https://shorturl.at/Ur7dh

KYOTO UNIVERSITY

統計的モデリング基礎①
～概要・導入～

鹿島久嗣
（情報学科 計算機科学コース）

DEPARTMENT OF INTELLIGENCE SCIENCE AND TECHNOLOGY

今学期の講義について

成績評価：
中間試験と期末試験

- 基本方針：中間試験と期末試験で成績をつける
- PandA上での確認クイズ：
  - 講義後1週間オープン
  - 出席の代わり
  - ただし、成績評価においては参考記録程度の扱い

導入

本講義の目的：
統計的モデル化の基礎を身につける

- 我々は、研究や業務で出会う様々な種類のデータから適切な判断を下したい（自動的なシステムあるいは、人間の意思決定をサポート）場面にしばしば遭遇する
  - 例：実験データ、社会調査データ、検査・診断データ、売り上げデータ、行動データ、Webサイトのログ等々
- そのために、観測されたデータに基づいて、不確実な現象の特性を捉え、まだ見ぬ将来の観測値の振る舞いを推定することで、予測や制御に資する統計的モデル化の基礎を学ぶ
  - 現在注目を浴びている機械学習（≒人工知能）の基礎でもある

統計的モデルが世の中で使われている例：顧客の購買行動の予測に基づく推薦

- Webショッピングサイトでの商品推薦の例を考える：
  - 誰に何を薦めると買ってくれるだろうか？下記はタコ焼き機を買った人に推薦される商品

- 消費者の購買行動を予測し、購入しそうなものを推薦する
  - 過去の購買履歴をもとに、ある商品を買ってくれるかどうか予測
    - これまでに購入した商品のリストから、将来ある商品を購入する確率を推定する
  - たとえば、最も購買可能性が高いものから提示すればよさそう

本講義のトピック（予定）：
データ解析の基礎的項目

1. 回帰モデル：線形回帰モデルと最小二乗法による推定など
2. モデル推定　：最尤推定、事後確率最大化等のモデル推定の枠組み
3. モデル選択　：情報量基準、交差確認等に基づくモデルの選択
4. 質的変数の予測モデル：ロジスティック回帰モデルなど
5. 複雑で大規模な予測モデル：ニューラルネットワーク
6. 様々なデータに対する確率モデル：時系列、テキスト、…
7. ベイズ推定：ベイズ統計の枠組みに基づく統計モデル推定
8. 因果推論　：相関関係と因果関係の違い、因果関係の推定法

データとはなにか：
たとえば表形式データ

- 項目と値の組で構成される

（各行が1つの企業、業種や会社規模などで表されている）

	タイプ	会社規模	売上($M)	利益($M)	従業員数	従業員一人あたりの利益	資産	利益/売上 (単位:%)
1	Computer	small	855.1	31.0	7523	4120.70	615.2	3.63
2	Pharmaceutical	big	5453.5	859.8	40929	21007.11	4851.6	15.77
3	Computer	small	2153.7	153.0	8200	18658.54	2233.7	7.10
4	Pharmaceutical	big	6747.0	1102.2	50816	21690.02	5681.5	16.34
5	Computer	small	5284.0	454.0	12068	37620.15	2743.9	8.59
6	Pharmaceutical	big	9422.0	747.0	54100	13807.76	8497.0	7.93
7	Computer	small	2876.1	333.3	9500	35084.21	2090.4	11.59
8	Computer	small	709.3	41.4	5000	8280.00	468.1	5.84
9	Computer	small	2952.1	-680.4	18000	-37800.0	1860.7	-23.05
10	Computer	small	784.7	89.0	4708	18903.99	955.8	11.34
11	Computer	small	1324.3	-119.7	13740	-8711.79	1040.2	-9.04
12	Pharmaceutical	medium	4175.6	939.5	28200	33315.60	5848.0	22.50
13	Computer	big	11899.0	829.0	95000	8726.32	10075.0	6.97
14	Computer	small	873.6	79.5	8200	9695.12	808.0	9.10
15	Pharmaceutical	big	9844.0	1082.0	83100	13020.46	7919.0	10.99
16	Pharmaceutical	small	969.2	227.4	3418	66530.13	784.0	23.46
17	Pharmaceutical	medium	6698.4	1495.4	34400	43470.93	6756.7	22.32
18	Computer	big	5956.0	412.0	56000	7357.14	4500.0	6.92

データをもとにやりたいことの例：予測や因果関係の抽出

- 前述のデータを利用してやりたいこととして、例えば：
  - 予測：会社の売り上げから利益を予測したい
  - モデル推定・選択 ：予測の式をデータからどのように得るか
  - 因果推論 ：従業員を減らすと、従業員ひとりあたり利益は伸びるか

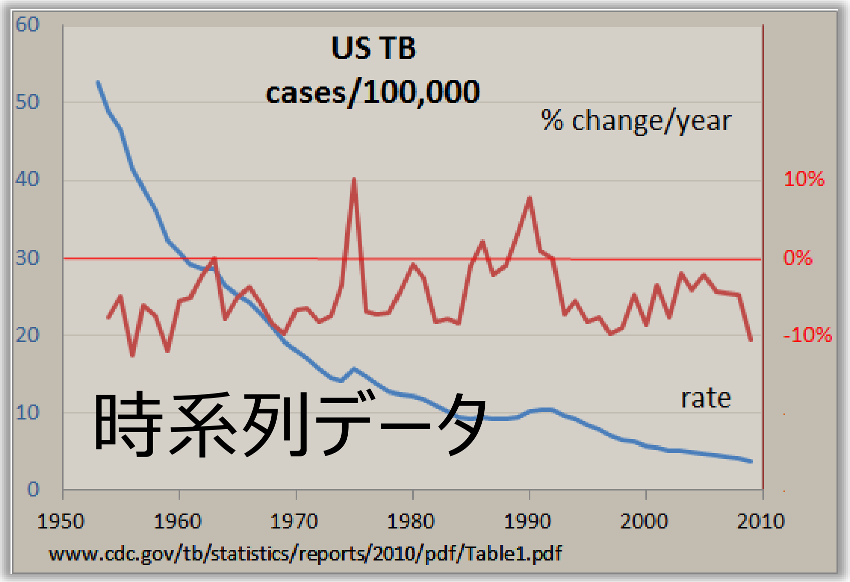
  などが考えられるだろう

- さらに進んで、以下のようなことも考えられるかもしれない：
  - ベイズ推定：データが少ないときにどうするか？
  - 様々なデータ：会社説明のテキストがあったらどうするか？

