भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Real Data Analysis

MA4740 - Introduction to Bayesian Statistics

GROUP - 8

April 21, 2023

Team Members

- Kethari Narasimha Vardhan - MA20BTECH11006
- Mulugu Vishwanath Sharma - MA20BTECH11010
- Prajwaldeep Kamble - MA20BTECH11013

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 Poisson Gamma Bayesian Analysis.

Introduction

Objective

The project includes:

- Real Data Analysis on Dataset:-
  - Rainfall in India from 1901 to 2015.
- Performing Poisson Gamma Bayesian Analysis on the collected Dataset.

Data Collection

Data Collection

The Dataset

- The Dataset, Rainfall in India from 1901 to 2015 is based on the Aggregate Rainfall of each state in India from 1901 to 2015. This dataset is collected from Kaggle.
- The data includes all the states, the rainfall in every state in every month throughout the years 1901 up till 2015.
- Amongst which, we are interested in the rainfall statistics of Telangana State.

Glimpse of the Dataset

Rainfall in India 1901-2015

SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL	Jan-Feb	Mar-May	Jun-Sep	Oct-Dec
TELANGANA	1901	6.90	41.80	7.80	45.20	22.00	123.60	237.80	177.20	77.70	75.50	12.20	0.00	827.70	48.70	75.00	616.40	87.70
TELANGANA	1902	0.00	0.00	0.20	10.70	7.30	52.40	146.30	142.80	190.50	41.70	31.20	7.30	630.40	0.00	18.20	532.00	80.20
TELANGANA	1903	12.90	4.60	0.00	9.90	40.70	99.20	505.20	246.70	191.90	155.80	15.50	1.10	1283.40	17.50	50.50	1042.90	172.40
TELANGANA	1904	0.00	0.00	10.80	0.80	14.70	104.20	139.50	50.00	162.30	44.40	0.00	0.00	526.70	0.00	26.30	456.00	44.40
TELANGANA	1905	0.00	4.30	12.80	27.60	32.20	129.50	82.40	237.30	179.10	19.60	0.00	0.00	724.90	4.30	72.60	628.40	19.60
TELANGANA	1906	22.50	1.20	13.40	2.40	0.70	211.10	210.80	226.70	96.30	20.50	14.90	34.80	855.20	23.70	16.50	744.80	70.20
TELANGANA	1907	1.00	3.30	10.20	61.90	0.20	217.50	160.50	263.30	116.80	0.30	3.60	5.00	843.70	4.30	72.30	758.10	8.90
TELANGANA	1908	35.60	2.60	5.20	0.30	6.50	107.40	254.90	168.30	401.20	0.10	0.00	0.70	982.80	38.30	12.00	931.80	0.80
TELANGANA	1909	0.50	5.90	0.50	26.40	2.20	133.80	288.30	168.60	138.50	4.60	0.00	0.20	769.50	6.30	29.10	729.20	4.80
TELANGANA	1910	0.00	0.00	0.00	4.20	25.00	220.90	198.20	150.30	230.50	101.40	45.30	0.00	975.90	0.00	29.20	799.90	146.80
TELANGANA	1911	0.00	0.00	7.90	0.70	7.80	133.90	122.10	176.30	174.40	23.00	6.00	9.90	662.00	0.00	16.40	606.70	39.00
TELANGANA	1912	0.00	37.50	0.00	20.40	6.60	28.50	263.30	196.80	149.50	7.80	33.40	0.00	743.80	37.50	26.90	638.10	41.30
TELANGANA	1913	0.00	13.40	0.00	8.00	34.60	83.50	337.20	108.80	101.70	35.20	0.00	9.70	732.20	13.40	42.60	631.30	44.90
TELANGANA	1914	0.00	0.00	1.20	34.10	41.50	233.10	271.30	195.10	278.60	16.20	10.30	2.00	1083.50	0.00	76.80	978.10	28.50
TELANGANA	1915	15.40	8.60	77.80	18.80	26.70	165.60	140.40	236.20	186.80	122.00	23.60	0.10	1022.00	24.00	123.30	729.00	145.70
TELANGANA	1916	0.00	4.10	0.00	15.90	16.10	233.20	216.40	137.00	291.80	153.10	95.10	0.00	1162.60	4.10	32.00	878.40	248.10
TELANGANA	1917	0.00	65.50	33.00	34.10	52.70	184.10	267.50	180.10	275.60	106.70	6.60	0.00	1206.00	65.50	119.90	907.30	113.40
TELANGANA	1918	20.20	0.00	32.60	8.30	55.50	97.80	106.10	89.60	96.90	6.40	7.50	33.90	554.70	20.20	96.40	390.30	47.80
TELANGANA	1919	5.50	39.20	31.30	35.00	23.60	158.80	139.40	115.30	129.90	130.60	78.50	7.40	894.60	44.80	89.90	543.40	216.60
TELANGANA	1920	7.40	0.00	1.70	24.40	31.40	62.80	138.80	66.50	78.90	24.90	0.00	0.00	437.00	7.40	57.60	347.10	24.90
TELANGANA	1921	7.10	0.00	0.10	5.40	5.50	200.90	296.70	161.10	240.70	68.00	17.90	0.00	1003.30	7.10	11.00	899.40	85.90
TELANGANA	1922	55.30	0.00	0.00	6.60	50.10	69.20	184.30	158.30	170.20	23.00	50.70	0.00	767.50	55.30	56.70	581.90	73.70
TELANGANA	1923	3.00	5.30	23.30	10.70	18.00	43.50	222.80	80.30	334.40	44.00	1.00	0.00	786.20	8.30	52.00	680.90	45.00
TELANGANA	1924	37.00	0.00	0.00	9.20	34.60	62.80	124.70	220.20	288.20	51.60	116.10	0.00	944.50	37.00	43.80	695.90	167.70

