第 3 章 3 節　相関分析

-二つの対象にどのくらいの関連性があるのかを指標化する方法-

山田　彬尭

1.　はじめに

この節はちょっとした頭の体操から話を始めたいと思います。今、架空の二つのクラス（A 組と B 組）の現代文（x 軸）と英語（y 軸）の試験結果が図 1 のように得られたとしましょう。この散布図から、みなさんは、どのような議論や分析をしてみたいか、少し考えてみてください。

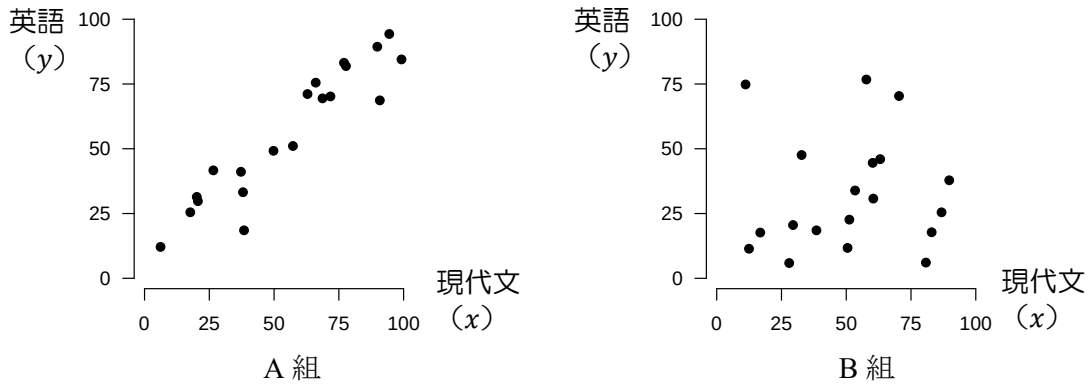

図 1　架空の二つのクラスの現代文と英語の試験結果を表した散布図

いかがでしょうか。もちろん、これらの散布図を分析する唯一の絶対的な方法というものはなく、様々な見方が可能です。しかし、ここでは、この後の節で扱う内容とのつながりも意識して、三つほど代表的な着眼点をお伝えしましょう。まず、図 2（左図）のように､データの真ん中を通るような線を引いて(紙幅の都合上 A 組のみ示します)、A 組と B 組でどちらの傾きが大きいのかを、議論するということが可能です。このようなアプローチは**回帰分析**と呼ばれ、本書では、第○○節で扱います。

次に、図 2（右図）のように、データに、●（成績上位者の集団)、▲（成績があまりよくない人たちの集団)、◇（成績中位者たちの集団）があるのではないかと、データを分割して、いくつかの集団を議論することも可能です。このような分析は**クラスター分析**と呼ばれ、第○○節で扱います。

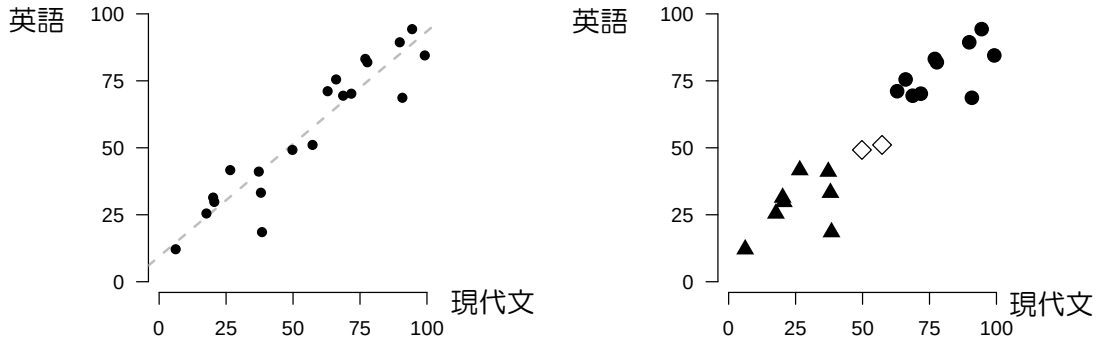

図 2　回帰分析（左図）とクラスター分析（右図）

最後に、現代文と英語の点数の間にどのくらい強い関連性があるのかを議論したいと思った人もいるかもしれません。これが、本節で紹介する相関分析です。例えば、B 組より A 組の方が、「現代文の点が高ければ英語の点が高くなる」という関係が顕著ですよね。このように、一方が上がると他方も上がる（または、下がる）という関係がはっきりしているとき、私たちは関連性、つまり、相関が強いと表現します。

