入門 統計学　第1章

データの整理

『入門 統計学 第2版　―検定から多変量解析・実験計画法・ベイズ統計学まで―』（オーム社）

※注：本書を購入された方へのサービスですので，教科書指定（参考図書は不可）していない授業での使用はお控えください。

統計学とは？

- 実験や調査によって観測されたデータの特性を調べる学問
- 図表にまとめたり，統計量（平均や分散）を計算することで特性を把握・推定する

データの特性を調べる

なぜ統計学を学ぶのか

- 大工さんが最初に学ぶのは，道具の使い方であるように，論文や報告書を書くための道具として統計学を学ぶ
- 実験結果の汎用性（偶然ではないこと）を示すことで，他人を納得させる

色々な場面で使われている統計学

- 地震の予測
- 選挙速報の当確
- テレビの視聴率
- 新薬の有効性
- アパートの家賃
- 品質の管理
- ビッグデータの解析…，他にも沢山！





統計学の種類

推測統計学の基本でもあるので，これから学ぶ

- 統計学
  - 記述統計学（第1～2章）
    ▸手元にあるデータの特性を調べる
  - 推測統計学
    標本から母集団の特性を推測
    - 頻度論的統計学（第3章以降）
      ▸区間推定や仮説検定など伝統的な手法
    - ベイズ統計学（第15章）
      ▸データが逐次更新される場合や複雑なモデル向き

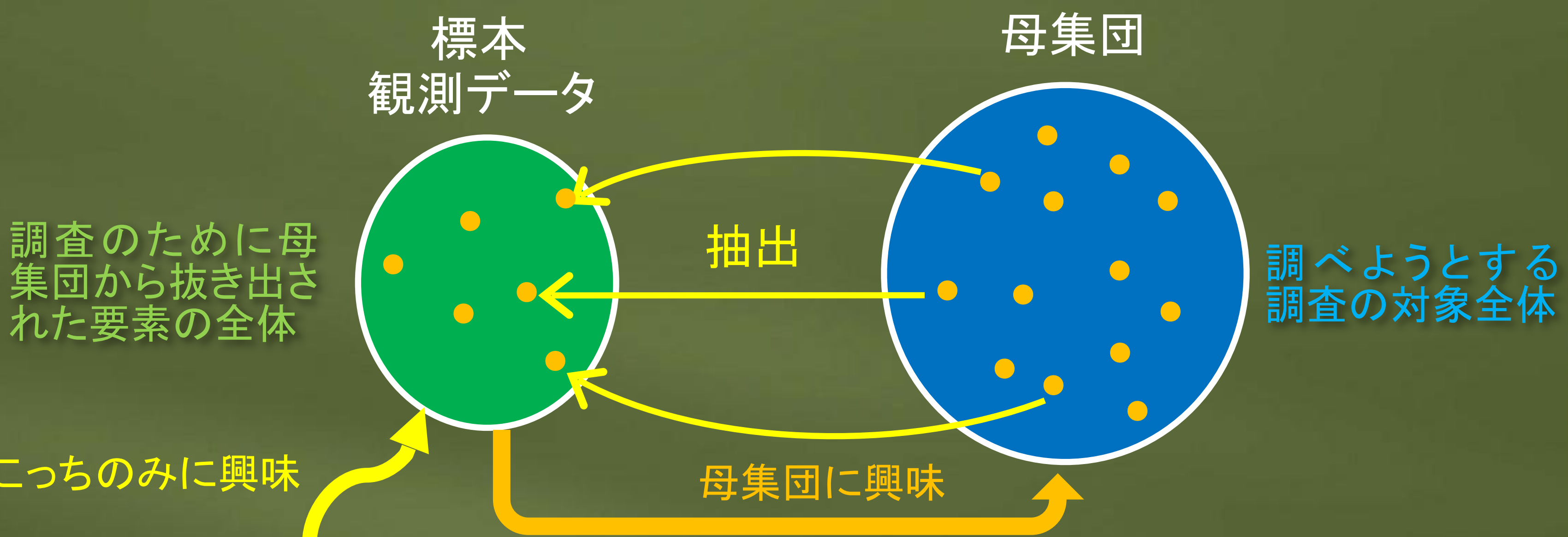
そのほかの統計学

- その他の統計学
  - 多変量解析（第12～14章）
    - ▸複数の変数の因果関係を明らかにしたり，個体を分類するための分析手法
  - 実験計画法（第10章）
    - ▸誤差の影響を受けないような実験や，効率的にデータを収集できる実験を計画する

