DAI assignment-1

23B0924-B.somanath vasanth
23B1032-V.Srikar nihal
23B1055-S.Anirudh

Files included in the zip file

For q6 code is in the file q6.py

For q7 code for plots is in q7violinplot.py,q7paretoplot.py,q7coxcomb.py,q7waterfallplot.py

For q8 code for q8monalisa.py

CS-215 ASSIGNMENT-1

1) Let's Gamble.

Given,

A has $n+1$ dice (fair)

B has n dice (fair)

Probability of getting a prime number on top = p

$$p = 3/6 = 1/2. \quad (\because \text{ each dice has 3 prime numbers i.e., } \{2,3,5\})$$

let p be the probability that A will have more wins than B

let x be the ~~wi~~ no. of wins of A and y be the no. of wins of B

$$p = P(X=1)P(Y<1) + P(X=2)P(Y<2) + P(X=3)P(Y<3) \ldots \cancel{P(X=n+1)P(Y<n+1)} \ldots P(X=n+1)P(Y<n+1)$$

$$P(X=r) = {}^{(n+1)}C_r \left(\tfrac{1}{2}\right)^r \left(\tfrac{1}{2}\right)^{n+1-r}$$

$$= {}^{n+1}C_r \left(\tfrac{1}{2}\right)^{n+1}$$

$$P(Y<r) = \sum_{x=0}^{r-1} P(Y=x) = \sum_{x=0}^{r-1} {}^{n}C_x \left(\tfrac{1}{2}\right)^x \left(\tfrac{1}{2}\right)^{n-x}$$

$$p = {}^{n+1}C_1 \left({}^{n}C_0 \left(\tfrac{1}{2}\right)^n\right)\left(\tfrac{1}{2}\right)^{n+1} + {}^{n+1}C_2 \left(\tfrac{1}{2}\right)^{n+1}\left[{}^{n}C_0 + {}^{n}C_1\right]\left(\tfrac{1}{2}\right)^n \ldots \ldots {}^{n+1}C_{n+1}\left({}^{n}C_0 + {}^{n}C_1 \ldots {}^{n}C_n\right)\left(\tfrac{1}{2}\right)^{2n+1}$$

$$= \left(\tfrac{1}{2}\right)^{2n+1}\left[\left({}^{n+1}C_1\,{}^{n}C_0 + {}^{n+1}C_2\,{}^{n}C_1 + \ldots {}^{n+1}C_{n+1}\,{}^{n}C_n\right) + \left({}^{n+1}C_2\,{}^{n}C_0 + {}^{n+1}C_3\,{}^{n}C_4 \ldots {}^{n+1}C_{n+1}\,{}^{n}C_{n-1}\right) + \ldots \left({}^{n+1}C_{n+1}\right)\left({}^{n}C_0\right)\right]$$

$$= \left(\tfrac{1}{2}\right)^{2n+1}\left[\left({}^{n+1}C_1\,{}^{n}C_n + {}^{n+1}C_2\,{}^{n}C_{n-1} \ldots {}^{n+1}C_{n+1}\,{}^{n}C_0\right) + \left({}^{n+1}C_2\,{}^{n}C_n + {}^{n+1}C_3\,{}^{n}C_{n-1} \ldots {}^{n+1}C_{n+1}\,{}^{n}C_1\right) + \ldots \left({}^{n+1}C_{n+1}\,{}^{n}C_n\right)\right]$$

$$(\because {}^{n}C_r = {}^{n}C_{n-r})$$

$$P = \left(\frac{1}{2}\right)^{2n+1}\left({}^{2n+1}C_{n+1} + {}^{2n+1}C_{n+2} \cdots {}^{2n+1}C_{2n+1}\right)$$

$\left(\because \sum {}^{a}C_r\, {}^{b}C_{n-r} = {}^{a+b}C_n\right)$

$$= \left(\tfrac{1}{2}\right)^{2n+1}\left(2^{2n}\right)$$

$\left(\because \sum_{r=0}^{2n+1} {}^{2n+1}C_r = 2^{2n+1}\right.$ and

$\sum_{r=0}^{n} {}^{2n+1}C_r = \sum^{n} {}^{2n+1}C_{2n+1-r}$.

$\Rightarrow \sum_{r=0} {}^{2n+1}C_r = \sum_{r=0} {}^{2n+1}C_{2n+1-r}\left.\right)$

$$= \frac{1}{2}$$

$\therefore$ The probability that A will have more wins than B $= \frac{1}{2}$

2) Two Trading Teams

Given, Team B is better at trading than Team A.

let $P(A)$ be the probability of we win against A. and $P(B)$ be the probability of we win against B.

So, $P(A) > P(B)$

Let ~~P(A-B-A)~~ Case 1: A - B - A

P_1 be the probability to win in two sets in a row.