しかし、図１は極端に違うケースをお見せしていたので、目視で A 組の方が関連性が強そうだと判断できたわけですが、印象で関連性の強さを議論してしまうと、意見が割れてしまうこともあることでしょう。そこで、何か二つの変数の関連度を測る客観的な指標が欲しいですよね。

そこで、英語と現代文は関連度が 0.53、英語と数学は 0.57、というように、二つのものの間にどのくらいの関連性があるのかを数字で議論することができたら、とても便利ですよね。本節で扱う相関分析とは、まさに、このように二つの対象の関連性の度合いを数値化して分析する手法です。

ただし、関連性（類似度）の指標には、対象の性質などに応じて様々なものが提案されていて、残念ながら紙幅の都合上、そのすべてはご紹介しきれません。そこで、本節では、最も代表的な**共分散**と呼ばれる指標とピアソンの**積率相関係数**というものに絞って説明をすることにします。なお、ピアソンの指標以外に相関係数と呼ばれる指標もありはするのですが、単に相関係数と呼ばれるときには、このピアソンの指標が意図されているのが普通です（このため、本節でも以下これを単に相関係数と呼ぶことにします）。

2. 相関分析とは

2.1 共分散

先ほど図１では、A 組の方が現代文と英語の関連度が高かったわけですが、これをもっと推し進めて、関連性度が最も強くなるケースを考えてみましょう。それは、現代文の点がこの値なら、英語の点はこの値になるというのが完全に予想できる場合です。図３のような状況がそのような状況を表しています。

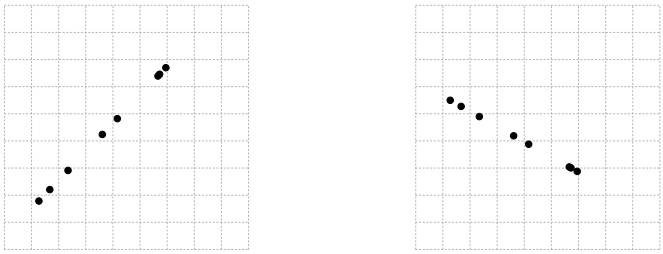

図３　関連度合いが最大化した場合：正の相関（右図）、負の相関（左図）

データに右肩上がりの関連性があるとき、データに**正の相関**、逆に右肩下がりの関係があるとき、**負の相関**があるという言い方をします。図３は正負という違いこそあれ、二変数の関連性が最も強くなっている状況だ、というわけです。

なお、二変数の関連性は、原点がどこにあるのかには依存しません。すべての球が同じように平行移動しても球のばらつきや関係は変わらないからです。そこで、図３を含め、今後は、原点を中心に引かれた x 軸、y 軸は省略します。しかし、軸がないというのも困るので、代わりに、x の値と y の値の平均をそれぞれ $\bar{x}$（エックス・バー）、$\bar{y}$（ワイ・バー）と表し、それらを新しい軸として採用することにします。図４（左図）は、図３（左図）に $\bar{x}$ と $\bar{y}$ という新しい二つの軸（点線）を加えて再表示させたものです。

