Insights Into Teaching
Basic Statistics

Anwar H Joarder

Contents

Preface

Six ways to look at linear interpolation*

Anwar Joarder
Department of Mathematical Sciences
King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
Email: anwarj@kfupm.edu.sa

Linear interpolation has been explained from different perspectives that are likely to be easily understood by students. Some methods seem to be more interesting in some situations. Apart from its suitability in classrooms, it has also a mnemonic value often expected by many readers.

1. Introduction

If a line is used to estimate a functional value between y-values for which the x-values are known, the process is called linear interpolation. In other words it consists of putting a line through two points over small regions on a curve, and then using the line to approximate the curve. Consider three points $A(x_1,\ y_1)$, $C(x,\ y)$, $B(x_2,\ y_2)$ on a line where y is unknown and $y_1 < y < y_2$. Let us represent them by

x_1	y_1
x	y
x_2	y_2

The value of y is obtained by equating slopes as follows:

$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1} \tag{1}$$

which is popularly known as linear interpolation.

In this note linear interpolation has been viewed in several ways. They have been labeled as the ratio method, the distance method, the weighing method, the determinant method, the least squares method and the expected value method. They are also illustrated with examples. An application is shown to derive the formula for median in the context of a frequency distribution. Finally linear interpolation has been characterized as the expected value of a random variable based on a uniform probability distribution.

* Published in *International Journal of Mathematical Education in Science and Technology*, 32(6), 932-937 [London, UK]

2. Some Representations of the Linear Interpolation

2.1 The Ratio Method

Since A, C and B are assumed to be on a line, it is easy to check that C divides AB at the ratio

$$\frac{AC}{BC} = \frac{x - x_1}{x_2 - x} = r \tag{2}$$

and that

$$\frac{y - y_1}{y_2 - y} = r .$$

Since all $x's$ are known, the value of r can be calculated by (2). It then follows from the above equation that

$$y = \frac{y_1 + r\, y_2}{1 + r} \ . \tag{3}$$

Example 2.1 Let A, C and B be given by

0.7486	0.67
0.75	y
0.7518	0.68

By (2) we have, $r = \dfrac{0.75 - 0.7486}{0.7518 - 0.75} = \dfrac{7}{9}$ so that it follows from (3) that

$$y = \frac{0.67 + r\, 0.68}{1 + r} = 0.674375 .$$

Readers acquainted with elementary statistics may recall that it is the 75^{th} percentile of the standard normal distribution. See e.g. [3, p 537].

2.2 The Distance Method

It follows from equation (1) that

$$y = y_1 + \frac{x - x_1}{x_2 - x_1}\ (y_2 - y_1) . \tag{4}$$

Since $y_1,\ y$ and y_2 $(y_1 < y < y_2)$ are points on the real line, one can make the following equivalent statements:

(i) y is $\dfrac{x - x_1}{x_2 - x_1}$ unit away from y_1 towards y_2.

(ii) y is $\dfrac{x - x_1}{x_2 - x_1}$ unit of the way between y_1 and y_2.

(iii) y is $\dfrac{x - x_1}{x_2 - x_1}$ of the way from y_1 to y_2 (cf. [4, p 45] or, [1, pp.10-11]).

Example 2.2 Let A, C and B be given by

0.9495	1.64
0.95	y
0.9505	1.65

Here $\dfrac{x - x_1}{x_2 - x_1} = \dfrac{0.0005}{0.0010} = 0.5$ so that y is 0.5 unit away from y_1 towards y_2. That is $y = 1.64 + 0.5(1.65 - 1.64) = 1.645$. Readers acquainted with elementary statistics may recall that 1.645 is the 95th percentile of the standard normal distribution. See e.g. [3, p 537].

2.3 The Weighing Method

The equation in (1) can also be written as

$$y = y_1 + \frac{x - x_1}{x_2 - x_1} y_2 - \frac{x - x_1}{x_2 - x_1} y_1 = (1 - w)\, y_1 + w\, y_2 \qquad (5)$$

where $w = \dfrac{x - x_1}{x_2 - x_1}$. If $x_1 < x < x_2$, then by subtracting x_1 from both sides of $x < x_2$, it follows that $0 < w < 1.$ Similarly it can be proved that $0 < w < 1$ if $x_1 > x > x_2$. Let us apply this method to Example 2.1 so that

$$w = \frac{x - x_1}{x_2 - x_1} = \frac{0.75 - 0.7486}{0.7518 - 0.7486} = 0.4375$$

which, by the Distance Method, means that y is 0.4375 unit away from y_1 towards y_2. Clearly y is closer to y_1 than it is to y_2. By the Weighing Method we then have

$$y = (1-w)\,y_1 + w\,y_2 = (1-0.4375)y_1 + 0.4375\,y_2$$
$$= (0.5625)\,0.67 + (0.4375)\,0.68 = 0.674375$$

It may be remarked here that in many situations $x_2 - x_1 = 1$ so that $w = x - x_1$. In those situations the weighing method is easily grasped by students. It is explained below by an example.

Example 2.3 Consider a sample [3, p 46] with n=10 observations with the second largest 5.4 and the third largest 5.7. The rank of the lower quartile (Q_1) is $(n+1)/4 = (10+1)/4 = 2 + 0.75$ so that the lower quartile is an observation between the 2nd and the 3rd as depicted in the following table:

Rank	Observation
2	5.4
2.75	Q_1
3	5.7

Here $w = x - x_1 = 2.75 - 2 = 0.75$ so that it follows from (5) that

$$Q_1 = (1-0.75)\,(\text{2nd observation}) + 0.75\,(\text{3rd observation})$$
$$= (1-0.75)(5.4) + 0.75(5.7) = 5.625.$$

It is interesting to note that the weights are intuitively appealing here. The rank 2.75 of the lower quartile implies that it is closer to the 3rd observation than it is to the 2nd . So the weights 0.75 and 0.25 must be attached to the 3rd and the 2nd observations respectively. This type of problems are frequently encountered in statistics and it seems this method is the best, especially in classrooms, for calculating quartiles, deciles, percentiles or in general for quantiles.

2.4 The Determinant Method

Since the three points are on a line, the area of the triangle made by the three points must vanish, hence in general we have

$$\begin{vmatrix} x_1 & y_1 & 1 \\ x & y & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0\,. \qquad (6)$$

Those who are acquainted with determinants can easily evaluate that

$$x_1(y - y_2) - y_1(x - x_2) + (xy_2 - yx_2) = 0$$

Consider interpolating the value of y from Example 2.1. We have

$$\begin{vmatrix} 0.7486 & 0.67 & 1 \\ 0.75 & y & 1 \\ 0.7518 & 0.68 & 1 \end{vmatrix} = 0$$

whence

$$0.7486\,(y - 0.68) - 0.67\,(0.75 - 0.7518) + 1[0.75\,(0.68) - 0.7518\,y] = 0$$

and consequently, $y = 0.674375$.

2.5 The Least Squares Method

The equation (1) can be written as

$$y = -\frac{x_1\,y_2 - y_1\,x_2}{x_2 - x_1} + \frac{y_2 - y_1}{x_2 - x_1}\,x\,. \tag{7}$$

But it is easy to check that the least squares estimate of β_0 and β_1 in the line $y = \beta_0 + \beta_1 x$ based on the points (x_1, y_1) and (x_2, y_2) are given by

$$\beta_1 = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2} = \frac{y_2 - y_1}{x_2 - x_1} \text{ and}$$

$$\beta_0 = \bar{y} - \beta_1\,\bar{x} = \frac{y_1\,x_2 - x_1\,y_2}{x_2 - x_1}$$

respectively, which are the slope and the intercept parameter. Note that the least square estimate of β_0 can simply be written as $\beta_0 = y_1 - \beta_1 x_1$ (or $\beta_0 = y_2 - \beta_1 x_2$).

Consider interpolating the value of y in Example 2.1. It is easily checked that $\beta_1 = 3.125$, $\beta_0 = -1.669375$ and consequently

$$y = -1.669375 + 3.125\,x\,. \tag{8}$$

If $x = 0.75$, then $y = 0.674375$. It is obvious from (8) that this method is the best if one needs to find several values of y corresponding to several values of x.

2.6 The Expected Value Method

Let us have the following three points

x_1	y_1
x	c
x_2	y_2

The following proposition is obvious from Sections 2.1 and 2.3.

Proposition 1: Linear interpolation to find c may be viewed as the expected value of a discrete random variable Y with the following probability mass function:

$$P(Y = y_1) = \frac{x_2 - x}{x_2 - x_1} = \frac{1}{1+r} = 1 - w$$
$$P(Y = y_2) = \frac{x - x_1}{x_2 - x_1} = \frac{r}{r+1} = w, \qquad 0 \le w \le 1$$

Proposition 2: Linear interpolation to find c may be viewed as the expected value of a discrete random variable Y with the following probability mass function :

$$Y = \begin{cases} y_1 & \text{if } X \ge x \\ y_2 & \text{if } X \le x \end{cases}$$

where X has a continuous uniform probability distribution $U(x_1, x_2)$.

Proof: The expected value of Y is given by

$$\begin{aligned} E(Y) &= y_1 \, P(Y = y_1) + y_2 \, P(Y = y_2) \\ &= y_1 \, P(X \ge x) + y_2 \, P(X \le x) \\ &= y_1 \, [1 - P(X \le x)] + y_2 \, P(X \le x) \\ &= y_1 \left(1 - \frac{x - x_1}{x_2 - x_1} \right) + y_2 \left(\frac{x - x_1}{x_2 - x_1} \right) \end{aligned}$$

which is the formula derived by the Weighing Method.

We conjecture that optimal probability model can be determined to improve upon the usual interpolation methods.

3. An Application to Statistics

Consider a frequency distribution having median class $[y,\ y+h]$ with relative frequency $F_{y+h} - F_y$ where F_{y+h} = cumulative relative frequency up to the median class and
F_y = cumulative relative frequency up to the class preceding the median class. Also let $y_{0.50}$ be the median. Then we have the following representation

F_y	y
0.50	$y_{0.50}$
F_{y+h}	$y+h$

Then by (4) we have

$$y_{0.50} = y + \frac{0.50 - F_y}{F_{y+h} - F_h}\ h$$

which is the well-known formula for median of grouped data in a frequency distribution. See e.g. [2, p 72].

Acknowledgement

The author acknowledges the excellent research facilities available at King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

References

[1] Briddle, D. F., 1982, Analytic Geometry. Wadsworth Publishing Co. California, USA.

[2] Kvanli, Alan H.; Guynes, C.H. and Pavur, Robert J., 1992, *Introduction to Business Statistics: A Computer Integrated Approach*. West Publishing Co. New York, USA.

[3] Lapin, L. L., 1997, *Modern Engineering Statistics*. Wadsworth Publishing Co., New York, USA.

[4] Newbold, P., 1995, *Statistics for Business and Economics*. Prentice Hall. New Jersey, USA.

The Halving Method for Sample Quartiles*

ANWAR H. JOARDER
Dept of Mathematical Sciences
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia 31261
Email: anwarj@kfupm.edu.sa

An attempt is made to put the notion of sample quartiles on a mathematical footing in the light of ranks of observations, and equisegmentation property that the number of ranks below that of the first quartile, that between the consecutive quartiles, and that above the third quartile are the same. Ranks of sample quartiles provided by the proposed Halving Method, based on hinges, does satisfy the property.

1. Introduction

There are many methods available for calculating sample quartiles in different elementary text books on statistics without any explanation. The most popular one, called Popular Method hereinafter, is described here. The rank of the $i(i=1,2,3)$ th quartile is given by

$$i\ (n+1)/4 = l + d,\quad \mathrm{i} = 1, 2, 3 \qquad (1.1)$$

where l is the largest integer not exceeding $i(n+1)/4$. Then the Popular Method uses
the following linear interpolation formula for the calculation of sample quartiles

$$Q_i = x_{(l)} + d\,(x_{(l+1)} - x_{(l)}) = (1-d)\,x_{(l)} + d\,x_{(l+1)},\quad (\mathrm{i} = 1, 2, 3), \qquad (1.2)$$

where $x_{(l)}$ is the l-th ordered observation (Ostle, Turner, Hicks and McElrath, 1996, 38).

However, students and instructors alike are curious to know why the formulae for quartiles in (1.1) contain the quantity $n+1$. Why not n or $n-1$? Though the formulae for the median in the literature appear to be different, they all are equivalent. It is given by $Q_2 = (n+1)/2$ th observation. In case n is odd, $(n+1)/2$ will be an integer so that the median will be an observation with integer rank. If however, n is even, $(n+1)/2$ will lie between $n/2$ and $n/2+1$. Then the median can be calculated by the use of linear interpolation. Because of the success of the quantity $(n+1)$ in equation (1.1) to find the median, the idea of proportional weight given by (1.1) or (1.2) has possibly been popular to find other quartiles by the above method.

* Published in *International Journal of Mathematical Education in Science and Education*, 34(4), 2003, [London, UK]

It would be clear down the road that $n+1$ is the total of the ranks for the largest and smallest observations in the sample, and that the rank of the median is the average of the ranks of the observations.

To write out the ranks exhaustively let us denote the sample size by the following remainder-modulus representation

$$n = r \bmod 4 = 4m + r,\ (r = 0,1,2,3), \tag{1.3}$$

so that the number of observations in each of the $4 \le n$ segments is given by $m = (n-r)/4$. With this representation of the sample size the ranks and quartiles of a sample will be denoted respectively by R_{ir} and Q_{ir} ; $i = 1,2,3;\ r = 0,1,2,3$. Though quartiles Q_{ir} ; $i = 1,2,3;\ r = 0,1,2,3$ are usually denoted by Q_i ; $i = 1,2,3$, we will not suppress r as it plays an important role in the proposed Halving Method for quartiles. The ranks in (1.1) given by the Popular Method can be rewritten as

$$R_{ir} = i(4m + r + 1)/4 = im + i(r+1)/4,\ i = 1,2,3;\ r = 0,1,2,3 \tag{1.4}$$

which is the the rank of the i th quartile corresponding to the sample size with remainder r. Then the ranks for sample quartiles provided by the Popular Method can be written out exhaustively as:

$$R_{10} = m + 1/4,\ R_{20} = 2m + 2/4,\ R_{30} = 3m + 3/4$$
$$R_{11} = m + 2/4,\ R_{21} = 2m + 1,\ R_{31} = 3m + 1 + 2/4$$
$$R_{12} = m + 3/4,\ R_{22} = 2m + 1 + 2/4,\ R_{32} = 3m + 2 + 1/4$$
$$R_{13} = m + 1,\ R_{23} = 2m + 2,\ R_{33} = 3m + 3$$

We propose the new criterion of equisegmentation property that the number of ranks below that of the first quartile, that between the consecutive quartiles, and that above the third quartile are the same. However this will divide the ordered sample observations into four segments leaving the same number of observations in each if all the observations are distinct. Let the number of integers in each segment be m_i $(i = 1,2,3,4)$. Then the equisegmentation property guarantees that $m_1 = m_2 = m_3 = m_4$. In case $1 \le n \le 3$, the above formulae can also be used to calculate quartiles with $m = 0$.

It is interesting to note that though the Popular Method is not based on good mathematical reasoning, the equisegmentation property is satisfied by the quartiles provided by this method for all sample sizes except for $n = 4m + r$, $m \ge 1,\ r = 2$. For $r = 2$, the number of observations in four segemnts are $m, (m+1), (m+1)$ and m respectively.

Thus it is essential to modify the formulae of ranks so that the equisegmentation property is satisfied by quartiles provided by the Popular Method for any sample size. It is observed that, whenever $n = 4m + 2$, simple arithmetic rounding of ranks provided by this method would satisfy the equisegmentation property.

The Halving Method discussed in this paper demonstrates in an accessible way that the set of formulae for quartiles offered by this method is based on good mathematical reasoning. The ranks provided by the Halving Method written out exhaustively by the remainder-modulus representation of the sample size help prove that the corresponding quartiles satisfy the equisegmentation property. Moreover, ranks provided by the Halving Method guarantee that the remainder r of the sample size is the number of quartiles having integer ranks. Linear interpolation should be used to find quartiles with noninteger ranks.

2. The Halving Method for Sample Quartiles

The method, developed in the spirit of Tukey (1977, p32-35), is based on hinges which finds the median first, and then finding the medians of upper and lower halves of the data. Usually median is included in both halves while calculating the hinges. But we observe that if median of the whole data set is ignored in the calculation of hinges, then the two extreme hinges and median enjoy equisegmentation property. We develop algebraic expressions for ranks of quartiles based on this argument and call this method the Halving Method. It thus resolves the difference between quartiles and hinges. The ranks for the quartiles given by the Halving Method are developed below in terms of r and m where the sample size $n = 4m + r,\ (r = 0, 1, 2, 3)$:

(a) Ranks of quartiles for $n = 4m$

The observations have ranks $1, 2, ...,\ 2m, 2m+1, ..., 4m$. The rank of the median is

$$R_{20} = \frac{1}{4m}(1 + 2 + \cdots + 4m) = \frac{1}{4m}\frac{4m(1+4m)}{2} = \frac{1+4m}{2} = 2m + 0.5$$

which is between $2m$ and $2m+1$ so that the ranks of extreme quartiles are given by

$$R_{10} = \frac{1+2m}{2} = m + 0.5,\ R_{30} = \frac{(2m+1)+4m}{2} = 3m + 0.5\,.$$

It is worth mentioning that in this case none of the quartiles has integer ranks.

(b) Ranks of quartiles for $n = 4m + 1$

The observations have ranks $1, 2, ...,\ 2m, 2m+1, 2m+2 ..., 4m+1$. The rank of the median is

$$R_{21} = \frac{1+(4m+1)}{2} = 2m + 1$$

which is between $2m$ and $2m+2$ so that the ranks of extreme quartiles are given by

$$R_{11} = \frac{1+2m}{2} = m + 0.5,\ R_{31} = \frac{(2m+2)+(4m+1)}{2} = 3m + 1.5\,.$$

It is worth mentioning that in this case the median has an integer rank.

(c) Ranks of quartiles for $n = 4m + 2$

The observations have ranks $1, 2, ..., 2m, 2m+1, 2m+2 ..., 4m+2$. The rank of the median is

$$R_{22} = \frac{1 + (4m + 2)}{2} = 2m + 1.5$$

which is between $2m+1$ and $2m+2$ so that the ranks of extreme quartiles

$$R_{12} = \frac{1 + (2m + 1)}{2} = m + 1, \; R_{32} = \frac{(2m + 2) + (4m + 2)}{2} = 3m + 2 .$$

It is worth mentioning that in this case the extreme quartiles have integer ranks.

(d) Ranks of quartiles for $n = 4m + 3$

The observations have ranks $1, 2, ..., 2m, 2m+1, 2m+2 , 2m+3, ..., 4m+3$. The rank of the median is

$$R_{23} = \frac{1 + (4m + 3)}{2} = 2m + 2$$

which is between $2m+1$ and $2m+2$ so that the ranks of extreme quartiles are

$$R_{13} = \frac{1 + (2m + 1)}{2} = m + 1, \; R_{33} = \frac{(2m + 3) + (4m + 3)}{2} = 3m + 3 .$$

In practice one may simply use the above argument to calculate ranks of quartiles. The other alternative is to find r and $m = (n - r)/4$ and then use the ranks of quartiles given below to calculate quartiles.

$$R_{10} = m + 2/4, \; R_{20} = 2m + 2/4, \; R_{30} = 3m + 2/4$$
$$R_{11} = m + 2/4, \; R_{21} = 2m + 1, \; R_{31} = 3m + 1 + 2/4$$
$$R_{12} = m + 1, \; R_{22} = 2m + 1 + 2/4, \; R_{32} = 3m + 2$$
$$R_{13} = m + 1, \; R_{23} = 2m + 2, \; R_{33} = 3m + 3$$

The rank of the median, in Popular Method as well as in Halving Method, is the average of the first and third quartiles. The remainder r here is also the number of quartiles having integer ranks in the Halving Method but not in the Popular Method. It is easy to check that equisegmentation property is satisfied by the quartiles offered by the Halving Method for $n = 4m + r, \; r = 0,1,2,3$ i.e. for all sample sizes. The explicit form of

the ranks of quartiles by the two methods help us compare them. In fact each of the rank $R_{10}, R_{30}, R_{12}, R_{32}$ given by the Popular Method differs from that given by the Halving Method by $1/4$.

We recommend to use the Halving Method as it is based on logic. The generalization of the method to deciles, percentiles or to any quantiles, in general, remains open.

3. An Illustration

The following ten value are sample weights (in grams) of coating materials used in a masking process:

5.3 5.4 5.7 6.0 6.1 6.2 6.3 6.4 6.5 6.6

(i) Calculation of Quartiles by Popular Method

Here the sample size $n = 10 = 4(2) + 2$ so that $m = 2$ and $r = 2$. Since $r = 2$ we will denote the ranks of quartiles by R_{i2} $(i = 1, 2, 3)$. The rank of the quartiles provided by the Popular Method are (see equation 1.1)

$$R_{12} = (n+1)/4 = 2.75,\ R_{22} = (n+1)/2 = 5.5,\ R_{32} = 3(n+1)/4 = 8.25$$

which can also be written equivalently as

$$R_{12} = m + 3/4 = 2.75,\ R_{22} = 2m + 1 + 2/4 = 5.5,\ R_{32} = 3m + 2 + 1/4 = 8.25$$

(see equation 1.4). Note that the consecutive ranks are apart by 3, and there are 2 ranks below R_{12} or above R_{32}. Then by linear interpolation (see equation 1.2) the quartiles are given by

$$Q_{12} = x_{(2.75)} = (1-0.75)\, x_{(2)} + 0.75\, x_{(3)} = 0.25(5.4) + 0.75(5.7) \approx 5.625$$
$$Q_{22} = x_{(5.5)} = (1-0.5)\, x_{(5)} + 0.5\, x_{(6)} = 0.5(6.1) + 0.5(6.2) = 6.15$$
$$Q_{32} = x_{(8.25)} = (1-0.25)\, x_{(8)} + 0.25\, x_{(9)} = 0.75(6.4) + 0.25(6.5.) \approx 6.425$$

To check the equisegmentation property, we show the ranks of the quartiles by downward arrows in the sample:

5.3 5.4 ↓ 5.7 6.0 6.1 ↓ 6.2 6.3 6.4 ↓ 6.5 6.6

We observe that there are $2\,(= m)$, $3(= m+1)$, $3(= m+1)$ and $2(= m)$ observations in the four segments i.e. the ranks of the quartiles do not satisfy the equisegmentation property.

(ii) *Calculation of Quartiles by Halving Method*

Instead of using the formulae provided by the Halving Method at the end of section 2, we prefer to use the idea of halving to find quartiles with the hope that it would provide more insight into the problem.

The rank of the median is $R_{22} = \frac{1+n}{2} = 5.5$ so that $Q_{22} = x_{(5.5)} = (1-0.5)\, x_{(5)} + 0.5\, x_{(6)} = 0.5(6.1) + 0.5(6.2) = 6.15$. The first quartile is the median of the observations below the median of the whole data set i.e. is $R_{12} = \frac{1+5}{2} = 3$ so that $Q_{12} = x_{(3)} = 5.7$. The third quartile is the median of the observations above the median of the whole data set i.e. is $R_{32} = \frac{6+10}{2} = 8$ so that $Q_{32} = x_{(8)} = 6.4$.

To check the equisegmentation property, we show the ranks of the quartiles by downward arrows in the sample:

We observe that there are $2(=m)$ observations in each of the four segments i.e. the ranks of the quartiles do satisfy the equisegmentation property. The author also thanks an anonymous referee for constructive suggestions that have improved the readability and presentation of an earlier draft of the paper.

Acknowledgements

The author is grateful to the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia. The author is also thankful to an anonymous referee for constructive suggestions on an earlier draft that have improved the readability and presentation of the paper.

References

Ostle, B., Turner, K.V. Hicks, C.R. and McElrath, G.W. (1996). *Engineering Statistics*: The Industrial Experience. New York: Duxbury Press.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

A Comparison and Contrast of Some Methods for Sample Quartiles*

Anwar H. Joarder and Raja M. Latif
Department of Mathematical Sciences
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia 31261; emails:
anwarj@kfupm.edu.sa and raja@kfupm.edu.sa.

ABSTRACT A remainder representation of the sample size $n = 4m + r$ ($r = 0, 1, 2, 3$) is exploited to write out the ranks of quartiles exhaustively which in turn help compare ranks for quartiles provided by different methods available in the literature. The criterion of the equisegmentation property that the number of integer ranks below the first quartile, that between the consecutive quartiles, and that above the third quartile are the same, has been used to compare and contrast different methods. Four segmentation identities can be obtained for each method of quartiles which show clearly the number of observations in each of the four quarters if the observations are distinct. The Halving Method and the Remainder Method have been proposed for the calculation of sample quartiles. The quartiles provided by each of these two methods satisfy the equi- segmentation property if the observations are distinct. More interestingly, in these two methods r also represents the number of quartiles having integer ranks.

Keywords : Quartiles; Remainders; Modulus; Quantiles.

1. Introduction

Quartiles, deciles, percentiles or more generally fractiles are uniquely determined for continuous random variables. A p^{th} quantile of a random variable X (continuous or discrete) is a value x_p such that $P(X < x_p) \le p$ and $P(X \le x_p) \ge p$. Let X be a continuous or discrete random variable with probability function $f(x)$ and the cumulative distribution function $F(x) = P(X \le x)$. If the distribution is continuous, then $P(X < x_p) = p$ and $P(X \le x_p) = p$ since $P(X = x_p) = 0$. Therefore, for the continuous case, $F(x_p) = p$.

The quartiles $Q_1 = x_{0.25}$, $Q_2 = x_{0.50}$ and $Q_3 = x_{0.75}$ for a continuous random variable with cumulative distribution function $F(x)$ are defined by $F(x_{0.25}) = 0.25$, $F(x_{0.50}) = 0.50$ and

* Published in *Journal of Probability and Statistical Science*, 2(1), 95-109. [Taiwan]

$F(x_{0.75})$ = 0.75 respectively. Let X follow an exponential distribution with the probability density function

$$f(x) = \beta^{-1} e^{-x/\beta}, \qquad x > 0$$

with the cumulative distribution function $F(x) = 1 - e^{-x/\beta}$. Then

$$1 - e^{-Q_1/\beta} = 1/4, \quad 1 - e^{-Q_2/\beta} = 2/4 \text{ and } 1 - e^{-Q_3/\beta} = 3/4$$

so that

$$Q_1 = \beta \ln(4/3), \; Q_2 = \beta \ln 2, \; Q_3 = \beta \ln 4.$$

However, for the discrete distribution, one has to use the basic definition. Consider the binomial distribution $B(n = 4, \pi = 1/2)$. The probability mass function is given by

$$f(x) = \begin{cases} \binom{4}{x}(1/2)^4, & x = 0, 1, \cdots, 4; \\ 0 & \text{elsewhere.} \end{cases}$$

Then $x_{0.25} = 1$, is the first quartile of the distribution since

$$P(X < 1) = P(X = 0) = 0.0625 \le 0.25,$$
$$P(X \le 1) = P(X = 0) + P(X = 1) = 0.3125 \ge 0.25.$$

Similarly $x_{0.50} = 2$, is the second quartile of the distribution since

$$P(X < 2) = 0.3125 \le 0.50, \quad P(X \le 2) = 0.6825 \ge 0.50.$$

Note that the median is the same as 0.5-quantile or the 50th percentile, or the 5th decile. It is not surprising that the 60th percentile, $x_{0.6} = 2$, since $P(X < 2) = 0.3125 \le 0.60$ and $P(X \le 2)$ = $0.6825 \ge 0.60$. Similarly it can be checked that the third quartile is given by $x_{0.75} = 3$.

