Data Analysis and Interpretation
ASSIGNMENT - 1
REPORT

Shreshtha Gupta (Roll No. 24B1033)
Manvi Mehta (Roll No. 24B1059)
Uma Kumari (Roll No. 24B1026)

Contents

1 Instructions to run our code

1.1 Submission Format

```
A1-24B1033-24B1059-24B1026/
├── Question1/
│   └── q1.m
├── Question2/
│   ├── UpdateMean.m
│   ├── UpdateMedian.m
│   └── UpdateStd.m
└── report.pdf
```

1.2 Detailed Instructions

- First unzip the submission folder.
- To see the results and solutions open report.pdf.
- To view the code for question 1 open the folder Question1.
- Similarly, to view the codes for question 2, open the folder Question2.
- The graph and results for question 1 and question 2 are given in report.pdf.

2 Question 1

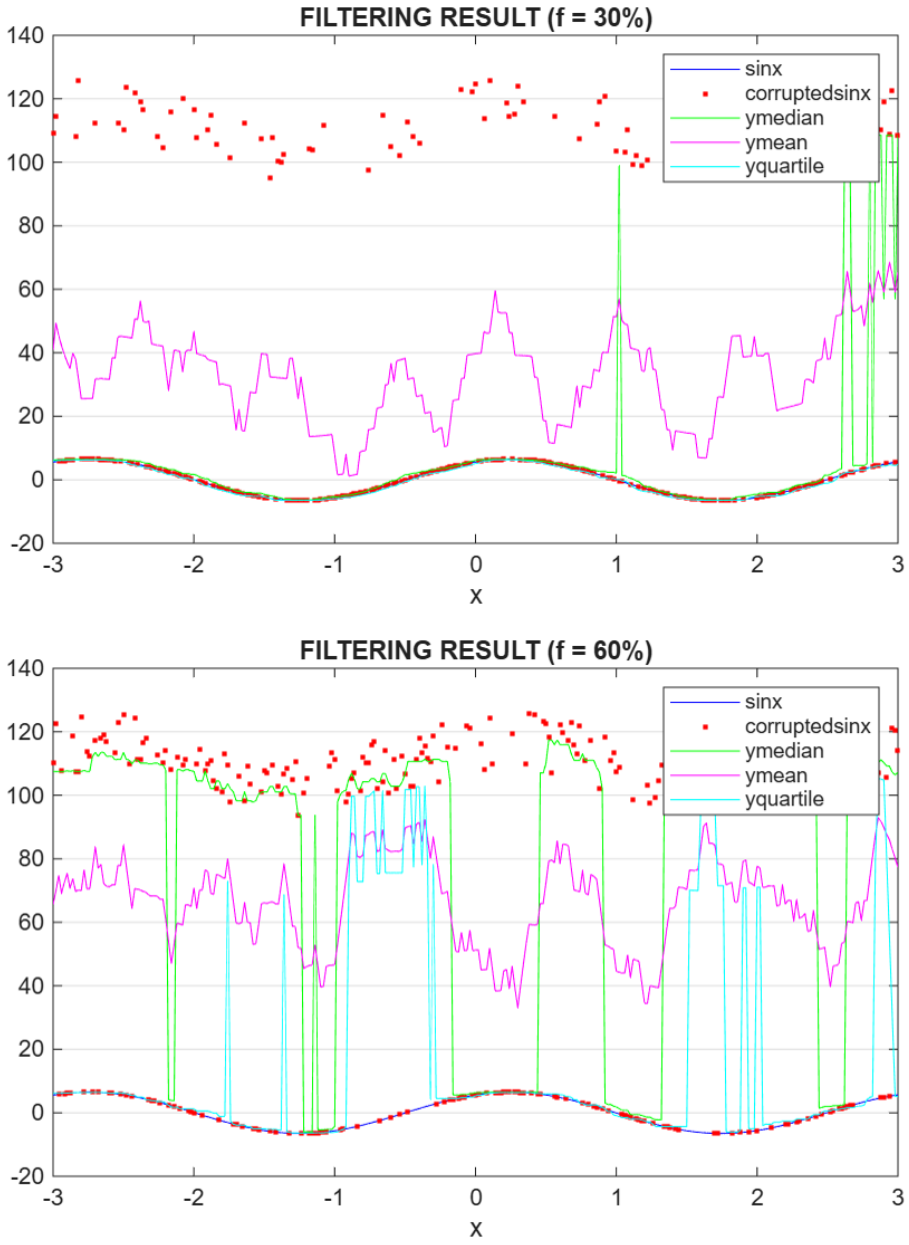
2.1 Graphs and Interpretation

2.1.1 Fraction f = 30%

- The mean squared value between y and y_{median} is **20.492**.
- The mean squared value between y and y_{mean} is **58.446**.
- The mean squared value between y and $y_{quartile}$ is **0.013**.

