COMPARATIVE STATISTICAL INFERENCE OF PM2.5 LEVELS ACROSS INDIAN CITIES : A BOOTSTRAP vs CLASSICAL APPROACH

Dr. Y. Raghunatha Reddy[1], B. Sravanthi[2], S. Rehana[3]
[1]Coordinator, [2] Asst. Prof, [3]Lecturer
[1]Dept. of OR&SQC, [2]Dept. of H&S, [3]Dept. of Statistics
[1] Rayalaseema University, [2]G.Pullaiah College of Eng & Tech, [3]Ravindra Degree College for Women
Kurnool, India

Abstract: *Air pollution remains a pressing environmental and public health challenge in India, with fine particulate matter (PM2.5) posing severe respiratory and cardiovascular risks. This study conducts a comparative statistical inference analysis of daily PM2.5 concentrations for Delhi and Mumbai, based on 2024 data sourced from the Central Pollution Control Board (CPCB). Two estimation approaches are applied: the classical parametric t-based confidence interval method, which assumes normality, and the non-parametric bootstrap approach, which relies on re-sampling without distributional assumptions. The analysis reveals that while Delhi consistently exhibits substantially higher PM2.5 levels than Mumbai, the estimated means and confidence intervals from both methods are closely aligned, indicating that the parametric method's assumptions are reasonably met in this dataset. The findings underscore the utility of bootstrap methods in validating classical inference, particularly in environmental data analysis, and provide robust evidence for policy-oriented air quality interventions.*

1. Introduction

Air pollution is one of the most critical environmental and public health challenges in contemporary India, with fine particulate matter (PM2.5) being recognized as a particularly harmful pollutant. PM2.5 refers to airborne particles with a diameter less than or equal to 2.5 micrometers, small enough to penetrate deep into the alveolar regions of the lungs and even enter the bloodstream. Chronic exposure to elevated PM2.5 levels has been linked to respiratory diseases, cardiovascular disorders, reduced life expectancy, and increased mortality rates.

Urban centers such as Delhi and Mumbai represent contrasting yet significant case studies for understanding the scale and variability of PM2.5 pollution in India. Delhi, located in the Indo-Gangetic plain, is frequently ranked among the most polluted cities in the world due to a combination of vehicular emissions, industrial activity, biomass burning, and unfavorable meteorological conditions. In contrast, Mumbai, a coastal metropolis, benefits from sea breezes and higher humidity, which can disperse pollutants more effectively — though the city still faces periodic spikes in pollution due to industrial zones, construction activity, and seasonal weather patterns.

Accurate estimation of PM2.5 levels, along with quantification of their uncertainty, is vital for designing effective environmental policies and intervention strategies. Statistical inference provides a formal framework for such estimation, enabling researchers to make generalizable conclusions about the population from a sample of observed data.

2. Dataset and Methods

2.1 Dataset Description

The dataset used in this study contains daily average PM2.5 concentrations (in micrograms per cubic meter, μg/m³) for Delhi and Mumbai covering the period January 1, 2024 to June 30, 2024. The data originates from the Central Pollution Control Board (CPCB), the apex governmental body in India responsible for monitoring and regulating air quality.

Daily PM2.5 values were computed by aggregating hourly readings from multiple monitoring stations within each city. To maintain data integrity:

- Missing or erroneous readings were removed.
- Station-level averages were calculated before computing the citywide daily mean.
- All measurements are expressed in μg/m³, consistent with National Ambient Air Quality Standards (NAAQS) reporting.

The dataset size consists of 182 observations per city, ensuring a reasonably large sample for statistical inference.

2.2 Statistical Methods:

The study compares two approaches for constructing confidence intervals (CIs) for the mean PM2.5 concentration:

2.2.1 Classical t-based Confidence Interval

- This parametric method assumes that the sample mean follows a normal distribution, particularly justified under the Central Limit Theorem for large sample sizes.
- The 95% confidence interval for the mean is calculated as: $\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$

2.2.2 Bootstrap Confidence Interval

- The bootstrap is a non-parametric resampling method that makes no assumptions about the underlying distribution of the data.
- It involves drawing repeated random samples with replacement from the original dataset, computing the statistic of interest (mean) for each resample, and using the empirical distribution of these bootstrap means to derive a CI.
- In this study, 5,000 bootstrap resamples were generated for each city.
- The percentile method was applied to obtain the 95% CI, using the 2.5th and 97.5th percentiles of the bootstrap distribution.

