第2講　推測統計学の四つの重要概念

学びのポイント

- □ **母集団**、**標本**、**統計量**、**標本分布**という四つの概念とその関係が分かる。
- □ 母集団分布に多く仮定される**正規分布の性質**が分かる。
- □ 標本は、母集団から**サンプリング（標本抽出）**されたものだと仮定されることがわかる。
- □ サンプリングに**確率的挙動**が関わることが分かる。
- □ **統計量**がデータから算出される量であることが理解でき、サンプリングに伴う確率的振る舞いをすることが分かる。
- □ **標本分布**が統計量の示す確率的挙動を表すことが分かる。
- □ サンプルサイズが大きい時、**標本平均**が従う標本分布は、**母集団分布にかかわらず正規分布**になることが分かる。
- □ サンプルサイズを大きくすれば、**標本分布の標準偏差**（標本誤差）が小さくなり、母平均の近似となることが分かる。
- □ 統計量の望ましい性質の一つに**不偏性**があることが分かり、定義を説明できる。
- □ 基本的な統計量を理解でき、その計算ができる；最頻値、中央値、平均；範囲、四分位範囲、平均偏差、標準偏差、分散；共分散、相関係数；回帰係数。

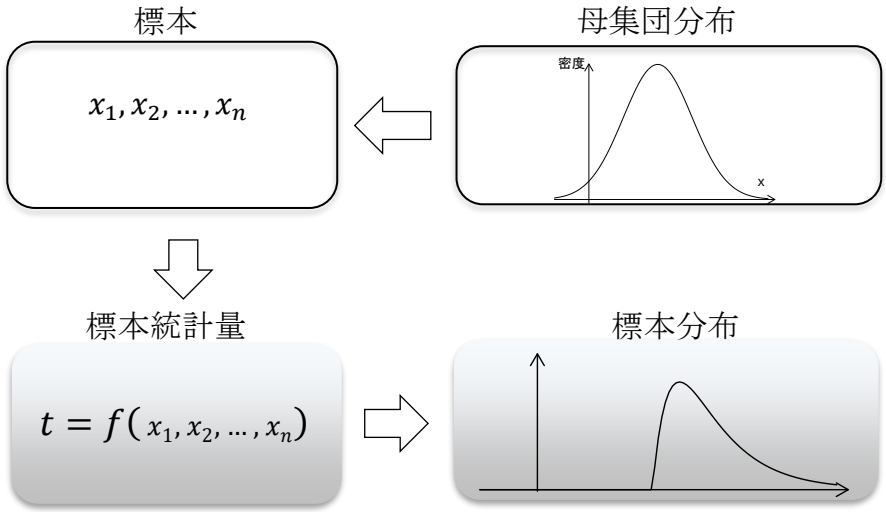
本講では、推測統計学という枠組みで習う四つの重要な概念を学びます。それは、母集団、標本、統計量、そして標本分布という概念です。

前半では、母集団から標本が得られ、そこから統計量が計算され、その統計量は標本分布と呼ばれる確率的挙動を示すことを説明します。これは、推測統計学の土台となる考え方なので、確実に自分の知識にしてください。

後半では、有名な統計量を学びます。第一は、データの中心を表すのに使う統計量（最頻値、中央値、平均）です。第二は、データのばらつきを表す統計量（範囲、四分位範囲、平均偏差、標準偏差、分散）です。第三は、二つの変数の関係を表す統計量（共分散、相関係数、回帰係数）です。

