データの分析と数列
～分散と数列の和～

西元（にしもと）　教善（のりよし）

§1. はじめに

x を変量とし，データの n 個の値 x_1，x_2，…，x_n が与えられているとき，データの平均値 $\overline{x}$ つまり $\overline{x}=\dfrac{1}{n}(x_1+x_2+\cdots+x_n)$ は，データの n 個の値 x_1，x_2，…，x_n をこの順の数列 $\{x_k\}_{k=1,2,\cdots,n}$ とみなすとき，初項から第 n 項までの和 S_n を項数 n で割ったものである。

このようにデータの値は，その順での数列とみなすことで，数列の考え方が使える。

本稿では，データの値としての $x：x_1$，x_2，…，x_n の分散と数列としての $\{x_n\}：x_1$，x_2，…，x_n の和を考察してみる。

§2. 1から n までの自然数の並び替え数列

x を変量とし，データの値として1から n までの自然数を1つずつとるとき，それがどのように並んでいたとしても平均値 $\overline{x}$ は，

$$\frac{1}{n}\sum_{k=1}^{n}k=\frac{1}{n}\cdot\frac{1}{2}n(n+1)=\frac{n+1}{2}$$

である。並び方を変えると数列としては変わるがその和は変わらない。

また，x の分散 ${\sigma_x}^2$ は

$$\begin{aligned}{\sigma_x}^2&=\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2\\&=\frac{1}{n}\sum_{k=1}^{n}\left(k-\frac{n+1}{2}\right)^2\\&=\frac{1}{n}\sum_{k=1}^{n}\left\{k^2-(n+1)k+\frac{(n+1)^2}{4}\right\}\\&=\frac{1}{n}\left\{\frac{1}{6}n(n+1)(2n+1)\right.\\&\qquad\left.-(n+1)\cdot\frac{1}{2}n(n+1)+n\cdot\frac{(n+1)^2}{4}\right\}\\&=\frac{1}{6}(n+1)(2n+1)-\frac{1}{2}(n+1)^2+\frac{1}{4}(n+1)^2\\&=\frac{1}{6}(n+1)(2n+1)-\frac{1}{4}(n+1)^2\\&=\frac{1}{12}(n+1)\{2(2n+1)-3(n+1)\}\\&=\frac{1}{12}(n+1)(n-1)\\&=\frac{1}{12}(n^2-1)\end{aligned}$$

である。

次に，1から n までの自然数を並び替えた数列（以後「**並び替え数列**」という）の1つを $\{a_k\}_{k=1,2,\cdots,n}$ とすれば，これをデータの値とみなすとき，変数 a_k の平均値 $\overline{a_k}$ は $\dfrac{n+1}{2}$ であり，変数 a_k の分散 ${\sigma_{a_k}}^2=\dfrac{1}{n}\sum_{k=1}^{n}(a_k-\overline{a_k})^2$ は $\dfrac{n^2-1}{12}$ である。

なお，$\overline{a_k}$ は $\dfrac{n+1}{2}$（一定）であり，どのような並び替え数列においても1から n までの自然数がもれなく1回だけ出てその総和をとることから，特定の並び替え数列だけに成り立つのではなくすべての並び替え数列について成り立つ。

しかし，$\dfrac{1}{n}\sum_{k=1}^{n}(a_k-k)^2$ を考えると並び替えによって値が変化する。そこで，これが最大になるとき数列としてどのような並びになっているか，またそのときの値を求めてみよう。これは並び替えによって元のデータの値との散らばり具合にどのような変化が出るのかを考えるものである。

(1) 互換数列

ここで，i，j を $1\leqq i<j\leqq n$ である自然数として，**互換数列** $\{\boldsymbol{a}_{i,j}\}$ を次のように定める。

> **数列 $\{\boldsymbol{k}\}_{k=1,2,\cdots,n}$ の互換数列 $\{\boldsymbol{a}_{i,j}\}$ $\Longleftrightarrow$ 数列 $\{\boldsymbol{k}\}_{k=1,2,\cdots,n}$ において，第 $\boldsymbol{i}$ 項と第 $\boldsymbol{j}$ 項のみを入れ替えた数列**

