# Assignment 3 - Data Analysis and Statistical Methods

Frans Simanjuntak - S3038971

September 28, 2017

## 1 Question 1

(A) The $\alpha$ value gives the probability of making a type I error.

(B) If the p-value of your hypothesis test is $< \alpha$, then you reject the null hypothesis

(C) If I cannot not reject the Null hypothesis based on my test results, this means the Null hypothesis is true.

(D) If we reject the Null Hypothesis, we think the Null Hypothesis is false.

(E) If I increase the confidence level 1-$\alpha$ for a t-test and don't change any of the other parameters, the power of the test will be affected

## 2 Question 2

Given:
n = 64
$\overline{X}$ = 150
s = 50

(A) Hypothesis:

- $H_0 : \mu = 165$ (The electricity cost for people in Groningen is similar to the national average)
- $H_1 : \mu \neq 165$ (The electricity cost for people in Groningen is different than the national average)

We want to test this on 95% confidence level ($\alpha$=0.05)
Rejection region: Reject $H_0$ if $z_{test} \geq z_{(1-\alpha/2)}$ or $z_{test} \leq z_{(\alpha/2)}$
$z_{(\alpha/2)}$ = -1.96
$z_{(1-\alpha/2)}$ = 1.96

(B) The statistic test:
$z_{test} = \frac{\overline{X}-\mu_0}{s/\sqrt{n}}$
$= \frac{150-165}{50/\sqrt{64}}$
= -2.4
Since $z_{test} \leq z_{(\alpha/2)}$ therefore we reject the null hypothesis. The electricity cost for people in Groningen is different than the national average.

(C) Here is the code in R

```
p_value = 2 * pnorm(−2.4)
```

The result of the p-value is 0.01639507. Since the p-value $\leq$ 0.05($\alpha$) therefore we must reject the null hypothesis.

