EE325 Assignment 1.

Wednesday, 18 August 2021 4:18 PM

Group Members

	Name	Roll no.
1.	Mitul Wankhede	20D070051
2.	Rohan Kalbag	20D170033
3.	Asifahmad Shaikh	20D070017
4.	Mihika Dhok.	20D070026

1.

(a) Using the given options for choosing K students, 3 plots were made for each case:

CASE 1: Choosing the first K entries

CASE 2: Choosing an arbitrary point, and taking the next K entries after that.

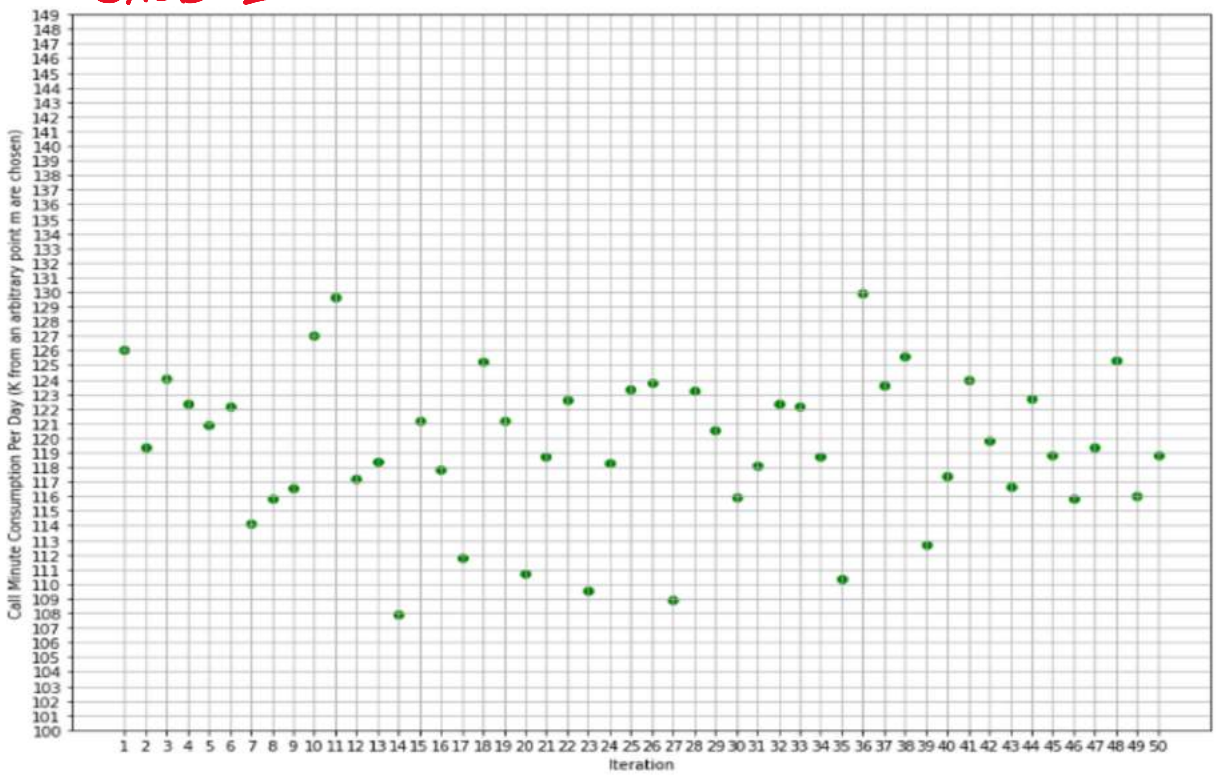
CASE 3: Choosing K random numbers from 0 to 9999. without replacement, and choosing the entry at that index.

For K = 10, 20, 50, 100, 200, 3 plots were made for each case, hence, 15 plots in total.

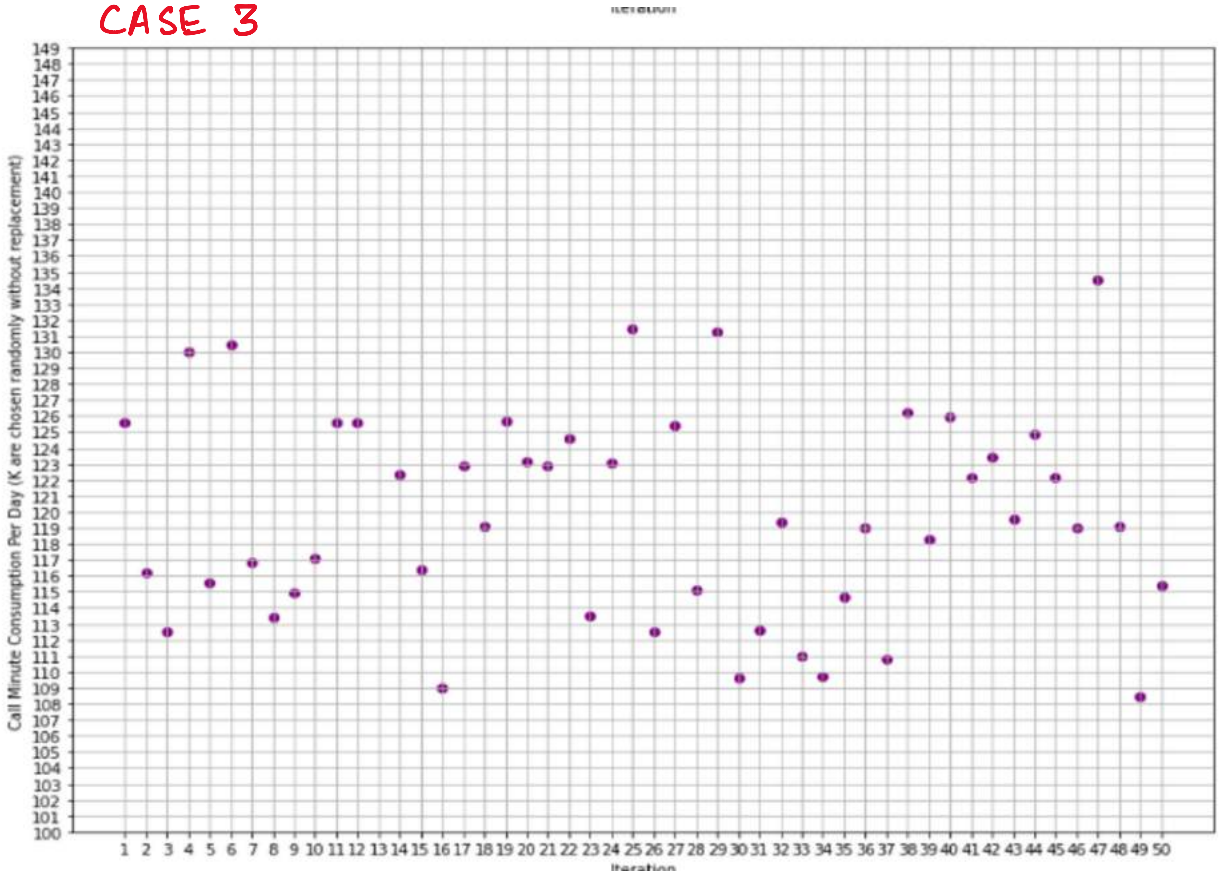
(b) For K = 10

CASE 1

CASE 2


Call Minute Consumption Per Day (K from an arbitrary point m are chosen)
Iteration

