Air-Quality Analysis: January in Beijing

Isabella Chittumuri
Professor Samantha Benedict
STAT 295: Intro to Complex Sampling
City University of New York, Hunter College

Abstract

Beijing, China is known to experience one of the worst air pollution worldwide. The World Health Organization (WHO) offers global guidance on limits for four major air pollutants: ozone (O_3), particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), and sulfur dioxide (SO_2). In this project we studied the statistical analysis of $PM_{2.5}$ concentration levels in Beijing through January 2017, to determine the spectrum of how unhealthy the air quality can be for that month.

1. Introduction

1.1 What is $PM_{2.5}$ and it's health risks?

$PM_{2.5}$ is particulate matter in the air that has a diameter of less than 2.5 micrometers. Because of its small size, $PM_{2.5}$ is able to penetrate the respiratory tract, deep into the lungs and sometimes enter the circulatory system. These particles are formed as a result of burning fuel, atmospheric chemical reactions, and even forest fires. Since potential health damage caused by air pollutants depends on both concentration and duration of exposure, $PM_{2.5}$ is measured by a 24-hour index. Long-term exposure to this particulate matter can lead to cardiovascular and respiratory diseases.

1.2 Data

This data was collected by the Beijing Municipal Environmental Monitoring Center from 12 nationally-controlled air-quality monitoring sites. It was then matched with the nearest weather station from the China Meteorological Administration, which established accuracy and validity. The sampling scheme used was systematic because it was collected at every hour of every day from March 1st, 2013 to February 28th, 2017, within all 12 monitoring sites. However, there were numerous missing values from all of the monitoring sites which made the data incomplete with unequal probability.

The data was available in csv format and it was consistent in the measurement of 18 variables and 420,768 observations. From these variables, we were most interested in the hourly $PM_{2.5}$ concentration levels (in $\mu g/m^3$) of January 2017 from the Wanshouxigong site. The data collected for this specific month had a total of 744 observations, with one missing value for day 25, hour 14. All of the analysis was done using R 3.6.3.

Since $PM_{2.5}$ is measured by a 24-hour index, we consolidated these 744 hourly observations into 31 daily observations. To do this we used the aggregate function to take the mean of every 24 observations. This created our target population of 24-hour $PM_{2.5}$ concentration levels (in μg/m^3) in January 2017. However, due to the one missing value in day 25, that day was considered missing.

2. Exploratory Analysis

In this report, we will explore how Beijing's January $PM_{2.5}$ concentration levels compare to that of the 2005 WHO Air quality guidelines.

2.1 Target Population

To do this we chose to observe and analyze Beijing's 24-hour $PM_{2.5}$ concentration levels through January of 2017. It should be noted that we removed day 25 during calculations to make the analysis simpler. In reference to the mathematical representation below, the mean was 129.98 μg/m^3, representing the average of 24-hour $PM_{2.5}$ concentration levels within the target population. The median was 74.21 μg/m^3, representing the middle value of the data once it was set in numeric order. There was no distinct mode, although there was high frequency of numbers between 20-40 μg/m^3. The considerable difference between the mean, median, and mode suggested a right-skewed distribution. In this case, the median was a better measure of central tendency, than the mean.

Target Population Mean

$$\mu_y = \frac{1}{30}(445.54 + \cdots + 101)$$

The 24-hour $PM_{2.5}$ concentration levels ranged from 13.08 to 445.54 μg/m^3, which gave a difference of 432.46 μg/m^3. Variance and standard deviation measure how dispersed the data values are around the mean. The target population had an unadjusted variance of 17016.67 (μg/m^3)^2, with an unadjusted standard deviation of 130.45 μg/m^3. It had an adjusted variance of 17583.9 (μg/m^3)^2, with an adjusted standard deviation of 132.60 μg/m^3.

Unadjusted Target Population Variance

$$\sigma_Y^2 = \frac{1}{30}[(445.54 - 129.98)^2 + \cdots + (101 - 129.98)^2$$

Unadjusted Target Population Standard Deviation

$$\sigma_Y = \sqrt{\frac{1}{30}[(445.54 - 129.98)^2 + \cdots + (101 - 129.98)^2]}$$

Adjusted Target Population Variance

$$S_Y^2 = \frac{1}{(30-1)}[(445.54 - 129.98)^2 + \cdots + (101 - 129.98)^2]$$

Adjusted Target Population Standard Deviation

$$S_Y = \sqrt{\frac{1}{(30-1)}[(445.54 - 129.98)^2 + \cdots + (101 - 129.98)^2]}$$

Figure 1 depicts a right-skewed unimodal frequency distribution. This distribution indicated the possibility of outliers within the target population. To determine which numbers were outliers, we found the upper fence, which was 400 μg/m^3. This meant that any numbers greater than 400 μg/m^3 were outliers.