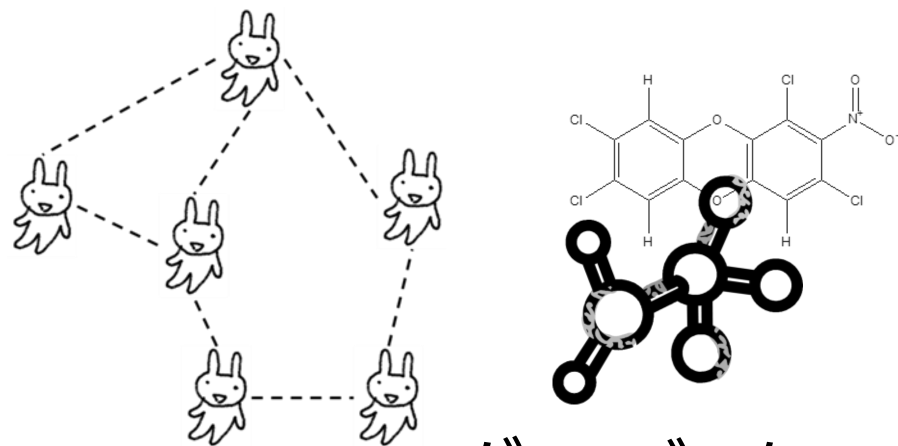
表形式以外のさまざまなデータ：時系列、テキスト、グラフなど…

- 時系列
- テキスト
- グラフ

https://en.wikipedia.org/wiki/Text_corpus



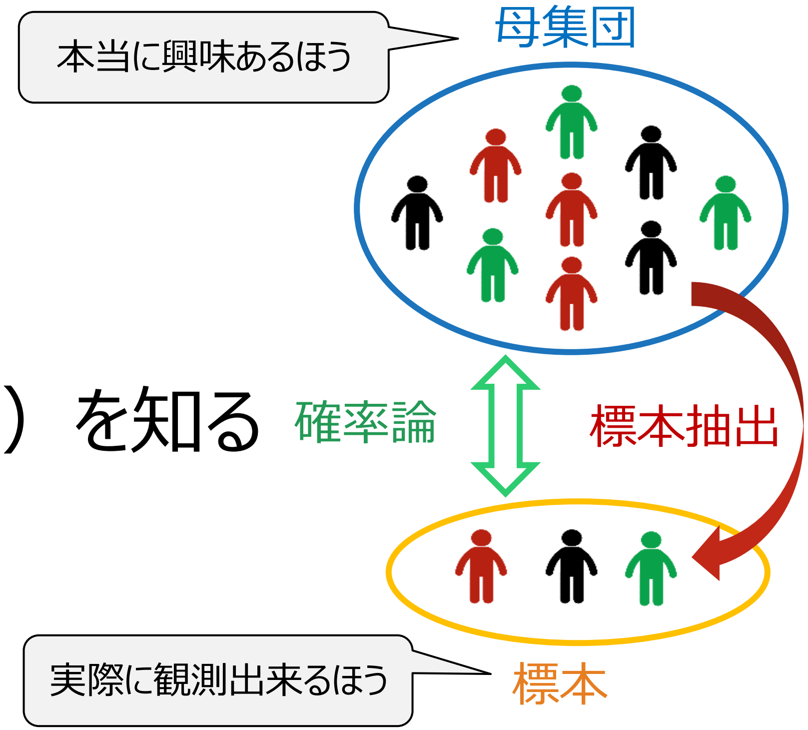
https://en.wikipedia.org/wiki/Time_series#/media/File:Tuberculosis_incidence_US_1953-2009.png



グラフデータ

統計的モデル化の目的：
「部分」から「全体」を知ること

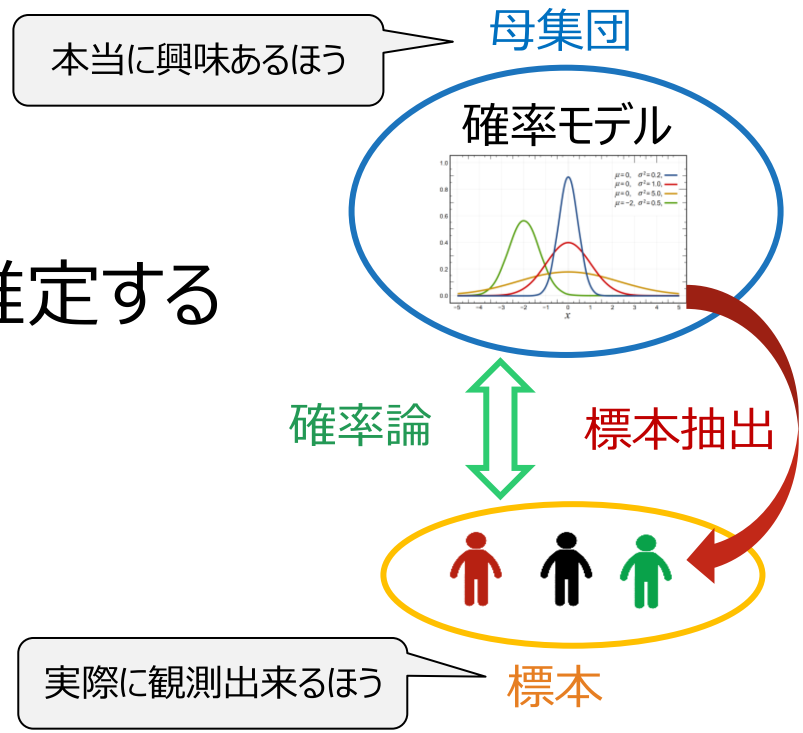
- すべての場合（母集団）を網羅的に観測できることは少ない
- 「記述統計」と「推測統計」
  - 記述統計：全数調査を前提とする
  - 推測統計：標本調査を前提とする
    - 部分（標本）から全体（母集団）を知る
    - 過去から未来を予測する



- 母集団と標本は「確率論」でつながる
  - 母集団は対象となる集合の要素すべて、あるいは、何らかの確率分布に従っていて、標本はそこから確率的に取り出されたと考える

