CS-215 Assignment-1 Report

22B0914 - Utkarsh Pant
22B0994 - Jaladhi Joshi
22B1005 - Anirudh Garg

Contents

PART

I

Question 1

(a)

The first person has to pick up his/her book from the n possible books. The second person has to pick up his/her book from the $n-1$ possible books, and so on. Hence the probability is

$$\frac{1}{n} \cdot \frac{1}{n-1} \cdot \ldots \cdot \frac{1}{1} = \frac{1}{n!} \tag{0.1}$$

(b)

Similar to part (a), the first person has to pick up his/her book from the n possible books. The second person has to pick up his/her book from the $n-1$ possible books, and so on until m persons have got their book. The last person has to get his/her book from the $n-m+1$ possible books. Hence the probability is

$$\frac{1}{n} \cdot \frac{1}{n-1} \cdot \ldots \cdot \frac{1}{n-m+1} = \frac{(n-m)!}{n!} \tag{0.2}$$

(c)

The first person has to pick up a book from a set of m books out of the total n books. The second person has to pick up his/her book from a set of $m-1$ books out of the total $n-1$ books and so on until m persons have got their book. The last person has to get his/her book from the $n-m+1$ possible books. Hence the probability is

$$\frac{m}{n} \cdot \frac{m-1}{n-1} \cdot \ldots \cdot \frac{1}{n-m+1} = \frac{(n-m)!(m)!}{n!} \tag{0.3}$$

(d)

The probability of picking a clean book is independent of the book which the person picks. So the probability of the first person picking a clean book is $(1-p)$, as is the probability of the second person picking a clean book and so on until the first m persons have got a clean book. Hence the probability is

$$\underbrace{(1-p) \cdot (1-p) \cdot \ldots \cdot (1-p)}_{m \text{ times}} = (1-p)^m \tag{0.4}$$

(e)

The probability of picking a clean book is independent of the book which the person picks. The number of sets of m persons that can be picking a clean book is given by $\binom{n}{m}$. The probability of each person in that set picking a clean book is given by $(1-p)$. The probability of a persons not in that group picking an unclean book is p. Hence the probability is

$$\begin{aligned}
&\underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{m\text{ times}}\cdot\binom{n}{m}\cdot\underbrace{p\cdot p\cdot\ldots\cdot p}_{n-m\text{ times}} \\
&=(1-p)^m\cdot\binom{n}{m}\cdot p^{n-m}
\end{aligned}\tag{0.5}$$

Question 2

PART

II

From the definition of Standard deviation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2} \tag{0.6}$$

$$\Rightarrow \sigma^2(n-1) = \sum_{i=1}^{n}(x_i-\mu)^2 \tag{0.7}$$

As all the squares are ≥ 0 for each $1 \leq i \leq n$, $(x_i-\mu)^2 \geq 0$.
$\Rightarrow \sum_{i=1}^{n}(x_i-\mu)^2 \geq (x_i-\mu)^2$ for each $1 \leq i \leq n$
Using Equation 0.7,
$\Rightarrow \sigma^2(n-1) \geq (x_i-\mu)^2$ for each $1 \leq i \leq n$
Taking square root of both sides, we get $\sigma\sqrt{(n-1)} \geq |x_i-\mu|$ for each $1 \leq i \leq n$

Chebyshev's inequality says that $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$. If we put $k = \sqrt{n-1}$ in Chebyshev's inequality, we get $P(|X-\mu| \geq \sqrt{n-1}\sigma) \leq \frac{1}{n-1}$. For large n, Chebyshev's inequality says that the probability of $|X-\mu| \geq \sqrt{n-1}\sigma$ is very small and it tends to zero as n increases, which follows from the proved inequality $|x_i-\mu| \leq \sigma\sqrt{(n-1)}$. Hence for large n, this inequality and Chebyshev's inequality convey the same meaning.

Question 3

PART

III

Given event F is $|Q1 + Q2| > \epsilon$ and given $E1$ is $|Q1| > \epsilon/2$ and $E2$ is $|Q2| > \epsilon/2$.

