Advanced Business Data Analysis (ABDA)

Higher Diploma in Data Analytics (HDSDA)

Student Name: Siobhan Purcell

Student Number: 18195342

Continuous Assignment 1

June 2019

Lecturer: Dr Giovanni Estrada

WILCOXON-SIGNED RANK SUM TEST (INDEPENDENT SAMPLES)

INTRODUCTION

The dataset reports the results of a clinical trial study from Freireich et al. (1963) which investigates the effects of a drug 6-mercaptopurine (6-MP) versus a placebo in children with acute leukemia. Patients were selected who had experienced complete or partial remission of their leukemia following prior treatment with the drug 6-MP. The purpose of this research is to evaluate the effect of 6-MP on the remission times of these patients to see if it could be used to prolong their survival. To conduct this trial study, a cohort of 42 patients were assigned into 21 pairs in accordance with their complete/partial remission status. Within each pair, one subject received 6-mercaptopurine (6-MP) while the other received placebo treatments as a control comparison. Although this trial was designed as matched patient pairs, this analysis will ignore the pairing and use the data to illustrate Wilcoxon's two-sample (non-paired) rank sum test.

To conduct this analysis, we will examine the difference in remission duration by months between children patients receiving 6-MP versus placebo. Patients were followed until their leukaemia returned or until the end of the study (Table 1). A link containing a full dataset description can be found in the references section

***Table 1:** Data shows the length of remission in months for two groups of leukemia patients, placebo and treated*

Time to relapse for Placebo Patients	Time to relapse for 6-MP Patients
1	10
22	7
3	32
12	23
8	22
17	6
2	16
11	34
8	32
12	25
2	11
5	20
4	19
15	6
8	17
23	35
5	6
11	13
4	9
1	6
8	10

NORMALITY ASSESSMENTS

As an initial brief assessment of the data, histograms and boxplots were created to explore the normality of time measurement distributions for each treatment population. The distributions of both populations appear to indicate signs non-normality, given the evident right skew in both histograms below (Figure 1), indicating that most patients relapse after a short period.

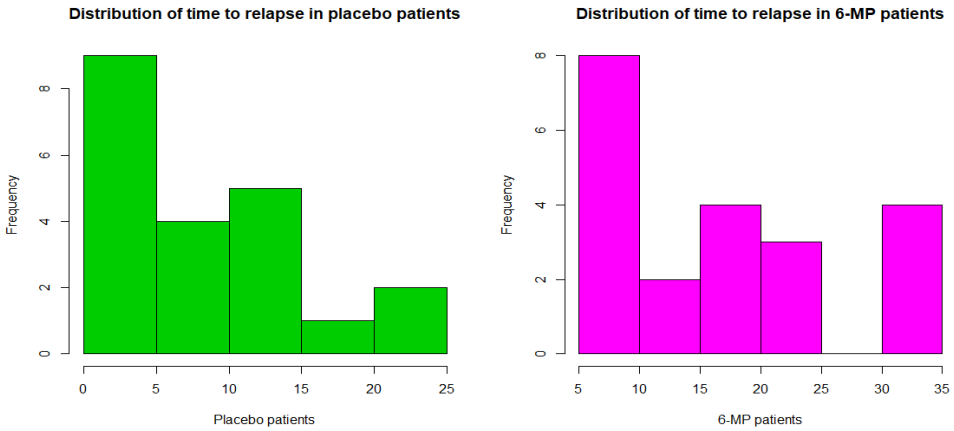


Figure 1: Histograms distributions of months to relapse for placebo and 6-MP treated patients

The irregularity of these distributions would suggest that there is considerable variability in the data which can be further represented by the use of boxplots as shown in figure 2 on the left. Whilst is also some degree of overlap between the interquartile ranges of the boxplots, the median line of the 6-MP drug treated population lies outside of the placebo treated boxplot, indicating that there is *likely* to be a difference between the two groups.