これらの概念は、今後の授業で何度も登場する“主要登場人物”です。あやふやな理解にならないよう、しっかり学習し、わからない点は次講に進む前に解消してください。

見取り図



（1）標本 Sample

得られたデータのことを標本（サンプル）と呼ぶ。多くの場合全てを報告するのではなく何らかの指標で標本全体を代表させる。

（2）母集団 Population

研究者が想定する「考えられる対象をすべて含む集団」。

※標本が抽出される際に、確率的な揺らぎが生じる。

（3）（標本）統計量 Sample Statistic

☞（標本）統計量

統計量とは、ある任意の関数fが標本$x_1, x_2, \ldots, x_n$を引数にとったときに返す返り値のことである。

$$T = f(x_1, x_2, \ldots, x_n)$$

標本から計算される指標を（標本）統計量と呼び、とりわけデータをうまく要約している統計量を 代表値 と呼ぶ。

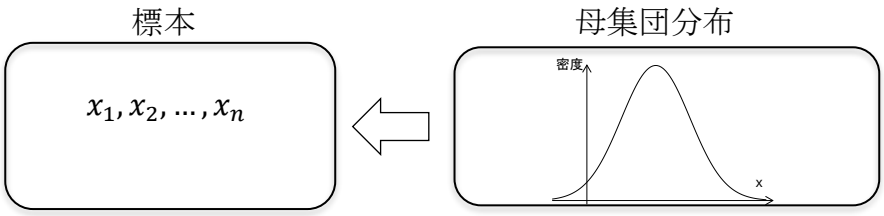
例：平均

$$\frac{x_1 + x_2 + \cdots + x_n}{n}$$

（4）標本分布 Sampling distribution

母集団から標本がサンプリングされる際の確率的な揺らぎのせいで、統計量は一つの値には決まらない。その結果統計量が見せる確率的挙動を表したもの。

🕮 ノート１　母集団分布



（１）標本とヒストグラム

① 標本（サンプル）

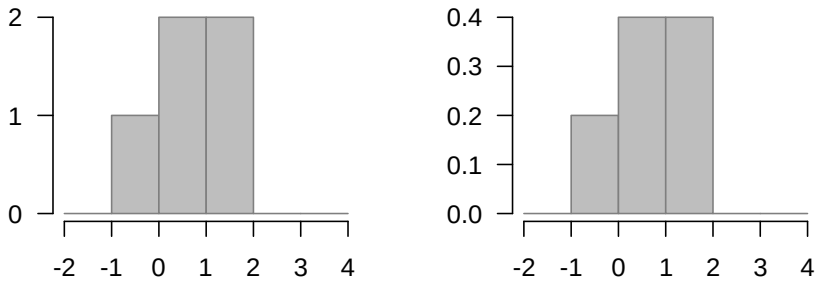
これは、研究者が観測するデータのこと。

② サンプルサイズ

これは、標本の中に含まれている観測値の数のこと。

③ ヒストグラム（柱状図）

これは、設定された各区間にどのくらいの**頻度**で観測値が存在しているのかを柱で表したグラフ。



ヒストグラムと棒グラフは同じですか？

