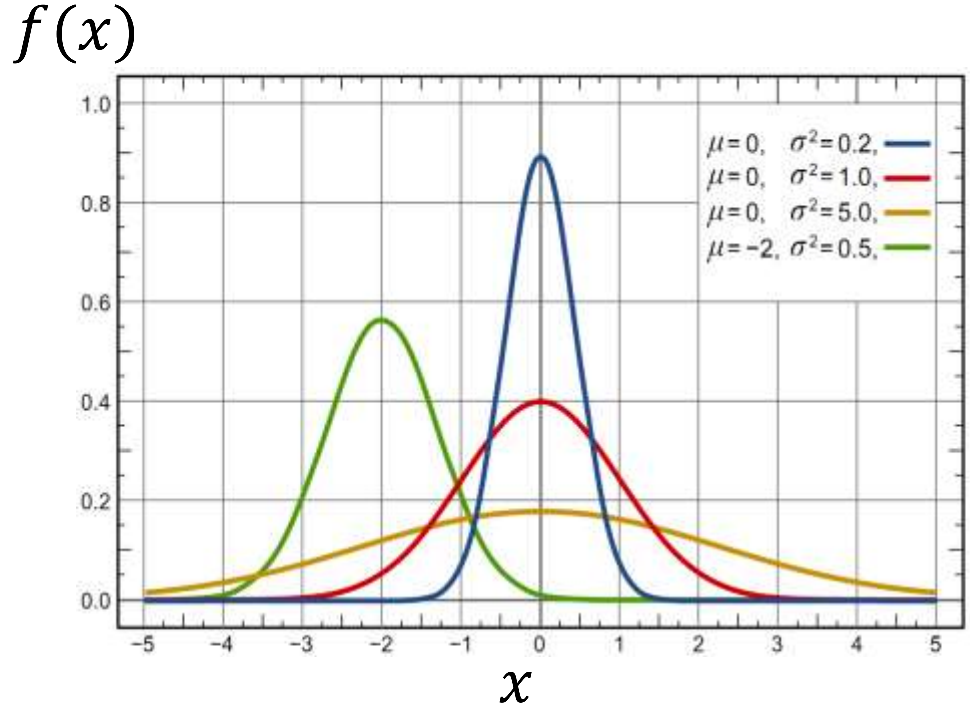
連続型確率変数の代表的な確率分布：確率密度関数で指定される

- 連続分布は確率密度関数$f(x)$で指定される
  - 確率 ＝ 確率密度の積分
    $[a, b]$内の値をとる確率：$P(a \leq X \leq b) = \int_a^b f(x)dx$
  - 連続変数がある特定の値をとる確率：$P(X = a) = 0$
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
- 一様分布：閉区間$[a, b]$上の一様分布は

$$f(x) = \begin{cases} \dfrac{1}{b-a} & (a \leq x \leq b) \\ 0 & (\text{その他}) \end{cases}$$

連続型確率変数の代表的な確率分布：正規分布

- 正規分布：$f(x) = N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - パラメータ：平均μと分散σ^2