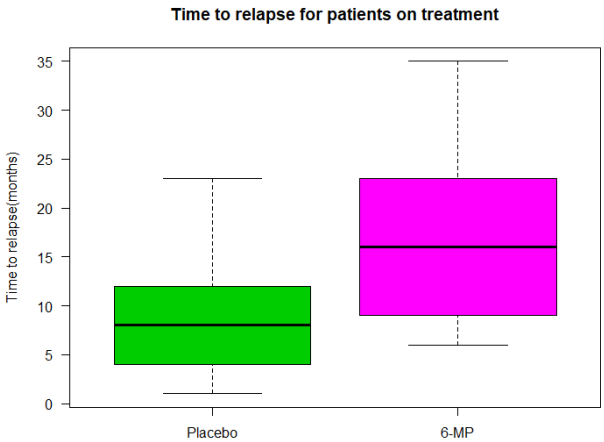
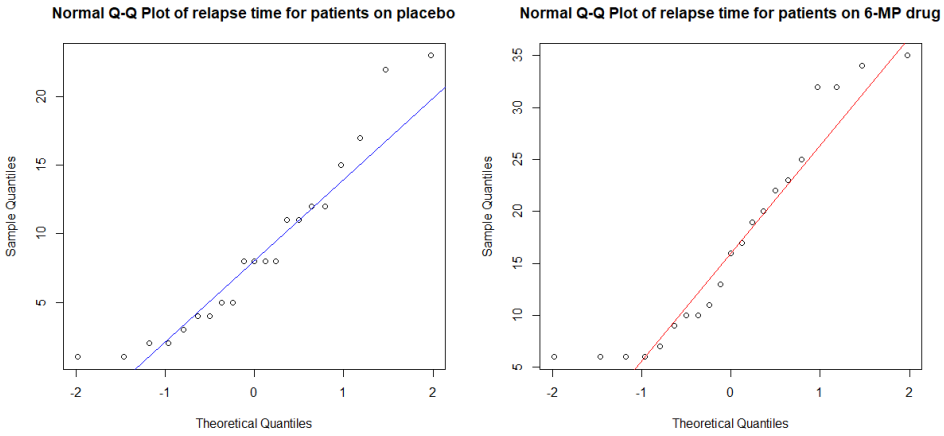


Figure 2: Boxplot distributions of time to relapse for plcaeboand 6-MP treated patients



A Q-Q plot was also applied to test the noramlity of these dsitrubtions. As illustrated in figure 3 on the left, the datapoints are not in a straight line. We observe a bit divergence from normality in the lower and upper tails tails of both treatment groups.

Figure 3: Q-Q plots demonstrating the divergence from noramlity for the distrutions of remission times in both patient populations

The Shapiro-Wilk test assumes a null hypothesis for a normal distribution. For this test, p-values were < 0.05 in the treated group (where p-value: 0.02618), thereby allowing the rejection of the null hypothesis and in favour of the alternative hypothesis that this distribution violate the assumption of normality.

```
> shapiro.test(drug6mp$t1)

        Shapiro-Wilk normality test

data:  drug6mp$t1
W = 0.91074, p-value = 0.0568
```

```
> shapiro.test(drug6mp$t2)

        Shapiro-Wilk normality test

data:  drug6mp$t2
W = 0.89345, p-value = 0.02618
```

WILCOXON SIGNED-RANK SUM TEST

Given that one of the two sample follow a non-normal distribution and are independent of each other, we will therefore conduct a non-parametric test known as Wilcoxon rank-sum test as an alternative to the parametric non-paired t-test for comparing the mean rank remission periods of the two treatment groups. Below are the null and alternate hypotheses:

H_0: The mean rank remission times of patients receiving placebo and patient receiving antileukemic test drug (6MP) are equal

H_A: The mean rank remission times of patients receiving placebo and patient receiving antileukemic test drug (6MP) are not equal

REPORT RESULTS

The remission duration between the placebo treated group and 6-MP treated group were remarkably different (medians: control =8 months vs treated =16 months). A Wilcoxon test showed that the observed differences in mean rank remission times between both treatment groups was statistically significant (N=42, Z= -2.859, p=0.004). Hence, these results provide strong evidence that the 6-MP drug could to be a viable maintenance therapy for children suffering from acute leukaemia by effectively prolonging their clinical remission periods.

```
> wilcox.test(drug6mp$t1, drug6mp$t2, paired =FALSE)

        Wilcoxon rank sum test with continuity correction

data:  drug6mp$t1 and drug6mp$t2
W = 106.5, p-value = 0.004248
alternative hypothesis: true location shift is not equal to 0

> qnorm(drug6mptest$p.value/2) #z=-2.859118, p =0.004
[1] -2.859118
```

IMPROVING STUDY DESIGN VIA POST HOC POWER ANALYSIS

Power analysis is an integral aspect of many experimental designs, particularly in clinical studies. It allows researchers to determine the sample size required to detect an effect of a given size with a given degree of confidence. An effect size can by calculating the difference between means of the two treatment groups and divided this by the average of their standard deviations. This was achieved using an R package called "effsize". Here, the effect size is 1.001, which would be considered a large effect according to Cohen's classification of effect size.

```
> cohen.d(Placebo, Drug6mp)

Cohen's d

d estimate: -1.000909 (large)
95 percent confidence interval:
     lower      upper
-1.6625276 -0.3392895
```

