Experiment 1
Error Measurements: σ and σ_s

Jamie Lee Somers,
B.Sc in Applied Physics.

Thursday 8$^{\text{th}}$ October, 2020

1 Introduction/Method:

The purpose of a numerical experiment like this one is to show visually, as well as mathematically the affects that increasing your sample can have on your overall accuracy due to how much more varied the data is, and how having more data can reveal irregularities that occur as well as giving a better sense of the actual measurement. The experiment involves working with four data sets, each set has different values. The first data set has 10 values, the second data set has 20 values, the third data set has 50 values and the final data set has 100 values. Our goal is to carry out the same analysis on all four of these data sets which includes making them into a histogram, finding their mean, finding their population standard deviation and sample standard deviation.

In the next part of the experiment we create 8 new sets by plucking 10 values from the original datasets 8 times. We calculate the mean and standard deviation of these newly made sets, as well as taking the mean and standard deviation of their mean values.

What we hope to gain from this entire experiment is a better understanding of what each of these data analysis means, as well as seeing the impact that increasing the number of measurements has on each of these analysis. The affects should be visually striking as well as mathematically apparent.

2 Results:

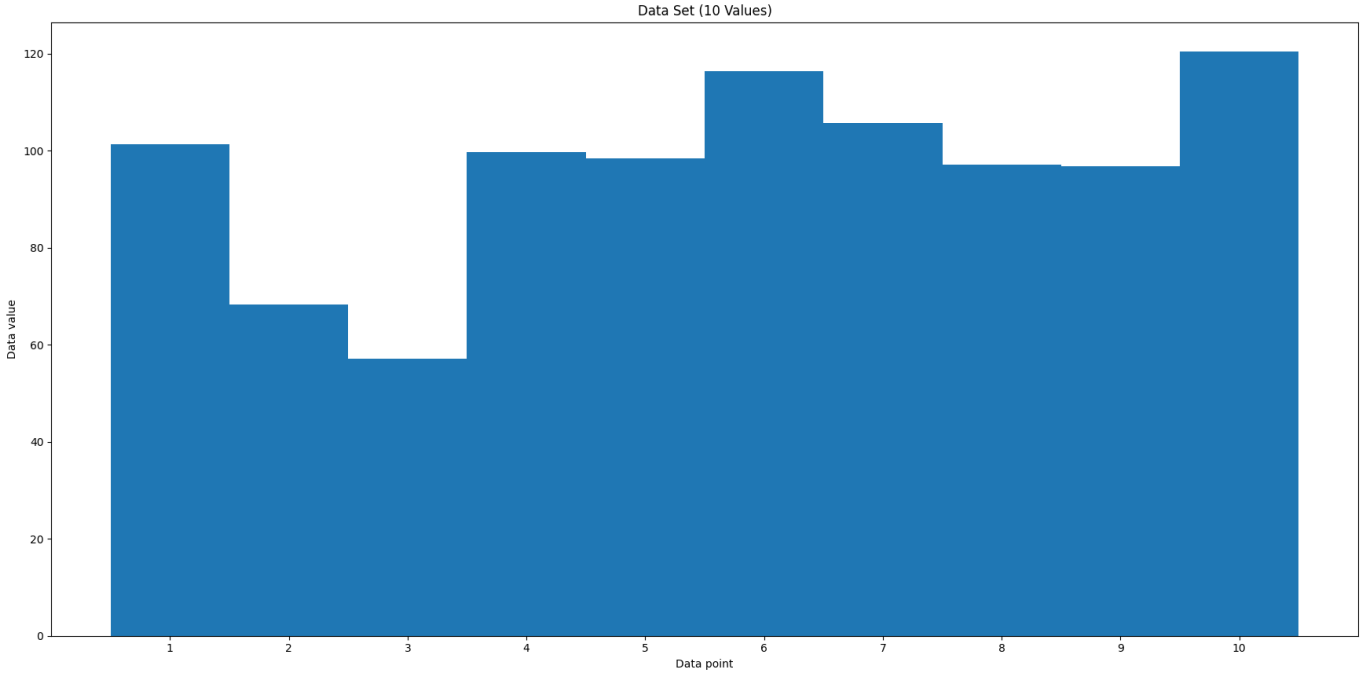

Figure 2.1: Data set using 10 points of data

This dataset shows a graph where the majority of the values are quite similar, it starts off around 100 before dipping and then jumps back up to around 100 for the rest of the data set. It has a peak in the middle and at the end.