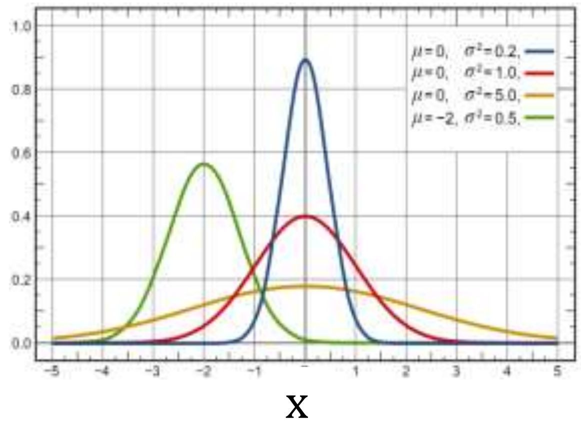


- 他、t分布、カイ2乗分布、ガンマ分布、ベータ分布、指数分布など

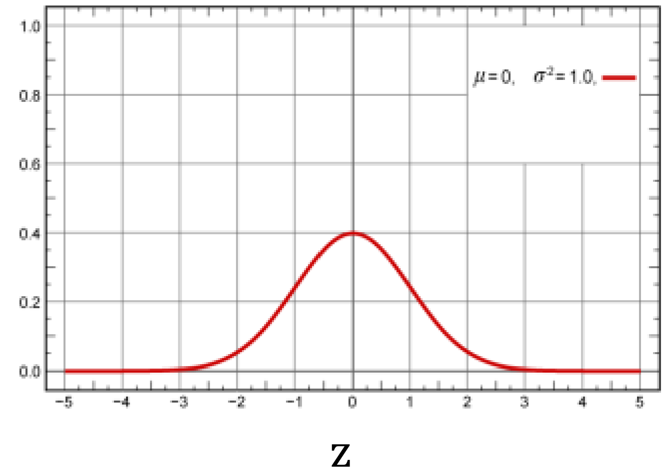
連続型確率変数の代表的な確率分布：標準正規分布

- $N(\mu, \sigma^2)$に従う確率変数Xを変数変換：$Z = \frac{X-\mu}{\sigma}$
- Zは平均0、標準偏差1の正規分布 $N(0,1)$に従う

確率密度関数: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ ➡ $f_Z(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$



標準化



確率分布の特性値：期待値は確率分布の代表値

- 確率変数Xの関数$g(X)$の期待値：確率での重みづけ平均

$$E[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x & (\text{連続型確率変数}) \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) & (\text{離散型確率変数}) \end{cases}$$

- さまざまな関数$g(X)$に対する期待値によって分布の特性を捉える
- 性質：
  - 線形性：$E[ag_1(X) + bg_2(X)] = aE[g_1(X)] + bE[g_2(X)]$
  - イェンセンの不等式：$E[g(X)] \geq g(E[X])$ (ただしgは凸関数)

さまざまな期待値：平均と分散

- 平均 $\mu = E[X]$：Xの期待値　（分布の“真ん中”）
- 分散 $\sigma^2 = \mathrm{Var}(X) = E[(X-\mu)^2]$：
  平均からの二乗偏差の期待値　（分布の“幅”）
  - $\mathrm{Var}(X) = E[X^2] - E[X]^2$
  - 標準偏差σ：分散の正の平方根
    - 正規分布なら$\mu \pm \sigma$：68%，$\pm 2\sigma$：95%，$\pm 3\sigma$：99.7%
- より一般的にはモーメント$E\left[X^k\right]$
- 例：厳密なサイコロ $P(X = i) = \frac{1}{6}$ の平均、分散を求めよ

平均の推定量：
標本平均

- 標本（部分）から平均（全体の性質）を知りたい
  - 標本 $S = \{x_1, s_2, \ldots, x_n\}$
- 標本平均：$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ を平均 $\mu = E[X]$の推定値として使う

推定量としての標本平均の好ましさ：標本平均は一致性と不偏性をもつ

- 標本平均は平均の推定値として好ましいか？
- 不偏性 $E_S[\bar{X}] = \mu$：標本平均の期待値は母集団の平均に一致する
  - E_Sは標本についての期待値　（何度も標本をとり直して、何度も標本平均を求めたときの、それらの平均）
- 一致性：標本サイズが大きくなるほど母集団の平均μに近づく
  - 標本平均の分散 $\mathrm{Var}_S[\bar{X}] = \dfrac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0$　　（対数の法則）

推定量としての標本平均の好ましさ：標本平均はBLUE（最良な線形不偏推定量）

- 効率性：推定値の分散が小さいこと
  - 標本平均$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$の代わりに最初の値を使う$\tilde{x} = x_1$とする
  - 標本平均のほうが効率的
    - 標本平均の分散$\frac{\sigma^2}{n}$ <最初の値の分散σ^2
- BLUE（最良な線形不偏推定量）：加重平均で表されるすべての不偏推定量のなかで、最も分散が小さい（効率的）なもの
  - 加重平均による推定量 $\hat{x} = \frac{1}{n}\sum_{i=1}^{n} a_i x_i$

分散の推定量：不偏分散

- 標本分散：$\frac{\left(x^{(1)}-\bar{x}\right)^2+\cdots+\left(x^{(n)}-\bar{x}\right)^2}{n}=\frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 不偏性をもたない： $E_S\left[\frac{1}{n}\sum_{i=1}^{N}\left(X^{(i)}-\bar{X}\right)^2\right]=\frac{n-1}{n}\sigma^2$
- 不偏分散：$\frac{1}{\boldsymbol{n-1}}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 不偏性をもつ：期待値が母集団の分散に一致する
- どちらも一致性はもつ：
  - 標本サイズが大きくなるほど母集団の分散に近づく
  - nが大きいところではnも$n-1$も大した違いはない