$P_1 = P(A\text{win})P(B\text{win}) + P(A\text{lose})\,P(B\text{win})P(A\text{win})$ $\left(\because P(A\text{win})\right.$ is probability we winning if we play with A similar for B$\left.\right)$

$= P(A)P(B) + (1-P(A))\,P(B)\cdot P(A)$

Case 2: B - A - B

$P_2 = P(B)P(A) + (1-P(B))(P(A))\,P(B)$

$1-P(A) < 1-P(B)$ $\quad (\because P(A) > P(B))$

$P(A)P(B) + (1-P(A))P(B)P(A) < P(A)P(B) + (1-P(B))P(A)P(B)$

$P_1 < P_2$.

Hence the probability of we win is more in the case-2

i.e., B-A-B order

3)

3.1) Given that,

~~P_x~~ Q_1, Q_2, q_1, q_2 are non-negative

$P(Q_1 < q_1) \geq 1 - P_1$

$P(Q_2 < q_2) \geq 1 - P_2$

We need to ~~prove~~ prove $P(Q_1 Q_2 < q_1 q_2) \geq 1 - P_1 P_2$

$P(Q_1 < q_1) \geq 1 - P_1 \Rightarrow P(Q_1 \geq q_1) \leq P_1$

similarly $P(Q_2 \geq q_2) \leq P_2$

we know that for $Q_1 Q_2 \geq q_1 q_2$ we need to have ~~q~~ $Q_1 \geq q_1$ or $Q_2 \geq q_2$. (if $Q_1 < q_1$ and $Q_2 < q_2$ then $Q_1 Q_2 < q_1 q_2$)

So let A be the event that $Q_1 \geq q_1$ and B be the Event that $Q_2 \geq q_2$ and C be the event that $Q_1 Q_2 \geq q_1 q_2$

we know that $P(A \cup B) \leq P(A) + P(B)$

$\Rightarrow P(Q_1 Q_2 \geq q_1 q_2) \leq P(Q_1 \geq q_1) + P(Q_2 \geq q_2)$

$\Rightarrow P(Q_1 Q_2 \geq q_1 q_2) \leq P_1 + P_2$

$\Rightarrow P(Q_1 Q_2 < q_1 q_2) > 1 - (P_1 + P_2)$ ($\because P(A) < K$ then $P(\bar{A}) \geq 1 - K$)

Hence Proved.

3.2) Given, for data values $\{x_i\}_{i=1}^{n}$, μ is mean and σ is standard deviation.

~~we~~ we need to prove that $|x - \mu_i| \leq \sigma\sqrt{n-1}$

Now let us consider $\sigma = \sqrt{\dfrac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots (x_n - \mu)^2}{n-1}}$

Scanned by TapScanner

from the formula we can say that

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

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

$$\Rightarrow \sigma\sqrt{n-1} \geq \sqrt{(x_i-\mu)^2}$$

$$\Rightarrow \sqrt{(x_i-\mu)^2} \leq \sigma\sqrt{n-1}$$

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

chebyshev's inequality states that

The proportion of sample points k or more than k ($k>0$) standard deviation away from the sample mean is less than or equal to $1/k^2$.

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

$$\frac{S_K}{N} \leq \frac{1}{K^2}$$

if we substitute $k=\sqrt{n-1}$ we have

$$S_K = \{x_i : |x_i-\mu| \geq \sigma\sqrt{n-1}\}$$

$$\frac{S_K}{N} \leq \frac{1}{n-1}$$

$$\Rightarrow S'_K = \{x_i : |x_i-\mu| \leq \sigma\sqrt{n-1}\}$$

$$\frac{S'_K}{N} \geq \frac{n-2}{n-1}$$

as n increases the chebyshev's inequality tends to the given inequality.

4) Staff Assitant

(a) Given,

Event E_i be that i^{th} candidate is best and we hire him.

$Pr(E_i) = 0 \quad 1 \le i \le m$ ($\because$ We reject the first m candidates)

$Pr(E_i)$ if $i > m$

let $i = m+r+1$ — i^{th} is the best candidate

$$\underbrace{- - - - - - - -}_{m}\ \overbrace{- - -}^{r}\ \underset{i^{th}}{-}\ \underbrace{- - - - - - -}_{n-m-r-1}$$

$$P_r(E_i) = \frac{{}^{n-1}C_{m+r} \cdot \times ({}^{m}C_1 \times (m+r-1)!) \times (n-m-r-1)!}{n!}$$

Selecting $m+r$ candidates from n excluding the best ones.