Mean: [96.135] The mean value of this data set is ≈ 96, which appears to be heavily motivated by the two numerical values in point 2 and 3 which are far lower than the rest.

Population Variance σ^2: [343.7969719]

Sample Variance σ_s^2: [381.9966355]

Population Standard Deviation σ: [18.54176291] The error for this data set is ≈ 19, which may appear like a high number but is directly in line a majority of our errors.

Sample Standard Deviation σ_s: [19.54473321] Seeing as our sample size here is 10, this is probably the more accurate measurement of our error, however this value is even higher than the population standard deviation at ≈ 20.

σ_m: [6.180587638] The error in the set itself is quite high, however the standard error in the mean is quite low at ≈ 6. We can still improve on this number by increasing the number of values in our data set.

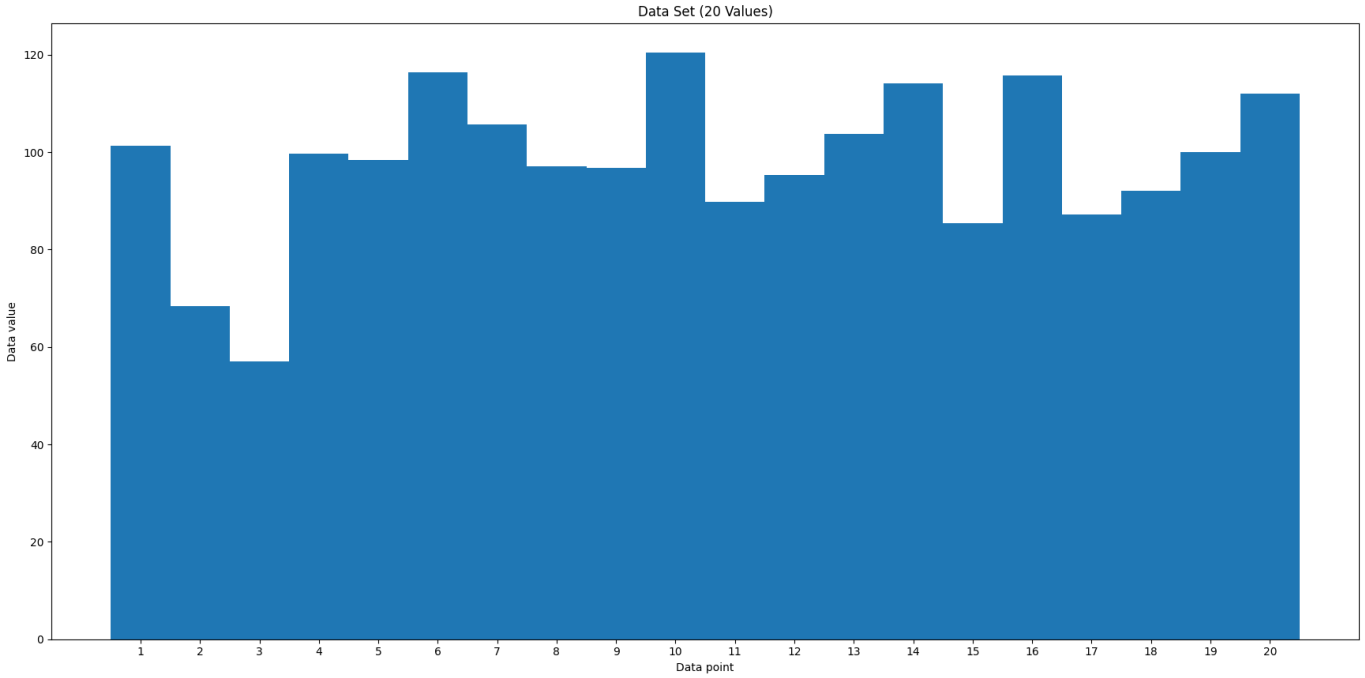

Figure 2.2: Data set using 20 points of data

This dataset shows a graph which is even more varied in terms of values, it still starts off at around 100 but we can now tell that it is slightly above, there is still a dip in values near the beginning of the data set however the peaks now appear to be more spread out as apposed to just being in the middle and at the end, there is now roughly one peak in the middle and 4 smaller peaks 1 to the left and 3 to the right.

Mean: [97.83745] Given how many more peaks are present in this graph, the expected outcome would be that the mean value has risen quite dramatically over the 10 value set. This isn't the case however as the mean value has moved up by less than 2. Even though most of the histogram appears to be at or above the 100 mark, it is clear that the two lower values are still having a big impact on the mean value.

Population Variance σ^2: [233.3617]

Sample Variance σ_s^2: [245.6439]

Population Standard Deviation σ: [15.27618081] While analysing the results I found this error to be quite irregular, my expectation for this experiment was that the standard deviations were going to remain largely the same with only minor variances between them. However this value is off by about 3 when compared to the three other values calculated.

Sample Standard Deviation σ_s: [15.67303098] As this calculation is determined by using the standard deviation shown above as one of the variables, it is to be expected that this value turned out to be equally as irregular and off by about as much.

σ_m: [3.504596231] Back to expected results, this result carries on what we would expect to see with an increase in number of values in our data set. As we increased the number of values in our data set from 10 to 20 we noticed the mean error reduce from 6.2 to 3.5. Its important to notice however that this is not a direct correlation, as we doubled our number of values we seen $\approx 43\%$ drop in the mean error. There are clearly still other errors contributing.

