import library

```
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------
------------------------------------------- tidyverse 1.2.1 --
```

```
## v tibble  2.1.1       v purrr   0.3.2
## v tidyr   0.8.3       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.4.0
## v tibble  2.1.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------
------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(statsr)
```

```
## Loading required package: BayesFactor
```

```
## Loading required package: coda
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Mo
rey (richarddmorey@gmail.com).
##
## Type BFManual() to open the manual.
## ************
```

import data

```
library(readxl)
changsha_pm2_5 <- read_excel("D:/changsha_pm2.5.xlsx",
     sheet = "Metadata", col_types = c("date",
         "numeric", "numeric", "numeric",
         "numeric", "numeric", "numeric",
         "numeric", "numeric", "text", "text",
         "text"))
View(changsha_pm2_5)
```

data manipulation and cleaning up

```
changsha_pm2_5<-changsha_pm2_5%>%
  mutate(Year = ifelse(Date<'2016-12-31',"2016",
                       ifelse(Date>"2018-12-31", "2019",
                              ifelse(Date>"2016-12-31" & Date<"2017-12-31", "2017","20
18"))))%>%
  mutate(Quality = ifelse(PM2.5 <= 12, "Good",
                          ifelse(PM2.5 > 12 & PM2.5 <=35, "Fair",
                            ifelse(PM2.5 > 55, "Severe", "Bad"))))
```

```
pm25<-changsha_pm2_5%>%
  select(1,3, 13)
```

```
y16vs17<-subset(pm25, Date < '2017-12-31')
```

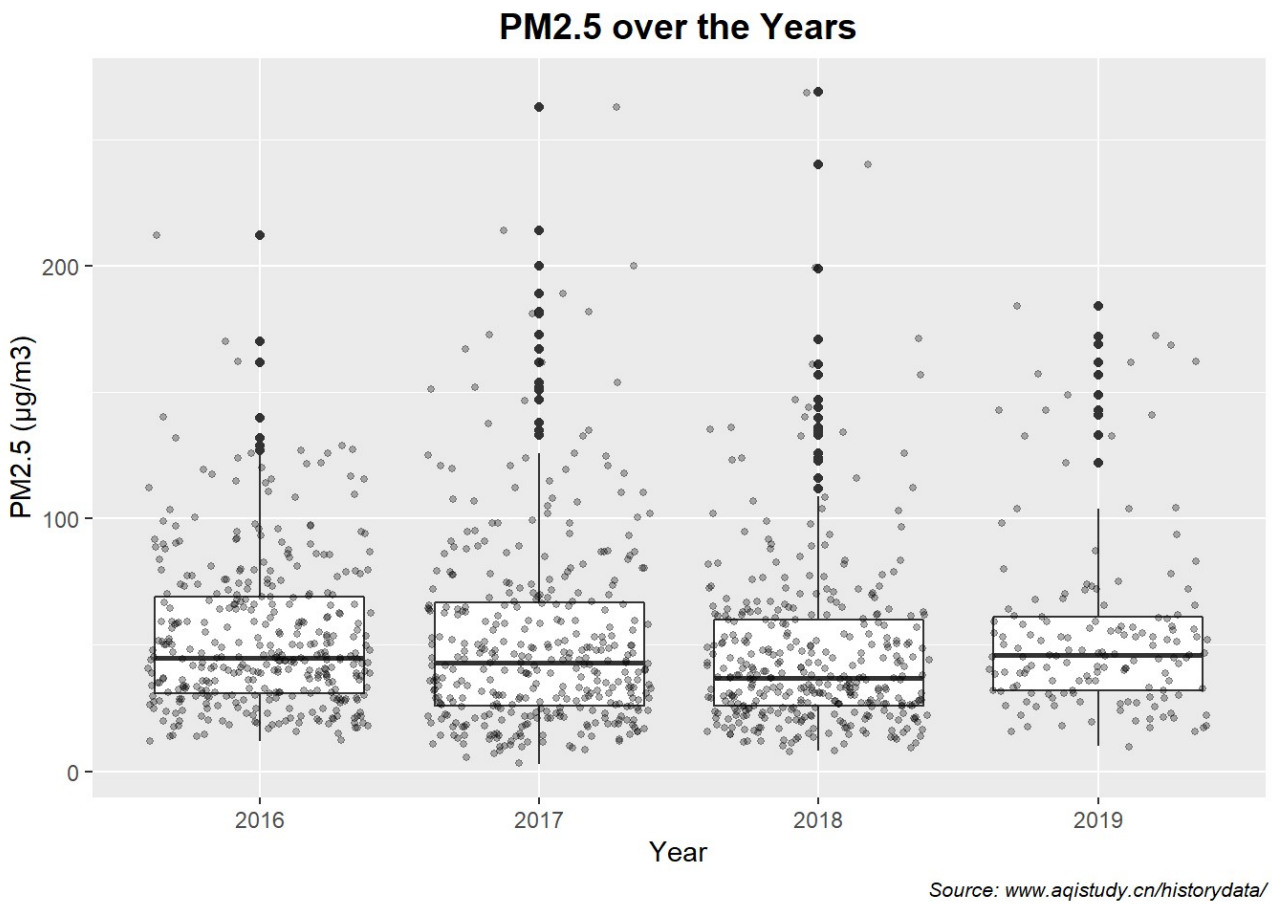
```
y17vs18<-subset(pm25, Year == "2017" | Year == "2018")
```

```
y18vs19<-subset(pm25, Year == "2018" | Year == "2019")
```

```
changsha_pm2_5%>%
  group_by(Year)%>%
  summarise(meanpm2.5=mean(PM2.5), standard_dev=sd(PM2.5), medianpm2.5=median(PM2.5),
minpm2.5=min(PM2.5, na.rm = TRUE), maxpm2.5=max(PM2.5, na.rm = TRUE), Q1 = quantile(PM
2.5, 0.25), Q3 = quantile(PM2.5, 0.75), IQR = IQR(PM2.5))
```

```
## # A tibble: 4 x 9
##   Year  meanpm2.5 standard_dev medianpm2.5 minpm2.5 maxpm2.5    Q1    Q3
##   <chr>     <dbl>        <dbl>       <dbl>    <dbl>    <dbl> <dbl> <dbl>
## 1 2016       53.5         29.8          45       12      212    31    69
## 2 2017       52.3         38.2          43        3      263    26    67
## 3 2018       46.7         33.6          37        8      269    26    60
## 4 2019       55.2         35.6          46       10      184    32    61
## # ... with 1 more variable: IQR <dbl>
```

```
changsha_pm2_5%>%
  ggplot(aes(x=Year, y=PM2.5)) +
  geom_boxplot()+
  geom_jitter(alpha = 0.3, size = 1)+
  labs(y = "PM2.5 (µg/m3)",
       title = "PM2.5 over the Years",
       caption = "Source: www.aqistudy.cn/historydata/")+
  theme(plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
        plot.caption = element_text(size = 8, face = "italic"))
```



