Real Data Analysis

MA 4740 – Introduction to Bayesian Statistics

Group 8
24 March 2023

TEAM MEMBERS:

- Varunaditya Singhal – MA20BTECH11021
- Supriya Rawat – MA21MSCST11004
- Routhu Prashanth – CS19BTECH11042
- Meghavath Rajnikanth – ES20BTECH11018

OBJECTIVE:

- Conduct real data analysis on a dataset using Method of Moments (MOM) and Maximum Likelihood approach (MLE).
- Performing Beta-Binomial Bayesian analysis on the same data by estimating a prior distribution.

ABOUT THE DATASET:

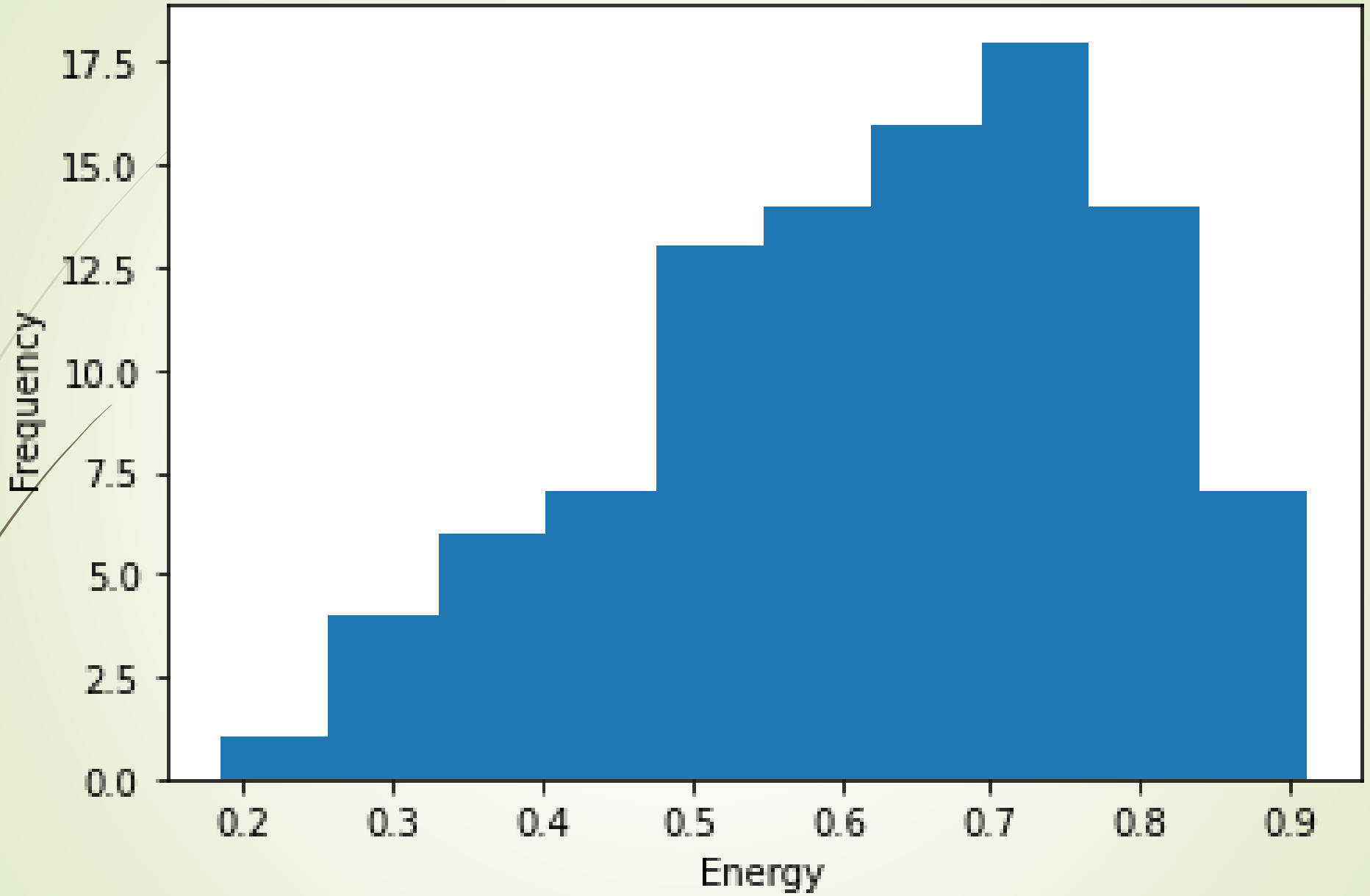
- The dataset has been taken from Kaggle website.
- It contains 100 most streamed songs on Spotify with their Features Extracted using the Spotify API .
- Link for the dataset- https://www.kaggle.com/datasets/amaanansari09/top-100-songs