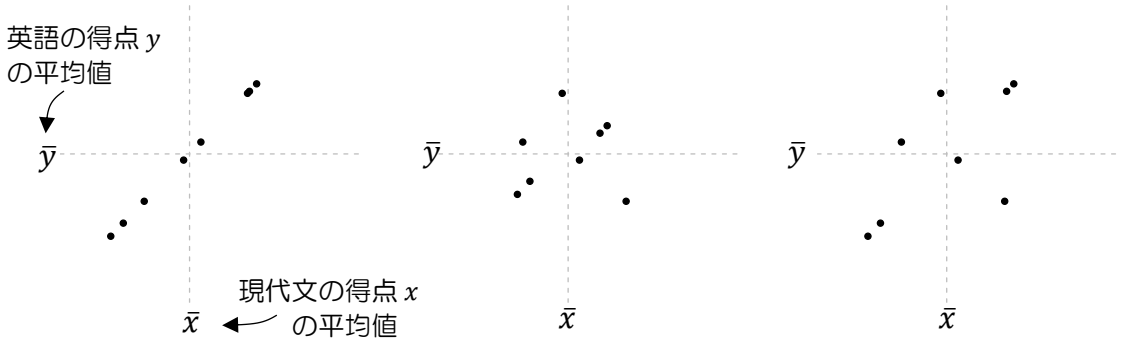

図４　関連性が強いもの（左図）、とても弱いもの（中央図）、弱いもの（右図）

さて、この図４（左図）と図４（中央図）を比べてみてください。後者は、見るからに二変数の関連性が弱そうです。先ほど述べた通り、この二つの差を、何らかの数字に基づいて議論したいのですが、みなさんならどういう指標を提案するでしょうか。この問題を考えるために、先ほど引いた点線に注目してみましょう。この点線によって、いま、散布図は四つの領域に区切られています。説明の都合上、右上から反時計回りに、それぞれの領域を第一象限、第二象限、第三象限、第四象限と呼ぶことにします。

図４（中央図）では、データがこの四つの領域すべてにまんべんなく分布しています。これに対し、図４（左図）では、第一象限と第三象限のみにデータが分布しています。どうも、二変数の関連性が強い状況というのは、対角線上の象限（つまり、第一象限＆第三象限か、第二象限＆第四象限）にデータが集中しているときであり、逆に、すべての象限にデータが分布しているときは、関連性が弱くなるのだ、と言えそうですね。

（第一案）球の個数の差を考える

そこで、第一案として、次の（式１）で計算される指標を考えてみましょう。

（式１）　　「第一／三象限の球の数」－「第二／四象限の球の数」

もし、第一項と第二項で測った数が同数であれば、相殺されて、この指標はゼロになります。前者が後者より大きい状況であれば正の値が、後者が前者より大きい状況であれば負の値が得られます。これによって、図４（中央図）は相関が弱いことが、そして、図４（左図）では、正の相関がある、という点を捉えられます。

ただし、この第一案では、図４（中央図）と図４（右図）の状況を区別できません。各象限に含まれている球の数は同じだからです。ですが、これは問題です。図４（中央図）と比べると図４（右図）の方が、より右肩上がりに見えますので、その分、強い関連性があるという判断を下していてほしいところです。

（第二案）中心からの離れ具合を「足し算」で捉える

そこで第二案です。実は、図４（中央図）と図４（右図）では、第二象限と第四象限の球は全く同じ位置に存在しています。一方で、図４（右図）では、図４（中央図）の第一象限と第三象限の球を、より中心から遠いところへ動かしました。再配置によって、つまり、中心からの離れ具合を変えることで、二変数の関連性が変わってしまう、ということを捉えるために、今度は、「球の個数」ではなく「中心からの球の遠さ」に注目してみましょう。図５を見てください。図４（中央図）と図４（右図）のみを取り出して表示させたものです。