Compare to 2016, in year 2017, the mean pm 2.5 and median pm 2.5 both decreased, but standard deviation and IQR surged. The boxplot visualized that for us.And a good side is that even though the IQR spanded, both the lower end and the upper end of it got lower than those of 2016's. This means,

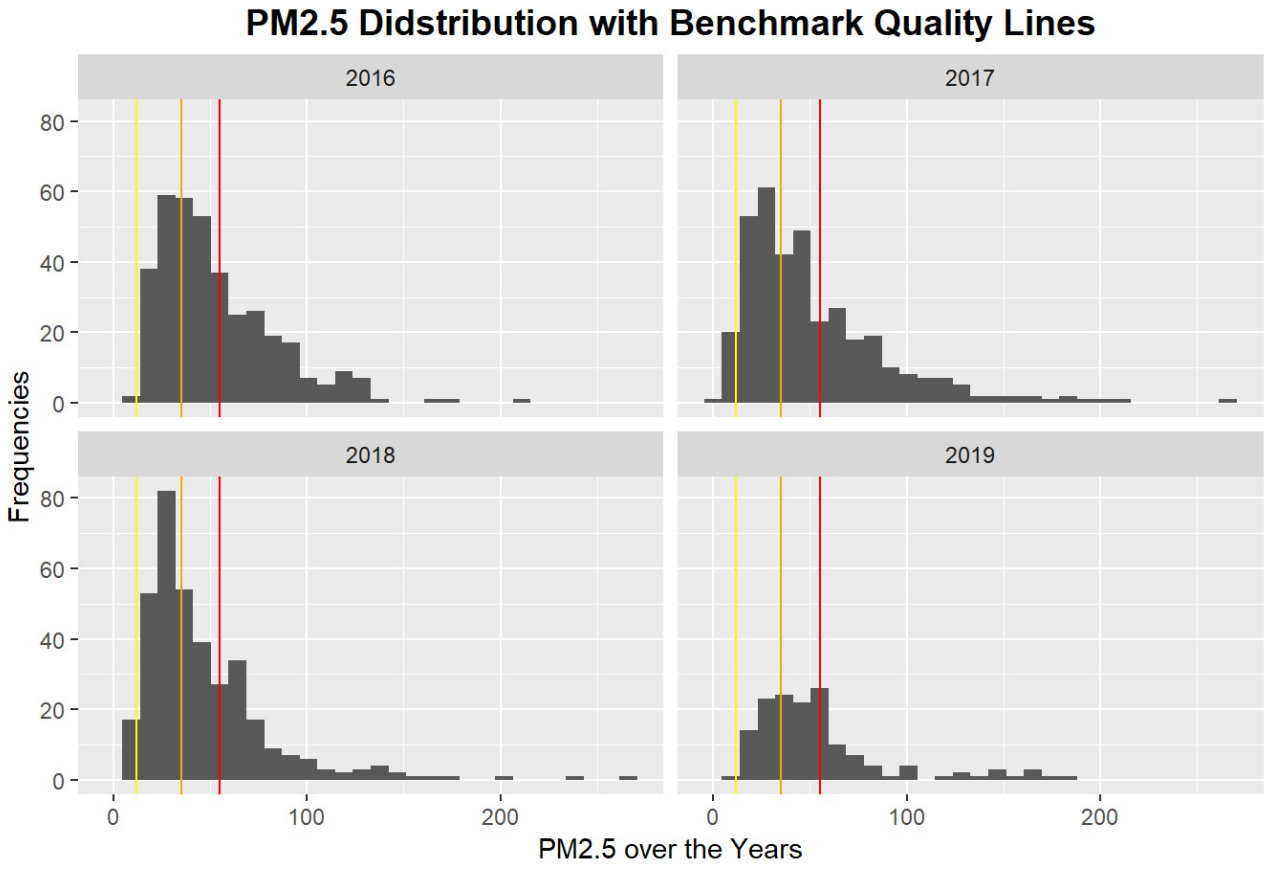
the 50% majorities shifted down, or the good days (low pm 2.5) gets better. However, with a much larger max point and bigger variation, the bad days gets worse and there are more bad days. In general, this is a good first step to have. But more efforts are needed to focus on fixing those really bad days.

