Standard Deviation for Small Samples[1]

Anwar H. Joarder and Raja M. Latif

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

Summary Neater representations for variance are given for small sample sizes especially for 3, 4 etc. With these representations, variance can be calculated without a calculator if sample sizes are small and observations are integers, and an upper bound for the standard deviation is immediate. Accessible proofs of lower and upper bounds are presented for broad spectrum of readers.

Key Words: Teaching; sample standard deviation; range; lower bound; upper bound.

1. INTRODUCTION

Throughout the paper we will assume that the observations in a sample of any size $n \geq 2$ are arranged in ascending order as $x_1 \leq x_2 \leq \cdots \leq x_n$ $(n \geq 2)$. Sample variance (s^2) is defined by $(n-1)s^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$ where $\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$ is the arithmetic mean of the sample observations. The sample standard deviation (s) is the positive square root of variance. In case the sample mean is not an integer, this formula for variance needs rounding off the mean which propagates error in the calculation. Prompted by this idea Joarder (2002) came up with many interesting representations of sample variance. In this note we provide neater representations for variance, implicit in Joarder (2003), for small sample sizes especially for 3, 4 etc. With these representations, variance can be calculated without a calculator especially if sample sizes are small and observations are integers.

Since the calculation of variance is difficult especially for beginners in a statistics course, bounds for variance may be useful in guarding against gross errors of calculation. Shiffler and Harsha (1980) have formulated an upper bound for the sample standard deviation (s) in terms of the sample range d, while Mcleod and Henderson (1984) have determined a lower bound for s in terms of d which also follows from Thomson (1955). Recently a better upper bound for standard deviation is derived by Croucher (2004) but for a sample of size 3. This is further discussed by Eisenhauer (2005) and Petocz (2005). The conjecture by Croucher (2004) that for a large sample, the standard deviation cannot exceed 60% of the range is argued to be true.

For broad spectrum of readers we present here the relevant representations of variance discussed by Joarder (2003). With these representations, the lower bound for standard deviation is also immediate. Accessible proofs of lower and upper bounds are presented for the same reason.

2. STANDARD DEVIATION FOR SMALL SAMPLES

(i) It is well known that for $n = 2$, the sample variance has a simpler form given by $s^2 = d^2/2$ where $d = x_2 - x_1$ is the sample range.

[1] Published in *Teaching Statistics*, **28**(2), 40-43.

(ii) The variance (s^2) for a sample of 3 observations can be written as

$$s^2 = \frac{1}{2}\left[x_1^2 + x_2^2 + x_3^2 - \frac{1}{3}\left(x_1 + x_2 + x_3\right)^2 \right].$$

Since $\left(x_1 + x_2 + x_3\right)^2 = x_1^2 + x_2^2 + x_3^2 + 2\left(x_1x_2 + x_1x_3 + x_2x_3\right)$, it follows that

$$3s^2 = x_1^2 + x_2^2 + x_3^2 - \left(x_1x_2 + x_1x_3 + x_2x_3\right). \tag{1}$$

Letting $x_2 - x_1 = d_1, x_3 - x_2 = d_2$, the first order differences of observations, we have

$$\begin{aligned} d_1^2 + d_2^2 + \left(d_1 + d_2\right)^2 &= (x_2 - x_1)^2 + (x_3 - x_2)^2 + (x_3 - x_1)^2 \\ &= 2\left[x_1^2 + x_2^2 + x_3^2 - \left(x_1x_2 + x_1x_3 + x_2x_3\right)\right]. \end{aligned} \tag{2}$$

Then from (1) and (2) we have the following equivalent forms

$$s^2 = \frac{1}{6}\left[d_1^2 + d_2^2 + (d_1 + d_2)^2 \right] = \frac{1}{3}\left(d_1^2 + d_2^2 + d_1d_2\right) = \frac{1}{3}\left(d_1^2 + dd_2\right) \tag{3}$$

where $d = d_1 + d_2$, the range of the 3 values in the sample.

Example 2.1 Consider calculating the variance of a sample 5, 6, 10. Since the first order differences are $6-5=1=d_1,\ 10-6=4=d_2$, and range $d=1+4=5$, by (3) we have

$s^2 = \left((1)^2 + d(4)\right)/3 = \left((1)^2 + 5(4)\right)/3 = 7$.

