# 強者の戦略

今回の問題は，2016 年度入試において京都府立医科大学で出題された問題でした．まず問題を確認してみましょう．

問題

$n$ は 2 以上の整数とする．変量 $x$ についてのデータの値を $x_k(1 \leqq k \leqq n)$ とし，変量 $y$ についてのデータの値を $y_k(1 \leqq k \leqq n)$ とする．変量 $z$ はデータの値が $x_k y_k$ $(1 \leqq k \leqq n)$ である変量を表す．

(1)　変量 $x$ と $y$ の $n$ 個の値の組を $(x_k,\ y_k)(1 \leqq k \leqq n)$ としたときの $x$ と $y$ の共分散 $s_{xy}$（偏差の積の平均）について

$$s_{xy} = \overline{z} - \overline{x}\,\overline{y}$$

が成り立つことを証明せよ．ここで $\overline{x}$，$\overline{y}$，$\overline{z}$ はそれぞれ変量 $x$，$y$，$z$ についてのデータの値の平均値を表す．

0 以上の整数 $a$ と 1 以上の整数 $b$ に対し，$a$ を $b$ で割った余りを $R_b(a)$ と表す．$l$，$m$ は 2 以上 $n$ 以下の整数とする．

変量 $x$ と $y$ の $n$ 個の値の組を

$(x_k,\ y_k) = (R_l(k-1)+1,\ R_m(k-1)+1)$　$(1 \leqq k \leqq n)$

としたときの $x$ と $y$ の相関係数を $r$ とする．

(2)　$l$ は $n$ の約数とし，$m = n$ であるとき，$r$ を求めよ．

続いて，(1) の解答です．

［解答 1］

(1)　共分散の定義より

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}) \\ &= \frac{1}{n}\sum_{k=1}^{n}(x_k y_k - \overline{x}y_k - x_k\overline{y} + \overline{x}\,\overline{y}) \\ &= \frac{1}{n}\sum_{k=1}^{n}x_k y_k - \overline{x}\cdot\frac{1}{n}\sum_{k=1}^{n}y_k - \overline{y}\cdot\frac{1}{n}\sum_{k=1}^{n}x_k + \frac{1}{n}\cdot n\overline{x}\,\overline{y} \\ &= \overline{z} - \overline{x}\,\overline{y} - \overline{y}\,\overline{x} + \overline{x}\,\overline{y} \\ &= \overline{z} - \overline{x}\,\overline{y} \end{aligned}$$

となる．以上より，示された．

（証明終）

《解説》

問題にもあるように

共分散：偏差の積の平均値

であり，また

偏差：データの値と平均値の差

なので，まずは定義に従って Σ 記号を用いた式を立式します．その後，Σ 計算のルールに従って 4 つの部分に分けて考えれば，$\overline{x}$，$\overline{y}$，$\overline{z}$ に相当する部分を作ることができ，証明することができます．高 3 生の強者メンバーにとっては，簡単な問題だったと思います．高 1，高 2 生でこれから強者にならんとしている方の場合は

- 「データの分析」は数学 I の単元だが，公式の証明において，数学 B「数列」の Σ 記号を用いると便利．
- Σ 計算では，変数（今回は $k$）に関係のない実数倍の部分は，Σ 記号の外に出せる．
- $n$ 個のデータの値 $x_k(1 \leqq k \leqq n)$ に対し

$$\frac{1}{n}\sum_{k=1}^{n}x_k = \frac{x_1 + x_2 + \cdots\cdots + x_n}{n}$$

が平均値に相当する．

という，3 点を押さえておくとよいでしょう．不安が残る方は，以下に挙げる分散を求める公式が有名ですので，証明も含めて覚えておきましょう．

$n$ 個のデータの値 $x_k(1 \leqq k \leqq n)$ の分散 $s_x{}^2$ は，平均値を $\overline{x}$ とおくと

$$\begin{aligned} s_x{}^2 &= \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2 \\ &= \frac{1}{n}\sum_{k=1}^{n}\{x_k^2 - 2x_k\overline{x} + (\overline{x})^2\} \\ &= \frac{1}{n}\sum_{k=1}^{n}x_k^2 - 2\overline{x}\cdot\frac{1}{n}\sum_{k=1}^{n}x_k + \frac{1}{n}\cdot n(\overline{x})^2 \\ &= \overline{x^2} - 2\overline{x}\cdot\overline{x} + (\overline{x})^2 \\ &= \overline{x^2} - (\overline{x})^2 \end{aligned}$$

となる.