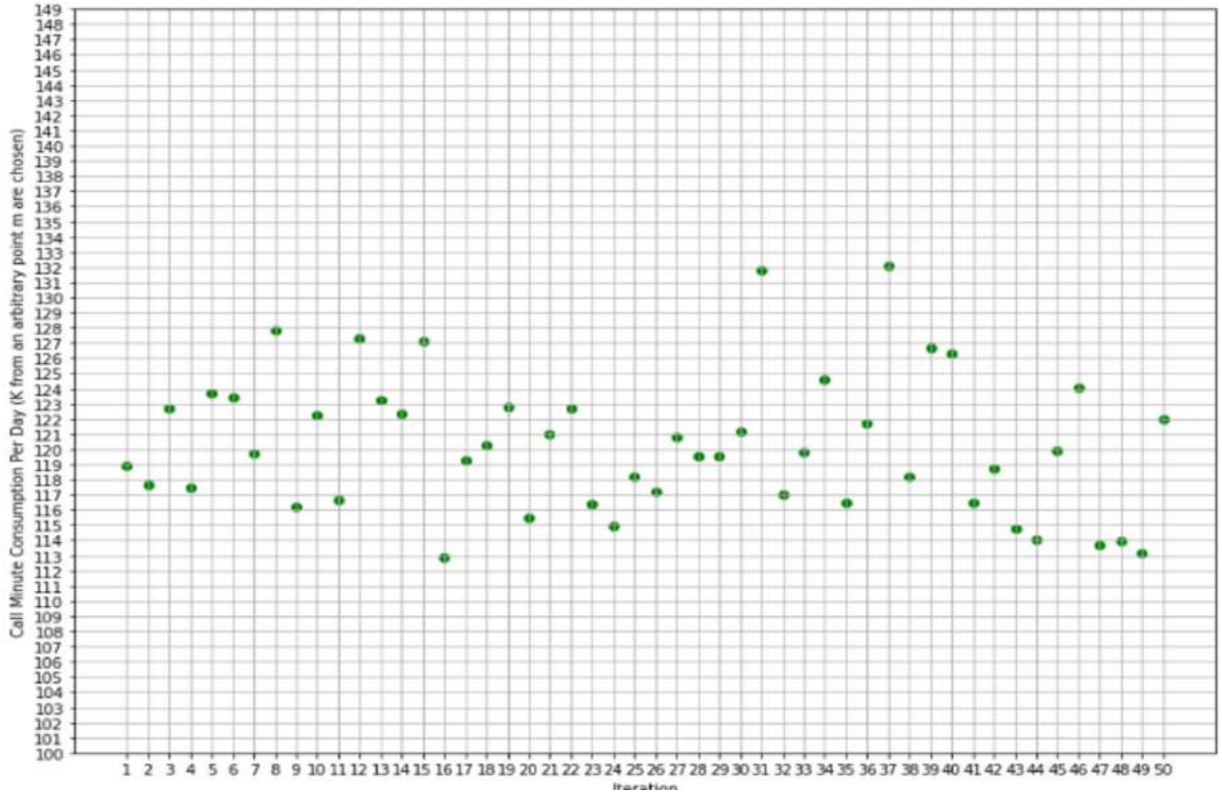
CASE 3


Call Minute Consumption Per Day (K are chosen randomly without replacement)
Iteration

