中央値絶対偏差を用いた分散の関係式の導出(別法)

－全二乗和の分解を用いた演習問題として－

松浦利治*

Deduction of Variances Relation Expression with Median Absolute Deviation

—— Square Sum Analysis Solution ——

Toshiharu MATSUURA

Abstract: Let n data $x_1, x_2, \cdots, x_n$ be $0 < x_1 \le x_2 \le \cdots \le x_{\frac{n}{2}} \le x_{\frac{n}{2}+1} \le \cdots \le x_n$ (n is even). Minimize the function $g(x) = |x_1 - x| + |x_2 - x| + \cdots + |x_n - x|$, then we get $x = med = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$, median, Minimum value $g(med) = | x_1 - med | + | x_2 - med | + \cdots + | x_n - med | = n \cdot d$. We call d Median Absolute Deviation. Let μ and σ^2 be the mean and variance, respectively, of n data $x_1, x_2, \cdots, x_n$. Let μ_L and σ_L^2 be the mean and variance, respectively, of lower group, n/2 data $x_1, x_2, \cdots, x_{\frac{n}{2}}$. Let μ_U and σ_U^2 be the mean and variance, respectively, of upper group, n/2 data $x_{\frac{n}{2}+1}, \cdots, x_n$.

Then we know we get $\mu_L = \mu - d, \mu_U = \mu + d$, $\sigma^2 = \frac{\sigma_L^2 + \sigma_U^2}{2} + d^2$ .In this paper, we deduce this σ^2 expression by analysis of square sum of n data $x_1, x_2, \cdots, x_n$.

Key words: Median Absolute Deviation, median, variance, mean

１．はじめに

データの代表値と散布度として、普通、平均と分散（あるいはその正の平方根である標準偏差）が用いられるが、これらは２次のノルムの世界のことであると考えると、１次のノルムの世界では何がいえるか、というのがここでの関心事である。

n 個のデータを $x_1, x_2, \cdots, x_n$ とする。関数 $f(x) = (x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2$ を最小にする x の値が平均 μ、その最小値が $n \cdot \sigma^2$ （σ^2 ：分散）である。

関数 $g(x) = |x_1 - x| + |x_2 - x| + \cdots + |x_n - x|$ を最小にす

（2008 年 11 月 28 日受理）

* 宇部工業高等専門学校　一般科

る x の値は中央値（メディアン median）であることは知られている。そのときの最小値を $n\cdot d$ とする。中央値 med より小さい値（より正確には大きくない値）のグループの平均と分散をそれぞれ μ_L 、$\sigma_L{}^2$、中央値 med より大きい値（より正確には小さくない値）のグループの平均と分散をそれぞれ μ_U 、$\sigma_U{}^2$ とすると、$\mu_L=\mu-d,\mu_U=\mu+d$ ，$\mu=\mu_L+d$ ，$\mu=\mu_U-d$ ，$\mu=\frac{\mu_L+\mu_U}{2}$ となることを、以前の研究報告[1)]で報告し、さらに、$\sigma^2=\frac{\sigma_L^2+\sigma_U^2}{2}+d^2$ となることを、前回の研究報告[2)]で報告した。

今回は、前回と異なったやり方でこの分散の関係式を導く。

２．中央値絶対偏差を用いた分散の表現の導出

２．１　前提条件

前回の研究報告[2)]でも述べたことであるが、本質的でない煩雑な議論を避けるため、以下の諸条件を設ける。

n を偶数とする。

n 個のデータを $x_1,x_2,\cdots,x_{\frac{n}{2}},x_{\frac{n}{2}+1},\cdots,x_n$ とし、さらに

$$0<x_1\le x_2\le\cdots\le x_{\frac{n}{2}}\le x_{\frac{n}{2}+1}\le\cdots\le x_n$$

