PPM Comparisons between Chengdu and Beijing

by Amber Benbow

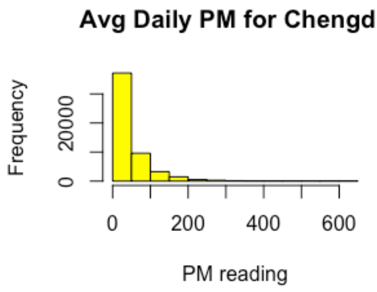
Introduction

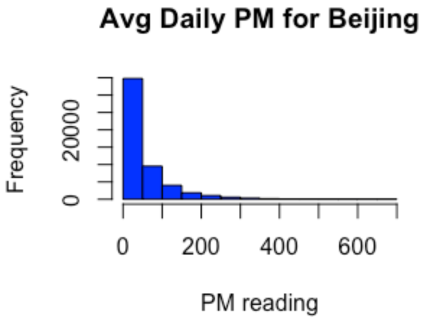
For this project I wanted to investigate if the particulate matter (PM) in two Chinese—Beijing and Chengdu—were significantly different. I have chosen this subject because I believe that there will be statistically significant rates of particulate matter in the two cities due to their population and locations within China. These PM readings would give a good indication of how poor the air quality would be. Since each workbook had PM readings from various locations within Chengdu or Beijing, I decided to create an "average PM reading" column. This provides representation for the whole of each city and simplifies my calculations. I obtained these datasets for free from Kaggle.

Graphs and Descriptive Stats

When performing a cursory look at each data set, we can see that the histograms are heavily skewed to the right and unimodal. These readings focus on the lower end of the spectrum. I did confirm that there were readings that exceeded 600 PM per day for each city and so would not want to reduce the length of the x-axis.

Below, Chengdu is in yellow and Beijing is in blue.

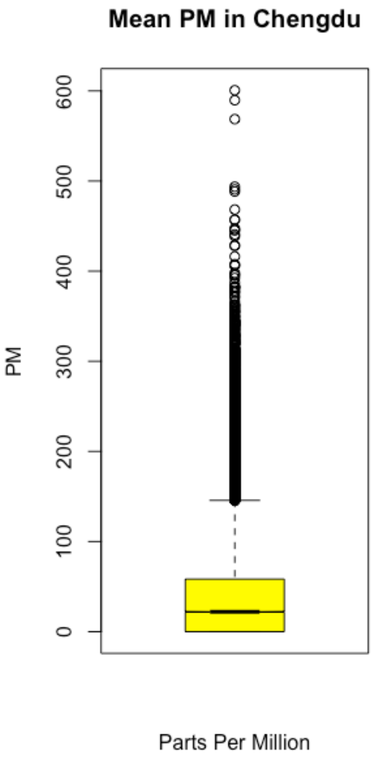


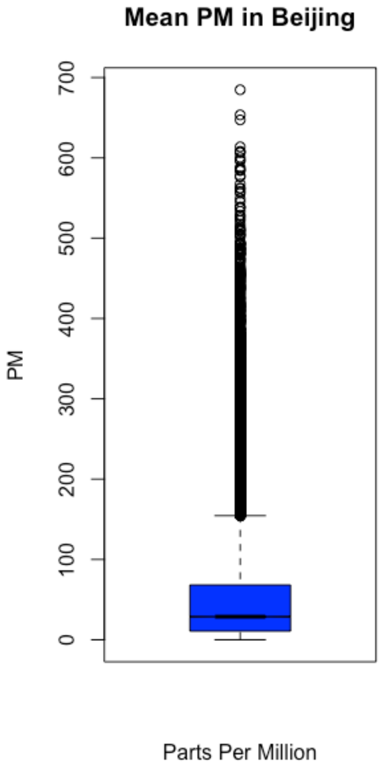


Here is the R code for the histograms:

```
> hist(chenguPM, main="Avg Daily PM for Chengdu", xlab="PM
reading", col="yellow")
> hist(beijingPM, main="Avg Daily PM for Beijing", xlab="PM
reading", col="blue")
```

Next, I created box plots in order to get an alternative look at each data set. Both boxplots seem to be similar in shape. Chengdu has fewer top-level outliers then Beijing does. Beijing overall seems to have a cluster that is shifted to be slightly higher then Chengdu with smaller numbers filling in the bottom whisker of the plot.





Here is the R code for the box plots:

```
> boxplot(chengduPM, main = "Mean PM in Chengdu", xlab = "Parts
Per Million", ylab = "PM", col = "yellow", notch = TRUE)
> boxplot(beijingPM, main = "Mean PM in Beijing", xlab = "Parts
Per Million", ylab = "PM", col = "blue", notch = TRUE)
```

In order to answer what the mean, median, and standard deviation are for each dataset I used the `Summary()` function in R.

