Question 1

(a) Total number of ways in which people can pick books $= \underbrace{n}_{\substack{\text{first}\\ \text{person's}\\ \text{choices}}} \cdot \underbrace{(n-1)}_{\substack{\text{second}\\ \text{"}\\ \text{"}}} \cdot (n-2) \cdots 2 \cdot 1$

$= n!$

Number of cases when every person picks his/her own book. $= 1$

$\therefore$ $\boxed{\text{Probability} = \frac{1}{n!}}$

(b) Number of ways, the first m people pick the books $= 1$

Number of ways in which the next $(n-m)$ people pick books $= (n-m)!$

$\therefore$ $\boxed{\text{Probability} = \frac{(n-m)!}{n!}}$

(c) Number of ways the first m people pick the books $= m!$

Number of ways the next $(n-m)$ people pick the ~~ways =~~ books $= (n-m)!$

$\boxed{\text{Probability} = \frac{((n-m)!)(m!)}{n!}}$

(d) For any person, the probability of picking a clean book is $(1-p)$.

$\therefore$ If the first m people pick clean books.

$\therefore$ $\boxed{\text{Probability} = (1-p)^m}$

(e) ~~m people picking clean books.~~

(e) Probability of m people picking clean books $= (1-p)^m$

Probability of (n-m) people picking unclean books $\cancel{(n-m)}^{m} = p^{(n-m)}$.

$\therefore \boxed{\text{Probability} = {}^nC_m (1-p)^m p^{(n-m)}}$

---

~~Question 2~~

$\cancel{|x_i - \mu| \le \sqrt{\phantom{x}}}$

Question 2

$$|x_i - \mu| = \sqrt{(x_i-\mu)^2} \le \sqrt{\sum_{i=1}^{N} (x_i-\mu)^2}$$

$$\sqrt{\sum_{i=1}^{n} (x_i-\mu)^2} = \frac{\sqrt{\sum_{i=1}^{N} (x_i-\mu)^2 \cdot \cancel{\beta}} \cdot \sqrt{N-1}}{\sqrt{N-1}}$$

$$= \sigma\sqrt{N-1}$$

$\therefore$

$$|x_i - \mu| \le \sigma\sqrt{N-1} \quad \forall \cancel{i \in \mathbb{N}} \; i \in \{1, 2 \ldots N\}$$

Hence proved. :)

Date:

Comparison with Chebyshev's →

We know that Chebyshev's inequality is

$$\frac{\left|\left\{|x_i - \mu| > k\sigma\right\}\right|}{m} \leq \frac{1}{k^2}$$

& our inequality derived from question is

$$|x_i - \mu| \leq \sigma\sqrt{n-1} \quad \forall\, i \in \{1, 2, \ldots, n\}$$

Hence our inequality states that →

$$\frac{\left|\left\{|x_i - \mu| \leq \sigma\sqrt{n-1}\right\}\right|}{m} = 1$$

(exact 1)

$|A|$ = cardinality(A)

But from Chebyshev's, {put $k = \sqrt{n-1}$}

$$\frac{\left|\left\{|x_i - \mu| \leq \sigma\sqrt{n-1}\right\}\right|}{m} \geq 1 - \frac{1}{\sqrt{n-1}^2} = \frac{n-2}{n-1}$$

Hence Chebyshev's gives a range of values for this $k = \sqrt{n-1}$. Hence our inequality is better as we get exact value

Also it shows the limit of use of Chebyshev, after $k \geq \sqrt{n-1}$, Chebyshev gives useless ranges when answers are exact.

## Solution for Q3

## Information provided to us :-

Event $F = \{|Q_1 + Q_2| > \epsilon\}$, $E_1 = \{|Q_1| > \frac{\epsilon}{2}\}$ and $E_2 = \{|Q_2| > \frac{\epsilon}{2}\}$

### Their Complementary forms :-

Event $F^c = \{|Q_1 + Q_2| \leq \epsilon\}$, $E_1^c = \{|Q_1| \leq \frac{\epsilon}{2}\}$ and $E_2^c = \{|Q_2| \leq \frac{\epsilon}{2}\}$

## Important Inequality :-

$$\forall x, y : |x + y| \leq |x| + |y| \tag{1}$$

Apply equation (1) on $Q_1$ and $Q_2$, hence we get : $|Q_1 + Q_2| \leq |Q_1| + |Q_2|$ for any $Q_1$, $Q_2$ ... (A)

