「データの分析」における分散のあれこれ

斎 木 清 治

1　はじめに

高等学校の数学Ⅰに「データの分析」が登場して，7〜8 年になるだろうか．大学入試への出題について，当初は「センター試験」くらいにしか出題されないだろうという声が多かったと記憶している．

「データ分析は，数学の論理構成能力を高めるには不向きな内容であり，難関私大，国公立二次試験に出題されることはほとんどないと考えられるため本書では割愛しました」と書いて，この内容を掲載しなかった割とメジャーな学習参考書まであったくらいである．

しかし，近年になって，本格的な出題が増えてきている．

試みに，2019 年福井大学医学部・前期の問題を載せてみる [(3)は省略した].

> 変数 x のデータの値を x_1, ……, x_n，変量 y のデータの値を y_1, ……, y_n とする．変量 x の標準偏差を s_x，変量 y の標準偏差を s_y とする．また，変量 x と変量 y の相関係数を r とする．このとき，以下の問いに答えよ．
>
> (1) 変量 x の最大値を $\max(x)$，最小値を $\min(x)$ とする．このとき
>
> $$s_x \leqq \max(x) - \min(x)$$
>
> が成り立つことを示せ．さらに，等号の成立条件を調べよ．
>
> (2) 変量 z のデータの値を $z_1 = x_1 - y_1$, ……, $z_n = x_n - y_n$ とする．このとき
>
> $$r = \frac{{s_x}^2 + {s_y}^2 - {s_z}^2}{2 s_x s_y}$$
>
> が成り立つことを示せ．ただし，s_z は変量 z の標準偏差とする．

なかなか，ハードである．(1)など，最初から鉛筆が止まる．

これは今年の問題だが，最近この分野からの出題を目にするようになり，先の参考書もさすがに看過できなくなったのか，増補版で「データの分析」を補わざるを得なくなっている．

そこで，データの分散に関して 2 つの内容を取り上げ考察したい．

2　データの追加と分散

変数 x のデータの値を x_1, ……, x_n，平均値を m，分散を σ^2 とする．ここに 1 個のデータ x_{n+1} を付け加えたとき，新たな平均値を m'，分散を ${\sigma'}^2$ とする．ただし，$\sigma^2 \neq 0$ とする．

このとき，${\sigma'}^2$ と σ^2 の大小関係はどうなるのだろうか．

平均については，$x_{n+1} = m$ であれば $m' = m$，$x_{n+1} < m$ であれば $m' < m$ などと言った関係は，計算しなくても分かる．

分散についても，$x_{n+1} = m$ であれば平均値の近くにデータが寄り集まったわけだから，${\sigma'}^2 < \sigma^2$ は感覚的に分かる．しかしそうでない場合はどうなのだろうか．${\sigma'}^2 = \sigma^2$ となることはあるのだろうか．

$x_{n+1} = m + (n+1)k$ とすると，$(n+1)m' = nm + \{m + (n+1)k\}$ から，$m' = m + k$ である．このとき，

$\sigma^2 = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} {x_j}^2 - m^2$ より，$\displaystyle\sum_{j=1}^{n} {x_j}^2 = n(\sigma^2 + m^2)$ であるから

$${\sigma'}^2 = \frac{1}{n+1}\sum_{j=1}^{n+1} {x_j}^2 - {m'}^2 = \frac{1}{n+1}\left\{n(\sigma^2 + m^2) + (m + (n+1)k)^2\right\} - (m+k)^2 = nk^2 + \frac{n\sigma^2}{n+1}$$

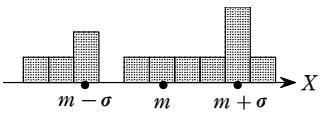
となるから，これと σ^2 を比較すると，${\sigma'}^2 - \sigma^2 = nk^2 - \dfrac{1}{n+1}\sigma^2$ より，

$k = \pm\dfrac{\sigma}{\sqrt{n(n+1)}}$ すなわち，$x_{n+1} = m \pm \sqrt{\dfrac{n+1}{n}}\sigma$ のとき，${\sigma'}^2 = \sigma^2$ である．

また，$-\dfrac{\sigma}{\sqrt{n(n+1)}} < k < \dfrac{\sigma}{\sqrt{n(n+1)}}$ すなわち，$m - \sqrt{\dfrac{n+1}{n}}\sigma < x_{n+1} < m + \sqrt{\dfrac{n+1}{n}}\sigma$ のとき，${\sigma'}^2 < \sigma^2$ であり，

