Assignment 1

Sagnik Nandi (23B0905)

Jainendrajeet (23B1008)

Sumedh S S (23B1079)

August 17, 2024

1 Let's Gamble

Let a, b be number of times A and B respectively win after n throws each. There are two cases when A has more wins overall than B :
Case 1 : $a > b$
In this case A has more wins than B irrespective of the $n+1$th throw. Since both have fair dice and an equal chance of winning in each throw,

$$\begin{aligned} P(a > b) &= P(a < b), \text{ (by symmetry)} \\ &= \frac{P(a > b) + P(a < b)}{2} \\ &= \frac{1 - P(a = b)}{2}, \text{ (sum of probability over all cases is 1)} \end{aligned} \tag{1}$$

Case 2 : $a = b$ and A wins $n+1$th round

$$\begin{aligned} P_{n+1}(a > b) &= P(a = b, A_{n+1}) \\ &= P(a = b)P(A_{n+1}), \text{ (independent events)} \\ &= \frac{P(a = b)}{2} \end{aligned} \tag{2}$$

Adding (1) and (2) we get $P(a > b) = \frac{1}{2}$

2 Two Trading Teams

Let P(A) and P(B) be the probability of a player to win against A and B respectively.Since it is given that B is better than A then $P(B) < P(A)$
Case 1 : A-B-A
The probability of winning at least two consecutive matches is that we win the first two match (and either Lose or Win the Last Match) or lose the first match and then win the last two matches.
So for the first case (win first two matches)

$$P(win) = P(A) * P(B) * 1$$

So for the second case (lose first and win last two matches)

$$P(win) = (1 - P(A)) * P(B) * p(A)$$

The two cases are mutually exclusive

$$P(win|ABA) = 2 * P(A) * P(B) - P(A) * P(B) * P(A)$$

Case 2 : B-A-B

The probability of winning at least two consecutive matches is that We win the first two match (and either lose or win the last match) or lose the first match and then win the last two matches.
So for the first case (win first two matches)

$$P(win) = P(B) * P(A) * 1$$

So for the second case (lose first and win last two matches)

$$P(win) = (1 - P(B)) * P(A) * P(B)$$

The two cases are mutually exclusive

$$P(win|BAB) = 2 * P(A) * P(B) - P(A) * P(B) * P(B)$$

Since

$$P(A) > P(B) : P(win|ABA) < P(win|BAB)$$

and we should choose the second option(B-A-B).

3 Random Variables

Part 1 : Let Q_1,Q_2 be non negative random variables. Given $P(Q_1 < q_1) \geq 1 - p_1$ and $P(Q_2 < q_2) \geq 1 - p_2$, where q_1, q_2 are non negative. Then show that$P(Q_1Q_2 < q_1q_2) \geq 1 - (p_1 + p_2)$
Taking complement of $P(Q_1 < q_1) \geq 1 - p_1$

$$P(Q_1 \geq q_1) \leq p_1$$

Taking complement of $P(Q_2 < q_2) \geq 1 - p_2$

$$P(Q_2 \geq q_2) \leq p_2$$

Taking union of $P(Q_1 \geq q_1)$ and $P(Q_2 \geq q_2)$

$$P(Q_1 \geq q_1 \text{ or } Q_2 \geq q_2) \leq P(Q_1 \geq q_1) + P(Q_2 \geq q_2) \leq p_1 + p_2$$

Now, $P(Q_1Q_2 \geq q_1q_2) \leq p_1 + p_2$ since the intersection is always smaller than union

Now taking complement again

$$P(Q_1Q_2 < q_1q_2) \geq 1 - (p_1 + p_2), \text{ Hence Proved}$$

Part 2 : Given n distinct values $\{x_i\}_{i=1}^n$ with mean μ and standard deviation σ, prove that for all i,we have $|x_i - \mu| \leq \sigma\sqrt{n-1}$. How does this inequality compare with Chebyshev's inequality as n increases? (give an informal answer)

By the definition of variance :

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

Taking positive square root on both sides,

$$\sigma(\sqrt{n-1}) = \sqrt{\sum_{i=1}^{i=n} |x_i - \mu|^2} \geq |x_i - \mu|$$

So we get

$$\sigma(\sqrt{n-1}) \geq |x_i - \mu|$$

Comparing with Chebyshev's inequality

As n increases $(\sigma\sqrt{n-1})$ also increases and tells us that x_i can deviate by a large value from the mean .

But chebyshev's inequality says that the probability $P(|x-\mu| \geq \sigma(\sqrt{n-1})) \leq \frac{1}{n-1}$ which is very low for large value of n.

