平成 23 年 3 月 18 日

メジアンと平均の関係について

新潟工科大学 情報電子工学科 竹野茂治

1 はじめに

先日、ある統計データとしてメジアンと平均を比較したものを見る機会があった。私は統計のことは良く知らないが、なんとなく珍しく感じた。

そこで、少しメジアンと平均の関係について考えてみたので、ここに簡単にまとめておく。

2 設定

メジアンは [1] にあるように中央値を指す。本稿では度数分布化されているデータを考察することとし、そのヒストグラムを考える (図 1)。階級幅を Δ、各階級の代表値を

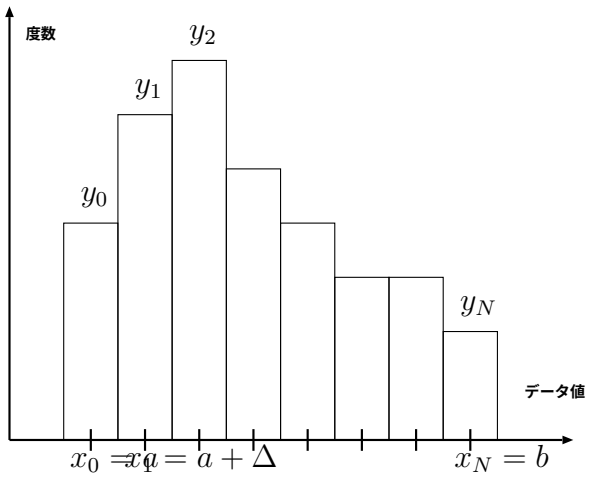

図 1: 考えるヒストグラム

$x_j = a + j\Delta$ $(j = 0, 1, 2, \ldots, N)$ とし、各階級の度数を y_j とする。

階級幅 Δ が $b-a$ $(b=a+N\Delta)$ に比べてかなり小さいときは、ヒストグラムは近似的に連続的なグラフ $y=f(x)$ と見ることもできる (図 2)。ただし、総量が決まってい

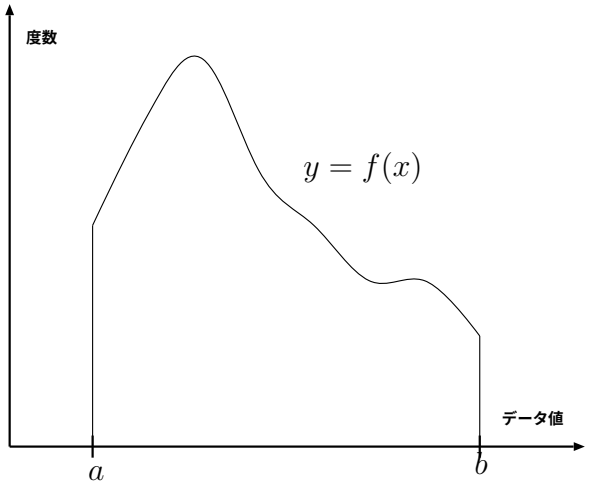

図 2: 連続的なヒストグラム

る現実のデータでは、単に階級幅 Δ を小さくするだけではこのような連続的なヒストグラムにはならず、むしろ階級幅を狭めることで各階級の度数が小さくなりヒストグラムの高さが下がっていって、逆に傾向のわかりにくいものになってしまうことに注意しなければならない (図 3)。

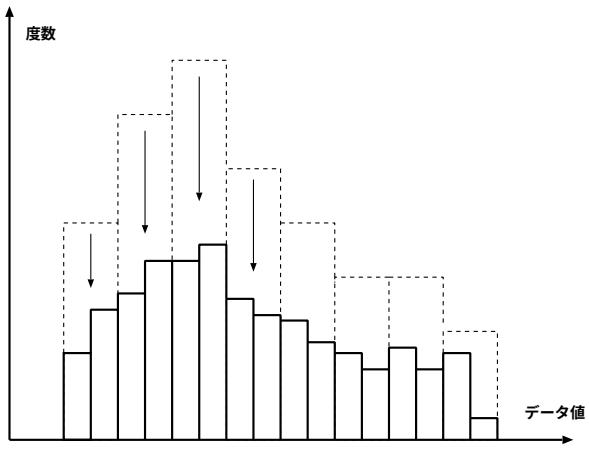

図 3: 階級幅を狭めた場合

3 メジアンと平均の表現

本節では 2 節の設定の元、平均値 M_n とメジアン M_e を式で表してみる。

各階級のデータの y_j 個のデータの値は、代表値 x_j の値を持つと考えれば、平均値 M_n は以下の式で表される。

$$M_n = \sum_{j=0}^{N} x_j y_j \Big/ \sum_{j=0}^{N} y_j \tag{1}$$

分母はデータの総量を表している。

これを、2 節の最後に述べた、連続的な $y = f(x)$ で表されるヒストグラムだと考えると、和は $f(x)$ に対する積分で表現され、

$$\sum_{j=0}^{N} y_j \Delta \approx \int_a^b f(x)dx \tag{2}$$

$$\sum_{j=0}^{N} x_j y_j \Delta \approx \int_a^b x f(x)dx \tag{3}$$

と近似されることになる。ここで、ヒストグラムの面積は、度数に底辺の Δ をかけたものになるので、左辺には Δ がつくことになる。このように考えると、結局 $y = f(x)$ で考えた平均 M_n は、

