Albert Wiryawan
NETID: avw2
UIN: 673431511
PHYS 466: Atomic Scale Simulation
Professor David Ceperly

Homework #1

Data Set #1

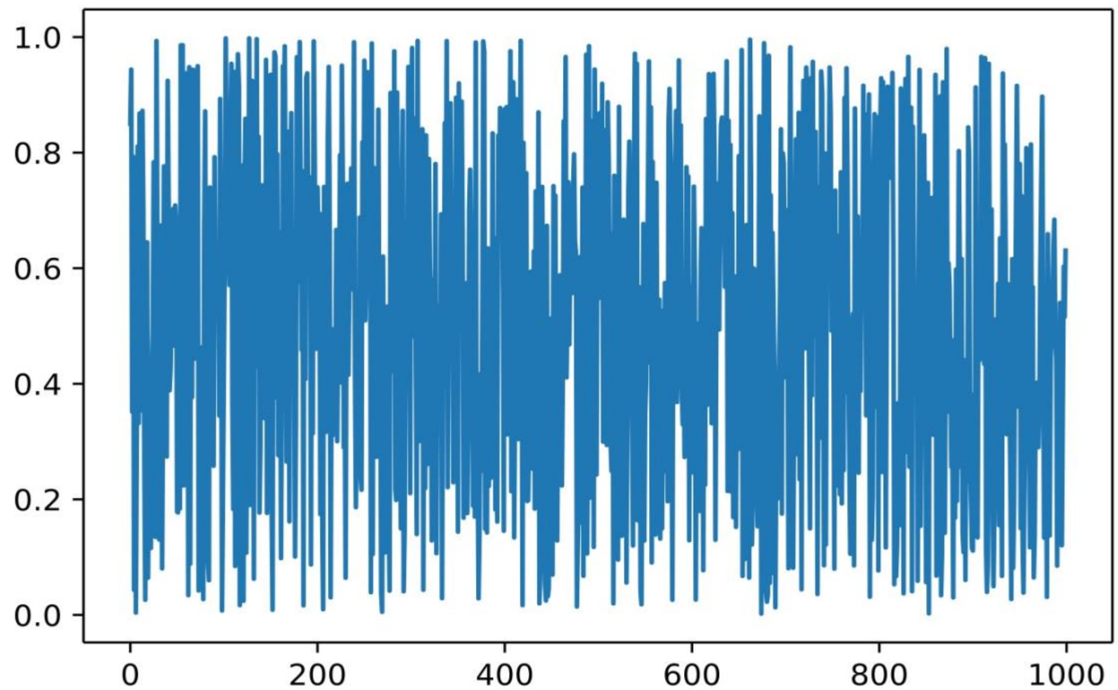

Figure 1. Plot of First Dataset: Time series data of points

By eye, its is hard to estimate an exact value of mean and standard deviation due to the fluctua- ions of the plotted data. However, based on the graph created, I would have to say that mean is close to 0.5 as the distrubtion seems to have a relatively similar number of maximums and mini mums with a lot of points generally around the value of 0.5. The standard deviation is also hard to estimate by eye. Using the central limit theorem (since we have a large data set) and what w e know about normality, the addition and subtraction of the standard deviation will encapsulat e about 68% of the data of the entire data set. As a result, I estimate the standard deviation to be around 0.23.

Mean of data = 0.5008547564015497
Standard Deviation of data = 0.28294102378178476

Comparing the computed standard deviation and mean produced with those that were estimated through eyeballing. The numbers are actually relatively similar to what was calculated indicating that my reasoning to the eyeball estimates based off the graph were relatively effective.