Compare to 2017, in year 2018, the mean pm 2.5 and median pm 2.5 both decreased, and even standard deviation and IQR shrinked. The boxplot visualized that for us. Rather than a shift, the 50% majority became more concentrated in 2018 than 2017, and pm 2.5 spread became more steady.And the max point jumped higher than that of 2017, however the min point was not as low as that of 2017. All the information in general shows a good sign to get pm 2.5 in control. However, the problem is, the bad days gets even worse, and the good days gets less good. This could be a sign of plateau, meaning, without more government forces in the game, the air quality will hardly gets better.

2019 has only several month's worth of data. And so far, we can see an obvious increase in median, mean, and standard deviation. Less good days, but also the bad days are not as bad.

```
changsha_pm2_5%>%
  ggplot(aes(x=PM2.5))+
  geom_histogram()+
  facet_wrap(~Year, nrow = 2, ncol = 2)+
  geom_vline(xintercept = 12, color = "yellow")+
  geom_vline(xintercept = 35., color = "orange")+
  geom_vline(xintercept = 55, color = "red")+
  labs(title = "PM2.5 Didstribution with Benchmark Quality Lines",
       x = "PM2.5 over the Years",
       y = "Frequencies",
       caption = "Source: www.aqistudy.cn/historydata/")+
  theme(plot.title = element_text(size = 14, face="bold", hjust = 0.5),
        plot.caption = element_text(size = 8, face = "italic"))
```

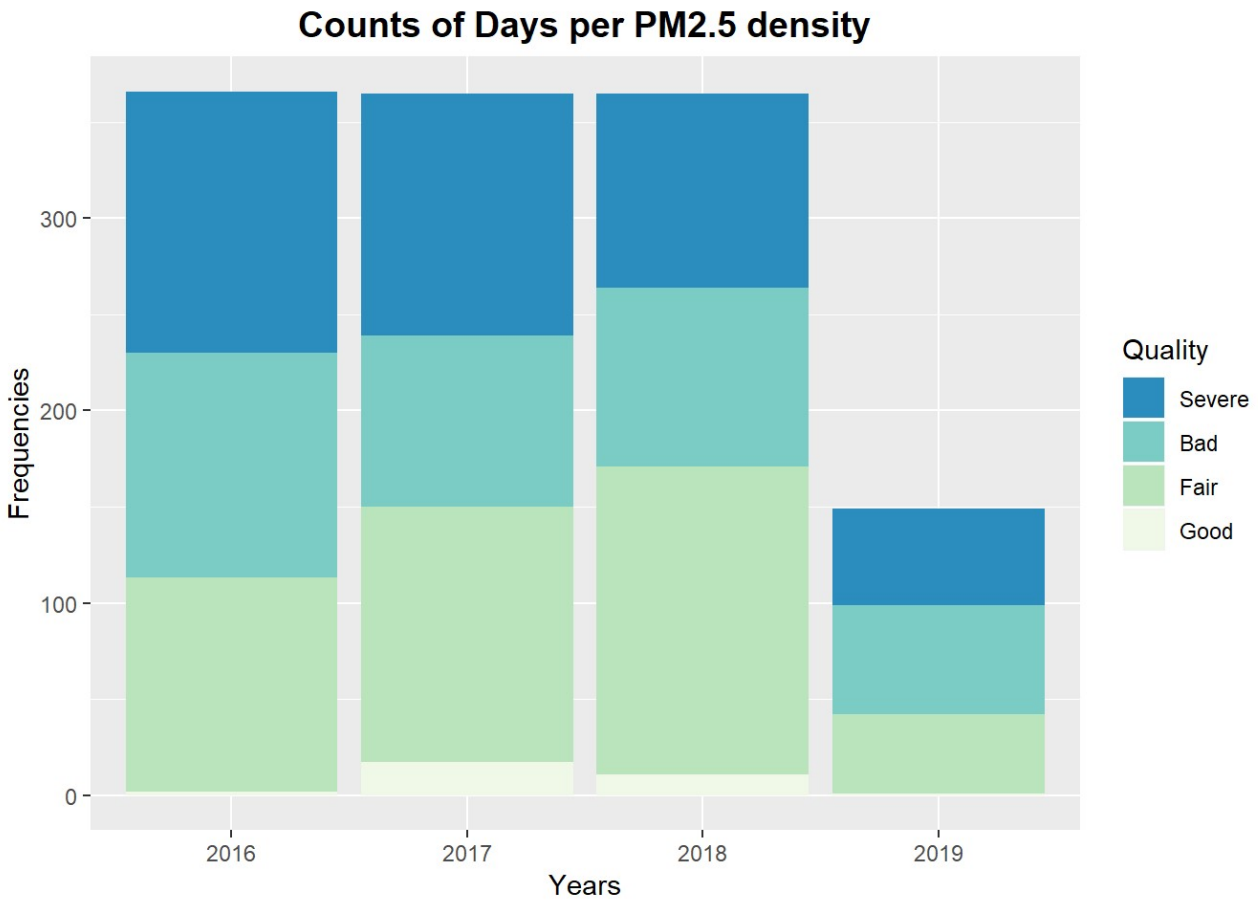
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
changsha_pm2_5%>%
  group_by(Year)%>%
  summarise(No_of_good_days = sum(PM2.5 <= 12), No_of_fair_days = sum(PM2.5 > 12 & PM
2.5 <=35),
            No_of_bad_days = sum(PM2.5 > 35 & PM2.5 <=55), No_of_severe_days = sum(PM
2.5 > 55))
```

```
## # A tibble: 4 x 5
##   Year  No_of_good_days No_of_fair_days No_of_bad_days No_of_severe_days
##   <chr>           <int>           <int>          <int>              <int>
## 1 2016                2             111            117                136
## 2 2017               17             133             89                126
## 3 2018               11             160             93                101
## 4 2019                1              41             57                 50
```

