CS215 : Home Work Assignment -1

Kona Pavan Sai Subhash
Roll No. 24B0955

Shaik Suraj
Roll No. 24B0907

August 18, 2025

Contents

Chapter 1

Filtering the corrupted sine wave

1.1 Introduction

This question introduces the concepts of moving average, median, and quartile filters, which help to remove outliers. A 30% and 60% corrupted sine wave is modified/filtered to get a proper sine wave. The plot 1.2 of all waves is shown in the next section

1.2 Running Instructions

Run the file HW_Q1.m in matlab, you will see the plots in a .png file. You can change the % corrpution by changing the percent variable in the file.

1.3 Plots for 30% filtered sine waves

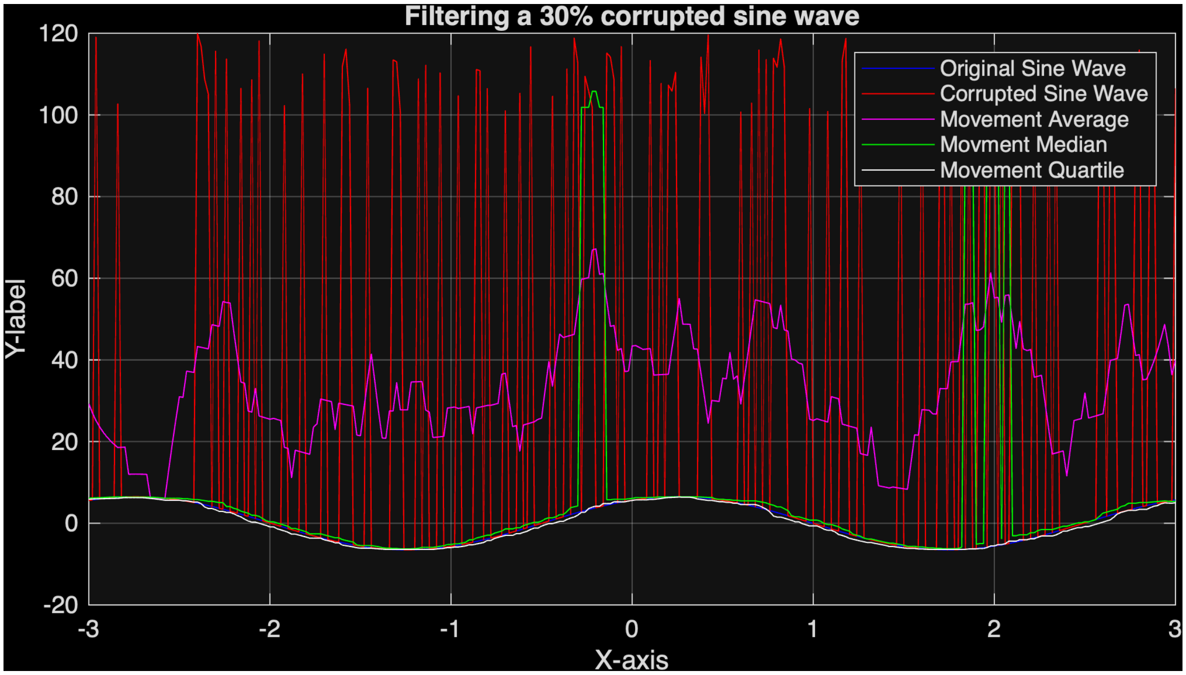

Figure 1.1: This plot is about filtering of 30% corrupted sine wave

1.4 Relative squared errors for all filtered plots

S.No	Filter type	Squared Error
1.	Mean	60.455
2.	Median	22.273
3.	1st Quartile	0.017

Table 1.1: Relative Sqaured Errors

The 1st quartile produced a better squared error compared to the other two. This behaviour can be justified by the fact that the outliers have a large value compared to the actual sine values, and the 1st quartile gives out the lower 25% of its neighbourhood. This decreases the error, whereas mean and median are affected by the large outliers(both large valued and large numbered).

1.5 Repeat for 60% corruption

The same process is repeated for 60% corrupted sine wave and the following results are obtained.

