Statistical Consulting Project

Project Title: A Device to Reduce Engine Emissions

Report prepared for Feifang Hu

By Xinning Chu

Statistical Consulting Program (SCP)

February 21, 2019

Summary

This article is about a statistical analysis on a device produced by the manufacturer. The objective of this project is to determine whether the device has the effect of reducing engine emissions when placed on a motor vehicle engine.

Some statistical procedures are used to analyze the data. Analysis of variances is used to detect the effect of some factors, and two-sample t-test is used to compare the population means of two samples. In addition, hotelling's T^2 test is performed because correlation between the responses is accounted for in this method.

According to the results, the p-values are larger than 0.05, so we cannot reject the null hypothesis that the sample means of groups. There is no statistically significant evidence that can prove the device has the effect to change these three kinds of emissions.

Finally, it can be concluded that the device makes no statistically significant difference to emissions. The device is not as useful as the company claimed.

1. Introduction:

The aim of this study was to investigate whether a device produced by the manufacturer would reduce the emissions when placed on a motor vehicle engine. As the manufacturer claimed, after using the device hydrocarbon and carbon monoxide (CO) emissions are supposed to be steadily decreased, while carbon dioxide (CO2) emissions should be steadily increased.

The main approach in this project is analysis of variances (ANOVA). But before this, it's necessary to do some tests to check if the data satisfies assumptions of one-way ANOVA. The emissions are probably dependent on other factors, such as day, replicates and spark plugs. The impact of these factors should also be considered.

2. Background and Dataset:

With the development of modern society, people in growing numbers are beginning to pay attention to engine emissions. In order to increase car sales and achieve a good reputation, some vehicle manufacturers often "claim" their product to reduce the emissions will achieve a certain result (if used as directed), based on laboratory tests. This may be due to conflicting results from independent tests, or there may be insufficient statistical evidence to support the manufacturer's claim.

2.1 Study Design:

The analysis was focused on using the car owned by the manufacturer. We operated the vehicle under the same conditions on 13 different days and performed 4 replicates of the experiment on each of the testing days, so the total number of observations is 52. The device was placed on the engine from the third day to the eleventh day. In the first two days and the last two days, there was no device on the engine. It took nearly six months to complete the tests.

2.2 Variables:

In this investigation, three emissions were observed for the manufacturer's car. Among them the unit of hydrocarbon emission is ppm, which is the abbreviation of parts per million. Other factors recorded were DAY, REP and DEV. Details concerning these variables are summarized below:

Table 1 Variables in the study

Variable	Definition	Type
HC	Hydrocarbon (ppm)	Quantitative Measures
CO	Carbon Monoxide (% of the Volume)	Quantitative Measures
CO2	Carbon Dioxide (% of the Volume)	Quantitative Measures
DAY	Test Day: 1, 2, … ,13	Categorical Factors
REP	Replicate: 1, 2, 3, 4	Categorical Factors
DEV	Device: 1/0 (Present/Absent)	Categorical Factors

3. Data Analysis and Results:

A statistical analysis of this project was performed using the statistical software R. Four statistical procedures were used in this analysis: exploratory data analysis (EDA) was used to summarized the data, two-sample t-tests were used to detect differences between groups, analysis of variance (ANOVA) was used to detect significant factor effects and Hotelling's T^2 test was used to test all three emissions simultaneously. Details about these approaches are briefly given below.

3.1 Exploratory Data Analysis:

EDA techniques are used to summarize the data. Table 2A and Table 2B presents some summary statistics for the HC, CO and CO2 variables overall, and with respect to DAY. Since the standard deviations is relatively large, so the accuracy of the measuring instrument has little impact on the statistical analysis.

Table 2A Mean Emissions for Manufacturer's Car

Device	Day	HC (ppm)	CO2 (%)	CO (%)
Absent	1	15.30	15.83	0.25
	2	10.75	14.41	0.11
Present	3	5.46	14.85	0.11
	4	6.04	13.48	0.16
	5	8.66	14.62	0.08
	6	13.18	13.38	0.13
	7	25.89	13.75	0.21
	8	4.45	15.95	0.20
	9	13.18	15.52	0.10
	10	1.11	15.42	0.11
	11	12.71	14.85	0.14
Absent	12	11.31	15.61	0.13
	13	13.54	14.24	0.15
Absent	Avg	12.72	15.02	0.16
Present	Avg	10.07	14.65	0.15

Table 2B Summary Statistics of Emissions

Total Observation Number = 52

	HC (ppm)	CO2 (%)	CO (%)
Min	0.170	11.14	0.0000
1st Qu.	5.438	13.94	0.0800
Median	10.130	14.75	0.1400
Mean	10.891	14.76	0.1460
3rd Qu.	13.932	15.54	0.2025
Max	29.440	17.11	0.3200
Std. Dev	6.9312	1.2472	0.0822

3.2 ANOVA:

The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values.

