Statistical Analysis on Shuttle Arrival Timings using R and SAS

Author: Vimalesh Raja Karupiah Ramachandran

Date12/15/2022

Introduction

The aim of the project is to find the arrival time of the shuttle to the bus stop and find the drivers efficiency, punctuality and the skill driving the shuttle. Here the population are the individual driver's time of arrival to the bus stop. The variables used in the problem will be unique driver number and the time of the shuttle by breaks.

This study helps in assessing the driver's punctuality and indirectly the kindness towards the students who are the most frequent travelers in the shuttle throughout the campus. This study also describes the procedure of two way Anova which compares the significant difference in the means of the groups. Moreover this study helps the students to prepare themselves when to be at the bus stop so that they don't miss the shuttle.

This is an interesting problem because most of the frequent travelers in the shuttle are the students who are busy all the day and want to travel around the campus every time. So this study might help plan their day so that they can reduce the pain of walking missing the shuttle. Also this study helps the transportation service to check the skill, punctuality and the efficiency of the driver driving the shuttle by analyzing the time they reach particular bus stop.

Methods

In this section we will see about the process of data collection, description of the data collected and samples of the data.

Since the time cannot be controlled and there are many lot of other variable like traffic, weather, travelers population which influence the arrival of the shuttle to a bus stop, this experiment is considered to be an observational study. This means the data is collected just by observing the happening of surroundings without controlling the nature.

Moving on to the sampling part, the data is collected at a random bus stop at a particular break periods. The shuttle timings are from morning 8.30 to evening 10. Therefore the 6 samples are taken on each day. The samples can be divided as 3 among two drivers(D1,D2). There are breaks given to the drivers during their shift and generally it is a 30 minutes break. Therefore the samples are taken from breaks considering that if the shuttle is running after a break or middle of two breaks or before a break that is going to happen. There are 3 categorical variable which are considered here (after break, before break, middle break).

By doing this sampling method we can find the efficiency of the two drivers who are driving the shuttle. This is because we are having the 1 sample of each break and 3 samples of each driver for over 5 days which are weekdays.

The weekdays were only considered because most number of travelers are active during the weekdays and the students need this study most during their working days. The shuttle also doesn't run during the weekends this also can be considered as a reason that the samples were taken only during the weekdays.

The response variable in our data is time the bus arrive at the bust stop and the independent variable here is the driver and the breaks period.