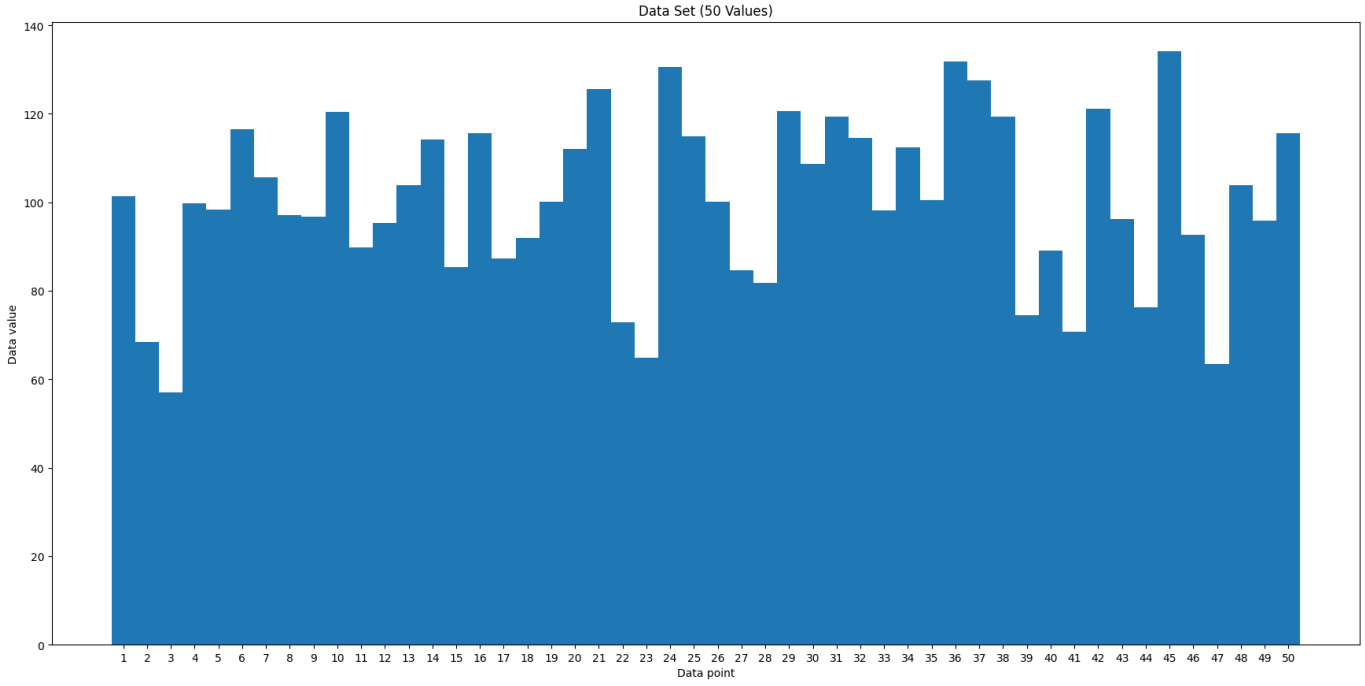

Figure 2.3: Data set using 50 points of data

It isn't until this dataset that we can see that there is actually a dip in values at the beginning, in the middle and also near the end of the data set. There now appears to be three peaks in the graph but they now appear in the middle, with two more closer to the end. By now we can see that the graph no longer shows the data as being quite similar, each of the values are actually quite distinct and the graph no longer seems to skew in any one direction.

Mean: [100.3686] At 50 values in our data set we finally start to see the number of peaks negate some of the affects that the large dips were having earlier on, as we finally see the mean value of 100 we were expecting to see come to fruition.

Population Variance σ^2: [359.6853]

Sample Variance σ_s^2: [367.0258]

Population Standard Deviation σ: [18.96537107] Once again our standard deviation returns to a similar value as we were seeing in our first graph, this gives credence to the idea that perhaps it was just a random sampling issue that caused the error in our 20 dataset to be so far off. Even with 50 values our standard deviation returns to within 0.4 of the first standard deviation we calculated.

Sample Standard Deviation σ_s: [19.15791742] Since the sample standard deviation calculation requires the use of the population standard deviation, this value has too returned to within range of the first one recorded. Now that our set contains 50 values it is possible to say that the sample standard deviation is no longer the most accurate and n may be too large for the equation to be accurate.

σ_m: [2.709338724] We continue to see a sizeable decline in the mean error as we continue to increase the size of our sets. This time we increased our set size from 20 to 50 and got our mean error down from 3.5 to 2.7.