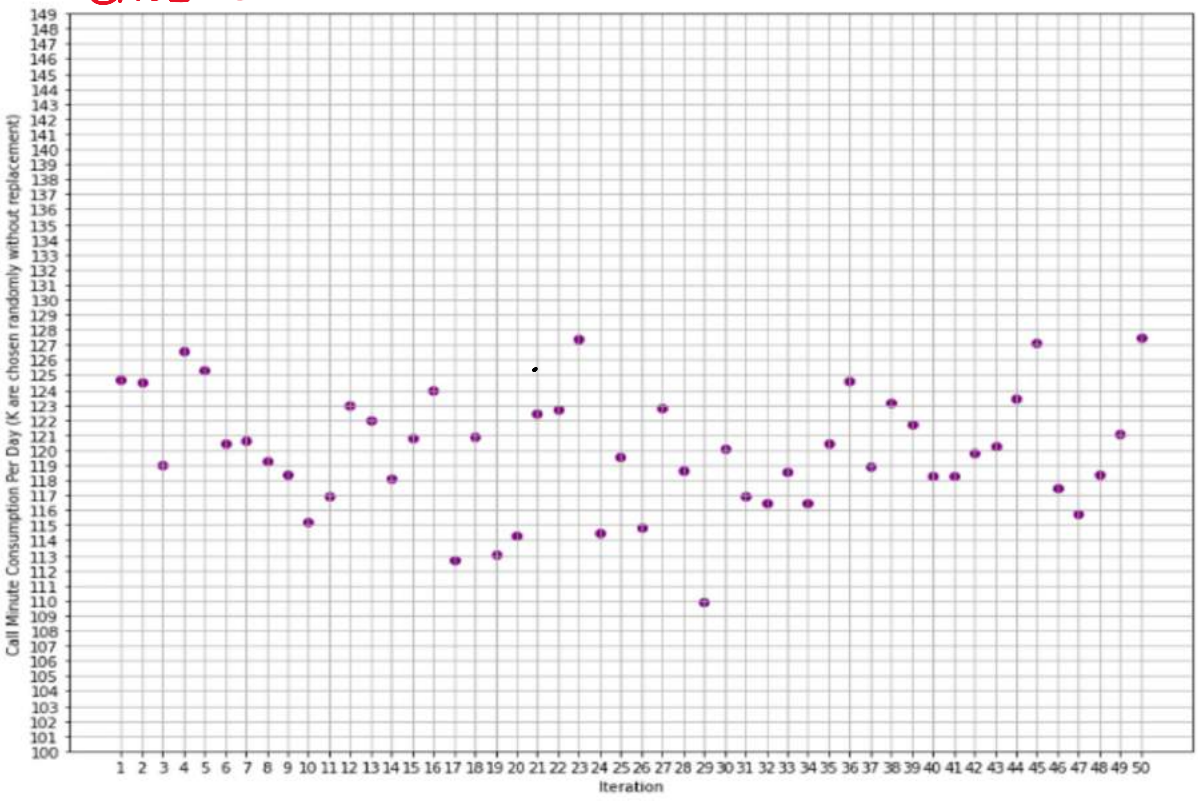
K = 20

CASE 1

CASE 2



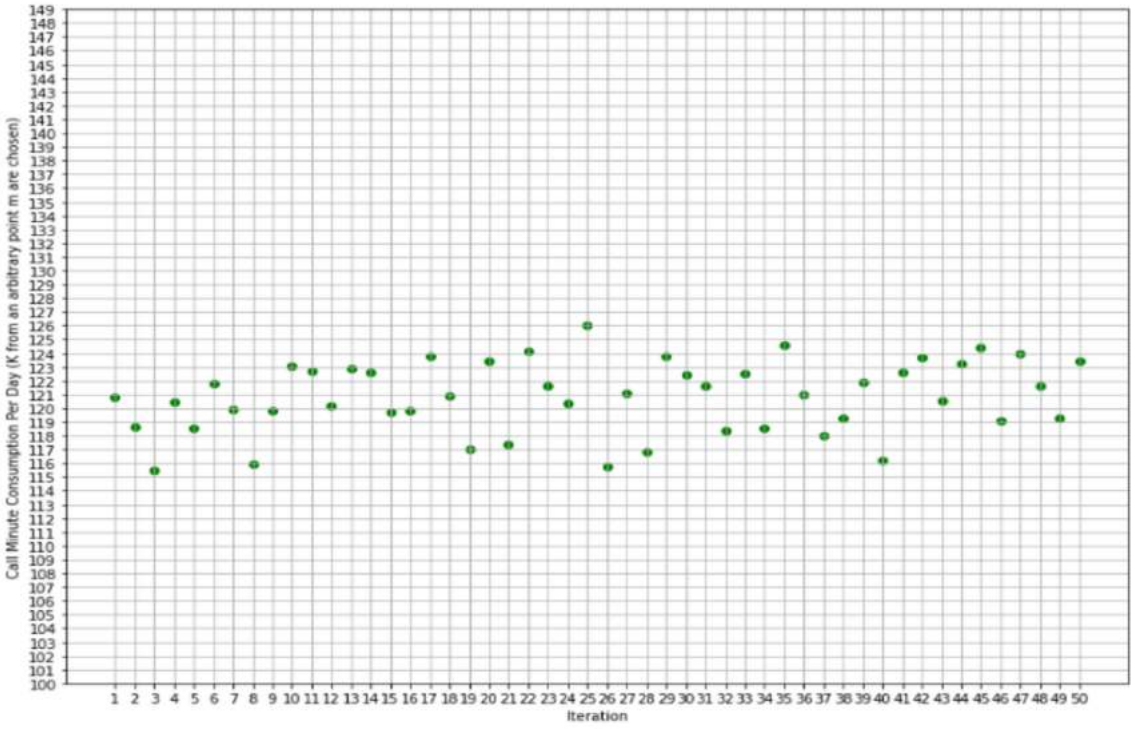
CASE 3



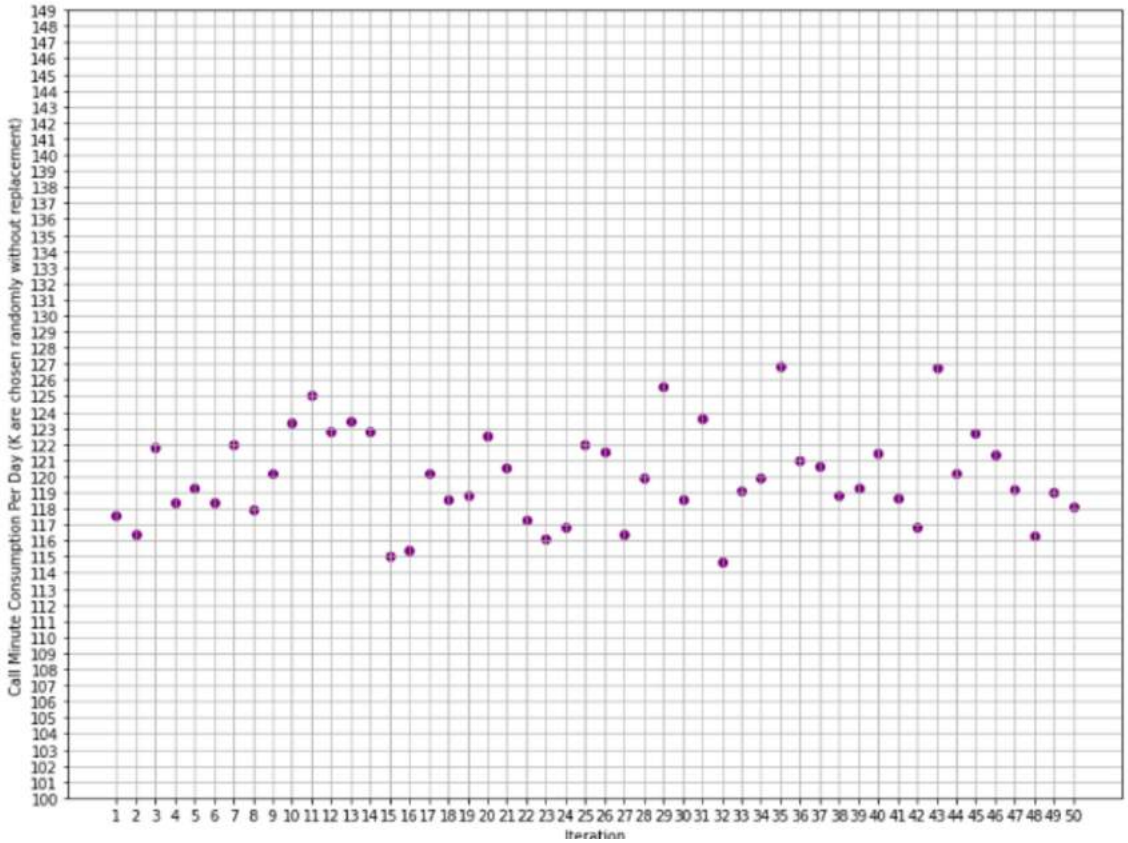
$K = 50$.

CASE 1

CASE 2


Call Minute Consumption Per Day (K from an arbitrary point m are chosen)
Iteration

CASE 3

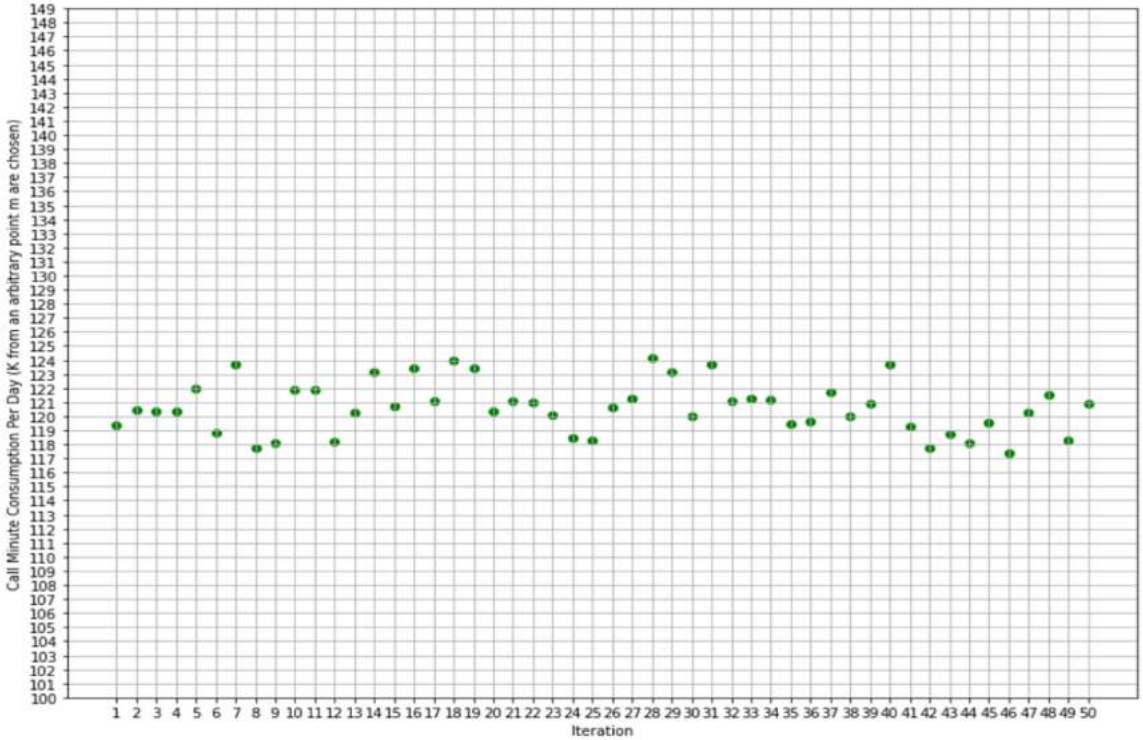

Call Minute Consumption Per Day (K are chosen randomly without replacement)
Iteration