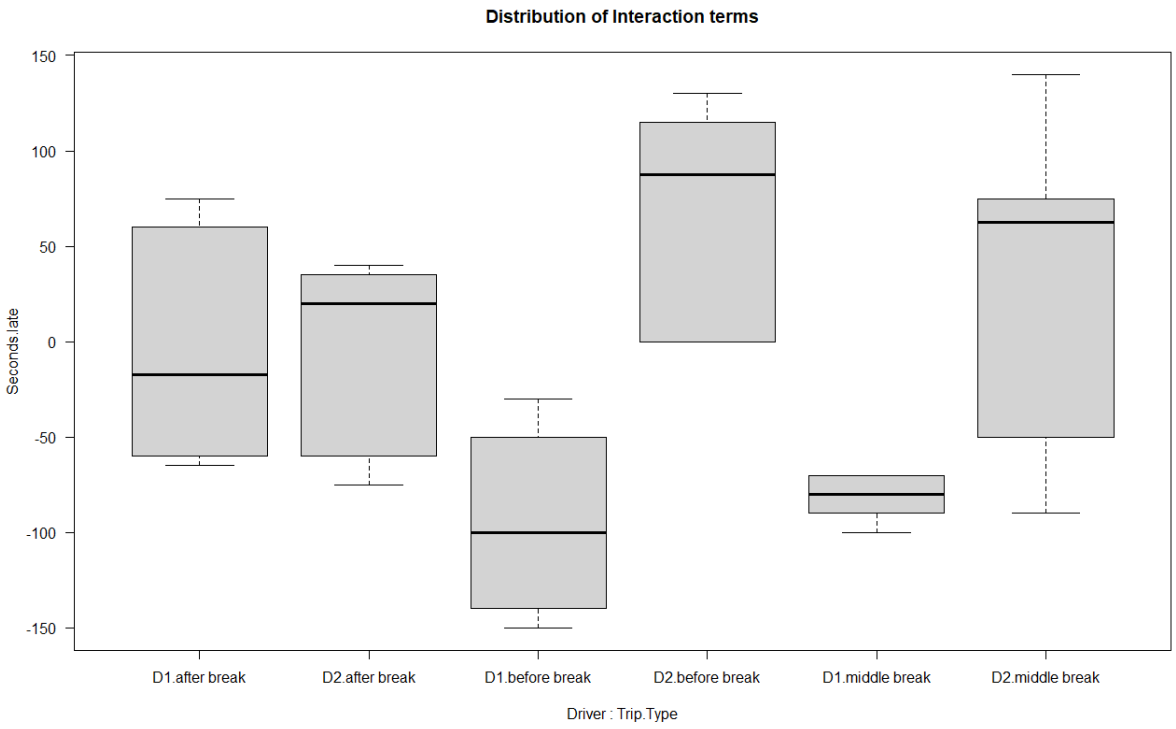
Husky Campus Shuttle					
Monday through Friday 7:30 AM - 5:05 PM					
MUB	**SDC**	**Upper DH**	**Lower DH**	**WADS**	**Library**
7:30 AM	7:35 AM	7:38 AM	7:40 AM	7:41 AM	7:44 AM
7:50 AM	7:55 AM	7:58 AM	8:00 AM	8:01 AM	8:04 AM
8:10 AM	8:15 AM	8:18 AM	8: 20AM	8:21 AM	8:24 AM
8:30 AM	8:35 AM	8:38 AM	8:40 AM	8:41 AM	8:44 AM
8:50 AM	8:55 AM	8:58 AM	9:00 AM	9:01 AM	9:04 AM
BREAK					
9:30 AM	9:35 AM	9:38 AM	9:40 AM	9:41 AM	9:44 AM
9:50 AM	9:55 AM	9:58 AM	10:00 AM	10:01 AM	10:04 AM
10:10 AM	10:15 AM	10:18 AM	10:20 AM	10:21 AM	10:24 AM
10:30 AM	10:35 AM	10:38 AM	10:40 AM	10:41 AM	10:44 AM
10:50 AM	10:55 AM	10:58 AM	11:00 AM	11:01 AM	11:04 AM
11:10 AM	11:15 AM	11:18 AM	11:20 AM	11:21 AM	11:24 AM
11:30 AM	11:35 AM	11:38 AM	11:40 AM	11:41 AM	11:44 AM
11:50 AM	11:55 AM	11:58 AM	12:00 PM	12:01 PM	12:04 PM
12:10 PM	12:15 PM	12:18 PM	12:20 PM	12:21 PM	12:24 PM
12:30 PM	12:35 PM	12:38 PM	12:40 PM	12:41 PM	12:44 PM
12:50 PM	12:55 PM	12:58 PM	1:00 PM	1:01 PM	1:04 PM
BREAK					
1:30 PM	1:35 PM	1:38 PM	1:40 PM	1:41 PM	1:44 PM
1:50 PM	1:55 PM	1:58 PM	2:00 PM	2:01 PM	2:04 PM
2:10 PM	2:15 PM	2:18 PM	2:20 PM	2:21 PM	2:24 PM
2:30 PM	2:35 PM	2:38 PM	2:40 PM	2:41 PM	2:44 PM
2:50 PM	2:55 PM	2:58 PM	3:00 PM	3:01 PM	3:04 PM
BREAK					
3:30 PM	3:35 PM	3:38 PM	3:40 PM	3:41 PM	3:44 PM
3:50 PM	3:55 PM	3:58 PM	4:00 PM	4:01 PM	4:04 PM
4:10 PM	4:15 PM	4:18 PM	4:20 PM	4:21 PM	4:24 PM
4:30 PM	4:35 PM	4:38 PM	4:40 PM	4:41 PM	4:44 PM
4:50 PM	4:55 PM	4:58 PM	5:00 PM	5:01 PM	5:04 PM
Drop Off Only					

Results

In this section the results of the analysis can be found. To begin the response variable is the time and the independent variable is driver id and the break period.