1.1　記述統計学

- 記述統計学：統計量（平均や分散）や図表に整理することで，観測したデータ（変数に値が入った状態）の特性を捉える統計学
- 推測統計学（第3章以降）との違い：記述統計学は手元のデータの特性を知ることが目的（次掲）

記述統計学 vs 推測統計学



記述統計学:手元にある観測データの特性(平均や分散)を知る
→本章で学ぶ

推測統計学:母集団の特性(平均や分散など)を標本から推測
→(第3章以降で学ぶ)

尺度水準の問題

- データを測定（観測）する尺度によって，許される計算や分析が異なる（尺度水準の問題）
- 手元のデータは，どの尺度水準で測定されたものなのかを区別できなければならない

- 測定データ
  - 量的データ
    - 比率データ（比率尺度）
    - 間隔データ（間隔尺度）
  - 質的データ
    - 順位データ（順序尺度）
    - カテゴリデータ（名義尺度）

尺度水準

高い水準ほど許される演算が増える

測定尺度①　ー量的データー

- 比率データ（質量，長さ，時間など）：絶対的なゼロを持っているため，値の間の比にも意味がある→四則演算が可能

- 間隔データ（摂氏温度，知能指数など）：相対的なゼロしか持たないが，値の間隔は等しい→足し算や引き算が可能

- これら２種類のデータ（量的データ）が統計学の主な対象となる

測定尺度②　－質的データ－

- 順位データ（満足度，選好度など）：
  値の大小関係にのみ意味がある
  →中央値の計算は可能

- カテゴリデータ（性別，職業など）：
  値は内容を区別するだけ
  →カウントしたり最頻値のみ可能

- これら２種類のデータ（質的データ）の分析には専用の手法が必要（第11章）

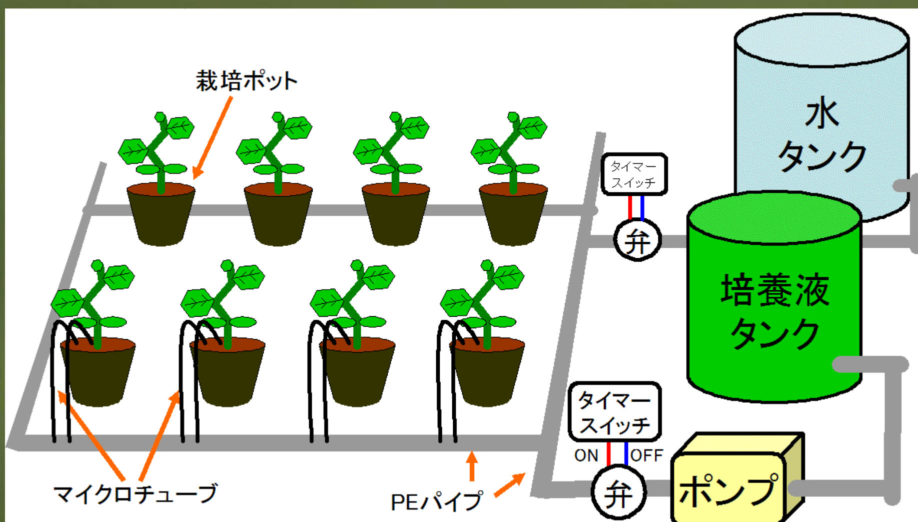
（まとめ）測定尺度の一覧

	データ名称	測定尺度	直接できる演算	主な代表値	事例
量的データ※	比率データ	比率尺度	＋－×÷	幾何平均	質量
	間隔データ	間隔尺度	＋－	算術平均	摂氏温度
質的データ	順位データ	順序尺度	＞＝	中央値	満足度
	ｶﾃｺﾞﾘﾃﾞｰﾀ	名義尺度	度数ｶｳﾝﾄ	最頻値	性別

