相関係数の値の意味の明確化

－回帰直線を用いて－

平　田　嘉　宏（北海道立教育研究所　研究・相談部）

要約

統計に関する学習の充実が重要とされている。本稿では、相関係数に着目し、回帰直線を用いて、相関係数の値の意味を説明する。併せて、回帰直線にまつわる回帰効果、２本の回帰直線とその留意点、回帰の意味と相関係数の値の意味について説明する。

キーワード：相関係数、回帰直線、回帰効果

Keywords:　correlation coefficient, regression line, regression effect

１　はじめに

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。高等学校の数学においても統計に関する学習を充実させていくことが重要であるとされている(2016 中央教育審議会)。そこで、本稿では、統計に関する学習内容の相関係数に着目した。相関係数は、必履修科目の数学Ⅰで学習し、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は取り扱っていない。

統計学的には、因果関係が認められる場合、相関係数rよりもrを二乗した決定係数に意味合いを持たせている。例えば$r^2 = 0.47$だと、結果を表す変量の変動の47%がその他の変量の変動に伴うと言ったり、結果を表す変量をその他の変量によって47%説明(予測)できると言ったりする(2015 神田)。理数数学Ⅰで発展、拡充させる内容として示されている回帰直線(2009 文部科学省)はこの因果関係が認められる場合に用いられるものである。以下、この回帰直線を用いて相関係数rの値の意味を説明する。

２　分布の標準化による相関係数の値の意味の明確化

簡便のため、相関係数は１より小さい正の値とする。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、２つの変量の$x,\ y$の平均値$\overline{x},\overline{y}$, 標準偏差$s_x, s_y$, 相関係数$r$を用いると、回帰直線の方程式は

$$y - \overline{y} = r\frac{s_y}{s_x}(x - \overline{x}) \ \cdots①$$

となる。

これを$z = \dfrac{x - \overline{x}}{s_x},\ w = \dfrac{y - \overline{y}}{s_y}$として標準化すると、$\overline{z} = \overline{w} = 0,\ s_z = s_w = 1$より、①は$w = rz$となる。

あらためて対応する２つの変量の$x,\ y$を$\overline{x} = \overline{y} = 0,\ s_x = s_y = 1$ とすると、回帰直線の方程式は

$y = rx$…②

となる。

回帰直線に基づくと、あるx_iに対するyの推定値はrx_i、即ちx_iよりr倍だけ0に近い値となる。

ここで注意すべきは、標準化した２つの変量は平均も標準偏差も等しいのにもかかわらず、推定値はx_iとは等しくならないことである。

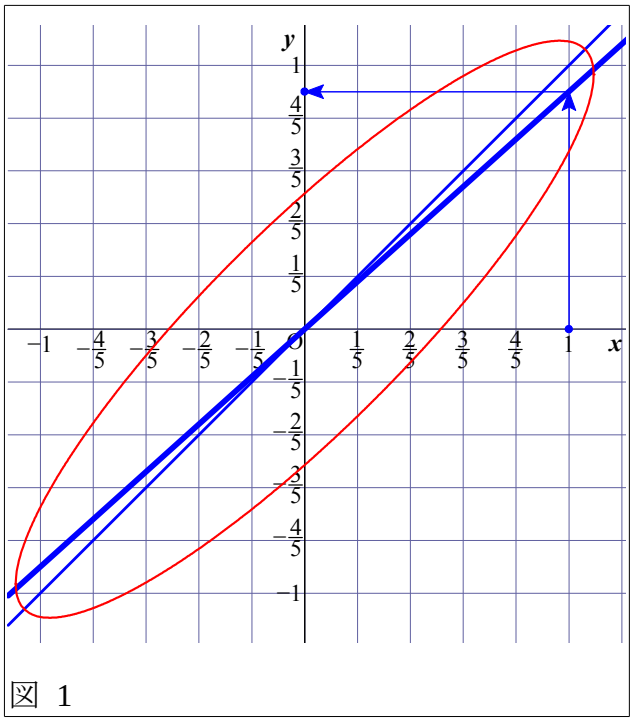

図 1

例えば、$r = 0.9$の場合、回帰直線は図１の太線のようになる（楕円は、対応する２つの変量のx, yの値の組$(x_i,\ y_i)$全体ががほぼ収まっている範囲）。このとき、矢印で例示したように、推定値の平均0との距離は、x_iの平均0との距離の0.9倍となる。