Taking the parameters of effect size (d) and sample size (n) into account, a two-sample t-test power analysis was performed to estimate the current power of the study at a significance level of 0.001. This was necessary step in order to determine how much of an improvement could be made to the power of the experiment by increasing its sample size. The results of the power calculation test reveal this study only had a 40% chance of finding a large effect size difference of d=1.001 at significance level of 0.001 with sample size that had been (n=21 per group). Therefore. it could be argued that this study was under-powered. However, an improvement could be made to the research design of this clinical study by requesting a larger sample size and in doing in so, achieve more power (95%) at a significance level of 0.001

```
> pwr.t.test(d=1.001, n=21, sig.level =0.001)

     Two-sample t test power calculation

              n = 21
              d = 1.001
      sig.level = 0.001
          power = 0.3957589
    alternative = two.sided

NOTE: n is number in *each* group
```

Using the study's previously calculated effect size, (Cohen's d =1.001), the minimum required sample size for achieving these specified measures was estimated. The results of the power calculation reveal that in order to appreciate a large difference between the placebo and 6MP treated subject with a power of 95% and p-value greater than an α of 0.001, we would need to accrue a cohort of n=51 patients in each sample group for future studies (Figure 4). Furthermore, as participant attrition is not uncommon during clinical trial studies, a potential 30% patient dropout rate was included for adjusting our final total sample size (N). Under this assumption, we would therefore need accrue at least 130 participants (n x k x 1.3) for this study in order to ensure an adequate sample number remaining at the end of the trial.

```
> pwr.t.test(d=1.001, power=0.95, sig.level=0.001)

     Two-sample t test power calculation

              n = 51.36005
              d = 1.001
      sig.level = 0.001
          power = 0.95
    alternative = two.sided

NOTE: n is number in *each* group
```

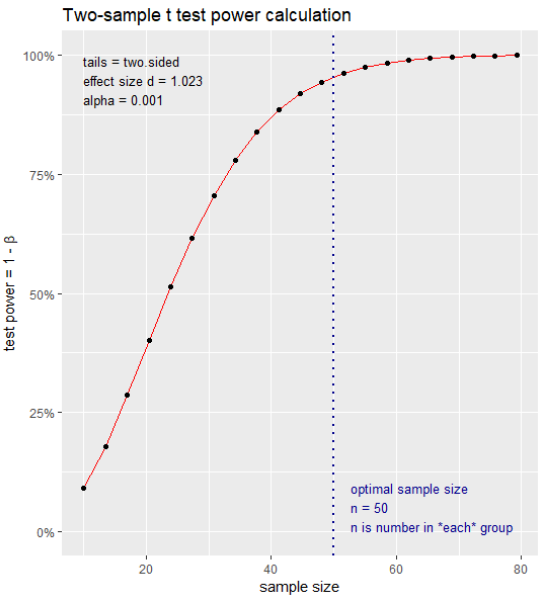


Figure 4: Results of t-test power calculation

FINDINGS AND CONCLUSIONS

Analysing survival data from clinical trial studies is critical step in development and validation process of new therapies. Our findings revealed that the antileukemic activity of 6-MP had a significant ability in prolonging the disease free period of children with acute leukemia This certainly delivers a profound impact on the science of systemic therapy for malignancies.

Furthermore, this analysis also revealed a more improved alternative design for this clinical trial by introducing more power for assessing the relative efficacy of 6MP aginast this disease. However, the limitations of this proposed strategy must also be noted regarding the major complications associated with recruiting such large number of patients or funding such costly studies.

KRUSKAL WALLIS TEST

INTRODUCTION

Serious and fatal road collisions take place each year in Ireland. By investigating into collision reports related to these incidents over recent years, it may be possible to identify key trends as to when a road user may be most at risk of injury during travel.

With this information, one may determine when is the most dangerous time of the day to be travelling on the road. Therefore, this analysis will investigate into the frequencies of killed and injured road casualties that have been classified by hour of day. Data on these occurrences were collected from the Central Statistics Office (CSO) for the years 2014-2016 in order to ensure a reasonable study sample size for analysis. As part of a data pre-processing step, all recorded observations were further grouped into three different 8-hour time buckets; 12:00 AM-7:00AM, 8:00PM -16:00 PM, and 17:00 PM -11:00 PM (k=3, n=24 per group). By doing so, we can further examine if there are any significant differences in the number incidents happening across three main time periods of the day, (night, morning/early afternoon and late afternoon/evening). Table 2 summarizes the cleaned and transformed data that after extraction.

NORMALITY ASSESSMENTS

Boxplots were created to explore the normality of the hourly reported incidents for each grouped time periods (Figure 5). The median number of reported casualties for between the hours of 12:00am-7:00am, 8:00AM-16:00PM and 17:00PM-11:00PM are 123.5, 413 and 459, respectively.

