Introduction to Bayesian Statistics Real Data Analysis

Anita Dash-MA20BTECH11001 Krishna Teja-MA20BTECH11005
KN Vardhan-MA20BTECH11006

Contents

***Abstract*—This report is based on a group project part of the MA4740 - Introduction to Bayesian Statistics course material. The primary goal of this project is to acquire a real-world data set (not synthetic) and execute Beta Binomial Bayesian Analysis and other approaches presented in class..**

1 Introduction

The project includes collecting a real-life data set and applying the statistical methods while also performing a beta-binomial bayesian analysis. The real-life data set that we have taken is based on the stocks of MSFT (Microsoft Corp) from the past 2000 to 2022, for the prior dataset.

The real-life data set, MSFT stocks have been taken from the Python3 *yfinance* package. The data includes opening price, closing price, maximum price, minimum price and date. Amongst which, the attributes that are of our interest are closing price and date.

2 Prior Data

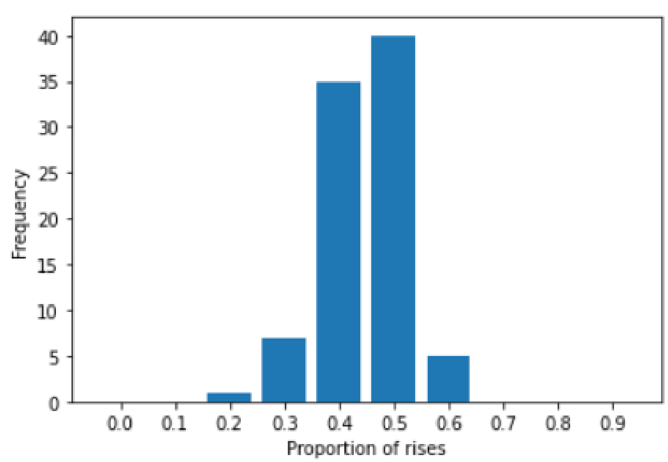
The prior data is on the stocks of MSFT from the year 2000 to 2022. The closing price of MSFT stocks on each day throughout the years is considered. Based on this prior data, we attempt to predict if MSFT's stock will rise or fall on a specific day after 2022.

The Prior Dataset is made up of the fraction of stock increases in each quarter from year 2000 to 2022.If the stock has increased from the previous day, the value is 1, otherwise it is 0. We choose a random variable X , where X is the proportion of days the stock price increased in n days, and P(X = x) representing the probability of the proportion, x

E.g. Suppose in the past n days, the MSFT stock has increased k times, then

$$x = \frac{k}{n} \tag{2.1}$$

The plot between these proportions with its frequency gives us a near normal distribution. This is our prior data and it's probability mass function has been depicted below.



3 Applying the Methods

We apply MOM and MLE to fit the normal distribution of our prior data.

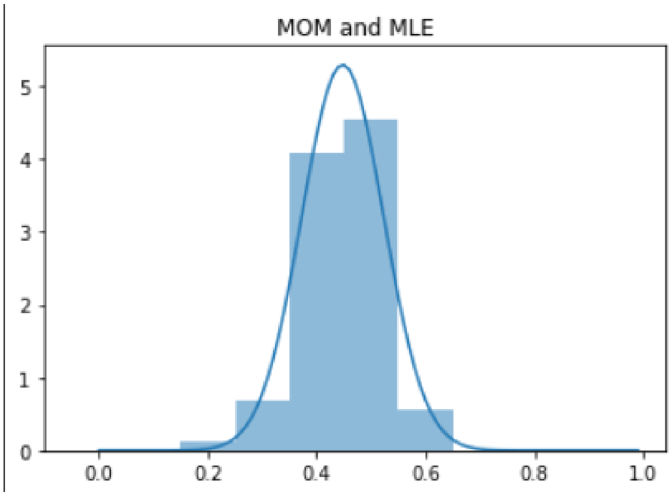
We know That,

$$E[X] = \mu = \frac{1}{n}\Sigma x_i \tag{3.1}$$

$$E[X^2] = \mu^2 + \sigma^2 = \frac{1}{n}\Sigma x_i^2 \tag{3.2}$$

where,
x_i are i.i.d realized values of X and
$N(\mu, \sigma)$ is the normal distribution that fits the prior data.