※量的データには，質量のように値が連続している連続型と，金額のように飛び飛びになっている離散型がある（質的データは全て離散型）。

1.2　度数分布表とヒストグラム

キュウリの養液土耕栽培の実験



観測

表1.2　キュウリの収量(g)

ポット番号	栽培法A（昼）	栽培法B（夜）
1	3,063	3,157
2	2,275	2,707
3	2,089	3,270
4	2,855	3,181
5	2,836	3,633
6	3,219	3,404
7	2,817	2,219
11	2,140	2,938
12	1,757	3,286
13	2,499	2,920
14	2,093	3,332
15	2,073	3,478

値を並べただけの状態から
特性を捉えるのは難しい…
そこで，表や図に整理！

実験の目的:
昼・夜どちらの時間帯に養液を与えた方が収量が多くなるのか（それとも変わらないのか）を検証する

度数分布表の作成手順

①データを昇順に並び替える
→Excelなどの表計算ソフトならば簡単
②階級の数と幅，階級値を決める
→階級数は$\sqrt{n}+1$ か $\log_2 n+1$などを参考　（nは総度数）
③各階級の度数を求める
④適宜，相対度数や累積相対度数を求める
→相対度数＝$\dfrac{\text{その階級の度数}}{\text{総度数}}$

度数分布表（キュウリの事例）

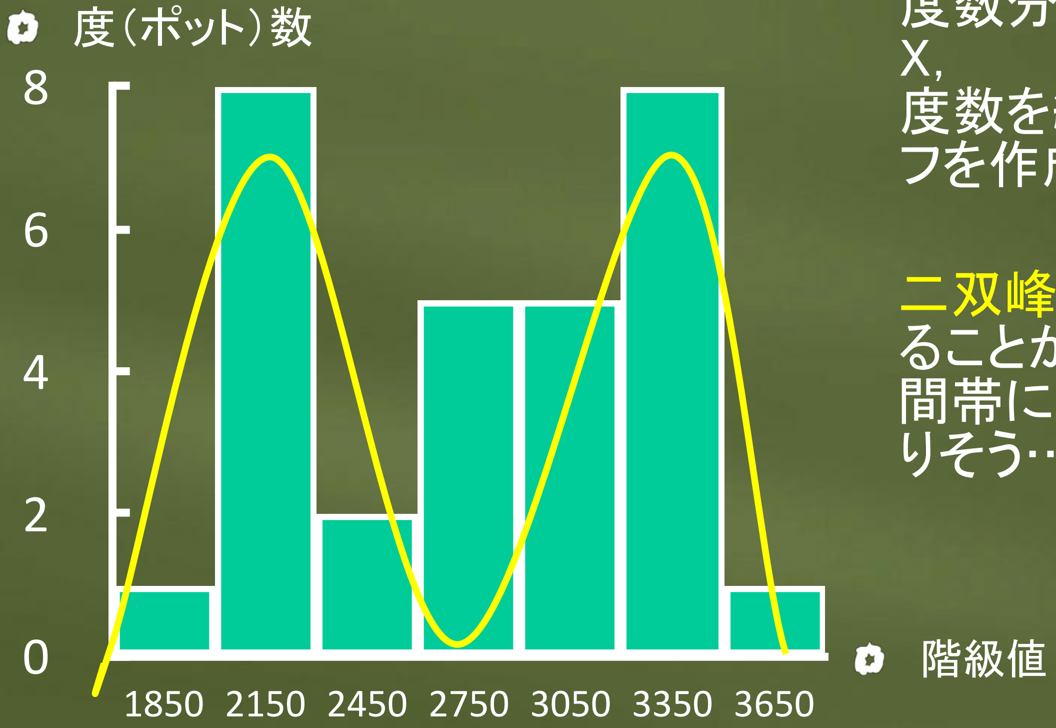
階級を代表する値で中央値を使うことが多い

収量（g）	階級値（g）	度数（ポット数）	相対度数（%）	累積相対度数（%）
1700以上～2000未満	1850	1	3.3	3.3
2000～2300	2150	8	26.7	30.0
2300～2600	2450	2	[illegible]	[illegible]
2600～2900	2750	5	16.7	53.3
2900～3200	3050	5	16.7	70.0
3200～3500	3350	8	26.7	96.7
3500～3800	3650	1	3.3	100.0

