標準偏差、中央値を巡る演習問題的一考察

―――偏差の絶対値和の最小化に関して―――

松浦利治*

A Consideration on Standard Deviation and Median
--- Minimization of Sum of Absolute Deviations ---

Toshiharu MATSUURA

Abstract：Let x1,x2, …,xn be n data, and F(x)=(x1－x)^2+(x2－x)^2+…+(xn－x)^2→min, then we get x=(x1+x2+…+xn)/n=μ and F(μ)=n*σ^2, where ^2:square, σ: the standard deviation. The Standard Deviation is not the mean distance from the Mean(Arithmetic Mean), one of the averages. If G(x)=|x1 – x|+|x2 – x|+…+|xn – x|→min, we can get the Median as the value of x, me, another average. What is the meaning of the minimum value of G(x), G(me)=|x1– me|+|x2 – me|+…+|xn – me|？ We will propose d, a Median Mean Deviation, and consider the meaning of d.

Keywords：Median, Standard Deviation Sum of Absolute Deviations, Median Mean Deviation

1．はじめに

標準偏差は、データのばらつきの程度を示す値であるが、「標準」とはいうものの、代表値の一つである平均値までの、普通の意味での平均的な隔たりを表すものではない。隔たりの算術平均ではない。二乗平均の平方根である。平均値との差の二乗の平均を開いたものである。

昔の手計算の時代に、なぜこのような面倒な計算をする「標準」偏差を定義したのであろうか。データのばらつきを平均的に表すのに、平均値までの隔たりの算術平均を用いるのが、私にとって自然であり、計算も簡単だと思われるにもかかわらず、である。

筆算・手計算の範囲では、負ではない隔たりすなわち絶対値の計算は簡単だが、絶対値というのは数学的に意外に扱いづらい1)。 絶対値の関数は微分できない点が存在するため、数学的に扱いづらいのである。

実は、算術平均値と分散は、偏差（誤差）の二乗和を最小にする値とその最小値に関係することに気がついた。即ち偏差の二乗和という評価関数が重要な役割を果たすことに気がついた。

そこで、評価関数として偏差（誤差）の絶対値和をとることにし、これを最小にする値とその最小値はどうなるだろうかと考えるにいたり、求めてみることにした。

そこで次の２つの演習問題を考えてみる。さらにその最小値について考察を加えてみる。

(2001年12月11日 受理)

* 宇部工業高等専門学校経営情報学科

2．演習問題1

x1,x2, …,xn を n 個のデータ、F(x)を次の評価関数とするとき、F(x)を最小とする x の値とその最小値を求めよ。ここで^2 は二乗を表す。

F(x)=(x1－x)^2+(x2－x)^2+…+(xn－x)^2

＜解＞

よく知られているように、微分を用いた簡単な計算により答は簡単に求まる。

dF(x)/dx=－2*(x1－x) －2*(x2－x) －…－2*(xn－x)=0

ゆえに x=(x1+x2+…+xn)/n=μ (算術平均値)のとき F(x)は最小となり、その最小値は

F(μ)=(x1－μ)^2+(x2－μ)^2+ …+(xn－μ)^2=n*σ^2

ここで*は乗算、^は指数演算を表す。したがってσ^2 は分散、σ は標準偏差を表す。

すなわち誤差の二乗和を最小とするのは平均値(算術平均値)であり、その最小値は分散の n 倍である、ということである。分散とは、各データと算術平均値との偏差の 2 乗の算術平均値（偏差の 2 乗平均）である。

3．演習問題2

x1,x2, …,xn を n 個のデータ、G(x)を次の評価関数とするとき、G(x)を最小とする x の値とその最小値を求めよ。ここで| |は絶対値を表す。ただし問題を簡単にするため、ここでは n は偶数とする。(なお n が奇数の場合は付録で述べることにする。)