In case we have a sample (discrete in nature), it is, however, difficult to define quartiles. One method, called the Hinge Method, is based on finding the median first and then finding the medians of the upper and lower halves (including original median in both halves) of the data. Done so, roughly 25% observations remain below the lower quartile and 25% above the upper quartile. A sample quantile is a point below which some specified proportion of the values of a data set lies. The median is the 0.50 quantile because approximately half of all the observations lie below this value. The name fractile for quantile is used by some authors (see Lapin [6], p. 52).

A remainder representation of the sample size $n = 4m + r$ $(r = 0,1,2,3)$ is exploited to write out the ranks of quartiles exhaustively which in turn help compare ranks for quartiles provided by different methods available in the literature. Some of them differ only by various rounding notions of the corresponding ranks for quartiles. We compare and contrast different methods of quartiles in the light of equisegmentation property that the number of integer ranks below the first quartile, that between the consecutive quartiles, and that above the third quartile are the same. For each method of quartiles, four segmentation identities are obtained which clearly show the number of observations in each of the four quarters if the observations are distinct. The Halving Method and the Remainder Method have been proposed for the calculation of sample quartiles. The quartiles provided by each of these two methods divide the ordered sample observations in four quarters with the same number of observations in each segment and provide the number of quartiles having integer ranks if the observations are distinct.

2. The Popular Method

There are many methods available for calculating sample quartiles in different elementary text books on statistics without any explanation. The most popular one, called the Popular Method hereinafter, is described below. The rank of the *i*(*i* = 1, 2, 3)th quartile is given by

$$i\,(n+1)/4 = l + d, \quad i = 1, 2, 3 \qquad (2.1)$$

where l is the largest integer not exceeding $i(n+1)/4$. Then the Popular Method uses the following linear interpolation formula for the calculation of sample quartiles

$$Q_i = x_{(l)} + d\,(x_{(l+1)} - x_{(l)}) = (1-d)\,x_{(l)} + d\,x_{(l+1)}, \quad (i = 1, 2, 3), \qquad (2.2)$$

where $x_{(l)}$ is the *l*th ordered observation (Ostle *et al*. [12], p. 38). To write out the ranks exhaustively let us denote the sample size $n(\geq 4)$ by the following remainder-modulus representation

$$n = 4m + r = r \pmod 4, \ (r = 0,\ 1,\ 2,\ 3), \qquad (2.3)$$

so that the number of observations in each of the four segments is given by $m = (n-r)/4.$ With this representation of the sample size, the ranks and the quartiles of a sample will be denoted respectively by R_{ir} and Q_{ir} ; $i = 1, 2, 3;\ r = 0, 1, 2, 3$. Though quartiles Q_{ir} ; $i = 1, 2, 3;\ r = 0, 1, 2, 3$ are

usually denoted by Q_i ; $i = 1, 2, 3$, we will not suppress r as it plays an important role in comparing the ranks of quarters given by different methods.

Let the number of observations in each segment be m_i (*i* = 1, 2, 3, 4). Then the equi- segmentation property guarantees that $m_1 = m_2 = m_3 = m_4$ if the observations are distinct. In case, $1 \le n \le 3$, the above formulae can also be used to calculate quartiles with $m = 0$.

It is interesting to note that though the Popular Method is not based on good mathematical reasoning, the equisegmentation property is satisfied by the quartiles provided by this method for all sample sizes $n = 4m + r \ \ (m \ge 1;\ r = 0, 1, 3)$ if the observations are distinct. For $n = 4m + 2 \ \ (m \ge 1),$ the number of observations in four segments are given by $m, (m+1), (m+1)$ and *m* respectively if the observations are distinct.

Thus it is essential to modify the formulae of ranks so that the equisegmentation property is satisfied by quartiles provided by the Popular Method for any sample size. It is observed that whenever $n = 4m + 2$, simple arithmetic rounding of ranks provided by this method would satisfy the equisegmentation property.

Example 2.1 The sizes of the police forces in the ten largest cities in the United States in 1993 (the numbers represent hundreds) are given below:

1.7 1.9 2.0 2.8 3.9 4.7 6.2 7.6 12.1 29.3

(Bluman [1], p.137). We now calculate quartiles by the Popular Method. Here the sample size is $n = 10 = 4(2) + 2$ so that $m = 2$ and $r = 2$. For $r = 2$ we will denote the ranks of quartiles by R_{i2} $(i = 1, 2, 3)$. The ranks of the quartiles provided by the Popular Method are (see equation 2.1) given by

$$R_{12} = (n+1)/4 = 2.75,\ R_{22} = (n+1)/2 = 5.5,\ R_{32} = 3(n+1)/4 = 8.25$$

so that by linear interpolation (see equation 2.2) the quartiles are given by

$$Q_{12} = x_{(2.75)} = (1 - 0.75)\, x_{(2)} + 0.75\, x_{(3)} = 0.25(1.9) + 0.75(2.0) = 1.975$$

$$Q_{22} = x_{(5.5)} = (1 - 0.5)\, x_{(5)} + 0.5\, x_{(6)} = 0.5(3.9) + 0.5(4.7) = 4.3$$

$$Q_{32} = x_{(8.25)} = (1 - 0.25)\, x_{(8)} + 0.25\, x_{(9)} = 0.75(7.6) + 0.25(12.1) = 8.725$$

To check the equisegmentation property, we show the position of the quartiles by downward arrows in the sample:

			⇓			⇓		⇓	
1.7	1.9	2.0	2.8	3.9	4.7	6.2	7.6	12.1	29.3

We observe that there are $2(=m)$, $3(=m+1)$, $3(=m+1)$ and $2(=m)$ observations in the four segments, i.e. the quartiles do not satisfy the equisegmentation property for $n=4m+2$.

3. A Review of the Well-known Formulae of Sample Quartiles

In this section we survey the formulae for quartiles available in the literature. We pro- vide algebraic expressions for quartiles by all existing methods in the literature. The use of remainder allows us to figure out the decimal part of the formulae of ranks for quartiles for a particular sample of size n. Let $m=(n-r)/4,\ n=4m+r\geq 4$, and R_{ir} be the rank of i th quartile with *m* observations in each segment. Then the rank of the i th quartile is given by

$$R_{ir}=i\ \frac{(4m+r)+1}{4}=im+\ i(r+1)/4=im+[u_{ir}]+d_{ir}/4 \qquad (3.1)$$

where *i* and *r* are integers with $1\leq i\leq 3$, $0\leq r\leq 3$, $[u_{ir}]$ is the largest integer less than or equal to $u=u_{ir}=i(r+1)/4$ and $i(r+1)=d_{ir}(\mathrm{mod}\,4)$. The quartiles can then be cal- culated by the simple linear interpolation as

$$Q_{ir}=(1-d/4)\,x_{(im+[u])}+(d/4)\,x_{(im+[u]+1)} \qquad (3.2)$$

where $x_{(i)}$ is the *i*th ordered observation, $u=u(i,r)=i(r+1)/4$, $[u]$ is the greatest integer not exceeding u and $d=d_{ir}=4(u-[u])$. The above method will be called the Popular Method.

Method 1 (Popular Method) The ranks for sample quartiles provided by the Popular Method can be written out exhaustively as (see 3.1):

$$\begin{gathered}R_{10}=m+1/4,\ R_{20}=2m+2/4,\ R_{30}=3m+3/4,\\ R_{11}=m+2/4,\ R_{21}=2m+1,\ R_{31}=3m+1+2/4,\\ R_{12}=m+3/4,\ R_{22}=2m+1+2/4,\ R_{32}=3m+2+1/4,\\ R_{13}=m+1,\ R_{23}=2m+2,\ R_{33}=3m+3.\end{gathered}$$

The ranks for different sample sizes provided by this method are tabulated below:

	$n=4m$	$n=4m+1$	$n=4m+2$	$n=4m+3$
R_{ir}	$r=0$	$r=1$	$r=2$	$r=3$
R_{1r}	$m+1/4$	$m+2/4$	$m+3/4$	$m+1$
R_{2r}	$2m+2/4$	$2m+1$	$2m+1+2/4$	$2m+2$
R_{3r}	$3m+3/4$	$3m+1+2/4$	$3m+2+1/4$	$3m+3$

Segmentation identities are given by

$$m+0R_{10}+m+0R_{20}+m+0R_{30}+m=4m,$$
$$m+0R_{11}+m+R_{21}^{0}+m+0R_{31}+m=4m+1,$$
$$m+0R_{12}+(m+1)+0R_{22}+(m+1)+0R_{32}+m=4m+2,$$
$$m+R_{13}^{0}+m+R_{23}^{0}+m+R_{33}^{0}+m=4m+3.$$

A rank R_{ir} appearing as $R_{ir}^{0}=1$ in the segmentation identity implies that the rank is an integer, and a rank R_{ir} appearing as $0R_{ir}=0$ implies that the corresponding rank is not an integer. It is seen that the equisegmentation property is satisfied by the Popular Method for *r* = 0, 1, 3 but not for $r=2$. Note that the Lapin Method (Lapin [7], 45-46) is a representation of the Popular Method accommodating simple linear interpolation.

Method 2 (Popular Method with Arithmetic Rounding) This method is based on arithmetic rounding applied to the ranks offered by the Popular Method. The ranks for different sample sizes provided by this method are tabulated below:

	$n=4m$	$n=4m+1$	$n=4m+2$	$n=4m+3$
R_{ir}	$r=0$	$r=1$	$r=2$	$r=3$
R_{1r}	m	$m+1$	$m+1$	$m+1$
R_{2r}	$2m+1$	$2m+1$	$2m+2$	$2m+2$
R_{3r}	$3m+1$	$3m+2$	$3m+2$	$3m+3$

Segmentation identities are given by

$$(m-1)+R_{10}^{0}+m+R_{20}^{0}+(m-1)+R_{30}^{0}+(m-1)=4m$$
$$m+R_{11}^{0}+(m-1)+R_{21}^{0}+(m-1)+R_{31}^{0}+m=4m+1$$
$$m+R_{12}^{0}+m+R_{22}^{0}+(m-1)+R_{32}^{0}+m=4m+2$$
$$m+R_{13}^{0}+m+R_{23}^{0}+m+R_{33}^{0}+m=4m+3$$

It is seen that the equisegmentation property is satisfied by the Popular Method only for *r* = 3.

Method 3 (Mendenhall and Sincich Method) This method suggests to round up the rank of the first quartile provided by the Popular Method if the rank is halfway between two integers. It also suggests rounding down the rank of the third quartile if the rank is halfway between two integers. It is easy to see that the suggestion by Mendenhall and Sincich ([9], p. 54) only applies to samples with size $n=4m+1$. For other sample sizes ranks offered by the Popular Method do not lie exactly in the halfway between two integers, and as such those ranks are the same in both the Popular Method and the Mendenhall and Sincich Method. The ranks for different sample sizes provided by this method are tabulated below:

$n=4m$	$n=4m+1$	$n=4m+2$	$n=4m+3$

R_{ir}	$r=0$	$r=1$	$r=2$	$r=3$
R_{1r}	$m+1/4$	$m+1$	$m+3/4$	$m+1$
R_{2r}	$2m+2/4$	$2m+1$	$2m+1+2/4$	$2m+2$
R_{3r}	$3m+3/4$	$3m+1$	$3m+2+1/4$	$3m+3$

Segmentation identities are given by

$$m+0R_{10}+m+0R_{20}+m+0R_{30}+m=4m$$
$$m+R_{11}^{0}+(m-1)+R_{21}^{0}+(m-1)+R_{31}^{0}+m=4m+1$$
$$m+0R_{12}+(m+1)+0R_{22}+(m+1)+0R_{32}+m=4m+2$$
$$m+R_{13}^{0}+m+R_{23}^{0}+m+R_{33}^{0}+m=4m+3$$

It is seen that the equisegmentation property is satisfied by the Mendenhall and Sincich Method for *r* = 0, 3 but not for *r* = 1, 2.

Method 4 By this method, the ranks of quartiles are given by $R_\alpha = \alpha n/4,\ \alpha = 1,\ 2,\ 3$. Separate the largest integer (i) not exceeding R_α, and decimal part (d) of R_α and write $R_\alpha = i + d$. The α th $(\alpha = 1, 2, 3)$ quartile is finally given by

$$Q_\alpha = x_{(i)} + d\,(x_{(i+1)} - x_{(i)}) = (1-d)\,x_{(i)} + d\,x_{(i+1)},\ \ (\alpha = 1, 2, 3)$$

where $x_{(i)}$ is the *i*th observation. This method is a slight variation of the Popular Method discussed in Section 2. The ranks for different sample sizes provided by this method are tabulated below:

	$n=4m$	$n=4m+1$	$n=4m+2$	$n=4m+3$
R_{ir}	$r=0$	$r=1$	$r=2$	$r=3$
R_{1r}	m	$m+1/4$	$m+1/2$	$m+3/4$
R_{2r}	$2m$	$2m+1/2$	$2m+1$	$2m+1+1/2$
R_{3r}	$3m$	$3m+3/4$	$3m+1+1/2$	$3m+2+1/4$

Segmentation identities are given by

$$(m-1)+R_{10}^{0}+(m-1)+R_{20}^{0}+(m-1)+R_{30}^{0}+m=4m,$$
$$m+0R_{11}+m+0R_{21}+m+0R_{31}+(m+1)=4m+1,$$
$$m+0R_{12}+m+R_{22}^{0}+m+0R_{32}+(m+1)=4m+2,$$
$$m+0R_{13}+(m+1)+0R_{23}+(m+1)+0R_{33}+(m+1)=4m+3.$$

It is seen that the equisegmentation property is not satisfied for any $r = 0,1,2,3$.

Method 5 (Hines and Montgomery [2], p. 18) Ranks of quartiles are given by $R_\alpha = \alpha n/4 + 0.5,\ \alpha = 1,\ 2,\ 3$. Separate the largest integer (*i*) not exceeding R_α, and decimal part (d) of R_α and write $R_\alpha = i + d$. The α th $(\alpha = 1, 2, 3)$ quartile is finally given by

$$Q_\alpha = x_{(i)} + d\,(x_{(i+1)} - x_{(i)}) = (1-d)\,x_{(i)} + d\,x_{(i+1)},\quad (\alpha = 1,2,3)$$

where $x_{(i)}$ is the *i*th observation. This method is a slight variation of Method 4. The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 1/2$	$m + 3/4$	$m + 1$	$m + 1 + 1/4$
R_{2r}	$2m + 1/2$	$2m + 1$	$2m + 1 + 1/2$	$2m + 2$
R_{3r}	$3m + 1/2$	$3m + 1 + 1/4$	$3m + 2$	$3m + 2 + 3/4$

Segmentation identities are given by

$$m + 0R_{10} + m + 0R_{20} + m + 0R_{30} + m = 4m$$
$$m + 0R_{11} + m + R^0_{21} + m + 0R_{31} + m = 4m + 1$$
$$m + R^0_{12} + m + 0R_{22} + m + R^0_{32} + m = 4m + 2$$
$$(m+1) + 0R_{13} + m + R^0_{23} + m + 0R_{33} + (m+1) = 4m + 3$$

It is seen that the equisegmentation property is satisfied by this method for *r* = 0, 1, 2 but not for *r* = 3.

Method 6 (Johnson [5], p. 32) The ranks of the quartiles are given by $R_\alpha = \alpha n/4,\ \alpha = 1,2,3$. Separate the largest integer (*i*) not exceeding R_α, and decimal part (*d*) of R_α and write $R_\alpha = i + d$. If $n/4$ is not an integer, round it up to the next integer and find the corres- ponding ordered observation. If $n/4$ is an integer, calculate the mean of the $(n/4)$th and the next observation. The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 1/2$	$m + 1$	$m + 1$	$m + 1$
R_{2r}	$2m + 1/2$	$2m + 1$	$2m + 1$	$2m + 2$
R_{3r}	$3m + 1/2$	$3m + 1$	$3m + 2$	$3m + 3$

Segmentation identities are given by

$$m + 0R_{10} + m + 0R_{20} + m + 0R_{30} + m = 4m,$$
$$m + R^0_{11} + (m-1) + R^0_{21} + (m-1) + R^0_{31} + m = 4m + 1,$$
$$m + R^0_{12} + (m-1) + R^0_{22} + m + R^0_{32} + m = 4m + 2,$$
$$m + R^0_{13} + m + R^0_{23} + m + R^0_{33} + m = 4m + 3.$$

It is seen that the equisegmentation property is satisfied by this method for *r* = 0, 3 but not for *r* = 1, 2.

Method 7 (Hinge Method) An interesting method to find extreme quartiles is based on finding the median first, and then finding the medians of upper and lower halves of the data . The tradition is to count the median in both halves (Mayer and Sykes [8], p. 25). Tukey ([16], pp. 32-35) called them hinges.

For $n = 4m$, it follows that the rank of the median is $R_{20} = (1+4m)/2 = 2m + 2/4$. Then by the Hinge Method $R_{10} = [1 + (2m + 2/4)]/2 = m + 3/4$ and $R_{30} = [(2m + 2/4) + 4m]/2 = 3m + 1/4$. For $n = 4m + 1$, it follows that the rank of the median is $R_{21} = [1 + (4m + 1)]/2 = 2m + 1$. Then by the Hinge Method $R_{11} = [1 + (2m + 1)]/2 = m + 1$ and $R_{31} = [(2m + 1) + (4m + 1)]/2 = 3m + 1$.

For $n = 4m + 2$, it follows that the rank of the median is $R_{22} = [1 + (4m + 2)]/2 = 2m + 1 + 2/4$. Then by the Hinge Method $R_{12} = [1 + (2m + 1 + 2/4)]/2 = m + 1 + 1/4$ and $R_{32} = [(2m + 1 + 2/4) + (4m + 2)]/2 = 3m + 1 + 3/4$. For $n = 4m + 3$, it follows that the median is $R_{23} = [1 + (4m + 3)]/2 = 2m + 2$. Then by the Hinge Method $R_{13} = [1 + (2m + 2)]/2 = m + 1 + 2/4$ and $R_{33} = [(2m + 2) + (4m + 3)]/2 = 3m + 2 + 2/4$.

The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 3/4$	$m + 1$	$m + 1 + 1/4$	$m + 1 + 2/4$
R_{2r}	$2m + 2/4$	$2m + 1$	$2m + 1 + 2/4$	$2m + 2$
R_{3r}	$3m + 1/4$	$3m + 1$	$3m + 1 + 3/4$	$3m + 2 + 2/4$

Segmentation identities are given by

$$m + 0R_{10} + m + 0R_{20} + m + 0R_{30} + m = 4m,$$
$$m + R^0_{11} + (m-1) + R^0_{21} + (m-1) + R^0_{31} + m = 4m + 1,$$
$$(m+1) + 0R_{12} + m + 0R_{22} + m + 0R_{32} + (m+1) = 4m + 2,$$
$$(m+1) + 0R_{13} + m + R^0_{23} + m + 0R_{33} + (m+1) = 4m + 3.$$

Clearly the equisegmentation property is satisfied by the Hinge Method only for $r = 0$.

Method 8 (Vinning Method) The formulae given by Vinning ([17], p. 44) can be simplified as

$$Q_1 = \begin{cases} (n+3)/4 \text{ th observation if } n \text{ is odd} \\ (n+2)/4 \text{ th observation if } n \text{ is even} \end{cases}$$

$$Q_3 = \begin{cases} (3n+1)/4 \text{ th observation if } n \text{ is odd} \\ (3n+2)/4 \text{ th observation if } n \text{ is even} \end{cases}$$

The example he provides with $n = 35$ divides the ordered sample observations into four segments with 9, 8, 8 and 9 observations among them. The median has an integer rank namely the 18th position. The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 2/4$	$m+1$	$m+1$	$m+1+2/4$
R_{2r}	$2m + 2/4$	$2m+1$	$2m+1+2/4$	$2m+2$
R_{3r}	$3m + 2/4$	$3m+1$	$3m+2$	$3m+2+2/4$

Segmentation identities are given by

$$m + 0R_{10} + m + 0R_{20} + m + 0R_{30} + m = 4m$$
$$m + R_{11}^0 + (m-1) + R_{21}^0 + (m-1) + R_{31}^0 + m = 4m+1$$
$$m + R_{12}^0 + m + 0R_{22} + m + R_{32}^0 + m = 4m+2$$
$$(m+1) + 0R_{13} + m + R_{23}^0 + m + 0R_{33} + (m+1) = 4m+3$$

Clearly the equisegmentation property is satisfied by the Vinning Method only for $r = 0$ and $r = 2$. Milton and Arnold Method ([11], pp. 207-208) suggested the ranks of extreme quartiles to be $R_{1r} = ([(n+1)/2]+1)/2$ and $R_{3r} = n+1-R_1$ but it turns out that they are exactly the same as the ranks of extreme quartiles given by the Vinning Method.

Method 9 (Siegel Method) Siegel ([14], p. 117) suggests the ranks of extreme quartiles to be $R_{1r} = ([(n+1)/2]+1)/2$ and $R_{3r} = n+1-R_1$ while, unlike any other method, he suggests the rank of the median to be $R_{2r} = [(n+1)/2]$ where $[a]$ is the largest integer not exceeding a. The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 2/4$	$m+1$	$m+1$	$m+1+2/4$
R_{2r}	$2m$	$2m+1$	$2m+1$	$2m+2$
R_{3r}	$3m + 2/4$	$3m+1$	$3m+2$	$3m+2+2/4$

Segmentation identities are given by

$$m + 0R_{10} + (m-1) + R_{20}^{0} + m + 0R_{30} + m = 4m$$
$$m + R_{11}^{0} + (m-1) + R_{21}^{0} + (m-1) + R_{31}^{0} + m = 4m + 1$$
$$m + R_{12}^{0} + (m-1) + R_{22}^{0} + m + R_{32}^{0} + m = 4m + 2$$
$$(m+1) + 0R_{13} + m + R_{23}^{0} + m + 0R_{33} + (m+1) = 4m + 3$$

Clearly the equisegmentation property is not satisfied for any value of r .

Method 10 (Smith Method) The formulae provided for percentiles by Smith ([15], pp. 36-38) can be specialized to quartiles as

$$Q_1 = \begin{cases} \dfrac{n+2}{4} \text{ th observation if } n/4 \text{ is not an integer} \\ \dfrac{1}{2}\left(\dfrac{n}{4}\text{th} + \dfrac{n+4}{4}\text{th} \right) \text{ observation if } n/4 \text{ is an integer} \end{cases}$$

$$Q_3 = \begin{cases} \dfrac{3n+2}{4} \text{ th observation if } 3n/4 \text{ is not an integer} \\ \dfrac{1}{2}\left(\dfrac{3n}{4}\text{th} + \dfrac{3n+4}{4}\text{th} \right) \text{ observation if } n/4 \text{ is an integer} \end{cases}$$

He suggests rounding the ranks to the nearest integer. The example he provides for $n = 12$ does satisfy the equisegmentation property with $m = 3$. The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 1/2$	$m + 3/4$	$m + 1$	$m + 1 + 1/4$
R_{2r}	$2m + 2/4$	$2m + 1$	$2m + 1 + 2/4$	$2m + 2$
R_{3r}	$3m + 2/4$	$3m + 1 + 1/4$	$3m + 2$	$3m + 2 + 3/4$

Segmentation identities are given by

$$m + 0R_{10} + m + 0R_{20} + m + 0R_{30} + m = 4m,$$
$$m + 0R_{11} + m + R_{21}^{0} + m + 0R_{31} + m = 4m + 1,$$
$$m + R_{12}^{0} + m + 0R_{22} + m + R_{32}^{0} + m = 4m + 2,$$
$$(m+1) + 0R_{13} + m + R_{23}^{0} + m + 0R_{33} + (m+1) = 4m + 3.$$

Clearly the equisegmentation property is satisfied by the Vinning Method for *r* = 0, 1, 2 but not for $r = 3$.

Method 11 (Shao Method) It is surprising that the method proposed by Shao ([13], 1976, pp.174-175) is the only method in the literature that enjoys the equisegmentation property.

a) If the sample size is divisible by 4, the quartiles can be easily determined. When a quartile is located between two values, the mid point of these two values is considered to be the quartile.

b) If the sample size is not divisible by 4, the quartiles can easily be determined in three steps:

(1) If the sample size is even, Q_1 is the median obtained from the lower 50% values of the sample.

(2) If the sample size is odd, Q_1 is the median obtained from the lower 50% values of the sample after having discarded the median of the complete sample.

(3) Locate Q_3 by the methods stated in (1) and (2) except that the upper 50% of the values of the sample are used in the process.

The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 2/4$	$m + 2/4$	$m+1$	$m+1$
R_{2r}	$2m + 2/4$	$2m+1$	$2m+1+2/4$	$2m+2$
R_{3r}	$3m + 2/4$	$3m+1+2/4$	$3m+2$	$3m+3$

Segmentation identities are given by

$$m + 0R_{10} + m + 0R_{20} + m + 0R_{30} + m = 4m,$$
$$m + 0R_{11} + m + R^0_{21} + m + 0R_{31} + m = 4m+1,$$
$$m + R^0_{12} + m + 0R_{22} + m + R^0_{32} + m = 4m+2,$$
$$m + R^0_{13} + m + R^0_{23} + m + R^0_{33} + m = 4m+3.$$

Observe that the equisegmentation property is satisfied by this method for any value of r.

4. Suggested Methods

We discuss below two methods namely the Halving Method and the Remainder Method each of which satisfies the equisegmentation property.

4.1 The Halving Method

We observe that Method 7 guarantees the equisegmentation property if the median of the whole data set is always ignored in the calculation of the lower and upper quartiles. Method 1 with this kind of adjustment will hereinafter be called the Halving Method (Joarder [3]).

Example 4.1 We calculate below the quartiles of the data in Example 2.1 by the Halving Method. The rank of the median is $R_{22} = (1+n)/2 = 5.5$ so that

$$Q_{22} = x_{(5.5)} = (1-0.5)\, x_{(5)} + 0.5\, x_{(6)} = 0.5(3.9) + 0.5(4.7) = 4.3 \ .$$

The first quartile is the median of the observations below the median of the whole sample, i.e. $R_{12} = (1+5)/2 = 3$ so that $Q_{12} = x_{(3)} = 2.$ The third quartile is the median of the observations above the median of the whole sample i.e. $R_{32} = (6+10)/2 = 8$ so that $Q_{32} = x_{(8)} = 7.6$. To check the equisegmentation property, we show the position of the quartiles by downward arrows in the sample:

		⇓		⇓		⇓				
1.7	1.9	2.0	2.8	3.9	4.7	6.2	7.6	12.1	29.3

We observe that there are $2(=m)$ observations in each of the four segments, i.e. the quartiles do satisfy the equisegmentation property. The ranks for different sample sizes provided by this method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 2/4$	$m + 2/4$	$m+1$	$m+1$
R_{2r}	$2m + 2/4$	$2m+1$	$2m+1+2/4$	$2m+2$
R_{3r}	$3m + 2/4$	$3m+1+2/4$	$3m+2$	$3m+3$

It may be remarked here that the first quartile is the median of the smallest *n*/2 observations if *n* is even, and that of the smallest $(n-1)/2$ observations if *n* is odd. Similarly the third quartile is the median of the largest *n*/2 observations if *n* is even, and that of the largest $(n-1)/2$ observations if *n* is odd.