(iii) For a sample of size $n=4$, the variance can be represented by

$$s^2 = \frac{1}{4(4-1)}\left[d_1^2 + d_2^2 + d_3^2 + (d_1 + d_2)^2 + (d_2 + d_3)^2 + (d_1 + d_2 + d_3)^2 \right]. \tag{4}$$

where $d_1 = x_2 - x_1,\ d_2 = x_3 - x_2,\ d_3 = x_4 - x_3$. Defining $d_{12} = d_1 + d_2, d_{23} = d_2 + d_3$, the above expression can also be written as

$12s^2 = 2\left(d_1^2 + d_{12}d_2\right) + 2\left(d_2^2 + d_{23}d_3\right) + \left(d^2 - d_2^2\right)$.

Example 2.2 Consider calculating the variance of a sample 60, 65, 71, 80. Since the first order differences are $65-60=5,\ 71-65=6,\ 80-71=9$, and range $d=5+6+9=20$, we have

$$s^2 = \frac{1}{4(3)}\left[5^2 + 6^2 + 9^2 + (5+6)^2 + (6+9)^2 + (5+6+9)^2 \right] = 74.$$

Note that by subtracting an arbitrary number (say the smallest observation), the observations in the sample can be coded as 0, 5, 11, 20, and then (4) can be applied.

(iv) For a sample of size $n=5$, the variance can be represented by

$$s^2 = \frac{1}{5(5-1)}\Big[d_1^2 + d_2^2 + d_3^2 + d_4^2 + (d_1+d_2)^2 + (d_2+d_3)^2 + (d_3+d_4)^2$$
$$+(d_1+d_2+d_3)^2 + (d_2+d_3+d_4)^2 + d^2 \Big].$$

Example 2.3 Consider calculating the variance of a sample 72, 71, 80, 73, 85. If the observations are ordered as 71, 72, 73, 80, 85, the first ordered differences would be 1, 1, 7, 5 and range $d = 1+1+7+5 = 14$ so that

$$s^2 = \frac{1}{5(4)}\Big[(1)^2 + (1)^2 + (7)^2 + (5)^2 + (1+1)^2 + (1+7)^2 + (7+5)^2 + (1+1+7)^2 + (1+7+5)^2 + 14^2 \Big]$$
$$= 36.7.$$

(v) In general for a sample of any size $n \geq 2$, let $x_{i+1} - x_i = d_i (i = 1, 2, \cdots, n-1)$ be the first order differences of observations arranged in ascending order. It is obvious that neither any of the d_i value is negative, nor does it exceed $d = d_1 + d_2 + \cdots + d_{n-1}$. Then in general for a sample of any size $n \geq 2$, we have the following.

The quantity $n(n-1)s^2$ is the sum of squares of (i) $(n-1)$ d_i values, (ii) the totals of all $(n-2)$ consecutive pairs of d_i values, (iii) the totals of all $(n-3)$ consecutive triplets of d_i values, and so on. The last term in the addendum should be the square of the range $d = d_1 + d_2 + \cdots + d_{n-1}$ (cf. Joarder, 2003).

The method is not a feasible alternative for calculating variance for large samples. For large samples with integer valued observations, Joarder (2003) argued that the above formula for variance would be an efficient method if computer programs are used for calculation. Note that in case observations are not arranged in ascending order, some d_i values would be negative, and $d_1 + d_2 + \cdots + d_{n-1}$ would be different from the range.

3. BOUNDS FOR STANDARD DEVIATION

(i) It is well known that for a sample of 2 observations $s = d/\sqrt{2} \approx 0.707d$. Since $d_1^2 + d_2^2 \leq (d_1 + d_2)^2 = d^2$, iIt follows from (3) that for 3 observations, $6s^2 = \left(d_1^2 + d_2^2\right) + d^2 \leq d^2 + d^2$ i.e. $s \leq d/\sqrt{3} \approx 0.577d$ which is proved recently by Croucher (2004). Again, since $d_1^2 + d_2^2 + d_3^2 \leq (d_1 + d_2 + d_3)^2 = d^2$, it follows from (4) that for 4 observations, $12s^2 \leq d^2 + (d_1+d_2)^2 + (d_2+d_3)^2 + d^2 \leq d^2 + \left(2d^2 + d^2\right)$ i.e. $s \leq d/\sqrt{3} \approx 0.577d$. Since $d_1^2 + d_2^2 + d_3^2 + d_4^2 \leq (d_1 + d_2 + d_3 + d_4)^2 = d^2$, it follows from (5) that for 5 observations, $20s^2 \leq d^2 + \left(3d^2 + 2d^2 + d^2\right)$ i.e. $s \leq \sqrt{\frac{7}{20}}d \approx 0.592d$. In general, it follows from Section 2 (v) that

$$n(n-1)s^2 \leq d^2 + \Big[(n-2)d^2 + (n-3)d^2 + \cdots + 2d^2 + d^2 \Big] \leq d^2 + (n-1)(n-2)d^2/2$$

i.e. $s \le \sqrt{\dfrac{n^2-3n+4}{2(n^2-n)}}\, d = \sqrt{\dfrac{1}{2}-\dfrac{n-2}{n(n-1)}}\, d, \quad n \ge 2.$ (6)

Simple proofs for lower and upper bounds for standard deviation are in order.