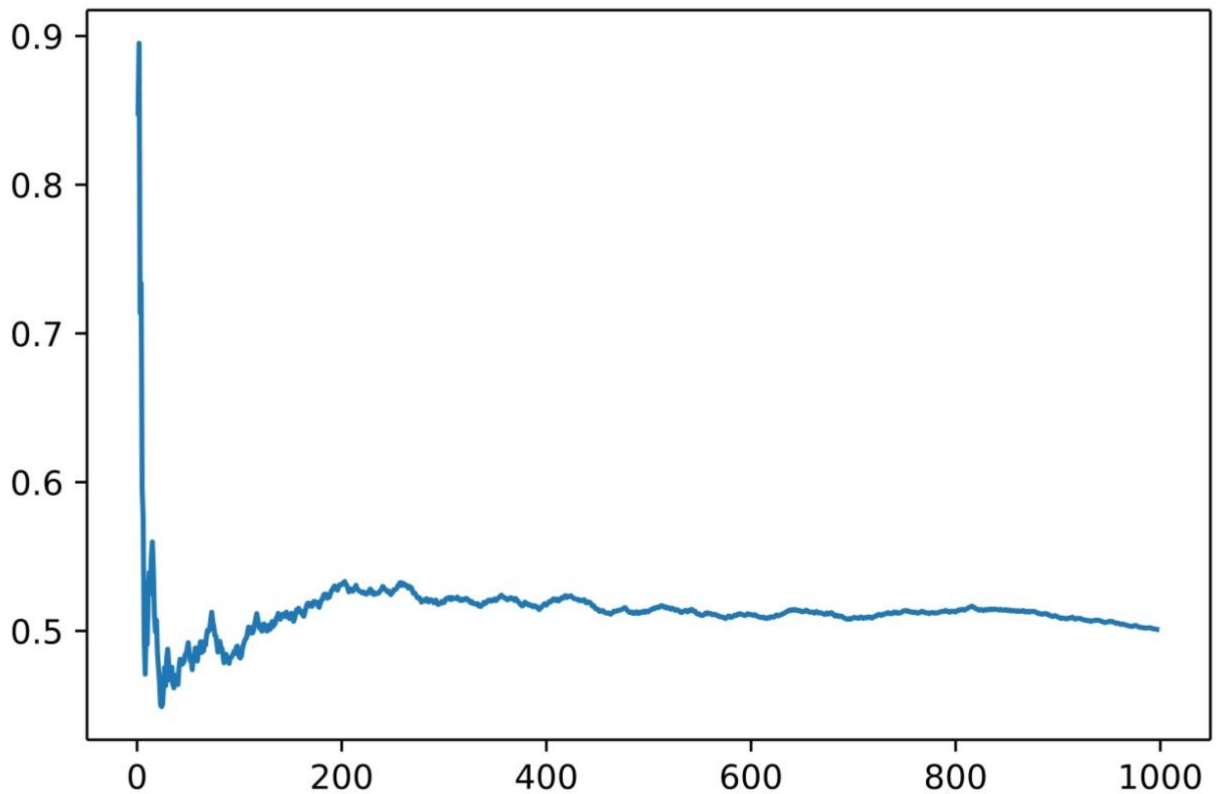

Figure 2. Plot of mean of data set vs cutoff of the number of points

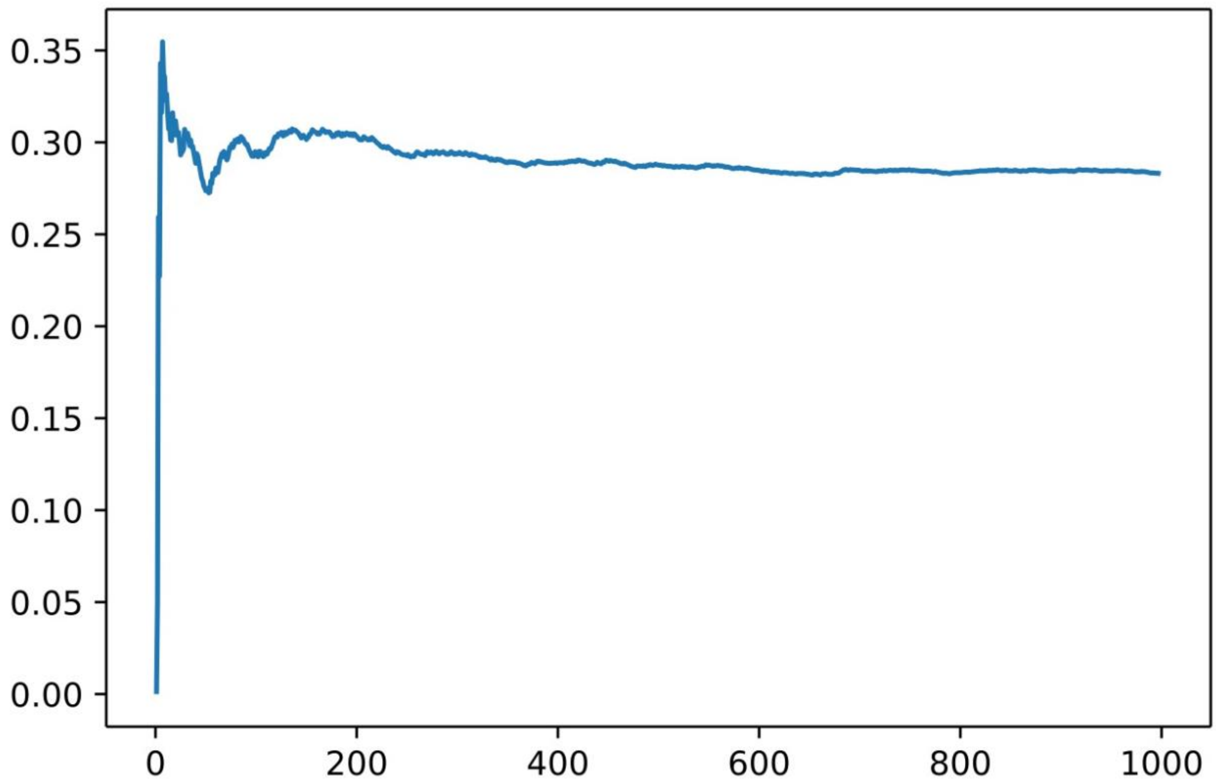

Figure 3. Plot of standard deviation of data set vs the cutoff number of data points used

Based on the graphs, the first 100 points will give a mean of approximately 0.48 and standard d-eviation of 0.27. As the cutoff of the number of points decrease it becomes