Statistics Project - 2

By :

Garvit Khedar
Vinaykarthik V
Tanmay Banjari

Describing our dataset

- To begin with, we have obtained the requisite data for the first two sections of this project from Investing.com.
- **Source :** “https://in.investing.com/indices/s-p-cnx-nifty-historical-data”
- The above dataset contains the daily trading volumes of the NIFTY 50 index from the year 2000 to 2024.
- The dataset for the remaining portion of the project has been obtained from Kaggle.
- **Source :** “https://www.kaggle.com/datasets/lainguyn123/student-performance-factors”
- The latter dataset consists of 6,607 records and 20 different attributes, which describe various aspects of students’ academic career.

Describing our dataset

- For the purpose of this project, we have chosen to work with the “Parental_Education” and “Exam_Score” attributes for the third section and the “Previous_Scores” and “Exam_score” attributes for the fourth section respectively.
- We shall describe the context of each attribute at the beginning of the corresponding section.
- Furthermore, we have neglected all records where the above mentioned attributes have null values for simplicity.
- To conclude, we treat both the datasets as the population of interest and randomly sample from them. This approach enables us to perform and verify our statistical analysis on the data.

Section - 1 :

Estimating parameters using MoM and MLE for a dataset with Gamma distribution

Extracting the necessary data

Initial Columns:

- Date
- Open, Close, High, Low
- Volume (our focus)

Steps Performed in segregation.py:

- Removed all columns except Date and Vol.
- Created a new column: Volume (in Million)
  - If value ends with 'B' → multiplied by 1000
  - If value ends with 'M' → kept as-is
- Finally removed the Vol. column
- Saved the modified dataset as: Final Data1.csv