(ii) Let $n=3$ ordered sample observations be denoted by $y_1 \le y_2 \le y_3$ and $d = y_3 - y_1$. Since $2\left(y_1^2 + y_3^2\right) \ge (y_3 - y_1)^2 = d^2$, it follows that $2\left(y_1^2 + y_2^2 + y_3^2\right) \ge 2\left(y_1^2 + y_3^2\right) \ge d^2$. Further, let $y_i = x_i - \bar{x}$ $(i = 1,2,3)$ and s^2 be the variance of 3 observations x_1, x_2 and x_3. Since the range of the y_i values is the same as that of the x_i values, and $2\left(y_1^2 + y_2^2 + y_3^2\right) = 2\left[(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2\right] = 2 \times 2s^2$, it follows that $4s^2 \ge d^2$ i.e. $s \ge 0.50\, d$. In general for a sample of any size $n \ge 2$, the lower bound for standard deviation given by

$$s \ge \frac{d}{\sqrt{2(n-1)}} \tag{7}$$

is derived by Mcleod and Henderson (1984) which also follows from Thompson (1955). Next consider $\left(x_1-\bar{x}\right)^2 + \left(x_2-\bar{x}\right)^2 + \left(x_3-\bar{x}\right)^2 \le \left(x_1-a\right)^2 + \left(x_2-a\right)^2 + \left(x_3-a\right)^2$. By choosing $a = x_1 + d/2$ we have $2s^2 \le (-d/2)^2 + (-d/2)^2 + (d/2)^2$, i.e. $s \le \sqrt{\frac{3}{8}}d \approx 0.612d$. In general, for a sample of any size $n \ge 2$, the upper bound for standard deviation is given by

$$s \le \frac{d}{2}\sqrt{\frac{n}{n-1}}. \tag{8}$$

(Shiffler and Harsha, 1980). Thus for a sample of size 3, the above bounds simplify approximately to $\frac{1}{2}d \le s \le \frac{1}{2}\sqrt{\frac{3}{2}}d \approx 0.612d$ while the upper bound in (6) is $d/\sqrt{3} \approx 0.577d$. For any sample of size 4, the upper bound in (6) is the same as that offered by Shifler and Harsha (1980). However as the sample size becomes larger than 4, the upper bound in (8) keeps uniformly dominating that in (6). The upper bound in (8) approaches to $0.5d$ whereas that in (6) approaches to $\sqrt{\frac{1}{2}} \approx 0.707d$ as n becomes very large.

iii) The Croucher Conjecture

The way Croucher (2004) improved the above upper bound, given in (8), for a sample of 3 values is described below with the help of (3). Since variance does not depend on the location, he considered the observations as $0, a, ka$ where $k \ge 1$. With notations in Section 2, $d_1 = a$, $d_2 = ka - a$ and $d = ka$, it follows from (3) that $3s^2 = a^2 + ka(ka-a) = a^2(k^2-k+1) = \left(d/k\right)^2(k^2-k+1)$. That is $s^2 = \dfrac{d^2}{3}\left(1 - \dfrac{1}{k} + \dfrac{1}{k^2}\right)$. Then given d, the variance for a sample of size 3 has a maximum value of $d^2/3$ occurring at $k=1$ i.e.

$$s \le d/\sqrt{3} \approx 0.577d \,. \qquad (9)$$

In view of (8), the conjecture by Croucher (2004) that for large data sets s cannot exceed 60% of the range, is just true. In what follows we argue that the upper bound of standard deviation given in (9) is true for any sample of size $n \ge 3$.

Example 3.1 To find the variance of the sample (60, 65, 71, 80**),** draw a sample without replacement of size 2 from the given sample. The $n(n-1)/2! = 4(4-1)/2 = 6$ samples (60, 65), (65, 71), (71, 80), (60, 71), (65, 80), (60, 80) have the ranges $d_1 = 5, d_2 = 6, d_3 = 9, d_4 = 11, d_5 = 15$ neither of which exceed $d = 20$, the range of the given sample. Then the variance of the given sample is

$$s^2 = \frac{1}{6}\left(s_1^2 + s_2^2 + \cdots + s_6^2\right) = \frac{1}{6}\left(\frac{5^2}{2} + \frac{6^2}{2} + \frac{9^2}{2} + \frac{11^2}{2} + \frac{15^2}{2} + \frac{20^2}{2}\right) = 74$$

where $s_i^2 (i = 1, 2, \cdots, 6)$ is the variance of the ith sample. Observe that none of the 6 sample variances exceeds $20^2/3$.

