# DAI
# Assignment 1

Gurnoor Sohal (24B1042)
Sanvi Jain (24B1046)
Rishika Dhiman (24B1090)

August 18, 2025

## Contents

**1 Instructions for running the code** **2**

**2 Question 1** **3**

**3 Question 2** **6**

3.1 Mean . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
3.2 Median . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
3.3 Standard Deviation . . . . . . . . . . . . . . . . . . . . . . . . . . 7
3.4 Updating Histogram . . . . . . . . . . . . . . . . . . . . . . . . . . 8

**4 Question 3** **9**

**5 Question 4** **10**

**6 Question 5** **11**

**7 Question 6** **12**

# 1 Instructions for running the code

The extracted folder contains six files:

```
A1-24B1042-24B1046-24B1090/
├── Assignment1_Report.pdf
├── Question1_30.m
├── Question1_60.m
├── UpdateMean.m
├── UpdateMedian.m
└── UpdateStd.m
```

**Description of files:**

- `Assignment1_Report.pdf`: The report file.
- `Question1_30.m`: Contains the solution for Question 1 with $f = 30\%$.
- `Question1_60.m`: Contains the solution for Question 1 with $f = 60\%$.
- `UpdateMean.m`: Function for updating the mean (Question 2).
- `UpdateMedian.m`: Function for updating the median (Question 2).
- `UpdateStd.m`: Function for updating the standard deviation (Question 2).

## 2 Question 1

The quartile method produced the least relative mean squared error.

The noise created were only in the positive direction and were high in magnitude compared to the original data. This gave positive spikes in the data.

The quartile method produces the least squares error as there is a lower chance of 75% of the data in the neighborhood to be corrupted. It discards the upper 75% of the data and thereby removing the noise along with it. This claim is further strengthened by decreasing the percentile and comparing the RMSE values. The lower percentiles give lower RMSE as it is able to suppress noise even better.

The other methods produce higher RMSE as:

- Mean filtering averages the noise and smoothens the wave. This produces the highest error as it averages the outliers.
- Median filtering removes symmetric noises and gives moderate error as our data has positively skewed noise. This would have been a better option when the data would have both positive and negative noise.

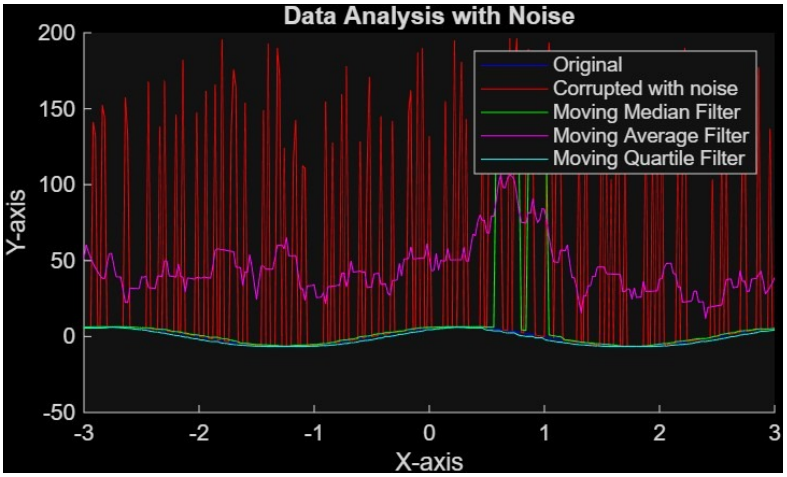

Figure 1: f=30%
Relative Mean Squared Errors:
Median filter: 43.305964
Mean filter: 108.072543
25th percentile filter: 0.087176