## Important Observation :-

If $|Q_1| \leq \frac{\epsilon}{2}$ **AND** $|Q_2| \leq \frac{\epsilon}{2}$ then from equation (A) we see that :-

$$|Q_1 + Q_2| \leq |Q_1| + |Q_2| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \leq \epsilon \implies |Q_1 + Q_2| \leq \epsilon \tag{2}$$

Hence, all values satisfying $|Q_1| \leq \frac{\epsilon}{2}$ **AND** $|Q_2| \leq \frac{\epsilon}{2}$ must also satisfy $|Q_1 + Q_2| \leq \epsilon$.

## Using this Observation :-

We see that the set...

$$\{|Q_1| \leq \frac{\epsilon}{2}\textbf{AND}|Q_2| \leq \frac{\epsilon}{2}\} = E_1^c \cap E_2^c \tag{3}$$

Because **AND** means that both the inequalities must follow at the same time, hence that set is **intersection** of the two complementary events $E_1^c$, $E_2^c$.

We already know $\{|Q_1 + Q_2| \leq \epsilon\} = F^c$.

### Important Remark :-

All the values belonging to $E_1^c \cap E_2^c$ must also belong to the set $F^c$, from our above observation, hence the former set is **subset** of the latter, or...

$$E_1^c \cap E_2^c \subseteq F^c \tag{4}$$

## Some Properties of Probability :-

We will use some properties of probability...

- a) If $A \subseteq B$ then $P(A) \leq P(B)$.
- b) $P(A^c) = 1 - P(A)$
- c) $P(A \cup B) = P(A) + P(B) - P(A \cap B) \implies P(A \cup B) \leq P(A) + P(B)$

Property (c) has been proved on the next page.

## Final Calculations :-

We have

$$E_1^c \cap E_2^c \subseteq F^c$$

then using (a)

$$\implies P(E_1^c \cap E_2^c) \leq P(F^c)$$

By **De-Morgan's Law**, $E_1^c \cap E_2^c = (E_1 \cup E_2)^c$, hence,

$$\implies P((E_1 \cup E_2)^c) \leq P(F^c)$$

Using (b),

$$\implies 1 - P(E_1 \cup E_2) \leq 1 - P(F)$$

$$\implies P(F) \leq P(E_1 \cup E_2)$$

Using (c),

$$P(F) \leq P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$$
$$\implies P(F) \leq P(E_1) + P(E_2)$$

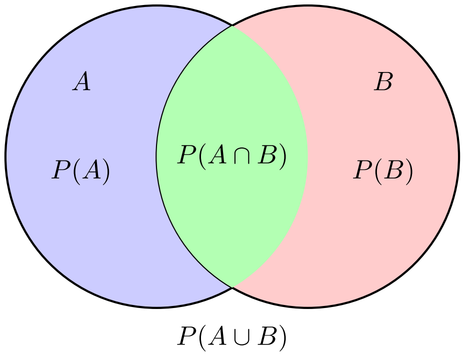
Hence Proved !!

## Proof of $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Let $A$ and $B$ be two events. We want to prove that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Consider the Venn diagram below:




The total area of the two circles represents the probability $P(A \cup B)$.

The area of circle $A$ represents $P(A)$, and the area of circle $B$ represents $P(B)$.

The overlapping region represents the intersection $A \cap B$, and its area is $P(A \cap B)$.

According to the Venn diagram, we can see that the overlapping area is counted twice when calculating $P(A \cup B)$. Thus, we need to subtract the area of $A \cap B$ once to avoid double counting.

Therefore, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

which completes the proof.

# Question 4
# Handwritten Solution
# By Kavya Gupta below

Let us define event $E_1 : \{ Q_1 < q_1 \}$
$E_2 : \{ Q_2 < q_2 \}$ and $E : \{ Q_1 Q_2 < q_1 q_2 \}$
Given : $Q_1$ and $Q_2$ and $q_1, q_2$ are non-negative
Hence for $Q_1 < q_1$ and $Q_2 < q_2$,
$Q_1 Q_2 < q_1 q_2$
So all values satisfying $Q_1 < q_1$ AND $Q_2 < q_2$ also satisfy $Q_1 Q_2 < q_1 q_2$
Now $\{ Q_1 < q_1 \text{ AND } Q_2 < q_2 \} \equiv E_1 \cap E_2$
and $\{ Q_1 Q_2 < q_1 q_2 \} \equiv E$
Hence all values belonging to $E_1 \cap E_2$ also belong to $E$

$\therefore$ $E_1 \cap E_2 \subseteq E$

$E_1 \cap E_2$ is subset of $E$.