確率モデルとは何か：
データとデータの「間」をつなぐもの

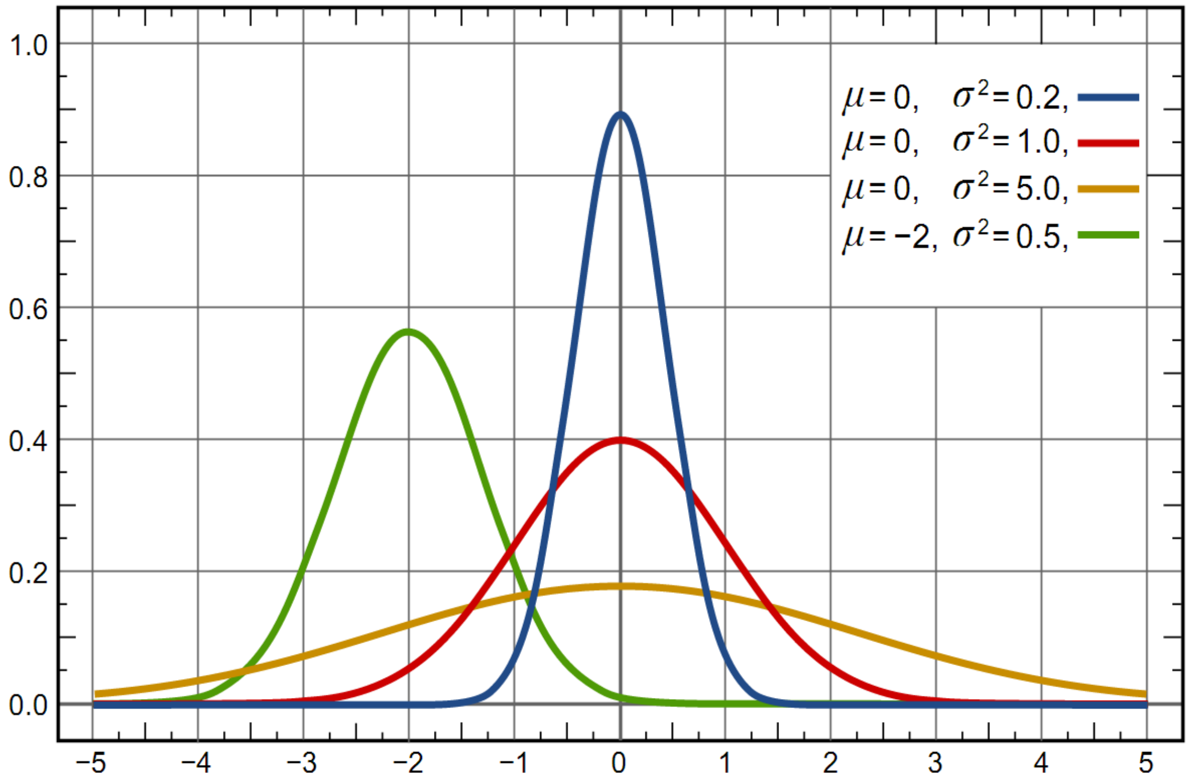
- 全数調査のかわりに、部分（限られたデータ）から全体を知るためには、データとデータの間を補間する必要がある
- そのためにはデータの分布に関する仮定が必要になる
  - 仮定＝確率モデル
- データから確率モデルを推定する
  - より具体的には、モデルパラメータを推定する
- モデルの利用法：
  - モデルを用いて全体の性質を知る
  - 未来のデータについて予測を行う



代表的な確率モデル：正規分布

- 量的な確率変数に関する最も基本的な確率分布の一つ
- データは平均値 μ を中心にバラつき度合σで散らばる

確率密度 $f(x)$



x

正規分布の確率密度関数

$$f(x) = N(x|\mu, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ただし以下を満たす

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

確率モデルとは：データの生成過程

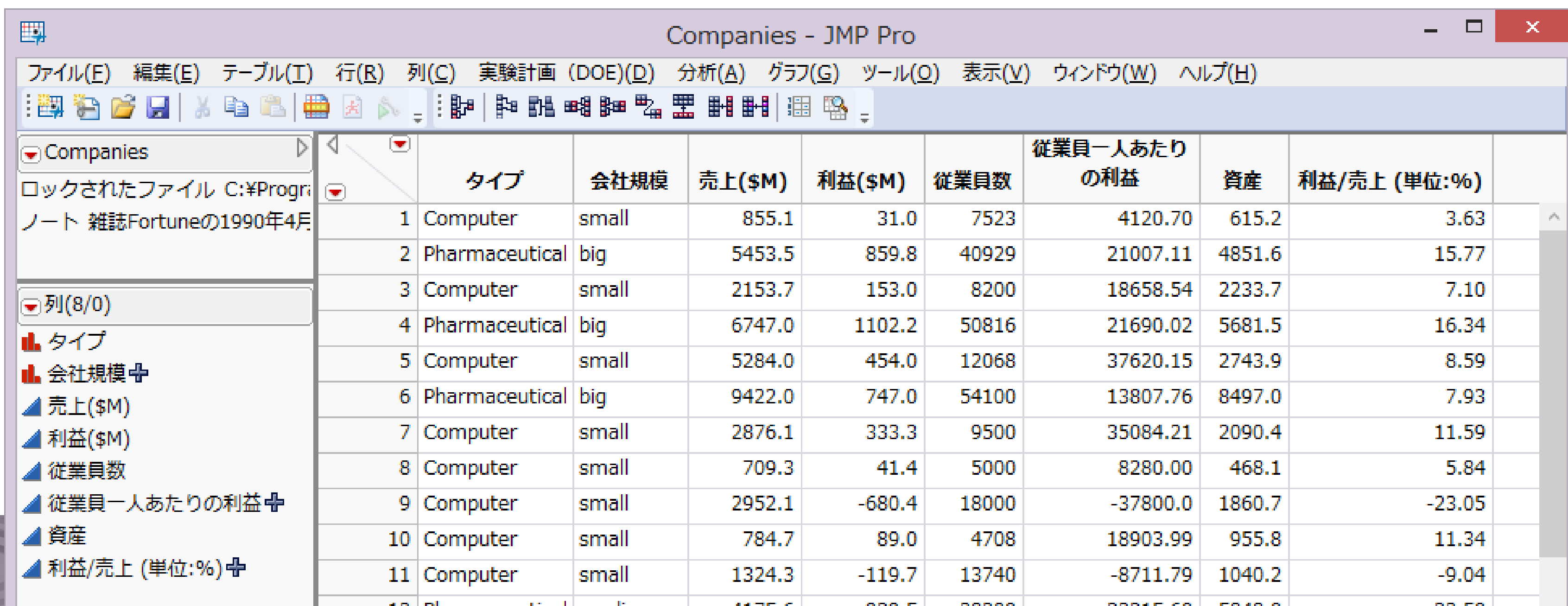
- 母集団は対象となる集合の要素すべて、あるいは、何らかの確率分布に従っていて、標本はそこから確率的に取り出されたと考える
- モデルはデータの生成器として理解できる
  - ボタンを押すとデータが出てくる機械（のようなもの）
- サイコロのモデル：出目Xの確率$P(X = i) = \frac{1}{6}$ $(i = 1,2, \ldots, 6)$
- ある行動をとるかどうかのモデル：
ある人のとる行動Xがaである確率$P(X = a) = 0.8$
- 多くの場合、個々のデータは同じ分布に従い、独立に生成されると仮定する （＝ i.i.d: identically & independently distributed）