candidate As the best of $m+r$ should be in first m if not then he will be selected

arranging of remaining $m+r-1$ candidates

$n!$ → Total no. of arrangements

arranging $(n-m-r-1)$ candidate

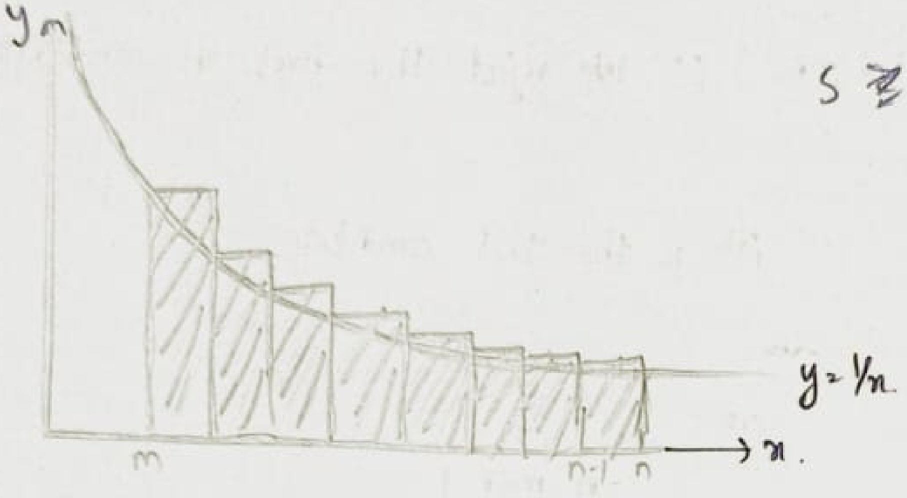
$$= \frac{(n-1)!}{(m+r)!\,\cancel{(n-m-r-1)!}} \times \frac{m \times (m+r-1)! \times \cancel{(n-m-r-1)!}}{n!}$$

$$= \frac{\cancel{(n-1)!} \times m \times \cancel{(m+r-1)!}}{n\,\cancel{(n-1)!}\,(m+r)\,\cancel{(m+r-1)!}}$$

$$= \frac{m}{n} \times \frac{1}{m+r} = \frac{m}{n} \times \frac{1}{i-1}$$

$$P_r(E_i) = \begin{cases} 0 & 1 \le i \le m \\ \frac{m}{n} \times \frac{1}{i-1} & m+1 \le i\ n \end{cases}$$

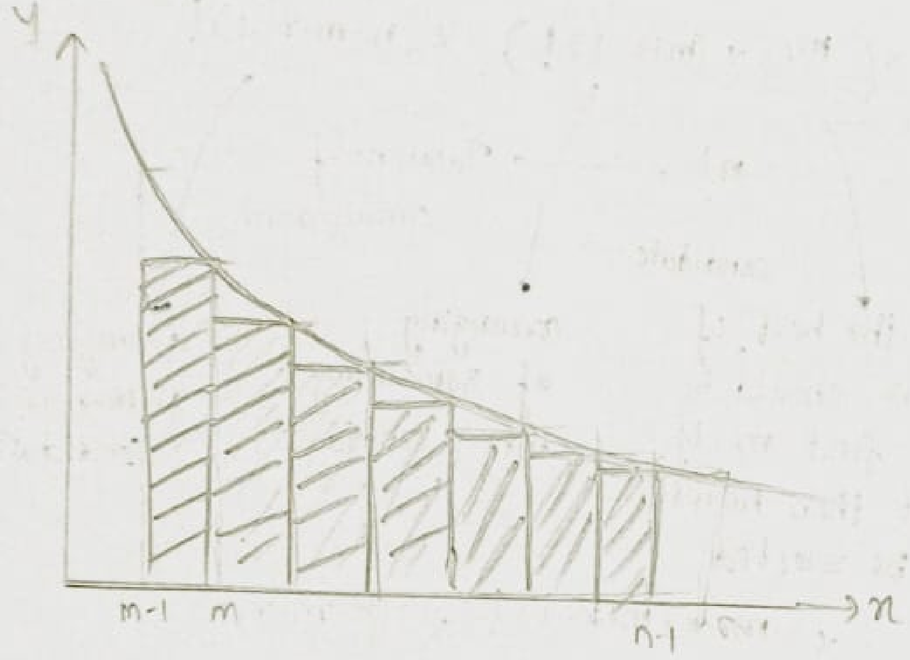
(b) $S = \sum_{j=m+1}^{n} \frac{1}{j-1} = \sum_{i=m}^{n-1} \frac{1}{i} = \frac{1}{m} + \frac{1}{m+1} \cdots \frac{1}{n-1}$;



$S \geq \int_m^n \frac{1}{x}\,dx$ $\left(\because \int \frac{1}{x}dx = \ln x\right)$

$S \geq [\ln x]_m^n$

$S \geq \ln n - \ln m$



$S \leq \int_{m-1}^{n-1} \frac{1}{x}\,dx$

$S \leq [\ln x]_{m-1}^{n-1}$

$S \leq \ln(n-1) - \ln(m-1)$

$Pr(t) = m/n \; S$

$\therefore \frac{m}{n}(\ln(n) - \ln(m)) \leq Pr(t) \leq \frac{m}{n}(\ln(n-1) - \ln(m-1))$

(c) Given equation is $\frac{m}{n}\{\ln(n) - \ln(m)\}$ now differentiate it with respective to m and equate it to zero