2.1.2 Fraction f = 60%

- The mean squared value between y and y_{median} is **393.333**.
- The mean squared value between y and y_{mean} is **216.909**.
- The mean squared value between y and $y_{quartile}$ is **66.253**.



2.2 Analysis

As seen with various random values, we saw **Quartile Mean** produced better relative mean squared error.

- **Why is quartile better than mean? - More robust to outliers**
  Mean is highly sensitive to extreme values or outliers as compared to quartile, and henceforth gives a large MSE.

- **Why is quartile better than median? - Upper Tail Contamination**
  If the fraction of the corrupted data is considerably low (approximately $< 20\%$), then the median produces better relative mean squared error. But, for the given values of fraction of corruption(f=30% and f=60%), quartile seemed to produce lower relative mean squared error.
  The main reason for this is that the corruption in the given problem is upper tail,

i.e. noise or outliers are big positive values (100 to 120 in our case). A lower quartile ignores upper tail outliers more so compared to the median, and thus is affected lesser. Median works better for a lower fraction of corruption because the median is closer to the original array's values and thus, when not significantly affected by upper tail contaminators, it performs more accurately.

3 Question 2

3.1 Formulaes Used

3.1.1 How is the Mean updated?

$$\text{OldMean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sum_{i=1}^{n} x_i = n \times \text{OldMean}$$

$$\text{New sum of all numbers} = n \times \text{OldMean} + \text{NewDataValue}$$

$$\text{NewMean} = \frac{\text{New sum of all numbers}}{n+1}$$

$$\text{NewMean} = \frac{n \times \text{OldMean} + \text{NewDataValue}}{n+1}$$

3.1.2 How is the Median updated?

If n is even:

$$\text{newMedian} = \begin{cases} A\left(\frac{n}{2}\right), & \text{if NewDataValue} \leq A\left(\frac{n}{2}\right) \\ A\left(\frac{n}{2}+1\right), & \text{if NewDataValue} \geq A\left(\frac{n}{2}+1\right) \\ \text{NewDataValue}, & \text{otherwise} \end{cases}$$

If n is odd:

$$\text{newMedian} = \begin{cases} \frac{A\left(\frac{n-1}{2}\right) + \text{oldMedian}}{2}, & \text{if NewDataValue} \leq A\left(\frac{n-1}{2}\right) \\ \frac{A\left(\frac{n+3}{2}\right) + \text{oldMedian}}{2}, & \text{if NewDataValue} \geq A\left(\frac{n+3}{2}\right) \\ \frac{\text{NewDataValue} + \text{oldMedian}}{2}, & \text{otherwise} \end{cases}$$

(Here A(i) has 1-based indexing)

3.1.3 How is the Standard Deviation updated?

$$(\text{OldStd})^2 = \frac{\sum_{i=1}^{n}(x_i - \text{OldMean})^2}{n-1}$$

where $x_i =$ numbers in original array A

$$(\text{OldStd})^2 = \frac{\sum_{i=1}^{n}(x_i^2 + (\text{OldMean})^2 - 2x_i(\text{OldMean}))}{n-1}$$

$$= \frac{(\sum_{i=1}^{n} x_i^2) + n(\text{OldMean})^2 - 2(\text{OldMean})\sum_{i=1}^{n} x_i}{n-1}$$

$$= \frac{(\sum_{i=1}^{n} x_i^2) + n(\text{OldMean})^2 - 2(\text{OldMean})n(\text{OldMean})}{n-1} \quad \left(\text{OldMean} = \frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n(\text{OldMean})^2}{n-1} \quad \cdots(1)$$

