高松大学紀要，46．11～19

Some Aspects Concerning the Evaluations of Sum of Squares by Range[1]

Toshihiko MENDORI[2], Yoshihisa ZINNO[3] and Hiroo FUKAISHI[4]

Abstract

In this paper, based on the study so far, we give the following consideration to the evaluation of the sum of squares by using the range R. Firstly another way is given for a proof of the Main Theorem, which describes the l.u.b. of s.s. by R. The proof by use of eigenvalues of matrices will contribute to mathematical education as a teaching material. Secondly the possibility of further investigations is suggested.

要　約

この論文では，これまでの研究を踏まえて，データのレインジ R による自乗和（$S.\,S.$）の評価について次の考察を行う。第１に，R による $S.\,S.$の上限に関する主定理の別証明を行列理論を用いて与えることであり，この方法はデータ空間の見方や固有空間の役割を示す例として数学教育にも寄与する。第２に，今後の研究発展の可能性について言及している。

Key Words ： l.u.b. of sum of squares, mathematical representation of data, matrix theory

キーワード： 自乗和の上限，　データの数学的表現，　行列論

§１. Introduction

In processing statistical data, the mean and the standard deviation（i.e., s.d.）need to be calculated as the representative values of them. The calculation of s.d. is rather complicated depending on the number of sample data. G. Snedecor（1989）studied the possibility to use the sample range in place of s.d. in §8.3. To sum up, he pointed out the usefullness of the range :

１. The value $C_n \cdot R$ is considered to be a quick and simple estimate of the population s.d. σ and a table of multiplier values C_n of the range R is computed for $2 \le n \le 20$.

---

[1]Received June 30, 2006
[2]Toshihiko MENDORI, Professor, Faculty of Human Development, Takamatsu University
Emeritus Professor of Kagawa University
[3]Yoshihisa ZINNO, Mathematics, Takamatsu Chūō Senior High School
[4]Hiroo FUKAISHI, Professor, Department of Mathematics, Faculty of Education, Kagawa University

2. The relative efficiency of the range estimate to the sample s.d. remains high in normal samples up to $n=8$ or 10, decreasing slowly but steadily as n increases.

3. Calculated from the extream values, the range estimate is very sensitive to erratic extream values (outliers). So the range gives a quick estimate of σ for distributions thought to be close to normal.

4. In controlling the quality of an industrial process, small samples of the product are taken out frequently to detect situations in which the industrial process becomes out of control. The estimate from the range computed on each sample is an easy and efficient alternate of the unbiased estimate s of σ.

5. The instructions given to the computer may contain errors. So the range estimator is also an easy check on the computation of s.

6. Samples of $n=5$ are often used, the range estimator being computed on each sample and plotted on a time chart. Often the purpose of the chart is to detect situations, as revealed by the appearance of increasing or sudden high values of the estimated σ [Raynolds et al. (1988)].

But he didn't describe much of the mathematical relation between s.s. and R. Therefore we have studied the mathematical structure of the sample variance s^2 (i.e., s.s.), and especially the relation with the range R. Namely the l.u.b. and the g.l.b. of s.s. are mathematically derived by using R, and the multiplier C_n of R is computed by Mendori et al. (2003). Starting from the problem of sample data, the same problems for the theoretical distribution of discrete and continuous types are investigated by Mendori et al. (2003a). Furthermore a relation between the classified groups of data and the whole data is studied and some illustrative examples are shown by Mendori et al. (2004).

In this paper we apply the matrix theory to another proof of the Main Theorem, and suggest the possibility of some extensions.

§ 2. Another approaches to proving the Main Theorem

In the previous paper [Mendori et al. (2003)] we gave the Main Theorem which estimates the sum of squares by using the data range. We also gave geometrical representations there. In this paper we shall give another proof of the Main Theorem by solving eigenvalue problems of the matrices induced by the quadratic form of s.s.. An application of linear algebra shown in the proof is a useful material in mathematics education.

Briefly we recall the notations. Let $X = \{x_1, x_2, \ldots, x_n\}$, $n > 1$, be a data set of n values. Define the data range R and the mean $\bar{x}$ of the data as follows :

$$R = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i ,$$

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

respectively. Each of the variance s^2 of the data and the unbiased estimate u^2 of the population variance is defined as follows :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad u^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

respectively.

Now define the function $g(x)$ of s.s. by the following :

$$g(x) = \sum_{i=1}^{n}(x_i - x)^2.$$

Assume that $x_1 \le x_2 \le \ldots \le x_n$. Then $R = x_n - x_1$.