4 Staff Assistant

(a)

The best candidate has to be in one of the positions from m+1 to n to get selected, therefore

$$Pr(E) = \sum_{i=m+1}^{n} Pr(E_i) \tag{3}$$

For calculating $Pr(E_i)$, note that the best candidate is in i^{th} position with probability $= \frac{1}{n}$. In each such arrangement where the i^{th} candidate is the best, the second best of the candidates lying between 1 and i-1 is among the first m candidates with probability $\frac{m}{i-1}$. Thus ,

$$Pr(E_i) = \frac{m}{n(i-1)} \tag{4}$$

Thus,

$$Pr(E) = \frac{m}{n} \sum_{j=m+1}^{n} \frac{1}{j-1} \tag{5}$$

(b)

By bounding the finite sum with 2 integrals calculated in a different way, we get that,

$$\int_m^n \frac{1}{x} dx \leq \sum_{m}^{n-1} \frac{1}{x} \leq \int_m^n \frac{1}{x-1} dx \tag{6}$$

By replacing the above summation with $Pr(E)$ and calculating the limits, we get

$$\frac{m}{n}(log(n) - log(m)) \leq Pr(E) \leq \frac{m}{n}(log(n-1) - log(m-1) \tag{7}$$

(c)

Consider the function $f = -xlogx$. Its derivate equals $-log(x) - 1$ which attains the value 0 when $x = \frac{1}{e}$. It is clear by the sign of its double differential that f attains its maxima at $x = \frac{1}{e}$. Replacing x by $\frac{m}{n}$, we notice that f becomes identical to $\frac{m}{n}(log(n) - log(m))$ and its value at $\frac{m}{n} = \frac{1}{e}$ equals $\frac{1}{e}$. Hence $Pr(E) \geq \frac{1}{e}$.

5 Free Trade

Let x be the position in queue and E(x) be the event that all the traders in the queue until $x-1$th place have distinct id's and at x the id repeats from the preceeding id's. Assuming the id's are distributed at random, probability of E(x) is

$$\begin{aligned} P(E(x)) &= 1\frac{199}{200}\frac{198}{200}...\frac{200-(x-2)}{200}\frac{x-1}{200} \\ &= \frac{200!}{(200-(x-1))!200^{x-1}}\frac{x-1}{200} \end{aligned} \tag{8}$$

To maximize P(E(x)), we have to solve $P(E(x)) > P(E(x+1))$ and $P(E(x)) > P(E(x-1))$

$$\begin{aligned} P(E(x)) > P(E(x+1)) &<=> \frac{x-1}{(200-(x-1))!200^x} > \frac{x}{(200-x)!200^{x+1}} \\ &<=> 200(x-1) > (201-x)x \\ &<=> x^2 - x - 200 > 0 \end{aligned} \tag{9}$$

which gives $x \geq 15$.
Solving the other equation in a similar way gives $x \leq 15$. Hence the value of x that maximizes desired probability is x=15.

6 Update Functions

Formula used to update mean:

$$\mu_{n+1} = \frac{\mu_n n + x_{n+1}}{n+1}$$

Formula used to update std:

$$\sigma_n^2 = \frac{\sum x^2}{n} - \mu_n^2$$
$$\frac{\sum x^2}{n} <= \frac{\sum x^2 + x_{n+1}^2}{n+1}$$
$$\mu_n <= \mu_{n+1}$$
$$\sigma_{n+1}^2 = \frac{\sum x^2}{n+1} - \mu_{n+1}^2$$

Formula used to update median (assuming A is sorted):
Case 1: n even — In this case, old median is $\frac{A_{\frac{n}{2}} + A_{\frac{n}{2}+1}}{2}$. If the new data is between $A_{\frac{n}{2}}$ and $A_{\frac{n}{2}+1}$ then new data becomes the median. If it is lesser than $A_{\frac{n}{2}}$ then $A_{\frac{n}{2}}$ becomes the median. Otherwise if it is greater than $A_{\frac{n}{2}+1}$ then $A_{\frac{n}{2}+1}$ becomes the new median.
Case 2: n odd — In this case, the median was $A_{\frac{n+1}{2}}$. If the new data lies between $A_{\frac{n+1}{2}-1}$ and $A_{\frac{n+1}{2}+1}$ then the new median is $\frac{newData + A_{\frac{n+1}{2}}}{2}$. If the new data is lesser than $A_{\frac{n+1}{2}-1}$ then median will be $\frac{A_{\frac{n+1}{2}} + A_{\frac{n+1}{2}-1}}{2}$. Otherwise if it is greater than $A_{\frac{n+1}{2}+1}$ then it will be $\frac{A_{\frac{n+1}{2}} + A_{\frac{n+1}{2}+1}}{2}$.