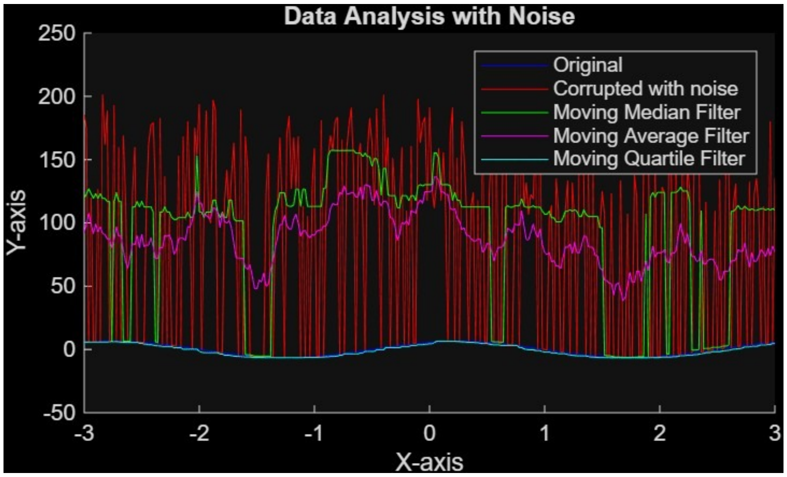

Figure 2: f=60%
Relative Mean Squared Errors:
Median filter: 533.963233
Mean filter: 379.750740
25th percentile filter: 0.059907

```
x = [-3:0.02:3];
y = 6.5*sin(2.1*x + pi/3);

z = y;

fraction = 0.3;
n = length(y);
num_points = round(n*fraction);
indices = randperm(n, num_points);
y_fraction = y(indices);

a = 100; b = 200;
r_noise = a + (b-a).*rand(1, num_points);

z(indices) = z(indices) + r_noise;

y_median = zeros(size(z));
y_average = zeros(size(z));
y_quartile = zeros(size(z));

for i = 1:n
    start = max(1, i-8);
    finish = min(n, i+8);

    Ni = z(start:finish);
    y_median(i) = median(Ni);
    y_average(i) = mean(Ni);
    y_quartile(i) = prctile(Ni, 5);
end

figure;
hold on;
plot(x, y, 'b');
plot(x, z, 'r');
plot(x, y_median, 'g');
plot(x, y_average, 'm');
plot(x, y_quartile, 'c');
xlabel('X-axis');
ylabel('Y-axis');
legend('Original', 'Corrupted with noise', 'Moving Median Filter', 'Moving Average Filter', 'Moving Quartile Filter');
title('Data Analysis with Noise');
hold off;

denominator = sum(y.^2);
rmse_median = sum((y - y_median).^2)/denominator;
rmse_average = sum((y - y_average).^2)/denominator;
rmse_quartile = sum((y - y_quartile).^2)/denominator;

fprintf('RMSE: median = %f, average = %f, quartile = %f', rmse_median, rmse_average, rmse_quartile)
```

# 3 Question 2

## 3.1 Mean

```
function newMean = UpdateMean(OldMean, NewDataValue, n)
    newMean = (OldMean * n + NewDataValue) / (n + 1);
end
```

Mean of a set of $n$ numbers is given by

$$Mean = \frac{S}{n} \tag{1}$$

Where S is the sum of all numbers in the set.

Let OldSum be the sum of the numbers in the array in A
then,

$$OldMean = \frac{OldSum}{n}$$

Therefore,

$$OldSum = OldMean * n \tag{2}$$

NewSum will be given by,

$$NewSum = OldSum + NewDataValue$$

Using equation(2),

$$NewSum = OldMean * n + NewDataValue \tag{3}$$

and by equation(1),

$$NewMean = \frac{NewSum}{(n+1)}$$

Therefore using equation(3) ,

$$NewMean = \frac{OldMean * n + NewDataValue}{(n+1)}$$

## 3.2 Median

```
function newMedian = UpdateMedian(oldMedian, NewDataValue, A, n)
    if mod(n,2) == 0
        if  NewDataValue < A(n/2)
            newMedian = A(n/2);
        elseif NewDataValue > A(n/2 + 1)
            newMedian = A(n/2 + 1);
        else
            newMedian = NewDataValue;
        end
    else
        if NewDataValue >= A((n+1)/2 - 1) && NewDataValue
            <= A((n+1)/2 + 1)
            newMedian = (A((n+1)/2) + NewDataValue) / 2;
        elseif N < A((n+1)/2 - 1)
            newMedian = (A((n+1)/2 -1 ) + oldMedian)/2;
        else
            newMedian = (oldMedian + A((n+1)/2 + 1)) / 2;
        end
    end
end
```

**Case 1: $n$ is even**

When $n$ is even, $n+1$ will be odd and in that case for $n+1$ values the median will be equal to the ${\frac{n+2}{2}}^{\text{th}}$ value. If the new value is less than the current ${\frac{n}{2}}^{\text{th}}$ value, then the current ${\frac{n}{2}}^{\text{th}}$ value will become the new $\left(\frac{n}{2}+1\right)^{\text{th}}$ value and hence will be the median. If the new value is greater than the original $\left(\frac{n}{2}+1\right)^{\text{th}}$ value, then the current and the new $\left(\frac{n}{2}+1\right)^{\text{th}}$ values will remain the same and it will be the median. Lastly, if the new value lies between the current ${\frac{n}{2}}^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ values, then it will be the new $\left(\frac{n}{2}+1\right)^{\text{th}}$ value and hence will be the median.

