CS 215 - Assignment 1

Niraj Mahajan 180050069

Raaghav Raaj 180050082

August 16, 2019

1 Question 1

Given:

$$\{x_i\}_{i=1}^{n} = \{x_1, x_2, x_3 ... x_n\}$$
$$Mean = \mu$$
$$StandardDeviation = \sigma$$

To Prove:

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

Proof:
We have that,

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

Now as we know from extended triangle inequality,

$$for \quad \sqrt{a^2 + b^2 + c^2 ...} = \alpha$$
$$|a| \leq \alpha$$
$$|b| \leq \alpha$$
$$|c| \leq \alpha$$
$$and \quad so \quad on...$$

Hence, for

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

$$\forall i \quad |x_i - \mu| \leq \sigma\sqrt{n-1} \tag{1}$$

The two sided Chebychev's Inequality says that,

$$\frac{|S_k|}{n} \leq \frac{1}{n-1} \quad S_k = \{x_i : |x_i - \mu| \geq \sigma\sqrt{n-1}\}$$

$$ie \quad |S_k| \leq 1 + \frac{1}{n-1}$$

Now as n increases, $1 + \frac{1}{n-1} \to 1$. This implies

$$\implies |S_k| \leq 1; \quad for \quad large \quad n$$

Hence, clearly, the first inequality is stronger than Chebychev's as it confirms the existence of no x_i, whereas, the Chebychev's inequality says that the number is less than or equal to 1.

2 Question 2

Given:

$$Mean = \mu$$

$$Median = \tau$$

$$StandardDeviation = \sigma$$

To Prove:

$$|\mu - \tau| \leq \sigma \tag{2}$$

Proof:
μ is given by

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

$$\implies |\mu - \tau| = \frac{1}{n}\left|\left(\sum_{x=1}^{n} x_i\right) - n\tau\right|$$

Now, by triangle inequality,

$$\frac{1}{n}\left|\left(\sum_{x=1}^{n} x_i\right) - n\tau\right| \leq \frac{1}{n}\sum_{i=1}^{n}|x_i - \tau|$$

Hence, we can concur,

$$|\mu - \tau| \leq \frac{1}{n}\sum_{i=1}^{n}|x_i - \tau|$$

We know that the median minimizes the sum of absolute differences.
Hence,

$$\sum_{i=1}^{n}|x_i - \tau| \leq \sum_{i=1}^{n}|x_i - \mu|$$

$$\implies |\mu - \tau| \leq \frac{1}{n}\sum_{i=1}^{n}|x_i - \mu|$$

Now, as $|x_i - \mu| \geq 0 \quad \forall i \in \{1, 2, 3...n\}$, we can use RMS-AM inequality.

$$\implies \frac{\sum_{i=1}^{n}|x_i - \mu|}{n} \leq \sqrt{\frac{\sum_{i=1}^{n}|x_i - \mu|^2}{n}}$$

$$\implies |\mu - \tau| \leq \sqrt{\frac{\sum_{i=1}^{n}|x_i - \mu|^2}{n}}$$

Now Clearly,

$$\sqrt{\frac{\sum_{i=1}^{n}|x_i-\mu|^2}{n}} \leq \sqrt{\frac{\sum_{i=1}^{n}|x_i-\mu|^2}{n-1}} = \sigma$$

And thus,

$$|\mu - \tau| \leq \sigma$$

Hence Proved!

3 Question 3

Number of rickshaws, $N_{red} = 1 \implies P(Red) = 0.01$
Number of rickshaws, $N_{blue} = 99 \implies P(Blue) = 0.99$
We have four different probabilities for person XYZ.

$$\begin{gathered}\text{P(seeing red as red)} = 0.99 = \text{P(RR)} \\ \text{P(seeing red as blue)} = 0.01 = \text{P(RR)} \\ \text{P(seeing blue as red)} = 0.02 = \text{P(RR)} \\ \text{P(seeing blue as blue)} = 0.98 = \text{P(RR)}\end{gathered}$$

We need to determine the probability of the auto being actually red when the person XYZ observed it Red.

