# 基礎的な統計学

Author: 小倉 康睦
Date: 15/12/2017

# 統計学

データの分析は難しい言葉で言うと「統計学」という分野になります。統計は確率と密接に関係しています。

統計において重要なのは「そのデータを見たときに何が知りたいのか」ということです。何かを知りたいからデータを集めるのであって、集めたデータから意味のある量を求めるのが統計です。

## 確率との関係

統計における平均は、実は確率における期待値によって定義されています。まずはその繋がりを見ていきましょう。

高校の教科書では平均は次のように定義されています。

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \tag{1}$$

一方で期待値の定義は、$x_i$が観測される確率を$p(x_i)$で表すことにして

$$E = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n) \tag{2}$$

と定義されています。二つの式がよく似ていることに気がついたでしょうか？

この二つの式を結びつけるためにちょっと考えてみましょう。そういえばどうしてサイコロの目が出る確率は、どの目も$\frac{1}{6}$なのでしょう。実際に100万回投げてみたって、出る目の割合はぴったり$\frac{1}{6}$ずつになるわけではありません。実はこの魔法の数字$\frac{1}{6}$は「同様に確からしい」という呪文によって導かれたものです。

「どの目が出る確率も同様に確からしい」とは「どの目が出る確率も同じということにしておきます」という意味です。要するに、実際に確率を調べるのが面倒臭いからサボったのです。$i$の目が出る確率を$p_i$と表すとき、どの目が出る確率も同様に確からしいとすると、ある確率$p$が存在して、

$$p_1 = p_2 = \cdots = p_6 = p \tag{3}$$

が成り立ちます。一方で、全ての事象の確率を足すとは1ですから、

$$p_1 + p_2 + \cdots + p_6 = 1 \tag{4}$$

となります。(3)式を(4)式に代入すると、

$$p + p + \cdots + p = 6p = 1 \tag{5}$$

$$\therefore p = \frac{1}{6} \tag{6}$$

が導かれます。同様にして、(2)式で$x_1, x_2, \cdots, x_n$が出る確率が同様に確からしいと仮定すると、

$$\begin{cases} p_1 = p_2 = \cdots = p_n = p \\ p_1 + p_2 + \cdots + p_n = 1 \end{cases} \tag{7}$$

より

$$p_1 = p_2 = \cdots = p_n = p = \frac{1}{n} \tag{8}$$

となります。これを(2)式に代入すると、

$$\begin{aligned} E &= x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n) \\ &= x_1 \frac{1}{n} + x_2 \frac{1}{n} + \cdots + x_n \frac{1}{n} \\ &= \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \\ &= \overline{x} \end{aligned} \tag{9}$$

となっており、確かに期待値と平均が一致することが確認できます。統計学では平均と期待値は同じものとして扱われ、どちらの名前で呼んでもよいことになっています。

## 度数分布と確率

先ほど平均と期待値が一致するための条件に「$x_1, x_2, \cdots, x_n$が観測される確率が同様に確からしい」という仮定を置きましたが、これは少し無理のある仮定だと思いませんか。たとえば日本全国の高校生からランダムにひとり選ぶとき、身長165cmと身長180cmが選ばれる確率は同じくらいと言えるでしょうか。

身長がぴったり165.000000…の人も180.000000…cmの人も地球上にいないので、厳密にはいずれも確率は等しく0（原子1個分くらいはズレているはず）なのですが、たとえば小数点以下第2位を四捨五入して身長$x_i$が$179.95 \leq x_i < 180.05$くらいまでの範囲にあるとき身長180cmとします。180cmぴったりかどうか測ることは難しいですが、ある範囲に入るかどうか測定すること、つまり度数分布表を作成することは比較的簡単です。

では度数分布表を作った上で、日本全国の高校生からランダムにひとり選ぶとき、階級が165の人と階級が180の人が選ばれる確率は同じくらいと言えるでしょうか。これはたまたま同じになることはありますが、普通は異なりますよね。

実際に例を見てみましょう。あるみかん農園で採れたみかんは下の表のようであったとします。

| $i$ | 1 | 2 | 3 | $\cdots$ | 100 |
|---|---|---|---|---|---|
| $x_i$ | 84.3 | 73.6 | 85.2 | $\cdots$ | 86.9 |