つまり，数列 1，2，…，i，…，j，…，n において 1，2，…，j，…，i，…，n としたもののことである。

第 i 項　第 j 項

$\{k\}$：1，2，…，i，…，j，…，n

入れ替える

$\{a_{i,j}\}$：1，2，…，j，…，i，…，n

どの 2 つを入れ替えたかを明示するときは数列 $\{k\}_{k=1,2,\cdots,n}$ における ***i*，*j* 互換数列** と呼ぶことにする。

つまり，互換数列とは，2 つの項だけを入れ替えた並び替え数列のことである。また，任意の並び替え数列は互換を何回かすることで得られる。

⑵　互換数列と自然数列の差の平方の和が最大になるとき

ここで，互換数列 $\{a_{i,j}\}$ と自然数列 $\{k\}_{k=1,2,\cdots,n}$ の差の平方の和が最大になるときは，どのような互換数列であるかを考える。

そこで，2 つの互換数列（i，j 互換数列と l，m 互換数列）$\{a_{i,j}\}$，$\{a_{l,m}\}$ に対して，i，j 互換数列 $\{a_{i,j}\}$ と自然数列 $\{k\}_{k=1,2,\cdots,n}$ の差の平方の和と l，m 互換数列 $\{a_{l,m}\}$ と自然数列 $\{k\}_{k=1,2,\cdots,n}$ の差の平方の和を考える。つまり，

次のような式を考える。

$$A=\sum_{k=1}^{n}(a_{i,j}-k)^2-\sum_{k=1}^{n}(a_{l,m}-k)^2$$

この式の右辺を変形すると

$$\begin{aligned}A&=\sum_{k=1}^{n}(a_{i,j}-k)^2-\sum_{k=1}^{n}(a_{l,m}-k)^2\\&=\sum_{k=1}^{n}\{(a_{i,j}{}^2-2ka_{i,j}+k^2)-(a_{l,m}{}^2-2ka_{l,m}+k^2)\}\\&=\sum_{k=1}^{n}\{a_{i,j}{}^2-a_{l,m}{}^2+2k(a_{l,m}-a_{i,j})\}\\&=\sum_{k=1}^{n}a_{i,j}{}^2-\sum_{k=1}^{n}a_{l,m}{}^2+2\sum_{k=1}^{n}k(a_{l,m}-a_{i,j})\end{aligned}$$

$\sum_{k=1}^{n}a_{i,j}{}^2$ は，数列 $\{k\}_{k=1,2,\cdots,n}$ において，第 i 項と第 j 項のみを入れ替えた数列の総和であるから，$\sum_{k=1}^{n}k^2$ に等しいし，$\sum_{k=1}^{n}a_{l,m}{}^2$ は，数列 $\{k\}_{k=1,2,\cdots,n}$ において，第 l 項と第 m 項のみを入れ替えた数列の総和であるから，これも $\sum_{k=1}^{n}k^2$ に等しいので，

$\sum_{k=1}^{n}a_{i,j}{}^2=\sum_{k=1}^{n}a_{l,m}{}^2$ である。

よって，$A=2\sum_{k=1}^{n}k(a_{l,m}-a_{i,j})$ である。

第 i 項　第 l 項　第 j 項　第m項

$\{a_{l,m}\}$：1，2，…，i，…，m，…，j，…，l，…，n

$\{a_{i,j}\}$：1，2，…，j，…，l，…，i，…，m，…，n

上のように，$\{a_{l,m}\}$ の第 l 項は m，第 m 項は l，$\{a_{i,j}\}$ の第 i 項は j，第 j 項は i であり，$\{a_{i,j}\}$，$\{a_{l,m}\}$ の i，j，l，m 以外の第 k 項は k で等しい。

よって，

$$\begin{aligned}A&=2\sum_{k=1}^{n}k(a_{l,m}-a_{i,j})\\&=2\{i(i-j)+l(m-l)+j(j-i)+m(l-m)\}\\&=2\{(i-j)^2-(l-m)^2\}\\&=2\{(i-j)+(l-m)\}\{(i-j)-(l-m)\}\\&=2\{(j-i)+(m-l)\}\{(j-i)-(m-l)\}\end{aligned}$$

