CS 215 - Data Analysis and Interpretation

Assignment 1

Saksham Rathi - 22B1003
Tanmay Gejapati - 22B0969
Aditya Singh - 22B1844

Contents

Question 1

Part (a)

The total number of possible ways in which the books can be picked out of the basket by people are $n!$. (The first person can pick a book in n ways, the second can do so in (n-1) way and for the n^{th} person, one book will be left. So, according to the product rule of counting, the total number of ways will be $n!$). Out of all these, there will be a single way in which everyone gets their own book.

$$P(\text{everyone picking up their own book}) = \frac{\text{Number of ways in which everyone gets their own book}}{\text{Total number of ways possible}} = \frac{\mathbf{1}}{\mathbf{n!}}$$

Part (b)

The total number of possible ways in which the books can be picked out of the basket by people are $n!$. The first m persons can get their own book back again in a single way. After this their will be (n-m) books left. These (n-m) books can be picked by the (n-m) persons left in (n-m)! ways. (The logic used is explained in the previous question.)

E = the first m persons who picked up a book receive their own book back again

$$\text{P(E)} = \frac{\text{Number of ways in which the first m persons received their own book back again}}{\text{Total Number of Ways}} = \frac{\textbf{(n-m)!}}{\textbf{n!}}$$

Part (c)

The total number of possible ways in which the books can be picked out of the basket by people are $n!$. The first person has m options (he has to pick a book which should belong to the last m people) to pick a book. After he picks a book, the second person will have (m-1) options to pick a book. Similarly, the m^{th} person will have only one option left. After this we have (n-m) people and (n-m) books left. Every person can pick any book of his choice, since we have no restrictions on them. So these (n-m) persons can pick the books in (n-m)! ways. The total number of ways will be m! × (n-m)!.

E = Each person among the first m persons to pick up the book gets back a book belonging to one of the last m persons to pick up the books

$$\text{Pr(E)} = \frac{\text{Number of ways in which the conditions are satisfied}}{\text{Total Number of Ways}} = \frac{\textbf{(n-m)!} \times \textbf{m!}}{\textbf{n!}}$$

Part (d)

The probability of the first person picking a clean book is (1- probability of the book being unclean) = (1-p). The second person picking a clean book is again (1-p) because of the independence of the two events. Similarly the m^{th} person has a probability of (1-p) of picking up a clean book. The other (n-m) persons can pick any books as we don't have any restrictions imposed on them. Hence their individual probabilities of picking a book will be 1. By product rule, we can say:
Total Probability = Product of Individual Probabilities = $(1-p) \times (1-p) \times \cdots \times (1-p) \times 1 =$ $\mathbf{(1\text{-}p)^m}$

Part (e)

In this part, we are asked to find the probability of exactly m persons will pick up clean books. Firstly, we have freedom to choose those m people. The number of ways in which we can choose m people out of n persons are ${}^{n}C_{m} = \frac{n!}{m!(n-m)!}$. Now, each m person out of these chosen people should choose a clean book. The probability of each one of them choosing a clean book is (1-p). By product rule, the probability of all of them to choose clean books will be $(1-p)^{m}$. Now, the remaining (n-m) persons should choose unclean books. (Since, we can have exactly m persons choosing clean books). The probability of each one of them choosing an unclean book is p. By product rule, the probability of all of them to choose unclean books will be $(p)^{(n-m)}$. The total probability calculated is:
$\mathbf{{}^{n}C_{m} \times (1\text{-}p)^{m} \times p^{(n\text{-}m)}}$

Question 2

Given: We have n distinct values $\{x_i\}_{i=1}^{n}$ with mean μ and standard deviation σ.
To Prove: $|x_i - \mu| \leq \sigma\sqrt{n-1} \quad \forall i \in \{1, 2, \ldots, n\}$
Proof: From the definition of σ, we have:

$$\text{Variance} = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \mu|^2 \tag{1}$$

Since,

$$|x_i - \mu|^2 \geq 0 \quad \forall i \in \{1, 2, \ldots, n\}$$

(because squares of real numbers are non-negative)
From this we can deduce:

$$\sum_{i=1}^{n}|x_i - \mu|^2 \geq |x_i - \mu|^2 \quad \forall i \in \{1, 2, \ldots, n\} \tag{2}$$