（証明終）

(2) の解答で分散を計算する際にも，この公式を用いますので，頭の片隅に置きつつ，続きを読んでください.

次に，(2) の解答を考えてみましょう．まずは，問題文の

$$(x_k,\ y_k)=(R_l(k-1)+1,\ R_m(k-1)+1) \qquad (1\leqq k\leqq n)$$

の意味を素早く読み取る必要があります．$x_k$ の式の一部である $R_l(k-1)$ は，問題より

$(k-1)$ を $l$ で割った余り

です．$k=1,\ 2,\ \cdots\cdots,\ n$ のとき

$$k-1=0,\ 1,\ \cdots\cdots,\ n-1$$

ですから，$l$ で割った余りは

$$0,\ 1,\ 2,\ \cdots\cdots,\ l-1$$

を繰り返します．これに1を足したものが $x_k$ なので，$x_k$ は

$$1,\ 2,\ \cdots\cdots,\ l$$

を $l$ 個周期で繰り返すわけです．しかも，$l$ が $n$ の約数なので，$x_n=l$ となり，$x_n$ が必ず周期の最後になります．（後に挙げる解答では，繰り返しの回数を $d$ 回と置くことで，計算を見やすくしています.）

ここまで読み取れれば，$y_k$ については $m=n$ のため，$k-1$ を $m\,(=n)$ で割った余りが

$$0,\ 1,\ 2,\ \cdots\cdots,\ n-1$$

で，$y_k$ はこれに1を加えて

$$1,\ 2,\ \cdots\cdots,\ n$$

となることが，すぐにわかります.

以上を踏まえて，解答を作成してみましょう.

［解答 2］

(2)　$x$，$y$ の標準偏差をそれぞれ $s_x$，$s_y$ とおく.

$l$ が $n$ の約数で，$m=n$ のとき，変量 $x$，$y$ の $n$ 個のデータの値は，それぞれ次のようになる.

| $x_1$ | $x_2$ | $\cdots$ | $x_l$ | $x_{l+1}$ | $x_{l+2}$ | $\cdots$ | $x_{n-1}$ | $x_n$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | $\cdots$ | $l$ | 1 | 2 | $\cdots$ | $l-1$ | $l$ |

| $y_1$ | $y_2$ | $\cdots$ | $y_l$ | $y_{l+1}$ | $y_{l+2}$ | $\cdots$ | $y_{n-1}$ | $y_n$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | $\cdots$ | $l$ | $l+1$ | $l+2$ | $\cdots$ | $n-1$ | $n$ |

つまり，$x_k\,(1\leqq k\leqq n)$ は1から $l$ の値を小さいものから順に $\dfrac{n}{l}$ 回繰り返し，$y_k\,(1\leqq k\leqq n)$ は1から $n$ までの値が小さいものから順に並ぶ．ここで，$d$ を自然数として

$$\frac{n}{l}=d \iff n=dl$$

とおくと，

$$\overline{x}=\frac{1}{n}\cdot d\sum_{k=1}^{l}k=\frac{d}{dl}\cdot\frac{l(l+1)}{2}=\frac{l+1}{2}$$

$$\overline{x^2}=\frac{1}{n}\cdot d\sum_{k=1}^{l}k^2=\frac{d}{dl}\cdot\frac{l(l+1)(2l+1)}{6}=\frac{(l+1)(2l+1)}{6}$$

ゆえ，変量 $x$ の分散は

$$\begin{aligned}s_x^{\,2}&=\overline{x^2}-(\overline{x})^2\\&=\frac{(l+1)(2l+1)}{6}-\frac{(l+1)^2}{4}\\&=\frac{(l+1)}{12}\{2(2l+1)-3(l+1)\}\\&=\frac{(l-1)(l+1)}{12}\end{aligned}$$