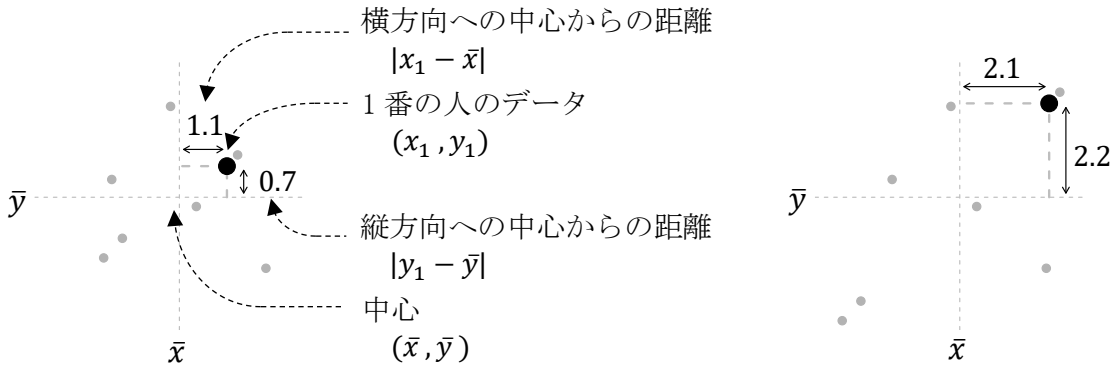

図５　データの中心から球が遠いケース（左図）と近いケース（右図）

いきなりすべての球について議論するのではなく、まず、出席番号 1 番の人に注目しましょう。図 5 で濃く大きく表示されている球が、この出席番号 1 番の生徒の成績です。この球は、データの中心から x 軸方向に 1.1、y 軸方向に 0.7 離れた位置にあります。そこで、この二つの長さを足した値を、中心からの離れ具合としましょう（ちなみに、このようにして計測された距離をマンハッタン距離と言います）。すると、図５（左図）では、$1.1 + 0.7 = 1.8$、図５（右図）では、$2.1 + 2.2 = 4.3$となります。4.3のほうが1.8より値が大きいので、球を動かした後の方が、データの中心から遠いところにいるということが、数字の上で捉えられるようになりました。

同じ要領で、すべての球についてマンハッタン距離を計算し、表示させたものが下の図６です（なお、説明のしやすさから、第一／第三象限を灰色にしておきました）。

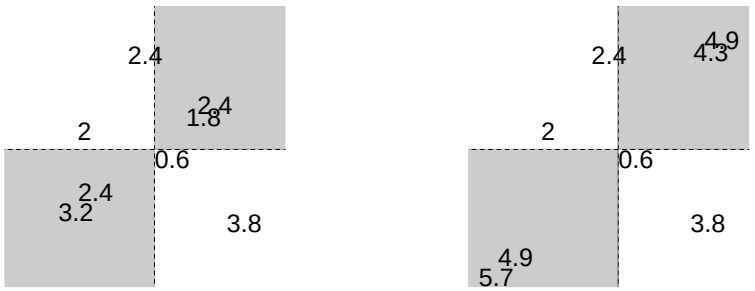

図６　$|x_i - \bar{x}| + |y_i - \bar{y}|$という和で計算した中心からの距離

そこで、（式１）のカギ括弧の部分を修正して、次の指標を考えることにしましょう。

（式２）「図６の『灰色』のエリアの数字の和」（①）－「図６の『白色』のエリアの数字の和」（②）

すると、図６（左図）で①−②を計算すると、$9.8 - 8.8 = 1.0$に、図６（右図）では$19.8 - 8.8 = 11.1$になります。11.1 の方が 1.0 よりも値が大きいので、データ全体では、図６（右図）の方が図６（左図）よりも関係性が強い、という結論が得られました。

今話したことを数式にしてみます。仮に灰色の領域にいるのは出席番号 1, 2, 7, 8 番の生徒で、白色のエリアには出席番号 3,4,5,6 番の生徒がいるものとします。すると我々は①の部分を計算するために、絶対値を使って、（式３）のような計算をします（シグマ記号の下部の$i \in \{1,2,7,8\}$は、i が 1,2,7,8 であるものについて計算しなさいという意味です）。一方で、②の部分を計算するためには、（式４）の量を計算することになります。

（式３）
$$\sum_{i\in\{1,2,7,8\}}(|x_i-\bar{x}|+|y_i-\bar{y}|) = \begin{array}{l}|x_1-\bar{x}|+|y_1-\bar{y}| \\ +|x_2-\bar{x}|+|y_2-\bar{y}| \\ +|x_7-\bar{x}|+|y_7-\bar{y}| \\ +|x_8-\bar{x}|+|y_8-\bar{y}|\end{array}$$

（式４）
$$\sum_{i\in\{3,4,5,6\}}(|x_i-\bar{x}|+|y_i-\bar{y}|) = \begin{array}{l}|x_3-\bar{x}|+|y_3-\bar{y}| \\ +|x_4-\bar{x}|+|y_4-\bar{y}| \\ +|x_5-\bar{x}|+|y_5-\bar{y}| \\ +|x_6-\bar{x}|+|y_6-\bar{y}|\end{array}$$

すると、最終的に計算する量は、次のような（式５）で表されることになります。

（式５）
$$\underbrace{\sum_{i\in\{1,2,7,8\}}(|x_i-\bar{x}|+|y_i-\bar{y}|)}_{\text{①第一／第三象限のデータ}} - \underbrace{\sum_{i\in\{3,4,5,6\}}(|x_i-\bar{x}|+|y_i-\bar{y}|)}_{\text{②第二／第四象限のデータ}}$$