初等的なデータ分析

基本的なデータの種類：質的データと量的データ

- 統計データには質的データと量的データがある

1. 質的データ：
   Yes/No、好き/普通/嫌い などの記号を値にとるデータ
2. 量的データ：
   温度や身長など 数値を値にとるデータ (連続尺度)



	タイプ	会社規模	売上($M)	利益($M)	従業員数	従業員一人あたりの利益	資産	利益/売上 (単位:%)
1	Computer	small	855.1	31.0	7523	4120.70	615.2	3.63
2	Pharmaceutical	big	5453.5	859.8	40929	21007.11	4851.6	15.77
3	Computer	small	2153.7	153.0	8200	18658.54	2233.7	7.10
4	Pharmaceutical	big	6747.0	1102.2	50816	21690.02	5681.5	16.34
5	Computer	small	5284.0	454.0	12068	37620.15	2743.9	8.59
6	Pharmaceutical	big	9422.0	747.0	54100	13807.76	8497.0	7.93
7	Computer	small	2876.1	333.3	9500	35084.21	2090.4	11.59
8	Computer	small	709.3	41.4	5000	8280.00	468.1	5.84
9	Computer	small	2952.1	-680.4	18000	-37800.0	1860.7	-23.05
10	Computer	small	784.7	89.0	4708	18903.99	955.8	11.34
11	Computer	small	1324.3	-119.7	13740	-8711.79	1040.2	-9.04

質的データと量的データの分類：さまざまな尺度

- 質的データ： 記号を値としてとるデータ
  - 名義尺度：値が単なるラベルとして扱われる（例：「Yes」「No」）
  - 順序尺度：順序に意味がある（例：「好き」>「普通」>「嫌い」）
- 量的データ： 数値を値としてとるデータ (連続尺度)
  - 間隔尺度：数の間隔に意味がある（例：温度）
  - 比例尺度：数の比に意味がある（例：身長）
    - 原点に意味があるともいえる

量的データの例：
体重データ

- 100名分の体重データ（1次元）：このままだとわかりにくい

No.	体重	No.	体重	No.	体重	No.	体重	No.	体重
1	48	21	52	41	52	61	55	81	54
2	48	22	50	42	57	62	54	82	55
3	40	23	55	43	56	63	55	83	52
4	52	24	53	44	50	64	52	84	49
5	60	25	49	45	49	65	50	85	51
6	55	26	56	46	52	66	50	86	55
7	52	27	52	47	51	67	48	87	50
8	55	28	56	48	45	68	52	88	51
9	53	29	50	49	46	69	52	89	45
10	50	30	52	50	50	70	50	90	56
11	53	31	50	51	49	71	55	91	53
12	62	32	55	52	50	72	50	92	50
13	48	33	50	53	53	73	56	93	53
14	55	34	56	54	58	74	54	94	55
15	45	35	66	55	52	75	48	95	55
16	48	36	49	56	48	76	54	96	51
17	50	37	55	57	65	77	50	97	48
18	50	38	58	58	56	78	49	98	52
19	50	39	48	59	50	79	52	99	63
20	48	40	58	60	60	80	52	100	68

量子化：
量的データを理解しやすくするための量子化

- 生データのままではデータを理解するのは困難
- 量子化：データがとりうる値の範囲を、予め定めた区間 （階級）に分け、観測される数値の入る階級 によって集計を行う
  - 観測される数値が実数 (連続値) の場合には、厳密な値は表現できないので必ず量子化を行う
  - 例：体重の場合
    - 観測する最小単位を1kgとし最小単位より小さい端数を丸める
    - あるいは、5kgずつの区間に分け、それぞれの区間で集計する

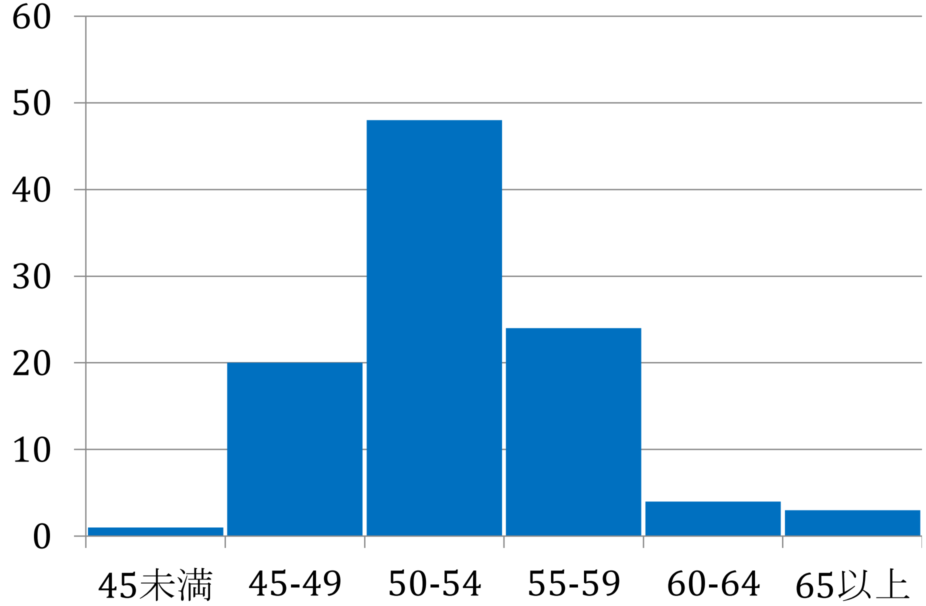
量的データの集計：度数分布表とヒストグラム

- ヒストグラムでデータ分布を視覚化
  - 度数分布表：各階級の度数をカウント
  - ヒストグラム： 度数分布のグラフ表現

度数分布表 （階級幅5kg）

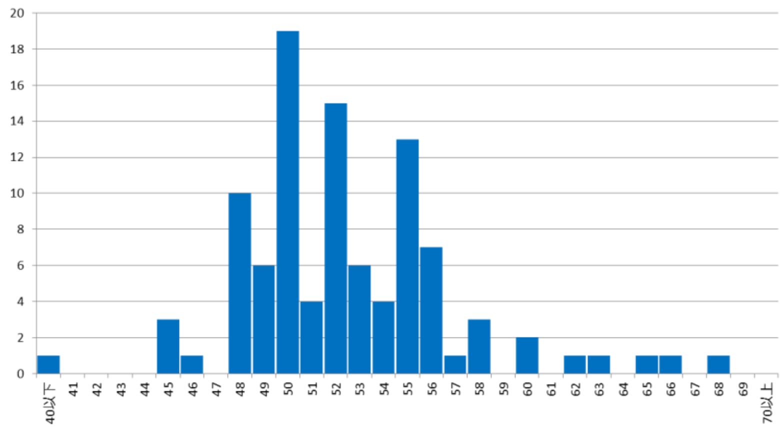
階級	度数
45未満	1
45～49	20
50～54	48
55～59	24
60～64	4
65以上	3