違います。ヒストグラムは「区間」に対してその頻度を長さで表現していますが、棒グラフは「点（値）」に対してその大きさを長さとして表しています。

（2）母集団分布

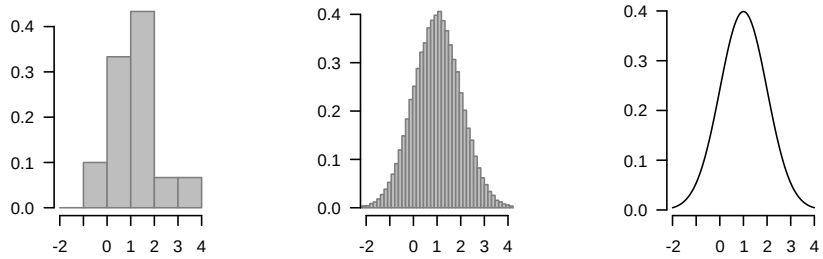
① 母集団

研究者が想定する「考えられる対象をすべて含む集団」。

※標本が抽出される際に、確率的な揺らぎが生じる。

② 母集団分布

標本のヒストグラムの行きつく先。母集団の中心やばらつきが表されている。



※ 確率分布

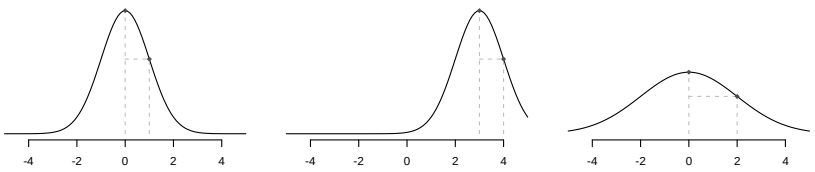
「区間」に対応する確率を示しているグラフのこと。

③ （確率）密度関数 Probability density function

確率分布に示された曲線を表すグラフの式。

（3）正規分布

ランダムな誤差が積み重なると生まれる左右対称の釣鐘状の形状をした確率分布。$N(\mu, \sigma^2)$のように表す。



（特徴１）数学的に扱いやすい！

（特徴２）無標な分布

📖 ノート２　統計量と標本分布

（１）統計量と標本分布

① 統計量 Statistic

これは、標本に基づいて計算された量。

② 標本分布 Sampling Distribution

これは、統計量が描くヒストグラムの行き着く先。

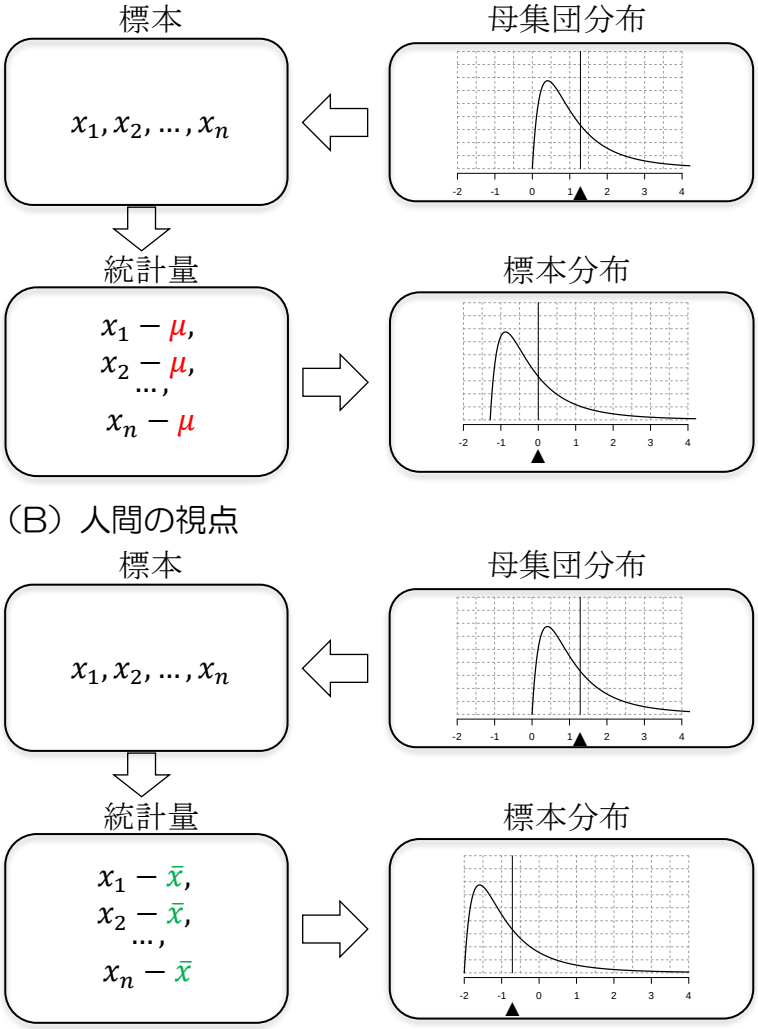
（２）統計量と標本分布の具体例

① 操作１：中心化

平均を引き、データの中心を動かす平行移動のこと。

データ　－ 中心

（A）全知全能の視点



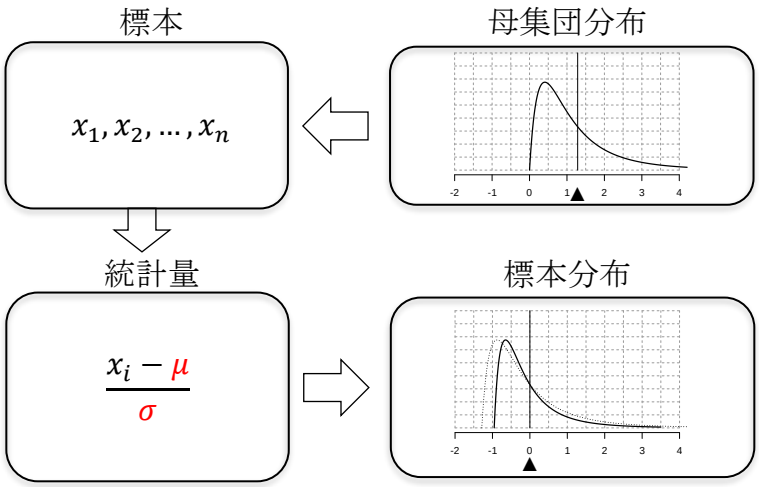
（B）人間の視点

② 操作２：標準化

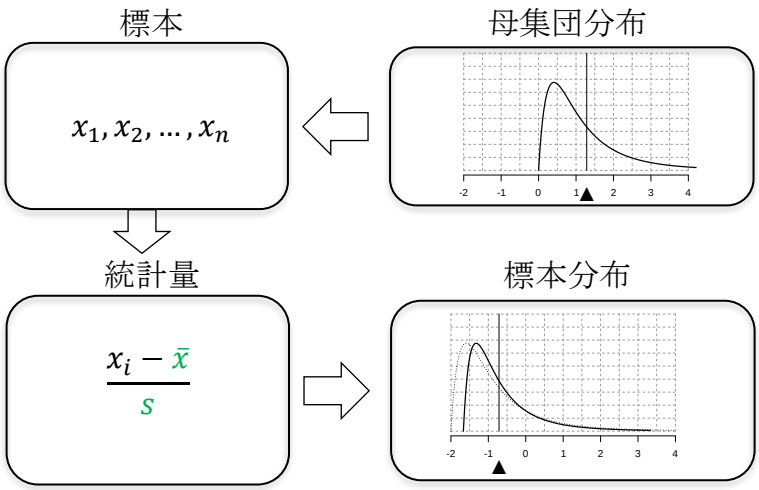
平均を引き、データの中心を動かす平行移動をした値を、ばらつきの尺度である標準偏差で割ること。