apparent that the mean and standard deviation of the data set begins to approach somewhat of a limit with more and more points of data.

Data set #2

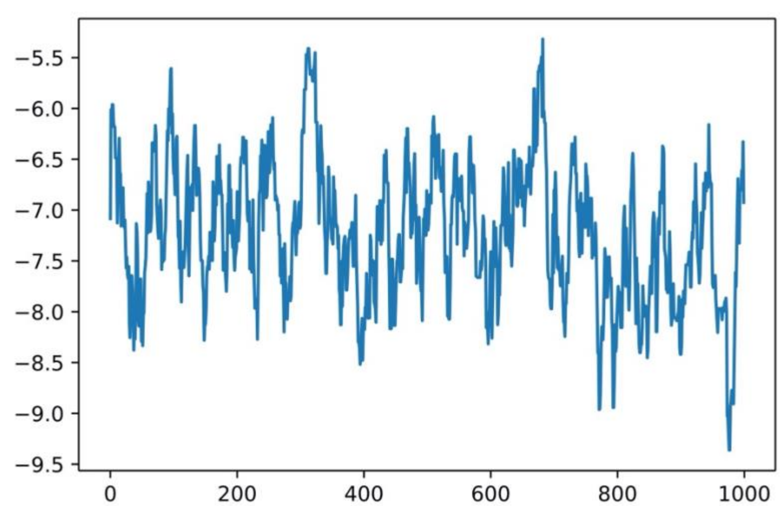

Figure 4. Plot of Time series data given by data set #2

Based on the graph above, the estimated correlation was 14 and the calculated value was 17.530834318271747 which indicates a percent error of … between the values.

