

高校数学の教科書に書いていないだろう統計学の話

～なぜデータの変換をするの？～

令和6年11月25日

北海道網走桂陽高等学校

教諭 阿部 卓朗

1. はじめに

昨年の発表後、私の専門が統計学だとバレてしまい、今年のネタは絶対に統計学に関することにしよう、と思ったことが拙稿を書くきっかけとなりました。30代を北海道でずっと過ごしていますが、統計学を専攻していた数学教員は少ないと思います。そこで、データを見て計算をしたり、処理したりするときに「それを何のためにするのか?」「統計的処理の背景にあるものは何か」という“問題を解くことの裏側”に焦点を当てて、統計学に興味を持っていただければと思い、拙稿を執筆しました。

一口に統計的処理といっても、様々な処理がありますが、今回はその中でも最もメジャーなものの一つである、データの変換について、具体例を交えながら考察したいと思います。

2. そもそもデータの変換とは? ～線形変換について～

データの変換は大きく2種類に分けられる。線形変換と非線形変換を次のように定義しよう。

一つ目は線形変換といい、1次関数を用いた変換である。二つ目は非線形変換といい、線形変換以外の変換のことを指す。

この定義に従い、一つの事例を考えてみることにする。

次のようなアンケート調査を行ったとする

Q1. あなたはサウナに入ることが好きですか? ①～⑤から最もあてはまるものを一つ選んでください。

① とても好き ② まあまあ好き ③ どちらともいえない

④ あまり好きではない ⑤ 好きではない

このアンケートの回答をデータとして入力するとき、各段階に5～1という数値を割り当てることとする。(とても好きが5, 好きではないが1というように)

ここで、吉田(2004)が行っているデータの変換のように、5~1の数値を割り当てた、もとの尺度上での値を X 、新しく+2~-2の数値を割り当てることにした尺度上での値を X' とすると、各段階に割り当てる数値を上記のように変えることは、すべてのデータに

$$X' = X - 3$$

という1次関数を用いた線形変換を行っているということと同義である。

もう一つ具体例を示そう。

cm単位で表されているサケの体長についてのデータ(X)を、mm単位の数値(X')で表し直すという場合は

$$X' = 10X$$

という1次関数を用いた線形変換を行っていることになる。

このように、一次関数を用いたデータの線形変換を行うことによるメリットは当然存在する。では、次にどのようなメリットがあるのかを考察する。

最初に考えられることは、データ処理が(数値的に)しやすくなるということであろう。例えば、1.7mという数値よりも170cmの方が(人にもよるかもしれないが)データ処理がしやすい。

この他にメリットは考えられるだろうか？

たとえば、データ処理がしやすいように数値を変換したとしても、統計量が変わってしまったのでは、結局のところ、データの解釈に手間取ってしまう。

3. データを線形変換することによるメリット

まず思いつくことは、先に述べたように、データ処理が(数値的に)しやすいことが挙げられる。他にはあるだろうか？単に、計算しやすいうというメリットだけならば、線形変換のメリットはほとんどない。

そこで、統計量に着目して、線形変換によるメリットを考えてみることにする。

以下のデータを用いて、考察する。なお、Mは中央値、SDは標準偏差を示す。

Table 1 線形変換による平均値と標準偏差の変化

	1	2	3	4	5	6	7	M	SD
X	2	5	3	4	3	5	3	3	1
X+3	5	8	6	7	6	8	6	6	1
2X	4	10	6	8	6	10	6	6	1
2X-2	2	8	4	6	4	8	4	4	1

Table 1 を見ると、例示した線形変換を行っても、SD の値に変化がないことが分かる。(もちろん、線形変換をすれば、M の値は変化するが) このことは何を意味しているのだろうか？ 標準偏差が変わらないということは、度数分布を示すグラフの概形が変わらないということを示す。つまり、データ処理をしやすくするために線形変換を施したとしても、母集団の(データの)特徴をつかむことができるという大きなメリットがある。

他にもメリットはあるだろうが、今回はここまで述べるにとどめておく。

4. さいごに

数理統計学や統計教育学、情報工学をしていると、たびたびデータと情報を混同して使われている学生を見かける(大学院生も、いわずもがな学部生も)。だが、両者には大きな違いがある。これを高校教育の段階から生徒に理解してほしいと思い、蛇足ながら最後に付け足させていただく。

簡単に言えば、データとは数字の羅列である。つまり、それ自体は意味をもたない。対して、データは意味づけをされた数値である。言い換えれば、解釈されて意味を与えられたデータが情報である。

例を挙げてみると、100, 200, 10000 はデータである。これ自体は何の意味も持たない数字である。だが、これに 100“円”, 200“円”, 10000“円”という単位(という名の設定)をつけると、“安い”, “高い”という評価をつけることができる。例えば、1 バック 10000 円の牛乳は高い。このように、データが何らかの意味づけをされたものを情報と呼ぶ。

これは価格だけに留まらない。点数、気温などあらゆる数量が情報となりうる。

目の前にある数量や数値がデータなのか情報なのか、考えながら日々を過ごしてみるのもまた面白いのではないだろうか。数理社会心理学とも繋がるが、我々がどれだけ多くの情報にさらされ、それらを解釈する必要に迫られているかが分かるだろう。

【参考・引用文献】

俣野・河野ほか 2023 数学 I Standard 東京書籍。

吉田寿夫 2004 本当にわかりやすいすごく大切なことが書いてあるごく初歩の統計の本 北大路書房。