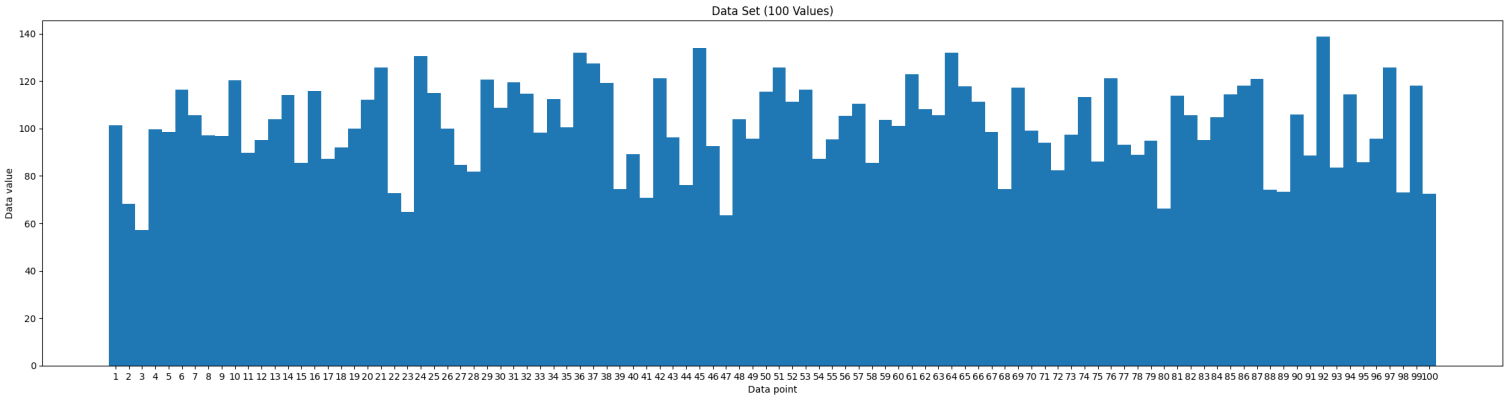

Figure 2.4: Data set using 100 points of data

What was made clear in the graph with 50 data points has now been completely fleshed out in what would be considered the most accurate representation of the data, the number of largest dips has increased to 5, while the number of peaks has also increased to 5. Values very rarely if ever equal each other which is something smaller data sets failed to show us

Mean: [101.0762] Once again as we increase our number of data points our mean appears to rise, this time there are many more peak values which contributes to the mean value increasing to ≈ 101.

Population Variance σ^2: [324.4987]

Sample Variance σ_s^2: [327.7765]

Population Standard Deviation σ: [18.01384745] Our error / standard deviation has stayed in the eighteen range, however this value is lower than the two previous errors in the eighteen range.

Sample Standard Deviation σ_s: [18.10459886] The sample standard deviation is inline with the two previous ones that we believe to be accurate, however as mentioned in the 50 data set, its likely that this value isn't reliable due to n being so high.

σ_m: [1.810459789] Finally, we see our lowest mean error yet as the number of data points is the highest it has ever been. This is a consistent relationship we have seen throughout all four mean errors.

Part 2:

Data Set 1

1	2	3	4	5	6	7	8	9	10
97.0893	96.7681	120.47	89.7525	95.2807	103.804	114.145	85.3891	115.673	87.2352