(*) Now property of probability:
$A \subseteq B \Rightarrow P(A) \leq P(B)$

so $P(E_1 \cap E_2) \leq P(E)$

(*) Now as $Q_1, Q_2$ are independent, events $Q_1 < q_1$ and $Q_2 < q_2$ are also independent $\Longrightarrow E_1, E_2$ are independent.
For any independent events A, B,
$P(A \cap B) = P(A) \cdot P(B)$

So, $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$

so $P(E) \geqslant P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$

or $P(E) \geqslant P(E_1) \cdot P(E_2)$

and given $\longrightarrow$ $P(E_1) \geqslant 1-p_1$ and

$P(E_2) \geqslant 1-p_2$

so $P(E_1) \cdot P(E_2) \geqslant (1-p_1)(1-p_2)$

This inequality is okay as both $P(E_1), P(E_2) \geqslant 0$ are non-negative and $1-p_1$ and $1-p_2$ also $\geqslant 0$

Hence $P(E) \geqslant P(E_1) \cdot P(E_2) \geqslant (1-p_1)(1-p_2)$

or $P(E) \geqslant (1-p_1)(1-p_2)$

$= 1 - p_1 - p_2 + p_1 p_2$

or $P(E) \geqslant 1 - p_1 - p_2 + p_1 p_2 \geqslant 1 - p_1 - p_2$

(as $p_1 p_2 \geqslant 0$)

so $P(E) \geqslant 1-(p_1+p_2)$

or $P\left(\theta_1 \theta_2 < q_1 q_2\right) \geqslant 1-(p_1+p_2)$

Hence proved.

22B1053 , 22B1205 , 22B1025

PAGE No.
DATE / /

**Q5** By definition of conditional probability of event E given event F, is $\frac{P(E\cap F)}{P(F)}$

(a) $$P(C_i|Z_1) = \frac{P(C_i\cap Z_1)}{P(Z_1)}$$

we assume that contestant is equally likely to choose any door

$$\therefore P(Z_1) = 1/3$$

$\because$ $C_i$ and $Z_1$ are indipedent events.

$$\therefore P(C_i\cap Z_1) = P(C_i)P(Z_1) = \frac{1}{3}\cdot\frac{1}{3}$$

$$\Rightarrow P(C_i|Z_1) = 1/3 \quad \text{for } i=1,2,3$$

(b) $$P(H_3|C_i,Z_1) = \frac{P(H_3\cap C_i\cap Z_1)}{P(C_i\cap Z_1)}$$ $\left(\because (C_i,Z_1) = \{C_i\cap Z_1\}\right)$

**Case I** $i=1$

In this case $\because$ Car is behind door 1

$\therefore$ Host is equally likely to open door 2 or door 3.

now, we since host opens either door 2 or door 3.

$$P(H_3\cap C_1\cap Z_1) + P(H_2\cap C_1\cap Z_1) = P(C_1\cap Z_1)$$ ($H_3, H_2$ are mutually exclusive)

$$\therefore 2P(H_3\cap C_1\cap Z_1) = P(C_1\cap Z_1)$$

$$\therefore P(H_3\cap C_1\cap Z_1) = \frac{P(C_1\cap Z_1)}{2}$$

$$\Rightarrow P(H_3|C_1,Z_1) = 1/2$$

**Case II** $i=2$

In this case, since car is behind door 2, host will never open door 2.

$$\therefore P(H_2\cap C_2\cap Z_1) = 0$$

now as $$P(H_3\cap C_2\cap Z_1) + P(H_2\cap C_2\cap Z_1) = P(C_2\cap Z_1)$$

$$\therefore P(H_3\cap C_2\cap Z_1) = P(C_2\cap Z_1)$$

$$\therefore P(H_3|C_2,Z_1) = 1$$

PAGE No.
DATE / /

Q.5 Case III $i=3$

In this case since Car is behind door 3 $\therefore$ host will never open door 3.