(D) Plotting

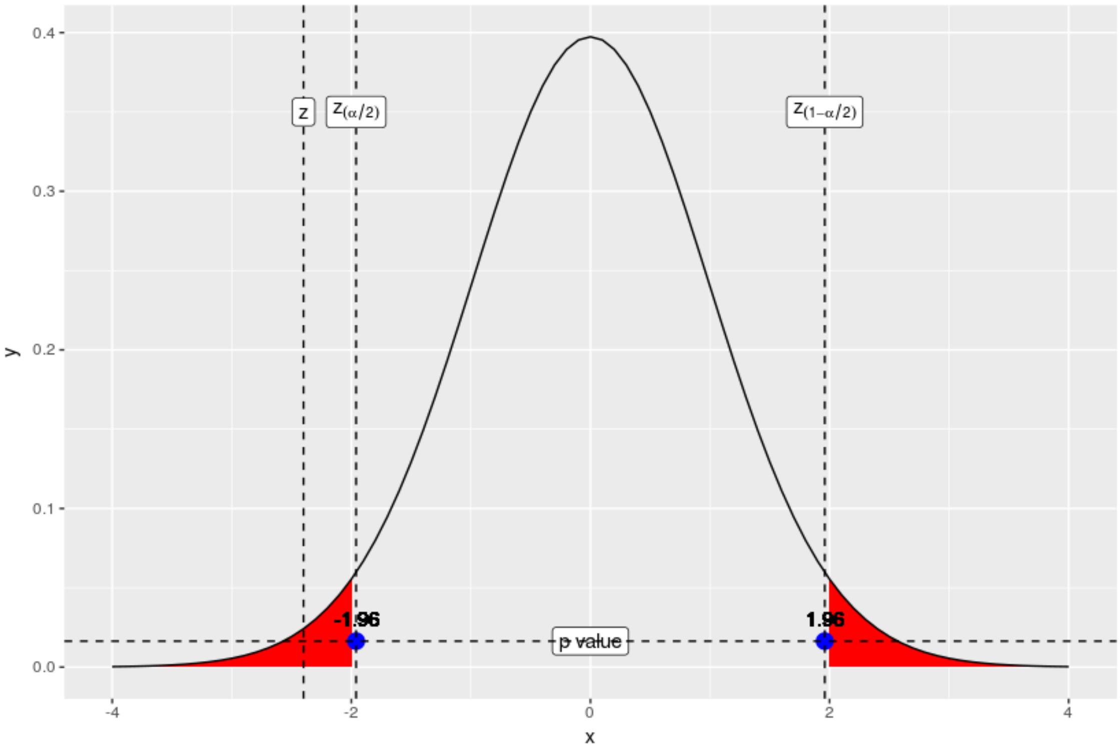


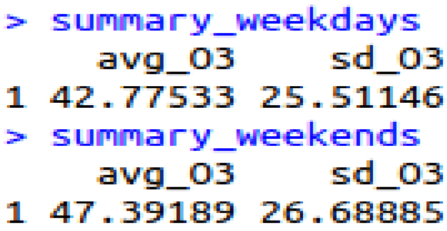
Figure 1: The hypothesis testing of the electricity cost for people in Groningen compared to the national average.

Here is the code in R:

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

n = 64 # sample size
s = 50 # standard deviation
x_bar = 150 # sample average
mu_0 = 165
alpha = 0.05

z = (x_bar - mu_0)/(s/sqrt(n))
z.half.alpha = qnorm(1-alpha/2)
p_value = 2 * pnorm(z)

x = seq(-4, 4, 0.1) # sequence from -4 to 4 in increments of 0.1
y = dt(x, df=n-1)   # probability density of the t distribution

dist = data.frame(x=x,y=y)

#Plot with ggplot
ggplot(dist, aes(x,y)) +
annotate(geom = "label", x = 0,y = p_value,label = "p␣value", parse = FALSE) +
geom_area(data = subset(dist, x <= -z.half.alpha), aes(x=x, y=y), fill="red")+
geom_area(data = subset(dist, x >= z.half.alpha), aes(x=x, y=y), fill="red")+
annotate(geom = "point", x=-z.half.alpha, y=p_value,color = "blue", size = 4)+
annotate(geom = "point", x=z.half.alpha , y=p_value , color = "blue", size = 4)+
geom_text(x=-z.half.alpha , y=0.03 , label="-1.96", col="Black")+
geom_text(x=z.half.alpha , y=0.03 , label="1.96", col="Black")+
geom_line() +
geom_vline(xintercept = -z.half.alpha ,linetype="dashed") +
annotate(geom = "label",x = -z.half.alpha ,y = 0.35 ,label = "z[(alpha/2)]"
,parse = TRUE) +
geom_vline(xintercept = z,linetype="dashed") +
annotate(geom = "label", x = z,y = 0.35 ,label = "z",parse = FALSE)+
geom_vline(xintercept = z.half.alpha ,linetype="dashed") +
annotate(geom = "label",x = z.half.alpha , y = 0.35 ,label = "z[(1-alpha/2)]",
parse = TRUE)+
geom_hline(yintercept = p_value , linetype="dashed")
```

## 3 Question 3

(A) Below are the statistic summary (mean and standard deviation) of O3 on weekdays and weekends.



```
> summary_weekdays
    avg_O3    sd_O3
1 42.77533 25.51146
> summary_weekends
    avg_O3    sd_O3
1 47.39189 26.68885
```

Figure 2: Statistic summary of O3 on weekdays and weekends for the entire year.

R-code to generate the mean and the standard deviation of O3:

```
library("dplyr")
setwd("~/Assignments/Assignment1")
df <- read.csv("data1(2).csv", header = TRUE)
df[, "date"] = lapply(df["date"],function(x){strptime(x, "%d/%m/%y")})

weekdays_data <- df %>% select(weekday, O3) %>%
filter(weekday %in% c(2:6))%>% select(O3)

weekends_data <- df %>% select(weekday, O3) %>%
filter(weekday %in% c(1,7)) %>% select(O3)

summary_weekdays <- weekdays_data %>% summarise(avg_O3 = mean(O3), sd_O3 =sd(O3))
summary_weekends <- weekends_data %>% summarise(avg_O3 = mean(O3), sd_O3 =sd(O3))

summary_weekdays
summary_weekends
```

From figure 2, it can be seen that the mean value and the standard deviation of O3 on weekend are higher than on weekdays.

Now, lets define our hypothesis testing:

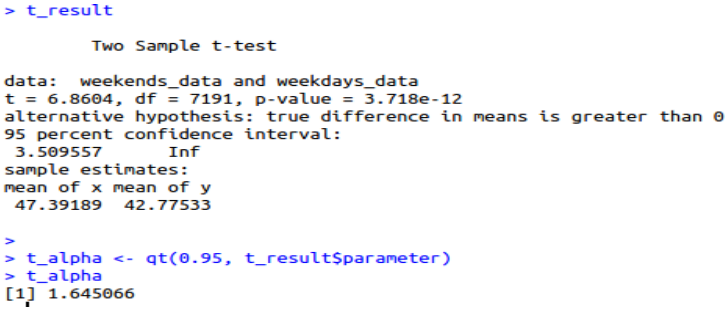
- $H_0$ : $\mu_1$ - $\mu_2 \leq 0$ (The $\mu$ value of O3 on weekends is equal or less than the value on weekdays).
- $H_1$ : $\mu_1$ - $\mu_2 > 0$ (The $\mu$ value of O3 on weekends is greater than the value on weekdays).

Assume that the confidence interval is 95% ($\alpha = 0.05$).
Rejection region: Reject $H_0$ if $t > t_\alpha$.
Assumption: Variance is equal.

Now lets perform statistic test using R:

```
library("dplyr")
setwd("~/Assignment1")
df <- read.csv("data1(2).csv", header = TRUE)
df[, "date"] = lapply(df["date"],function(x){strptime(x, "%d/%m/%y")})

weekdays_data <- df %>% select(weekday, O3) %>%
filter(weekday %in% c(2:6)) %>% select(O3)

weekends_data <- df %>% select(weekday, O3) %>%
filter(weekday %in% c(1,7)) %>% select(O3)

t_result <-t.test(weekends_data, weekdays_data, alternative = "greater",
conf.level = 0.95, var.equal = TRUE, paired = FALSE)
t_result

t_alpha <- qt(0.95, t_result$parameter)
t_alpha
```

The result is:



```
> t_result

        Two Sample t-test

data:  weekends_data and weekdays_data
t = 6.8604, df = 7191, p-value = 3.718e-12
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.509557      Inf
sample estimates:
mean of x mean of y 
 47.39189  42.77533 

> 
> t_alpha <- qt(0.95, t_result$parameter)
> t_alpha
[1] 1.645066
```

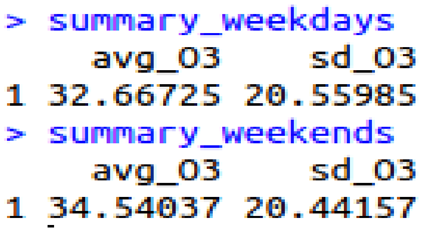
Figure 3: The two sample test of O3 on weekdays and weekends for the entire year.

From figure 3 we can see:

- The $t_{test}$ is 6.8604. This value tells us the statistic test.
- The degree of freedom is 7191.
- The p-value is 3.718e-12.
- The $t_\alpha$ is 1.645866.
- The rejection region is $t > 1.645866$.
- The average of O3 on weekends is 47.39189 and on weekdays is 42.77533.

From the above results, we can draw a conclusion that we reject $H_0$ since the $t_{test} > t_\alpha$, $6.8604 > 1.645866$.

(B) For January data, the mean value and the standard deviation of O3 are as follows:



```
> summary_weekdays
    avg_O3    sd_O3
1 32.66725 20.55985
> summary_weekends
    avg_O3    sd_O3
1 34.54037 20.44157
```

Figure 4: Statistic summary of O3 on weekdays and weekends in January.

R-code to generate the mean and the standard deviation of O3:

```
library("dplyr")
setwd("~/Assignment1")
df <- read.csv("data1(2).csv", header = TRUE)
df[, "date"] = lapply(df["date"],function(x){strptime(x, "%d/%m/%y")})

weekdays_data <- df %>% select(weekday,month, O3) %>%
filter(weekday %in% c(2:6) & month ==1) %>% select(O3)

weekends_data <- df %>% select(weekday,month, O3) %>%
filter(weekday %in% c(1,7) & month ==1) %>% select(O3)

summary_weekdays <- weekdays_data %>% summarise(avg_O3 = mean(O3), sd_O3 =sd(O3))
summary_weekends <- weekends_data %>% summarise(avg_O3 = mean(O3), sd_O3 =sd(O3))

summary_weekdays
summary_weekends
```

From figure 4, it can be seen that the mean value of O3 on weekend is higher than on weekdays, however the standard deviation is the other way around.

Now, lets define our hypothesis testing:

- $H_0$ : $\mu_1$ - $\mu_2 \leq 0$ (There $\mu$ value of O3 on weekends is equal or less than the value on weekdays).
- $H_1$ : $\mu_1$ - $\mu_2 > 0$ (There $\mu$ value of O3 on weekends is greater than the value on weekdays).

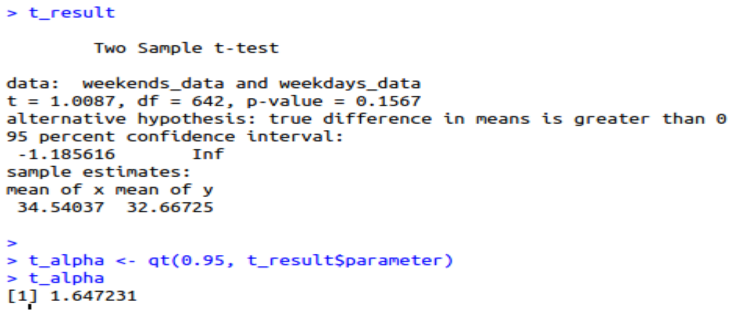
Assume that the confidence interval is 95% ($\alpha = 0.05$).
Rejection region: Reject $H_0$ if $t_{test} > t_\alpha$.
Assumption: Variance is equal.

Lets perform statistic test using R:

```
library("dplyr")
setwd("~/Assignment1")
df <- read.csv("data1(2).csv", header = TRUE)
df[, "date"] = lapply(df["date"],function(x){strptime(x, "%d/%m/%y")})

weekdays_data <- df %>% select(weekday,month, O3) %>%
filter(weekday %in% c(2:6) & month ==1) %>% select(O3)

weekends_data <- df %>% select(weekday,month, O3) %>%
filter(weekday %in% c(1,7) & month ==1) %>% select(O3)
```

```
t_result <- t.test(weekends_data, weekdays_data, alternative = "greater",
conf.level = 0.95, var.equal = TRUE, paired = FALSE)

t_result

t_alpha <- qt(0.95, t_result$parameter)
t_alpha
```

The result is:



```
> t_result

        Two Sample t-test

data:  weekends_data and weekdays_data
t = 1.0087, df = 642, p-value = 0.1567
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.185616       Inf
sample estimates:
mean of x mean of y
 34.54037  32.66725

