ASSIGNMENT 1

CS 215

Details of students in the group for the assignment :

1.) Name : TATHAGAT VERMA Roll Number : 180050111.

2.) Name : NEEL ARYAN GUPTA Roll Number : 180050067.

ANSWERS : Follow from next page.

- Instructions for running the MATLAB code is specified in the INSTRUCIONS.txt file.

Question - 1

Given n distinct values $\{x_i\}_{i=1}^{n}$

$\therefore (x_1-\mu)^2 + (x_2-\mu)^2 \ldots + (x_{k-1}-\mu)^2 + (x_{k+1}-\mu)^2 \ldots$
$\ldots + (x_n-\mu)^2 \geq 0$

$\Rightarrow \sum(x_i-\mu)^2 \geq (x_k-\mu)^2$

$\therefore \sigma^2 \text{(variance)} = \dfrac{\sum\limits^{n}(x_i-\mu)^2}{n-1}$

$\Rightarrow \sigma^2(n-1) = \sum\limits^{n}(x_i-\mu)^2 \geq (x_k-\mu)^2$

$\Rightarrow \sigma^2(n-1) \geq (x_k-\mu)^2$

$\Rightarrow \sigma\sqrt{n-1} \geq |x_k-\mu|$

Chebyshev's inequality: $P(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$

$\Rightarrow P(|X-\mu| \leq k\sigma) \geq 1 - \dfrac{1}{k^2}$

Put $k = \sqrt{n-1}$

$\Rightarrow P(|X-\mu| \leq \sigma\sqrt{n-1}) \geq 1 - \dfrac{1}{n-1}$

As n increases, the probability will approach the value 1 (the probability should be 1 but isn't)

$\Rightarrow$ Chebyshev's inequality agrees with the given inequality only for large n

This means that Chebyshev's inequality gives correct bounds but the bounds are very loose.

Question - 2

μ = mean

τ = median

σ = standard deviation

To prove: $|\mu - \tau| \le \sigma$

By Chebyshev-Cantelli inequality,

$$P(X-\mu \ge k\sigma) \le \frac{1}{1+k^2}$$

Also, $P(X-\mu \le -k\sigma) \le \dfrac{1}{1+k^2}$

Put $k=1$

$\Rightarrow P(X-\mu \ge \sigma) \le \frac{1}{2}$ & $P(X-\mu \le -\sigma) \le \frac{1}{2}$

$\therefore \frac{1}{2} = P(X \ge \tau) = P(X \le \tau)$

$\Rightarrow P(X \ge \sigma+\mu) \le P(X \ge \tau)$ & $P(X \le \mu-\sigma) \le P(X \le \tau)$

Since $P(X \ge \tau)$ is greater than $P(X \ge \sigma+\mu)$, it means that ~~τ must be as greater bound than $\sigma+\mu$~~ values greater than τ are more ~~than~~ in number than values greater than $\sigma+\mu$.

$\Rightarrow \tau \le \sigma+\mu \Rightarrow \tau-\mu \le \sigma$

Similarly from 2^{nd} inequality, we have

$\tau \ge -\sigma+\mu$

$\Rightarrow \tau-\mu \ge -\sigma$

So, $-\sigma \le \tau-\mu \le \sigma$

$\Rightarrow |\tau-\mu| \le \sigma$

Hence, proved.

3.] <u>Given</u>: There exist 100 rickshaws of which 1 is red and 99 are blue.

XYZ sees red objects as red 99% of the time & blue objects as red 2% of the time.

∴ P(rickshaw was really red | XYZ observed it to be red) {using ①}

$$= \frac{P(\text{rickshaw was red, XYZ observed it to be red})}{P(\text{XYZ observed it to be red})}$$ {Refer NOTE}

$$= \frac{P(\text{rickshaw was red}) \times P(\text{XYZ observed red rickshaw as red})}{P(\text{XYZ observed a rickshaw to be red})}$$

$$= \frac{1/100 \times 99/100}{P\left(\begin{array}{c}\text{rickshaw was blue} \\ \text{\& XYZ observed red}\end{array}\right) + P\left(\begin{array}{c}\text{rickshaw was red} \\ \text{\& XYZ observed red}\end{array}\right)}$$