To test if the device has the effect to reduce engine emissions, one-way ANOVA can be performed. The data would be separated into 6 groups: HC(DEV=0), HC(DEV=1), CO(DEV=0), CO(DEV=1), CO2(DEV=0) and CO2(DEV=1). However, there are three main assumptions that the test makes. Firstly,

the dependent variable is normally distributed in each group that is being compared in the one-way ANOVA. Secondly, there is homogeneity of variance, which means that the population variance in each group are equal. Thirdly, observations are independent.

Thus, before performing ANOVA, we did Shapiro-Wilk normality test six times to check if the data of each group is normally distributed. Similarly, Bartlett's test was used three times to check if the variance of two groups that should be compared is same. For the third assumption, it's necessary to determine that our observations are not independent on the study design (e.g. day/rep), so we can use ANOVA to determine whether the DAY and REP are important.

The results of normality tests and variance tests are presented in Table 3A and Table 3B. It was found that these assumptions are true, except that one group of hydrocarbon seemed not to be normal. However, the one-way ANOVA is considered a robust test against the normality assumption, which means that it tolerates violations to its normality assumption rather well.

Table 3A Results of Shapiro-Wilk normality Tests

	DEV = 0	DEV = 1
HC	W = 0.95 p-value = 0.47	W = 0.90, p-value = 0.0038
CO2	W = 0.97, p-value = 0.88	W = 0.95, p-value = 0.12
CO	W = 0.93, p-value = 0.28	W = 0.98, p-value = 0.88

Table 3B Results of Bartlett's Tests

HC	Bartlett's K-squared = 6.47, p-value = 0.01096
CO2	Bartlett's K-squared = 0.033, p-value = 0.856
CO	Bartlett's K-squared = 1.70, p-value = 0.1918

The following may have an impact on the measures. The test days were not necessarily contiguous, or the car was driven for approximately 700 to 2,000 miles between tests. In addition, testing began in November and was completed the following May. But from the results of analysis of the effects of DAY and REP using ANOVA shown in the Table 4, DAY and REP have no statistically significant effect since p-values are larger than 0.05. In other words, when we calculate the ANOVA using DEV as the factor, it can be considered as a completely randomized design. Perhaps some factors can affect our measures, but the impact is very insignificant. Hence, ANOVA is an efficient way.

Table 4 Results of ANOVA with DAY and REP

CO1 ~ DAY	Df Sum Sq Mean Sq F value Pr(>F) 1 0.0028 0.0028 0.417 0.522 50 0.3418 0.0068
CO1 ~ REP	Df Sum Sq Mean Sq F value Pr(>F) 1 0.0004 0.0004 0.061 0.086 50 0.3442 0.0069
CO2 ~ DAY	Df Sum Sq Mean Sq F value Pr(>F) 1 0.9800 0.9807 0.626 0.433 50 78.35 1.5670

CO2 ~ REP	Df Sum Sq Mean Sq F value Pr(>F) 1 1.50 1.497 0.962 0.331 50 77.83 1.557
HC ~ DAY	Df Sum Sq Mean Sq F value Pr(>F) 1 1 1.02 0.021 0.886 50 2449 48.98
HC ~ REP	Df Sum Sq Mean Sq F value Pr(>F) 1 43.6 43.65 0.907 0.346 50 2406.5 48.13

Table 5 shows the results of ANOVA to check the effect of DEV respectively to each emission. Since all the p-values are greater than 0.05, we cannot reject the null hypothesis. The device has no statistically significant effect to affect three emissions, which is not as same as the company claimed.

Table 5 Results of ANOVA with DEV

HC ~ DEV	Df Sum Sq Mean Sq F value Pr(>F) 1 77.6 77.61 1.636 0.207 50 2372.5 47.45
CO ~ DEV	Df Sum Sq Mean Sq F value Pr(>F) 1 0.0050 0.00497 0.731 0.396 50 0.3397 0.006794
CO2~ DEV	Df Sum Sq Mean Sq F value Pr(>F) 1 1.55 1.546 0.993 0.324 50 77.79 1.556

3.3 Two-sample t-tests:

To investigate the manufacturer's claim that once the device is placed on an engine, the engine is irrevocably changed, we can do a two-sample t-test for equality of mean emissions using the data from Days 1 and 2, versus those of Days 12 and T13. Then we can find if there is evidence to support this claim. In addition, we should consider the fact that new spark plugs were placed on the car after Day 7, which may have an impact on the experimental results. So another two-sample t-test is performed for equality of mean emissions using the data from Day 3, 4, 5, 6 and 7, versus those of Day 8, 9, 10 and 11.

The results of two-sample t-tests are shown in Table 6A and Table 6B. Since p-values are greater than 0.05, there is no statistically significant difference between the emissions of the first two days and the last two days. The claim made by the manufacturer that the device could reduce the emissions once placed on the engine can be rejected. Also, from Table 5B the effect of changing spark plugs is not statistically significant except CO2.