Interestingly, as illustrated in figure --, the median number of reported road incidents are quite visually different between the early hours of the morning and the two later time period of the day. Note the existence of an outlier in the 12:00AM-7:00AM time period which further exhibits indications of non-normality which shall be further investigated to determine a appropriate statistical test.

Table 2: Number of killed and injured causalities recorded per hour and grouped by three different time periods of the day. Data collected for the years 2014-2016

12:00 AM-7:00 AM	8:00 AM-16:00 PM	17:00 PM-11:00 PM
145	369	549
146	364	657
132	326	550
119	333	477
97	367	448
70	416	306
87	481	237
166	565	202
184	365	547
144	397	659
160	295	609
128	387	482
97	446	364
80	484	279
119	527	248
221	546	234
171	426	623
146	410	643
119	299	588
119	314	470
84	467	369
69	504	301
112	446	278
268	469	201

Source: Road Safety Authority

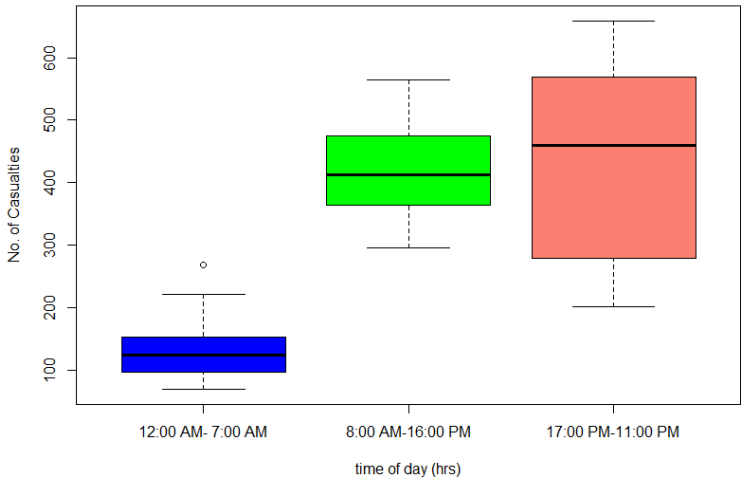


Figure 5: Boxplot of reported number of road incidents against 3 across three time periods of the day

Histograms were also used to further investigate the normality of these population distributions (Figure 6). A clear visual indicator of a non-normality can be observed in 12:00am -7:00am and 17:00pm-11:00pm. The effect of the previously mentioned outlier in the first time period is evident from the right skewed shape of it's distribution. However, this would interestinglt indicates that there a relatively low number of incidents reported during this time

This sense of non-normality can be further supported With the Q-Q plot, points follow a relatively straight line for the 8:00AM-16:00PM group (Figure 6). However, we can see we observe a bit divergence from normality in the upper tails of groups 12:00AM -7:00AM and 17:00PM-11:00PM.

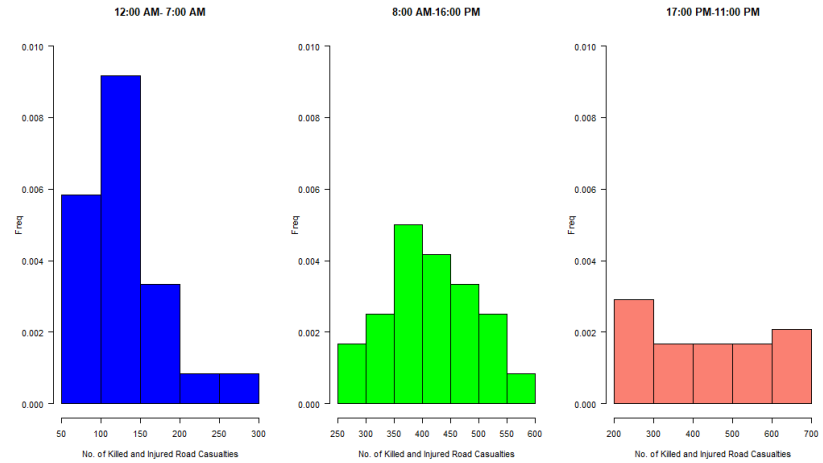


Figure 6: Histograms of hourly incidents reported at different time periods

```
> shapiro.test(RoadCasualties$'12:00 AM- 7:00 AM')

        Shapiro-Wilk normality test

data:  RoadCasualties$"12:00 AM- 7:00 AM"
W = 0.92504, p-value = 0.07555

> shapiro.test(RoadCasualties$'8:00 AM-16:00 PM'

        Shapiro-Wilk normality test

data:  RoadCasualties$"8:00 AM-16:00 PM"
W = 0.96887, p-value = 0.6393

> shapiro.test(RoadCasualties$'17:00 PM-11:00 PM')

        Shapiro-Wilk normality test

data:  RoadCasualties$"17:00 PM-11:00 PM"
W = 0.90873, p-value = 0.03311
```