しかし、実演しておいてなんですが、この数式は計算が面倒ですよね。なぜならば、①と②というようにデータを場合分けしなければならないからです。

（第三案）中心からの離れ具合を「掛け算」で捉える

そこで、第三案です。実は、あるトリックを使うと、今まで述べてきた基本的な考え方を維持したまま、場合分けをせずに、関連性を表す指標が作れるのです。そのあることとは、$1.1+0.7$というふうに「足す」のではなく、1.1×0.7というふうに「かける」ということなのですが、なぜこれが場合分けを無くすことにつながるのかを理解してもらうために、図７を用意しましたのでご覧ください。これは、すべての生徒について、足し算ではなく、掛け算、つまり$(x_i-\bar{x})(y_i-\bar{y})$で、中心からの遠さを測ったものです。

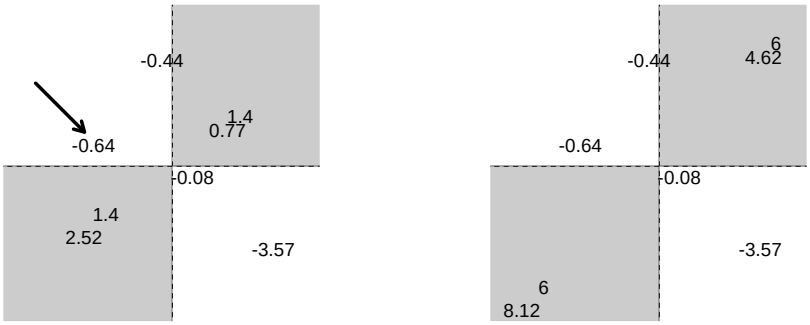

図７　$(x_i-\bar{x})(y_i-\bar{y})$という積で計測した中心からの距離

第一に、灰色のエリアを見てください。右図の方が左図に比べて大きな値が出ています。そもそも第二案を作ったのは図４（中央図）と図４（右図）を区別できるようにするためでした。今回の指標でも、右の方が関連性が強いことは表せていますので、先ほどの第二案で実現したかった目標はこの第三案でも実現されている、というわけです。

第二に、白色のエリアを見てください。例えば、矢印で記されたデータは、$(-1.6, 0.4)$という座標で、この二つの値を掛け合わせると-0.64という負の値が得られます。図６では白色エリアも灰色エリアもすべてが正の値でしたので、それらを①の側と②の側に

分類しなければならなかったのですが、ちょうど②、つまり「引く数」に当たる数字の前に自動的にマイナスの記号がついてくれているので、今回は、灰色と白色のエリアの数字をただ足し合わせるだけで、勝手に白いエリアの数字が引き算されることになります。もはや二つのシグマ記号を使う必要はなくなり、場合分けをしなくていい、その意味で扱いやすい指標が誕生しました。これを数式で表すと次のようになります。

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

（第四案）サンプルサイズの大きさの影響をなくす

最後に、この第三案に、あとひと手間加えたいと思います。図８（右図）は、図８（左図）のそれぞれのデータの周りに似たような値のデータを 3 個ずつ追加したものです。したがって、この二つの図では二変数の関連性の強さは大して変わっていないわけです。しかし、第三案の指標を使って計算すると、左図では 2.3、右図では 8.2 という全く違う値になってしまいます。単純にデータが多くなることで足される$(x_i-\bar{x})(y_i-\bar{y})$の数が増えるためです。

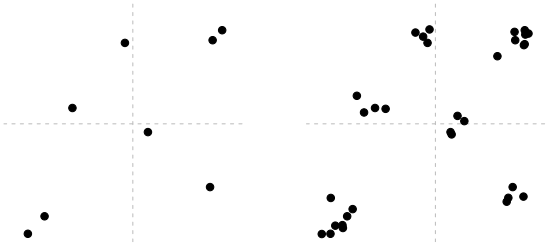

図８　データを増加させた散布図

ですが、分布傾向は同じなのに、球が増えただけで関連度が上がっては困ります。そこで、サンプルサイズの影響を除去するために、n個足し合わせた分、nで割って、平均を出します（すると、左が 0.26、右が 0.29 と似た値になります）。これが、x と y の**共分散**と呼ばれる指標で、統計学で用いられる最も代表的な関連性を測る指標です。記号ではS_{xy}あるいは$Cov(x,y)$などと表記します（なお、不偏性というより望ましい性質を出すために$n-1$で割る定義もありますが、本節ではその定義には深入りしません）。

$$S_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

余談ですが、異なる二変数の関連性ではなく、自分自身（例：x と x）との共分散（つまり、S_{xx}）を求めたものが第〇〇節で見た**分散**（S_x^2）です。当然、自分自身と自分自身をプロットするので、散布図の上ではデータは一直線に並びます。自分自身との関連度など英語でも現代文でも必ず最大化しているに決まっていますから、分散が表す数字の大きさはもはや関連度とは無関係になります。そうではなく、分散は、単に、データの中心から延びる対角線上において、どのくらい遠くまで球が分布しているのか、を教える指標になっています。そこで、分散は、ばらつきの指標として使われます。また、分散の平方根が**標準偏差** S_xです（第〇〇節参照）。

