散布図

これまでに色々なデータの指標を見てきたが、それらはデータの差異を比べるには都合がよかった。ただ、データの差異を知るだけでは展望がひらけないものだ。たとえば、現在あるデータをもとにして予測を立てる場合などだ。予測を立てるにはデータの関連を知る必要がある。そこで、次のような 2 種類のデータを考えてみよう。

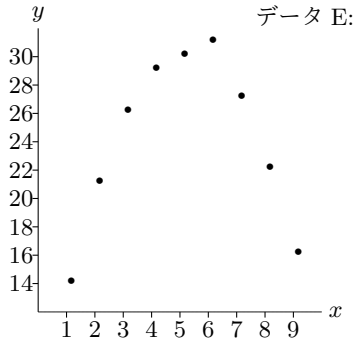
データ E:

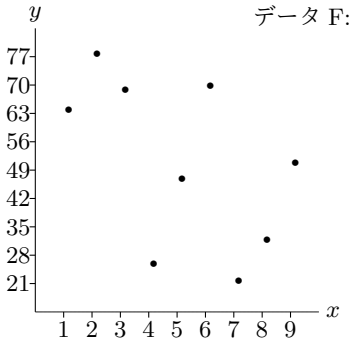
x	1	2	3	4	5	6	7	8	9
y	14	21	26	29	30	31	27	22	16

データ F:

x	1	2	3	4	5	6	7	8	9
y	63	77	68	25	46	69	21	31	50

一見して際立った規則があるわけではない。それというのも、数値というものは見た目で違いをはっきり認識できるほど、目に優しいわけではないからだ。やはり、このような場合はグラフ化する方がよい。次の図は**散布図**と呼ばれる。





すると一目瞭然であろう。データ E には傾向のようなものが見て取れるが、データ F には目立った傾向のようなものは見えない。この場合、データ E では、$x = 3.5$ 付近で $y = 28$ 付近の値をとるように思える。データ E の x、y には強い関係性があるとも言える。しかしデータ F では、$x = 3.5$ 付近の y の値といっても判然としない。データ F の x、y にはほとんど関連性がないとも言える。

データの相関

視覚的にデータに関連性がある・なしを区別することはできたとしても、一体どれほどの関連性を認めることができるだろうか。やはり、そういう場合は数値化するのがよいだろう。では、どのような考えに基づいて数値化できるだろうか。

今まで通り、偏差を基準に考えることにしよう。まず、データ E、F は共に x、y の平均値を求めることができる。データ E は $\overline{x}=5.0$、$\overline{y}=24.0$ で、データ F は $\overline{x}=5.0$、$\overline{y}=50.0$ であることはすぐ分かる。すると、各々の値 x_i、y_i に対する偏差 $\Delta x=|x_i-\overline{x}|$、$\Delta y=|y_i-\overline{y}|$ は

データ E:

Δx	4	3	2	1	0	1	2	3	4
Δy	10	3	2	5	6	7	3	2	8

データ F:

Δx	4	3	2	1	0	1	2	3	4
Δy	13	27	18	25	4	19	29	19	0

となっている。しかし、この表を見ても何かはっきりした特徴をつかむのは難しい。Δx が大きければ Δy も大きいなどの関連が見てとれないからである。

実は、データ E やデータ F のような場合は、関連性を数値化するのが少々困難なのである。それをやるのが数学だろうという考えもあろうが、数学の基本は『まずは単純化』である。数学で単純な関連性を示すものは**比例関係**であろう。そう考えると、データ E は前半と後半に分けて見た場合、非常に強い関連があることになる。そのため、このような関連性の強さを指標にするのが適当であると思われる。

そこで、比例には正の値で比例する場合と負の値で比例する場合があることを踏まえて、絶対値をとらない偏差を考えよう。データ E の前半は $\overline{x}=3.0$、$\overline{y}=24.0$ だから、偏差の対応に直すと

データ E:（前半）

x	1	2	3	4	5
y	14	21	26	29	30

$\Rightarrow$

データ E:（前半）

$(x_i-\overline{x})$	-2	-1	0	1	2
$(y_i-\overline{y})$	-10	-3	2	5	6

となるのに対し、データ E の後半は $\overline{x}=7.5$、$\overline{y}=24.0$ だから、偏差の対応に直すと

データ E:（後半）

x	6	7	8	9
y	31	27	22	16

$\Rightarrow$

データ E:（後半）

$(x_i-\overline{x})$	-1.5	-0.5	0.5	1.5
$(y_i-\overline{y})$	7	3	-2	-8

である。偏差の対応を見た場合、データ E の前半は x が大きくなるほど y は大きく、後半は x が大きくなるほど y は小さくなっている。このような関係を**相関がある**という。