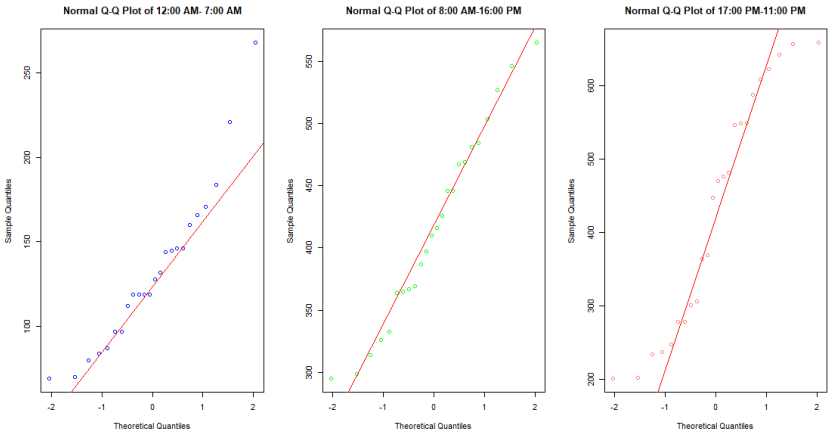


Figure 6: QQ plot of the three time group distributions

The Shapiro-Wilks test above was also applied to test for normality for which all statistical test above were examined at an $\alpha = 0.05$. With a p=0.076 for the '12:00 AM-7:00 AM" and a p= 0.639 for the "8:00 AM- 16:00 PM" group, we can conclude that both populations are normally distributed. However, given that the "17:00PM-11:00PM" group (p=0.033), has a $p<0.05$, we must therefore reject of the null hypothesis in favour of the alternative that the assumption of normality has not been met for this group.

KRUSKAL WALLIS TEST

As exploratory analyses revealed a non-normal distribution for one of the three groups, the Kruskal-Wallis test was adopted as a non-parametric alternative to the ANOVA for studying their differences between each other, setting α the default value, α=0.05. Below are the null and alternate hypotheses:

H_0: the mean rank number of road incidents are equal across all three time groups of the day

H_1: the mean rank number of road incidents are not equal across all three time group of the day and at least one group is different

FORMAL REPORT OF RESULTS

```
> RoadCasualties.kw = kruskal.test(RoadCasualties)
> RoadCasualties.kw

        Kruskal-Wallis rank sum test

data:  RoadCasualties
Kruskal-Wallis chi-squared = 46.231, df = 2, p-value = 9.143e-11
```

As we can see from the R output to the left, a Kruskal-Wallis test revealed a significant difference in the mean rank number of reported road accident between different time groups of the day, $\chi^2(2, N=24) = 46.23$, $p<0.001$. Thus, with a p-value less than α, the H_0 could therefore be rejected. The test therefore provides very strong evidence the group means are significantly different and that the distribution of reported road incidents is not equal across different time periods of the day.

POST HOC ANALYSIS (DUNN'S TEST)

A post hoc analysis was performed in order to identify these group differences which will test pairs of groups and adjust the p-value for multiple comparisons. The Dunn's test was selected as an appropriate method for nonparametric pairwise multiple-comparisons of medians between the independent groups

```
> dunn.test(RoadCasualties, kw=FALSE, method ='bh')

                    Comparison of RoadCasualties by group
                            (Benjamini-Hochberg)
Col Mean-|
Row Mean |          1          2
---------+----------------------
       2 |  -5.814632
         |    0.0000*
         |
       3 |  -5.959481  -0.144848
         |    0.0000*     0.4424

alpha = 0.05
Reject Ho if p <= alpha/2
```

From the post-hoc output on the left we see that the Dunn's pairwise test showed a significant difference ($p<0.05$) in the mean rank number of road fatalities of the "12:00AM-7:00AM" and "8:00AM-16:00PM" time groups ($p<0.001$), and of the "12:00AM-7:00AM" and "17:00PM-11:00PM" time groups ($p<0.001$).

However, it did not provide evidence of a significant difference in the mean rank number of road fatalities of the "8:00AM-16:00PM" and the "17:00PM-11:00PM" time group ($p=0.4424$).