{Refer NOTE}

$$= \frac{1/100 \times 99/100}{P(\text{rickshaw was blue}) \times P\left(\begin{array}{c}\text{XYZ observed} \\ \text{blue object} \\ \text{as red}\end{array}\right) + P\left(\begin{array}{c}\text{rickshaw} \\ \text{was red}\end{array}\right) \times P\left(\begin{array}{c}\text{XYZ} \\ \text{observed} \\ \text{red as} \\ \text{red}\end{array}\right)}$$

$$= \frac{1/100 \times 99/100}{99/100 \times 2/100 + 1/100 \times 99/100} = \frac{99}{99 \times 2 + 99} = \boxed{1/3}$$

Answer: probability that the rickshaw was really a red when XYZ observed it to be a red one is $\underline{1/3}$.

used in solution

① $P(X|Y) = \dfrac{P(X,Y)}{P(Y)}$.

NOTE: we also have considered that the color of the rickshaw is independent of the probability of XYZ seeing red objects as red or blue objects as red.
Hence we used $P(X,Y) = P(X)P(Y)$.

4.] (a) $P(C_i|Z_1) = \dfrac{P(C_i, Z_1)}{P(Z_1)}$.

$= \dfrac{P(C_i)\,P(Z_1)}{P(Z_1)} = \dfrac{1/3 \times \cancel{1/3}}{\cancel{1/3}}$ for all $i = 1,2,3$

~~as~~ ($P(C_i, Z_1) = P(C_i) \times P(Z_1)$ as the car being ~~being~~ behind a particular door is independent of the door being chosen first)

$P(C_i|Z_1) = \boxed{1/3}$ for $i = 1,2,3$. [ANS].

(b.) $P(H_3 | C_i, Z_1) = \dfrac{P(H_3, C_i, Z_1)}{P(C_i, Z_1)}$

$i=1$; $P(H_3 | C_1, Z_1) = \dfrac{P(H_3, C_1, Z_1)}{P(C_1, Z_1)} = \dfrac{\cancel{1/3} \times \cancel{1/3} \times 1/2}{\cancel{1/3} \times \cancel{1/3}}$

(as the host can open door 2 or 3 with equal probability)
& $P(C_1) = 1/3$.
$P(Z_1) = 1/3$

$i=1.$ $\Rightarrow$ $\boxed{P(H_3 | C_1, Z_1) = 1/2}$

$i=2$; $P(H_3 | C_2, Z_1) = \dfrac{P(H_3, C_2, Z_1)}{P(C_2, Z_1)} = \dfrac{1/3 \times 1/3 \times 1}{1/3 \times 1/3} = \underline{1}$.

(as when car is behind door 2 & contestant chose door 1, the host will open door 3 with probability 1).
& $P(C_2) = P(Z_1) = 1/3$.

$i=2.$ $\Rightarrow$ $\boxed{P(H_3 | C_2, Z_1) = 1}$

$i=3$; $\boxed{P(H_3 | C_3, Z_1) = 0}$ (As when the car is behind door 3, the host will NOT open door 3).

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

$= \dfrac{1 \times 1/9}{P(H_3, C_1, Z_1) + P(H_3, C_2, Z_1) + P(H_3, C_3, Z_1)}$ [as C_1, C_2, C_3 are mutually exclusive & exhaustive]

$= \dfrac{1/9}{\sum_{1}^{3} P(H_3 | C_i, Z_1) \times P(C_i, Z_1)}$

Scanned by CamScanner

$$= \frac{1/9}{1/9 \times 1/2 + 1/9 \times 1 + 1/9 \times 0} = \frac{\cancel{1/9}}{\cancel{1/9} \times 3/2}$$

$$= 2/3$$

$$\Rightarrow \boxed{P(C_2 \mid H_3, Z_1) = 2/3}$$

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

$$= \frac{(1/2) \times (1/9)}{1/9 \times 1/2 + 1/9 \times 1 + 1/9 \times 0}$$

{ same explanation as that of part (C) }

$$= \frac{1/2 \times \cancel{1/9}}{3/2 \times \cancel{1/9}} = 1/3$$

$$\Rightarrow \boxed{P(C_1 \mid H_3, Z_1) = 1/3}$$

(e.) Since probability of winning by switching i.e. $P(C_2 \mid H_3, Z_1)$ $= 2/3$ $>$ probability of winning by not switching i.e. $P(C_1 \mid H_3, Z_1) = 1/3$, we conclude that switching is indeed beneficial.

(f.) Repeating calculations when the host chooses to open doors with equal probability.