Mean:

$$\frac{97.0893+96.7681+120.47+89.7525+95.2807+103.804+114.145+85.3891+115.673+87.2352}{10} = 100.56069$$

Variance / Standard Deviation:

$$\frac{(97.0893-100.56069)^2}{10} + \frac{(96.7681-100.56069)^2}{10} + \frac{(120.47-100.56069)^2}{10} + \frac{(89.7525-100.56069)^2}{10} + \frac{(95.2807-100.56069)^2}{10} + \frac{(103.804-100.56069)^2}{10} + \frac{(114.145-100.56069)^2}{10} + \frac{(85.3891-100.56069)^2}{10} + \frac{(115.673-100.56069)^2}{10} + \frac{(87.2352-100.56069)^2}{10} = 139.8690456$$

$$\mu = 100.56069 \quad \sigma = \sqrt{139.8690456} = 11.82662444$$

Data Set 2

1	2	3	4	5	6	7	8	9	10
86.0732	121.064	93.2652	88.8322	94.9117	66.3984	113.864	105.708	95.0351	104.807

Mean:

$$\frac{86.0732+121.064+93.2652+88.8322+94.9117+66.3984+113.864+105.708+95.0351+104.807}{10} = 96.99588$$

Variance / Standard Deviation:

$$\frac{(86.0732-96.99588)^2}{10} + \frac{(121.064-96.99588)^2}{10} + \frac{(93.2652-96.99588)^2}{10} + \frac{(88.8322-96.99588)^2}{10} + \frac{(94.9117-96.99588)^2}{10} + \frac{(66.3984-96.99588)^2}{10} + \frac{(113.864-96.99588)^2}{10} + \frac{(105.708-96.99588)^2}{10} + \frac{(95.0351-96.99588)^2}{10} + \frac{(104.807-96.99588)^2}{10} = 214.4985789$$

$$\mu = 96.99588 \quad \sigma = \sqrt{214.4985789} = 14.64577$$

Data Set 3

1	2	3	4	5	6	7	8	9	10
96.1787	76.2106	134.07	92.6653	63.3877	103.883	95.8431	115.605	125.794	111.393

Mean:

$$\frac{96.1787+76.2106+134.07+92.6653+63.3877+103.8834+95.8431+115.605+125.794+111.393}{10} = 101.503$$