$x_{n+1} < m-\sqrt{\dfrac{n+1}{n}}\sigma$, $m+\sqrt{\dfrac{n+1}{n}}\sigma < x_{n+1}$ のとき， $\sigma'^2 > \sigma^2$ である．

n が一定大きければ， $\sqrt{\dfrac{n+1}{n}} \fallingdotseq 1$ であるから，平均値から標準偏差だけ離れた値を追加した場合，(平均値は変わるものの) 分散はほぼ変わらない．平均値から標準偏差だけ離れた値より平均値に近い値を追加した場合は，分散は小さくなると言うことである．



では，2 個以上のデータを追加した場合はどうなるのか．これは，2 つの集団の合併の問題である．

2013年11月に行われた統計検定第3級に次のような問題がある．

ある学習塾にはA 組とB 組の2 クラスがある。先日行った数学の模擬試験の結果は次の通りであった。

A 組： 受験者19名 平均点65.0点 標準偏差9.6点

B 組： 受験者25名 平均点57.0点 標準偏差14.4点

2 つのクラスをまとめた44名のこの試験の標準偏差の求め方として適切なものを，次の①～⑤ のうちから一つ選べ。

① $(9.6+14.4)\div 2$

② $\sqrt{(9.6\times 19+14.4\times 25)\div 44}$

③ $\sqrt{(9.6^2\times 19+14.4^2\times 25)\div 44}$

④ $\sqrt{9.6^2\times 19+14.4^2\times 25+(65.0-60.5)^2\times 19+(57.0-60.5)^2\times 25}$

⑤ $\sqrt{\{9.6^2\times 19+14.4^2\times 25+(65.0-60.5)^2\times 19+(57.0-60.5)^2\times 25\}\div 44}$

具体的には，例えばこういった問題になる．

選択肢があるので，⑤あたりだなと見当はつくものの，こんな複雑な「公式」は知らない．

X ：データ数 m，平均値 $\bar{x}$ ，標準偏差 σ_X ；Y ：データ数 n，平均値 $\bar{y}$ ，標準偏差 σ_Y とし，X, Y を合併した Z に対して，

Z ：データ数 $m+n$，平均値 M，標準偏差 σ

とするとき，

$$\sigma^2=\frac{1}{m+n}\{m\sigma_X{}^2+n\sigma_Y{}^2+m(\bar{x}-M)^2+n(\bar{y}-M)^2\}$$

であることを示す．

$x_j=\bar{x}+k_j\sigma_X$，$y_j=\bar{y}+l_j\sigma_Y$ と表すと，明らかに $\sum_{j=1}^{m}k_j=\sum_{j=1}^{n}l_j=0$ である．

また，$m\sigma_X{}^2=\sum_{j=1}^{m}(x_j-\bar{x})^2=\sum_{j=1}^{m}(k_j\sigma_X)^2=\sigma_X{}^2\sum_{j=1}^{m}k_j{}^2$ より，$\sum_{j=1}^{m}k_j{}^2=m$．同様にして $\sum_{j=1}^{n}l_j^2=n$ である．

すると $\sigma^2=\dfrac{1}{m+n}\left\{\sum_{j=1}^{m}x_j{}^2+\sum_{j=1}^{n}y_j{}^2\right\}-M^2$ より

$$(m+n)\sigma^2=\sum_{j=1}^{m}x_j{}^2+\sum_{j=1}^{n}y_j{}^2-(m+n)M^2=\sum_{j=1}^{m}(\bar{x}+k_j\sigma_X)^2+\sum_{j=1}^{n}(\bar{y}+l_j\sigma_Y)^2-(m+n)M^2$$

$$=\sum_{j=1}^{m}(\bar{x})^2+2\sum_{j=1}^{m}\bar{x}k_j\sigma_X+\sum_{j=1}^{m}(k_j\sigma_X)^2+\sum_{j=1}^{n}(\bar{y})^2+2\sum_{j=1}^{n}\bar{y}l_j\sigma_Y+\sum_{j=1}^{n}(l_j\sigma_Y)^2-(m+n)M^2$$

$$=m(\bar{x})^2+0+m\sigma_X{}^2+n(\bar{y})^2+0+n\sigma_Y{}^2-(m+n)M^2=m\sigma_X{}^2+n\sigma_Y{}^2+m\{(\bar{x})^2-M^2\}+n\{(\bar{y})^2-M^2\}$$

ここで，

$$m\{(\bar{x})^2-M^2\}=m\{(\bar{x})^2-2\bar{x}M+M^2\}+m(2\bar{x}M-2M^2)=m(\bar{x}-M)^2+2mM(\bar{x}-M)$$

