相関係数の視覚化

平　田　嘉　宏（北海道立教育研究所　研究・相談部）

要約

社会的に統計に関するリテラシーの重要性が高まってきている。学校教育においても、統計に関する学習を充実させていく方向にある。本稿では、相関係数に着目し、数学的な観点に立ち、高校生が理解できる範囲の知識や技能を用いて相関係数が視覚化できることを示す。まず、正方形の等積変形による平行四辺形とベクトルのなす角を用いた視覚化について述べる。次に、相関係数の定義そのものと偏差を成分とするn次元ベクトルを用いた視覚化について述べる。最後に、統計的な観点に立ち、例示したデータの検定結果を示すことで、数学的な観点と統計的な観点との違いの留意点を明らかにする。

キーワード：相関係数の視覚化、ベクトルのなす角

Keywords:　visualization of correlation coefficient, angle between two vectors

１　はじめに

（１）統計教育の重要性

社会生活などの様々な場面において、必要なデータを収集して分析し、その傾向を踏まえて課題を解決したり意思決定をしたりすることが求められている(2016 中央教育審議会)。「近年ではあらゆる学問分野で統計やデータを使った分析が行われるようになり、また企業社会でも基本的なリテラシーとして必須条件となって」いる(2011 原)。日本マイクロソフト株式会社の文教部門の責任者によれば、人類が始まって以来、数千年の歴史で排出されたデータの全部のうちの９割がここ直近の２年間で排出されており、大量にあるデータを整理して分析をする能力というのは、ある程度文系でも素養がないと仕事にならない時代になってきているという(2017 小野田)。数学の本の売れ筋ベストテンのうち５冊が統計学の本だったり(2017 Amazon)、今年度の数学セミナー４月号掲載の「数学関連の本の売れ行きベストテン」は統計関連の本が７冊ランクインしたりしている。このように、社会全般においても、数学関連の分野においても、統計に関するリテラシーの重要性が確実に高まってきている状況にある。

学校教育においては、この３月告示された小中学校学習指導要領や、５月に明らかになった同移行措置案によって、現在高等学校で指導している四分位範囲及び箱ひげ図が中学校で指導され、中学校からは中央値や最頻値が小学校に移ることとなった。大学生は文系理系を問わず多くの学生が研究、卒論あるいは将来のためにと、様々な必要性があって統計学を履修している。高校数学の次期学習指導要領の改訂に向けた改善事項では「統計に関する学習を充実させていくことが重要である」(2016 中央教育審議会)とされており、高校数学での統計に関する学習は、今後ますます充実させていく方向にある。

（２）相関係数の視覚化を示す理由と視覚化の概要

こうした状況にあって、統計に関する学習の充実に資するため、本稿では、相関係数に着目した。相関係数は、相関関係の強弱に用いられてはいるもの

の、その値が具体的に何を意味するのかということまでは取り扱わない。同時に取り扱う散布図が視覚的に捉えやすいのとは対照的である。

高校数学において取り扱う２次元の対象のうち、この相関係数のように、その値が何かの意味を直接示しはしないものとしては、平面ベクトルの内積がある。これも散布図と同様に、同時に取り扱う２つのベクトルは視覚的に捉えやすく、相関係数と同様に、内積の値が具体的に何を意味するのかということまでは取り扱わない。ただ、内積は様々な活用ができることを学習する中で、その値を視覚化して説明することは可能であり、その方法も既知である。そこで、内積のような、相関係数の値の視覚化について２通り明らかにすることとする。その際、統計学的な観点ではなく、数学的な観点に立ち、高校生が理解できる範囲の知識や技能を極力用いるように配慮する。

一つは、正方形を、ある条件の下で等積変形した平行四辺形の頂点の座標を変量として用いると、内角の余弦の値が相関係数と等しくなることを利用し、相関係数が平行四辺形の内角の余弦の値として視覚化できることを示すものである。

もう一つは、一般に、相関係数が各変量ごとの偏差を成分とするn次元ベクトルのなす角の余弦の値と等しいことを利用し、相関係数はベクトルのなす角としてイメージすることが可能であり、特に、３次元では視覚化できることを示すものである。

なお、最後に、統計的な観点に立ち、例示する具体的なデータについて統計的な検定を行う。こうして、数学的な観点と統計的な観点との違いを明らかにする。

２　特別な形の平行四辺形による相関係数の視覚化

相関係数は、−1 から 1 までの値をとるため、同様に−1 から 1 までの値をとる余弦との関係をもたせられる可能性がある。その一つの方法として次のように座標平面上で視覚化することができた。