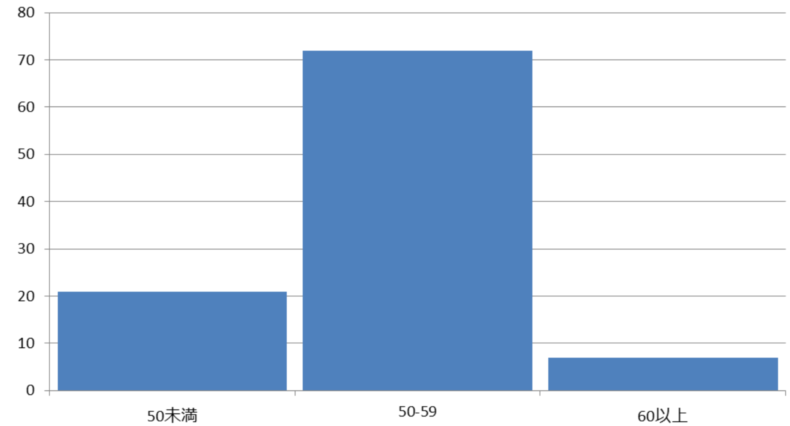
ヒストグラム



ヒストグラムと階級幅の関係：ヒストグラムでは幅の決め方で見た目が大きく変わる

- 階級幅1の場合と10の場合でヒストグラムの形が変わる





- スタージェス (Sturges) の方法：$K = \log_2 N + 1$
  - データが100個：$\log_2 100 + 1 = 7.643856 \rightarrow$ 8階級ぐらい
  - データが50個：$\log_2 50 + 1 = 6.643856 \rightarrow$ 7階級ぐらい
  - データが25個：$\log_2 25 + 1 = 5.643856 \rightarrow$ 6階級ぐらい

そのほかの集計：
度数・累積度数・相対度数・累積相対度数

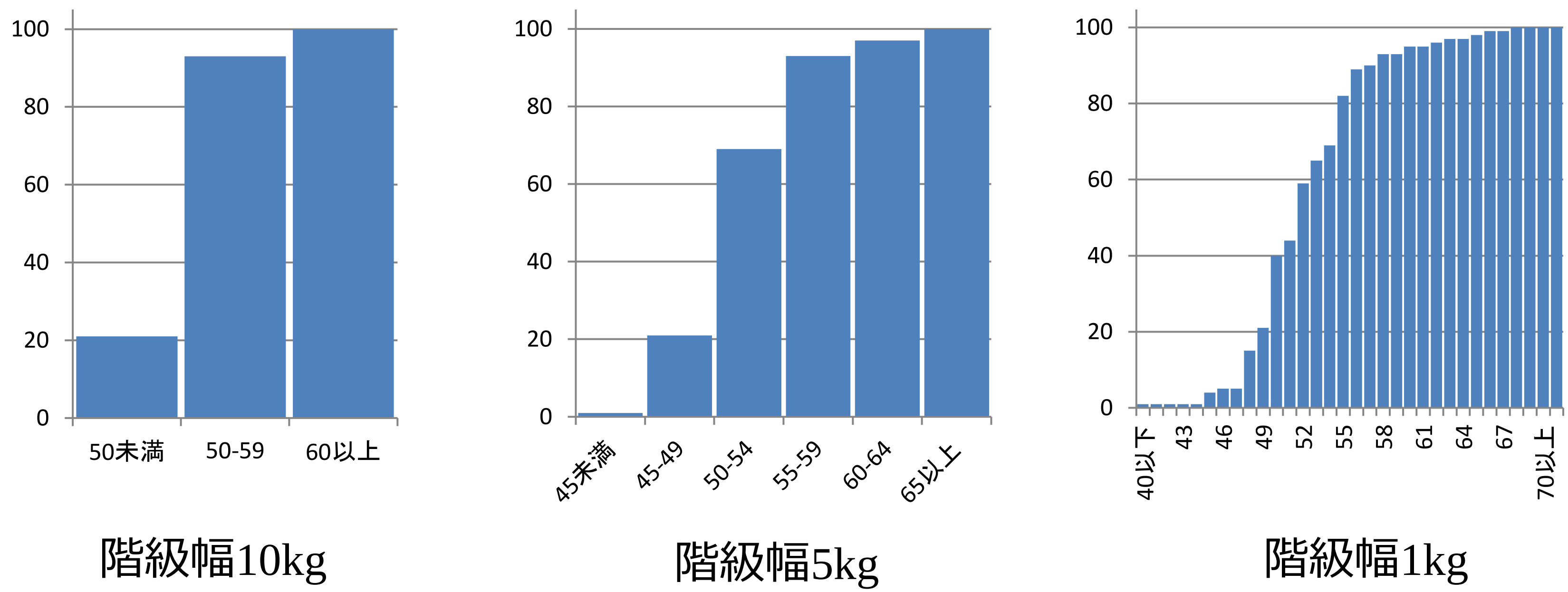
- データ：$x^{(1)}, x^{(2)}, \cdots, x^{(n)}$ を いくつかの階級：$I_1, I_2, I_3, \cdots, I_K$ に分割する

- 度数：$f_1, f_2, f_3, \cdots, f_K$
  - $x_i \in I_k$を満たすiの個数
  - 累積度数：$F_k = \sum_{i=1}^{k} f_k$
  - 相対度数：$\frac{f_k}{N}$
  - 相対累積度数：$\frac{F_k}{N}$

階級	度数	累積度数	相対度数	累積 相対度数
45未満	1	1	1%	1%
45-49	20	21	20%	21%
50-54	48	69	48%	69%
55-59	24	93	24%	93%
60-64	4	97	4%	97%
65以上	3	100	3%	100%