Variance / Standard Deviation:

$$\frac{(96.1787-101.503)^2}{10}+\frac{(76.2106-101.503)^2}{10}+\frac{(134.07-101.503)^2}{10}+\frac{(92.6653-101.503)^2}{10}+\frac{(63.3877-101.503)^2}{10}+\frac{(103.883-101.503)^2}{10}+\frac{(95.8431-101.503)^2}{10}+\frac{(115.605-101.503)^2}{10}+\frac{(125.794-101.503)^2}{10}+\frac{(111.393-101.503)^2}{10}=418.4782114$$

$$\mu = 101.503 \quad \sigma = \sqrt{418.4782114} = 20.45674$$

Data Set 4

1	2	3	4	5	6	7	8	9	10
105.712	97.0893	96.7681	120.47	89.7525	95.2807	103.804	114.145	85.3891	115.673

Mean:

$$\frac{105.712+97.0893+96.7681+120.47+89.7525+95.2807+103.804+114.145+85.3891+115.673}{10} = 102.4084$$

Variance / Standard Deviation:

$$\frac{(105.712-102.4084)^2}{10}+\frac{(97.0893-102.4084)^2}{10}+\frac{(96.7681-102.4084)^2}{10}+\frac{(120.47-102.4084)^2}{10}+\frac{(89.7525-102.4084)^2}{10}+\frac{(95.2807-102.4084)^2}{10}+\frac{(103.804-102.4084)^2}{10}+\frac{(114.145-102.4084)^2}{10}+\frac{(85.3891-102.4084)^2}{10}+\frac{(115.673-102.4084)^2}{10}=124.0149504$$

$$\mu = 102.4084 \quad \sigma = \sqrt{124.0149504} = 11.1362$$

Data Set 5

1	2	3	4	5	6	7	8	9	10
89.7525	95.2807	103.804	114.145	85.3891	115.673	87.2352	91.9985	100.059	112.062

Mean:

$$\frac{89.7525+95.2807+103.804+114.145+85.3891+115.673+87.2352+91.9985+100.059+112.062}{10} = 99.5399$$

Variance / Standard Deviation:

$$\frac{(89.7525-99.5399)^2}{10}+\frac{(95.2807-99.5399)^2}{10}+\frac{(103.804-99.5399)^2}{10}+\frac{(114.145-99.5399)^2}{10}+\frac{(85.3891-99.5399)^2}{10}+\frac{(115.673-99.5399)^2}{10}+\frac{(87.2352-99.5399)^2}{10}+\frac{(91.9985-99.5399)^2}{10}+\frac{(100.059-99.5399)^2}{10}+\frac{(112.062-99.5399)^2}{10}=117.2245786$$

$$\mu = 99.5399 \quad \sigma = \sqrt{117.2245786} = 10.82703$$

Data Set 6

1	2	3	4	5	6	7	8	9	10
99.7427	98.4174	116.425	105.712	97.0893	96.7681	120.47	89.7525	95.2807	103.804

Mean:

$$\frac{99.7427+98.4174+116.425+105.712+97.0893+96.7681+120.47+89.7525+95.2807+103.804}{10} = 102.3462$$

Variance / Standard Deviation:

$$\frac{(99.7427-102.3462)^2}{10} + \frac{(98.4174-102.3462)^2}{10} + \frac{(116.425-102.3462)^2}{10} + \frac{(105.712-102.3462)^2}{10} + \frac{(97.0893-102.3462)^2}{10} + \frac{(96.7681-102.3462)^2}{10} + \frac{(120.47-102.3462)^2}{10} + \frac{(89.7525-102.3462)^2}{10} + \frac{(95.2807-102.3462)^2}{10} + \frac{(103.804-102.3462)^2}{10} = 85.56483102$$

$$\mu = 102.3462 \quad \sigma = \sqrt{85.56483102} = 9.250126$$

Data Set 7

1	2	3	4	5	6	7	8	9	10
111.393	116.484	87.3041	95.3729	105.29	110.308	85.4532	103.599	101.003	122.887