$\therefore P(H_3 \cap C_3 \cap Z_1) = 0$

$\therefore P(H_3 | C_3, Z_1) = 0$

(c) $P(C_2 | H_3, Z_1) = \dfrac{P(C_2 \cap H_3 \cap Z_1)}{P(H_3 \cap Z_1)}$ $\left( \because (H_3, Z_1) = \{H_3 \cap Z_1\} \right)$

now $\because C_1, C_2, C_3$ are mutually exclusive and exhaustive collection of events for car being behind any door.

$\therefore P(H_3 \cap Z_1) = P(H_3 \cap C_1 \cap Z_1) + P(H_3 \cap C_2 \cap Z_1) + P(H_3 \cap C_3 \cap Z_1)$

$= \dfrac{P(C_1 \cap Z_1)}{2} + P(C_2 \cap Z_1) + 0$ (from (b))

$= \dfrac{3}{2} P(C_2 \cap Z_1)$ $\left( \because P(C_1 \cap Z_1) = P(C_2 \cap Z_1) = 1/9 \text{ from (a)} \right)$

$\therefore P(C_2 | H_3, Z_1) = \dfrac{2}{3} \dfrac{P(C_2 \cap H_3 \cap Z_1)}{P(C_2 \cap Z_1)} = \dfrac{2}{3} \dfrac{P(C_2 \cap Z_1)}{P(C_2 \cap Z_1)} = 2/3$

(d) $P(C_1 | H_3, Z_1) = \dfrac{P(C_1 \cap H_3 \cap Z_1)}{P(H_3 \cap Z_1)} = \dfrac{P(C_1 \cap Z_1)}{2 P(H_3 \cap Z_1)} = \dfrac{1}{2} \cdot \dfrac{2}{3} \dfrac{P(C_1 \cap Z_1)}{P(C_2 \cap Z_1)} = \dfrac{1}{2} \cdot \dfrac{2}{3} \dfrac{P(C_2 \cap Z_1)}{P(C_2 \cap Z_1)} = 1/3$

(e) $\because P(C_1 | H_3, Z_1) < P(C_2 | H_3, Z_1)$

$\therefore$ There is more probability that the car is behind door 2 if contestant has chosen door 1 and host had opened door 3.

$\therefore$ The contestant must switch in order to increase his/her chances for wining.

NOTE:- This extra probability in case $C_2$ is due to host baisness, not to open the door with car $\therefore$ ~~sk~~ skiping some possibilities.

PAGE No.

DATE / /

**Q5** (f) we start with calculating $P(H_3 \cap C_i \cap Z_1)$

$\because$ Host is equal likely to open door 2 or door 3, irrespective of where car is.

$\therefore P(H_3 \cap C_i \cap Z_1) = P(H_2 \cap C_i \cap Z_1)$

$\therefore$ as $P(H_3 \cap C_i \cap Z_1) + P(H_2 \cap C_i \cap Z_1) = P(C_i \cap Z_1)$

$\therefore \quad 2P(H_3 \cap C_i \cap Z_1) = P(C_i \cap Z_1)$

$\therefore P(H_3 \cap C_i \cap Z_1) = \dfrac{P(C_i \cap Z_1)}{2}$

now we calculate,

$$P(C_2 | H_3, Z_1) = \frac{P(C_2 \cap H_3 \cap Z_1)}{P(H_3 \cap Z_1)}$$

$$= \frac{P(C_2 \cap Z_1)/2}{\sum_{i=1}^{3} P(H_3 \cap C_i \cap Z_1)} = \frac{P(C_2 \cap Z_1)/2}{3P(C_2 \cap Z_1)/2} \qquad \left(\because P(C_i \cap Z_1) = 1/9, i = 1,2,3\right)$$

$$= 1/3$$

similarly,

$$P(C_1 | H_3, Z_1) = \frac{P(C_1 \cap H_3 \cap Z_1)}{P(H_3 \cap Z_1)} = \frac{P(C_1 \cap Z_1)/2}{3P(C_2 \cap Z_1)/2} = \frac{P(C_1 \cap Z_1)/2}{3P(C_2 \cap Z_1)/2} = 1/3$$

$\because P(C_2 | H_3, Z_1) = P(C_1 | H_3, Z_1)$

$\therefore$ There is equal probability that the car is behind door 2 or door 3 if contestant has chosen door 1 and host hand opened door 3.

