

# 統計学の整理帳

tomixy

2025 年 7 月 26 日

## 目次

第 1 章	一次元データの代表値と散らばり	2
	データの中心の指標：平均値 . . . . .	2
	データのばらつきの指標：偏差 . . . . .	2
	データのばらつきの指標：分散と標準偏差 . . . . .	4
	分散公式 . . . . .	5
	データの変換による平均と分散の変化 . . . . .	5
	データの標準化 . . . . .	7
第 2 章	二次元データの相関	9
	変量の相互関係：相関 . . . . .	9
	相関の数値化：共分散 . . . . .	10

# 第 1 章


## 一次元データの代表値と散らばり

### データの中心の指標：平均値

「データを 1 つの値で要約するならばこれ」といった指標を代表値という。

最もよく使われる代表値が平均値 (mean) であり、データの中心を表す指標として広く用いられる。

平均値は、データをすべて足し合わせて、データの数で割ることで求まる。

 平均  $N$  個の観測値  $x_1, \dots, x_N$  の総和をデータのサイズ  $N$  で割ったものを平均値という。


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

### データのばらつきの指標：偏差

代表値はデータを 1 つの値で要約する指標であり、データのばらつきや偏りは表現しきれない。

そこで、新たにデータのばらつきを表す指標を考える。

各データが、平均からどれくらい離れているかを表す指標を**偏差** (deviation) という。

 **偏差**  $N$  個の観測値  $x_1, \dots, x_N$  の平均値を  $\bar{x}$  とするとき、各観測値  $x_i$  の**偏差**は次のように定義される。

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

ここで、 $d_i$  は  $i$  番目のデータの偏差を表す。

## 偏差の平均値で全体をみる


全データの偏差  $d_1, \dots, d_N$  の平均値を求めることで、データ全体が平均からどれくらい離れて分布しているか（どれくらいばらついているか）を表すことができそうである。

しかし、偏差の平均値は、次のように常に 0 になってしまう。

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N d_i &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \\ &= \frac{1}{N} \left( \sum_{i=1}^N x_i - \sum_{i=1}^N \bar{x} \right) = \frac{1}{N} \left( \sum_{i=1}^N x_i - N\bar{x} \right) \\ &= \sum_{i=1}^N \frac{1}{N} x_i - \bar{x} = \bar{x} - \bar{x} = 0 \end{aligned}$$

そこで、単なる平均との差ではなく、平均との距離を考えることにする。

偏差に絶対値をつけたものの平均を**平均偏差**という。

 **平均偏差**  $N$  個の観測値  $x_1, \dots, x_N$  の平均値を  $\bar{x}$  とするとき、**平均偏差**を次のように定義する。

$$d := \frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|$$




## データのばらつきの指標：分散と標準偏差

平均偏差では、データと平均値の距離として絶対値を用いたが、絶対値は次のような理由で計算が面倒である。

- 絶対値は微分できない点がある
- 正負を判定する条件分岐処理が入り、コンピュータでの計算速度が落ちる

そこで、絶対値の代わりに二乗を用いた、**分散** (**variance**) という指標を定義する。

 **分散**  $N$  個の観測値  $x_1, \dots, x_N$  の平均値を  $\bar{x}$  とするとき、**分散**を次のように定義する。

$$\sigma^2 := \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

### もとのデータと同じ単位を持ったばらつきの指標


分散では二乗を用いるため、単位に注意が必要である。


もとのデータの単位が  $[x]$  であれば、分散の単位は  $[x]^2$  となる。

たとえば、点数を表すデータを扱っているとすると、その分散の単位は「点<sup>2</sup>」となり、直観的に理解しづらい。

そこで、単位をもとのデータと揃えるために、分散の平方根をとった形がよく用いられる。

分散の平方根を**標準偏差** (**standard deviation**) という。


 **標準偏差** 分散  $\sigma^2$  の平方根をとったものを**標準偏差**として定義する。

$$\sigma := \sqrt{\sigma^2}$$


## 分散公式

分散は、次のように計算することもできる。

分散 = データの二乗平均 - 平均の二乗

 分散公式  $N$  個の観測値  $x_1, \dots, x_N$  の平均を  $\bar{x}$ 、分散を  $\sigma^2$  とすると、次の関係が成り立つ。