ここから度数分布表を作ると次のようになります。たとえばこの農園では製品として扱えるみかんを重さで超小玉（75g付近）、小玉（80g付近）、中玉（85g付近）、大玉（90g付近）、超大玉（95g付近）の5階級に分類しています。$q_j$は階級$X_j$に属するみかんの個数を表すとします。

| $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X_j$ | 75 | 80 | 85 | 90 | 95 |
| $q_j$ | 15 | 23 | 37 | 20 | 5 |

$q_j$を、取り出した集団の大きさ100で$q_j$を割ってやることによって、割合としておおよその確率$p(X_j)$が求まるので、これも次に書いておきます。高校の確率の問題で期待値を求めるときに確率分布表というものを作ったはずですが、度数分布と確率分布はほとんど同じものだということがわかるでしょう。

| $j$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $X_j$ | 75 | 80 | 85 | 90 | 95 |
| $p(X_j)$ | $\frac{15}{100}$ | $\frac{23}{100}$ | $\frac{37}{100}$ | $\frac{20}{100}$ | $\frac{5}{100}$ |

このみかん農園で採れるみかんの重さの平均を求めよと言われた場合、本来ならば次のようにして求めることになります。

$$\overline{x} = \frac{1}{100}(x_1 + x_2 + \cdots + x_n) \tag{10}$$

これがとても面倒臭いということがわかるでしょうか。この式を実際に計算しようとすると、みかんの重さを1個1個量った上で、同じみかんの重さを2回量ってしまわないよう、それぞれに名前をつけて慎重に管理しなければなりません。普通はみかんごときにそこまでしません。そこで代わりに度数分布を見て、

$$\overline{X} = X_1 p(X_1) + X_2 p(X_2) + \cdots + X_5 p(X_5) \tag{11}$$

としたらどうでしょうか。実際に計算してみましょう。

$$\begin{aligned}\overline{X} &= X_1 p(X_1) + X_2 p(X_2) + \cdots + X_5 p(X_5) \\ &= 75 \times \frac{15}{100} + 80 \times \frac{23}{100} + \cdots + 95 \times \frac{5}{100} \\ &= \frac{1}{100}(75 \times 15 + 80 \times 23 + 85 \times 37 + 90 \times 20 + 95 \times 5)\end{aligned} \tag{12}$$

これは簡単です。1回量ったみかんはその階級の箱に入れておいて、あとで個数を数え直せばすぐに計算できます。そして(12)式は(10)式と見比べると大雑把に一致していることが次の式によってわかります。

$$\begin{aligned}\overline{x} &= \frac{1}{100}(x_1 + x_2 + \cdots + x_n) \\ &= \frac{1}{100}(84.3 + 87.6 + 85.2 + \cdots + 86.9) \\ &\fallingdotseq \frac{1}{100}(85 + 90 + 85 + \cdots + 85) \\ &= \frac{1}{100}(X_3 + X_4 + \cdots + X_3) \\ &= \frac{1}{100}(X_1 q_1 + X_2 q_2 + X_3 q_3 + X_4 q_4 + X_5 q_5) \\ &= \frac{1}{100}(75 \times 15 + 80 \times 23 + 85 \times 37 + 90 \times 20 + 95 \times 5) \\ &= \overline{X}\end{aligned}$$

要するに$\overline{X}$は、$x_1, x_2, \ldots, x_n$のそれぞれを$X_1, X_2, \ldots, X_5$のうちでもっとも近い値に置き直してから平均を取ったものと考えることができるのです。

## 期待値と分散

次の二つのみかん農園$A, B$を比べることを考えてみましょう。

| $A$ | | | | | |
|---|---|---|---|---|---|
| $j$ | 1 | 2 | 3 | 4 | 5 |
| $X_j$ | 75 | 80 | 85 | 90 | 95 |
| $q_j$ | 15 | 23 | 37 | 20 | 5 |
| $p(X_j)$ | $\frac{15}{100}$ | $\frac{23}{100}$ | $\frac{37}{100}$ | $\frac{20}{100}$ | $\frac{5}{100}$ |

| $B$ | | | | | |
|---|---|---|---|---|---|
| $j$ | 1 | 2 | 3 | 4 | 5 |
| $X_j$ | 75 | 80 | 85 | 90 | 95 |
| $q_j$ | 21 | 23 | 22 | 20 | 14 |
| $p(X_j)$ | $\frac{21}{100}$ | $\frac{23}{100}$ | $\frac{22}{100}$ | $\frac{20}{100}$ | $\frac{14}{100}$ |

