# 分散はなぜ２乗するのか？　～標準偏差と平均偏差の違い～

旭川南高等学校
岡崎知之

## ０．はじめに

「データの分析」が導入され、「四分位数」や「相関係数」等の新たな代表値が導入された。センター試験でもこれらの代表値を利用し、正誤判断をする問題が定番となっているが、その代表値の中でも「標準偏差」を直接利用する問題は少ないように見受けられる。

実際生徒に、代表値の定義や利用方法を尋ねてみても、「標準偏差」についてはほぼノーコメントであった。教えている私も、そのことについてあまり語ったことがない。新学習指導要領では、更に新たな代表値が導入されることが予測される。その前に「標準偏差」について、深堀りしてみる。

## １．現行教科書における「標準偏差」

教科書では、次のような記述で「分散」が説明されている。

> データの平均値の周りに、データの各値がどのように分布しているかを示す値として、まず各値と平均値との差を考えてみる。
>
> 変量$x$のデータの値を$x_1$，$x_2$，…，$x_n$、その平均値を$\bar{x}$とするとき、…偏差の総和は
>
> $(x_1-\bar{x})+(x_2-\bar{x})+\cdots+(x_n-\bar{x})=0$ になるから、偏差の平均値も０である。
>
> よって偏差の平均値では散らばりの度合いを表すことはできない。
>
> そこで、偏差をそのまま用いずに、偏差の２乗の平均を考えよう。
>
> （数研出版　改訂版　高等学校　数学Ⅰ　より抜粋）

ここでは「標準偏差」は「平均値の周りのデータ分布を示す値」として、「偏差の２乗平均」を利用することが書かれている。すなわち「標準偏差」を、母集団が複数存在したときに、「平均値を基準とした散らばりの度合い」が比較できる代表値としている。

## ２．「標準偏差」を、どう教えるか

しかし、標準偏差を求めて散らばり度合いを調べる機会は少ない。なぜなら、散らばりを調べる代表値は他にも「範囲」「四分位範囲」等があり、「平均値を基準とした…」というフレーズは意味が伝わりにくいからである（多分）。その結果「標準偏差」は「相関係数」を計算する材料や、変数変換後の値の変化を調べる問題に利用される程度に留まっている。

そこで、私は標準偏差を以下のように説明していた。

> Ｔ　「テストで平均点が出ても、得点が平均点ピッタリということは少ないよね。じゃあ自分の点数が全員の中でちょっと優秀だとか、すごく優秀だというのは、どうやって判断するの？」
>
> Ｓ　「大体１０点くらい上だったら、すごいなぁ…とか。」「順位で考える。」
>
> Ｔ　「テストの平均も変化するから１０点は曖昧だし、順位も点数にばらつきがあるから、必ずしも判断材料にならないよね。そこで、平均点からどの程度離れた点数までが『普通』なのかを考える。そのために、『離れ具合の平均』を求める。偏差の平均値を求めると…」
>
> Ｓ　「０になった。」
>
> Ｔ　「そりゃそうだよね。＋－の差を埋めたのが平均なんだから。そこで、すべての値を正に直すのが面倒なので、とりあえず２乗する。で、元に戻すためにルートをつける。」
>
> Ｔ　「もし平均点が６０点で、標準偏差が５なら、５５～６５点の人は大勢いて、普通だということだね。」

＊数学の有用性を感じてもらいたいので、厳密な言語使用はしていない。

ところが、最近こんな疑問が湧いてきた。

データの数値が分かっていれば、絶対値を計算するほうが、２乗するより労力が少ないのではないか。
そうであれば、絶対値の方が「簡潔」かつ「平均値との関係が分かりやすくなる」のではないだろうか。

## ３．平均偏差

教科書には掲載されていないが、偏差の絶対値の平均は「平均偏差」（＊１）と呼ばれる。

☆平均偏差（AD：Average Deviation）

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

ここで、$\bar{x}$…代表値 であり、平均値とは限らない。

## ４．標準偏差との比較

（例）

| | 得点 | 偏差 | 絶対値 | ２乗 |
|---|---|---|---|---|
| 1 | 66 | 15 | 15 | 225 |
| 2 | 22 | -29 | 29 | 841 |
| 3 | 44 | -7 | 7 | 49 |
| 4 | 60 | 9 | 9 | 81 |
| 5 | 71 | 20 | 20 | 400 |
| 6 | 28 | -23 | 23 | 529 |
| 7 | 56 | 5 | 5 | 25 |
| 8 | 55 | 4 | 4 | 16 |
| 9 | 40 | -11 | 11 | 121 |
| 10 | 68 | 17 | 17 | 289 |
| 平均 | 51 | 0 | 14 | 257.6 |

| 標準偏差 |
|---|
| 16.9 |

- （標準偏差）－（平均偏差）＝16.9-14=2.9
- 平均からの差が 14 点以内は普通の（平均的な）点数。
  実際、10 個のデータのうち５個は 37～65 点以内に含まれている。