```
changsha_pm2_5%>%
  ggplot(aes(x = Year, fill = factor(Quality, levels = c("Severe", "Bad", "Fair", "Goo
d"))))+
  geom_bar()+
  scale_fill_brewer(palette = 4, direction = -1)+
  labs(title = "Counts of Days per PM2.5 density",
       x = "Years",
       y = "Frequencies",
       fill = "Quality")+
  theme(plot.title = element_text(size = 14, face="bold", hjust = 0.5),
        plot.caption = element_text(size = 8, face = "italic"))
```

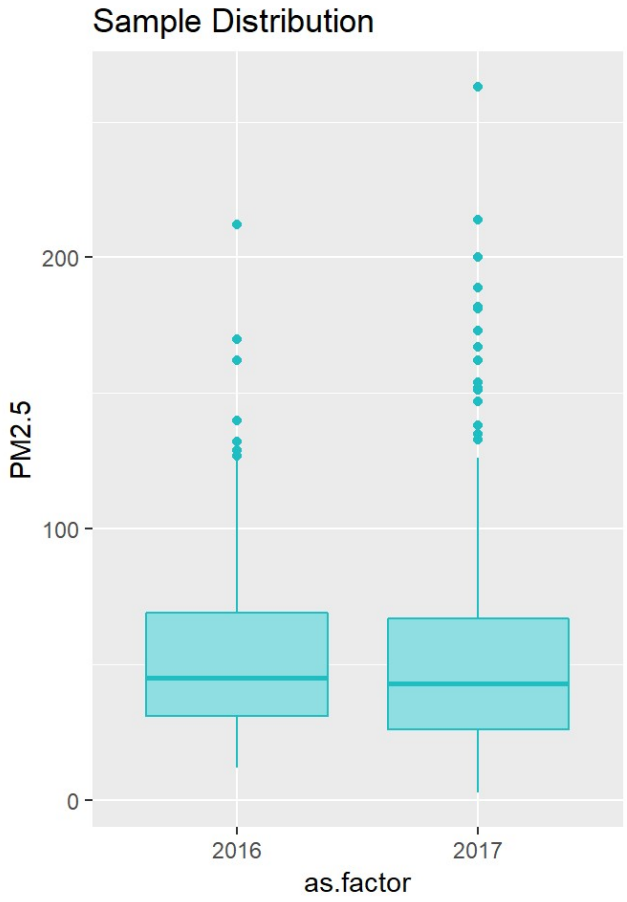


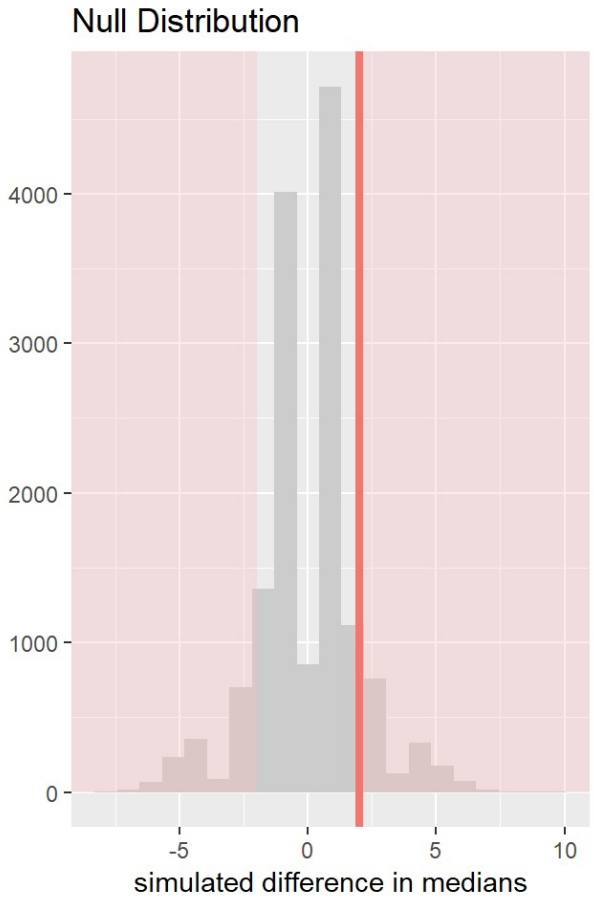
In the distributions of the four years, the three lines represent good, fair, and bad air quality breakpoint per US air quality standard. We further visualized the comparison among years with summary table and filled bar chart. The bar chart comparison sent out mixed messages that, on the good side, total of bad and severe days are decreasing year to year from 2016 to 2018, however, on the bad side, the good days are decreasing from 2017 to 2019 (which only has partial years, so there is some hope for good days in the rest of the year).

```
inference(data = y16vs17, y = PM2.5, x = as.factor(Year), type = "ht", statistic = "me
dian", method = "simulation",
    null = 0, success = "atheist", alternative = "twosided", conf_level = 0.95)
```