実際に計算してもらえればわかることですが、この二つの農園で採れるみかんの重さの期待値を比べたとき、次のようになっています。

| | $\overline{X}$ |
|---|---|
| $A$ | 83.85 |
| $B$ | 84.15 |

どちらのみかん農園でも採れるみかんの重さの期待値は84g程度です。すなわち、期待値を見た限りではどちらのみかん農園からも中玉（85g付近）のみかんが採れそうだと思うでしょう。しかし実際は、$B$の農園で採れる確率がもっとも高いのは小玉（80g付近）です。したがって、中玉のみかんを求めて$B$の農園に行っても、$A$の農園より中玉のみかんは手に入りにくいのです。

なぜこんなことが起こったのでしょうか。実は、$A$の農園では中玉のみかんがとてもよくなる品種を育てているのですが、$B$の農園では小玉のみかんから大玉のみかんまでまんべんなくなる品種を育てているのです。

こういった現象を比較するための尺度が分散です。分散は、観測したデータがどれくらい期待値の付近から散らばっているかを表す値です。分散が大きいほど、観測されるデータは期待値から離れたものになる確率が高くなります。高校の教科書では分散は次のように定義されています。

$$\sigma^2 = \frac{1}{n}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\} \tag{13}$$

統計学における分散は、$x_i$が観測される確率を$p(x_i)$で表すことにして

$$\sigma^2 = (x_1 - \overline{x})^2 p(x_1) + (x_2 - \overline{x})^2 p(x_2) + \cdots + (x_n - \overline{x})^2 p(x_n) \tag{14}$$

と定義されています。期待値のときと同様にして、$x_1, x_2, \cdots, x_n$が観測される確率が同様に確からしいという、特殊な状況を仮定すれば(13)式と(14)式は一致します。また、期待値のときと同様にして、$x_1, x_2, \ldots, x_n$のそれぞれを$X_1, X_2, \ldots, X_5$のうちでもっとも近い値に置き直してから、

$$\sigma^2 = (X_1 - \overline{X})^2 p(X_1) + (X_2 - \overline{X})^2 p(X_2) + \cdots + (X_n - \overline{X})^2 p(X_n) \tag{15}$$

とすることで大雑把な計算ができます。この定義を用いて実際に$A$の農園と$B$の農園を比較してみると、

| | $\overline{X}$ | $\sigma^2$ | $\sigma$ |
|---|---|---|---|
| $A$ | 83.85 | 29.43 | 5.42 |
| $B$ | 84.15 | 45.03 | 6.71 |

となり、$B$の農園のほうが分散が大きくなっていることがわかります。分散が小さいほど、その集団からランダムに取り出したデータは期待値に近い値を取る確率が大きいので、中玉のみかんを求めるならば$A$の農園に行くのが正解だということがわかります。

$\sigma$は分散$\sigma^2$のルートを取ったもので、標準偏差と呼ばれています。しかしこれではまだどうして分散でデータの散らばり具合を測ることができるのか、また、どうして分散のルートを取った標準偏差をわざわざ考える必要があるのか分かりません。分散と標準偏差についてはもう少しじっくりと考える必要がありそうです。

## 分散と標準偏差の意味

分散は観測されたデータが期待値からどれだけばらついているかを示す数値であると言いましたが、距離のようなものであると考えることもできます。分散を表す(13)式をもう一度書いておきます。

$$\sigma^2 = \frac{1}{n}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\}$$

ここで、{}の中の$i$番目の項だけを取り出して

$$r^2 = (x_i - \overline{x})^2 \tag{16}$$

という式と対比してみます。すると、標準偏差$\sigma$に対応するのは

$$r = \sqrt{(x_i - \overline{x})^2} = |x_i - \overline{x}| \tag{17}$$

です。$r$はデータ点$x_i$と期待値$\overline{x}$との距離を表しています。(16)式は(17)式を2乗したものですから、距離を2乗したような概念であることがわかります。このイメージを保ったまま(13)式を見直してみれば、分散$\sigma^2$は「各データ点$x_1, x_2, \ldots, x_n$のそれぞれと、データの平均$\overline{x}$との距離を2乗したものの期待値」を表していることがわかります。つまりその集団から適当なデータ点$x$を選んだとき、そのデータ点$x$はデータの平均$\overline{x}$から（距離を2乗した尺度で）$\sigma^2$くらい離れているだろうと期待できるのです。

距離を2乗した尺度は、私たちの直感を歪めたものになっています。$0.1^2 = 0.01$や、$100^2 = 10000$となることを見ればわかる通り、近くにあるデータはより近くに、遠くにあるデータはより遠くにあるように見えてしまいます。その歪みを補正するために、ルートを取って元に戻した尺度が標準偏差$\sigma$というわけです。