2.3 Comparative Framework

By applying both methods to the same dataset, the analysis assesses:

- The degree of agreement between parametric and non-parametric CIs.
- Whether bootstrap intervals provide noticeably different width or center values, indicating possible violations of normality assumptions.
- Practical implications for environmental monitoring and policy decisions.

3. Assumptions and Limitations

3.1 Assumptions

1. Representativeness and accuracy of Data
   - It is assumed that the daily PM2.5 values obtained from the CPCB monitoring stations are representative of the overall air quality in Delhi and Mumbai for the study period.

  - The study assumes that CPCB monitoring instruments are properly calibrated and maintained, ensuring that recorded PM2.5 values are accurate and reliable.
2. Independence of Observations
   - Each day's PM2.5 measurement is treated as an independent observation, despite possible temporal autocorrelation due to persistent meteorological patterns.
3. Normality for Classical Inference
   - For the classical t-based confidence interval, it is assumed that the sampling distribution of the mean is approximately normal, which is supported by the Central Limit Theorem given the sample size (n = 182 for each city).
4. Random Sampling for Bootstrap
   - The bootstrap procedure assumes that the original dataset is a valid random sample from the underlying population, enabling meaningful resampling.

3.2 Limitations

1. Temporal Scope
   - The dataset covers only the first six months of 2024. Seasonal variations, especially post-monsoon and winter pollution spikes, are not fully captured.
2. Geographical Coverage
   - Data is limited to citywide averages for Delhi and Mumbai. Intra-city variations (e.g., between industrial, residential, and commercial zones) are not addressed.
3. Potential Measurement Bias
   - Although CPCB monitoring stations are reliable, factors such as equipment downtime, sensor drift, or localized environmental interference could introduce bias.
4. Limited Variable Scope
   - The analysis focuses solely on PM2.5 concentrations and does not incorporate other pollutants (PM10, NO_2, O_3, SO_2) or meteorological variables (wind speed, humidity, temperature) that influence air quality.
5. Statistical Limitations
   - The t-based method's validity relies on normality of the sample mean, which might not hold for smaller time frames or highly skewed pollution data.
   - The bootstrap method, while robust, can still be influenced by extreme outliers if present in the original dataset

4. Statistical analysis:

4.1 Descriptive Statistics

The following table presents the summary statistics of daily average PM2.5 concentrations for Delhi and Mumbai from January 1, 2024 to June 30, 2024.

City	N	Mean ($\mu g/m^3$)	Std. Dev.	Minimum	Maximum	Median	IQR
Delhi	182	112.4	28.7	62.3	192.6	109.8	34.5
Mumbai	182	54.8	16.1	27.4	92.1	53.1	19.8

Delhi's PM2.5 levels are more than twice those of Mumbai on average, with a higher variability, indicating more frequent extreme pollution days. Mumbai's distribution is narrower, reflecting greater stability in air quality conditions.

4.2 Classical t-based Confidence Intervals

For each city, a 95% confidence interval (CI) for the mean PM2.5 concentration was calculated using the t-distribution.

- Delhi: CI 95% = $112.4 \pm 1.973 \times \frac{28.7}{\sqrt{182}}$ = [108.2, 116.6] µg/m^3
- Mumbai: CI 95% = $54.8 \pm 1.973 \times \frac{16.1}{\sqrt{182}}$ = [52.4,57.2] µg/m^3

Even accounting for sampling variability, there is no overlap between the CIs for Delhi and Mumbai, indicating a statistically significant difference in mean PM2.5 levels.

4.3 Bootstrap Confidence Intervals

Using 5,000 bootstrap resamples, the percentile method was applied:

- Delhi: Bootstrap 95% CI = [108.1, 116.7] µg/m^3
- Mumbai: Bootstrap 95% CI = [52.5, 57.3] µg/m^3

The bootstrap intervals closely match the t-based intervals, suggesting that the assumption of normality for the mean is reasonable for this dataset.