DATASET

id	name	duration	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	danceability
0VjIjW4GlUZAMYd2vXMi3b	Blinding Lights	3.33	0.730	1	-5.934	1	0.0598	1.46e-03	9.54e-05	0.0897	0.3340	171.005	0.514
7qiZfU4dY1lWllzX7mPBI3	Shape of You	3.90	0.652	1	-3.183	0	0.0802	5.81e-01	0.00e+00	0.0931	0.9310	95.977	0.825
2XU0oxnq2qxCpomAAuJY8K	Dance Monkey	3.49	0.588	6	-6.400	0	0.0924	6.92e-01	1.04e-04	0.1490	0.5130	98.027	0.824
7qEHsqek33rTcFNT9PFqLf	Someone You Loved	3.04	0.405	1	-5.679	1	0.0319	7.51e-01	0.00e+00	0.1050	0.4460	109.891	0.501
0e7ipj03S05BNilyu5bRzt	Rockstar	3.64	0.520	5	-6.136	0	0.0712	1.24e-01	7.01e-05	0.1310	0.1290	159.801	0.585
3KkXRkHbMCARz0aVfEt68P	Sunflower	2.63	0.479	2	-5.574	1	0.0466	5.56e-01	0.00e+00	0.0703	0.9130	89.911	0.760
1zi7xx7UVEFkmKfv06H8x0	One Dance	2.90	0.625	1	-5.609	1	0.0536	7.76e-03	1.80e-03	0.3290	0.3700	103.967	0.792
7BKLCZ1jbUBVqRi2FVlTVw	Closer	4.08	0.524	8	-5.599	1	0.0338	4.14e-01	0.00e+00	0.1110	0.6610	95.010	0.748
789CxjEOtO76BVD1A9yJQH	Stay	4.01	0.310	9	-10.164	0	0.0283	9.45e-01	6.12e-05	0.1170	0.1250	111.893	0.621
0pqnGHJpmpxLKifKRmU6WP	Believer	3.41	0.780	10	-4.374	0	0.1280	6.22e-02	0.00e+00	0.0810	0.6660	124.949	0.776
6v3KW9xbzN5yKLt9YKDYA2	Señorita	3.18	0.548	9	-6.049	0	0.0290	3.92e-02	0.00e+00	0.0828	0.7490	116.967	0.759
41P6Tnd8KIHqON0Qlydx6a	Perfect	2.96	0.663	11	-6.818	1	0.0331	4.33e-01	1.24e-01	0.1020	0.6000	146.053	0.634
02MWAaffLxlfxAUY7c5dvx	Heat Waves	3.98	0.525	11	-6.900	1	0.0944	4.40e-01	6.70e-06	0.0921	0.5310	80.870	0.761
5uCax9HTNlzGybIStD3vDh	Say You Won't Let Go	3.52	0.557	10	-7.398	1	0.0590	6.95e-01	0.00e+00	0.0902	0.4940	85.043	0.358
7MXVkk9YMctZqd1Srtv4MB	Starboy	3.84	0.587	7	-7.015	1	0.2760	1.41e-01	6.35e-06	0.1370	0.4860	186.003	0.679
2Fxmhks0bxGSBdJ92vM42m	Bad Guy	3.23	0.425	7	-10.965	1	0.3750	3.28e-01	1.30e-01	0.1000	0.5620	135.128	0.701
0UduxLm2YxJcRMMX1K6E7E	Thinking Out Loud	4.69	0.445	2	-6.061	1	0.0296	4.74e-01	0.00e+00	0.1840	0.5910	78.997	0.781
285pBltuF7vW8TeWk8hdRR	Lucid Dreams	4.00	0.566	6	-7.230	0	0.2000	3.49e-01	0.00e+00	0.3400	0.2180	83.903	0.511
6o6urrgOlVbIzzmO4KJHCO	Don't Start Now	3.02	0.185	11	-12.205	0	0.0484	8.76e-01	4.59e-01	0.2370	0.7070	78.775	0.498
0u2P5u6lvoDfwTYjAADbn4	Lovely	3.34	0.296	4	-10.109	0	0.0333	9.34e-01	0.00e+00	0.0950	0.1200	115.284	0.351
6UelLqGlWMcVH1E5c4H7lY	Watermelon Sugar	2.90	0.816	0	-4.209	1	0.0465	1.22e-01	0.00e+00	0.3350	0.5570	95.390	0.548
6DCZcSspjsKoFjzjrWoCdn	God's Plan	3.32	0.449	7	-9.211	1	0.1090	3.32e-02	8.29e-05	0.5520	0.3570	77.169	0.754
4cj6Ti4wOXZ5ZdWlUZxRSP	Photograph	4.32	0.379	4	-10.480	1	0.0359	6.07e-01	4.72e-04	0.0986	0.2200	108.033	0.718
7tFiyTwD0nx5a1eklYtX2J	Bohemian Rhapsody	5.91	0.402	0	-9.961	0	0.0536	2.88e-01	0.00e+00	0.2430	0.2280	143.883	0.392
6RUKPb4LETWmmr3iAEQktW	Something Just Like This	4.12	0.635	11	-6.769	0	0.0317	4.98e-02	1.44e-05	0.1640	0.4460	103.019	0.617
2QjOHCTQ1Jl3zawyYOpxh6	Sweater Weather	4.01	0.807	10	-2.810	1	0.0336	4.95e-02	1.77e-02	0.1010	0.3980	124.053	0.612