Hence, as per Baye's Theorem,

$$P(actually\ red|observed\ red) = \frac{P(actually\ red\ \cap\ observed\ red)}{P(observed\ red)}$$

$$\implies P(actually\ red|observed\ red) = \frac{P(Red).P(RR)}{P(Blue).P(BR) + P(Red).P(RR)}$$

$$\implies P(actually\ red|observed\ red) = \frac{(0.01)(0.99)}{(0.99)(0.02) + (0.01)(0.99)} = \frac{1}{3}$$

Thus the probability for the rickshaw to be actually red is 1/3 whereas it is 2/3 for the rickshaw to be blue.

Hence the foremost important argument of the defense lawyer must be that the probability of the rickshaw being actually red is less than that of it being blue.

4 Question 4

We have

$$P(C_i) = \frac{1}{3} \quad , i \in \{1, 2, 3\}$$

Part A

Let Z_1 be the event that the contestant chose door 1

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

where $P(C_i \cap Z_1) = P(C_i).P(Z_1)$ as the probability of choosing door 1 is independent of the car being behind door i.
Hence,

$$P(C_i|Z_1) = P(Ci) = \frac{1}{3}$$

Part B

Let H3 be the event that the host opened door 3.

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

Now for i = 1,

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

(This is because if car is behind door 1, host can open door 2, 3 with equal probability)

Now for i = 2,

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

(This is because if door 1 is chosen and car is behind door 2, host will open door 3 only)

Now for i = 3,

$$P(H_3|C_3, Z_1) = \frac{P(H_3 \cap C_3, Z_1)}{P(C_3, Z_1)} = \frac{0.\frac{1}{3}.\frac{1}{3}}{\frac{1}{3}.\frac{1}{3}} = 0$$

(This is because if door 1 is chosen and car is behind door 3,host will never open the door 3)

Part C

The conditional probability of winning after switching is $P(C_2|H_3, Z_1)$

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

$$P(C_2|H_3, Z_1) = \frac{P(H_3|C_2, Z_1).P(C_2, Z_1)}{\sum_i [P(H_3|C_i, Z_1).P(C_i, Z_1)]}$$

$$P(C_2|H_3, Z_1) = \frac{\frac{1}{3}.\frac{1}{3}}{\frac{1}{9}.(1 + \frac{1}{2} + 0)} = \frac{2}{3}$$

Part D

The conditional probability of winning without switching is $P(C_1|H_3, Z_1)$

$$P(C_1|H_3, Z_1) = \frac{P(H_3|C_1, Z_1).P(C_1, Z_1)}{P(H_3, Z_1)}$$

$$P(C_1|H_3, Z_1) = \frac{P(H_3|C_1, Z_1).P(C_1, Z_1)}{\sum_i [P(H_3|C_i, Z_1).P(C_i, Z_1)]}$$

$$P(C_1|H_3, Z_1) = \frac{\frac{1}{2}.\frac{1}{9}}{\frac{1}{9}.(1 + \frac{1}{2} + 0)} = \frac{1}{3}$$

Part E

Clearly the probability of winning after switching is higher than without switching. Hence **it's always better to switch**.

Part F

In the case, where the host opens any of the two remaining doors, we have

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

$$P(H_3|C_i, Z_1) = \frac{\frac{1}{2}.\frac{1}{3}.\frac{1}{3}}{\frac{1}{3}.\frac{1}{3}} = \frac{1}{2}$$

This will be independent of the fact that the car is behind which door.