Since our distribution is a normal distribution, the MOM and MLE yield the same result. Below is the plot depicted for MOM and MLE.



4 Beta Distribution

We try to fit the prior data distribution to a beta distribution.
In a beta distribution, we get $X \sim Beta(\alpha, \beta)$, where,

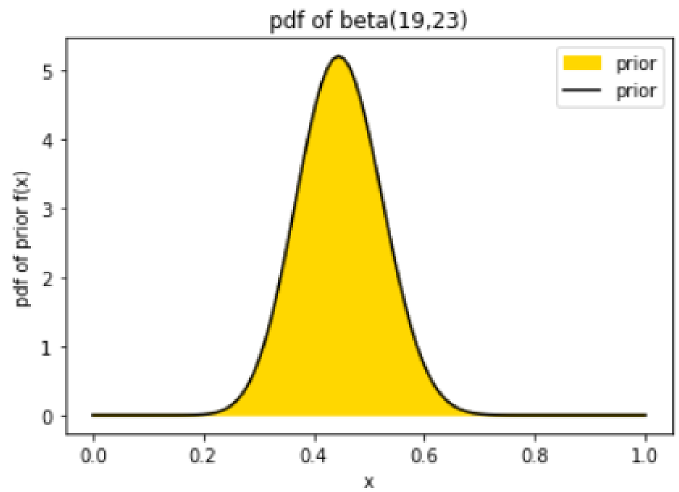
$$mean = \frac{\alpha}{\alpha + \beta} \tag{4.1}$$

$$var = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{4.2}$$

The calculations yield us the results

First moment = Mean (μ) = 0.4647727,
Second moment = 0.2235227,
Variance = 0.0075090,
Standard Deviation = 0.08665477,

And the values of α, β are 19.055, 23.504 respectively. Given below is the plot of pdf of beta distribution of the prior data.



5 Data-Likelihood Function

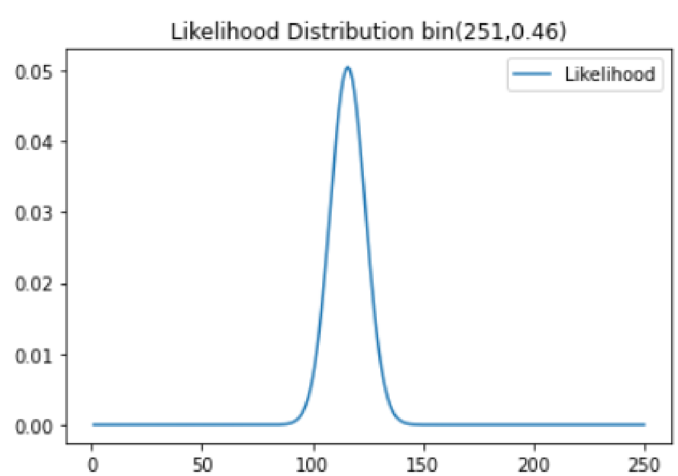
Now we require a likelihood function to perform a beta-binomial analysis on the generated beta distribution.
We choose $L \mid \pi \sim Bin(n, \pi)$ where,

