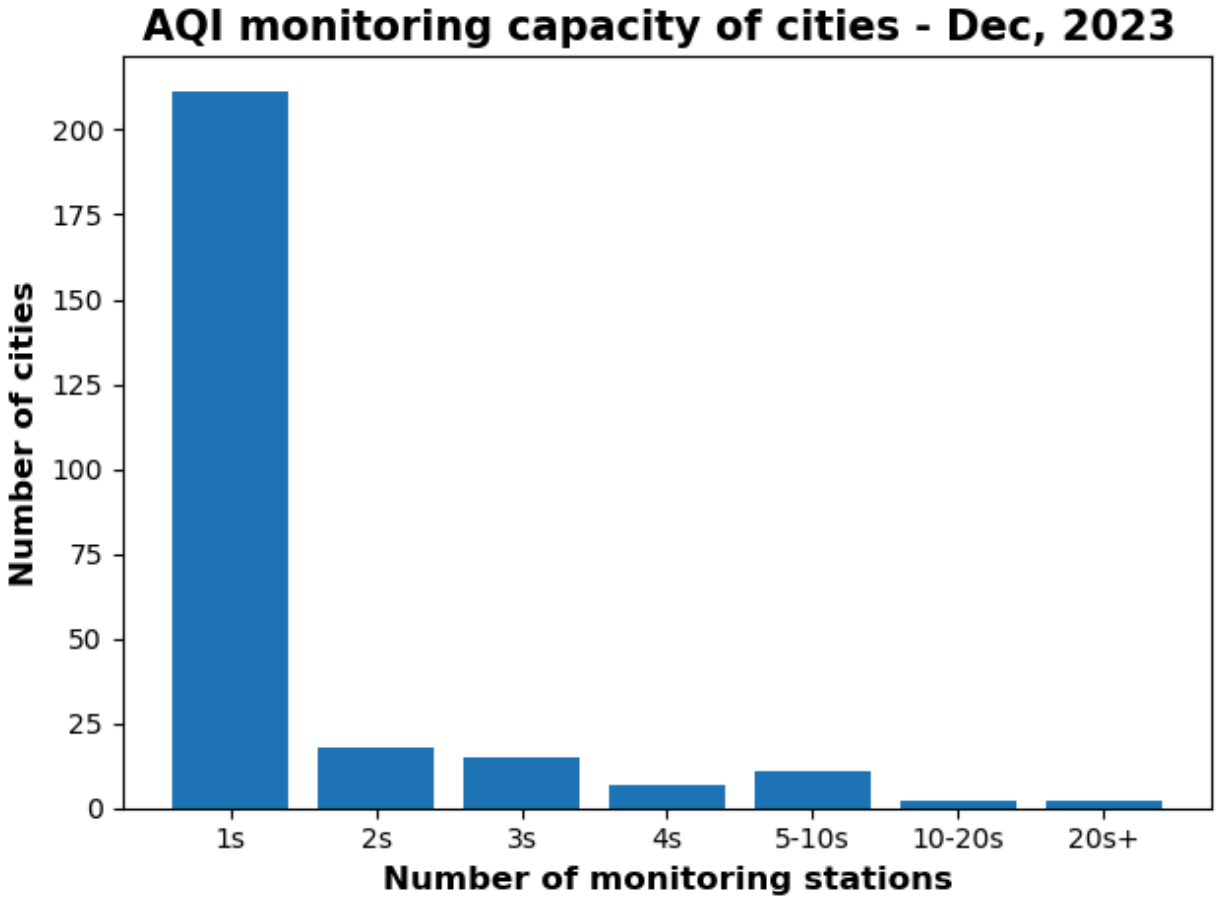
Issue with less number of air quality monitoring stations

Dammalapati Sai Krishna

As per the CPCB's daily AQI Bulletins data of December, 2023, 211 cities have only one air quality monitoring station. Only four cities – Delhi, Mumbai, Bengaluru and Hyderabad – have more than 10 monitoring stations.



However, often in the media, the AQI values from these few monitoring stations are presented as the air quality of the entire city. Such reportage could be untrue, statistically, for the following two reasons:

1. **Sampling Bias:** When there are very few monitoring stations in a city, they may not be a representative sample of the entire city. The placement of these monitoring stations may not be "random" (in a statistical sense). Hence, any inference made on the air quality of the entire city based on this unrandom sample could be biased.

