CS 215 Assignment 1

Nama N V S S Hari Krishna - 170050077 Srikakulapu Rohan Abhishek - 170050078

Question 1

Given : n distinct values $\{x_i\}_{i=1}^n$ with mean μ and standard deviation σ.
To prove : For all i,

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

Proof:

$$\because x^2 \leq x^2 + y^2 \ \forall x, y \in \mathbf{R}$$

$\therefore$ For any j = 1 to n

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

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

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

$$\frac{\sqrt{(x_j - \mu)^2}}{\sqrt{n-1}} \leq \sigma$$

$$|x_j - \mu| \leq \sigma\sqrt{n-1}$$

$\therefore$ For any $i = 1$ to n ,

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

Hence, Proved.

Question 2

Given : n values $\{x_i\}_{i=1}^n$ having mean μ, median τ and standard deviation σ.
To prove :

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

Proof : We know that

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$\left|\frac{x_1 + x_2 + x_3 + ... + x_n}{n} - \tau\right| \leq \sigma$$

$$|x_1 + x_2 + x_3 + ... + x_n - n\tau| \leq n\sigma$$

$$|(x_1 - \tau) + (x_2 - \tau) + (x_3 - \tau) + ... + (x_n - \tau)| \leq n\sigma$$

Let us consider two cases if n is odd and n is even.
Say,if n is **odd**.

$$|(x_1 - \tau) + (x_2 - \tau) + (x_3 - \tau) + ... + (x_n - \tau)| \leq \sum_{i=1}^{n} |x_i - \tau|$$

based on definition of median

$$\leq (\tau - x_1) + (\tau - x_2) + ... + (\tau - x_{\frac{n-1}{2}}) + 0 + (x_{\frac{n+3}{2}} - \tau) + (x_{\frac{n+5}{2}} - \tau) + ... + (x_n - \tau)$$

all the τ will be cancelled,now add and subtract μ in place of each τ

$$\leq (\mu - x_1) + (\mu - x_2) + ... + (\mu - x_{\frac{n-1}{2}}) + 0 + (x_{\frac{n+3}{2}} - \mu) + (x_{\frac{n+5}{2}} - \mu) + ... + (x_n - \mu)$$

$$\leq |\mu - x_1| + |\mu - x_2| + ... + |\mu - x_{\frac{n-1}{2}}| + 0 + |\mu - x_{\frac{n+3}{2}}| + |\mu - x_{\frac{n+5}{2}}| + ... + |\mu - x_n|$$

On applying AM $\leq$ RMS

$$\leq n\sqrt{\frac{(\mu - x_1)^2 + (\mu - x_2)^2 + ... + (\mu - x_{\frac{n-1}{2}})^2 + 0 + (x_{\frac{n+3}{2}} - \mu)^2 + (x_{\frac{n+5}{2}} - \mu)^2 + ... + (x_n - \mu)^2}{n}}$$

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

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

$$\leq \sqrt{n}\sqrt{n-1}\sigma$$

Since n-1 is less than n

$$\leq n\sigma$$

Hence, the given inequality is true if n is odd
Similarly repeat this, if n is **even**

$$|(x_1 - \tau) + (x_2 - \tau) + (x_3 - \tau) + ... + (x_n - \tau)| \leq \sum_{i=1}^{n} |x_i - \tau|$$

based on definition of median

$$\leq (\tau - x_1) + (\tau - x_2) + ... + (\tau - x_{\frac{n}{2}}) + (x_{\frac{n+2}{2}} - \tau) + (x_{\frac{n+4}{2}} - \tau) + ... + (x_n - \tau)$$

all the τ will be cancelled,now add and subtract μ in place of each τ

$$\leq (\mu - x_1) + (\mu - x_2) + ... + (\mu - x_{\frac{n}{2}}) + (x_{\frac{n+2}{2}} - \mu) + (x_{\frac{n+4}{2}} - \mu) + ... + (x_n - \mu)$$

$$\leq |\mu - x_1| + |\mu - x_2| + ... + |\mu - x_{\frac{n}{2}}| + |\mu - x_{\frac{n+2}{2}}| + |\mu - x_{\frac{n+4}{2}}| + ... + |\mu - x_n|$$

On applying AM $\leq$ RMS

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

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

$$\leq \sqrt{n}\sqrt{n-1}\sigma$$

Since n-1 is less than n

$$\leq n\sigma$$

Hence, the given inequality is true if n is even
So it is proved that

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

Question 3

(a)