Figure 1: Rainfall Statistics in Telangana from 1901 - 1924

Defining variables

The below representations of the data we used regarding the variables we learned in class.

$$\lambda \sim \mathit{Gamma}(\alpha, \beta) \tag{1}$$

$$Y \mid \lambda \sim \mathit{Poisson}(\lambda) \tag{2}$$

$$\lambda \mid Y \propto f(Y \mid \lambda) * f(\lambda) \tag{3}$$

Where,

$$\lambda = \text{Amount of rainfall in a year}$$
$$Y = \text{Data of rainfall over the } n \text{ years}$$
$$(\text{Random sample of size } n,\ Y_i \mid \lambda \sim \mathit{Poisson}(\lambda))$$

Poisson Gamma Bayesian Data Analysis

Prior Distribution

Prior Distribution

- The prior data is taken from the dataset from the years 1901 to 1980.
- We choose the amount of rainfall in a year (λ) for the prior data.
- We try to fit the prior data to a Gamma Distribution.
  We get $\lambda \sim \mathit{Gamma}(\alpha, \beta)$.

Formula

$$\mu = \frac{\alpha}{\beta}, \quad \sigma = \frac{\alpha}{\beta^2}$$

Where, α is the shape hyper-parameter and β is the scale hyper-parameter of $\mathit{Gamma}(\alpha, \beta)$.

Prior Distribution

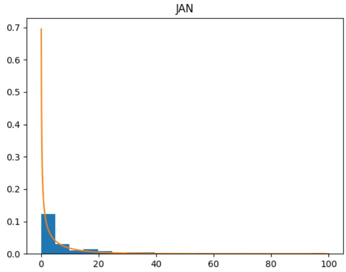
- From the 12 months of data that is collected, the Bayesian Analysis is performed on four months of the year.
  Namely, January, February, March, and April.
- After performing the necessary calculations, we get the values of the hyper-parameters α and β for these months as

Calculations

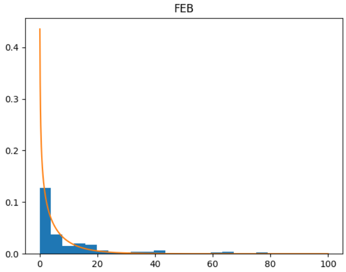
$$\alpha_{Jan} = 0.37, \quad \beta_{Jan} = 12.07, \quad PriorMean_{Jan} = 4.51$$
$$\alpha_{Feb} = 0.62, \quad \beta_{Feb} = 8.39, \quad PriorMean_{Feb} = 5.22$$
$$\alpha_{Mar} = 0.41, \quad \beta_{Mar} = 23.29, \quad PriorMean_{Mar} = 9.44$$
$$\alpha_{Apr} = 1.1, \quad \beta_{Apr} = 16.44, \quad PriorMean_{Apr} = 18.11$$