    a. This can be assessed by checking the spatial representation of the monitors. By ascertaining the land-use characteristics of the location in which monitors exist in a city, we can comment on the spatial representation of the monitors.
2. **Wide Confidence Intervals:** Even if the assumption of "randomness" in the placement of air quality monitors is considered, there is another issue of wide confidence intervals.
    a. The true air quality value of a city from the monitors can only be determined by installing a monitor for every 4 sq.km.
    b. It is an exercise of statistical inference to estimate the air quality of a city with less number of monitors (a sample)
    c. There are two ways to perform this statistical inference: non-parametric and parametric. Non-parametric methods like bootstrap estimation are performed when we are unaware of the underlying population distribution. Parametric estimations are performed when we have knowledge of underlying population distributions from prior research. Prior research indicates that the pollution concentration data and AQI data is in a log-normal distribution.
        i. Statistical Analysis of Air Quality Indices: A Study | Nimesh | International Journal of Ecological Economics and Statistics™ (ceser.in)
        ii. The lognormal distribution, environmental data, and radiological monitoring | Environmental Monitoring and Assessment (springer.com)
        iii. A physical explanation of the lognormality of pollutant concentrations - PubMed (nih.gov)
        iv. 23_Air-quality-predictions-using-log-normal-distribution-functions-of-particulate-matter-in-Kuala-Lumpur.pdf (ukm.my)
    d. ~~Given the small sample size, a Student's t-distribution would be used for the purposes of statistical inference. Confidence Intervals of the mean air quality of the city built using the t-distribution would be wide for small sample sizes. For instance, if a city only has 2 monitors and the media reports the mean AQI value of these two monitors, the margin of error would be 12.71 times the standard error of the mean (for a 95% Confidence Interval). More monitoring stations would be needed to address this issue. The CPCB guidelines start with a minimum of 4 stations for any airshed. Even if this is achieved, the margin of error would reduce to 3.18 times the standard error of the mean.~~

Example:

Kolhapur

Kolhapur in Maharashtra has only two air quality monitoring stations. As this note is being written on April 01, 2024, the AQI values reported by these two stations at 16:00 (N=2) are: 185

and 227. The official AQI bulletin reported the average AQI as 206 and attributed “Poor” AQI category to Kolhapur accordingly.

A non-parametric bootstrap statistical inference on such a small sample would estimate that the true AQI mean value would lie between 185 and 227.

As mentioned above, parametric inference of Kolhapur’s true AQI value can be performed considering that AQI data would be in log-normal distribution. But given the sample size, a Student’s t-statistic would be used. This is because we consider that the sampling distribution of log-means would converge to log-normal distribution at higher sample sizes. But at smaller sample sizes, it would converge to log Student-t distribution. Inference with this assumption would give an extremely wide 95% confidence interval for the true AQI of Kolhapur – (55, 751). Given below is the table of statistical inference done for Kolhapur with various assumptions.

Statistical Inference	95% Percentile Confidence Interval of mean	AQI Categories
Non-Parametric Bootstrap	(185, 227)	Moderate-Poor
Parametric: log Student-t Distribution	(55, 751)	Satisfactory-Moderate-Poor-VeryPoor-Severe
Parametric: log Normal Distribution	(167, 250)	Moderate-Poor
Parametric: Normal Distribution	(164, 247)	Moderate-Poor
Parametric: Students’ t- Distribution	(-60, 472)	Good-Satisfactory-Moderate-Poor-VeryPoor-Severe

This wide confidence interval for Kolhapur does not help in definitely assigning the AQI category. It spans from “Satisfactory” to “Moderate” categories.

~~Standard Error of the mean (SEM) is 30 ($s/\sqrt{N}$). For N=2 (dof = 1), the margin of error is 12.71 times SEM for a 95% Confidence Interval, which equals 381! So the true AQI value of Kolhapur would be anywhere between 0 to 501.~~

Delhi

Delhi has 36 air quality monitoring stations (N=36) reporting AQI on April 01, 2024 at 4PM. The AQI values reported are: 105, 144, 148, 150, 118, 179, 120, 156, 147, 87, 133, 83, 158, 109, 288, 94, 104, 118, 195, 170, 97, 123, 116, 119, 120, 130, 139, 136, 120, 118, 108, 199, 112,

106, 111, 131. The official AQI bulletin reported the average AQI as 133 and attributed “Moderate” AQI category to Delhi accordingly.

A non-parametric bootstrap statistical inference on this sample would estimate that the true AQI mean value of Delhi would be in the (121, 146) interval with 95% confidence.

A parametric inference considering that AQI data would be in log-normal distribution would estimate that the true AQI value of Delhi would be in (118, 139) interval with 95% confidence. This is a narrower band compared to that of Kolhapur. It also helps in placing Delhi’s AQI category definitely in the “Moderate” category on April 01, 2024. Given below is the table of statistical inference done for Delhi with various assumptions.

Statistical Inference	95% Percentile Confidence Interval of mean	AQI Categories
Non-Parametric Bootstrap	(121, 146)	Moderate
Parametric: log Student-t Distribution	(118, 140)	Moderate
Parametric: log Normal Distribution	(118, 139)	Moderate
Parametric: Normal Distribution	(120, 145)	Moderate
Parametric: Students’ t- Distribution	(120, 146)	Moderate

~~Standard Error of the mean (SEM) is 6.6 (s/√N). For N=37 (dof = 36), the margin of error is ~2 times the SEM for a 95% Confidence Interval, which equals 13.22! So, the true AQI value of Delhi would be between 232 to 258. A narrower band.~~

Hyderabad

Hyderabad has 11 air quality monitoring stations (N=11) reporting to AQI on April 01, 2024 at 4PM. The AQI values reported are: 90, 78, 181, 79, 78, 76, 55, 82, 84, 58, 102. The official AQI bulletin reported the average AQI as 88 and attributed “Satisfactory” AQI category to Hyderabad accordingly.