For F to occur, following three cases are possible:

1. $|Q1| > \epsilon/2$ and $|Q2| > \epsilon/2$
2. $|Q1| > \epsilon/2$ and $|Q2| \leq \epsilon/2$
3. $|Q1| \leq \epsilon/2$ and $|Q2| > \epsilon/2$

Hence the event F is given by

$$F \subseteq (E1 \cap E2) \cup (E1 \cap E2') \cup (E1' \cap E2).$$

The probability $P(F)$ can be bounded:

$$P(F) \leq P((E1 \cap E2) \cup (E1 \cap E2') \cup (E1' \cap E2))$$

Since the events in the union are mutually exclusive and the events E_1 and E_2 are independent:

$$\leq P(E1 \cap E2) + P(E1 \cap E2') + P(E1' \cap E2)$$

$$\leq P(E1)P(E2) + P(E1)P(E2') + P(E1')P(E2)$$

$$\leq P(E1) + P(E2) - P(E1)P(E2)$$

$$P(F) \leq P(E1) + P(E2)$$

Question 4

PART

IV

Let E_1 and E_2 be the events that $Q_1 < q_1$ and $Q_2 < q_2$ respectively. Let E be the event that $Q_1 Q_2 < q_1 q_2$. Since all Q_1, Q_2, q_1 and q_2 are non-negative, the inequality $Q_1 Q_2 < q_1 q_2$ will hold if:

- $Q_1 < q_1$ and $Q_2 < q_2$ or,
- $Q_1 \geq q_1$ and $Q_2 < q_2$ and $Q_1 Q_2 < q_1 q_2$ or,
- $Q_1 < q_1$ and $Q_2 \geq q_2$ and $Q_1 Q_2 < q_1 q_2$

Since all of the above possibilities are mutually exclusive and exhaustive for E, mathematically we can write it as:

$$P(E) = P(E_1 \cap E_2) + P(E_1^c \cap E_2 \cap E) + P(E_1 \cap E_2^c \cap E)$$

Which implies,

$$P(E) \geq P(E_1 \cap E_2)$$

E_1 and E_2 are independent events since Q_1 being less than q_1 does not affect the probability of Q_2 being less than q_2. Hence,

$$P(E) \geq P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

Since,

$$P(E_1) \geq 1 - p_1 \text{ and } P(E_2) \geq 1 - p_2$$

$$P(E) \geq P(E_1) \cdot P(E_2) \geq (1 - p_1)(1 - p_2) \geq 1 - p_1 - p_2 + p_1 p_2$$

Since $p_1 p_2$ is non-negative, we have

$$P(E) \geq 1 - (p_1 + p_2)$$

or

$$P(Q_1 Q_2 < q_1 q_2) \geq 1 - (p_1 + p_2)$$

Question 5

PART

V

(a)

Since the probability of a car being behind any door is equal and independent of the choice of the contestant, $P(C_1|Z_1) = P(C_2|Z_1) = P(C_3|Z_1) = \frac{1}{3}$.

(b)

If the car is behind door 1 (the door chosen by the contestant), the host will open doors 2 and 3 with equal probability. Hence $P(H_3|C_1, Z_1) = \frac{1}{2}$.
If the car is behind door 2, then the host will only open door 3, so $P(H_3|C_2, Z_1) = 1$.
If the car is behind door 3, the host will never open door 3, so $P(H_3|C_3, Z_1) = 0$.

(c)

$$\begin{aligned} P(C_2|H_3, Z_1) &= \frac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{P(H_3, Z_1)} \\ &= \frac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{\sum_{i=1}^{3} P(H_3|C_i, Z_1) \cdot P(C_i, Z_1)} \end{aligned}$$

Since C_i and Z_i are independent events, $P(C_i, Z_j) = P(C_i)P(Z_j) = \frac{1}{9}$. Putting in values from above subparts, we get:

$$\begin{aligned} P(C_2|H_3, Z_1) &= \frac{1 \cdot \frac{1}{9}}{\frac{1}{2} \cdot \frac{1}{9} + 1 \cdot \frac{1}{9} + 0 \cdot \frac{1}{9}} \\ &= \frac{2}{3} \end{aligned}$$