FINDINGS AND CONCLUSIONS

Based on our statistical findings, there is strong evidence that between the hours 12:00AM-7:00AM appear be safest time of the day to travel for road users. In contrast, our analysis shows that the morning and evening times tend to be the most dangerous times for travel. Whilst most incidents were recorded during late afternoon/evening hours there is no statistically significant difference between this and the morning/early afternoon hours. This information could be leverage upon by the Irish Road Safety Authority (RSA) by informing Irish citizens to be more vigilant while travelling during the day. Perhaps for future analysis, these time groups could be narrowed into smaller windows so that we could deliver a more targeted and insightful view as to why road fatalities happen, when are they most frequent. Such results might help with proposing and evaluating wat to prevent collision and injuries.

MULTIPLE LINEAR REGRESSION

INTRODUCTION

The Pioneer Valley Planning Commission (PVPC) collected data on the volume of train users on the Northampton Rail Trail in Florence, Massachusetts for a ninety-day period. Data collectors set up a laser sensor that recorded when a rail-trail user passed the data collection station. The variables in this dataset provide historical information on various meteorological data including rainfall, cloud cover and temperature as well as data pertaining to the types of days in the work-holiday calendar. These will be applied as potential explanatory variables (independent variable, IV) for building a multiple linear regression model in order to predict the volume of train *riders* (dependant variable, DV). The task of forecasting the number of train passengers has direct business relevance as it can ensure the availability of adequate transport for customers by estimating when the demand may be high. A summary and brief description of these variable is highlighted in Table 3 below.

Table 3. Volume of users of a Massachusetts Rail Trail

Dependent variable:		
***riders*: estimated number of trail crossings that day (number of breaks recorded)**		
Independent variables (predictor variables)		
day: a factor with the levels: *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*	*highT*: high temperature for the day	*lowT*: low temperature for the day
precip: inches of rainfall	*clouds*: Measure of cloud cover	*weekday*: types of days in the work-holiday calendar. a factor with levels N (weekend or holiday) and Y (non-holiday weekday)

http://vincentarelbundock.github.io/Rdatasets/doc/mosaicData/Riders.html

GRAPHIC EXPLORATIONS

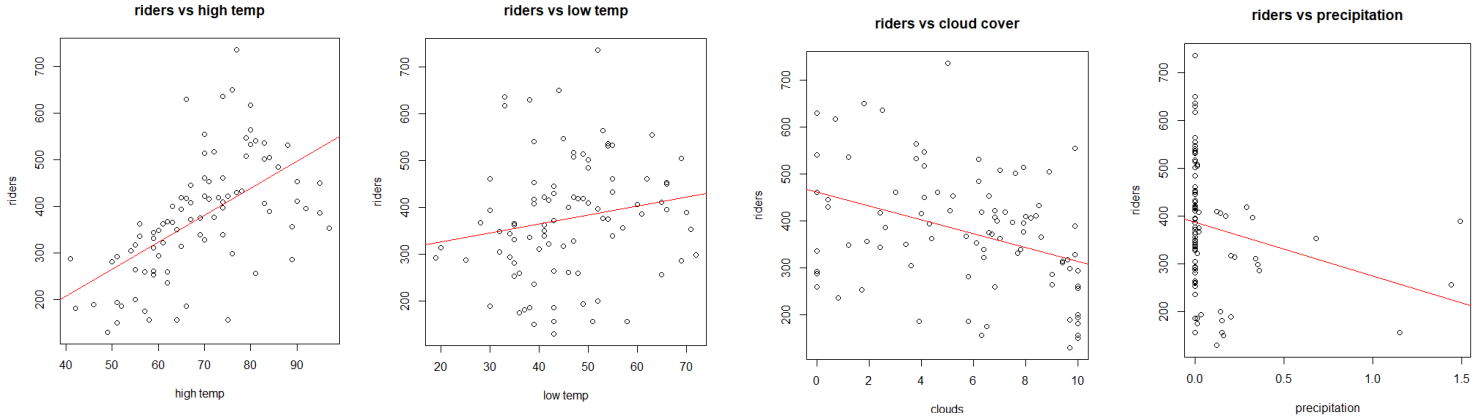


Figure 7: Scatterplots of voume of riders as a function of high temp, cloud cover precipitation and low

.

