

CS 215  
**Homework 1**

Omkar Shirpure  
22B0910

Krish Rakholiya  
22B0927

## 1. Question 1 :

### 1.1

The probability of the every person picking their own notebook is (for the first person to get his own notebook is  $\frac{1}{n}$ , for the second person  $\frac{1}{n-1}$  and so on... )

So their overall probability is

$$P = \frac{1}{n} * \frac{1}{n-1} \dots \frac{1}{2} * \frac{1}{1}$$

Hence,

$$P = \frac{1}{n!}$$

### 1.2

The probability of getting the correct notebooks for the first  $m(<n)$  people could be calculated by same method. W.K.T. (for the first person to get his own notebook is  $\frac{1}{n}$ , for the second person  $\frac{1}{n-1}$  up to  $m$ th person  $\frac{1}{n-m+1}$ )

The overall probability will be

$$P = \left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right)\left(\frac{1}{n-2}\right)\dots\left(\frac{1}{n-m+1}\right)$$

$$P = \frac{1}{(n-m)!}$$

$$P = \binom{n}{m}^{-1}$$

### 1.3

This is conceptually same as the previous one, where just the set of  $m$  changes from first to last. Hence,

$$P = \frac{1}{(n-m)!}$$

$$P = \binom{n}{m}^{-1}$$

### 1.4

The probability of getting clean books for first  $m$  persons is

$$P = (1-p)^m$$

since the probability of book being clean is independent of any other factors.

### 1.5

Since it is given that first  $m$  persons pick up clean books (with probability  $(1-p)$ ) and rest of the  $(n-m)$  people pick up unclean books (with probability  $p$ ). The total probability is

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

## 2. Question 2 :

We are given  $n$  distinct values  $x_i$  with their mean  $\mu$  and standard deviation  $\sigma$  and to prove

$$|x_i - \mu| \leq \sigma \sqrt{n-1}$$

Since we defined the variance as

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

From the equation to prove, squaring on both sides and expanding, it gives,

$$(x_i - \mu)^2 \leq \sigma^2(n-1)$$

$$x_i^2 - 2x_i\mu + \mu^2 \leq \sigma^2(n-1)$$

Taking expectation of both sides,

$$E(x_i^2) - 2\mu E(x_i) + \mu^2 \leq$$

$$E(x_i^2) - \mu^2 \leq \sigma^2(n-1)$$

Since,  $\text{Var} = E(x_i^2) - \mu^2$ , substituting we get,

$$\text{var} \leq \sigma^2(n-1)$$

$$\text{var} \leq \text{var}(n-1)$$

And since, variance is always non-negative, this inequality holds true. And hence the inequality  $|x_i - \mu| \leq \sigma \sqrt{n-1}$  holds true for all individual  $x_i$ .

Coming to comparison with Chebyshev's Inequality as  $n$  increases, the above becomes a more stronger statement compared to Chebyshev's Inequality. Chebyshev's Inequality in general being just the probabilistic bound for all distributions, but the above inequality gives a direct relation between all the individual data points, and other details such as mean, standard deviation, etc. And as  $n$  increases, the proven inequality becomes more accurate and informative for this specific data set as compared to Chebyshev's inequality.

### 3. Question 3 :

We are given that  $\varepsilon > 0$  and two values  $Q_1$  and  $Q_2$  alongside the events  $F$  and  $E$  , ie.

$$F : |Q_1 + Q_2| > \varepsilon$$

$$E : |Q_1| + |Q_2| > \varepsilon$$

$E_1$  and  $E_2$  be the events  $|Q_1| > \varepsilon/2$  and  $|Q_2| > \varepsilon/2$  respectively .

Now, By simple Triangle law of inequality,

$$\varepsilon < |Q_1 + Q_2| < |Q_1| + |Q_2|$$

Thus the values of  $Q_1$  and  $Q_2$  satisfying  $F$  will also satisfy  $E$ . In other words  $F \subset E$  . So,

$$P(E) \geq P(F) \tag{3.1}$$

Now,

$$\begin{aligned} P(E) \text{ or } P(|Q_1| + |Q_2| > \varepsilon) &\leq P(\text{Either of } Q \text{ has the absolute value } > \varepsilon/2) \\ &= P(|Q_1| > \varepsilon/2 \cup |Q_2| > \varepsilon/2) \\ &= P(|Q_1| > \varepsilon/2) + P(|Q_2| > \varepsilon/2) - P(|Q_1| > \varepsilon/2) * P(|Q_2| > \varepsilon/2) \\ &= P(E_1) + P(E_2) - P(E_1) * P(E_2) \\ &\leq P(E_1) + P(E_2) \end{aligned}$$

Thus ,

$$P(E) \leq P(E_1) + P(E_2) \tag{3.2}$$

Thus , combining eq. 3.1 and 3.2 , We get

$$P(F) \leq P(E) \leq P(E_1) + P(E_2)$$

Finally ,

$$P(F) \leq P(E_1) + P(E_2)$$

$\therefore$  Hence, Proved .