2変量データの解析

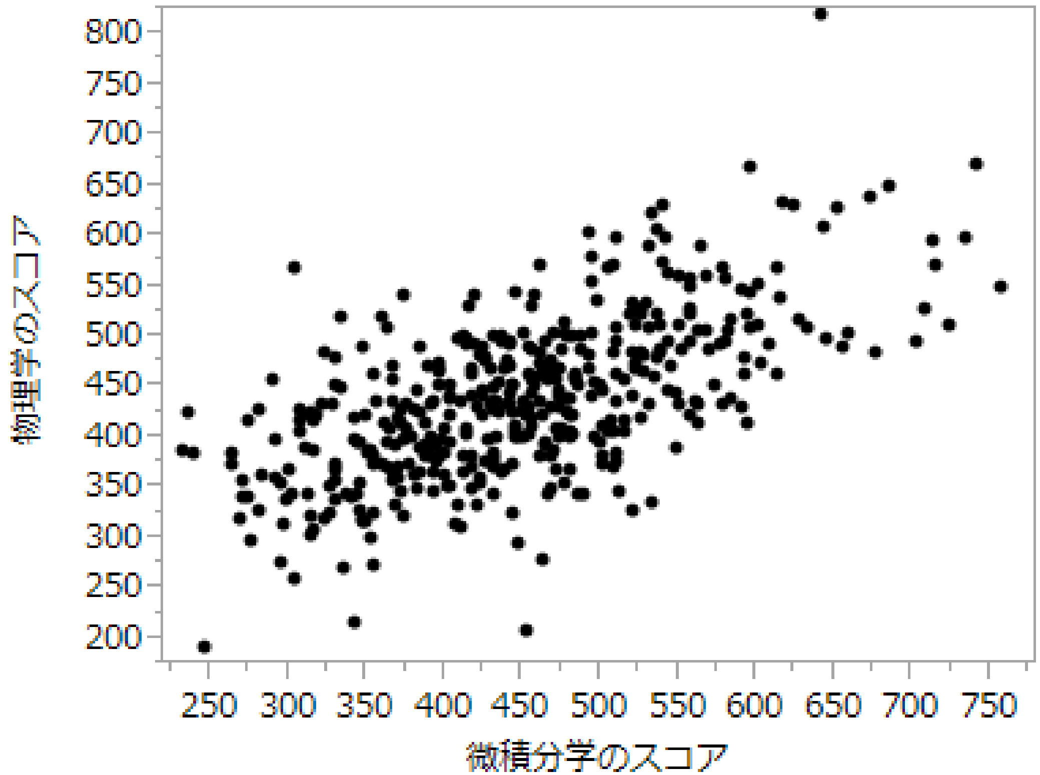
2つ以上の変量のデータ分析：変量間の関係を調べることでより深い分析が可能

- 前回は、1変数の単純分析について考えた
- 2つ（もしくはもっと多く）の変数の関係に興味があることが多い
- ２変量（あるいはさらに多く）の間の関係を調べることで、より積極的なデータ利活用が可能になる
  – ある属性をもった人は、ある商品を買いやすいのか？
  – ある薬を飲むと、ある病気に効果があるのか？
- 変数の種類によって、さまざまな分析手法がある
  – 量的変数：散布図、相関、回帰
  – 質的変数：クロス表、リスク差・比、オッズ比

2変量の単純な分析：散布図による視覚化

- 例：微積分の点数と物理の点数の関係

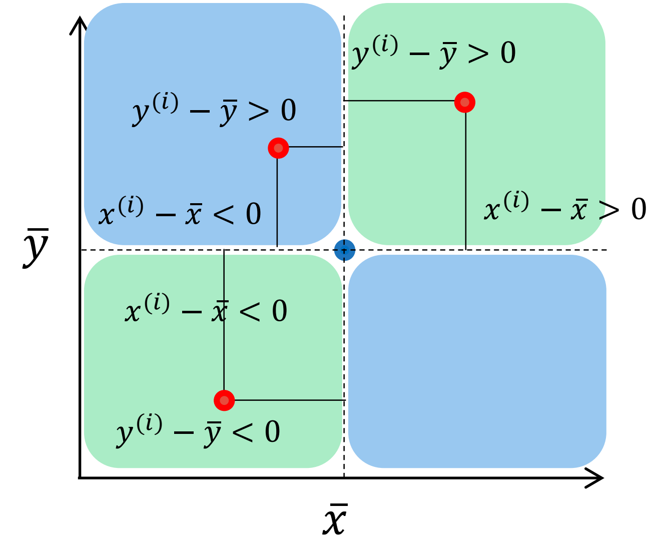
	微積分学のスコア	物理学のスコア
1	441.4	470.7
2	632.16	508.44
3	361.56	412.75
4	479.39	425.47
5	476.32	408.27
6	446.92	400.99
7	394.2	390.62
8	645.76	496.97
9	329.75	367.39
10	496.07	453.41
11	487.91	498.97
12	403.82	441.7
13	480.21	400.41
14	460.33	460.72
15	303.72	259.66
16	463.01	278.04
17	428.98	396.21
18	412.12	380.53
19	366.84	355.72