Statistical Inference	95% Percentile Confidence Interval of mean	AQI Categories
Non-Parametric Bootstrap	(72, 108)	Satisfactory-Moderate
Parametric: log Student-t Distribution	(67, 102)	Satisfactory-Moderate

Parametric: log Normal Distribution	(69, 100)	Satisfactory
Parametric: Normal Distribution	(67, 107)	Satisfactory-Moderate
Parametric: Students' t- Distribution	(64, 110)	Satisfactory-Moderate

Jabalpur

Jabalpur has 4 air quality monitoring stations (N=4) reporting to AQI on April 01, 2024 at 4PM. The AQI values reported are: 98, 150, 133, 193. The official AQI bulletin reported the average AQI as 144 and attributed "Moderate" AQI category to Jabalpur accordingly.

Statistical Inference	95% Percentile Confidence Interval of mean	AQI Categories
Non-Parametric Bootstrap	(111, 178)	Moderate
Parametric: log Student-t Distribution	(89, 218)	Satisfactory-Moderate-Poor
Parametric: log Normal Distribution	(105, 183)	Moderate
Parametric: Normal Distribution	(104, 182)	Moderate
Parametric: Students' t- Distribution	(80, 206)	Satisfactory-Moderate-Poor

Comment

1. With more monitors, the confidence intervals are narrower, helping in definite attribution of AQI category for the city.
2. With more monitors, sensitivity to the type of statistical inference reduces.

Methods

Non-parametric bootstrap

The non-parametric bootstrap method is a resampling technique used to estimate the distribution of a statistic by repeatedly sampling with replacement from the observed data. This method is particularly useful for making statistical inferences when the underlying distribution is unknown.

Let X={x1,x2,…,xn} be the original sample consisting of n observations. Then we resample with replacement from this original sample a number of times, say 10000 times. Thus we obtain 10000 resampled samples and thus 10000 means or any other statistic of interest θ. The collection of these statistics (θ) is then used to infer the true statistic. 95% Confidence Interval of the statistic can be built by building the interval from 2.5 percentile to 97.5 percentile of the collection of these statistics.

Parametric inference using Normal Distribution and Students' t-Distribution

Parametric inference involves making statistical inferences about population parameters based on assumptions about the underlying distribution of the data. When the data is assumed to follow a normal distribution, parametric inference is performed by first estimating the parameters of the normal distribution (mean, standard deviation) using the sample data.

Let X={x1,x2,…,xn} be the original sample consisting of n observations. Then, the maximum likelihood estimates of the mean ($\bar{x}$) and standard deviation (s) are:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{1=n} (x_i - \overline{x})^2}$$

Once the estimates are calculated, then confidence interval of the mean can be calculated by

$$\overline{x} \pm z\frac{s}{\sqrt{n}}$$

where z is the confidence level value.

When the sample size is small, the sampling distribution of means doesn't converge to a normal distribution and thus a Students' t-distribution is used. The confidence interval of the mean can then be calculated by

$$\overline{x} \pm t\frac{s}{\sqrt{n}}$$

where t is the critical value of t-distribution at desired confidence level.

Parametric inference using log-Normal Distribution and log-Normal Students' t-Distribution

The log-normal distribution also has the same parameters like a normal distribution – mean, standard deviation. However, these are calculated after log transformation of the original sample data.

Let X={x1,x2,…,xn} be the original sample consisting of n observations. Then this sample data is transformed by applying natural logarithm. Y = {ln(x1), ln(x2), …, ln(xn)}. Then, the maximum likelihood estimates of the mean ($\overline{y}$) and standard deviation (s) are:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{i=n} ln(x_i)$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{1=n} (ln(x_i) - \overline{y})^2}$$

Once the estimates are calculated, then confidence interval of the log-mean can be calculated by

$$\overline{y} \pm z\frac{s}{\sqrt{n}}$$

where z is the confidence level value.

The confidence interval of the mean can be then calculated by applying exponential transformation to the lower and upper bounds.

$$(e^{\overline{y} - z\frac{s}{\sqrt{n}}}, e^{\overline{y} + z\frac{s}{\sqrt{n}}})$$

When the sample size is small, the sampling distribution of log-means doesn't converge to a normal distribution and thus a Students' t-distribution is used. The confidence interval of the log-mean can then be calculated by

$$\overline{y} \pm t\frac{s}{\sqrt{n}}$$

where t is the critical value of t-distribution at desired confidence level.

The confidence interval of the mean can be then calculated by applying exponential transformation to the lower and upper bounds.

$$\left(e^{\overline{y} - t\frac{s}{\sqrt{n}}}, e^{\overline{y} + t\frac{s}{\sqrt{n}}}\right)$$