4.2 The Remainder Method

We observe that each of the ranks R_{10} and R_{30} given by the Halving Method is smaller than that given by the Remainder Method by $1/4$. We also observe that the ranks of the quartiles given by the Popular Method satisfy the equisegmentation property if the rank is rounded down for $(r = 2, d = 1)$ and rounded up for $(r = 2, d = 3)$. A special kind of rounding applied to the ranks provided by the Popular Method for quantiles of even order has been discussed by Joarder [4]. The ranks obtained this way, called the Remainder Method, satisfy the equisegmentation property.

Let $[u]$ be the largest integer not exceeding u, and $\lceil u \rceil$ the smallest integer exceeding u. Again let $u = u_{ir} = i(r+1)/4$ and $i(r+1) = d_{ir} \pmod 4$. Then we have the following theorem.

Theorem 4.1 Let $m = (n-r)/4,\ n = 4m + r \geq 4$, and R_{ir} be the rank of the ith quartile with m observations in each segment. Then the ranks given by

$$R_{ir} = \begin{cases} im + [u_{ir}] & \text{if } (r,d) \in A \text{ and } d \leq 2, \quad (4.1a) \\ im + \lceil u_{ir} \rceil & \text{if } (r,d) \in A \text{ and } d > 2, \quad (4.1b) \\ im + u_{ir} & \text{if } (r,d) \notin A \quad (4.1c) \end{cases}$$

where i and r are integers with $1 \leq i \leq 3$ and $0 \leq r \leq 3$, and $A = \{(r,d) : r = 2,\ d = 1,3\}$, satisfy the equisegmentation property. If $(r,d) \notin A$, then the quartiles can be calculated by the simple linear interpolation as

$$Q_{ir} = (1 - d/4)\, x_{(im+[u])} + (d/4)\, x_{(im+[u]+1)},$$

where $x_{(i)}$ is the *i*th ordered observation.

Example 4.2 Let us now calculate the quartiles for the sample in Example 2.1 by the Remainder Method. Here $n = 10 = 4(2) + 2$ i.e. $m = 2,\ r = 2$. Since $u_{12} = 1(2+1)/4 = 3/4$ (i.e. $r = 2,\ d = 3 > 2$), the rank of the first quartile is $R_{12} = 1(m) + \lceil u_{12} \rceil = 2 + \lceil 3/4 \rceil = 3$ (see 4.1b). Again since $u_{22} = 2(2+1)/4 = 1 + 2/4$ (i.e. $r = 2,\ d = 2 \leq 2$), the rank of the second quartile is $R_{22} = 2(m) + u_{22} = 2(2) + 1 + 2/4 = 5.5$ (see 4.1 c). Finally since $u_{32} = 3(2+1)/4 = 2 + 1/4$ (i.e. $r = 2,\ d = 1 \leq 2$), the rank of the third quartile is $R_{32} = 3(m) + [u_{12}] = 6 + 2 + [1/4] = 8$ (See 4.1a). So the quartiles are $Q_{12} = x_{(3)} = 2.0$, $R_{22} = (1 - 0.5)x_{(5)} + 0.5\, x_{(6)} = (3.9 + 4.7)/2 = 4.3$, and $R_{32} = x_{(8)} = 7.6$.

To check the equisegmentation property, we show the position of the quartiles by downward arrows in the sample:

$$\begin{array}{cccccccccc} & & \Downarrow & & & \Downarrow & & & \Downarrow & \\ 1.7 & 1.9 & 2.0 & 2.8 & 3.9 & & 4.7 & 6.2 & 7.6 & 12.1 & 29.3 \end{array}$$

We observe that the equisegmentation property is satisfied here with $m = 2$. The ranks of the quartiles for different sample sizes given by the Remainder Method are tabulated below:

	$n = 4m$	$n = 4m+1$	$n = 4m+2$	$n = 4m+3$
R_{ir}	$r = 0$	$r = 1$	$r = 2$	$r = 3$
R_{1r}	$m + 1/4$	$m + 2/4$	$m + 1$	$m + 1$

R_{2r}	$2m+2/4$	$2m+1$	$2m+1+2/4$	$2m+2$
R_{3r}	$3m+3/4$	$3m+1+2/4$	$3m+2$	$3m+3$

The Halving Method as well as the Remainder Method satisfies the equisegmentation property. It is worth noting that in each of the two methods the value of $r\,(n=4m+r)$ is the number of quartiles with integer ranks. The Shao Method, however, doesn't have algebraic expression for the ranks and hence may not be suitable for using it or generalizing it to other quantiles. Though the Halving Method is the simplest one and satisfies the equisegmentation property, it seems to be difficult to generalize the notion to quantiles in general. Note that the Remainder Method for quartiles happens to be the Popular Method with arithmetic rounding for outer quartiles when $r=2$. The Remainder Method is generalized to quantiles of even order by Joarder [4].

It remains open to check the equisegmentation property for samples with ties. Finally it is worth remarking that for a sample of large size, the empirical cumulative distribution function may be used to calculate sample quartiles (Mendenhall *et al*. [10], Section 15.1.1).

Acknowledgements

The authors acknowledge King Fahd University of Petroleum and Minerals, Saudi Arabia, for providing excellent research facilities. The authors are grateful to Prof. M. M. Ali, Ball State University, USA, for constructive suggestions that have improved the quality of the paper.

References

[1] Bluman, A. G. (2001). Elementary Statistics: A Step by Step Approach. McGraw Hill, New York.

[2] Hines, W. and Montgomery, D. C. (1990). Probability and Statistics in Engineering and Management Sciences, New York: John Wiley.

[3] Joarder, A. H. (2003). The halving method for sample quartiles, *International Journal of Mathematical Education in Science and Technolog*, 34(4), 629-633.

[4] Joarder, A. H. (2002). The remainder method for sample quantiles of even order, *Technical Report No*. 274, King Fahd University of Petroleum and Minerals, Saudi Arabia.

[5] Johnson, R. (2000). Miller and Freund's Probability and Statistics for Engineers. Prentice Hall.

[6] Lapin, L. L. (1975). Statistics: Meaning and Method, New York: Harcourt Brace Jovanovich, Inc.

[7] Lapin, L. L. (1997). Modern Engineering Statistics, Duxbury Press.

[8] Mayer, A. D. and Sykes, A. M. (1996). Statistics, London: Edward Arnold.

[9] Mendenhall, W. and Sincich, T. (1995). Statistics for Engineering and the Sciences. Englewood Cliffs, NJ: Prentice Hall.

[10] Mendenhall, W., Beaver, R. J. and Beaver, B. M. (2002). A Brief Introduction to Probability and Statistics, United States: Duxbury.

[11] Milton, J. S. and Arnold, J. C. (2003). Introduction to Probability and Statistics, New
York: McGraw-Hill.

[12] Ostle, B., Turner, K. V., Hicks, C. R. and McElrath, G. W. (1996). Engineering Statistics: The Industrial Experience, New York: Duxbury Press.

[13] Shao, S. P. (1976). Statistics for Business and Economics, Columbus, Ohio: Charles
E. Merill Publishing Co.

[14] Siegel, A. F. (2003). Practical Business Statistics, New York :McGraw-Hill.

[15] Smith, P. J. (1997). Into Statistics, Singapore: Springer-Verlag.

[16] Tukey, J. W. (1977). Exploratory Data Analysis, Reading, MA: Addison Wesley.

[17] Vinning, G. G. (1998). Statistical Methods for Engineers, New York: Duxbury Press.

The remainder method for sample quantiles of even order*

A.H. JOARDER and M. FIROZZAMAN[1]

Dept of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia 31261, Emails: anwarj@kfupm.edu.sa , firoz@mathla.asu.edu

A new method called the Remainder Method is proposed for the calculation of sample quantiles of a given order, for example, quartiles, deciles etc assuming that the observations are all distinct. Proof is given for a special case of deciles. We propose the new criterion of equisegmentation property that the number of observations below the first quantile, that between the consecutive quantiles, and that above the last quantile are the same. The proposed method satisfies the equisegmentation property , and more interestingly provides the number of quantiles having integer ranks. Some open problems are indicated.

2. Introduction

A sample quantile is a point below which some specified proportion of the values of a data set lies. The median is the 0.50 quantile because approximately half of all observations lie below this value. The name fractile for quantile is used by some authors (see Lapin, 1975, 52). Quartiles, hexatiles, octatiles, deciles, percentiles are special cases of quantiles.

One method for quartiles, called the hinges (Tukey, 1976), is based on finding the median first and then finding the medians of the upper and lower halves of the data each time including the median of the whole data set. Done so, approximately 25% observations remain below the lower quartile and 25% above the upper quartile. The literature is full of different formulae for sample quartiles with various rounding notions of the corresponding ranks of quartiles . See for example Mendenhall and Sincich (1995, 54), and Joarder and Firozzaman (2001) for a detailed survey and illustrations. Joarder (2003) discussed halving method of sample quartiles that satisfies equisegmentation property but it seems rather difficult to generalize it to quantiles of higher order.

A new method called the Remainder Method has been proposed for the calculation of sample quantiles of some order say quartiles, hexatiles, octatiles, deciles etc. In this note we propose, by the use of remainders, the correct form of rounding of the ranks of quantiles of even order based on any sample with distinct n values. Proof is given for a special case of deciles.

We propose the new criterion of equisegmentation property that the number of observations below the first quantile, that between the consecutive quantiles, and

---

* Present Address: Dept of Mathematics, Arizona University, Tempe, USA

that above the last quantile are the same. Let the number of observations in each segment be $m_i \ \ (i = 1,2,3,...,f)$. Then the equisegmentation property guarantees that $m_1 = m_2 = \cdots = m_f$. However this will divide the ordered sample observations into desired number of segments leaving the same number of observations in each if the observations are distinct.

Consider the quantiles of even order say $f = 2, 4, 6, ...,$ that divides the ordered sample observations in f divisions with m observations in each segment. Since the sample size can be represented by

$$n = r \bmod f = fm + r, \ (r = 0, 1, 2, \cdots, f-1), \tag{1.1}$$

the number of observations in each of the $f \le n$ segments is given by

$$m(r) = (n - r)/f \tag{1.2}$$

or m for short. The ranks $R_{ir} \ (i = 1, 2, \cdots, f-1; \ r = 0, 1, \cdots, f-1)$ for quantiles of order f satisfies equisegmentation property if

$$(i) \ \lceil R_{1r} \rceil - 1 = m \tag{1.3a}$$

$$(ii) \ \lceil R_{1r} \rceil - [R_{i-1,r}] - 1 = m, \ i = 2,3, \cdots, f \tag{1.3b}$$

$$(iii) \ fm + r - [R_{f-1,r}] = m \tag{1.3c}$$

where $[x]$ and $\lceil x \rceil$ are the floor function (largest integer not exceeding x) and the ceiling function (smallest integer at least as large as x) of x. The equation $(1.3a)$ states that the number of observations below the first quantile is m while the equation $(1.3c)$
states that the number of observations above the third quantile is m. The equation $(1.3b)$ states that the number of observations between two consecutive quantiles is m. Interestingly the quantity r is also the number of quantiles with integer ranks.

It remains open to come up with a general method for ranks of quantiles of even order. The Remainder Method for quartiles $(f = 4)$, hexatiles $(f = 6)$, octatiles $(f = 8)$ and deciles $(f = 10)$ have been discussed in Sections 2 and 3. The method has been proved for a special case of deciles in Section 2.4, and argued that proofs for all other cases are similar. A set of general formula for quantiles of even order, in particular, or quantiles of any order, in general, remains open to be addressed.

Sample quartiles are popularly interpolated linearly by the observations corresponding to the ranks $i(n+1)/4 \ , \ (i = 1,2,3)$. This method will hereinafter be called the Popular Method. Joarder (2003) observed that the ranks provided by this method do not satisfy equisegmentation property if sample sizes are $n = 6,10,14$ etc. This led us to

conjecture that the remainder of the sample size with respect modulus 4 may play a role in the determination of the ranks for quartiles.

Let R_{ir} be the rank of i th quartile with m observations in each of the f (which is 4 for quartiles) segments. Then

$$R_{ir} = i\ \frac{n+1}{f} = \frac{(fm+r)+1}{f} = im + \ i(r+1)/f = im + [u_{ir}] + d/f \tag{1.1}$$

where i and r are integers with $1 \le i \le f-1$, $0 \le r \le f-1$, $[u_{ir}]$ is the largest integer not exceeding (less than or equal to) $u_{ir} = i(r+1)/f = [u_{ir}] + d/f$. For simplicity we will often use u for u_{ir}. The quartiles can then be calculated by the simple linear interpolation as

$$Q_{ir} = (1-d/f)\, x_{(im+[u])} + (d/f)\, x_{(im+[u]+1)}\ , \tag{1.2}$$

where $x_{(i)}$ is the *i* th ordered observation. Note that $u = d/f$ if $[u_{ir}] = 0$ i.e. if $u < 1.$

2. The remainder method for sample quartiles, hexatiles, octatiles and deciles

2.1 The remainder method for quartiles

The refinement of the formulae for quartiles is based on the equisegmentation property discussed in Section 1. With a view to improving upon the rank of quartiles given by the Popular Method so that equisegmentation property is satisfied, a special notion of rounding depending on the remainder r and d of the ranks of quartiles is considered. The following theorem is considered in Joarder and Latif (2004).

Theorem 2.1 Let $m = (n-r)/4,\ n = 4m + r \ge 4$, and R_{ir} be the rank of the i th quartile with m observations in each segment. Then the ranks

$$R_{ir} = \begin{cases} im + [u_{ir}] \text{ if } (r,d) \in A, \text{and } d \le 2 & (2.1a) \\ im + \lceil u_{ir} \rceil \text{ if } (r,d) \in A, \text{and } d > 2 & (2.1b) \\ im + u_{ir} \quad \text{if } (r,d) \notin A & (2.1c) \end{cases}$$

where i and r are integers with $1 \le i \le 3$ and $0 \le r \le 3$, $u_{ir} = i(r+1)/4 = [u_{ir}] + d/4$, and $A = \{(r,d) : (2,\ 1),\ (2,\ 3)\}$ satisfy equisegmentation property.

Example 2.1 An independent consumer group tested radial tires from a major brand to determine expected tread life. The data (in thousands of miles) are given below:

50 54 52 47 61 54.5 50.5 51

48 55 53 43 56 58 42

(cf. Vinning, 1998, 193). The ordered sample observations are given by

42 43 47 48 50 50.5 51 52
53 54 54.5 55 56 58 61

To illustrate the proposed formulae we make four different data sets with the first $n=12,\ n=13,\ n=14,\ n=15$ observations and label them as **Data 1**, **Data 2**, **Data 3** and **Data 4** respectively. We observe that the popular method satisfies equisegmentation property for all the above data sets except **Data 3**. We show below how the Remainder Method can be applied for **Data 3** for sample quartiles and that it satisfies the equisegmentation property:

Here the sample size is $n=14=4(3)+2$ i.e. $m=3, r=2$. The ranks for the quartiles are given by R_{12}, R_{22}, R_{32}. Since $u_{12}=1(2+1)/4=3/4$, $(r,d)=(2,3)\in A$ and $d=3>2$, it follows from $(2.1b)$ that $R_{12}=1m+\lceil u_{12}\rceil=3+1$. Again since $u_{22}=2(2+1)/4=1+2/4$ and $(r,d)=(2,2)\notin A$, it follows from $(2.1c)$ that $R_{22}=2m+u_{22}=6+1+2/4=7+2/4$. Also since $u_{32}=3(2+1)/4=2+1/4$, $(r,d)=(2,1)\in A$ and $d=1<2$, it follows from $(2.1a)$ that $R_{32}=3m+[u_{32}]=9+2$. Thus by the Remainder Method ranks for quartiles are given by $R_{12}=4,\ R_{22}=7+2/4,\ R_{32}=11$. Clearly the ranks satisfy equisegmentation property. The position of quartiles for **Data 3** given by

$Q_{12}=4\text{ th obs}=48$.

$Q_{22}=7+2/4\text{ th obs}=(1-2/4)(7\text{ th obs})+(2/4)(8\text{ th obs})=0.50(51)+0.50(52)=51.5$

$Q_{32}=11\text{ th obs}=56$

are shown below:

$$x_{(1)}, x_{(2)}, x_{(3)}, \underset{\Downarrow}{x_{(4)}}, x_{(5)}, x_{(6)}, x_{(7)}, \Downarrow x_{(8)}, x_{(9)}, x_{(10)}, \underset{\Downarrow}{x_{(11)}}, x_{(12)}, x_{(13)}, x_{(14)}$$

where $x_{(i)}$'s are the ordered sample observations. The quartiles $Q_{ir}\ (i=1,2,3)$ for a particular r are the usual quartiles and are popularly denoted by simply Q_1, Q_2, Q_3. Clearly the above rounding of ranks guarantees the desirable equisegmentation property.

Here there are $m=3$ observations in each segment. The remainder $r(=2)$ is also, as expected, the number of quartiles having integer ranks for any sample of size $n\geq 4$.

2.2 The remainder method for hexatiles

Hexatiles are five numbers that divide ordered sample observations into six segments. The following theorem guarantees that the ranks for hexatiles given by the Remainder Method satisfy equisegmentation property.

Theorem 2.2 Let $m=(n-r)/6,\ n=6m+r\geq 6$, and R_{ir} be the rank of the ith hexatile with m observations in each segment. Then the ranks of hexatiles given by

$$R_{ir}=\begin{cases} im+[u_{ir}] \text{ if } (r,d)\in A, \text{and } d\leq 3 & (2.2a)\\ im+\lceil u_{ir}\rceil \text{if } (r,d)\in A, \text{and } d>3 & (2.2b)\\ im+u_{ir} \text{ if } (r,d)\notin A & (2.2c)\end{cases}$$

where i and r are integers with $1\leq i\leq 5$ and $0\leq r\leq 5$, $u_{ir}=i(r+1)/6=[u_{ir}]+d/6$ and
$A=\{(r,d):(3,2),(4,1),(4,2),(4,4),(4,5)\}$ satisfy the equisegmentation property.

Example 2.2 Consider **Data 4** with sample size $n=15=6(2)+r$ i.e. $m=2,\ r=3$. The ranks for the hexatiles are given by R_{13} , R_{23} , R_{33}, R_{43} , R_{53} . Since $u_{13}=1(3+1)/6=4/6$, $(r,d)=(3,4)\notin A$, it follows from (2.2c) that $R_{13}=1m+u_{13}=2+4/6$. Again since $u_{23}=2(3+1)/6=1+2/6$ and $(r,d)=(3,2)\in A$ and $d<3$, it follows from ($2.2a$) that $R_{23}=2m+[u_{23}]=2(2)+1=5$. Also since $u_{33}=3(3+1)/6=2+0/6$, $(r,d)=(3,0)\notin A$, it follows from ($2.2c$) that $R_{33}=3m+u_{33}=3(2)+2=8$. Thus by the Remainder Method ranks for hexatiles are given by $R_{13}=2+4/6,\ R_{23}=5,\ R_{33}=8, R_{43}=10+4/6, R_{53}=13$. Clearly the ranks satisfy equisegmentation property. The hexatiles for **Data 4** given by
$Q_{13}=(2/6)(43)+(4/6)(47)\approx 45.67,\ Q_{23}=50,\ \mathrm{Q}_{33}=52$
$Q_{43}=(2/6)(54)+(4/6)(56)\approx 55.33,\ Q_{53}=56$

are shown below:

$$x_{(1)},x_{(2)},\Downarrow x_{(3)},x_{(4)},\overset{\Downarrow}{x_{(5)}},x_{(6)},x_{(7)},\overset{\Downarrow}{x_{(8)}},x_{(9)},x_{(10)},\Downarrow x_{(11)},x_{(12)},\overset{\Downarrow}{x_{(13)}},x_{(14)},x_{(15)}$$

where $x_{(i)}$'s are the ordered sample observations. Clearly the above rounding of ranks guarantees the desirable equisegmentation property. There are $m=2$ observations in each segment here. The remainder $r(=3)$ is also, as expected, the number of hexatiles having integer ranks for any sample of size $n\geq 6$.

2.3 The remainder method for octatiles

Octatiles are seven numbers that divide ordered sample observations into eight segments. The following theorem guarantees that the ranks for octatiles given by the Remainder Method satisfy equisegmentation property.

Theorem 2.3 Let $m=(n-r)/8,\ n=8m+r\geq 8$, and R_{ir} be the rank of the ith quartile with m observations in each segment. Then the ranks for octatiles are given by

$$R_{ir} = \begin{cases} im + [u_{ir}] \text{ if } (r,d) \in A, \text{and } d \le 4 & (2.3a) \\ im + \lceil u_{ir} \rceil \text{ if } (r,d) \in A, \text{and } d > 4 & (2.3b) \\ im + u_{ir} \text{ if } (r,d) \notin A & (2.3c) \end{cases}$$

where i and r are integers with $1 \le i \le 7$ and $0 \le r \le 7$, $u_{ir} = i(r+1)/8 = [u_{ir}] + d/8$ and

$$A = \{(r,d) : (2,1), (2,2), (4,1), (4,2), (4,3), (4,4), (5,2), (5,6), (6,1), (6,2), (6,3), (6,5), (6,6), (7,7)\}$$

satisfy the equisegmentation property.

2.4 The remainder method for deciles

Deciles are nine numbers that divide ordered sample observations into ten segments. The following theorem guarantees that the ranks for deciles given by the Remainder Method satisfy the equisegmentation property.

Theorem 2.4 Let $m = (n-r)/10,\ n = 10m + r \ge 10$, and R_{ir} be the rank of the ith decile with m observations in each segment. Then the ranks for deciles are given by

$$R_{ir} = \begin{cases} im + [u_{ir}] \text{ if } (r,d) \in A, \text{ and } d \le 5 & (2.4a) \\ im + \lceil u_{ir} \rceil \text{ if } (r,d) \in A, \text{ and } d > 5 & (2.4b) \\ im + u_{ir} \text{ if } (r,d) \notin A & (2.4c) \end{cases}$$

where i and r are integers with $1 \le i \le 9$, $0 \le r \le 9$ and $u_{ir} = i(r+1)/10 = [u_{ir}] + d/10$ and

$$A = \{(r,d) : (2,1), (2,2), (3,2), (5,2), (5,4), (6,1), (6,2), (6,3), (6,4)(6,8), (6,9), (7,2), (7,4), (7,8), (8,1), (8,2), (8,3), (8,4), (8,6), (8,7), (8,8), (8,9)\}$$

satisfy equisegmentation property.

Proof. By writing out the ranks for deciles by (1.1) with $f = 10$, it is easy to observe that no rounding is needed for $r = 0, 1, 4, 9$ i.e. the ranks are given by 2.4 (c). For other cases of $r = 2,5,6,7,8$, some ranks need to be rounded so that the deciles satisfy equisegmentation property . Since proofs are similar in all cases of $r = 2,5,6,7,8$ we prove the theorem for the special case of $r = 6$.

Let $n = 10(m) + 6$ so that $r = 6$. Then by Theorem 2.4 the ranks for deciles are given by

$R_{16} = 1m + 1(6+1)/10 = m + 7/10\,,\ \text{since}\, d = 7,\ (r,d) = (6,7) \notin A$
$R_{26} = 2m + [2(6+1)/10] = 2m + [1 + 4/10] = 2m + 1,\ \text{since}\, d = 4 < 5,\ (r,d) = (6,4) \in A$

$R_{36} = 3m + [3(6+1)/10] = 3m + [2 + 1/10] = 3m + 2,\ \text{since}\, d = 1 < 5,\ (r,d) = (6,1) \in A$
$R_{46} = 4m + \lceil 4(6+1)/10 \rceil = 4m + \lceil 2 + 8/10 \rceil = 4m + 3,\ \text{since}\, d = 8 > 5, (r,d) = (6,8) \in A$

$R_{56} = 5m + 5(6+1)/10 = 5m + 3 + 5/10, \text{since}\, d = 5 \le 5, (r,d) = (6,5) \notin A$
$R_{66} = 6m + [6(6+1)/10] = 6m + 4, \text{since}\, d = 2 < 5, (r,d) = (6,2) \in A$

$R_{76} = 7m + \lceil 7(6+1)/10 \rceil = 7m + 5, \text{since}\, d = 9 > 5, (r,d) = (6,9) \in A$
$R_{86} = 8m + 8(6+1)/10 = 8m + 5 + 6/10,\ \text{since}\, d = 6, (r,d) = (6,6) \notin A$

$R_{96} = 9m + [9(6+1)/10] = 9m + 6,\ \ \text{since}\, d = 3 < 6, (r,d) = (6,3) \in A$

Then it is easy to check from the above that

$(i)\ \lceil R_{16} \rceil - 1 = \lceil m + 7/10 \rceil - 1 = (m+1) - 1 = m$

$(ii)\ \lceil R_{26} \rceil - [R_{16}] - 1 = \lceil 2m + 1 \rceil - [m + 7/10] - 1 = (2m+1) - (m) - 1 = m,$
$\lceil R_{36} \rceil - [R_{26}] - 1 = \lceil 3m + 2 \rceil - [2m + 1] - 1 = 3m + 2 - (2m+1) - 1 = m,$
$\lceil R_{46} \rceil - [R_{36}] - 1 = \lceil 4m + 3 \rceil - [3m + 2] - 1 = 4m + 3 - (3m+2) - 1 = m,$
$\lceil R_{56} \rceil - [R_{46}] - 1 = \lceil 5m + 3 + 5/10 \rceil - [4m + 3] - 1 = 5m + 4 - (4m+3) - 1 = m,$
$\lceil R_{66} \rceil - [R_{56}] - 1 = \lceil 6m + 4 \rceil - [5m + 3 + 5/10] - 1 = 6m + 4 - (5m+3) - 1 = m,$
$\lceil R_{76} \rceil - [R_{66}] - 1 = \lceil 7m + 5 \rceil - [6m + 4] - 1 = 7m + 5 - (6m+4) - 1 = m,$
$\lceil R_{86} \rceil - [R_{76}] - 1 = \lceil 8m + 5 + 6/10 \rceil - [7m + 5] - 1 = 8m + 6 - (7m+5) - 1 = m,$
$\lceil R_{96} \rceil - [R_{86}] - 1 = \lceil 9m + 6 \rceil - [8m + 5 + 6/10] - 1 = 9m + 6 - (8m+5) - 1 = m$

$(iii)\ 10m + 6 - [R_{96}] = 10m + 6 - [9m + 6] = 10m + 6 - (9m + 6) = m\,.$

Thus it is proved that ranks of deciles given by the Remainder Method satisfy equisegmentation property.

The idea is also illustrated with a complete example in Firozzaman and Joarder (2001).
The remainder r is also, as expected, the number of deciles having integer ranks for any sample of size $n \ge 10$.