G(x)=|x1－x|+|x2－x|+…+|xn－x|

＜解＞

x1,x2, …,xn を昇順（小さいものから大きなものへ）に並べ替え、その結果を、x(1),x(2),…,x(n) と表すと（即ち、x(i) <= x(j), i<j)、G(x)は x=中央値のときに最小になることが知られている 2) 。(なお中央値とは x(1),x(2),…,x(n)の中央の値であり、n=2*h+1 即ち奇数のときは x(h+1)であり、n=2*h 即ち偶数のときは中央が一つに決まらないので(x(h)+x(h+1))/2 とする。) ゆえに

x(k)=<x<=x(k+1)のとき

G(x)=－(x(1)－x)－(x(2)－x)－…－(x(k)－x)
+(x(k+1)－x) +(x(k+2)－x)+ …+(x(n)－x)
={k*x－(x(1)+x(2) +…+x(k))}
+ {(x(k+1) +x(k+2)+…+x(n))－(n－k)*x}
={(x(k+1) +x(k+2)+…+x(n))
－(x(1)+x(2) +…+x(k))}
+ (2*k－n)*x

$$=\left(\sum_{i=k+1}^{n} x(i) - \sum_{i=1}^{k} x(i)\right) + (2*k-n)*x$$

よって G(x)は連続で、k=1,2,3,…とふえていくとき、負のきつい勾配から次第に緩やかになり、正に転じて次第にきつい勾配になる。x の係数(2*k–n)が負から正に反転するところで最小となる。

n は偶数であるから n=2*h とおくと、k=h のとき即ち x(h)=<x<=x(h+1)のとき G(x)は最小となりその最小値は

$$G(x) = \sum_{i=h+1}^{n} x(i) - \sum_{i=1}^{h} x(i)$$

ただし x(h)<=x<=x(h+1), n=2*h

ここで x の代表値として、範囲の区間の算術平均値(x(h)+x(h+1))/2 をとるものとし、それを中央値 me とする。

4．考察： G(x)の最小値の解釈

G(x)の最小値

$$G(me)=|x1-me|+|x2-me|+\cdots+|xn-me|$$

$$= \sum_{i=h+1}^{n} x(i) - \sum_{i=1}^{h} x(i)$$

について考察する。

G(x)の最小値はデータのバラツキを表す一つのメジャーと考えて、これを n*d とおく。d は各データと中央値 me との距離の算術平均値を表す（中央値平均偏差とでも呼ぶのがふさわしい）。

即ち $|x1-me|+|x2-me|+\cdots+|xn-me|=n*d$

$$\sum_{i=h+1}^{n} x(i) - \sum_{i=1}^{h} x(i) = n*d \qquad (4.1)$$

n=2*h であるから、(4.1)の両辺を n/2 で割ると

$$\sum_{i=n/2+1}^{n} x(i)/(n/2) - \sum_{i=1}^{n/2} x(i)/(n/2) = 2*d \qquad (4.2)$$

(4.2)の左辺第一項は、n 個のデータを小さいものから大きなものへと並べて 2 グループに分けたとき、上位のグループの算術平均値を表しているので、これを上位平均値(Upper Mean)と名づけ、UMと表すことにする。同様に(4.2)の左辺第二項は、下位のグループの算術平均値を表しているので、これを下位平均値(Lower Mean)と名づけ、LMと表すことにする。よって

$$UM-LM=2*d$$

$$\therefore\ d=(UM-LM)/2 \qquad (4.3)$$

ところで、(4.1)より

$$n*d = \sum_{i=h+1}^{n} x(i) - \sum_{i=1}^{h} x(i)$$

$$= \sum_{i=h+1}^{n} x(i) + \left(\sum_{i=1}^{h} x(i) - \sum_{i=1}^{h} x(i)\right) - \sum_{i=1}^{h} x(i)$$

$$= \left(\sum_{i=h+1}^{n} x(i) + \sum_{i=1}^{h} x(i)\right) - \sum_{i=1}^{h} x(i) - \sum_{i=1}^{h} x(i)$$

$$= \sum_{i=1}^{n} x(i) - 2^{*} \sum_{i=1}^{n/2} x(i)$$

$$\therefore \quad d = \frac{\sum_{i=1}^{n} x(i)}{n} - \frac{\sum_{i=1}^{n/2} x(i)}{(n/2)}$$

$$\therefore \quad d = \mu - LM \tag{4.4}$$

ここでμはn個のデータの算術平均値Mean、即ち全体の算術平均値 Total Mean (このような言葉があるかわからないが)を表す。

(4.3)と (4.4)とから

$$\mu = (UM + LM)/2 \tag{4.5}$$

さらに(4.3)と (4.5)とから

$$UM = \mu + d \tag{4.6}$$

$$LM = \mu - d \tag{4.7}$$

(4.3)、 (4.5)、 (4.6) 、(4.7)は何を意味するのであろうか。

なるほど絶対値の計算がわずらわしいことがよくわかった。

5．おわりに

（1） 既知のことがらの整理

評価関数	偏差の二乗和 Σ(xi−x)^2→min	偏差の絶対値和 Σ|xi−x|→min
最小を与えるx値	算術平均値	中央値 me
最小値	分散のn倍　n*σ^2	中央値平均偏差のn倍 n*d（提案）