一辺の長さがaの正方形 ABCD の辺 CD を、その辺と平行な方向に平行移動（即ち等積変形）してできる平行四辺形の４つの頂点の座標をx, yの組としたときの相関係数は、平行四辺形の隣り合う２辺のなす角の余弦の値の一つと等しい。

この一般的な場合を示す前に、具体的な数値で計算すると次のようになる。

例１　座標平面上の４点 A(1,4), B(1,2), C(3,5), D(3,7) を頂点とする平行四辺形がある。これら４点を、対応する２つの変量のx, yの値の組としたときの相関係数と、$\angle ABC$の余弦の値を求める。

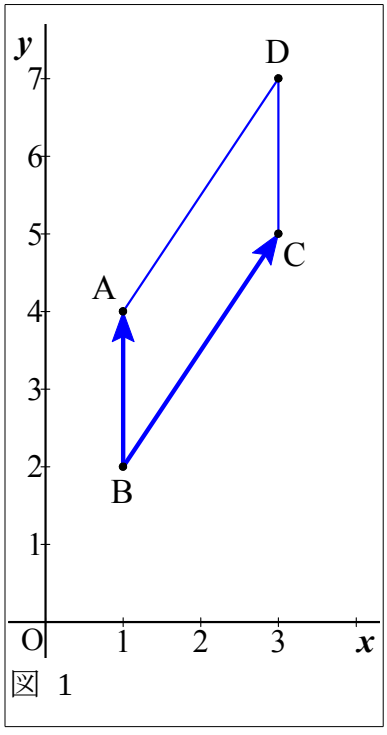

図 1

$\overline{x}=2$, $\overline{y}=\dfrac{9}{2}$であるから、共分散$s_{xy}$は

$$s_{xy}=\frac{1}{4}\left\{(1-2)\left(4-\frac{9}{2}\right)+(1-2)\left(2-\frac{9}{2}\right)\right.$$
$$\left.+(3-2)\left(5-\frac{9}{2}\right)+(3-2)\left(7-\frac{9}{2}\right)\right\}=\frac{3}{2}$$

x, yの標準偏差はそれぞれ

s_x

$$= \sqrt{\frac{1}{4}\left\{(1-2)^2+(1-2)^2+(3-2)^2+(3-2)^2\right\}}$$

$$= 1$$

s_y

$$= \sqrt{\frac{1}{4}\left\{\left(4-\frac{9}{2}\right)^2+\left(2-\frac{9}{2}\right)^2+\left(5-\frac{9}{2}\right)^2+\left(7-\frac{9}{2}\right)^2\right\}}$$

$$= \frac{\sqrt{13}}{2}$$

相関係数は

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\frac{3}{2}}{\frac{\sqrt{13}}{2}} = \frac{3\sqrt{13}}{13} \fallingdotseq 0.8321$$

となる。

次に、余弦の値を求めると

$\overrightarrow{BA} = (0,2),\ \overrightarrow{BC} = (2,3)$ より

$\overrightarrow{BA} \cdot \overrightarrow{BC} = 0 \cdot 2 + 2 \cdot 3 = 6$

$|\overrightarrow{BA}| = 2,\ |\overrightarrow{BC}| = \sqrt{13}$

$$\cos\angle ABC = \frac{\overrightarrow{BA} \cdot \overrightarrow{BC}}{|\overrightarrow{BA}||\overrightarrow{BC}|} = \frac{6}{2\sqrt{13}} = r$$

よって相関係数は$\angle ABC$の余弦の値に等しい。

次に、一般的な場合を示す。

i)　平行四辺形の1組の辺がy軸に等しい場合

4点A, B, C, Dを、それぞれ$(x_1,\ y_1+a)$, $(x_1,\ y_1)$,$(x_1+a,\ y_1+c)$,$(x_1+a,\ y_1+a+c)$(ただし、いずれの文字も正)とすると

$$\overline{x} = x_1 + \frac{a}{2}, \overline{y} = y_1 + \frac{a+c}{2}$$

共分散s_{xy}は

$$s_{xy} = \frac{1}{4}\sum_{k=1}^{4}(x_k-\overline{x})(y_k-\overline{y})$$

$$= \frac{1}{4}\left(-\frac{a}{2}\cdot\frac{a-c}{2} - \frac{a}{2}\cdot\frac{-a-c}{2} + \frac{a}{2}\cdot\frac{-a+c}{2} + \frac{a}{2}\cdot\frac{a+c}{2}\right)$$