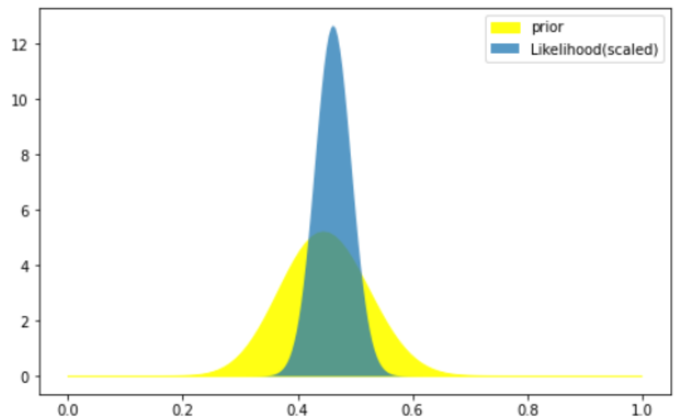
n: Number of days we are checking if the stock prize increased or decreased
π: Probability of stock getting increased

the Data consists of the realized proportion of increase in stocks in the year 2022, where

n = 251
y = 116
p = 0.46215

And the plot of the likelihood function and the prior distributions are as follows





6 Posterior Distribution

We find the posterior distribution by using the prior data and the likelihood function.
In a posterior distribution, we have,

$$\pi \mid L = y \sim beta(\alpha', \beta') \tag{6.1}$$

where,

$$\alpha' = \alpha + y,$$
$$\beta' = \beta + n - y$$

y is the realized value of number of times stocks increased in n days

Thus, after performing the necessary calculations, we get

$$\text{posterior alpha} = 135.0546860782528$$
$$\text{posterior beta} = 158.50400363967225$$
$$\text{Posterior mean} = 0.46006025646191656$$

And the plot is as shown below

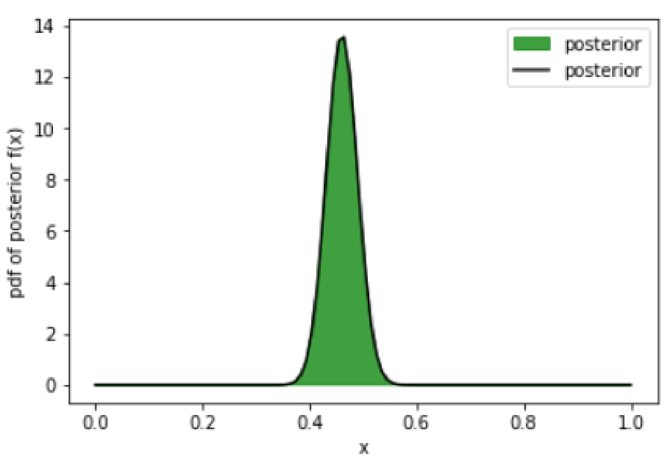


Fig. 0: Posterior Distribution

7 Conclusion

From the below combined plot we can see that the mean of our prior and posterior differ by 0.02 which is not that high. But the variance of prior and posterior differ.

$$\textit{Prior std} = 0.07534356570114019$$
$$\textit{Post std} = 0.029039831153004316$$

$$\textit{Prior estimation of proportion} \approx (29.704\%, 59.841\%)$$
$$\textit{Posterir estimation of proportion} \approx (40.198\%, 51.813\%)$$

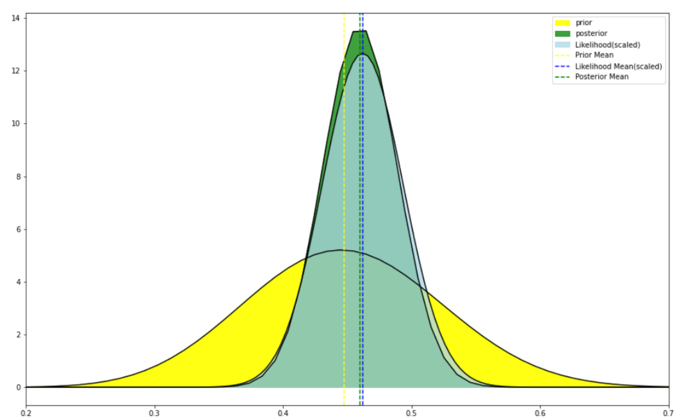


Fig. 0: Combined Plot

Link-Source Code