Set a vector $\boldsymbol{x}$ as follows[5] :

$$\boldsymbol{x} = [x_i] \equiv (x_1, x_2, \ldots, x_n)^T \quad \text{where} \quad x_1 \le x_2 \le \ldots \le x_n .$$

Then we have

$$g(x) = [x_i - x]^T [x_i - x] = [x_i - x]^T \left(\frac{1}{n}E + I - \frac{1}{n}E\right)[x_i - x],$$

where I is the identity matrix and E is the $n \times n$ matrix as follows :

$$E = \begin{bmatrix} 1 & \cdots & 1 \\ & \ddots & \\ 1 & \cdots & 1 \end{bmatrix} .$$

[5]A small latin letter of a bold face is used for a column vector. The transpose of a matrix A is denoted by A^T.

For the matrices $\frac{1}{n}E$ and $I-\frac{1}{n}E$, we have

$$\left(\frac{1}{n}E\right)^2 = \frac{1}{n}E, \quad \frac{1}{n}E\left(I-\frac{1}{n}E\right) = O, \quad \left(I-\frac{1}{n}E\right)^2 = I-\frac{1}{n}E.$$

To solve the eigenvalue problem for each of matrices, tE and $I-tE$ simultaneously, we introduce the following polynomials $F_n(t)$ and $G_n(t,\ \lambda)$:

$$F_n(t) = \det{}_n(I-tE),$$
$$G_n(\lambda,\ t) = \det{}_n\{(1-\lambda)I-tE\}.$$

Then we have the following[6] :

Lemma 1. $$F_n(t) = 1-nt,$$
$$G_n(\lambda,\ t) = (1-\lambda)^{n-1}(1-\lambda-nt).$$

Proof.

$$det_n(I-tE) = \begin{vmatrix} 1-t & -t & \cdots & -t \\ -t & 1-t & \cdots & -t \\ \vdots & \vdots & \ddots & \vdots \\ -t & -t & \cdots & 1-t \end{vmatrix} = \begin{vmatrix} 1-nt & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{vmatrix}.$$

Therefore we have

$$F_n(t) = 1-nt,$$

$$G_n(\lambda,\ t) = (1-\lambda)^n F_n\left(\frac{t}{1-\lambda}\right) = (1-\lambda)^{n-1}(1-\lambda-nt). \qquad \square$$

Since $G_n(1+\lambda,\ -t)$ and $G_n(\lambda,\ t)$ are the characteristic polynomials of tE and $I-tE$, respectively, we have the following :

---

[6]We treated the determinant of the same type

$$\underbrace{\begin{vmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & n-1 \end{vmatrix}}_{n-2} = n^{n-2}F_{n-2}\left(\frac{1}{n}\right) = 2n^{n-3},$$

in the proof of Main Theorem [Mendori et al. (2003), Theorem 2.1].

Lemma 2.

1. *The eigenvalues of the matrix* tE *are* $\lambda = nt$ *and* $\lambda = 0$, *a root of multiplicity* $n-1$.
2. *The eigenvalues of the matrix* $(I - tE)$ *are* $\lambda = 1 - nt$ *and* $\lambda = 1$, *a root of multiplicity* $n-1$.

In the case of $t = \frac{1}{n}$, $G_n\left(1+\lambda,\ -\frac{1}{n}\right)$ and $G_n\left(\lambda,\ \frac{1}{n}\right)$ are the characteristic polynomials of $\frac{1}{n}E$ and $I - \frac{1}{n}E$, respectively, i.e.,

$$G_n\left(1+\lambda,\ -\frac{1}{n}\right) = (-\lambda)^{n-1}(1-\lambda),\quad G_n\left(\lambda,\ \frac{1}{n}\right) = -\lambda\,(1-\lambda)^{n-1}.$$

Then we have the following :

Lemma 3.

1. *The eigenvalues of the matrix* $\frac{1}{n}E$ *are* $\lambda = 1$ *and* $\lambda = 0$, *a root of multiplicity* $n-1$.

2. *The eigenvalues of the matrix* $I - \frac{1}{n}E$ *are* $\lambda = 0$ *and* $\lambda = 1$, *a root of multiplicity* $n-1$.

For the matrices $\frac{1}{n}E$ and $I - \frac{1}{n}E$, we have

$$\frac{1}{n}E\boldsymbol{f}_1 = \boldsymbol{f}_1,\qquad \frac{1}{n}E\boldsymbol{f}_2 = \boldsymbol{o},$$
$$\left(I - \frac{1}{n}E\right)\boldsymbol{f}_1 = \boldsymbol{o},\qquad \left(I - \frac{1}{n}E\right)\boldsymbol{f}_2 = \boldsymbol{f}_2,$$

where

$$\boldsymbol{f}_1 = \frac{1}{\sqrt{n}}(1, 1, \ldots, 1, 1)^T,\quad \boldsymbol{f}_2 = \frac{1}{\sqrt{2}}(-1, 0, \ldots, 0, 1)^T.$$

These formulae indicate that

(1) the vector $\boldsymbol{f}_1$ is the eigenvector of $\frac{1}{n}E$ associated with the eigenvalue $\lambda = 1$, and the eigenvector of $I - \frac{1}{n}E$ associated with $\lambda = 0$, and

(2) the vector $\boldsymbol{f}_2$ is an eigenvector of $\frac{1}{n}E$ associated with the eigenvalue $\lambda = 0$, and an eigenvector of $I - \frac{1}{n}E$ asociated with $\lambda = 1$.