S.No	Filter type	SquaredError
1.	Mean	213.880
2.	Median	426.554
3.	1st Quartile	26.158

Table 1.2: Relative Sqaured Errors

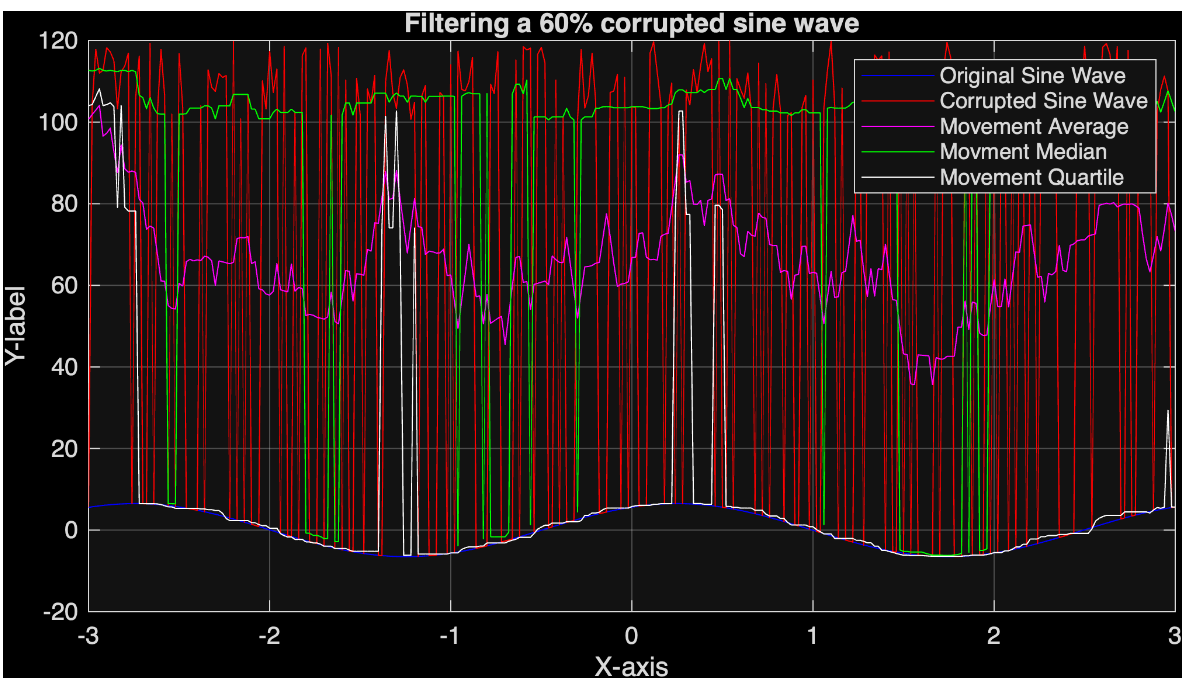

Figure 1.2: This plot is about filtering of 30% corrupted sine wave

Here also we observed the same results as before but may be even less percentile gives better results.

P.S: *I tried to find the best percentile, but it fluctuated between 1% and 15%*

Chapter 2

Updating the Statistical measures

2.1 Running Instructions

Using the functions in the file HW1_Q2.m, give the respective arguments to each function like new data value, old mean etc.
After giving arguments, run the code in matlab to get the updated values of mean, median, standard devation.

2.2 Mean

new mean:

Given old mean = μ_0, new data value = α, ~~old~~ length of old dataset = n.

let new mean be μ_n, the dataset is $\{x_1, x_2, --- x_n\}$

$$\mu_0 = \frac{x_1 + x_2 + x_3 + --- x_n}{n} \quad -①$$ according to definition of mean

$$\mu_n = \frac{x_1 + x_2 + --- x_n + \alpha}{n+1} \quad -②$$, as α is added to data set

$x_1 + x_2 + -- + x_n = n\mu_0$ from ①

$\underbrace{x_1 + x_2 + --- + x_n} + \alpha = (n+1)\mu_n$ from ②

$n\mu_0 + \alpha = (n+1)\mu_n$, replacing $x_1 + x_2 + -- x_n = n\mu_0$ by ①

$$\Rightarrow \mu_n = \frac{n\mu_0 + \alpha}{n+1}$$

So $\boxed{\text{new mean} = \frac{n \times (\text{old mean}) + \text{new data value}}{n+1}}$

2.3 Median

<u>new median</u>:

let old median $= m$, newdata value $= \alpha$. and old datavalues $= \{x_1, x_2, --- x_n\}$ in ascending order.