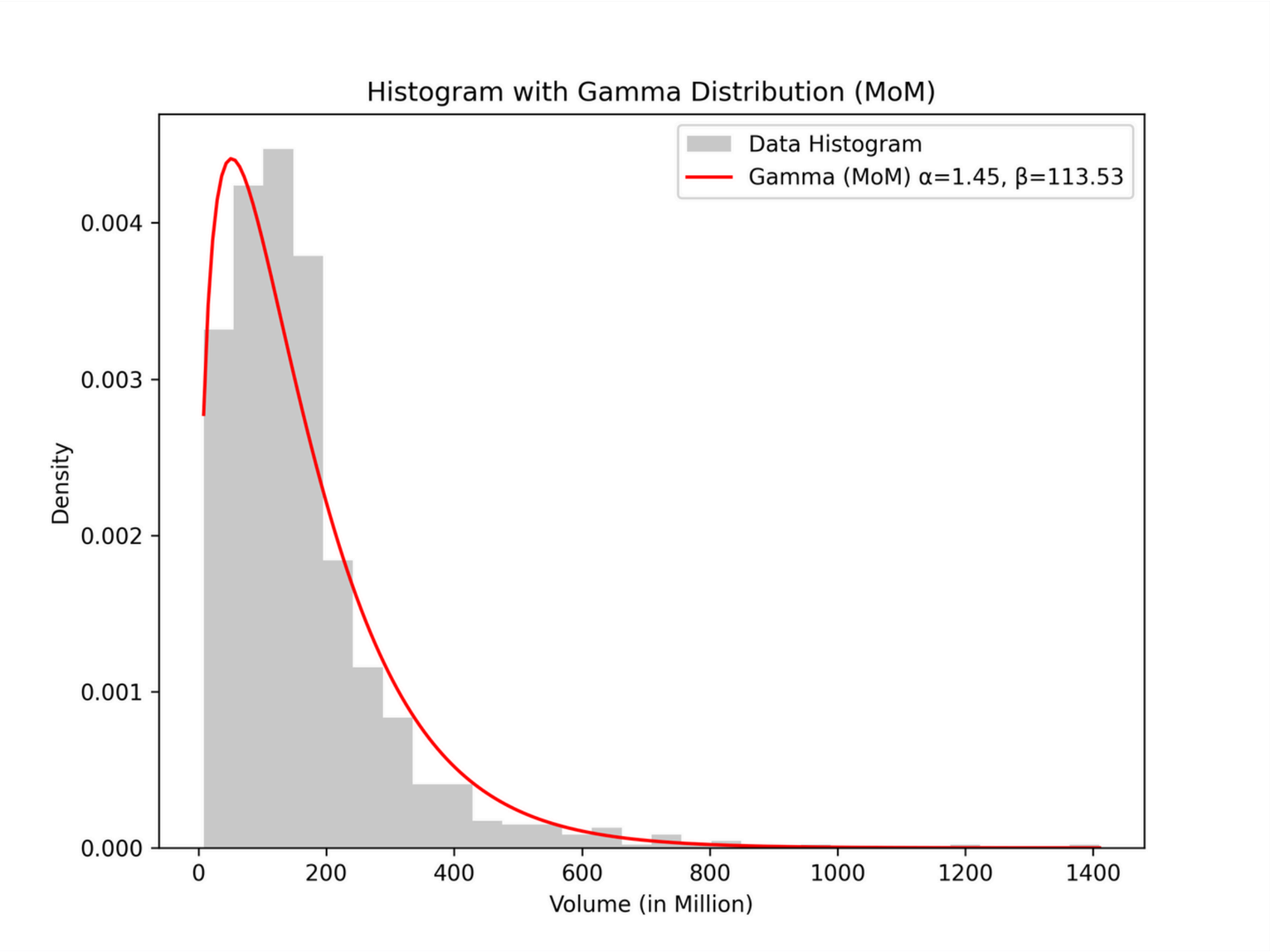
Approach Overview

1. Extracted a random 20% sample from the dataset.
2. Assumed the sampled data follows a Gamma distribution with parameters α and β.
3. Estimated the parameters using the Method of Moments (MoM).
4. Applied Maximum Likelihood Estimation (MLE) using the digamma function (explained later)

Method of Moments (MoM) estimator

- Sample Moments (First and Second Moments):
  - E[X] = sampled_data.mean() = m1
  - E[X^2] = (sampled_data^2).mean() = m2
- Population Moments for Gamma(α, β) (First and Second Moments):
  - E[X] = α * β
  - E[X^2] = α * (β^2) + (α * β)^2 (Var(X) = E(X^2)-(E(X))^2)
- Method of Moments Equating:
  - Equating sample and population moments:
    - beta_mom = (m2 - m1^2) / m1
    - alpha_mom = m1 / beta_mom

Gamma Distribution Fit Using Previously Calculated MoM Parameters



Finding the MLE estimator

- There is no closed-form solution for Maximum Likelihood Estimation (MLE) of Gamma distribution parameters. To overcome this dilemma, we use the digamma function as defined below

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

- Let the likelihood function be denoted by L,

$$\begin{aligned} \log L(\alpha, \beta \mid x_1, x_2, \ldots, x_n) &= \sum_{i=1}^{n} \log f(\alpha, \beta \mid x_1, x_2, \ldots, x_n) \\ &= -n(\alpha \log \beta + \log \Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \frac{1}{\beta} \sum_{i=1}^{n} x_i. \end{aligned}$$

- Differentiating with respect to β and setting the derivative to zero gives

$$\begin{aligned} \frac{\partial \log L(\alpha, \beta \mid x_1, x_2, \ldots, x_n)}{\partial \beta} &= -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i \\ n\alpha\beta &= \sum_{i=1}^{n} x_i \\ \hat{\beta} &= \frac{\bar{x}}{\hat{\alpha}} \end{aligned}$$

Finding the MLE estimator

- Similarly, we equate the partial derivative of the log likelihood function w.r.t. to obtain

$$\frac{\partial \log L(\alpha, \beta \mid x_1, x_2, \ldots, x_n)}{\partial \alpha} = -n\psi(\alpha) - n \log \beta + \sum_{i=1}^{n} \log x_i = 0$$

- Substituting β = x̄ / α into the above equation yields

$$\psi(\alpha) + \log\left(\frac{\bar{x}}{\alpha}\right) = \frac{\sum_{i=1}^{n} \log x_i}{n} = \sum_{i=1}^{n} \log x_i^{\frac{1}{n}}$$