**Case 2: $n$ is odd**

When $n$ is odd, $n+1$ will be even, and in that case for $n+1$ values the median will be the average of the ${\frac{n+1}{2}}^{\text{th}}$ and $\left(\frac{n+1}{2}+1\right)^{\text{th}}$ values. If the new value lies between the current $\left(\frac{n+1}{2}-1\right)^{\text{th}}$ and $\left(\frac{n+1}{2}+1\right)^{\text{th}}$ values (inclusive), then the middle two values will be the current ${\frac{n+1}{2}}^{\text{th}}$ value and the new value, so the median will be their average. If the new value is less than the current $\left(\frac{n+1}{2}-1\right)^{\text{th}}$ value, then the two middle values will be the current $\left(\frac{n+1}{2}-1\right)^{\text{th}}$ value and the old median, and their average will be the new median. If the new value is greater than the current $\left(\frac{n+1}{2}+1\right)^{\text{th}}$ value, then the two middle values will be the old median and the current $\left(\frac{n+1}{2}+1\right)^{\text{th}}$ value, and their average will be the new median.

## 3.3 Standard Deviation

```
function newStd = UpdateStd (OldMean, OldStd, NewMean,
    NewDataValue, n)
    var=((OldStd.^2)*(n-1) + (OldMean.^2)*(n) +
        NewDataValue.^2)/n- NewMean.^2*(n+1)/n;
    newStd=sqrt(var);
end
```

Standard Deviation of a set of $n$ numbers is given by

$$\sigma = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \tag{4}$$

Variance is square of Standard deviation given as

$$\sigma^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} \tag{5}$$

Old sum of squares of $x_i$ is

$$\sum x_i^2 = \sigma^2 * (n-1) + n * \bar{x}^2$$

New sum of squares of $x_i'$ is

$$\sum x_i'^2 = \sigma^2 * (n-1) + n * \bar{x}^2 + newData^2$$

Therefore, new variance is

$$\sigma'^2 = \frac{\sum x_i'^2 - (n+1) * \bar{x'}^2}{n} \tag{6}$$

New standard deviation is

$$\sigma' = \sqrt{\frac{\sigma^2 * (n-1) + n * \bar{x}^2 + newData^2 - (n+1) * \bar{x'}^2}{n}} \tag{7}$$

### 3.4 Updating Histogram

Suppose the original values lie in the interval $[a, b]$. Then there will be two cases based on where the new value lies. In both cases, we do not need to change the bin width as $n + 1 \approx n$ for large $n$.

1. **The new value lies in $[a, b]$:** In this case, the height of the bin where the value lies would just increase by one unit.

2. **The new value does not lie in the interval:** In this case, we need to create another bin in which its value will lie. Suppose the bin width is $d$, then:

   - If it is greater than $b$, we need to find an $n$ such that:

   $$b + n \cdot d \leq \text{newValue} < b + (n+1) \cdot d$$

   - If it is less than $a$, we need to find an $n$ such that:

   $$a - (n+1) \cdot d \leq \text{newValue} < a - n \cdot d$$

   The height of that bin will be one unit.

# 4 Question 3

$$\begin{aligned}
&\text{Given,}\quad P(A) \geq 1 - q_1 \quad \text{and} \quad P(B) \geq 1 - q_2,\\
&\implies P(A^C) = 1 - P(A) \leq 1 - (1 - q_1) = q_1,\\
&\implies P(B^C) = 1 - P(B) \leq 1 - (1 - q_2) = q_2.
\end{aligned}$$

From Bonferroni's identity, taking n=2:

$$\begin{aligned}
&P(A \cap B) \geq 1 - P(A^C) - P(B^C),\\
&\implies P(A \cap B) \geq 1 - q_1 - q_2,\\
&\implies P(A \cap B) \geq 1 - (q_1 + q_2).
\end{aligned}$$

# 5 Question 4

Given a total of 100 buses out of which 1 is red and 99 are blue. Let the event that a bus is red be denoted as $B_{red}$ and that a bus is blue be denoted as $B_{blue}$.

