相関係数はなぜ標準偏差で割るのか

－統計分野の「理不尽」を考察する－

旭川南高校
岡崎　知之

０．はじめに

「新学習指導要領」では、数学科カリキュラムの大幅な変更がなされ、特に「データサイエンス」を目的とした単元に力点がおかれている。しかしながら現在の学習指導要領の下では、数学Ⅰ「データの分析」・数学Ｂ「確率分布と統計的な推測」、いずれも履修することを肯定的に捉える教員は少ないように感じる。私はその原因を次の２点だと考えている。

１つは、教員自身に「統計」の知識がないことである。「統計」は大学入試の必須分野でないことから、高校時代に習っていない教員が多く、また数学の教員免許が取得できる大学の学科では、「統計」の講義が必須となっていない場合が大半である。数学Ⅰで「データの分析」が導入される際にも、現場では大きな混乱が見られた。

もう１つは、「理不尽」だからである。高校数学で紹介される公式や定理は、証明という保障の上で活用されている。ところが、「統計」は式の成り立ちや、統計量についての説明が不十分であるものが少なくない。仕組みが分からないブラックボックスを、半信半疑で活用することに数学教員は大きな抵抗感を抱えるのではないだろうか。

「知識がないこと」は新たに知識を獲得すればよいが、「理不尽であること」はどう解消すれば良いのだろう。

本レポートではその事例として、教科書の中で読み過ごしているような「統計」における２つの理不尽を挙げる。

１．相関係数の「$s_x s_y$」

教科書に掲載されている「相関係数」の内容を、以下のようにまとめてみた。

☆定義

$$r = \frac{s_{xy}}{s_x s_y}$$

☆性質

・相関のようす（正・負）を判断することができる。

・$-1 \leq r \leq 1$

・rの値が-1 に近いほど負の相関が強く、１に近いほど正の相関が強い。

☆根拠

散布図をx, yの平均値の前後で４つの領域に区分すると、偏差の積が正（負）となるデータは正（負）の相関をなす領域に多く存在するため、共分散の値で相関を判断することができる。

この説明では、以下の点が述べられていない。

(1) 何故、共分散を標準偏差の積（$s_x s_y$）で割るのか。
(2) 何故、$-1 \leq r \leq 1$が成立するのか。

この２点を考察してみた。

(1) 何故、共分散を標準偏差の積（$s_x s_y$）で割るのか。

⇒相関係数の範囲が不定ならば、複数のデータで相関の度合いが比較できないから。

（例）もし、標準偏差の積で割らなかったら…

データＡの共分散が３，データＢの共分散が１０

⇒データＢの方が正の相関が強い、とは言い切れない。

(2) 何故、$-1 \leq r \leq 1$が成立するのか。

⇒$-s_x s_y \leq s_{xy} \leq s_x s_y$が成立するから

（証明）$s_{xy} \geq 0$のとき

$$\left(s_x s_y\right)^2 = (s_x)^2\left(s_y\right)^2 = \left(\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2\right)\left(\frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2\right) = \frac{1}{n^2}\left(\sum_{k=1}^{n}\widetilde{x_k}^2\right)\left(\sum_{k=1}^{n}\widetilde{y_k}^2\right)$$

$$\left(s_{xy}\right)^2 = \left(\frac{1}{n}\sum_{k=1}^{n}(x - \bar{x})\,(y - \bar{y})\right)^2 = \frac{1}{n^2}\left(\sum_{k=1}^{n}(x - \bar{x})\,(y - \bar{y})\right)^2 = \frac{1}{n^2}\left(\sum_{k=1}^{n}\widetilde{x_k}\,\widetilde{y_k}\right)^2$$

ここで、Schwaltz の不等式から、

$\left(\sum_{k=1}^{n}\widetilde{x_k}\,\widetilde{y_k}\right)^2 \leq \left(\sum_{k=1}^{n}\widetilde{x_k}^2\right)\left(\sum_{k=1}^{n}\widetilde{y_k}^2\right)$が成立する。

よって、

$\left(s_{xy}\right)^2 \leq \left(s_x s_y\right)^2$

$\therefore\ s_{xy} \leq s_x s_y$

ちなみに、等号成立は

$x_1 : x_2 : x_3 : \cdots : x_n = y_1 : y_2 : y_3 : \cdots : y_n$となり、任意の$k$に対し

$y_k = ax_k$　（aは定数）　が成立し、各データが一直線上に並ぶときに、

$r = 1$となることが分かる。

＊$s_{xy} < 0$の場合も同様に証明可能。

２．「正規分布は確率密度関数」

教科書の「確率密度関数」「正規分布」についての記述をまとめてみた。

☆確率密度関数の性質

・常に$f(x) \geq 0$で　$P(a \leq X \leq b) = \int_a^b f(x)dx$

・Xのとる値の範囲が　$\alpha \leq x \leq \beta$のとき　$\int_\alpha^\beta f(x) = 1$

☆正規分布の定義

mを実数、σを正の定数とするとき、

関数 $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$　は連続型確率変数 X の確率密度関数になることが**知られている**。

このとき Xは正規分布$N(m, \sigma^2)$に従うという。

この説明では、関数$f(x)$が確率密度関数であることの説明が省略されている。

試しに確率密度関数の性質　$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}dx = 1$を確認してみた。

（予備知識）

$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$ は「ガウス積分」と呼ばれている。

この式を認めれば、

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

ここで、$t = \frac{x-m}{\sqrt{2\sigma}}$とすると

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-t^2} \sqrt{2\sigma} dt = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt = 1$$となる。

（「ガウス積分」の証明）

面積要素dAを

xy一直交座標系を用いて　$dA = dxdy$　，$r\theta$-極座標系では　$dA = rdrd\theta$で表されるので、

$$\int_{R^2} exp\{-(x^2+y^2)\} dA = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dxdy = \left(\int_{-\infty}^{\infty} e^{-t^2}\,dt\right)^2 \quad \cdots①$$

$$\int_{R^2} exp\{-(x^2+y^2)\} dA = \int_0^{2\pi} \int_0^{\infty} exp(-r^2) rdrd\theta = 2\pi \int_0^{\infty} r\, exp(-r^2) dr$$

ここで、$s = -r^2$とすると、

$$\int_{R^2} exp\{-(x^2+y^2)\} dA = 2\pi \int_0^{\infty} r\, exp(-r^2) dr = \pi \int_{-\infty}^{0} e^s ds = \pi \quad \cdots②$$

①，②より

$$\left(\int_{-\infty}^{\infty} e^{-t^2}\,dt\right)^2 = \int_{-\infty}^{\infty} e^{-x^2}\,dx = \pi$$

$$\therefore \int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

3．考察を通して

教科書において「統計」が他の単元と異なり、「理不尽」な部分が多い原因は、そもそも理論が高校数学の範囲を（一般の高校生が履修するレベルで）超えているからであると言える。

個人的な意見だが、生徒に説明できない部分が多い点や、情報機器が必要な点からも、「統計」分野は独自に科目を設けたり、他教科に取り入れた方が良いのではないだろうか。

4．参考資料

「ガウス関数」（Wikipedia）

「シュワルツの不等式とそのエレガントな証明」（Web 版「高校数学の美しい物語」）

（第 115 回数学教育実践研究会にて発表）