$$M_n = \int_a^b x f(x)dx \Big/ \int_a^b f(x)dx \tag{4}$$

となる。この (4) は、良く知られているように、

> 「M_n は、$y = f(x)$ の下の $a \leq x \leq b$ の範囲を板のように考えたときの重心の x 座標に等しい」

ということを意味している。

一方、メジアン M_e は中央値なので、$Me = k$ とすると

$$\sum_{j=0}^{k-1} y_j \approx \sum_{j=k+1}^{N} y_j \tag{5}$$

であることになる。より厳密に言えば、

$$\left|\sum_{j=0}^{k-1} y_j - \sum_{j=k+1}^{N} y_j\right| \leq y_k \tag{6}$$

となる。

これも連続的な $y = f(x)$ で近似して考えると、(2) より (6) は

$$\left|\int_a^{M_e} f(x)dx - \int_{M_e}^{b} f(x)dx\right| \leq f(M_e)\Delta$$

と書けるが、連続的な方では $\Delta \approx 0$ と見れるので、

$$\int_a^{M_e} f(x)dx = \int_{M_e}^{b} f(x)dx \tag{7}$$

となるとみなすことができる。この (7) は、(5) に対応すると見ることもできるが、

> 「$x = M_e$ で $y = f(x)$ の下の $a \leq x \leq b$ の範囲を左右に分けると、面積が等分される」

ということを意味している。

4 メジアンと平均の関係

本節では、主に 3 節で求めた連続的な、そして幾何学的な意味を持つ表現である (4), (7) を元に、メジアンと平均の簡単な関係を考えてみる。とりあえずは、「$M_n < M_e$ のときにどのような分布になっているかを知ること」を目標とする。

もし分布が左右対称であれば、M_n も M_e もその中心に一致することは、3 節で説明した M_n, M_e の意味からすぐにわかる。

左右対称とは限らない分布で考えると、M_e は中央値であるから、$x = M_e$ で左右に分けたときにその左右の部分の面積は同じであり、その右側の一部分を切り離して、その右側の範囲を移動した分布を考えても M_e の値は変化しない (図 5)。

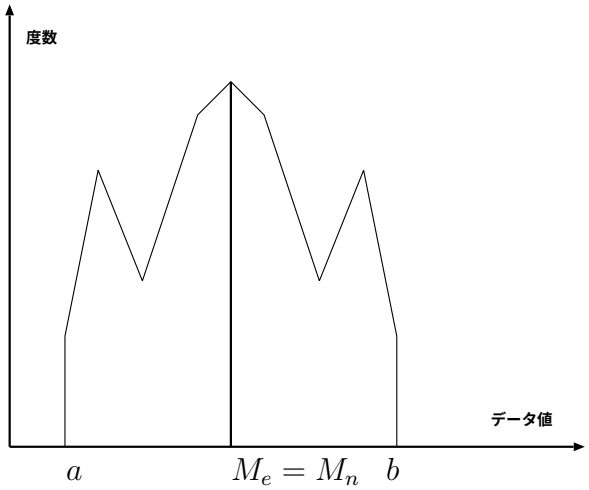

図 4: 左右対称なヒストグラム

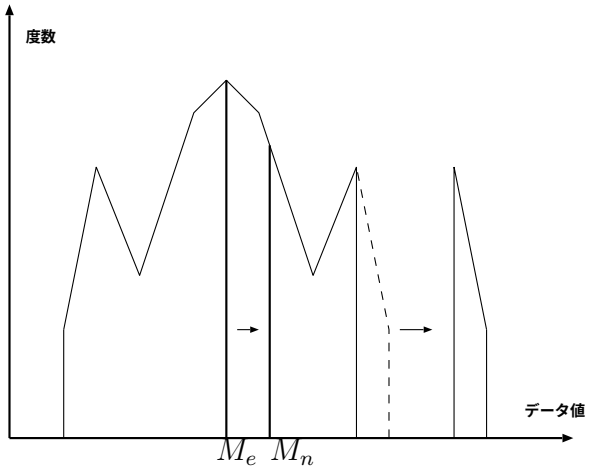

図 5: M_e の右側の一部を移動

一方、M_n の方は、図形の重心の x 座標であるから、そのような操作によって M_n の位置は変化し、一部分を右へ移動すれば重心も右へ移動するから、M_n の値も大きくなることになる。

