「資料の整理をベクトルで理解してみる」

札幌創成高校　数学科教諭

外山尚生

§０　はじめに

先日、インターネットの記事でこんな事実を見つけた。「相関係数はコサインである」。

一体何を言っているのだと思いながらも、調べてみた。資料の整理をベクトルを使って考えてみると確かに「相関係数はコサイン」であり、資料の整理で出てくるいろいろなことが簡単に理解できた。

§１　資料の整理をベクトルで理解する

１．１　偏差をベクトルで定義する。

n 個のデータ $x_1, x_2, \cdots\cdots, x_n$ と $y_1, y_2, \cdots\cdots, y_n$ を座標 $X(x_1, x_2, \cdots\cdots, x_n)$ 、 $Y(y_1, y_2, \cdots\cdots, y_n)$ で表す。

$x_1, x_2, \cdots\cdots, x_n$ の平均を $\overline{x} = \dfrac{1}{n}\sum_{k=1}^{n} x_k$ 、 $y_1, y_2, \cdots\cdots, y_n$ の平均を $\overline{y} = \dfrac{1}{n}\sum_{k=1}^{n} y_k$ で表す。

すべての要素が $\overline{x}$ 、 $\overline{y}$ である点を O_x 、 O_y とする。つまり $O_x(\overline{x}, \overline{x}, \cdots\cdots, \overline{x})$ 、 $O_y(\overline{y}, \overline{y}, \cdots\cdots, \overline{y})$ と定義すると、 $\overrightarrow{\mathrm{O}_x\mathrm{X}} = (x_1 - \overline{x}, x_2 - \overline{x}, \cdots\cdots, x_n - \overline{x})$ は偏差を表すベクトルとなる。これを $\overrightarrow{\mathrm{O}_x\mathrm{X}} = \vec{x}$ 、 $\overrightarrow{\mathrm{O}_y\mathrm{Y}} = \vec{y}$ として議論をしていく。

１．２　分散と標準偏差をベクトルで考える。

n 個のデータ $x_1, x_2, \cdots\cdots, x_n$ の分散は $V(X) = \dfrac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2$ で表される。

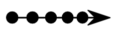
ベクトルの大きさ $|\vec{x}|^2 = \sum_{k=1}^{n}(x_k - \overline{x})^2$ であるから、 $V(X) = \dfrac{|\vec{x}|^2}{n}$ で表すことができる。

これに平方根をつけたものが標準偏差であるから $s_x = \dfrac{|\vec{x}|}{\sqrt{n}}$ と表すことができる。

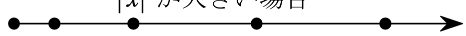
ここから分散は $|\vec{x}|^2$ 、標準偏差は $|\vec{x}|$ と関連があることがわかる。

実際、分散や標準偏差をベクトルの大きさとみなすと、分散や標準偏差が大きいほうがデータのばらつきが大きいことがわかる。

$|\vec{x}|$ が小さい場合。



$|\vec{x}|$ が大きい場合



１．３　共分散と相関係数をベクトルで考える。

n 個のデータ $x_1, x_2, \cdots\cdots, x_n$ と $y_1, y_2, \cdots\cdots, y_n$ の共分散は $s_{xy} = \dfrac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$ で表される。

ここでベクトルの内積は $\vec{x}\cdot\vec{y} = \sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$ と表されるから、共分散は $s_{xy} = \dfrac{\vec{x}\cdot\vec{y}}{n}$ で表すことができる。

さらに相関係数は $r = \dfrac{s_{xy}}{s_x \times s_y} = \dfrac{\dfrac{\vec{x}\cdot\vec{y}}{n}}{\dfrac{|\vec{x}|}{\sqrt{n}} \times \dfrac{|\vec{y}|}{\sqrt{n}}} = \dfrac{\vec{x}\cdot\vec{y}}{|\vec{x}| \times |\vec{y}|}$ であり、 $\vec{x}$ と $\vec{y}$ のなす角を θ とすると