$$\frac{\text{データ} \quad - \text{中心}}{\text{ばらつきの尺度}}$$

（A）全知全能の視点



（B）人間の視点

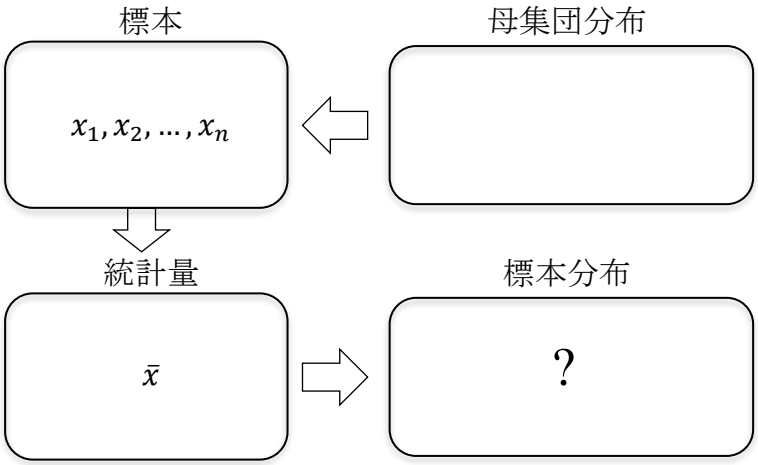


（3）中心極限定理 Central Limit Theorem

これは、平均という統計量がしたがう標本分布が正規分布になるという定理。母集団の形状で二種類のものがある。

① 具体例

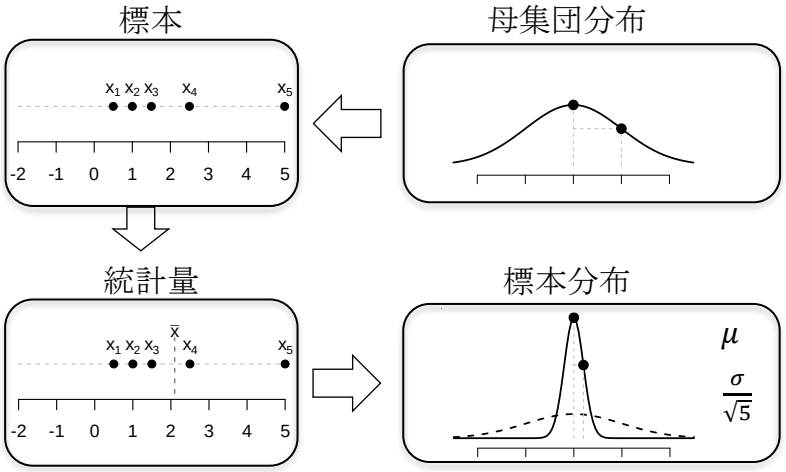
「大学入試共通テスト」を受けたi番目の生徒の英語の点をx_iとする。このときクラス平均がしたがう分布は？



② バージョン１：母集団分布が正規分布のとき

任意の母集団では、サンプルサイズ n が大きくなくても、$\bar{x}$ の標本分布は$N\left(\mu, \frac{\sigma^2}{n}\right)$になる。

サンプルサイズｎが小さいケース　　$n = 5$



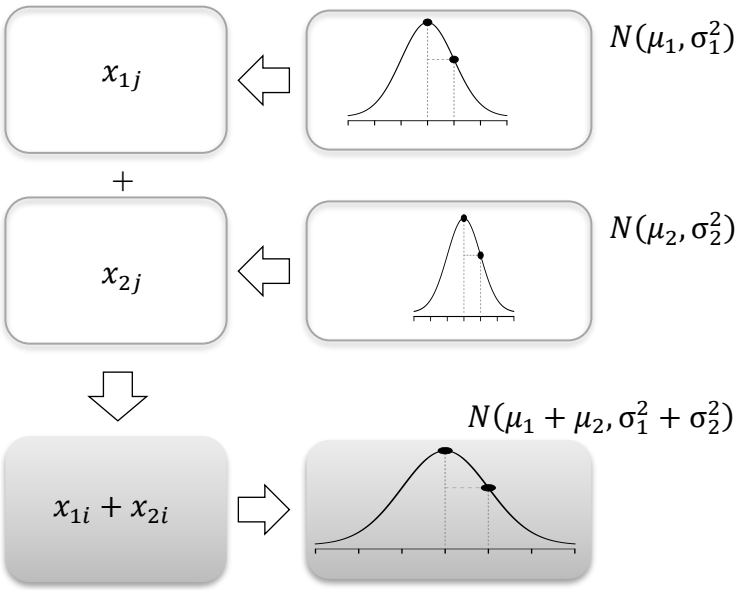
再生性（正規分布）

（ケース１）別々の正規分布に従う場合

$$x_{1j} \sim N(\mu_1, \sigma_1^2)$$
$$x_{2j} \sim N(\mu_2, \sigma_2^2)$$
$$\vdots$$
$$x_{nj} \sim N(\mu_n, \sigma_n^2)$$

+)

$$x_{1j} + x_{2j} + \cdots + x_{nj} \sim N(\mu_1 + \mu_2 + \cdots + \mu_n,\ \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$$



（ケース２）同一の正規分布に従う場合

$$x_{1j} \sim N(\mu, \sigma^2)$$
$$x_{2j} \sim N(\mu, \sigma^2)$$
$$\vdots$$
$$x_{nj} \sim N(\mu, \sigma^2)$$

+)

$$x_{1j} + x_{2j} + \cdots + x_{nj} \sim N(n\mu, n\sigma^2)$$

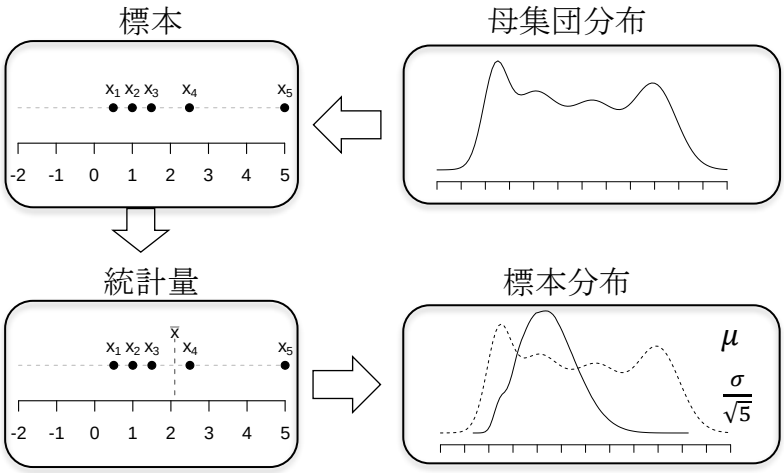
↓サンプルサイズ n で割る

$$\frac{1}{n}(x_{1j} + x_{2j} + \cdots + x_{nj}) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

