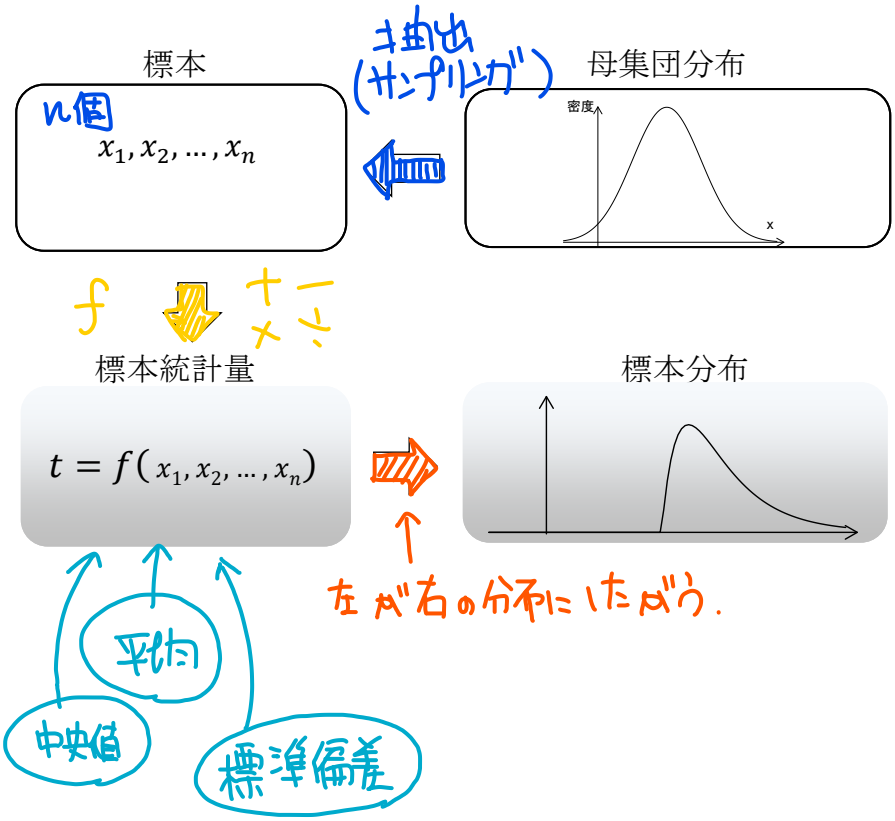
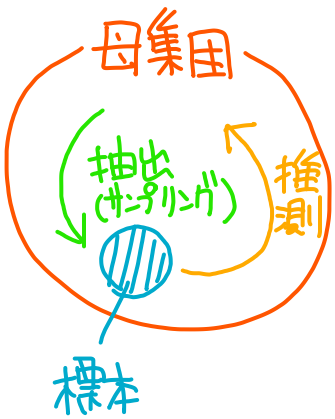
見取り図


標本
n個
$x_1, x_2, \dots, x_n$
抽出
(サンプリング)
母集団分布
密度
x
f
+ − × ÷
標本統計量
$t = f(x_1, x_2, \dots, x_n)$
標本分布
左が右の分布にしたがう.
平均
中央値
標準偏差



（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。

サンプル
標本（観測データ）を ＋ − × ÷ してつくった量のこと.

（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 $x_1, x_2, \dots, x_n$ を引数にとったときに返す返り値のことである。

データ
入力値.

出力

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

統計量　変換　入力値.

関数f
これは，入力値がどのように出力値に変換されるかを示したもののこと.

$f(x) = 2 \times x$

入力値　出力値.

$f(x) = 5 + x$

データという入力値　統計量.

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

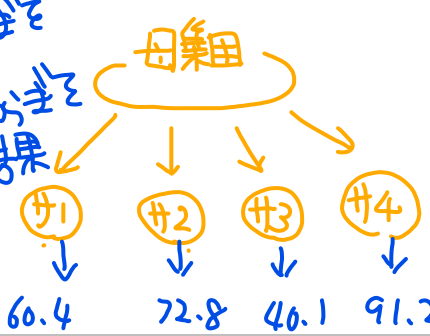
$$\text{平均} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

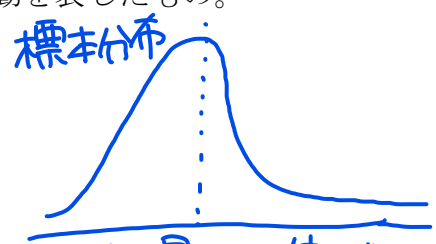
これは，データに ＋ と ÷ をあてがって新しい量を生成したもの.

（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。

サンプリングというプロセスが確率的ゆらぎを招くので，結果的に統計量も確率的ゆらぎを内包してしまう．その結果分布がでまるという意味

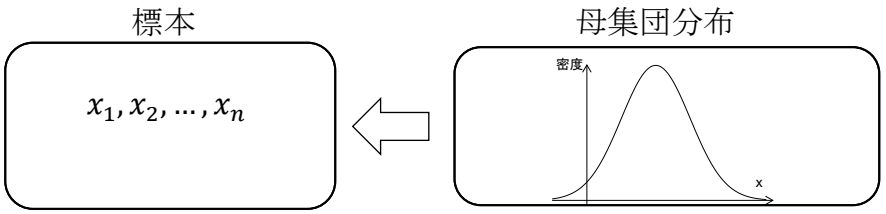


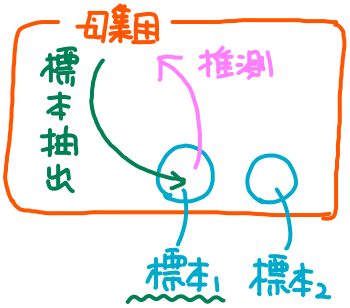


統計量のとる値がどのような挙動を示すのか確率分布にしたもの.

📖 ノート１　母集団分布

Ⓟ 母集団と標本





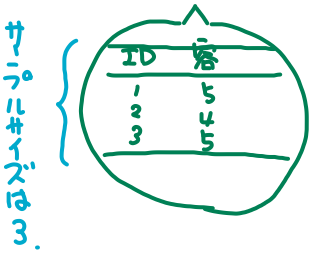
（１）標本とヒストグラム

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**

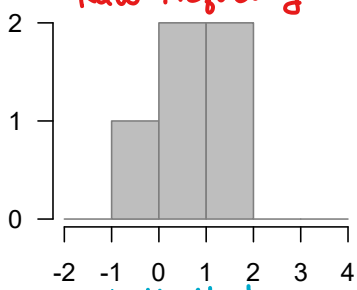
これは、標本の中に含まれている観測値の数のこと。



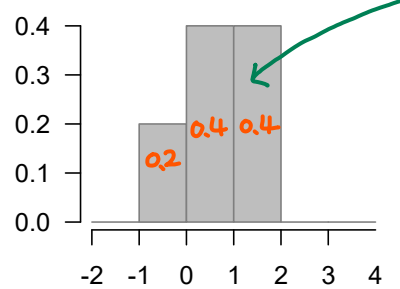
③ **ヒストグラム（柱状図）**

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。

(1) 粗頻度：回数を y 軸
Raw Frequency



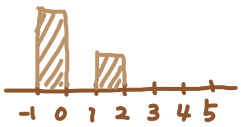
(2) 相対頻度：比率を y 軸



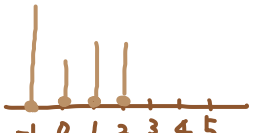
1~2の区間に2回 → 2/5 = 0.4
0~1 〃 2回 → 2/5 = 0.4
-1~0 〃 1回 → 1/5 = 0.2
⇒ 全区間 5回