内積の定義 $\vec{x}\cdot\vec{y} = |\vec{x}||\vec{y}|\cos\theta$ から $r = \dfrac{|\vec{x}||\vec{y}|\cos\theta}{|\vec{x}||\vec{y}|} = \cos\theta$ であることがわかる。

１．４　まとめ

１．１から１．３をまとめると資料の整理とベクトルには表のような関連があることが分かった。

資料の整理	ベクトル
偏差	$\vec{x}$
分散	$\lvert\vec{x}\rvert^2$
標準偏差	$\lvert\vec{x}\rvert$
共分散	$\vec{x}\cdot\vec{y}$
相関係数	$\cos\theta$

資料の整理の用語を上の表のベクトルとみなして資料の整理の用語が持つ意味を考えていきたい。

§２　相関係数をベクトルで理解する。

§１の結果から相関係数 r は $\cos\theta$ であることがわかった。これをもとに相関係数 r のもつ意味を考えてみよう。

２．１　$-1 \leqq r \leqq 1$ である

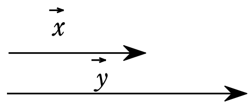
これは $-1 \leqq \cos\theta \leqq 1$ であることからあたりまえである。

２．２　r が１に近いと正の相関であり、r が -1 に近いと負の相関である。

θはなす角であるから $\vec{x}$ と $\vec{y}$ は次のようになる。

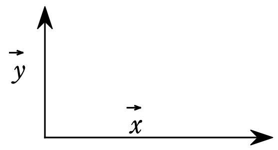
［１］　$r=1$ のとき $\cos\theta=1$ より $\theta=0°$ だから $\vec{x}$ と $\vec{y}$ は同じ方向を向く。

つまり x が増加すると y も増加する。



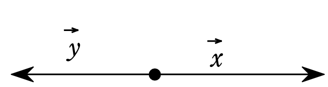
［２］　$r=0$ のとき $\cos\theta=0$ より $\theta=90°$ だから $\vec{x}$ と $\vec{y}$ は別の方向を向く。

つまり x と y に相関はない。



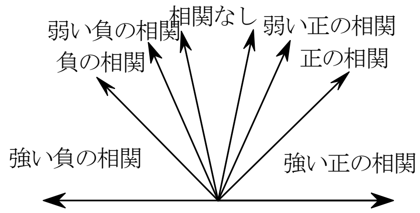
［３］　$r=-1$ のとき $\cos\theta=-1$ より $\theta=180°$ だから $\vec{x}$ と $\vec{y}$ は逆向きになる。

つまり x が増加すると y は減少する。



２．３　相関係数の強さと角度の関係

相関係数の強さを相関係数の値から角度に照らし合わせてみた。



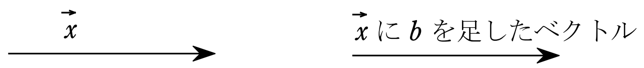
相関なし	弱い正の相関	正の相関	強い正の相関
$0 \leqq r \leqq 0.2$	$0.2 \leqq r \leqq 0.4$	$0.4 \leqq r \leqq 0.7$	$0.7 \leqq r \leqq 1$
$78° \leqq \theta \leqq 90°$	$66° \leqq \theta \leqq 78°$	$45° \leqq \theta \leqq 66°$	$0° \leqq \theta \leqq 45°$

強い負の相関	負の相関	弱い負の相関	相関なし
$-1 \leqq r \leqq -0.7$	$-0.7 \leqq r \leqq -0.4$	$-0.4 \leqq r \leqq -0.2$	$-0.2 \leqq r \leqq 0$
$135° \leqq \theta \leqq 180°$	$114° \leqq \theta \leqq 135°$	$102° \leqq \theta \leqq 114°$	$90° \leqq \theta \leqq 102°$

§３　$ax+b$ の変換における値の変化

３．１　すべての値に b を足したときの分散、標準偏差、共分散、相関係数の変化。

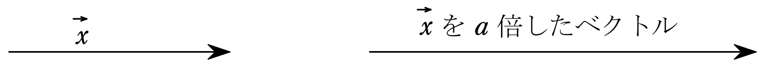
すべての値に b を足すということは $\vec{x}$ を平行移動しただけである。