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2$$

### 証明

分散の定義に基づいて、次のように計算する。

$$\begin{aligned}\sigma^2 &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ &= \frac{1}{N} \sum_{i=1}^N (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 - 2\bar{x} \frac{1}{N} \sum_{i=1}^N x_i + \bar{x}^2 \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 - 2\bar{x}^2 + \bar{x}^2 \\ &= \frac{1}{N} \sum_{i=1}^N x_i^2 - \bar{x}^2\end{aligned}$$

## データの変換による平均と分散の変化


データの変換（スケーリングやシフト）を行うと、平均や分散はどのように変化するのだろうか。

## データのスケールリング

まず、データを定数  $a$  倍する変換を考える。

すなわち、各データ  $x_i$  を  $y_i = ax_i$  に変換すると、平均と分散は次のように変化する。

$$\begin{aligned}\bar{y} &= a\bar{x} \\ \sigma_y^2 &= a^2\sigma_x^2\end{aligned}$$

 データのスケールリングによる平均と分散の変化    観測値が  $a$  倍されると、平均は  $a$  倍、分散は  $a^2$  倍される。

### 証明

各データ  $x_i$  を  $y_i = ax_i$  に変換すると、平均は次のように変化する。

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N ay_i = a \cdot \frac{1}{N} \sum_{i=1}^N y_i = a\bar{y}$$

また、分散は次のように変化する。

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N} \sum_{i=1}^N (ax_i - a\bar{x})^2 = \frac{1}{N} \sum_{i=1}^N (a(x_i - \bar{x}))^2 \\ &= a^2 \cdot \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = a^2\sigma_x^2\end{aligned}$$

## データのシフト

次に、データを定数  $b$  だけシフトする変換を考える。

すなわち、各データ  $x_i$  を  $y_i = x_i + b$  に変換すると、平均と分散は次のように変化する。

$$\begin{aligned}\bar{y} &= \bar{x} + b \\ \sigma_y^2 &= \sigma_x^2\end{aligned}$$

📌 データのシフトによる平均と分散の変化 観測値に  $b$  を加えると、平均は  $b$  だけ増えるが、分散は変化しない。

#### 🔪 証明

各データ  $x_i$  を  $y_i = x_i + b$  に変換すると、平均は次のように変化する。

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N (x_i + b) = \frac{1}{N} \sum_{i=1}^N x_i + b = \bar{x} + b$$

また、分散は次のように変化しない。

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^N (x_i + b - (\bar{x} + b))^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \sigma_x^2$$

このように、データの変換によって平均と分散は異なる影響を受けることがわかる。

## データの標準化

たとえば、平均点が 30 点のテストでとった 60 点と、平均点が 80 点のテストでとった 60 点とでは、相対的な出来が異なる。このように、点数というデータはそのテストの平均や分散によって評価が変わってしまう。