Accounted correlation mean error = 0.021359134950472643
Unaccounted correlation mean error = 0.08943035451115007

Based on the values calculated above the mean error is larger when the correlation is not accounted

<u>Data set #3</u>

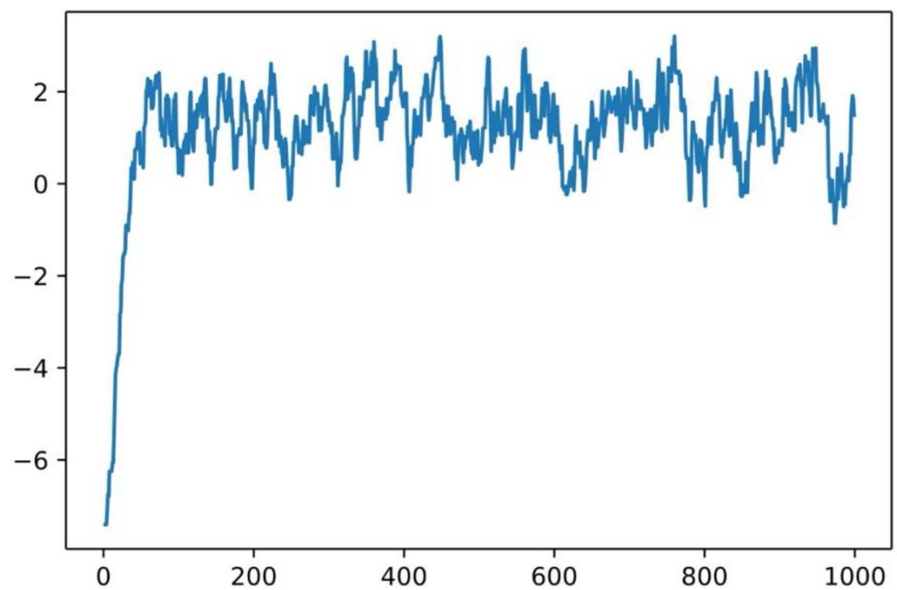

Figure 5. Plot of Time series data given by data set 3

Without Cutoff
Mean = 1.1445021326705114
Standard Error of the mean = 0.22031785960176142

Based on the graph, the initial cutoff should be around 60 points in order to get rid of outliers.