-

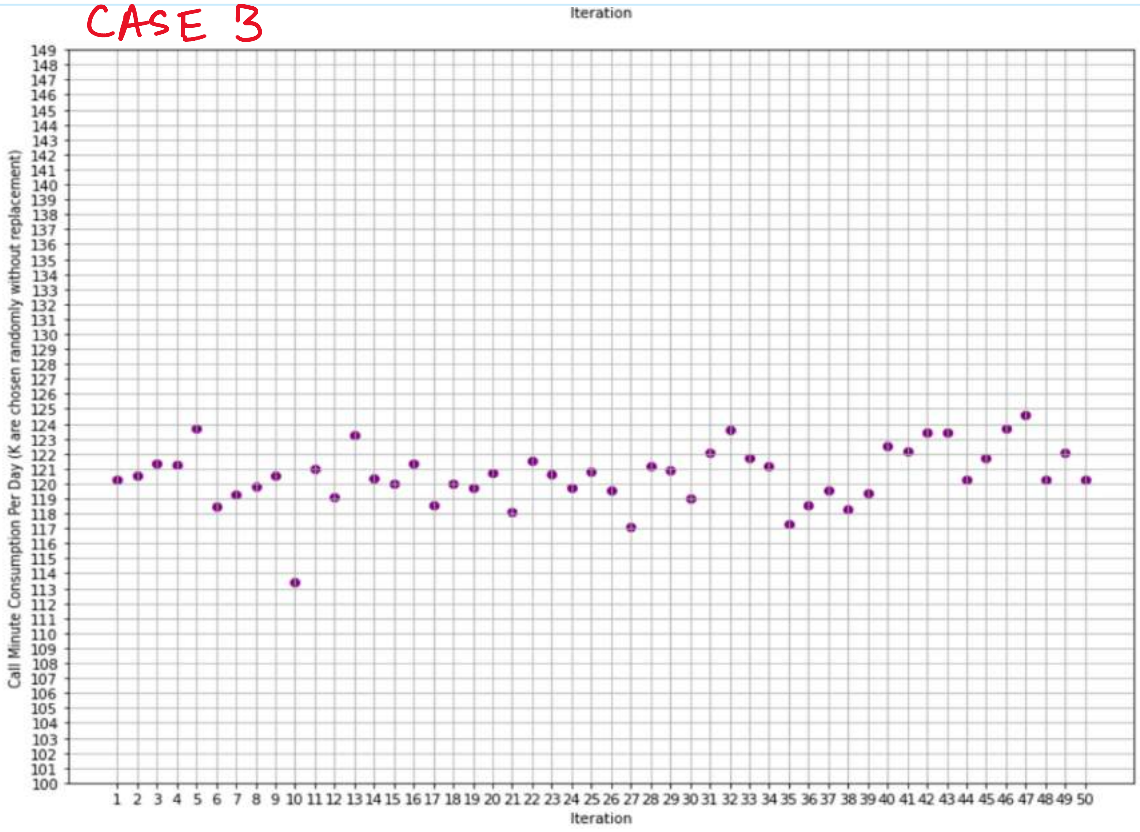
K = 100

CASE 1

CASE 2

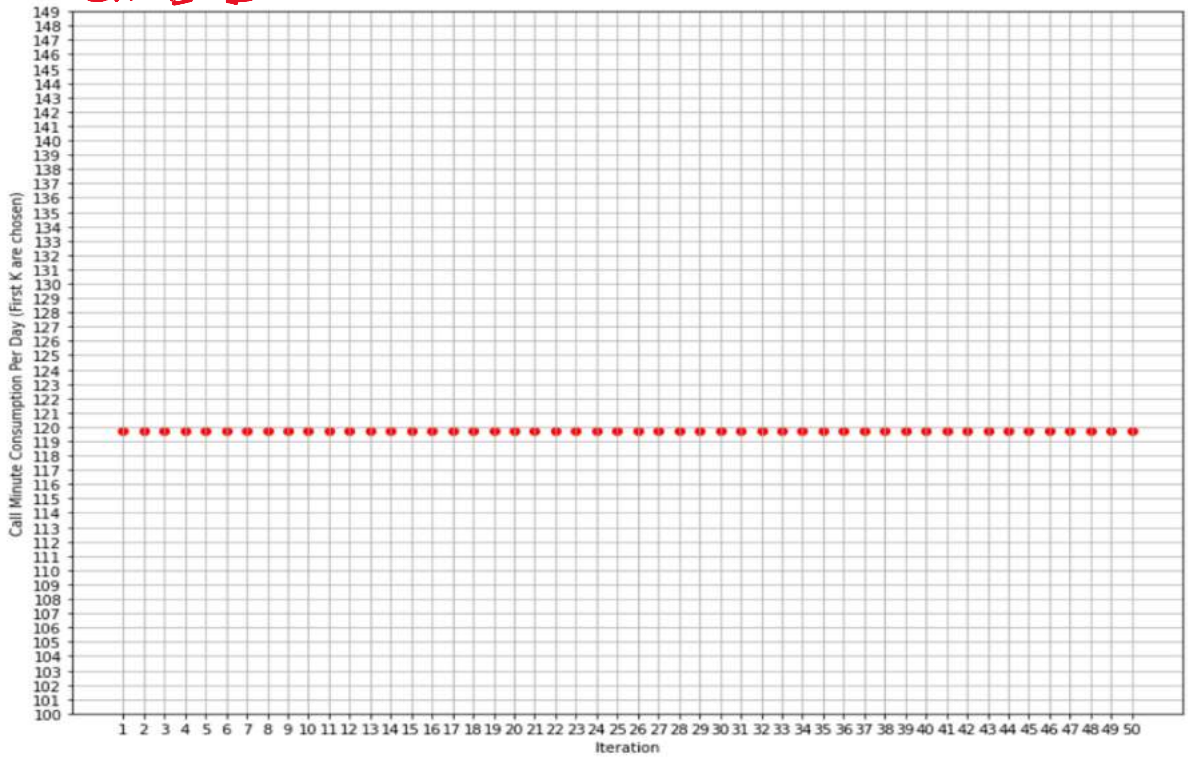


CASE 3

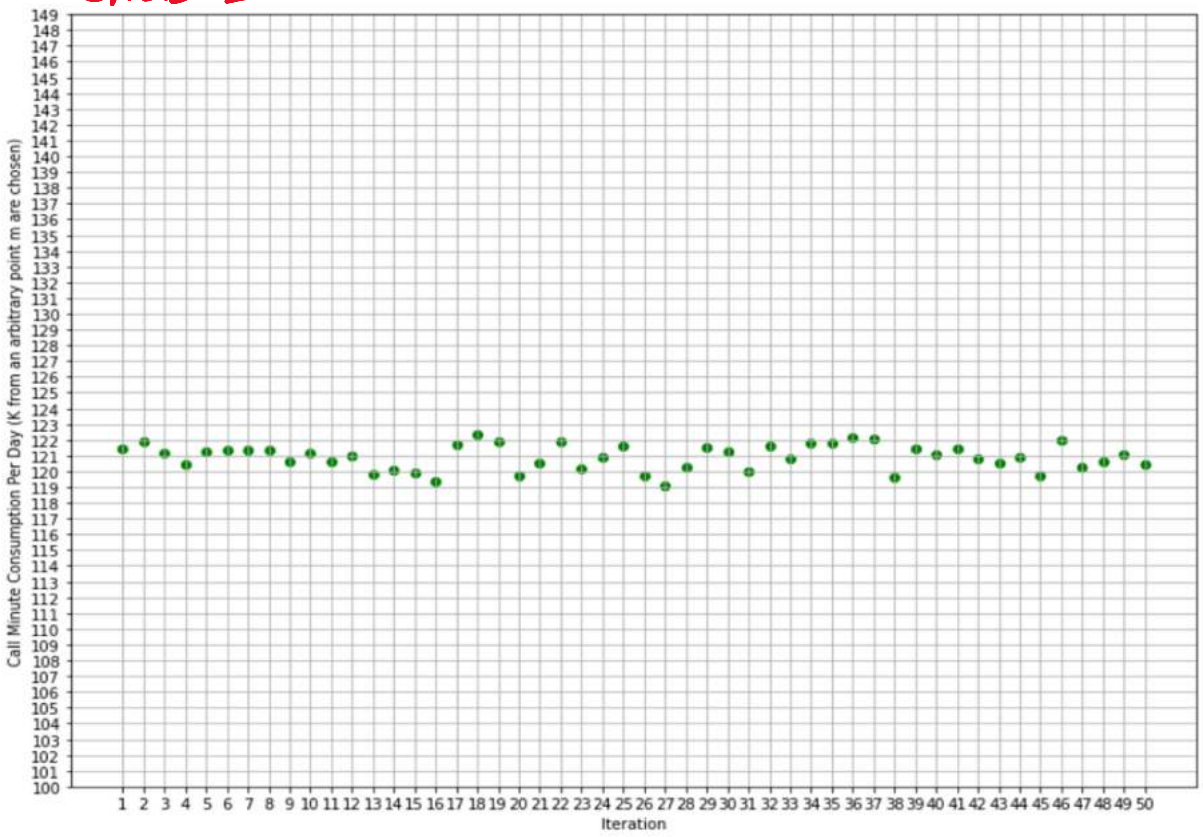


K = 200

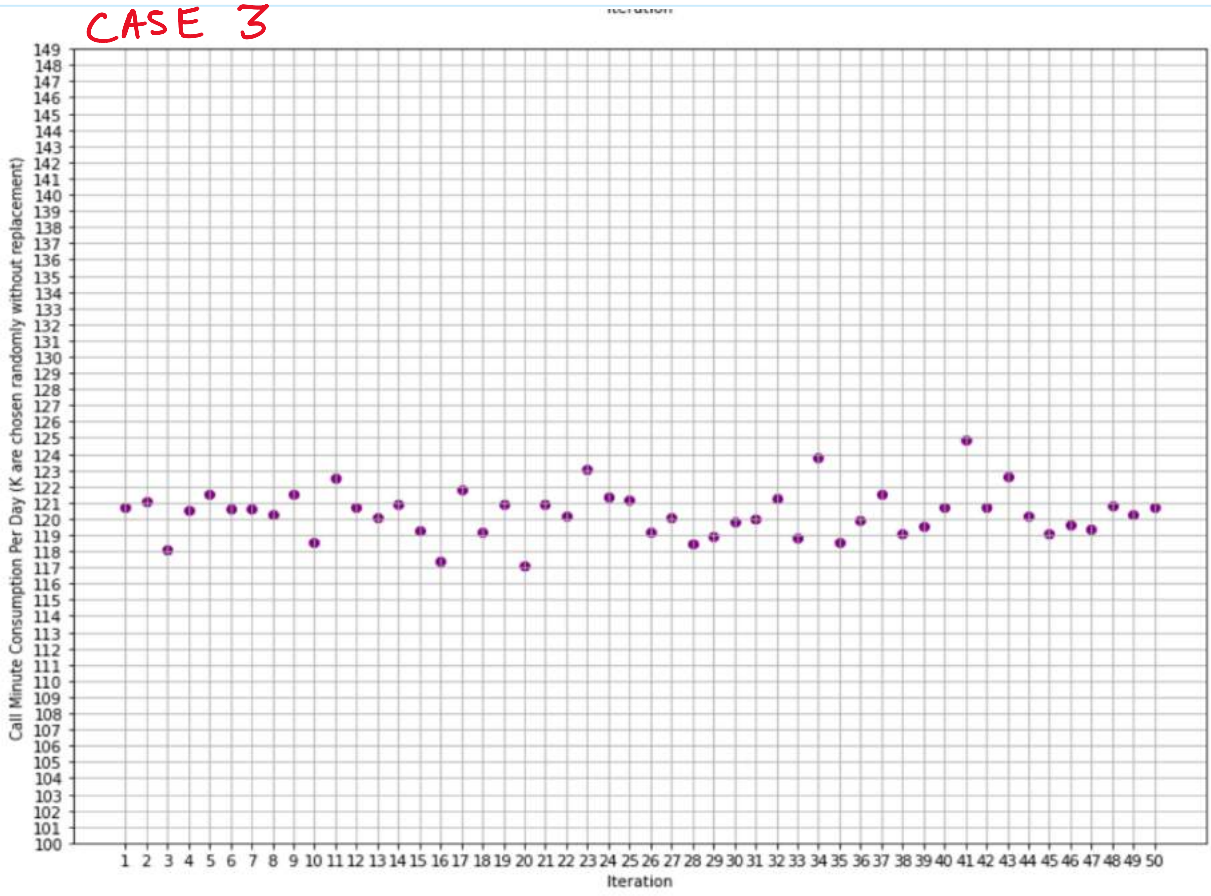
CASE 1



CASE 2


Call Minute Consumption Per Day (K from an arbitrary point m are chosen)
Iteration

CASE 3


Call Minute Consumption Per Day (K are chosen randomly without replacement)
Iteration

For CASE 1: We obtained the graph to be a straight line, which gradually approached 120, as the value of K increased.