```
## Warning: Ignoring success since y is numerical
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2016 = 366, y_med_2016 = 45, IQR_2016 = 38
## n_2017 = 365, y_med_2017 = 43, IQR_2017 = 41
## H0: mu_2016 =  mu_2017
## HA: mu_2016 != mu_2017
## p_value = 0.3316
```

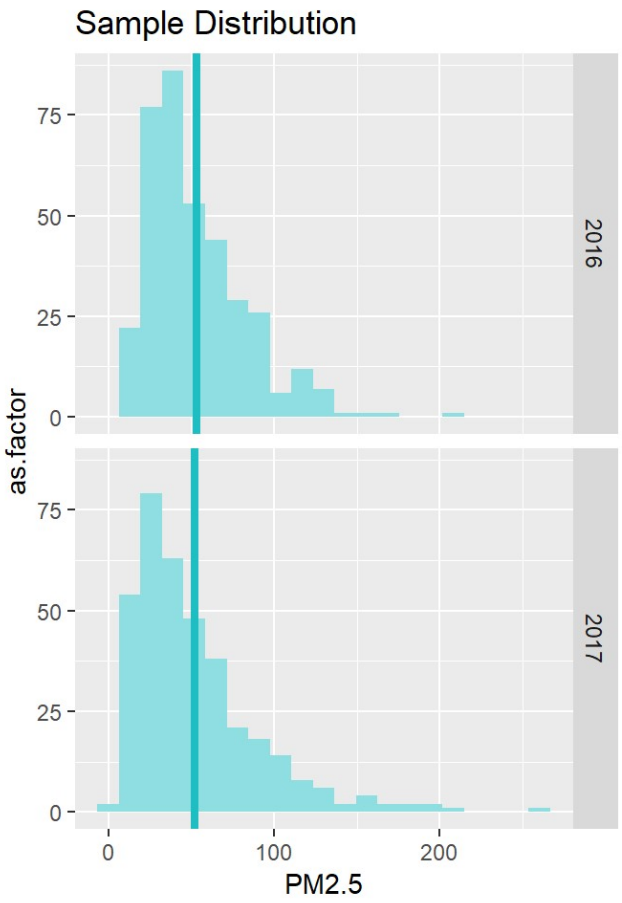


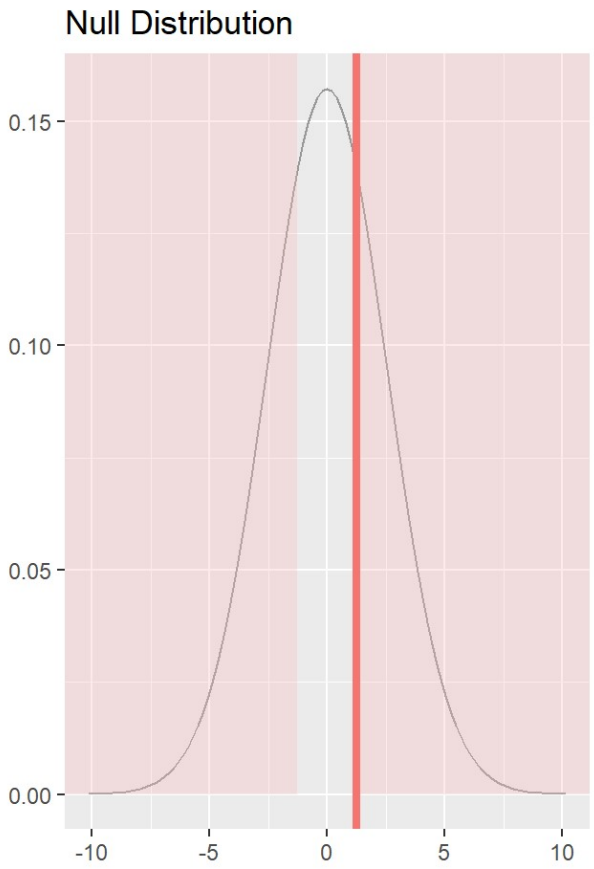



```
inference(data = y16vs17, y = PM2.5, x = as.factor(Year), type = "ht", statistic = "me
an", method = "theoretical",
    null = 0, success = "atheist", alternative = "twosided", conf_level = 0.95)
```