ここに相関係数の値の意味がある。

つまり、**推定値は平均値にr倍近づく。**

ところで、散布図で、変量の散らばり具合の真ん中を貫く直線は、$y = rx$ではなく$y = x$である。推定値を求めるのになぜ散らばり具合の真ん中にある$y = x$を用いないかは、散らばり具合をどの方向から見るかによって次のように説明できる。

確かに平均の値の組の原点からみれば、$y = x$は散らばり具合の真ん中を貫いている。一方、回帰直線は、最小二乗法の考え方に基づいている。つまり、一般に、対応する２つの変量のx, yの値の組$(x_i,\ y_i)$のそれぞれのx_iに対して、ある直線$y = ax + b$で求められる$ax_i + b$とy_iとの差（座標平面上では同じx座標である２点の「y座標の差」）を平方した値の総和が最小のときの「ある直線」が回帰直線 $y - \overline{y} = r\dfrac{s_y}{s_x}(x - \overline{x})$ である。

回帰直線上の点からのいわゆる「y座標の差」を図示すると、図２のようになる。図２は、図１における$x_i = 1$の場合を拡大して４点を定めたもので、点Ｂは$y = x$上の点、点Ｃは回帰直線$y = rx$上の点とする。

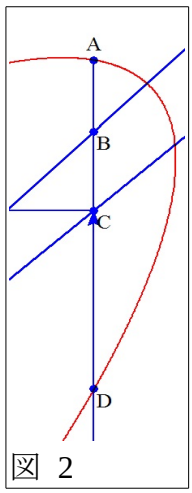

図 2

図２において、「線分ＡＣあたりに分布する$(x_i,\ y_i)$の数と線分ＣＤあたりに分布する$(x_i,\ y_i)$の数との差」よりも、「線分ＡＢあたりに分布する$(x_i,\ y_i)$の数と線分ＢＤあたりに分布する$(x_i,\ y_i)$の数との差」の方が大きい。これは、回帰直線がそうした差を他の直線の場合よりも最も小さくなるようにしたものだからである。

よって、x軸上の点からy軸に平行な方向にみれば、散らばり具合の真ん中を貫いている直線は、$y = x$ではなく$y = rx$の方である。これは任意のx_iに対して成り立つ。

したがって、推定値を求めるのに用いる直線は回帰直線$y = rx$であり、$y = x$ではない。

３　一般的な分布の場合

推定値は平均値にr倍近づくことが、標準化された対応する２つの変量について成立したが、これを一般化すると、次のようになる。

（１）$s_x = s_y$の場合

回帰直線の方程式は①より$y - \overline{y} = r(x - \overline{x})$となる。このとき、ある$x_i$に対する推定値は、$\overline{y}$より $x_i - \overline{x}$ のr倍だけずれた値になる。$x_i - \overline{x}$だけずれるわけではない。よって、**推定値は平均値にr倍近づく。**

なお、散布図で、変量の散らばり具合の真ん中を貫く直線の方程式は$y - \overline{y} = x - \overline{x}$ である。

（２）$s_x \neq s_y$の場合（一般化）

回帰直線の方程式は①であり、この直線は、

（１）の一つ目の直線をx軸からy軸方向に$\dfrac{s_y}{s_x}$倍拡大または縮小させたものである。このとき、あるx_iに対する推定値は、$\overline{y}$より $x_i - \overline{x}$ の$\dfrac{s_y}{s_x}$倍だけずれるわけではなくそのさらにr倍、つまり$x_i - \overline{x}$の$r\dfrac{s_y}{s_x}$倍だけずれた値になる。**よって、推定値は平均値にr倍近づく。**

なお、散布図で、変量の散らばり具合の真ん中を貫く直線の方程式は $y - \overline{y} = \dfrac{s_y}{s_x}(x - \overline{x})$ である。

