

Applied Statistics

Computational Project 2

Statistical Analysis of Weather Data for Major Indian Cities

Saran Konala - AI23BTECH11023

Ashwath E - AI23BTECH11001

D Sai Durga Rishi - AI23BTECH11004

Index

1. Part 1: Point Estimation
2. Part 2: Interval Estimation
3. Part 3: Interval Estimation on the difference between sample means
4. Part 4: Hypothesis Testing

Datasets used

The dataset provides comprehensive **daily meteorological data** across multiple Indian cities, spanning a period from **January 1, 1990, to July 20, 2022**. The data has been curated to support exploratory data analysis, statistical inference, and climate trend investigations over more than three decades.

[Dataset](#)

Units:

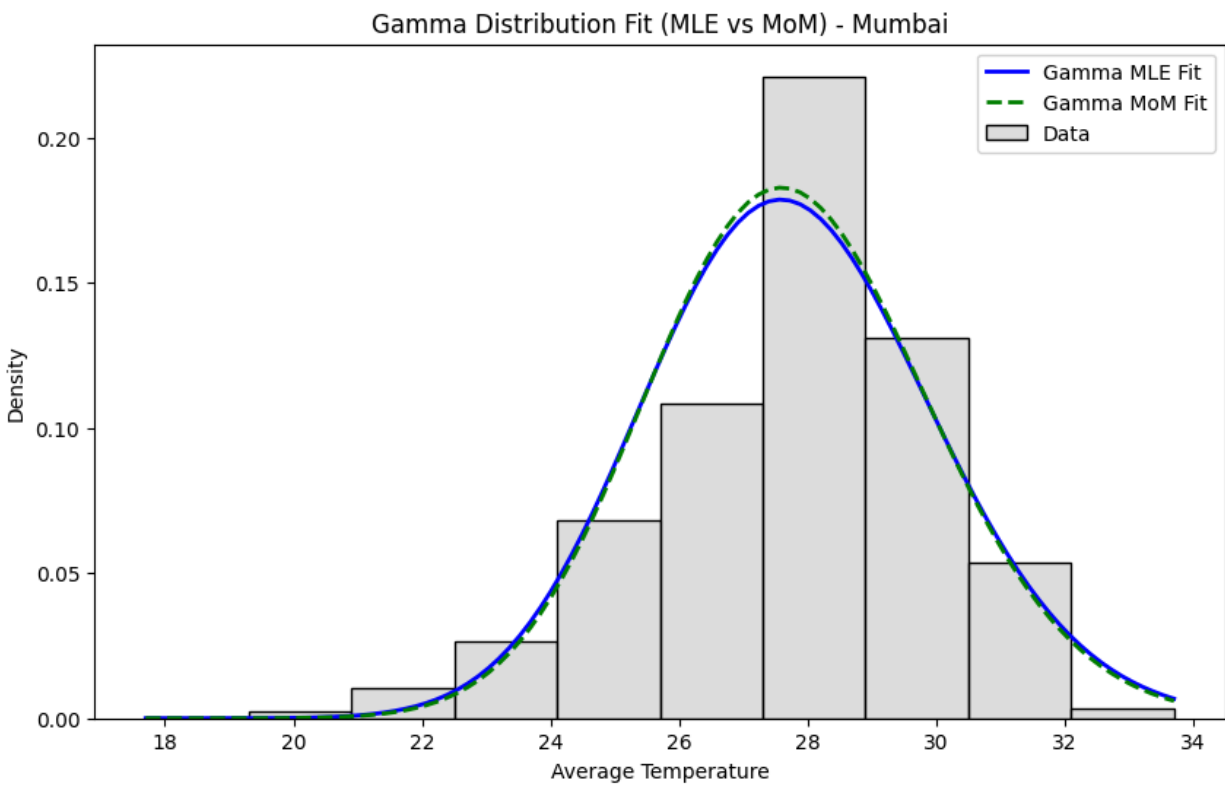
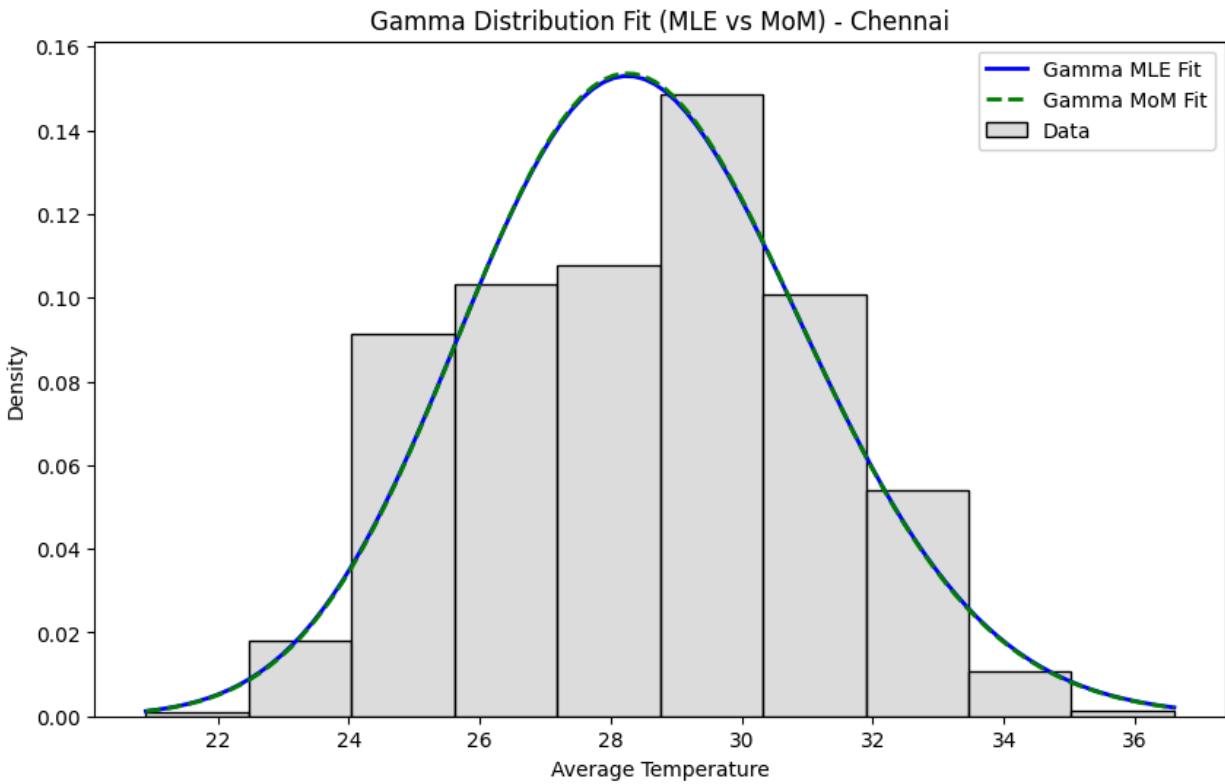
Temperature - °C