Table 6A Two-sample t-tests for DAY 1 ~ 2 and DAY 12 ~ 13

	t	p-value	Mean in group 1 / 2
HC	0.28	0.79	13.02 / 12.42
CO	0.67	0.51	0.18 / 0.14
CO2	0.31	0.76	15.12 / 14.92

Table 6B Two-sample t-tests for old spark plugs and new spark plugs

	t	p-value
HC	1.56	0.13
CO	0.0049	0.99
CO2	-4.03	0.0003

3.4 Hotelling's T^2 test:

An alternative approach is to consider all three emissions simultaneously, then perform a two-sample Hotelling's T^2 test. The potential advantage of this approach is that the correlation between the responses is accounted for, which is expected according to the manufacturer's claim. Thus, we can see that if the multivariate approach makes any difference. Also, before Hotelling's T^2 test multivariate Shapiro-Wilk tests should be performed to check the normality. Table 7 shows the results of Hotelling's T^2 test. The p-value is 0.34, which is much larger than 0.05. The null hypothesis is rejected as well, so the result makes no difference. When consider all three emissions simultaneously, there is still no statistically significant effect of the device to emissions.

Table 7 Result of Hotelling's two-sample T2-test

Hotelling's two sample T2-test	T2=1.1363	df1 = 3, df2 = 48	p-value=0.3439

4. Conclusions

Firstly, the SCP found that there were no significant results for the variables and REP, which suggests that the days and replicates did not differ significantly with respect to these variables. Then, there was a no significant DEV effect for the variables HC, CO and CO2, suggesting that hydrocarbon and carbon monoxide emissions cannot be steadily decreased and carbon dioxide emissions cannot be steadily increased after the device was placed on the engine. The certain results claimed by the manufacturer was not achieved by us. Therefore, a conclusion can be safely drawn that the claim made by the manufacturer should be rejected because there was no statistically significant difference that the device made to the emissions.

References:

[1]https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide-4.php

[2]https://www.statmethods.net/stats/anova.html

Appendix 1: R codes

```
setwd("C:/Users/derri/Onedrive/r")
data <- read.csv('emission.csv',header=TRUE)
x=data$hc
y=data$co1
z=data$co2
summary(x)
summary(y)
summary(z)

model11 <- aov(co1 ~ day, data=data)
model12 <- aov(co1 ~ rep, data=data)
summary(model11)
summary(model12)
model21 <- aov(co2 ~ day, data=data)
model22 <- aov(co2 ~ rep, data=data)
summary(model21)
summary(model22)
model31 <- aov(hc ~ day, data=data)
model32 <- aov(hc ~ rep, data=data)
summary(model31)
summary(model32)

modela <- aov(hc ~ dev, data=data)
modelb <- aov(co1 ~ dev, data=data)
modelc <- aov(co2 ~ dev, data=data)
summary(modela)
summary(modelb)
summary(modelc)
```

```
hc_a <- c(8.24,20.2,12.48,20.27,7.73,13.1,12.13,10.03,18.08,11.16,10.76,5.22,13.63,14,14.04,12.5)
shapiro.test(hc_a)

ee <-as.matrix(data)
dd <- as.vector(ee)

data <- read.table('1.txt',header=TRUE)
ee <-as.matrix(data)
co2_a <- as.vector(ee)
co2_a
shapiro.test(co2_a)

data <- read.table('2.txt',header=TRUE)
ee <-as.matrix(data)
co2_b <- as.vector(ee)
co2_b
shapiro.test(co2_b)

data <- read.table('3.txt',header=TRUE)
ee <-as.matrix(data)
co1_a <- as.vector(ee)
co1_a
shapiro.test(co1_a)

data <- read.table('4.txt',header=TRUE)
ee <-as.matrix(data)
co1_b <- as.vector(ee)
co1_b
shapiro.test(co1_b)
```

```
data <- read.csv('emission.csv',header=TRUE)
bartlett.test(hc~dev, data)
bartlett.test(co2~dev, data)
bartlett.test(co1~dev, data)
t.test(hc ~ co3, var.equal = TRUE, data)
t.test(co1 ~ co3, var.equal = TRUE, data)
t.test(co2 ~ co3, var.equal = TRUE, data)

data <- read.csv('4.csv',header=TRUE)
attach(data)
t.test(hc ~ co3, var.equal = TRUE, data)
t.test(co1 ~ co3, var.equal = TRUE, data)
t.test(co2 ~ co3, var.equal = TRUE, data)
data <- read.csv('multi.csv',header=TRUE)
x1=data$hca
x2=data$co1a
x3=data$co2a
x=cbind(x1,x2,x3)

install.packages("mvnormtest")
library(mvnormtest)
mshapiro.test(t(x))

data <- read.csv('multi2.csv',header=TRUE)
y1=data$hcb
y2=data$co1b
y3=data$co2b
y=cbind(y1,y2,y3)
mshapiro.test(t(y))
```

```
install.packages("ICSNP")
library(ICSNP)
HotellingsT2(x,y)
```

Appendix 2:

Table 8 Results of Multivariate Shapiro-Wilk normality test

	W	p-value
The group with device	0.97011	0.8404
The group without device	0.95308	0.1307