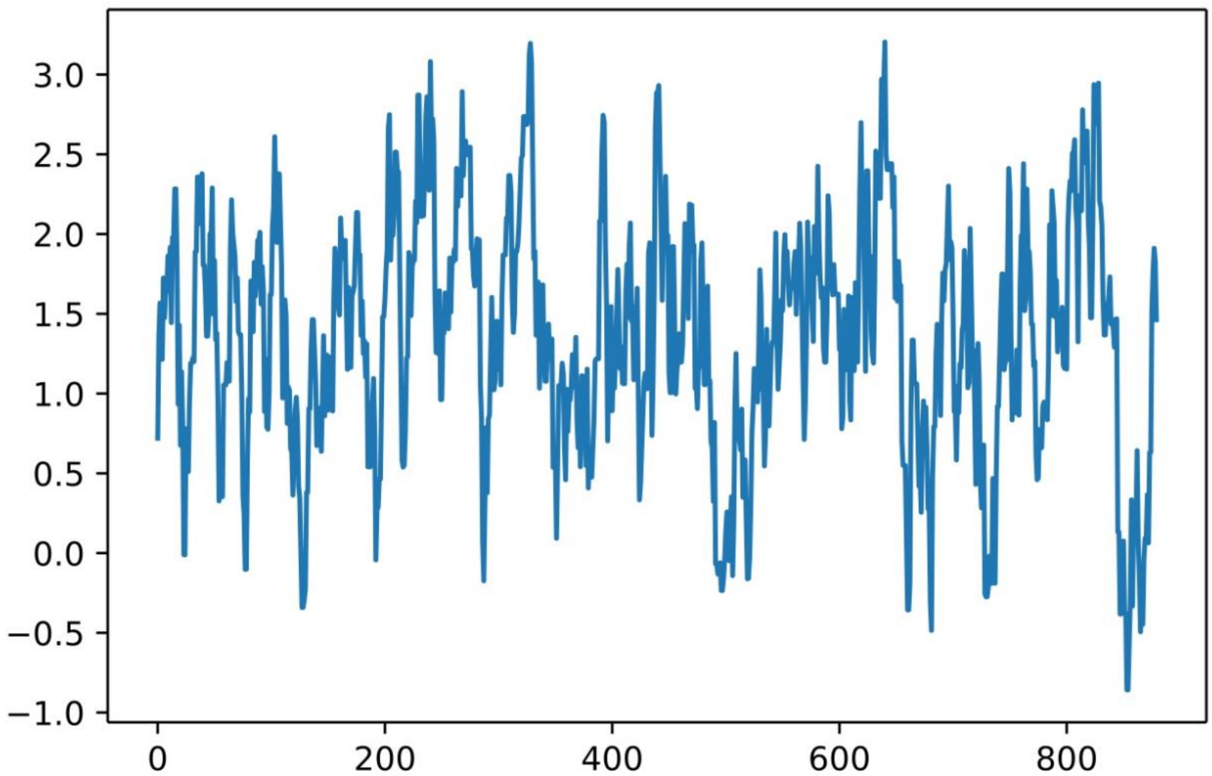

Figure 6. Plot of Revised data

By cutting off the initial outlier from the data the new mean and mean error was found to be 1.3545941304427673 and 0.09766396407896263 for mean and mean error respectively.

Data set #4

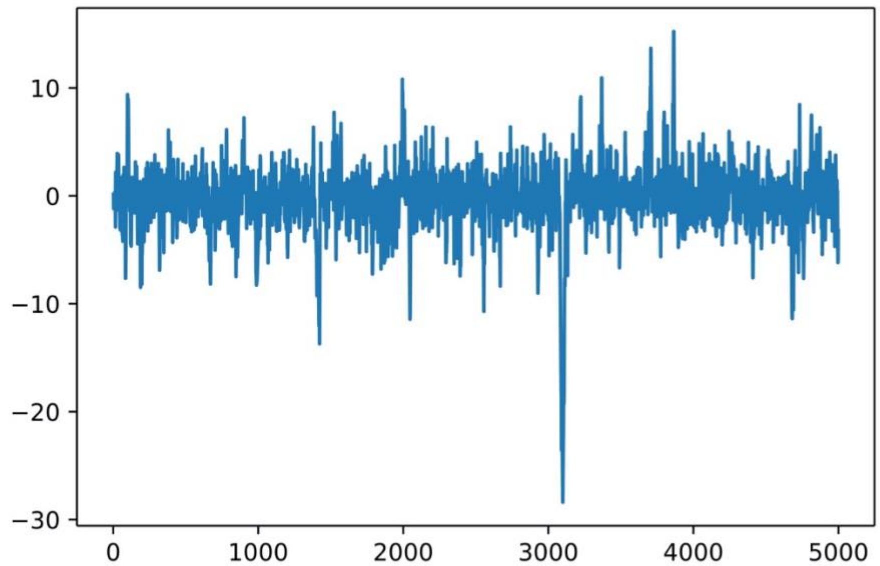

Figure 7. Plot of Times Series data given by data set 4

Based on the graph above the and the analytical expression shown, I expect the mean to reside around -5. This should also be influenced by the large negative outliers. At large x, values the variance is expected to decrease considering the equation given should converge.