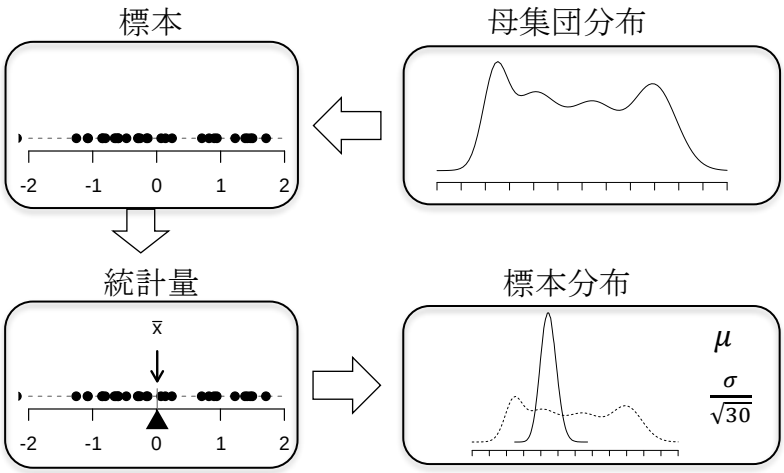
③ バージョン２：母集団分布が正規分布以外のとき

任意の母集団では、サンプルサイズ n が大きければ大きいほど、$\bar{x}$ の標本分布は$N\left(\mu, \frac{\sigma^2}{n}\right)$ に近づいていく。

サンプルサイズｎが小さいとき　　$n = 5$



サンプルサイズｎが大きくなってきたとき　　$n = 30$



数学が苦手な人のために

（ 英語 ）単語・句・節

I remember [↑] .

that

visiting my friends

that I visited my friends

（ 数学 ）数式の中における埋め込み

ここに入る表現が複雑化する

$$\frac{\boxed{\text{データ}} - \boxed{\text{中心}}}{\boxed{\text{ばらつきの尺度}}}$$

この後何度も使うことになる操作

（1）標準化
データから中心を引いて、それをばらつきの尺度で割る

（2）中心極限定理
母集団が正規分布だと、平均の標本分布は正規分布になる

（3）（正規分布の）再生性
正規分布に従う変数を足した和も（差も）正規分布になる

🕮 ノート３　基本的な統計量

■　基本的な統計量Ⅰ（データの　中心）

（１）（標本）最頻値 Sample Mode

これは、 頻度 という観点からデータの真ん中を定めたもので、得られたデータの中で最も多く得られた値のこと。

（２）（標本）中央値　Semple Median

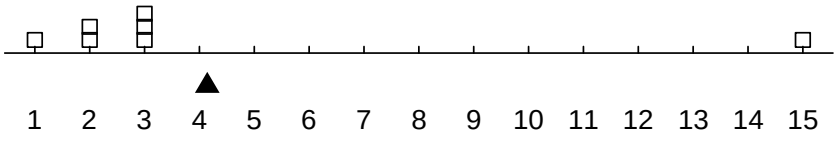
これは、 順位 という観点からデータの真ん中を定めたもので、小さい順に並べた時の真ん中にくる値のこと。

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{\left(\frac{n}{2}\right)} \leq x_{\left(\frac{n}{2}+1\right)} \leq \cdots \leq x_{(n)}$$

$$MD = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\text{が奇数のとき} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & n\text{が偶数のとき} \end{cases}$$

（３）（標本）平均 Sample Mean

これは、バランス という観点から真ん中を定めたもので、値をすべて足して総数で割った値のこと。



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

データとの距離の最小化

① 中央値：一次の距離の最小化

「t と各点x_iの　差の絶対値　の総和を最小化する」という基準を満たすデータの中心。

⇒ すなわち、次の基準T_1を最小化する t が中央値。

$$T_1 = |x_1 - t| + |x_2 - t| + \cdots + |x_n - t|$$

$$= \sum_{i=1}^{n} |x_i - t|$$

② 平均：二次の距離の最小化

「t と各点x_iの　差の二乗　の総和を最小化する」という最小二乗基準 を満たすデータの中心。

⇒ すなわち、次の基準T_2を最小化する t が平均。

$$T_2 = (x_1 - t)^2 + (x_2 - t)^2 + \cdots + (x_n - t)^2$$

$$= \sum_{i=1}^{n} (x_i - t)^2$$

それぞれの指標のメリットとデメリット

（特徴１）中央値、平均と違い最頻値は存在しないときがある。

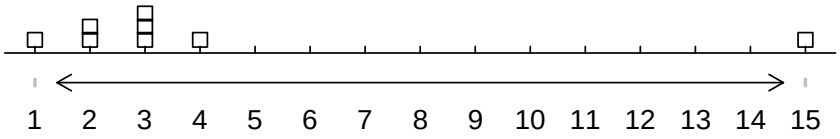
（特徴２）中央値と違い平均は外れ値の影響を受けやすい。

（特徴３）中央値と違い平均は数学的な取り扱いが楽であり、
かつ、いろいろときれいな性質がある。

■ 基本的な統計量Ⅱ（1 次データの ばらつき）

（1）標本範囲（最小統計量と最大統計量）Sample Range

これは、最大値（$x_{(n)}$）と最小値（$x_{(1)}$）に注目し、その差を測って得られる値のこと。

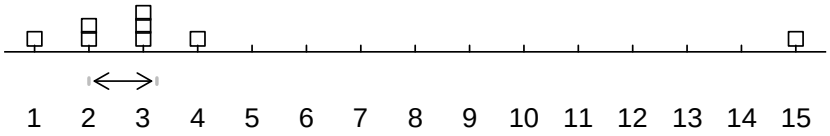


$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

$$Rg = x_{(n)} - x_{(1)}$$

（2）四分位範囲 Interquartile Range

これは、データを小さい順に並べたときの下位 25%（第 1 四分位点 Q1）、75%（第 3 四分位点 Q3）に位置するデータの距離（L1 距離）を出したものです。