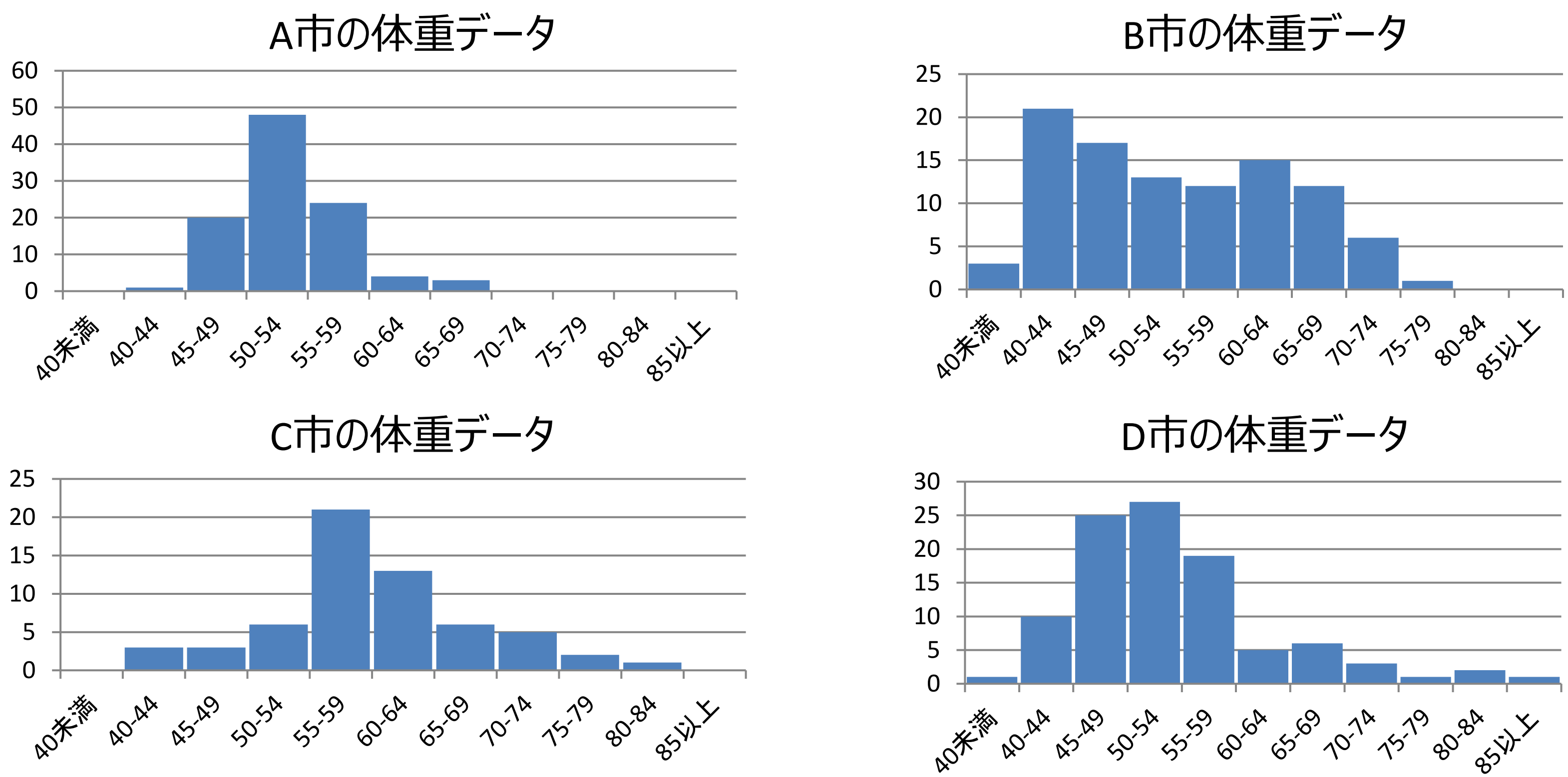
累積度数と階級幅の関係：累積度数は階級幅にそれほど左右されない

- 累積度数は階級幅にそれほど左右されない
  - –むしろ階級幅が小さいほうが分布の様子がよくわかるくらい…



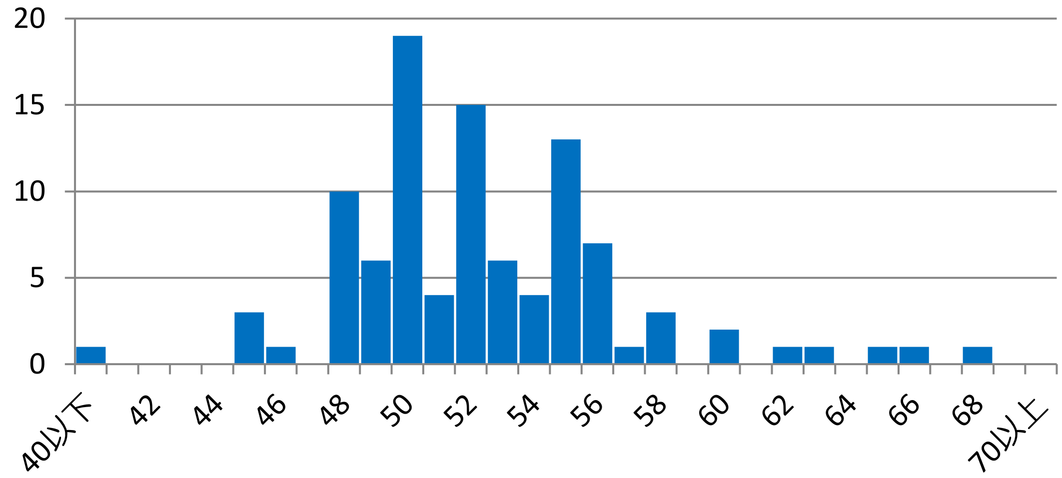
複数種類のデータを比較したい場合：ヒストグラムの形を表す指標がほしい

- ヒストグラムから分布の形状はよくわかるが、一覧性には欠ける
- ヒストグラムの特徴を表す少数の指標で代表したい



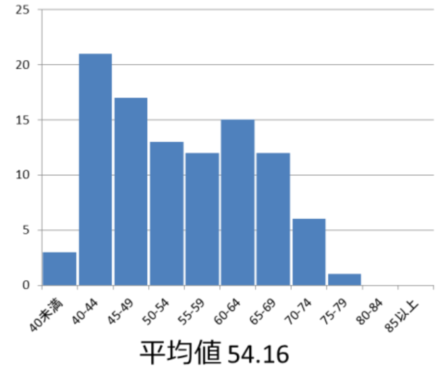
データの代表値：
標本平均・中央値

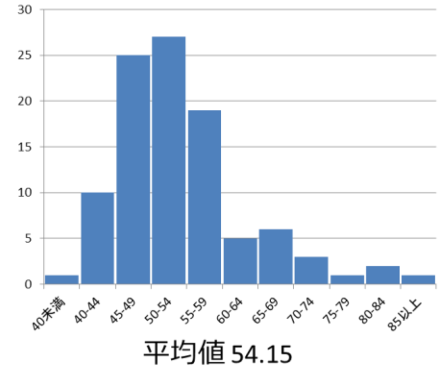
- データ$x^{(1)}, x^{(2)}, \cdots, x^{(n)}$の特徴を表す数値
  - 標本平均：$\bar{x} = \frac{1}{N}\left(x^{(1)} + x^{(2)} + \cdots + x^{(n)}\right)$
    - $\mathrm{argmin}_x f(x) = \left(x^{(1)} - x\right)^2 + \left(x^{(2)} - x\right)^2 + \cdots + \left(x^{(n)} - x\right)^2$
  - 中央値 (median) : 大きいほうからだいたい $\frac{n}{2}$ 番目の値
    - 外れ値の影響を受けにくい