$\frac{1}{n}\{\ln(n) - \ln(m)\} + \frac{m}{n}\left(0 - \frac{1}{m}\right) = 0$

$\Rightarrow \frac{\ln(n)}{n} - \frac{\ln(m)}{n} - \frac{1}{n} = 0$

$$\Rightarrow \quad \frac{\ln(n) - (1+\ln(m))}{n} = 0$$

$$\Rightarrow \quad (\ln(n) - 1) = \ln(m)$$

$$\Rightarrow \quad m = n/e$$

Now if we differentiate it again with respective to m we get

$$\frac{1}{n}(-1/m) + 0$$

after substituting $m = \frac{n}{e}$ we get $-e/n^2$ which is negative so n/e is where m is maximising the curve.

now if we substitute $m = n/e$ in

$$\frac{m}{n}\ln(n/m) \leq P_r(E) \quad \text{(from previous part)}$$

we get

$$P_r(E) \geq 1/e$$

7) <u>Uses of Violin-Plot</u>

① Violet plots allow for a direct comparison of distributions across multiple catagories. ~~For instance~~

② Violet plot can reveal if a dataset has multiple peaks or modes which might indicate the presence of subgroups within the data

③ By visualizing density, violen plots help asses whether the data is symmetrically distributed or skewed

② Uses of Pareto Plots

① They are used to identify the most frequent defects or issues in manfacturing and service industries

② They help in prioritizing resources by highlighting the few Critical areas that cause the most significant impact

③ These plots are commonly used to determine the primary Causes of Problems

③ Uses of Coxcomb Plots

① This plot is particularly effective for visualizing cyclic or seaso-nal data, where the cyclical nature of the data can be naturally represented.

② The Coxcomb plot was famously used by Florence Nightingale to present mortality data, making it easier to understand the impact of different causes of death.

④ Uses of Waterfall plots:

① Waterfall plots are commonly used in finance to break down Changes in finance metrics like profit or revenue

② Waterfall plots are used for analizing the step-by-step impact of changes in business strategies, cost reductions, or sales initia-tives, helping to identify which factors had most significant effect.