Data Visualization:

Boxplot:



This plot shows the distribution of time along the different groups of interaction terms. There seem to be no outliers in the dataset but the last group seems to be skewed heavily.

As we can see that the Driver 1 before break will always leave early. Driver 2 before breake has either on time or late by average of 100 seconds.

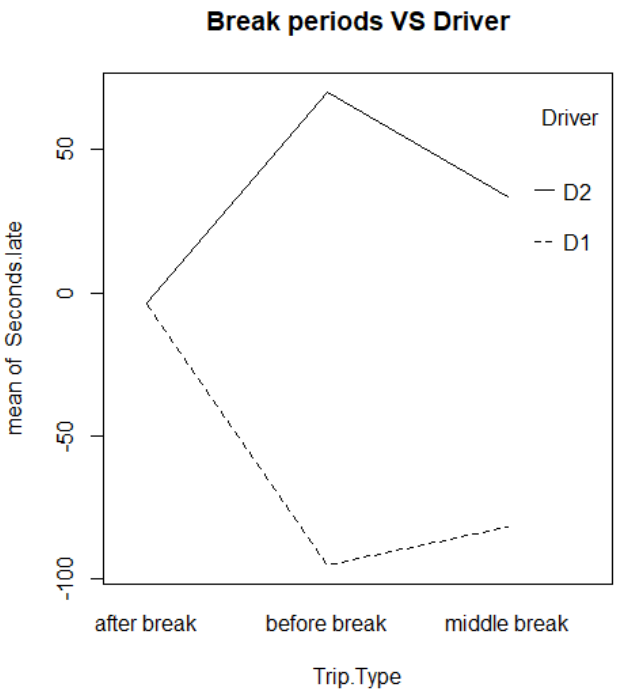
	Driver	Trip.Type	Seconds.late.Mean	Seconds.late.SD	Seconds.late.Min	Seconds.late.0.25
1	D1	after break	-4.166667	60.284050	-65.000000	-53.750000
2	D2	after break	-3.333333	50.563491	-75.000000	-40.000000
3	D1	before break	-95.000000	53.944416	-150.000000	-140.000000
4	D2	before break	70.000000	56.833089	0.000000	20.000000
5	D1	middle break	-81.666667	12.110601	-100.000000	-88.750000
6	D2	middle break	33.333333	86.120071	-90.000000	-23.750000

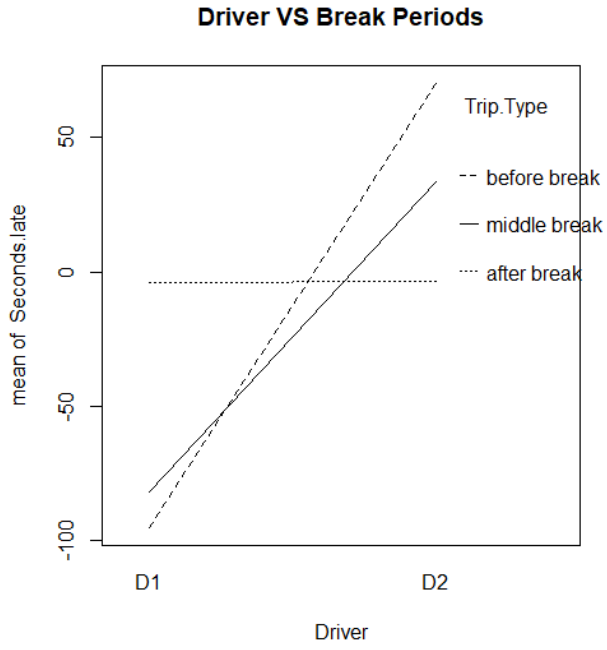
	Seconds.late.Median	Seconds.late.0.75	Seconds.late.Max
1	-17.500000	45.000000	75.000000
2	20.000000	31.250000	40.000000
3	-100.000000	-52.500000	-30.000000
4	87.500000	110.000000	130.000000
5	-80.000000	-71.250000	-70.000000
6	62.500000	73.750000	140.000000

In the above table we can see the summary statistics of the data by the group. This table explains the numbers which describes the boxplot numerically.