Mean:

$$\frac{111.393+116.484+87.3041+95.3729+105.29+110.308+85.4532+103.599+101.003+122.887}{10} = 103.9094$$

Variance / Standard Deviation:

$$\frac{(111.393-103.9094)^2}{10} + \frac{(116.484-103.9094)^2}{10} + \frac{(87.3041-103.9094)^2}{10} + \frac{(95.3729-103.9094)^2}{10} + \frac{(105.29-103.9094)^2}{10} + \frac{(110.308-103.9094)^2}{10} + \frac{(85.4532-103.9094)^2}{10} + \frac{(103.599-103.9094)^2}{10} + \frac{(101.003-103.9094)^2}{10} + \frac{(122.887-103.9094)^2}{10} = 132.5100277$$

$$\mu = 102.3462 \quad \sigma = \sqrt{132.5100277} = 11.5113$$

Data Set 8

1	2	3	4	5	6	7	8	9	10
57.1007	99.7427	98.4174	116.425	105.712	97.0893	96.7681	120.47	89.7525	95.2807

Mean:

$$\frac{57.1007+99.7427+98.4174+116.425+105.712+97.0893+96.7681+120.47+89.7525+95.2807}{10} = 97.67584$$

Variance / Standard Deviation:

$$\frac{(57.1007-97.67584)^2}{10} + \frac{(99.7427-97.67584)^2}{10} + \frac{(98.4174-97.67584)^2}{10} + \frac{(116.425-97.67584)^2}{10} + \frac{(105.712-97.67584)^2}{10} + \frac{(97.0893-97.67584)^2}{10} + \frac{(96.7681-97.67584)^2}{10} + \frac{(120.47-97.67584)^2}{10} + \frac{(89.7525-97.67584)^2}{10} + \frac{(95.2807-97.67584)^2}{10} = 265.695216$$

$$\mu = 97.67584 \quad \sigma = \sqrt{265.695216} = 16.30016$$

Mean of the eight mean values:

$$\frac{100.56069+96.99588+101.503+102.4084+99.5399+102.3462+103.9094+97.67584}{10} = 80.493931$$

Standard Deviation of 8 mean values:

$$\frac{(100.56069-80.493931)^2}{8} + \frac{(96.99588-80.493931)^2}{8} + \frac{(101.503-80.493931)^2}{8} + \frac{(102.4084-80.493931)^2}{8} + \frac{(99.5399-80.493931)^2}{8} + \frac{(102.3462-80.493931)^2}{8} + \frac{(103.9094-80.493931)^2}{8} + \frac{(97.67584-80.493931)^2}{8} = 50.08874811$$

$$\mu = 80.493931 \quad \sigma_m = \sqrt{265.695216} = 7.077340469$$

3 Observations/Conclusions:

An overall theme throughout the numerical analysis is that the increase in sample size improves the accuracy of the data, both visually and mathematically. The first graph which represented a sample size of 10 is way different and less varied than the final data set which had a sample size of 100. The increase in sample size allows for more values which can be included in the analysis and can shape the output depending on whether it is higher, lower or on par with the mean value.

However this doesn't always work as intended, as seen in the data set with a sample size of 20, the mean value stayed relatively the same while the variance and standard deviation differed wildly, it is likely that the 10 extra values added to the sample size were not evenly spread out like all the others, so instead we got 10 extra values that were all really close to the mean value. As a result of this the mean value moved vary little but the standard deviation improved quite a lot as the 20 data points weren't spread out from the mean as much. This was a statistical anomaly that didn't occur in any of the other data sets, and was corrected by increasing the sample size even further.

As expected, as the number of values increased, the mean increased slightly, the standard deviation rarely differed and the standard error always decreased. When comparing our 8 data sets with 10 chosen values the mean varied even less, the standard deviation varied quite a bit more and the standard error of the mean values was approximately similar to the standard error in the first data set which also had 10 values, i.e 6.1 being near 7

The biggest takeaway from this experiment is that the best way to represent a statistic is by having as large of a sample size as possible, which will help in reducing things like your standard error and will help account for more variability in graphical representations.