Now,

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

$$P(C_2|H_3, Z_1) = \frac{\frac{1}{2}.(\frac{1}{3}.\frac{1}{3})}{\frac{1}{9}.(\frac{1}{2}+\frac{1}{2}+\frac{1}{2})} = \frac{1}{3}$$

And,

$$P(C_1|H_3, Z_1) = \frac{P(H_3|C_1, Z_1).P(C_1, Z_1)}{P(H_3, Z_1)}$$

$$P(C_1|H_3, Z_1) = \frac{\frac{1}{2}.\frac{1}{9}}{\frac{1}{9}.(\frac{1}{2}+\frac{1}{2}+\frac{1}{2})} = \frac{1}{3}$$

Since the probability of winning after switching and not switching is same, it is not beneficial to change the choice

5 Question 5

A sine wave of the form $y = 5sin(2.2x + pi/3)$ is plotted, corrupted and filtered by three methods - moving median filtering, moving average filtering and moving quartile filtering and the **root mean squared errors** of the outputs of each method are :

- 30 percent data corrupted
    1. **Moving Median Filtering** = 28.121810
    2. **Moving Mean Filtering** = 99.607732
    3. **Moving Quartile Filtering** = 0.017173
- 60 percent data corrupted
    1. **Moving Median Filtering** = 692.400109
    2. **Moving Mean Filtering** = 351.490789
    3. **Moving Quartile Filtering** = 59.808615

The clean sine wave, corrupted wave, and the filtered waves using all three methods are plotted in Figure 1 and Figure 2.