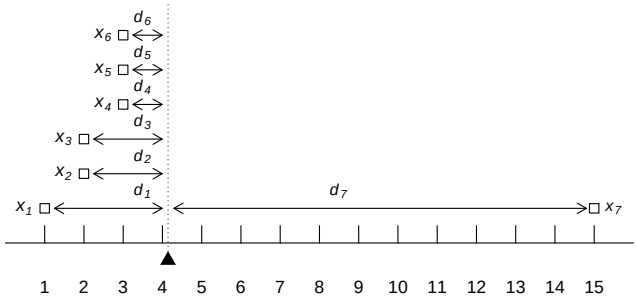


$$\mathrm{IQR} = Q_3 - Q_1$$

※ 箱ひげ図

これは、最大値、最小値、第 1～3 位四分位点を表す図。**外れ値**　を示すために、IQR の 1.5 倍の範囲の外にあるものは、独立して表す。

（3）平均偏差（一次の距離に基づく指標）



① 偏差 Deviation

これは、一次距離　で測った平均からの距離。

$$d_i = x_i - \bar{x}$$

② 平均偏差 Mean Deviation

これは、偏差の平均。一次の距離　に基づいた指標。

$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

（4）標本分散（二次の距離に基づく指標）

① 偏差平方和

これは、二次距離　で測った平均からの距離。

$$d_i^2 = (x_i - \bar{x})^2$$

$$SSD = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

② 分散 Variance

これは、偏差平方和の平均。

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

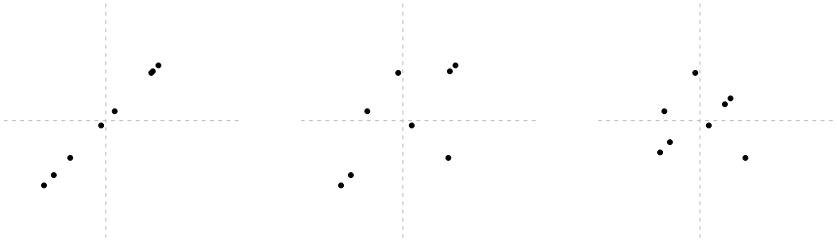
③ 標準偏差 Standard Deviation

これは、分散の平方根。

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

■　基本的な統計量 III（二つの変数の 相関）

【目的】x 軸と y 軸の関連性の方向と大きさを測る指標を作りたい。



（1）共分散

これは、平均と各点との間の **二次の距離（面積）** の平均。

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

※共分散には最大値と最小値が存在する。

$$-s_x s_y \leq s_{xy} \leq s_x s_y$$

（2）相関係数（ピアソンの積率相関係数）

これは、共分散が最大値と比較したときに占める割合。

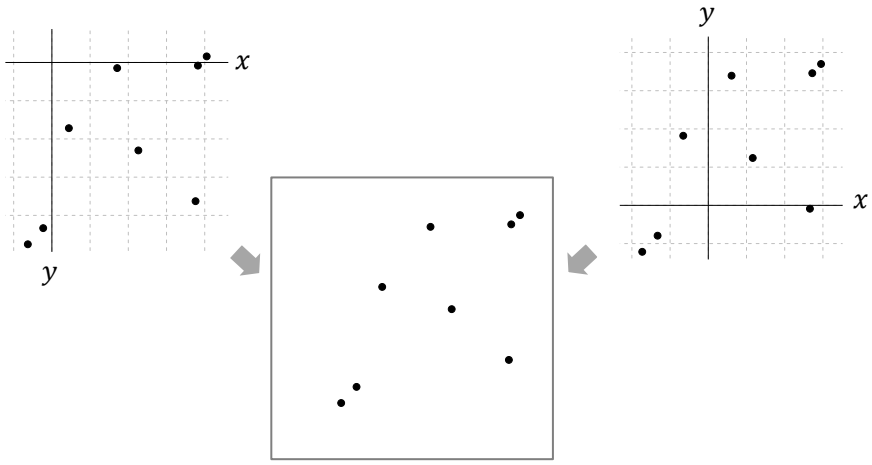
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

※相関係数には最大値と最小値が存在する。

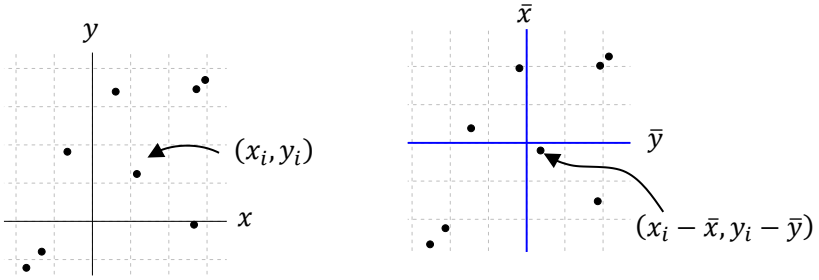
$$-1 \leq r_{xy} \leq 1$$

（準備）軸を取り払う

もともとの x 軸、y 軸の位置は関連度合いには無関係なので、取っ払ってしまいたい。

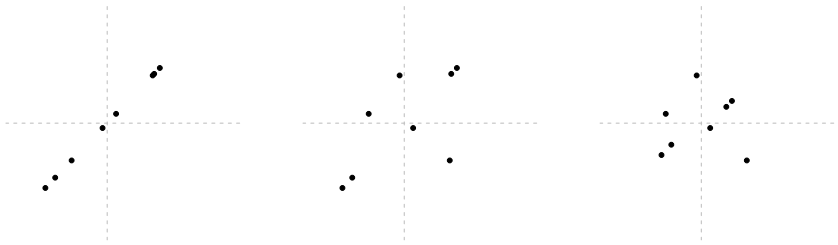


しかし、座標自体は維持したいので、新しく、横軸を$\bar{y}$、縦軸を$\bar{x}$の位置に据えた新しい座標を考える。これはもともとの座標を平行移動したことに相当する。