Given this information, we know

$$P(B_{red}) = \frac{1}{100} \quad \& \quad P(B_{blue}) = \frac{99}{100}$$

Let the event that the person sees a red object be $S_{red}$ and a blue object be $S_{blue}$. It is given that, the person XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. This is the case of conditional probability, which can be represented as:

$$P(S_{red} \mid B_{red}) = \frac{99}{100} \quad \& \quad P(S_{red} \mid B_{blue}) = \frac{2}{100}$$

Now, the probability that the bus is really a red one given that XYZ claims to have seen a red one on that fateful night is $P(B_{red} \mid S_{red})$. Using Bayes' Theorem which is given as:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{\sum_i P(B \mid A_i)\, P(A_i)}$$
$$\text{Here,} \quad A = B_{\text{red}}, \quad B = S_{\text{red}},$$
$$\text{and} \quad A_i \in \{B_{\text{red}}, B_{\text{blue}}\}.$$

We get:

$$\begin{aligned} P(B_{\text{red}} \mid S_{\text{red}}) &= \frac{P(S_{\text{red}} \mid B_{\text{red}})\, P(B_{\text{red}})}{P(S_{\text{red}} \mid B_{\text{red}})\, P(B_{\text{red}}) + P(S_{\text{red}} \mid B_{\text{blue}})\, P(B_{\text{blue}})} \\ &= \frac{\frac{99}{100} \cdot \frac{1}{100}}{\frac{99}{100} \cdot \frac{1}{100} + \frac{2}{100} \cdot \frac{99}{100}} \\ &= \frac{\frac{99}{10000}}{\frac{297}{10000}} = \frac{99}{297} = \frac{1}{3}. \end{aligned}$$

The defense lawyer could contend that **there is only a 33% probability that the bus XYZ claims to have seen on that fateful night was, in fact, a red bus.**

# 6 Question 5

Let the probability that a resident favours A be $p_A$ and that of B be $p_B$. Then

$$p_A = 0.95 \tag{8}$$

$$p_B = 0.05 \tag{9}$$

Let $P(n)$ be the probability that n out of the three voters favoured A over B, then the probability that the exit poll declared a majority for A is given by,

$$P = P(3) + P(2) \tag{10}$$

Calculating $P(3)$ and $P(2)$
For P(3) all the voters must prefer A,

$$P(3) = (p_A)^3 \tag{11}$$

$$P(3) = (0.95)^3$$

For P(2), one voter can be chosen in $\binom{3}{1}$ ways, who favours B, and the other two favour A,

$$P(2) = (p_A)^2.(p_B).\binom{3}{1} \tag{12}$$

$$P(2) = (0.95)^2.(0.05).3$$

Plugging in the value of $P(3)$ and $P(2)$ in equation(3),

$$P = (0.95)^3 + (0.95)^2.(0.05).3$$

$$P = 0.992750 \tag{13}$$

The accuracy of the poll does not depend on the number of residents in the village, as long as the value of $p_A$ remains the same. The answer would be the same for 100 and 10,000 residents.

# 7 Question 6

Given that there are $m$ voters in the village and the probability that voters prefer A to b is $p = \frac{k}{m}$.
Also each voter has a unique index number $i \in [1, m]$ and:

$$x_i = \begin{cases} 1 & \text{if the } i^{th} \text{ voter voted for A} \\ 0 & \text{if the } i^{th} \text{ voter voted for B} \end{cases} \tag{14}$$

Also, $q(S)$ is defined as:

$$q(S) = \Sigma_{i \in I(S)} \frac{x_i}{n} \tag{15}$$

## (A)

$$\sum_S \frac{q(S)}{m^n}$$

Note that the number of all possible sets $S$ (where $|S| = n$) is $m^n$. Since we are sampling a set $S$ from these $m^n$ possibilities at random, hence $P(S) = \frac{1}{m^n}$, we get $E[q(S)] = \sum_{\text{all possibilities of S}} q(S) \cdot P(S) = \sum_S q(S) \cdot \frac{1}{m^n}$. Therefore, the quantity we are determining is the expected value of $q(S)$:

$$E[q(S)] = E\left[\frac{\sum_{i=1}^n x_i}{n}\right] = \frac{\sum_{i=1}^n E[x_i]}{n} \quad \text{(linearity of expectation)}$$

