雑感　相関係数に関する注意

■　昨年 11 月に大学入試センターが発表した 2015 年からのセンター試験の試作問題の「解答・解説」を「雑感 111」に載せたのだが，恥ずかしいことに問題文を読み間違えるという初歩的なミスで答を間違えてしまっていて，修正した．

恥ずかしいことである．

どの部分のどのような誤読かというと，次の通りである．

> (3)　相関係数の一般的な性質に関する次の [A] から [C] の説明について，［ス］ということがいえる。［ス］に当てはまるものを，次の⓪～④のうちから一つ選べ。
>
> [A]　相関係数 r は，常に $-1 \leqq r \leqq 1$ であり，すべてのデータが 1 つの曲線上に存在するときには，いつでも $r=1$ または $r=-1$ である。
>
> [B]　もとのデータを定数倍しても，相関係数の値は変わらないが，もとのデータに定数を加えると相関係数の値は変わる。
>
> [C]　2 つの変量間の相関係数の値が高い場合には，これらの 2 つの変量には因果関係があるといえる。
>
> ⓪　[A] だけが正しい　　①　[B] だけが正しい
>
> ②　[C] だけが正しい　　③　[A] だけが間違っている
>
> ④　⓪～③のどれでもない

この[A]の「曲線上」という部分を「直線上」と誤読して，この文章は「正しい」としてしまったのである．もちろん，「正しくない」が正解である．

■　さて，「直線」だったら，「正しい」で良いのかと言うことを考察してみたい．

雑感に記したように，データ $\{x_k\}$, $\{y_k\}$ に $y_k=ax_k+b$ の関係がある（データが直線上にある）場合，$\sigma_x{}^2=\sum(x_k-\overline{x})^2/n$,

$\sigma_y{}^2=\sum(y_k-\overline{y})^2/n=\sum\{ax_k+b-(a\overline{x}+b)\}^2/n=a^2\sum(x_k-\overline{x})^2/n=a^2\sigma_x$,

$\sigma_{xy}=\sum(x_k-\overline{x})(y_k-\overline{y})/n=\sum(x_k-\overline{x})a(x_k-\overline{x})/n=a\sigma_x$ から，相関係数 $r=\dfrac{\sigma_{xy}}{\sigma_x\sigma_y}=\dfrac{a\sigma_x}{\sigma_x\sqrt{a^2\sigma_x}}=\dfrac{a}{|a|}=\pm1$ である．

として良かったのかである．

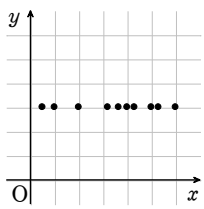
■　実は，相関係数が存在しない場合がある．

上の式変形で言えば，$r=\dfrac{a}{|a|}$ で，$a=0$ の場合と，直線の方程式を $y=ax+b$ でなく，$x=c$ とおくべき場合である．

そのケースは，データ数が $n\,(\geqq 2)$ のデータ $\{x_k\}$, $\{y_k\}$ で，$x_1=x_2=\cdots=x_n$ または $y_1=y_2=\cdots=y_n$ の場合である．

この場合は，$\sigma_x=0$ または $\sigma_y=0$ なので，相関係数 r の定義式　$r=\dfrac{\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sigma_x\sigma_y}$　の分母が 0 となってしまい，値が存在しないのである．

このとき，相関図でデータは x 軸に平行な直線または y 軸に平行な直線上に並ぶことになる．



すると「すべてのデータが 1 つの直線上に存在するとき，$r=1$ または $r=-1$ または，相関係数が存在しない」ことになる．

■　これは正直，想定外のことであった．

相関係数 r が $r=1$ または $r=-1$ ならば，データがすべて一直線上に存在するは正しいが，その逆は真ではないのであった．

2014 年 2 月 4 日