MOM & MLE

- We took the attribute – ‘energy’ from the dataset and performed MOM & MLE on it.
- Energy is a measure that takes values from 0.0 to 1.0 and represents a perceptual measure of intensity and activity in the song.


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Frequency
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Energy

MOM & MLE

- We first conducted test of normality on our data to confirm the distribution of the data.
- We performed Anderson-Darling test and Shapiro-Wilk test for Normality on the dataset.
- We could confirm that our dataset follows a normal distribution.

Anderson-Darling Normality Test

```
> View(data)
> ad.test(data)

        Anderson-Darling normality test

data:  data
A = 0.50128, p-value = 0.2027
```

p-value(0.2027)>0.05

Shapiro-Wilk test for normality

```
> shapiro.test(data)

        Shapiro-Wilk normality test

data:  data
W = 0.97952, p-value = 0.122
```

p-value(0.2027)>0.05

We found that X ~ N(μ, σ) and X are i.i.d. realized values of X.
Here,X represents the different energy levels in the dataset.
We know that:

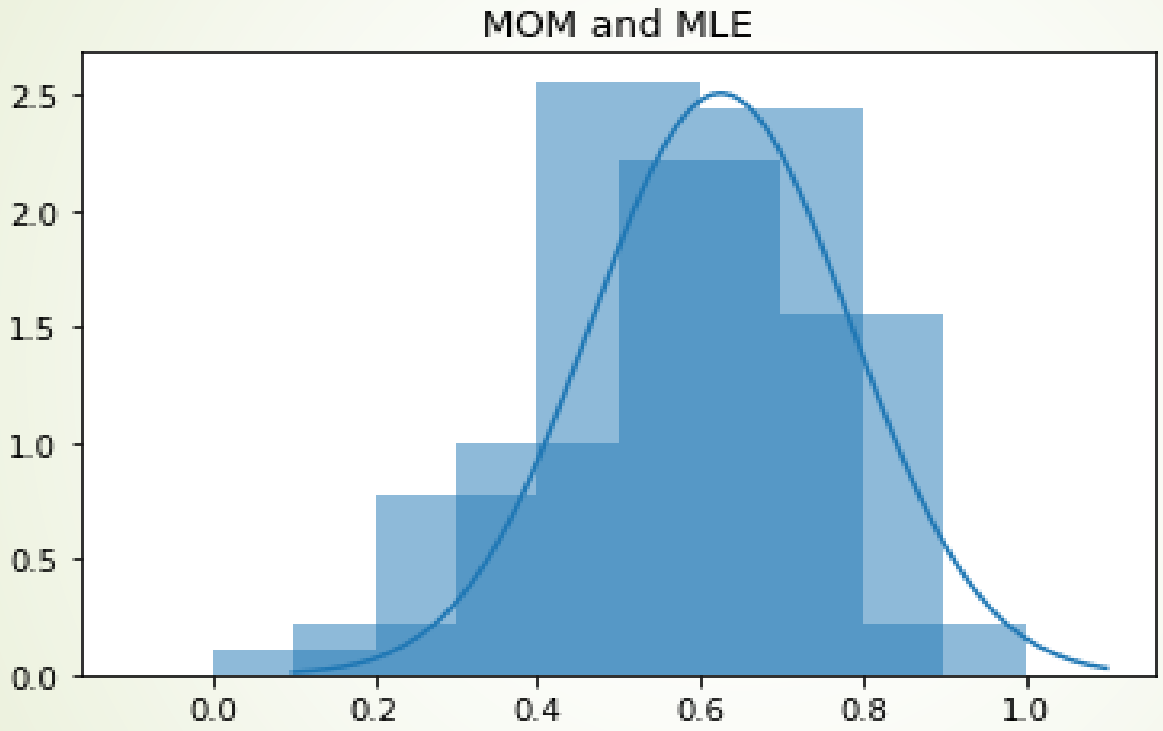
$$m_1 = E[X] = \mu = \frac{1}{n}\sum X_i$$

$$m_2 = E[X^2] = \mu^2 + \sigma^2 = \frac{1}{n}\sum X_i^2$$

The following results have been obtained from the observation:

- m_1 = Mean (μ) = 0.6255
- m_2 = Second moment =0.41660846
- Variance = 0.02535821
- Standard Deviation(σ) = 0.1592426

Since our distribution is a normal distribution, the MOM and MLE yield the same result.



It is clearly visible that the graph peaks around the same value as the original data.

Beta-Binomial Bayesian Analysis

PRIOR DATA:

First ,we fit our prior data to a Beta distribution ,$\pi \sim Beta(\alpha,ß)$.
Based on the data and calculations done,we could obtain the values of α and ß.
Formula for mean and variance of Beta distribution

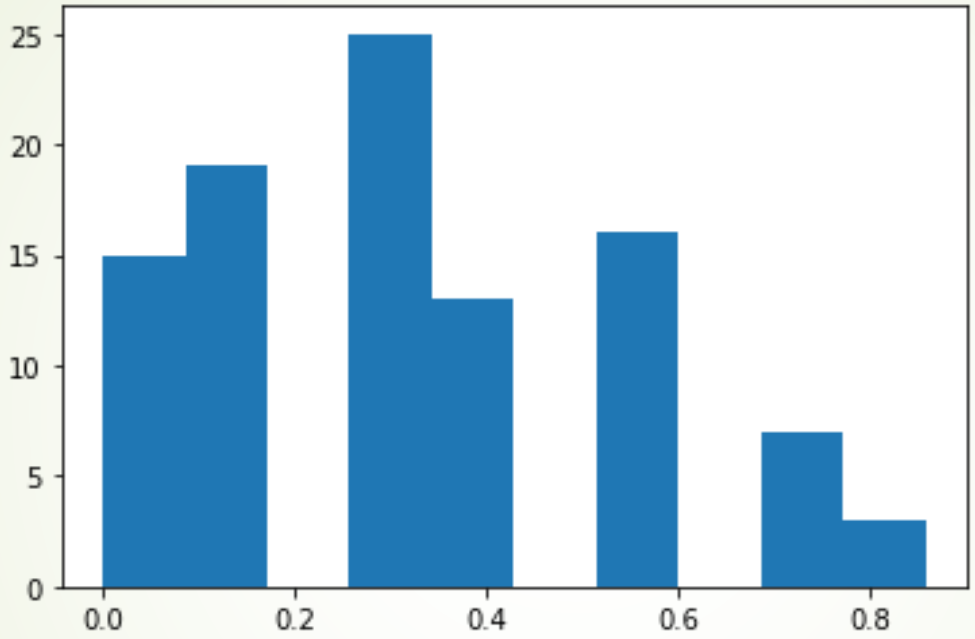
$$mean(\mu) = \alpha/\alpha+\beta$$
$$var(\sigma) = \alpha\beta(\alpha+\beta)/2(\alpha+\beta+1)$$

```
alpha = ((1-mean)/var - 1/mean)*pow(mean,2)
print("alpha = "+str(alpha))

beta = alpha*(1/mean - 1)
print("beta = "+str(beta))

prior = st.beta(alpha,beta)
```

```
alpha = 5.152637282757734
beta = 3.0849922660156226
```