```
## Warning: Ignoring success since y is numerical
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2016 = 366, y_bar_2016 = 53.5191, s_2016 = 29.8365
## n_2017 = 365, y_bar_2017 = 52.263, s_2017 = 38.2288
## H0: mu_2016 =  mu_2017
## HA: mu_2016 != mu_2017
## t = 0.4951, df = 364
## p_value = 0.6208
```



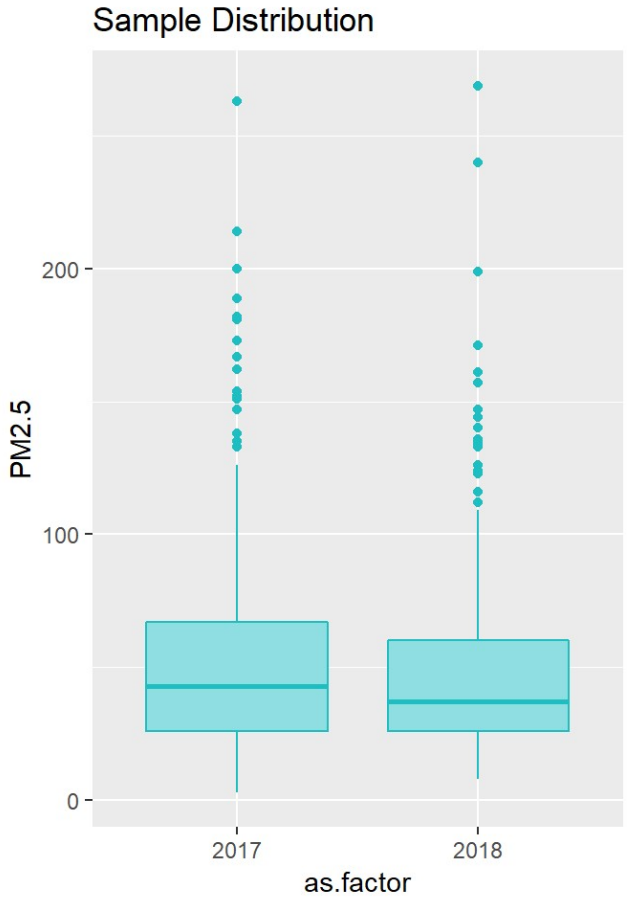


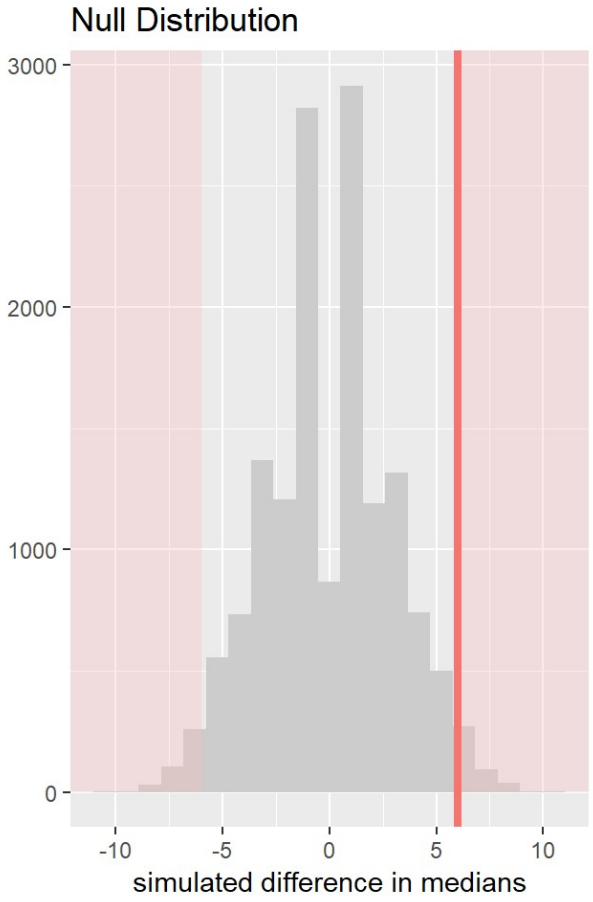
Both p values for median and mean of the two years are fairly big, meaning the difference between the means or medians of the two years are not statistically significant. Therefore, even though the mean and median decreased, we would still conclude that there is no difference between the two years. Or in other words, there is not noticable improvement or setback of average air quality between the two years, it is more of making bad days worse and good days better but total poplution is somehow consistant.

```
inference(data = y17vs18, y = PM2.5, x = as.factor(Year), type = "ht", statistic = "me
dian", method = "simulation",
    null = 0, success = "atheist", alternative = "twosided", conf_level = 0.95)
```

```
## Warning: Ignoring success since y is numerical
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2017 = 365, y_med_2017 = 43, IQR_2017 = 41
## n_2018 = 365, y_med_2018 = 37, IQR_2018 = 34
## H0: mu_2017 =  mu_2018
## HA: mu_2017 != mu_2018
## p_value = 0.0535
```

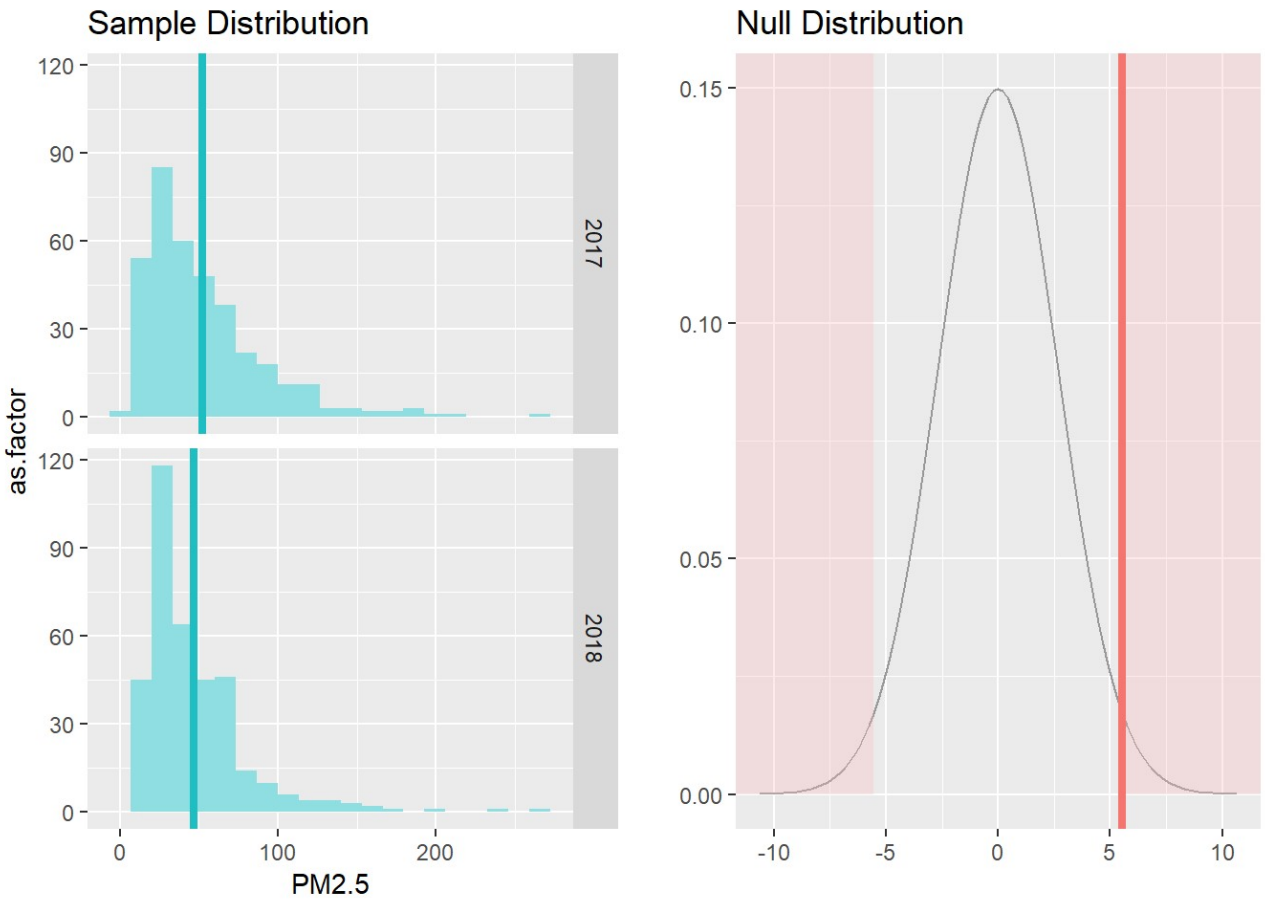