Case wise analysis:

1) n is even: ~~then~~ $m =$ mean of $(A[n/2], A[n/2+1])$

→ if $\alpha > A[(n/2)+1]$ α (will fall in this range)

$A[1], A[2], --- A[n/2], A[(n/2)+1] (--- A[n])$

the length of new dataset is odd, so the middle most element is median

$\boxed{A[(n/2)+1] \text{ is new median}}$ as number of elements left to it is $n/2$ and right to it is $n/2$.

$\underbrace{---}_{n/2}, A[(n/2)+1], \underbrace{---}_{n/2}$

→ if $\alpha < A[(n/2)]$

$\underbrace{(A[1], A[2], ---)}_{n/2 \text{ values}} A[n/2], \underbrace{A[(n/2)+1], ---, A[n]}_{n/2 \text{ values}}$

so the middle most element = $\boxed{\text{median} = A[n/2]}$

→ or $A[(n/2)] \le \alpha \le A[(n/2)+1]$

$\underbrace{A[1], A[2], ---, A[n/2]}_{n/2 \text{ values}}, \alpha, \underbrace{A[(n/2)+1], --- A[n]}_{n/2 \text{ value}}$

so the middle most element = $\boxed{\text{median} = \alpha}$

2) if n is odd,

old median $m = A[n+1/2]$ $\rightarrow$ 1

$\rightarrow$ if $\alpha \geq A((n+3)/2)$

$$\underbrace{A[1], A[2], \ldots}_{(n-1)/2 \text{ values}} (A[(n+1)/2], A[(n+3)/2]), \underbrace{(\ldots A[n])}_{(n-1)/2 \text{ values}}$$

median $=$ mean$(A[(n+1)/2], A[(n+3)/2])$

as now we have even number of elements, ~~we~~ median = average of the two middle elements

median $= \dfrac{m + A[(n+3)/2]}{2}$ as $m = A[(n+1)/2]$

$\rightarrow$ or $\alpha \leq A[(n-1)/2]$

$$\underbrace{(A[1], A[2] \ldots)}_{(n-1)/2 \text{ values}} A[(n-1)/2], A[(n+1)/2] \underbrace{(\ldots A[n])}_{(n-1)/2 \text{ values}}$$

So median $= \dfrac{A[(n+1)/2] + A[(n-1)/2]}{2}$

median $= \dfrac{m + A[(n-1)/2]}{2}$

$\rightarrow$ or if $A[(n-1)/2] \leq \alpha < A[(n+1)/2]$

$$\underbrace{A[1], A[2], \ldots A[(n-1)/2]}_{(n-1)/2 \text{ values}}, \alpha, A[(n+1)/2] \underbrace{(\ldots A[n])}_{(n-1)/2 \text{ values}}$$

So median $= \dfrac{\alpha + A[(n+1)/2]}{2}$

median $= \dfrac{\alpha + m}{2}$

2.4 Std. deviation(σ)

new standard deviation (σ_n):

let old data set is $\{x_1, x_2, \cdots x_n\}$

let old standard deviation be σ, old nean be μ, new mean be (μ_n), new data value be α ~~data~~ is

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

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

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

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

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

$$\sigma_n^2 = \frac{\sum_{i=1}^{n}(x_i-\mu_n)^2 + (\alpha-\mu_n)^2}{n}$$

$$(\sigma_n^2)n = \sum_{i=1}^{n}(x_i^2+\mu_n^2-2x_i\mu_n) + (\alpha-\mu_n)^2$$

$$(\sigma_n^2)n = \sum_{i=1}^{n}x_i^2 + n\mu_n^2 - 2\mu_n\sum_{i=1}^{n}x_i + (\alpha-\mu_n)^2$$

$$(\sigma_n^2)n = \underbrace{(\sigma^2)(n-1) + n\mu^2} + n\mu_n^2 - 2\mu_n(n\mu) + (\alpha-\mu_n)^2 \qquad \text{from } (1)$$

$$(\sigma_n^2)n = \sigma^2(n-1) + n(\mu^2+\mu_n^2-2\mu_n\mu) + (\alpha-\mu_n)^2$$

$$\sigma_n = \sqrt{\frac{\sigma^2(n-1)+n(\mu-\mu_n)^2+(\alpha-\mu_n)^2}{n}}$$

So, the new standard deviation $= \sigma_n = \sqrt{\frac{\sigma^2(n-1)+n(\mu-\mu_n)^2+(\alpha-\mu_n)^2}{n}}$

2.5 Updating Histogram

For updating the histogram, first we will look for the bin into which the new element will fall into. After identifying the bin, increase the count of the bin by 1. In this way, we can update the histogram.