For CASE 2: The graph appears to be a bit scattered, and as the value of K increases the value of standard deviation of the calculated means decreases.

For CASE 3: For small value of K, the graphs are more scattered than those in CASE 2. But, as K increases, the values appear very similar to CASE 2.

Hence from the above inferences, we can guess the average to be 120.

From the above graphs, it is difficult to guess the value of standard deviation by just mere speculation.

Using the value of K and method obtained later in part (A) and (B), we get:

Mean = 120.63
Standard deviation = 17.18

⇒ The actual values of mean and standard deviation (for all entries) are:

Mean = 120.133
Standard deviation = 19.97

To find the quantitative measure to describe the sureness of the estimate from the single survey of K samples:

For a single value of K, and a particular scheme, let $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_{50}$ be the values of the means obtained and let $\bar{x}$ be the actual mean of all the 10,000 entries.
$\bar{x}'$ and σ are the mean and standard deviation respectively, of the fifty values $\bar{x}_1, \ldots, \bar{x}_{50}$.

We consider it to be a Gaussian distribution.

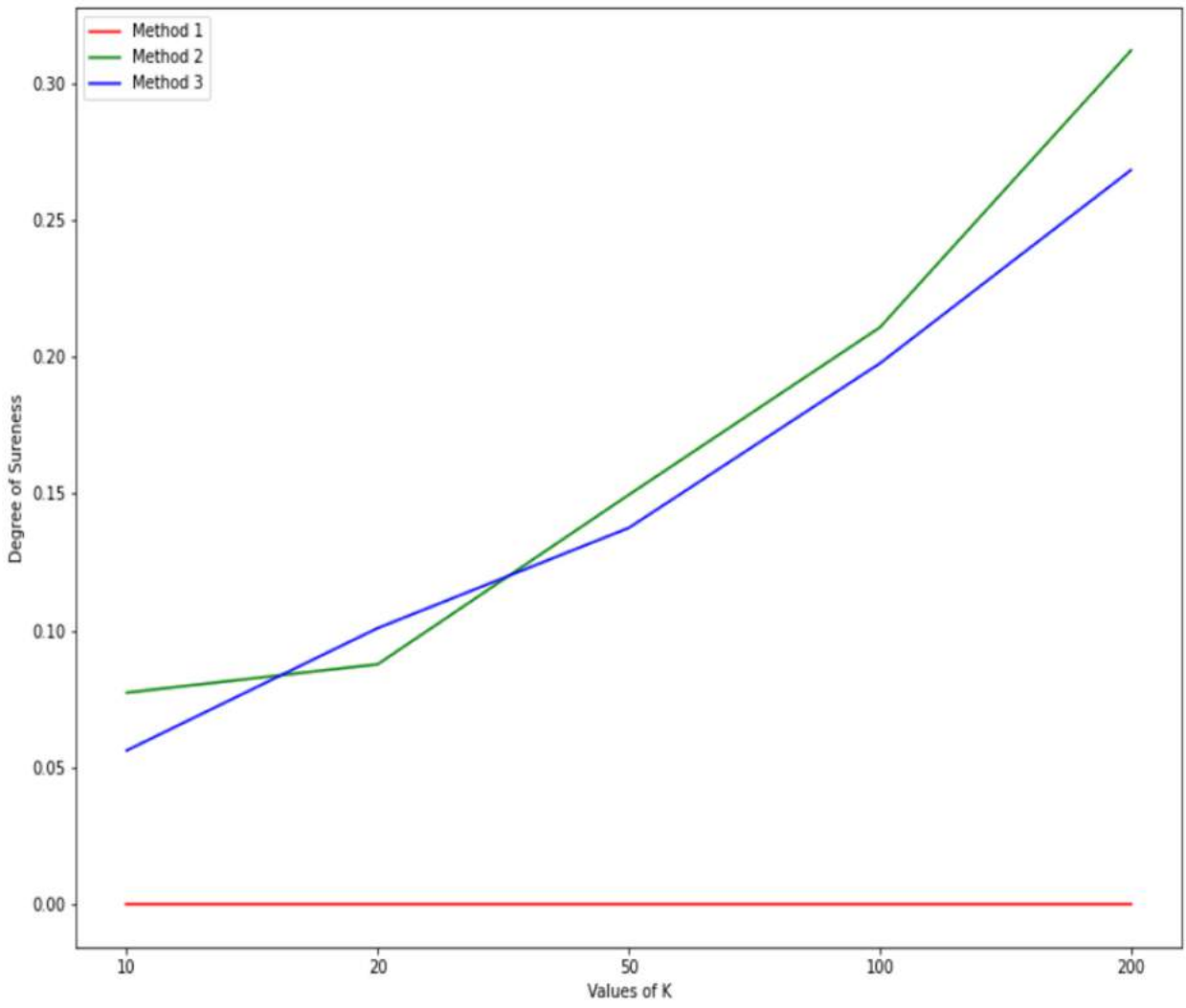
Hence, we define $$L = N(\bar{x} / \bar{x}', \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(\bar{x}-\bar{x}')^2}{2\sigma^2}}$$

Here, we use the Central Limit Theorem (CLT)

In probability theory, CLT states that, in many situations, when independent random variables are added, their properly normalized sum tends towards a normal distribution (Gaussian distribution), even if the original variables themselves are not normally distributed.

In the above formula, when the value of σ is very high, $L \to 0$.
When $\sigma = 0$, in this case too, $L \to 0$.

We plot a graph for Degree of Sureness vs Value of K for each of the 3 methods:



(A) In this graph we observe that, L increases with the value of K for CASE 2 and 3, while it remains zero for CASE 1 (since standard deviation is zero for CASE 1)

We get the highest L for CASE 2. Hence, we choose scheme 2 to determine the best guess.

choose scheme 2 to determine the best guess.

(B) We choose the best value of K to be 200, as it gives the highest value of L.

We define the value of L as written above to be the quantitative measure to describe the sureness of the estimate from the single survey of K samples.

---

2. Kolhi will assume the probability of the coin toss coming up heads to be 0.5, since he assumes that the coin is fair.

He will begin to doubt his assumption if a particular outcome is occurring more frequently than the other.

In order to measure the likelihood of an event, given that the coin is fair, we define the Likelihood Function (F_L) on the sample space.

Let n = no. of tosses
i = no. of heads

$$F_L(n,i) = \frac{\text{Prob. of current event}}{\text{Prob. of most likely event}} = \frac{{}^nC_i}{{}^nC_{[n/2]}} \times \frac{(1/2)^n}{(1/2)^n}$$

$$= \frac{{}^nC_i}{{}^nC_{[n/2]}}$$

where, [ ] = g.i.f

Here, the most likely event refers to :
Equal no. of heads and tails → when n = even.
$\frac{n+1}{2}$ and $\frac{n-1}{2}$ → when n = odd.

$$\Rightarrow {}^{n}C_{n+1/2} = {}^{n}C_{n-1/2} = {}^{n}C_{[n/2]}$$

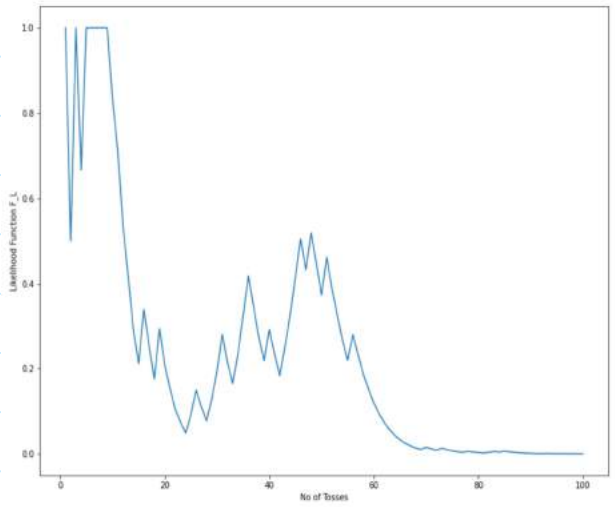
We define the log likelihood as $F_{LL} = -\log(F_L)$.

