Assignment 1

170050025 – Yash Khemchandani

170070015- Anshul Nasery

Q1.

1. By definition

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

where σ is the standard deviation of $\{x_i\}_{i=1}^{n}$
& μ is the mean of $\{x_i\}_{i=1}^{n}$

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

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

For any $1 \le i \le n$ $\quad |x_i - \mu|^2 \le \sum_{i=1}^{n} |x_i - \mu|^2$

$\therefore\ \sigma^2(n-1) \ge |x_i - \mu|^2$ for all i

Since both sides are positive, we can take the root directly

$\therefore\ |x_i - \mu| \le \sigma\sqrt{n-1}$ for all i

Q2.

By definition,

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

Since $n-1 < n$

(i)— $$\sigma > \sqrt{\frac{\sum_{i=1}^{n}|x_i-\mu|^2}{n}} \geq \frac{\sum|x_i-\mu|}{n}$$ [Since Quadratic Mean $\geq$ Arithmetic Mean]

(ii)— $$\frac{\sum|x_i-\mu|}{n} \geq \frac{\sum|x_i-\tau|}{n}$$ [$\sum_{i=1}^{n}|x_i-y|$ is minimum ~~for~~ when y is median]

(iii)— $$\frac{\sum|x_i-\tau|}{n} \geq \left|\frac{\sum(x_i-\tau)}{n}\right|$$ [$|a_1|+|a_2|+\dots|a_n| \geq |a_1+a_2+\dots a_n|$]

$$= \left|\frac{\sum x_i}{n} - \tau\right|$$

$$= |\mu-\tau|$$

Hence $\boxed{\sigma \geq |\mu-\tau|}$

Q3.

3.

(a) $P(C_1/Z_1) = 1/3$

$P(C_2/Z_1) = 1/3$

$P(C_3/Z_1) = 1/3$

(b) $P(H_3/C_1Z_1) = 1/2$

$P(H_3/C_2Z_1) = 1$

$P(H_3/C_3Z_1) = 0$

(c) $\dfrac{P(H_3|C_2,Z_1) \cdot P(C_2,Z_1)}{P(H_3,Z_1)}$

$= \dfrac{1 \cdot 1/3}{1/2} = 2/3$

(d) $P(C_1|H_3,Z_1) = \dfrac{P(H_3|C_1,Z_1)\, P(C_1,Z_1)}{P(H_3,Z_1)}$

$= \dfrac{1/2 \times 1/3}{1/2} = 1/3$

(e) Since the probability of winning when switched > probability of winning when not, we can conclude that switching is indeed beneficial.

Q4.
f=0.6

err_median =
733.0434
err_mean =
359.5526
err_quartile =
80.9587

f=0.3

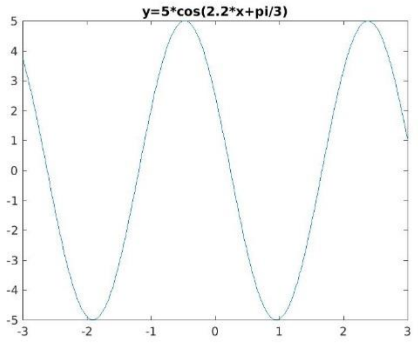
err_median =
22.0039
err_mean =
95.2669
err_quartile =
0.0173

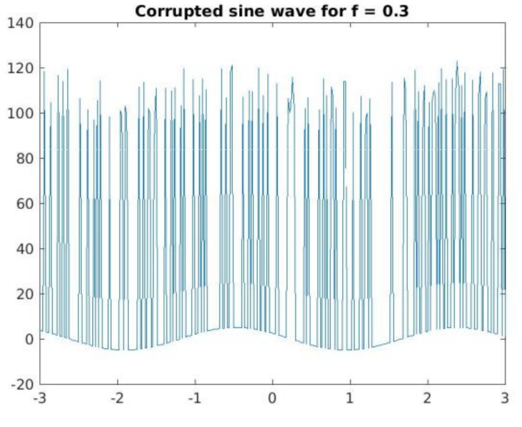
For both the cases, **the quartile provides the best filtering**, because in both the cases, on an average, more than 25% of the values are uncorrupted, hence the first quartile is more or less unchanged from the non-noisy data