#### 4. Question 4 :

We are given 2 non-negative random variable  $Q_1$  and  $Q_2$  alongside  $q_1$  and  $q_2$ , both are non-negative. Given ,  $P(Q_1 < q_1) \geq 1 - p_1$  and  $P(Q_2 < q_2) \geq 1 - p_2$ . It can be written that

$$P(Q_1 * Q_2 < q_1 * q_2) = P(Q_1 < q_1 \cap Q_2 < q_2) + P(Q_1 > q_1 \text{ but } Q_1 * Q_2 < q_1 * q_2) + P(Q_2 > q_2 \text{ but } Q_1 * Q_2 < q_1 * q_2)$$

Or in other words,

$$P(Q_1 * Q_2 < q_1 * q_2) = P(Q_1 < q_1 \cap Q_2 < q_2) + P_1 + P_2$$

where  $P_1$  and  $P_2$  are the respective probabilities from the above equation.

So , We can write that

$$P(Q_1 * Q_2 < q_1 * q_2) \geq P(Q_1 < q_1 \cap Q_2 < q_2)$$

As selection of  $Q_1$  and  $Q_2$  are independent, So

$$\begin{aligned} P(Q_1 * Q_2 < q_1 * q_2) &\geq P(Q_1 < q_1) * P(Q_2 < q_2) \\ &\geq (1 - p_1) * (1 - p_2) \\ &= 1 - p_1 - p_2 + p_1 * p_2 \\ &\geq 1 - p_1 - p_2 \end{aligned}$$

$\therefore$  Hence ,Proved that

$$P(Q_1 * Q_2 < q_1 * q_2) \geq 1 - p_1 - p_2$$

## 5. Question 5 :

Given ,  $P(C_i)$  be the probability of having a car behind the door  $i$  ( $i = 1, 2, 3$ ) =  $1/3$  .

$Z_i$  be the event that the contestant chose door  $i$  ( $i = 1, 2, 3$ )

$H_i$  be the event that the host opened door  $i$  ( $i = 1, 2, 3$ ).

### 5.1

$$P(C_i / Z_i) = \frac{P(C_i \cap Z_i)}{P(Z_i)}$$

As the event of choosing a door and of having a car behind a door are independent , So

$$\begin{aligned} P(C_i / Z_i) &= \frac{P(C_i) * P(Z_i)}{P(Z_i)} \\ &= P(C_i) \\ &= \frac{1}{3} \end{aligned}$$

for all  $i = 1, 2, 3$

### 5.2

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

- Now , for  $i=1$

The expression  $P(H_3 / C_1, Z_1)$  asks the probability that the host opens door 3 given that the person opens the door 1 and the car is also in door 1.

The person opening the door 1 and the car is also in door 1 leads to 2 cases , ie. the host opens either door 2 or the door 3 .

As both the event has equally likely to happen.

So, the Probability of him opening the door 3 will turn out to be  $\frac{1}{2}$

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

- Now , for  $i=2$

The expression  $P(H_3 / C_2, Z_1)$  asks the probability that the host opens door 3 given that the person opens the door 1 and the car is also in door 2.

The person opening the door 1 and the car is in door 2 leads to only 1 case , ie. the host will open door 3 , as he knows that door 2 has the car in it ,thus can't open it .

So , the Probability of the host opening the door 3 is 1(the only possibility).

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

- Now , for  $i=3$

The expression  $P(H_3 / C_3, Z_1)$  asks the probability that the host opens door 3 given that the person opens the door 1 and the car is also in door 3.

The person opening the door 1 and the car is in door 3 leads to only 1 case , ie. the host will open door 2 , as he knows that door 3 has the car in it ,thus can't open it .

So , the Probability of the host opening the door 3 is 0 (not possibility).