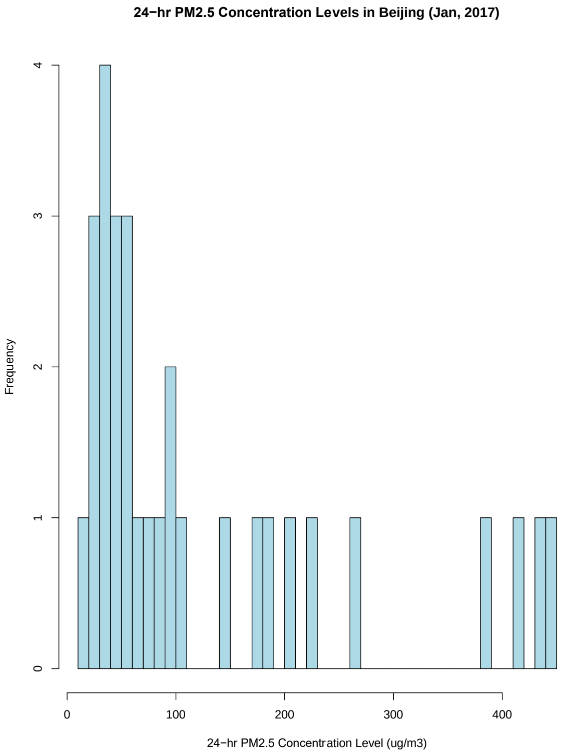

Figure 1 Frequency Plot of 24-hour $PM_{2.5}$ Concentration Levels in Beijing (Jan, 2017)

2.2 Reduced Target Population

In composing a reduced target population of size eight, we used numbers from the five-number summary, removed outliers, and chose three additional numbers from the target population. The three additional numbers were based on the frequency distribution in **Figure 1** to gain an accurate representation of the target population. The 24-hour $PM_{2.5}$ concentration levels (in μg/m^3) for our reduced target population were: 13.08, 36.67, 38.22, 50.5, 74.21, 178.25, 185.31, and 268.92. **Figure 2** represents the frequency distribution of these eight numbers.

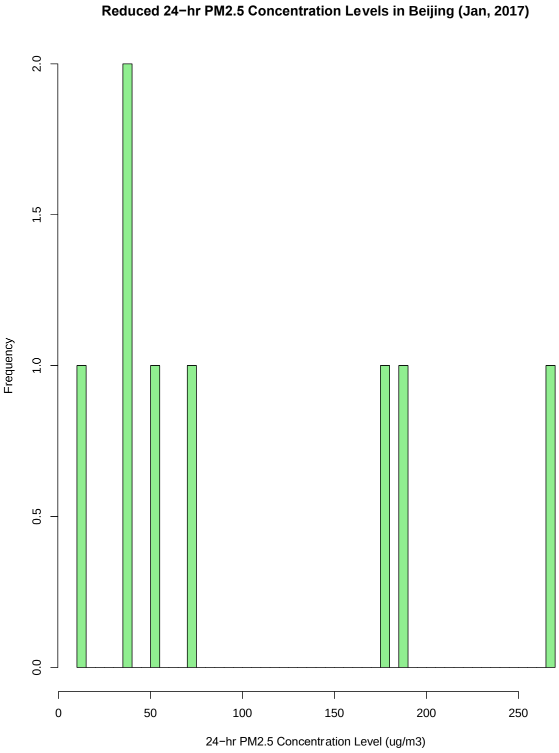

Figure 2 Frequency of Reduced 24-hour $PM_{2.5}$ Concentration Levels in Beijing (Jan, 2017)

The reduced target population mean was 105.65 μg/m^3. The unadjusted variance was 7522.23 (μg/m^3)^2, and the unadjusted standard deviation was 86.73 μg/m^3. The adjusted variance was 8596.83 (μg/m^3)^2, and the adjusted standard deviation was 92.72 μg/m^3.