Finding the MLE estimator

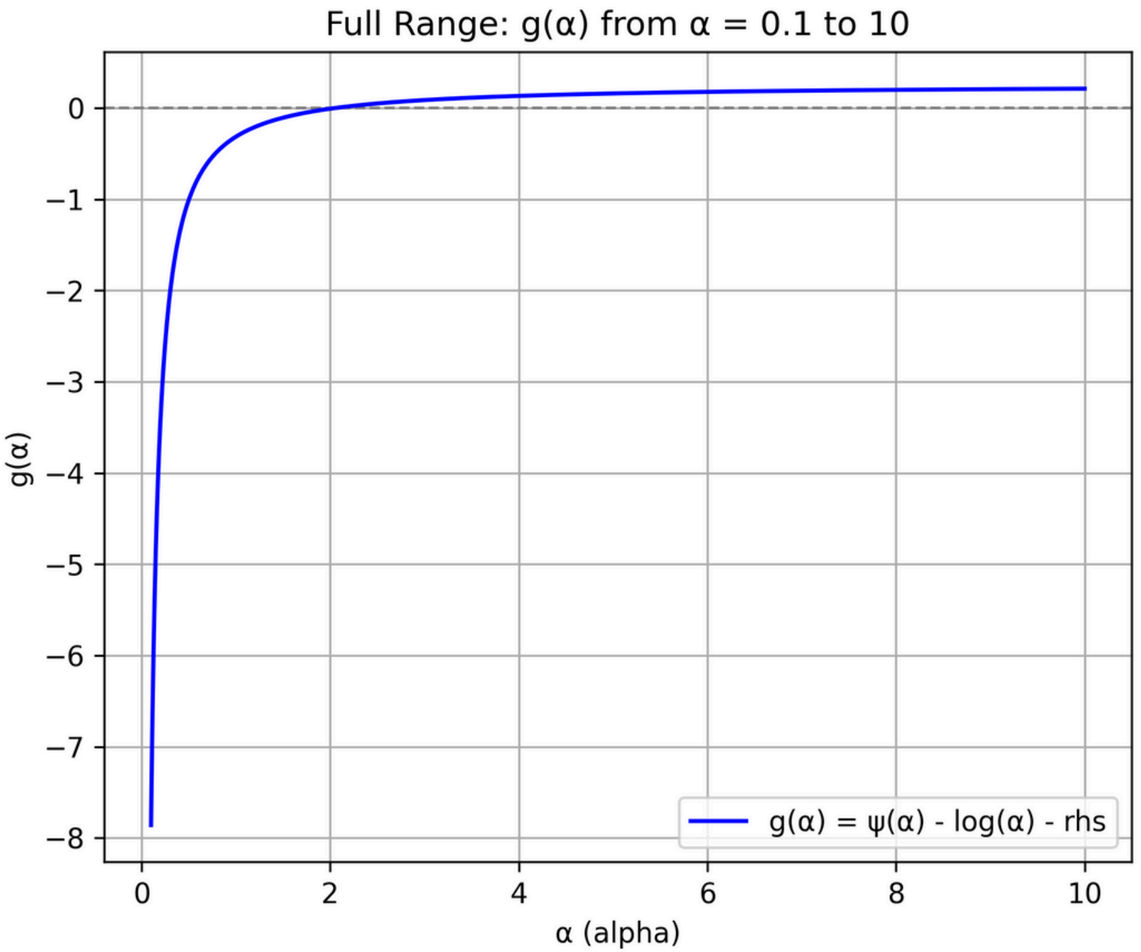
- Upon further simplification, we obtain the following