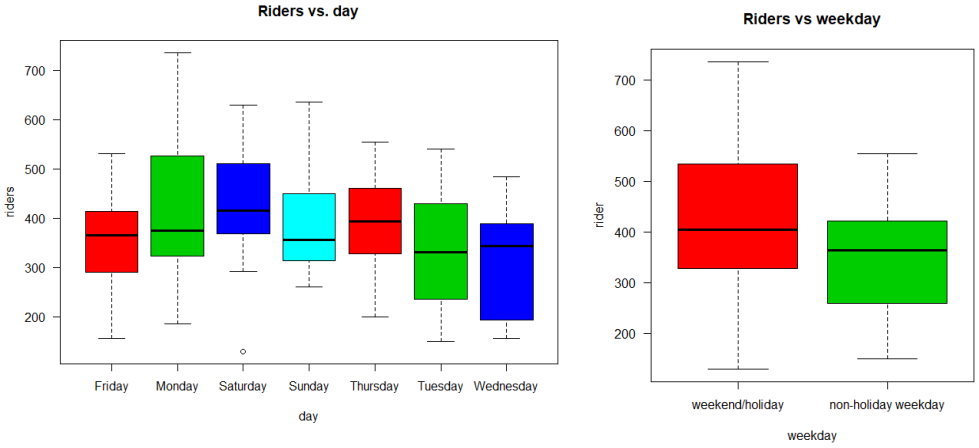


Figure 8: Boxplot of two categorical IVs, day and weekday, against the volume of riders (DV)

Scatterplots were employed as descriptive means to examine the relationship between the continuous numeric IVs and the DV (*riders*) (Figure 8). This was done to assess linearity, as well as an initial understanding of the distribution, strength and direction of the relationships. Visually there is a relatively strong and positive linear relationship between *riders & highT* which would seem reasonable, as commuters may wish to travel outdoors on a hot days. Less evident signs of correlation can be observed between the DV and the other IVs(*lowT*, *clouds* and *precip*.).

However, it is worth observing the similar negative direction of riders against *clouds* and *precip*. Given the close association between these two weather variables in nature this shared direction would make sense from a contextual perspective.

Boxplots were created for each independent categorical variable (*day* and *weekday*) against the DV in order to determine if there was an obvious trend between the number of *riders* travelling and a particular type of day in the week and work-holiday calendar. For *day*, we observe a higher number of users travel on a *Monday* and *Saturday*. However, note the heavy overlap between the interquartile ranges (IQR) for all of the boxplots and their close proximity between their median bars which would suggest that their differences are not significant. For *weekday*, it can be seen the IQR for the number of commuters on a weekend/holiday is much higher than the number recorded during weekdays.

The Shapiro-Wilks test above was also applied to test for normality in the distribution of the dependant variable, *riders* ($\alpha = 0.05$). With a p-value<0.05 (p=0.3981) the population is normally distributed.

```
> shapiro.test(Riders$riders)

        Shapiro-Wilk normality test

data:  Riders$riders
W = 0.98514, p-value = 0.3981
```

INVESTIGATION OF REGRESSION MODELS

The formula for a multiple regression equation is as follows:

$$\mathbf{y = a + b_1x_1 + b_2\,x_2 + b_3x_3 + \ldots + b_kx_k}$$

where a is the y-intercept which is dependant variable we wish to predict and $b_1 \ldots b_k$ are the regression coefficients or parameters. For the construction of these model, we will set the α=0.05

MODEL1: Simple linear regression model

For the first initial model, a simple linear regression model was overlaid on the relationship between *riders* and *highT*, given their previously demonstrated strong linearity to each other. The R output for the performance results of model1 are shown to the right.

From this, the regression equation for this model can be reported as follows:

riders = - 24.7468 + 5.794*highT

F(104.9 ,88) =43.31, P=3.256e-09, R2 =0.33

With a p-value p < .05, we can confirm that *highT* is be statistically significant at predicting an increase for *riders*. However, with a multiple R-Squared valuing at 33%, the explanatory power of this variable alone for the DV would be characterized as being relatively weak.

```
> summary(model1)

Call:
lm(formula = Riders$riders ~ highT, data = Riders)

Residuals:
     Min       1Q   Median       3Q      Max
-253.776  -65.054    9.066   58.780  314.637

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.7468    61.8034  -0.400     0.69
highT         5.7936     0.8804   6.581 3.26e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 104.9 on 88 degrees of freedom
Multiple R-squared:  0.3298,    Adjusted R-squared:  0.3222
F-statistic: 43.31 on 1 and 88 DF,  p-value: 3.256e-09
```

MODEL2: Addition of more variables to the model

Therefore, to improve upon our first model for *riders* , a number of other variables were added to as predcitors including the amount of rainfall (*precip*), the amount of cloud coverage (*clouds*) day of the week (*day*) and types of days in the work-holiday calendar(*weekday*). This was to highlight any issues with including all independent variables and to provide a basis for further optimisation later on. The R output for the performance results of model2 is shown to the right

From this, the regression equation for this model can be reported as follows:

riders = -175.93 +5.55 *highT -92.59precip -147.2692weekdayY - 2.07dayMonday -116.83 daySaturday -125.89 daySunday 36.2914 dayThursday -4.8890dayTuesday - 13.9960dayWednesday -7.6175clouds