分散の意味を理解するために、非常に大雑把な評価ではありますが、どんな確率分布に対しても成り立つチェビシェフの不等式という評価尺度が存在します。

$$\frac{1}{a^2} \geq (P(|x - \overline{x}| \geq a\sigma)$$

この式は高校では習わない記号法で書かれているので理解するのは難しいかもしれませんが「平均$\overline{x}$から$a\sigma$以上離れた場所にある点が見出される確率は$\frac{1}{a^2}$以下である」という意味を持った式です。たとえば$a = 2$のときを考えれば、平均$\overline{x}$から$2\sigma$以上離れたところで値が見つかる確率は$\frac{1}{4}$以下であるということを主張しています。対偶を取れば「平均$\overline{x}$から$2\sigma$より近いところでデータが見つかる確率は$\frac{3}{4}$よりも大きい」ということになります。証明は大学の統計の教科書に譲ることにします。

先ほどのみかん農園$A$で考えれば、平均$\overline{X} = 83.85$から$2\sigma = 2 \times 5.42 = 10.84$以上離れたみかんが見つかる確率は$\frac{1}{4}$以下であるということで、実際に$X_j \leq 73.01,\ 94.69 \leq X_j$の範囲にある$X_5 = 95$のみかんが見つかる確率は$p(X_j) = \frac{5}{100} = \frac{1}{20} \leq \frac{1}{4}$なので、確かに成り立っています。

ここまでの話を非常にざっくりとまとめれば、分散は「観測されたデータがどれだけ期待値の近くに集まっているか」を見るための尺度であるといえます。

## 分散と共分散

分散$V(x)$の定義と共分散$Cov(x, y)$の定義を見比べてみましょう。

$$V(x) = \frac{1}{n}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\} \tag{18}$$

$$Cov(x, y) = \frac{1}{n}\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\} \tag{19}$$

(18)式は(13)式で$V(x) = \sigma^2$と置いただけの式です。$(x - \overline{x})^2 = (x - \overline{x})(x - \overline{x})$であることを考えれば、共分散の式は分散の式とほとんど同じであることがわかります。唯一の違いは、共分散が定義できるようなデータは2次元以上のデータ（みかんを特徴付けるための2つ以上の尺度）を持っていることです。たとえば今までみかん農園ではみかんの重さ$x$だけを量っていましたが、それぞれのみかんの糖度$y$と、そのみかんを収穫した週の平均気温$z$も測ることにした状況を考えます（ただしこれらのデータは気温にバラつきを与えるため、などの理由でわざと過去10年分のデータをごちゃまぜにして記録してあります）。さて、このみかん農園で採れたみかんについて記録した表は下のようになっていたとします。

| $i$ | 1 | 2 | 3 | $\cdots$ | $n$ |
|---|---|---|---|---|---|
| $x_i$ | 75 | 80 | 85 | $\cdots$ | 95 |
| $y_i$ | 18 | 21 | 19 | $\cdots$ | 20 |
| $z_i$ | 13 | 10 | 12 | $\cdots$ | 12 |

分散は$x, y, z$それぞれの尺度に対して考えることができるので、この表からは三種類の分散が計算できます。

$$
\begin{aligned}
V(x) &= \frac{1}{n}\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\} \\
V(y) &= \frac{1}{n}\{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2\} \\
V(z) &= \frac{1}{n}\{(z_1 - \overline{z})^2 + (z_2 - \overline{z})^2 + \cdots + (z_n - \overline{z})^2\}
\end{aligned} \tag{20}
$$

簡単にいえば、みかんの重さ、みかんの糖度、収穫時の気温のそれぞれについて、バラつきを測ってやることができるということです。また、少し考えれば$Cov(x, y) = Cov(y, x)$であることはすぐにわかるので、この表から計算できる共分散は三種類あるとわかります。

$$
\begin{aligned}
Cov(x, y) = Cov(y, x) &= \frac{1}{n}\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\} \\
Cov(y, z) = Cov(z, y) &= \frac{1}{n}\{(y_1 - \overline{y})(z_1 - \overline{z}) + (y_2 - \overline{y})(z_2 - \overline{z}) + \cdots + (y_n - \overline{y})(z_n - \overline{z})\} \\
Cov(z, x) = Cov(x, z) &= \frac{1}{n}\{(z_1 - \overline{z})(x_1 - \overline{x}) + (z_2 - \overline{z})(x_2 - \overline{x}) + \cdots + (z_n - \overline{z})(x_n - \overline{x})\}
\end{aligned} \tag{21}
$$