Prior Distribution

The corresponding Prior Distribution graphs are as follows



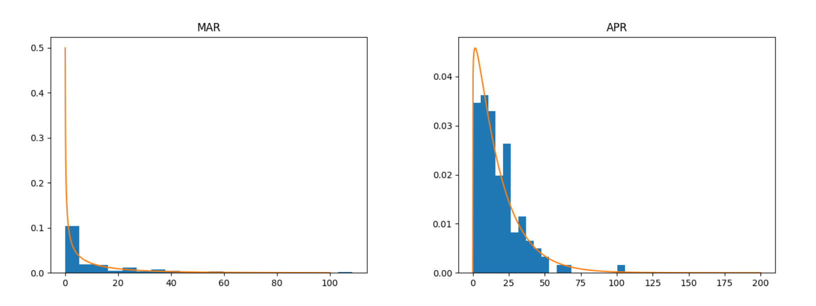
(a) Prior Distribution of January Month



(b) Prior Distribution of February Month

Prior Distribution

The corresponding Prior Distribution graphs are as follows



(a) Prior Distribution of March Month

(b) Prior Distribution of April Month

Data-Likelihood Function

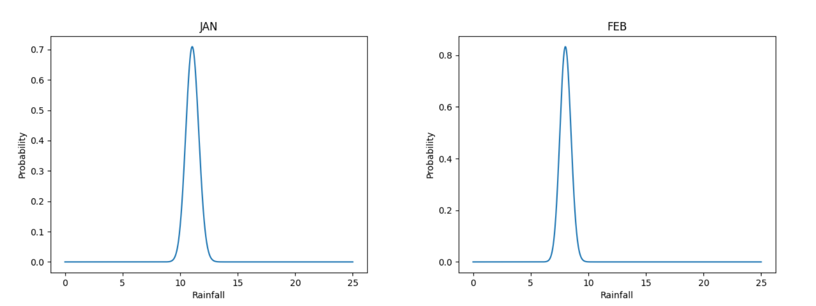
Data-Likelihood Function

- We require a likelihood function to perform a Poisson Gamma Bayesian Analysis.
- We choose $Y \mid \lambda \sim Poisson(\lambda)$
- Our data consists of the realized values of average rainfall from the years 1981 to 2015.

Joint Likelihood

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_{35} = y_{35} \mid \lambda) = \frac{e^{-n\lambda}\lambda^{\Sigma y_i}}{\prod_{i=1}^{i=15} y_i!}$$

The corresponding Likelihood Function graphs are as follows

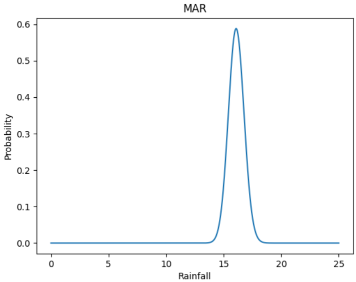


(a) Likelihood for January Month

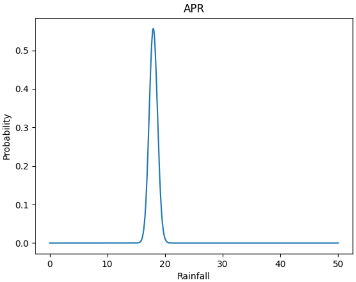
(b) Likelihood for February Month

Data-Likelihood Function

The corresponding Likelihood Function graphs are as follows



(a) Likelihood for March Month



(b) Likelihood for April Month

Posterior Distribution

Posterior Distribution

Using the prior distribution and likelihood function, we try to calculate the posterior distribution. We have,

Definition

If,

$$\text{Prior: } \lambda \sim \mathit{Gamma}(\alpha, \beta)$$
$$\text{Likelihood: } Y \mid \lambda \sim \mathit{Poisson}(\lambda)$$

Then,

$$\text{Posterior: } \lambda \mid Y \sim \mathit{Gamma}(\alpha + \Sigma y_i, \beta + n)$$

Posterior Distribution

Thus, after performing the necessary calculations, we get the following values for $Gamma(\alpha', \beta')$, with