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

### 5.3

The conditional probability of winning by switching is  $P(C_2/H_3, Z_1)$ . Also

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

As we already know ,  $P(H_3/C_2, Z_1) = 1$  So,

$$\begin{aligned} P(C_2/H_3, Z_1) &= \frac{1 * P(C_2, Z_1)}{P(H_3, Z_1)} \\ &= \frac{P(C_2 \cap Z_1)}{P(H_3 \cap Z_1)} \\ &= \frac{P(C_2) * P(Z_1)}{P(H_3/Z_1) * P(Z_1)} \\ &= \frac{P(C_2)}{P(H_3/Z_1)} \end{aligned}$$

Now to compute the probability  $P(H_3/Z_1)$  which is the Probability that the host will open door 3 if the contestant chooses door1 ,

$$P(H_3/Z_1) = P(C_1)P(H_3/C_1, Z_1) + P(C_2)P(H_3/C_2, Z_1) + P(C_3)P(H_3/C_3, Z_1)$$

As all the RHS values are already computed,

$$P(H_3/Z_1) = \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * 1 + \frac{1}{3} * 0 = \frac{1}{2}$$

So, finally

$$\begin{aligned} P(C_2/H_3, Z_1) &= \frac{P(C_2)}{P(H_3/Z_1)} = \frac{1/3}{1/2} \\ P(C_2/H_3, Z_1) &= \frac{2}{3} \end{aligned}$$

### 5.4

The conditional probability of winning by switching is  $P(C_1/H_3, Z_1)$ . Also

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

As we already know ,  $P(H_3/C_1, Z_1) = 1/2$  So,

$$\begin{aligned} P(C_1/H_3, Z_1) &= \frac{1/2 * P(C_1, Z_1)}{P(H_3, Z_1)} \\ &= \frac{P(C_1 \cap Z_1)}{2 * P(H_3 \cap Z_1)} \\ &= \frac{P(C_1) * P(Z_1)}{2 * P(H_3/Z_1) * P(Z_1)} \\ &= \frac{P(C_1)}{2 * P(H_3/Z_1)} \end{aligned}$$

As we already know the value of  $P(H_3/Z_1)$  which is  $\frac{1}{2}$  So,

$$\begin{aligned} P(C_1/H_3, Z_1) &= \frac{P(C_1)}{2 * P(H_3/Z_1)} = \frac{1/3}{2 * 1/2} \\ P(C_1/H_3, Z_1) &= \frac{2}{3} \end{aligned}$$

## 5.5

Looking at the above parts of this question , as the probability of winning by switching (part 5.3) =  $\frac{2}{3}$  is greater than the probability of winning by not switching (part 5.4) =  $\frac{1}{3}$ .

So , We can conclude that switching the door here is more beneficial than remaining on the previously selected door.

## 5.6

Considering that the person chose door 1 .Then,

- Switching the door (to door 2)

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

As there are 2 remaining door , so opening either of them will be equally likely and irrespective of car in any of the door and thus , equal to 1/2.

So ,  $P(H_3/C_2, Z_1) = P(H_3/Z_1) = 1/2$

$$\begin{aligned} P(C_2/H_3, Z_1) &= \frac{1/2 * P(C_2, Z_1)}{P(H_3, Z_1)} \\ &= \frac{1/2 * P(C_2 \cap Z_1)}{P(H_3 \cap Z_1)} \\ &= \frac{1/2 * P(C_2) * P(Z_1)}{P(H_3/Z_1) * P(Z_1)} \\ &= \frac{P(C_2)}{2 * P(H_3/Z_1)} \\ &= \frac{1/3}{2 * 1/2} = \frac{1}{3} \end{aligned}$$

- Switching the door (to door 3)

$$P(C_3/H_3, Z_1) = \frac{P(H_3/C_3, Z_1)P(C_3, Z_1)}{P(H_3, Z_1)}$$

As there are 2 remaining door , so opening either of them will be equally likely and irrespective of car in any of the door and thus , equal to 1/2.