For updating the histogram:
Supposing the data is divided into n bins, if the data does not fall in any of these bins then we create a new bin with frequency one; otherwise we increment the frequency of the bin containing it by one.

Running instructions:
Enter the submission folder and run "python3 Q6.py"

7 Plots

7.1 Violin Plot

Uses of Violin Plot: Violin plots can be used to compare distribution of heights between different states or city. Violin plots can also be used to compare distribution of grades between sections of the institute, Since it provides information on density and median also, we can easily compare multiple data.

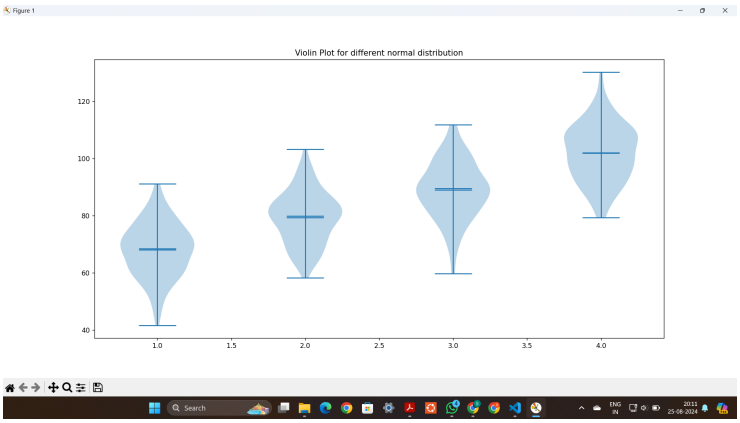


Figure 1: Violin Plot

7.2 Pareto Chart

Uses of Pareto chart Pareto chart are basically based on *80 20 rule* , which means 80 percent of the effects arise from the 20 percent of the causes . For example 80 percent of the sales comes from the 20 percent of the consumer. For example we can create a Pareto chart to see frequency of students are struggling in each topic. In DLDCA ,course for now most of the people are struggling in Sequential implementation of Karatsuba :(.

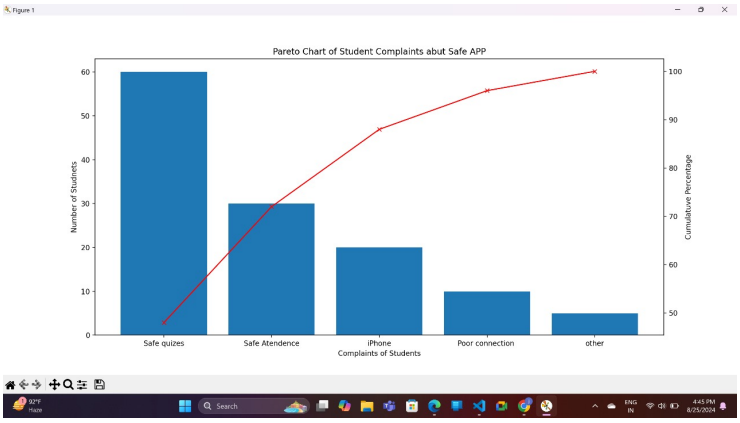


Figure 2: Pareto Chart

7.3 Coxcomb

Uses of coxcomb chart This chart can be used for visualising the monthly data .For example : The number of people died due to corona virus before and after the vaccination / wearing mask and following social distancing .

Since the radius can also change we can actually compare the data(by calculating the area) just by looking at it carefully.

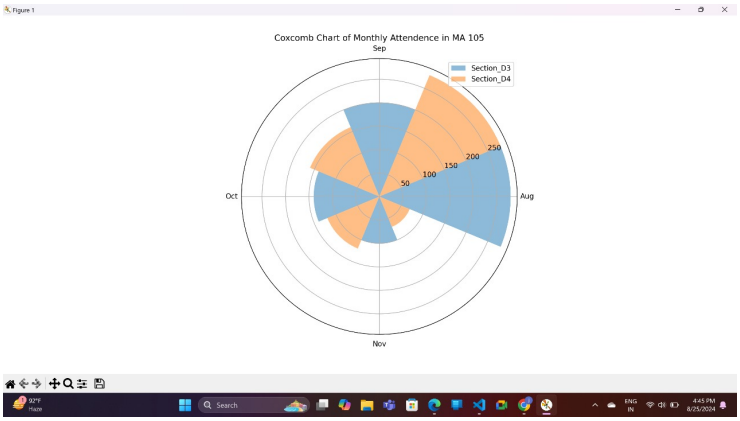


Figure 3: Coxcomb Plot

7.4 Waterfall Plot

Uses of Waterfall Plot Waterfall plot can be used to visualise the increase or decrease of net income of any person and the cause associated with it , and finally compare the present annual income with the last year income .

