

# Notes on

## TWO SIMPLE STATISTICAL CALCULATIONS

Derek O'Connor

Started : February 14, 2008,

Latest : February 23, 2009

### 1 Introduction

This note is prompted by reports of errors in the calculation of two simple statistics: the mean of a vector  $x$ , or  $\bar{x} = \frac{1}{n} \sum_i x_i$ , and its variance  $\text{Var}(x) = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ .

Errors in such simple calculations usually indicate that a problem is ill-conditioned, or an unstable algorithm is being used, or a combination of both.

We illustrate these ideas by analyzing two problems where these errors arise: the Sea Surface Heights mean problem which is ill-conditioned, and the Microsoft Excel variance problem where an unstable algorithm is used.

## 1.1 The *Sea Surface Heights* Problem

The following information was taken from a paper by Yun He & Chris Ding of the NERSC-Lawrence Berkeley Labs who were doing a large-scale simulation of ocean circulation.

At each step of the simulation the following is done

1. *Sea Surface Heights* are calculated at each point on a  $64 \times 120$  latitude-longitude grid.
2. The average of these  $64 \times 120 = 7680$  numbers is then calculated.
3. This average is compared with satellite data.

Table 1 shows the results that He & Ding got with the Fortran code shown in section ??, on a single processor, using IEEE double precision (~16 decimal digits). He & Ding point out that these results are completely wrong — not one digit is correct.

Table 1: He and Ding's Results using 16-digit Precision

Sum Order	Value	Rel. Error
Longitude First	34.4147682189941410	95.1
Latitude First	0.67326545715332031	0.88
Reverse Longitude First	32.302734375	89.2
Reverse Latitude First	0.734375	1.05
Correct Value	0.35798583924770355224609375	

We will analyse this problem in Section ?? and explain why He & Ding got such inaccurate results.

Here is the original paper : <http://www.derekroconnor.net/NA/Notes/HeDing--OceanHeights.pdf>

Here is the data : <http://www.derekroconnor.net/NA/Notes/etaana.dat>. Use this data to check your favourite software. Does it give  $\bar{x} = 0.35798583924770355224609375$ ?

## 1.2 The Microsoft Excel Problem

Microsoft's Excel spreadsheet has been in use for many years and has gone through many versions. Many millions in business, government, and universities, use some version of Excel.

Table 2 shows the results of Excel 2000's calculations on the following data :

Column 2 =  $\{x_1, x_2, \dots, x_i, \dots, x_{10}\} = \{1, 2, 1, 2, \dots, 1, 2\}$ .

Columns 3 to 6 are derived from this:  $x + M$ , with  $M = 10^8, 10^{10}, 10^{14}, 10^{15}$ .

The exact values of the mean and variance are

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \sum_{i=1}^{10} (M + x_i) = 10M + 15 \frac{10M + 15}{10} = M + 1.5 \\ \text{Var}(x) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{10-1} \sum_{i=1}^{10} (M + x_i - M - 1.5)^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - 1.5)^2 \\ &= \frac{1}{9} \sum_{i=1}^{10} (\pm 0.5)^2 = \frac{1}{9} 2.5 = 0.27\dot{7} \dots,\end{aligned}$$

$$\text{SDev}(x) = \sqrt{0.27\dot{7} \dots} = 0.5270462766947299 \text{ rounded to 16 digits and does not involve } M.$$

The last line of the Table 2 contains Excel 2000's values for the standard deviation. None of these values is correct. We will analyse this problem in Section ?? and explain why Excel 2000 and later versions get such inaccurate results.

Table 2: Excel 2000 Results

Row $i$	$x_i$	$x_i + 10^8$	$x_i + 10^{10}$	$x_i + 10^{14}$	$x_i + 10^{15}$
1	1	10000001	1000000001	100000000000001	1000000000000001
2	2	10000002	1000000002	100000000000002	1000000000000002
3	1	10000001	1000000001	100000000000001	1000000000000001
4	2	10000002	1000000002	100000000000002	1000000000000002
5	1	10000001	1000000001	100000000000001	1000000000000001
6	2	10000002	1000000002	100000000000002	1000000000000002
7	1	10000001	1000000001	100000000000001	1000000000000001
8	2	10000002	1000000002	100000000000002	1000000000000002
9	1	10000001	1000000001	100000000000001	1000000000000001
10	2	10000002	1000000002	100000000000002	1000000000000002
Sum	15	100000015.0	10000000015.0	1000000000000015.0	10000000000000016.0
Mean	3/2	10000001.5	1000000001.5	100000000000001.5	1000000000000001.6
Sdev	$\sqrt{0.2\dot{7}}$	0.54...20E+00	0.00...00E+00	1.39...30E+06	0.00...00E+00

The values in the last row should be 0.5270462766947299