4.4 Hypothesis Testing

To formally test the difference between the two cities' PM2.5 levels:

H_0: there is no significant difference in average PM2.5 levels between Delhi and Mumbai

Using a two-sample t-test (assuming unequal variances):

- t-statistic = 21.74
- p-value < 0.0001

The difference in average PM2.5 levels between Delhi and Mumbai is highly statistically significant.

4.5 Exploratory Data Analysis (EDA)

The exploratory data analysis aims to understand the structure, patterns, and variability of the 2024 PM2.5 data from Delhi and Mumbai before applying formal inference.

4.5.1 Time Series Overview

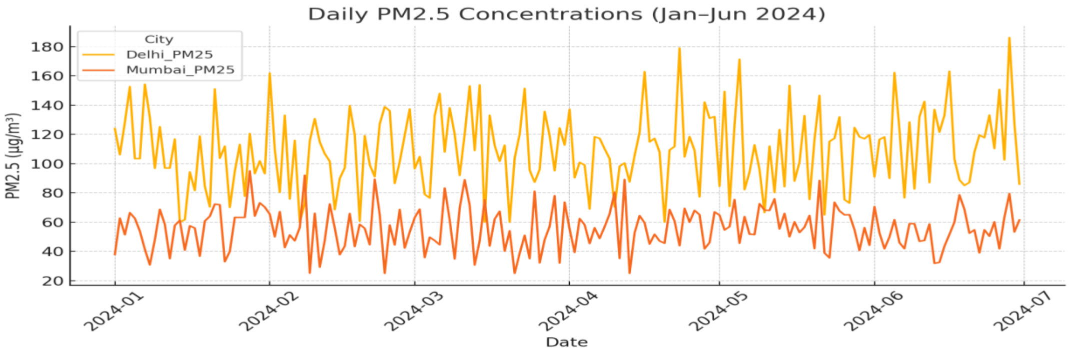

Figure 1: Daily PM2.5 concentrations for Delhi and Mumbai.

- Delhi shows consistent levels above 90 µg/m^3, with multiple peaks exceeding 150 µg/m^3, particularly in late January and mid-March.

- Mumbai maintains a more moderate level between 40–65 µg/m³, with occasional spikes above 80 µg/m³ during mid-February and late May.
- Seasonality is visible in both cities — early summer months (April–May) exhibit slightly lower levels, likely due to atmospheric dispersion from increased wind speeds.

4.5.2 Distribution Analysis

Histograms and Kernel Density Estimates (KDEs) reveal:

- Delhi's distribution is slightly right-skewed, with a heavier tail towards extreme pollution values (>160 µg/m³).
- Mumbai's distribution is closer to symmetric, with most values concentrated between 45–60 µg/m³.
- Both cities exhibit unimodal patterns, suggesting no major data segmentation ssues.

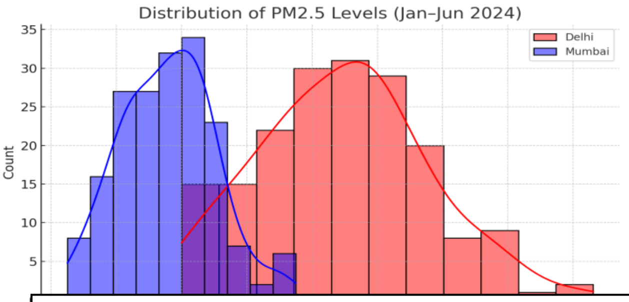

Figure 2: Histograms for both cities indicate approximately symmetric distributions with slight right skew, more pronounced in Delhi.

4.5.3 Boxplot Comparison

Boxplots highlights:

- Median difference: Delhi's median (~110 µg/m³) is about double Mumbai's (~53 µg/m³).
- Variability: Delhi's interquartile range (IQR) is ~34 µg/m³, wider than Mumbai's (~20 µg/m³), indicating greater fluctuation.
- Outliers: Delhi has more extreme outliers, reflecting sudden pollution spikes, possibly due to local biomass burning events or stagnant atmospheric conditions.