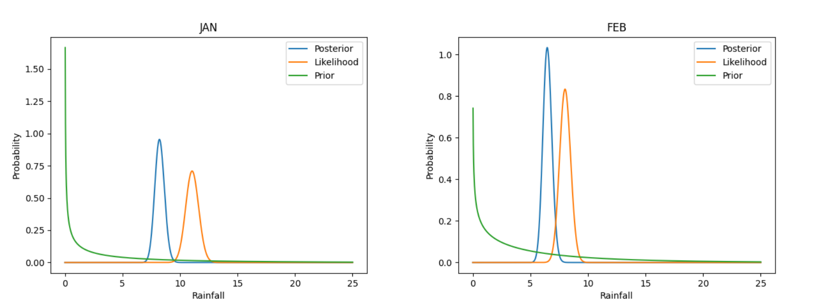
$$\Sigma y_i = y_1 + y_2 + ... + y_n$$
$$\text{and } n = 35$$

Calculations

$$\alpha'_{Jan} = 387.77, \quad \beta'_{Jan} = 47.07, \quad PosteriorMean = 8.24$$
$$\alpha'_{Feb} = 281.52, \quad \beta'_{Feb} = 43.39, \quad PosteriorMean = 6.49$$
$$\alpha'_{Mar} = 563.91, \quad \beta'_{Mar} = 58.29, \quad PosteriorMean = 9.67$$
$$\alpha'_{Apr} = 630.4, \quad \beta'_{Apr} = 51.44, \quad PosteriorMean = 12.26$$

Posterior Distribution

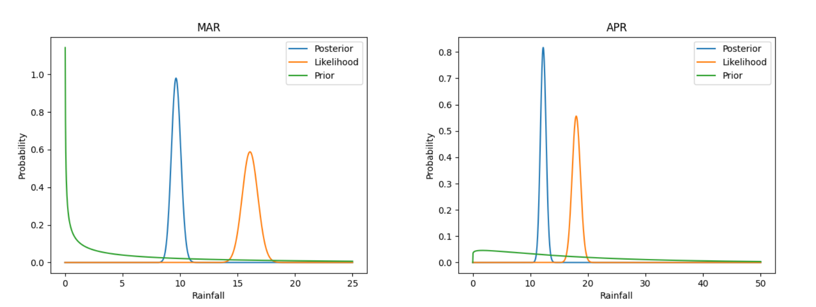
The combined graphs of Prior, Likelihood, and Posterior are as follows



(a) Combined Graph for January Month **(b)** Combined Graph for February Month

Posterior Distribution

The combined graphs of Prior, Likelihood, and Posterior are as follows



(a) Combined Graph for March Month **(b)** Combined Graph for April Month

Conclusion

If we calculate the means and variances, we can see that the mean of our prior and posterior differ very slightly whereas the variance of prior and posterior differ by a very large margin.

For Jan

Prior std = 7.38, Posterior std = 0.42

For Feb

Prior std = 6.62, Posterior std = 0.39

For Mar

Prior std = 14.83, Posterior std = 0.41

For Apr

Prior std = 17.25, Posterior std = 0.49

Conclusion

Below are the 95% confidence intervals for each month.

$$\text{For Jan}$$
$$\text{Prior interval} \approx (0.0, 25.87),$$
$$\text{Posterior interval} \approx (7.44, 9.08)$$
$$\text{For Feb}$$
$$\text{Prior interval}\approx (0.02, 23.75),$$
$$\text{Posterior interval} \approx (5.75, 7.27)$$
$$\text{For Mar}$$
$$\text{Prior interval} \approx (0.0, 52.19),$$
$$\text{Posterior interval}\approx (8.89, 10.49)$$
$$\text{For Apr}$$
$$\text{Prior interval}\approx (0.61, 64.10),$$
$$\text{Posterior interval}\approx (11.32, 13.23)$$

Inference

1. The confidence interval for the posterior distribution is narrower than the prior's interval and the variance of the posterior distribution is very less than the prior variance so our goal of getting a value more precise than the expectation of the prior is achieved.
2. And now if you observe the values of the mean Posterior distributions you can observe that the expected rainfall in Telangana is increasing gradually and if you plot the mean of the posterior distribution of all the months you will encounter a rise in expected rainfall and a drop.
3. the peak is attained in the rainy season and followed by a drop which is attained in the winter.

Thank You