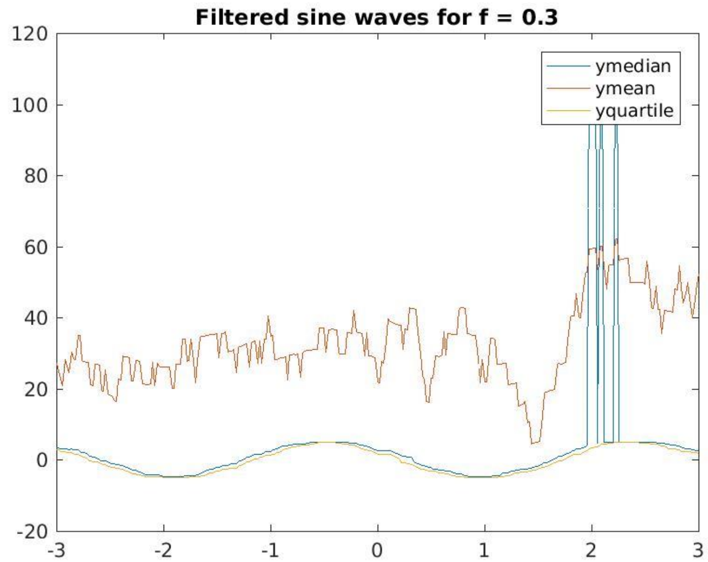
For a smaller corruption in values, the median provides better filtering than mean because the noise is in only 1 direction, and the median's robustness means that for most intervals, only some points are much above the actual value, and this assures the mid value to be approximately the same as before.
For the mean on the other hand, even the small number of extreme values wreaks havoc with the value.

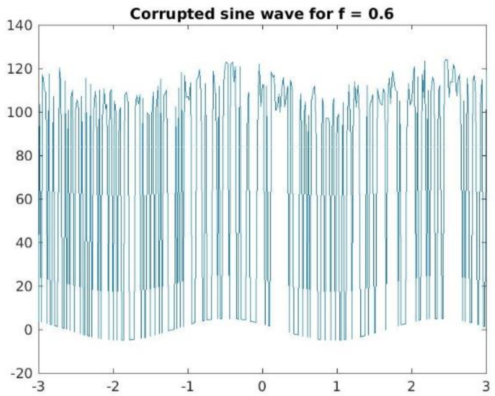
For a higher corruption, more intervals have more than half number of values being corrupted, making the median much bigger than the actual value of y, while the mean performs marginally better, since about 40% values are very small compared to the noise, which keeps the mean of the interval small.

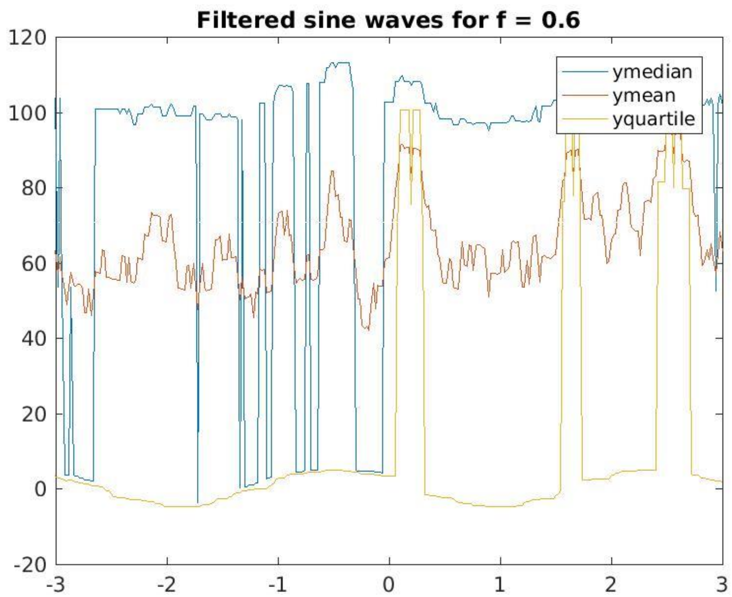
Plots-






Filtered sine waves for f = 0.3
ymedian
ymean
yquartile
120
100
80
60
40
20
0
-20
-3
-2
-1
0
1
2
3


Corrupted sine wave for f = 0.6
140
120
100
80
60
40
20
0
-20
-3
-2
-1
0
1
2
3


Filtered sine waves for f = 0.6
ymedian
ymean
yquartile
120
100
80
60
40
20
0
-20
-3
-2
-1
0
1
2
3

Q.5

$$\sigma^2_{new} = \frac{\sum_1^{n+1} x_i^2}{n} - \left(\frac{n+1}{n}\right)(NewMean)^2$$

$$\sigma^2_{old} = \frac{\sum_1^{n} x_i^2}{n-1} - \frac{n}{n-1}(OldMean)^2$$

$$\sum_1^n x_i^2 = (n-1)\left(\sigma^2_{old} + \left(\frac{n}{n-1}\right)(\mu_{old})^2\right)$$

$$\Rightarrow \sigma^2_{new} = \frac{(n-1)}{n}\left(\sigma^2_{old} + \left(\frac{n}{n-1}\right)\mu^2_{old}\right) + \frac{(NewData)^2}{n} - \left(\frac{n+1}{n}\right)(\mu^2_{new}$$

$$mean_{new} = \frac{\sum_1^{n+1} x_i}{n+1}$$

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

$$= \left(\frac{n}{n+1}\right) OldMean + \frac{NewDataValue}{n+1}$$

In order to update the histogram, one would need the bin size, and then add one value to the required bin, by taking $\lfloor (newDataValue - a_{min}/bin\_size \rfloor$ as the bin into the which the value goes. If the newDataValue is less than a_{min} a new bin will have to be created.