2.2 相関係数

共分散は、シグマ記号が一つで済むという数式上の簡便性だけでなく、数学的に便利な性質も知られていて、統計学では、大変重宝されます。しかし、同時に「単位に依存してしまう」という注意が必要な性質も持ちあわせて。図９を見てください。

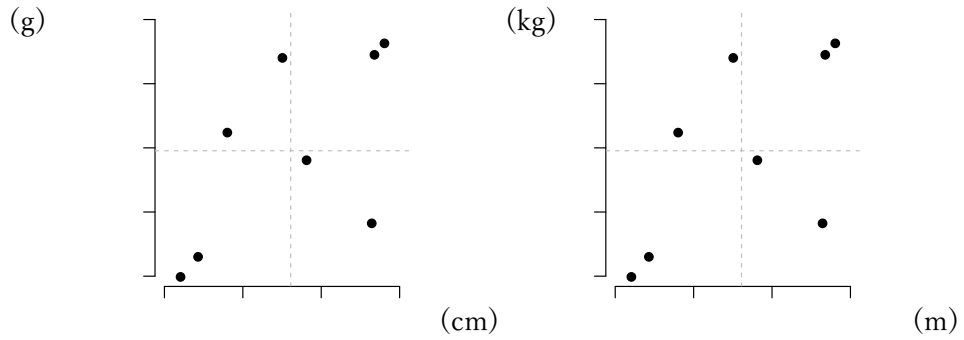

図９　異なる単位で表現された散布図

この二つの散布図の差は単位にあります。左は cm と g で計測をしたもの、右は m と kg で計測をしたものです。単位こそ違えど、両者は本質的には同じものです。ですので、関連度としては両方とも同じ値になってほしいところですが、共分散を計算すると前者は 240,504.5 になるのに対し、後者は 2.405045 となり、10 万倍も値が違ってしまっています。このように、共分散は、単位に影響されて値が変わってしまうので、そこで、共分散を修正して単位に依存しない指標を作ってみることにしましょう。

証明は少しややこしいので割愛しますが、実は、共分散は図３のように球が一直線上になり、関連度が最大化するときに、x 軸（つまり現代文）の標準偏差 S_x と y 軸（つまり英語）の標準偏差 S_y の積に等しくなるという性質があり、次式が成り立ちます。

$$S_{xy} = S_x S_y \quad \text{（右肩上がりの時）} \qquad \text{あるいは} \quad S_{xy} = -S_x S_y \quad \text{（右肩下がりの時）}$$

それ以外、つまり、一直線にならないケースでは、共分散は、$S_x S_y$ の値よりも（絶対値の上で）値が小さくなります。このため、一般的に次のような関係が成り立ちます。

$$-S_x S_y \leq S_{xy} \leq S_x S_y$$

そこで「理論上の最大値である $S_x S_y$」に占める「今回採取したデータの共分散 S_{xy}」の割合を計算し、ありえたかもしれない最大の共分散の値に対して、今回の共分散がどのくらい「いい線いっているのか」を測る指標を作ります。これが**相関係数**で、記号では r_{xy} で表されます。

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

つまり、今回のデータの関連性がどのくらい一直線という理想的な状態に近いのかを議論しているのが相関係数だとも言えます。割合なので、もともとが cm だったのか m だったのかということには関係しません。単位に依存しない指標が出来上がりました。

また、上記の不等式を変形すると次のようになりますので、相関係数は、必ず -1 から 1 までの値を取るということも分かります。

$$-1 \leq r_{xy} \leq 1$$

これもうれしい性質です。1に近づけば近づくほど正の相関（右肩上がりの分布傾向）を、逆に、−1に近づけば近づくほど負の相関（右肩下がりの分布傾向）を表し、0を取るとときに二変数が無相関であることを示すので、解釈もしやすいわけです。