Example 3.2 To find the variance of the sample (60, 65, 71, 80**),** draw a sample without replacement of size 3 from the given sample**.** The $n(n-1)(n-2)/3! = 4(4-1)(4-2)/6 = 4$ samples are (60, 65, 71), (60, 65, 80), (60, 71, 80), (65, 71, 80) with variances $30\frac{1}{3}, 108\frac{1}{3}, 100\frac{1}{3}$ and 57 respectively. Then the variance of the given sample is $s^2 = \left(s_1^2 + s_2^2 + s_3^2 + s_4^2\right)/4 = \left(30\frac{1}{3} + 108\frac{1}{3} + 100\frac{1}{3} + 57\right)/4 = 74$. Observe that none of the 4 sample variances exceeds $20^2/3$.

In general if 3 observations are drawn without replacement from a given sample of size $n \ge 3$, there would be $m = n(n-1)(n-2)/3!$ samples with ranges $d_1, d_2, \cdots, d_m$ and variances $s_1^2, s_2^2, \cdots, s_m^2$ respectively. It is obvious that neither any of the $d_i (i = 1, 2, \cdots, m)$ value is negative, nor does it exceed d, the range of the given sample. It then follows from the above considerations (or see Cochran, 1977) that $s^2 = \left(s_1^2 + s_2^2 + \cdots + s_m^2\right)/m$, and then by (6) or (9) we have

$$s^2 \le \frac{1}{m}\left(\frac{d_1^2}{3} + \frac{d_2^2}{3} + \cdots + \frac{d_m^2}{3}\right) \le \frac{d^2}{3}\,.$$

That is $s \le d/\sqrt{3} \approx 0.577d$. Thus the conjecture by Croucher (2004) is proved to be true for a sample of any size $n \ge 3$.

4. BETTER BOUNDS FOR STANDADR DEVIATION

For a sample of any size $n \ge 2$, Joarder and Laradji (2004) improved the lower bound of standard deviation given by Mcleod and Henderson (1984) in terms of range and the difference between mean $(\bar{x})$ and median $(\tilde{x})$ to account for the skewness in the sample. They also improved the upper bound of standard deviation given by Shiffler and Harsha (1980) or Croucher (2004) in terms of range and some other statistics. In

passing we mention that the lower bound for standard deviation is
$s \geq \sqrt{d^2/(2n-2) + (\tilde{x} - \bar{x})^2/2}$.
Clearly the improvement considered in Joarder and Laradji (2004) for the bounds of standard deviation by incorporating mean, median and other sample statistics looses the simplicity offered in (9) by Croucher (2004) or others.

Acknowledgement

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

References

Cochran, W.G. (1977) *Sampling Techniques*. Wiley, New York, USA.

Croucher, J.S. (2004). An upper bound on the value of the standard deviation. *Teaching Statistics*, 26(2), 54-55.

Eisenhauer, J.G. (2005). Letter to the Editor. Teaching Statistics, 27(1), 15.

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.

Joarder, A.H. and Laradji, A. (2004). Some inequalities in descriptive statistics. Technical Report No. 321, Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Saudi Arabia.

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

Petocz, P. (2005). An upper bound on standard deviation as a function of range. Teaching Statistics, 27(2), 42-4.

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\sforsmalla.doc

Get these papers:

Petocz, Peter (2005). An upper bound on standard deviation as a function of range, Teaching Statistics, Summer 2005, 27(2), 42-44.

Eisenhauer, J.G. (2005). Letter to the Editor, 27(1), 15.

[[Problem for Dr Munir Mahmood

Refer to the end of Example 3.2: ([urther Better Bound along Croucher (2004)]

In general if k observations are drawn without replacement from a given sample of size $n \geq 3$, there would be $m = n(n-1)\cdots(n-k)/k!$ samples with ranges $d_1, d_2, \cdots, d_m$ and variances $s_1^2, s_2^2, \cdots, s_m^2$ respectively. It is obvious that neither any of the $d_i (i = 1, 2, \cdots, m)$ value is negative, nor does it exceed d, the range of the given sample. If $s_i^2 \leq b_i$, $(i = 1, 2, \cdots, m)$ it then follows (see Cochran, 1977) that

$s^2 = \left(s_1^2 + s_2^2 + \cdots + s_m^2\right)/m$ and

$s^2 \leq \left(b_1 + b_2 + \cdots + b_m\right)/m$.

Research Directions: Assume $k = 4$ and find b_i, $(i = 1, 2, \cdots, m)$**.]]**