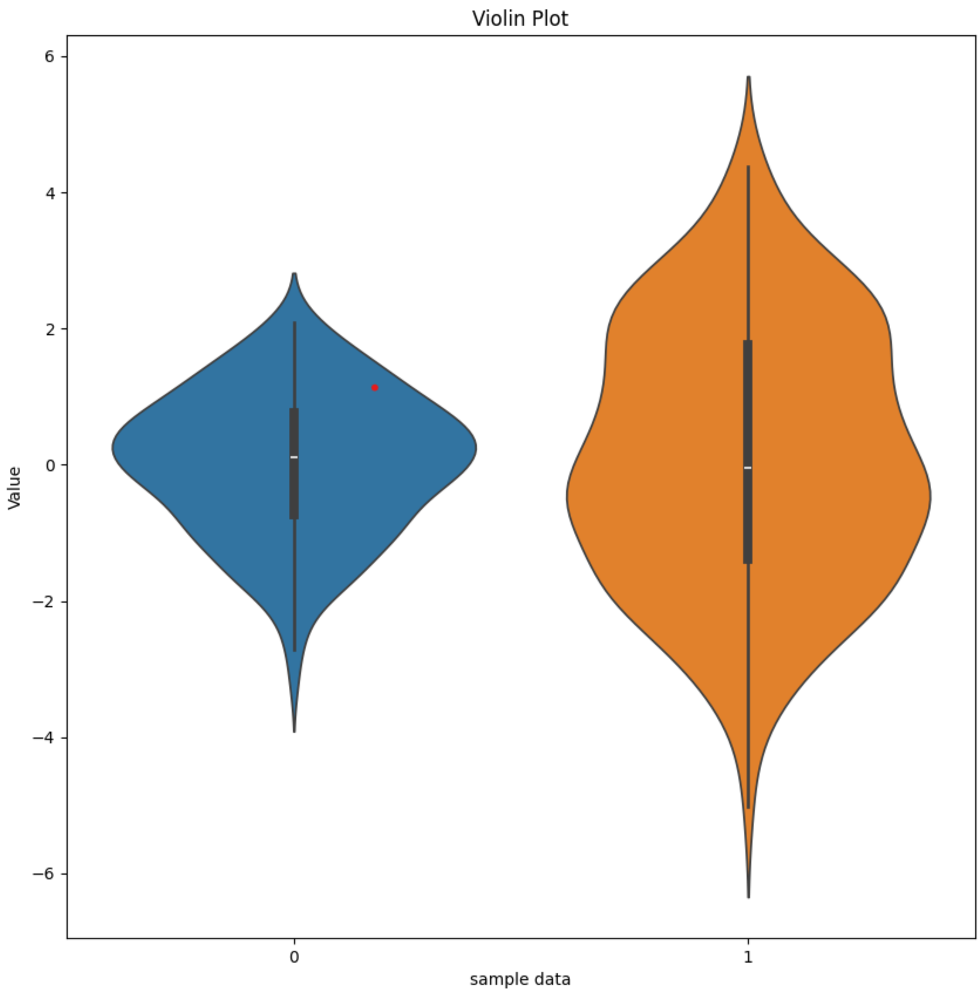


Figure 1: violin plot generated for sample data

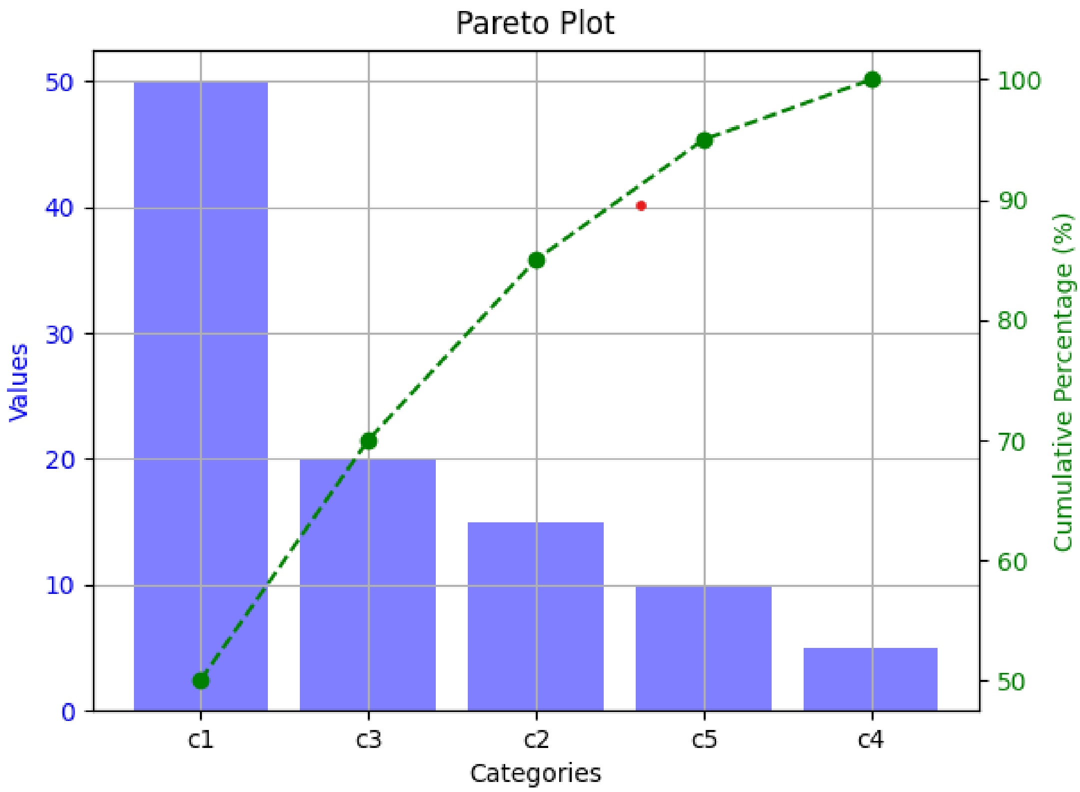


Figure 2: pareto plot generated for sample data

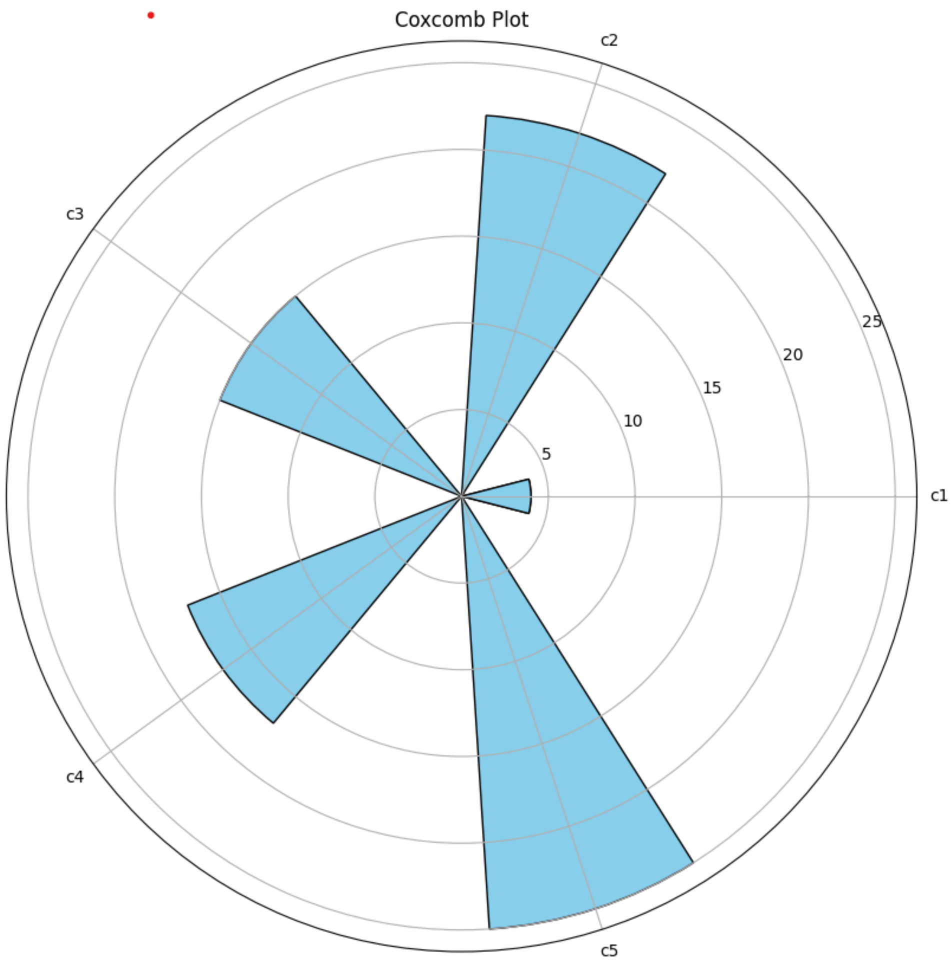


Figure 3: coxcomplot generated for sample data