This is because, the sum will be greater than equal to the individual elements because all the terms are non-negative. The equality will hold when all elements are equal.
From the above two equations 1 and 2, we can deduce:

$$\begin{aligned}
\sigma^2 &= \frac{1}{n-1}\sum_{i=1}^{n}|x_i - \mu|^2 && \text{from equation 1} \\
&\geq \frac{1}{n-1}|x_i - \mu|^2 \quad \forall i \in \{1, 2, \ldots, n\} && \text{from equation 2}
\end{aligned}$$

$$\sigma \geq \frac{1}{\sqrt{n-1}}|x_i - \mu| \qquad \text{Taking square root on both the sides}$$

$$\sigma\sqrt{n-1} \geq |x_i - \mu| \qquad \text{Transferring one term from R.H.S. to L.H.S.}$$

Hence we have proved that $|x_i - \mu| \leq \sigma\sqrt{n-1} \quad \forall i \in \{1, 2, \dots, n\}$

The next part of the question is to compare this result with Chebyshev's inequality. According to the above result, all x have to lie in range $[\mu - \sigma\sqrt{n-1}, \mu + \sigma\sqrt{n-1}]$. and there are no elements outside this range, So the probability to find them outside this range is 0. Let us apply the Chebyshev's inequality.

$$S_k = x_i : |x_i - \bar{x}| \geq k\sigma$$

$$\frac{|S_k|}{N} < \frac{1}{k^2}$$

In our case the value of k is $\sqrt{n-1}$. So, the probability that $|x_i - \bar{x}| \geq k\sigma$ will be $\frac{|S_k|}{N}$ (number of elements satisfying the constraint divided by the total number of elements). The value of k in this case is $\sqrt{n-1}$. (comparable with the previous result). Putting this value of k, we will get:

$$P(x_i : |x_i - \bar{x}| \geq \sigma\sqrt{n-1}) < \frac{1}{k^2} = \frac{1}{n-1}$$

The above probability comes out to be zero when calculated from the previous inequality.(as no elements can be out of this range). But with Chebyshev's inequality, the probability is non-zero. Thus, we can say that the Chebyshev's inequality produces a weaker result as compared to the inequality given in the question. Let us see the result as n increases,

$$\lim_{n\to\infty} \frac{1}{n-1} = 0$$

(The probability is upper bounded by $\frac{1}{n-1}$ and lower bounded by 0. Hence, we can apply Sandwich theorem to prove the limit of probability as n increases.) So, as we can see the Chebyshev's inequality approaches the exact behaviour (as proved by the earlier result) as n increases.

Question 3

Given: $\epsilon > 0$, two values Q_1 and Q_2. F be the event that $\{|Q_1 + Q_2| > \epsilon\}$ and let E be the event that $\{|Q_1| + |Q_2| > \epsilon\}$. And, let E_1 and E_2 be the events $\{|Q_1| > \epsilon/2\}$ and $\{|Q_2| > \epsilon/2\}$ respectively.

To Prove: $P(F) \leq P(E_1) + P(E_2)$

Proof: From Venn Diagram, one can easily prove

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) \tag{3}$$

Now, probability of a quantity is always greater than or equal to zero. This will imply

$$P(E_1 \cap E_2) \geq 0$$

Combining this with equation 3, we get,

$$P(E_1) + P(E_2) \geq P(E_1 \cup E_2) \tag{4}$$

We also have,

$$|Q_1| + |Q_2| \geq |Q_1 + Q_2| \tag{5}$$

This will imply that whenever $|Q_1 + Q_2|$ is greater than ϵ, then $|Q_1| + |Q_2|$ will also be greater than ϵ. Hence whenever event F occurs, event E should also occur. This implies F is a subset of E. Hence, we have:

$$P(F) \leq P(E) \quad \text{(Because } F \subseteq E\text{)} \tag{6}$$

Now, if $|Q_1|+|Q_2|$ is greater than ϵ, then atleast one of the quantities $|Q_1|$ or $|Q_2|$ should be greater than $\epsilon/2$. (Because if this does not happen, i.e. both the quantities are less than $\epsilon/2$, the sum will be less than ϵ). This implies, whenever event E happens, one of the events E_1 or E_2 occurs. In other words $E_1 \cup E_2$ occurs. Therefore, we have:

$$P(E) \leq P(E_1 \cup E_2) \quad \text{(Because } E \subseteq E_1 \cup E_2\text{)} \tag{7}$$

Combining the two equations, we get:

$$P(F) \leq P(E_1 \cup E_2) \tag{8}$$

Combining equation 4 with equation 8, we get:

$$P(F) \leq P(E_1) + P(E_2) \tag{9}$$

Hence, we have proved the required result.