以上で一般化ができた。

４　回帰直線と回帰効果

以上の結果を用いると、例えば「同一集団で、ある科目の実力テストを２回行ったとき、偏差値が 50 を超える層では２回目の方が偏差値が下がる者の方が多く、偏差値が 50 を下回る層では２回目の方が偏差値が上がる者の方が多いことは統計的には自然にあり得ることである。よって、その事実のみをもって、上位層への指導の仕方に何か問題があったとか、下位層への指導の効果があったなどと評価するのは誤りである。」ということがいえる。

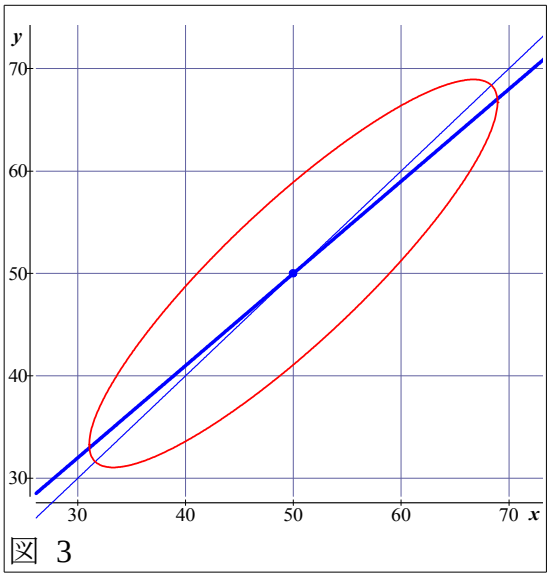

図 3

図３は図１を拡大し、x軸, y軸の正の方向にともに50だけ平行移動させたものである。$x = 65$や$x = 35$で２の図２のような検討をすれば、上記の例えのようなことがいえることは明らかである。これは**回帰効果**といわれている。回帰効果は気がつきにくく、回帰効果を無視した誤りは専門家にすら見受けられるといわれている(1985 村上)。

５　２本の回帰直線とその留意点

対応する２つの変量のx, yの値の組の因果関係が逆ならば、あるy_iに対するxの推定値を求めることになる。例えば図１の矢印の先の$y_i = 0.9$に対する推定値を求めることになるが、このとき、矢印を逆にたどって推定値を１とするのは誤りであり、注意が必要である。x, yはともに標準化された２つの変量であるから、推定値は正しくは$0.9^2 = 0.81$となる。

①とは逆にあるy_iに対するxの推定値を求めるための回帰直線の方程式は、①におけるxとyの文字を入れ替えればよい。

$$x - \overline{x} = r\frac{s_x}{s_y}(y - \overline{y})$$

図１では$x = 0.9y$である。

因果関係は通常xが要因となってyとなる関係なので、この回帰直線はあまり意味がない。しかし、予測する方向が違うと回帰直線が($r = \pm 1$でなければ)異なることには留意しなければならない。

Geogebra では、２本の回帰直線をワンクリックで切り替えて示すことができて非常に便利である。図４の右上の楕円部分をクリックするだけで、図５に切り替わり、グラフとグラフの下の直線の方程式により、回帰直線が２通りあることが確認できる。

なお、変量の散らばり具合の真ん中を貫く直線と併せて３本を描くと図６のようになる。$s_x = s_y$であれば、２本の回帰直線は、真ん中を貫く直線に関して線対称となる。