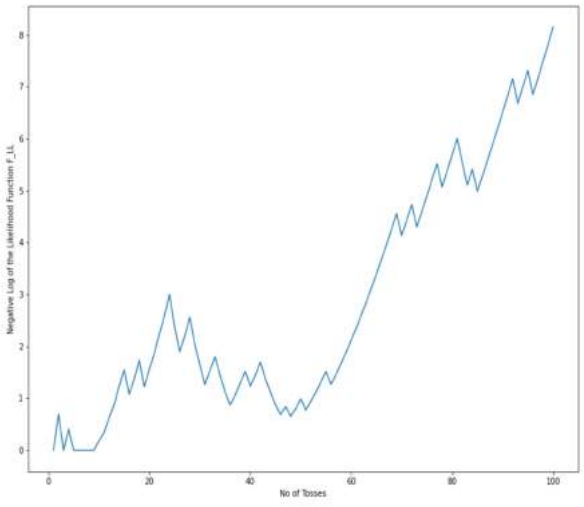
For an unbiased coin, the value of F_L will be close to 1, and hence, value of F_{LL} will be close to 0.

We say that Kolhi will begin to doubt his initial assumption when F_{LL} exceeds a threshold, say ε_1.

(b) He can be sure that his initial assumption is wrong when F_{LL} exceeds ε_2, where $\varepsilon_2 > \varepsilon_1 > 0$.

The following graphs depict the likelihood functions and the log likelihood functions :

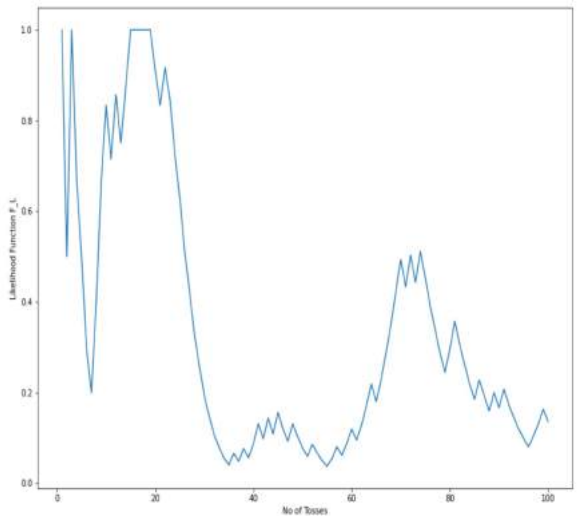


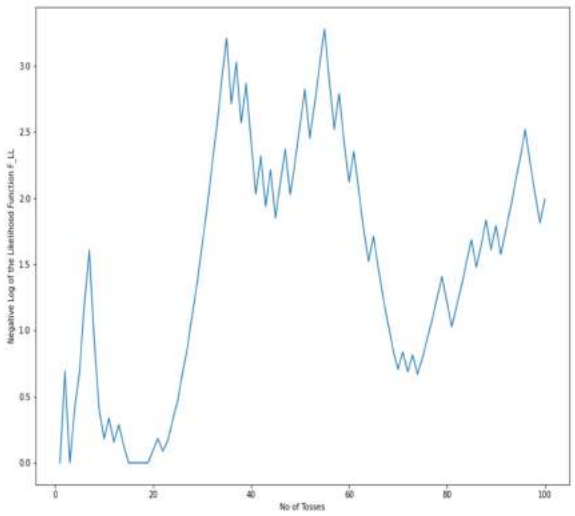


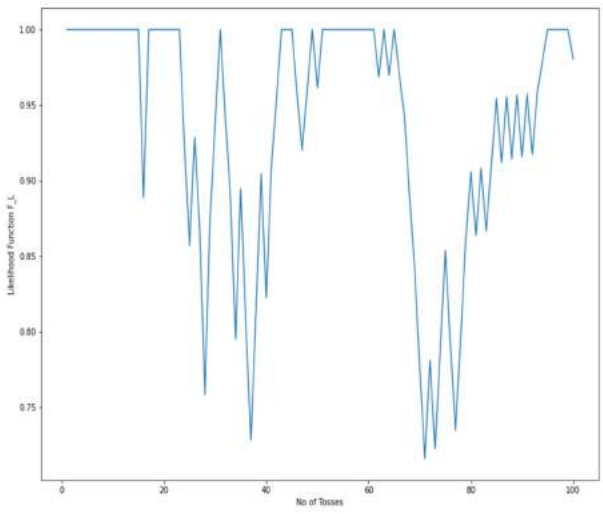
From these graphs, we choose the value of ε_1 to be 2 and ε_2 to be 4. (The choice of ε_1 & ε_2 will depend on the personality of the observer).

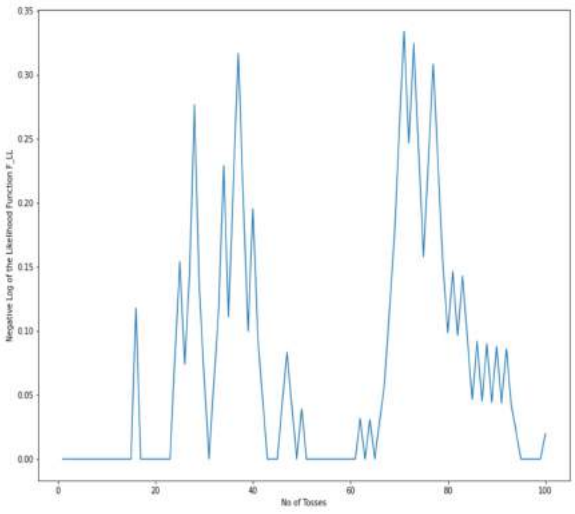
For these values of ε_1 & ε_2, Kolhi is doubtful at the 22nd toss, and is sure that his assumption is wrong, at the 68th toss.

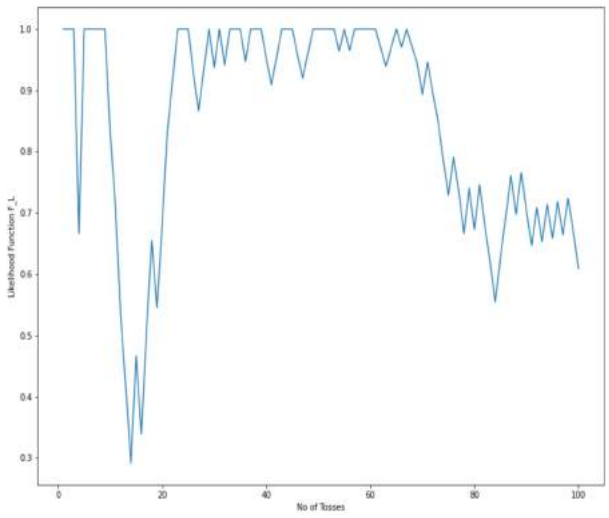
(c)

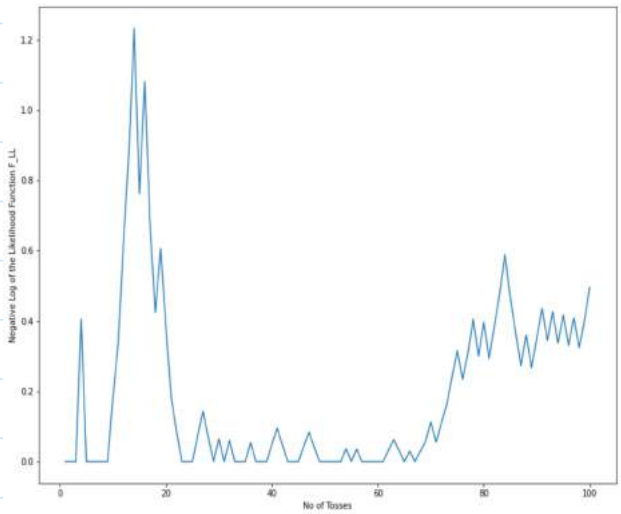












In the above graphs, for the same values of ϵ_1 and ϵ_2,

⇒ In file 2, he gets doubtful at the 32nd toss and is never sure that his assumption is wrong.

⇒ In file 3 and 4, he never doubts his assumption.

(d) The degree of sureness can be determined by the same values of F_L and F_{LL} defined above, and we use this concept as a quantitative measure as a description to our hunch.

---

3. We are supposed to minimize the root mean square error: $\sqrt{\sum_{i=1}^{N} (y_i - ax_i - b)^2}$ (Least Squares Error)

by determining a and b, to do straight line fitting on given data, of the form $y = ax + b$.