Ⓟ 面積が割合！
灰色の面積の合計
＝１（100%）

Ⓟ 棒グラフとの違い
(1) ヒストグラム



区間に対して定まる
(2) 棒グラフ



点（値）に対して定まる

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。

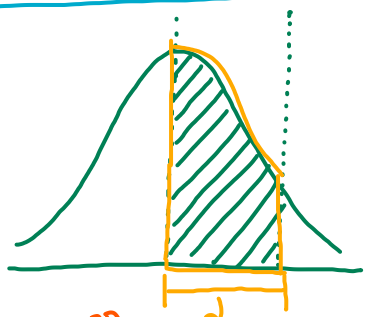
（2）母集団分布

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。

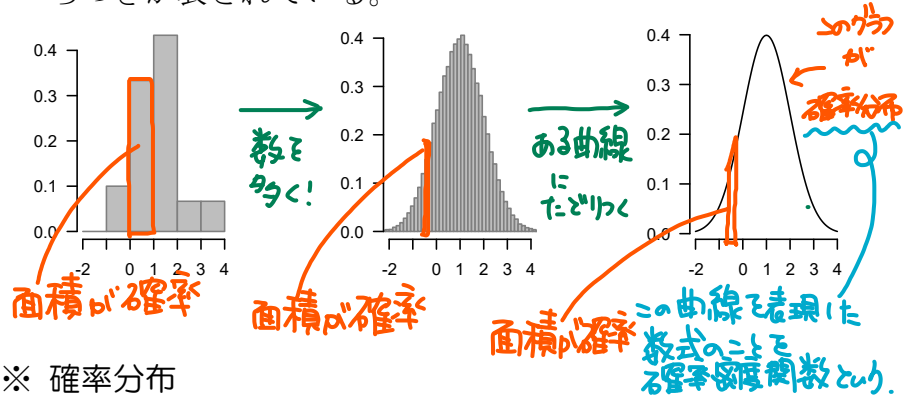
Ⓟ確率密度関数の見方



この区間に入る割合(確率)を面積で表現している！

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。

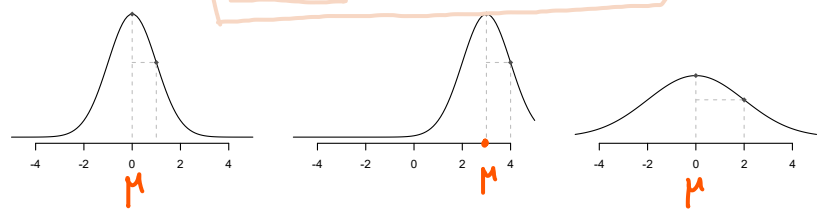
（3）正規分布

ほーっておくと、勝手に生まれている分布

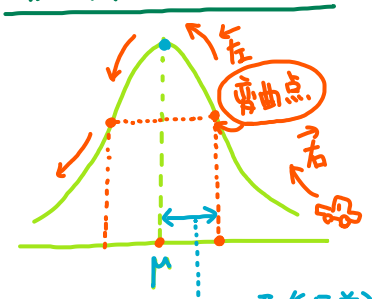
ランダムな誤差が積み重なると生まれる左右対称の釣鐘状の形状をした確率分布。$N(\mu, \sigma^2)$のように表す。

Normal Distribution　シグマ（ロ-マ字のSに対応）　ミュー（ロ-マ字のmに対応）

(例1) $\mu = 0$, $\sigma = 1$　(例2) $\mu = 3$, $\sigma = 1$　(例3) $\mu = 0$, $\sigma = 2$