というわけで、共分散と相関係数の紹介が終わりました。単位への依存度という点で、違いはあるものの、どちらも、関連性の指標として望ましい以下の性質を持っているということを押さえてください。

（性質１） 指標の「正負」は、「右肩上がり（正の相関）」なのか「右肩下がり（負の相関）」なのかを表す
（性質２） 指標の値の「（絶対値の）大きさ」は、関連性の強さを表す
（性質３） 指標が０になるときは、二変数が無相関であることを表す
（性質４） 計算するときに場合分けをしなくて済む
（性質５） サンプルサイズに影響を受けない

3. 実践事例

ここからは、これらの性質を踏まえつつ、それが実践でどのように使われるのか、そして、仕様上どういう点に気を付けておくべきかというお話しをしていきましょう。

なお、SPSS と R では次のような操作を行いますが、いずれもその実装の詳細は演習用コンパニオンサイトに譲りたいと思いますので、興味がある方は HP をご覧ください。

表○. SPSS を実施してみよう

≪SPSS を実施してみよう！≫
Step 1. 「分析→分類→階層クラスタ」をクリックします。 Step 2.
ここは西田先生に埋めてもらえましたら幸いです。
*詳細は「人文学系学生のためのはじめての量的・質的研究」の演習用コンパニオンサイトを確認してください。

表○. R を実施してみよう

≪R を実施してみよう！≫
共分散の計算には cov 関数を、相関係数の係数には cor 関数を使用します。
*詳細は「人文学系学生のためのはじめての量的・質的研究」の演習用コンパニオンサイトを確認してください。

3.1 実践例

ある架空の学校で五教科の期末試験が行われました。そこで、任意の二つの教科の点数の間に関連性がないか、相関係数を出してみました。ですが、五教科の中から二教科を選ぶには 10 通りの選び方があり、それを「英語と数学は 0.1」「現代文と数学は 0.3」と地の文で報告してしまうと、記号や数字が羅列され、読むのは少々大変です。

そこで、論文では計算された共分散や相関係数を、次のような表形式でまとめて提示したりします。このような表は、**分散共分散行列**、**相関係数行列**と呼ばれます。

	現代文	数学	理科	社会	英語
現代文	266.7	-400.0	-66.7	100.0	226.7
数学	-400.0	1522.9	841.7	-114.6	-313.3
理科	-66.7	841.7	966.7	-75.0	-113.3
社会	100.0	-114.6	-75.0	56.3	103.3
英語	226.7	-313.3	-113.3	103.3	210.7

	現代文	数学	理科	社会	英語
現代文	1.0	-0.6	-0.1	0.8	1.0
数学	-0.6	1.0	0.7	-0.4	-0.6
理科	-0.1	0.7	1.0	-0.3	-0.3
社会	0.8	-0.4	-0.3	1.0	0.9
英語	1.0	-0.6	-0.3	0.9	1.0

表● 分散共分散行列（左）と相関係数行列（右）

表の読み方は難しくありません。行と列の変数の相関係数がその行と列が交差するセルに記されています。例えば、現代文と数学の共分散は−400.0で、その相関係数は−0.6という具合です。なお、現代文と英語の相関係数は、英語と現代文の相関係数でもありますので、ちょうど左上から右下にかけての対角線を境に、同じ値が登場します。このため、冗長さを避けるため灰色の上三角部分に何も書かず報告する研究もあります。

また、左上から右下に向かう対角線の位置には、自分自身との共分散や相関係数が記されます。自分自身との関連性は最大化しますので、相関係数行列の対角線には 1 が並びます。また、自分自身との共分散は分散ですので、分散共分散行列の対角線上には分散が記されます。分散と共分散が記されているので分散共分散行列と呼ぶわけです。

3.2 注意点

さて、みなさんが論文を読んでいて、次のような架空の記述にあったとしましょう。実はこの記述にはよくないところがあるのですが、それがどこだか分かるでしょうか。

「音楽の得点と英語の得点の相関係数は 0.8 だった。ここから、英語の得点を上げるためには、音楽の力を伸ばす必要があることが分かった。」

相関というのはあくまでも「関連性」があったということを示すだけで、「因果関係」は証明できません。次のような三つの可能性が存在するからです。

（1）音楽のスコアが英語のスコアに影響を与えるので相関係数が 0.8 になった。
（2）英語のスコアが音楽のスコアに影響を与えるので相関係数が 0.8 になった。
（3）音楽と英語のスコアには、互いに直接的な関係はないが、音楽と英語にともに影響を与える第三の変数（例：まじめさ、教科に関係ない学習意欲の高さ）があるので、見かけ上、0.8 という高い相関係数が観察された。