$$\sum_{i=1}^{n} x_i^2 = \text{sum of squares of numbers in original array}$$

$$= (n-1)(\text{OldStd})^2 + n(\text{OldMean})^2$$

New sum of squares of numbers in original array

$$= \sum_{i=1}^{n} x_i^2 + (\text{NewDataValue})^2$$

$$= (n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{NewDataValue})^2$$

$$(\text{NewStd})^2 = \frac{\sum_{i=1}^{n+1} x_i^2 - (n+1)(\text{NewMean})^2}{n} \quad \text{(replacing } n \text{ by } n+1 \text{ in eq.(1))}$$

$$= \frac{(n-1)(\text{OldStd})^2 + n(\text{OldMean})^2 + (\text{NewDataValue})^2 - (n+1)(\text{NewMean})^2}{n}$$

3.2 An Example

For example, consider the set of 5 numbers:

$$[1, 3, 5, 7, 9]$$

and a new data value added:

$$\{6\}$$

Old mean:

$$\text{OldMean} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1+3+5+7+9}{5} = 5.00$$

New mean:

$$\text{newMean} = \frac{n \times \text{OldMean} + \text{NewDataValue}}{n+1}$$

$$= \frac{5 \times 5 + 6}{6} = \frac{31}{6} = 5.1667$$

Old median:

$$\text{OldMedian} = 5$$

Since $3 < 5 < 7$, the new median is:

$$\text{newMedian} = \frac{\text{NewDataValue} + \text{OldMedian}}{2} = \frac{6+5}{2} = 5.5$$

Old standard deviation:

$$(\text{Old Std})^2 = \frac{\sum_{i=1}^{n}(x_i - \text{OldMean})^2}{n-1}$$
$$= \frac{4^2 + 2^2 + 0^2 + 2^2 + 4^2}{4} = 10$$
$$\text{OldStd} = \sqrt{10} = 3.1623$$

New standard deviation:

$$(\text{New Std})^2 = \frac{(n-1)(\text{Old Std})^2 + n(\text{OldMean})^2 + (\text{NewDataValue})^2 - (n+1)(\text{NewMean})^2}{n}$$
$$= \frac{4(10) + 5(5)^2 + 6^2 - 6(5.1667)^2}{5}$$
$$= 2.8577$$

```
>> hw1_q2
Old Mean: 5.0000, New Mean: 5.1667
Old Median: 5.0000, New Median: 5.5000
Old Std: 3.1623, New Std: 2.8577
```

Figure 1: Command window output of our code

3.3 How is the Histogram of A updated?

When we update a histogram to include a new value, then we have to loop through the lower bounds of every bin in the histogram and when we find a lower bound of a bin which is larger than the value, then we add the value to the bin below that. If the value is lower than the lower limits of all the bins, then we can create a bin below all others for this element. If the element is larger than all the lower limits, then we have to check if it is smaller than the highest limit of the highest bin. If it is lesser, then we add the element to the highest bin otherwise we create a bin higher than all other bins for this element

4 Question 3

Considering two events A and B.
Given:

$$P(A) \geq 1 - q_1 \tag{1}$$

$$P(B) \geq 1 - q_2 \tag{2}$$

To prove:

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

Proof:

From the axioms of probability, we know:

$$P(A) + P(A^c) = 1 \tag{3}$$

Here, A^c denotes the complement of A.

From Eq. (1) and Eq. (3):

$$\begin{aligned} & 1 - P(A^c) \geq 1 - q_1 \\ \Rightarrow \quad & P(A^c) \leq q_1 \end{aligned} \tag{4}$$

Similarly, from Eq. (2) and Eq. (3):

$$\begin{aligned} & 1 - P(B^c) \geq 1 - q_2 \\ \Rightarrow \quad & P(B^c) \leq q_2 \end{aligned} \tag{5}$$

From De Morgan's law:

$$(A \cap B)^c = A^c \cup B^c$$

Equating the probabilities of RHS and LHS:

$$P\big((A \cap B)^c\big) = P(A^c \cup B^c)$$

But from Eq. (3):

$$P\big((A \cap B)^c\big) = 1 - P(A \cap B)$$

Thus:

$$P(A^c \cup B^c) = 1 - P(A \cap B) \tag{6}$$

From Boole's inequality:

$$P(A^c \cup B^c) \leq P(A^c) + P(B^c) \tag{7}$$

Adding Eq. (4) and Eq. (5):

$$P(A^c) + P(B^c) \leq q_1 + q_2 \tag{8}$$

So, from Eq. (7) and Eq. (8), We have:

$$\begin{aligned} & P(A^c \cup B^c) \leq P(A^c) + P(B^c) \leq q_1 + q_2 \\ \Rightarrow \quad & P(A^c \cup B^c) \leq q_1 + q_2 \end{aligned}$$

Substituting value from Eq. (6):

$$\begin{aligned} & 1 - P(A \cap B) \leq q_1 + q_2 \\ \Rightarrow \quad & P(A \cap B) \geq 1 - (q_1 + q_2) \end{aligned}$$

Hence proved.

