CS215: Data Analysis and Interpretation
Assignment 1

Anilesh Bansal, 22b0928
Arihant Vashista, 22b0958
Sanskar Shaurya, 22b0985

14 August 2023

Contents

1 Question 1

The first person picking a book has n choices, the next person has $(n-1)$ choices, and so on ... hence the size of the sample space is $(n) \cdot (n-1) \ldots (2) \cdot (1) = n!$

1.1 (a)

There is only one case where each and every person picks back his/her book hence the probability of that happening is $\frac{1}{n!}$

1.2 (b)

The first m persons pick up their own book, now the $(m+1)^{th}$ person has $(n-m)$ choices, the $(m+2)^{th}$ person has $(n-m-1)$ and so on till the last person has only 1 choice, therefore the number of favorable outcomes $(n-m) \cdot (n-m-1) \ldots (2) \cdot (1) = (n-m)!$. Hence the required probability is $\frac{(n-m)!}{n!}$

1.3 (c)

According to the condition in the question, the first m persons are allowed to pick up books from the set $\{n, n-1, n-2, \ldots n-m+1\}$ while the rest of the $(n-m)$ are allowed to pick books from the set $\{1, 2, 3 \ldots (n-m)\}$. Now the first person has m choices, the next person has $(m-1)$ choices, and so on, till the m^{th} person has 1 choices. The $(m+1)^{th}$ person will again have $(n-m)$ choices, the $(m+2)^{th}$ person has $(n-m-1)$ and so on till the last person has only 1 choice, therefore the number of favorable outcomes is $(m) \cdot (m-1) \ldots (2) \cdot (1)(n-m) \cdot (n-m-1) \ldots (2) \cdot (1) = (n-m)!m!$. Hence the required probability is $\frac{m!(n-m)!}{n!}$

1.4 (d)

Since the probability of picking a clean book is $(1-p)$ irrespective of what other people are choosing (the probability is independent) hence the required probability will be simply $(1-p) \cdot (1-p) \ldots (1-p) = (1-p)^m$

1.5 (e)

Since exactly m person picks up the clean book, we can first select m persons out of the n persons in $\binom{n}{m}$ ways.Now, the rest of the $(n-m)$ people must have picked up an unclean book which will similarly have a probability of p^{n-m} using similar arguments. And since all these events are independent, the final probability is the product that is $\binom{n}{m} \cdot p^{n-m} \cdot (1-p)^m$.

Note: Here we are using the fact that for independent events A and B, $P(A \cap B) = P(A)P(B)$

2 Question 2

It can be easily seen that :-

$$\begin{aligned} \sigma^2(n-1) &= \sum_{i=0}^{n}(x_i - \mu)^2 \\ &\geq (x_i - \mu)^2 \quad \forall i \text{ since rest of the terms are positive} \\ \implies & |x_i - \mu| \leq \sigma\sqrt{n-1} \end{aligned}$$

The *Chebyshev's Inequality* states that $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$. Now putting $k = \sqrt{n-1}$ in this equation to get

$$P(|X-\mu| \geq \sigma\sqrt{n-1}) \leq \frac{1}{n-1}$$

. Since as $n \to \infty$, $\frac{1}{n-1} \to 0$.Now, The above inequality states that

$$P(|X-\mu| \geq \sigma\sqrt{n-1}) = 0$$

. Therefore, the above inequality is just a special case of *Chebyshev's Inequality* as $n \to \infty$. In other words, we can say that as n increases, a special case of *Chebyshev's inequality* takes the form of the inequality in the question.

3 Question 3

Claim: $E \subseteq E_1 \cup E_2$.
Now consider Q_1, Q_2 such that E is true, i.e., $|Q_1| + |Q_2| \geq \epsilon$. Now assume $\bar{E_1} \cap \bar{E_2}$. This means $|Q_1| < \frac{\epsilon}{2}$ and $|Q_2| < \frac{\epsilon}{2}$ which implies $|Q_1| + |Q_2| < \epsilon$ which is a contradiction. Hence the event $E_1 \cup E_2$ is true. Since event E occurs implies event $E_1 \cup E_2$ occurs, we can say that $E \subseteq E_1 \cup E_2$.