これら2つの階級で度数が多くなっていることが容易に確認できる

階級数は$\sqrt{30}+1=6.5$, $\log_2 30+1=5.9$だが，切りの良い区間（300g）とするため7階級と

ヒストグラム



度数分布表の階級を横軸X,
度数を縦軸Yとして縦棒グラフを作成

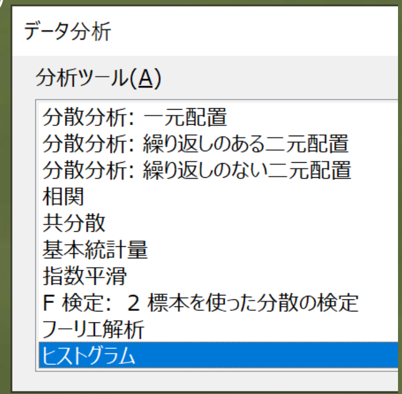
二双峰性の分布になっていることから，養液を与える時間帯によって収量に差がありそう…

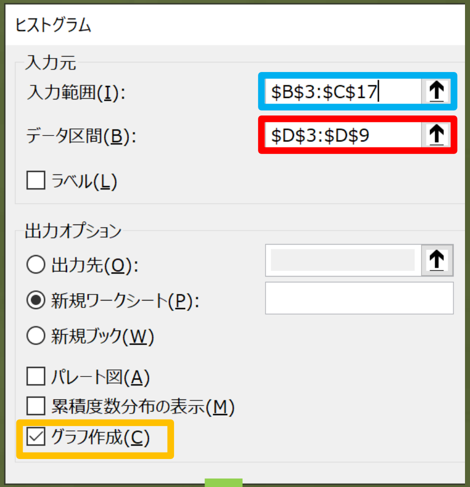
Excel分析ツールによる
度数分布表とヒストグラムの作成

階級値（最大値）を適当な場所に入力
しておく（最大値の数が階級数になる）

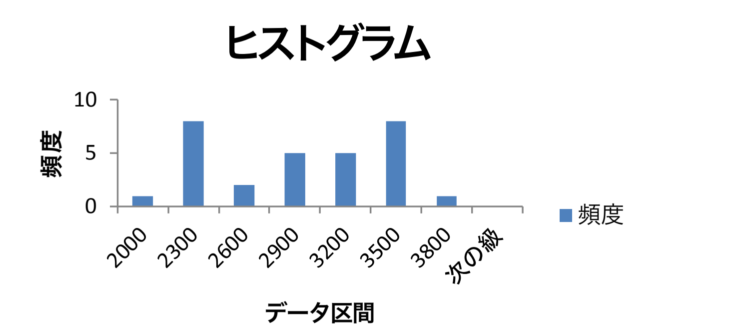
栽培法A	栽培法B	
3,063	3,157	2000
2,275	2,707	2300
2,089	3,270	2600
2,855	3,181	2900
2,836	3,633	3200
3,219	3,404	3500
2,817	2,219	3800
2,136	2,730	
2,540	3,408	
2,263	3,203	
2,140	2,938	
1,757	3,286	
2,499	2,920	
2,093	3,332	
2,073	3,478	