Since the interaction between the driver and the break period is to be found we go with the two way ANOVA test. Which seem to be appropriate in our case.

We can also see another graphical representation of this groups. Interaction plot which helps to compare the drivers and the break periods.





The Charts explain the interaction between the plots. In this case the left plot shows drivers take similar starts in the after break period.

ANOVA Results:

The 2-way ANOVA model have been applied to the data and the results are shown below.

```
Analysis of Variance Table

Response: Seconds.late
                 Df  Sum Sq Mean Sq F value    Pr(>F)
Driver            1   78867   78867 23.7861 3.296e-05 ***
Trip.Type         2    2518    1259  0.3797  0.687295
Driver:Trip.Type  2   42485   21242  6.4066  0.004821 **
Residuals        30   99471    3316
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above table we can see that the p-values is significantly less than significance value which means there is a significant difference between the mean value of the Driver and the interaction term. Also the trip.type has insignificant effect on the arrival time.

(alpha/3 taken as significance level which is 0.05/3 =0.017)

SAS Output:

The GLM Procedure

Dependent Variable: Secondslate

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	123870.1389	24774.0278	7.47	0.0001
Error	30	99470.8333	3315.6944		
Corrected Total	35	223340.9722			

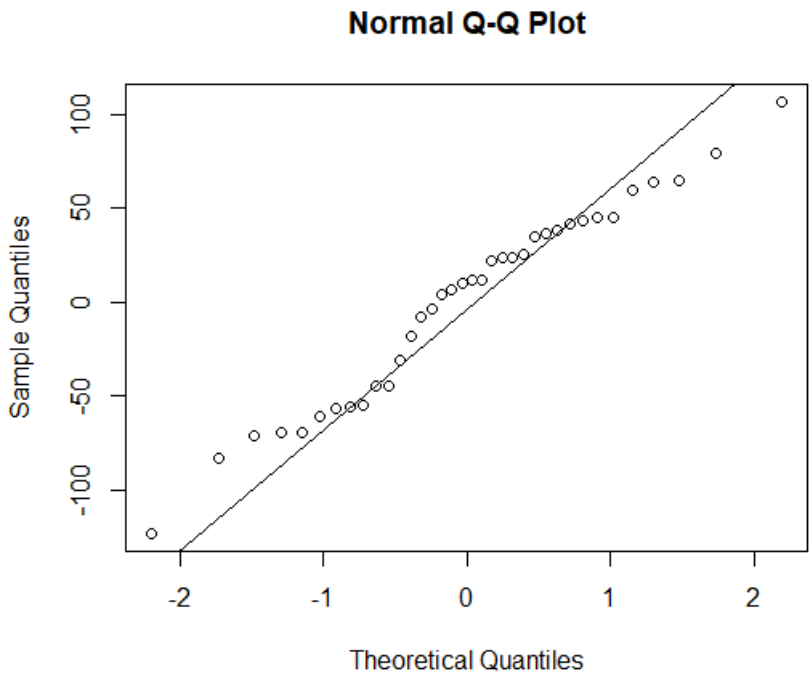
R-Square	Coeff Var	Root MSE	Secondslate Mean
0.554623	-427.4133	57.58207	-13.47222

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Driver	1	78867.36111	78867.36111	23.79	<.0001
TripType	2	2518.05556	1259.02778	0.38	0.6873
Driver*TripType	2	42484.72222	21242.36111	6.41	0.0048

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Driver	1	78867.36111	78867.36111	23.79	<.0001
TripType	2	2518.05556	1259.02778	0.38	0.6873
Driver*TripType	2	42484.72222	21242.36111	6.41	0.0048

Normality assumption:

For evaluating the normality q-q plot is drawn and Shapiro-Wilk test is done.



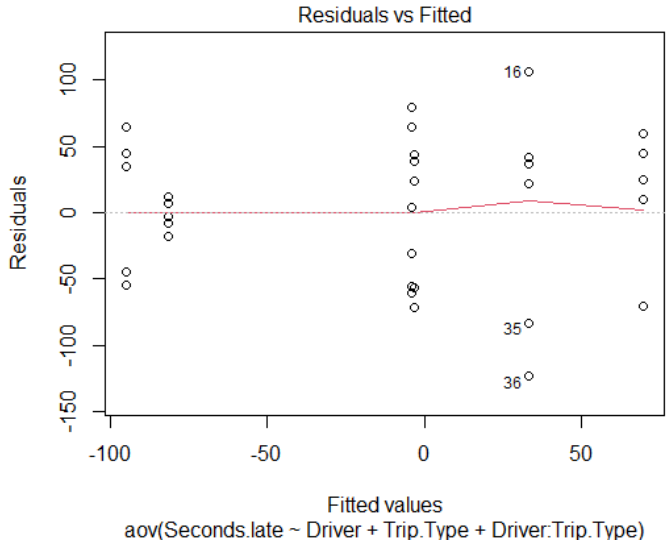
```
        Shapiro-Wilk normality test