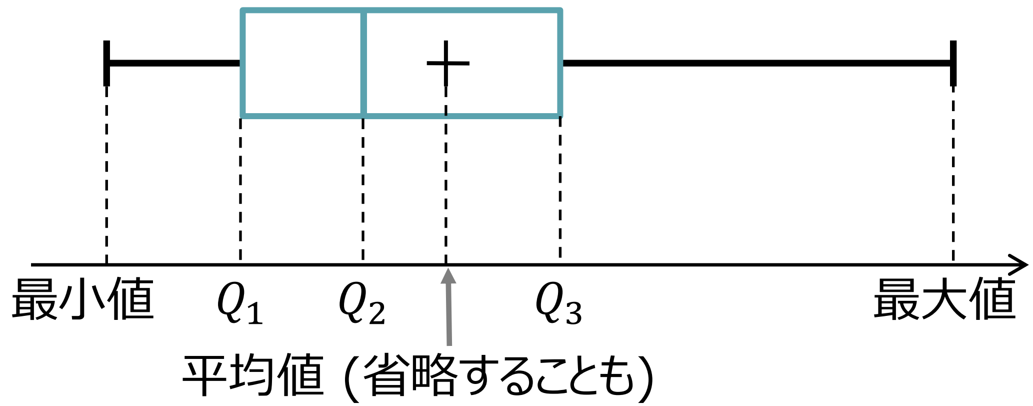
データ分布の代表値：分散・四分位点・箱ひげ図

- 平均だけでは不十分な場合もある
- 分布の形も知りたい
  - データのばらつき ： 分散
  - ４分位点：整列したデータを四等分する位置にある値
    - Q_1：25%点、Q_2：50%点（中央値）、Q_3：75%点、





- 箱ひげ図による可視化



不偏分散：データのばらつきをあらわす

- 不偏分散$\hat{\sigma}^2$：データのばらつきを表す
  - $\hat{\sigma}^2 = \frac{\left(x^{(1)}-\bar{x}\right)^2+\left(x^{(2)}-\bar{x}\right)^2+\cdots+\left(x^{(n)}-\bar{x}\right)^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2$
  - 平均と分散でデータを捉える ＝ 背後に正規分布を仮定
- ばらつきを表す類似の指標：
  - 変動係数CV (coefficient of variation)：$\frac{\hat{\sigma}}{\bar{x}}$
    - 相対標準偏差 (relative standard deviation: RSD) とも呼ばれる
    - 平均値が異なる二つの集団のばらつきを比較するのに用いる
  - 偏差値T_i：$x^{(i)}$を平均値50・標準偏差10となるようにスケールした値

練習問題：
ストリームデータの平均・分散の計算

- ストリームデータ：時々刻々到着するデータ
  - 時刻tにおいてデータ$x^{(t)}$が観測される
  - 例：センサーデータ
- これまでに観測されたデータの平均・分散を、各時刻で$O(1)$で保持したい
  - 定義に従って素朴に計算すると$O(t)$