Precipitation - mm

Part 1: Point Estimation

For this section, we estimated the mean and variance of the average daily temperature of Chennai and Mumbai (independently) from 1990 to 2022. The method of moments was used to reach this estimation.

Then, we fitted a gamma distribution on the data and estimated its parameters (shape parameter and scale parameter), using which we calculated the mean and variance. The ML estimate was used to calculate the values of the parameters.



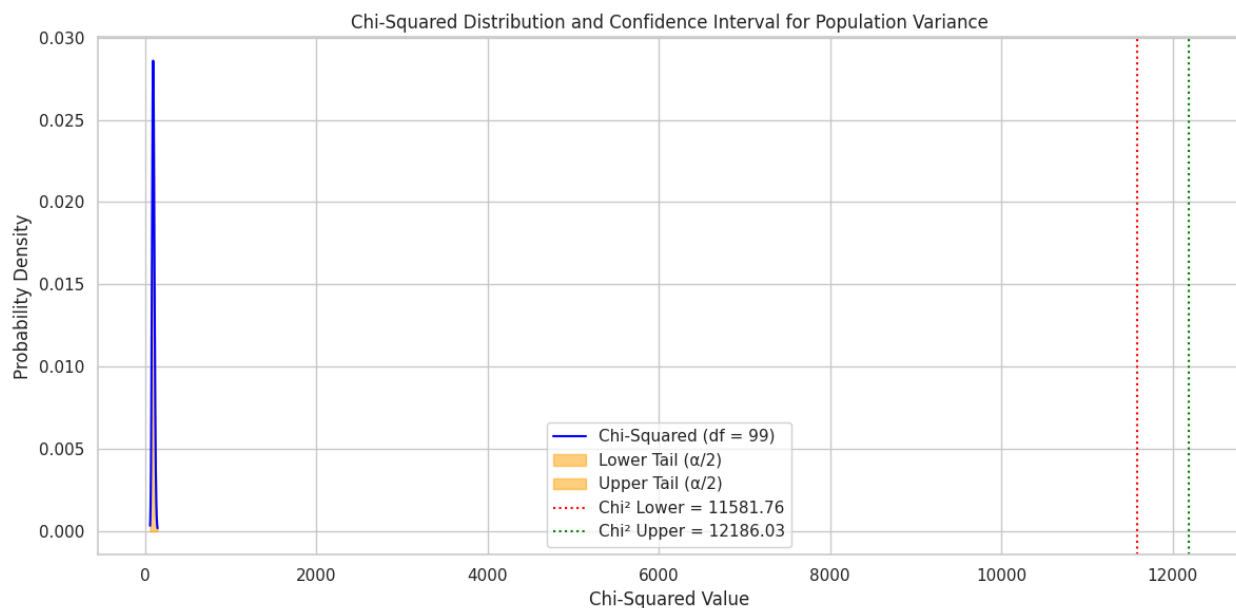
Part 2: Interval Estimation

In this section, we estimated a 95% confidence interval for the variance of the average daily temperature data from Chennai over the period 1990 to 2022. Assuming the data is drawn from a normally distributed population, we utilized the Chi-squared distribution to derive the confidence bounds. The interval was computed based on the sample variance and the number of observations, providing a range within which the true population variance is expected to lie with 95% confidence. This method accounts for sampling variability and offers a statistically rigorous way to quantify uncertainty around the spread of the temperature data.

The 95% confidence intervals for the variances are:

$$T_{Chennai} = [6.6208, 6.9665]$$

$$T_{Mumbai} = [4.6728, 4.9166]$$



The blue region of the graph shows the chi-square distribution, and the orange region showcases the rejection region of the distribution.

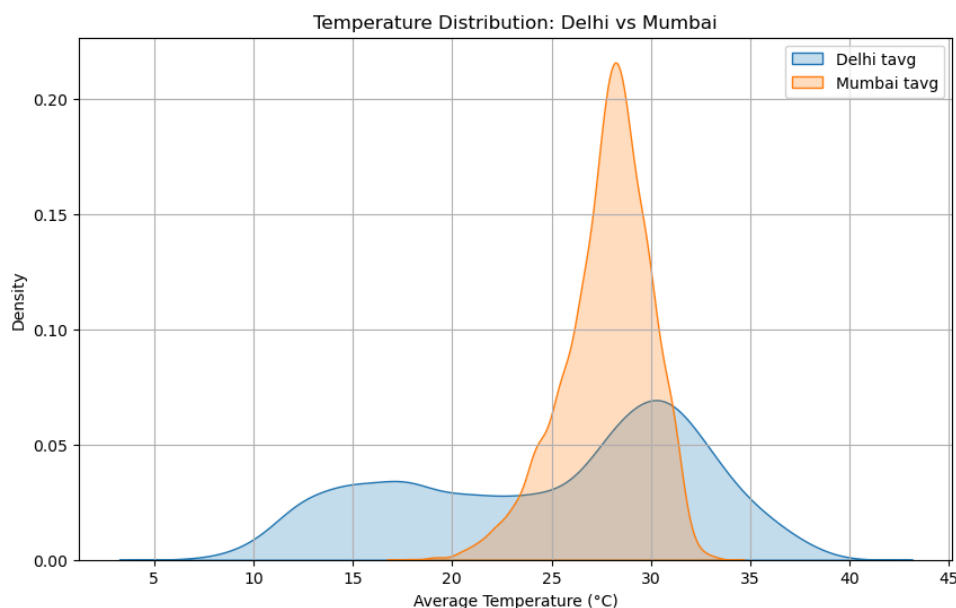
Part 3: Interval Estimation on the difference between sample means