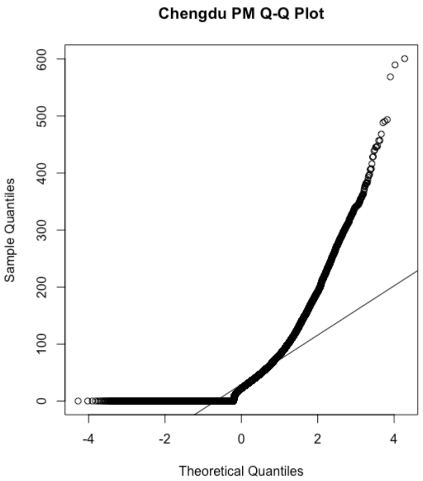
```
> summary(chengduPM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00   22.00   39.06   58.33  600.67
> summary(beijingPM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   10.75   28.75   53.12   68.25  684.75
```

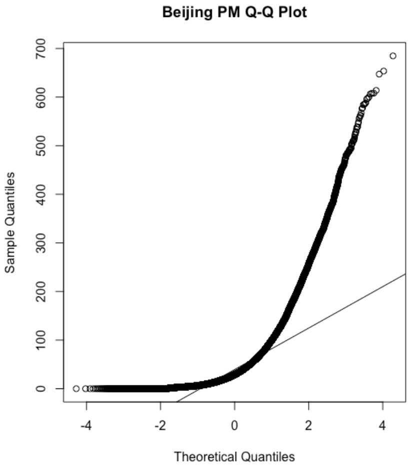
The IQR for each dataset can also be calculated using the specialized `IQR()` R function. Interestingly, it looks like when the outliers are excluded Chengdu seems to have the slightly larger PM reading.

```
> IQR(beijingPM)
[1] 57.5
> IQR(chengduPM)
[1] 58.33
```

Two sample t-test

It is unknown if the samples were collected randomly, however, since the sample size is significantly larger then 30, it is possible to say that the information collected would be an accurate representation of the population as a whole. To verify that the populations are normally distributed, we will look at their QQ plots. Both QQ plots confirm that these data sets are right skewed. Again, since the sample size is so large, we will use the Welch T-Test method.





Below is the R output I used to compute the QQ plot.

```
> qqnorm(beijingPM, main = "Beijing PM Q-Q Plot")
> qqline(beijingPM)
> qqnorm(chengduPM, main = "Chengdu PM Q-Q Plot")
> qqline(chengduPM)
```

Welch T-Test

In order to test to see if the PM levels are statistically different between these two cities, we must perform a statistical analysis through point estimation, interval estimation, and then hypothesis testing. This will help us to ensure that we are accurately measuring if random forces are at play or if there truly is a difference between each data set.

1. The first step of Welch T-Test is to ensure that it’s assumptions are met. If these assumptions are not met, then the analysis will not be useful. The assumptions are:

- the sample is collected randomly (which we cannot confirm)
- has a normal distribution
- or whose $n > 30$

Since n is larger than 30 we will proceed with the Welch T-Test knowing that our results will be worthwhile.

2. Next, we state our hypotheses. One must always be the null hypothesis which assumes that the primary force at work is random chance. The other is the alternative hypothesis which generally supports what the researcher is looking to prove. Here we are looking to prove that the difference between the two population means are different enough as to be statistically significant. Below are the hypotheses written in their statistical annotation.

- H_0: $\mu_1 - \mu_2 = 0$
- H_a: $\mu_1 - \mu_2 \neq 0$

3. To find the difference between the two sample means and solve for the Standard Error we must first know the basic stats which were discovered during the exploration phase of this analysis.

Group	$\bar{x}$	s	n
Beijing	28.75	66.80	52585
Chengdu	22	53.55	52585

To find Standard Error for the difference between sample means, we will use R to compute the following statistical equation.

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

```
> sqrt((66.80^2/52585)+(53.55^2/52585))
[1] 0.3733502
```

4. Now that we know the Standard Error, we will now calculate the Confidence Interval. For this project, we wish to know with a 98% confidence what the difference is between the two population means.
The formula for this looks like

$$(\bar{x}_1 - \bar{x}_2) \mp t_{0.01}, \text{df= min}(n_1\text{-1}, n_2\text{-1}) * \text{SE}$$

$$(28.75\text{-}22) \mp t_{0.01}, \text{df= } 52584 * 0.37$$

```
>  qt(1-0.01, df=52584)
[1] 2.326419
```

$$(6.75) \mp 2.33 * 0.37 = 0.86$$

Thus we know with a 98% confidence that the difference between the PM population mean of Beijing and Chengdu is between 5.89 and 7.61.

5. Since the difference between the two populations means does not include zero in its range we can successfully reject the null hypothesis in favor of the alternative hypothesis.

Conclusion and summary

From our analysis we can discover the difference between the two means and that they are statistically significant. It is clear that Beijing has a slightly higher PM rating then Chengdu. This conclusion makes sense because the Beijing is a larger city close to the coast (which can prevent PM from dissipating) and Chengdu is more rural and overall smaller in population. Another thing to consider when doing this kind of testing is the type of errors you may make when you draw a conclusion. In this example, we rejected the null hypothesis when in reality it could still be true. This would be a Type I Error and would wrongly attribute an effect when there is none.