5 Question 4

Given:

Total number of buses in the town = 100,
Total number of red buses in the town = 1,
Total number of blue buses in the town = 99,
% times XYZ sees red object as red = 99%,
% times XYZ sees blue object as red = 2%.

Let the bus which caused the accident be B_x.

B_{xR} : B_x is a red bus,
B_{xB} : B_x is a blue bus,
XYZ_{RR} : XYZ saw a red bus as red,
XYZ_{BR} : XYZ saw a blue bus as red.

Assuming chances of an accident caused by any bus are equiprobable. So,

$$P(\text{A random bus is } B_x) = \frac{1}{100}$$

As there are 99 blue buses:

$$P(B_{xB}) = 99 \times \frac{1}{100} = \frac{99}{100}.$$

Similarly, for a red bus:

$$P(B_{xR}) = 1 \times \frac{1}{100} = \frac{1}{100}.$$

By Bayes' Rule, We know:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

We need:

$$P(B_{xR} \mid XYZ_{\text{saw a red bus}}) = \frac{P(B_{xR} \cap XYZ_{\text{saw a red bus}})}{P(XYZ\text{saw a red bus})}.$$

To find $P(XYZ\text{saw a red bus})$:

$$P(XYZ\text{saw a red bus}) = P(B_{xR} \cap XYZ_{RR}) + P(B_{xB} \cap XYZ_{BR}).$$

Since bus color and XYZ's observation are independent events. We know for two independent events,

$$P(A \cap B) = P(A) \cdot P(B).$$

Thus:

$$P(B_{xR} \mid XYZ_{\text{saw a red bus}}) = \frac{P(B_{xR}) \cdot P(XYZ_{RR} \mid B_{xR})}{P(B_{xR}) \cdot P(XYZ_{RR}) + P(B_{xB}) \cdot P(XYZ_{BR})}$$

Substituting the values:

$$\begin{aligned} P(B_{xR} \mid XYZ_{\text{saw a red bus}}) &= \frac{\frac{1}{100} \times \frac{99}{100}}{\frac{1}{100} \times \frac{99}{100} + \frac{99}{100} \times \frac{2}{100}} \\ &= \frac{\frac{99}{10^4}}{\frac{99}{10^4} + \frac{198}{10^4}} \\ &= \frac{1}{1+2} \\ &= \frac{1}{3} \\ &\approx 0.33. \end{aligned}$$

Answer: The probability that the bus was really a red one, when XYZ observed it to be red, is 0.33.

6 Question 5

Given:

Number of residents in village = 100,
% residents favouring candidate A = 95%,
% residents favouring candidate B = 5%.

Probability that a randomly chosen villager favours candidate A:

$$P(A) = \frac{95}{100}$$

Probability that a randomly chosen villager favours candidate B:

$$P(B) = \frac{5}{100}$$

We want:

$$P(\text{exit poll is accurate}) = P(\text{expected winner of exit poll is A})$$

For P(expected winner of exit poll is A):

$$P(\text{2 out of 3 people asked favoured A}) + P(\text{3 out of 3 people asked favoured A})$$

$$\begin{aligned} P(\text{accurate}) &= \left(\frac{95}{100} \cdot \frac{95}{100} \cdot \frac{5}{100} \cdot 3\right) + \left(\frac{95}{100} \cdot \frac{95}{100} \cdot \frac{95}{100}\right) \\ &= \left(\frac{95}{100}\right)^2 \left(\frac{15}{100} + \frac{95}{100}\right) \\ &= \left(\frac{95}{100}\right)^2 \cdot \frac{110}{100} \\ &= 0.99275 \end{aligned}$$

Accuracy of the exit poll with 100 residents = 0.99275

—

Now, if

Number of residents in village = 10000, % favouring A = 95%, % favouring B = 5%.