Prior data frequency plot

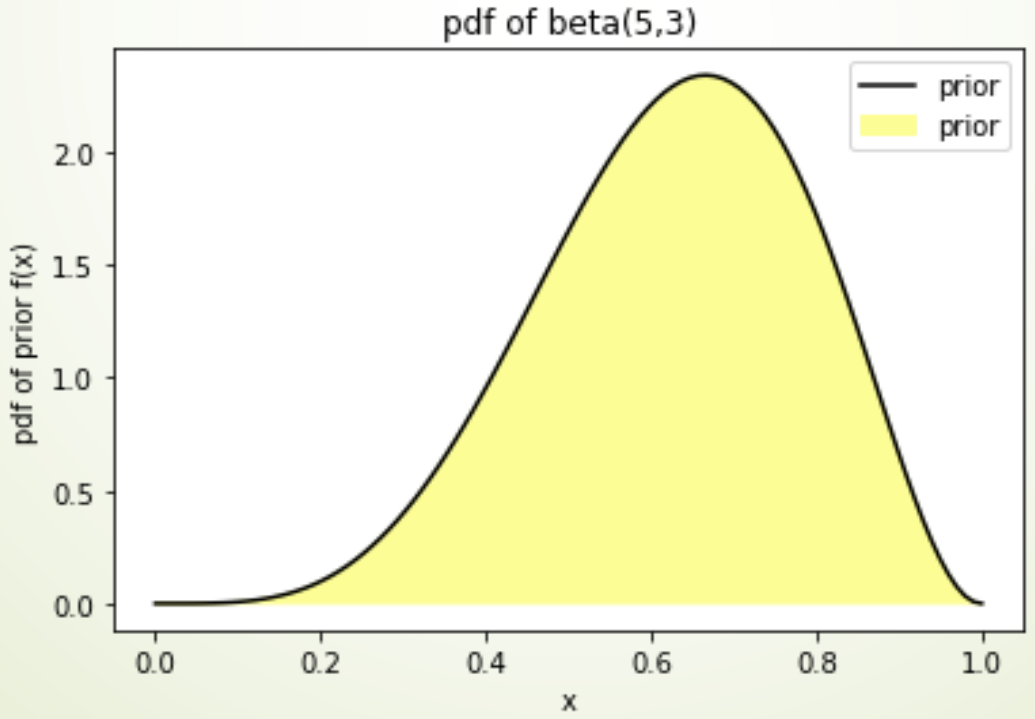
NECESSARY CALCULATIONS

Using the values obtained during MLE & MOM,we can estimate values of α and ß.

