Computational Project-2

Applied Statistics - MA4240

Sai Kaushik .P - AI23BTECH11017
Pranav JVS - AI23BTECH11009
Samagalla Sukesh Kumar - AI23BTECH11024

1 Introduction

The Data file used for this project was `Admission_Predict_Ver1.1.csv`. It's Source being (Source) from www.kaggle.com.

2 Observations from each question of Problem statement

Question 1

1. Data Overview

- **Variable**: `CGPA` from *Admission_Predict_Ver1.1.csv*
- **Sample Size (n)**: 500
- **Sample Mean**: ≈ 8.576
- **Sample Variance**: ≈ 0.366

2. Estimation Methods

2.1 Method of Moments (MoM) Using moment equations:

$$a = \frac{\bar{x}^2}{s^2}, \quad b = \frac{\bar{x}}{s^2}$$

- **Estimates**: $\hat{a}_{\text{MoM}} \approx 201.082, \quad \hat{b}_{\text{MoM}} \approx 23.446$

2.2 Maximum Likelihood Estimation (MLE) **a) Using SciPy's `gamma.fit` with `loc = 0`:**

$$\hat{a}_{\text{MLE}} \approx 200.169, \quad \hat{b}_{\text{MLE}} = \frac{1}{\text{scale}} \approx 23.339$$

b) Numerical Optimization
Minimizing the negative log-likelihood with bounds $a, b > 0$

c) Stirling's Approximation
Using:

$$a \approx \frac{1}{2(\log \bar{x} - \overline{\log x})}, \quad b = \frac{a}{\bar{x}}$$

$$\hat{a} \approx 200.002, \quad \hat{b} \approx 23.320$$

d) Newton-Raphson Iteration
Solving:

$$\log a - \psi(a) = \log \bar{x} - \overline{\log x}$$

$$\text{Converges to: } \hat{a} \approx 200.169, \quad \hat{b} \approx 23.339$$

- All methods yield **consistent estimates**.
- **MLE methods** (fit, optimization, iteration) align closely, $\Rightarrow$ **stable likelihood surface**.
- **Stirling's approximation** slightly overestimates but close.
- **MoM** is computationally simpler.