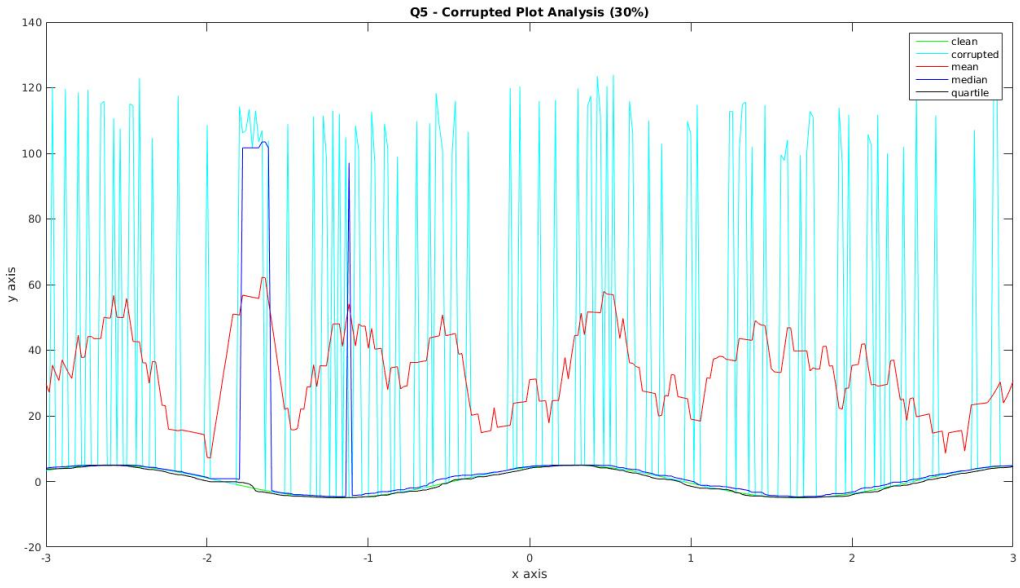

Figure 1: 30% data corrupted

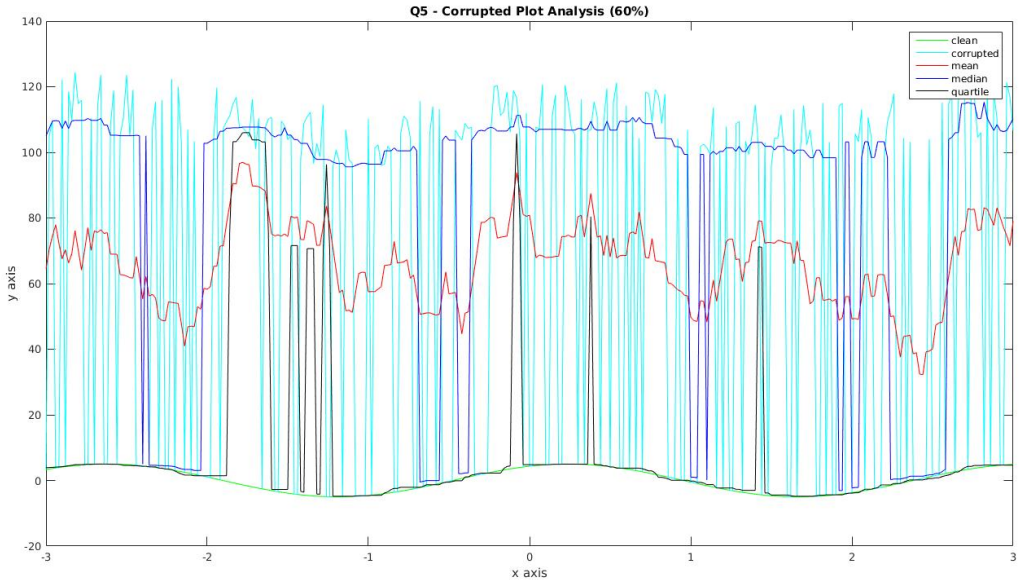

Figure 2: 60% data corrupted

From the plots and the root mean square error data, it is clearly evident that the **moving quartile filtering** method has the **least root mean square error** and, hence, is the **best** method of filtering corrupted data.

This is because since the data is increased by around 100 when it is corrupted, we need to consider the smallest values in any interval while filtering.

Mean of a data considers even the **corrupted values** and is heavily **influenced** by them. Hence we can safely rule out 'mean' as an optimum method for filtering.

Median of a data considers the middle term and is not influenced as heavily as mean by corruption but still, if **more than 50 %** of the data is corrupted then the median will also be a **corrupted value**, and so we can rule out median as well.

Quartile of a dataset focusses on the lower 25% data by value and we can say that it will **prioritize the clean (non corrupted) data** in any interval and hence will not get influenced by the corrupted data untill more than 75% data in an interval has been corrupted.

Hence moving quartile filtering is the best way to filter our data.

Usage of MATLAB Code

- Load code in the following path 'matlab_code/q5/q5.m'
- Run the code
- This should output all values of root mean squared error for all three methods, both for 30% corruption and 60% corruption.
- This should also generate two figures, one containing all the required plots for 30%corruption and the other for 60% corruption

6 Question 6

Formulae for Updating the mean, median and standard deviation are :

1. The **updated Mean** is given by the following formula.

$$Mean_{new} = \left(\frac{n.Mean_{old} + NewDataValue}{n+1}\right)$$

2. For the **Updated Median**, given the variables , $Median_{old}$, number of Data values(n), DataSet in increasing order, we consider the following cases.

<u>Case I : n is even</u>

If NewDataValue $\geq$ $Median_{old}$
  If NewDataValue $\geq$ $A_{\frac{n}{2}+1}$
    Median$_{new}$ = $A_{\frac{n}{2}+1}$
  Else
    Median$_{new}$ = NewDataValue
Else
  If NewDataValue $\geq$ $A_{\frac{n}{2}}$
    Median$_{new}$ = NewDataValue
  Else
    Median$_{new}$ = $A_{\frac{n}{2}}$

<u>Case II : n is odd</u>

If NewDataValue $\geq$ $Median_{old}$
  If NewDataValue $\geq$ $A_{\frac{n+1}{2}+1}$
    Median$_{new}$ = $\left(\frac{A_{\frac{n+1}{2}}+A_{\frac{n+1}{2}+1}}{2}\right)$
  Else
    Median$_{new}$ = $\left(\frac{A_{\frac{n+1}{2}}+NewDataValue}{2}\right)$
Else
  If NewDataValue $\leq$ $A_{\frac{n+1}{2}-1}$
    Median$_{new}$ = $\left(\frac{A_{\frac{n+1}{2}}+A_{\frac{n+1}{2}11}}{2}\right)$
  Else
    Median$_{new}$ = $\left(\frac{A_{\frac{n+1}{2}}+NewDataValue}{2}\right)$

3. The updated **Standard Deviation** is given by the following formula.

$$StD_{new} = \sqrt{\left(\frac{n(Mean_{old} - Mean_{new})^2 + (n-1)StD_{old}^2 + (NewDataValue - Mean_{old})^2}{n}\right)}$$