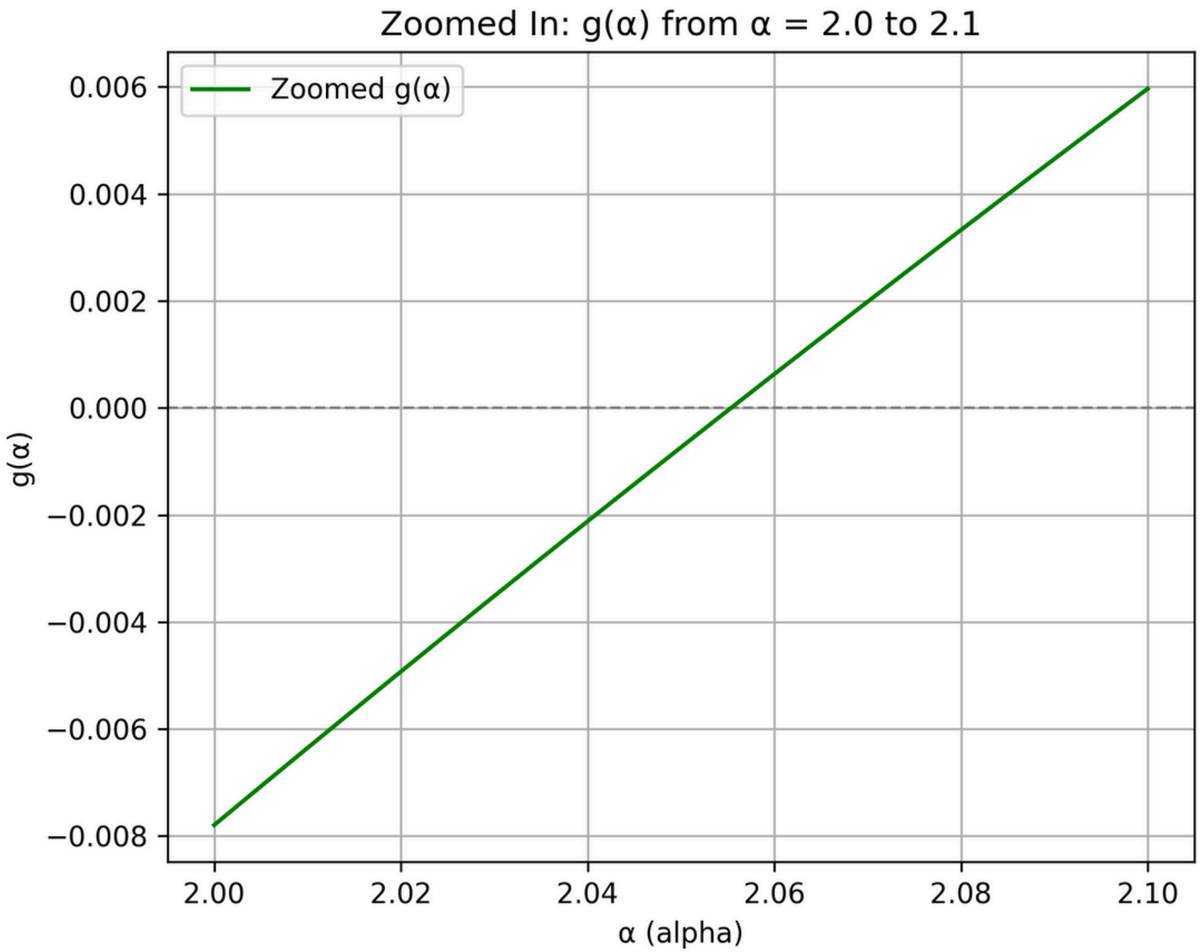
$$\psi(\alpha) - \log \alpha = \sum_{i=1}^{n} \log x_i^{\frac{1}{n}} - \log \bar{x}$$

$$\psi(\alpha) - \log \alpha = \sum_{i=1}^{n} \log \frac{x_i^{\frac{1}{n}}}{\bar{x}}$$

- To proceed further, define g(α) := ψ(α) − log α - (RHS in the above equation). We now plot the graph of g(α) and check where it intersects the x - axis to obtain an estimate for the value of α.