である。

$i<j$，$l<m$ であるから，$j-i>0$，$m-l>0$ である。

よって，$(j-i)+(m-l)>0$ であるから，$j-i>m-l$ であれば，$A>0$ である。

つまり，自然数列 1，2，…，i，…，j，…，n において 1，2，…，j，…，i，…，n とした互換数列 $\{a_{i,j}\}$ と自然数列 1，2，…，l，…，m，…，n において 1，2，…，m，…，l，…，n とした互換数列 $\{a_{l,m}\}$ において，$j-i>m-l$ であれば，

$\sum_{k=1}^{n}(a_{i,j}-k)^2>\sum_{k=1}^{n}(a_{l,m}-k)^2$ である。

これより，$j=n$，$i=1$，$1<l$，$m<n$ とすると，$j-i>m-l$ であるから，

$\sum_{k=1}^{n}(a_{i,j}-k)^2>\sum_{k=1}^{n}(a_{l,m}-k)^2$ であり，また，

$\{a_{1,n}\}$：n，2，3，……，$n-2$，$n-1$，1

である。

次に，2 から $n-1$ までの数列 2，3，……，$n-1$ について，数列 $\{a'_{i,j}\}$ は，i，j を $2\leqq i<j\leqq n-1$ である自然数として，数列 $\{k\}_{k=1,2,\cdots,n}$ において，第 i 項と第 j 項のみを入れ替えた数列，つまり，2，…，i，…，j，…，$n-1$ において 2，…，j，…，i，…，$n-1$ とした互換数列 $\{a'_{l,m}\}$ は l，m を $2\leqq l<m\leqq n-1$ である自然数として，数列 $\{k\}_{k=1,2,\cdots,n}$ において，第 l 項と第 m 項のみを入れ替えた数列，つまり，数列 2，…，l，…，m，…，$n-1$ において 2，…，m，…，l，…，$n-1$ とした互換数列である。

⑶　並び替え数列と自然数列の差の平方の和が最大になるとき

次に，$A'=2\sum_{k=2}^{n-1}k(a_{l,m}-a_{i,j})$ を考えると，同様の議論から $j=n-1$，$i=2$，$2<l$，$m<n-1$ とすると $j-i>m-l$ であるから，

$\sum_{k=2}^{n-1}(a'_{2,n-1}-k)^2>\sum_{k=2}^{n-1}(a_{l,m}-k)^2$ である。

また，$\{a'_{2,n-1}\}$：$n-1$，3，……，$n-2$，2 であり，これより数列 $\{k\}_{k=1,2,\cdots,n}$ の並び替え数列

$\{d_k\}$：n，$n-1$，……，2，1

(……には 3 以上 $n-2$ 以下の自然数が並ぶ)，

$\{c_k\}$：n，……，1

(……には 2 以上 $n-1$ 以下の自然数が並ぶ)，

$\{b_k\}$：初項が n でない または 第 n 項が 1 ではない数列に対して

$\frac{1}{n}\sum_{k=1}^{n}(d_k-k)^2>\frac{1}{n}\sum_{k=1}^{n}(c_k-k)^2>\frac{1}{n}\sum_{k=1}^{n}(b_k-k)^2$

である。

これを繰り返せば，$\frac{1}{n}\sum_{k=1}^{n}(a_k-k)^2$ が最大になる並び替え数列は n，$n-1$，……，2，1 つまり，$\{n-k+1\}_{k=1,2,\cdots,n}$ であることがわかる。

⑷　並び替え数列と自然数列の差の平方の和の最大値

次に，$\frac{1}{n}\sum_{k=1}^{n}(a_k-k)^2$ の最大値 M を求める。

$\frac{1}{n}\sum_{k=1}^{n}(a_k-k)^2$ が最大になる並び替え数列は n，$n-1$，……，2，1　つまり $a_k=n-k+1$ であることから