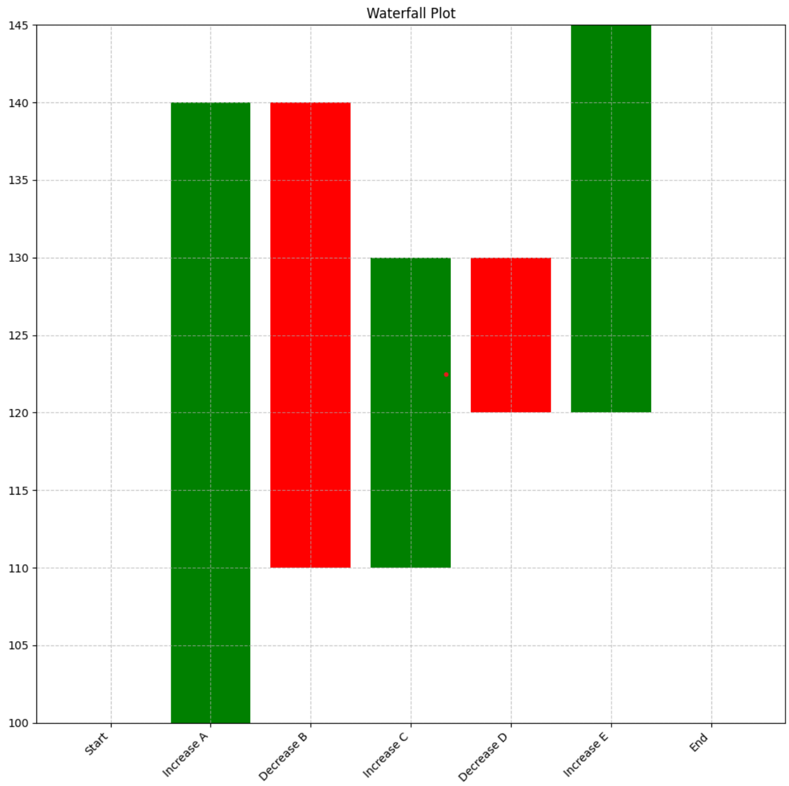
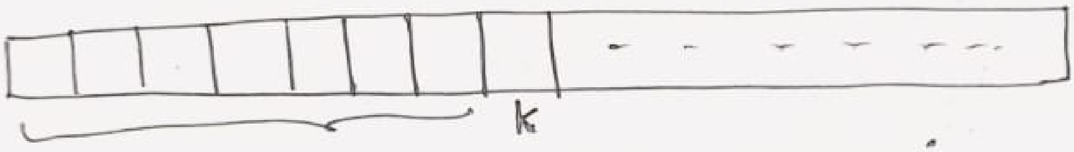


Figure 4: pareto plot generated for sample data

5) given 200 tokens.

we have to find the psn in which probability of getting a free tradeoff is maximum.

let the psn be k^{th} psn.