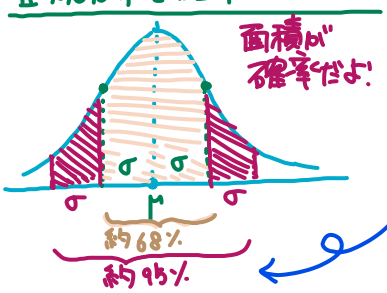
Ⓟ変曲点とσ



① μ：位置パラメータ（分布の中心）

② σ：尺度 〃　（標準偏差）（分布の広がり）

Ⓟ正規分布と確率



（特徴１）数学的に扱いやすい！

（特徴２）無標な分布

注 より厳密には、1.96×σの区間が全体の95%に対応しているヨ。

📖 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。

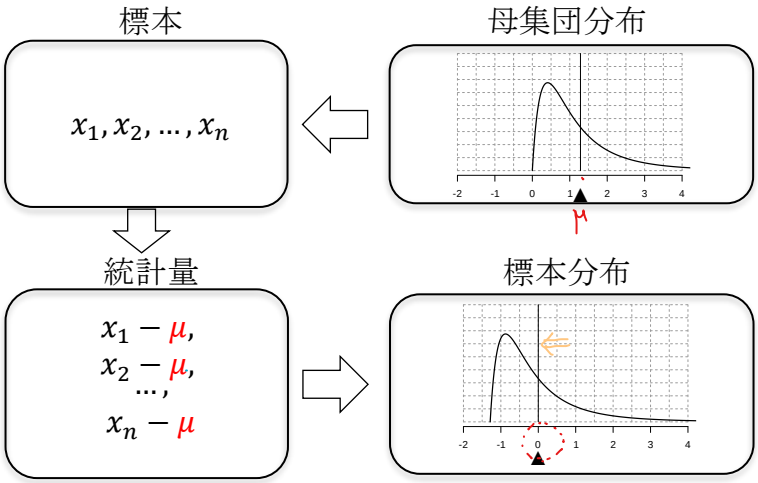
（２）統計量と標本分布の具体例

① **操作１**：中心化

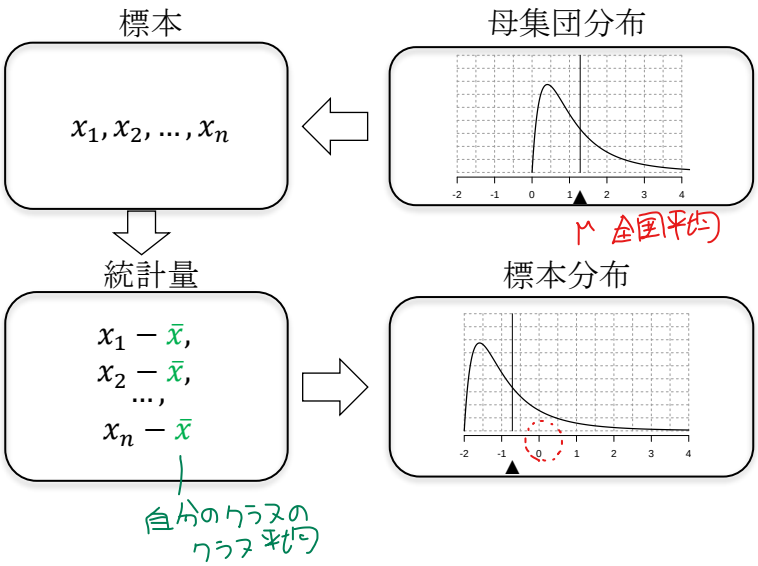
平均を引き、データの中心を動かす平行移動のこと。

データ　－ 中心

（A）全知全能の視点



（B）人間の視点



Ⓟ具体例

x_1：出席番号１の人の点数

x_2：〃　２　〃

⋮

x_n：〃　n　〃

↓

標本のサンプルサイズ

Ⓟ記法：$\bar{\ }$（バー）

ある変数の平均を

$\bar{\ }$（バー）をつけて表す

（例）$x_1, x_2, \cdots, x_n$

↓

$\bar{x}$：エックス・バー

Ⓟ 割り算をすること

これは、分母にきている「単位」何個分なのかを計算すること．

120個のアルミ缶

1ケース12個はいる

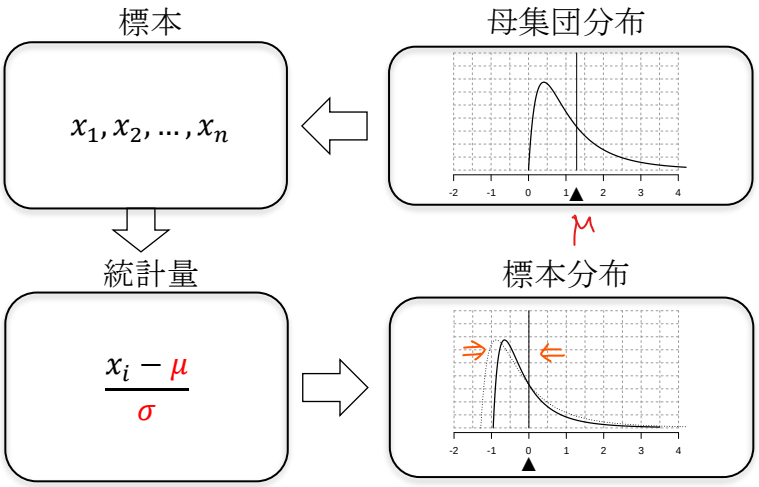
＝10ケース分の缶がある

② 操作２：標準化

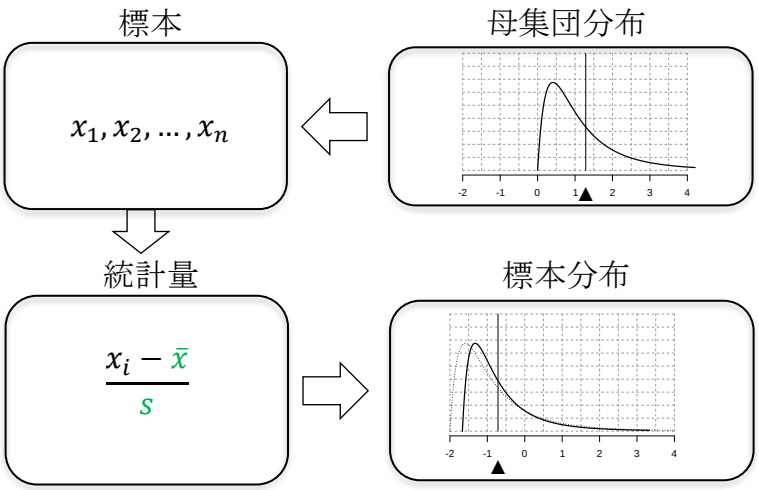
平均を引き、データの中心を動かす平行移動をした値を、ばらつきの尺度である標準偏差で割ること。

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

（A）全知全能の視点



（B）人間の視点

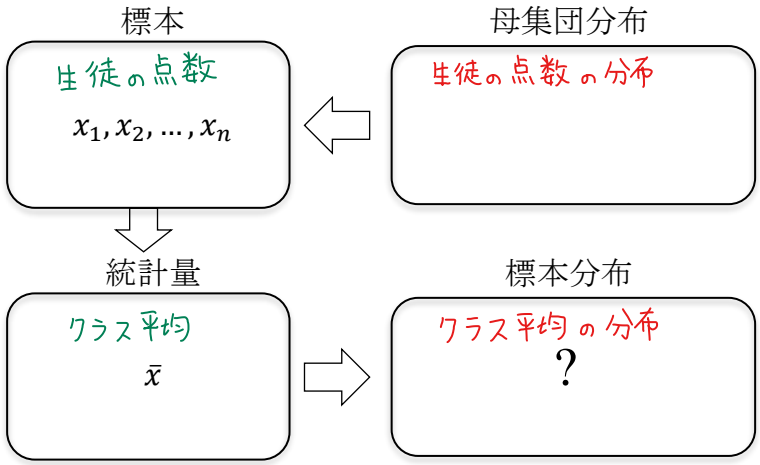


（3）中心極限定理 Central Limit Theorem

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=1

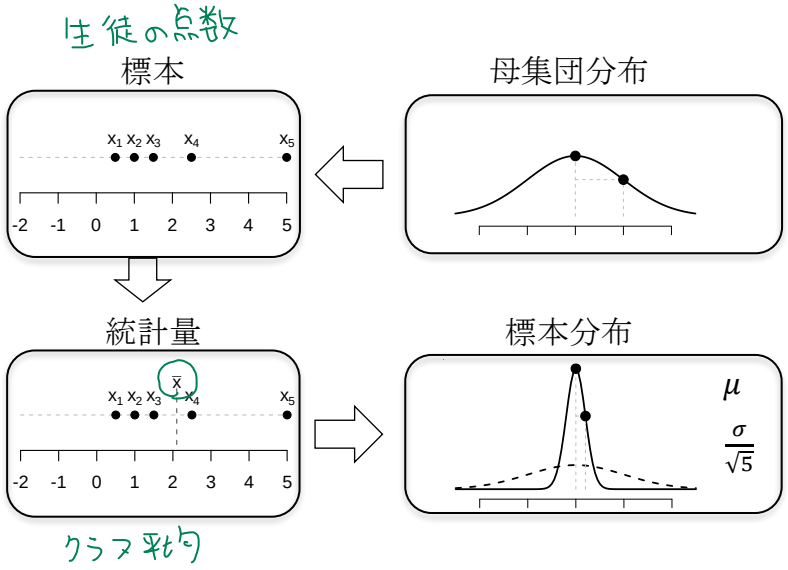
これは，すべてのクラスサイズがマンツーマンだという個別指導の状況を表わしている．