$$= \frac{ac}{4}$$

標準偏差は

$$s_x = \sqrt{\frac{1}{4}\sum_{k=1}^{4}(x_k-\overline{x})^2} = \sqrt{\frac{1}{4}\cdot 4\cdot\left(\frac{a}{2}\right)^2} = \frac{a}{2}$$

$$s_y = \sqrt{\frac{1}{4}\sum_{k=1}^{4}(y_k-\overline{y})^2}$$

$$= \sqrt{\frac{1}{4}\left\{\left(\frac{a-c}{2}\right)^2+\left(\frac{-a-c}{2}\right)^2+\left(\frac{-a+c}{2}\right)^2+\left(\frac{a+c}{2}\right)^2\right\}}$$

$$= \frac{\sqrt{a+c}}{2}$$

よって、

$$r = \frac{c}{\sqrt{a^2+c^2}} \cdots ①$$

一方、

$\overrightarrow{BA} = (0,a),\ \overrightarrow{BC} = (a,c)$より

$\overrightarrow{BA} \cdot \overrightarrow{BC} = ac,\ |\overrightarrow{BA}| = a,\ |\overrightarrow{BC}| = \sqrt{a^2+c^2}$

$$\cos\angle ABC = \frac{\overrightarrow{BA} \cdot \overrightarrow{BC}}{|\overrightarrow{BA}||\overrightarrow{BC}|} = \frac{c}{\sqrt{a^2+c^2}} \cdots ②$$

①，②より

$r = \cos\angle ABC$

ii)　i)と合同な平行四辺形の辺がいずれも座標軸と平行でない場合

平行移動や回転移動によって第1象限で1組の辺がy軸と平行となるようにすれば同じ結論が得られる。

iii)　i), ii)より$\angle ABC$が鋭角の場合が示された。

iv)　$\angle ABC$が直角か鈍角の場合

$c \leqq 0$であり、i)～iii)と同様に示すことができる（詳細は省略）。

v)　i)～iv)より、相関係数は$\angle ABC$の余弦の値の一つと等しいことが成り立つ。

なお、本稿では、統計的な観点ではなく数学的な観点で議論を進めているため、相関関係の有意性の有無は検証していないことをことわっておく。実は、例1において相関関係の有意性に関する検定を行う

と、ここで求めた相関係数は使えないことになる（詳しくは４で述べる）。

３　相関係数の視覚化の一般化

（１）相関係数の定義とベクトルの余弦の値

対応する２つの変量のx, yの値の組が任意に与えられたとき、それを用いた２つのベクトルのなす角の余弦が相関係数と等しくなる関係を導くために、対応する２つの変量のx, yの値の組を$(x_1,\ y_1)$, $(x_2,\ y_2)$, $(x_3,\ y_3)$, $\cdots$, $(x_n,\ y_n)$とすると、相関係数rは

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

$$= \frac{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})^2}\sqrt{\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(y_k-\overline{y})^2}}$$

$$r = \frac{\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\displaystyle\sum_{k=1}^{n}(x_k-\overline{x})^2}\sqrt{\displaystyle\sum_{k=1}^{n}(y_k-\overline{y})^2}}\cdots③$$

となる。

対応する２つの変量のx, yの値の組$(x_1,\ y_1)$, $(x_2,\ y_2)$, $(x_3,\ y_3)$, $\cdots$, $(x_n,\ y_n)$を用いた２つのベクトル$\vec{p}$, $\vec{q}$を、各変量ごとの偏差を成分として

$\vec{p} = (x_1-\overline{x},\ x_2-\overline{x},\ x_3-\overline{x},\ \cdots,\ x_n-\overline{x})$,

$\vec{q} = (y_1-\overline{y},\ y_2-\overline{y},\ y_3-\overline{y},\ \cdots,\ y_n-\overline{y})\cdots$(※)

と定めると、

$$\vec{p}\cdot\vec{q} = \sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})$$

$$|\vec{p}| = \sqrt{\sum_{k=1}^{n}(x_k-\overline{x})^2},\ |\vec{q}| = \sqrt{\sum_{k=1}^{n}(y_k-\overline{y})^2}$$