Prior mean = 0.6255

ß = 3.084~3

α =5.152~5



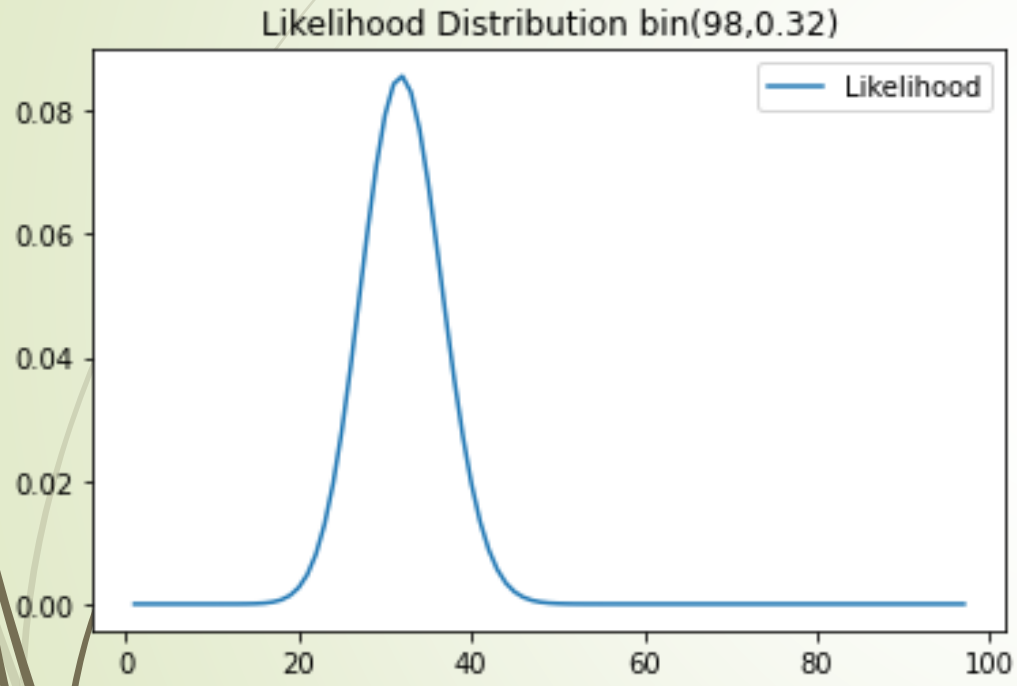
We require a likelihood function to perform a beta-binomial analysis on the generated beta distribution.

• We choose $Y \mid \pi \sim Bin(n, \pi)$ where,
n:-The number of songs on which we wish to analyse the energy levels.
Π=The probability that song's energy level is above the mean energy level

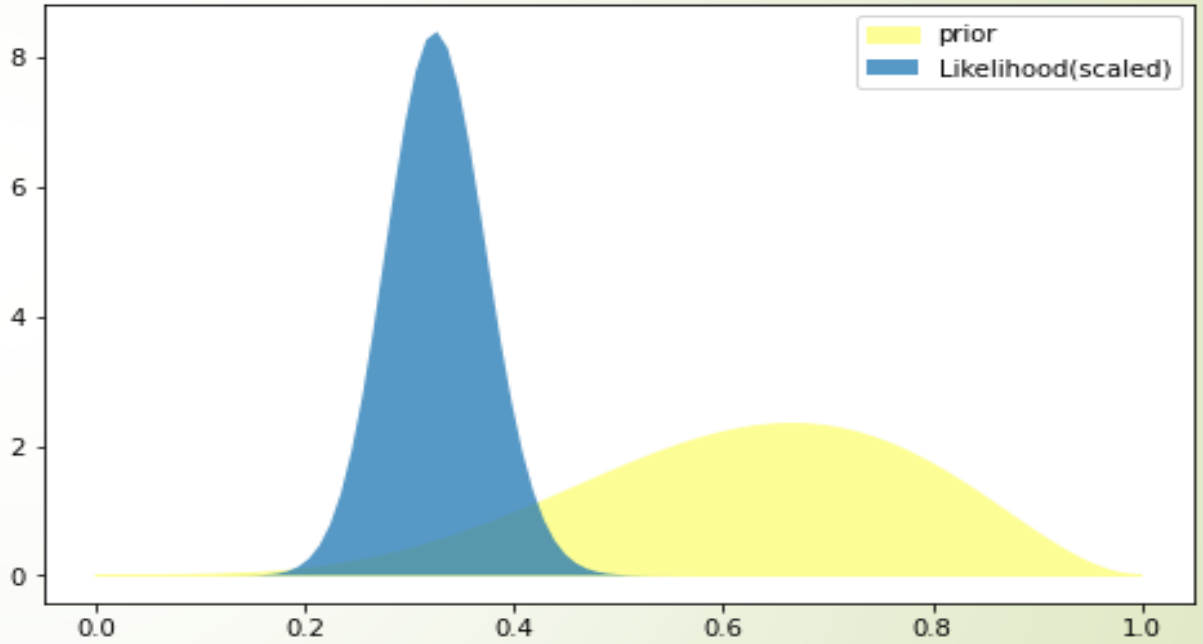
• Our data consists of the realized proportion of energy levels, where we have done the Bayesian analysis keeping the number of songs to be 32,which is the number of songs having energy level greater than the mean 'energy level'.