Estimating the value of α



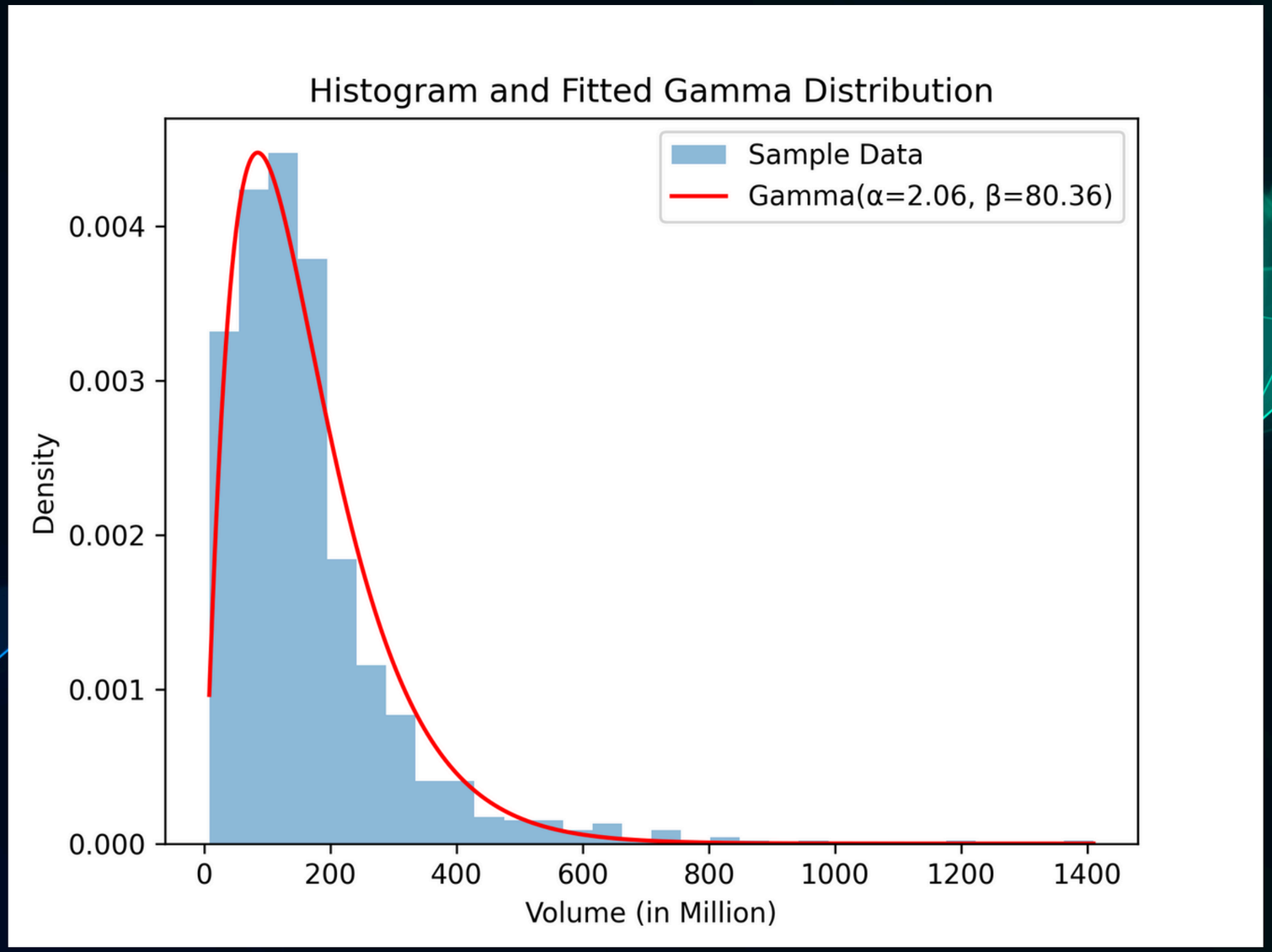


Finding the value of α and β

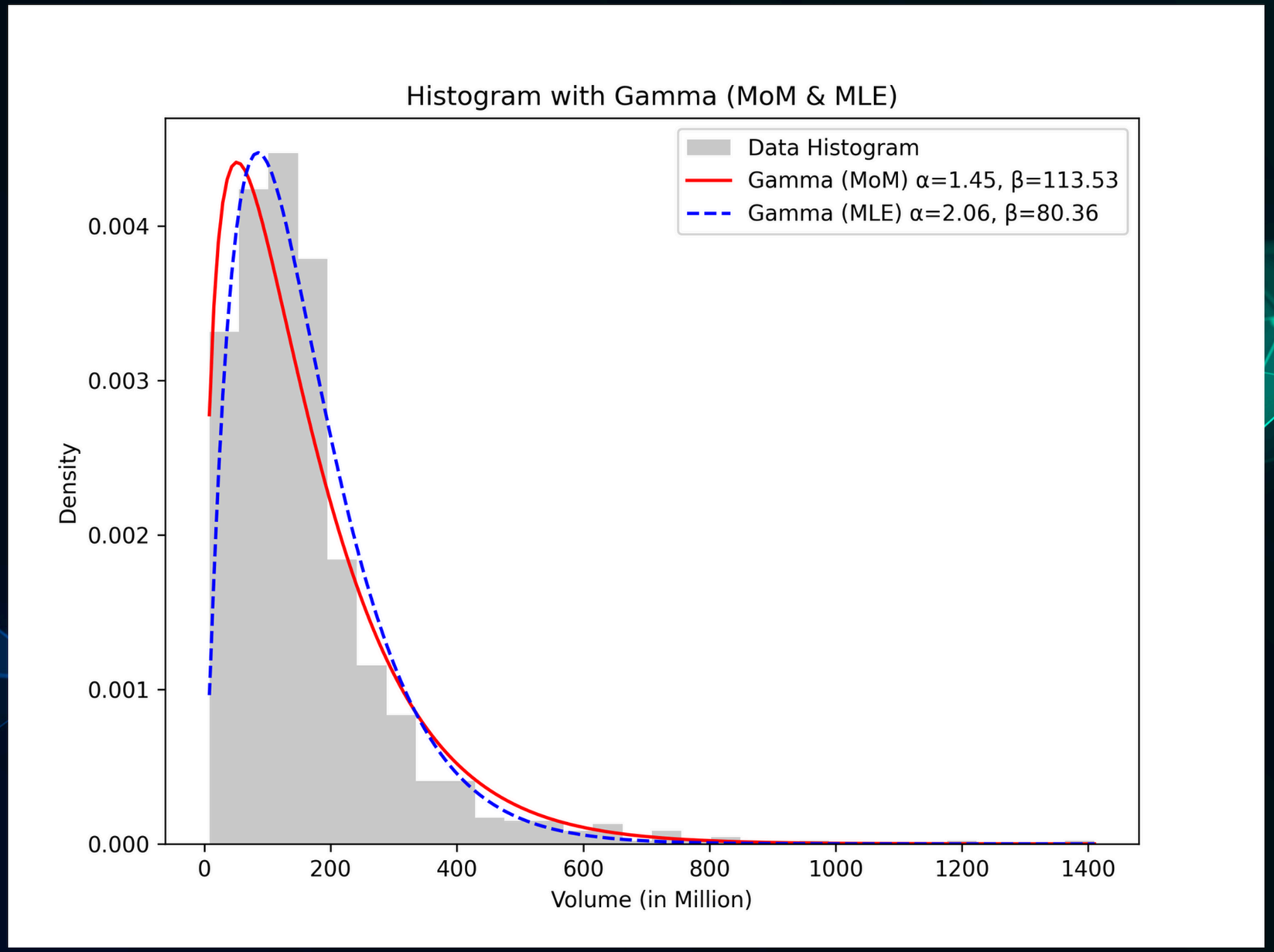
- The full plot shows that g(α) crosses the y = 0 line around α ∈ [2.0, 2.1].
- So we zoomed in on this region and found the root at α ≈ 2.055, which is accurate to three decimal places.
- With the help of α_{mle}, we computed β_{mle} using:

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}$$

Gamma Distribution Fit Using Previously Calculated MLE Parameters



Summary


Histogram with Gamma (MoM & MLE)
Data Histogram
Gamma (MoM) α=1.45, β=113.53
Gamma (MLE) α=2.06, β=80.36
Density
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
1000
1200
1400
Volume (in Million)