Figure 4: Waterfall Plot in 2D

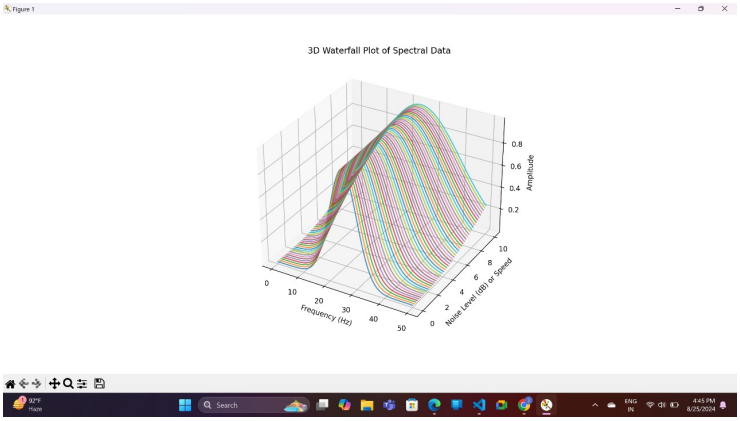
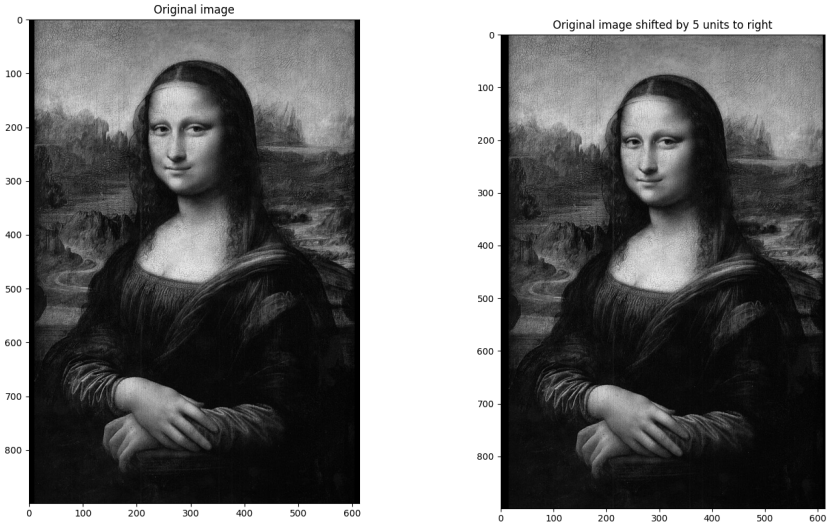


Figure 5: Waterfall Plot in 3D

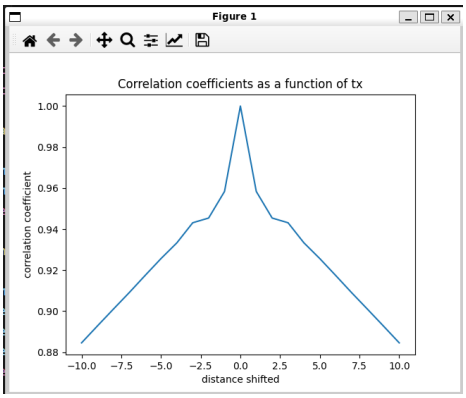
Running Instructions:
Enter the submission folder and run “python3 Q7.py”

8 Monalisa

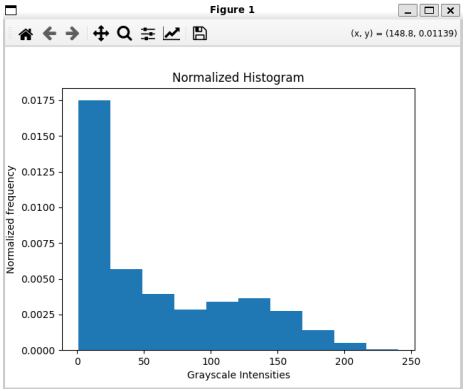


(a) Original Image

(b) Transformed Image



(a) Correlation coefficients as a function of distance shifted



(b) Normalized histogram generated using matplotlib.pyplot.hist

Running Instructions:
Enter the submission folder and run “python3 Q8.py”