したがって、分散、標準偏差、共分散、相関係数は変わらない。

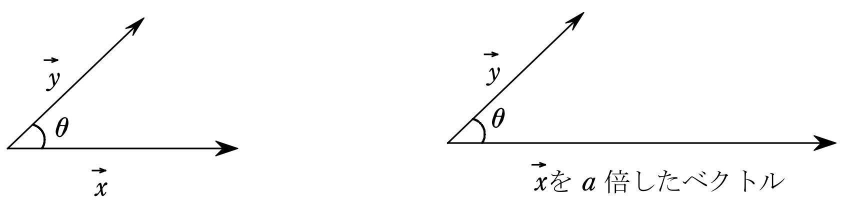
３．２　すべての値を a 倍（ $a>0$ ）したときの分散、標準偏差、共分散、相関係数の変化。

すべての値を a 倍するとは $\vec{x}$ に対して $a\vec{x}$ を考えることである。



ベクトルの大きさが a 倍されるのだから、$|\vec{x}|^2$ である分散は a^2 倍され、$|\vec{x}|$ である標準偏差は a 倍になる。

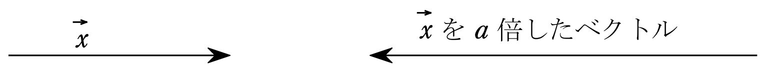
次に $\vec{x}$ を a 倍して、$\vec{y}$ は変化しない場合を考えてみよう。



ベクトルの大きさが a 倍されるのだから、$\vec{x}\cdot\vec{y}$ である共分散は a 倍されるが、 $\cos\theta$ である相関係数は変化しない。

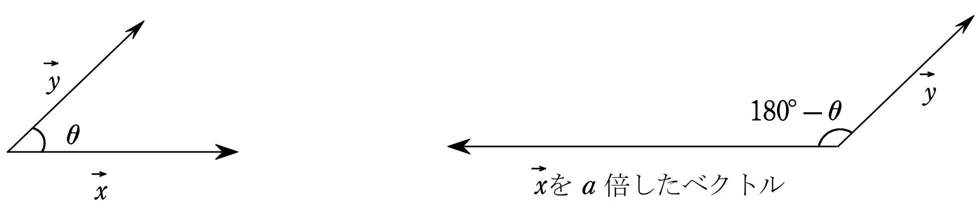
３．３　すべての値を a 倍（ $a<0$ ）したときの分散、標準偏差、共分散、相関係数の変化。

すべての値を a 倍（ $a<0$ ）するとは $\vec{x}$ に対して $a\vec{x}$ を考えることである。



ベクトルの大きさが $|a|$ 倍されるのだから、$|\vec{x}|^2$ である分散は a^2 倍され、$|\vec{x}|$ である標準偏差は $|a|$ 倍になる。

次に $\vec{x}$ を a 倍して、$\vec{y}$ は変化しない場合を考えてみよう。



ベクトルの大きさが a 倍されるのだから、$\vec{x}\cdot\vec{y}$ である共分散は $-a$ 倍されるが、 $\cos\theta$ である相関係数は $-\cos\theta$ となる。つまり $-\cos\theta=\cos(180°-\theta)$ より相関の強さは変わらないが、正負が変わる。

§４　まとめ

資料の整理をベクトルを使って考えるなど、思いつきもしないアイディアであったが、実際調べてみると資料の整理とベクトルにはいろいろな関連性があることが分かり、数学って面白いなあと改めて思うことができた。

資料の整理をベクトルで考えることで、今まであやふやであった知識がより感覚的に目に見える形で表すことができたと思う。それにしても図形に始まり、いろいろなところに応用できるベクトルってすごい！

□参考資料

・相関係数のイメージを解説したい
https://qiita.com/epppJones/items/44ddc2d96f35900174e2

・【暗記しない数学】相関係数を内積と同じように考えるとかなりわかりやすい件
https://www.yukisako.xyz/entry/correlation－coefficient