(d)

Similarly, the conditional probability of the car being behind door 1 given that host opened door 3 ($P(C_1|H3, Z1)$) is:

$$\begin{aligned} P(C_1|H_3, Z_1) &= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3, Z_1)} \\ &= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{\sum_{i=1}^{3} P(H_3|C_i, Z_1) \cdot P(C_i, Z_1)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{9}}{\frac{1}{2} \cdot \frac{1}{9} + 1 \cdot \frac{1}{9} + 0 \cdot \frac{1}{9}} \\ &= \frac{1}{3} \end{aligned}$$

(e)

Using the probabilities obtained in the last two subparts, we can conclude that: If a contestant chooses a particular door and then the host opens a second door which is guaranteed to not have the car, we can conclude that the probability of the car being behind the third door ($\frac{2}{3}$) is greater than the probability of the car being behind the door which was selected by the contestant earlier ($\frac{1}{3}$).
Hence, switching is indeed proven to be beneficial to the contestant.

(f)

Let's recalculate all the probabilities for the whimsical host condition. Since the probability of a car being behind any door is equal and independent of the choice of the contestant, $P(C_i|Z_1) = \frac{1}{3}$ for all $i \in \{1, 2, 3\}$. The difference here will be in $P(H_3|C_i, Z_1)$. This probability will always be $\frac{1}{2}$, irrespective of whether the car is behind that door or not because the host is whimsical.

The conditional probability of winning by switching, $P(C_2|H3, Z1)$ is,

$$\begin{aligned} P(C_2|H_3, Z_1) &= \frac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{P(H_3, Z_1)} \\ &= \frac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{\sum_{i=1}^{3} P(H_3|C_i, Z_1) \cdot P(C_i, Z_1)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{9}}{\frac{1}{2} \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9}} \\ &= \frac{1}{3} \end{aligned}$$

The conditional probability of winning by not switching, $P(C_1|H_3, Z_1)$ is,

$$
\begin{aligned}
P(C_1|H3, Z1) &= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3, Z_1)} \\
&= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{\sum_{i=1}^{3} P(H_3|C_i, Z_1) \cdot P(C_i, Z_1)} \\
&= \frac{\frac{1}{2} \cdot \frac{1}{9}}{\frac{1}{2} \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9}} \\
&= \frac{1}{3}
\end{aligned}
$$

Hence, the conditional probability of winning by switching is the same as winning by not switching. Therefore, it is not beneficial to switch choices in this case.

There is also a possibility that because the host is whimsical, they may actually open the door containing the car, which is given by $P(C_3|H3, Z1)$:

$$
\begin{aligned}
P(C_3|H3, Z1) &= \frac{P(H_3|C_3, Z_1)P(C_3, Z_1)}{P(H_3, Z_1)} \\
&= \frac{P(H_3|C_3, Z_1)P(C_3, Z_1)}{\sum_{i=1}^{3} P(H_3|C_i, Z_1) \cdot P(C_i, Z_1)} \\
&= \frac{\frac{1}{2} \cdot \frac{1}{9}}{\frac{1}{2} \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9} + \frac{1}{2} \cdot \frac{1}{9}} \\
&= \frac{1}{3}
\end{aligned}
$$

Question 6

PART

VI

The data for the relative mean squared error between each result and the original sine wave is

Table 1. Relative Mean Square Error (RMSE) Comparison

Method	**f = 0.3**	**f = 0.6**
Median Filtering	3.1984	447.2337
Mean Filtering	55.1707	211.6589
Quartile Filtering	0.0125	33.5588

Note: f denotes the corruption percentage.

From the data, it's evident that the Quartile Filtering method consistently produces the lowest RMSE values for both f = 0.3 and f = 0.6. This means that the quartile-based filtering method results in predictions that are closer to the actual values, on average, compared to the other two methods.
This happens because after random 30% of the values have been corrupted, to get the median or quartile in the neighbourhood of any point we sort the values around the point and the corrupted values have very less probability of affecting the median or quartile at that point. Hence these methods are more robust to corruption in data.
However, from the data we can also see that at 60% corrupted values, the RMSE of median is even higher than the RMSE of mean while quartile filtering is still much better off. This is because at f>50% even median is susceptible to errors as the corrupted values might be the median in a neighbourhood. Quartile remains robust as it is adjusting the values to 25 percentile keeping it relatively secure from the 60% error profile.
The quartile-based filtering method may perform better because it's more robust to outliers compared to the median and mean methods. Outliers can significantly affect the accuracy of the predictions as they do in case of mean, but quartiles provide a measure of central tendency that is less sensitive to extreme values.