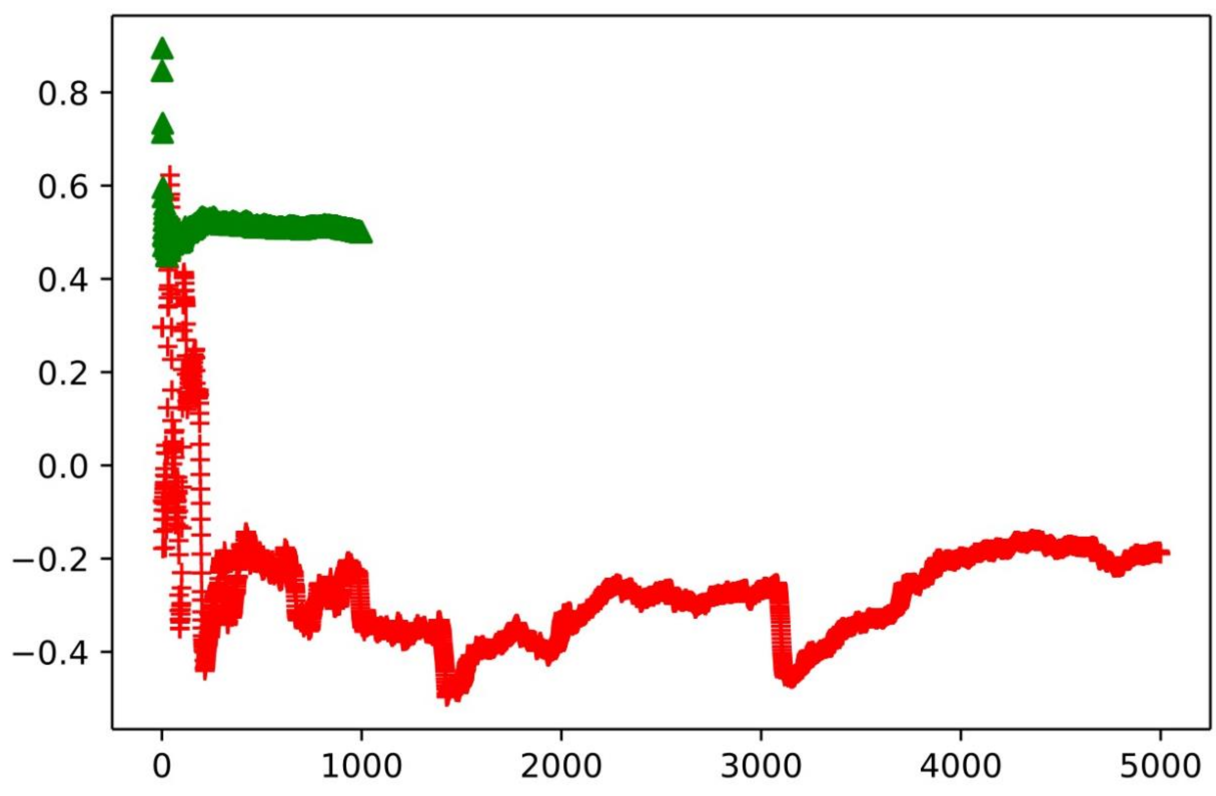

Figure 8. Mean vs cutoff for dataset #1(green triangles) and dataset #2(red plus)