data:  res
W = 0.96793, p-value = 0.3717
```

The p-value is greater than significance value and hence the normality assumption is not violated.

Homoscedasticity Assumption:

For this assumption residual value is plotted against the fitted value and check for outliers. Seems to be not following the assumption. Non-parametric test to be done.



Post-Hoc Analysis:

Tukey test is conducted to find the best group.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Seconds.late ~ Driver + Trip.Type + Driver:Trip.Type)

$Driver
          diff      lwr      upr   p adj
D2-D1 93.61111 54.41169 132.8105 3.3e-05

$Trip.Type
                               diff       lwr      upr     p adj
before break-after break   -8.75000 -66.70306 49.20306 0.9266616
middle break-after break  -20.41667 -78.36972 37.53639 0.6639513
middle break-before break -11.66667 -69.61972 46.28639 0.8736204

$`Driver:Trip.Type`
                                          diff        lwr        upr     p adj
D2:after break-D1:after break       0.8333333 -100.28453 101.951201 1.0000000
D1:before break-D1:after break     -90.8333333 -191.95120  10.284534 0.0981113
D2:before break-D1:after break      74.1666667  -26.95120 175.284534 0.2539057
D1:middle break-D1:after break     -77.5000000 -178.61787  23.617868 0.2133605
D2:middle break-D1:after break      37.5000000  -63.61787 138.617868 0.8658496
D1:before break-D2:after break     -91.6666667 -192.78453   9.451201 0.0931011
D2:before break-D2:after break      73.3333333  -27.78453 174.451201 0.2648227
D1:middle break-D2:after break     -78.3333333 -179.45120  22.784534 0.2040018
D2:middle break-D2:after break      36.6666667  -64.45120 137.784534 0.8763000
D2:before break-D1:before break    165.0000000   63.88213 266.117868 0.0003435
D1:middle break-D1:before break     13.3333333  -87.78453 114.451201 0.9985228
D2:middle break-D1:before break    128.3333333   27.21547 229.451201 0.0067359
D1:middle break-D2:before break   -151.6666667 -252.78453 -50.548799 0.0010345
D2:middle break-D2:before break    -36.6666667 -137.78453  64.451201 0.8763000
D2:middle break-D1:middle break    115.0000000   13.88213 216.117868 0.0185867
```

There is significance difference in the drivers and the before break group, middle break group.

SAS Output:

The GLM Procedure

Tukey's Studentized Range (HSD) Test for Secondslate

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	30
Error Mean Square	3315.694
Critical Value of Studentized Range	2.88818
Minimum Significant Difference	39.199

Comparisons significant at the 0.05 level are indicated by ***.

Driver Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
D2 - D1	93.61	54.41	132.81	***
D1 - D2	-93.61	-132.81	-54.41	***

CONCLUSION:

To conclude there is significance difference in the driver and the driver 2 (D2) seems to be more kind in terms waiting for the students in the bus stop. The Driver 1 (D1) seems to be punctual and skilled in the drive in all kind of weather. This report helps the students in conclude that they can do their work and can come upto some seconds late to bus stop if Driver 2 is driving the bus and can reach the lectures early if Driver 1 is driving the bus.