भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Real Data Analysis

MA4740 - Introduction to Bayesian Statistics

GROUP-2

March 24, 2023

Team Members

- Anita Dash - MA20BTECH11001
- Chilukuri Krishna Teja - MA20BTECH11005
- Kethari Narasimha Vardhan - MA20BTECH11006

Introduction

Introduction

Abstract

This presentation 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.

Introduction

Objective

The project includes:

- Real Data Analysis on Dataset MSFT stocks using method of moments and Maximum Likelihood approach
- Performing Beta-Binomial Bayesian Analysis on the Dataset MSFT stocks

Data Collection

Data Collection

The Dataset

- The Dataset, MSFT stocks which is based on the stocks of MSFT (Microsoft Corp) has 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.

Date	closing_price
2022-01-03 00:00:00-05:00	330.8138732910156
2022-01-04 00:00:00-05:00	325.141357421875
2022-01-05 00:00:00-05:00	312.6598815917969
2022-01-06 00:00:00-05:00	310.1892395019531
2022-01-07 00:00:00-05:00	310.3473815917969

Figure 1: Glimpse of the Dataset

Beta-Binomial Bayesian Data Analysis

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.

Prior-Data

- 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.

Example

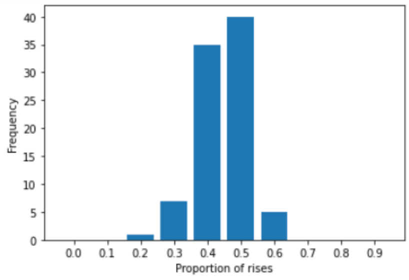
Suppose in the past n days, the MSFT stock has increased k times, then

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

Prior-Data

Year	Quarter	Prop_Inc
2000	1	0.4
2000	2	0.4
2000	3	0.4
2000	4	0.4
2001	1	0.5
2001	2	0.4
2001	3	0.4
2001	4	0.6
2002	1	0.3
2002	2	0.4

(a) Glimpse of Prior Dataset



(b) prior data

We can see that the plot between the proportions with its frequency gives us a near normal distribution.

Data Analysis on Prior-Data using MOM and MLE

Method Of Moments (MOM)

Formula

We can see that $X \sim N(\mu, \sigma)$ and x_i are i.i.d realized values of X and hence we know that,

$$M1 = E[X] = \mu = \frac{1}{n}\Sigma x_i$$

$$M2 = E[X^2] = \mu^2 + \sigma^2 = \frac{1}{n}\Sigma x_i^2$$

The calculations yield us the results

- M1 = Mean (μ) = 0.447727
- M2 = 0.206136
- Variance = 0.0057
- Standard Deviation(σ) = 0.075344

Maximum Likelihood Estimators (MLE)

definition

Suppose $X \sim N(\mu, \sigma)$ and x_i are i.i.d realized values of X

Then the likelihood function is,

$$L(\theta|k) = \prod_{i=1}^{n} f(x_i|\theta_1, ..., \theta_k)$$

And the Maximum Likelihood Estimator(MLE) is,

$$\hat{\theta}_{mle} = argmax_{\theta} L(\theta|x)$$

Maximum Likelihood Estimators (MLE)

MLE of normal distribution

The likelihood function,

$$L(\theta, \sigma^2|X) = \prod_{i=1}^{n} f(x_i|\theta, \sigma^2)$$
$$\implies L(\theta, \sigma^2|X) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{\frac{-(x-\theta)^2}{2\sigma^2}}\right)$$
$$\log(L(\theta, \sigma^2|X)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2$$

For a given σ^2, the log likelihood function is maximized when,

$$\frac{\partial \log(L(\theta, \sigma^2|X))}{\partial \theta} = 0$$
$$\implies \sum_{i=1}^{n}(x_i - \theta) = 0 \implies \theta = \frac{\sum x_i}{n}$$

Therefore, $\theta = \bar{X}$

Maximum Likelihood Estimators (MLE)

MLE of normal distribution

For a given θ, the log likelihood function is maximized when,

$$
\begin{aligned}
\frac{\partial \log(L(\theta,\sigma^2|X))}{\partial\sigma^2} &= 0 \\
\implies -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_{i=1}^{n}(x_i-\theta)^2\right) &= 0 \\
\implies \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2 \\
\implies \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2)-\theta^2
\end{aligned}
$$

Therefore, $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2)-\theta^2$

Maximum Likelihood Estimators (MLE)

Our distribution is a normal distribution, and we have shown that the MOM and MLE yield the same result.

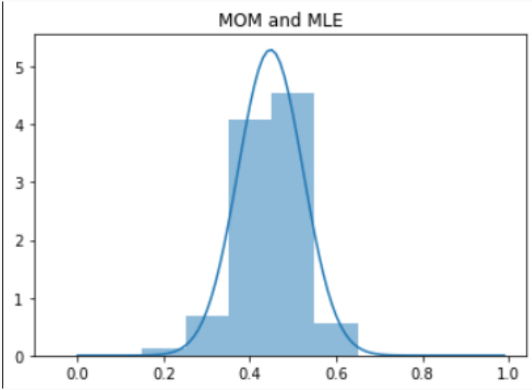


Figure 3: MOM and MLE

Beta Distribution

Beta Distribution

We try to fit our prior-data distribution to a beta distribution.