>
> t_alpha <- qt(0.95, t_result$parameter)
> t_alpha
[1] 1.647231
```

Figure 5: The two sample test of O3 on weekdays and weekends in January.

From figure 5 we can see that:

- The $t_{test}$ is 1.0087. This value tells us the result of statistic test.
- The degree of freedom is 642.
- The p-value is 0.1567
- The $t_{\alpha}$ is 1.647231.
- The rejection region is $t > 1.647231$.
- the average of O3 on weekends is 34.54037 and on weekdays is 32.66725.

From the above results, we can draw a conclusion that we accept $H_0$ since the $t_{test} \leq t_{\alpha}$, $1.0087 \leq 1.647231$.

In the comparison to the statistic test in question 3.A, the value of $t_{test}$ of the January data is lower than the data from the entire year. However, when it comes to p-value, the result is the other way around. Moreover, the value of the 95% confidence interval of January data is greater than the value of the whole year.

This phenomenon might be caused by the different size of each samples. The difference size in between two samples leads to different degree of freedom which influences the statistic test.

(C) Test by hand:

Given:
$\overline{x_1} = 34.54037$
$\overline{x_2} = 32.66725$
$n_1 = 164$
$n_2 = 480$
Here are the hypothesis:

- $H_0$ : $\mu_1$ - $\mu_2 \leq 0$ (There $\mu$ value of O3 on weekends is equal or less than the value on weekdays).
- $H_1$ : $\mu_1$ - $\mu_2 > 0$ (There $\mu$ value of O3 on weekends is greater than the value on weekdays).

Confidence Interval = 95% ($\alpha = 0.05$)= $t_\alpha$=1.645
Assume standard deviation (s) is equal = 20.55985
Rejection region: Reject $H_0$ if $t_{test}$<$t_\alpha$

Statistic Test:

- df = $n_1$+$n_2$-2 = 164+480-2= 642
- $S_{x_1-x_2} = \sqrt{\frac{(n_1-1)s^2+(n_2-1)s^2}{n_1+n_2-2}}$
  $= \sqrt{\frac{(480-1)20.55985^2+(164-1)20.55985^2}{164+480-2}}$
  = 20.55985
- $t_{test} = \frac{(\overline{x_1}-\overline{x_1})-D_0}{S_{x1-x2}\sqrt{\frac{1}{n1}+\frac{1}{n2}}}$
  $= \frac{(34.54037-32.66725)-0}{20.55985\sqrt{\frac{1}{164}+\frac{1}{480}}}$
  = 1.007268536

Since the value of $t_{test} \leq t_\alpha$, therefore we accept $H_0$, 1.007268536 $\leq$ 1.645.

# 4 Question 4

(A) For this question **we should use** paired t-test because:

- In terms of the sample size, both of them are equal.
- The data tells that the re-exam can only be done after the exam. So they are related to each other.
- This observation measures the same sample before and after an event.
- They are dependent

We **can't apply** paired t-test when:

- The events are not dependent
- It is collected from two different and independent subjects
- The size between the two samples is not equal

(B) **Paired t-test**
Before we calculate the statistic test, first we have to define the hypothesis:

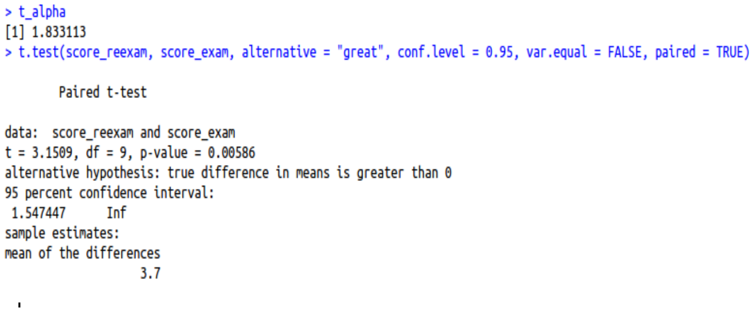
- $H_0 : \mu_d \leq 0$ (There is not any significant improvement of the score after re-exam)
- $H_1 : \mu_d > 0$ (There is significant improvement of the score after re-exam)

The confidence interval is 95% ($\alpha$=0.05)
df = 9
$t_\alpha$ = 1.833
Rejection region: Reject $H_0$ if $t_{test}$>$t_\alpha$

Using below R-codes, we can find the $t_{test}$:

```
library("dplyr")
score_reexam = c(45,39,10,25,15,49,30,32,22,41)
score_exam = c(39,35,13,22,16,41,27,25,20,33)
df = 9
confidence_interval = 0.95

t_alpha = qt(confidence_interval, df)
t_alpha
t.test(score_reexam, score_exam, alternative = "greater",
conf.level = 0.95, var.equal = FALSE, paired = TRUE)
```

The result of the t-test is:



```
> t_alpha
[1] 1.833113
> t.test(score_reexam, score_exam, alternative = "great", conf.level = 0.95, var.equal = FALSE, paired = TRUE)

	Paired t-test

data:  score_reexam and score_exam
t = 3.1509, df = 9, p-value = 0.00586
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.547447      Inf
sample estimates:
mean of the differences
                    3.7
```

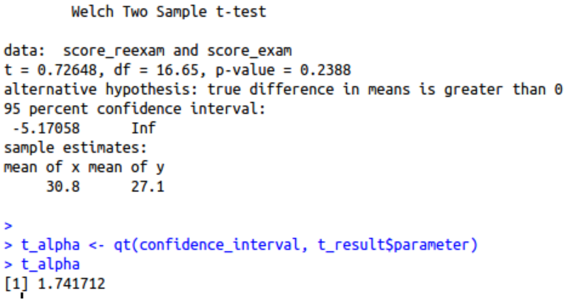
Figure 6: The paired statistic test, score re-exam vs exam.

From the above result it can be seen that the value of $t_{test} > t_\alpha$, (3.1509 > 1.833) therefore we reject $H_0$.

(C) **Unpaired t-test**
Given the hypothesis explained in 4b.
In order to get the $t_{test}$, we should execute the below R-code for unpaired data:

```
library("dplyr")
score_reexam = c(45,39,10,25,15,49,30,32,22,41)
score_exam = c(39,35,13,22,16,41,27,25,20,33)
confidence_interval = 0.95

t_result <- t.test(score_reexam, score_exam, alternative = "greater",
conf.level = 0.95, var.equal = FALSE, paired = FALSE)
t_result

t_alpha <- qt(confidence_interval, t_result$parameter)
t_alpha
```

The result of the t-test:



```
	Welch Two Sample t-test

data:  score_reexam and score_exam
t = 0.72648, df = 16.65, p-value = 0.2388
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -5.17058      Inf
sample estimates:
mean of x mean of y
     30.8      27.1

>
> t_alpha <- qt(confidence_interval, t_result$parameter)
> t_alpha
[1] 1.741712
```

Figure 7: The unpaired statistic test, score re-exam vs exam.

From the above result it can be seen that the value of $t_{test} \leq t_\alpha$, (0.72640 $\leq$ 1.741712) therefore we accept $H_0$. In comparison to question 4b, this statistic result might influence the administrator decision.

# 5 Question 5

(A) Given the mean and the standard deviation of O3 for the entire year during the weekdays and the weekends as described in figure 2.

$n_1 = 5118$<br>
$n_2 = 2075$

Before we calculate the statistic test, first lets define our hypothesis testing:

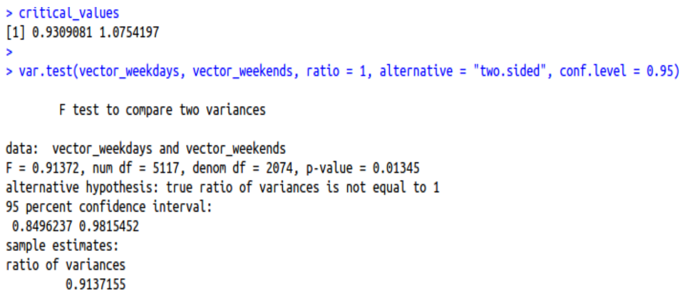
- $H_0 : \sigma_1^2 - \sigma_2^2 = 0$ (O3 variances on weekends and on weekdays are equal).
- $H_1 : \sigma_1^2 - \sigma_2^2 \neq 0$ (O3 variances on weekends and on weekdays are not equal).

Confidence level = 95% ($\alpha = 0.05$)<br>
Rejection Region: Reject $H_0$ if $F_{test} <= F_{\alpha/2,n_1-1,n_2-1}$ or $F_{test} >= F_{1-\alpha/2,n_1-1,n_2-1}$

The R-code to perform the statistic test:

```
library("dplyr")
setwd("~/Assignment1")
df <- read.csv("data1(2).csv", header = TRUE)
df[, "date"] = lapply(df["date"],function(x){strptime(x, "%d/%m/%y")})

weekdays_data <- df %>% select(weekday, O3) %>%
filter(weekday %in% c(2:6)) %>% select(O3)

weekends_data <- df %>% select(weekday, O3) %>%
filter(weekday %in% c(1,7)) %>% select(O3)

vector_weekdays <- unlist(weekdays_data['O3'])
vector_weekends <- unlist(weekends_data['O3'])

critical_values <- qf(c(0.025,0.975),5118,2075)
critical_values

var.test(vector_weekdays, vector_weekends, ratio = 1,
alternative = "two.sided", conf.level = 0.95)
```

The result of statistic test:



```
> critical_values
[1] 0.9309081 1.0754197
>
> var.test(vector_weekdays, vector_weekends, ratio = 1, alternative = "two.sided", conf.level = 0.95)

        F test to compare two variances

data:  vector_weekdays and vector_weekends
F = 0.91372, num df = 5117, denom df = 2074, p-value = 0.01345
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8496237 0.9815452
sample estimates:
ratio of variances
         0.9137155
```

Figure 8: The F statistic test of O3 on weekdays and weekends for the entire year.

From figure 8 we can see:

- The $F_{test}$ is 0.91372. This value tells us the statistic test.
- The degree of freedom is 5117 and the denominator degree of freedom is 2074.
- The p-value is 0.01345.
- The ratio of variances is 0.9137155.

Since the value of F$_{test} \leq$ F$_{\alpha/2}$, ($0.91372 \leq 0.9309081$) therefore we reject the null hypothesis.

(B) Test by hand for January data.
Given the mean and the standard deviation of O3 in January during the weekdays and the weekends as described in figure 4.

n$_1$ = 480
n$_2$ = 164
Now, lets define our hypothesis:

- H$_0$ : $\sigma_1^2$ - $\sigma_2^2$ = 0 (O3 variances on weekends and on weekdays are equal).
- H$_1$ : $\sigma_1^2$ - $\sigma_2^2 \neq 0$ (O3 variances on weekends and on weekdays are not equal).

Confidence level = 95% ($\alpha$ = 0.05)
Rejection Region: Reject F$_{test}$ <= F$_{\alpha/2,n_1-1,n_2-1}$ or F$_{test}$ >= F$_{1-\alpha/2,n_1-1,n_2-1}$

Now, lets calculate the statistic test:
F$_{test}$ = $\frac{s_1^2}{s_2^2}$
F$_{test}$= $\frac{20.55985^2}{20.44157^2}$
F$_{test}$= 1.0116059776

In order to get the value of F table, we must execute the below R-code:

```
qf(c(0.025,0.975),480,164)
```

Below are the F values returned by R:

- $\alpha/2$ = 0.7838436
- 1-$\alpha/2$ = 1.2959456

Since the value of F$_{test}$ < 1.2959456 and F$_{test}$ > 0.7838436 therefore we accept the null hypothesis.

# References