Question 4

Given: $P(Q_1 < q_1) \geq 1 - p_1$ and $P(Q_2 < q_2) \geq 1 - p_2$.
Let E be the event that $Q_1 \geq q_1$ and let F be the event that $Q_2 \geq q_2$.
E and F are complementary events of the given events so we know that, $P(E) < p_1$ and $P(F) < p_2$.
$P(E \cup F) \leq P(E) + P(F) < p_1 + p_2$.
Let X be the event that $Q_1Q_2 \geq q_1q_2$ then we can see that $X \subseteq E \cup F$. This is because for $D \in (E \cup F)^c$, $Q_1 < q_1$ and $Q_2 < q_2$, so then $Q_1Q_2 < q_1q_2$ which violates the condition for X.
Thus $P(Q_1Q_2 \geq q_1q_2) = P(X) \leq P(E \cup F) < p_1 + p_2$. Finally we can state that $P(Q_1Q_2 < q_1q_2) = 1 - P(Q_1Q_2 \geq q_1q_2) \geq 1 - (p_1 + p_2)$. Therefore, we have proved the result.

Question 5

Part (a)

In this part we need to calculate the values for $P(C_i|Z_1)$ for $i \in \{1, 2, 3\}$. We observe that for a car being behind any of the doors is independent of the choice of the contestant. So C_i and Z_1 are independent. Thus $P(C_i|Z_1) = \frac{P(C_i \cap Z_i)}{P(Z_i)} = \frac{P(C_i) \times P(Z_i)}{P(Z_i)} = P(C_i)$. C_1, C_2 and C_3 are equally probable, so

$$P(C_1) = \frac{1}{3}$$

$$P(C_2) = \frac{1}{3}$$

$$P(C_3) = \frac{1}{3}$$

Part (b)

Here the host has to open either door 2 or door 3.
For $P(H_3|C_1, Z_1)$ both doors 2 and 3 have stones behind them, so the host may choose either of them with equal probability, thus

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

The car is behind door 2, so the host has to open the door 3 only, hence

$$P(H_3|C_2, Z_1) = 1$$

As the car is behind door 3, the host won't open door 3 thus the probability is

$$P(H_3|C_3, Z_1) = 0$$

Part (c)

We are given that

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

From part (b), we know that $P(H_3|C_2, Z_1) = 1$.
C_2 and Z_1 are independent events, thus $P(C_2, Z_1) = P(C_2) \times P(Z_1)$ and we know that $P(C_2) = \frac{1}{3}$ and $P(Z_1) = \frac{1}{3}$. Thus $P(C_2, Z_1) = \frac{1}{9}$.
For $P(H_3, Z_1)$ we need to use basic probability rules. Total number of possible ways: The contestant can choose a door in 3 ways. After that the host can choose a door in 2 ways(the doors which the contestant did not choose). By the product rule of counting the total number of ways is 6. Now H_3 and Z_1 is one of those cases. So, $P(H_3, Z_1) = \frac{1}{6}$.
Combining all these we can calculate the required probability,

$$P(C_2|H_3, Z_1) = \frac{1 \times \frac{1}{9}}{\frac{1}{6}} = \frac{2}{3}$$

So, the probability of winning by switching our choice is $\frac{2}{3}$.

Part (d)

We need to calculate,

$$P(C_1|H_3, Z_1) = \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3, Z_1)}$$

From part (b), $P(H_3|C_1, Z_1) = \frac{1}{2}$.
With a similar argument as given in part (c), $P(C_1, Z_1) = \frac{1}{9}$.
From part (c), $P(H_3, Z_1) = \frac{1}{6}$.
Combining these we get,

$$P(C_1|H_3, Z_1) = \frac{\frac{1}{2} \times \frac{1}{9}}{\frac{1}{6}} = \frac{1}{3}$$

So, the probability of winning by sticking to our original choice is $\frac{1}{3}$.

Part (e)

From parts (c) and (d), we can see that switching our choice gives us a much higher chance of winning the car. Thus switching is beneficial.

Part (f)

Assumption: The contestant can even switch to a door which the host has opened.