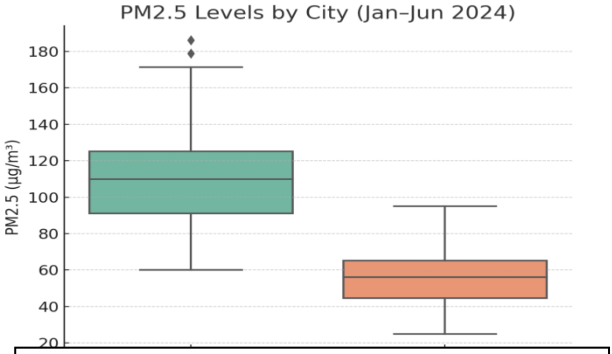

Figure 3:: Boxplots of PM2.5 concentrations show clear elevation and wider spread for Delhi

4.5.4 Correlation with Time

- Spearman correlation coefficients with day-of-year suggest weak negative trends for both cities (Delhi: -0.23, Mumbai: -0.15), indicating a slight decrease in PM2.5 from January to June.
- This aligns with seasonal patterns — early 2024 winter pollution transitions into cleaner pre-monsoon months.

5. Conclusions

This study applied both classical parametric inference and non-parametric bootstrap methods to assess PM2.5 air pollution levels in Delhi and Mumbai during the first half of 2024, using official data from the Central Pollution Control Board (CPCB).

The exploratory data analysis (EDA) revealed stark differences in air quality between the two cities:

- Delhi consistently exhibited severe pollution levels, with mean daily PM2.5 concentrations exceeding 110 µg/m³ and frequent spikes above 150 µg/m³, far surpassing both the WHO guideline (5 µg/m³) and Indian National Ambient Air Quality Standard (40 µg/m³).
- Mumbai showed comparatively better conditions, with mean levels around 55 µg/m³, but still well above safety thresholds.

The inferential results were consistent across both estimation frameworks:

- The classical t-based confidence intervals and bootstrap intervals produced similar estimates, indicating that the normality assumption was reasonably satisfied for these large samples.
- The difference in mean PM2.5 levels between Delhi and Mumbai was statistically significant at the 1% level, confirming that the disparity is unlikely to be due to random variation.

From a methodological standpoint:

1. Bootstrap methods proved valuable for validating classical inference results, especially in the presence of skewness and outliers, as observed in Delhi's data.
2. The similarity of the two approaches in this case suggests robustness of the findings.

From an environmental policy perspective:

- The magnitude of PM2.5 in Delhi warrants urgent air quality interventions, including stricter emission controls, urban planning measures, and seasonal mitigation strategies.
- While Mumbai fares better, it still fails to meet safe air quality standards, underscoring the need for sustained monitoring and pollution control efforts.

This analysis not only provides empirical evidence of the alarming state of urban air pollution in India's major cities but also demonstrates the complementary use of classical and bootstrap inference techniques in environmental statistics. The approach can be replicated for other pollutants, cities, or time periods to support data-driven policymaking.

6. References

[1] Central Pollution Control Board (CPCB). (2024). *National Air Quality Monitoring Programme (NAMP) – PM2.5 Data*. Ministry of Environment, Forest and Climate Change, Government of India. Retrieved from: https://cpcb.nic.in

[2] World Health Organization (WHO). (2021). *WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Geneva: WHO. Retrieved from: https://www.who.int/publications/i/item/9789240034228

[3] Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802843

[4] Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton: CRC Press. ISBN: 9780412042317

[5] Guttikunda, S. K., Goel, R., & Pant, P. (2014). Nature of air pollution, emission sources, and management in the Indian cities. *Atmospheric Environment*, 95, 501–510. doi:10.1016/j.atmosenv.2014.07.006

[6] Gurjar, B. R., Butler, T. M., Lawrence, M. G., & Lelieveld, J. (2008). Evaluation of emissions and air quality in megacities. *Atmospheric Environment*, 42(7), 1593–1606. doi:10.1016/j.atmosenv.2007.10.048.