$\bar{x} = x_1$

$x_1 \sim N(\mu, \sigma^2)$

$\bar{x} \sim N(\mu, \sigma^2)$

サンプルサイズｎが小さいケース　　$n = 5$



再生性（正規分布）

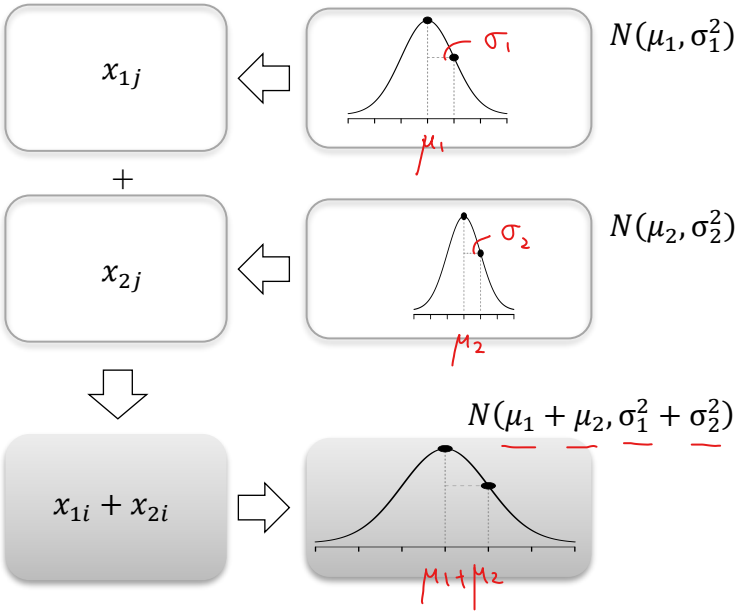
（ケース１）別々の正規分布に従う場合

$$x_{1j} \sim N(\mu_1, \sigma_1^2) \quad \text{英語の点}$$
$$x_{2j} \sim N(\mu_2, \sigma_2^2) \quad \text{数学の点}$$
$$\vdots \qquad \vdots$$
$$x_{nj} \sim N(\mu_n, \sigma_n^2) \quad \text{美術の点}$$

+)

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

総合点



（ケース２）同一の正規分布に従う場合

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

+)

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

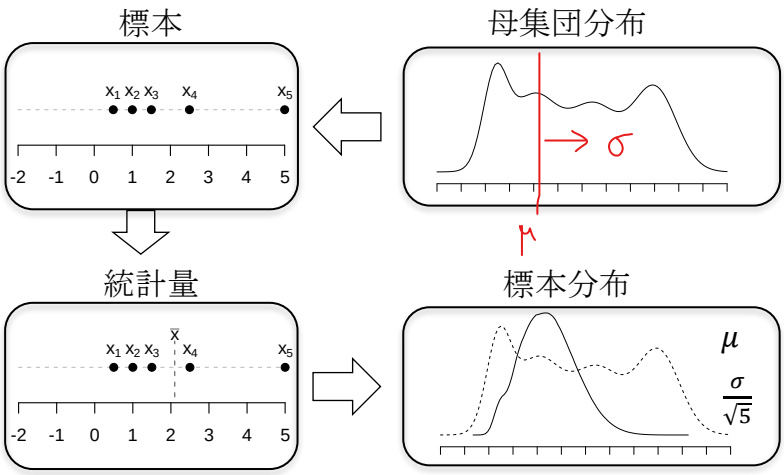
↓サンプルサイズ n で割る

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

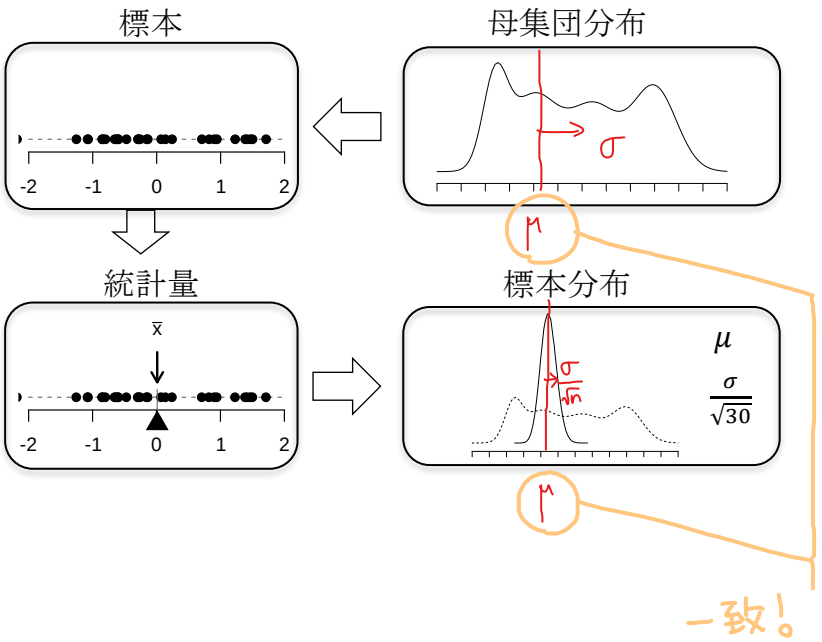
③ バージョン２：母集団分布が正規分布以外のとき

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

S　V　　O

I remember [　　] .

that　　単語（名詞）

visiting my friends　　句（名詞句）

that I visited my friends　　節（名詞節）

（ 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

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

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

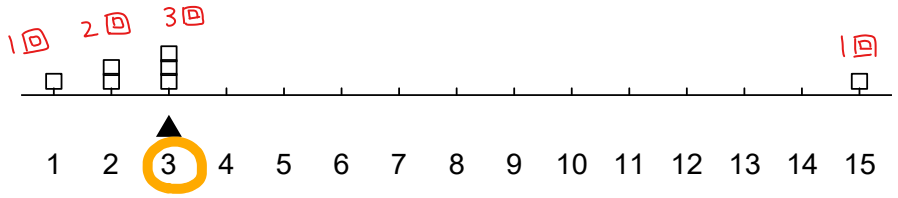
正規分布に従う変数を足した和も（差も）正規分布になる

📖 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）

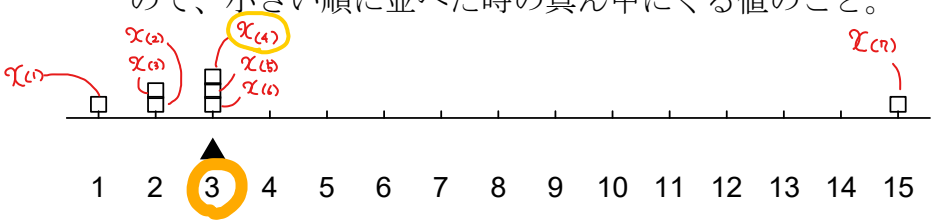
（１）（標本）最頻値 Sample Mode

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。



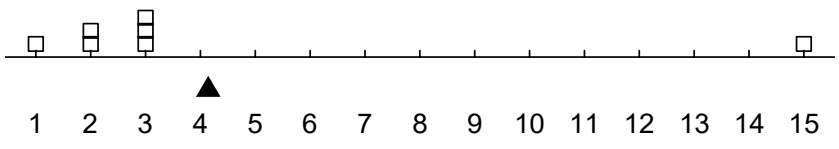
$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{\left(\frac{n}{2}\right)} \leq x_{\left(\frac{n}{2}+1\right)} \leq \cdots \leq x_{(n)}$$