Probability that a randomly chosen villager favours candidate A:

$$P(A) = \frac{95}{100} \cdot \frac{10000}{10000} = \frac{95}{100}$$

Probability that a randomly chosen villager favours candidate B:

$$P(B) = \frac{5}{100} \cdot \frac{10000}{10000} = \frac{5}{100}$$

As the probability of a randomly chosen villager favouring candidate A or B remains the same, i.e., does not depend on the number of villagers:

Accuracy of exit poll is always 0.99275 regardless of number of residents.

7 Question 6

7.1 6(a)

Given:
No. of voters in the village = m.

Probability that people prefer A over $B = p = \frac{k}{m}$.

$$q(S) = \frac{\sum\limits_{i \in S} x_i}{n}$$

To prove:

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

Let A_i denote the subset of S where i people voted for A.

Hence, S is union of all such subsets:

$$S = \bigcup_{i=0}^{n} A_i$$

As given in question:

$$q(S) = \frac{\sum\limits_{i \in S} x_i}{n}$$

where $x_i = 1$ if the i^{th} voter voted for A and 0 if he/she voted for B.

So,

$$q(A_i) = \frac{i}{n}$$

No. of people who prefer A = total voters $\times$ probability of people preferring A:

$$= m \times p$$

$$= pm$$

No. of people who prefer $B = m - pm$.
No. of sets in $A_i = \binom{n}{i}(pm)^i(m-pm)^{n-i}$

$$|A_i| = \binom{n}{i}(pm)^i(m-pm)^{n-i}$$

(because we need to choose i people who prefer A from the n voters of set S)

—

So for all these sets in A_i have:

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

$$\begin{aligned}
\sum_S q(S) &= \sum_{A_0} q(A_0) + \sum_{A_1} q(A_1) + \cdots + \sum_{A_n} q(A_n) \\
&= \frac{0}{n} \cdot \frac{|A_0|}{m^n} + \frac{1}{n} \cdot \frac{|A_1|}{m^n} + \cdots + \frac{n}{n} \cdot \frac{|A_n|}{m^n} \\
&= \frac{0}{n} \cdot \binom{n}{0}(pm)^0(m-pm)^n + \frac{1}{n} \cdot \binom{n}{1}(pm)^1(m-pm)^{n-1} + \cdots + \frac{n}{n} \cdot \binom{n}{n}(pm)^n(m-pm)^0 \\
&= \frac{\sum_{r=0}^{n} r\binom{n}{r}(pm)^r(m-pm)^{n-r}}{n \times m^n} \\
&= \frac{n\sum_{r=1}^{n}\binom{n-1}{r-1}(pm)^r(m-pm)^{n-r}}{n \times m^n} (\text{by } r\binom{n}{r} = n\binom{n-1}{r-1}) \\
&= \frac{n(pm)\sum_{r=1}^{n}\binom{n-1}{r-1}(pm)^{r-1}(m-pm)^{n-r}}{n \times m^n} \\
&= \frac{n(pm)\cdot(pm+m-pm)^{n-1}}{n \times m^n} (\text{by } (y+z)^n = \sum_{r=0}^{n}\binom{n}{r}y^r z^{n-r}) \\
&= \frac{(pm)\cdot m^{n-1}}{m^n} \\
&= \frac{p \cdot m^n}{m^n} \\
&= p
\end{aligned}$$

Hence proved.

7.2 6(b)

To prove:

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

Here, for all sets in A_i,

$$q^2(S) = \big(q(S)\big)^2 = \left(\frac{i}{n}\right)^2 = \frac{i^2}{n^2}$$

$$\begin{aligned}
\frac{\sum\limits_S q^2(S)}{m^n} &= \frac{\sum\limits_{A_0} q^2(A_0)}{m^n} + \frac{\sum\limits_{A_1} q^2(A_1)}{m^n} + \cdots + \frac{\sum\limits_{A_n} q^2(A_n)}{m^n} \\
&= \frac{\sum\limits_{A_0} \frac{0^2}{n^2} + \sum\limits_{A_1} \frac{1^2}{n^2} + \cdots + \sum\limits_{A_n} \frac{n^2}{n^2}}{m^n} \\
&= \frac{\frac{0^2}{n^2}|A_0| + \frac{1^2}{n^2}|A_1| + \cdots + \frac{n^2}{n^2}|A_n|}{m^n} \\
&= \frac{0^2\binom{n}{0}(pm)^0(m-mp)^n + 1^2\binom{n}{1}(pm)^1(m-mp)^{n-1} + \cdots + n^2\binom{n}{n}(pm)^n(m-mp)^0}{n^2m^n} \\
&= \frac{\sum\limits_{r=0}^{n}\binom{n}{r}(pm)^r(m-mp)^{n-r}\cdot r^2}{n^2m^n} \\
&= \frac{n\sum\limits_{r=1}^{n}\binom{n-1}{r-1}(pm)^r(m-mp)^{n-r}\cdot r}{n^2m^n} (\text{by } r\binom{n}{r} = n\binom{n-1}{r-1}) \\
&= \frac{\sum\limits_{r=1}^{n}\binom{n-1}{r-1}(r-1+1)(pm)^r(m-mp)^{n-r}}{nm^n} \\
&= \frac{\sum\limits_{r=1}^{n}\binom{n-1}{r-1}(r-1)(pm)^r(m-mp)^{n-r} + \sum\limits_{r=1}^{n}\binom{n-1}{r-1}(pm)^r\cdot(m-pm)^{n-r}}{nm^n} \\
&= \frac{(n-1)\sum\limits_{r=2}^{n}\binom{n-2}{r-2}(pm)^r(m-mp)^{n-r} + (pm)\sum\limits_{r=1}^{n}\binom{n-1}{r-1}(pm)^{r-1}(m-mp)^{n-r}}{nm^n} \\
&= \frac{(n-1)(pm)^2\sum\limits_{r=2}^{n}\binom{n-2}{r-2}(pm)^{r-2}(m-mp)^{n-r} + (pm)(pm+m-pm)^{n-1}}{nm^n} \\
&= \frac{(n-1)p^2m^2(pm+m-mp)^{n-2} + pm\times m^{n-1}}{nm^n} \\
&= \frac{(n-1)p^2m^2m^{n-2} + pm^n}{nm^n} \\
&= \frac{(n-1)p^2 + p}{n} \\
&= \frac{p}{n} + \frac{(n-1)p^2}{n}
\end{aligned}$$

Hence proved.

7.3 6(c)

To prove:

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

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

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

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

Total no. of subsets containing n voters$(S) = m^n$ (from m people we need to choose n values with replacement)

Using part (a) and (b):

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

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

$$\begin{aligned}\frac{\sum_S (q(S)-p)^2}{m^n} &= \frac{p}{n} + \frac{p^2(n-1)}{n} + \frac{p^2(m^n)}{m^n} - 2p \times p \\ &= \frac{p}{n} + \frac{p^2(n-1)}{n} - p^2\end{aligned}$$

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

Hence proved.

7.4 6(d)

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

$$S_\delta = \{\, |q(S)-p| > \delta \,\}$$

$$\sum_S \frac{(q(S)-p)^2}{m^n} = \sum_{S \in S_\delta} \frac{(q(S)-p)^2}{m^n} + \sum_{S \notin S_\delta} \frac{(q(S)-p)^2}{m^n}$$

$$\sum_{S} \frac{(q(S)-p)^2}{m^n} \geq \sum_{S \in S_\delta} \frac{(q(S)-p)^2}{m^n}$$

$$\frac{p(1-p)}{n} \geq \sum_{S \in S_\delta} \frac{(q(S)-p)^2}{m^n}$$

$$(q(S)-p) > \delta$$

$$\frac{p(1-p)}{n} \geq \frac{\delta^2 \left|S_\delta\right|}{m^n}$$

$$\frac{|S_\delta|}{m^n} \leq \frac{p(1-p)}{n\ \delta^2}$$

The result that the proportion of n-sized subsets S (out of m^n) for which $|q(s)-p| > \delta$ is less than or equal to $\frac{1}{\delta^2}\frac{p(1-p)}{n}$ can be concluded from Chebyshev's inequality. This shows that the number of sets S such that the proportion $q(s)$ differs from p by more than δ is inversely proportional to n. Therefore, for a large value of n, the sample proportion will be very close to p, implying that increasing the number of samples increases the accuracy.