中央値絶対偏差を用いた平均と分散の表現

松浦利治*

Mean and Variance Expressions using Median Absolute Deviation

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

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）のグループの平均を μ_L、中央値 med より大きい値（より正確には小さくない値）のグループの平均を μ_U 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、前の研究報告[1]で報告した。

今回は、分散がどう表されるかについて主に述べる。

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

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。

（2007年12月4日受理）

* 宇部工業高等専門学校　一般科

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))と名付けるものとする。前の研究報告[1]では、中央値平均偏差といおうとしていたが、このようにする。MAD としなかったのは、今野先生[4]がすでに平均と（2乗平均即ち分散、標準偏差でなく）絶対偏差の意味で 平均・絶対偏差 Mean-Absolute Deviation を使っているからである。

2．2　中央値絶対偏差を用いた平均の表現

そうすると以上のような条件のもとで、前の研究報告[1]で述べたことではあるが、表記法を変えたので、以下の通り再掲する。

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

下位グループの平均(lower mean とでも名付けるべきか) μ_L を

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

上位グループの平均（upper mean とでも名付けるべきか） μ_U を

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

とすると、

$$\mu = \mu_L + d$$

$$\mu = \mu_U - d$$

$$\mu = \frac{\mu_L + \mu_U}{2}$$

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

が成り立つ。

これらの導出については、補足 1 を参照されたい。

2．3　中央値絶対偏差を用いた分散の表現の導出

本報告では、さらに次のことが成り立つことを述べる。

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

下位グループの分散 ${\sigma_L}^2$ を　${\sigma_L}^2 = \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}(x_i - \mu_L)^2$

上位グループの分散 ${\sigma_U}^2$ 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、

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

が成り立つ。

以下にその導出過程を示す。

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

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

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

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

$\mu = \mu_L + d$ であるから、

右辺第１項

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

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

$$= \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}\{(x_i-\mu_L)^2-2d(x_i-\mu_L)+d^2\}$$

$$= \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}\{(x_i-\mu_L)^2-2dx_i+2d\mu_L+d^2\}$$

$$= \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}(x_i-\mu_L)^2 - 2d\cdot\frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i + (2d\mu_L+d^2)\cdot\frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}1$$

$$= \sigma_L^2 - 2d\mu_L + (2d\mu_L + d^2)$$

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

同様に、$\mu = \mu_U - d$ であるから、

右辺第２項

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

よって

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

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

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

＜補足１．$\mu = \mu_L + d$, $\mu = \mu_U - d$ の導出＞

前の研究報告ですでに述べていることではあるが、本報告では表記法を変えたので、再度述べることにする。

$$g(med) = |x_1 - med| + |x_2 - med| + \cdots + \left|x_{\frac{n}{2}} - med\right| + \left|x_{\frac{n}{2}+1} - med\right| + \cdots + |x_n - med| = n\cdot d$$

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

$med = \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$　であるから、

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

よって

$$(med - x_1) + (med - x_2) + \cdots + \left(med - x_{\frac{n}{2}}\right) + \left(x_{\frac{n}{2}+1} - med\right) + \cdots + (x_n - med) = n\cdot d$$

$$\left(x_{\frac{n}{2}+1} + x_{\frac{n}{2}+2} + \cdots x_n\right) - \frac{n}{2}\cdot med + \frac{n}{2}\cdot med - \left(x_1 + x_2 + \cdots + x_{\frac{n}{2}}\right) = n\cdot d$$

$$\therefore \sum_{i=\frac{n}{2}+1}^{n}x_i - \sum_{i=1}^{\frac{n}{2}}x_i = n\cdot d$$

両辺を $\frac{n}{2}$ で割ると

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

$$\therefore \mu_U - \mu_L = 2d$$

$$d = \frac{\mu_U - \mu_L}{2}$$

一方

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

より

$$\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}x_i + \left(\frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i - \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i\right) - \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i = 2d$$

$$\left(\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}x_i + \frac{1}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i\right) - \frac{2}{\frac{n}{2}}\sum_{i=1}^{\frac{n}{2}}x_i = 2d$$

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

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

$\therefore \mu-\mu_L=d$

$\mu=\mu_L+d$

$\mu_U-\mu_L=2d$ より $\mu_L=\mu_U-2d$

$\therefore \mu=\mu_U-d$

＜補足２．右辺第２項 $\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}(x_i-\mu)^2=\sigma_U^2+d^2$ の導出＞

$\mu=\mu_U-d$ であるから

右辺第２項 $\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}(x_i-\mu)^2$

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

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

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

$$=\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}\{(x_i-\mu_U)^2+2dx_i-2d\mu_U+d^2\}$$

$$=\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}(x_i-\mu_U)^2+2d\cdot\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}x_i+(-2d\mu_U+d^2)\cdot\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}1$$

$$=\sigma_U^2+2d\mu_U+(-2d\mu_U+d^2)$$

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

３．なぜこのようなことを考えたか

（１）　データの集まりがあるとき、代表値と散布度でその特徴を表現する。一言で言うといくらで、バラツキ（データの散らばり具合）はこの程度、という具合に。そして、普通、代表値として平均、散布度として分散（またはその正の平方根をとった標準偏差）が用いられる。代表値として平均を用いるのはまあ妥当なところであろう。データのバラツキになぜ分散を用いるのか。