とする。

関数 $g(x)$ を次のようにする。

$g(x)=|x_1-x|+|x_2-x|+\cdots+|x_n-x|$

med を中央値（メディアン）とする。n は偶数だから

$$med=\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$$

上記関数 $g(x)$ を最小にする x の値はメディアンであることが知られている。よって

$g(x)$ の最小値を次のようにおく。

$$g(med)=|x_1-med|+|x_2-med|+\cdots+|x_n-med|$$
$$=n\cdot d$$

d を中央値絶対偏差(MediAD(Median Absolute Deviation))ということにしている。

２．２　中央値絶対偏差を用いた平均と分散の表現

そうすると以上のような条件のもとで、既に報告したことではあるが、以下の議論のために、まとめてここに再掲する。

平均 μ を　$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i$

分散 σ^2 を　$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$

下位グループについて、

平均（lower mean とでも名付けるべきか） μ_L を

$$\mu_L=\frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i$$

分散 $\sigma_L{}^2$ を　$\sigma_L{}^2=\frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}(x_i-\mu_L)^2$

上位グループについて、

平均（upper mean とでも名付けるべきか） μ_U を

$$\mu_U=\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}x_i$$

分散 $\sigma_U{}^2$ を　$\sigma_U{}^2=\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}(x_i-\mu_U)^2$

とすると、

$$\mu=\mu_L+d$$
$$\mu=\mu_U-d$$
$$\mu=\frac{\mu_L+\mu_U}{2}$$
$$\mu^2=\frac{\mu_L^2+\mu_U^2}{2}-d^2$$
$$\sigma^2=\frac{\sigma_L^2+\sigma_U^2}{2}+d^2$$

が成り立つ。

２．３　全二乗和の分解を用いた分散の関係式の導出

ここでの目的は、全二乗和を分解するというやり方で、分散の関係式 $\sigma^2 = \frac{\sigma_L^2 + \sigma_U^2}{2} + d^2$ を導き出すことである。

表記を見やすくするために、下位グループのデータの個数を n_L 、上位グループのデータの個数を n_U とすると、

$$n_L = n_U = \frac{n}{2}$$

データの総和を考える。

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{\frac{n}{2}} x_i + \sum_{i=\frac{n}{2}+1}^{n} x_i$$

ところで、

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ より } \sum_{i=1}^{n} x_i = n\mu$$

$$\mu_L = \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}} x_i \text{ より } \sum_{i=1}^{\frac{n}{2}} x_i = \frac{n}{2}\mu_L = n_L\mu_L$$

$$\mu_U = \frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n} x_i \text{ より } \sum_{i=\frac{n}{2}+1}^{n} x_i = \frac{n}{2}\mu_U = n_U\mu_U$$

よって　$n\mu = n_L\mu_L + n_U\mu_U$

次にデータの２乗の和(全２乗和)を考える。

$$\sum_{i=1}^{n} x^2{}_i = \sum_{i=1}^{\frac{n}{2}} x^2{}_i + \sum_{i=\frac{n}{2}+1}^{n} x^2{}_i$$

ところで、

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \text{ より}$$

$$n\sigma^2 = \sum_{i=1}^{n}(x_i - \mu)^2$$

$$= \sum_{i=1}^{n} x^2{}_i - 2\mu\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\mu^2$$

$$= \sum_{i=1}^{n} x^2{}_i - 2n\mu^2 + n\mu^2$$

$$= \sum_{i=1}^{n} x^2{}_i - n\mu^2$$

よって　$\sum_{i=1}^{n} x^2{}_i = n\mu^2 + n\sigma^2 = n(\mu^2 + \sigma^2)$

同様に

$$\sum_{i=1}^{\frac{n}{2}} x^2{}_i = n_L(\mu_L^2 + \sigma_L^2)$$

$$\sum_{i=\frac{n}{2}+1}^{n} x^2{}_i = n_U(\mu_U^2 + \sigma_U^2)$$

よって

$$n(\mu^2 + \sigma^2) = n_L(\mu_L^2 + \sigma_L^2) + n_U(\mu_U^2 + \sigma_U^2)$$

ここで　$n_L = n_U = \frac{n}{2}$　であるから

$$n(\mu^2 + \sigma^2) = \frac{n}{2}(\mu_L^2 + \sigma_L^2) + \frac{n}{2}(\mu_U^2 + \sigma_U^2)$$

