Project Part 2

Michael Song, Morgan Wang

Team's Github:https://github.com/anv7wj-gif/Project-Part-2

Fine particulate matter (PM2.5) is an important public health concern because small particles can travel deep into the lungs and bloodstream. In Part 1, we summarized daily PM2.5 levels in Los Angeles County from December 2023 through August 2024 using EPA AirData. Winter months appeared to show slightly higher and more variable PM2.5 than summer months, but descriptive summaries alone cannot determine whether these differences reflect true seasonal effects or natural day-to-day variability. In this Part 2 analysis, we formally address the question: **"Is the mean daily PM2.5 concentration lower in summer (June–August) than in winter (December–February) in Los Angeles County?"** We apply hypothesis testing and evaluate the assumptions required for inference.

Methods

Let
- μ_S: true mean daily PM2.5 in summer
- μ_W: true mean daily PM2.5 in winter

We test:

$$H_0 : \mu_S - \mu_W = 0, \quad H_A : \mu_S - \mu_W < 0.$$

Because the data consist of independent daily averages for two groups, and PM2.5 is continuous, we use a Welch two-sample t-test, which does not assume equal variances. Assumptions are checked using histograms and Q–Q plots. With sample sizes near 90 per group, the t-test is generally robust to moderate skewness.

Data Processing

```r
df <- read.csv("LA_2024_PM25.csv")
df$Date <- as.Date(df$Date, "%m/%d/%Y")
df_clean <- df[, c("Date", "Daily.Mean.PM2.5.Concentration", "County")]
names(df_clean)[2] <- "pm25"
df_clean$month <- as.numeric(format(df_clean$Date, "%m"))
df_clean$season <- ifelse(df_clean$month %in% c(12,1,2), "Winter",
                   ifelse(df_clean$month %in% c(6,7,8), "Summer", NA))
df_clean <- df_clean[!is.na(df_clean$season) & !is.na(df_clean$pm25), ]
by_day <- aggregate(pm25 ~ Date + season, data = df_clean, FUN = mean)
table(by_day$season)
```

```
##
## Summer Winter
##     92     91
```

Assumption Checks

To evaluate whether the t-test is appropriate, we examine one histogram and one Q–Q plot for each season.

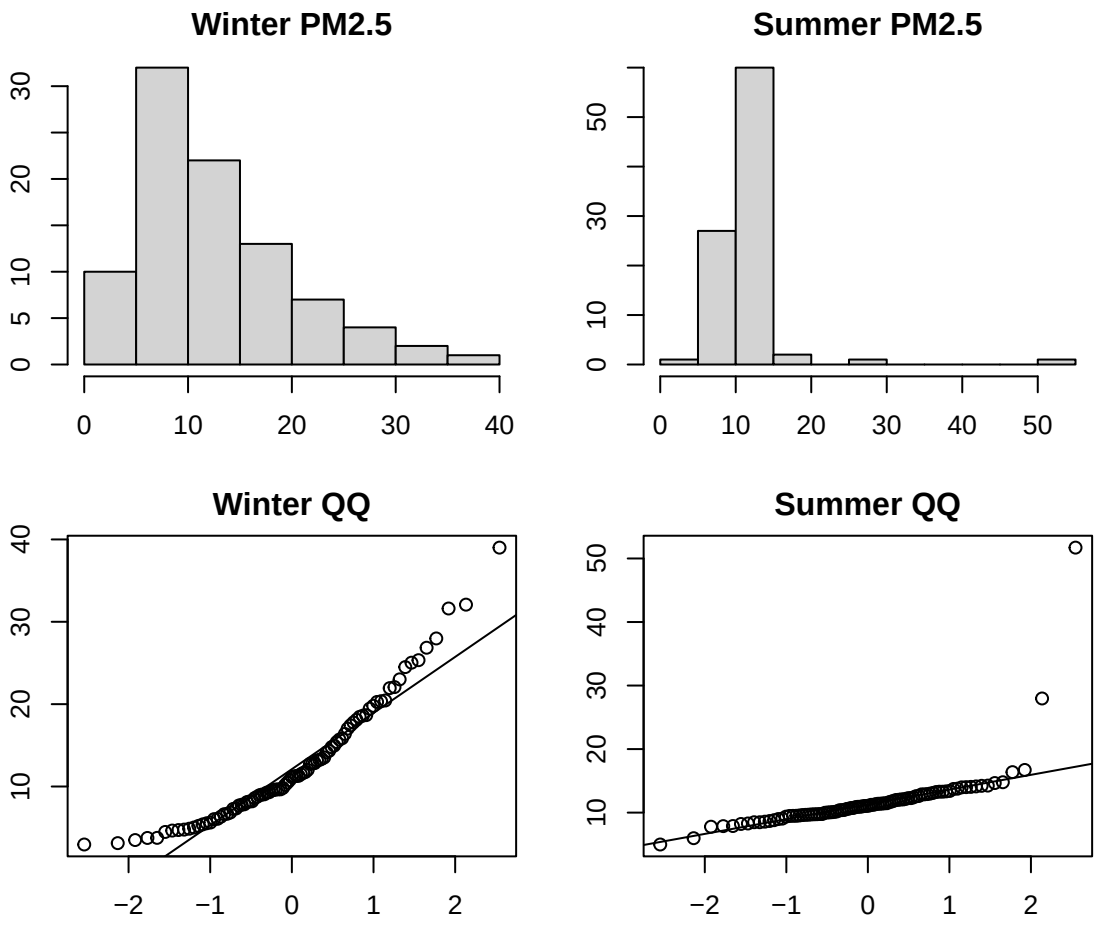
```r
par(mfrow = c(2, 2), mar=c(3,3,2,1))

hist(by_day$pm25[by_day$season == "Winter"],
     main = "Winter PM2.5", xlab = "")

hist(by_day$pm25[by_day$season == "Summer"],
     main = "Summer PM2.5", xlab = "")

qqnorm(by_day$pm25[by_day$season == "Winter"], main = "Winter QQ")
qqline(by_day$pm25[by_day$season == "Winter"])

qqnorm(by_day$pm25[by_day$season == "Summer"], main = "Summer QQ")
qqline(by_day$pm25[by_day$season == "Summer"])
```