Here is the code used for the $f = 0.6$ case

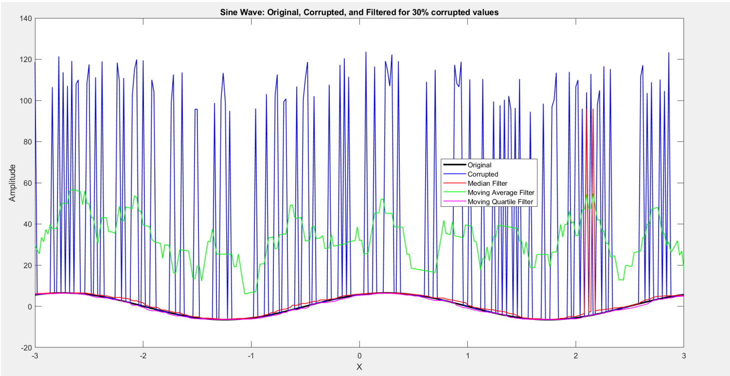
```
clear;
clc;
x = -3:0.02:3;

y = 6.5 * sin(2.1 * x + pi/3);

f = 0.6;

```

```
num_samples = length(y);
num_corrupted = round(f * num_samples);
corrupted_indices = randperm(num_samples, num_corrupted);

corruption_values = rand(1, num_corrupted) * 20 + 100;

z = y;
z(corrupted_indices) = z(corrupted_indices) + corruption_values;

ymedian=zeros(size(z));
ymean=zeros(size(z));
yquartile=zeros(size(z));
nb=8;
for i=1:length(z);
    starti = max(1, i - nb);
    endi = min(length(z), i + nb);

    neighbourhood = z(starti:endi);

    ymedian(i) = median(neighbourhood);
    yquartile(i) = quantile(neighbourhood, 0.25);
    ymean(i) = mean(neighbourhood);
end
rmse_median= sum((y - ymedian).^2)/sum(y.^2);
rmse_mean= sum((y - ymean).^2)/sum(y.^2);
rmse_quartile= sum((y - yquartile).^2)/sum(y.^2);
figure;
plot(x, y, 'k', 'LineWidth', 2);
hold on;
plot(x, z, 'b', 'LineWidth', 1);
plot(x, ymedian, 'r', 'LineWidth', 1);
plot(x, ymean, 'g', 'LineWidth', 1);
plot(x, yquartile, 'm', 'LineWidth', 1);
hold off;
xlabel('X');
ylabel('Amplitude');
title('Sine Wave: Original, Corrupted, and Filtered for 60% corrupted values');
legend('Original', 'Corrupted', 'Median Filter', 'Moving Average Filter', 'Moving Quartile Filter', 'Location', 'best');
```

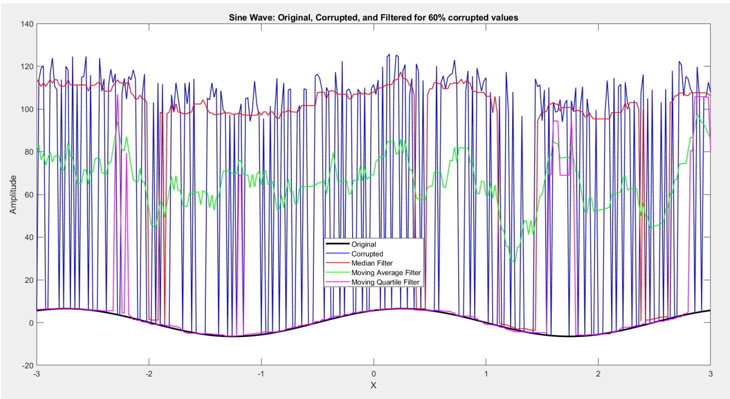
Listing 1. MATLAB Code

Here are the plots generated for both cases

- At $f = 0.3$:



- At $f = 0.6$



Question 7

PART

VII

NewMean

$OldMean = \frac{\sum_{i=1}^{n} Ai}{n}$

$\Rightarrow \sum_{i=1}^{n} Ai = OldMean \times n$

$NewMean = \frac{\sum_{i=1}^{n+1} Ai}{n+1}$

$NewMean = \frac{\sum_{i=1}^{n} Ai + NewDataValue}{n+1}$

$NewMean = \frac{OldMean \times n + NewDataValue}{n+1}$

```
function newMean = UpdateMean(OldMean, NewDataValue, n)
    newMean = (n * OldMean + NewDataValue) / (n + 1);
end
```

Listing 2. Update Mean Algorithm

NewMedian

1. When n is even:

$$\text{OldMedian} = \frac{A\left[\frac{n}{2}\right] + A\left[\frac{n+2}{2}\right]}{2}$$

On adding a NewDataValue,the number of terms become n+1 which is odd

- If NewDataValue is greater than or equal to $A[\frac{n+2}{2}]$ then the middle term will be $A[\frac{n+2}{2}]$ and hence it will be the new median.
- If NewDataValue is lesser than or equal to $A[\frac{n}{2}]$ then the middle term will be $A[\frac{n}{2}]$ and hence it will be the new median.
- The other case where the NewDataValue lies between $A[\frac{n}{2}]$ and $A[\frac{n+2}{2}]$, the middle term will be the NewDataValue inserted and hence it will be the new median.

2. When n is odd:

$$\text{OldMedian} = A\left[\frac{n+1}{2}\right]$$

On adding a NewDataValue,the number of terms become n+1 which is even

- If NewDataValue is greater than or equal to $A[\frac{n+1}{2}]$ then the two middle terms are the old median and the term just right to it, which is the minimum of NewDataValue and $A[\frac{n+3}{2}]$
- The other condition in which the NewDataValue is lesser than $A[\frac{n+1}{2}]$,then the two middle terms are the old median and the term just left to it, which is the maximum of NewDataValue and $A[\frac{n-1}{2}]$

```
function newMedian = UpdateMedian(oldMedian, NewDataValue, A,
    n)
    if mod(n, 2) == 0
        if NewDataValue >= A(n/2 + 1)
            newMedian = A(n/2 + 1);
        elseif NewDataValue <= A(n/2)
            newMedian = A(n/2);
        else
            newMedian = NewDataValue;
        end
    else
        if NewDataValue >= oldMedian
            newMedian = (min(NewDataValue, A((n+3)/2)) +
                oldMedian) / 2;
        else
            newMedian = (max(NewDataValue, A((n-1)/2)) +
                oldMedian) / 2;
        end
    end
end
```

Listing 3. Update Median Algorithm

Derivation of New Standard Deviation

Starting with the original standard deviation formula:

$$\sigma_{old} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_{old})^2}{n-1}}$$

Where: - x_i is the ith data value.
- μ_{old} is the mean (average) of the data values.
- n is the number of data values.

We want to modify this formula for $n+1$ terms.

$$\mu_{new} = \frac{\sum_{i=1}^{n+1} x_i}{n+1}$$

So,

$$\sum_{i=1}^{n+1} x_i = \mu_{new} \times (n+1)$$

$$\Rightarrow \mu_{new} \times (n+1) = \sum_{i=1}^{n} x_i + A$$

$$\Rightarrow \sum_{i=1}^{n} x_i = \mu_{new} \times (n+1) - A$$

where A is the NewDataValue;

$$\mu_{old} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\mu_{old} = \frac{\mu_{new} \times (n+1) - A}{n}$$

$$\sigma_{old}^2 \times (n-1) = \sum_{i=1}^{n} (x_i - \mu_{old})^2$$

$$\sigma_{old}^2 \times (n-1) = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} \mu_{old}^2 - \sum_{i=1}^{n} 2 \times \mu_{old} \times x_i$$