⇒「平均偏差」も使えるのでは…？だったら、何故「標準偏差」ばかり使われる？

## ５．統計的決定理論

「平均値の周りの分布」を評価する値として、「標準偏差」と「平均偏差」のどちらがふさわしいのか？
そのためには、統計的決定理論に基づき、「損失関数」を考え、その値を最小化する代表値を調べる。
例えば…

「分散」の損失関数（代表値との差を評価する関数）を$\sum_{i=1}^{n}(x_i - \bar{x})^2$とするとき

$\bar{x}$ がどの代表値（平均，中央値，最頻値，etc…）であれば、損失関数が最小になるのかを考えると、

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

ここで注意したいのが、$x_i$は定数で、$\bar{x}$は変数だということ。

簡潔に表示するために、全体を $\mathrm{f}(\bar{x})$，$\sum_{i=1}^{n}{x_i}^2 = a$，$\sum_{i=1}^{n}x_i = b$と置いてみる。

すると、

$\mathrm{f}(\bar{x}) = a - 2b\bar{x} + n\bar{x}^2$

$\mathrm{f}'(\bar{x}) = -2b + 2n\bar{x}$

よって、

$-2\mathrm{b} + 2\mathrm{n}\bar{x} = 0$ のとき、すなわち $\bar{x} = \frac{b}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$、$\bar{x}$が「平均値」のとき、最小となる。

では、「平均偏差」のときはどうだろう？

実は、「平均偏差」の損失関数　$\sum_{i=1}^{n}|x_i - \bar{x}|$ は、$\bar{x}$ が「中央値」のとき、最小となる。

一例として、データの個数が奇数の場合

データを小さい順から

$x_1$，$x_2$,…，$x_c$，…，$x_n$とすると、

（ｉ）代表値＝中央値の場合　（$\bar{x} = x_c$）

$$\begin{aligned}\sum_{i=1}^{n}|x_i - \bar{x}| &= |x_1 - \bar{x}| + \cdots + |x_{c-1} - \bar{x}| + |x_c - \bar{x}| + |x_{c+1} - \bar{x}| \ldots + |x_n - \bar{x}| \\ &= (\bar{x} - x_1) + \cdots + (\bar{x} - x_{c-1}) + 0 + (x_{c+1} - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= -(x_1 + \cdots + x_{c-1}) + (x_{c+1} + \cdots + x_n) = M \text{ とすると}\end{aligned}$$

（ⅱ）代表値≦中央値の場合　（$x_k \le \bar{x} \le x_{k+1}$）

$$\begin{aligned}\sum_{i=1}^{n}|x_i - \bar{x}| &= |x_1 - \bar{x}| + \cdots + |x_k - \bar{x}| + |x_{k+1} - \bar{x}| + \cdots + |x_n - \bar{x}| \\ &= (\bar{x} - x_1) + \cdots + (\bar{x} - x_k) + (x_{k+1} - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= -(x_1 + \cdots + x_{c-1}) + (x_{c+1} + \cdots + x_n) + 2(x_{k+1} - \bar{x}) + \cdots + 2(x_n - \bar{x}) > M\end{aligned}$$

同様に、代表値≧中央値、等の場合も証明できる。

６．結論

平均値の周りの分布を評価する際に、適切な代表値は、やっぱり「標準偏差」だった。

７．「標準偏差」を、どう教えるか（再）

以上の論理を、高１に説明するのは難しいだろう。だが、標準偏差の簡易的な活用法は紹介したい。

やはり、信頼区間の考えが使えるだろうか。

（平均点）±σ　の範囲…データの 68.3%が含まれる

±２σの範囲…データの 95.4%が含まれる

８．感想

「なぜ分散はデータを２乗するのか？」について、多くの入門書では「絶対値の計算は煩雑だから」と述べられている。これまでその記述に曖昧さを感じていたが、ようやく納得することができた。

今後、統計教育が進んでいく予定だが、より研鑽を積んでいかなくてはならないと感じた。

９．参考サイト

「tsujimotter のノートブック　統計的決定理論　～平均値・中央値・最頻値って何？～」

「inobuyuki のブログ　標準偏差と平均偏差の使い分け」

「総合案内所／五捨五超入と四分位数と放射線のサイト　基礎統計学」

おまけ：もっと上手な「標準偏差」の教え方があれば、教えてください…。

（2019 年８月 31 日　第 110 回数学教育実践研究会にて発表）