分散は距離の比喩を用いて既にイメージできていると思いますが、共分散とはどんなものでしょう。イメージを掴めるでしょうか。

## 共分散の意味

ここで共分散の意味を考えておくことにしましょう。少し見辛いので(18)式と(19)式の$i$番目の項だけをそれぞれ取り出して比較してみます。

$$r^2 = (x_i - \overline{x})(x_i - \overline{x}) \tag{22}$$
$$c = (x_i - \overline{x})(y_i - \overline{y}) \tag{23}$$

$c$が2乗されていないのは誤植ではありません。(23)式の右辺は負の値を取りうるので、左辺を勝手に2乗して負ではない値の範囲で考えるわけにはいかないのです。(22)式の解釈については『分散の意味』の部分を見返してもらうことにして、今は(23)式に集中しましょう。

$c$がどういう性質を持った数であるかを考えると、次のようになっています。

- $x_i$の平均$\overline{x}$からのズレが正 かつ $y_i$の平均$\overline{y}$からのズレが正 $\Rightarrow c > 0$
- $x_i$の平均$\overline{x}$からのズレが正 かつ $y_i$の平均$\overline{y}$からのズレが負 $\Rightarrow c < 0$
- $x_i$の平均$\overline{x}$からのズレが負 かつ $y_i$の平均$\overline{y}$からのズレが正 $\Rightarrow c < 0$
- $x_i$の平均$\overline{x}$からのズレが負 かつ $y_i$の平均$\overline{y}$からのズレが負 $\Rightarrow c > 0$

すなわち、$c$は$x$と$y$でそれぞれ平均からのズレが同符号ならば正、異符号ならば負の値を取る数です。共分散とはこれをすべてのデータ点について足し合わせた値になります。したがって共分散$Cov(x,y)$が大きな正の値を取るときは、次のような傾向があります。

- $x_i$が大きくなるほど$y_i$も大きくなる（$x_i$が小さくなるほど$y_i$も小さくなる）。

反対に、共分散$Cov(x,y)$が大きな負の値を取るときは次のような傾向があります。

- $x_i$が大きくなるほど$y_i$は小さくなる（$x_i$が小さくなるほど$y_i$は大きくなる）。

そして共分散$Cov(x,y)$が0に近いときは、次のような傾向が考えられます。

- $x_i$の増減は$y_i$の増減に関係ない（$c$が正の値を取ったり負の値を取ったりするので足し合わせると打ち消しあう）。

ただし$x_i$,$y_i$のいずれかの分散がそもそも小さい場合は$x_i-\overline{x}, y_i-\overline{y}$の値がそもそも小さくなるため、共分散は小さくなります。逆に$x_i$,$y_i$のいずれかの分散がもともと大きい場合は$x_i-\overline{x}, y_i-\overline{y}$の値が大きくなるため、共分散は大きくなります。したがって共分散の大小を見ただけでは本当に「$x_i$が大きくなるほど$y_i$は小さくなる」という関係があるのか、それとも$x_i$と$y_i$の増減にはなんの関係もないのかを正確に判断することはできません。そこで用いるのが相関係数です。

## 相関係数

相関係数は2つのデータがどれだけ似通っているかを表す指標です。$x$と$y$の相関係数は分散と共分散を用いて次のように定義されます。

$$r=\frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} \tag{24}$$

この式を理解するために今一度共分散の定義を見てみましょう。

$$Cov(x,y)=\frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})\}$$

この式でもしもそれぞれの$y_i$に、対応する$x_i$とまったく同じ値が入っていたとしたら、共分散は$x$の分散に一致します。すなわち$Cov(x,y)=Var(x)$です。このときさらに$Var(x)=Var(y)$となっていますから、相関係数を求めると、

$$r=\frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}=\frac{Var(x)}{\sqrt{Var(x)}\sqrt{Var(x)}}=1$$

となります。すなわちまったく同じデータどうしの相関係数は1になるのです。ではそれぞれの$y_i$に、対応する$x_i$を$a$倍した（ただし$a$は正の実数）データを入れてみたらどうなるでしょうか。

$$\begin{aligned}Cov(x,ax)&=\frac{1}{n}\{(x_1-\overline{x})(ax_1-a\overline{x})+(x_2-\overline{x})(ax_2-a\overline{x})+\cdots+(x_n-\overline{x})(ax_n-a\overline{x})\}\\&=a\cdot Cov(x,x)\\&=a\cdot Var(x)\end{aligned}$$