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

Since, The two events, contestant choosing the first door is independent with the car being placed behind i^{th} door

$$P(C_i, Z_1) = P(C_i)P(Z_1)$$
$$\therefore P(C_i|Z_1) = P(C_i)$$

As mentioned in the question $P(C_i) = \frac{1}{3}$for all $i \in \{1, 2, 3\}$
$\therefore P(C_i|Z_1) = \frac{1}{3}$ for all $i \in \{1, 2, 3\}$

(b)

For i = 1,$P(H_3|C_1, Z_1)$ Probability that host opens door 3 given that contestant opened door 1 and car is behind door 1 equals to $\frac{1}{2}$.
Here host knows the car is behind the door 1, he can either open the door 2 or door 3 with equal probability , so the probability will be $\frac{1}{2}$.
For i = 2,$P(H_3|C_2, Z_1)$ Probability that host opens door 3 given that contestant opened door 1 and car is behind door 2 equals to 1.
In this case,the host knows the car is behind the door 2,so he cannot open the door 2 then the probability for this case is 1.
For i = 3,$P(H_3|C_3, Z_1)$ Probability that the host opens door 3 given that contestant opened door 1 and car is behind door 3 equals to 0.
As the host knows the car is behind the door 3,so he will not open door 3 then probability will be 0.

$$\therefore P(H_3|C_1, Z_1) = \frac{1}{2}$$
$$P(H_3|C_2, Z_1) = 1$$
$$P(H_3|C_3, Z_1) = 0$$

(c)

The probability of winning by switching is $P(C_2|H_3, Z_1)$
According to **Bayes theorem**

$$\begin{aligned}
P(C_2|H_3, Z_1) &= \frac{P(C_2, H_3, Z_1)}{P(H_3, Z_1)} \\
&= \frac{P(C_2, H_3, Z_1)}{P(H_3|C_1, Z_1)P(C_1, Z_1) + P(H_3|C_2, Z_1)P(C_2, Z_1) + P(H_3|C_3, Z_1)P(C_3, Z_1)} \\
&= \frac{(\frac{1}{3} * \frac{1}{3})}{(\frac{1}{2}) * (\frac{1}{3} * \frac{1}{3}) + (1) * (\frac{1}{3} * \frac{1}{3}) + (0) * (\frac{1}{3} * \frac{1}{3})} \\
&= \frac{2}{3} \\
P(C_2, Z_1) &= \frac{1}{3} * \frac{1}{3} \qquad \textit{Independent events} \\
P(H_3, Z_1) &= \left(\frac{1}{3} * \frac{1}{3} * \frac{1}{2}\right) + \left(\frac{1}{3} * \frac{1}{3} * 1\right) + \left(\frac{1}{3} * \frac{1}{3} * 0\right) = \frac{1}{3} * \frac{1}{2} \\
P(H_3|C_2, Z_1) &= 1 \qquad \textit{by the second part of the question}
\end{aligned}$$

Now as per the given formula

$$P(C_2|H3, Z1) = \frac{P(H_3|C_2, Z_1)P(C2, Z1)}{P(H_3, Z_1)}$$

LHS :

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

RHS :

$$= \frac{1 * (\frac{1}{3} * \frac{1}{3})}{\frac{1}{3} * \frac{1}{2}}$$
$$= \frac{2}{3}$$

$$\therefore \textbf{LHS} = \textbf{RHS} = \tfrac{2}{3}$$

So, both the expressions have same values

(d)

Similarly

$$\begin{aligned} P(C_1|H_3, Z_1) &= \frac{P(C_1, H_3, Z_1)}{P(H_3, Z_1)} \\ &= \frac{P(C_1, H_3, Z_1)}{P(H_3|C_1, Z_1)P(C_1, Z_1) + P(H_3|C_2, Z_1)P(C_2, Z_1) + P(H_3|C_3, Z_1)P(C_3, Z_1)} \\ &= \frac{(\frac{1}{3} * \frac{1}{3} * \frac{1}{2})}{(\frac{1}{3} * \frac{1}{3} * \frac{1}{2}) + (\frac{1}{3} * \frac{1}{3} * 1) + (\frac{1}{3} * \frac{1}{3} * 0)} \\ &= \frac{1}{3} \end{aligned}$$

(e)

Since, probability of switching is double than sticking to original choice.
Switching is indeed beneficial.