（2）わかったこと

各データと中央値との距離の算術平均値 d（ただしデータ個数が奇数のときは中央値となるデータを除いて）は、中央値より上位のグループの算術平均値と中央値より下位のグループの算術平均値との差の1／2である。

（3）今後の課題

1）中央値と、その概念の拡張であるという50%分位点との関係を明確にすること

2）非対称の分布では、算術平均値より中央値が代表値としてより適切である3)とのことなので、中央値を中心にした観点から統計学を眺めて見ること

引用文献

1) 今野紀雄「図解雑学　統計」ナツメ社、p.64、1999年12月

2) 林　周二「統計学講義」丸善、p.22、昭和38年12月

3) 東京大学教養学部統計学教室編「統計学入門」東京大学出版会、p.32、1991年7月

付録――演習問題2でnが奇数のときおよびその考察

n=2*h+1とおくと、G(x)の勾配即ちxの係数2*k－nは

k=hのとき、2*k－n＝2*h－(2*h+1)= －1<0

k=h+1のとき、2*k－n＝2*(h+1)－(2*h+1)= 1>0

よってk=h+1のとき即ちx=x(h+1)のときG(x)は最小となり、その最小値は

$$G(x(h+1)) = \left(\sum_{i=h+2}^{n} x(i) - \sum_{i=1}^{h+1} x(i)\right) + x(h+1)$$

なおここでx(h+1)は中央値である。

さらに

$$G(x(h+1)) = \left(\sum_{i=h+2}^{n} x(i) - \sum_{i=1}^{h} x(i) - x(h+1)\right) + x(h+1)$$

$$= \sum_{i=h+2}^{n} x(i) - \sum_{i=1}^{h} x(i)$$

$$= \sum_{i=(h+1)+1}^{(h+1)+h} x(i) - \sum_{i=1}^{h} x(i)$$

これを(n－1)*d'とおくと、

$$\sum_{i=(h+1)+1}^{(h+1)+h} x(i) - \sum_{i=1}^{h} x(i) = (n-1)*d' \qquad (A.1)$$

両辺を(n－1)/2で割ると、

$$\sum_{i=(h+1)+1}^{(h+1)+h} x(i)/((n-1)/2) - \sum_{i=1}^{h} x(i)/((n-1)/2) = 2*d' \qquad (A.2)$$

n=2*h+1より(n－1)/2=hであるから

$$\sum_{i=(h+1)+1}^{(h+1)+h} x(i)/h - \sum_{i=1}^{h} x(i)/h = 2*d'$$

nが偶数のときと同様に、左辺第一項をUM'、第二項をLM'とおくと

$$UM' - LM' = 2*d'$$

$$\therefore \quad d' = (UM' - LM')/2 \qquad (A.3)$$

(A.1)より

$$(n-1)*d' = \sum_{i=(h+1)+1}^{(h+1)+h} x(i) - \sum_{i=1}^{h} x(i)$$

$$= \sum_{i=(h+1)+1}^{(h+1)+h} x(i) + \sum_{i=1}^{h} x(i) - 2*\sum_{i=1}^{h} x(i)$$

$$\therefore \quad d' = \left(\sum_{i=(h+1)+1}^{(h+1)+h} x(i) + \sum_{i=1}^{h} x(i) \right) / (n-1) - \sum_{i=1}^{h} x(i) / ((n-1)/2)$$

$$\therefore \quad d' = \mu' - LM' \tag{A.4}$$

ここでμ'は中央値x(h+1)を除いた(n－1)個のデータの算術平均値を表す。即ち

$$(n-1)^* \ \mu' + x(h+1) = n^* \mu \tag{A.5}$$

ここでμはn個の全データの算術平均値を表す。

(A.3)と (A.4)とから

$$\mu' = (UM' + LM')/2 \tag{A.6}$$

さらに(A.3)と (A.6)とから

$$UM' = \mu' + d' \tag{A.7}$$

$$LM' = \mu' - d' \tag{A.8}$$

(A.3)、 (A.6)、 (A.7) 、(A.8)は何を意味するのであろうか。