$n = 98$,$y = 32$ & $p = 0.3265$ $\left(p = \frac{y}{n}\right)$

LIKELIHOOD DISTRIBUTION



Likelihood distribution plot



Prior distribution and data likelihood plot

POSTERIOR DISTRIBUTION

To find the posterior distribution using the prior data and the likelihood function.

We know that if

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

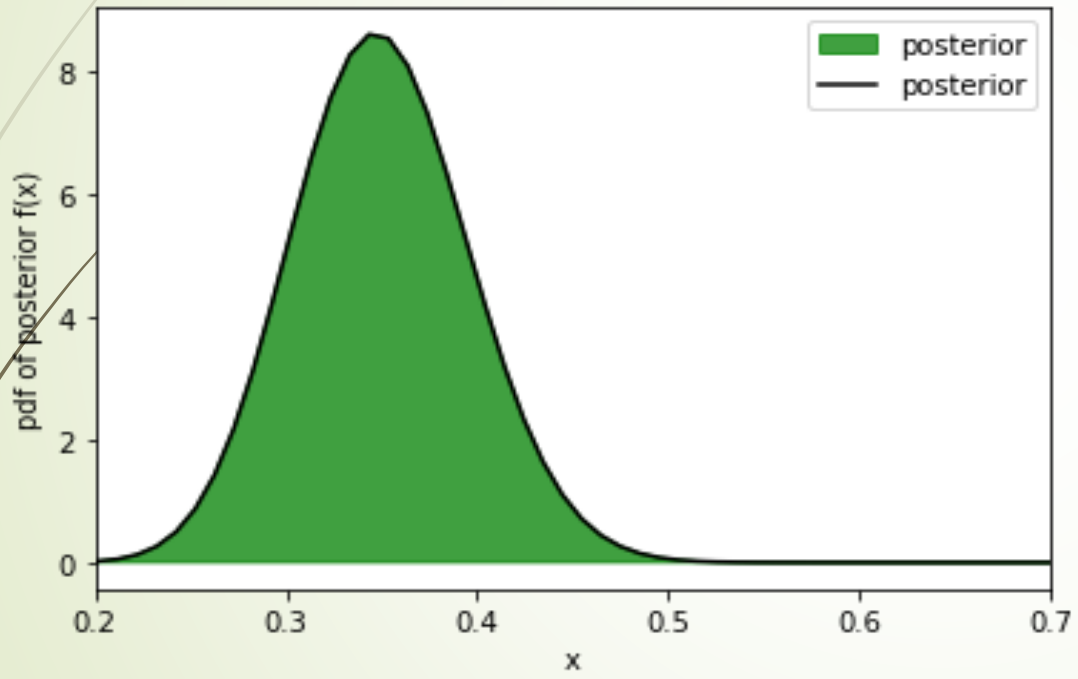
Data-Likelihood: $Y|\pi \sim \text{Bin}(n, \pi)$