最小値　　最大値

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

（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。



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

$$= \frac{1}{n}\left(x_1 + x_2 + \cdots + x_i + \cdots + x_n\right)$$

Ⓟ 代表値

これは，データ(=標本)を1つの値に代表させて示すときに用いる統計量のこと。

① データの中心
② データのばらつき
③ データ間の関連度

Ⓟ 順序統計量

これは，順位によって定義される統計量のこと。

(例)　最大値
　　中央値
　　最小値

Ⓟ 記法 ： x_1 と $x_{(1)}$

① 生のデータ

出席番号	英語
x_1 → 1	60
x_2 → 2	80
3	26
⋮	⋮

② 並べかえたデータ

出席番号	英語
$x_{(1)}$ → 3	26
$x_{(2)}$ → 17	35
⋮ 9	36
	⋮

Ⓟ 総和記号 $\sum_{i=1}^{n} x_i$

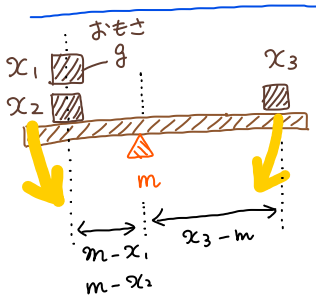
これは シグマ と読み 英語の S に対応するもの。
（Σ：大文字，σ：小文字）

the sum of x_i from $i=1$ to n.

cf. 総積記号 $\prod_{i=1}^{n} x_i$

$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_n$

Ⓟ 平均が $\frac{1}{n}\sum_{i=1}^{n} x_i$ な理由



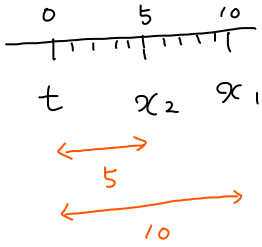
$g \times (m-x_1) + g \times (m-x_2) = g \times (x_3-m)$

$3m = x_1+x_2+x_3$

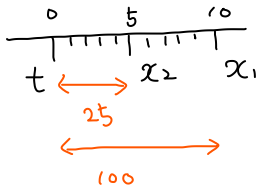
$m = \frac{x_1+x_2+x_3}{3}$

Ⓟ "距離"の考え方を拡張

① ユークリッド距離



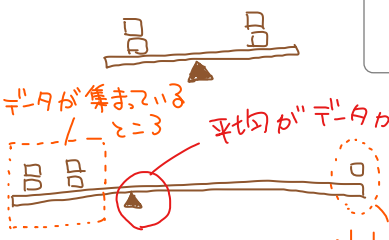
② 平方ユークリッド距離



どちらのアプローチをとっても x_1 が x_2 よりも t から遠いという点は捉えられる.

→ そこで②のケースも,「ものの遠さ」を測かる指標として採用してもいいじゃん。

Ⓟ 外れ値 (outlier)



外れ値：データが分布しているトコロから離れたトコロに位置する点のこと.

データとの距離の最小化

① 中央値：一次の距離の最小化

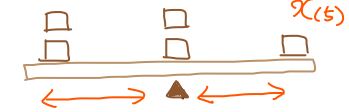
「t と各点x_iの 差の絶対値 の総和を最小化する」という基準を満たすデータの中心。

定規で測った距離
ユークリッド距離
(＝我々が日常的に使っている距離)

⇒ すなわち、次の基準T_1を最小化する t が中央値。

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

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



絶対値を計算するとき場合わけが必要(＝手間)

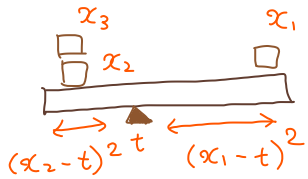
② 平均：二次の距離の最小化

「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。

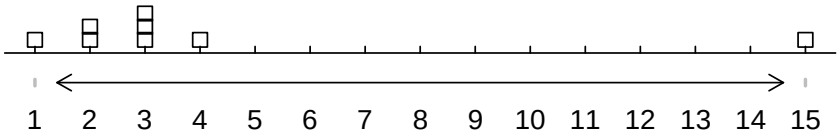
(特徴２) 中央値と違い平均は外れ値の影響を受けやすい。

(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）

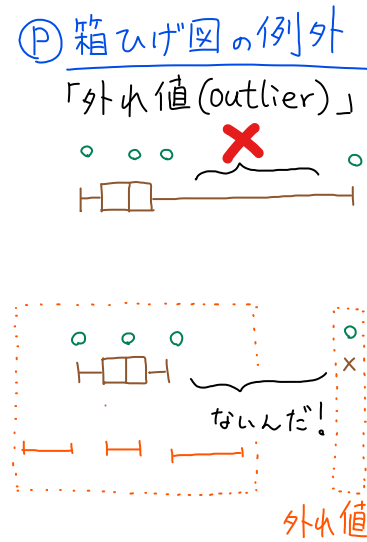
（1）標本範囲（最小統計量と最大統計量）Sample Range

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。



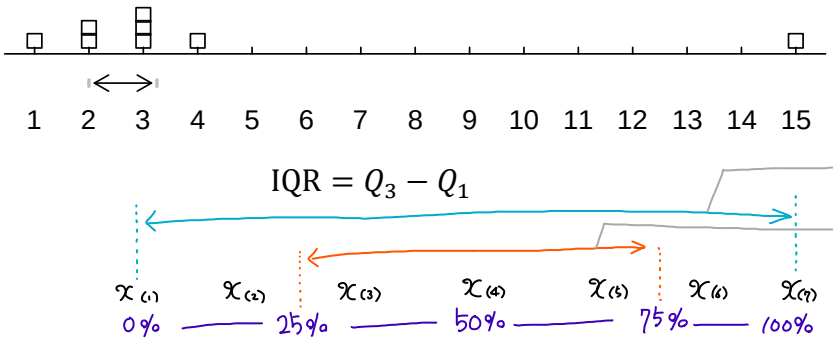
$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