now let the token of this k^{th} person be i

the tokens from 1^{st} to $k-1^{th}$ psn should have 1 token of i & $k-2$ tokens of other 199 tokens (no token should be with more than 1 person)

selecting $k-2$ tokens from $199 = {}^{199}C_{k-2}$

now arranging these token in different orders

no. of orders = $(k-1)!$

The probability of k^{th} person getting ~~trade~~ a free trade

$$= \frac{\text{no. of orders satisfying the given condition}}{\text{Total no. of orders}}$$

no. of order satisfying given condition $= {}^{199}C_{k-2} \times (k-1)!$

Total no. of orders $= 200^{k-1}$ (each psn can get any token)

$$\text{Probability} = \frac{{}^{199}C_{k-2} \times (k-1)!}{200^{k-1}} = \frac{\frac{199!}{(k-2)!\,(201-k)!} \times (k-1)!}{200^{k-1}}$$

$$= \frac{(199!) \times (k-1)}{(201-k)!\,(200^{k-1})}$$

now we have to maximize this probability

$$P(K) = \frac{199!\,(k-1)}{(201-k)!\,(200)^{k-1}}$$

calculating $P(i)/P(i+1)$

$$\frac{P(i)}{P(i+1)} = \frac{\dfrac{i-1}{(201-i)!\;200^{i-1}}}{\dfrac{i}{(200-i)!\;200^{i}}} = \left(\frac{i-1}{i}\right)\left(200\right)\left(\frac{(200-i)!}{(201-i)!}\right)$$

$$= \left(\frac{i-1}{i}\right)\left(200\right)\frac{1}{201-i}$$

$$\frac{P(i)}{P(i+1)} = \frac{200\,(i-1)}{(i)(201-i)}$$

now calculate when

$$\frac{P(i)}{P(i+1)} \leq 1 \qquad \frac{200\,(i-1)}{i\,(201-i)} \leq 1$$

$$200i - 200 \leq 201i - i^2$$

$$i^2 - i - 200 \leq 0 \qquad \&$$

$$\frac{1-\sqrt{1+4(200)}}{2} < i < \frac{1+\sqrt{1+4(200)}}{2}$$

$$-13.6525 \leq i \leq 14.6525$$

$$\frac{P(14)}{P(15)} \leq 1$$

$$P(14) \leq P(15)$$

$$\frac{P(15)}{P(16)} \geq 1 \qquad P(15) \geq P(16)$$

15 is maximum

by this $\frac{P(i)}{P(i+1)} \leq 1$ for 0 to 14 & $\frac{P(i)}{P(i+1)} \geq 1$ for greater than 15

by this calculation we can say.

P(i) increases from 0-14 & reaches maximum

at $i = 15$

So we should choose the 15th posn of queue to get maximum probability.

6) given $\boxed{\text{Assuming A to be sorted for calculating new median}}$

mean, median, standard deviation. of a set of n number.

MEAN

$$\text{mean} = \frac{\sum_{x=1}^{n} x_i}{n} \rightarrow (1)$$

$$\text{new mean} = \frac{\sum_{x=1}^{n} x_i + \text{newelement}}{n+1}$$

$$\text{from}(1) \Rightarrow \sum_{x=1}^{n} x_i = n(\text{mean})$$

$$\boxed{\text{new mean} = \frac{n(\text{mean}) + \text{newelement}}{n+1}}$$

without calculating mean from individual elements we can use this formula.

STANDARD DEVIATION:

$$\text{st dev} = \sqrt{\left(\frac{1}{n-1}\right) \sum(x_i-\mu)^2}.$$

$$\sum(x_i-\mu)^2 = \sum x_i^2 + \mu^2 - 2\mu x^i$$

$$= \sum x_i^2 + n\mu^2 - 2\mu^2(n)$$

$$= \sum x_i^2 - \mu^2 n$$

$$\text{st dev} = \sqrt{\left(\frac{1}{n-1}\right)\left(\sum_{1}^{n} x_i^2 - \mu^2 n\right)} \rightarrow (1)$$

$$\text{new st dev} = \sqrt{\left(\frac{1}{n}\right)\left(\sum_{1}^{n} x_i^2 + (\text{newelem})^2 - (\text{newmean})^2(n+1)\right)}$$

by squaring (1)

$$(n-1)(\text{st dev})^2 = \sum_{1}^{n} x_i^2 - \mu^2 n \qquad \sum_{1}^{n} x_i^2 = \mu^2 n + (n-1)(\text{st dev})^2$$

by substituting this in formula

new stddev $= \sqrt{\left(\frac{1}{n}\right)\left((newelem)^2 + (oldmean)^2 n + (n-1)(oldstddev)^2 - (n+1)(newmean)^2\right)}$

by using this formula we can get new stddev.

Median

$\boxed{\text{Assumtion: Assuming A to be sorted}}$

if n iseven: $(n/2 - 1)$ $(n/2)$. $\boxed{\text{Indexing is done from 0.}}$

— — — — . . — — . . . — — — — — } n numbers

0 1 2 3

→ if the new element is less than $(n/2 - 1)^{th}$ element
it will be inserted somewhere before it & $\left(\frac{n}{2} - 1\right)^{th}$ element
will become median/ middle element

→ like this if new element is greater tha $\frac{n}{2}$ th element
it will be inserted after $n/2$ th element & $\left(\frac{n}{2}\right)^{th}$ element
will be come median.

→ if new element is b/w $n/2$ & $\frac{n}{2} - 1$ elements both including then
the new element will become median.