よって、$\vec{p}$, $\vec{q}$のなす角をθとし、余弦の値を求めると

$$\cos\theta = \frac{\vec{p}\cdot\vec{q}}{|\vec{p}||\vec{q}|} = r\ (\because ③)$$

したがって、(※)で定義する２つのベクトルを用いると、ベクトルのなす角の余弦は相関係数に等しくなる。

（２）相関係数の定義による視覚化の具体例

例１を用いて具体的に計算してみる。

対応する２つの変量のx, yの値の組を用いた２つのベクトル$\vec{p}$, $\vec{q}$を、各変量ごとの偏差を成分として定義すると、

$$\vec{p} = (-1,\ -1,\ 1,\ 1),\ \vec{q} = \left(-\frac{1}{2},\ -\frac{5}{2},\ \frac{1}{2},\ \frac{5}{2}\right)$$

この２つのベクトルのなす角をθとし、余弦の値を求めると

$$\cos\theta = \frac{\vec{p}\cdot\vec{q}}{|\vec{p}||\vec{q}|} = \frac{6}{2\sqrt{13}} = \frac{3\sqrt{13}}{13} = r \fallingdotseq 0.8321$$

このベクトルは４次元空間なので、生徒にとっては視覚化は難しいが、少しでもイメージだけでも持てるようにするためには例えば、次のような説明をすることが考えられる。

「$\cos\theta = \dfrac{3\sqrt{13}}{13}$となる$\theta$は約$34°$である。２次元空間と３次元空間で、長さが$2$と$\sqrt{13}$、なす角が$34°$のベクトルの形状ならそれぞれ視覚化することはできる。どちらも、２つのベクトルはある一つの平面上にあるから、同じ形である。したがって、４次元における２つのベクトルも、ある一つの平面上にあって、形も先の２つのベクトルと同じだとイメージしてもよいだろう。」

このように、例１の場合、明確な視覚化はできないが、イメージはもてる。とはいえ、明確に視覚化するには３次元空間での例が望ましいため、次の例２でそれを示す。

例２　対応する２つの変量のx, yの値の組 (1,2), (2,4), (3,5) を用いて相関係数などを計算する。

$\overline{x}=2,\ \overline{y}=\dfrac{11}{3}$であるから、$\vec{p},\ \vec{q}$を、各変量ごとの偏差を成分として

$\vec{p}=(-1,\ 0,\ 1),\ \vec{q}=\left(-\dfrac{5}{3},\ \dfrac{1}{3},\ \dfrac{4}{3}\right)$と定義する。

この例はこれらのベクトルを視覚化することが目的ではあるが、この２つのベクトルのなす角をθとし、相関係数即ち余弦の値を求めておくと

$$r=\cos\theta=\frac{\vec{p}\cdot\vec{q}}{|\vec{p}||\vec{q}|}=\frac{3}{\sqrt{2}\sqrt{\dfrac{14}{3}}}=\frac{3\sqrt{21}}{14}\fallingdotseq 0.9820$$

２つのベクトルを視覚化すると、図２のようになる（ただし、見やすくするため、ともに負だったx成分はその絶対値を用いた。それでも検証に実質的には影響はない。）。

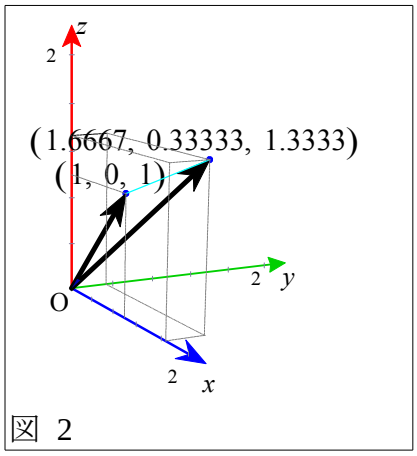

図 2

この例２によって、「相関係数は、各変量ごとの偏差を成分とするベクトルのなす角の余弦の値に等しい」ことの視覚化ができた。

なお、(1,2), (2,4), (3,5) とこれらのベクトルとが視覚的にはどのような関連があるかについては、次のとおりである。

「(1,2), (2,4), (3,5) の各点の$\left(2,\ \dfrac{11}{3}\right)$からの$x$の偏差を３つの成分とするベクトルと、$y$の偏差を３つの成分とするベクトルは、$x$成分、$y$成分、$z$成分それぞれを２つのベクトルで比較したとき、３次元空間でのベクトルのなす角が小さいほど、２つのベクトルの方向は近くなる。この近い度合いを３次元空間におけるベクトルのなす角の余弦の値で表したものが相関係数である。」