So ,  $P(H_3/C_3, Z_1) = P(H_3/Z_1) = 1/2$

$$\begin{aligned} P(C_3/H_3, Z_1) &= \frac{1/2 * P(C_3, Z_1)}{P(H_3, Z_1)} \\ &= \frac{1/2 * P(C_3 \cap Z_1)}{P(H_3 \cap Z_1)} \\ &= \frac{1/2 * P(C_3) * P(Z_1)}{P(H_3/Z_1) * P(Z_1)} \\ &= \frac{P(C_3)}{2 * P(H_3/Z_1)} \\ &= \frac{1/3}{2 * 1/2} = \frac{1}{3} \end{aligned}$$



- Not switching the door (remaining in door 1)

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

As there are 2 remaining door , so opening either of them will be equally likely and irrespective of car in any of the door and thus , equal to 1/2.

So ,  $P(H_3/C_1, Z_1) = P(H_3/Z_1) = 1/2$

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

$$= \frac{1/2 * P(C_1 \cap Z_1)}{P(H_3 \cap Z_1)}$$

$$= \frac{1/2 * P(C_1) * P(Z_1)}{P(H_3/Z_1) * P(Z_1)}$$

$$= \frac{P(C_1)}{2 * P(H_3/Z_1)}$$

$$= \frac{1/3}{2 * 1/2} = \frac{1}{3}$$

Thus ,We can conclude that in this case switching will have no further benefit for the contestant. Winning in both the cases is equally likely.

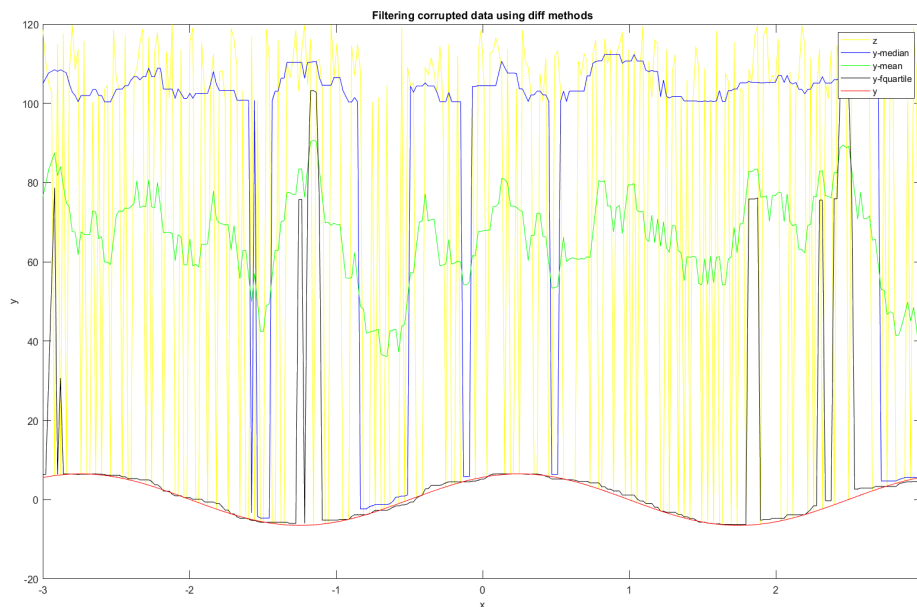
## 6. Question 6 :

The code and plots for this question has been included in the folder.

- at  $f = 30\%$ 
  - The RMS error of moving median filtering is 15.186.
  - The RMS error of moving mean filtering is 58.9177.
  - The RMS error of moving quartile filtering is 0.015119.



- at  $f = 60\%$ 
  - The RMS error of moving median filtering is 420.5754.
  - The RMS error of moving mean filtering is 215.2678.
  - The RMS error of moving quartile filtering is 45.1039.



The moving quartile method seems to be more accurate and produces better results in RMS error.

The noise reduction using *mean method* is least efficient since mean is more sensitive to larger noise values. The *median method* in this case being more accurate since we consider the middle element by sequence, it is robust to the larger noise values. But may be inefficient in the cases of extreme noisy conditions (eg.  $f = 60\%$ ).

And in the end, the *quartile method*, which focuses on the middle 50% of the data, is most robust against noise and outliers.

## 7. Question 7 :

- Updating the value of the bin

When we receive the new data value to be added into A, that new value or the data-point will fall into its corresponding bin by finding the range into which the value belongs to. And then increment the value or count of that bin by 1.

Recalculating the values for the updated data consisted in the A is computationally inefficient. Hence, defining functions to recalculate the data in a smart way becomes necessary. Also in histogram's case, some times the number of bins might be much smaller than the total number of data points, updating the histogram is generally more efficient than recomputing the entire histogram from scratch. This is particularly important when dealing with large data-sets.