分析ツールを起動して[ヒストグラム]を選択





データ区間	頻度
2000	1
2300	8
2600	2
2900	5
3200	5
3500	8
3800	1
次の級	0



1.3　代表値①　―平均―

- 代表値：データの特性を表す統計量※

  ※データに対して何らかの計算をして得られた値

- 平均や，バラツキを表す分散などがある
- 色々あるなかから，まずは平均をいくつか解説する（算術平均，加重平均，幾何平均，移動平均の順）

算術平均

- 小学校で学んだ誰もが知っている平均

平均を表す記号※
（エックスバー）

第n番目のデータ値の意味
$_n$は添字（そえじ）・添数（てんすう）

算術平均 $\boxed{\bar{x}} = \dfrac{x_1 + \cdots + \boxed{x_n}}{\boxed{n}}$

データの数（標本サイズ）←標本数ではない

総和記号Σ（シグマ）を使う

i=1〜nまでのxを足し合わせる
（Σの上下記号は以降省略）

$$\bar{x} = \frac{\boxed{\sum_{i=1}^{n} x_i}}{n}$$

Excel関数＝AVERAGE

※第3章以降で学ぶ推測統計学では，$\bar{x}$は標本の平均という意味で，母集団の平均であるμと区別する。

例題（キュウリ収量）

算術平均をExcel関数で求める

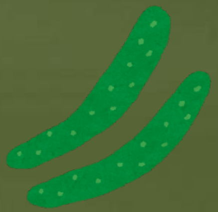


表1.2　キュウリの収量（g）

ポット番号	栽培法A（昼）	栽培法B（夜）
1	3,063	3,157
2	2,275	2,707
3	2,089	3,270
4	2,855	3,181
5	2,836	3,633
6	3,219	3,404
7	2,817	2,219
8	2,136	2,730
9	2,540	3,408
10	2,263	3,203
11	2,140	2,938
12	1,757	3,286
13	2,499	2,920
14	2,093	3,332
15	2,073	3,478

両栽培法合わせた算術平均

=AVERAGE(B3:C17)＝2784

栽培法A（昼）の算術平均

=AVERAGE(B3:B17)＝2444

栽培法B（夜）の算術平均

=AVERAGE(C3:C17)＝3124

加重平均

（群別標本サイズなどを重みとした平均）

- 観測された値x_iに重みw_iをかけて平均を計算

加重平均 $\bar{x}_w = \dfrac{w_1 x_1 + \cdots + w_n x_n}{w_1 + \cdots + w_n} = \dfrac{\sum w_i x_i}{\sum w_i}$

例：女子20名，男子10名のクラスで，女子の平均が65点，男子が80点のクラスの平均点は？（算術平均は72.5点）

$$\bar{x}_w = \frac{20 \times 65 + 10 \times 80}{20 + 10} = 70$$

（女子の人数が多かったため）算術平均72.5点よりも下がる

幾何平均

（外れ値のあるデータや変化率の平均）

- 観測値をかけ合わせた積のn乗根

幾何平均 $\bar{x}_g = \sqrt[n]{x_1 \times \cdots \times x_n} = \sqrt[n]{\prod x_n} = \left(\prod x_n\right)^{\frac{1}{n}}$

総乗記号（パイ）

特徴：外れ値に引っ張られないため，極端な値の出やすい細菌数の測定（の平均）によく使われる

Excel関数＝GEOMEAN（数値）

幾何平均で変化率の平均を求める

- 掛け算の累乗根である幾何平均は，掛け算である「変化率の平均」にも適している
- 例（物価）：2年前から昨年にかけて物価が2倍になり，昨年から今年にかけては8倍になりました。この2年間の物価の対前年比の平均は何倍でしょうか？

算術平均：（2+8）÷2=5 or 幾何平均：$\sqrt{2\times 8}$=4

答え：2年前に100円のものが昨年は200円になり，今年は1600円（16倍）になっています。算術平均を2回（2年分）掛けると25倍，幾何平均を2回だと16倍なので，後者の方が妥当でしょう。