(i.) $P(C_i|Z_1) = \underline{1/3}$ for $i = 1, 2, 3$.

remains unchanged as this doesn't depend on which door the host opens.

(ii) $P(H_3|C_i, Z_1) = 1/2$ for $i = 1, 2, 3$

as no-matter where the car is, the host will ~~choose~~ open any one of doors 2 & 3, with equal probability i.e. $\underline{1/2}$.

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

$$= \frac{1/2 \times 1/3 \times 1/3}{\sum_1^3 P(H_3|C_i, Z_1)\, P(C_i, Z_1)}$$

{~~sim~~ same logic as part (c)}

$$= \frac{1/2 \times 1/9}{1/2 \times 1/9 \times 3} = 1/3$$

$$\Rightarrow \boxed{P(C_2|H_3, Z_1) = 1/3}$$

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

$$= \frac{\cancel{1/2} \times \cancel{1/9}}{\cancel{1/2} \times \cancel{1/9} \times 3}$$

{similar calculations as that of previous part}

$$= 1/3$$

$$\Rightarrow \boxed{P(C_1|H_3, Z_1) = 1/3}$$

Since probability of winning with & without switching is the same i.e. $P(C_2|H_3, Z_1) = P(C_1|H_3, Z_1) = 1/3$, It is <u>NOT</u> <u>Beneficial</u> to switch. (as chances to win remains unchanged).

5.] For $f = 30\%$, Relative mean squared error with:

Median $= 2.99 \times 10^1$.

Mean $= 9.22 \times 10^1$

1st Quartile $= 1.177 \times 10^{-2}$.

For $f = 60\%$, Relative mean squared error with:

Median $= 6.38 \times 10^2$

Mean $= 3.54 \times 10^2$

1st Quartile $= 1.05 \times 10^2$.

<u>NOTE</u>: These error values are for one iteration (one time), with every iteration, different random values and positions will be generated leading to different error values. However the range would be around these values.

Now; For <u>$f = 30\%$ and $f = 60\%$</u>, We can observe Relative Mean Squared Error to be much less for 1st Quartile than for Mean & Median. <u>Hence quartile filtering produces best relative mean squared error</u>.

This can be justified as;

The error caused is mainly due to addition of random values in the 100–120 range. (Taking mean/median/quartile also generates error but this is magnified only due to the large random values).

• In mean filtering, every ~~rand~~ random value added is ~~co~~ taken into account,

while for median, the random value would be accounted for only if number of random values added in a particular interval of $i-8$ to $i+8$ was >8. (This has less probability)

Eg. original values of an interval are (we consider length 5 subarray)

$y = 0.10 \quad 0.15 \quad 0.2 \quad 0.17 \quad 0.12$

If only 1 random value was added;

mean would change, median & 25% quantile would remain unchanged.

• Median would ~~can~~ change only if number of random values added was $>\underline{2}$. (As values added are greater than original values of y.)

• 25% quantile would change only if number of random values added was $> \lfloor 75\% \times (\text{length of subarray}) \rfloor$

$$= \left\lfloor \frac{3}{4} \times 5 \right\rfloor = \lfloor 3.75 \rfloor = \underline{3}.$$

• Due to this reason, 1st quartile produces least mean squared error.

• For $f = 60\%$, we observe Relative Mean squared error to be least for 25% quartile, then <u>Mean</u>, then Median.

In this case, since 60% (<u>many</u>) random values are added, in many intervals of $i-8$ to $i+8$, there would be >8 random values added due to which median went in the range of 100-120 causing lot of error, while the mean in this case uses the un-corrupted values as well to reduce the filtered value to some extent. Due to this relative mean squared error is less for mean than for median.

In this case, still 1st quartile performs best due to the same reason stated before (i.e. $f = 30\%$), but we can observe the extent to which 1st quartile performs better than the mean & median has reduced. The reason for this is same as the reason why median has started performing poorer than the mean.