We get $X \sim \mathit{Beta}(\alpha, \beta)$, where,

Formula

$$\text{mean}(\mu) = \frac{\alpha}{\alpha+\beta}$$

$$\text{var}(\sigma) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Beta Distribution

Using the values of μ and σ found earlier using MOM and MLE.
We get the following values of α and β

$$\alpha = 19.055, \beta = 23.504$$
$$\text{prior mean} = 0.4477$$

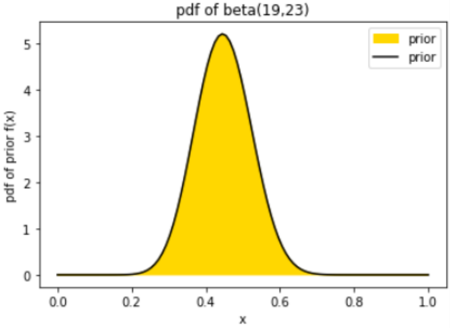


Figure 4: PDF of $\beta(19, 23)$

Data-Likelihood Function

Data-Likelihood Function

- 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,

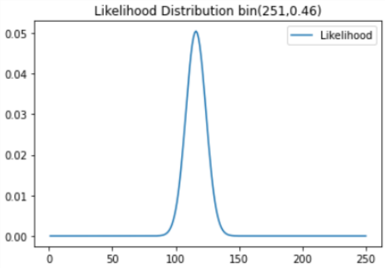
$$n\text{: The number of days we are looking at to see if the stock price has increased or decreased}$$

$$\pi\text{: the Probability that the stock price will increase}$$

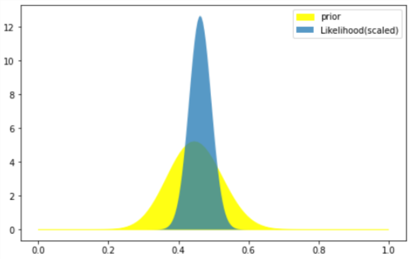
- Our data consists of the realized proportion of increase in stocks in the year 2022, where

$$n = 251$$
$$y = 116$$
$$p = 0.46215$$

Plot of Distributions



(a) Plot of Likelihood Function



(b) Plot of Prior Distribution and Likelihood Function (Scaled)

Posterior Distribution

Posterior Distribution

We Find the posterior distribution using the prior data and the likelihood function.

Definition

If the,

$$\text{Prior: } \pi \sim \textit{Beta}(\alpha, \beta)$$
$$\text{Data-Likelihood: } Y|\pi \sim \textit{Bin}(n, \pi)$$

Then the,

$$\text{Posterior: } \pi|(Y = y) \sim \textit{Beta}(\alpha + y, \beta + n - y)$$

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

Posterior Distribution

Thus, after performing the necessary calculations, we get

posterior alpha = 135.055

posterior beta = 158.504

Posterior mean = 0.46006

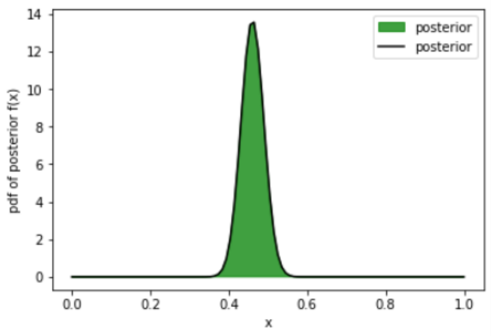


Figure 6: PDF of posterior distribution

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 by a lot.

$$\begin{aligned} \textit{Prior std} &= 0.07534356570114019 \\ \textit{Post std} &= 0.029039831153004316 \end{aligned}$$

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

Conclusion

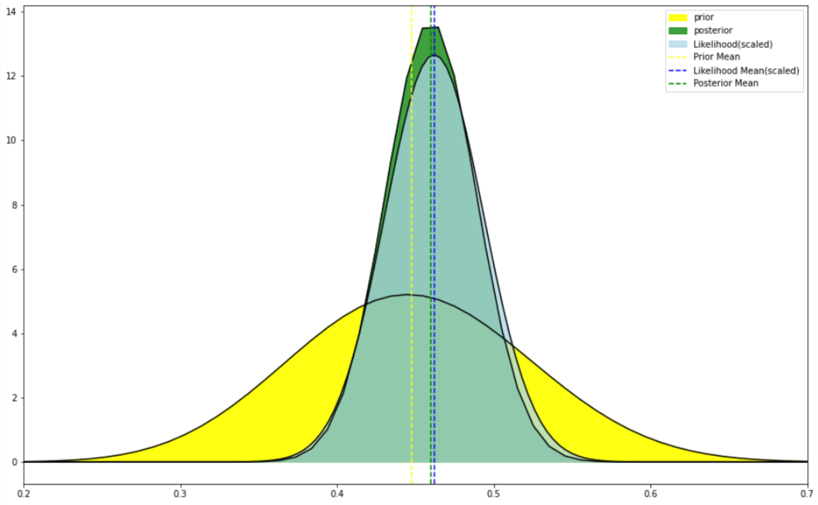


Figure 7

Thank You

▸ Source Code