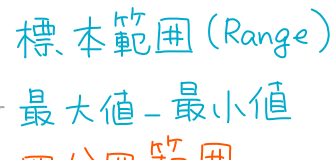
$$Rg = x_{(n)} - x_{(1)}$$



（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。

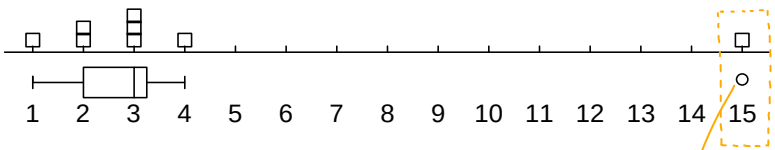


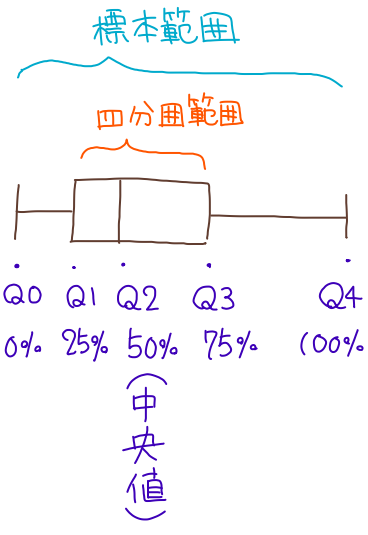


※ 箱ひげ図

これは、最大値、最小値、第 1～3 位四分位点を表す図。

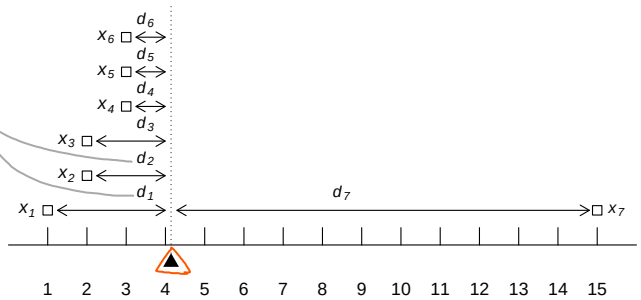
外れ値　を示すために、IQR の 1.5 倍の範囲の外にあるものは、独立して表す。





Ⓟ偏差

$|x_1-\bar{x}|=d_1$
$|x_2-\bar{x}|=d_2$
$\vdots$
$|x_7-\bar{x}|=d_7$



Ⓟ偏差を足し合わす

$$|x_1-\bar{x}| + |x_2-\bar{x}| + \cdots + |x_7-\bar{x}|$$
$$= |x_1-\bar{x}|+|x_2-\bar{x}|+\cdots+|x_7-\bar{x}|$$
$$= \sum_{i=1}^{7} |x_i-\bar{x}|$$

（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。

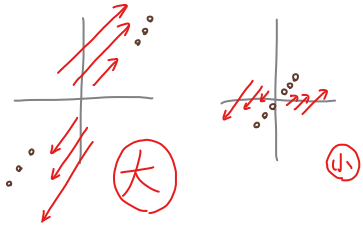
$$MD = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