であるから，

$$m\{(\bar{x})^2-M^2\}+n\{(\bar{y})^2-M^2\}=m(\bar{x}-M)^2+2mM(\bar{x}-M)+n(\bar{y}-M)^2+2nM(\bar{y}-M)$$

$= m(\bar{x}-M)^2 + n(\bar{y}-M)^2 + 2M(m\bar{x}-mM+n\bar{y}-nM) = m(\bar{x}-M)^2 + n(\bar{y}-M)^2 + 2M\{(m\bar{x}+n\bar{y})-(m+n)M\}$
$= m(\bar{x}-M)^2 + n(\bar{y}-M)^2 + 2M\{(m+n)M-(m+n)M\} = m(\bar{x}-M)^2 + n(\bar{y}-M)^2$
となる．

したがって，

$$\sigma^2 = \frac{1}{m+n}\{m{\sigma_X}^2 + n{\sigma_Y}^2 + m(\bar{x}-M)^2 + n(\bar{y}-M)^2\}$$

であり，□の公式が示された．

複雑な式だが，${\sigma_X}^2$ と ${\sigma_Y}^2$ の加重平均と $(\bar{x}-M)^2$ と $(\bar{y}-M)^2$ の加重平均の和である．

なお，データ Y が，p 個の $\bar{x}+\sigma_X$ と p 個の $\bar{x}-\sigma_X$ からなるとき，

$\bar{y}=\bar{x}=M$，${\sigma_Y}^2 = \frac{1}{2p}(p{\sigma_X}^2 + p{\sigma_X}^2) = {\sigma_X}^2$ であるから，X と Y を合併した Z について，□の公式により，分散は（平均値も）X のそれと等しい．つまり，X に複数のデータを追加して，平均値も分散も X のそれらと変えないようなデータ（セット）が存在する．

3　分散とデータの存在範囲可能性

次は，ある大学の推薦入試の問題だそうである．

> 変数を x とし，n 個のデータ x_1, ……, x_n が与えられている．データの平均値を $\bar{x}$，標準偏差を s とする．
>
> (1) データの平均値 $\bar{x}$，分散 s^2，標準偏差 s の定義と意味を述べなさい．
>
> (2) $n=8$ のとき，$|x_i-\bar{x}|>3s$ となるデータの値 x_i は存在する可能性はあるか．もし存在するとしたら，最大何個のデータの値 x_i が存在するか．また，$|x_i-\bar{x}|>2s$ の場合はどのように考えられるか，あなたの考えを述べなさい．

(2)は興味深い問題である．

標準偏差 s は 0 でないとしてよいであろう．

$x_i=\bar{x}+k_i s$ と表すと，$|x_i-\bar{x}|>3s \iff |k_i|>3$ である．

また，分散 $s^2 = \frac{1}{8}\sum_{i=1}^{8}(x_i-\bar{x})^2 = \frac{1}{8}\sum_{i=1}^{8}(k_i s)^2 = \frac{s^2}{8}\sum_{i=1}^{8}{k_i}^2$ で，$s^2\neq 0$ より $\sum_{i=1}^{8}{k_i}^2 = 8$ … ①

が成り立つ．

$\max|k_i|>3$ のとき，①の左辺 >9 であるから，①は成り立たない．したがって，$|x_i-\bar{x}|>3s$ となるデータの値 x_i は存在しない．

また，$|x_i-\bar{x}|>2s \iff |k_i|>2$ であり，$\max|k_i|>2$ のとき，①の左辺>4 であるから，①を成り立たせるデータの値 x_i は存在しうる．よって，$|x_i-\bar{x}|>2s$ となるデータの値 x_i は存在しうる．ただし，このようなデータが 2 個以上の場合，①の左辺>8 となるから，2 個以上ということはありえず，存在したとしても 1 個である．

平均値から 3σ 以上離れたデータが存在するには，データ数が 9 個以上でなければならないという結果は，データ数が少ない特殊なケースであるとは言え，分散の 1 つの特質と言って良い．

外れ値は（最近は(?)）四分位範囲を基準に設定されることが多いのかも知れないが，標準偏差を用いた設定もあり，その場合，データ数が少なければ外れ値が絶対に存在しないことがあると言うことである．

4　終わりに

世は統計ブームだと言うが，計算はなかなか煩雑である．

ここでは，幾つかの問題について，データを **平均＋係数･標準偏差** の形に表して計算してみたが，この方法でスッキリ見やすくなる部分があることは事実であるように思われる．

（2019 年 10 月 17 日）