Applying the Schmidt's orthogonalization process for a system $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_{n-2}\}$, we have the rest of the eigenvectors $\boldsymbol{f}_k,\ 3 \le k \le n$; $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_k, \dots, \boldsymbol{f}_n\}$ is the orthonormal system such that[7]

$$e_j = \left[\, \delta_{ij} \,\right], \quad 1 \le j \le n-2 \,,$$

$$\boldsymbol{f}_k = \frac{1}{\sqrt{(k-1)k}} \left(\underbrace{-1, -1, \dots, -1}_{k-2}, k-1, 0, \dots, 0, -1 \right)^T$$

and

$$\frac{1}{n} E \boldsymbol{f}_k = \boldsymbol{o}, \qquad \left(I - \frac{1}{n} E \right) \boldsymbol{f}_k = \boldsymbol{f}_k \quad \text{for} \quad k \in \{3, \dots, n\} \,.$$

Since it holds

$$\boldsymbol{f}_i^T \boldsymbol{f}_j = \delta_{ij} \quad \text{for} \quad i,\ j \in \{1, \dots, n\},$$

we have an orthogonal matrix $P = [\, \boldsymbol{f}_1\ \boldsymbol{f}_2 \dots \boldsymbol{f}_k \dots \boldsymbol{f}_n \,]$, which satisfies

$$\frac{1}{n} EP = PD_1, \qquad \left(I - \frac{1}{n} E \right) P = PD_2,$$

where D_1 and D_2 denote the $n \times n$ diagonal matrices such that

$$D_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Then we have

$$P^T \begin{bmatrix} x_1 - x \\ x_2 - x \\ x_3 - x \\ \vdots \\ x_k - x \\ \vdots \\ x_n - x \end{bmatrix} = \begin{bmatrix} \sqrt{n} \left(\bar{x} - x \right) \\ \frac{1}{\sqrt{2}} (x_n - x_1) \\ \frac{1}{\sqrt{6}} (-x_1 + 2x_2 - x_n) \\ \vdots \\ \frac{1}{\sqrt{(k-1)k}} \left\{ -x_1 - x_2 - \cdots - x_{k-2} + (k-1) x_{k-1} - x_n \right\} \\ \vdots \\ \frac{1}{\sqrt{(n-1)n}} \left\{ -x_1 - x_2 - \cdots - x_{n-2} + (n-1) x_{n-1} - n_n \right\} \end{bmatrix}.$$

[7]The symbol δ_{ij} denotes the Kronecker's delta : $\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \,, \\ 0 & \text{otherwise.} \end{cases}$

Therefore we have

$$g(x) = [x_i - x]^T P^T P [x_i - x]$$

$$= n(\overline{x} - x)^2 + \frac{1}{2}(x_n - x_1)^2 + \sum_{k=3}^{n} \frac{1}{(k-1)k} \{-x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n\}^2,$$

which shows the function $g(x)$ takes the minimum at $x = \overline{x}$:

$$g(\overline{x}) = \frac{1}{2}(x_n - x_1)^2 + \sum_{k=3}^{n} \frac{1}{(k-1)k} \{-x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n\}^2.$$

Main Theorem [Mendori et al. (2003), Theorem 2.1]. The minimum of the sum of squares $g(\overline{x})$ is evaluated by using the data range R as follows :

$$\frac{1}{2}R^2 \leq g(\overline{x}) \leq \left\{ \frac{n}{4} - \frac{1-(-1)^n}{8n} \right\} R^2.$$

Proof. The minimum of the s.s. $g(\overline{x})$ takes the minimum at the point $[x_i]$ if and only if x_j, $2 \leq j \leq n-1$, satisfy the condition

$$-x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n = 0 \quad \text{for all} \quad k \in \{3, \ldots, n\}.$$

It follows

$$x_j = \frac{1}{2}(x_1 + x_n) \quad \text{for all} \quad j \in \{2, \ldots, n-1\}$$