となります。同様にして$Var(ax)=a^2\cdot Var(x)$が導かれるので、

$$r=\frac{Cov(x,ax)}{\sqrt{Var(x)}\sqrt{Var(ax)}}=\frac{a\cdot Var(x)}{\sqrt{Var(x)}\sqrt{a^2\cdot Var(x)}}=1$$

となって、やはり相関係数は1になります。今度は$-a$倍してみると$Cov(x,-ax)=-a\cdot Cov(x,x)=-a\cdot Var(x)$

と$Var(-ax) = a^2 \cdot Var(x)$が求まるので、

$$r = \frac{Cov(x, -ax)}{\sqrt{Var(x)}\sqrt{Var(-ax)}} = \frac{-a \cdot Var(x)}{\sqrt{Var(x)}\sqrt{a^2 \cdot Var(x)}} = -1$$

となります。また、それぞれのy_iに、対応する$x_i$にbを足した（ただし$b$は任意の実数）データを入れてみたらどうなるでしょう。

$$\begin{aligned} Cov(x, x+b) &= \frac{1}{n}\{(x_1 - \bar{x})(x_1 + b - (\bar{x} + b)) + (x_2 - \bar{x})(x_2 + b - (\bar{x} + b)) + \cdots + (x_n - \bar{x})(x_n + b - (\bar{x} + b))\} \\ &= Cov(x, x) \\ &= Var(x) \end{aligned}$$

同様にして$Var(x + b) = Var(x)$が求まるので、

$$r = \frac{Cov(x, x + b)}{\sqrt{Var(x)}\sqrt{Var(x + b)}} = \frac{a \cdot Var(x)}{\sqrt{Var(x)}\sqrt{a^2 \cdot Var(x)}} = 1$$

となります。以上のことをまとめると、相関係数には次のような性質があります。

- データ$x, y$が互いに比例関係にあり、比例定数が正ならば1に近づく。
- データ$x, y$が互いに比例関係にあり、比例定数が負ならば$-1$に近づく。
- データ$x, y$が互いに比例関係にないとき、0に近づく（分子の共分散が0に近づくため）。

つまり相関係数を見るだけで、観測した範囲での$x$と$y$がどのくらい比例関係に近いかを確認できるというわけです。

## 相関係数の解釈に関する注意

相関係数は、あくまでも観測した範囲での$x$と$y$がどのくらい比例関係に近いかを表す尺度であって、それ以上でも以下でもありません。

よく言われるのは因果関係との違いです。データ$x$とデータ$y$の相関係数が1に近いからといって「$x$のせいで$y$が起こった」とか、その逆に「$y$のせいで$x$が起こった」とかは言えないのです。因果関係はないのに、いろんな理由でたまたま相関係数が大きくなることを「疑似相関」といいます。疑似相関ばかりを集めたサイト『Spurious Correlations』には、たとえば「一人当たりの年間チーズ消費量と、年間でベッドシーツに絡まって死ぬ人の数には相関がある」とか「年間でプールに落っこちて溺れ死ぬ人の数と、ニコラス・ケイジの年間出演映画数には相関がある」とかグラフ付きで書いてあります。しかしこれらの間に因果関係があるとするには、いくらなんでも相関係数だけでは根拠薄弱でしょう。

また、見過ごされがちですが、相関係数が0に近いからといって「$x$と$y$の間にはまったく関係はない」ということもできません。もっとも簡単な例を以下に示します。次の表に示したデータ$x$と$y$の間で相関係数を求めてみてください。

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | −4 | −3 | −2 | −1 | 0 | 1 | 2 | 3 | 4 |
| $y_i$ | 16 | 9 | 4 | 1 | 0 | 1 | 4 | 9 | 16 |

実際に求めてみると相関係数は0になっていますが、種明かしをすればこのデータは$y = x^2$という関係に基づいて生成されています。相関係数で確かめることのできる関係性はあくまでも比例関係だけであるということには常に注意しておいたほうがよいでしょう。

# 参考文献

理系の大学生向けに書かれていますが、微分積分には学部1年程度の解析学の知識が必要となりますが、大部分は高校数学の範囲でも理解できるように丁寧に書かれた書籍です。

[1] 『理工系の数学入門コース7 確率・統計』薩摩順吉 著