```
inference(data = y17vs18, y = PM2.5, x = as.factor(Year), type = "ht", statistic = "me
an", method = "theoretical",
    null = 0, success = "atheist", alternative = "twosided", conf_level = 0.95)
```

```
## Warning: Ignoring success since y is numerical
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2017 = 365, y_bar_2017 = 52.263, s_2017 = 38.2288
## n_2018 = 365, y_bar_2018 = 46.7068, s_2018 = 33.5663
## H0: mu_2017 =  mu_2018
## HA: mu_2017 != mu_2018
## t = 2.0865, df = 364
## p_value = 0.0376
```

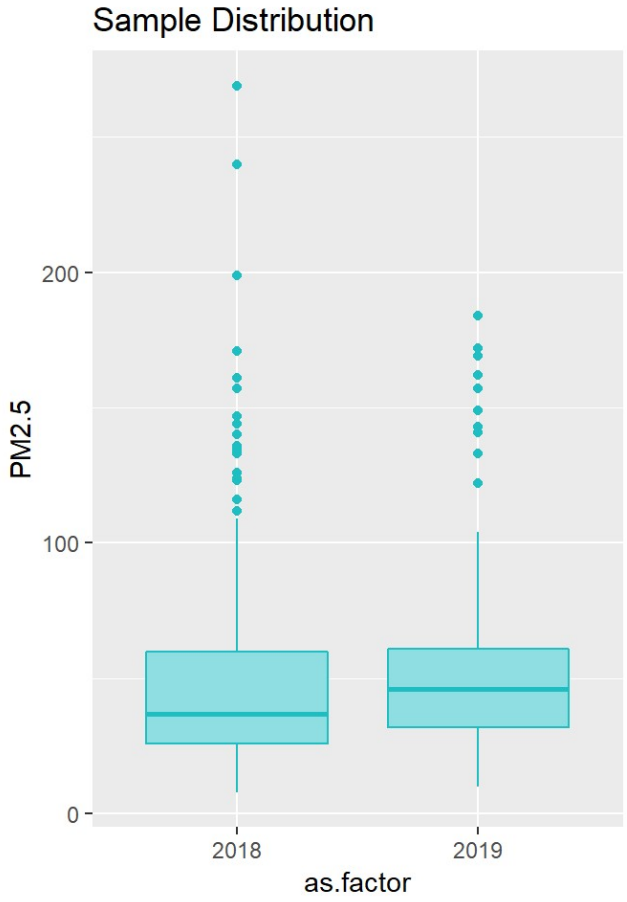


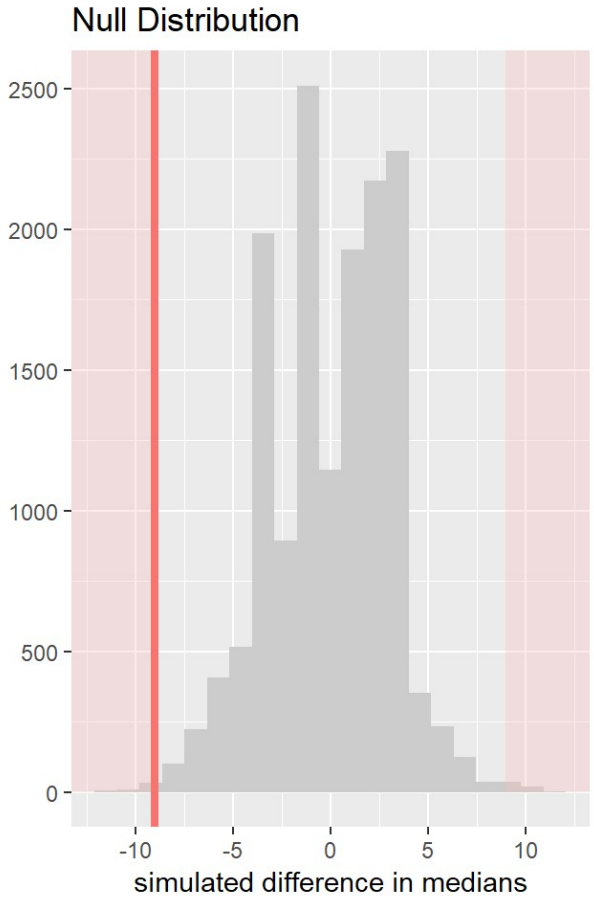
The median is indifferent between the years, but the mean has statistically decreased from 2017 to 2018.This agrees with the box plot comparison that 2018 distribution became more densed towards the lower end than 2017's that brought the mean down but did not change the robust median as much.