また、次のような表現もできる。

「図３において (1,2), (2,4), (3,5) とのxの偏差を３回、(1,2), (2,4), (3,5) とのyの偏差を３回、順にみていく。３回のxの偏差　$-1, 0, 1$　の数の並びと、３回のyの偏差

$-\dfrac{5}{3},\dfrac{1}{3},\dfrac{4}{3}$　の数の並びの共通点は、ベクトルとしてとらえると、ほぼ同じ方向を向いている点である。

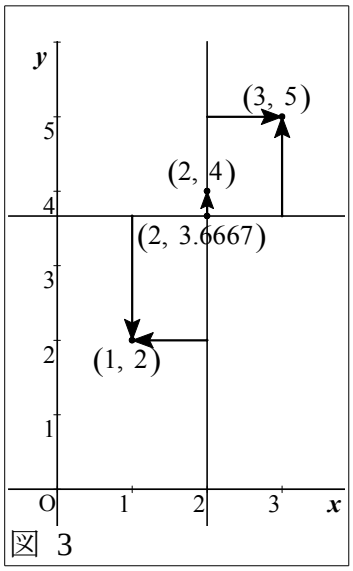

図 3

相関係数は、ベクトルの向きが一致したときが１、全く逆の向きのときが－１となり、その間は、できたベクトルのなす角の余弦の値である。」

４　相関関係の有意性に関する検定について

本稿では、統計的な観点ではなく数学的な観点で議論を進めてきたため、ここまで相関関係の有意性の有無は検証してこなかった。しかし、実社会において相関関係を用いる場合には、統計学的見地から相関関係の有意性に関する検定を行う必要がある。

例１と例２において無相関検定を、Benjamini & Hochberg 法による p 値の調整を含めてを行うと、

例１では　p=0.168 > 0.1　(df= 1 & 2, F=4.50)

例２では　p=0.121 > 0.1　(df= 1 & 1, F=27.00)

通常 p<0.05 が有意性があると認められる範囲であるから、それに則れば、例１、例２ともに相関関係があるとすることはできず、rの値は使えない。

つまり、「たまたま３～４つのデータを集めたらこういうこともしばしばあるから、rが 0.8 より大きいから『正の相関が強い』と結論づけてしまうと、

事実に反してしまう可能性が結構あるから危ない」ということである。

では、もし相関関係が有意な例を用いて数学的な議論を進めようとすると、今度は計算量が膨大になり、生徒が計算するには適当でなくなる恐れがある。

よって、統計的な観点と数学的な観点の両方を成立させるのではなく、状況に応じてどちらかの観点を優先するのが適当である。

なお、検定の具体的な過程は省略したが、検定の計算は、ブラウザ上で動作する無償のソフトウェアjs-STAR(version 8.0.1j)及びオープンソース・ソフトウェアR(version 3.2.5)により行った。

５　まとめ

生徒が相関係数の値とは具体的に何を指しているかと本質的な質問をしてきた際、本稿の相関係数の視覚化は、説明に使える可能性がある。ただ、生徒の興味・関心が相関係数の論理的な意味合いである場合などには使えない。この相関係数の値の意味の明確化は、回帰直線を用いて説明することができるが、その説明は、本稿とは別に行う。

引用文献等

小野田哲也（2017）．これからの社会と求められる能力.北海道高等学校長協会会報，179:5-24

中央教育審議会（2016年12月）．幼稚園、小学校、中学校、高等学校及び特別支援学校の学習指導要領等の改善及び必要な方策等について（答申）〈http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/1380731.htm〉（2017-5-21アクセス）

原俊彦（2011）．統計の世界－物の見方・考え方・心構え.原書房

数学の本の売れ筋ベストテン〈https://www.amazon.co.jp/gp/bestsellers/books/492168/ref=zg_bs_nav_b_2_466290〉（2017-5-21アクセス）

参考文献

岩井勇児・鈴木眞雄（1985）．教師のための統計法入門〔第２版〕.福村出版

中野博幸・田中敏（2011）．フリーソフトjs-STARでかんたん統計データ分析.技術評論社

舟尾暢男（2008）.「R」Commanderハンドブック.オーム社