Claim: $F \subseteq E_1 \cup E_2$
Assume that event F occurs.This means $|Q_1 + Q_2| \geq \epsilon$.Now, $|Q_1| + |Q_2| \geq |Q_1 + Q_2| \geq \epsilon \implies$ event E occurs. Which means $F \subseteq E \subseteq E_1 \cup E_2$.
Now we can simply show the following:

$$\begin{aligned} & F \subseteq E \subseteq E_1 \cup E_2 \\ \implies & P(F) \leq P(E) \leq P(E_1 \cup E_2) \\ \implies & P(F) \leq P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ & \leq P(E_1) + P(E_2) \end{aligned}$$

4 Question 4

Let E_1 denote the event $Q1 < q1$, E_2 the event $Q2 < q2$ and E the event $Q1 \cdot Q2 < q1 \cdot q2$. Then

$$\begin{aligned} E_1 \cap E_2 & \implies Q1 < q1 \text{ and } Q2 < q2 \\ & \implies Q1 \cdot Q2 < q1 \cdot q2 \text{[1]} \end{aligned}$$

Which means $E_1 \cap E_2 \subseteq E$ and hence $P(E_1 \cap E_2) \leq P(E)$ Now,

$$\begin{aligned} P(E_1) + P(E_2) - 1 & \leq P(E_1 \cap E_2) \text{[2]} \\ \implies 1 - p_1 + 1 - p_2 - 1 & \leq P(E_1 \cap E_2) \leq P(E) \\ \implies P(Q1 \cdot Q2 < q1 \cdot q2) & \geq 1 - (p1 + p2) \end{aligned}$$

5 Question 5

[1] As Q1 and Q2 are non-negative
[2] Bonferroni's Inequality

5.1 (a)

Since the contestant has no way of knowing which door might hold the car and no extra information whatsoever before starting the competition, we may assume that the events Z_1 is **independent of the event** C_i for all i. Hence

$$P(C_i|Z_1) = P(C_i) = \frac{1}{3} \quad \forall i$$

5.2 (b)

Let us consider the 3 cases:
Case I: $i = 1$ In this case, the host can choose H_3 and H_2, and since both of them have stones behind them, either of them will work(i.e. satisfy the condition given in the question) hence

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

Case II: $i = 2$ Here, since the contestant has already chosen the first door and door 2 contains the car, the only choice the host has is to select the 3^{rd} door, the probability,

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

Case III: $i = 3$ Since the third door has the car and the host can't choose the door with the car, this probability has to 0. That is:

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

5.3 (c)

Let us evaluate the terms individually.From part(b), we know that $P(H_3|C_2, Z_1) = 1$. Now,

$$P(C_2, Z_1) = P(C_2|Z_1)P(Z_1) = (\frac{1}{3})(\frac{1}{3}) = \frac{1}{9} \quad \text{(from part (a))} \tag{1}$$

Now we need to calculate $P(H_3, Z_1)$, Now,

$$\begin{aligned}
P(H_3 \cap Z_1) &= P(H_3 \cap Z_1 \cap (\bigcup_i C_i)) \quad \text{Since } C_i \text{ are mutually exclusive and exhaustive} \\
&= \sum_i P(H_3 \cap Z_1 \cap C_i) \\
&= \sum_i P(H3|C_i, Z_1)P(C_i \cap Z_1) \quad \text{from conditional probability definition} \\
&= \sum_i P(H3|C_1, Z_i)P(C_i|Z_1)P(Z_1) \\
&= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{6}
\end{aligned}$$

It is given that the probability of winning by switching is given by

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

5.4 (d)

It is easy to observe that in this case we need to replace C_2 by C_1 in the above formula. From the above parts we can conclude:

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

5.5 (e)

From the above analysis, it is quite clear that switching is beneficial with a probability of winning $\frac{2}{3}$.