$$\sum_{i=1}^{n} x_i^2 = \sigma_{old}^2 \times (n-1) - \sum_{i=1}^{n} \mu_{old}^2 + \sum_{i=1}^{n} 2 \times \mu_{old} \times x_i$$

$$\text{New Standard Deviation (newStd)} \sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n+1} (x_i - \mu_{new})^2}{n}}$$

$$\text{New Standard Deviation (newStd)} \sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n+1} x_i^2 + \sum_{i=1}^{n+1} \mu_{new}^2 - \sum_{i=1}^{n+1} 2\mu_{new} x_i}{n}}$$

$$\sigma_{new} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 + A^2 + \mu_{new}^2 \times (n+1) - 2\mu_{new}(\sum_{i=1}^{n} x_i + A)}{n}}$$

$$\sigma_{new} = \sqrt{\frac{\sigma_{old}^2 \times (n-1) - \sum_{i=1}^{n} \mu_{old}^2 + \sum_{i=1}^{n} 2 \times \mu_{old} \times x_i + A^2 + \mu_{new}^2 \times (n+1) - 2\mu_{new}(\sum_{i=1}^{n} x_i + A)}{n}}$$

$$\sigma_{new} = \sqrt{\frac{\sigma_{old}^2 \times (n-1) - \mu_{old}^2 \times n + 2 \times \mu_{old} \sum_{i=1}^{n} x_i + A^2 + \mu_{new}^2 \times (n+1) - 2\mu_{new}(\sum_{i=1}^{n} x_i + A)}{n}}$$

Using the value of $\sum_{i=1}^{n} x_i$

$$\sigma_{new} = \sqrt{\frac{\sigma_{old}^2 \times (n-1) - \mu_{old}^2 \times n + 2 \times \mu_{old}(\mu_{new} \times (n+1) - A) + A^2 + \mu_{new}^2 \times (n+1) - 2\mu_{new}(\mu_{new} \times (n+1))}{n}}$$

$$\sigma_{new} = \sqrt{\frac{\sigma_{old}^2 \times (n-1) - \mu_{old}^2 \times n + 2 \times \mu_{old}(\mu_{new} \times (n+1) - A) + A^2 - \mu_{new}^2 \times (n+1)}{n}}$$

Using the value of μ_{old}

$$\sigma_{new} = \sqrt{\frac{\sigma_{old}^2 \times (n-1) - (\frac{\mu_{new} \times (n+1) - A}{n})^2 \times n + 2 \times (\frac{\mu_{new} \times (n+1) - A}{n})(\mu_{new} \times (n+1) - A) + A^2 - \mu_{new}^2 \times (n+1)}{n}}$$

$$\sigma_{\text{new}} = \sqrt{\frac{\sigma_{\text{old}}^2 \times n \times (n-1) + \mu_{\text{new}}^2 \times (n+1)^2 + A^2 - 2 \times A \times \mu_{\text{new}} \times (n+1) + n \times A^2 - \mu_{\text{new}}^2 \times n \times (n+1)}{n^2}}$$

$$\sigma_{\text{new}} = \sqrt{\frac{\sigma_{\text{old}}^2 \times n \times (n-1) + \mu_{\text{new}}^2 \times (n+1) - 2 \times A \times \mu_{\text{new}} \times (n+1) + (n+1) \times A^2}{n^2}}$$

$$\sigma_{\text{new}} = \sqrt{\frac{\sigma_{\text{old}}^2 \times n \times (n-1) + (n+1) \times (\mu_{new} - A)^2}{n^2}}$$

End of Derivation

```
function newStd = UpdateStd(OldMean, OldStd, NewMean, NewDataValue, n)
    newStd = sqrt(((n - 1) *n* OldStd^2 + (n+1)*(NewDataValue - NewMean)^2) / n^2);
end
```

Listing 4. Update Standard Deviation Algorithm

Updating the Histogram To update the histogram after adding a new value, we would need to determine the appropriate bin for the new value and increment the count for that bin.

- Identify the bin or range where the new value falls.
- Increment the count of that bin in the histogram by 1.