となる．また，$m=n$ のとき，$m$ も $n$ の約数なので，変量 $y$ の分散について，変量 $x$ の分散と同様に考えることができ

$$\overline{y}=\frac{m+1}{2}=\frac{n+1}{2}$$

$$s_y^{\,2}=\frac{(m-1)(m+1)}{12}=\frac{(n-1)(n+1)}{12}$$

となる．次に，$\overline{z}$ を計算すると

$$\begin{aligned}\overline{z}&=\frac{1}{n}\sum_{k=1}^{n}x_ky_k\\&=\frac{1}{n}\sum_{j=0}^{d-1}\left\{\sum_{i=1}^{l}i(jl+i)\right\}\quad\cdots\cdots\cdots(*)\\&=\frac{1}{n}\sum_{j=0}^{d-1}\left\{jl\cdot\frac{l(l+1)}{2}+\frac{l(l+1)(2l+1)}{6}\right\}\end{aligned}$$

$$=\frac{1}{n}\cdot\frac{l(l+1)}{6}\sum_{j=0}^{d-1}\{3jl+(2l+1)\}$$
$$=\frac{1}{n}\cdot\frac{l(l+1)}{6}\left\{3l\sum_{j=1}^{d-1}j+(2l+1)d\right\}$$
$$=\frac{1}{n}\cdot\frac{l(l+1)}{6}\left\{3l\cdot\frac{(d-1)d}{2}+(2l+1)d\right\}$$
$$=\frac{1}{n}\cdot\frac{l(l+1)}{6}\cdot\frac{d}{2}\{3l(d-1)+2(2l+1)\}$$
$$=\frac{l+1}{12}\{3(n-l)+2(2l+1)\}\quad(\because\ dl=n)$$
$$=\frac{(l+1)(n-l)}{4}+\frac{(l+1)(2l+1)}{6}$$

となる．さらに (1) より，変量 $x$ と $y$ の共分散は

$$s_{xy}=\overline{z}-\overline{x}\,\overline{y}$$
$$=\frac{(l+1)(n-l)}{4}+\frac{(l+1)(2l+1)}{6}-\frac{l+1}{2}\cdot\frac{n+1}{2}$$
$$=\frac{l+1}{12}\{3(n-l)+2(2l+1)-3(n+1)\}$$
$$=\frac{(l+1)(l-1)}{12}$$
$$=s_x^{\,2}$$

となる．よって，求める相関係数 $r$ は

$$r=\frac{s_{xy}}{s_x s_y}=\frac{s_x^{\,2}}{s_x s_y}=\frac{s_x}{s_y}$$
$$=\sqrt{\frac{\frac{(l-1)(l+1)}{12}}{\frac{(n-1)(n+1)}{12}}}$$
$$=\sqrt{\frac{l^2-1}{n^2-1}}$$

である．

《補足 1》

相関係数は，共分散を標準偏差（＝分散の正の平方根）の積で割ったものですから，共分散と分散を求めにいきます．分散を求める際は，(1) の《解説》の部分で証明した

$$s_x^{\,2}=\overline{x^2}-(\overline{x})^2$$

の公式を使っていきましょう．

《補足 2》

$\overline{z}$ を計算する際の，(*) 部分を省略せずに書くと，以下のようになります．

$$\overline{z}=\frac{1}{n}\sum_{k=1}^{n}x_k y_k$$
$$=\frac{1}{n}[1\cdot1+2\cdot2+\cdots\cdots+l\cdot l$$
$$+1(l+1)+2(l+2)+\cdots\cdots+l(l+l)$$
$$+\cdots\cdots$$
$$+1\{(d-1)l+1\}+2\{(d-1)l+2\}+\cdots\cdots+l\cdot n\,]$$
$$=\frac{1}{n}\sum_{j=0}^{d-1}\{1(jl+1)+2(jl+2)+\cdots\cdots+l(jl+l)\}$$
$$=\frac{1}{n}\sum_{j=0}^{d-1}\left\{\sum_{i=1}^{l}i(jl+i)\right\}$$

また，上に挙げた部分を Σ を二重に用いて手早く書くことができたとしても，その後の計算部分がどうしても長くなり，最後の答えが綺麗な式になるかどうか不明なまま，先の見えない計算を続けないといけなくなります．実は，次の《補足 3》の事実に気づいていれば，もう少しだけ見通しが立てやすくなります．