Question 4

Sample output and figures in the report.

Submitted code is for $f = 30\%$ and for getting plots of $f = 60\%$ replace f=0.6 in the submitted code.

For $\mathbf{f = 30\%}$

Relative Mean Squared Error for Median = 69.6310

Relative Mean Squared Error for Mean = 106.6195

Relative Mean Squared Error for Quartile = 0.0142

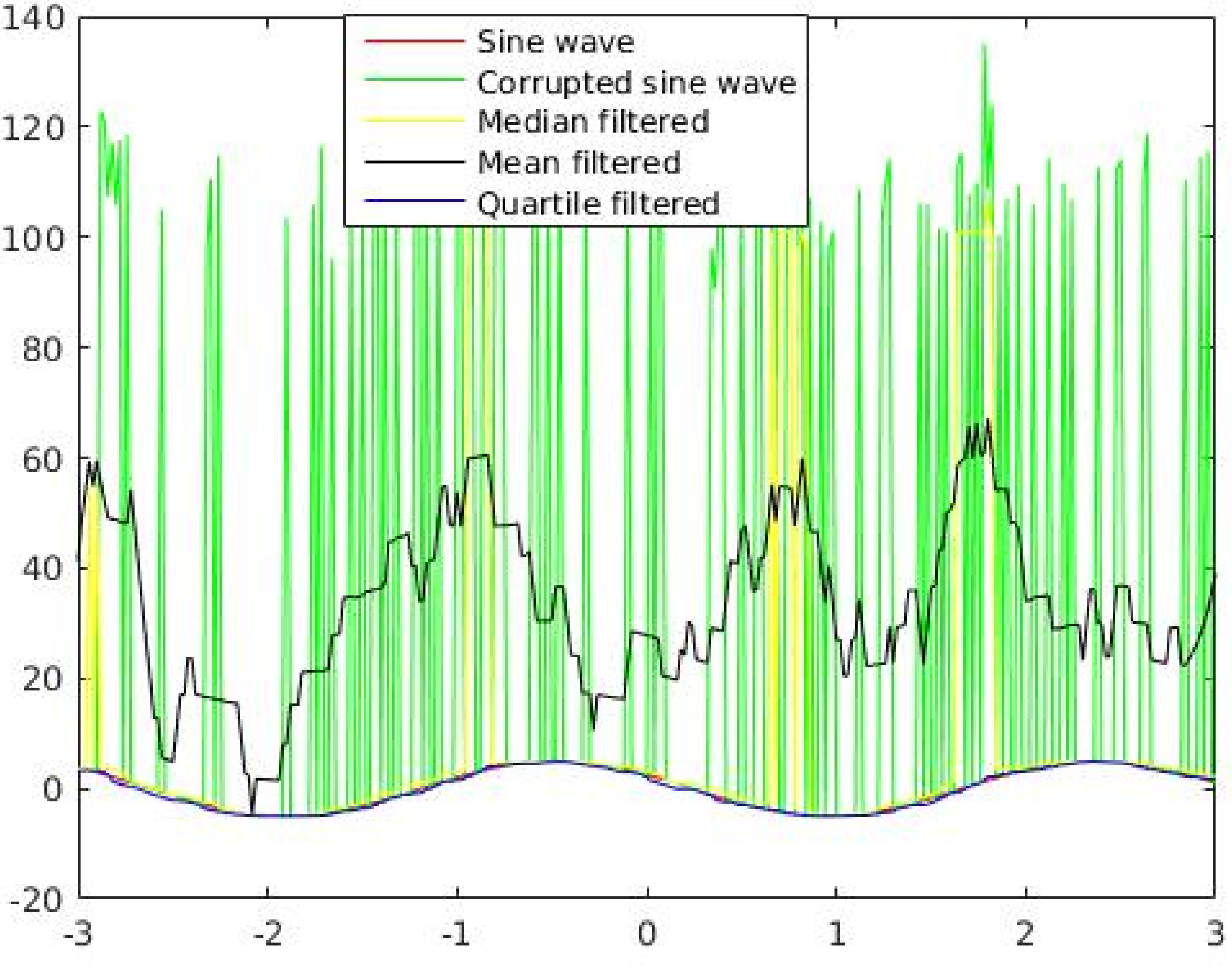

Figure 1: All the plots for f = 30%

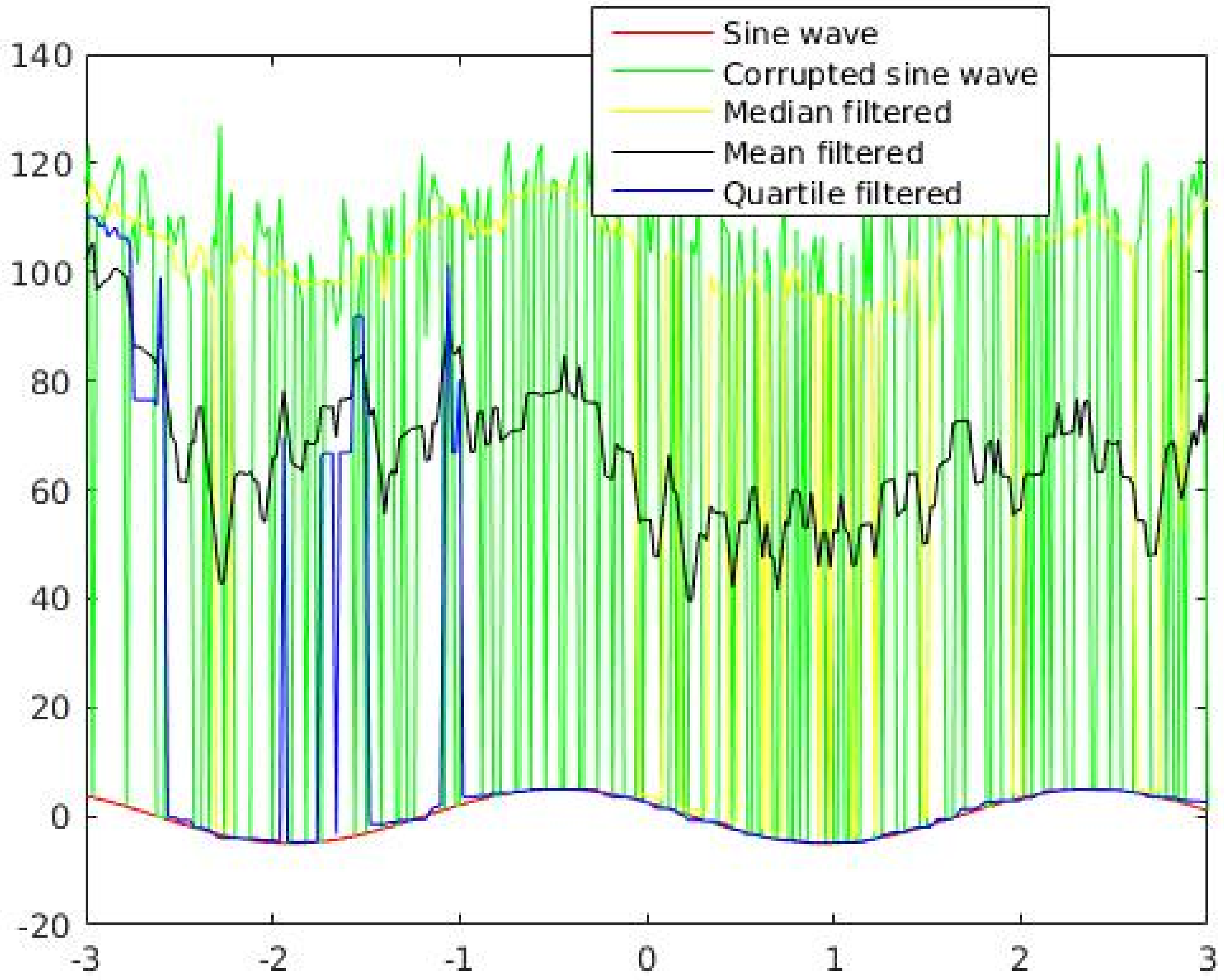

Figure 2: All the plots for f = 60%

For **f = 60%**

Relative Mean Squared Error for Median = 763.2747

Relative Mean Squared Error for Mean = 369.9397

Relative Mean Squared Error for Quartile = 83.3783

Quartile method produced best results in both the cases.Corrupted sine wave is formed after adding a random number between 100 to 120 which is too large compared to -5 and 5. If we take the mean method, mean calculates the average,average value is deviated from original, since corrupted values are largely deviated from original values. If we take the median method, median measures the middle value(50 percentile).Whereas in Quartile method(measures 25 percentile), most probably gives the values closer to required than median.Median is better than mean and works (good) for low f but Quartile is better than Median.