移動平均

（変動を抑えて傾向を見る）

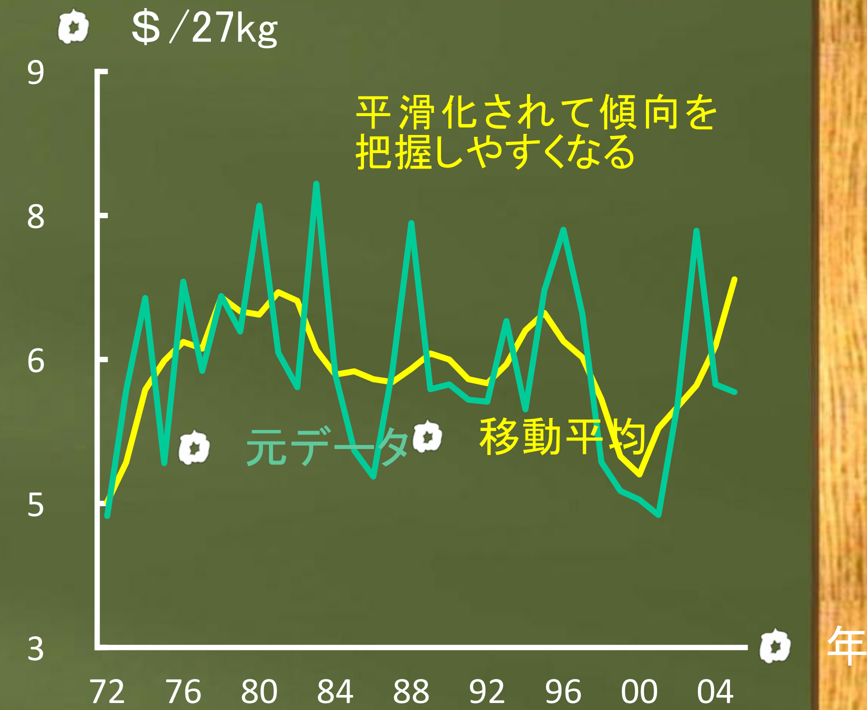
	大豆価格（シカゴ）	移動平均（5項）
1970	2.85	
1971	3.03	
1972	4.37	4.51
1973	5.68	4.93
1974	6.64	5.68
1975	4.92	5.99
…	…	…
2003	7.34	5.73
2004	5.74	6.14
2005	5.66	6.83
2006	6.43	
2007	9.00	

1970～1974の5年間※の平均

1年ずらした5年間の平均

同様に，1年ずらした平均

最初と最後の2年ずつ失う



※何項（年）の平均を取るかは自分で決める

Excel分析ツールにも「移動平均」が搭載

1.4　代表値②　ーバラツキの指標ー

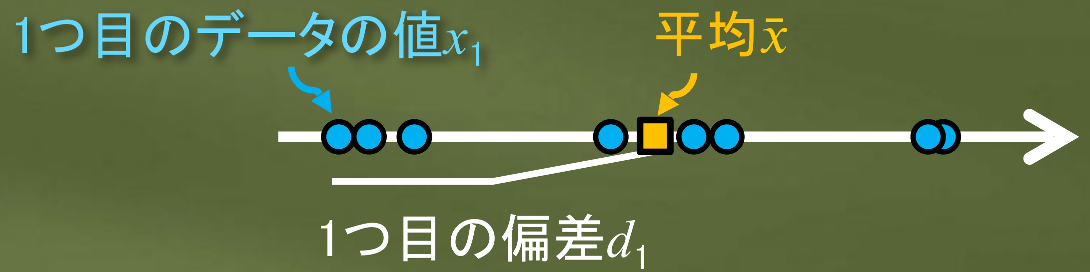
- データの特性として平均だけでは情報不足
- どのぐらいバラついているのかも知りたい
- バラツキの指標（統計量）として,
  偏差 → 偏差平方和 → 分散 → 標準偏差
  → 変動係数,　の5つを順に解説