5.6 (f)

Let's assume again that the contestant chooses Door 1 , now the probability $P(H_3|C_i, Z_1) = \frac{1}{2}$ for all three cases according to the question because the host can choose any of the remaining doors. Now we can easily calculate the rest of the values as before:

$$\begin{aligned}
P(H_3 \cap Z_1) &= \sum_i P(H3|C_1, Z_i)P(C_i|Z_1)P(Z_1) \\
&= \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3} + \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3} + \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3} = \frac{1}{6} \\
P(C_2|H_3, Z_1) &= \frac{P(H_3|C_2, Z_1)P(C_2, Z_1)}{P(H_3, Z_1)} = \frac{\frac{1}{2}\cdot\frac{1}{9}}{\frac{1}{6}} = \frac{1}{3} \\
P(C_1|H_3, Z_1) &= \frac{P(H_3|C_1, Z_1)P(C_1, Z_1)}{P(H_3, Z_1)} = \frac{\frac{1}{2}\cdot\frac{1}{9}}{\frac{1}{6}} = \frac{1}{3}
\end{aligned}$$

Since switching and not switching have equal probbility of winning , switching is not beneficial to the contestant.

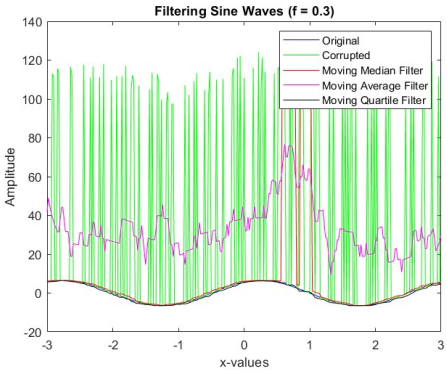
Note: *Here the probability of switching and not switching do not sum up to 1. This is because these events are* ***not exhaustive****, i.e., there is a third case , when the host opens the door with car, in that case neither switching nor retaining the door will lead to a win.*

6 Question 6

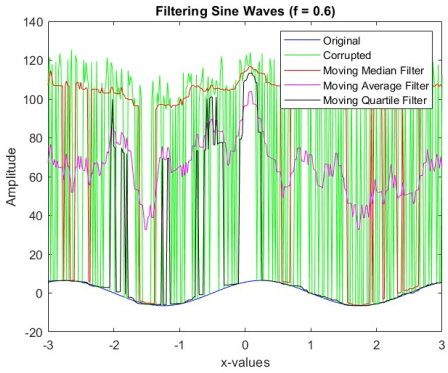
In this question, we created a MATLAB function that takes in a parameter 'f' (fraction of corrupted values in the sine graph) and corrupts a sine wave of the form $y = 6.5 \cdot sin(2.1x + \pi/3)$ where x ranges from -3 to 3 in steps of 0.02. Then it uses a *for* loop to filter the graph in 3 different ways:

1. Moving Median Filtering
2. Moving Average Filtering
3. Moving Quartile Filtering

After it gets all the 3 filtered arrays, it plots them on the graph alongside the original sine wave and the corrupted sine wave. In the script file (Q6Script.m), we have called the function two times for f = 0.3 and f = 0.6. Running the script file, we get the following two plots for the two f-values:



(a) Plot for f = 0.3



(b) Plot for f = 0.6

Figure 1: Plot of filtered sine-waves for different values of 'f'

The function also outputs the relative-mean squared error for each of the three filtering processes. For this run we got:

Filtering Method	f = 0.3	f = 0.6
Moving Median Filter	5.7903	19.9543
Moving Average Filter	7.6036	14.4402
Moving Quartile Filter	0.12015	7.3209

Table 1: Relative-Mean Squared Error for f = 0.3 and f = 0.6

The Moving Average Filter performs the poorest for f = 30%. This is because the mean value is very much affected by outliers and here the way the data is corrupted, we have large outliers and hence even a few outliers in a neighborhood of an element can heavily influence the filtered value we are getting.