Chapter 3

Short Proof

Given:

$$P(A) \geq 1 - q_1 \quad \text{and} \quad P(B) \geq 1 - q_2$$

Show that:

$$P(A \cap B) \geq 1 - (q_1 + q_2)$$

We know that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Substituting the given inequalities:

$$P(A \cup B) = (1 - q_1) + (1 - q_2) - P(A \cap B)$$

$$P(A \cup B) = 2 - (q_1 + q_2) - P(A \cap B)$$

We also know that for any event E:

$$0 \leq P(E) \leq 1$$

Hence:

$$P(A \cup B) \leq 1$$

So:

$$1 \geq 2 - (q_1 + q_2) - P(A \cap B)$$

Rearranging:

$$P(A \cap B) \geq 1 - (q_1 + q_2)$$

Conclusion:

Therefore, we have shown that:

$$\boxed{P(A \cap B) \geq 1 - (q_1 + q_2)}$$

Chapter 4

Bayes Theorem

In a town there are 100 buses: 1 red and 99 blue. An eyewitness (XYZ) saw a bus at night and reported it was *red*. Under test conditions similar to that night, XYZ

$$P(\text{SR} \mid \text{OR}) = 0.99, \qquad P(\text{SR} \mid \text{OB}) = 0.02.$$

We have to find $P(\text{OR} \mid \text{SR})$

Explanation of parameters.

Let

$\text{OR} = \{\text{bus is originally red}\}, \quad \text{OB} = \{\text{bus is originally blue}\}, \quad \text{SR} = \{\text{witness reports “red”}\}, \quad \text{SB} = \{\text{witness reports}$

Given Conditions:

$$P(\text{OR}) = \frac{1}{100} = 0.01, \qquad P(\text{OB}) = \frac{99}{100} = 0.99.$$
$$P(\text{SR} \mid \text{OR}) = 0.99, \qquad P(\text{SR} \mid \text{OB}) = 0.02.$$

Computation (Bayes' Theorem)

$$P(\text{OR} \mid \text{SR}) = \frac{P(\text{SR} \mid \text{OR})P(\text{OR})}{P(\text{SR} \mid \text{OR})P(\text{OR}) + P(\text{SR} \mid \text{OB})P(\text{OB})}.$$

Keeping the values:

$$P(\text{OR} \mid \text{SR}) = \frac{0.99 \cdot 0.01}{0.99 \cdot 0.01 + 0.02 \cdot 0.99} = \frac{0.0099}{0.0099 + 0.0198} = \frac{0.0099}{0.0297} = \frac{1}{3} \approx 0.3333.$$

Conclusion & Lawyer's Argument

Given the report, the probability that the bus was actually red is only

$$\boxed{P(\text{OR} \mid \text{SR}) = \tfrac{1}{3}}.$$

Therefore it is still twice as likely the bus was originally blue ($P(\text{OB} \mid \text{SR}) = \frac{2}{3}$). So we cannot guarentee that the bus was actually red based on the eyewitness report alone.

Chapter 5

Exit Poll Prediction

Let event M_A be "declared majority for A".

5.1 Probability for event M_A with 100 voters

Total voters = 100. Supporters: 95 for A and 5 for B. An exit poll samples 3 voters *with replacement*.

If A should be the majority among the three voters, then there will be two cases possible,
case 1: All the three voters vote for A.
case 2: Two voters vote for A and one voter vote for B.

$$P(M_A) = P(3\text{A}) + P(2\text{A,1B}) = \left(\frac{95}{100}\right)^3 + \binom{3}{2}\left(\frac{95}{100}\right)^2\left(\frac{5}{100}\right).$$

Computing each term

$$P(3\text{A}) = \left(\frac{95}{100}\right)^3 = \frac{95^3}{100^3} = 0.857375, \qquad P(2\text{A,1B}) = \binom{3}{2}\left(\frac{95}{100}\right)^2\left(\frac{5}{100}\right) = 0.135375.$$

Result:

$$\boxed{P(M_A) = 0.857375 + 0.135375 = 0.99275 \approx 99.275\%}.$$

5.2 Probability of event M_A with 10,000 voters

Total eligible voters = 10,000. Fraction favouring A: 95%, favouring B: 5%. An exit poll samples 3 voters *with replacement*. Let event M_A be "declared majority for A".