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

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

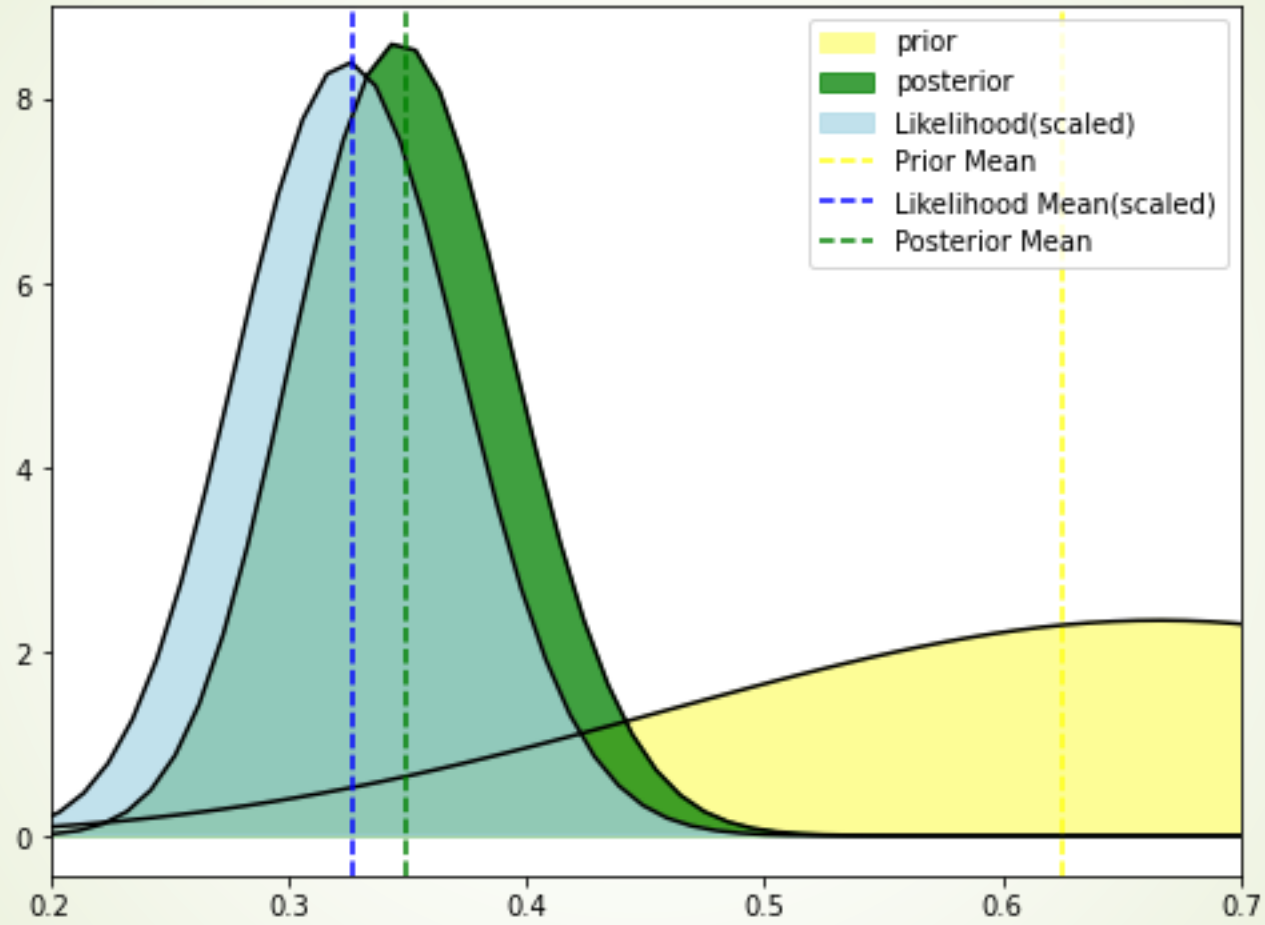
$n = 98$, $y = 32$ & $p=0.3265$

POSTERIOR DISTRIBUTION



Obtained values :

- Posterior α = 37.152
- Posterior ß = 69.084
- Posterior μ = 0.3497



Combined plot

INFERENCES:

- As 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 now conclude that given the value of y=32,there is 34% chance that the music would have higher energy level than the average level.

Thank You!!