3. Quantiles of even order

Quantiles of order f (even) are $f - 1$ numbers that divide ordered sample observations into f segments. The notion is now conjectured to generalize to

quantiles of even order. Let the sample size be denoted by $n = r \bmod f,\ \ (r = 0, 1, 2, \cdots, f-1;\ f = 2, 3, \cdots)$ and the number of observations in each segment $m = (n-r)/f,\ n = fm + r \geq f > 4$, and $R_{ir}(f, m)$ be the rank of the i th quantile with m observations in each segment.

Theorem 3.1 Let $m = (n-r)/f,\ n = fm + r \geq f$, where f is even and R_{ir} be the rank of the i th quartile with m observations in each segment. Then the ranks for quantiles are given by

$$R_{ir} = \begin{cases} im + [u_{ir}] \ \text{ if } \ (r,d) \in A, \text{ and } d \leq f/2 & (3.1a) \\ im + \lceil u_{ir} \rceil \ \text{ if } \ (r,d) \in A, \text{ and } d > f/2 & (3.1b) \\ im + u_{ir} \ \text{ if } \ (r,d) \notin A & (3.1c) \end{cases}$$

where i and r are integers with $1 \leq i \leq f-1,\ 0 \leq r \leq f-1$
$u_{ir} = i(r+1)/f = [u_{ir}] + d/f$ and $A = \{(r, d_{ir})\}$, a set yet to be determined so that the ranks satisfy the equisegmentation property discussed in Section 1.

(i) **The set A for hexatiles $f = 6$**

The set A for hexatiles (See Theorem 2.2) given by

$$A = \{(r, d_{ir}) : (r = 3, d_{23} = d_{53} = 2), (r = 4, d_{54} = 1), (r = 4, d_{44} = 2), \\ (r = 4, d_{24} = 4), (r = 4, d_{14} = 5)\}$$

can be simply written as

$(a)\ r = 2, 3;\ 1 \leq d \leq r,\ (b)\ r = 4,\ \ d = 1, 2\ \ (c)\ \ r = 4,\ \ d = 4,5$
by suppressing the suffixes of d_{ir}.

(ii) **The set A for octatiles $(f = 8)$**

The set A for octatiles (See Theorem 2.3) given by

$$A = \{(r, d_{ir}) : (r = 2, d_{32} = 1), (r = 2, d_{62} = 2), (r = 4,\ d_{54} = 1), (r = 4, d_{24} = 2), (r = 4, d_{74} = 3), \\ (r = 4, d_{44} = 4), (r = 5, d_{35} = d_{75} = 2), (r = 5, d_{15} = d_{55} = 6), (r = 6, d_{76} = 1), \\ (r = 6, d_{66} = 2), (r = 6, d_{56} = 3), (r = 6, d_{36} = 5), (r = 6, d_{26} = 6), (r = 6, d_{16} = 7)\}$$

can be simply written as

$$(a)\ r = 2, 3, 4;\ 1 \leq d \leq r,\ (b)\ r = 5, 6;\ \ d = 1, 2, 3\ \ (c)\ \ r = 5, 6;\ \ 14 - r \leq d \leq 9$$

(iii) **The set A for deciles $(f = 10)$**

The value of corresponding to any r for which the ranks of deciles needs to be rounded to ensure equisegmentation property are provided in the following table:

Table 3.1

r		Type
2	$d_{72}=1, d_{42}=2$	$(3.1a)$
3	$d_{32}=d_{83}=2$	$(3.1a)$
5	$(d_{25}=d_{75}=2, d_{45}=d_{95}=4)$	$(3.1a)$
6	$(d_{36}=1, d_{66}=2, d_{96}=3, d_{26}=4),$ $(d_{46}=8, d_{76}=9)$	$(3.1b)$ $(3.1c)$
7	$(d_{47}=d_{97}=2, d_{37}=d_{87}=4),$ $(d_{17}=d_{67}=8)$	$(3.1b)$ $(3.1c)$
8	$(d_{98}=1, d_{88}=2, d_{78}=3, d_{68}=4),$ $(d_{46}=6, d_{38}=7, d_{28}=8, d_{18}=9)$	$(3.1b)$ $(3.1c)$

We prepare below Table 4.2 whose 2nd column shows possible values of d_{ir} corresponding to different values of r. The third column shows values d of d_{ir} for which rounding is needed to ensure equisegmentation property.

Table 3.2

r	d_{ir}		Adjoining set	Larger set	Type	Count
0	$1 \le d \le 9$	No need				
1	$0, 2, 4, 6, 8$	No need				
2	$1 \le d \le 9$	1, 2		$1 \le d \le r$	3.1 (a)	2
3	$0, 2, 4, 6, 8$	2	3	$1 \le d \le r$	3.1 (a)	1
4	0, 5	No need	1, 2, 3, 4	$1 \le d \le r$	3.1 (a)	
5	$0, 2, 4, 6, 8$	2, 4	1, 3	$1 \le d \le r$	3.1 (a)	2
6	$1 \le d \le 9$	1, 2, 3, 4 8, 9		$1 \le d \le 4$ $14 - r \le d \le 9$	3.1 (b) 3.1 (c)	4 2
7	$0, 2, 4, 6, 8$	$2,4$ 8	$1,3$ $7,9$	$1 \le d \le 4$ $14 - r \le d \le 9$	3.1 (b) 3.1 (c)	2 1
8	$1 \le d \le 9$	1, 2, 3, 4 6, 7, 8, 9		$1 \le d \le 4$ $14 - r \le d \le 9$	3.1 (b) 3.1 (c)	4 4
9	0	No need				**22**

In view of these relationships between r and d as seen in Table 3.2 the superset of $A = \{(r, d)\}$ defined by

$$
\begin{aligned}
&(a)\ 2 \le r \le 5,\ 1 \le d \le r\ ,\ \text{or} \\
&(b)\ 6 \le r \le 8,\ \ 1 \le d \le 4\ \ \text{or} \\
&(c)\ 6 \le r \le 8,\ 14 - r \le d \le 9
\end{aligned}
\tag{3.2}
$$

(See Column 5 and Column 6 of Table 3.2 for specific explanations) also does the same rounding as done by the set given in Table 3.1 (See Column 2 and Column 3 of Table 4.1). There are 22 distinct sets of (r,d) corresponding to 28 ranks for which rounding are needed in the formulae for ranks of deciles to ensure equisegmentation property.

(iv) **The set A for $f = 20$**

We have also checked the method for $f = 20$ divisions, and determined the set A of $\{(r,d)\}$ points for which rounding is needed to satisfy equisegmentation property (See Appendix 1). The results reported here are based on Joarder (2002) who has written out explicitly the ranks for all quantiles of order $f = 4, 6, 8, 10, 20$ by the formula given by (1.1) and inspected manually the set of $A = \{(r,d)\}$ where

$$
\begin{aligned}
&(a)\ 2 \le r \le 10,\ 1 \le d \le r\ ,\ \text{or} \\
&(b)\ 11 \le r \le 18,\ \ 1 \le d \le 9\ , \text{or} \\
&(c)\ 11 \le r \le 18,\ \ 29 - r \le d \le 19
\end{aligned}
\tag{3.3}
$$

With the above consideration we conjecture that the set $A = \{(r,d)\}$ is given by

$$
\begin{aligned}
&(a)\ 2 \le r \le f/2,\ 1 \le d \le r\ , \\
&(b)\ \ f/2+1 \le r \le f-2,\ \ 1 \le d \le f/2-1, \\
&(c)\ \ f/2+1 \le r \le f-2,\ \ 3f/2-1-r \le d \le f-1
\end{aligned}
\tag{3.4}
$$

It is very interesting to note that the remainder r is also, as expected, the number of quantiles having integer ranks. One limitation of the Remainder Method is that one needs $n \ge f$ observations to satisfy equisegmentation property. We recommend to use the two popular quantiles namely, quartiles or deciles (unless $n \ge f$), and not any other quartiles to overcome the above difficulty. It has been checked that the distinct number of (r,d) points for which rounding is nedded for quartiles, hexatiles, octatiles, deciles, and quantiles with $f = 20$ divisions are given by 2, 6, 14, 22 and 95 respectively. We conclude the article with two open problems:

(1) Prove the conjecture for quantiles with even number of divisions, given in (3.4), or odd number of divisions that satisfy equisegmentation property for a sample with distinct observations, and also generalize it for non-distinct observations.

(2) Determination of the number of distinct (r, d) points for which rounding is essential by the Remainder Method to ensure equisegmentation property for quantiles.

Acknowledgements

The authors are grateful to the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia.

References

Firozzaman, M. and Joarder, A.H. (2001). A refinement over the usual formulae for deciles. *International Journal of Mathematical Education in Science and Technology*. **32** (5), 761-765.

Joarder, A.H. (2002). Sample Quantiles of Even Order. Unpublished Notes.

Joarder, A.H. (2003). The halving method for sample quartiles. *International Journal of Mathematical Education in Science and Technology*. **34**(4), 629-633.

Joarder, A.H. and Firozzaman, M. (2001). Quartiles for discrete data. *Teaching Statistics*, **23** (3), 86-89.

Joarder, A.H. and Latif, R.M. (2004). A comparison and contrast of some methods for sample quartiles. *Journal of Probability and Statistical Science*. 2(1), 95-109.

Lapin, L. (1975). *Statistics: Meaning and Method*. Harcourt Brace Jovanovich, Inc. New York.

Mendenhall, W. and Sincich, T. (1995). *Statistics for Engineering and the Sciences*. John Wiley. New York.

Tukey, J .W. (1977). Exploratory Data Analysis. Reading, MA: Addison Wesley.

Vinning, G.G. (1998). *Statistical Methods for Engineers*. Duxbury Press. New York

Appendix 1 (Set A for Quantiles of order $f = 20$)

r	d	d^*	Type	Count
0	$1 \le d \le 19$			0
1	$0 \le d \le 18$, all even			0
2	$1 \le d \le 19$	$1, 2$	(3.1a)	2
3	$0, 4, 8, 12, 16$			0
4	$0, 5, 10, 15$			0
5	$0 \le d \le 18$, all even	$2, 4$	(3.1a)	2
6	$1 \le d \le 19$	$1 \le d \le 6$	(3.1a)	6
7	$0, 4, 8, 12, 16$	4	(3.1a)	1
8	$1 \le d \le 19$	$1 \le d \le 8$	(3.1a)	8
9	$0, 10$			0
10	$1 \le d \le 19$	$1 \le d \le 9$	(3.1a)	9
11	$0, 4, 8, 12, 16$	$4, 8$	(3.1b)	2
12	$1 \le d \le 19$	$1 \le d \le 9$ $d = 17,18,19$	$(3.1b)$ $(3.1c)$	9 3
13	$0 \le d \le 18$, all even	$2, 4, 6, 8$ $16,18$	$(3.1b)$ $(3.1c)$	4 2
14	$0, 5, 10, 15$	5 15	$(3.1b)$ $(3.1c)$	1 1
15	$0, 4, 8, 12, 16$	$4, 8$ 16	$(3.1b)$ $(3.1c)$	2 1
16	$1 \le d \le 19$	$1 \le d \le 9$ $13 \le d \le 19$	$(3.1b)$ $(3.1c)$	9 7
17	$0 \le d \le 18$, all even	$2, 4, 6, 8,$ $12, 14, 16,18$	$(3.1b)$ $(3.1c)$	4 4
18	$1 \le d \le 19$	$1 \le d \le 9$ $11 \le d \le 19$	$(3.1b)$ $(3.1c)$	9 9
19	0			95

There are a total of 95 (r, d) points for which rounding is essential.

Some Representations of Sample Variance*

ANWAR H JOARDER

Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals Dhahran 31261, Saudi Arabia, Email: anwarj@ kfupm.edu.sa

The usual formula of variance depending on the rounding off the sample mean lacks in precision especially when computer programs are used for the calculation. The well known simplification of the total sums of squares does not always benefit. Since the variance of two observations is easily calculated without the use of sample mean, and the variance of a sample of n observations is the average of variances of observations based on $n(n-1)/2$ distinct subsets of units of size 2 from the sample, it is argued that this sense of pairing may result in precision. Some other forms of variance have been presented which provide some insight into it. Contribution of a new observation to variance is highlighted which is important in sequential sampling. Notions are illustrated with examples.

1. Introduction

The variance is a measure of variability that exists in a sample. There are two important reasons for measuring variability. The first reason is how well the average value depicts the data. A second reason is to learn the extent of scatter so that steps may be taken to control the existing variation. For example, while maintaining a long average mileage is the most important objective of the manufacturer of a tire, he tries to improve the uniformity in the mileage of it through better inspection and other quality control procedures; otherwise some customers would be satisfied and some would remain upset.
This is desired in many real world situations (Kolman, Anton and Averbach, 1992, 312).

The sample variance is one of the very basic notions a student learns in the beginning week of a statistics course. But to many it is a tongue twister, and most frustratingly, its meaning has nothing to do with the mathematical expression of the definition or the way it is calculated. Can we explain it in easy-to-understand terms? It is based on deviations of observations from the sample mean denoted by $x_i - \bar{x},\ (i = 1, 2, ..., n)$. These do help understand the variation in the sample observations. For a sample of (4,5,11,14), the sample mean $\bar{x} = 8.5$ so that the deviations are given by $x_1 - \bar{x} = 4 - 8.5 = -4.5,\ x_2 - \bar{x} = 5 - 8.5 = -3.5,\ x_3 - \bar{x} = 11 - 8.5 = 2.5,$ $x_4 - \bar{x} = 14 - 8.5 = 5.5$. Another sample with the same mean of 8.5 may have different variability e.g. the sample (7, 10, 8, 9) also has a mean of 8.5 but the deviations are $7 - 8.5 = -1.5,\ 10 - 8.5 = 1.5,\ 8 - 8.5 = -0.5,\ 9 - 8.5 = 0.5$ which do not exhibit as much variability as they do in the previous sample.

* *Published in International Journal of Mathemaatical Education in Science and Technology, 33(5), 772-784.(London, UK)*

What information do these distances or deviations $x_i - \bar{x},\ (i = 1, 2, \cdots, n)$ contain? If they tend to be large in absolute values, the data are spread out or highly variable. If they are mostly small in the absolute sense, the data are clustered around the sample mean and therefore do not exhibit much variability. A deviation indicates the amount by which an observation is away from the sample mean. Thus the deviation '-4.5' indicates that there is an observation in the sample which is 4.5 units below the sample mean. Similarly
the deviation '5.5' indicates that there is an observation which is 5.5 units above the sample mean.

Now the question is how to condense the information on deviations so as to form a single numerical measure of variability. Note that the deviations always add to zero and as such we do not gain any information with the total $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

To gauge the variability of the observations in a sample all we care about is whether an observation is away from the sample mean, be it below or above it. Thus we may use the absolute deviations or the squared deviations. The measure of variability produced by the absolute deviations did not gain popularity because on the one hand it presents analytic difficulties, and, on the other hand, it does not bring any benefit while compared with its counterpart. If we square the deviation -4.5, it would be 20.25. The latter implies there is an observation which is $\sqrt{20.25} = 4.5$ units away from the mean. The measure of variability produced by squared deviations, known as variance, indicates the variability of the sample observations around their mean.

The sum of the squared deviations is variously known as Total Sums of Squares (TSS), Corrected Sums of Squares (CSS) or simply as Sums of Squares (SS), and can be mathematically written as:

$$TSS = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

where the sum is over all the sample observations. This denotes the total variation among observations in a sample. For the sample (4, 5, 11, 14), TSS is given by:

$$TSS = (-4.5)^2 + (-3.5)^2 + (2.5)^2 + (5.5)^2 = 69\,.$$

The deviations are not always symmetric around zero though they add to zero. However, because of round-off error, the sum of the deviations may not be exactly zero. It may be remarked here that the fact that deviations add to zero implies that if $n-1$ of them are known, the other one is automatically determined. This number $n-1$ is called the degrees of freedom of the sample or of the sample mean or of *TSS*.

The variance $(s_n^{\,2})$ of the observations in a sample of size n is just the ratio of the total squared deviations to the degrees of freedom as defined below:

$$s_n^2 = \frac{TSS}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ n \geq 2\,. \qquad (1.1)$$

Obviously $0 \le s_n^2 < \infty$. In case $n = 1,$ the variance is usually defined to be 0. If all the observations were the same, each deviation would have been zero, so would have been the variance. If , however, the observations are widely apart, so will be the deviations producing positive TSS or positive variance. Thus the smaller (larger) the deviations in absolute value, the smaller (larger) is the variance, and vice versa.

The variance of the sample (4, 5, 11, 14) is $s_4^2 = 69/(4-1) = 23$. The variance of the second sample, producing deviations that are relatively less widely apart compared to that of the first sample, is approximately 1.67 which is, as expected, much lesser than that of the first sample.

Most statisticians use a simplified form of variance given by (2.1). In this paper some different forms of variance have been represented with the hope of shedding more light into the nature of variance. Though most of them are scattered in different text books, neater proofs of related theorems have been presented. Most importantly a new direction is emphasized for calculating variance that avoids using the sample mean and thereby guarantees least rounding off error.

Since the variance of two observations is easily calculated without the use of sample mean, and the variance of a sample of n observations is the average of variances of observations based on $n(n-1)/2$ distinct subsets of units of size 2 from the sample, it is argued that this sense of pairing may result in precision. The result is implicit in many texts (see e.g. Lindgren , 1993, 206). Recurrence relation of variance, which is important in sequential sampling for quality control in industry, is also emphasized for calculation. A grouping or pairing technique is introduced. Notions are illustrated with hypothetical examples.

2. Some Representations of Sample Variance

In this section we present seven different forms of variance. Though they are algebraically the same, they do differ in precision and time it takes to calculate them. Let $\bar{x}_n$ and s_n^2 be sample mean and variance of n observations respectively.

2.1 A Simplified Formula

$(n-1)s_n^2 = TSS$ can be represented by the following equivalent forms:

$$(n-1)s_n^2 = \frac{1}{n}\sum_{1\le i<j\le n}(x_i - x_j)^2 = \frac{1}{n}\sum_{1\le j<i\le n}(x_i - x_j)^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$$
$$= \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i - x_j)^2 = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=1}^{i+1}(x_i - x_j)^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 . \qquad (2.1)$$

If sample observations are integers but not large in size, the last representation allows one to do the calculation mentally. Since $\sum_{i=1}^{n} x_i = n\bar{x}_n$, it follows from (2.1) that

$$\sum_{i=1}^{n} x_i^2 = (n-1)s_n^2 + n\,\bar{x}_n^2 . \tag{2.2}$$

2.2 *Recurrence Relation Depending on Sample Mean*

A representation of variance due to Ross (1987, p 143) is presented below with an elegant proof.

Theorem 2.1 For $n \geq 2$ the following recurrence relation holds:

$$s_{n+1}^2 = \left(1 - \frac{1}{n}\right) s_n^2 + \frac{1}{n+1}(x_{n+1} - \bar{x}_n)^2 \quad . \tag{2.3}$$

Proof: It is easy to check that :

$$\begin{aligned} n\,s_{n+1}^2 &= \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2 \\ &= \left[(x_1 - \bar{x}_{n+1})^2 + (x_2 - \bar{x}_{n+1})^2 + \cdots + (x_n - \bar{x}_{n+1})^2\right] + (x_{n+1} - \bar{x}_{n+1})^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x}_{n+1})^2 + (x_{n+1} - \bar{x}_{n+1})^2 \end{aligned} \tag{2.4}$$

where $\bar{x}_{n+1}$ is the sample mean based on $n+1$ observations. Since $(n+1)\bar{x}_{n+1} = \sum_{i=1}^{n+1} x_i = n\bar{x}_n + x_{n+1}$ it follows that $\bar{x}_{n+1} = \frac{1}{n+1}(n\bar{x}_n + x_{n+1}) = \bar{x}_n + \frac{x_{n+1} - \bar{x}_n}{n+1}$ and consequently we have:

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x}_{n+1})^2 &= \sum_{i=1}^{n} \left[(x_i - \bar{x}_n) - \frac{x_{n+1} - \bar{x}_n}{n+1} \right]^2 \\ &= (n-1)s_n^2 + \frac{n}{(n+1)^2}\ (x_{n+1} - \bar{x}_n)^2 \end{aligned}$$

and $x_{n+1} - \bar{x}_{n+1} = x_{n+1} - \frac{1}{n+1}(n\bar{x}_n + x_{n+1}) = \frac{n}{n+1}(x_{n+1} - \bar{x}_n)$.

The proof is immediate by plugging the above two identities back in (2.4).

Thus if the first n observations are known, a value x_{n+1} can be obtained if a particular variance s_{n+1}^2 is desired.

2.3 Distinct Pairing (Variance Without Sample Mean)

Intuitively, the variability of a set of two observations say x_1 and x_2 should be reflected in the difference $|x_1 - x_2|$. Indeed for $n = 2$, it follows from (1.1) that:

$$s_2^2 = \frac{(x_1 - x_2)^2}{2}$$

which is just $1/2$ times the square of the range. In what follows we present a neater proof of a theorem implicit in many texts (see e.g. Lindgren, 1993, 206) that the variance of a sample of n observations can be easily calculated by calculating the variances of $\binom{n}{2}$ distinct pairs of observations and then averaging them.

Theorem 2.2 For a sample of size $n \geq 2$ the following result hold:

$$s_n^2 = \frac{1}{\binom{n}{2}} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{(x_i - x_j)^2}{2} \tag{2.5}$$

Proof. Since

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} (x_i - x_j)^2 = (n-1) \sum_{i=1}^{n} {x_i}^2 - 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j \tag{2.6}$$

and $$\left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{i=1}^{n} {x_i}^2 + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j \,, \tag{2.7}$$

it follows that :

$$\begin{aligned} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (x_i - x_j)^2 &= (n-1) \sum_{i=1}^{n} {x_i}^2 - \left[\left(\sum_{i=1}^{n} x_i \right)^2 - \sum_{i=1}^{n} {x_i}^2 \right] \\ &= n \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i \right)^2 = n \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= n(n-1) s_n^2 \end{aligned} \tag{2.8}$$

The proof is then completed by dividing both sides of (2.8) by $(n-1)$.

It follows from Theorem 2.2 that a table showing the differences among observations can be prepared whose entries are $w_{ij} = x_i - x_j \;\; (i, j = 1, 2, \ldots, n)$. Then

$$s_n^2 = \frac{1}{\binom{n}{2}} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{w_{ij}^{\ 2}}{2} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij}^{\ 2} = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^{\ 2} \tag{2.9}$$

where the factor $n(n-1) = n^2 - n$ is the number of off-diagonal elements of the matrix with elements w_{ij} $(i, j = 1, 2, ..., n)$. Note that $\frac{(n-1)s_n^2}{n} = V$ (see 2.8), is the second sample moment reported by Lindgren (1993, 206).

2.4 Variance Depending on Distinct Pairing and Sample Mean

The following theorem is a direct consequence of (2.3) and (2.5) .

Theorem 2.3 $s_{n+1}^2 = \frac{1}{n^2} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (x_i - x_j)^2 + \frac{1}{n+1} (x_{n+1} - \bar{x}_n)^2$. (2.10)

2.5 Recurrence Relation of Variance Without Sample Mean

The following Recurrence Relation follows from Theorem 2.2.:

$$\begin{aligned} \binom{n+1}{2} s_{n+1}^2 &= \sum_{i=2}^{n+1} \sum_{j=1}^{i-1} \frac{(x_i - x_j)^2}{2} \\ &= \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{(x_i - x_j)^2}{2} + \sum_{j=1}^{n} \frac{1}{2} (x_{n+1} - x_j)^2 \\ &= \binom{n}{2} s_n^2 + \frac{1}{2} \sum_{i=1}^{n} (x_{n+1} - x_i)^2. \end{aligned} \tag{2.11}$$

If the above recurrence relation is used in conjunction with Distinct Pairing (Theorem 2.2), i.e the expression in the middle of (2.11) is used, the sample variance is calculated without the sample mean. Avoidance of sample mean may result in precision.

2.6 Variance by Grouping

The variance or TSS can be calculated by grouping the sample observations, calculating variance of different groups and finally combining them by the following theorem. However, it is usually proved by labeling the observations with two suffixes which can be avoided since means and variances of groups are all that we need. The proof is made further neater by the use of identity in (2.2).

Theorem 2.4 Let n observations be divided into k groups containing $n_1, n_2, ..., n_k$ observations with means $\bar{x}_{(1)}, \bar{x}_{(2)}, \cdots, \bar{x}_{(k)}$ respectively. Then TSS will be given by:

$$TSS = \sum_{i=1}^{k} (n_i - 1)\, s_{(i)}^2 + \sum_{i(<l)=1}^{k} \frac{n_i\, n_l}{n}\, (\bar{x}_{(i)} - \bar{x}_{(l)})^2 \,. \tag{2.12}$$

Proof: Let the observations be divided into two groups $(i.e.\, k = 2)$ containing n_1 and n_2 observations with means $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$, and variances $s_{(1)}^2$ and $s_{(2)}^2$ respectively. Then:

$$\begin{aligned} TSS &= \sum_{i=1}^{n_1+n_2} (x_i - \bar{x}_n)^2 = \sum_{i=1}^{n_1+n_2} x_i^2 - (n_1 + n_2)\bar{x}_n^2 = \sum_{i=1}^{n_1} x_i^2 + \sum_{i=n_1+1}^{n_1+n_2} x_i^2 - (n_1 + n_2)\bar{x}_n^2 \\ &= \left\{(n_1 - 1)\, s_{(1)}^2 + n_1\, \bar{x}_{(1)}^2\right\} + \left\{(n_2 - 1)\, s_{(2)}^2 + n_2\, \bar{x}_{(2)}^2\right\} - \frac{\left(n_1\, \bar{x}_{(1)} + n_2\, \bar{x}_{(2)}\right)^2}{n_1 + n_2} \\ &= (n_1 - 1)\, s_{(1)}^2 + (n_2 - 1)\, s_{(2)}^2 + \frac{n_1\, n_2}{n_1 + n_2} (\bar{x}_{(1)} - \bar{x}_{(2)})^2 . \end{aligned}$$

Similarly for 3 groups we have:

$$\begin{aligned} TSS &= \sum_{i=1}^{n_1+n_2+n_3} (x_i - \bar{x}_n)^2 = \sum_{i=1}^{n_1} x_i^2 + \sum_{i=n_1+1}^{n_1+n_2} x_i^2 + \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} x_i^2 - (n_1 + n_2 + n_3)\bar{x}_n^2 \\ &= (n_2 - 1)\, s_{(1)}^2 + (n_2 - 1)\, s_{(2)}^2 + (n_3 - 1)\, s_{(3)}^2 \\ &+ \frac{n_1\, n_2}{n_1 + n_2} (\bar{x}_{(1)} - \bar{x}_{(2)})^2 + \frac{n_1\, n_3}{n_1 + n_3} (\bar{x}_{(1)} - \bar{x}_{(3)})^2 + \frac{n_2\, n_3}{n_2 + n_3} (\bar{x}_{(2)} - \bar{x}_{(3)})^2 . \end{aligned}$$

The proof for k groups is thus obvious.

