Prepared by group 28

Applied Stats Project-2

April, 2025

Team 28 members:

1. Devesh Gautam (MA23BTECH11009)
2. Sivanesan (MA23BTECH11024)
3. Ansh Bhatia (MA23BTECH11003)

Table of Contents:

Gamma & Normal Distributions

Comparing MLE & MOM

Question 1 & 2

Dataset

Heights & Weights of students

BMI:
A new attribute we added to reflect the BMI of every student using their height and weight

Height:
Height of the student in inches

Gender:
Whether the student is male or female

Weight:
Weight of the student in pounds

Method of moments

As per the theory, we have to equate sample raw moments and population raw moments to get mom estimators for a and b(sacle parameter)

Sample raw moments:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Population raw moments:

$$u_1 = E[X] \qquad u_2 = E[X^2]$$

Method of moments

On substituting values and equating corresponding raw moments we get,

$$\hat{a}_{\text{MoM}} = \frac{m_1^2}{m_2 - m_1^2}$$

$$\hat{b}_{\text{MoM}} = \frac{m_2 - m_1^2}{m_1}$$

m1 : 25.4754

m2: 656.2547

Substituting values of m1 and m2 from our dataset we get

$\hat{a}_{\text{MoM}}$ 89.4444733183219 $\hat{b}_{\text{MoM}}$ 0.28481867861657806

Maximum Likelihood Estimator

Differentiating partially the log-likelihood function wrt a and b and then equating those to 0 we get the equations.

$$\log a - \psi(a) = \log \bar{X} - \frac{1}{n}\sum \log X_i$$

$$\hat{b} = \frac{\bar{X}}{a}$$

- $\psi(a)$ is the digamma function

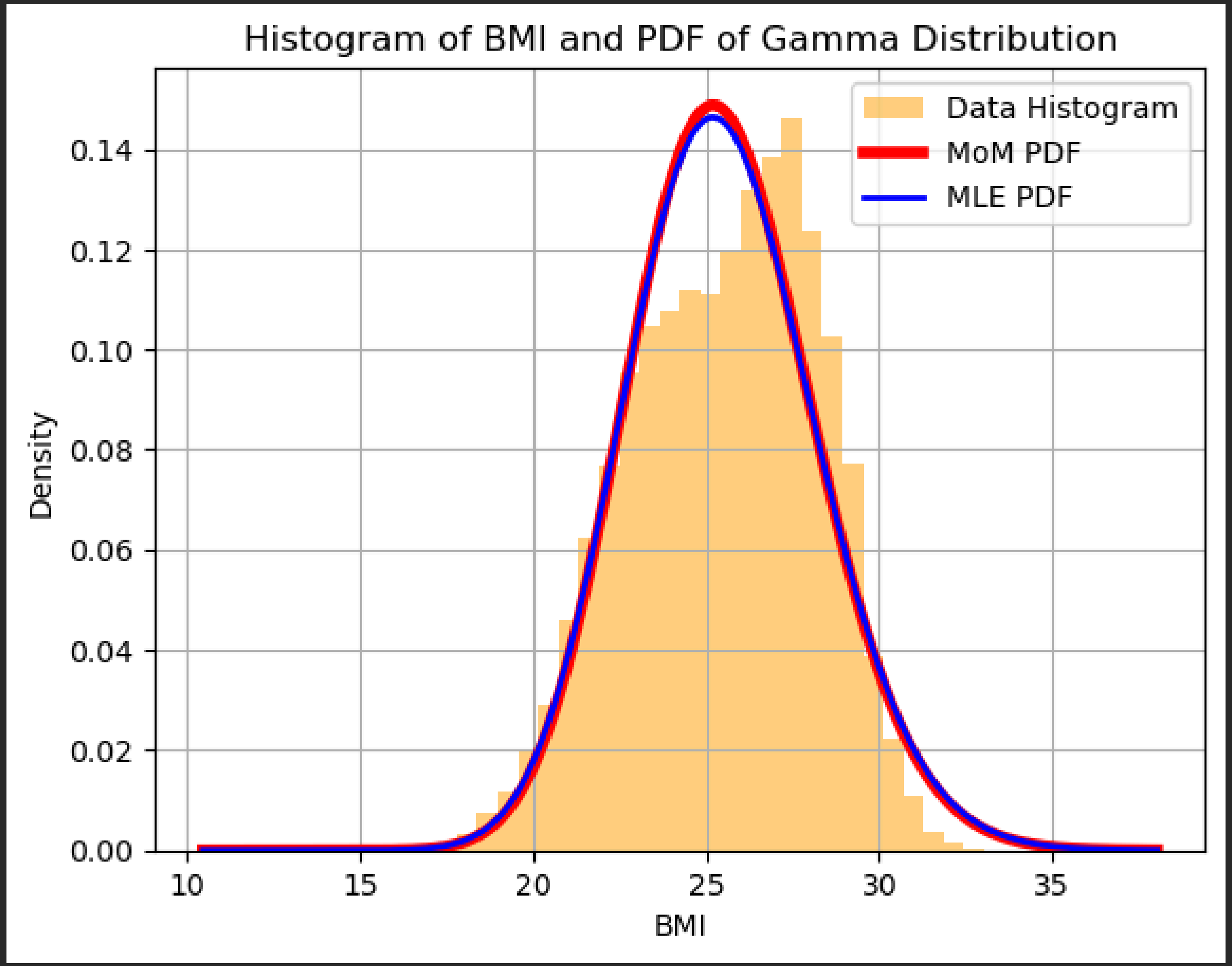
Define $g(a) = \log a - \psi(a) - \left[\log(\bar{x}) - \frac{1}{n}\sum \log x_i\right]$ and find the point of intersection with the a-axis this will be MLE estimator of a.