《補足 3》

今回の問題では，共分散 $s_{xy}$ を計算してみた結果，$x$ の分散 $s_x^{\,2}$ と一致しました．この仕組みについて，以下のように説明することができます．

変量 $x$ について，$1\leqq i\leqq l$ においては

$$x_i=i$$

が成り立つので

$$\overline{z}=\frac{1}{n}\sum_{k=1}^{n}x_k y_k$$
$$=\frac{1}{n}\sum_{j=0}^{d-1}\left\{\sum_{i=1}^{l}i(jl+i)\right\}$$
$$=\frac{1}{n}\sum_{j=0}^{d-1}\left\{\sum_{i=1}^{l}x_i(jl+x_i)\right\}$$
$$=\frac{1}{n}\sum_{j=0}^{d-1}jl\left(\sum_{i=1}^{l}x_i\right)+\frac{1}{n}\sum_{j=0}^{d-1}\left(\sum_{i=1}^{l}x_i^{\,2}\right)$$
$$=\frac{l}{n}\sum_{j=0}^{d-1}j\cdot\frac{n\overline{x}}{d}+\frac{1}{n}\cdot d(1^2+2^2+\cdots\cdots+l^2)\quad\cdots\cdots(**)$$
$$=\frac{l}{n}\cdot\frac{n\overline{x}}{d}\cdot\frac{(d-1)d}{2}+\overline{x^2}$$
$$=\frac{l(d-1)}{2}\overline{x}+\overline{x^2}$$

$$=\frac{n-l}{2}\overline{x}+\overline{x^2} \quad (\because dl=n)$$
$$=\left(\frac{n+1}{2}-\frac{l+1}{2}\right)\overline{x}+\overline{x^2}$$
$$=(\overline{y}-\overline{x})\overline{x}+\overline{x^2}$$
$$=\overline{x}\,\overline{y}+\overline{x^2}-(\overline{x})^2$$
$$=\overline{x}\,\overline{y}+s_x^2$$
$$\therefore\ s_x^2=\overline{z}-\overline{x}\,\overline{y}=s_{xy}\ (\because\ (1))$$

【注意】

計算中の (**) の部分では

$$d(x_1+x_2+\cdots\cdots+x_l)=n\overline{x}$$

の関係を用いています．各辺ともに，変量 $x$ の $n$ 個の総和

$$x_1+x_2+\cdots\cdots+x_n$$

を計算しており，左辺は変量 $x$ が 1，2，……，$l$ を $d$ 回繰り返す，つまり，$x_1$，$x_2$，……，$x_l$ を $d$ 回繰り返すことから，右辺は

$$\frac{x_1+x_2+\cdots\cdots\cdots+x_n}{n}=\overline{x}$$
$$\Longleftrightarrow\ x_1+x_2+\cdots\cdots+x_n=n\overline{x}$$

という式から求めています．

ただ，試験会場でこの問題を初めて見た場合，この事実に気づくことは困難だと思われますので，実際には解答で用いているような，気合で計算する解法が無難な攻め方になるでしょう．特に医科大学の入試では，膨大な量の問題文章の中から短時間で必要な情報を読み取ったり，時間内に解けそうな問題を見つけ出したりして，いかに適切に部分点を稼ぐか，という戦い方が大事になる場合もあります．今回も細部にこだわり過ぎることなく，まずは結論を導き出すことに重点を置きたいところです．

（最後に）

今年度の「強者の戦略」の冊子を作成するにあたり，受験生の体験談に目を通していた際，「京都府立医科大学の入試において，問題冊子が配られる．表紙から問題がうっすらと透けて見えたので１問目を確認すると，「データの分析」の問題と分かり青ざめる」という言葉を見つけたときから，『この問題は取り上げておかなければ！』と考えていました．医歯薬系，特に医学部を目指す場合は，どうしても小さな失点が不合格に繋がる場合があります．医学部を志す強者の皆さんは，単元によって食わず嫌いなどをせず，試験範囲内に含まれている問題については，是非，公式・解法を確認し，自分の考えを答案に描き出せるよう，心構えをしておいてください．

（数学科　中西）