雑感　データの分析のオリジナル問題

■　数学Ⅰに『データの分析』という新しい内容が設定された．
参考書や問題集を見ると，どういう問題を扱ったらいいか編集者が苦慮しているように思われる．

■　試みに問題を 1 つ作ってみた．生徒にとっては取っつきがよくはないだろうが，内容的に難しいものではない．実力考査などにいかがであろうか．

> 標本数 $n(\geqq 2)$ のデータ $X=\{x_1,\ x_2,\ x_3,\ \cdots,\ x_n\}$ の平均値が m，標準偏差が $\sigma(>0)$ であるとする．
> (1)　X に 1 つの標本 x を付け加えたデータでは，「平均値も標準偏差も変わらない」ということはないことを示せ．
> (2)　X に 2 つの標本 $y_1,\ y_2\ (y_1 \geqq y_2)$ を付け加えたデータで，平均値も標準偏差も変わらないとき，$y_1,\ y_2$ を求めよ．
> (3)　X に 3 つの標本 $a,\ b,\ c\ (a \geqq b \geqq c)$ を付け加えたデータで，平均値も標準偏差も変わらないとする．
> 　$a=m+s\sigma,\ b=m+t\sigma,\ c=m+u\sigma$ と表すとき，$t,\ u$ を s で表せ．

＜解＞　$m=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k,\ \sigma^2=\dfrac{1}{n}\sum_{k=1}^{n}(x_k-m)^2$ である．

(1)　平均値が m で変わらないから，付け加えたデータ x は m に等しい．このとき，新たな分散の値は

$\dfrac{1}{n+1}\left\{\displaystyle\sum_{k=1}^{n}(x_k-m)^2+(m-m)^2\right\}=\dfrac{1}{n+1}(n\sigma^2)=\dfrac{n}{n+1}\sigma^2 \neq \sigma^2$ となり，分散は一致しない．

よって，平均値も標準偏差も変わらないことはない．

(2)　平均値が m で変わらず，$y_1 \geqq y_2$ であるから，$d \geqq 0$ として，$y_1=m+d\sigma,\ y_2=m-d\sigma$ と表せる．

このとき，新たな分散の値は

$\dfrac{1}{n+2}\left\{\displaystyle\sum_{k=1}^{n}(x_k-m)^2+(d\sigma)^2+(-d\sigma)^2\right\}=\dfrac{1}{n+2}(n\sigma^2+2d^2\sigma^2)$ である．これが σ^2 に等しいから，$\dfrac{\sigma^2(n+2d^2)}{n+2}=\sigma^2$．

$\sigma^2>0$ より $n+2d^2=n+2$．$d \geqq 0$ から $d=1$．

よって，$y_1=m+\sigma,\ y_2=m-\sigma$．

(3)　平均値が m で変わらないから，付け加えた $a,\ b,\ c$ の平均値が m である．したがって，

$(m+s\sigma)+(m+t\sigma)+(m+u\sigma)=3m$ である．

$\sigma>0$ から $s+t+u=0$．

また，新たな分散の値は

$$\frac{1}{n+3}\left\{\sum_{k=1}^{n}(x_k-m)^2+(s\sigma)^2+(t\sigma)^2+(u\sigma)^2\right\}$$

$$=\frac{1}{n+3}\{n\sigma^2+s^2\sigma^2+t^2\sigma^2+u^2\sigma^2\}$$

で，これが σ^2 に等しいから，$\dfrac{\sigma^2}{n+3}(n+s^2+t^2+u^2)=\sigma^2$．

$\sigma^2>0$ より $n+s^2+t^2+u^2=n+3$．よって $s^2+t^2+u^2=3$．

$u=-s-t$ から $s^2+t^2+(-s-t)^2=3$．よって $t=\dfrac{-s\pm\sqrt{6-3s^2}}{2}$，

$u=\dfrac{-s\mp\sqrt{6-3s^2}}{2}$（複号同順）．

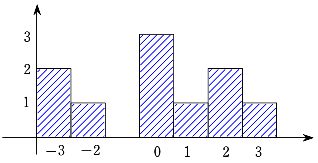
$a \geqq b \geqq c$ から $s \geqq t \geqq u$ で，$s+t+u=0$ から $s \geqq 0$ より

$t=\dfrac{-s+\sqrt{6-3s^2}}{2}$，$u=\dfrac{-s-\sqrt{6-3s^2}}{2}$ である．

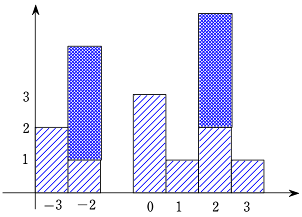
■　この問題で注意してほしいのは，このデータ X は特殊なデータではなく，一般的なデータであると言うことである．ヒストグラムが左右対称であるなどと言うことも必要ない．

とりわけ，(2)の結果は極めて興味深い．この論を進めれば，データ X に $m-\sigma$ の値を持つ標本を p 個と $m+\sigma$ の値を持つ標本を p 個付け加えても，平均値も標準偏差も変わらないということになる．

例えば，$X=\{-3,\ -3,\ -2,\ 0,\ 0,\ 0,\ 1,\ 2,\ 2,\ 3\}$ の平均値は 0，標準偏差は 2 である．ここに，$\{-2,\ -2,\ -2,\ 2,\ 2,\ 2\}$ という標本を付け加えたデータについても，（平均値が変わらないのは当然だが）標準偏差は 2 であり，変わらない（下のヒストグラムの右が追加後である）．



$m=0,\quad \sigma=2$



$m=0,\quad \sigma=2$

σ という値の「本質」のようなものを垣間見る気がする．

■　また，(3)の $t=\dfrac{-s+\sqrt{6-3s^2}}{2}$，$u=\dfrac{-s-\sqrt{6-3s^2}}{2}$ は実数だから，$6-3s^2 \geqq 0$ であり，$s \geqq 0$ と併せて，$0 \leqq s \leqq \sqrt{2}$ である．

これがどういうことかおわかりであろうか．

データ X に 3 つの標本 $a,\ b,\ c$ を付け加えて，平均と標準偏差が変わらないようにするとき，付け加える標本の最大値は $m+\sqrt{2}\sigma$ でなければならないということである．もし，$m+2\sigma$ という標本を付け加えてしまったら，付け加えるあと 2 つの標本の値をどのように設定しても，平均値と標準偏差を変わらないようにはできないということなのである．

もちろん，付け加える標本数を多くすれば話は別である．

こういったことも，「問題」に仕立て上げることもできるだろう．

2013 年 3 月 8 日