偏　差

（もっとも基本的なバラツキの指標）

- 各データ値と平均との差（個別データが平均からの偏り度）

偏差　$d_i = x_i - \overline{x}$



偏差の欠点：
①データの数だけ示さなければならない
②統計量として1つの値にまとめようと足し合わせると，±値が混在
しているため，相殺されて0になってしまう

偏差平方和と分散

- 偏差を2乗して足し合わせれば０にならない

偏差平方和（変動） $S = \sum (x_i - \bar{x})^2$ Excel関数＝DEVSQ

欠点：標本サイズnが大きいと巨大な値になってしまう

- 標本サイズで割れば解決（サイズに応じて小さくなる）

分散※ $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n}$ Excel関数＝VAR.P

欠点：単位が2乗のままなので，値はまだ大きい

※推測統計学では，母集団の分散はσ^2で表して区別します。

標準偏差

- 分散の平方根を取ることで，元の単位に戻り，大きさもちょうど良くなる

標準偏差（SD） $s = \sqrt{\text{分散}} \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

Excel関数＝STDEV.P

欠点：値の大きさが極端に異なる集団間や，単位が異なる集団間のバラツキは比較できない

平均値の大きな集団の標準偏差が大きくなってしまう

変動係数

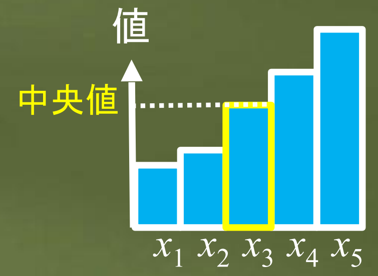
- 標準偏差を平均で割ることで，集団間の平均を揃えて，単位を取り去る（無名数にする）

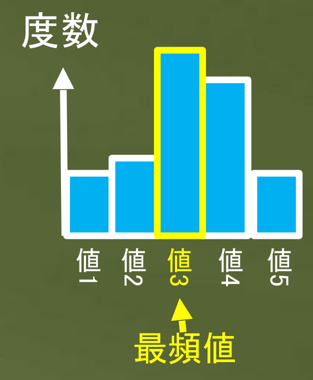
変動係数 $$CV = \frac{\text{標準偏差}}{\text{平均}} = \frac{s}{\bar{x}} = \frac{\sqrt{\sum (x_i - \bar{x})^2 / n}}{\bar{x}}$$

注意：5種類のバラツキの統計量は，それぞれ有する情報量が異なるので，分析に適したものを使い分ける　（例：偏差ならば分布形もわかるが，変動係数は単位さえわからない）

1.5　質的データの代表値

- 質的データは平均$\bar{x}$の計算が許されない（分散も不可）
- 平均に代えて，カテゴリカルデータでは最頻値を，順位データでは最頻値や中央値を計算する
- 中央値：データを大きさ順に並べたとき中央にくる値
  Excel関数＝MEDIAN
- 最頻値：最も多く現れる（最大度数を持つ）値
  Excel関数＝MODE
- バラツキは図や表に表して視覚的に確認する