But $E[x_i] = p \quad \forall\ 1 \leq i \leq n$, so

$$\frac{\sum_{i=1}^n p}{n} = \frac{n \cdot p}{n} = p$$

**Alternate Method:**

We want to compute:

$$\frac{\sum\limits_S q(S)}{m^n}.$$

Note that:

$$\frac{\sum\limits_S q(S)}{m^n} = \sum_{i=0}^n \frac{i}{n} \times \frac{\text{(number of subsets possible with } i \text{ votes for A)}}{m^n}.$$

A value of $q(S)$ can range from 0 to $\frac{1}{n}$, so it can be represented as $\frac{i}{n}$ where $0 \leq i \leq n$. The summation of $q(S)$ for all subsets $S$ is obtained by multiplying $\frac{i}{n}$ by the number of subsets having that value, and division of the number of subsets by $m^n$ (the total number of subsets) gives the probability of such a subset.

For some $i$, the probability of such a subset is given by:

$$P(i) = \binom{n}{i} p^i (1-p)^{n-i},$$

where $p$ is the probability that a person favours A over B.

Hence:

$$\frac{\sum\limits_S q(S)}{m^n} = \sum_{i=0}^{n} \frac{i}{n} \binom{n}{i} p^i (1-p)^{n-i}.$$

We can write:

$$= \sum_{i=1}^{n} \binom{n-1}{i-1} p^{\,i-1} (1-p)^{n-i} \cdot p$$

(using the identity $\frac{i}{n}\binom{n}{i} = \binom{n-1}{i-1}$).

This becomes:

$$= p \sum_{i=1}^{n} \binom{n-1}{i-1} p^{\,i-1} (1-p)^{n-i}.$$

Using the binomial theorem:

$$= p\left[(p + (1-p))^{n-1}\right] = p \cdot 1^{n-1} = p.$$

Thus:

$$\frac{\sum\limits_S q(S)}{m^n} = p.$$

## (B)

$$\sum_S \frac{q^2(S)}{m^n}$$

Using a similar argument as in (A), we observe that the required quantity is the expected value of $q^2(S)$:

$$E[q^2(S)] = E\left[\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2\right] = E\left[\frac{\sum_{i=1}^n x_i^2 + 2\sum_{1\le i<j\le n} x_i x_j}{n^2}\right]$$

$$= \frac{E\left[\sum_{i=1}^n x_i^2\right]}{n^2} + \frac{E\left[2\sum_{1\le i<j\le n} x_i x_j\right]}{n^2}$$

$$= \frac{\sum_{i=1}^n E[x_i^2]}{n^2} + \frac{2\sum_{1\le i<j\le n} E[x_i x_j]}{n^2}$$

Since $x_i \in \{0,1\}$, we have $x_i^2 = x_i$, and $E[x_i] = p$:

$$= \frac{np}{n^2} + \frac{2\sum_{1\le i<j\le n} E[x_i x_j]}{n^2}$$

Note that $x_i x_j = 1$ only if both are 1, which happens with probability $p \cdot p = p^2$ (sampling with replacement). Also number of pairs of $(i, j)$ are $\binom{n}{2}$ (since $i \neq j$).

Thus:

$$\begin{aligned}
&= \frac{p}{n} + \frac{2 \cdot \binom{n}{2} \cdot p^2}{n^2} \\
&= \frac{p}{n} + \frac{n(n-1)p^2}{n^2} \\
&= p + p^2 \cdot \frac{n-1}{n}
\end{aligned}$$

**Alternate Method:**

We want to prove:

$$\frac{\sum_S q^2(S)}{m^n} = \frac{p}{n} + \frac{p^2(n-1)}{n}.$$

From part (a), the probability of a subset with $i$ voters favouring A is:

$$P(i) = \binom{n}{i} p^i (1-p)^{n-i},$$

and for such a subset:

$$q(S) = \frac{i}{n}.$$

Hence:

$$\frac{\sum_S q^2(S)}{m^n} = \sum_{i=0}^{n} \left(\frac{i}{n}\right)^2 \binom{n}{i} p^i (1-p)^{n-i}.$$