This least squares problem can be solved by the following method:

Let $A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$, $B = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$, $X = \begin{bmatrix} a \\ b \end{bmatrix}$

We have to find $X \in \mathbb{R}^{2\times 1}$ such that:

$$\|AX - B\|^2 = \sum_{i=1}^{N} |y_i - ax_i - b|^2$$

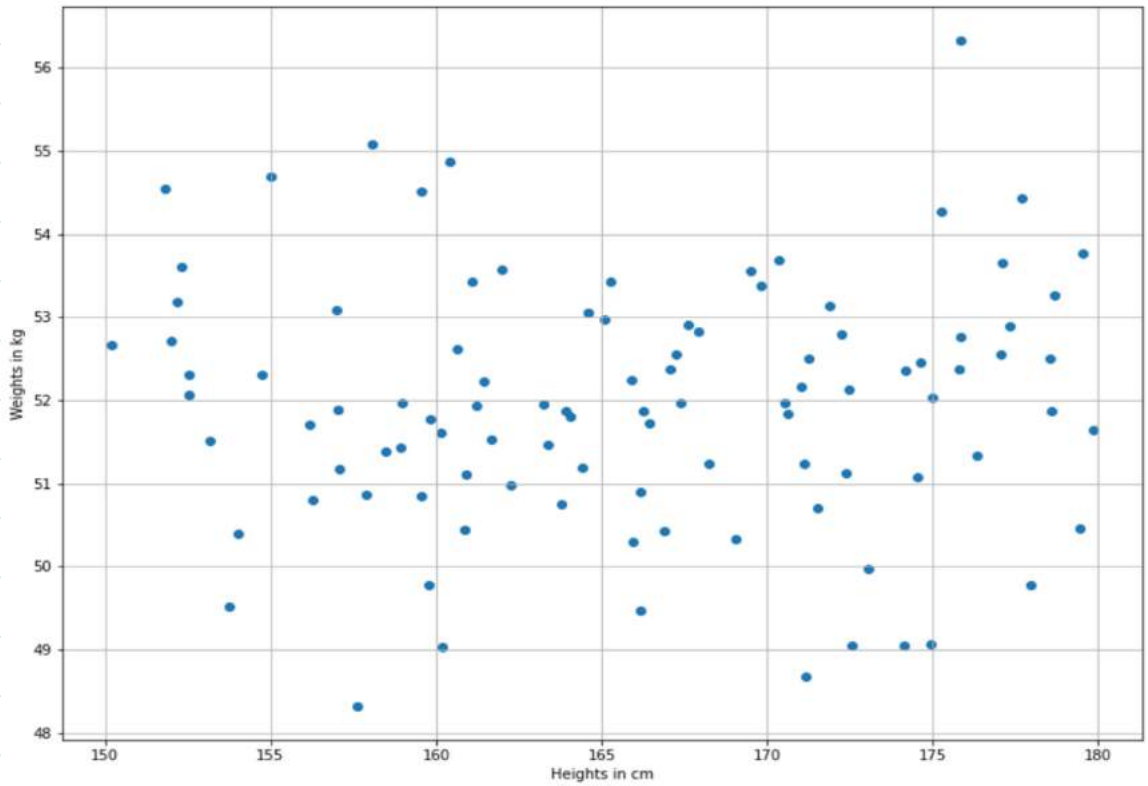
is minimized.

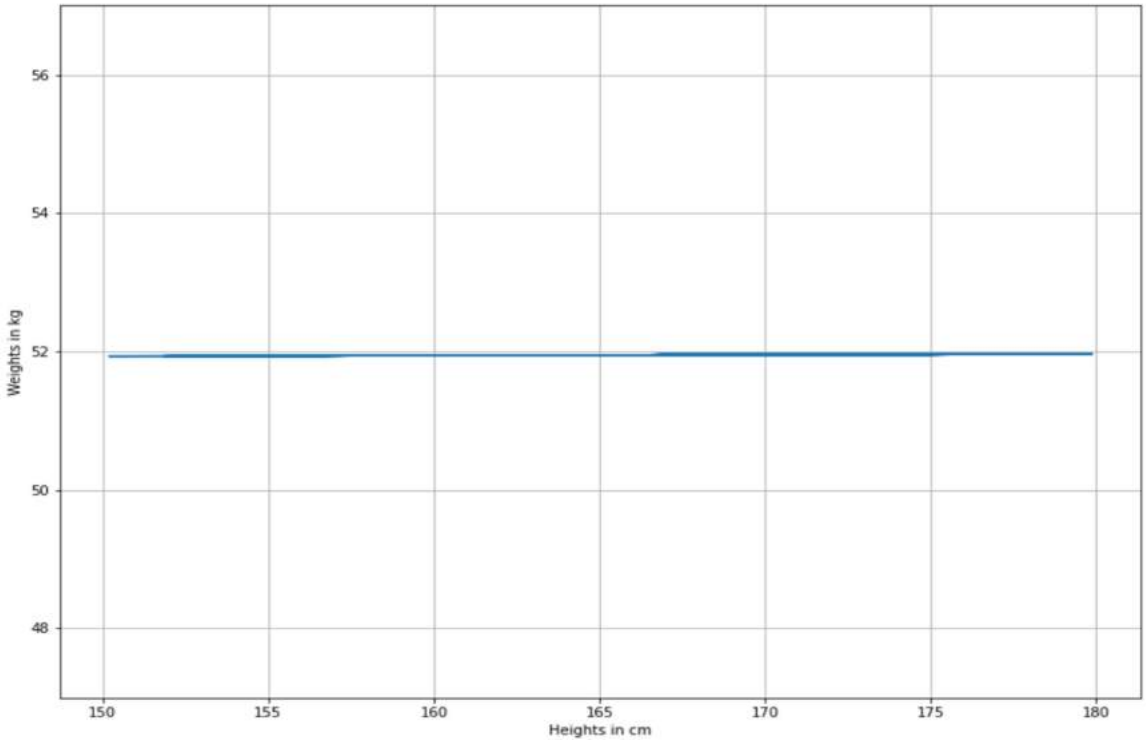
We find the best approximation to the vector B from the column space $C(A) = \{AX : X \in \mathbb{R}^{2\times 1}\}$ of A.

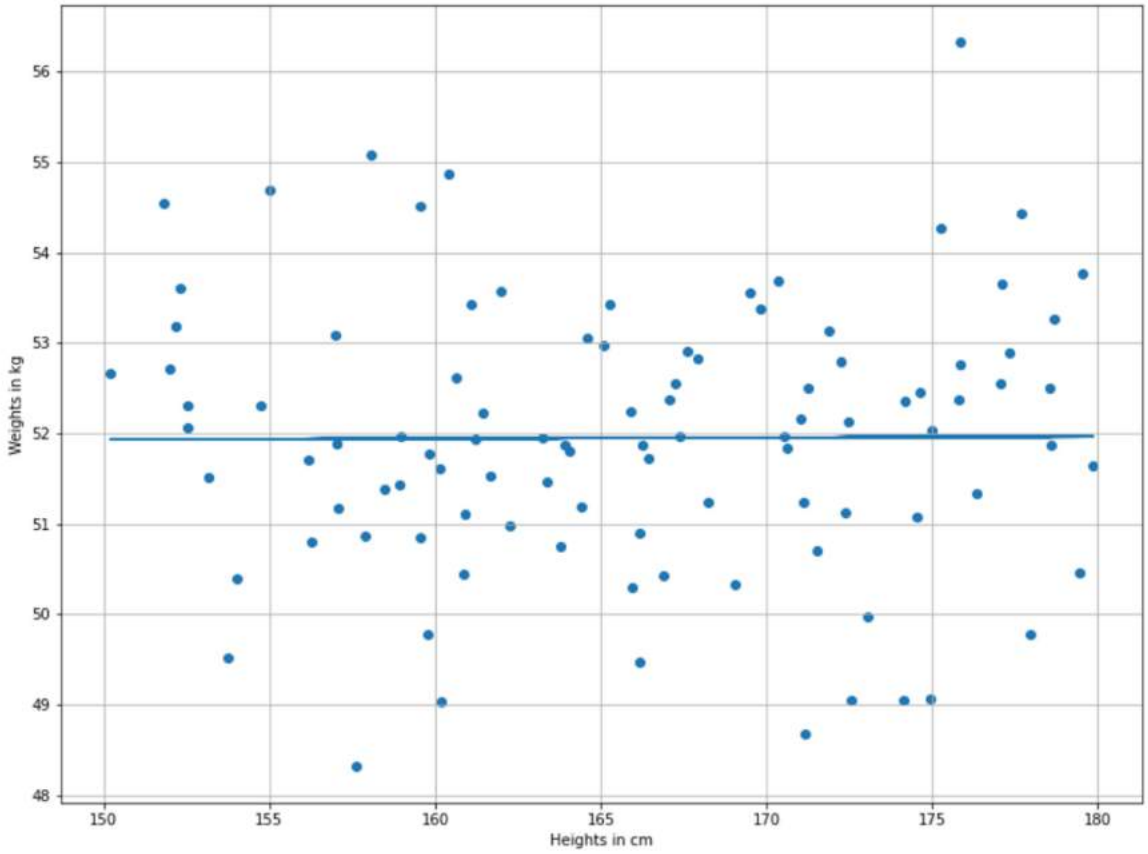
$$\Rightarrow X = (A^T . A)^{-1} . (A^T . B)$$

Using the above method, the values of a and b were calculated as follows..