バラツキとして、各データと平均との差（これを偏差という）の平均で表すことを考えてみよう。

$$E=\frac{1}{n}\{(x_1-\mu)+(x_2-\mu)+\cdots+(x_n-\mu)\},$$

ただし μ は平均で、 $\mu=\frac{1}{n}(x_1+x_2+\cdots+x_n)$

そうすると、

$$E=\frac{1}{n}\{(x_1+x_2+\cdots+x_n)-n\cdot\mu\}$$

$$=\frac{1}{n}(x_1+x_2+\cdots+x_n)-\frac{n}{n}\cdot\mu$$

$$=\mu-\mu$$

$$=0$$

となってしまい、E はバラツキ（散布度）のメジャーとはならない。なぜこうなるかというと、各偏差 $x_1-\mu, x_2-\mu, \cdots, x_n-\mu,$ には正負があって、これらが互いに打ち消しあってしまうからである。

ならば偏差の 2 乗を用いて負の値を正にすれば打ち消しあうこともないであろう。

$$\sigma^2=\frac{1}{n}\{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_n-\mu)^2\}$$

偏差の 2 乗の平均を分散 σ^2 として、これを散布度のメジャーとする。これはこれでよいとしよう。

ただし偏差の正負を打ち消しあわないようにするには、偏差の絶対値をとってもよいはずである。

$$H=\frac{1}{n}(|x_1-\mu|+|x_2-\mu|+\cdots+|x_n-\mu|)$$

これも散布度のメジャーである。平均偏差という。今野先生[4]は絶対偏差といっているようだ。

しかしこれは普通にはあまり使われない。絶対値は数学的に扱いにくいからという。だがこの理由はおかしい。手計算の範囲では、2 乗の計算より絶対値の計算の方が手間がかからない。今日のように数値計算にコンピュータを用

いるのであれば、2 乗の計算も絶対値の計算も大差ない。絶対値は微分操作がやりにくいということであろう。絶対値の中身が符号が反転するところで、微分不能になるからである。数値計算ではどうでもよいのではないか。

意味的にはどうか。絶対値を用いた平均偏差は、平均との、普通の意味での距離の平均を表しており、わかりやすい。平均との隔たりの平均を表している。一方分散は、平均との偏差の 2 乗平均を表している。日常、2 乗平均はなじみがうすい。直感的にわかりやすいわけではない。

（2）　分散は偏差の 2 乗平均だから、もとのデータの 2 乗のディメンションをもつ。cm 単位の身長の議論をしているとき、そのバラツキ（分散）はなぜ面積 cm^2 なのだ？！

（3）　（2）のようではやはり困るので、分散の正の平方根をとってそれを標準偏差とする。そうすれば（2）の例ではこのディメンションは cm となり、議論している身長のディメンションと同じになって都合がよい。

だがなぜこれが「標準」なのだ？！　平均との、普通の意味での隔たりの平均を表している平均偏差の方がわかりやすくこれが本来の標準であるべきだが、日常なじみのうすい偏差の 2 乗平均の平方根を「強引に」「標準」偏差としているのであろう。

（4）　ラプラスは偏差の絶対値を考えたが、ガウスは偏差の 2 乗で考察すべきことを提案し[5]、正規分布や最小二乗法などの成果を出した。

（5）　平均と分散（偏差の 2 乗平均）が物理系を適切に表現しているようである。エネルギーが種々の物理量の 2 乗で表される場合が多いからか[2]。

（6）　実験計画法、分散分析、タグチメソッド[3]では、2 乗和を用いている。2 乗和以外は無意味といったら、言い過ぎか。

（7）　一方、数理計画法の分野では、平均と標準偏差のモデルではなく、平均・絶対偏差モデル（MAD(Mean-Absolute Deviation)モデル）[4]が効率的であるということである。

（8）　はじめにも述べたが、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。

しかも、ある場合には、平均よりも中央値の方が実態をよく表すといわれているにもかかわらず、中央値の観点からあまり議論がなされていないように思われる。私の調査・勉強不足ならばよいが。中央値から統計現象を見てもあまり成果が期待できないとすれば、それはなぜか。

4．おわりに

中央値絶対偏差と、平均と分散との関係式を導いた。昇順に並べたデータを 2 等分割したそれぞれのグループの平均と分散と、全データの平均と分散との間に、中央値絶対偏差が介在していることがわかった。中央値絶対偏差の定義からすれば、自明に近い結論かも知れないが、上記のように明確に表現できたことは、一つの成果であると考える。

なお、得られた結果を順次下位のグループに適用していくとどうなるかは、興味あるところである。

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、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

3）田口玄一、実験計画法 下、第１８章 pp.529-550、とくに 18.2 ２乗和を用いる理由、東京、丸善、昭和５２年８月

4）今野浩、MAD モデルあれこれ、http://www.kier.kyoto-u.ac.jp/fe-tokyo/symposium_/syuppankinen/konno(11.01).pdf #search='今野浩%20MAD'

5）世界大百科事典第 2 版、「誤差」の項目、日立デジタル平凡社