$\therefore$ It is bad nor good to switch.

**NOTE:-** $P(C_2 | H_3, Z_1) + P(C_1 | H_3, Z_1) \neq 1$

Because $P(C_3 | H_3, Z_1) \neq 0$, i.e., the case when contestant loses if car is behind door 3.

# Question 6 : Comparison of Filtering Methods for Corrupted Sine Wave

Kavya Gupta

August 21, 2023

## 1 Instructions for Running the Program

We have provided two MATLAB `.m` files for this question.

1. `Q6.m` MATLAB file for fraction f=0.3
2. `Q6_2.m` MATLAB file for fraction f=0.6

Running the code on MATLAB Editor will provide you with the appropriate graph and also the relative mean squared error for the 3 different methods (median/mean/quartile), will be printed in the output terminal.

We have provided two graph images of `.png` type.

1. `Q6.png` graph for fraction f=0.3
2. `Q6_2 (1).png` graph for fraction f=0.6

These two graphs are already attached in this file.

## 2 Introduction

In this report, we compare three different filtering methods applied to a corrupted sine wave. The corrupted sine wave is obtained by adding random noise to the original clean sine wave. The filtering methods examined are:

1. Moving Median Filtering
2. Moving Mean (Arithmetic Mean) Filtering
3. Moving Quartile Filtering

The goal is to evaluate the effectiveness of these methods in reducing the impact of corruption on the sine wave.

# 3 Methodology

The following steps were performed to conduct the comparison:

1. Generate the original clean sine wave and introduce corruption by adding random noise.
2. Apply each filtering method to the corrupted sine wave to obtain filtered signals.
3. Calculate the relative mean squared error (RMSE) between each filtered signal and the original clean sine wave.

# 4 Results

## 4.1 Effectiveness of Different Filtering Methods

The effectiveness of the filtering methods was evaluated by calculating the RMSE between each filtered signal and the original clean sine wave. The RMSE values for different corruption fractions (f = 0.3 and f = 0.6) are presented in Table 1.

| **Method** | **f = 0.3** | **f = 0.6** |
|---|---|---|
| Moving Median Filtering | 11.3022 | 424.6993 |
| Moving Mean Filtering | 58.7516 | 211.4631 |
| Moving Quartile Filtering | 0.0128 | 40.0488 |

Table 1: Relative Mean Squared Error (RMSE) for Different Filtering Methods

## 4.2 Graphical Comparison

Figures 1 and 2 show graphical comparisons of the original clean sine wave, corrupted sine wave, and the filtered signals using each method for corruption fractions f = 0.3 and f = 0.6, respectively.

# 5 Discussion

Based on the results presented in Table 1 and the graphical comparisons in Figures 1 and 2, it is evident that the Moving Quartile Filtering method consistently produces the lowest relative mean squared error (RMSE) when compared to the original clean sine wave. This indicates that the Moving Quartile Filtering method is better at preserving the underlying signal structure while removing corruption.

The Moving Mean Filtering method, on the other hand, tends to have higher RMSE values, especially for higher corruption fractions (f = 0.6), which suggests that it might not perform as well in scenarios with significant corruption.

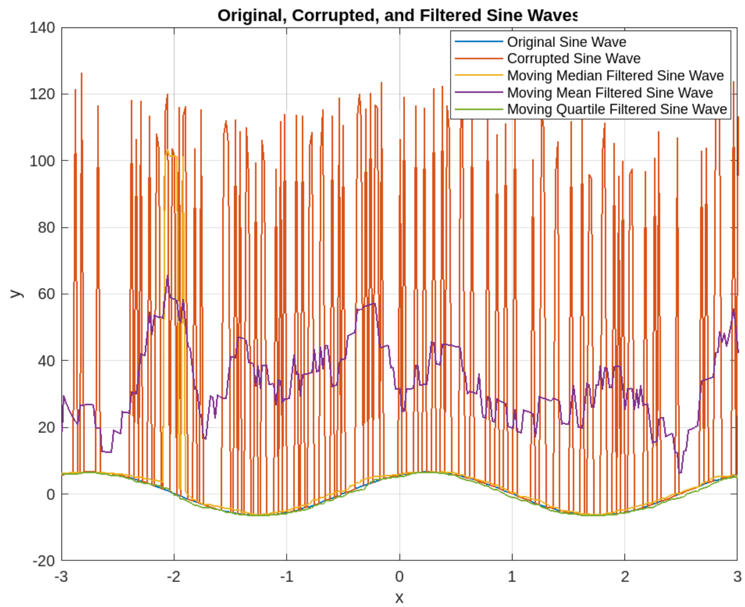



Figure 1: Comparison of Filtering Methods for f = 0.3

The Moving Median Filtering method also performs well in reducing corruption, as evidenced by its lower RMSE values, but it does not consistently outperform the Moving Quartile Filtering method.

