雑感　データの追加と分散の変化

■　あるデータの平均値や分散などが分かっているとする．そこに 1 つのデータを付け加えたとき，平均や分散はどのように変化するだろうか．

もとのデータを $\{x_1, x_2, \cdots, x_n\}$ とし，平均値を $\overline{x}$，分散を σ^2 とする．そこに，データ x_{n+1} を付け加えたときの平均を $\overline{x}'$，分散を σ'^2 とする．

■　平均値については容易であり，$(n+1)\overline{x}' = n\overline{x} + x_{n+1}$ であるから，$\overline{x}' = \dfrac{n\overline{x} + x_{n+1}}{n+1}$ である．

■　分散は容易ではない．ちなみに，$\sigma^2 = \dfrac{1}{n}\sum_{k=1}^{n} x_k{}^2 - \left(\dfrac{1}{n}\sum_{k=1}^{n} x_k\right)^2$ であり，$\sigma'^2 = \dfrac{1}{n+1}\sum_{k=1}^{n+1} x_k{}^2 - \left(\dfrac{1}{n+1}\sum_{k=1}^{n+1} x_k\right)^2$ である．

これらの関係を明らかにするには，こつこつ計算してもよいのだが，以前の「雑感 128　合併 2 集団の分散」の次の結果を利用するとよい．

> 集団 X ：データ数 m，平均値 $\overline{x}$，標準偏差 σ_X
> 集団 Y ：データ数 n，平均値 $\overline{y}$，標準偏差 σ_Y
> とし，2 集団 X, Y を合併した集団 Z に対して，
> 集団 Z ：データ数 $m+n$，平均値 M，標準偏差 σ
> とするとき，
> $$\sigma^2 = \frac{1}{m+n}\{\sigma_X{}^2 m + \sigma_Y{}^2 n + (\overline{x} - M)^2 m + (\overline{y} - M)^2 n\}$$

これによれば，集団 X を元の集団，$Y = \{x_{n+1}\}$ とし，$\sigma_Y = 0$ に注意して，

$$\sigma'^2 = \frac{1}{n+1}\{\sigma^2 n + (\overline{x} - \overline{x}')^2 n + (x_{n+1} - \overline{x}')^2\}$$
$$= \frac{n}{n+1}\left\{\sigma^2 + \left(\frac{\overline{x} - x_{n+1}}{n+1}\right)^2 + \left(\frac{\overline{x} - x_{n+1}}{n+1}\right)^2 n\right\}$$ となる．

■　さて，データ x_{n+1} によって，分散が増加する場合と減少する場合があるが，それはどのような場合であろうか．

$x_{n+1} = \overline{x} + k\sigma$ $(k > 0)$ とおいて，k についての条件を求める．

$\sigma'^2 - \sigma^2$ を計算すると，$\sigma'^2 = \dfrac{\sigma^2 n}{(n+1)^2}(n+1+k^2)$ であるから，

$\sigma'^2 - \sigma^2 = \dfrac{n(n+1+k^2) - (n+1)^2}{(n+1)^2}\sigma^2$ より

$\sigma' > \sigma \iff k > \sqrt{\dfrac{n+1}{n}}$，$\sigma' < \sigma \iff k < \sqrt{\dfrac{n+1}{n}}$

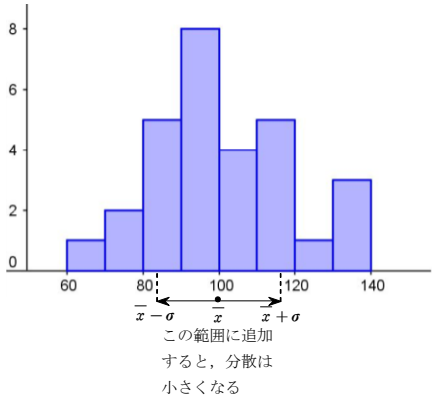
を得る．

■　なるほど！　素晴らしい結果が得られたことになる．

n が一定大きければ，$\sqrt{\dfrac{n+1}{n}} \fallingdotseq 1$ であるから，右のようなイメージになる．

区間 $[\overline{x} - \sigma, \overline{x} + \sigma]$ に追加すれば分散は小さくなり，それ以外（両サイド）に追加すれば分散は大きくなる．

この結果は，標準偏差 σ の意味を再認識させる．



■　なお，この結果は，以前の「雑感 93　データの分析のオリジナル問題」の(2)も関連しているものと考えられる．