Assumption discussion:
Both seasons show right-skewed distributions and several higher-pollution days. The Q–Q plots display upper-tail deviations consistent with skewness. With sample sizes of 92 summer days and 91 winter days, the Welch t-test is sufficiently robust, so we proceed while noting skewness as a limitation.

Hypothesis Test

```
by_day$season <- factor(by_day$season, levels=c("Summer","Winter"))
tapply(by_day$pm25, by_day$season, summary)
```

```
## $Summer
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.982   9.733  11.077  11.768  12.868  51.711
##
```

```
## $Winter
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.958   7.513  11.142  12.654  16.705  39.007
```

```
t_out <- t.test(pm25 ~ season, data=by_day, alternative="less")
t_out
```

```
##
##  Welch Two Sample t-test
##
## data:  pm25 by season
## t = -0.95848, df = 159.84, p-value = 0.1696
## alternative hypothesis: true difference in means between group Summer and group Winte
## 95 percent confidence interval:
##       -Inf 0.6429449
## sample estimates:
## mean in group Summer mean in group Winter
##             11.76828             12.65374
```

Results

Summer days averaged about 11.77 µg/m^3, while winter days averaged about 12.65 µg/m^3, a sample difference of roughly -0.89 µg/m^3. The Welch t-test produced a test statistic of -0.96 with about 160 degrees of freedom and a one-sided p-value of 0.17. The one-sided 95% confidence interval for $\mu_S - \mu_W$ was $(-\infty, 0.64)$, which includes 0. Since the p-value exceeds 0.05, we fail to reject H_0. The evidence is insufficient to conclude that true mean daily PM2.5 is lower in summer than in winter during this time period.

Conclusions

Although summer days had a lower sample mean PM2.5 than winter days, the difference was not statistically significant. The observed seasonal difference could reasonably be due to natural daily variability. Limitations include: the dataset spans less than one year; important factors such as weather and wildfire activity are not controlled; and daily PM2.5 may be temporally correlated. Within these constraints, we do not find strong evidence of a true seasonal difference in mean PM2.5.

References

- U.S. Environmental Protection Agency. "AirData: Air Quality Data Collected at Outdoor Monitors Across the US." https://www.epa.gov/outdoor-air-quality-data/download-daily-data

- Day-of-week and seasonal patterns of PM2.5 concentrations over the United States (Zhao et al., 2018) https://www.sciencedirect.com/science/article/pii/S1352231018305715

- Seasonal dynamics and trends in air pollutants: A comprehensive analysis of PM2.5, NO2, CO, SO2 and O3 in Houston, USA (Alam, Karim & Uz Zaman, 2025) https://link.springer.com/article/10.1007/s11869-025-01790-9