Using the identity:

$$\frac{i}{n}\binom{n}{i} = \binom{n-1}{i-1},$$

so:

$$\frac{\sum_S q^2(S)}{m^n} = \frac{p}{n} \sum_{i=1}^{n} i \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}.$$

$$= \frac{p}{n} \sum_{i=1}^{n} (i-1) \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} + \frac{p}{n} \sum_{i=2}^{n} i \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}.$$

The second term gives:

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

The first term can be simplified using:

$$(i-1)\binom{n-1}{i-1} = (n-1)\binom{n-2}{i-2},$$

which leads to:

$$\frac{p^2}{n} \cdot (n-1) \sum_{i=2}^{n} \binom{n-2}{i-2} p^{i-2}(1-p)^{n-i}.$$

Using the binomial theorem:

$$\sum_{j=2}^{n-2} \binom{n-2}{j-2} p^{j-2}(1-p)^{n-j} = 1,$$

Thus the first term simplifies to:

$$\frac{p^2(n-1)}{n}.$$

Combining both terms:

$$\frac{\sum_S q^2(S)}{m^n} = \frac{p}{n} + \frac{p^2(n-1)}{n}.$$

## (C)

$$\sum_S \frac{(q(S)-p)^2}{m^n}$$

Once again, we are asked to compute the expected value of $(q(S)-p)^2$:

$$E\left[(q(S)-p)^2\right] = E\left[q^2(S) + p^2 - 2p\,q(S)\right]$$

Using linearity of expectation:

$$= E[q^2(S)] + E[p^2] - 2p\,E[q(S)]$$

From (A) and (B):

$$E[q(S)] = p, \quad E[q^2(S)] = \frac{p}{n} + \frac{p^2(n-1)}{n}$$

Thus:

$$E\left[(q(S)-p)^2\right] = \frac{p}{n} + \frac{p^2(n-1)}{n} + p^2 - 2p \cdot p$$
$$= \frac{p}{n} + \frac{np^2 - p^2}{n} + p^2 - 2p^2$$
$$= \frac{p}{n} + \frac{np^2 - p^2}{n} - p^2$$

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

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

**Alternate Method:**

We want to prove:

$$\frac{\sum\limits_{S}(q^2(S) - p)^2}{m^n} = \frac{p}{n} + \frac{p^2(n-1)}{n}.$$

Expanding the squared term:

$$= \frac{\sum_S q(S)^2 + p^2 - 2\,q(S)\,p}{m^n}$$

$$= \frac{\sum_S q(S)^2}{m^n} + \frac{\sum_S p^2}{m^n} - 2p \cdot \frac{\sum_S q(S)}{m^n}$$

$$= \frac{\sum_S q(S)^2}{m^n} + m^n \cdot \frac{p^2}{m^n} - 2p \cdot \frac{\sum_S q(S)}{m^n}$$

From parts (A) and (B):

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

Simplifying:

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

$$\frac{\sum\limits_{S}(q^2(S) - p)^2}{m^n} = \frac{p(1-p)}{n}$$

## (D)

Let $\mathcal{A}$ be the set

$$\mathcal{A} = \{S : |q(S) - p| > \delta\}.$$

Then $|\mathcal{A}|$ is the number of such subsets, and $\frac{|\mathcal{A}|}{m^n}$ will be the proportion (since $m^n$ is the total number of subsets).

We want to prove:

$$\frac{|\mathcal{A}|}{m^n} \leq \frac{1}{\delta^2} \cdot \frac{p(1-p)}{n}.$$

Chebyshev's inequality states that:

> The proportion of sample points $k$ or more standard deviations away from the sample mean is less than or equal to $\frac{1}{k^2}$.

In our case,

$$\epsilon = k \cdot \sigma \quad \Rightarrow \quad k = \frac{\epsilon}{\sigma}.$$

Hence,

$$\frac{|\mathcal{A}|}{m^n} \leq \frac{\sigma^2}{k^2}.$$

Now, $\sigma^2$ is nothing but the variance:

$$\sigma^2 = \sum_S \frac{(q(S) - p)^2}{m^n},$$

which we already proved in part (c) to be equal to:

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

Therefore, we conclude:

$$\frac{|\mathcal{A}|}{m^n} \leq \frac{1}{\delta^2} \cdot \frac{p(1-p)}{n}.$$

Indeed, this proportion is quite small. Thus, it reinforces the **Weak Law of Large Numbers**, which states that the sample average converges to the true population average. This result is also relevant in real-life exit polls, where a large random sample allows us to make more confident inferences about an entire population without having to survey everyone.