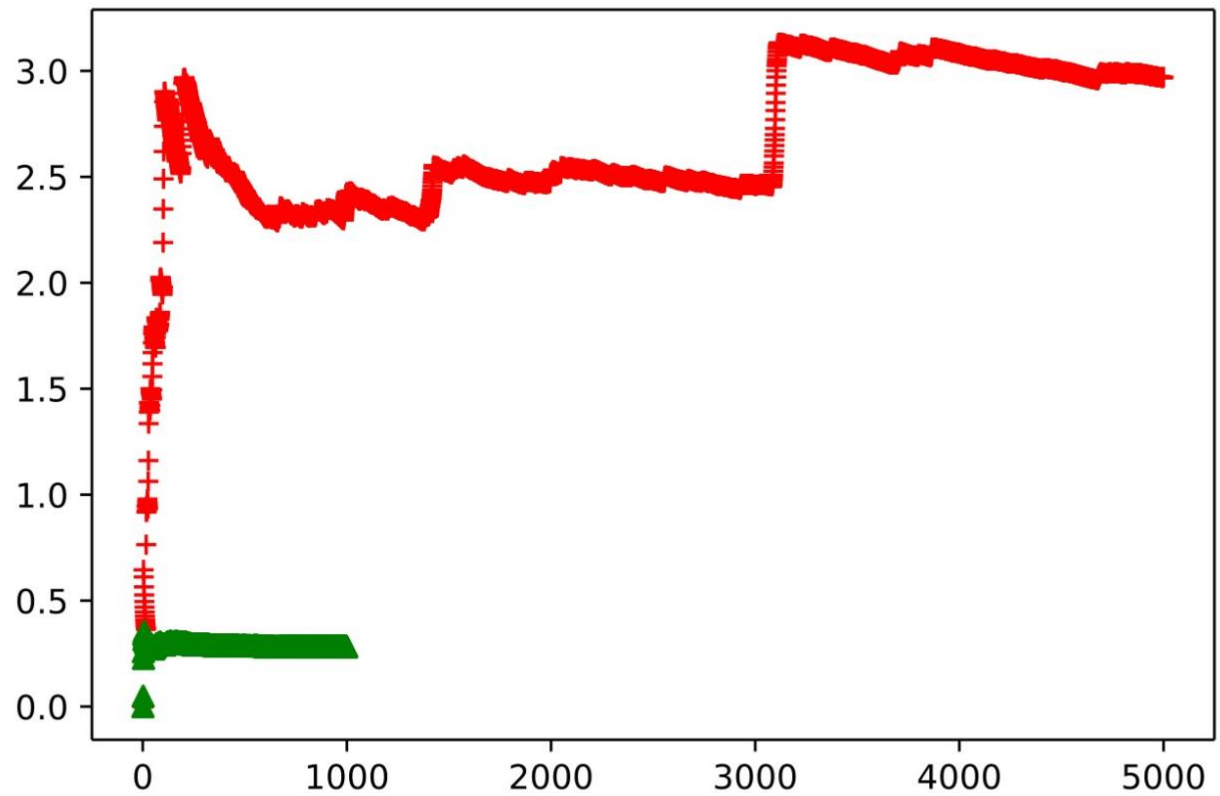

Figure 9. Standard Deviation vs Cut off for data set #1(green triangles) data set #2 (red plus)

By comparing the plots for the two data sets, its looks as though the values for mean and standard deviation roughly converge for all plots. However, we can see how the large negative peak that is created at around x = 3000 in data set 4 effects largely effect the mean and standard deviations.

<u>Part #2: Central Limit Theorem</u>

A Version:
Mean: 1.2533333333333334
Variance: 0.024866666666666666
Standard error:0.06437735971942656

B Version
Mean:1.3516666666666666

Variance:0.027256666666666676

Standard error:0.06740013188249545

```
#estimate of the difference
C = abs(statistics.mean(B) - statistics.mean(A))
print(C)
#estimate of the error of the difference
D = math.sqrt((my_stderr(A)**2) + (my_stderr(B)**2))
print(D)
print(C/D)
#The t-statistic ---> 1.0550190358247307
#Normal Probability is 71% meaning tail ends are the wanted 29%
```

Performing the calculations stated by the homework provides us with a t statistic of 1.0550190358247307 this can then be used to find the probability distribution from the normal curve which is 71%. However, due to the nature of this question, we are interested in the tails and thus 1-0.71 = 0.29 or 29% of data coming from the same population.