Proofs

1. # Mean

$$Mean_{old} = \left(\frac{\sum_i a_i}{n}\right) \tag{3}$$

Similarly

$$Mean_{new} = \left(\frac{\sum_i a_i + NewDataValue}{n+1}\right) \tag{4}$$

Substituting value of $\sum_i a_i$ in eqn(4)

$$Mean_{new} = \left(\frac{n.Mean_{old} + NewDataValue}{n+1}\right) \tag{5}$$

2. # Median

$$Median_{old} = \left\{ \begin{array}{ll} \left(\frac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2}\right) & ;2|n \\ a_{\frac{n+1}{2}} & ;otherwise \end{array} \right\}$$

Now lets analyse all cases

- If the new value is equal to $Median_{old}$ then the new Median will be equal to $Median_{old}$ since Median is the 'Middle Element' and effectively, we are not adding any element greater or less than the median, so it will remain unchanged.
- Lets now see for **even n**
  - On adding an element, the number of elements will become odd, and hence the new median will be unique.
  - The NewDataValue can either lie between the values of $a_{\frac{n}{2}}$ and $a_{\frac{n}{2}+1}$, or lie outside this range.
  - If the NewDataValue lies between $a_{\frac{n}{2}}$ and $a_{\frac{n}{2}+1}$ then we can concur that in the new sequence (let it be b), the NewDataPoint will be $b_{\frac{n}{2}+1}$ or $b_{\frac{n+1+1}{2}}$ which is the middle term, and hence the median.
  - If the NewDataValue lies outside this range, then either of $a_{\frac{n}{2}}$ or $a_{\frac{n}{2}+1}$ will be the new median. Subsequently, if the new value is greater than the $Median_{old}$, then, $a_{\frac{n}{2}+1}$ will be the new Median, else, $a_{\frac{n}{2}}$ will be the new Median.
- Lets now see for **odd n**
  - On adding an element, the number of elements will become even, and hence the new median will not be unique. (Lets say average of middle two terms)
  - The NewDataValue can either lie between the values of $a_{\frac{n}{2}+1}$ and $a_{\frac{n}{2}-1}$, or lie outside this range.

- If the NewDataValue lies between $a_{\frac{n}{2}+1}$ and $a_{\frac{n}{2}-1}$, then the two middle values will comprise of the $Median_{old}$ and the NewDataValue itself, and hence the $Median_{new}$ will be an average of these two.
- If the NewDataValue lies outside this range, then the two middle values will comprise of the $Median_{old}$ and either of $a_{\frac{n}{2}+1}$ or $a_{\frac{n}{2}-1}$. Now if the NewDataValue is greater than $Median_{old}$ then the median will be the mean of $Median_{old}$ and $a_{\frac{n}{2}+1}$, else the $Median_{new}$ will be the mean of $a_{\frac{n}{2}-1}$ and $Median_{old}$.

3. Standard Deviation

$$StD_{old} = \left(\sqrt{\frac{\sum_i (a_i - Mean_{old})^2}{n-1}}\right) \tag{6}$$

Similarly

$$StD_{new} = \left(\sqrt{\frac{[\sum_i (a_i - Mean_{new})^2] + (NewDataValue - Mean_{new})^2}{n}}\right) \tag{7}$$

Each term $(a_i - Mean_{new})$ can be written as $((a_i - Mean_{old}) + (Mean_{old} - Mean_{new}))$

$$StD_{new} = \left(\sqrt{\frac{[\sum_i ((a_i - Mean_{old}) + (Mean_{old} - Mean_{new}))^2] + (NewDataValue - Mean_{new})^2}{n}}\right) \tag{8}$$

We can simplify $\sum_i ((a_i - Mean_{old}) + (Mean_{old} - Mean_{new}))^2$ as

$$= \sum_i [(a_i - Mean_{old})^2 + (Mean_{old} - Mean_{new})^2 + 2(Mean_{old} - Mean_{new})(a_i - Mean_{old})]$$

$$= \sum_i [(a_i - Mean_{old})^2] + \sum_i [(Mean_{old} - Mean_{new})^2] + \sum_i [2(Mean_{old} - Mean_{new})(a_i - Mean_{old})]$$

$$= \sum_i [(a_i - Mean_{old})^2] + n.(Mean_{old} - Mean_{new})^2 + \Big\{2(Mean_{old} - Mean_{new}) \sum [(a_i - Mean_{old})]\Big\}$$