If A should be the majority among the three voters, then there are two cases:
case 1: All the three voters vote for A.
case 2: Exactly two voters vote for A and one voter for B.

$$P(M_A) = P(3\text{A}) + P(2\text{A,1B}) = \left(\frac{95}{100}\right)^3 + \binom{3}{2}\left(\frac{95}{100}\right)^2\left(\frac{5}{100}\right).$$

Computing each term

$$P(3\text{A}) = \left(\frac{95}{100}\right)^3 = \frac{95^3}{100^3} = 0.857375, \qquad P(2\text{A,1B}) = \binom{3}{2}\left(\frac{95}{100}\right)^2\left(\frac{5}{100}\right) = 0.135375.$$

Result:

$$\boxed{P(M_A) = 0.857375 + 0.135375 = 0.99275 \approx 99.275\%}.$$

So, with three truthful responses sampled *with replacement*, the exit poll declares a majority for A with probability about 99.275%.

Chapter 6

The Math of Exit Polls

6) Given, probability p that voters prefer A over B is k/m, ie, k people out of m prefered A over B.

* $q(S) = \sum_{i \in I(S)} x_i/n$ where * $x_i = 1$ if i^{th} voter voted for A, else 0.

  * $I(S)$ is the index set of each voter in S.

  * S is all randomly chosen (with replacement) subset containing n voters.

a) Prove that $\sum_S \frac{q(S)}{m^n} = p$.

let,

The value of $q(S) = \frac{n-i}{n}$ for $i \in [0, n]$.

The no. of times $q(S)$ appears in the summation $= {}^nC_i\,(m-k)^i \cdot k^{n-i}$ — ①

Justification of ①,

We have to chose $(n-i)$ voters who prefers A & i voters who prefers B.

$(m-k)^i \Rightarrow$ chosing voters who prefered B

$k^i \Rightarrow$ Chosing voters who prefered A.

${}^nC_i \Rightarrow$ The order in which A, B voters are chosen.

So,

$$\sum_S \frac{q(S)}{m^n} = \frac{\sum_{i=0}^{n} \left(\frac{n-i}{n}\right) \cdot {}^nC_i \cdot (m-k)^i \cdot k^{n-i}}{m^n}$$

$$(x+y)^n = \sum_{i=0}^{n} {}^nC_i \cdot x^{n-i} \cdot y^i$$

Differentiate on both sides w.r.t x.

$$n \cdot (x+y)^{n-1} = \sum (n-i) \cdot {}^nC_i \, x^{n-i-1} \cdot y^i$$

$$n \cdot x \cdot (x+y)^{n-1} = \sum_{i=0}^{n} (n-i) \cdot {}^nC_i \cdot x^{n-i} \cdot y^{i}$$

Substitute $x = K$ and $y = m-K$

$$n \cdot K \cdot (m-\cancel{K}+\cancel{K})^{n-1} = \sum_{i=0}^{n} (n-i) \cdot {}^nC_i \, K^{n-i} \cdot (m-K)^{i}$$

$$\frac{K \cdot m^n}{m} = \sum_{i=0}^{n} \left(\frac{n-i}{n}\right) \cdot {}^nC_i \cdot K^{n-i} \cdot (m-K)^{i}$$

$$\frac{\sum_{i=0}^{n} \left(\frac{n-i}{n}\right) \cdot {}^nC_i \cdot K^{n-i} \cdot (m-K)^{i}}{m^n} = \sum_{S} \frac{q(s)}{m^n} = \frac{K}{m} = P.$$

) Prove that,

$$\sum_{S} \frac{q^2(s)}{m^n} = \frac{P}{n} + \frac{P^2(n-1)}{n}.$$

Using the results from a,

$$\rightarrow \sum_{S} \frac{q^2(s)}{m^n} = \frac{\sum_{i=0}^{n} \left(\frac{n-i}{n}\right)^2 \times {}^nC_i \times (m-i)^{k} \times K^{n-i}}{m^n}$$

$$\rightarrow n \cdot x \cdot (x+y)^{n-1} = \sum_{i=0}^{n} (n-i) \cdot {}^nC_i \cdot x^{n-i} \cdot y^{i}$$