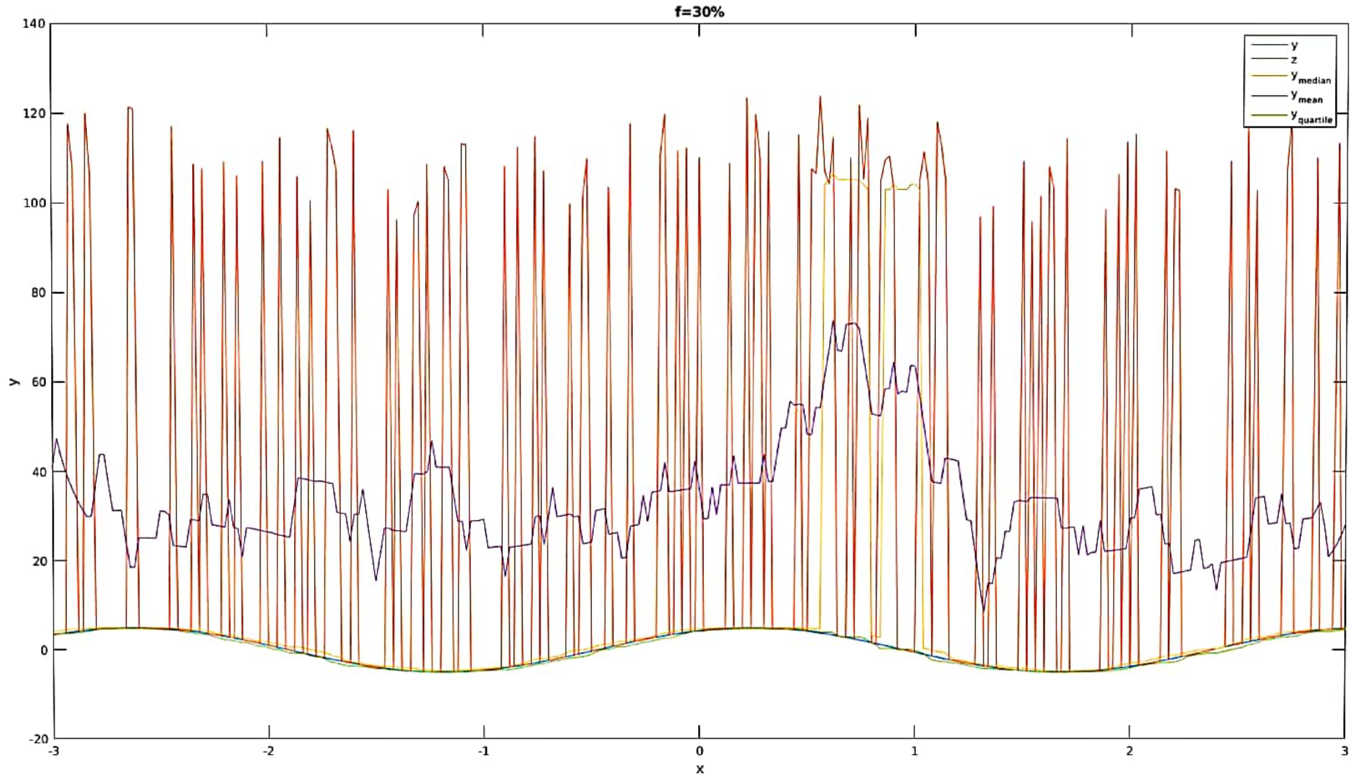
- For $f = 30\%$, what we can see is the robustness of median, and more robustness of 1st quartile.

<u>Instructions to execute</u>: Go in the directory of the zip folder & execute :-