Reduced Target Population Mean

$$\mu_Z = \frac{1}{8}(13.08 + \cdots + 268.92)$$

Unadjusted Reduced Target Population Variance

$$\sigma_Z^2 = \frac{1}{8}[(13.08 - 105.645)^2 + \cdots + (268.92 - 105.65)^2]$$

Unadjusted Reduced Target Population Standard Deviation

$$\sigma_Z = \sqrt{\frac{1}{8}[(13.08 - 105.65)^2 + \cdots + (268.92 - 105.65)^2]}$$

Adjusted Reduced Target Population Variance

$$S_Z^2 = \frac{1}{(8-1)}[(13.08 - 105.65)^2 + \cdots + (268.92 - 105.65)^2]$$

Adjusted Reduced Target Population Standard Deviation

$$S_Z = \sqrt{\frac{1}{(8-1)}[13.08 - 105.65)^2 + \cdots + (268.92 - 105.65)^2]}$$

In this case, μ_Z was lower than μ_Y, with a difference of 24.34 μg/m^3. It was also much closer in value to the target population median than μ_Y was. Both σ_Z^2 and S_Z^2 were substantially lower than σ_Y^2 and S_Y^2, with respective differences of 9494.44 (μg/m^3)^2 and 8987.07 (μg/m^3)^2 . These disparities verified that the outliers highly impacted our target population statistics.

2.3 Sampling Distribution of the Sample Means

To construct a sampling distribution of the sample means from the reduced target population, we created samples of size three with equal probability and without replacement. This resulted in 56 samples; we then found the mean for each sample. **Table 1** displays the first three and last three samples, along with their mean. **Figure 3** portrays the frequency of the sampling distribution of the sample means, depicting a normal distribution.

Table 1. First 3 and last 3 samples and their mean

Sample #	X_1	X_2	X_3	ȳ
1	13.08	36.67	38.22	29.32
2	13.08	36.67	50.50	33.42
3	13.08	36.67	74.21	41.32
54	74.21	178.25	268.92	173.79
55	74.21	185.31	268.92	176.15
56	178.25	185.31	268.92	210.83

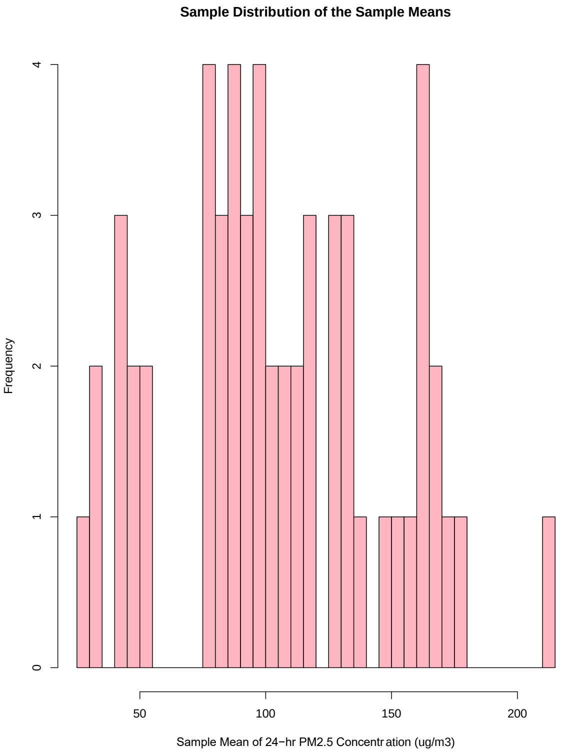

Figure 3 Frequency of Sampling Distribution of the Sample Means

The expectation of $\bar{y}$ was 105.65 µg/m^3, which represents the average value expected over repeated trials of our sampling procedure. This value was equal to the true mean of our reduced target population, μ_Z. As a result, $\bar{y}$ was an unbiased estimator of μ_Z.

When we factored in the adjusted reduced target population variance, the expectation of the squared deviations was 1791.01 (µg/m^3)^2, which represents the sampling distribution variance.

Mean of the Sampling Distribution of the Sample Means

$$E[\,\bar{y}] = \frac{1}{56}(29.33 + \cdots + 210.83)$$

Adjusted Variance of the Sampling Distribution of the Sample Means

$$Var[\,\bar{y}] = \left(1 - \frac{n}{N}\right)\left(\frac{S_Z^2}{n}\right) = \left(1 - \frac{3}{8}\right)\left(\frac{8596.83}{3}\right)$$

We were able to claim accuracy in our sampling distribution because we found that $\bar{y}$ was an unbiased estimator of μ_Z. And since the expectation of s^2 is generally equal to S^2, we considered s^2 as an unbiased estimator of S_Z^2.

2.4 Three Samples from the Sampling Distribution

After establishing accuracy in our sampling distribution, we were confident that a single sample mean can be used to estimate the true mean. However, we needed to verify that a single sample can be used to estimate the variance of the distribution of the sample means. To do this we randomly generated three samples from our sampling distribution and found their adjusted variances, displayed in **Table 2**. We then factored in each s^2 to calculate three different variances of the distribution of the sample means, shown in the mathematical representation below.