$a = 0.001$
$b = 51.78$


Weights in kg
Heights in cm
56
55
54
53
52
51
50
49
48
150
155
160
165
170
175
180


Weights in kg
Heights in cm
56
54
52
50
48
150
155
160
165
170
175
180



To plot the error, each entry in the text file was taken.
For every (x_i, y_i) in the text file, we found $\hat{y}_i = ax_i + b$ from the values of a and b we calculated before.

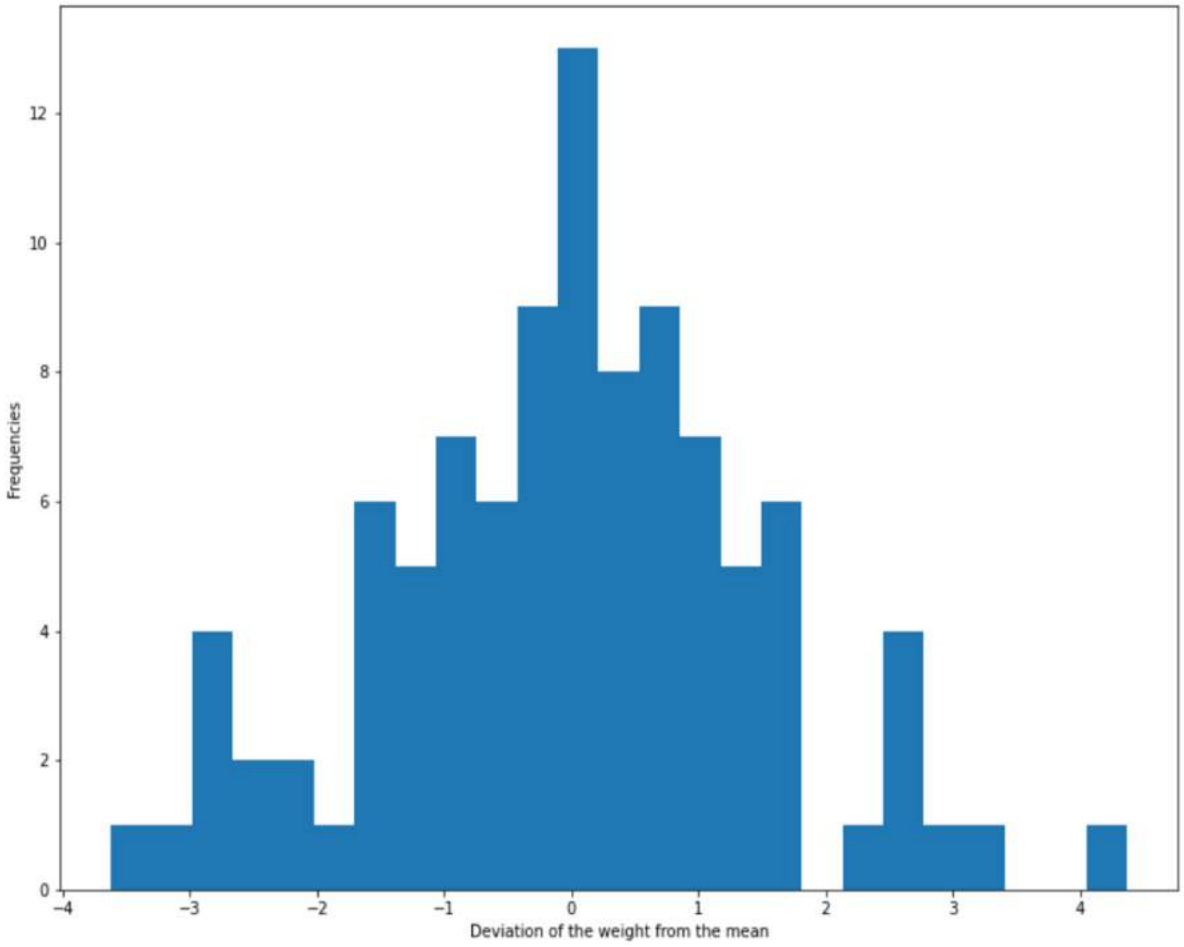
So, the error for the i^{th} entry is $\boxed{E_i = y_i - \hat{y}_i}$

(a) We took the entire set of E_i's, where $i = 1, \ldots, 25$. Then we plotted the histogram, and calculated the values of mean and standard deviation.

$$\overline{E_i} = 4.2 \times 10^{-16}$$

$$\text{Std. dev}(E_i) = 1.49$$

The range of values that E_i's took was divided into 25 subdivisions (d_i) and the y-coordinate represents the frequencies of each subdivision.



The histogram resembles a Gaussian function Hence the random variable E_i has Gaussian distribution and it is independently and identically distributed.

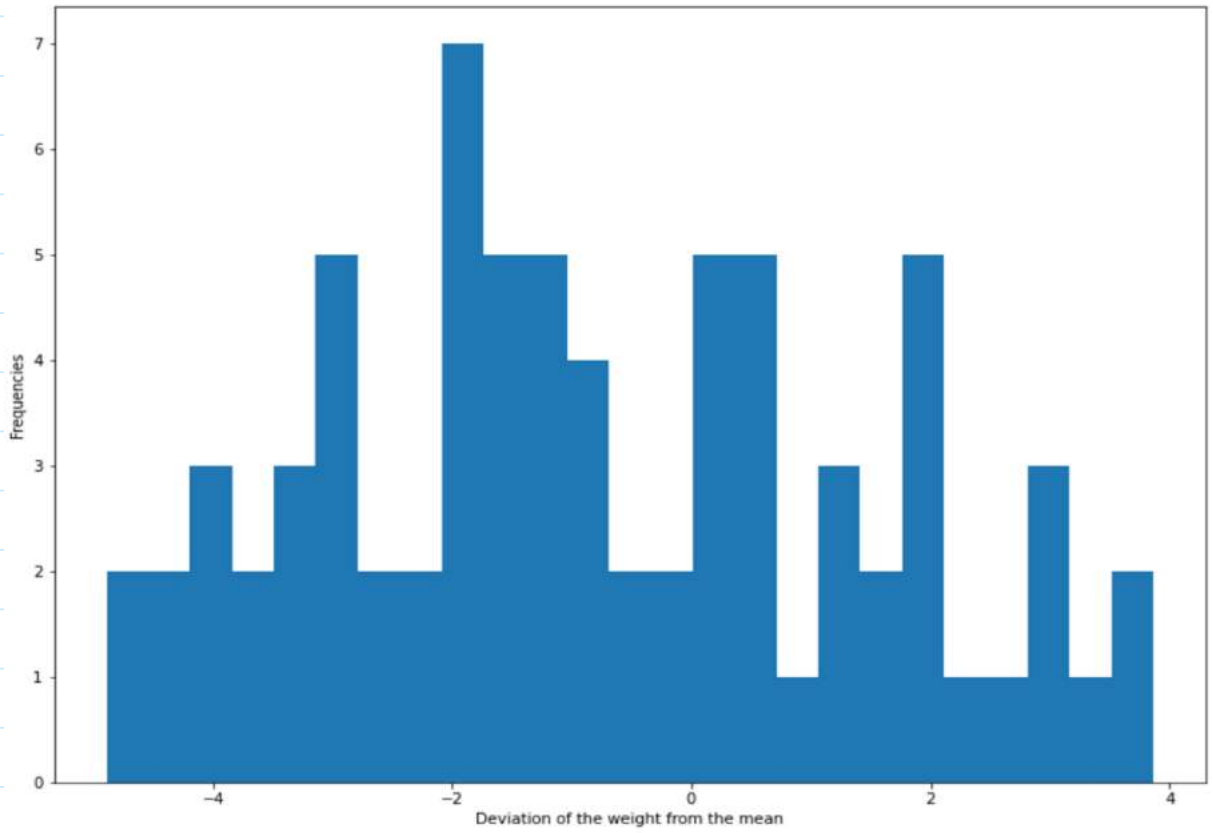
(b) The properties preferred would be:

⇒ we would like it to resemble the distribution of errors (E_i's) we obtained from our model.

⇒ The standard deviation should be comparable.

The standard deviation is the performance measure of our model on a dataset.

For our model to be good, it must show similar performance on unseen data and the data which it has been optimised upon. That is why, standard deviation should be comparable.



For this plot,

Average error (E_i') = −0.772

Standard dev. = 2.229