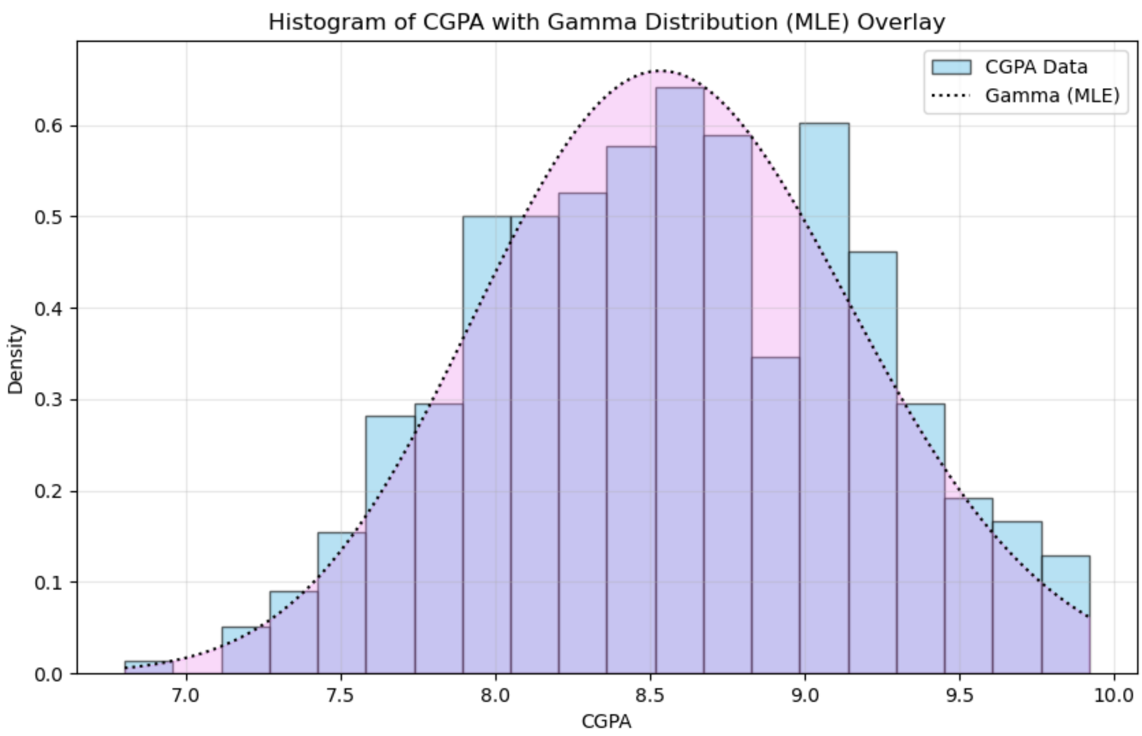

Figure 1: Histogram of CGPA with Gamma Distribution (MLE) Overlay

---

Question 2

1. Data Overview

- **Variable**: `CGPA`
- **Sample Size (n)**: 500
- **Sample Variance**: ≈ 0.366
- Assumed distribution: **Normal** with unknown mean and variance

2. Estimation Method

A 95% confidence interval for the **variance** σ^2, using the χ^2 distribution:

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} \right)$$

With:

- $\alpha = 0.05$
- Degrees of freedom: $df = n - 1 = 499$
- $\chi^2_{0.975} \approx 439.00$
- $\chi^2_{0.025} \approx 562.79$
- **Lower Bound** ≈ 0.32
- **Upper Bound** ≈ 0.42

- The interval quantifies uncertainty in the estimate of CGPA variance.
- Assuming the underlying data is approximately **normal**, given the central limit theorem at $n = 400$

Question 3

1. Data Overview

- **Populations**:
  - Group 1: *TOEFL Score*
  - Group 2: *GRE Score*
- **Sample Sizes**: $n = m = 500$
- **Sample Means**:
  - $\bar{x}_1 \approx 107.192$ (TOEFL)
  - $\bar{x}_2 \approx 316.472$ (GRE)
- **Sample Variances**:
  - TOEFL ≈ 36.989
  - GRE ≈ 127.500

2. Estimation Method

Assuming both populations follow **independent Normal distributions** with **unknown and equal variances**, the confidence interval for the difference of means $\mu_1 - \mu_2$ is computed using the **pooled variance**:

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}, n+m-2} \cdot s_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

- $t_{0.975,998} \approx 1.962$
- **Lower Bound** ≈ -210.41
- **Upper Bound** ≈ -208.15

- The confidence interval does **not contain 0**, implying a **statistically significant** difference in means.
- The negative interval indicates TOEFL scores are **significantly lower** in magnitude than GRE scores, consistent with their respective numerical scales.

Question 4

1. Data Overview

- **Variable**: `Research` (binary: 0 = No, 1 = Yes)
- **Distribution Assumed**: Bernoulli with success probability p
- **Sample Size**: 500
- **Observed Proportion (Sample Mean)**: ≈ 0.56

2. Hypothesis Test

We test:

$$H_0 : p \leq \frac{1}{2} \quad \text{vs.} \quad H_1 : p > \frac{1}{2}$$

Test Statistic: Standardized z-score using:

$$\text{threshold} = p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$$

- $\alpha = 0.05 \Rightarrow z_\alpha \approx 1.645$
- threshold ≈ 0.537

Rule:

- Reject H_0 if $\bar{x} > 0.537$

- Sample mean $\approx 0.56 \Rightarrow$ **reject** H_0

- Indicates **statistically significant evidence** that $p > 0.5$.
- The large sample size $\Rightarrow$ the **normal approximation** to binomial is correct.
- This result implies that applicants with research experience are slightly **more prevalent than not** in the dataset.