Differentiate w.r.t x,

$$n \cdot (x+y)^{n-1} + n(n-1) \cdot (x+y)^{n-2} \cdot x = \sum_{i=0}^{n} (n-i)^2 \cdot {}^nC_i \cdot x^{n-i-1} \cdot y^{i}$$

Substitute $x = K$ and $y = m-K$,

$$K \left[n \cdot m^{n-1} + n(n-1) K \cdot m^{n-2} \right] = \sum_{i=0}^{n} (n-i)^2 \cdot {}^nC_i \cdot K^{n-i} \cdot (m-K)^{i}.$$

$$\frac{\sum_{i=0}^{n} (n-i)^2 \cdot {}^nC_i \cdot (m-k)^i \cdot k^{n-i}}{n^2 \cdot m^n} = \frac{k\left[n \cdot m^{n-1} + n(n-1)\cdot m^{n-2} \cdot k\right]}{n^2 \cdot m^n}$$

$$= \left(\frac{k}{m}\right)/n + \left(\frac{n-1}{n}\right)\frac{k^2}{m^2}$$

$$= \frac{p}{n} + p^2\left(\frac{n-1}{n}\right).$$

$$\therefore \sum_s \frac{q^2(s)}{m^n} = \sum_{i=0}^{n} \frac{\left(\frac{n-i}{n}\right)^2 \cdot {}^nC_i \cdot (m-k)^i \cdot k^{n-i}}{m^n} = \frac{p}{n} + \frac{p^2(n-1)}{n}$$

) Prove that. $\sum_s \frac{(q(s)-p)^2}{m^n} = \frac{p(1-p)}{n}$.

$$\sum_s \left(\frac{q^2(s)}{m^n} + \frac{p^2}{m^n} - \frac{2pq(s)}{m^n}\right) = \left[\frac{p}{n} + \frac{p^2(n-1)}{n}\right] + \frac{p^2}{m^n} \times m^n - 2p \cdot p$$

$$= \frac{p}{n} + \cancel{p^2} - \frac{p^2}{n} + \cancel{p^2} - \cancel{2p^2}$$

$$= \frac{p(1-p)}{n}$$

$$\therefore \sum_s \frac{(q(s)-p)^2}{m^n} = \frac{p(1-p)}{n}$$

3) P.T, proposition of n-sized subsets S, for which $|q(S) - p| > \delta$ is less than or equal to $\frac{1}{\delta^2} \cdot \frac{p(1-p)}{n}$.

Using Two sided Chebyshev's inequality.

If $S_k = \{ x_i : |x_i - \mu| > k\sigma \}$ then $\frac{S_k}{n} \le \frac{1}{k^2}$.

Here,

let $P_\delta = \{ q(S) : |q(S) - p| > \delta \}$ where S is a n-sized subset.

From ⓐ, p is the mean of $q(S)$ over all subsets.

Variance (σ^2) of $q(S)$ = $$\frac{\sum_S (q(S) - p)^2}{m^n - 1}$$

$$\boxed{\sigma^2 = \frac{p(1-p)}{n(m^n - 1)}}$$

Replacing δ with $\left(\frac{\delta}{\sigma}\right)\sigma$, $P_\delta = \left\{ q(S) : |q(S) - p| > \left(\frac{\delta}{\sigma}\right)\sigma \right\}$

From Chebyshev's inequality,

proportion of S for which $|q(S) - p| > \delta$ is

$$\le \frac{1}{k^2} = \frac{\sigma^2}{\delta^2} = \frac{p(1-p)}{\delta^2 \cdot n \cdot (m^n - 1)}$$

$$\le \frac{1}{\delta^2} \cdot \frac{p(1-p)}{n}$$

Significance of this result:

$q(S)$ is the probability that voters in a randomly taken subset S (size n) prefers A over B. According to this result, $|q(S) - p| \gg \delta$ has very less proportion.

So, $q(S)$ will be close to p in most of the chosen subsets.

∴ The exit poll may have a good prediction of the election result.