モーメントと代表値

鬼頭幸助

2020 年 5 月 31 日

実数のデータ $x_1, \ldots, x_n$ の代表値として有名なのは、平均、中央値、最頻値の 3 種類。これらの統一的な理解と、これらを「データを代表する値」として採用することの正当化について。

主張. 平均、中央値、最頻値は、それぞれ、2 次、1 次、0 次の中心化モーメントを最小化する値である。

嬉しいポイント.

1. 平均、中央値、最頻値という、一見全然別物だが、横並びで出てきがちな概念を統一的に理解できる。
2. 外れ値に対する強さ (頑健性、robustness) の違いを、「重みの付け方」という観点で理解できる。
   モーメントの $\sum$ の中を損失関数だと思うと、次数が高いものほど、外れ値に大きなペナルティを課していることが分かる。逆に、次数が低いものほど、外れ値を軽視もしくは無視するように働く。これが、「頑健さ」を生んでいる。

定義. 実数データ $x_1, \ldots, x_n$ と $c \in \mathcal{R}$ 及び $p > 0$ に対して、点 c における p 次中心化モーメントとは、以下で定まる実数 $\mu_p(c)$。

$$\mu_p(c) = \sum_{i=1}^{n} |x_i - c|^p$$

また、点 c における 0 次中心化モーメント $\mu_0(c)$ を以下によって定める。

$$\mu_0(c) = \sum_{i=1}^{n} 1 - \delta_{x_i, c}$$

ただし、$\delta_{i,j}$ は Kronecker のデルタである。これは、$p > 0$ のときの定義において、$p \to 0$ としたときの極限である。

定理. 平均 $\frac{1}{n}\sum_{i=1}^{n} x_i$ は、2 次中心化モーメントを最小化する。すなわち、

$$\operatorname*{argmin}_{c} \mu_2(c) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Proof. 2 次中心化モーメントを、中心の値の関数とみると、微分可能なので、微分すればよい。

$$\mu_2(c) = \sum_{i=1}^{n} (x_i - c)^2$$

となるが、これは c に関する 2 次式で、c^2 の係数は $n > 0$ なので、最小値を持つ。これを最小化する点を $\bar{x}$ とすると、

$$\frac{d}{dc}\mu_2(c) = \sum_{1=1}^{n} 2(x_i - \bar{x}) = 0$$

となるので、$\bar{x}$ について解いて、

$$\bar{x} = \sum_{i=1}^{n} x_i$$

を得る。 □

定理. 中央値 $\mathrm{median}(\{x_i \mid i = 1, \ldots, n\})$ は、1 次中心化モーメントを最小化する。すなわち、

$$\begin{aligned}\operatorname*{argmin}_c \mu_1(c) &= \mathrm{median}(\{x_i \mid i = 1, \ldots, n\}) \\ &= \begin{cases} x_{(\frac{n+1}{2})} & (n \text{ が奇数のとき}) \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & (n \text{ が偶数のとき}) \end{cases}\end{aligned}$$

となる。ただし、$x_{(i)}$ は、$x_1, \ldots, x_n$ を昇順に並べた時の i 番目の値とする。

Proof. $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n$ としてよい。

$x_k \leq c \leq x_{k+1}$ のときを考える。このとき、

$$\begin{aligned}\mu_1(c) &= \sum_{i=1}^{n} |x_i - c| \\ &= \sum_{i=1}^{k} (c - x_i) + \sum_{i=k+1}^{n} (x_i - c) \\ &= (2k - n)c - \sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i\end{aligned}$$

となる。よって、

$$\operatorname*{argmin}_{x_k \leq c \leq x_{k+1}} = \begin{cases} x_k & (2k - n > 0 \text{ のとき}) \\ x_{k+1} & (2k - n < 0 \text{ のとき}) \\ \text{任意の } c & (2k - n = 0 \text{ のとき}) \end{cases}$$

と分かる。よって、n が奇数のとき、$2k < n$ ならば、

$$\mu_1(c) \geq \mu_1(x_{k+1}) \geq \mu_1(x_{k+1}) \geq \cdots \geq \mu_1(x_{\frac{n+1}{2}})$$

となり、$2k > n$ ならば、

$$\mu_1(c) \geq \mu_1(x_k) \geq \mu_1(x_{k-1}) \geq \cdots \geq \mu_1(x_{\frac{n+1}{2}})$$

となる。よって、n が奇数のとき、

$$\operatorname*{argmin}_c \mu_1(c) = x_{\frac{n+1}{2}} = \mathrm{median}(\{x_i \mid i = 1, \ldots, n\})$$

と分かる。また、n が偶数のとき、奇数のときと同様にして、

$$\begin{cases} \mu_1(c) \geq \mu_1(x_{\frac{n}{2}}) & (2k < n \text{ のとき}) \\ \mu_1(c) \geq \mu_1(x_{\frac{n}{2}+1}) & (2k > n \text{ のとき}) \end{cases}$$

と分かり、$2k = n$ のとき、$\mu_1(c)$ は定数になるので、

$$\operatorname*{argmin}_c \mu_1(c) = (x_{\frac{n}{2}} \leq c \leq x_{\frac{n}{2}+1} \text{を満たす任意の値})$$

となるが、もちろん中央値 $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$ はこの区間に含まれる。

□

定理. 最頻値 $\mathrm{mode}(\{x_i \mid i = 1, \ldots, n\})$ は、0 次中心化モーメントを最小化する。すなわち、

$$\begin{aligned}\operatorname*{argmin}_c \mu_0(c) &= \mathrm{mode}(\{x_i \mid i = 1, \ldots, n\}) \\ &= \operatorname*{argmax}_c \#\{i \mid x_i = c\}\end{aligned}$$

ただし、$\#$ は、集合の要素の数を表す。

Proof.

$$\begin{aligned}\mu_0(c) &= \sum_{i=1}^n 1 - \delta_{x_i, c} \\ &= \#\{i \mid x_i \neq c\}\end{aligned}$$

となるので、

$$\begin{aligned}\operatorname*{argmin}_c \mu_0(c) &= \operatorname*{argmin}_c \#\{i \mid x_i \neq c\} \\ &= \operatorname*{argmax}_c \#\{i \mid x_i = c\} \\ &= \mathrm{mode}(\{x_i \mid i = 1, \ldots, n\})\end{aligned}$$

と分かる。 □