Table 2. Three samples and their adjusted variance

Random Sample#	V_1	V_2	V_3	s^2
1	13.08	185.31	268.92	17017.99
2	36.67	74.21	178.25	5379.75
3	13.08	38.22	185.31	8655.11

Variance of the Distribution of the Sample Means from Three Single Samples

$$var_1[\,\bar{y}] = \left(1 - \frac{n}{N}\right)\left(\frac{s^2}{n}\right) = \left(1 - \frac{3}{8}\right)\left(\frac{17017.99}{3}\right) = 3545.41$$

$$var_2[\,\bar{y}] = \left(1 - \frac{n}{N}\right)\left(\frac{s^2}{n}\right) = \left(1 - \frac{3}{8}\right)\left(\frac{5379.75}{3}\right) = 1120.78$$

$$var_3[\,\bar{y}] = \left(1 - \frac{n}{N}\right)\left(\frac{s^2}{n}\right) = \left(1 - \frac{3}{8}\right)\left(\frac{8655.11}{3}\right) = 1803.15$$

Recall, our true variance of the distribution of the sample means was 1791.01(μg/m^3)^2. In comparison to this value, $var_1[\,\bar{y}]$ was an overestimate, while $var_2[\,\bar{y}]$ was an underestimate. However, $var_3[\,\bar{y}]$ was relatively close to $Var[\,\bar{y}]$. Regardless of the discrepancies, $var[\,\bar{y}]$ of a single sample can be used to estimate $Var[\,\bar{y}]$.

3. Results and Conclusion

3.1 Summary of Results

The 2005 WHO Air quality guidelines for PM$_{2.5}$ concentration recommended the 24-hour mean to be about 25-35 μg/m^3. Our target population had a high frequency of 24-hour PM$_{2.5}$ concentration levels ranging between 20-40 μg/m^3. Based on this one parameter and the WHO guidelines, it would seem that our target population's PM$_{2.5}$ concentration levels were adequate. However, this is misleading since μ_Y was equal to 129.98 μg/m^3 and S_Z^2 was equal to 17583.9 (μg/m^3)^2, indicating that PM$_{2.5}$ concentration levels greatly varied.

We found that our target population consisted of outliers, which were any concentration level greater than 400 μg/m^3. Once they were removed, we created a reduced target population consisting of eight numbers ranging from 13.08 μg/m^3 to 268.92 μg/m^3. Even though there was a high frequency for concentration levels between 30-40 μg/m^3, we found μ_Z equal to 105.65 μg/m^3 and S_Z^2 equal to 8596.83 (μg/m^3)^2. These values were drastically lower than that of the target population, verifying that outliers were present.

From the reduced target population, we created a sampling distribution of the sample means. We found $E[\,\bar{y}]$ equal to μ_Z, which verified $\bar{y}$ as an unbiased estimator of μ_Z. We also assumed s^2 as an unbiased estimator of S_Z^2. Then we found $Var[\,\bar{y}]$ equal to 1791.01(μg/m^3)^2, which was much lower than S_Z^2. As a result, we were able to substantially lower the variance in the sampling distribution of the sample means while maintaining accuracy of the reduced target population.

Finally, we did a trial of choosing three random samples from the sampling distribution and finding their $var[\,\bar{y}]$. In comparison to $Var[\,\bar{y}]$, we found $var_1[\,\bar{y}]$ to be an overestimate and $var_2[\,\bar{y}]$ to be an underestimate. However, even though we found $var_3[\,\bar{y}]$ to be a slight overestimate, it only had a difference of 12.14 (μg/m^3)^2. This made it relatively close to $Var[\,\bar{y}]$. This trial verified that $var[\,\bar{y}]$ of a single sample can be used to estimate $Var[\,\bar{y}]$.

3.2 Conclusion and Implications

By developing a comprehensive analysis on our sampling distribution, our results do not suggest bias. Therefore, our sampling distribution can be used as an accurate representation for Beijing's January 24-hour PM$_{2.5}$ concentration levels. These levels range from 29.32 to 201.827 μg/m^3, with a variance of 1791.01(μg/m^3)^2. According to the WHO Air quality guidelines, Beijing's January air quality fluctuates from unhealthy for sensitive groups to very unhealthy.

Since the distribution of our sampling mean was normal, the mean for that distribution was considered normal. Based on this claim, Beijing's January average of PM$_{2.5}$ concentration is 105.65 μg/m^3. In reference to the WHO's suggestion of the 24-hour mean of PM$_{2.5}$ concentration to be about 25-35 μg/m^3, Beijing's average air quality for January is considered unhealthy.

References

[1] *UCI Machine Learning Repository: Beijing Multi-Site Air-Quality Data Data Set,* https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data

[2] “Ambient (Outdoor) Air Pollution.” *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health.