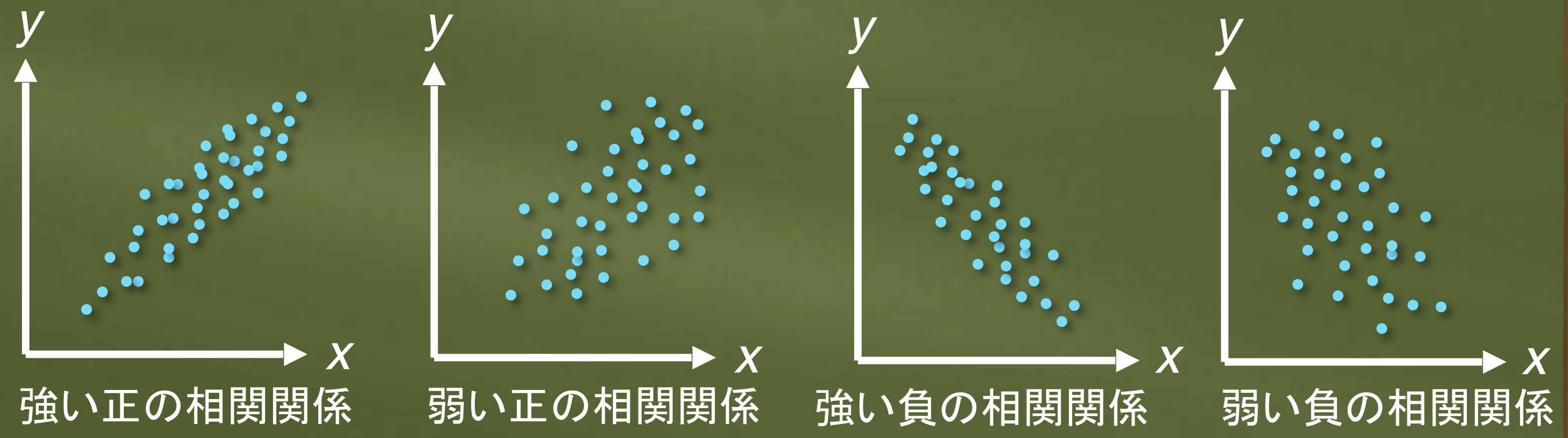




1.6　相関係数と共分散

－２つの変数の関係－

- ２変数間で，片方が大きくなるに従ってもう片方も大きくなる場合，「正の相関関係がある」という（小さくなる場合は負の相関関係）
- 相関関係の強さを示す統計量が相関係数と共分散



共分散

- 2変数の偏差の積の平均で相関関係を捉える

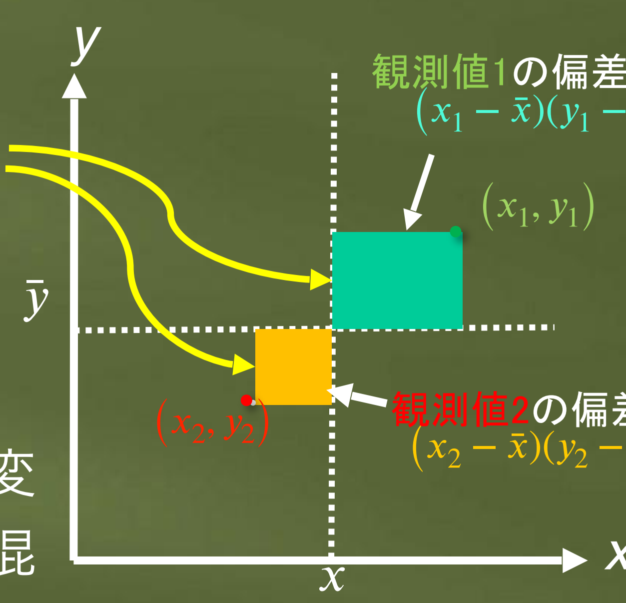
偏差の積の総和は，正の相関があるときには正となる（負の相関があるときには負となる）

$$\sum (x_i - \overline{x})(y_i - \overline{y})$$

標本サイズに左右されないようにnで割る

共分散 $$s_{xy} = \frac{1}{n}\sum (x_i - \overline{x})(y_i - \overline{y})$$

欠点：バラツキによって値が大きく変わってしまう（異なる単位も混在）



Excel関数：COVARIANCE.P

相関係数

- どちらの偏差も標準偏差sで割ってバラツキをそろえる

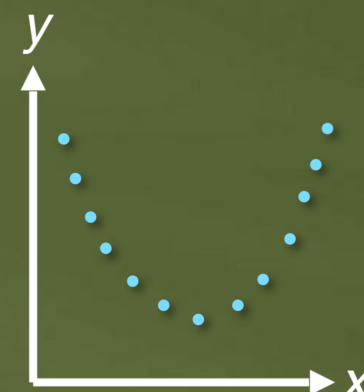
相関係数 $$r_{xy}=\frac{1}{n}\sum\frac{(x_i-\bar{x})}{\boxed{s_x}}\cdot\frac{(y_i-\bar{y})}{\boxed{s_y}}=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

相関係数の性質：

①-1から1までの値を取る（0は無相関）が,
いくつ以上が強い相関という基準はない

②あくまで2変数の直線的な関係を捉えるだけ
（右図のような関係があっても捉えられない→）

Excel関数＝CORREL



相関係数r_{xy}≒0

以上で第１章は終了です。