In this part, along with $P(C_1|H_3, Z_1)$ and $P(C_2|H_3, Z_1)$, there is a third case as well, which is $P(C_3|H_3, Z_1)$. Here the first one indicates the case where we stick to our choice, and the other two indicate the cases where we switch our choice. The third comes due to the new case where the host might even open door 3 when the car is behind it.
From the previous parts we know that,

$$P(C_i|H_3, Z_1) = \frac{P(H_3|C_i, Z_1)P(C_i, Z_1)}{P(H_3, Z_1)}$$

Here $P(C_i, Z_1) = \frac{1}{9}$ ∀ i and $P(H_3, Z_1) = \frac{1}{6}$.
$P(H_3|C_i, Z_1) = \frac{1}{2}$ ∀ i because now the host chooses any of the doors 2 and 3 with equal probability.
So, we see that each of the $P(C_i|H_3, Z_1)$ are,

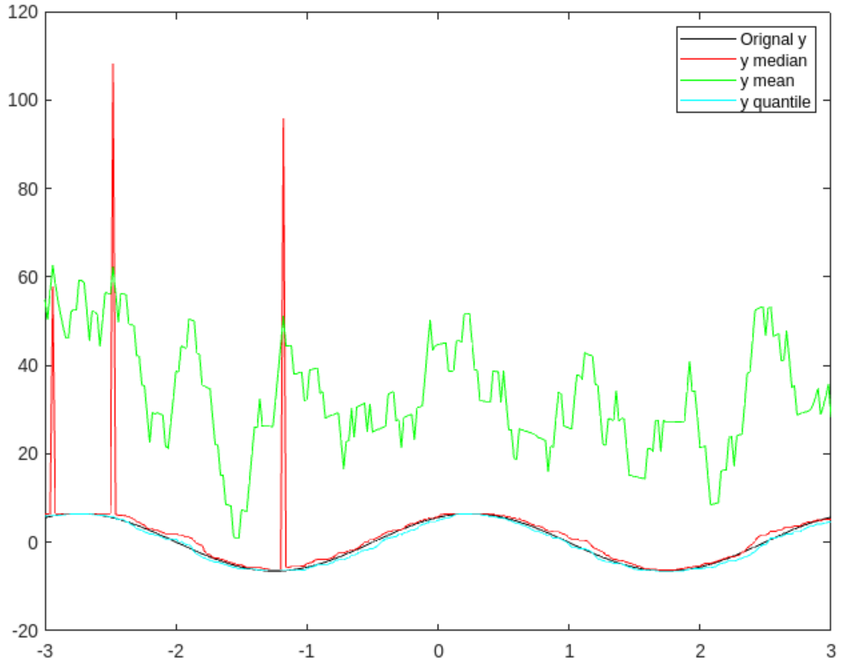
$$P(C_i|H_3, Z_1) = \frac{\frac{1}{2} \times \frac{1}{9}}{\frac{1}{6}} = \frac{1}{3}$$

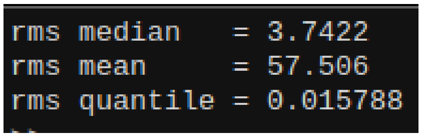
The total probability of winning by switching is: $P(C_2|H_3, Z_1) + P(C_3|H_3, Z_1) = \frac{2}{3}$.
The total probability of winning by sticking to original choice is: $P(C_1|H_3, Z_1) = \frac{1}{3}$.
So we can see that switching of choice is still beneficial and gives a higher chance of winning the car.

Question 6

Code is in file: *a1_p6.m* In this problem, a pure sine wave was first corrupted upto varying degrees and then used some filtering techniques to understand which recovers the most amount of information. For that matter filters such as moving median filter, moving average filter and moving quartile (25th percentile) filter in order to filter out noise in z, which is a corrupted version of the original sine wave y.
In order to analyse their effectiveness, we used the relative mean squared error metric on the filtered value in comparison to the original clean sine wave.
These results and their explanations are given below.

