Real Data Analysis

Group 8 (MA4740- Introduction to Bayesian Statistics)

Megavath Rajnikanth *ES20BTECH11018*

Routhu Prashanth *CS19BTECH11042*

Varunaditya Singhal *MA20BTECH11021*

Supriya Rawat *MA21MSCST11004*

Contents

***Abstract*—This project explores the idea of analyzing a real-world data set to demonstrate maximum likelihood estimation and method of moments approach as well as to demonstrate Beta Binomial Bayesian analysis. The project is part of the group project MA4740 - Introduction to Bayesian Statistics.**

I. 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 about the 100 most streamed songs on Spotify with their features extracted using spotify api. The dataset was taken from kaggle and is of the year of 2021. The real-life data set, most streamed songs of all time contains attributes name, duration, energy, key, loudness, mode, acoustic ness etc.

We first conducted a test of normality on our data to confirm the distribution of data. We could confirm that our dataset approximately follows a normal distribution.

Going further we perform MOM and MLE on the attribute, ”energy” which is a perpetual measure of intensity and activity in the song

II. Applying The Methods

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

A. MOM

For selecting an appropriate attribute to perform MOM and MLE on it,we performed **Anderson Darling test** and **Shapiro-Wilk test** for normality on the **Energy** attribute.Both of these are inbuilt tests in **R**.

In both the cases,the obtained value of **p** was greater than 0.05,which confirms that the **energy** attribute follows normal distribution at 5% level of significance.

We know that, since our data follows Normal distribution, equating population moment and raw moment,$m_k = \mu_k$,we get

First moment :

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \mu_1 \Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \text{ (estimator of } \theta)$$

Second moment :

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu_2 = \theta^2 + \sigma^2$$

$$\Rightarrow \hat{\sigma}^2 = m_2 - \hat{\theta}^2 = m_2 - \bar{X}^2$$

Method of moments (MOM) estimator for the parameters of a normal distribution:

$$\hat{\theta} = \frac{1}{n}\sum X_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum \left(X_i - \bar{X}\right)^2$$

where, X_i are i.i.d realized values of X

B. MLE

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f\left(x \mid \theta_1, \ldots, \theta_k\right)$

$$f\left(x_1, \ldots, x_n \mid \theta_1, \ldots, \theta_k\right) = \prod_{i=1}^{n} f\left(x_i \mid \theta_1, \ldots, \theta_k\right).$$

$$likelihood\ function.$$

It is a function of the parameters, which have a true fixed unknown value.

For each sample point x, L(θ | x) is a function of θ parameter, because x is fixed.

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta} L(\theta \mid x)$$

Given x, the value of θ which maximizes the likelihood function. Equivalently, the value of θ, for which the likelihood of observing x is maximum.

Finding MLE is a **maximization** problem.We use **Calculus** to solve for obtaining **Maximum Likelihood Estimator**.

Since,our data follows Normal distribution,**Method of Moments** and **Maximum Likelihood Estimation** approach gives same result.

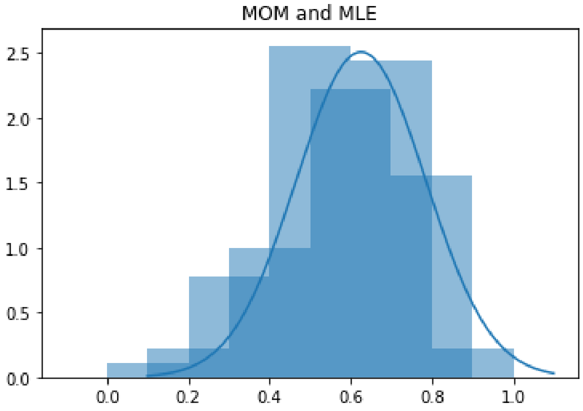

Fig. 1. Figure of MOM and MLE

III. Prior Data

The prior data for the application of beta-binomial distribution was taken from our main data-set and attribute energy. The prior data is on the energy levels of the top 100 Spotify songs of 2021. Based on this prior data, we attempt to predict the types of songs and their energy levels are likely to decrease or increase.

If the energy levels of the a particular song is greater then the mean value, the value is 1, otherwise it is 0. We denote a random variable X, where X is the proportion of songs with higher energy in set of n songs.

P(X = x) representing the probability of the proportion x. Suppose in the top n songs the energy is greater than the mean k times, then the proportion of $\frac{k}{n}$ is taken and is our random variable.

IV. Beta Distribution

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

$$mean = \frac{\alpha}{\alpha+\beta} \tag{1}$$

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

The calculations yield us the results

First moment = Mean (μ) = 0.6255,
Second moment = 0.41660,
Variance = 0.025,
Standard Deviation = 0.159

The values obtained for α and β are 5.15 and 3.08 respectively. Fiven below is the pdf of beta distribution of the prior data.

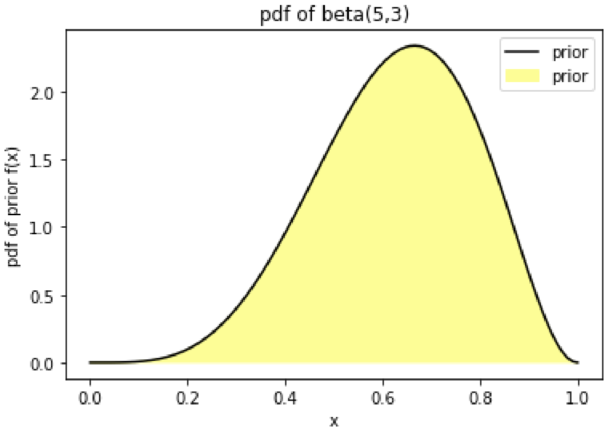

Fig. 2. PDF of beta(5,3)

V. Data-Likelihood Function

To perform beta-binomial analysis we will be requiring likelihood function.