Question 5

Mean:

Let Oldmean = m_o , newMean = m_n , newDataValue = d

By the definition of mean,

$$m_o = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } m_n = \frac{\sum_{i=1}^{n} x_i + d}{n+1}$$

By the first equation,

$$\sum_{i=1}^{n} x_i = n * m_o$$

Substituting in the second equation,

$$m_n = \frac{n * m_o + d}{n+1}$$

Median:

Let oldMedian = M_o , NewDataValue = d , newMedian = M_n

Assuming that the array A is sorted in ascending order and all the numbers are unique.
We consider two cases, n is even and n is odd

When n is **even**,
Let $a = A_{\frac{n}{2}}$, $b = A_{\frac{n}{2}+1}$

- $d \leq a$ then $M_n = a$
- $d \geq b$ then $M_n = b$
- $a < d < b$ then $M_n =$ d

When n is **odd**,
Let $x = A_{\frac{n+3}{2}}$, $y = A_{\frac{n-1}{2}}$

- $d \geq x \implies M_n = \frac{M_o+d}{2}$
- $d \leq y \implies M_n = \frac{d+M_o}{2}$
- $y < d < x \implies M_n = \frac{M_o+d}{2}$

We have assumed that when the number of elements are even , then median is the average of the two middle values i.e. Matlab's convention.

Standard Deviation:

Let newStd = σ_n , oldStd = σ_o , NewDataValue = d , Oldmean = m_o , newMean = m_n.
$\sigma_o = \sqrt{\frac{\sum_{i=1}^{n}(x_i - m_o)^2}{n-1}}$, $\sigma_n = \sqrt{\frac{\sum_{i=1}^{n+1}(x_i - m_n)^2}{n}}$
Add and subtract m_o

$$\sum_{i=1}^{n+1}(x_i - m_n)^2 = \sum_{i=1}^{n+1}((x_i - m_o) + (m_o - m_n))^2$$

$$= \sum_{i=1}^{n+1}(x_i - m_o)^2 + (d - m_o)^2 + \sum_{i=1}^{n+1}(m_o - m_n)^2 + \sum_{i=1}^{n+1} 2 * (m_o - m_n) * (x_i - m_o)$$

$\because m_o - m_n$ is a constant,

$$= (n-1)\sigma_o^2 + (d-m_o)^2 + (n+1)(m_o-m_n)^2 + 2(m_o-m_n)\sum_{i=1}^{n+1}(x_i-m_o)$$

$\because \sum_{i=1}^{n+1}(x_i-m_o) = d-m_o$

$$n*(\sigma_n)^2 = (n-1)\sigma_o^2 + (d-m_o)^2 + (n+1)(m_o-m_n)^2 + 2(m_o-m_n)(d-m_o)$$

By putting $d = (n+1)*m_n - n*m_o$,
$(d-m_o)^2 + (n+1)(m_o-m_s) + 2(m_o-m_s)(d-m_o) = d^2-(n+1)m_n^2+m_n^2n$

$$\therefore \sigma_n = \sqrt{\frac{(n-1)\sigma_o^2 + d^2 - (n+1)m_n^2 + m_n^2 n}{n}}$$

Histogram :

If we receive a new value to be added to Histogram of A,then we search for the appropriate bin in the Histogram of A, then update the bin by one value. If the new value doesn't fall into any of the given bins then we may have to modify bins(depending on the requirements) in such a way that the new value falls into some bin.
Ex:
Suppose we initially have 0 to 100 values and bin length of 5 in histogram of A.If we want to add 2500 value, then we can increase the width of the new bin i.e., the 21st one bin to appropriate length so that we can store the new value to histogram so that we don't lose the initial data.

Question 6

Let p_1 be the probability that among n people, two of share their birthday.
Let us find the probability that no two of them share their birthday which equals to $1-p_1$.
Each person can have his/her birthday on any of the 365 days.(Assuming as a non-leap year)
$\therefore$ For n people,the total possible outcomes are 365^n.
So, for the event that no two of them share birthday
The first person has 365 ways
The second person will have 364 ways (All days except the birthday of first)
Similarly,
The n^{th} person will have $(365-(n-1)) = (366-n)$ ways.

$$\therefore 1-p_1 = \frac{(365)(364)(363)...(366-n)}{365^n}$$
$$\implies p_1 = 1 - \frac{(365)(364)(363)...(366-n)}{365^n}$$