平均している

紫色の矢印の長さを全部足しあわせたのがここ

Ⓟ分散

これは，自分自身との共分散であり，ばらつきの指標である

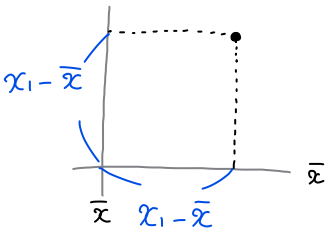


（4）標本分散（二次の距離に基づく指標）

① 偏差平方和
これは、二次距離　で測った平均からの距離。

x軸 × y軸

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



② 分散 Variance
これは、偏差平方和の平均。

二乗

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

面積

③ 標準偏差 Standard Deviation
これは、分散の平方根。

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

$\sqrt{\ }$をとります

長さ

Ⓟ変数 x の分散の表記

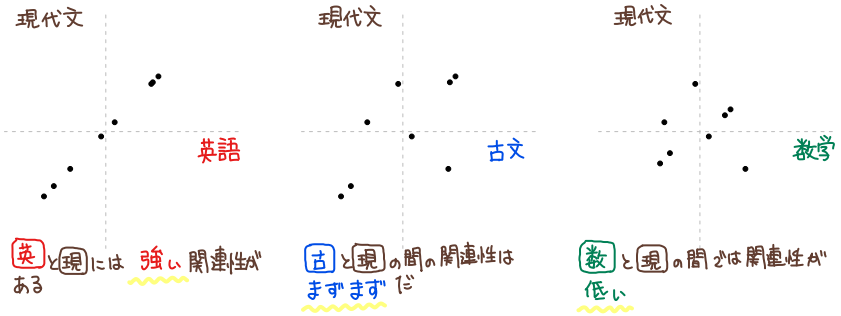
s^2（文脈上明白）

あるいは

s_x^2（文脈上変数が2つ以上あって，どれを指すのかはっきりさせたいとき）

■　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。

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

（s_{xy}：今回とった共分散、$s_x s_y$：関連度がMAXのときの共分散）

※相関係数には最大値と最小値が存在する。

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

（-1 ↑ −100%、1 ↑ 100%）

Ⓟ最大値で割ること

Aさん：90点（990点満点）
Bさん：90点（200点満点）

$\frac{90}{990} \fallingdotseq 0.09$（9%）

$\frac{90}{200} = 0.45$（45%）

最大値を分母

Ⓟ 共分散の弱点

単位に依存してしまう。

（ケース1）g, 1000, 1000×100 $= 10^5$, 100, cm

（ケース2）kg, 1, $|x| = 1$, m

Ⓟ 共分散の上限と下限

共分散 S_{xy} は，関連性が最大化するとき，次の値をとる

$S_{xy} = S_x S_y$（右肩あがり）

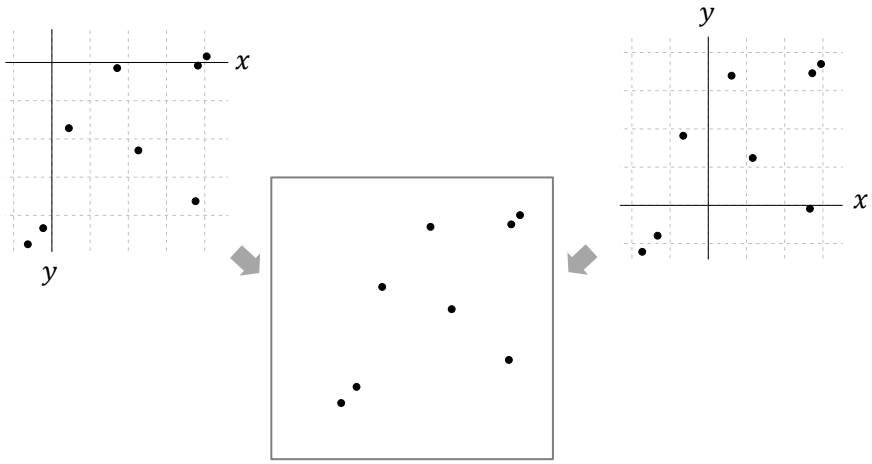
$S_{xy} = -S_x S_y$（右肩下がり）

Ⓟ 相関係数の良いところ

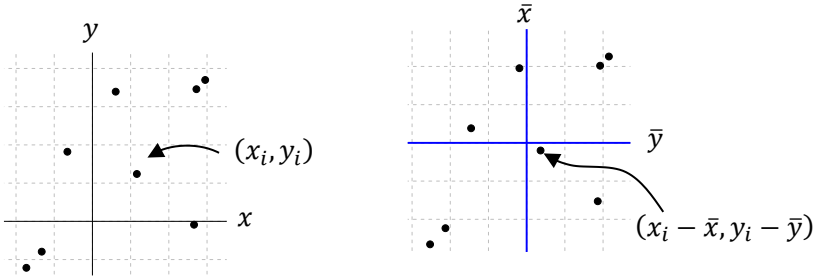
① 正負：関連の方向
② 値：関連度合の強さ
　1：相関が（正の方向に）MAX
　0：相関がない
　−1：相関が（負の方向に）MAX
③ 場合分け不用
④ サンプルサイズに依存しない
⑤ データの単位に依存しない

（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。

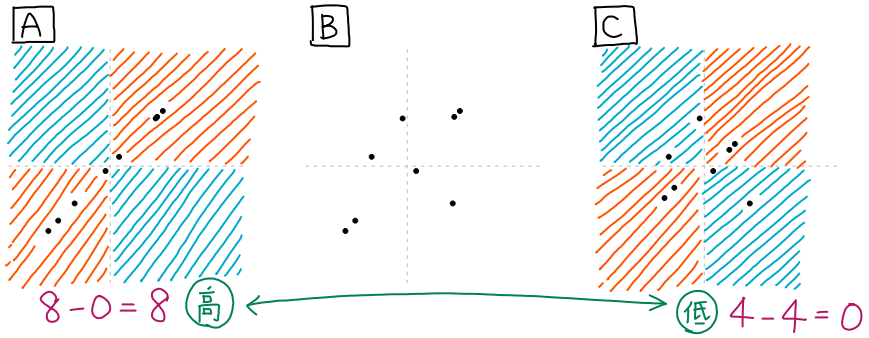


しかし、座標自体は維持したいので、新しく、横軸を$\bar{y}$、縦軸を$\bar{x}$の位置に据えた新しい座標を考える。これはもともとの座標を平行移動したことに相当する。



（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。

Ⓟ 象限

第2象限　第1象限

第3象限　第4象限

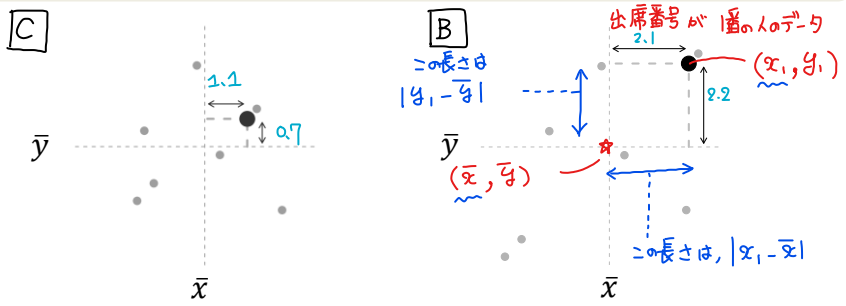
Ⓟ 第1案の良いところ

① 正負：関連の方向

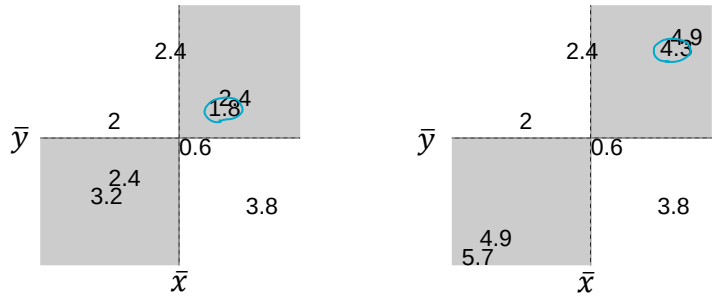
- 正：↗
- 0：↔
- 負：↘

② 値の大きさ：関連の強さ

（第二案）中心からの離れ具合を「足し算」で捉える

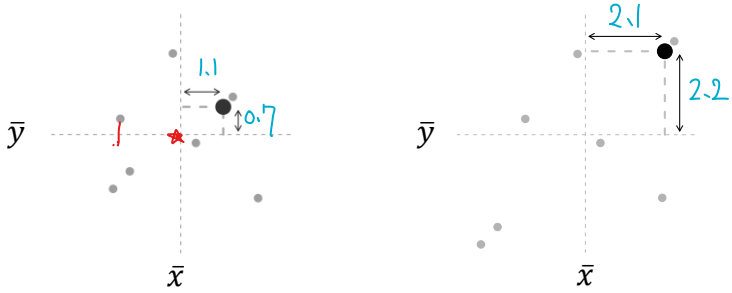


（式２）「灰色の領域の数字の和」－「白色の領域の数字の和」

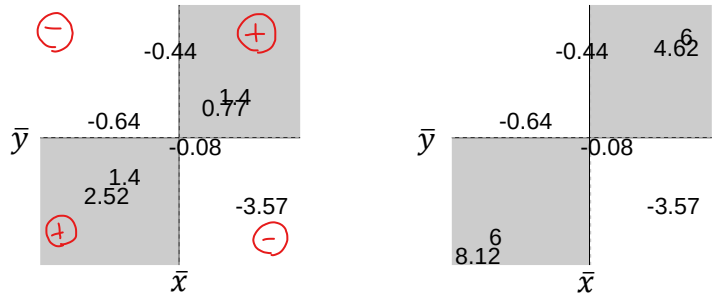


問題点：数式に場合分けが存在してしまう。

（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2案の良いところ

① 正負：関連性の方向
② 値　：関連性の強さ
③ Bと Cを区別できる

Ⓟ数式で表現（第2案）

① 出席番号1の人の中心からの遠さ

$2.1 + 2.2$
$= |x_1-\bar{x}| + |y_1-\bar{y}|$

② 全ての人をまとめて，中心からの遠さ

$4.3 + 4.9 + 4.9 + 5.7 - 2.4 - 2 - 0.6 - 3.8 = 11.1$

$|x_1-\bar{x}| + |y_1-\bar{y}|$
$+ |x_2-\bar{x}| + |y_2-\bar{y}|$
$+ |x_5-\bar{x}| + |y_5-\bar{y}|$
$+ |x_8-\bar{x}| + |y_8-\bar{y}|$
$-(|x_3-\bar{x}| + |y_3-\bar{y}|)$
$-(|x_4-\bar{x}| + |y_4-\bar{y}|)$
$-(|x_6-\bar{x}| + |y_6-\bar{y}|)$
$-(|x_7-\bar{x}| + |y_7-\bar{y}|)$