For the median, several cases exist, which have been coded in the function. These are-

1. If n is odd, and new value is greater than the next-to-median element, or lesser than previous-to-median, the new_median will be the mean of the old_median and next/previous element. If it is between the old_median and next element/previous element, the new_median is the mean of the old_median and the newDataValue.

2. If n is even, and the new value is less than a[n/2], median is a[n/2], or if it is greater than a[n/2 + 1] median is a[n/2 + 1], while if it is between these two, the median is newDataValue

In the programming assignment, the three files contain functions for updating mean, std and median, but driver code to test them, or for I/O is not written.

Q.6

Number of people = n

Let us take the case where none of the birthdays clash.

~~Assuming~~ taking the number of days in a year to be 365.

No. of days on which ~~first~~ Birthday of p_1 can fall $= 365$

$\exists$

No. of days for p_2's birthday $= 364$.

Similarly

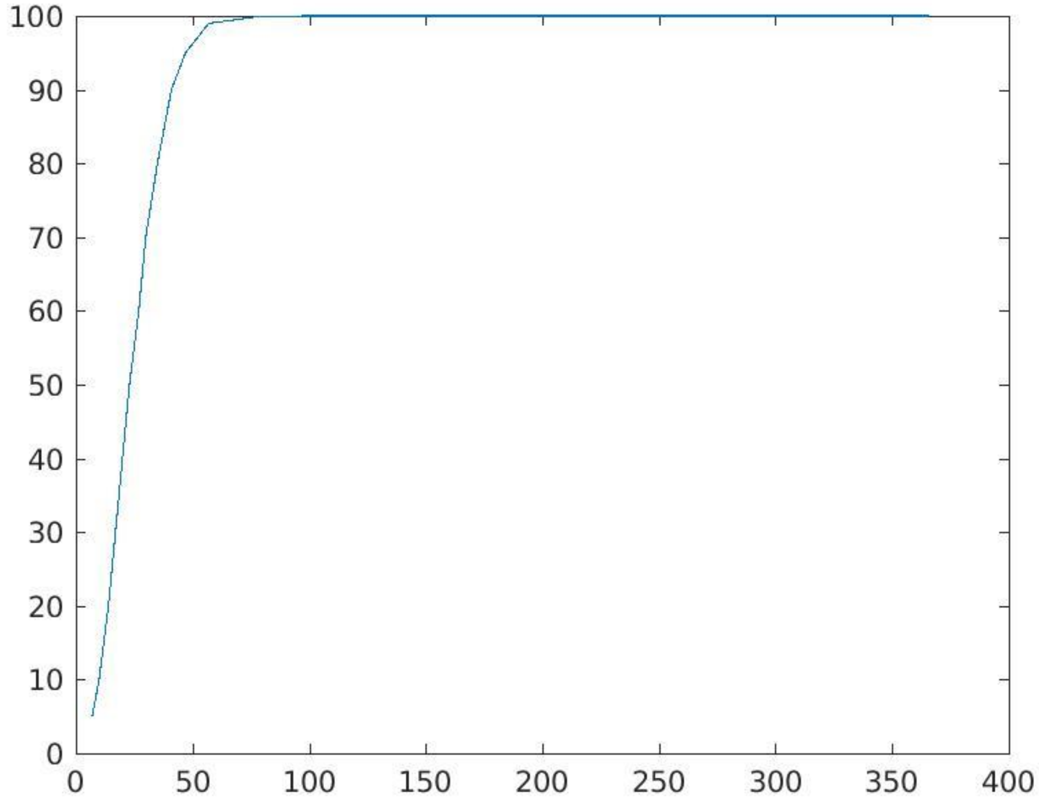
no. of days for p_n's birthday $= (365-n+1)$

$\therefore$ Probability that no birthday falls on same day

$$= \frac{365 \times 364 \times 363 \times \cdots (365-n+1)}{(365)^n}$$

Probability that atleast two birthdays coincide

$$= 1 - \left(\frac{365 \times 364 \times \cdots (365-n+1)}{(365)^n} \right)$$



Plot of probability (in percent) vs No. of people

The programming assignment has been done such that the file **Problem6.m** contains the main code for generating this plot, while the other two files contain helper functions.