The Moving Median Filter performs better than the average filter for f = 30% but worse for f = 60%. This is because for a small number of outliers, the median is completely unaffected no matter how extreme they are, but as soon as the number of corruptions increases, the median value in a neighborhood can be heavily altered from the original value. Hence for f = 60%, the Moving Median Filter performs the poorest.

The Moving Quartile Filter performs the best here as the bottom 25-percentile in a neighborhood are not affected a lot by this corruption. The way we have corrupted this sine curve, the outliers are huge compared to the original values and hence the probability they lie in the bottom 25-percentile is very low, which is why this filtering method has the least relative-mean squared error in both the cases.

Note: The .m files for this question are "FilteringSineWaves.m" and "Q6Script.m".

7 Question 7

Consider the array $A = [x_1, x_2, \cdots, x_n]$ with mean $\bar{x}_{old}$, standard deviation σ_{old} and median M_{old} Let the newDataValue inserted be x_{n+1}. Then we can define functions for new mean, medain and standard deviation as follows :-

7.1 UpdateMean

$$\bar{x}_{old} = \frac{1}{n}\sum_{i=1}^{n} x_i \implies n\bar{x}_{old} = \sum_{i=1}^{n} x_i$$

$$\bar{x}_{new} = \frac{1}{n+1}\sum_{i=1z}^{n+1} x_i = \frac{n\bar{x}_{old} + x_{n+1}}{n+1}$$

7.2 UpdateMedian

$$A = [x_1, x_2, \cdots, x_n]$$

- If n == even

  $$M_{old} = (A[n/2] + A[n/2+1])/2$$

  Depending on the value of x_{n+1} as compared to the values of $A[n/2]$ and $A[n/2+1]$, we can make 3 cases and find M_{new} as

  $$M_{new} = \begin{cases} A[n/2+1] & \text{if } x_{n+1} \geq A[n/2+1] \\ x_{n+1} & \text{if } A[n/2+1] > x_{n+1} > A[n/2] \\ A[n/2] & \text{if } A[n/2] \geq x_{n+1} \end{cases}$$

- If n == odd

  $$M_{old} = A[(n+1)/2]$$

  Depending on the value of x_{n+1} as compared to the values of $A[(n+3)/2]$ and $A[(n-1)/2]$, we can make 3 cases and find M_{new} as

  $$M_{new} = \begin{cases} \frac{A[(n+1)/2]+A[(n+3)/2]}{2} & \text{if } x_{n+1} \geq A[(n+3)/2] \\ \frac{x_{n+1}+A[(n+1)/2]}{2} & \text{if } A[(n+3)/2] > x_{n+1} > A[(n-1)/2] \\ \frac{A[(n-1)/2]+A[(n+1)/2]}{2} & \text{if } A[(n-1)/2] \geq x_{n+1} \end{cases}$$

7.3 UpdateStd

$$(n-1)\sigma^2_{old} = \sum_{i=1}^{n}(x_i - \bar{x}_{old})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2_{old}$$

$$\implies \sum_{i=1}^{n} x_i^2 = n\bar{x}^2_{old} + (n-1)\sigma^2_{old}$$

Now,

$$n\sigma^2_{new} = \sum_{i=1}^{n+1}(x_i - \bar{x}_{new})^2 = \sum_{i=1}^{n+1} x_i^2 - (n+1)\bar{x}^2_{new}$$

$$n\sigma^2_{new} = n\bar{x}^2_{old} + (n-1)\sigma^2_{old} + x^2_{n+1} - (n+1)\bar{x}^2_{new}$$

$$\implies \sigma^2_{new} = \frac{n\bar{x}^2_{old} + (n-1)\sigma^2_{old} + x^2_{n+1} - (n+1)\bar{x}^2_{new}}{n}$$

Note: The .m files for this question are "UpdateMean.m", "UpdateMedian.m","UpdateStd.m" and "q7Script.m".

To update the histogram of A, we can find the bin to which the newDataValue x_{n+1} belong to and can update that value by 1. Alternatively, one can insert the new value in A and calculate the histogram.