Ⓟ数式で表現（第3案）

① 出席番号1の人

2.1×2.2
$= (x_1-\bar{x})(y_1-\bar{y})$

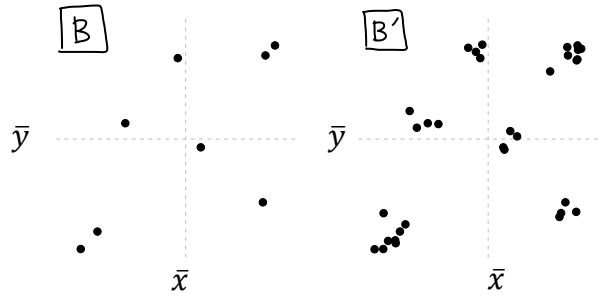
② 全ての人をまとめて

$6 + 4.62 + 6 + 8.12 + (-0.44) + (-0.64) + (-0.08) + (-3.57)$

$(x_1-\bar{x})(y_1-\bar{y})$
$+(x_2-\bar{x})(y_2-\bar{y})$
$+(x_5-\bar{x})(y_5-\bar{y})$
$+(x_8-\bar{x})(y_8-\bar{y})$
$+(x_3-\bar{x})(y_3-\bar{y})$
$+(x_4-\bar{x})(y_4-\bar{y})$
$+(x_7-\bar{x})(y_7-\bar{y})$
$+(x_8-\bar{x})(y_8-\bar{y})$

$= \sum_{i=1}^{8} (x_i-\bar{x})(y_i-\bar{y})$

（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」）

共分散

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

準備：平行移動をする

第2案：中心からの遠さを考慮する

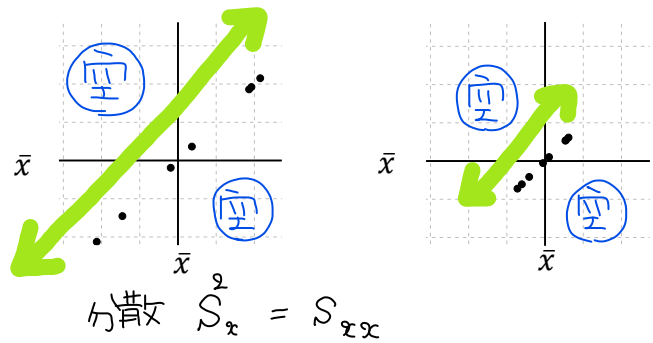
第3案：場合わけを回避

第4案：サンプルサイズの影響を取りのぞく

1つのデータではなくデータ全体を見よう

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。



分散　$S_x^2 = S_{xx}$

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。

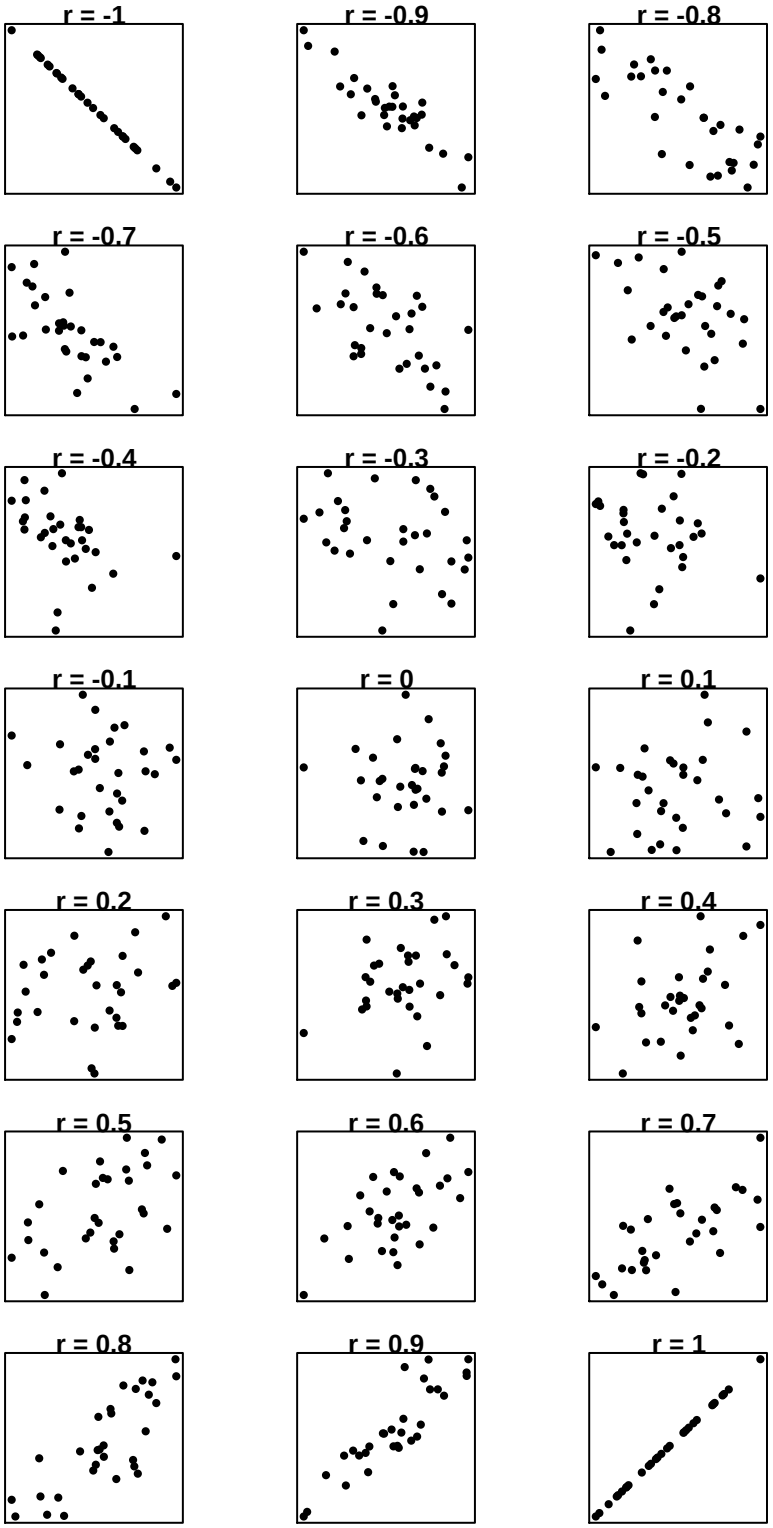
$$\bar{x}_{調} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}$$

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】

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