For this we are using Newton-Rhapson method in which we start from some initial ao = 1 and keep iterating (using a_(n+1) = a_n - g'(a_n)/g(a_n)) until we get close enough to the exact zero (root).

Results

$\hat{a}_{\text{MLE}}$: 86.51220 $\hat{b}_{\text{MLE}}$: 0.2944

$\hat{a}_{\text{MoM}}$: 89..4444 $\hat{b}_{\text{MoM}}$: 0.2848



If X1,...,Xn is a sample from a normal distribution having unknown parameters μ and σ^{2}, then we can construct a confidence interval for variance σ^{2} by using the fact that:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hence,

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

Hence when, a 100*(1−α) percent confidence interval for σ^{2} is:

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

95% Confidence Interval for Variance

Normal Population

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

Substituting the values we get our 95% confidence interval as:

[7.0595, 7.4620]

Two Normal Populations

& the 95% confidence interval in the difference of their means

Question 3

Standford Open Policing Data for the city of Nashville

Overview

Data regarding the routine traffic stops conducted by the Nashville Police department from 2010 to 2017

Date:
Date of the stop

Time:
Time of the stop

Location:
The location at which the traffic stop took place

Nashville Statistics:

Overall Population - 689,000
White Population - 385,840
African-American Population - 172,250

Subject Age:
Age of the person being stopped

Subject Race:
The Race of the person being stopped

Subject Sex:
The Sex of the person being stopped

Modifications & Final Data we'll be working with

Date	Number of Stops per 10,000 White Americans	Number of Stops per 10,000 African-Americans

Lets say, on a given day, there were 1000 white Americans stopped by the police, and 800 African-Americans stopped by the police. A surface level look at this data would lead one to believe that the police may be biased against white Americans. But we must not forget that the number of white people in the city is more than twice the number of black people in the city.

So to get a more proportionate understanding of the data wrt race, we decided to go with "number of white/black people stopped by the Nashville police per 10,000 white/black people.

Calculating the Sample Mean and Sample Variance

$X_1, X_2, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ (White population) $Y_1, Y_2, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma^2)$ (Black population)

Sample Mean:

$$\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i \quad \text{and} \quad \bar{Y} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j$$

Sample Variance:

$$S_X^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \quad \text{and} \quad S_Y^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2$$

We know that:

$$(n_1 - 1)\frac{S_1^2}{\sigma^2} \sim \chi^2_{n_1-1} \qquad (n_2 - 1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{n_2-1}$$

And by properties of chi-squared distribution:

$$(n_1 - 1)\frac{S_1^2}{\sigma^2} + (n_2 - 1)\frac{S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

Also,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

Pooled Estimator:

$$S_p^2 = \frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}$$

By the definition of a t-distribution, we know the ratio of two independent random variables, with the numerator being a standard normal variable, and denominator being a chi-squared random variable with n degrees of freedom, is a t-random variable with n degrees of freedom. And hence,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \div \sqrt{\frac{S_p^2}{\sigma^2}} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is a t-random variable with n1 + n2 degrees of freedom.

And hence,

$$P\left\{-t_{\alpha/2,n_1+n_2-2} \le \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le t_{\alpha/2,n_1+n_2-2}\right\} = 1-\alpha$$

And in turn, $(\mu_1-\mu_2)$ belongs to the corresponding interval:

$$\left(\overline{X}-\overline{Y}-t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\ \overline{X}-\overline{Y}+t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

For our dataset, the values of the variables come out to

$$\overline{X} = 13.21 \quad \text{and} \quad \overline{Y} = 20.35$$
$$\alpha = 0.05$$
$$n_1 = n_2 = 455$$
$$S_p^2 = 46.40$$

Hence,

$$\textbf{95\% Confidence Interval:} \quad \mu_1 - \mu_2 \in (-8.02, \; -6.25)$$

Inference

Since the entire interval lies in the negative (and by a very substantial amount), we can infer that:

$$\mu_1 - \mu_2 < 0 \quad \Rightarrow \quad \mu_1 < \mu_2$$

There is strong statistical evidence to suggest that, on average, Black individuals are stopped more frequently than White individuals — assuming the samples are random, the assumptions of normality and equal variance hold, and the data is representative.

Bernoulli Distribution

& Hypothesis testing on it

Question 4

Dataset

Cancer Data

Data of thousands of cells, which has several attributes of the cell (its radius, concavity, etc.) and an extra attribute which tells us whether the cell is cancerous or not (diagnosis).

Diagnosis:
Tells whether a cell is benignant or malign, the former of which we treat as 1, and the latter we treat as 0.

We are testing the hypothesis:

$$H_0 : p \leq 0.5 \quad \text{vs} \quad H_1 : p > 0.5$$

When n is large, the test statistic Z defined as,

$$Z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \qquad \hat{p} = \bar{X}$$

Approximately follows a standard normal distribution N(0, 1), where:

- p̂ is the MLE estimator for bernoulli parameter
- Po is the value under H_0
- n is the sample size

For our dataset:

n = 569

Number of malignant cases = 212

$\hat{p}$ = 212 / 569 ≈ 0.3725

Under H_o, p = p_{o} = 0.5

Substituting in the formula:

$$Z \approx \frac{0.3725 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{569}}} \approx -6.707$$

For a **significance level α = 0.05**, the critical value is **Z_a = 1.645**

Since **$Z < Z_a$**, we fail to reject the null hypothesis.

Conclusion: There is insufficient evidence to suggest that the proportion of malignant cases is greater than 0.5

Thank you