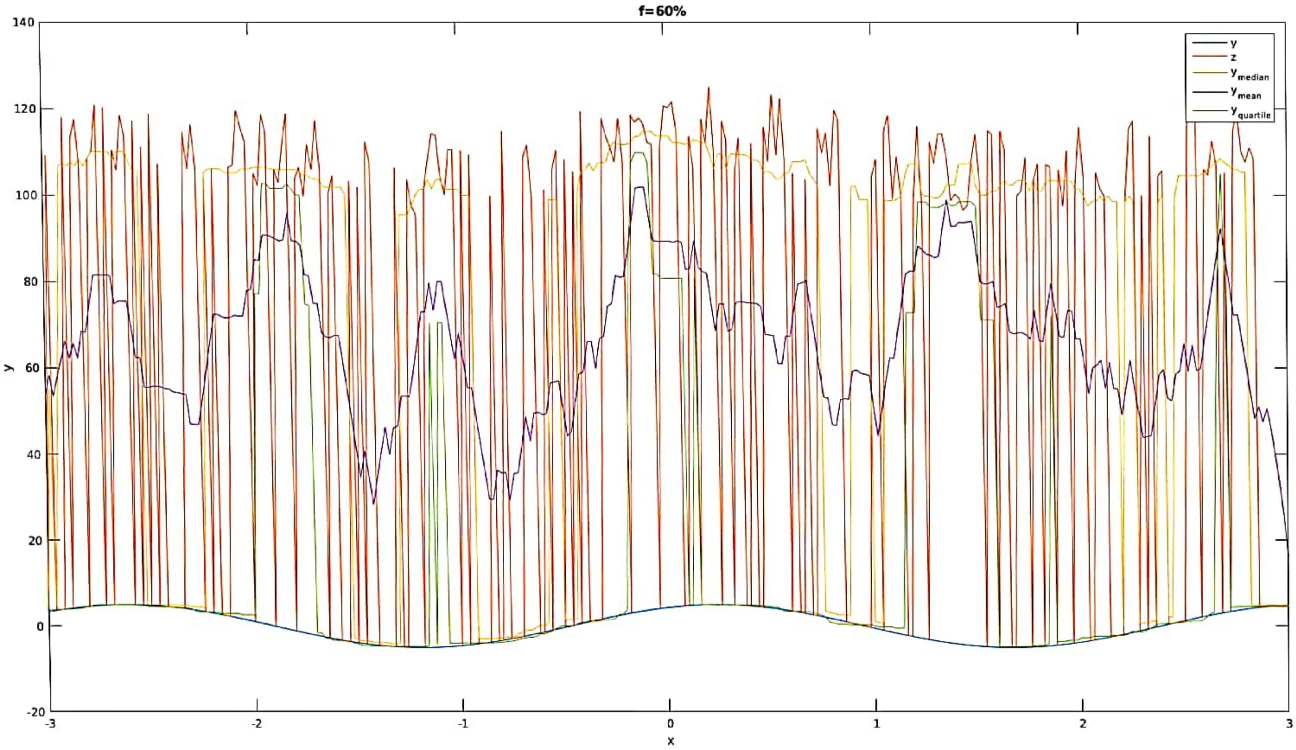
run('a1q5.m');

in the command window.

- Graphs are supplied on the next pages.



f=30%
y
z
y_{median}
y_{mean}
$y_{quartile}$
140
120
100
80
60
40
20
0
-20
-3
-2
-1
0
1
2
3
x
y


f=60%
140
120
100
80
60
40
20
0
-20
y
-3
-2
-1
0
1
2
3
x
y
z
y_{median}
y_{mean}
$y_{quartile}$

Question - 6 REPORT

Let $\{x_i\}_{i=1}^{n}$ denote the old array elements and x be the new element

→ Updating the mean

$$\mu_{old} = \frac{\sum x_i}{n} \Rightarrow \sum x_i = n\mu_{old}$$

$$\mu_{new} = \frac{\sum x_i + x}{n+1} = \frac{n\mu_{old} + x}{n+1}$$

→ Updating the median

Case 1: n is odd

$x_1, x_2 \text{ ------ } x_{k-1}, x_k, x_{k+1} \text{ ------ } x_n$

be the sorted array, without loss of generality

$\therefore k = (n+1)/2$, $M_{old} = x_k$, since $n+1$ is even,

If $x \geq x_{k+1}$ then $M_{new} = (x_k + x_{k+1})/2$

If $x < x_{k-1}$ then $M_{new} = (x_k + x_{k-1})/2$

else $M_{new} = (x_k + x)/2$ $[x_{k-1} \leq x \leq x_{k+1}]$

Case 2: n is even

$x_1, x_2 \text{ ------ } x_k, x_{k+1} \text{ ------ } x_n$

be the sorted array, WLOG

$\therefore k = n/2$, $M_{old} = (x_k + x_{k+1})/2$, since $n+1$ is odd

If $x \geq x_{k+1}$ then $M_{new} = x_{k+1}$

If $x \leq x_k$ then $M_{new} = x_k$

else $M_{new} = x$ $[x_k < x < x_{k+1}]$

→ Updating the standard deviation

Let $\mu_{new} - \mu_{old} = d$

i.e. $\mu_{new} = \mu_{old} + d$

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

$$\Rightarrow \sum (x_i - \mu_{old})^2 = (n-1)\sigma_{old}^2$$

$$\therefore\ (x_i - \mu_{new})^2 = (x_i - \mu_{old} - d)^2$$

$$\sum^n (x_i - \mu_{new})^2 = \sum^n (x_i - \mu_{old} - d)^2$$

$$= \sum^n (x_i - \mu_{old})^2 + \sum^n d^2 - 2d \sum^n (x_i - \mu_{old})$$

Also $\sum^n x_i = n\mu_{old} = \sum^n \mu_{old}$

$$\Rightarrow \sum^n (x_i - \mu_{old}) = 0$$

$$\Rightarrow \sum^n (x_i - \mu_{new})^2 = \sum^n (x_i - \mu_{old})^2 + nd^2 - 0$$