$M=\frac{1}{n}\sum_{k=1}^{n}\{(n-k+1)-k\}^2$ である。

$$
\begin{aligned}
M&=\frac{1}{n}\sum_{k=1}^{n}\{(n-k+1)-k\}^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\{(n+1)-2k\}^2\\
&=\frac{1}{n}\sum_{k=1}^{n}\{(n+1)^2-4(n+1)k+4k^2\}\\
&=\frac{1}{n}\left\{n(n+1)^2-4(n+1)\cdot\frac{1}{2}n(n+1)+4\cdot\frac{1}{6}n(n+1)(2n+1)\right\}\\
&=(n+1)^2-2(n+1)^2+\frac{2}{3}(n+1)(2n+1)\\
&=\frac{2}{3}(n+1)(2n+1)-(n+1)^2\\
&=\frac{1}{3}(n+1)\{2(2n+1)-3(n+1)\}\\
&=\frac{1}{3}(n+1)(n-1)\\
&=\frac{1}{3}(n^2-1)
\end{aligned}
$$

$\sigma_x{}^2=\frac{n^2-1}{12}$ であるから，$M=4\sigma_x{}^2$ あるいは $\sigma_x=\frac{\sqrt{M}}{2}$ である。

⑸　共分散について

対応する 2 つの変量 x，y について，その値の組が $(1,\ n)$，$(2,\ n-1)$，$(3,\ n-2)$，…，$(n,\ 1)$ であるとき，

$\overline{x}=\overline{y}=\frac{n+1}{2}$ であるから，x，y の共分散 σ_{xy} は

$$
\begin{aligned}
\sigma_{xy}&=\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})\\
&=\frac{1}{n}\sum_{k=1}^{n}\left(k-\frac{n+1}{2}\right)\left\{(n-k+1)-\frac{n+1}{2}\right\}\\
&=\frac{1}{n}\sum_{k=1}^{n}\frac{2k-(n+1)}{2}\cdot\frac{(n+1)-2k}{2}\\
&=-\frac{1}{4n}\sum_{k=1}^{n}\{(n+1)-2k\}^2\\
&=-\frac{M}{4}
\end{aligned}
$$

$\sigma_x=\sigma_y=\frac{\sqrt{M}}{2}$ であるから　$\sigma_x\sigma_y=\frac{M}{4}$

よって，変量 x，y の相関係数 r は，

$r=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\frac{-\frac{M}{4}}{\frac{M}{4}}=-1$ である。

つまり，$\frac{1}{n}\sum_{k=1}^{n}(a_k-k)^2$ が最大になる並び替え数列 $\{a_k\}_{k=1,2,\cdots,n}$ は数列 $\{k\}_{k=1,2,\cdots,n}$ を逆に並び替えた数列であり，このとき，2 つの変量

x：1，2，3，…，k，…，n と

y：n，$n-1$，$n-2$，…，$n-k+1$，…，1

の相関係数は -1 である。

§3. まとめ

n 個の値 1, 2, …, n をとる 2 つの変量 x, y について，x は x：1, 2, 3, …, n つまり数列 $\{k\}_{k=1,2,\cdots,n}$ とみなし，y は n 個の値 1, 2, …, n を並び替えた数列（並び替え数列）$\{a_k\}_{k=1,2,\cdots,n}$ とみなすと，差の 2 乗の平均である $\frac{1}{n}\sum_{k=1}^{n}(a_k-k)^2$ は $a_k=n-k+1$ のとき，つまり数列 $\{k\}_{k=1,2,\cdots,n}$ を逆に並べたときに最大となり，その値は $\frac{n^2-1}{3}$ になる。また，このとき変量 x，y の相関係数は -1 になるが，点 $(k,\ n-k+1)$ は直線 $y=-x+n+1$ 上にあるので当然といえばそれまでである。

データの分析を，数列を使って行うことは履修順序からみて無理であるが，数列を扱ったときにここでの考察に言及すれば，2 つの学習内容の理解において相乗効果が期待できる。

（山口県立岩国高等学校）