例えば一定に増加するようなヒストグラム、すなわち $y = f(x)$ が傾きが正の直線の場合を考えてみる (図 6)。

この場合、度数は右に行くにつれて増えているので、M_e は、データ値の中央である $(a+b)/2$ よりも右にある。式で表すこともでき、ある 2 次方程式の解になることがわかるが、ここでは省略する。もちろん、重心も $x = (a+b)/2$ よりも右にあるので、M_n

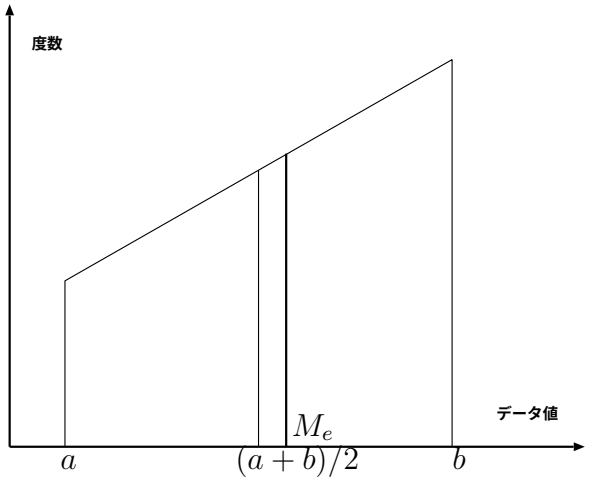

図 6: 右上がりの直線のような分布

は $(a+b)/2$ よりも大きいことがわかる。

一方、この図の M_e より右の部分を、面積を変えずに左右対称になるように変形しても、M_e の位置は変わらない (図 7)。

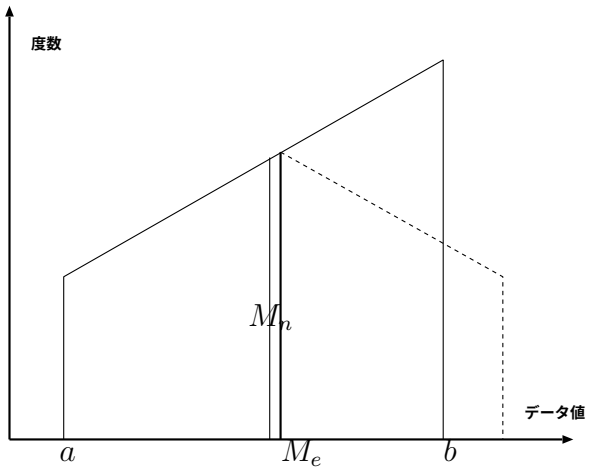

図 7: 左右対称に変形

しかし、元の右上がりの直線図形と、この左右対称な図形では、元の図形の方が移動した部分は M_e の近くにあるので、図形の重心は元の図形の方が左側にある。左右対称な図形では重心は $x = M_e$ 上にあるので、よって元の図形の重心は $x = M_e$ よりも左側にあることになるから、結局、右上がりの直線の分布に対しては、以下の不等式

が成り立つことになる。

$$\left(\frac{a+b}{2} <\right) < M_n < M_e \tag{8}$$

この例と同じ考え方を用いれば、逆に直線とは限らない分布で

$$M_n < M_e$$

である場合を考えてみると、上と同じように $x = M_e$ で左右に分けて、右側を左右対称な形に変形して考えると、その場合よりも元の分布は重心が左にあることになる。

つまり、そのような左右対称な図よりも、元の分布は $x = M_e$ の右側の部分 (の重心) が $x = M_e$ に近い方にある、ということが言える。おおざっぱに言えば、$x = M_e$ の右と左では、右の方が $x = M_e$ の近くにあり、左の方が遠くにある、ということになる。

しかし、これは「右の分散が左の分散よりも小さい」ことを表すかというと、必ずしもそうではない。上の「遠い近い」はあくまで平均、図形でいえば重心のことを意味していて、いわゆる「1 次モーメントの意味で」ということになるが、分散とは「2 次モーメント量」であるからである。1 次モーメントの大小と 2 次モーメントの大小とは一般には関係がない。

5 最後に

もしかしたら、統計の分野では平均とメジアンの値から分布に関してより詳しいことがわかるのかもしれないが、単純な考察ではこの程度のことまでしかわからなかった。

しかし、その比較である程度の傾向が、しかもある種の幾何学的な説明ができることを知ることができたのは、個人的には良かった。

参考文献

[1] 竹野茂治、「体積の相対誤差とデータ評価」、2002 年 7 月
http://takeno.iee.niit.ac.jp/~shige/math/lecture/excsiee/excsiee.html#error