# 6 Why ?

So in both the cases, Mean will be directly affected by adding positive bias to the data, hence even for f=0.3 or f=0.6, the moving mean filtered graph is almost equally affected.

For f=0.3, moving median filtering works better, this is because for f=0.3 there are lesser no. of points affected, hence the probability that in a neighbourhood the median of that data gets affected is lesser, but when f=0.6, then there are more points corrupted hence the density of points corrupted in a neighbourhood increases and chances of median being affected increases, leading almost all the points of `ymedian` getting corrupted, which is apparent from the table.

**Median is present above 50 percentile of the data. So, corrupting around 50 percent of data can lead the median point to the corrupted data input and hence median is also corrupted. That's why at f=0.6 (60 percent) median filtering is poor.**

Now the moving quartile method works the best because it finds the lowest

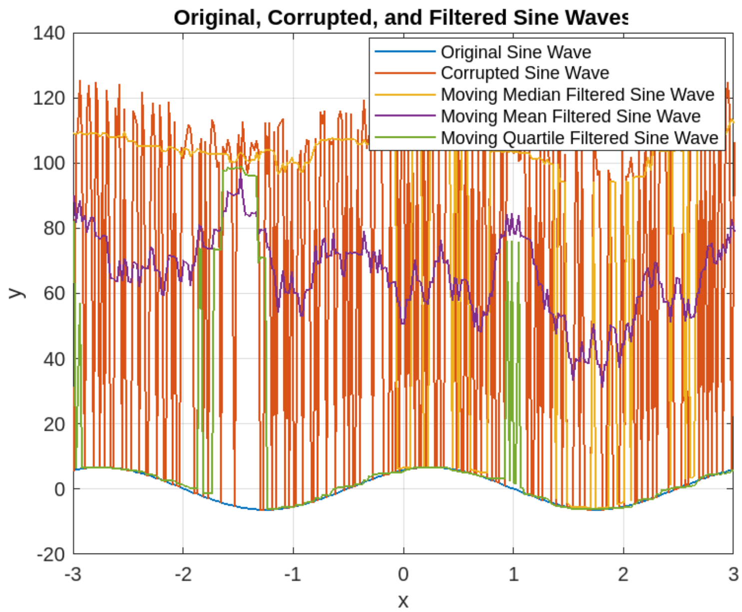



Figure 2: Comparison of Filtering Methods for f = 0.6

25 percent of the data in a neighbourhood, which are least likely to get affected for f=0.3 or f=0.6, median is nothing but 50 percent quartile. Lower percentile values will require a lot higher corruption values to get corrupted, hence 25 percent quartile remains almost same throughout.

**Quartile is present at 25 percentile of the data. So, corrupting around 30 or 60 percent of data least likely affects 25 percentile point and hence it is not likely corrupted. That's why even at f=0.6 (60 percent) quartile filtering is fine. We would have to corrupt around 75 percent of data for the quartile to be corrupted as well.**

# 7 Conclusion

In this study, we compared three filtering methods for removing corruption from a sine wave signal. The Moving Quartile Filtering method demonstrated superior performance in terms of relative mean squared error (RMSE) when compared to the original clean sine wave. This suggests that the Moving Quartile Filtering method is a robust choice for effectively reducing corruption while preserving the signal's underlying characteristics.

# Question 7 : Updating Mean, Median, and Standard Deviation in MATLAB

Kavya Gupta

August 21, 2023

## 1 Instructions

All the MATLAB functions `UpdateMean, UpdateMedian, UpdateStd` are present in `Q7.m` file.
Also note that `IMG_4045.jpeg` and `IMG_4046.jpeg` in the zip are the solutions to UpdateStd derivation which is also attached in this pdf.

## 2 Introduction

In this report, we will derive and implement MATLAB functions to update the mean, median, and standard deviation of a dataset without recalculating them from scratch. This will allow us to efficiently incorporate new data points into our existing statistics.