ここで、絶対値をとらない偏差を考えた理由がはっきりしたと思う。絶対値をとってしまうと、偏差の大きさだけが際立つことになってしまうからである。そして、このようにしたために、データ E の前半の相関は右上がりの比例に近く、後半の相関は右下がりの比例に近くなっている。そこで、データ E の前半のような相関を**正の相関**、データ E の後半のような相関を**負の相関**と呼んで区別しておこう。

共分散

では、正の相関と負の相関を数値化するにはどうすればよいだろうか。それには、x、y それぞれの**偏差積**を求めるとよい。

データ E:（前半）

$(x_i-\overline{x})$	-2	-1	0	1	2
$(y_i-\overline{y})$	-10	-3	2	5	6
$(x_i-\overline{x})(y_i-\overline{y})$	20	3	0	5	12

データ E:（後半）

$(x_i-\overline{x})$	-1.5	-0.5	0.5	1.5
$(y_i-\overline{y})$	7	3	-2	-8
$(x_i-\overline{x})(y_i-\overline{y})$	-10.5	-1.5	-1	-12.5

すると、正の相関の場合には偏差積がより大きな正の値に、負の相関の場合には偏差積がより小さな負の値になる。したがって、指標となる数値を考えるなら、それらの合計を求めるとよい。ただし、単に和を求めるだけではデータ数の多少で値が左右されてしまうので、1 あたりの平均にするためにデータ数で割る必要がある。この値を、変量 x と変量 y の**共分散**といい、s_{xy} で表す。いまの場合、

$$\begin{aligned}\text{データ E の前半の共分散}\quad s_{xy} &= \frac{20+3+0+5+12}{5}=8.0\\ \text{データ E の後半の共分散}\quad s_{xy} &= \frac{-10.5-1.5-1-12.5}{4}=-6.25\end{aligned}$$

となっている。

2 変量からなる n 個のデータ $(x_1,\ y_1)$、$(x_2,\ y_2)$、…、$(x_n,\ y_n)$ に対し

$$\text{共分散 } s_{xy}=\frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}$$

ここでも、共分散の式は複雑で覚えるのは大変である。しかし

$$\text{共分散} := \text{偏差積、の平均}$$

と捉えれば、何も式だけに頼る必要はない。

* * *

これでデータ E については、その前半と後半で明らかに正の相関と負の相関が見られることが分かった。ところが全体として見た場合、データ E にはそれほど強い相関はないのである。実際にデータ E に関して共分散を求めてみよう。

データ E:

$(x_i-\overline{x})$	-4	-3	-2	-1	0	1	2	3	4
$(y_i-\overline{y})$	-10	-3	2	5	6	7	3	-2	-8
$(x_i-\overline{x})(y_i-\overline{y})$	40	9	-4	-5	0	7	6	-6	-32

これより、$s_{xy} = (40 + 9 - 4 - 5 + 0 + 7 + 6 - 6 - 32)/9 \approx 1.67$ である。これがどの程度の相関であるかは、データ F の共分散と比較するのがよいだろう。

データ F:

$(x_i - \overline{x})$	-4	-3	-2	-1	0	1	2	3	4
$(y_i - \overline{y})$	13	27	18	-25	-4	19	-29	-19	0
$(x_i - \overline{x})(y_i - \overline{y})$	-52	-81	-36	25	0	19	-58	-38	0

これより、$s_{xy} = (-52 - 81 - 36 + 25 + 0 + 19 - 58 - 38 + 0)/9 \approx -24.6$ である。データ E に比べてだいぶ強い負の相関があることが分かる。データの “関連性” という点ではデータ E はかなりはっきりした特徴を持っていたが、データの “相関” という点ではむしろデータ F の方が強い相関があるのである。このことから、共分散の考えは必ずしも万能とは言えないのである。▌

相関係数

共分散が万能でないのは仕方ないとしても、共分散にはちょっとした欠点がある。偏差積の総和を求めているので、単位も 2 変量の積になることである。たとえばデータ E において、仮に x がある年の立春から経過した月数、y がそのときの最高気温とすると、共分散の単位は（月 · 度）という訳の分からないものになっている。何らかの指標とするには、その都度単位が異なるのは良いことではない。