Since group means in this context need to be rounded, we prefer to use totals of the groups to avoid rounding errors as much as possible. Let $T_{(i)} = n_i \bar{x}_{(i)}$, the total of the ith group. Then the following form may be helpful in calculating TSS:

$$\begin{aligned} TSS &= \sum_{i=1}^{k} (n_i - 1)\, s_{(i)}^2 + \sum_{i(>l)=1}^{k} \frac{n_i n_l}{n} \left(\frac{T_{(i)}}{n_i} - \frac{T_{(l)}}{n_l} \right)^2 \\ &= \sum_{i=1}^{k} (n_i - 1)\, s_{(i)}^2 + \sum_{i(>l)=1}^{k} \frac{\left(n_l T_{(i)} - n_i T_{(l)}\right)^2}{n\, n_i\, n_l} \end{aligned} \tag{2.13}$$

which will be reduced to $(n-1)\sum_{i=1}^{k} s_{(i)}^2 + \frac{1}{n}\sum_{i(>l)=1}^{k}\left(T_{(i)} - T_{(l)}\right)^2$ if the group sizes are the same. For ease of calculation by hand the following representation may be better:

$$TSS = \sum_{i=1}^{k}(n_i - 1)\, s_{(i)}^2 + \sum_{i(>l)=1}^{k} \frac{1}{n\, n_i\, n_l} \begin{vmatrix} n_i & T_{(i)} \\ n_l & T_{(l)} \end{vmatrix}^2 \qquad (2.14)$$

The first k terms in the Theorem is the contribution of the observations due to variation within groups (VWG), while the next $\binom{k}{2}$ terms can be attributed to the variation between groups (VBG). Groups having less variation among the observations within the groups may be used to have smaller contribution by VWG and more by VBG. Groups having more variation among the observations within the groups may be used to have larger contribution by variation within the groups (VWG) and less by VBG. Samples having the modal observation with high frequencies may be a good example for the first case (See Section 4.5). The idea of attributing the variation here is much similar to what led Fisher (1947) to discover the analysis of variance.

If variance of every group vanishes, the overall variance will be given by the second summand in (2.14). If , on the other hand, the sample observations are grouped in a way that the group means are the same, then the sample variance is given by :

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{k}(n_i - 1)\, s_{(i)}^2$$

which is a weighted sum of group variances.

2.7 Variance by Pairing

If sample size is 2, the sample variance is ½ times the square of the range. This suggests us that the grouping technique in (2.14) can be used to calculate variance by choosing $n_i = 2\,(i = 1, 2, \cdots, k-1)$ and $n_k = 2$ or 1 depending on whether sample size is even or odd. Let $T_{(i)} = n_i\, \bar{x}_{(i)}$, the total of the i th group, and $s_{(i)}^2 = w_i^2 / 2$, the variance of the i th pair. For ease of calculation by hand the following representation that follows from (2.14), may be better:

$$TSS = \sum_{i=1}^{k}\frac{w_i^2}{2} + \sum_{i(>l)=1}^{k} \frac{1}{n\, n_i\, n_l} \begin{vmatrix} n_i & T_{(i)} \\ n_l & T_{(l)} \end{vmatrix}^2 \qquad (2.15)$$

If sample observations are paired in a way that the means of pairs are the same then

$TSS = \sum_{i=1}^{k} \frac{w_i^2}{2}$ so that the variance will be $s^2 = \frac{1}{n-1} \sum_{i=1}^{k} \frac{w_i^2}{2}$. Note that $w_i = 0$ if $n_i = 1$.

2.8 Geometric Interpretation

The pair $\bar{x}_n$ and s_n^2 can be derived from the Euclidean minimization problem. Suppose that for observations $x_i \ (i = 1,2,\cdots,n)$, we want to find the value of t that minimizes $\sum_{i=1}^{n}(x_i - t)^2$. Note that the sums of squares $\sum_{i=1}^{n}(x_i - t)^2$ is the square of the Euclidean distance in n dimensions between n-dimensional point $(t,t,\cdots,t)$ and the observations expressed as the point $(x_1, x_2, \cdots, x_n)$. The minimization problem is amenable to calculus, but algebra is all that is needed here. Since x_i's are known sample values, the expression $\sum_{i=1}^{n}(x_i - t)^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\, t + t^2) = \sum_{i=1}^{n} x_i^2 - 2t\sum_{i=1}^{n} x_i + nt^2$ is quadratic in t. Its minimum occurs at its vertex and algebra shows that the minimum is $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$. This is the TSS we discussed in Section 1. Interested readers may go through an stimulating paper by Farnsworth (2000).

3. Contribution of a New Observation

Let c_{n+1} be the contribution of a new observation x_{n+1} to any variance formula based on n observations so that

$$s_{n+1}^2 = s_n^2 + c_{n+1}.$$
(3.1)

3.1 The contribution of a new observation $x_{n+1} = \bar{x}_n \mp d$ to the variance formula of Theorem 2.1 is given by:

$$c_{n+1} = -\frac{s_n^{\,2}}{n} + \frac{d^2}{n+1}$$
(3.2)

which is minimized if $d = 0$ i.e. if $x_{n+1} = \bar{x}_n$. If the new observation is at most d units away from $\bar{x}_n$, i.e. $\bar{x}_n - d \le x_{n+1} \le \bar{x}_n + d$, then it follows from (3.2) that c_{n+1} satisfies the following bounds:

$$-\frac{1}{n}\ s_n^2 \le c_{n+1} \le -\frac{s_n{}^2}{n} + \frac{d^2}{n+1}\,.$$
(3.3)

3.2 The contribution of a new observation x_{n+1} to the variance formula of Theorem 2.2 can be calculated easily from (2.11) as follows:

$$c_{n+1} = \frac{1}{(n+1)n}\ \sum_{i=1}^{n}(x_{n+1} - x_i)^2\ -\ \frac{2}{n+1}\ s_n^2$$
(3.4)

which also satisfies the bounds given by (3.3).

Theorem 3.1 If the new observation is $x_{n+1} = \bar{x}_n \mp d$, then for any expression of variance s_n^2 , the following recurrence relation holds:

$$s_{n+1}^2 = \frac{n-1}{n}\ \ s_n^2 + \frac{d^2}{n+1} = \frac{n-1}{n}\ \ s_n^2 + \frac{\left(x_{n+1} - \bar{x}_n\right)^2}{n+1}\,.$$
(3.5)

The contribution $c_{n+1} = s_{n+1}^2 - s_n^2$ of a new observation x_{n+1} is

(ii) negative if $\bar{x}_n - \sqrt{\frac{n+1}{n}}\ s_n < x_{n+1} < \bar{x}_n + \sqrt{\frac{n+1}{n}}\ s_n$,

(ii) zero if $x_{n+1} = \bar{x}_n \mp \sqrt{\frac{n+1}{n}}\ s_n$ and

(iii) positive elsewhere.

and is minimized at $x_{n+1} = \bar{x}_n$ (in which case it is $(n-1)n^{-1}\ s_n^2 < s_n^2$)

4. Some Illustrations

4.1 Variance by Recurrence Relation Depending on Sample Mean

To calculate the variance of the sample (4, 5, 11, 14) sequentially by Theorem 2.1, we have:

$$s_2^2 = \frac{1}{2}\left[(4 - 9/2)^2 + (5 - 9/2)^2\right] = 1/2\,,\quad \bar{x}_2 = 9/2\,,$$

$$s_{2+1}^2 = \left(1-\frac{1}{2}\right)s_2^2 + \frac{1}{2+1}(x_{2+1}-\bar{x}_2)^2 = \frac{1}{2}\frac{1}{2} + \frac{1}{3}\left(11-\frac{9}{2}\right)^2 = 43/3\,,\quad \bar{x}_3 = 20/3\,,$$

$$s_{3+1}^2 = \left(1-\frac{1}{3}\right)s_3^2 + \frac{1}{3+1}(x_{3+1}-\bar{x}_3)^2 = \frac{2}{3}\frac{43}{3} + \frac{1}{4}\left(14-\frac{20}{3}\right)^2 = 23\,.$$

To see the contribution of a new observation to the variance formula in Theorem 2.1 let us assume that we already have 3 observations (4, 5, 11) with variance $s_3^2 = 43/3$ and a new observation say $x_{3+1} = 14$. It follows from (3.2) that:

$$c_{3+1} = s_{3+1}^2 - s_3^2 = -\frac{1}{3}\,s_3^2 + \frac{1}{3+1}(x_{3+1}-\bar{x}_3)^2 = -\frac{1}{3}\frac{43}{3} + \frac{1}{4}\left(14-\frac{20}{3}\right)^2 = 26/3$$

so that by (3.1) we have $s_{3+1}^2 = s_3^2 + c_{3+1} = 43/3 + 26/3 = 23.$

4.2 Variance by Distinct Pairing (Variance Without Sample Mean)

To calculate the variance of the sample (4, 5, 11, 14) sequentially by Theorem 2.2, we have:

$$s_2^2 = \frac{(5-4)^2}{2} = \frac{1}{2}\,,$$

$$s_3^2 = \frac{1}{3}\left[\frac{(1)^2}{2} + \frac{(7)^2}{2} + \frac{(6)^2}{2}\right] = \frac{1}{3}\left(\frac{86}{2}\right) = 43/3\,,$$

$$s_4^2 = \frac{1}{6}\left[\frac{(1)^2}{2} + \frac{(7)^2}{2} + \frac{(6)^2}{2} + \frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2}\right] = \frac{1}{6}\left(\frac{276}{2}\right) = 23\,.$$

The differences can better be calculated by preparing the following difference table:

x	4	5	11	14
4				
5	$5-4=1$			
11	$11-4=7$	$11-5=6$		
14	$14-4=10$	$14-5=9$	$14-11=3$	

The arrangement of the sample observations in ascending order results in nonnegative entries (differences) in the table.

To see the contribution of a new observation $x_{3+1} = 14$ in the variance formula in Theorem 2.2 let us again assume that we already have 3 observations (4, 5, 11) with variance $s_3^2 = 43/3$ and a new observation say $x_{3+1} = 14$. It follows from (3.4) that:

$$c_{3+1} = \frac{1}{4(3)}\left[\frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2}\right] - \frac{2}{3+1} s_3{}^2 = 104/12 = 26/3$$ so that by (3.1) we have

$$s_{3+1}^2 = 43/3 + 26/3 = 23.$$

4.3 Variance Depending on Distinct Pairing and Sample Mean

To calculate the variance of the sample (4, 5, 11, 14) sequentially by Theorem 2.3, we have:

$$s_2^2 = 0 + \frac{1}{2}(5-4)^2 = \frac{1}{2},$$

$$s_{2+1}^2 = \frac{1}{2^2}(x_2 - x_1)^2 + \frac{1}{2+1}(x_3 - \bar{x}_2)^2$$
$$= \frac{1}{4}(5-4)^2 + \frac{1}{3}\left(11 - \frac{9}{2}\right)^2 = 43/3,$$

$$s_{3+1}^2 = \frac{1}{3^2}\left[(x_2 - x_1)^2 + (x_3 - x_1)^2 + (x_3 - x_2)^2\right] + \frac{1}{3+1}(x_4 - \bar{x}_3)^2$$
$$= \frac{1}{4}\left(1^2 + 7^2 + 6^2\right) + \frac{1}{4}\left(14 - \frac{20}{3}\right)^2 = 23.$$

4.4 Variance by Recurrence Relation Without Sample Mean (see equation 2.11)

To calculate the variance of the sample (4, 5, 11, 14) sequentially by equation 2.11, we have

$$s_2^2 = 0 + \frac{(5-4)^2}{2},$$

$$\binom{2+1}{2} s_{2+1}^2 = \binom{2}{2} s_2^2 + \frac{1}{2}\sum_{i=1}^{3}(x_{2+1} - x_i)^2$$
$$= \frac{1}{2} + \frac{1}{2}\left[(x_3 - x_1)^2 + (x_3 - x_2)^2\right] = \frac{1}{2} + \frac{1}{2}(7^2 + 6^2) = 43,$$

$$\binom{3+1}{2} s_{3+1}^2 = \binom{3}{2} s_3^2 + \frac{1}{2}\sum_{i=1}^{3}(x_{n+1} - x_i)^2$$

$$= 3\left(\frac{43}{3}\right) + \left[\frac{(10)^2}{2} + \frac{(9)^2}{2} + \frac{(3)^2}{2}\right] = 43 + 95$$

so that $s_2^2 = 1/2$, $s_3^2 = 43/3$ and $s_{3+1}^2 = \frac{1}{6}(43+95) = 23$.

4.5 Variance by Grouping

To calculate the variance of grades of 9 students (40, 70, 95, 70, 50, 70, 90, 70, 70) by (2.14), the sample may be grouped as (40, 50), (90, 95) and (70, 70, 70, 70, 70) for smaller VWG.

$s_{(i)}^2$ (=VWG)	i	groups	n_i	$T_{(i)}$	VBG
$50 = (50-40)^2/2$	1	(40, 50)	2	90	$\frac{1}{9(2)(2)}\begin{vmatrix}2 & 90\\ 2 & 185\end{vmatrix}^2 = 1002.77..$
$12.5 = (95-90)^2/2$	2	(90, 95)	2	185	$\frac{1}{9(2)(5)}\begin{vmatrix}2 & 90\\ 5 & 350\end{vmatrix}^2 = 694.44...$
0	3	70, 70, 70, 70, 70	5	350	$\frac{1}{9(2)(5)}\begin{vmatrix}2 & 185\\ 5 & 350\end{vmatrix}^2 = 0562.5$
62.5					2259.722...

The variance is given by $s^2 = (62.5 + 2259.722...)/8 \approx 290.278$.

4.6 Variance by Pairing

To calculate the variance of (4,5,11,14,20) by (2.15), we group them as (4, 20), (5, 14), (11) for larger contribution by variation within groups (VWG). The following table is prepared to apply the formula in (2.15).

$s_{(i)}^2$ (=VWG)	i	pairs	n_i	$T_{(i)}$	VBG
$128 = (4-20)^2/2$	1	(4, 20)	2	24	$\frac{1}{5(2)(2)}\begin{vmatrix}2 & 24\\ 2 & 19\end{vmatrix}^2 = 5$
$40.5 = (5-14)^2/2$	2	(5, 14)	2	19	$\frac{1}{5(2)(1)}\begin{vmatrix}2 & 24\\ 1 & 11\end{vmatrix}^2 = 0.4$

0	3	(11)	1	11	$\frac{1}{5(2)(1)}\begin{vmatrix}2 & 19\\ 1 & 11\end{vmatrix}^2 = 0.9$
168.5					6.3

The variance is given by $s^2 = \frac{1}{4}(168.5 + 6.3) = 43.7$.

5. Conclusion

When calculating variance by hand some representations may prove to be much efficient. However if the sample size is large and the computation is performed on a computer, then because of 'round-off error" some methods will be more efficient than the others. It is not surprising if the last representation of equation (2.2) provides negative value for sample variance (Ross, 1987, 143). Methods that avoid using sample mean (say equation 2.5 or 2.11) to the extent possible may result in much precision. Different grouping or pairing techniques along the line may also be devised for the same. It remains open to check the relative efficiency of various methods by computer programs.

Acknowledgements

The author acknowledges the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia. The author also thanks an anonymous referee, and my colleague Dr. A. Laradji for constructive suggestions that have improved the presentation of the paper.

References

Farnsworth, D.L., 2000, The Geometry of Statistics. *The College Mathematics Journal*. **31**(3), 200-204.

Fisher, R.A., 1947, *The Design of Experiments* (4th ed.), (Edinburgh :Oliver and Boyd).

Kolman, B., Anton , H. and Averbach, B. ,1992, *Mathematics with Applications for the Management, Life and Social Sciences,* (Philadelphia: Saunders College Publishing)

Lindgren, B.W. ,1993, *Statistical Theory.* (London: Chapman and Hall).

Ross, S.M. ,1987, *Introduction to Probability and Statistics for Engineers and Scientists.* (New York: Wiley).

File: c \ paper\ p30oc

Sample Variance and First-Order Differences of Sample Observations

Anwar H. Joarder

Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia, Email: anwarj@ kfupm.edu.sa

Abstract It is proved that sample variance can be calculated by the first order differences of sample observations via a matrix which is constant for any sample of a particular size. The constant matrix itself is open for further study. An alternative method is presented for the calculation of sample variance from a frequency distribution.

Key Words: Teaching; sample variance; difference table; reflection table; first order differences; pattern matrix.

1. Introduction

The variance $(s_n^{\,2})$ of n observations in a sample is just the ratio of $TSS(n)$ (Total squared deviations corrected by the mean) to the degrees of freedom where

$$TSS(n) = (n-1)\, s_n^{\,2} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \quad n \geq 2 . \tag{1.1}$$

If sample observations are integers but not large in size, the last representation in (1.1) allows you to do the calculation mentally. The usual formula for variance depending on rounding off the sample mean lacks in precision, especially when computer programs are used for the calculation. The problem of calculating sample variance by avoiding the use of sample mean was posed by Ross (1987, 143-144) who offered a recurrence relation of sample variance. In the spirit of Ross (1987), some solutions to the problem were discussed by Joarder (2002).

The quantity $TSS(n)$ can also be represented by the following equivalent forms

$$\frac{1}{n} \sum_{1 \leq i < j \leq n} (x_i - x_j)^2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 = \frac{1}{n} \sum_{i=2}^{n} \sum_{j=1}^{i-1} (x_i - x_j)^2 \tag{1.2}$$

(see e.g. Kotz, Kozubowski and Podgoriski, 2001, 186). Intuitively, the variability of a set of two observations say x_1 and x_2 should be reflected in the difference $|x_1 - x_2| = d$. Indeed for $n = 2$, it follows from (1.2) that $s_2^2 = (x_1 - x_2)^2 / 2 = d^2 / 2$ which is just $1/2$ times the square of the range.

The implication of the result in (1.2) is that the variance of a sample of n observations can be easily calculated by calculating the variances of $\binom{n}{2}$ distinct pairs and then averaging them. That is for a sample of size $n \geq 2$ the sample variance is given by:

$$s_n^2 = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \frac{w_{ij}^{\ 2}}{2} \text{ where } w_{ij} = x_i - x_j \quad (i, j = 1, 2, \ldots, n). \tag{1.3}$$

This note presents some tabular way for the calculation of sample variance with some mathematical insight. In the spirit of Ross (1987), this note resorts to the first order differences of sample observations to calculate sample variance. The main result is presented in Theorem 3.1. The first order differences of sample observations via a symmetric constant matrix C (depending on the finite sample size n) can also be used for the calculation of variance. Notions are illustrated with hypothetical examples. The pattern matrix C itself is open for further study. An alternative method for the calculation of sample variance from a frequency distribution with equal class widths is presented which is only good if number of classes is small.

2. The Difference Table and the Reflection Table for the Calculation of Sample Variance

It follows from equation (1.3) that a table showing the differences among observations can be made whose entries are $w_{ij} = x_i - x_j \quad (i, j = 1, 2, \ldots, n;\ i > j)$.

Example 2.1 To calculate variances of first $n = 2,3,4,5$ ordered observations of the sample (104, 94, 95, 101, 111), we may prepare the following Difference Table:

	94	95	101	104	111
94					
95	$1 = w_{21} = d_1$				
101	$7 = w_{31}$	$6 = w_{32} = d_2$			
104	$10 = w_{41}$	$9 = w_{42}$	$3 = w_{43} = d_3$		
111	$17 = w_{51}$	$16 = w_{52}$	$10 = w_{53}$	$7 = w_{54} = d_4$	

where we have used the notation $w_{i+1,i} = d_i \quad (i = 1,2,3,4)$ which are in fact the first order differences of the ordered observations of the sample.

A table called **Reflection Table** can be prepared by the use of the ordered observations (in ascending order) in a column followed by columns where each element in a column is obtained by subtracting it from the smallest observation in the

previous column. The following reflection table also provides the same set of differences of observations.

$x_{(i)}$	$r_i^{(1)} = x_{(i)} - x_{(1)}$ $(i = 2,3,4,5)$	$r_i^{(2)} = r_i^{(1)} - r_2^{(1)}$ $(i = 3,4,5)$	$r_i^{(3)} = r_i^{(2)} - r_3^{(2)}$ $(i = 4,5)$	$r_i^{(4)} = r_i^{(3)} - r_4^{(3)}$ $(i = 5)$
94				
95	$1 = 95 - 94$			
101	$7 = 101 - 94$	$6 = 7 - 1$		
104	$10 = 104 - 94$	$9 = 10 - 1$	$3 = 9 - 6$	
111	$17 = 111 - 94$	$16 = 17 - 1$	$10 = 16 - 6$	$7 = 10 - 3$

The variance of $n = 2,3,4,5$ observations is calculated below:

$s_2^2 = w_{21}^2 / 2 = d_1^2 / 2 = 1/2$,

$s_3^2 = \frac{1}{3(2)}\left(w_{21}^2 + w_{31}^2 + w_{32}^2\right) = \frac{1}{3(2)}(1^2 + 7^2 + 6^2) = 43/3$ and

$s_4^2 = \frac{1}{4(3)}\left(w_{21}^2 + w_{31}^2 + w_{32}^2 + w_{41}^2 + w_{42}^2 + w_{43}^2\right)$

$= \frac{1}{4(3)}(1^2 + 7^2 + 6^2 + 10^2 + 9^2 + 3^2) = 276/12 = 23.$

Since $\sum_{i=2}^{5}\sum_{j=1}^{4} w_{ij}^2 = 1^2 + 7^2 + 10^2 + \ldots + 7^2 = 970$, it follows that

$s_5^2 = \frac{1}{5(4)}\ (970) = \frac{970}{5(4)} = 48.5$.

An arrangement of observations into ascending order while preparing the above tables produces nonnegative entries.

3. Variance and the First Order Differences of Observations

Consider a triangular matrix $W = ((w_{ij}))$, with elements $w_{ij} = x_i - x_j$, $(i = 1,2,\ldots,n;$ $j = 1,2,\ldots,n;\ i > j)$ as shown in the Difference Table in Section 2. Further consider imaginary right angled triangles with vertices $w_{ij}(i > j)'s$ and right angle at the bottom left corner of the lower triangular matrix, and diagonal as the hypotenuse. Then any element, excluding the elements on the hypotenuse, in the right angled vertex of any imaginary triangle is the sum of the elements in the corresponding part of the hypotenuse. For example for a sample of size $n = 5$, we have

$w_{31} = w_{21} + w_{32} = d_1 + d_2$

$w_{41} = w_{21} + w_{32} + w_{43} = d_1 + d_2 + d_3,\ w_{42} = w_{32} + w_{43} = d_2 + d_3$

$w_{51} = \sum_{i=1}^{4} d_i\ ,\ w_{52} = d_2 + d_3 + d_4,\ w_{53} = d_3 + d_4$

In general let $w_{ij} = x_i - x_j\ (i, j = 1,2,\cdots,n; i > j)$ and the first order differences $w_{i+1,i} = d_i (i = 1,2,\cdots,n)$. The elements of the l $(l = 1,2,...,n-1)$ th diagonal line is thus given by

$$w_{i+l,l} = x_{i+l} - x_i = (x_{i+l} - x_{i+l-1}) + (x_{i+l-1} - x_{i+l-2}) + \cdots + (x_{i+2} - x_{i+1}) + (x_{i+1} - x_i)$$
$$= d_{i+l-1} + d_{i+l-2} + \cdots + d_{i+1} + d_i = \sum_{j=i}^{i+l-1} d_j,\ i = 1,2,...,n-l \quad .$$

Then the elements in the $n-1$ diagonal lines (making the lower triangular matrix) can be represented by

$w_{i+1,i} = d_i$ (let), $i = 1,2,...,n-1$ (say, $n-1$ th diagonal line or the hypotenuse line),

$w_{i+2,i} = \sum_{j=i}^{i+1} d_j,\ i = 1,2,...,n-2$ (say, $n-2$ th diagonal line i.e. line below the hypotenuse line),

$w_{i+3,i} = \sum_{j=i}^{i+2} d_j,\ i = 1,2,...,n-3$ (say, $n-3$ th diagonal line),

… … …

$w_{i+n-2,i} = \sum_{j=i}^{i+n-3} d_j,\ i = 1,2$ (say, the second diagonal line),

$w_{i+n-1,i} = \sum_{j=i}^{i+n-2} d_j,\ i = 1$ (say, the first diagonal line).

Theorem 3.1 Let $d_i (i = 1,2,\cdots,n)$ be the first order difference of observations, and $w_{ij} = \sum_{k=j}^{i-1} d_k\ ,\ (i > j)$. Then the variance of $n \geq 2$ observations is given by

$$s_n^2 = \frac{TSS(n)}{n-1} = \frac{1}{(n-1)n} \sum_{i=2}^{n} \sum_{j=1}^{i-1} w_{ij}^{\ 2} = \frac{1}{n(n-1)}\ d'Cd$$

where $d' = (d_1, d_2, \cdots, d_{n-1})$ and $C = ((c_{ij}))$ is a $(n-1) \times (n-1)$ symmetric matrix with $c_{ij} = (n-i)j$ if $i \geq j$; $(i, j = 1,2,\cdots,n-1)$.

Proof. It follows from (1.2) and the above notations that

$$
\begin{aligned}
n\,TSS(n) &= \sum_{i=2}^{n}\sum_{j=1}^{i-1} w_{ij}{}^{2} \\
&= w_{21}^2 + (w_{31}^2 + w_{32}^2) + (w_{41}^2 + w_{42}^2 + w_{43}^2) + \cdots \\
&\quad + (w_{n-1,1}^2 + w_{n-1,2}^2 + \cdots + w_{n-1,n-3}^2 + w_{n-1,n-2}^2) + \cdots \\
&\quad + (w_{n1}^2 + w_{n2}^2 + \cdots + w_{n,n-2}^2 + w_{n,n-1}^2)
\end{aligned}
$$

$$
\begin{aligned}
nTSS(n) &= w_{21}^2 + w_{32}^2 + w_{43}^2 + \cdots + w_{n-1,n-2}^2 + w_{n,n-1}^2) + \cdots \\
&\quad + (w_{31}^2 + w_{42}^2 + \cdots + w_{n-1,n-3}^2 + w_{n,n-2}^2) + \cdots \\
&\quad + (w_{n-1,1}^2 + w_{n2}^2) + w_{n1}^2 \\
&= \sum_{j=1}^{n-1} d_j^2 + \left[(d_1 + d_2)^2 + (d_2 + d_3)^2 + \cdots + (d_{n-2} + d_{n-1})^2 \right] \\
&\quad + \left[(d_1 + d_2 + d_3)^2 + (d_2 + d_3 + d_4)^2 + \cdots + (d_{n-3} + d_{n-2} + d_{n-1})^2 \right] \\
&\quad + \cdots \\
&\quad + \left[\left(\sum_{i=1}^{n-2} d_i \right)^2 + \left(\sum_{i=2}^{n-1} d_i \right)^2 \right] + \left(\sum_{i=1}^{n-1} d_i \right)^2
\end{aligned}
$$

$$
nTSS(n) = \sum_{j=1}^{n-1} d_j^2 + \sum_{i=1}^{n-2} \left(\sum_{j=i}^{i+1} d_j \right)^2 + \sum_{i=1}^{n-3} \left(\sum_{j=i}^{i+2} d_j \right)^2 + \cdots + \sum_{i=1}^{2} \left(\sum_{j=i}^{i+n-3} d_j \right)^2 + \sum_{i=1}^{} \left(\sum_{j=i}^{i+n-2} d_j \right)^2 \quad (3.1)
$$

Since $nTSS(n) = n(n-1)s_n^2$, it follows from (3.1) for $n = 2,3,4,5,\cdots$ that

$2(1)s_2^2 = d_1^2$

$3(2)s_3^2 = d_1^2 + d_2^2 + (d_1 + d_2)^2 = 2\left(d_1^2 + d_2^2 + d_1 d_2\right) = d'Cd$ where $d' = (d_1, d_2)$ and $C = ((c_{ij}))$ with $c_{11} = c_{22} = 2,\ c_{12} = 1$,

$4(3)s_4^2 = d_1^2 + d_2^2 + d_3^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_1 + d_2 + d_3)^2 = d'Cd$ where $d' = (d_1, d_2, d_3)$ and $C = ((c_{ij}))$ with $c_{11} = c_{33} = 3,\ c_{12} = c_{23} = 2,\ c_{13} = 1$, and

$$
\begin{aligned}
5(4)s_5^2 &= d_1^2 + d_2^2 + d_3^2 + d_4^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_3 + d_4)^2 \\
&\quad + (d_1 + d_2 + d_3)^2 + (d_2 + d_3 + d_4)^2 + (d_1 + d_2 + d_3 + d_4)^2 = d'Cd
\end{aligned}
$$

where $d' = (d_1, d_2, d_3, d_4)$ and $C = ((c_{ij}))$ with $c_{11} = c_{23} = c_{44} = 4$,
$c_{12} = c_{34} = 3,\quad c_{13} = c_{24} = 2,\ c_{14} = 1$.