For this section, we compared the temperature difference between the average temperatures of major Indian cities. Specifically, we aim to compute the **95% confidence interval** for the difference in the population means under the assumption of normally distributed data with **equal variances**.

$$\bar{X} - \bar{Y} \pm t_{\alpha/2, n+m-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \cdot \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}$$

Temperature difference between Delhi and Mumbai

First, we plotted the empirical PDFs for each city's temperature individually.



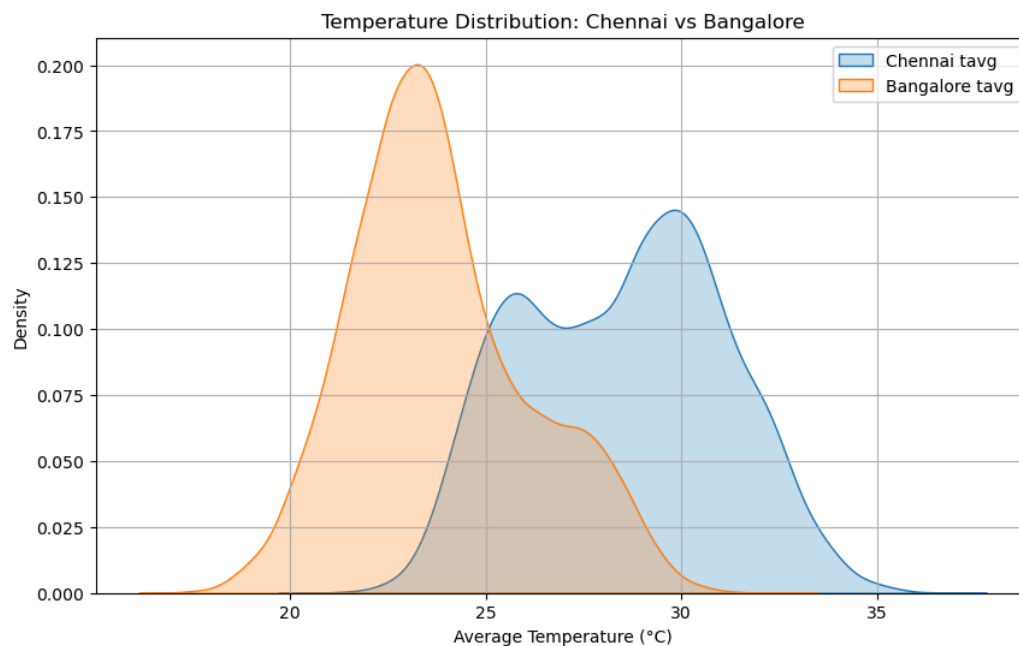
The 95% confidence interval for the difference between temperatures

$$T_{Delhi} - T_{Mumbai} = [-2.90, -2.63]$$

with the mean difference being -2.77.

Temperature difference between Chennai and Bangalore

First, we plotted the empirical PDFs for each city's temperature individually.



The 95% confidence interval for the difference between temperatures

$$T_{Delhi} - T_{Mumbai} = [-4.58, -4.71]$$

with the mean difference being 4.65.

Part 4: Hypothesis Testing

The purpose of this analysis is to test whether the probability of rainfall in Delhi is greater than 50%. This is assessed using a one-sided binomial hypothesis test based on sampled binary rainfall data, where:

- 1 represents a rainy day
- 0 represents a non-rainy day

Let p_0 be the probability of rain on a given day in Delhi. We aim to test:

- **Null Hypothesis (H_0):** $p \leq 0.5$
- **Alternative Hypothesis (H_1):** $p > 0.5$

This is a **right-tailed binomial test** where we check if the number of rainy days in a sample provides enough evidence to reject the null hypothesis.