そのため、平均や分散に依存せずにデータの相対的な位置関係がわかるようにできたら便利である。

特に、平均が 0、標準偏差が 1 になるようにデータを変換することを**標準化 (standardization)**という。

$y_i = ax_i + b$  というデータの変換を考えよう。

これは、データを  $a$  倍して  $b$  だけシフトする変換である。

このとき、平均と標準偏差は次のように変化する。

- 平均は  $a$  倍され、 $b$  だけ増える
- 標準偏差は  $a$  倍される（分散が  $a^2$  倍される）

数式で表すと、

$$\begin{aligned}\bar{y} &= a\bar{x} + b \\ \sigma_y &= a\sigma_x\end{aligned}$$


そこで、平均  $\bar{y}$  が 0、標準偏差  $\sigma_y$  が 1 になるように、 $a$  と  $b$  を次のように設定する。

$$a = \frac{1}{\sigma_x}, \quad b = -\frac{\bar{x}}{\sigma_x}$$

このようにすると、たしかに  $\bar{y} = 0$ 、 $\sigma_y = 1$  となる。

このとき、変換後のデータ  $y_i$  は次のように表される。

$$y_i = ax_i + b = \frac{x_i - \bar{x}}{\sigma_x}$$

 **標準化** 各データから平均を引き、標準偏差で割ることで、平均が 0、標準偏差が 1 になるように変換することを**標準化**という。

各データ  $x_i$  を標準化したデータを  $y_i$  とすると、次の関係が成り立つ。

$$y_i = \frac{x_i - \bar{x}}{\sigma_x}$$



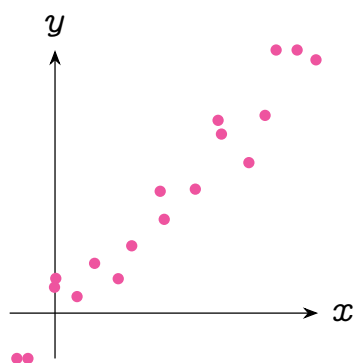
## 第 2 章

# 二次元データの相関

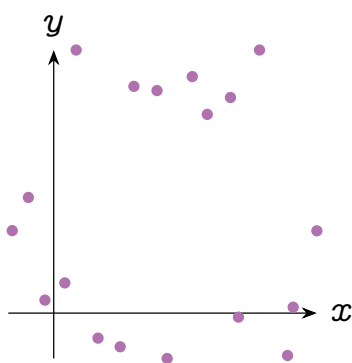
### 変量の相互関係：相関

2次元データ  $(x, y)$  において、2つの変量  $x$  と  $y$  の間に相互関係がみられるとき、 $x$  と  $y$  の間には**相関**（correlation）関係があるという。

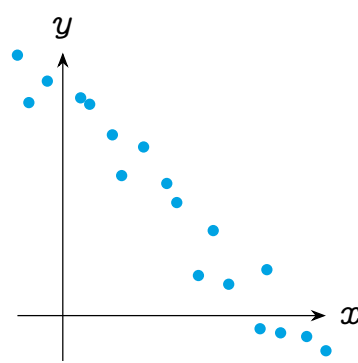
- **正の相関**： $x$  が大きいほど、 $y$  も大きくなる傾向がある
- **負の相関**： $x$  が大きいほど、 $y$  は小さくなる傾向がある
- **無相関**：どちらにも当てはまらない（直線的な関係がない）