Proceeding thus we have the general expression as stated in the theorem.

Note that $d' = (w_{21},\ w_{32}\ , \cdots, w_{i,i-1}, \cdots, w_{n,n-1}) = (d_1, d_2, \cdots, d_{n-1})$. Though the algebra in (3.1) looks complicated, the entire calculation can be made simple by ordering the sample observations and then preparing a table showing d_i's (the first order differences), in a column, followed by totals of consecutive pairs of $d_i's$ in the next column followed by totals of consecutive triplets of differences $d_i's$ in the next column and so on.

Corollary 3.1 For a sample of size $n \geq 2$ the following recurrence relation holds:

$$(n+1)\ TSS(n+1) - nTSS(n)$$
$$= \left(\sum_{i=1}^{n} d_i\right)^2 + \left(\sum_{i=2}^{n} d_i\right)^2 + \cdots + \left(d_{n-2} + d_{n-1} + d_n\right)^2 + \left(d_{n-1} + d_n\right)^2 + d_n^2$$

Proof. It follows from (3.1) that

$$(n+1)TSS$$
$$= \sum_{j=1}^{n} d_j^2 + \sum_{i=1}^{n-1}\left(\sum_{j=i}^{i+1} d_j\right)^2 + \sum_{i=1}^{n-2}\left(\sum_{j=i}^{i+2} d_j\right)^2 + \cdots + \sum_{i=1}^{2}\left(\sum_{j=i}^{i+n-2} d_j\right)^2 + \sum_{i=1}^{}\left(\sum_{j=i}^{i+n-1} d_j\right)^2$$
$$= \left[\sum_{j=1}^{n-1} d_j^2 + d_n^2\right] + \left[\sum_{i=1}^{n-2}\left(\sum_{j=i}^{i+1} d_j\right)^2 + (d_{n-1} + d_n)^2\right] + \left[\sum_{i=1}^{n-3}\left(\sum_{j=i}^{i+1} d_j\right)^2 + (d_{n-2} + d_{n-1} + d_n)^2\right] + \cdots$$
$$+ \left[\left(\sum_{i=1}^{n-1} d_i\right)^2 + \left(\sum_{i=2}^{n} d_i\right)^2\right] + \left(\sum_{i=1}^{n} d_i\right)^2$$
$$= nTSS(n) + d_n^2 + (d_{n-1} + d_n)^2 + (d_{n-2} + d_{n-1} + d_n)^2 + \cdots + \left(\sum_{i=2}^{n} d_i\right)^2 + \left(\sum_{i=1}^{n} d_i\right)^2$$

Example 3.1 To calculate variances of first $n = 2,3,4,5$ ordered observations of the sample (104, 94, 95, 101, 111), we may prepare the following table:

$x_{(i)}$	d_i $(i = 1,2,3,4)$	$d_i^{(2)} = d_i + d_{i+1}$ $(i = 1,2,3)$	$d_i^{(3)}$ $= d_i + d_{i+1} + d_{i+2} (i = 1,2)$	$d_i^{(4)} = \sum_{i=1}^{4} d_i$

94				
95	$1 = 95 - 94$			
101	$6 = 101 - 95$	$7 = 1 + 6$		
104	$3 = 104 - 103$	$9 = 6 + 3$	$10 = 1 + 6 + 3$	
111	$7 = 111 - 104$	$10 = 3 + 7$	$16 = 6 + 3 + 7$	$17 = 1 + 6 + 3 + 7$

The variance for sample sizes $n = 2,3,4,5$ are given below:

(i) Calculation of variance by first order differences and their "totals"

$2\, s_2^2 = d_1^2 = 1$,
$3(2)\, s_3^2 = (1^2 + 6^2) + 7^2 = 86$
$4(3)\, s_4^2 = (1^2 + 6^2 + 3^2) + (7^2 + 9^2) + 10^2 = 276$
$5(4)\, s_5^2 = (1^2 + 6^2 + 3^2 + 7^2) + (7^2 + 9^2 + 10^2) + (10^2 + 16^2) + 17^2 = 970$
so that variances are $s_2^2 = 1/2,\ s_3^2 = 43/3,\ s_4^2 = 23,\ s_5^2 = 97/2$.

(ii) Calculation of variance by first order differences and a constant matrix

$2s_2^2 = d_1^2 = 1$,
$3(2)\, s_3^2 = d'Cd = 86$ where $d' = (d_1, d_2) = (1,6)$ and $C = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$,
$4(3)\ \ s_4^2 = d'Cd = 276$ where $d' = (d_1, d_2, d_3) = (1,6,3)$ and
$C = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}$, and

$5(4)\ \ s_5^2 = d'Cd = 970$ where $d' = (d_1, d_2, d_3, d_4) = (1,6,3,7)$ and

$$C = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 1 & 2 & 3 & 4 \end{bmatrix}.$$

(iii) *Calculation of variance by the Recurrence Relation*

$2TSS(2) = d_1^2 = 1$
$3TSS(3) = 2TSS(2) + (d_1 + d_2)^2 + d_2^2 = 1 + 7^2 + 6^2 = 86$
$4TSS(4) = 3TSS(3) + (d_1 + d_2 + d_3)^2 + (d_2 + d_3)^2 + d_3^2 = 86 + 10^2 + 9^2 + 3^2 = 276$

$$5TSS(5) = 4TSS(4) + (d_1 + d_2 + d_3 + d_4)^2 + (d_2 + d_3 + d_4)^2 + (d_3 + d_4)^2 + d_4^2$$
$$= 276 + 17^2 + 16^2 + 10^2 + 7^2 = 970$$

The variance can then be calculated as before.

4. Correlation Coefficient

The sum of products between two variables x and y can be variously written as

$$s_{xy} = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)(y_i - y_j)$$
$$= \frac{1}{n}\sum_{1\le i<j\le n}(x_i - x_j)(y_i - y_j) = \frac{1}{n}\sum_{i=2}^{n}\sum_{j=1}^{i-1}(x_i - x_j)(y_i - y_j)$$

(see e.g. Kotz, Kozubowski and Podgoriski, 2001, 186). The calculation can be done by difference table or preferably reflection table described Section 2. Then the correlation coefficient r can be calculated by $r\sqrt{s_{xx}}\sqrt{s_{ss}} = s_{xy}$.

Example 4.1 Consider the following grades of 5 students in Midterm and final exam:

$$(50, 60),\ (60, 70),\ (75, 90),\ (80,\ 80),\ (85,\ 90)$$

To calculate the correlation coefficient we prepare the following Reflection Table for x and y values:

	$(50,\ 60)$	$(60, 70)$	$(75,\ 90)$	$(80,\ 80)$	$(85,\ 90)$
$(50, 60)$					
$(60, 70)$	$(10,\ 10)$				
$(75,\ 90)$	$(25,\ 30)$	$(15,\ 20)$			
$(80,\ 80)$	$(30,\ 20)$	$(20,\ 10)$	$(5,\ -10)$		
$(85,\ 90)$	$(35,\ 30)$	$(25,\ 20)$	$(10,\ 0)$	$(5,\ 10)$	

Then $s_{xx} = \left(10^2 + 25^2 + \cdots + 5^2\right)/5 = 4250/5$, $s_{yy} = \left(10^2 + 30^2 + \cdots + 10^2\right)/5 = 3400/5$ and $s_{xy} = \left[10(10) + 25(30) + \cdots + 5(10)\right]/5 = 3500/5$. The correlation coefficient is then given by $r = \dfrac{3500}{\sqrt{(4250)(3400)}} \approx 0.92$.

5. Variance of Observations with frequencies

It is well known that variance of n consecutive integers is given by $n(n+1)/12$. Interestingly, the variance of any three consecutive integers is 1, and that of any 11 consecutive numbers is 11. This kind of result is sometimes important to construct examples quickly in classrooms. In this section we present an alternative method for calculating sample variance (s^2) from a frequency distribution. Note that the variance of a sample classified into k classes in a frequency distribution with mid values $y_1, y_2, \cdots, y_k$ and frequencies $f_1, f_2, \cdots, f_k$ is given by $(n-1)s^2 = \sum_{i=1}^{k} (y_i - \bar{y})^2 f_i$. The following theorem is well known (Joarder, 2002).

Theorem 5.1 Let n observations be divided into k groups containing $f_1, f_2, \cdots, f_k$ observations with means $\bar{x}_{(1)}, \bar{x}_{(2)}, \cdots, \bar{x}_{(k)}$ and variances $s_i^2 (i = 1,2,\cdots,k)$ respectively. Then

$$TSS(n) = \sum_{i=1}^{k} (f_i - 1)^2 \; s_i^2 + \sum_{i(<l)=1}^{k} (\bar{x}_i - \bar{x}_l)^2 \; \frac{f_i f_l}{n} \text{ where } \quad n = f_1 + f_2 + \cdots + f_k .$$

In case observations in every group are the same, the means of k groups may be denoted by $y_1, y_2, \cdots, y_k$ and the variance of the ith group $s_i^2 = 0, (i = 1,2,\cdots,k)$ so that the first summand in the $TSS(n)$ in the above expression vanishes, and we have the following corollaries.

Corollary 5.1 Let $y_1, y_2, \cdots, y_k$ have frequencies $f_1, f_2, \cdots, f_k$ with $n = f_1 + f_2 + \cdots + f_k$ then

$$TSS(n) = \sum_{i(<l)=1}^{k} (y_i - y_l)^2 \; \frac{f_i f_l}{n} .$$

The following corollary is obvious from Corollary 5.1. See e.g. Joarder (2003)

Corollary 5.2 If k observations follow arithmetic series with common difference w and frequencies $f_i \;\; (i = 1,2,\cdots,k)$ then $TSS(n)$ is given by

$$TSS(n) = \frac{w^2}{n} \sum_{i(<l)=1}^{k} (i-l)^2 \; f_i f_l \tag{5.1}$$

$$= w^2 \left[\sum_{i=1}^{k} i^2 f_i - \frac{1}{n} \left(\sum_{i=1}^{k} i f_i \right)^2 \right] \tag{5.2}$$

The expression in (5.2) follows from the basic definition of sample variance by noting that the variance does not depend on the origin of transformation. Note that in this case variance depends on the observations only through the common difference, first k positive integers and corresponding frequencies.

Example 5.1 (Bluman, 2001, 113) Thirty automobiles were tested for fuel efficiency (in miles per gallon). The following frequency distribution was obtained.

Class boundaries	frequencies
$7.5-12.5$	3
$12.5-17.5$	5
$17.5-22.5$	15
$22.5-27.5$	5
$27.5-32.5$	2

Since $\sum_{i=1}^{5} if_i = 1(3)+2(5)+3(15)+4(5)+5(2) = 88$ and

$\sum_{i=1}^{5} i^2 f_i = (1)^2(3)+(2)^2(5)+(3)^2(15)+(4)^2(5)+(5)^2(2) = 288$

it follows from (5.2) that $TSS(30) = 5^2\left[288-\left(88\right)^2/30\right] = 746\frac{2}{3}$ so that the variance is 25.747 approximately.

Joarder (2003) deduced many special cases from (5.1) may be of much use in constructing examples quickly in classrooms. The following corollary follows from (5.2) or directly by the basic definition of variance.

Corollary 5.3 If $n = kf$ observations follow arithmetic series with common difference w and common frequency f, then the variance is then given by

$$s_n^2 = \frac{k(k+1)}{12}\ \frac{n-f}{n-1}\ w^2 \qquad (5.3)$$

where $k(k+1)/12$, the first factor in (5.3), is the variance of k natural integers. The attachment of common frequency to each of the k integers is contributing to the second factor in the above expression. The variance of the first k positive integers each with common frequency f is given by (5.3) with $w = 1$

Acknowledgement

The author gratefully acknowledges the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia.

References

Bluman, A.G. (2001). *Elementary Statistics: A Step by Step Approach*. McGraw Hill, New York.

Joarder, A.H. (2002). On some representations of sample variance. *International Journal of Mathematical Education in Science and Technology*. 33(5), 772-784.

Joarder, A.H. (2003). Sample Variance and the first order differences. Technical Report. Dept of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Saudi Arabia.

Kotz, S.; Kozubowski, T and Podgoriski, K. (2001). *The Laplace Distribution and Generalizations*. Birkhauser, Boston, USA.

Ross, S.M. (1987). *Introduction to Probability and Statistics for Engineers and Scientists*. Wiley, New York.

*Inequalities Among Some Measures of Location**

A. LARADJI and A. H. JOARDER

Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals Dhahran 31261, Saudi Arabia. Emails: alaradji@kfupm.edu.sa; anwarj@kfupm.edu.sa

Abstract Some inequalities involving sample means, sample median, the smallest and the largest observations are established. An upper bound of the absolute difference between the sample mean and median are also derived. Interesting inequalities among sample mean and median are obtained for cases when all the observations have the same sign. Some other inequalities are derived by taking expected values of the sample results and then applying them to some continuous distributions.

1. Introduction

Inequalities involving measure of location namely, sample means, median and extreme observations do not appear to be generally known. This note is inspired by Shiffer and Harsha (1980) and Macleod and Henderson (1984) who worked on the bounds of sample standard deviation. Some inequalities involving sample means, sample median, the smallest and the largest observations are established. An upper bound of the absolute difference between the sample mean and median are also derived. Interesting inequalities are deduced for cases when all the observations are nonnegative or have the same sign. We believe that the inequalities will, in particular, provide additional information to students in statistics, and, in general, open a new direction of further research to refine inequalities on other sample statistics along the line of Shiffer and Harsha (1980), Macleod and Henderson (1984) and Eisenhauer (1983).

Let $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ be the order statistics corresponding to the sample $(x_1, x_2, \cdots, x_n)$ with median $\tilde{x} = \frac{1}{2}\left(x_{([n/2+1/2])} + x_{([n/2+1])}\right)$ where $[m]$ is the bracket function denoting largest integer not exceeding m. Also let the arithmetic, geometric and harmonic mean of a sample $(x_1, x_2, \cdots, x_n)$ be denoted by $a(x_1, x_2, \cdots, x_n) = \bar{x}$, $g(x_1, x_2, \cdots, x_n)$ and $h(x_1, x_2, \cdots, x_n)$ respectively. In this paper we establish interesting inequalities involving some of the sample characteristics, namely, $\bar{x}$, $g(x_1, x_2, \cdots, x_n)$, $h(x_1, x_2, \cdots, x_n)$, $\tilde{x}$, $x_{(1)}$ and $x_{(n)}$. An upper bound of the absolute difference between the sample mean and median are also derived. Interesting inequalities among sample mean and median are obtained for cases when all the observations have the same sign. Some other inequalities are derived by taking expected values of the sample results and then applying them to some continuous distributions.

---

[*] [2] The paper is based on Technical Report 283, Department of mathematical Sciences, King Fahd University of Petroleum and Minerals, Saudi Arabia.

2. Main Results

The following lemma is obvious.

Lemma 2.1 Let $x_i \le y_i \ (1 \le i \le n)$. Then

$(i)\ \sum_{i=1}^{n} x_i \le \sum_{i=1}^{n} y_i$

$(ii)\ \prod_{i=1}^{n} x_i \le \prod_{i=1}^{n} y_i$ if $x_{(1)} \ge 0$.

$(iii)\quad \sum_{i=1}^{n} \frac{1}{x_i} \ge \sum_{i=1}^{n} \frac{1}{y_i}$ if $x_{(1)} > 0$.

Consider the three sequences $A = \{a_1, a_2, \cdots, a_{2n}\},\ B = \{b_1, b_2, \cdots, b_{2n}\}$ and $C = \{c_1, c_2, \cdots, c_{2n}\}$ each having $2n$ terms defined by

$$a_k = \begin{cases} x_{(1)} & \text{if } 1 \le k \le n \\ \tilde{x} & \text{if } n+1 \le k \le 2n \end{cases},$$

$b_k = x_{([k/2+1/2])}$ and $c_k = \begin{cases} \tilde{x} & \text{if } 1 \le k \le n \\ x_{(n)} & \text{if } n+1 \le k \le 2n \end{cases}$

These sequences are then

$A = \{x_{(1)}, x_{(1)}, \cdots, x_{(1)}, \tilde{x}, \tilde{x}, \cdots, \tilde{x}\},\ B = \{x_{(1)}, x_{(1)}, x_{(2)}, x_{(2)}, \cdots, x_{(n)}, x_{(n)}\}$ and $C = \{\tilde{x}, \tilde{x}, \cdots, \tilde{x}, x_{(n)}, x_{(n)}, \cdots, x_{(n)}\}$ where A and C contain n medians $(\tilde{x})$. For $1 \le k \le n$,

$a_k = x_{(1)} \le x_{([k/2+1/2])} = b_k \le \frac{1}{2}\left(x_{([n/2+1/2])} + x_{([n/2+1])}\right) = \tilde{x} = c_k$ and for $n+1 \le k \le 2n$,

$a_k = \tilde{x} \le \frac{1}{2}\left(x_{([n/2+1/2])} + x_{([n/2+1])}\right) \le x_{([k/2+1/2])} = b_k \le x_{(n)} = c_k$. Since the elements of the three sets satisfy the conditions of Lemma 2.1, we have the following theorem .

Theorem 2.1 For any sample of $n \ge 2$ observations $x_1, x_2, \cdots, x_n$ with $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, the following inequalities hold:

$(i)\ \frac{x_{(1)} + \tilde{x}}{2} \le \bar{x} \le \frac{\tilde{x} + x_{(n)}}{2}$

$(ii)\ \sqrt{x_{(1)}\,\tilde{x}} \le g(x_1, x_2, \cdots, x_n) \le \sqrt{\tilde{x}\,x_{(n)}}$ if $x_{(1)} \ge 0$ and

$$(iii)\ \frac{2}{\dfrac{1}{x_{(1)}}+\dfrac{1}{\tilde{x}}} \le h(x_1, x_2, \cdots, x_n) \le \frac{2}{\dfrac{1}{\tilde{x}}+\dfrac{1}{x_{(n)}}} \text{ if } x_{(1)} > 0 \text{ and}$$

where $g(x)$ and $h(x)$ are the geometric and harmonic means of a sample of n observations.

Proof. Applying Lemma 2.1 *(i)* to the sets A and B, and then to B and C we have $nx_{(1)} + n\tilde{x} \le 2n\bar{x}$ and $2n\bar{x} \le nx_{(n)} + n\tilde{x}$ which implies Theorem 2.1 (i). The other two parts of the theorem are deduced from Lemma 2.1 (ii) and Lemma 2.1 (iii) respectively in a similar manner.

Since, in many real world situations observations are nonnegative, the following corollary may be useful.

Corollary 2.1 If $x_{(1)} > 0,$ then $\dfrac{1}{2}h(x_1, x_2, \cdots, x_n) \le \tilde{x} \le 2\bar{x}$

Proof. It follows from Theorem 2.1 (iii) that
$\dfrac{1}{2}h(x_1, x_2, \cdots, x_n) \le \dfrac{1}{1/\tilde{x} + 1/x_{(n)}} \le \dfrac{1}{1/\tilde{x}} = \tilde{x} \le x_{(1)} + \tilde{x}$ which, by virtue of Theorem 2.1(i), cannot exceed $2\bar{x}$.

Theorem 2.2 For any sample of $n \ge 2$ observations $x_1, x_2, \cdots, x_n$ with $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, the following inequalities hold:

$$(i)\ (i-1)x_{(1)} + x_{(i)} + (n-i)x_{(i+1)} \le n\bar{x}, \quad \text{for} \quad 1 \le i \le n-1$$

$$(ii)\ \frac{1}{2}\left[\left(1-\frac{1}{n}\right)x_{(1)} + \left(1+\frac{1}{n}\right)\tilde{x}\right] \le \bar{x} \le \frac{1}{2}\left[\left(1+\frac{1}{n}\right)\tilde{x} + \left(1-\frac{1}{n}\right)x_{(n)}\right],$$

$$(iii)\ |\tilde{x} - \bar{x}| \le \frac{n-1}{n+1}\ \max(\bar{x} - x_{(1)}, x_{(n)} - \bar{x})$$

Proof. (i) For $1 \le i \le n-1$, we have

$$\begin{aligned} n\bar{x} &= \left(x_{(1)} + x_{(2)} + \cdots + x_{(i-1)}\right) + x_{(i)} + \left(x_{(i+1)} + \cdots + x_{(n)}\right) \\ &\ge (i-1)x_{(1)} + x_{(i)} + (n-i)x_{(i+1)} \end{aligned} \qquad (2.1)$$

(ii) For odd n and $i = (n+1)/2$, it follows from (2.1), by virtue of $\tilde{x} = x_{((n+1)/2)} \le x_{((n+3)/2)}$, that

$\frac{n-1}{2}x_{(1)} + \tilde{x} + \frac{n-1}{2}\ \tilde{x} \le n\bar{x}$ so that

$(n-1)x_{(1)} + (n+1)\tilde{x} \le 2n\bar{x}$.

(2.2)

When n is even, letting $i = n/2$ and $i = n/2+1$, it follows from (2.1) that

$\left(\frac{n}{2}-1\right)x_{(1)} + x_{(n/2)} + \frac{n}{2}\ x_{(n/2+1)} \le n\bar{x}$ and $\frac{n}{2}x_{(1)} + x_{(n/2+1)} + \left(\frac{n}{2}-1\right)\ x_{(n/2+2)} \le n\bar{x}$.

By adding the two inequalities the and using the fact that $\tilde{x} \le x_{(n/2+1)} \le x_{(n/2+2)}$ for even n, we have

$(n-1)x_{(1)} + \frac{n}{2}\ \tilde{x} + \left(\frac{n}{2}-1\right)\ \tilde{x} + 2\tilde{x} + \le 2n\bar{x}$

(2.3)

so that the inequality (2.2) also follows from (2.3). Hence for any sample of size $n \ge 2$, we have

$(n-1)x_{(1)} + (n+1)\tilde{x} \le 2n\bar{x}$.

(2.4)

Next from $-x_{(n)} \le -x_{(n-1)} \le \cdots \le -x_{(1)}$, similarly we obtain
$(n+1)(-\tilde{x}) + (n-1)(-x_{(n)}) \le 2n(-\bar{x})$ or, $(n+1)\tilde{x} + (n-1)x_{(n)} \ge 2n\bar{x}$.
The proof is thus complete.

(iii) By writing $2\bar{x} = \left(1-\frac{1}{n}\right)\bar{x} + \left(1+\frac{1}{n}\right)\bar{x}$, it follows from Theorem 2.2 (ii) that

$\left(1-\frac{1}{n}\right)x_{(1)} + \left(1+\frac{1}{n}\right)\tilde{x} \le \left(1-\frac{1}{n}\right)\bar{x} + \left(1+\frac{1}{n}\right)\bar{x} \le \left(1+\frac{1}{n}\right)\tilde{x} + \left(1-\frac{1}{n}\right)x_{(n)}$

or, $\left(1-\frac{1}{n}\right)(\bar{x} - x_{(n)}) \le \left(1+\frac{1}{n}\right)(\tilde{x} - \bar{x}) \le \left(1-\frac{1}{n}\right)(\bar{x} - x_{(1)})$

or, $-\frac{n-1}{n+1}\ \left(x_{(n)} - \bar{x}\right) \le \tilde{x} - \bar{x} \le \frac{n-1}{n+1}\ \left(\bar{x} - x_{(1)}\right)$

or, $|\tilde{x} - \bar{x}| \le \frac{n-1}{n+1}\ \max(\bar{x} - x_{(1)}, x_{(n)} - \bar{x})$.

It is worth noting that the inequalities $\frac{1}{2}(x_{(1)}+\tilde{x}) \le \bar{x} \le \frac{1}{2}(\tilde{x}+x_{(n)})$

in Theorem 2.1 (i) can be deduced from Theorem 2.2 (ii) in the following way:

$$\frac{1}{2}(x_{(1)}+\tilde{x}) \le \frac{1}{2}\left(x_{(1)}+\tilde{x}+\frac{1}{n}\left(\tilde{x}-x_{(1)}\right)\right) = \frac{1}{2}\left(\left(1-\frac{1}{n}\right)x_{(1)}+\left(1+\frac{1}{n}\right)\tilde{x}\right) \le \bar{x}$$
$$\le \frac{1}{2}\left[\left(1+\frac{1}{n}\right)\tilde{x}+\left(1-\frac{1}{n}\right)x_{(n)}\right] = \frac{1}{2}\left[\tilde{x}+x_{(n)}-\frac{1}{n}\left(x_{(n)}-\tilde{x}\right)\right] \le \frac{1}{2}(\tilde{x}+x_{(n)})$$

(2.5)

Corollary 2.2 The following inequalities hold for any sample of $n \ge 2$ observations:

(i) $2|\bar{x}| \ge |\tilde{x}|$, if the observations have the same sign.

(2.6)

(ii) $$x_{(1)} \le \frac{2n}{n-1}\,\bar{x}+\left(1-\frac{2n}{n-1}\right)\tilde{x} \le x_{(n)}$$

(2.7)

Proof. (i) If $x_{(1)} \ge 0$, then both $\bar{x}$ and $\tilde{x}$ are nonnegative $\tilde{x}/2 \le \left(x_{(1)}+\tilde{x}\right)/2$ which cannot exceed $\bar{x}$ by (2.5). If $x_{(n)} \le 0$ then both $\bar{x}$ and $\tilde{x}$ are nonpositive and $\bar{x} \le \left(\tilde{x}+x_{(n)}\right)/2$ which cannot exceed $\tilde{x}/2$ by (2.5). Taking absolute values we have the inequality in (i).

(ii) The inequalities follow directly from Theorem 2.2 (ii).