The probability is dependent on n.
Theoretically, if the number of the persons increase the probability indeed increases.
Even Mathematically, since $\frac{366-n}{365} \leq 1 \; \forall n \geq 1$which decreases the negative term, makes probability increase.
The probability that at least two of them share their birthday is at least p,

$$\implies p_1 \geq p$$

$$\implies 1 - \frac{(365)(364)(363)...(366-n)}{365^n} \geq p$$

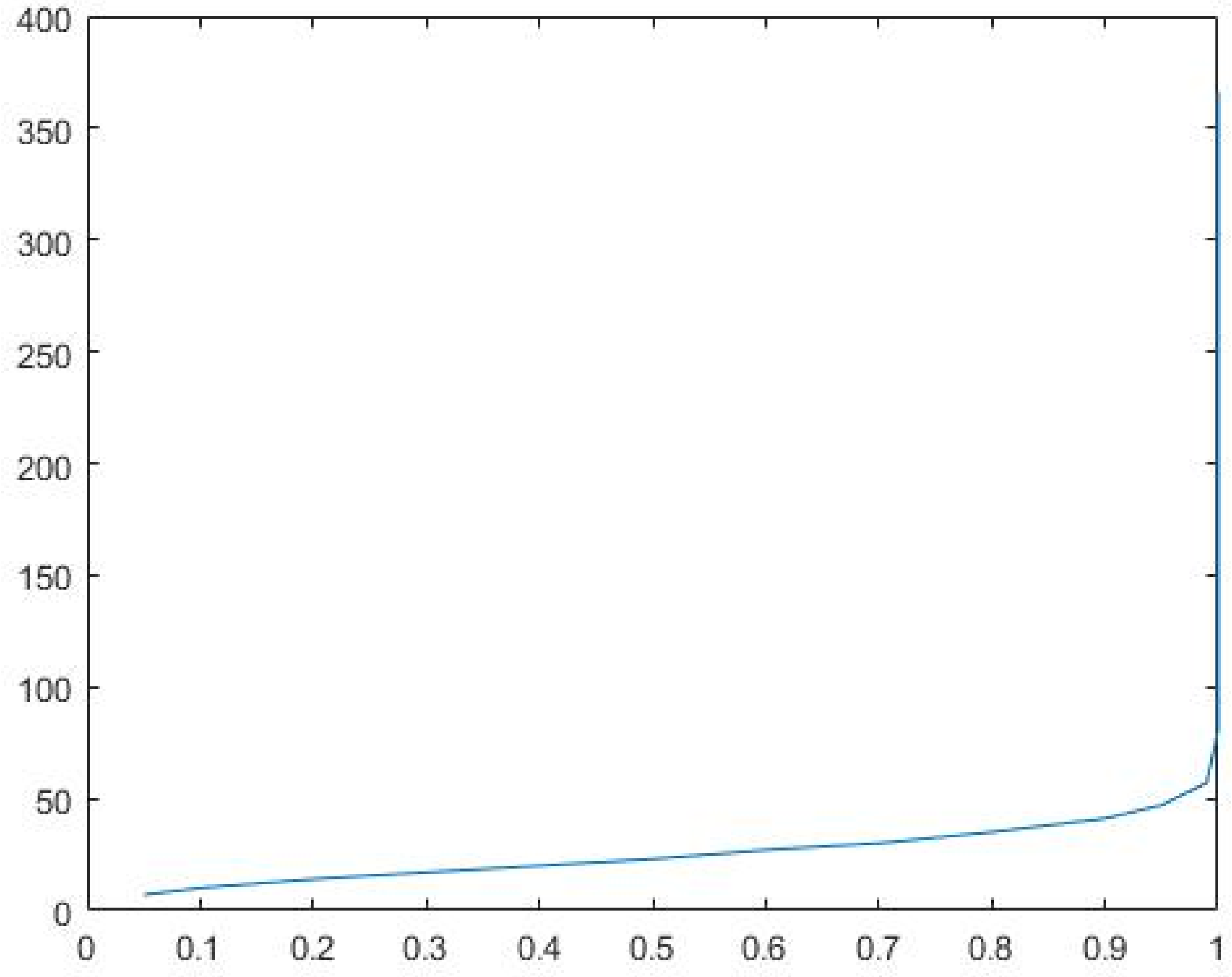

Figure 3: Smallest n versus p