Methodology

- A **random sample of 100 days** was drawn from the binary rain data for Delhi.
- The number of rainy days X in the sample was counted.
- The data is assumed to follow a **Binomial distribution** $B(n=100, p=0.5)$
- The **p-value** was calculated as:

$$p_value = P(X \geq \text{observed value}) = 1 - \text{Binomial CDF}(X - 1)$$

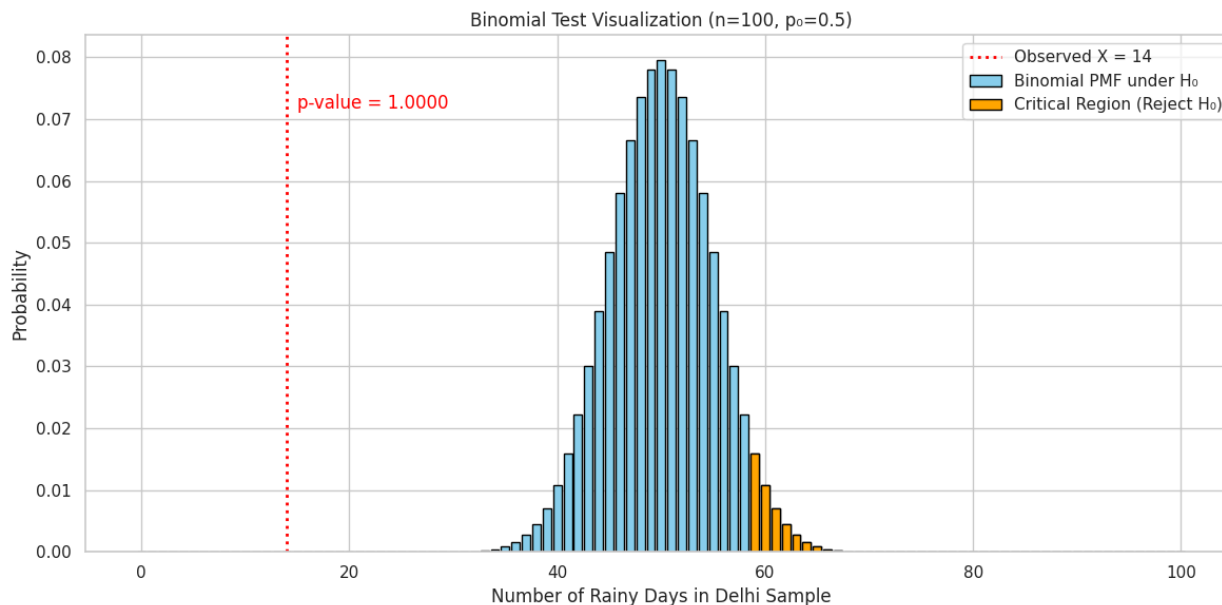
- The **critical value k^*** was determined such that:

$$P(X \geq k^*) \leq \alpha = 0.05$$

Visualization


The visualization consists of the probability mass function (PMF) of the binomial distribution $B(n=100, p=0.5)$:

- The **blue bars** represent the likelihood of each number of rainy days under the null hypothesis.



- The **orange bars** ($X \geq k^*$) highlight the **critical region**, which corresponds to outcomes considered statistically unlikely under H_0 . If the observed value falls in this region, we reject H_0 .
- A **red dashed line** marks the **observed number of rainy days (X)** in the Delhi sample.
- An annotation displays the **computed p-value**, which represents the probability of observing at least as many rainy days as X under the null hypothesis.

If the observed value X falls within or beyond the critical region (orange area), it provides **significant evidence** to reject the null hypothesis in favor of the alternative $H_a : p > 0.5$. The visualization makes it clear whether the observed result lies in a high-probability region under H_0 , or in the tail, supporting rejection.



This graphical interpretation enhances statistical understanding by showing both the expected distribution under H_0 and how extreme the observed value is in that context.