正の相関



無相関

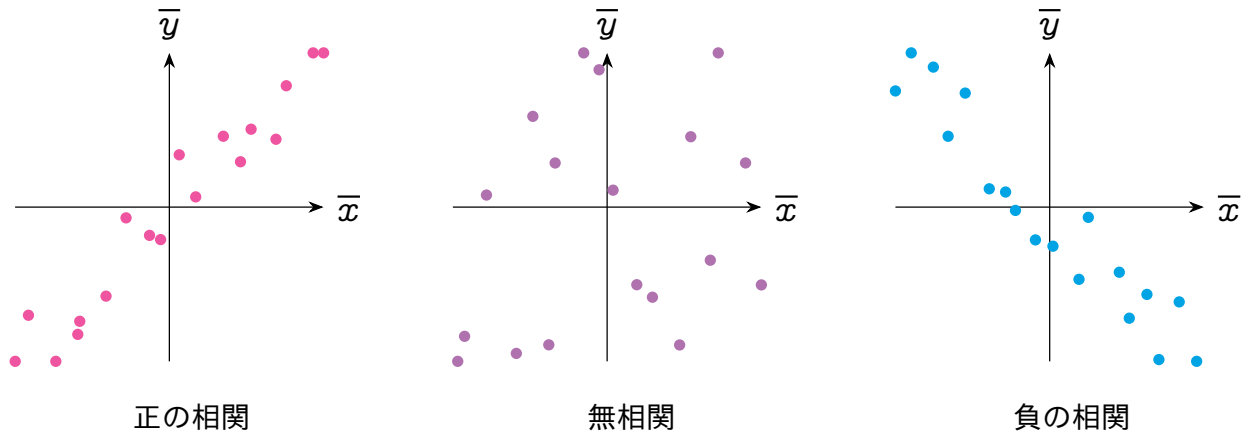


負の相関

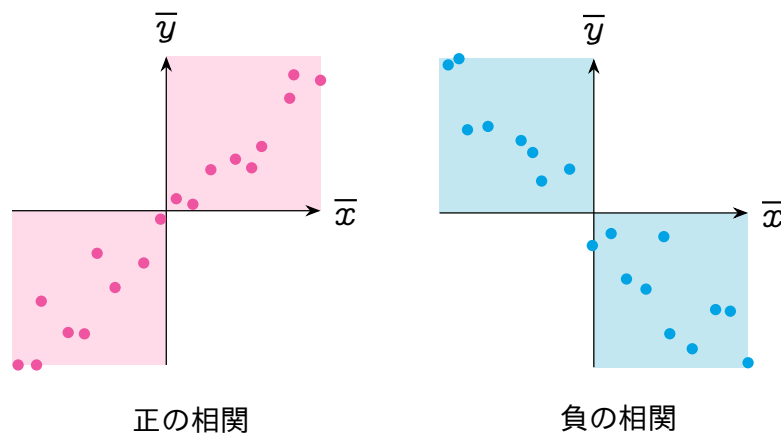
## 相関の数値化：共分散

グラフを描いて視覚的に相関を確認することはできるが、客観的に表現するために、数値で判断できるようにしたい。

そのために、 $x, y$  の平均 ( $\bar{x}, \bar{y}$ ) を原点とする新たな座標軸を考える。



すると、正の相関か負の相関かに応じて、データが多く分布する象限（座標軸で切り分けた領域）が異なることがわかる。



**正の相関**の場合は、第一象限と第三象限にデータが多く分布することがわかる。

- 第一象限： $x > \bar{x}$  かつ  $y > \bar{y}$  である範囲
- 第三象限： $x < \bar{x}$  かつ  $y < \bar{y}$  である範囲

負の相関の場合は、第二象限と第四象限にデータが多く分布することがわかる。

- 第二象限： $x < \bar{x}$  かつ  $y > \bar{y}$  である範囲
- 第四象限： $x > \bar{x}$  かつ  $y < \bar{y}$  である範囲


この場合分けは、次のようにまとめることができる。

- 正の相関の場合、 $x - \bar{x}$  と  $y - \bar{y}$  の符号が同じになる点が多い
- 負の相関の場合、 $x - \bar{x}$  と  $y - \bar{y}$  の符号が反対になる点が多い

さらに、符号が同じものの積は正、符号が反対のものの積は負になることから、

- 正の相関の場合、 $(x - \bar{x})(y - \bar{y}) > 0$  となる点が多い
- 負の相関の場合、 $(x - \bar{x})(y - \bar{y}) < 0$  となる点が多い

各データについて  $(x_i - \bar{x})(y_i - \bar{y})$  を求め、全データの平均をとることで、相関を判定できそうである。このような考え方で相関を数値化したものを共分散 (covariance) という。

 共分散  $N$  個の観測値  $(x_1, y_1), \dots, (x_N, y_N)$  の平均をそれぞれ  $\bar{x}, \bar{y}$  とするとき、共分散を次のように定義する。

$$\sigma_{xy} := \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$