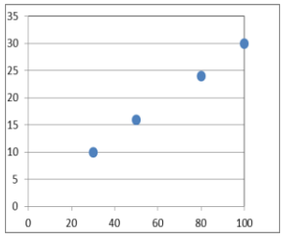
盛岡三高数学科通信 How do you solve ? How do you teach ?

第 19 号
発行責任者
盛岡第三高等学校
下町壽男

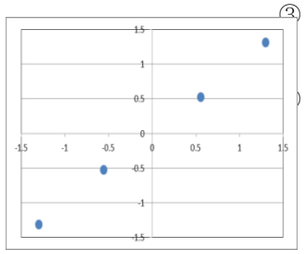
データの相関　②

	テスト1	テスト2
学者うさぎ	100	30
マドンナうさぎ	80	24
元気ネコ	50	16
うかれぎつね	30	10
平均	65	20

テスト5	テスト6
1.30	1.31
0.56	0.53
−0.56	−0.53
−1.30	−1.31



テスト１vs テスト２



テスト５vs テスト６

前回は，テスト１とテスト２の相関を考える代わりに，テスト５とテスト６の相関を考えても良さそうだという話までいきました．

テスト５とテスト６は，テスト１，テスト２をそれぞれ「平均０，標準偏差１」に平行移動し，拡大縮小したものであることに注意しください．このようなデータの変換を「標準化」または「正規化」する，と呼ぶことがあります．

ここで，少し脱線します．偏差値について考えてみたいと思います．

例えば，模試で数学の得点が０点だったのに，偏差値が 30 もついてくることがあります．この，偏差値の計算のメカニズムはどうなっているのか，おさらいしておきましょう．

ある集団でテストを行ったとき，得点を X とし，それに対応する偏差値を Z とします．全体の平均点を $\overline{X}$，標準偏差を σ とおけば

$Z=\dfrac{X-\overline{X}}{\sigma}\times 10+50$　と表されます．

これが偏差値です．

この式の意味を考えてみましょう．

①　$X-\overline{X}$　‥‥平均点を 0 点にする

②　$\dfrac{X-\overline{X}}{\sigma}$ ‥‥標準偏差を 1 にする

$\dfrac{X-\overline{X}}{\sigma}\times 10$　‥‥標準偏差を 10 にする

$\dfrac{X-\overline{X}}{\sigma}\times 10+50$　‥平均を 50 にする

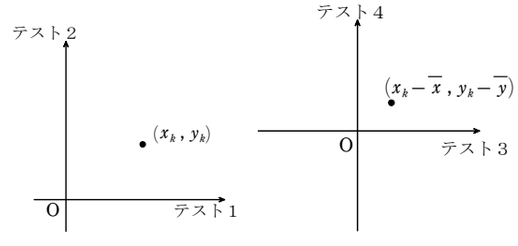
つまり，偏差値とは，得点の分布を，平均 50，標準偏差 10 になるように平行移動したり，拡大縮小した値にすぎないということです．

このような処理を日本中で行えば，他と比較して自分の位置がどのくらいのところにあるかわかる一つの指標になるというわけですね．

話をもどします．相関について，次のようにまとめておきましょう．

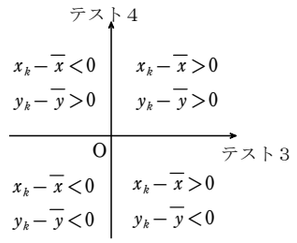
> 変量 x, y の相関を考えることは，各変量を平均０，標準偏差１になるように標準化したデータの相関を考えることと同じである．

では，これまで述べてきたテスト１とテスト２の相関の具合を数量化する式を考えてみましょう．



テスト１とテスト２に関するデータは，左下の図の点 (x_k, y_k) で表されます．一方，それぞれの平均点を引いた得点に変換した，テスト３とテスト４のデータは，平均点分それぞれ平行移動しているので，左下図の，$(x_k-\overline{x}, y_k-\overline{y})$ という点に対応しています．

このとき，次のことに注意します．



上図で，第 1 象限は２つの変量がともに平均を超えている領域，第 2 象限はテスト３の平均は越えず，テスト４の平均は超えている領域，第 3 象限は，２つの変量がともに平均を下回っている領域，第 4 象限は，テスト３の平均は越え，テスト４の平均が下回っている領域ということになります．

このように，散布図は，２つの変量の平均のところで線を引いて，４分割して考えると分析しやすくなります．

ここで，あるデータ (x_k, y_k) に対し，平均との差の積 $(x_k-\overline{x})(y_k-\overline{y})$ を考えると，

データが第 1・第 3 象限にあるとき，積は正になり，第 2・第 4 象限にあるとき，積は負になります．

第 1・第 3 象限にデータがあるということは，正の相関が強いときであり，逆に，データが第 2・第 4 象限にあるときは負の相関が強いときですね．

そこで，すべてのデータに対して，平均との差の積を加えて，その値が大きいほど，データが第 1・第 3 象限にたくさんある，つまり正の相関が強いということがいえます．この考えを利用すると，相関の強さを数量で表すことができるわけです．平均との差の積の平均

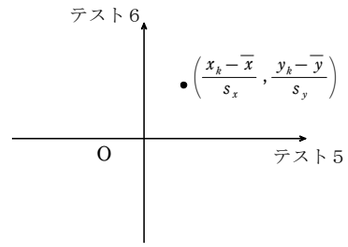
$\dfrac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$

を**共分散**といって，s_{xy} という文字で表します．

まとめましょう．

> $s_{xy}>0$ ⇒①③にデータがある⇒正の相関が強い
>
> $s_{xy}<0$ ⇒②④にデータがある⇒負の相関が強い

さて，ではこの共分散を，テスト５とテスト６で考えてみましょう．



$\dfrac{1}{n}\cdot\dfrac{1}{s_x s_y}\cdot\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_1-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$

これは，$\dfrac{s_{xy}}{s_x s_y}$　というシンプルな式になりますね．

この値を**「相関係数」**といい，r で表します．

このように表すことによって，相関の強さを−1 から 1 までの数で標準化できるので，イメージがしやすくなります．

相関係数をベクトルから眺めてみましょう．

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

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

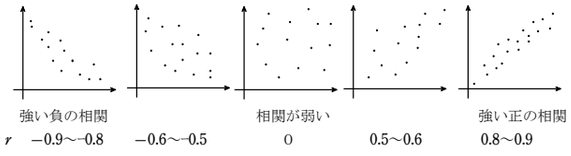
とすると，　$r=\dfrac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|}$　ということです．

つまり，２つのベクトルのなす角を θ としたとき，相関係数とは２つのベクトルの $\cos\theta$ を表すものと考えることができます．

コサインは，２つベクトルの「近さ加減」を表す量なので，これを相関係数と考えても納得がいきますね．

基本的に，相関係数の算出はコンピュータ等にやらせればよくて，問題は，相関係数の見方です．

散布図と相関係数の関係を図に表しておきましょう．



これを基に，データの相関の強弱を判断できればよいと思います．大切なのは，相関や相関係数とは，データ分析の目を持つために重要な事柄であるということで，決してセンター試験のためにあるのではありません．