JMPサンプルデータ

2変数間の関係の指標：共分散と相関

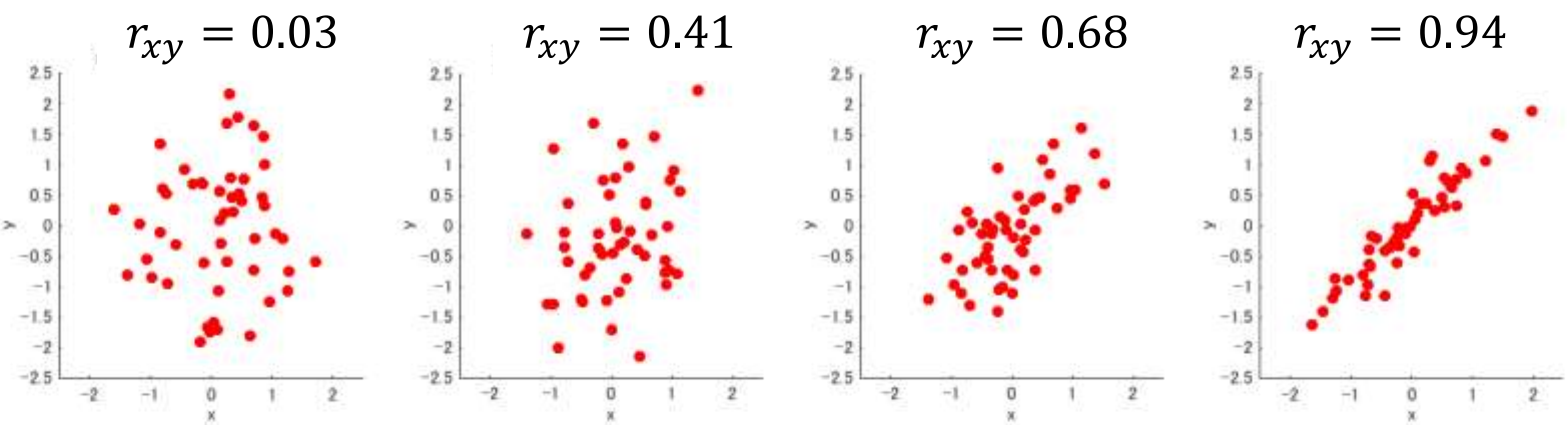
- 一方が増えたときに他方が増える（減る）関係性を表す指標
- 共分散(covariance)：$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$
  - ただし、$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$， $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
  - 偏差積の平均 (データのバラツキを表現)
    - 偏差 $(x^{(i)} - \bar{x})$ と偏差 $(y^{(i)} - \bar{y})$ の符号が一致する（緑領域）なら正の値をとる
    - 偏差 $(x^{(i)} - \bar{x})$ と偏差 $(y^{(i)} - \bar{y})$ の符号が不一致である（青領域）なら負の値をとる



- ただし、x, y の単位やスケールに影響されるため共分散の絶対的な大きさのみでは関係の強さを評価できない

2変数間の関係の指標：共分散と相関

- 相関 (correlation)：$r_{xy}=\dfrac{\sum_{i=1}^{n}(x^{(i)}-\bar{x})(y^{(i)}-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x^{(i)}-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y^{(i)}-\bar{y})^2}}$
  - $r_{xy}>0$：正の相関　$r_{xy}<0$：負の相関　$r_{xy}=0$：無相関
  - $-1\le r_{xy}\le 1$の値を取る



相関についての注意：
相関関係と因果関係はイコールではない

- 相関関係 (correlation) があるからといって必ずしも因果関係 (causality) があるわけではない
  - 体重と身長の相関は高いが片方が他方を決めるともいえない
  - 因果関係を示すことは難しい
- 見かけ上の因果関係に注意
  - 背後に共通原因が存在する場合もある
  - 例：「明かりをつけたまま眠る子供は近視になりやすい」？
    - 両者に「親が近視」という別の原因がある
  - そのほか、原因と結果が逆、互いに一方が他方の原因になっている、といったケースあり

相関と因果の違い：介入の効果があるかどうか

- 相関は必ずしも因果を意味しない
  - 相関：片方の変数が変化**する**と、もう片方の変数も変化する
  - 因果：片方の変数を変化**させる**と、もう片方の変数も変化する（介入する）

- 原因？結果？