したがって、他に独立した証拠がない場合には、先ほどの記述は、次のような記述に修正し、因果関係があるような表現を避けることが望ましいと言えます。

「音楽の得点と英語の得点の相関係数は 0.8 だった。ここから、<u>両者にはそれなりに強い関連性があることが分かった</u>。」

しかし、「関連性」だけではなく、「因果関係」にこそ興味を持っている人にとっては「ものたりないな」と思ったかもしれません。気持ちはわかります。ですが、相関係数や共分散は、あくまで関連性の指標であって、因果関係の指標ではありません。にもかかわらず、学生さんレポートを採点していると、数字が示唆している以上の意味合いを（何の予備的な考察なく）読み取ってしまっているな、と思うことがしばしばあります。

今の英語と音楽の事例では、常識からおかしいと気づいた方もいたかもしれませんが、実際の研究では、二つの変数の関係が分からないからこそ研究をしているので、分からないがゆえに、相関係数に大きな値が出ると「あ、因果関係があるんだ」と誤解をしてしまうという間違いが起こりやすくなります。得られた数値データから「過剰に主張しすぎないことがないように心がける」という姿勢を忘れないでください。

でも、「自分が本当にやりたいのは、関連性ではなく因果関係を見ることなのに…」と残念に思った人もいたかもしれません。入門の範囲を超えてしまうので、本書で紹介することはできませんが、幸いにして、因果関係を議論できる統計手法というものは別に存在します。興味がある方は専門の授業などで学んでみてください。

最後に、相関係数の実践において、もう一つ、コメントをしておきましょう。それは、ここで扱った共分散や相関係数は、あくまで私たちが採取した標本（サンプル）に対する値であって、母集団における関連性を測っているわけではないということです。このため、新しい標本を採取したら値は毎回変わりますし、母集団の値が今回みなさんが計算した値と大きくずれていることもありえます。

研究者が興味を持っているのは、自分がとった標本の共分散や相関係数ではなく、母集団全体における共分散や相関係数であることがほとんどですので、第〇〇節でみた二群の差の検定（t 検定）のように、母集団相関係数が 0 であるかどうか（つまり、母集団において無相関ではないのかどうか）を検証したり、その信頼区間を出したりして、限られた標本から計算された標本相関係数をもとに母集団の相関について分析することも多くあります。詳細は紙幅の都合上割愛しますが、興味のある方は下に紹介する南風原（2002）などを参考に学びを深めてください。

4. おわりに

ぜひ自分が採ったデータを用いて、共分散や相関係数を計算し、この二つの指標に慣れ親しんでみてください。また、冒頭でお話ししたように、同じデータに対しても、相関係数を出すだけでなく、回帰分析やクラスター分析など別の角度からデータを吟味することも可能です。そこで、他の節の話もしっかり理解して多角的なデータ分析ができるように、ぜひ広い視野を持ちながら、定量的な手法をマスターしていってください。

≪考えてみよう！≫

・自分の研究分野ではどのような変数たちの間の関連性を分析してみたいですか？
・相関分析を用いた研究のリサーチクエスチョンを考えてみてください。

≪もっと詳しく読んでみよう！≫

1) 南風原朝和 (著) (2002).『心理統計学の基礎』有斐閣アルマ.
一つ一つ統計の基礎を説明するその丁寧な筆致は、確実にみなさんの統計に関する理解を広げてくれることでしょう。相関係数の検定についても詳しく説明しています。

2) 石川慎一郎・前田忠彦・山崎誠 (編) (2010).『言語研究のための統計入門』くろしお出版.
Excel を使いながら様々な多変量解析の手法が学べる書籍です。相関分析だけでなく、言語研究の事例を具体例に、実践的な統計利用のノウハウも学べます。

3) 日本統計学会 (編) (2023).『日本統計学会公式認定　統計検定３級・４級　公式問題集［CBT 対応版］』実務教育出版.
理解の定着度の確認に、統計検定３，４級の問題を解くというのもお勧めです。