Remarks

(i) If the observations $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ have the same sign then $2|\bar{x}| = |\tilde{x}|$ occurs exactly when all x's are equal to 0. If $x_{(1)} \ge 0$, then $2|\bar{x}| = |\tilde{x}|$ implies $2\bar{x} = \tilde{x}$ so that we have $0 \le \left(1-\frac{1}{n}\right)x_{(1)}+\left(1+\frac{1}{n}\right)\tilde{x} \le \tilde{x}$ by Theorem 2.2 (ii) and hence $\frac{1}{n}\,\tilde{x}+\left(1-\frac{1}{n}\right)x_{(1)} = 0$ which happens only if $\tilde{x} = 0$ i.e. if $2\bar{x} = 0$ and so all observations are 0's. A similar argument applies when $x_{(n)} \le 0$.

(ii) If $x_{(1)} > 0$, then $2\bar{x} \ge \frac{n+1}{n}\,\tilde{x} > \tilde{x}$ by Theorem 2.2 (ii). Similarly, $2\bar{x} < \tilde{x}$ if $x_{(n)} < 0$.

(iii) In case not all the observations have the same sign, an example of a sample showing $2\bar{x} = \tilde{x}$ may be: $n = 3,\ x_{(1)} = -10,\ x_{(2)} = 10,\ x_{(3)} = 15$ which could be average temperatures of three days in a city.

(iv) If all the observations are nonnegative, then for a negatively skewed distribution we have $\tilde{x}/2 \le \bar{x} \le \tilde{x}$ but for a positively skewed distribution we have $\tilde{x} \le \bar{x} \le (\tilde{x} + x_{(n)})/2$.

Corollary 2.3 If $n \ge 2$ observations have the same sign, then $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| \le 1$.

Proof. Since x's have the same sign, it follows from (2.5) that

$\frac{\tilde{x}}{\bar{x}} = \left|\frac{\tilde{x}}{\bar{x}}\right| \le 2$. Then consider the following cases:

If $\frac{\tilde{x}}{\bar{x}} \ge 1$, then $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| = \frac{\tilde{x}}{\bar{x}} - 1 \le 1$, and if $\frac{\tilde{x}}{\bar{x}} < 1$, then $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| = 1 - \frac{\tilde{x}}{\bar{x}} < 1$. Hence the proof.

3. Inequalities Involving Expected Values

The following theorem follows from Corollary 2.2.

Theorem 3.1 For any nonnegative random variable X, the inequality $E(\tilde{X}) \le 2E(\bar{X})$ holds whenever the expected values exist.

Example 3.1 Let the random variable X have the probability density function (pdf)

$$f(x) = \beta^{-1} e^{-x/\beta} \text{ where } 0 < x,\ 0 < \beta \qquad (3.1)$$

The expected value of the r th order statistics $X_{(r)}$ is

$$E(X_{(r)}) = \frac{1}{\beta B(r, n-r+1)} \int_0^\infty x_{(r)}\ e^{-(n-r+1)x_{(r)}} (1 - e^{-x_{(r)}/\beta}) dx_{(r)} \qquad (3.2)$$

where $B(a,b)$ is the usual beta function. To check theorem 3.1 let us assume first that $n = 2m+1$ so. Then by replacing r in (3.2) by $m+1$we have

$$E(\tilde{X}) = \frac{(2m+1)!}{(m!)^2} \beta \int_0^\infty e^{-(m+1)y} (1 - e^{-y})^m dy = \frac{(2m+1)!}{(m!)^2} \beta I(m)$$

where $I(m) = \int_0^\infty u e^{-(m+1)u} (1 - e^{-u})^m du$

Since $X_1, X_2, \cdots, X_n$ are identically distributed, it follows that $E(\bar{X}) = \beta$ and hence we have to prove

$$\frac{(2m+1)!}{(m!)^2}\beta I(m) \le 2\beta \tag{3.3}$$

Let $g(u) = ue^{-u}$, then $g(u)$ has its absolute maximum at $u = 1$ on $[0, \infty)$, its value is $u(1) = 1/e$. So the integral $I(m)$ satisfies

$I(m) = \int_0^\infty ue^{-u}e^{-mu}(1-e^{-u})^m du \le \frac{1}{e}\int_0^\infty e^{-mu}(1-e^{-u})^m du = \frac{1}{e}B(m, m+1)$, and it follows from (3.3) that

$$\frac{(2m+1)!}{(m!)^2}\beta I(m) \le \frac{(2m+1)!}{(m!)^2}\ \frac{\beta}{e}\ B(m, m+1) = \frac{(2m+1)!}{(m!)^2}\ \frac{\beta}{e}\frac{(m-1)!m!}{(2m)!} = \frac{2m+1}{m}\beta \le 2\beta$$

for all $m \ge 2$. Note that

$I(1) = \int_0^\infty ue^{-2u}(1-e^{-u})du = \frac{5}{36}$ so that (3.2) holds. Hence for all $m \ge 1$ and $n = 2m+1$, we obtain $E(\tilde{X}) \le \beta$.

If the sample size n is even let $n = 2m$ so that $2E(\tilde{X}) = E(X_{(m)}) + E(X_{(m+1)})$. Then by replacing r in (3.2) by m and $m+1$ we have

$$2E(\tilde{X}) = \frac{1}{\beta B(m, m+1)}\int_0^\infty ye^{-my/\beta}(1-e^{-y/\beta})^{m-1}dy\ .$$

Letting $y = \beta u$ we have

$$\begin{aligned} 2E(\tilde{X}) &= \frac{\beta}{B(m, m+1)}\int_0^\infty ue^{-mu}(1-e^{-u})^{m-1}du = \frac{\beta}{B(m, m+1)}I(m-1) \\ &\le \frac{\beta}{B(m, m+1)}\frac{1}{e}B(m-1, m) = \frac{2(2m-1)\beta}{(m-1)e} \le 2\beta \end{aligned} \tag{3.4}$$

since $m \ge 3$. If $m = 1, 2$ it is obvious from (3.4) that

$$2E(\tilde{X}) = \frac{\beta}{B(1,2)}\int_0^\infty ue^{-u}du = \frac{\beta}{B(1,2)} \le 2\beta \quad \text{and}$$

$$2E(\tilde{X}) = \frac{\beta}{B(2,3)}\int_0^\infty ue^{-2u}(1-e^{-u})du = \frac{5\beta}{36B(2,3)} \le 2\beta$$

Hence for all $m \geq 1$ and $n = 2m$ we obtain $E(\tilde{X}) \leq \beta$. Note that examples can be multiplied for other distributions. It is expected that interesting inequalities would result in for distributions involving multiple parameters.

We now apply Theorem 3.1 to different continuous distributions and obtain interesting inequalities:

1. For the above exponential distribution the expected value of the ith order statistic $X_{(i)}$ is given by $E\left(X_{(i)}\right) = \beta \sum_{j=1}^{i} (n-j+1)^{-1}$ (see Harter and Balakrishnan, 1996, 42), and $E(\bar{X}) = E(X) = \beta$ so that for $n = 2m+1$, it follows from Theorem 3.1 that

$$\sum_{j=1}^{m+1} (2m+2-j)^{-1} \leq 2 \tag{3.5}$$

2. For the above gamma distribution with p.d.f.

$$f(x) = \frac{1}{\beta^{\alpha}\,\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-x/\beta} \quad 0 \leq x,\ 0 < \beta,\ 0 < \alpha$$

the expected value of the ith order statistic $X_{(i)}$ is given by

$$E\left(X_{(i)}\right) = \frac{n\beta}{\Gamma(\alpha)} \binom{n-1}{i-1} \int_0^{\infty} \left(\frac{\Gamma(\alpha;x)}{\Gamma(\alpha)}\right)^{i-1} \int_0^{\infty} \left(1 - \frac{\Gamma(\alpha;x)}{\Gamma(\alpha)}\right)^{n-i} x^{\alpha} e^{-x} dx$$

(see Harter and Balakrishnan, 1996, 45), and $E(\bar{X}) = E(X) = \alpha\beta$ so that for $n = 2m+1$, it follows from Theorem 3.1 that

$$\int_0^{\infty} \left(\frac{\Gamma(\alpha;x)}{\Gamma(\alpha)} - \frac{\Gamma^2(\alpha;x)}{\Gamma^2(\alpha)}\right)^{m} x^{\alpha} e^{-x} dx \leq 2\Gamma(\alpha+1)\, B(m+2, m) \tag{3.6}$$

3. For the above Weibull distribution with p.d.f.

$$f(x) = \frac{\alpha}{\beta^{\alpha}}\, x^{\alpha-1}\, e^{-(x/\beta)^{\alpha}},\ 0 \leq x,\ 0 < \beta,\ 0 < \alpha$$

the expected value of the ith order statistic $X_{(i)}$ is given by

$$E\left(X_{(i)}\right) = n\beta\, \Gamma(1+1/k) \sum_{j=0}^{i-1} (-1)^{i-1+j} \binom{i-1}{j} (n-j)^{-1-1/k}$$

(see Harter and Balakrishnan, 1996, 44), and $E(\bar{X}) = E(X) = \beta\,\Gamma(1+1/k)$ so that for $n = 2m+1$, it follows from Theorem 3.1 that

$$\sum_{j=0}^{m} (-1)^{m+j} \binom{m}{j} (2m+1-j)^{-1-1/k} \le 2B(m+1, m+1) \tag{3.7}$$

4. For the above Pareto distribution with p.d.f.

$$f(x) = \alpha\theta^{\alpha}\, x^{-\alpha-1}\ ,\ 0 < \theta \le x\ ,\ \ 0 < \alpha$$

the expected value of the ith order statistic $X_{(i)}$ is given by

$$E\left(X_{(i)}\right) = \frac{\Gamma(n+1)\Gamma(n-i+1-1/\alpha)}{\Gamma(n-i+1)\Gamma(n+1-1/\alpha)}\ \theta$$

(see Harter and Balakrishnan, 1996, 71), and $E(\bar{X}) = E(X) = \alpha\theta(\alpha-1)^{-1}$, $1 < \alpha$ so that for $n = 2m+1$, it follows from Theorem 3.1 that

$$\frac{\Gamma(2m+2)}{\Gamma(m+2)} \frac{\Gamma(m+2-1/\alpha)}{\Gamma(2m+2-1/\alpha)} \le \frac{2\alpha}{\alpha-1} \tag{3.8}$$

5. For the two parameter exponential distribution with p.d.f.

$$f(x) = \frac{1}{\sigma}\ \mathrm{e}^{-(x-\alpha)/\sigma}\ ,\ 0 \le \alpha \le x\ ,\ \ 0 < \sigma$$

the expected value of the ith order statistic $X_{(i)}$ is given by

$$E\left(X_{(i)}\right) = \alpha + \sigma \sum_{j=1}^{i} (n-j+1)^{-1}$$

(see Harter and Balakrishnan, 1996, 92), and $E(\bar{X}) = E(X) = 2(\alpha+\sigma)\theta$ so that for $n = 2m+1$, it follows from Theorem 3.1 that

$$\sum_{j=1}^{m+1} (2m+2-j)^{-1} \le 2 + \alpha/\sigma \tag{3.9}$$

4. Acknowledgements

The authors gratefully acknowledge the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia.

References

Eisenhauer, J. G. (1993). A measure of relative dispersion. *Teaching Statistics*, **15**(2), 37-39.

Harter, H.L. and Balakrishnan, N. (1996). Tables for the Use of Order Statistics in Estimation. CRC Press, New York.

Mcleod, A.J. and Henderson, G.R. (1984). Bounds for the sample standard deviation. *Teaching Statistics*, **6**(3), 72-76.

Shiffler, R.E. and Harsha, P.D. (1980). Upper and lower bounds for the sample standard deviation. *Teaching Statistics*, **2**(3), 84-86.

File: c\ pedastat \ q06cstma.doc

*Algebraic Inequalities for Standard Deviation**

Anwar H. Joarder and A. Laradji

King Fahd University of Petroleum and Minerals

ABSTRACT Some upper and lower bounds for sample standard deviation are established in terms of sample mean, sample median, sample range, the smallest order statistic and the largest order statistic. Upper bounds for variance are also derived for odd and even sample sizes whenever the sample observations are of the same sign. They are used to find bounds for some well-known sample statistics: z-scores, coefficient of variation, coefficient of skewness and the least squares estimator of the slope parameter in the context of a simple linear regression. Statistical estimation of related parameters can be improved on the basis of these fixed sample properties.

Keywords: Inequalities in statistics; sample mean; sample median; standard deviation; z -score, coefficient of variation; coefficient of skewness; regression parameters.

5. Introduction

Let X be a random variable with mean μ and standard deviation σ. For $0 < p < 1$, the p th quantile x_p of X is defined by $P(X \le x_p) \ge p$ and $P(X \ge x_p) \ge 1-p$. For example if $p = 1/2$, then $x_p = \tilde{\mu}$, the median of the random variable X. Page and Murty (1983) published an elementary proof of the inequality $|\tilde{\mu} - \mu| \le \sigma$. O'Cinneide (1990) presented a new proof for $|\tilde{\mu} - \mu| \le \sigma$ and stated the following generalization.

Proposition 1.1 Let X be a random variable with mean μ and standard deviation σ. Then for $0 < p < 1$ and $q = 1 - p$, the following inequality holds

$|x_p - \mu| \le \sigma \quad \max\left(\sqrt{p/q},\ \sqrt{q/p}\right)$ where x_p is the p th quantile.

For $p = 1/2$, it follows from the above proposition that $|\tilde{\mu} - \mu| \le \sigma$. Dharmadhikari (1991) noted that for $p \ne 1/2$, the inequality is somewhat unsatisfactory. The refined inequality proved by her with the help of one-sided Chebyshev inequality is stated in the following theorem.

Proposition 1.2 Let X be a random variable with mean μ and standard deviation σ. Then for $0 < p < 1$ and $q = 1 - p$, the following inequalities hold

---

* The paper is based on a Technical report "Inequalities in descriptive statistics", King Fahd University of Petroleum and Minerals, Saudi Arabia. ???

$$\mu - \sigma\sqrt{q/p} \le x_p \le \mu + \sigma\sqrt{q/p} \ .$$

Both Anwar H. Joarder and A. Laradji are Associate Professors in the Department of Mathematical Sciences at King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia, Emails: anwarj@ kfupm.edu.sa, alaradji@kfupm.edu.sa

For stimulating discussions, readers may go through Mallows (1991) and the references therein. A
more general inequality than that in Proposition 1.2 relating sample standard deviation to mean and the i-th order statistic discussed by David (1988) and David (1991) is presented in Theorem 1.2. Interested readers can go through the references in David (1988) for bounds of order statistics.

Sample standard deviation (s) or variance (s^2) is nonnegative, and is defined by $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$. But for most data sets, the range of s is much narrower than the nonnegative part of the real line. Some representations of sample variance are discussed in Joarder (2002). Further it has been proved by Joarder (2003) that if a computer program is used to calculate sample variance, then it can be efficiently calculated by the representation based on the first order differences of observations.

It is well known that $\sqrt{n-1}\ s \ge \sqrt{n}\ u$ where u is the mean absolute deviation of sample values around the mean defined by $nu = \sum_{i=1}^{n} |x_i - \bar{x}|$. Let n ordered sample observations be denoted by $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. It is also well known that for $n = 2$, the sample variance has a simpler form given by $s^2 = w^2/2$ where $w = x_{(n)} - x_{(1)}$ is the sample range. Shiffler and Harsha (1980) have formulated an upper bound for the sample standard deviation (s) in terms of the sample range w, while Mcleod and Henderson (1984) have determined a lower bound for s in terms of w which appeared originally in Thomson (1955). Eisenhauer (1993) combined them. A stronger version of these results with more transparent arguments is provided in Theorem 2.1.

Theorem 1.1 (Macleod and Henderson (1984) and (Shiffler and Harsha (1980)). Let w and s denote, respectively, the range and standard deviation of a sample of size $n \ge 2$. Then

$$\frac{w}{\sqrt{2(n-1)}} \le s \le \frac{w}{2}\sqrt{\frac{n}{n-1}} \ .$$

Theorem 1.2 (David, 1991) For $1 \le i \le n$, let $x_{(i)}$ be the ith order statistic and s, the standard deviation based on a sample of size $n \ge 2$. Then

$$|x_{(i)}-\bar{x}| \le s \max\left(\sqrt{\frac{(n-1)(i-1)}{n(n+1-i)}}, \sqrt{\frac{(n-1)(n-i)}{ni}}\right).$$

By the use of Theorems 1.1 and 1.2 we immediately obtain the following corollaries:

Corollary 1.1 For $1 \le i \le n$, let $x_{(i)}$ be the i th order statistic from a sample of size $n \ge 2$. Then

$$\sqrt{\frac{n}{\max(i-1,n-i)}-1}\ \sqrt{\frac{n}{n-1}}\ |x_{(i)}-\bar{x}| \le s \le \frac{w}{2}\sqrt{\frac{n}{n-1}}.$$

Corollary 1.2 (Eisenhauer, 1993) Let w and s denote the range and standard deviation of a sample of size n. Then

(a) $\frac{1}{\sqrt{2(n-1)}} \le \frac{s}{w} \le \frac{1}{2}\sqrt{\frac{n}{n-1}}$,

(b) $0 \le s/w \le 1/2$ as $n \to \infty$.

6. Some Results on the Inequalities in Descriptive Statistics

The following result is a refined version of Theorem 1.1.

Theorem 2.1 Let $\bar{x}, \tilde{x}, w$ and s respectively denote the mean, median, range and standard deviation of a sample of size n. Then

$$\frac{w}{\sqrt{2\ (n-1)}} \le \sqrt{\frac{w^2}{2(n-1)}+\frac{(\tilde{x}-\bar{x})^2}{2}} \le s \le \sqrt{\frac{n(\bar{x}-x_{(1)})(x_{(n)}-\bar{x})}{n-1}} \le \frac{w}{2}\sqrt{\frac{n}{n-1}}.$$

Proof. For any a and $n \ge 2$, $(n-1)s^2 = \sum_{i=1}^{n}(x_i-a)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}(x_i-a)\right)^2$ so that for the ordered sample observations $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, we have $(n-1)s^2 \le n\max\left((x_{(1)}-a)^2, (x_{(n)}-a)^2\right) - n(\bar{x}-a)^2$. In particular, for $a = (x_{(1)}+x_{(n)})/2 = x_{(1)} + w/2$, we have

$$(n-1)s^2 \le nw^2/4 - n\left((\bar{x}-x_{(1)})-w/2\right)^2 = n(\bar{x}-x_{(1)})(x_{(n)}-\bar{x}). \quad (2.1)$$

Since $ab \le (a+b)^2/4$ for any real numbers a and b, we deduce by putting $a = \bar{x}-x_{(1)} \ge 0,\ b = x_{(n)}-\bar{x} \ge 0$ that $(\bar{x}-x_{(1)})(x_{(n)}-\bar{x}) \le w^2/4$. Then the two

inequalities in the theorem are evident by (2.1). Next, by using $2(a^2+b^2) \geq (a-b)^2$, for any $y_1 \leq y_2 \leq \cdots \leq y_n$ with $w_y = y_n - y_1$, we have

$$2\sum_{i=1}^{n} y_i^2 = 2\left(y_n^2 + y_1^2\right) + 2\left(y_2^2 + \cdots + y_{n-1}^2\right) \geq w_y^2 + 2\left(y_2^2 + \cdots + y_{n-1}^2\right). \quad (2.2)$$

For n odd, rewrite (2.2) as

$$2\sum_{i=1}^{n} y_i^2 = w_y^2 + 2\left(y_2^2 + \cdots + y_{(n+1)/2}^2\right) + 2\left(y_{(n+3)/2}^2 + \cdots + y_{n-1}^2\right). \quad (2.3)$$

If $\tilde{y} < 0$ then for $n \geq 3$ and $2 \leq i \leq (n+1)/2$, we have $y_i \leq \tilde{y}$, or $y_i^2 \geq \tilde{y}^2$, so it follows from (2.3) that $2\sum_{i=1}^{n} y_i^2 \geq w_y^2 + 2\left((n+1)/2 - 2 + 1\right)\tilde{y}^2 = w_y^2 + (n-1)\tilde{y}^2$. If $\tilde{y} \geq 0$ then for $n \geq 3$ and $(n+1)/2) \leq i \leq (n-1)$, we have $\tilde{y} \leq y_i$, so that it follows from (2.3) that $2\sum_{i=1}^{n} y_i^2 \geq w_y^2 + 2(n-1-(n+1)/2+1)\tilde{y}^2 = w_y^2 + (n-1)\tilde{y}^2$.

Next, let n be even. For $n = 2$ we have
$2\left(y_1^2 + y_2^2\right) = (y_2 - y_1)^2 + (y_2 + y_1)^2 = w_y^2 + (2\tilde{y})^2 \geq w_y^2 + 2\tilde{y}^2$.

For $n = 4$, we have $2\sum_{i=1}^{4} y_i^2 = (y_4 - y_1)^2 + (y_3 - y_2)^2 + (y_4 + y_1)^2 + (y_3 + y_2)^2 \geq w_y^2 + 4\tilde{y}^2$.

Since $2(a^2+b^2) \geq (a+b)^2$, rewrite (2.2) for even $n \geq 2$ as

$$2\sum_{i=1}^{n} y_i^2 \geq w_y^2 + 2\left(y_2^2 + \cdots + y_{n/2-1}^2\right) + \left(y_{n/2} + y_{n/2+1}\right)^2 + 2\left(y_{n/2+2}^2 + \cdots + y_{n-1}^2\right)$$
(2.4)

If $\tilde{y} < 0$, then for even $n \geq 6$ and $n/2 \leq i \leq n/2 - 1$, we have $y_i \leq \tilde{y}$, i.e. $y_i^2 \geq \tilde{y}^2$, and it follows from (2.4) $2\sum_{i=1}^{n} y_i^2 \geq w_y^2 + 2(n/2 - 1 - 2 + 1)\tilde{y}^2 + (2\tilde{y})^2 \geq w_y^2 + n\tilde{y}^2$.

Similarly, if $\tilde{y} \geq 0$, then for even $n \geq 6$ and $n/2+2 \leq i \leq n-1$, we have $y_i \leq \tilde{y}$, and it follows from (2.4) that
$2\sum_{i=1}^{n} y_i^2 \geq w_y^2 + 2\left(y_2^2 + \cdots + y_{n/2-1}^2\right) + \left(2\tilde{y}\right)^2 + 2\left(n - 1 - (n/2+2) + 1\right)\tilde{y}^2\right) \geq w_y^2 + n\tilde{y}^2$. So for even $n \geq 2$, $2\sum_{i=1}^{n} y_i^2 \geq w^2 + n\tilde{y}^2$. In all cases for $n \geq 2$, putting $y_i = x_i - \bar{x}$ $(1 \leq i \leq n)$, we have

$$2(n-1)s^2 \geq \begin{cases} w^2 + (n-1)(\tilde{x}-\bar{x})^2 & \text{if } n \text{ is odd} \\ w^2 + n(\tilde{x}-\bar{x})^2 & \text{if } n \text{ is even} \end{cases}$$

as required.

It may be remarked here that the two sides of the rightmost inequality in the theorem are equal in case $x_{(1)} + x_{(n)} = 2\bar{x}$, otherwise the sharper inequality $\sqrt{(\bar{x} - x_{(1)})(x_{(n)} - \bar{x})} < w/2$ holds. The following result improves the bound for s described in Theorem 1.2. It is also in agreement with the known result that for any two observations $s = w/\sqrt{2}$ where $w = x_{(n)} - x_{(1)}$. In what follows let $[n]$ be the greatest integer function i.e. it is the largest integer not exceeding n.

Corollary 2.1 For any sample of $n \geq 2$ observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$,

$$s^2 \geq \frac{1}{2(n-1)} \sum_{i=1}^{[n/2]} w_i^2 = \frac{w_1^2}{2(n-1)} \text{ where } w_i = x_{(n-i+1)} - x_{(i)}, \ (1 \leq i \leq m).$$

Proof. For any real numbers $y_1, y_2, \cdots, y_n$ and $m = [n/2]$ we have

$$\sum_{i=1}^{n} y_i^2 \geq \left(y_1^2 + y_n^2\right) + \left(y_2^2 + y_{n-1}^2\right) + \cdots + \left(y_m^2 + y_{n-m+1}^2\right) \geq \frac{1}{2}\left[\left(y_n - y_1\right)^2 + \cdots + \left(y_{n-m+1} - y_m\right)^2\right]$$

Letting $y_i = x_{(i)} - \bar{x}$, $(i = 1, 2, \cdots, m)$ and $w_j = y_{n-j+1} - y_j$, $(1 \leq j \leq m)$, it follows from the above inequality that $2(n-1)s^2 \geq w_1^2 + w_2^2 + \cdots + w_m^2 \geq w_1^2$.

Theorem 2.2 For $1 \leq i \leq n$, let $x_{(i)}$ be the i th order statistic from a sample of size $n \geq 2$. Then

$(i) \ |x_{(i)} - \bar{x}| \leq s \ \frac{n-1}{\sqrt{n}}$ for each i

$(ii) \ |x_{(i)} - \bar{x}| \leq s \ \frac{n-1}{\sqrt{n(n+1)}} \leq s$ if $i = \frac{n+1}{2}$

$(iii) \ |\tilde{x} - \bar{x}| \leq s\sqrt{\frac{n-1}{n}} < s$

Proof. (i) $\max_i |x_{(i)} - \bar{x}| \leq \max(x_{(n)} - \bar{x}, \ \bar{x} - x_{(1)})$. But by Theorem 1.2 we have

$$x_{(n)} - \bar{x} \leq s \ \max\left(\frac{n-1}{\sqrt{n}}, \ 0\right) = s \ \frac{n-1}{\sqrt{n}} \text{ and } \bar{x} - x_{(1)} \leq s \ \max\left(0, \ \frac{n-1}{\sqrt{n}}\right) = s \ \frac{n-1}{\sqrt{n}}.$$

(ii) Use Theorem 1.2

(iii) If n is odd and $i=(n+1)/2$, then $x_{(i)}=\tilde{x}$, and $\frac{n-1}{\sqrt{n(n+1)}}\le\sqrt{\frac{n-1}{n}}$ for any $n\ge 1$, the inequality follows from *(ii)*. If n even, then $\tilde{x}=\left(x_{(n/2)}+x_{(n/2+1)}\right)/2$, and the leftmost inequality in (iii) follows from

$$|\tilde{x}-\bar{x}| \le 2^{-1}\left|\left(x_{(n/2)}-\bar{x}\right)+\left(x_{(n/2+1)}-\bar{x}\right)\right| \le 2^{-1}\left(|x_{(n/2)}-\bar{x}|+|x_{(n/2+1)}-\bar{x}|\right)$$

by virtue of Theorem 1.2.

Let the z-scores be defined by $z_i=(x_i-\bar{x})/s,\ i=1,2,\cdots,n$. Then it follows from $\sum_{i=1}^{n} z_i^2=n-1$ that $n\max_i\{|z_i|\}^2\ge n-1$ so that, by virtue of (i),

$\min_i\{|z_i|\}\le\sqrt{\frac{n-1}{n}}<\max_i\{|z_i|\}\le\frac{n-1}{\sqrt{n}}$ as in Hayes(2004). The upper bound is originally by Pearson and Chandrashekhar (1936).

The inequality in (iii) tells us that $\tilde{x}-s<\bar{x}<\tilde{x}+s$, or, $\bar{x}-s<\tilde{x}<\bar{x}+s$. That is sample mean and median lie within one standard deviation of each other. The following corollary is obvious from Theorem 2.1 and Theorem 2.2.