Then we have

$$\min_{x_1 \leq x_2 \leq \ldots \leq x_n} g(\overline{x}) = \frac{1}{2}R^2.$$

To determine the maximum of $g(\overline{x})$, set

$$f(\boldsymbol{x}) = g(\overline{x}),$$

and

$$\boldsymbol{p}_j = (\underbrace{x_1, \ldots, x_1}_{j}, x_n, \ldots, x_n)^T, \qquad 1 \leq j \leq n-1.$$

For each vector $\boldsymbol{x} = \boldsymbol{p}_j$ we have

$$-x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n = \begin{cases} jR & \text{if} \quad j < k-1, \\ -R & \text{if} \quad j \geq k-1. \end{cases}$$

Since it holds

$$\sum_{k=3}^{n} \frac{1}{(k-1)k} \left\{ -x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n \right\}^2$$

$$= \sum_{k=3}^{j+1} \frac{1}{(k-1)\,k} \left\{ -x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n \right\}^2$$

$$+ \sum_{k=j+2}^{n} \frac{1}{(k-1)\,k} \left\{ -x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n \right\}^2$$

$$= \left\{ \frac{1}{2} - \frac{1}{j+1} + \left(\frac{1}{j+1} - \frac{1}{n} \right) j^2 \right\} R^2,$$

we have

$$\sum_{k=3}^{n} \frac{1}{(k-1)\,k} \left\{ -x_1 - x_2 - \cdots - x_{k-2} + (k-1)\, x_{k-1} - x_n \right\}^2 = -\frac{1}{2} R^2 + \frac{j(n-j)}{n} R^2.$$

Thus we have

$$f(\boldsymbol{p}_j) = \frac{j(n-j)}{n} R^2 = -\frac{1}{n} \left(j - \frac{n}{2} \right)^2 R^2 + \frac{n}{4} R^2.$$

It follows

$$\max_{x_1 \le x_2 \le \ldots \le x_n} g(\bar{x}) = \max_{1 \le j \le n-1} f(\boldsymbol{p}_j) = \begin{cases} \dfrac{n}{4} R^2 & \text{if } n \text{ is even}, \\ \left\{ \dfrac{n}{4} - \dfrac{1}{4n} \right\} R^2 & \text{if } n \text{ is odd}. \end{cases} \quad \square$$

§ 3. Further Problems

In this paper we gave a proof of the Main Theorem by a purely mathematical technique. One of the remained problems is to extend the method to the case of the ranges with k stratifications.

From the point of view of the statistical inference it is interesting and practically useful to investigate the validity of an estimation based on the distribution of the range R in the order statistics.

References

Gauss, C. F.（1823）: Theoria combinationis observationum erroribus minimis obnoxiae pars prior, *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores*, 5.（in the book C. F. Gauss: *Gosa-ron*, edited and translated into Japanese by T. Hida and T. Ishikawa, Kinokuniya Shoten, Tokyo 1981.）

Healy, M. J. R.（1986）: *Matrices for Statistics*, Oxford University Press, New York.

Hogg, R. V. and Craig, A. T.（1978）: *Introduction to Mathematical Statistics, 4th ed.*, Collier Macmillan International Editions.

Kitagawa, T.（1958）: *Suisoku-tôkei-gaku I*（in Japanese）, Iwanami Shoten, Tokyo.

Mendori, T., Zinno, Y. and Fukaishi, H.（2003）: On the evaluations of sum of squares by using the range, *Mem. Fac. Educ., Kagawa Univ. II*, 53, 1-12.

Mendori, T., Zinno, Y. and Fukaishi, H.（2003a）: On the evaluations of sum of squares by using the range II, *Mem. Fac. Educ., Kagawa Univ. II*, 53, 73-83.

Mendori, T., Zinno, Y. and Fukaishi, H.（2004）: On the evaluations of sum of squares by using the range III, *Mem. Fac. Educ., Kagawa Univ. II*, 54, 53-74.

Mood, A. M., Graybill, F. A. and Boes, D. C.（1974）: *Introduction to the Theory of Statistics, 3rd ed.*, McGraw-Hill.

Raynolds, M. R., Jr., Amin, R. W., Arnold, J. C. and Nachlas, J. A.（1988）: $\overline{X}$ charts with variable sampling intervals, *Technometrics*, 30-2, 181-192.

Satake, I.（1975）: *Linear Algebra*, Marcel Dekker Inc., New York.

Snedecor, G. W., Cochran, W. G. and Cox, D. F.（1989）: *Statistical Methods, 8th ed.*, The Iowa State University Press, Ames.

高 松 大 学 紀 要
第　46　号
平成18年９月25日　印刷
平成18年９月28日　発行

編集発行　高　　松　　大　　学
高　松　短　期　大　学
〒761-0194　高松市春日町960番地
TEL（087）841－3255
FAX（087）841－3064

印　　刷　株式会社　美巧社
高松市多賀町１－８－10
TEL（087）833－5811