2変量データの解析

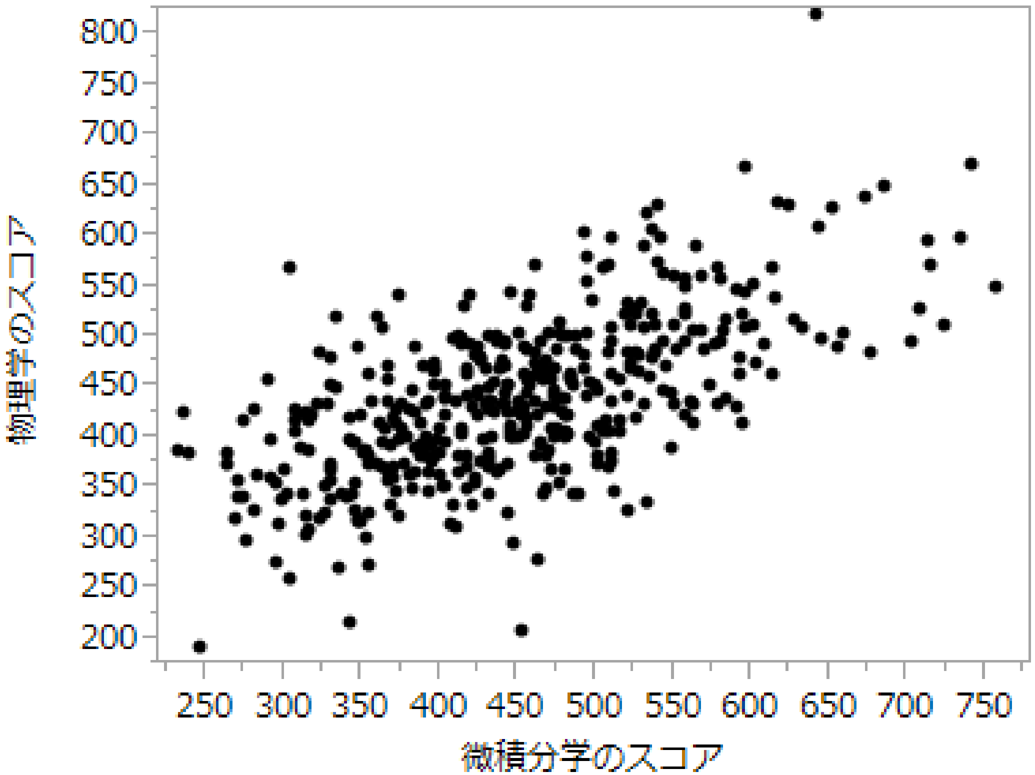
2つ以上の変量のデータ分析：
変量間の関係を調べることでより深い分析が可能

- 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

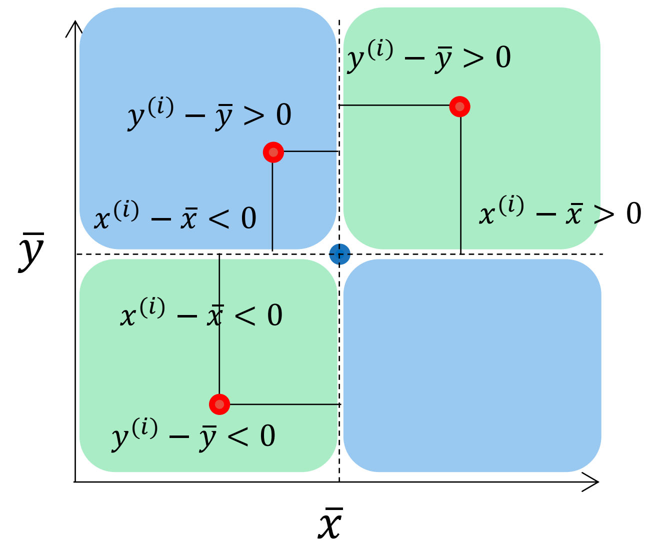
	微積分学のスコア	物理学のスコア
1	441.4	470.7
2	632.16	508.44
3	361.56	412.75
4	479.39	425.47
5	476.32	408.27
6	446.92	400.99
7	394.2	390.62
8	645.76	496.97
9	329.75	367.39
10	496.07	453.41
11	487.91	498.97
12	403.82	441.7
13	480.21	400.41
14	460.33	460.72
15	303.72	259.66
16	463.01	278.04
17	428.98	396.21
18	412.12	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：
共分散によって２つの変数の増減の関係が測れる

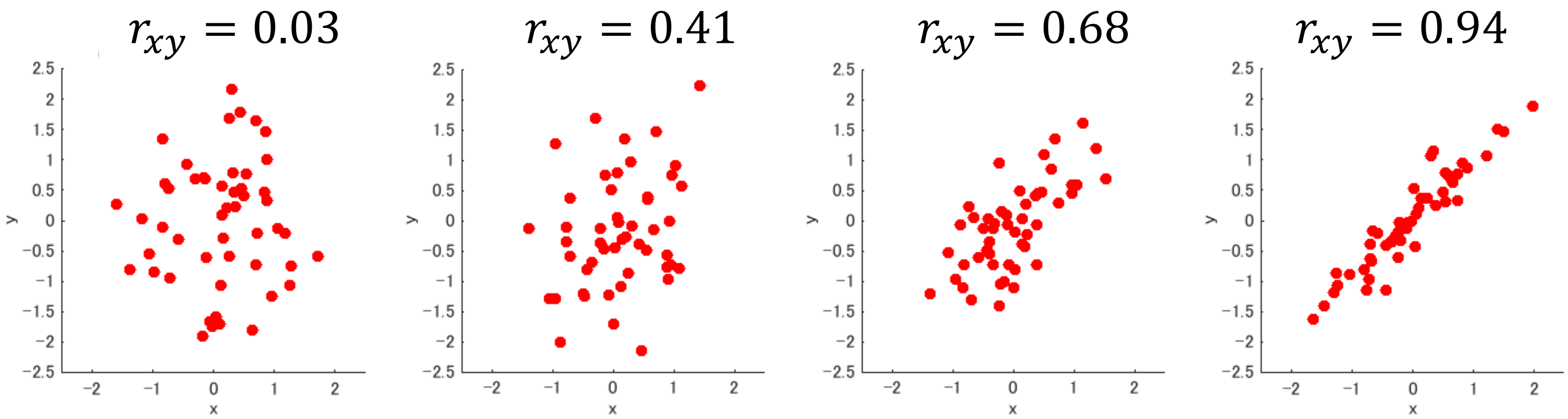
- 一方が増えたときに他方が増える （減る） 関係性を表す指標
- 共分散 (covariance)：$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$
  - ただし、$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$， $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
  - 偏差積の平均 (データのバラツキを表現)
    - 偏差 $(x^{(i)} - \bar{x})$ と偏差 $(y^{(i)} - \bar{y})$ 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：
相関は共分散のスケールを正規化したもの

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



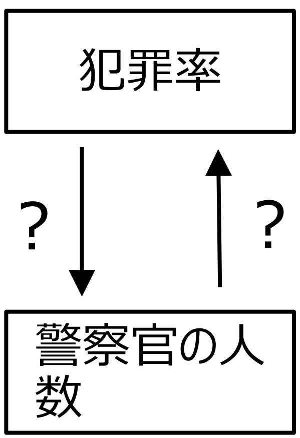
相関についての注意：
相関関係と因果関係はイコールではない

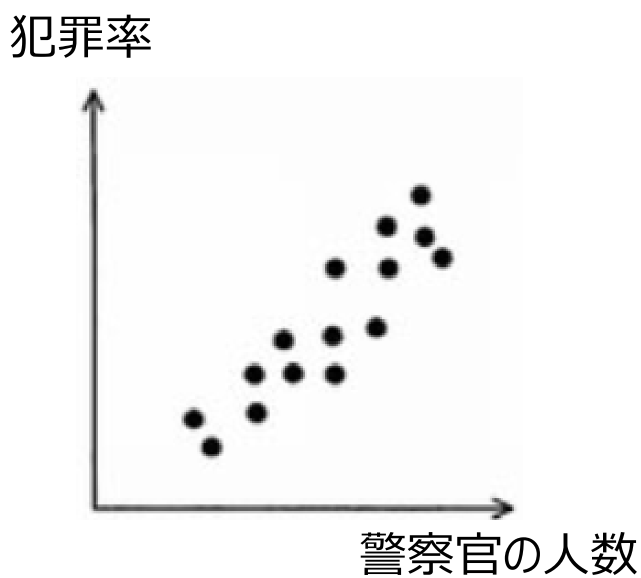
- 相関関係 (correlation) があるからといって必ずしも因果関係 (causality) があるわけではない
  - 体重と身長の相関は高いが片方が他方を決めるともいえない
  - 因果関係を示すことは難しい
- 見かけ上の因果関係に注意
  - 背後に共通原因が存在する場合もある
  - 例：「明かりをつけたまま眠る子供は近視になりやすい」？
    - 両者に「親が近視」という別の原因がある
  - そのほか、原因と結果が逆、互いに一方が他方の原因になっている、といったケースあり

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**<u>する</u>**と、もう片方の変数も変化する
  - 因果：片方の変数を変化**<u>させる</u>**と、もう片方の変数も変化する（介入する）

- 原因？結果？





まとめ：
統計的モデル化の導入と量的データの初等的分析

- 観測されたデータを理解し、予測をおこなうために、データの背後でデータを生み出す確率モデルを考える
- 限られたデータ（標本）から（母集団の）モデルを推定する
- データには量的データ、質的データがある
- 1変量の量的データの初等的分析には、ヒストグラム等を用いて可視化したり、平均・分散などの指標でとらえる
- 2変量の間の関係を捉えるためには、散布図や相関係数等を用いる
- 相関と因果は異なる