Corollary 2.2 Let $\bar{x}$ and $\tilde{x}$ be the sample mean and median based on a sample of size $n\ge 2$. Then

$$\max\left(\sqrt{\frac{n}{n-1}}\,|\bar{x}-\tilde{x}|\ ,\ \sqrt{\frac{w^2}{2(n-1)}+\frac{(\bar{x}-\tilde{x})^2}{2}}\right)\le s\le\ \frac{w}{2}\sqrt{\frac{n}{n-1}}$$

Theorem 2.3 If the observations are of the same sign, then for any sample size $n\ge 2$, the following inequalities hold:

(i) $s^2\le n\bar{x}^2-\frac{n+1}{4}\tilde{x}^2$ if n is odd,

(ii) $s^2\le n\bar{x}^2-\frac{n(n-2)}{4(n-1)}\tilde{x}^2$ if n is even.

Proof. Without loss of generality we assume that all the observations are nonnegative. If n is odd, there are $\binom{(n+1)/2}{2}$ products of the form x_ix_j where $1\le i<j\le n$ and $x_i,x_j\ge\tilde{x}$. Then $(n\bar{x})^2=\sum_{i=1}^{n}x_i^2+2\sum_{i<j}x_ix_j\ge\sum_{i=1}^{n}x_i^2+2\left(\frac{n^2-1}{8}\right)\tilde{x}^2$. If n is even, there are $\binom{n/2}{2}$ products of the form x_ix_j where $1\le i<j\le n$ and $x_i,x_j\ge\tilde{x}$. Then

$(n\bar{x})^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{i<j} x_i x_j \geq \sum_{i=1}^{n} x_i^2 + 2\left(\frac{(n/2)(n/2-1)}{2}\right)\tilde{x}^2$. The rest of the proof is immediate.

The following corollary follows from Theorem 2.1 and Theorem 2.3.

Corollary 2.3 For any sample of $n \geq 2$ observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, if all the observations are of the same sign. the following inequalities hold:

$$(i)\ s \leq \min\left(\sqrt{\frac{n(\bar{x}-x_{(1)})(x_{(n)}-\bar{x})}{n-1}}, \sqrt{n\bar{x}^2 - \frac{n+1}{4}\tilde{x}^2}\right) \text{ for odd } n,$$

$$(ii)\ s \leq \min\left(\sqrt{\frac{n(\bar{x}-x_{(1)})(x_{(n)}-\bar{x})}{n-1}}, \sqrt{n\bar{x}^2 - \frac{n(n-2)}{4(n-1)}\tilde{x}^2}\right) \text{ for even } n.$$

Corollary 2.4 For any sample size $n \geq 2$, if all the observations are of the same sign, the following inequalities hold:

$$\frac{n}{n-1}(\bar{x}-\tilde{x})^2 \ \leq\ s^2 \leq n\bar{x}^2 - \frac{n(n-2)}{4(n-1)}\tilde{x}^2 \leq n\bar{x}^2 \quad .$$

Proof. For any $n \geq 2$, it follows from Theorem 2.3 that

$n(n-1)\bar{x}^2 \geq (n-1)s^2 + \frac{n(n-2)}{4}\tilde{x}^2$ i.e.

$n\bar{x}^2 \geq s^2 + \frac{n(n-2)}{4(n-1)}\tilde{x}^2$. The leftmost inequality is by virtue of Theorem 2.2 (iii).

Note that if all the observations are of the same sign, a less sharper but simpler than the above inequality is given by $|\bar{x}-\tilde{x}| \ \leq\ s \leq \sqrt{n}\ |\bar{x}|$. The following corollary is by virtue of Theorem 2.1 and Corollary 2.4.

Corollary 2.5 For any sample size $n \geq 2$, if all the observations are of the same sign, the following inequalities hold:

$$\max\left(\sqrt{\frac{w^2}{2(n-1)} + \frac{(\tilde{x}-\bar{x})^2}{2}}, \sqrt{\frac{n}{n-1}}\,|\tilde{x}-\bar{x}|\right)$$

$$\leq s \leq \sqrt{\frac{n}{n-1}}\min\left(\sqrt{(n-1)\bar{x}^2 - \frac{n-2}{4}\tilde{x}^2},\ \sqrt{(\bar{x}-x_{(1)})(x_{(n)}-\bar{x})}\right)$$

The following corollary is obvious by virtue of Theorem 1.2 and Corollary 2.3.

Corollary 2.6 For any sample of *n* nonnegative observations ($n \geq 2$), the following inequalities hold:

$$\frac{1}{\sqrt{n-1}} \max\left(\frac{w}{\sqrt{2}},\ \sqrt{n}\,|\,\bar{x}-\tilde{x}\,|\right) \leq s \leq\ \min\left(\sqrt{n}\ \ |\bar{x}\,|,\quad \frac{w}{2}\sqrt{\frac{n}{n-1}}\right).$$

The following result is due to Laradji and Joarder (2002).

Theorem 2.4 For any sample of $n \geq 2$ observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the following inequalities hold:

$$(i)\ \frac{1}{2}\left[\left(1+\frac{1}{n}\right)\tilde{x}+\left(1-\frac{1}{n}\right)x_{(1)}\right] \leq \bar{x} \leq \frac{1}{2}\left[\left(1+\frac{1}{n}\right)\tilde{x}+\left(1-\frac{1}{n}\right)x_{(n)}\right]$$

$$(ii)\ |\,\tilde{x}-\bar{x}\,| \leq \frac{n-1}{n+1}\ \max\left(\bar{x}-x_{(1)}, x_{(n)}-\bar{x}\right)$$

$$(iii)\ \left|\frac{\tilde{x}}{\bar{x}}-1\right| \leq 1$$

7. Inequalities for some Useful Statistics

The following corollaries are obvious from Theorem 2.2 and Corollary 2.4 respectively.

Corollary 3.1 If $n \geq 2$ observations are positive, then the coefficient of variation $CV(x) = s/\bar{x}$ satisfies the following inequalities:

$$\left|\frac{\tilde{x}}{\bar{x}}-1\right| \leq \sqrt{\frac{n}{n-1}}\left|\frac{\tilde{x}}{\bar{x}}-1\right| \leq CV(x) \leq \sqrt{n}\ .$$

Corollary 3.2 If $n \geq 2$ observations are positive, then the coefficient of skewness $CS(x) = \dfrac{\bar{x}-\tilde{x}}{s/3}$ satisfies the following inequalities:

$$-3\sqrt{\frac{n-1}{n}} \leq CS(x) \leq 3\sqrt{\frac{n-1}{n}}$$

which is slightly narrower than the known interval $[-3, 3]$.

Theorem 3.1. Let $w_y = y_{(n)} - y_{(1),} w_x = x_{(n)} - x_{(1)}, s_{xy} = \sum (x-\bar{x})(y-\bar{y}), s_{xx} = s_x^2$**.** Then the regression coefficient $\hat{\beta}_1 = s_{xy}/s_{xx}$ satisfies the following inequalities:

$$(i)\quad -\frac{s_y}{s_x} \leq \hat{\beta}_1 \leq \frac{s_y}{s_x}$$

$$(ii) \quad -\sqrt{\frac{n}{2}} \quad \frac{w_y}{w_x} \le \hat{\beta}_1 \le \sqrt{\frac{n}{2}} \quad \frac{w_y}{w_x}$$

Proof. The sample correlation coefficient (r) is defined by $rs_x s_y = s_{xy} = \hat{\beta}_1 s_{xx}$ so that the inequality in (i) follows by virtue of $-1 \le r \le 1$. The proof for part (ii) is immediate by virtue of Theorem 1.2.

Acknowledgements

The authors gratefully acknowledge an anonymous referee, an associate editor and Professor K.C. Chang for constructive suggestions that have improved the presentation and readability of the paper. The authors also take this opportunity to acknowledge King Fahd University of Petroleum and Minerals, Saudi Arabia for providing excellent research facilities.

References

David, H.A. (1988). General bounds and inequalities in order statistics. *Communications in Statistics – Theory and Methods*, **17**, 2119-2134.

David, H.A. (1991). Mean minus median: A comment on O'Cinneide. *American Statistician*. **45**(3), 257.

Dharmadhikari, S. (1991). Bounds on quantiles: A comment on O'Cinnneide. *American Statistician*. **45** (3), 258.

Eisenhauer, J. G. (1993). A measure of relative dispersion. *Teaching Statistics*, **15**(2), 37-39.

Hayes, K. (2004). A lower bound for the most deviant z-score. *Teaching Statistics*, 26, 3, 89-91.

Joarder, A.H. (2002). On some representations of sample variance. *International Journal of Mathematical Education in Science and Technology*, 33(5), 772-784.

Joarder, A.H. (2003). Sample variance and first-order differences of observations. *Mathematical Scientist*, 28, 129-133.

Laradji, A. and Joarder, A.H. (2002). Inequalities involving sample mean, sample median and extreme observations. *Technical Report*, No. 283. Department of Mathematical Sciences. King Fahd University of Petroleum and Minerals. Saudi Arabia.

Mallows, C. (1991). Another comment on O'Cinneide. *American Statistician*. 1991, **45** (3), 257.

Mcleod, A.J. and Henderson, G.R. (1984). Bounds for the sample standard deviation. *Teaching Statistics*, **6**(3), 72-76.

O'Cinneide, C. A. (1990). The mean is within one standard deviation of any median. *American Statistician*, **44** (4), 292.

Page, W. and Murty, V.N. (1983). Nearness relations among measures of central tendency and dispersion: part 2. *Two Year College Mathematics Journal*. 14, 8-17.

Pearson, E.S. and Chandra Sekhar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28, 308-320.

Shiffler, R.E. and Harsha, P.D. (1980). Upper and Lower Bounds for the sample standard deviation. *Teaching Statistics*, **2**(3), 84-86.

Thomson, G.W. (1955). Bounds for the ratio of range to standard deviation. Biometrika, 42, 268-269.

File: c\ pedastat \ p47ijmesta.doc

*The dependence structure of conditional probabilities in a contingency table**

ANWAR H. JOARDER and WALID S. AL-SABAH
Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Dhahran 312361, Saudi Arabia, Emails: anwarj@kfupm.edu.sa and walid@kfupm.edu.sa

Conditional probability and statistical independence can better be explained with contingency tables. In this note some special cases of 2×2 contingency table is considered. In turn an interesting insight into statistical dependence as well as independence of events is obtained.

Keywords: Conditional probability; contingency table; incidence matrix; singularity; statistical independence

1. Introduction

Elementary probabilities are obtained for the outcomes of situations conveniently called random experiments. They are usually taught with the help of examples of dice, coins and cards. Not everybody feels comfortable with these approaches. Experience shows that
conditional probability and statistical independence can better be explained with contingency tables often encountered by them in real life. Consider a general 2×2 contingency table

	B_1	B_2
A_1	n_{11}	n_{12}
A_2	n_{21}	n_{22}

* [3] Published in *International Journal of Mathematical Education in Science and Technology*, 33(3), 475-480 [London, UK]

The matrix given by

$$N = \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$$

will hereinafter be called incidence matrix. In this note some special cases of 2×2 contingency table is considered. In turn a relation is observed between the dependence structure of conditional probabilities, nonsingularity of the incidence matrix N formed by the square contingency table, and statistical dependence of events. The properties that are going to be discussed here will also be true for any $r \times c$ contingency table collapsed as a 2×2 contingency table.

The notion of statistical independence is closely related to conditional probability. Given that B happens, the probability is

$$\frac{P(A \cap B)}{P(B)}$$

that the event A happens. The above ratio is usually denoted by $P(A \mid B)$ i.e.

$$\frac{P(A \cap B)}{P(B)} = P(A \mid B)\,. \qquad (1.1)$$

The left hand side of (1.1) should be emphasized to the students as the right hand side is usually misunderstood by them. If the ratio is the same as $P(A)$, it implies that B does not affect the occurrence of $A.$ In other word, A is statistically independent of B. Thus in this case it follows from (1.1) that

$$P(A \cap B) = P(A)P(B) \qquad (1.2)$$

which is used as the definition of statistical independence in many books. It follows from (1.2) that if A is statistically independent of B, then B is statistically independent of A.

Consider the independence of the categories of two attributes A and $B.$ By definition each pair of events $(i)\ A_1$ and B_1 $(ii)\ A_1$ and B_2 $(iii)\ A_2$ and B_1 and $(iv)\ A_2$ and B_2 are

independent if the following conditions hold:

$$
\begin{aligned}
&(i)\ P(A_1 \mid B_1) = P(A_1) \\
&(ii)\ P(A_1 \mid B_2) = P(A_1) \\
&(iii)\ P(A_2 \mid B_1) = P(A_2) \text{ and} \\
&(iv)\ P(A_2 \mid B_2) = P(A_2)
\end{aligned}
\qquad (1.3)
$$

respectively. But it is straightforward to prove that the above four $(=2^2)$ conditions are equivalent (Hines and Montgomery, 1990, p.51). Thus if A_1 and B_1 are independent, then so are (a) A_1 and B_2 , (b) A_2 and B_1 and (c) A_2 and B_2. That is if any pair of events are independent in a 2×2 table, then other three pair of events in (1.3) are independent and not mutually exclusive.

In what follows we provide two interesting results that provide some insight into statistical independence. They follow from the rearrangement of the equations in (1.3).

(1) For any contingency table having attributes A and B with categories $A_1,\ A_2$ and categories $B_1,\ B_2$ respectively, the events A_1 and B_1 are independent *if and only* if $A \mid B_1$ and $A \mid B_2$ have the same probability distribution i.e.

$$
\begin{aligned}
&(i) \quad P(A_1 \mid B_1) = P(A_1 \mid B_2) \\
&(ii) \quad P(A_2 \mid B_1) = P(A_2 \mid B_2)
\end{aligned}
$$

(1.4)

In a 2×2 contingency table it is conventional to write $A_1 = A$ and $B_1 = B$ so that $A_2 = \overline{A}$ and $B_2 = \overline{B}$. To explain (1.4), consider the following example of the breakdown of computers having circuit boards for a modem (A) or for a printer (B):

	A	$\overline{A}$	
B	10	15	25

$\overline{B}$	30	45	75
	40	60	100

The events A and B are independent *if and only* if $A \mid B$ and $A \mid \overline{B}$ have the same probability distribution i.e.

(i) $P(A \mid B),\ P(\overline{A} \mid B)$ and

(ii) $P(A \mid \overline{B}),\ P(\overline{A} \mid \overline{B})$

are the same. Since the two sets of probabilities

$$(i) \quad P(A \mid B) = \frac{10}{25} = 0.40,\ P(\overline{A} \mid B) = \frac{15}{25} = 0.60 \text{ and}$$

$$(ii) \quad P(A \mid \overline{B}) = \frac{30}{75} = 0.40,\ P(\overline{A} \mid \overline{B}) = \frac{45}{75} = 0.60$$

are the same, the events A and B are independent .

(2) For any contingency table having attributes A and B with categories $A_1,\ A_2$ and categories $B_1,\ B_2$ respectively, the events A_1 and B_1 are independent if and only

$$(i)\ P(A_1 \mid B_1) = P(A_1 \mid B_2) = P(A_1) \text{ and}$$

$$(ii)\ P(A_2 \mid B_1) = P(A_2 \mid B_2) = P(A_2)$$

(1.5)

The equation (i) of (1.5) says that neither the occurrence of B_1 nor B_2 affects the occurrence of A_1. Similarly the equation *(ii)* of (1.5) indicates that neither the occurrence of B_1 nor B_2 affects the occurrence of A_2.

In what follows we provide two other interesting results that are special cases of a 2×2 contingency table:

(1) For any contingency table having attributes A and B with categories $A_1,\ A_2$ and categories $B_1,\ B_2$ respectively, the following holds:

$P(A_1 \cap B_1) = P(A_2 \cap B_2)$ if and only if $P(A_1) = P(A_2),\ P(B_1) = P(B_2)$.

This means that the 2×2 incidence matrix has equal diagonal elements.

(2) For any contingency table having attributes A and B with categories $A_1,\ A_2$ and categories $B_1,\ B_2$ respectively, the following holds:

$$\frac{P(A_1)}{P(A_2)} = \frac{P(B_1)}{P(B_2)}$$

if and only if $P(A_1 \cap B_1) + P(A_2 \cap B_2) = P(A_1 \cap B_2) + P(A_2 \cap B_1)$.

This implies that the sum of the diagonal elements is the same as that of the off-diagonal elements. Thus the probability of having exactly one of the two attributes is the same as having none or both the attributes.

2. The Main Result

The main result is presented below in the form of a theorem.

Theorem 2.1 For any contingency table having attributes A and B with categories $A_1,\ A_2$ and $B_1,\ B_2$ respectively, the incidence matrix has the following implications:

(a) $P(A_1 \mid B_1) < P(A_1) < P(A_1 \mid B_2) \text{ iff } |N| < 0$ (2.1)

(b) $P(A_1 \mid B_1) = P(A_1) = P(A_1 \mid B_2) \text{ iff } |N| = 0$ (2.2)

(c) $P(A_1 \mid B_1) > P(A_1) > P(A_1 \mid B_2) \text{ iff } |N| > 0$ (2.3)

Proof: *(a)* Let $P(A_1 \mid B_1) < P(A_1) < P(A_1 \mid B_2)$. Then

$$\frac{n_{11}}{n_{11}+n_{21}} < \frac{n_{11}+n_{12}}{n} \text{ and } \frac{n_{11}+n_{12}}{n} < \frac{n_{12}}{n_{12}+n_{22}} .$$

Writing out $n = n_{11}+n_{12}+n_{21}+n_{22}$ and simplifying, we have from each of the inequality

$$n_{11}n_{22} - n_{12}n_{21} < 0$$

or $n_{11}n_{22} < n_{12}n_{21}$ (i.e. $|N| < 0$). (2.4)

Again let $|N| < 0$, i.e. $n_{11}n_{22} < n_{12}n_{21}$. Now adding $n_{11}(n_{11}+n_{12}+n_{21})$ to both sides of this inequality, we have

$$n_{11}n_{22} + n_{11}(n_{11}+n_{12}+n_{21}) < n_{12}n_{21} + n_{11}(n_{11}+n_{12}+n_{21})$$

i.e. $n_{11}\, n < (n_{11}+n_{12})(n_{11}+n_{21})$.

Dividing both sides by $n(n_{11}+n_{21})$, we have

$$\frac{n_{11}}{n_{11}+n_{21}} < \frac{n_{11}+n_{12}}{n}, \quad \text{i.e. } P(A_1 \mid B_1) < P(A_1) .$$

Similarly by adding $n_{12}(n_{11}+n_{12}+n_{22})$ to both sides of (2.4), we have

$$n_{11}n_{22} + n_{12}(n_{11}+n_{12}+n_{22}) < n_{12}n_{21} + n_{12}(n_{11}+n_{12}+n_{22})$$

or, $(n_{11}+n_{12})(n_{12}+n_{22}) < n_{12}(n_{11}+n_{12}+n_{21}+n_{22})$

or, $(n_{11}+n_{12})(n_{12}+n_{22}) < n\, n_{12}$.

Dividing both sides of the resulting inequality by $n(n_{12}+n_{22})$, we have

$$\frac{n_{11}+n_{12}}{n} < \frac{n_{12}}{n_{12}+n_{22}}, \quad \text{i.e.} \quad P(A_1) < P(A_1 \mid B_2) \ .$$

(b) See Joarder (1998).

(c) The proof is similar to that in part (a) above.

The result in *(a)* here means that A_1 is less likely to happen if B_1 happens, while A_1 is more likely to happen if B_1 does not happen. The result in *(c)* similarly means that A_1 is more likely to happen if B_1 happens, while A_1 is less likely to happen if B_1 does not happen. The result in *(b)* means that the occurrence of B_1 does not affect the occurrence of A_1 and vice versa.

Part *(b)* implies that the events A_1 and B_1 are independent *if and only if* any of the following equivalent conditions is satisfied:

(i) rows are linearly dependent
(ii) columns are linearly dependent
(iii) the incidence matrix N is singular
(iv) $n_{ij} = \frac{n_{i.}n_{.j}}{n}$ where $n_{i.} = n_{i1} + n_{i2}$ and $n_{.j} = n_{1j} + n_{2j}$ $(i = 1, 2; j = 1, 2)$.

3. Some Illustrations

As earlier let $A_1 = A$ and $B_1 = B$ so that $A_2 = \overline{A}$ and $B_2 = \overline{B}$. To explain *(a)* of Theorem 2.1, consider the following the breakdown of a computer having modem boards (A) or printer boards (B):

	A	$\overline{A}$	
B	4	16	20
$\overline{B}$	36	44	80

	40	60	100

Here the following three probabilities

$$P(A \mid B) = \frac{4}{20} = 0.20,\ P(A) = \frac{40}{100} = 0.40,\ P(A \mid \overline{B}) = \frac{36}{80} = 0.45$$

are not the same. Observe that $|N| < 0$ and $P(A \mid B) < P(A) < P(A \mid \overline{B})$. This means that that computers without printer boards are more likely to have modem boards than computers with printer boards. In other words, they are statistically dependent.

Similarly, the probabilities

$$P(B \mid A) = \frac{4}{40} = 0.10,\ P(B) = \frac{20}{100} = 0.20,\ P(B \mid \overline{A}) = \frac{16}{60} \approx 0.26$$

are not the same. Observe that $|N| < 0$ and $P(B \mid A) < P(B) < P(B \mid \overline{A})$. This means that that computers without modem boards are more likely to have printer boards than computers with modem boards. In other words, they are statistically dependent.

To explain *(c)* of Theorem 2.1, consider the following the breakdown of a computer having a modem board (A) or a circuit board (B):

	A	$\overline{A}$	
B	12	8	20
$\overline{B}$	28	52	80
	40	60	100

Here the following three probabilities

$$P(A \mid B) = \frac{12}{20} = 0.60,\ P(A) = \frac{40}{100} = 0.40,\ P(A \mid \overline{B}) = \frac{28}{80} = 0.35$$

are not the same. Observe that $|N| > 0$ and $P(A \mid B) > P(A) > P(A \mid \overline{B})$. This means that that computers with printer boards are more likely to have modem boards than computers without printer boards. In other words, they are statistically dependent.

Similarly, the following three probabilities

$$P(B \mid A) = \frac{12}{40} = 0.30,\ P(B) = \frac{20}{100} = 0.20,\ P(B \mid \overline{A}) = \frac{8}{60} \approx 0.13,$$

are not the same. Observe that $|N| > 0$ and $P(B \mid A) > P(B) > P(B \mid \overline{A})$. This means that that computers with modem boards are more likely to have printer boards than computers without modem boards. In other words, they are statistically dependent.

To explain *(b)* of Theorem 2.1, consider the following the breakdown of a computer having a modem board (A) or a circuit board (B):

	A	$\overline{A}$	
B	10	15	25
$\overline{B}$	30	45	75
	40	60	100

Here the following three probabilities

$$P(A \mid B) = \frac{10}{25} = 0.40,\ P(A) = \frac{40}{100} = 0.40,\ P(A \mid \overline{B}) = \frac{30}{75} = 0.40$$

are the same. Observe that $|N| = 0$ and $P(A \mid B) = P(A) = P(A \mid \overline{B})$. The same is true for the following three probabilities:

$$P(B \mid A) = \frac{10}{40} = 0.25,\ P(B) = \frac{2}{100} = 0.25,\ P(B \mid \overline{A}) = \frac{15}{60} = 0.25,$$

Observe that $|N| = 0$ and $P(B \mid A) = P(B) = P(B \mid \overline{A})$. Since the above three probabilities are the same, it follows that having a modem has nothing to do with having a printer or vice versa. In other words, they are statistically independent.

The notions discussed here are also true for any $r \times c$ contingency table collapsed into an appropriate 2×2 contingency table with categories of interest. We remark that though Theorem 2.1 is proved in the context of contingency table, it is true for any two events

A and B where $A_1 = A, B_1 = B$ so that $A_2 = \overline{A}$, $B_2 = \overline{B}$ and

$$N = \begin{pmatrix} P(AB) & P(A\overline{B} \\ P(\overline{A}B) & P(\overline{A}\overline{B}) \end{pmatrix}.$$

Acknowledgements

The authors acknowledge the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia.

References

[1] Hines, W.W. and Montgomery, D.C., 1990, *Probability and Statistics in Engineering and Management Sciences*. John Wiley and Sons, New York.

[2] Joarder, A.H., 1998, On the statistical independence in a contingency table. *International Journal of Mathematical Education in Science and Technology*. 29, No.5, 780-782

References

1. Joarder, A.H.(2002). Six ways to look at linear interpolation, *International*

Journal of Mathematical Education in Science and Technology. 32, 6, 932-937. (file: c: \pedastat\ p31a)

2. Joarder, A.H.(2003). The halving method for sample quartiles. *International Journal of Mathematical Education in Science and Technology*. 34**(4), 629-633.** (file: c\ pedastat\ p96a)

3. Joarder, A.H. and Latif, R.M. (2004). A comparison and contrast of some methods for sample quartiles. *Journal of Probability and Statistical science. 2(1), 99-105.* (file: c\ pedastat\ p83a.doc)

4. Joarder, A.H. and Firozzaman, M. (12004). The Remainder Method for Sample Quantiles of Even Order (c: \pedastat \ p95a)

5. Joarder, A.H. (2002). On some representations of sample variance. *International Journal of Mathematics Education for Science and Technology,* 33 (5), 772-784**.** (file: c \ pedastat \ p30a .doc)

6. Joarder, A.H.(2003). The sample variance and first-order differences of observations. *Mathematical Scientist*. 28, 129-133. [London, UK] (c: \pedastat \ p34a.doc)

6. Laradji, A. and Joarder, A.H. (2004). Inequalities among Some Measures of Location (file: c \ pedastat \ q06a.doc)

7. Joarder, A.H. and Laradji, A. (2004). Algebraic Inequalities for Standard Deviation (file: c \ pedastat \ q06a.doc)

8**.** Joarder, A.H. and Al-Sabah, Walid S. (2002) The dependence structure of conditional probabilities in a contingency table. To appear in ***International*** Journal of Mathematics Education for Science and Technology. 33 (3), 475-480. (file: c\ pedastat \ p61a.doc)

9. Joarder, A. H. (2004). An Expository Note on Confidence Interval

10. Joarder, A.H. (2004). The logic of Testing of Hypotheses

11. Joarder, A.H. (2004). Tips on Linear Correlation and Regression

12. Joarder, A.H. (2004). Formulae of Statistics

RESEARCH PAPERS (*Descriptive Statistics: 10 Papers*)

- Joarder, A.H. and Firozzaman, M. (2001). A refinement over the usual formulae for deciles. ***International Journal of Mathematical Education in Science and Technology***, **32**(5), 761-765. [London, UK]

- Joarder, A.H. and Firozzaman, M. (2001). Quartiles for discrete data. ***Teaching Statistics***, 23 (3), 86-89. [London, UK]

- Joarder, A.H. and Mahmood, M. (1997). An inductive derivation of Stirling numbers of the second kind and their applications in statistics. ***Journal of Applied Mathematics and Decision Sciences***, **1,** 151--157. [**New Zealand**]

- Joarder, A.H. (1998a). On the statistical independence in a contingency table**.** ***International Journal of Mathematical Education in Science and Technology***, **29**, 780--782. [London, UK]