プログラミングの背景：統計計算

標準偏差の計算

tbasic.org [*1]

[2014 年 6 月版]

標準偏差はデータのばらつきの度合いを表す，統計計算に表れる基本的な量です。偏差値の計算にも使われます。ここでは，その計算法について説明します。

目次

[*1] http://www.tbasic.org

1 標準偏差

統計調査の目的は，多量のデータがあったとき，その全体的特徴を知ることにあります。少ないデータであればすべてのデータを表示して，よく観察することで，その特徴を見出すのが最も直接的です。しかしデータが多数ある場合は，すべてのデータを一度に直接観察するのは限界があります。

そのために色々な工夫を行い，それらのデータから，特徴的な値を求め，それからそのデータの特徴を見出すことになります。

最も基本的な量として，平均値があります。

定義 1.1 (平均)**.** 自然数 n 個のデータ $x_1, x_2, \ldots, x_n$ に対して，

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

を平均 (Mean)，平均値という。[*2]

平均はデータの特徴を表す量ですが，勿論これだけではデータの概要を見出すことはできません。例えば，同じ 3 個のデータでも，

$$2,\ 2,\ 2 \quad と \quad 1,\ 2,\ 3$$

は同じ平均値 2 を持ちますが，分布の様子は異なります。このように平均値を求めただけでは，データの全体像を掴むことはできません。

標準偏差はータ $x_1, x_2, \ldots, x_n$ の値が，平均値からどのくらいの範囲に広がっているかを示す指標で，次のように定義されます。

定義 1.2 (分散，標準偏差)**.** 自然数 n 個のデータ $x_1, x_2, \ldots, x_n$ が与えられた時，

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

を標準偏差 (Standard deviation) と言います[*3]。

ここで，$\bar{x}$ はデータ $x_1, x_2, \ldots, x_n$ の平均値です。標準偏差の 2 乗 σ^2 を分散（Variance）と言います。即ち，

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

です。

例 1.1.

(1) 3 個のデータ，$2, 2, 2$ の平均値は 2，標準偏差は 0

(2) 3 個のデータ，$1, 2, 3$ の平均値は 2，標準偏差は $\sqrt{2/3}$ [*4]

[*2] 算術平均 (Arithmetic Mean) とも言います。

[*3] この標準偏差は Excel では，STDEVP 関数で求めることができます。

[*4] 良く使われる似た量として，n 個の標本データ $x_1, x_2, \ldots, x_n$ から計算される

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

この統計量は重要な量で，例えば学力偏差値の計算にも表れます。学力偏差値は次のように定義されます。母集団（全体のデータ）が正規分布に近い場合，この偏差値からそのデータの大体での位置知ることが出来ます。

定義 1.3 (偏差値)**.** 自然数 n 個のデータ $x_1, x_2, \ldots, x_n$ に対して，$\sigma \neq 0$ のとき，

$$T_i = \frac{10(x_i - \bar{x})}{\sigma} + 50$$

をデータ x_i の偏差値と言う。

2 計算法 1

この標準偏差の値をプログラミングで求めることを考えてみましょう。まず，次のように計算するのが普通でしょう。

(1) 平均値 $\bar{x}$ を $x_1, x_2, \ldots, x_n$ から計算する。

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

適当な変数 `S=0` を用意し，その `S` に順次 $x_1, x_2, \ldots, x_n$ を加え，最後の x_n になったら，`S` を n で割り，平均値 $\bar{x}$ を求める。

(2) 分散 σ^2 を $\bar{x}$ と $x_1, x_2, \ldots, x_n$ から計算する。

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

適当な変数 `Ssq=0` を用意し，その `Ssq` に順次 $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$ を加え，最後の $(x_n - \bar{x})^2$ になったら，`Ssq` を n で割り，分散 σ^2 を求める。

(3)

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

分散 σ^2 の平方根をとり，σ を求める。

これはこれで良いのですが，状況によっては工夫したい場合もあります。この計算では，$x_1, x_2, \ldots, x_n$ を (1) と (2) でそれぞれデータの最初から合計２回使っていますので，例えば，次の様な状況が考えられます。

- n が非常に大きく，それぞれ，$x_1, x_2, \ldots, x_n$ の入力に時間がかかり，できれば 1 回の入力で計算したい[*5]。

を普遍分散と言います。この平方根 s も標準偏差と言われることがあります。この標準偏差を使って，3 個のデータ 1, 2, 3 の分散を計算すると，1 になります。こちらの方が値は自然な感じがします。

[*5] 入力データをコンピューターのメモリに蓄えれば，何度でも利用可能ですから，この問題は本質的には，n が大きくて，内部メモリに入らない，或いは入れるのを避けたい状況で生じます。

3　計算法２

実は，分散について次の等式が成立することが比較的簡単に分かります。

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \qquad (*)$$

証明. 実際,

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

の両辺を n で割ることで $(*)$ を求めることが出来ます。 □

この式 $(*)$ を利用すると，上の問題が解決できます。

(1) 和 S と平方和 Ssq を $x_1, x_2, \ldots, x_n$ から順次計算する。

$$\mathtt{S} = x_1 + x_2 + \cdots + x_n, \quad \mathtt{Ssq} = x_1^2 + x_2^2 + \cdots + x_n^2$$

和を表す変数 `S=0` と平方和を表す変数 `Ssq =0` を用意する。S には x_i を Ssq には x_i^2 を順次加える。これを，$i = 1$ から n まで実行して，和と平方和を計算する。

(2) 上の式 $(*)$ を使って，和 S と平方和 Ssq から分散を計算する。

$$\sigma^2 = \frac{1}{n}\mathtt{Ssq} - \left(\frac{1}{n}\mathtt{S}\right)^2$$

(3) 分散の平方根をとり，σ を求める。

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

4　まとめ

計算法１と計算法２を比較すると，計算法１が定義に従って，ある意味分かりやすいものです。一方，計算法２は，定義からは少し異なりますが，色々な状況で使える方法です。その意味で計算法２の方がやや良い方法とも言えます。

ただ，計算方法の実装は，状況に応じて最適なものを採用するのが基本です。ですから，常に計算法２を使うべきと考える必要はありません。しかし，与えられた状況で最適な方法を選ぶために，対象をある程度詳しく分析することは常に必要でしょう。