## 3 Derivation: Updating Mean

Let's consider a dataset with $n$ data points: $x_1, x_2, \ldots, x_n$.

The mean of the dataset is calculated as:

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

We want to update the mean when adding a new data point $x_{new}$ to the dataset.

### 3.1 Step 1: Original Mean

The original mean of the dataset with $n$ data points is represented as:

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

We can also express the sum of all data points as:

$$oldSum = \sum_{i=1}^{n} x_i$$

Since $mean = \frac{oldSum}{n}$, we have:

$$oldSum = oldMean \cdot n$$

### 3.2 Step 2: Updating with New Data

When we add a new data point $x_{new}$ to the dataset, we need to update the mean. The updated sum can be calculated as:

$$newSum = oldSum + x_{new}$$

The updated mean, denoted as $newMean$, is then given by:

$$newMean = \frac{newSum}{n+1}$$

### 3.3 Step 3: Final Formula

Substituting the expression for $oldSum$ and $newSum$, we get:

$$newMean = \frac{(oldMean \cdot n) + x_{new}}{n+1}$$

This is the formula to update the mean when a new data point is added to the dataset.

## 4 Derivation: Updating Median

Let's consider a sorted dataset with $n$ unique data points: $A = [x_1, x_2, \ldots, x_n]$.

We want to update the median when adding a new data point $x_{new}$ to the dataset.

### 4.1 Step 1: Original Median

The original median of the dataset is the middle value of the sorted array $A$. For an odd number of data points ($n$ is odd):

$$oldMedian = A\left(\frac{n+1}{2}\right)$$

For an even number of data points ($n$ is even):

$$oldMedian = \frac{A\left(\frac{n}{2}\right) + A\left(\frac{n}{2}+1\right)}{2}$$

### 4.2 Step 2: Updating with New Data

When we add a new data point $x_{new}$ to the sorted array $A$, we need to update the median.

We insert $x_{new}$ into the array $A$ to get a new sorted array $sortedA$.

### 4.3 Step 3: Calculating the Updated Median

After inserting $x_{new}$ into the array, the total number of data points becomes $n+1$.

If $n+1$ is odd, the new median is the middle value of the updated array:

$$newMedian = sortedA\left(\frac{n+2}{2}\right)$$

If $n+1$ is even, the new median is the average of the two middle values in the updated array:

$$newMedian = \frac{sortedA\left(\frac{n+1}{2}\right) + sortedA\left(\frac{n+1}{2}+1\right)}{2}$$

## 5 Derivation : Update Standard Deviation

Handwritten Solutions given :-
`IMG_4045.jpeg` and `IMG_4046.jpeg` in the zip are same photos of my handwritten solution, which are also attached in this pdf.
**Final Formula :-**

$$newStd = \sqrt{\frac{(n-1)OldStd^2 + nOldMean^2 + NewDataValue^2 - (n+1)NewMean^2}{n}} \tag{1}$$

## 6 Histogram

We will check in which bin (interval) the `NewDataValue` lies in and then increase the frequency of that bin by 1.
To check in which bin the value will be in, we will compare it with lower and upper value of the bin.

# Updating Standard deviation

$$\sigma = OldStd = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

so $$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - oldMean)^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n}\left(x_i^2 + (\bar{x})^2 - 2(x_i.\bar{x})\right)}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + n \times \bar{x}^2 - 2 \times \sum_{i=1}^{n} x_i.\bar{x}}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}\bar{x}}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{n-1}$$

so $$\sum_{i=1}^{n} x_i^2 = (n-1)(OldStd)^2 + n(OldMean)^2$$

Now restart newStd, same formula as above

$$newStd^2 = \frac{\sum_{i=1}^{n+1} x_i^2 - (n+1)(newMean)^2}{n}$$

Cas

Now $\sum_{i=1}^{n+1} x_i^2 = \sum_{i=1}^{n} x_i^2 + (\text{newdata})^2$

$= (n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{newdata})^2$

so $(\text{newStd})^2 = \dfrac{(n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{newdata})^2 - (n+1)(\text{new Mean})^2}{n}$

so, Final Formula.

$$\text{newStd} = \sqrt{\dfrac{(n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{newdata})^2 - (n+1)(\text{new Mean})^2}{n}}$$