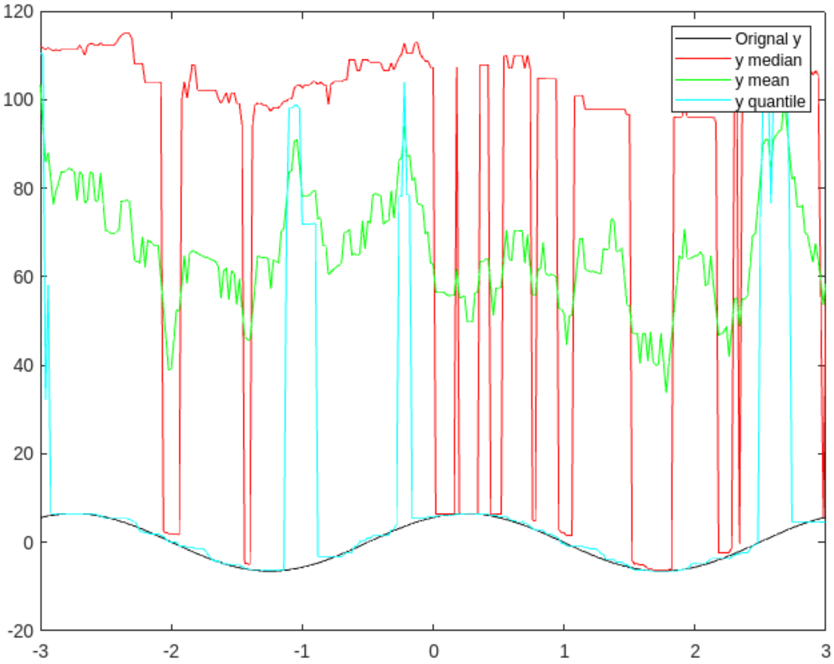
For $f = 30\%$

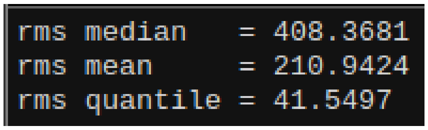




```
rms median   = 3.7422
rms mean     = 57.506
rms quantile = 0.015788
```

For $f = 60\%$





```
rms median   = 408.3681
rms mean     = 210.9424
rms quantile = 41.5497
```

It can be inferred from the computed rms values that the moving quartile filter performs the best, as compared to moving mean or median filters.
Then we observe that moving median filter performed well when $f = 30\%$ but performs really bad when $f = 60\%$. This is because moving median filter is good at eliminating outliers when the data-set is $< 50\%$ corrupted, else it performs bad. This is because once the data-set is corrupted beyond 50%, the moving median essentially captures noise and hence the poor results.
The moving average filter is much more sensitive to outliers than moving median filter hence it performs bad in either of the cases.
The moving quartile filter performs well as it selects the bottom $25th$ percentile, which is less sensitive to outliers than the moving median filter in this case as the outliers are large positive values.

Question 7

Code is in file: $a1\_p7.m$

Updating mean:

$$\bar{x}_{old} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{10}$$

$$\sum_{i=1}^{N} x_i = \bar{x}_{old} \cdot N \tag{11}$$

$$\bar{x}_{new} = \frac{x_{new} + \sum_{i=1}^{N} x_i}{N+1} \tag{12}$$

Using 11 and 12:

$$\bar{x}_{new} = \frac{x_{new} + \bar{x}_{old} \cdot N}{N+1} \tag{13}$$

Updating standard deviation:

$$\sigma_{old}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x}_{old})^2 \tag{14}$$

$$\sigma_{old}^2 = \frac{(\sum_{i=1}^{N} x_i^2) - \bar{x}_{old}^2 \cdot N}{N-1} \tag{15}$$

$$\sum_{i=1}^{N} x_i^2 = \sigma_{old}^2 \cdot (N-1) + \bar{x}_{old}^2 \cdot N \tag{16}$$

$$\sigma_{new}^2 = \frac{(\sum_{i=1}^{N} x_i^2) + x_{new}^2 - \bar{x}_{new}^2 \cdot (N+1)}{N} \tag{17}$$

Using 16 and 17:

$$\sigma_{new}^2 = \frac{(\sigma_{old}^2 \cdot (N-1) + \bar{x}_{old}^2 \cdot N) + x_{new}^2 - \bar{x}_{new}^2 \cdot (N+1)}{N} \tag{18}$$

Updating median:

There are two cases to this problem: when N is even and when it's odd.

N is even

Define two central elements:

```
mid_left = A(n/2);
mid_right = A(n/2 + 1);
```

- If the new element is greater than mid_right, then the new median will be mid_right.
- If the new element is less than mid_left, then the new median will be mid_left.
- Otherwise, the new element itself is the new median.

N is odd

```
i_mid = (n+1)/2;

mid_left = A(i_mid - 1);
mid = A(i_mid);
mid_right = A(i_mid + 1);
```

- If the new element is greater than *mid_right*, then the new median will be the median of *mid* and *mid_right*.
- If the new element is less than *mid_left*, then the new median will be the median of *mid* and *mid_left*.
- Otherwise, if median is between *mid_left* and *mid_right*, then the new median will be the median of itself and *mid*.