そこで、単位に関係ない値にするためには、共分散を（月 · 度）で割る必要がある。その際、何で割るのが適当だろうか。共分散は偏差積の和なので、やはり偏差で割るのがよいだろう。幸い x、y それぞれについて標準偏差というモノサシがあり、単位は（月）と（度）であるから都合がよい。データ E の x、y それぞれの標準偏差 s_x、s_y は

$$\begin{aligned} s_x &= \sqrt{\frac{(-4)^2 + (-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 + 4^2}{9}} = \sqrt{60} \\ s_y &= \sqrt{\frac{(-10)^2 + (-3)^2 + 2^2 + 5^2 + 6^2 + 7^2 + 3^2 + (-2)^2 + (-8)^2}{9}} = \sqrt{300} \end{aligned}$$

であるから、データ E の共分散から単位を取ると $\dfrac{1.67}{\sqrt{60}\sqrt{300}} \approx 0.012$ となる。また、データ F の x、y それぞれの標準偏差 s_x、s_y は

$$\begin{aligned} s_x &= \sqrt{\frac{(-4)^2 + (-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 + 4^2}{9}} = \sqrt{60} \\ s_y &= \sqrt{\frac{13^2 + 27^2 + 18^2 + (-25)^2 + (-4)^2 + 19^2 + (-29)^2 + (-19)^2 + 0^2}{9}} = \sqrt{3906} \end{aligned}$$

であるから、データ F の共分散から単位を取ると $\dfrac{-24.6}{\sqrt{60}\sqrt{3906}} \approx -0.051$ となる。いずれも、かなり 0 に近い値となった。

このように、共分散から単位を取り除いた値を**相関係数**と呼ぶが、それではこの値はどんな意味を持つのだろうか。相関係数を一般化した式

$$\frac{s_{xy}}{s_x s_y} = \frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}\sqrt{(y_1-\overline{y})^2+(y_2-\overline{x})^2+\cdots+(y_n-\overline{y})^2}}$$

から読み解いてみよう。ちなみに共分散、標準偏差は n で割っていたはずだが、上に示した式は分子・分母共に n で約分されていることに注意されたい。

さて、分子のひとつの項 $(x_i-\overline{x})(y_i-\overline{y})$ と分母のひとつの項 $(x_i-\overline{x})^2$ および $(y_i-\overline{y})^2$ に注目してもらいたい。これらの項の大小関係は

$$|(x_i-\overline{x})(y_i-\overline{y})| \leqq (x_i-\overline{x})^2 \text{ または } (y_i-\overline{y})^2$$

である。なぜなら、$|(x_i-\overline{x})(y_i-\overline{y})|$ は、$(x_i-\overline{x})$ と $(y_i-\overline{y})$ の大きい方の 2 乗より小さいに決まっているからだ。すると $|(x_i-\overline{x})(y_i-\overline{y})|$ は、$(x_i-\overline{x})=(y_i-\overline{y})$ のときが最大になるから、

$$\frac{s_{xy}}{s_x s_y} = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{\sqrt{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}\sqrt{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}} = 1$$

が最大である。同様に、最小は -1 である。したがって

$$\boxed{-1 \leqq \frac{s_{xy}}{s_x s_y} \leqq 1}$$

が分かる。つまり相関係数は、一番強い正の相関を 1、一番強い負の相関を -1 とする指標ということである。そして、0 なら相関がないものと考える。

そうすると、データ E もデータ F も相関係数の指標のもとでは、ほとんど相関がないという結果になるのである。

* * *

先ほど “関連性” と “相関” とを区別して用いたが、似たような言葉遣いに “因果関係” と “相関関係” がある。たとえば、冬には灯油の売れ行きが増えるし風邪をひく人の数も増えるので、灯油の売れ行きと風邪をひく人の数には正の相関があると言える。しかし、灯油を買って家に置いたとき揮発成分が風邪を誘発しているわけではないので、灯油が原因で風邪の結果をもたらすわけではない。すなわち、灯油と風邪に因果関係はない。灯油の売れ行きと風邪をひく人の数が増える原因は寒さである。あるものに相関関係が認められたとしても、一概に因果関係があるとは限らない。相関関係の裏に原因があることもあるし、単に偶然による関連のこともあるだろう。相関関係から何らかの結論を導くときは注意が要る。▌