We choose L|π $\sim$ Bin(n,π) where,

n : The number of songs we are checking if the energy is higher or not.

π : Probability of energy level increasing

The data consists of the proportion of songs with high energy levels in the year 2021, where

n = 98
y = 32
p = 0.326

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

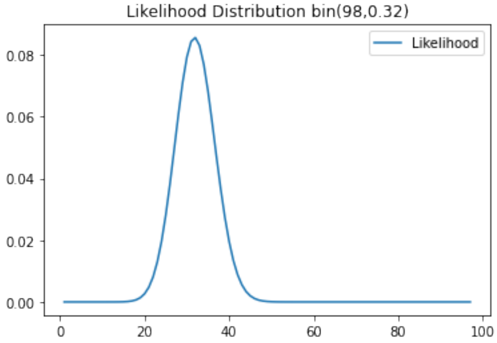

Fig. 3. Likelihood distribution

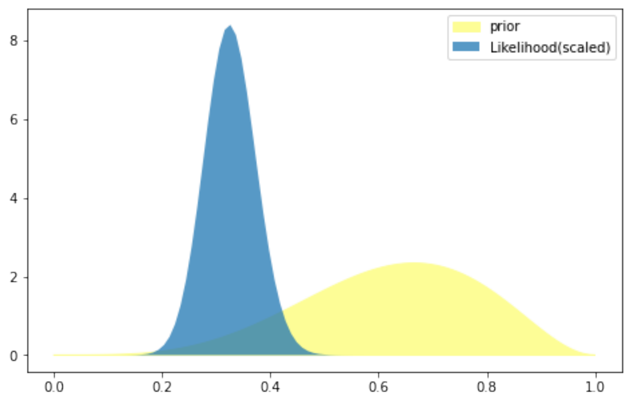

Fig. 4. Likelihood and Prior

VI. Posterior Distribution

We combine the results from our prior data on Spotify top 100 songs and after finding an appropriate data likelihood function,we know that if

Prior: $\pi\sim Beta(\alpha,\beta)$

Data-Likelihood: Y$|\pi \sim$ Bin(n, π)

Then

$$L(\pi \mid y)=\binom{n}{y} \pi^{y}(1-\pi)^{n-y} \text { for } \pi \in[0,1]$$

The conjugate Beta prior combined with the Binomial data model produce a posterior model for π. The updated Beta posterior parameters $(\alpha+y, \beta+n-y)$ reflect the influence of the prior (via α and β) and the observed data (via y and n).

$$f(\pi \mid y) \propto f(\pi) L(\pi \mid y) \propto \pi^{(\alpha+y)-1}(1-\pi)^{(\beta+n-y)-1}.$$

After substituting the values for α and β calculated from Python and substituting in the equation for posterior distribution,we obtain:

Then, Posterior distribution:

$$\pi \mid(Y=y) \sim Beta(\alpha+y, \beta+n-y)$$

$$\pi \mid(Y=y) \sim Beta(37.152,69.084)$$

where, y is the realised value of number of songs having energy level more than the average.

$$E(\pi \mid(Y=y)=0.3497 \tag{3}$$

$$\operatorname{Var}(\pi) \mid(Y=y)=0.00212 \tag{4}$$

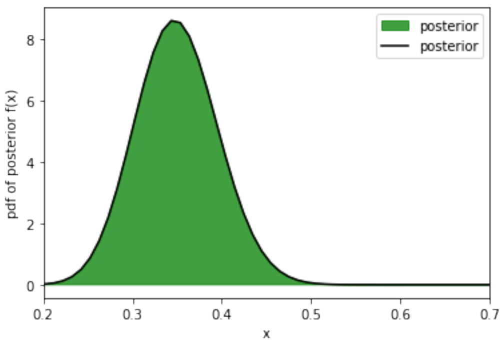

Fig. 5. Posterior Distribution

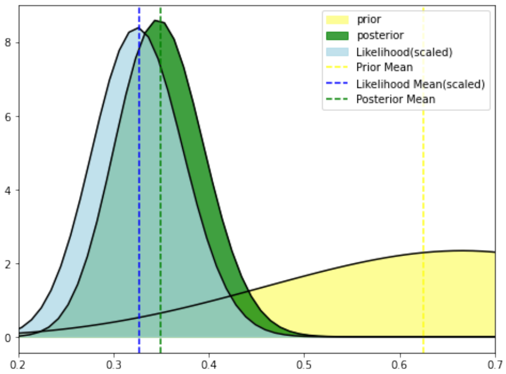

Fig. 6. Prior, Likelihood and Posterior Distribution

VII. Conclusion

We have based our conclusions on the value of y being 32,the final posterior mean obtained is equal to 0.3497. It shows that there is a significant difference between prior and posterior mean. Hence, our conclusions are based more on the likelihood data than the prior distribution.We can conclude that the new mean energy level has reduced to 0.34. We can also observe that energy alone is not the deciding factor our statistical analysis.

Hence,the newly obtained results imply that after performing Bayesian analysis on the data,the chances of picking a top 100 song would be more likely be of less energy and songs corresponding to that like Lofi and slow songs are more likely to come more in top 100.