Design Document-

Made By – Tanay Gupta (2018AAPS0343H), Kriti Jethlia (2018A7PS0223H), Anusha Agrawal (2018A3PS0032H)

1. Dataset Generation

After assuming some value of μ_{ML} such that $\mu_{ML} \notin (0.4,0.6)$ the required number of heads and tails were stored in an array and shuffled. We have assumed $\mu_{ML} = 0.65$.

2. Prior Distribution

The given prior distribution to us in Beta distribution with mean as 0.4. As we know, the mean for Beta distribution is a / (a+b) . So we choose a=2 and b=3.

3. Posterior Distribution

- Lisa's Method

  According to Lisa's method likelihood is calculated for the entire dataset and then the posterior is calculated. This method is possible in real time only if the entire dataset is known beforehand and the dataset is small enough to fit into the memory.

  Obtained Mean : 0.6424

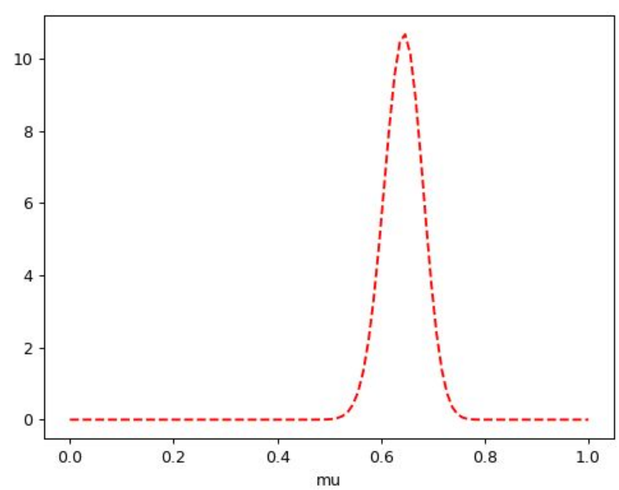


Fig 1 Distribution obtained for μ according to Lisa's method

- Bob's Method

According to Bob’s method likelihood is calculated at each step of the dataset . The calculated posterior probability is the prior for the next iteration. This is carried out for the entire dataset. This is a sequential learning model.
While iterating through this model we observe that the peak becomes more sharper after every iteration. This is because as we observe more and more data the uncertainty of the posterior for the Bayesian approach reduces.

Obtained Mean : 0.6424

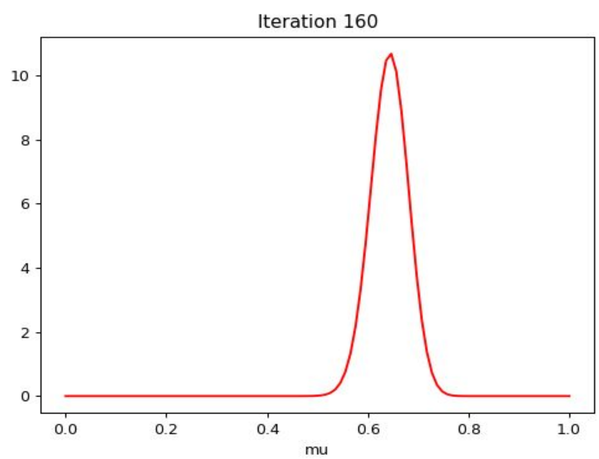


Fig 2 Final distribution obtained for μ according to Bob’s method

- **Comparison and Observations**
  The posterior distribution obtained by both the methods is the same, however, Bob’s method is computationally more heavy than Lisa’s method. Since Lisa’s method uses vectorized implementation whereas Bob’s method is a step by step implementation.
  The mean of Bayesian calculation approach for posterior and MLE approaches vary a bit because they agree for infinitely large dataset . In general posterior mean lies between prior mean and MLE of the dataset

4. Further Observations

Q1. The size of the dataset has been restricted to 160 data points. What happens if more points are added(say ~10_5)? What would the posterior distribution look like if μ_{ML}= 0.5? Which model, Bob’s or Lisa’s, would be more helpful and easier while working with large real time data and why?

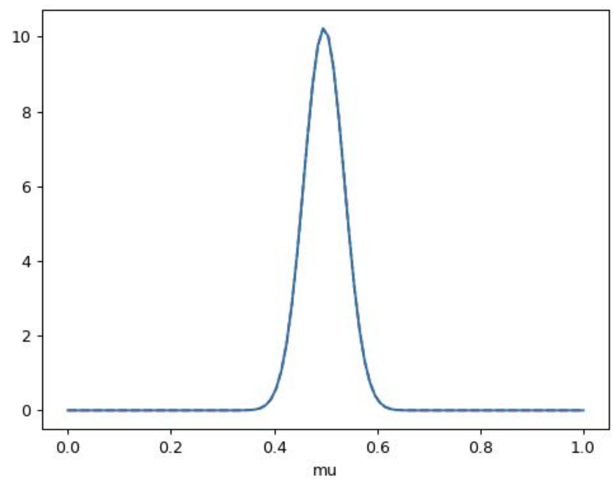
Ans :

- As we know the time complexity for Bob’s method is O(n) so for substantially small values of n such that the data can be loaded in the memory, for eg n = 10^5

the computation time for Bob is greater than Lisa. So if the hardware supports data to be loaded in memory then Lisa's approach is faster. However for very large values of n the hardware might not support the loading of entire data in that case even if Bob's approach takes longer time it is better and viable.

- In the real world scenario we cannot have restrictions on input data, as discussed above Bob's solution seems more appropriate for real world problems. Another factor to be considered while taking into account real world calculations is that we might not have the entire dataset in one go, rather it might be a steady stream of incoming data on which predictions are to be made so in that case as well Bob's approach works better.

- When μ_{ML} = 0.5 the distribution shifts towards the left and the mean so obtained is 0.497.



Q2. What if another distribution like Gamma, Guassian or Pareto were to be chosen as the prior? Would the posterior computation be easier or difficult and why?

- The computations have turned out to be easier because the prior distribution i.e Beta distribution is conjugate prior to the likelihood function.The likelihood function takes the form of the product of factors of the form $\mu^x(1 - \mu)^{1-x}$. We choose beta as a prior which is also proportional to powers of μ and (1 − μ), then the posterior distribution, which is proportional to the product of the prior and the likelihood function, will have the same functional form as the prior. If the priors are conjugate priors, we already know the parametric form of the posterior distribution so the computation is less complicated and they can be interpreted as additional data. However in real world problems it is not

necessary that conjugate prior exists for all models and represent the situation completely.
Therefore if we have another distribution like Gamma, Guassian or Pareto to be chosen as the prior then the computation might become a bit heavier as compared to Beta distribution.