```
inference(data = y18vs19, y = PM2.5, x = as.factor(Year), type = "ht", statistic = "me
dian", method = "simulation",
    null = 0, success = "atheist", alternative = "twosided", conf_level = 0.95)
```

```
## Warning: Ignoring success since y is numerical
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2018 = 365, y_med_2018 = 37, IQR_2018 = 34
## n_2019 = 149, y_med_2019 = 46, IQR_2019 = 29
## H0: mu_2018 =  mu_2019
## HA: mu_2018 != mu_2019
## p_value = 0.006
```

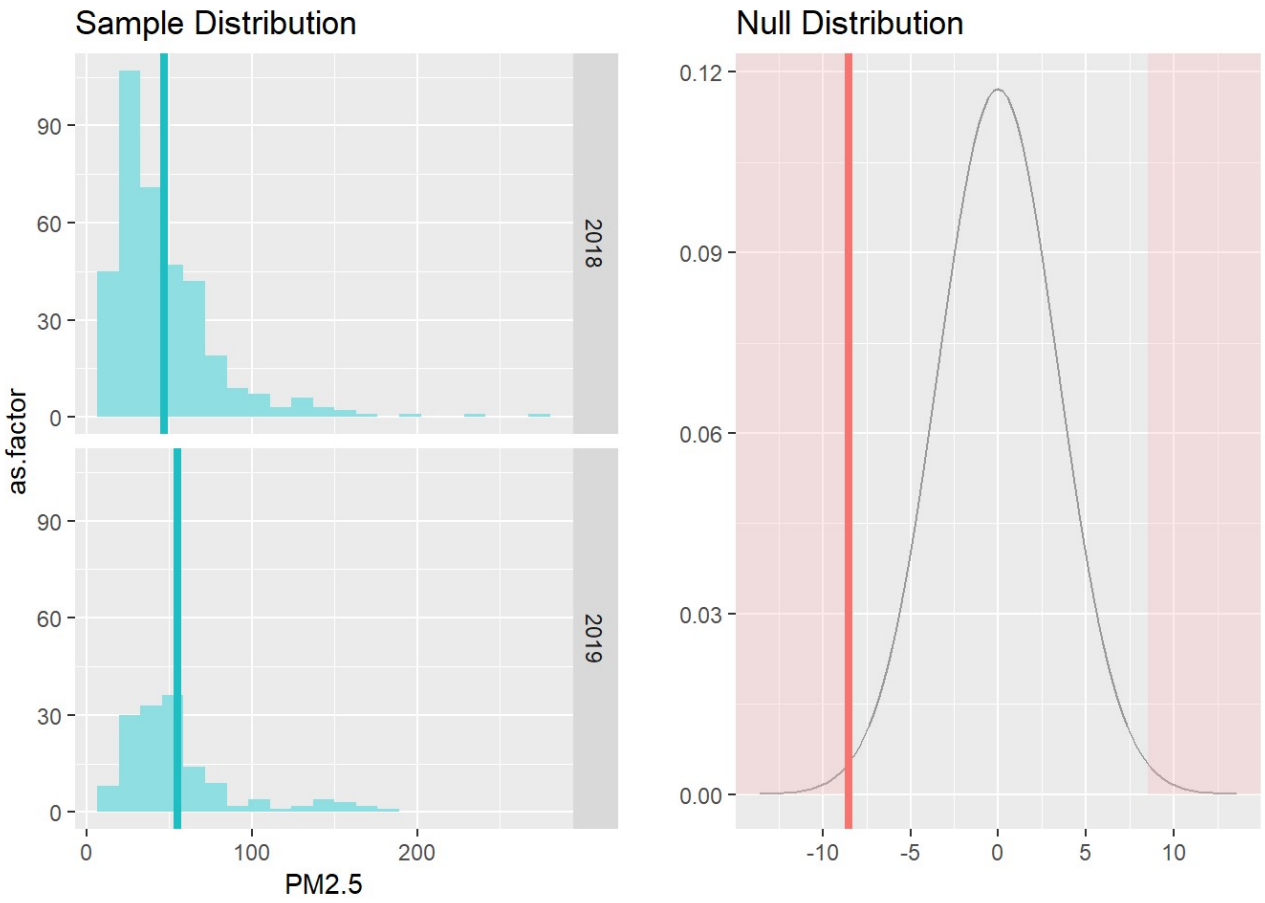





```
inference(data = y18vs19, y = PM2.5, x = as.factor(Year), type = "ht", statistic = "me
an", method = "theoretical",
    null = 0, success = "atheist", alternative = "twosided", conf_level = 0.95)
```

```
## Warning: Ignoring success since y is numerical
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2018 = 365, y_bar_2018 = 46.7068, s_2018 = 33.5663
## n_2019 = 149, y_bar_2019 = 55.2483, s_2019 = 35.5687
## H0: mu_2018 =  mu_2019
## HA: mu_2018 != mu_2019
## t = -2.5103, df = 148
## p_value = 0.0131
```



Even though we only have partial 2019 data, it would not compromize our capability to compare means and medians. For data we have for 2019 so far, per the p value, it is clear that 2019 mean and median of pm 2.5 is significantly higher than those of 2018. However, as pm 2.5 density could be seasonable, we cannot make conclusion yet.

In conclusion, we can see some progress from 2016 to 2018 on pm 2.5 control but will for sure need more efforts to bring it down, especially effort to get under control of the worst days. From 2016 to 2017, the overall pm 2.5 for the whole year did not change much, but bad days became less and fair days became more, or in other words, bad days got worse but good days got better. From 2017 to 2018, almost all indicators went to the positive direction of improvement. However, the good days were not as good as 2017, and the worst several days were worse. All the change from 2016 to 2017 then to 2018 almost makes me feel there was only actions like sprinkling water to try to keep PM 2.5 done, but rarely any action to control PM 2.5 from the source.