If n is odd: $\left(\frac{n-3}{2}\right)^{th}$ $\left(\frac{n+1}{2}\right)^{th}$ [Indexing done from 0]

$\underline{\ }_0\ \underline{\ }_1\ \underline{\ }_2\ \cdots\ \underset{\uparrow}{\underline{\ }}\ \underset{\downarrow}{\underline{\ }}\ \underset{\uparrow}{\underline{\ }}\ \cdots$

$\left(\frac{n-1}{2}\right)^{th}$

if new element is greater than $\left(\frac{n+1}{2}\right)^{th}$ element.

then $\left(\frac{n-1}{2}\right)^{th}$ & $\left(\frac{n+1}{2}\right)^{th}$ elements will become middle elements

and median is $\dfrac{\left(\frac{n-1}{2}\right)^{th}\text{element} + \left(\frac{n+1}{2}\right)^{th}\text{element}}{2}$

if new element is lesser than $\left(\frac{n-3}{2}\right)^{th}$ element.

then $\left(\frac{n-3}{2}\right)^{th}$ element & $\left(\frac{n-1}{2}\right)^{th}$ element will become middle elements

and median is $\dfrac{\left(\frac{n-3}{2}\right)^{th}\text{elem} + \left(\frac{n-1}{2}\right)^{th}\text{elem}}{2}$

if new element is b/w $\left(\frac{n-3}{2}\right)^{th}$ & $\left(\frac{n+1}{2}\right)^{th}$ element & (both including)

then $\left(\frac{n-1}{2}\right)^{th}$ & newelement will become middle elements.

and median is $\dfrac{\text{newelement} + \left(\frac{n-1}{2}\right)^{th}\text{element}}{2}$

Like this we find median.

<u>How to update histogram of A:</u>

we can check for the bin which contains the range in which new element lies & increase its count of that bin frequency by 1. every thing else remains the same

If the new value doesn't fall in any bin we should create a new bin for this ~~new element~~ new element that is added

Question 8

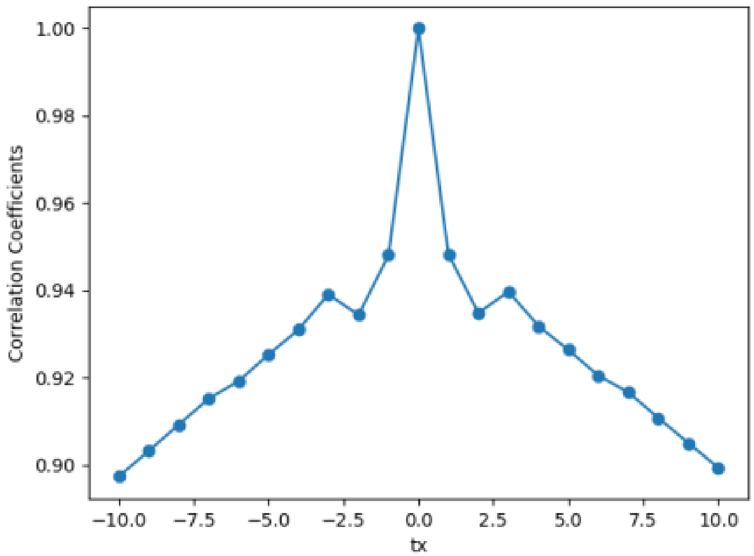


Figure 1: correlation coefficient vs tx

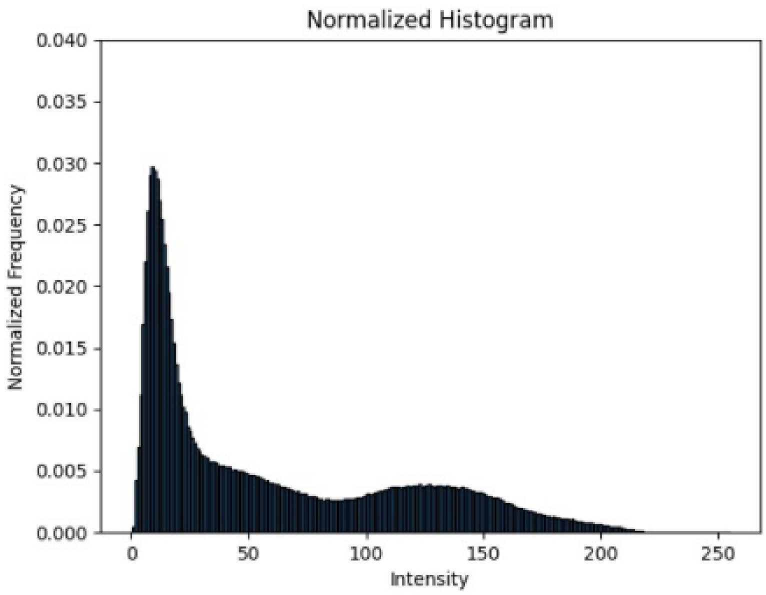


Figure 2: Normalized frequency vs intensity