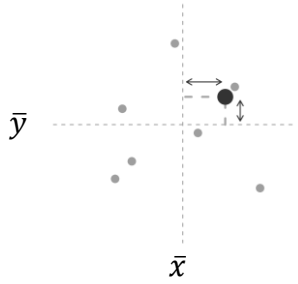
（第一案）球の個数の差を考える

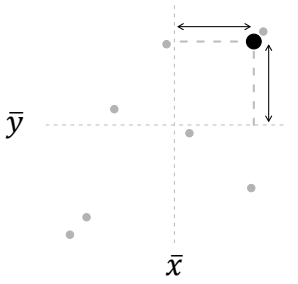
（式１）　「第一／三象限の球の数」－「第二／四象限の球の数」



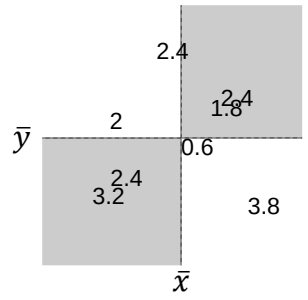
問題点：ケース B とケース C の違いを捉えられない。

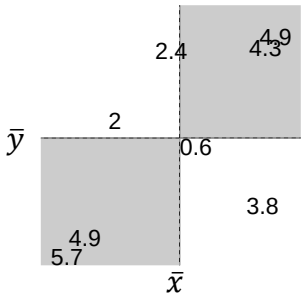
（第二案）中心からの離れ具合を「足し算」で捉える





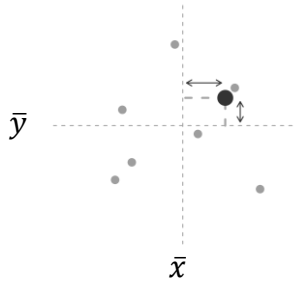
（式２）「灰色の領域の数字の和」－「白色の領域の数字の和」

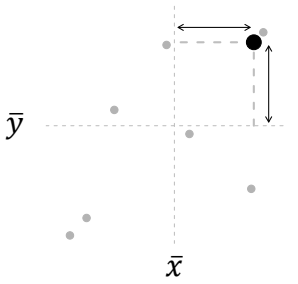




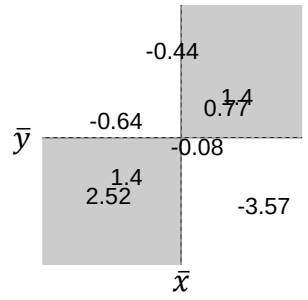
問題点：数式に場合分けが存在してしまう。

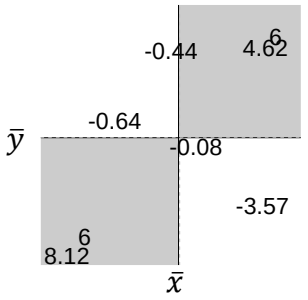
（第三案）中心からの離れ具合を「掛け算」で捉える





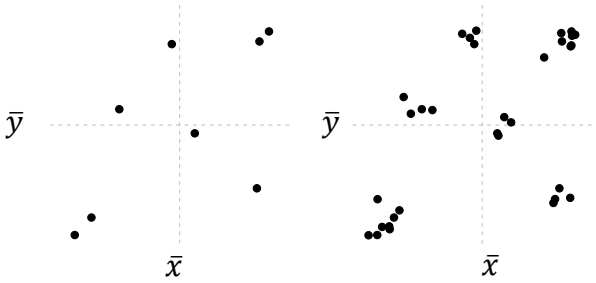
（式２）　「灰色の領域の数字の和」－「白色の領域の数字の和」





問題点：サンプルサイズに依存してしまう。

（第四案）サンプルサイズの影響を無くす　採用！

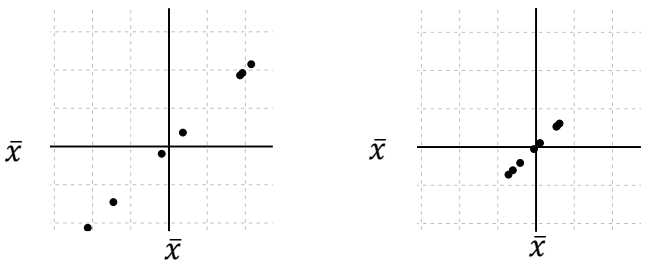


（式３）　$\frac{1}{n}\times$(「奇数象限の数字の和」－「偶数象限の数字の和」)