よって

$$2(\mu^2 + \sigma^2) = (\mu_L^2 + \sigma_L^2) + (\mu_U^2 + \sigma_U^2)$$

ところで、

$\mu = \mu_L + d$　より　$\mu_L = \mu - d$

$\mu = \mu_U - d$　より　$\mu_U = \mu + d$

であるから、これらを代入すると

$$2(\mu^2 + \sigma^2) = (\mu - d)^2 + \sigma_L^2 + (\mu + d)^2 + \sigma_U^2$$

$$= 2\mu^2 + 2d^2 + \sigma_L^2 + \sigma_U^2$$

よって　$2\sigma^2 = 2d^2 + \sigma_L^2 + \sigma_U^2$

$$\therefore \quad \sigma^2 = \frac{\sigma_L^2 + \sigma_U^2}{2} + d^2$$

３．全二乗和の分解に関する一考察

（１）私は以前に偏差の二乗和について考察したことがある[3)]が、全二乗和は何を表しているか。エネルギーであると考える。（数学の議論に物理の概念を持ち込むのはよくないかもしれないが。）　種々の各要因のエネルギーの総和が全二乗和である。これはエネルギーの加法性による。

（２）逆に全二乗和を各要因のエネルギーに分解することにより、全エネルギー＝全二乗和に対する各要因の貢献度がわかる。（全二乗和を分散＝残差平方和とした分解が分散分析である。）

（３）エネルギーの加法性に基づく評価により、設計した

ものの良し悪しがわかる。

（4）一方、機能とはエネルギー変換であるとすると、エネルギーが効率よく（無駄が少なく）変換されるのが良い設計ということになる。

（5）また、環境に関する要因の貢献度が小さいとすれば、環境が多少変動しても、機能にはあまり影響しないことを意味し、安定性が高い、頑健であることを意味する。

（6）このように、エネルギーの加法性に基づく分解、全二乗和の分解を活用することにより、技術開発、良い設計が可能になると考えられる。こういうことが、タグチメソッド[4) 5)]の基本思想であると解釈できよう。

（7）ところで、私は平均に対して、単に、すべてのデータが同じ値であったとしたときのその値という感覚を持っている。データにバラツキが全くなかったとしたらその値は？ということである。関数

$$f(x) = (x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2$$

を最小にする x の値を求めようという意識はない。

バラツキが全くないとした値と、上記評価関数を最小にするような、バラツキ最小の値とは、感覚的には別物であると受け取っている。

中央値（メディアン median）にしても、我々は関数 $g(x) = |x_1 - x| + |x_2 - x| + \cdots + |x_n - x|$ を最小にする x の値、ということを意識して、中央値を使っているわけではない。単に大きさの順番が真中のデータの値ということである。

平均がたまたま、偏差の二乗和を最小にする値であることから、偏差の二乗和を小さくすること、分解することが意味あることになるのであろう。偏差の二乗和を小さくすることから、例えば最小二乗法が提案されるのであろう。

よってデータの二乗和を分解することが意味あることになる。そして全二乗和の分解は一般的な方法となる。

4．おわりに

全二乗和を分解するというやり方で、分散の関係式 $\sigma^2 = \dfrac{\sigma_L^2 + \sigma_U^2}{2} + d^2$ を導き出した。さらに全二乗和の分解に関する一つの考え方を述べた。全二乗和の分解にはもっと深い意味がありそうであるが、今後の課題とする。

参考文献

1）松浦利治「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――　」宇部工業高等専門学校研究報告　第４８号、pp.61-66、平成１４年３月

2）松浦利治、中央値絶対偏差を用いた平均と分散の表現、宇部工業高等専門学校研究報告　第５４号、pp.87-91、平成２０年３月

3）松浦利治、偏差の二乗に関する一考察、宇部工業高等専門学校研究報告　第５０号、pp.19-20、平成１６年３月

4）田口玄一：タグチメソッド わが発想法、経済界、東京、1999 年 11 月

5）松浦利治、タグチメソッド（品質工学）の基本思想に関する一考察、宇部工業高等専門学校研究報告　第５２号、pp.89-92、平成１８年３月