$$= (n-1)\sigma_{old}^2 + nd^2$$

$$\sigma_{new}^2 = \frac{\sum^n (x_i - \mu_{new})^2 + (x - \mu_{new})^2}{n}$$

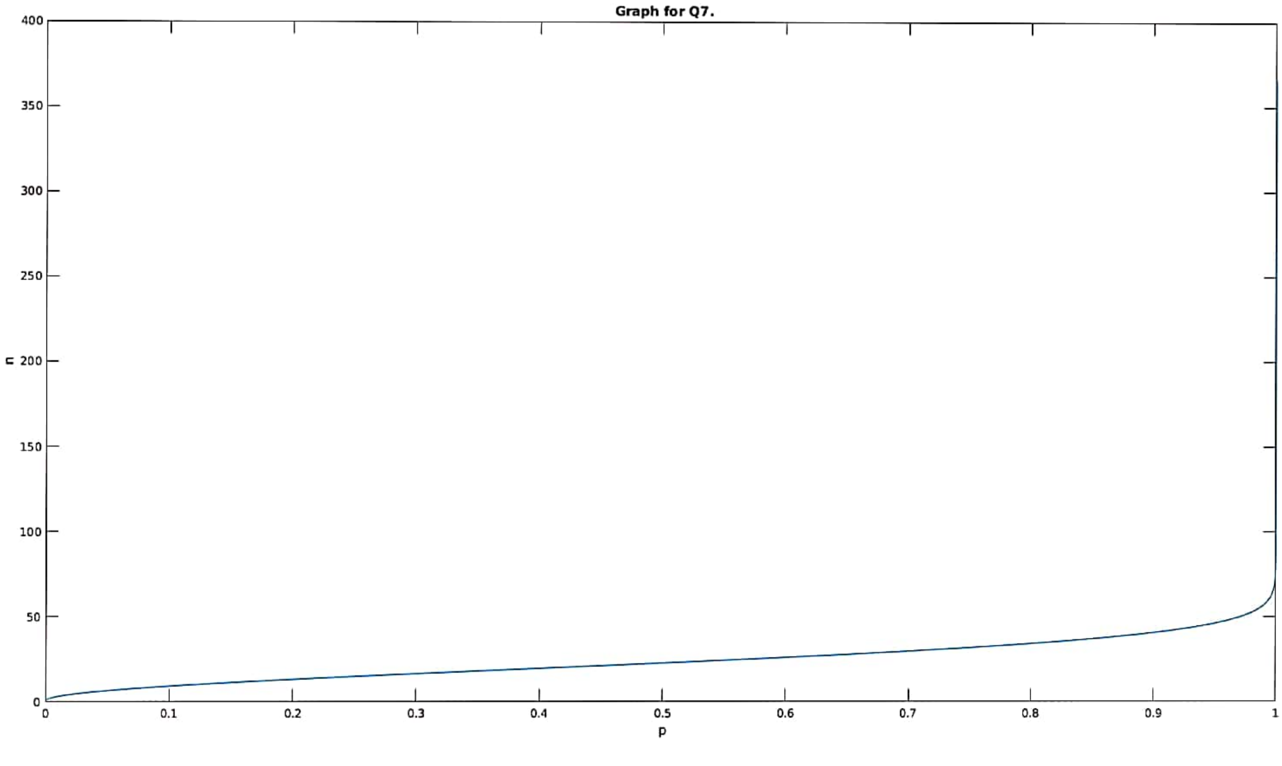
$$= \frac{(n-1)\sigma_{old}^2 + nd^2 + (x - \mu_{new})^2}{n}$$

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

→ Updating the histogram

We will just increment the value of the bin by 1 in which the new value/element lies.

Scanned by CamScanner


Graph for Q7.
n
p
0
50
100
150
200
250
300
350
400
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1