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

共分散の特殊なケースとしての分散

xとyという二変数の間ではなく、xとxという自分自身との間の共分散を考えてみたものが分散である。



この値が大きいということは、その変数のばらつきが大きいということ。そこで、分散はばらつきの指標として使われる。

資料2-1　　様々な「平均」

①算術平均 (arithmetic mean)、別名：相加平均
これは、もっとも一般的な「平均」で、すべての値を足し、総数で割ったもの。

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

②加重平均 (weighted average)
これは、各値にそれに対応する重みを付けたもの。

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

算術平均は、次のように書けるので、加重平均の特殊な場合とみなすことができる。あるいは、加重平均は、重みが$w_i = 1/n$だという制約を取り払った、一般化された算術平均だとも言える。

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

③調和平均 (harmonic average)
これは、逆数の平均。統計学では、分散の逆数（これを精度と呼ぶ）という概念がしばしば登場する。

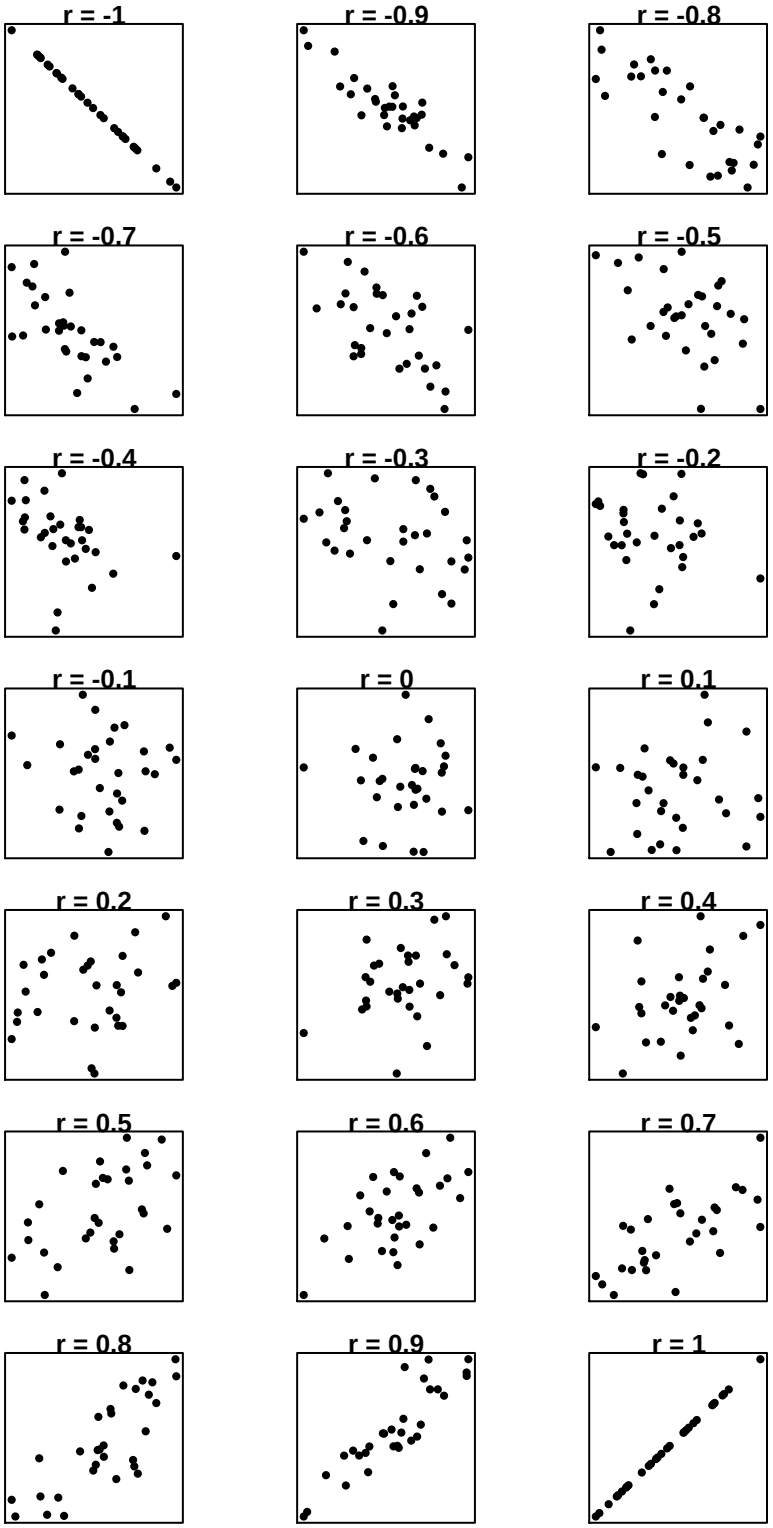
$$\bar{x}_{調} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}$$

④幾何平均 (geometric average)、別名：相乗平均
これは、積に関する平均。つまり、一回当たり、平均してどれくらいの値をかけていたのかを表す。

$$\bar{x}_{幾} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

例：$(2 \times 4 \times 8)^{\frac{1}{3}} = (4 \times 4 \times 4)^{\frac{1}{3}} = 4$

資料2-2　　共分散／相関係数と散らばり



資料2-3　　分散の分解

$$s_{x+y}^2 = s_x^2 + 2s_{xy} + s_y^2$$

【証明】

$$\begin{aligned}
s_{x+y} &= \frac{1}{n}\sum_{i=1}^{n}\{(x_i + y_i) - \overline{(x+y)}\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \bar{x}) + (y_i - \bar{y})\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(x_i - \bar{x})^2 + 2(x_i - \bar{x})(y_i - \bar{y}) + (y_i - \bar{y})^2\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}2(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= s_x^2 + 2s_{xy} + s_y^2
\end{aligned}$$