Section - 2

Finding the 95% confidence interval for variance using the dataset in Section - 1, assuming it follows the normal distribution

95% Confidence Interval for Variance

- Assume that the data points (X_i) are independent and identically distributed and follow a normal distribution with unknown mean and variance.
- Now, we take a random sample of 20% size from the dataset (population).
- Goal: To construct a 95% confidence interval for the variance (alpha = 0.05).
- Based on the property that $(n - 1) * s^2 / \sigma^2$ follows a chi-square distribution with (n - 1) degrees of freedom, the 100(1 - α)% confidence Interval for Variance is given by

$$P\{\chi^2_{(1-\alpha/2,n-1)} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{(\alpha/2,n-1)}\} = 1-\alpha$$

95% Confidence Interval for Variance is:
(17225.452, 20529.576)

Section – 3:

Calculating 95% confidence interval for the difference in sample means of two datasets

95% Confidence Interval

- We have taken a sample of 1,000 exam scores from our dataset and divided it into two groups based on parental education: sample X (students whose parents have upto high school education) and sample Y (students whose parents have a college or postgraduate education).
- Assuming that the variances are unknown, we have the 100(1-α)% confidence interval as

$$P\{-t_{\alpha/2,n+m-2} \leq \frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq t_{\alpha/2,n+m-2}\} = 1-\alpha$$

where ‘t’ denotes the t-distribution, n - number of elements in X, m - number of elements in Y and Sp is the pooled estimator of X and Y given

$$S_p^2 = \frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}$$

where S_X and S_Y denote sample standard deviations of X and Y.

Calculations

The values of X are $n = 485$, $\bar{X} = 66.913$, $S_x^2 = 15.373$.
The values of Y are $m = 515$, $\bar{Y} = 67.445$, $S_y^2 = 13.835$.
On substituting the values given above, we get $S_p = 3.819$,
and the 95% confidence interval is $(-1.006, -0.058)$

Since the 95% confidence interval for the difference in means is (-1.006, -0.058), we can conclude that there is a slight difference in exam scores between the two groups. Specifically, students whose parents have a college or postgraduate education tend to perform better on average.

Section – 4 :

Hypothesis testing for bernoulli distributed data

Modelling the Bernoulli random variable

- We use the attributes “Previous_scores” (average of all exam scores except the final one) and “Exam_Score” (final exam score) to construct a Bernoulli random variable, as described below.
- Let us define a Bernoulli random variable X_i corresponding to the i^{th} student, whose value is 1 if the final score of the student is more than the previous score, and is 0 otherwise.
- Since the students with $X_i = 1$ have shown improvement in their scores, the probability of success p therefore represents the proportion of students who have shown improvement in their final exams.
- Now, we draw a random sample of 100 students from the original dataset and record their X_i values to construct the Bernoulli sample for further analysis.

Formulating the testing procedure

- We aim to test the hypothesis H_0 : p <= 1/2 vs H_1 : p > 1/2, which is equivalent to testing H_0 : p = 1/2 vs H_1 : p > 1/2. For doing so, we calculate the p-value (or the critical value) and test this equivalent form of the hypothesis.
- To begin with, we compute the value of the following test statistic

$$t = \frac{(succ - 0.5 - np_0)}{\sqrt{np_0(1-p_0)}}$$

- Here, succ denotes the number of favorable outcomes (X_i = 1), p_0 = 1/2 is the proposed value of p in the hypothesis, and n is the sample size. Also, we subtract 0.5 to incorporate the continuity correction factor. For our particular sample, this yields t = -3.5.

Inferences

- Since **p** = P(Z > t), the **p**-value is **p** = 0.999767, which is much higher than the given level of significance α = 0.05, implying that H_0 should be accepted.

- Additionally, since the **p**-value is quite high, we can almost always accept H_0 for all practical purposes, regardless of the level of significance α, since α typically lies between 0.01 and 0.1.

- The success probability for the selected sample is p = 0.33, indicating that roughly one out of three randomly selected students tend to show improvement in their final exam scores.

THANK YOU!