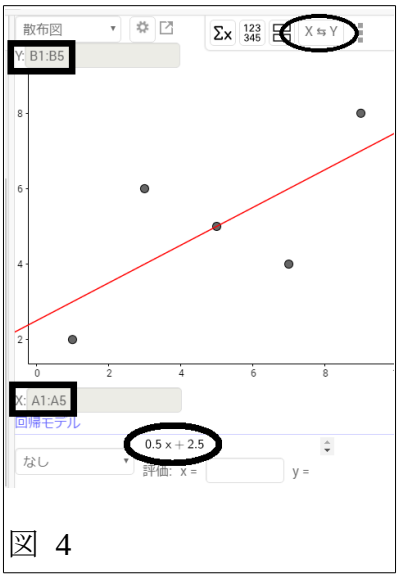

図 4

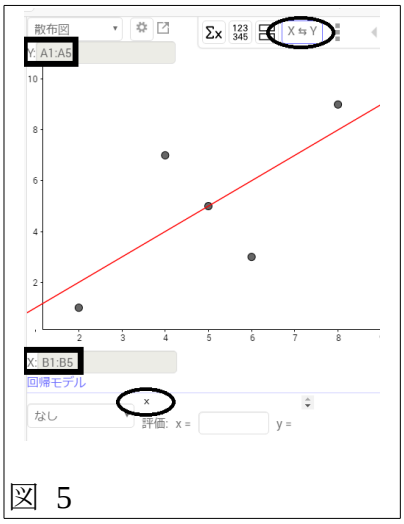

図 5

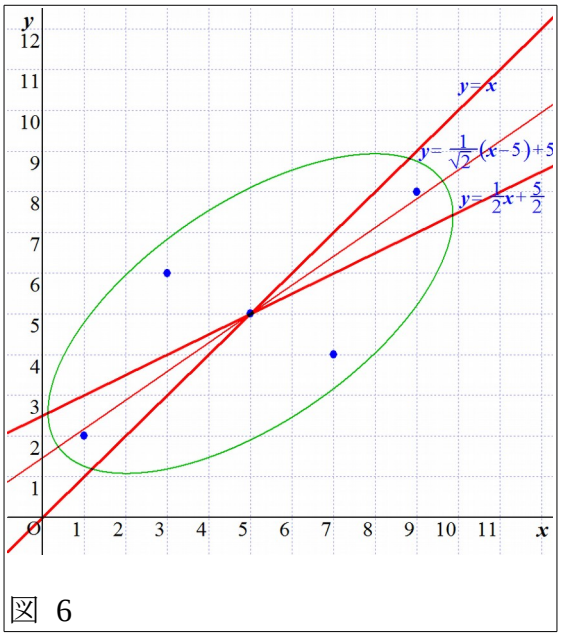

図 6

６　「回帰」の意味と相関係数の値の意味

回帰直線による推定値は平均に一定の割合で近づくが、この「回帰」という言葉は、19 世紀末にイギリスの Galton が最初に用いた。Galton は、数百人の親の身長とその子供の成人になった時の身長を、性差によらない補正をしながら全体平均と比較したところ、親の身長が全体平均の身長より高くても子供の身長は成人した時には全体平均の身長よりは高い傾向にあるものの、それほど高くならず、同様に、親が全体平均より低くても子供は成人したときには全体平均よりは低い傾向にあるものの、それほど低くならないことを発見した。Galton は、子供の成人した時の身長の全体平均との差は、親の身長の全体平均との差の３分の２倍と推定し、これを「回帰」の程度とした(2011 大澤)。

この身長の分析は、ちょうど３（１）にあてはまり、$s_x = s_y$に加えて$\overline{x} = \overline{y}, r = \frac{2}{3}$とした「推定値が平均値に３分の２近づく」場合である。

また、子供の身長が親ほど高く（低く）ならないことは、４の回帰効果の一つの例でもある。

このように、回帰効果はどこでも生じるものと認識できる。よって、日常生活の中でこの考え方が役立つことがある。例えば、「より努力してよい成果をあげた取組に満足して努力を怠ると、成果が平凡なものになる危険性が高いから、常に工夫改善する姿勢をもつことが大切だ」といえる。

こうしてみると、本稿の相関係数rの値の意味は、生徒が相関係数の論理的な意味合いを質問してきたときに説明に使うのは難しいが、捉えようによっては常日頃の工夫改善の姿勢への示唆にもつながるなど、大変奥が深い。

引用・参考文献

Galton, F. (1886) . *Regression towards mediocrity in hereditary stature.* The Journal of the Anthropological Institute of Great Britain and Ireland. 15: 246-263. doi :10.2307/2841583 <http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf>（2017-5-27 アクセス）

大澤清二 (2011). 生活の統計学.建帛社

神田善伸 (2015). ＥＺＲでやさしく学ぶ統計学－ＥＢＭの実践から臨床研究まで.中外医学社

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）

村上隆 (1985) .海保博之(編著). 心理・教育データの解析法 10 講－基礎編.55-56

文部科学省 (2009). 高等学校学習指導要領解説数学編理数編