Now since $\sum [(a_i - Mean_{old})] = 0$, we get

$$= \sum_i [(a_i - Mean_{old})^2] + n.(Mean_{old} - Mean_{new})^2$$

Substituting eqn(6), we get

$$= (n-1) StD_{old}^2 + n.(Mean_{old} - Mean_{new})^2 \tag{9}$$

Substituting eqn(9) in eqn(8), we get,

$$StD_{new} = \left(\sqrt{\frac{(n-1) StD_{old}^2 + n(Mean_{old} - Mean_{new})^2 + (NewDataValue - Mean_{new})^2}{n}}\right) \tag{10}$$

Hence Proved!

Usage of MATLAB Code

- Load code in the following path 'matlab_code/q6/'
- There should be three different matlab files corresponding to the required three fucntions
- This is done since MATLAB versions ¡2016 do not support function declarations in a single script file
- Load and run every function in the command line window for testing

Updating the Histogram

Let the frequency be plotted on the y axis.
Then increase the value of frequency corresponding to the new data value by 1.

7 Question 7

Formula
The number of people 'n' such that any two of them have the same birthday with atleast 'p' probability is given by the inequality

$$\left(1 - \frac{^{365}C_n}{365^n}\right) \leq \frac{p}{100}$$

The situation can be visualized as the negation of the event where all n people have unique birthdays.
The plot of values of n vs p is given in Figure 3

$$\forall\ \mathrm{p} \in \{5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 95, 99, 99.99, 99.9999, 100\}$$

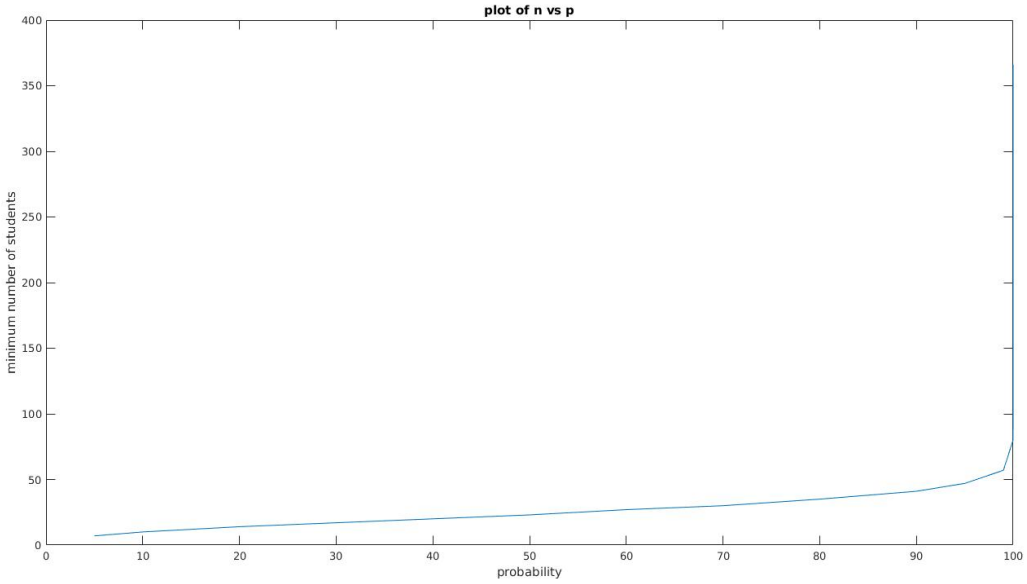

Figure 3: plot of n vs p

Usage of the MATLAB Code

- Load code in the following path 'matlab_code/q5/q5.m'
- Run the code
- This should output all values of n for corresponding values of p
- This should also plot the required plot of n vs p