F(10,79) = 9.39, p=4.463e-10, R= 0.48

```
summary(model2)

Call:
lm(formula = Riders$riders ~ highT + precip + weekday + day +
    clouds, data = Riders)

Residuals:
    Min      1Q  Median      3Q     Max
-227.15  -49.28   11.18   52.15  203.40

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   175.9297    89.0551   1.976   0.0517 .
highT           5.5530     0.7996   6.944 9.54e-10 ***
precip        -92.5889    41.8192  -2.214   0.0297 *
weekdayY     -147.2692    58.2786  -2.527   0.0135 *
dayMonday      -2.0784    42.1981  -0.049   0.9608
daySaturday  -116.8274    67.9850  -1.718   0.0896 .
daySunday    -125.8930    67.5293  -1.864   0.0660 .
dayThursday    36.2914    34.1456   1.063   0.2911
dayTuesday     -4.8890    34.8119  -0.140   0.8887
dayWednesday  -13.9960    35.4431  -0.395   0.6940
clouds         -7.6175     3.4528  -2.206   0.0303 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.44 on 79 degrees of freedom
Multiple R-squared:  0.5432,    Adjusted R-squared:  0.4853
F-statistic: 9.392 on 10 and 79 DF,  p-value: 4.463e-10
```

With a p-value>0.05, we can conclude that the difference between the DV and IV are statistically significant. With an R2 value of 48%, adding more variables appearing to have increased the predictive performance of our model

MODEL3: Model Optimization

As *precip*, *weekday* and *clouds* revealed to be statistically significant at predicting a drop in the volume of riders (-92.58, -157.27 and -7.62, respectively) in our previous model2 we will include these in model3 alongside *highT* and exclude the categorical variable *day*. We will test and see if this selection of variables are a bitter fit for our regression optimisation. The R output of these results are shown below.

From this, the regression equation for this model can be reported as follows:

riders = 66.96 + 5.89*highT -104.82precip -35.65weekdayY - 9.03clouds

F(4, 85) = 21.03, p<0.001, R2 =0.47 -

```
> summary(model3)

Call:
lm(formula = Riders$riders ~ highT + precip + weekday + clouds,
    data = Riders)

Residuals:
     Min       1Q   Median       3Q      Max
-247.260  -40.908    6.478   50.015  273.995

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   69.7573    63.9654   1.091  0.27855
highT          5.6802     0.8035   7.069 4.06e-10 ***
precip      -104.8155    40.7674  -2.571  0.01188 *
weekdayY     -35.6514    21.7855  -1.636  0.10544
clouds        -9.0255     3.3376  -2.704  0.00827 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.47 on 85 degrees of freedom
Multiple R-squared:  0.4974,    Adjusted R-squared:  0.4737
F-statistic: 21.03 on 4 and 85 DF,  p-value: 4.441e-12
```

However, with an R-Squared value of 0.47, removing *day* variable did not seem to improve the R-squared value of our model

MODEL4: Multiple Regression with Interaction

Following a number of variable combinations with several test models, a final model was created that included all of the same variables used for model3 but also inputting an additional interaction term between the *weekday* and *precip*. Not only did this slight adjustment improve our model's performance by having a higher adjusted R-squared value compared with model2 but it also by achieves this with a smaller difference between its multiple and adjusted R squared values in model4. This would indicate model4 could maybe fit the data better

From this, the regression equation for this model can be reported as follows:

riders = 66.96 + 5.89*highT -7.97clouds -377.79precip - 55.97weekdayY +296.25precip:weekdayY

P<alpha), R2 =0.50 - F(5, 84) = 18.93

```
> summary(model4)

Call:
lm(formula = Riders$riders ~ highT + clouds + precip * weekday,
    data = Riders)

Residuals:
    Min      1Q  Median      3Q     Max
-246.47  -51.99   10.12   51.44  255.42

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.9588    62.2483   1.076  0.28515
highT             5.8893     0.7867   7.486  6.4e-11 ***
clouds           -7.9703     3.2769  -2.432  0.01713 *
precip         -377.7860   120.2193  -3.142  0.00231 **
weekdayY        -55.9712    22.8184  -2.453  0.01624 *
precip:weekdayY 296.2524   123.1663   2.405  0.01836 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.97 on 84 degrees of freedom
Multiple R-squared:  0.5298,    Adjusted R-squared:  0.5018
F-statistic: 18.93 on 5 and 84 DF,  p-value: 1.468e-12
```

This statistically significantly interaction (p<0.05) could suggest that an increase in the volume of *riders* is noticed when there is rainfall during weekdays

HYPOTHESIS TESTING TO COMPARE REGRESSION MODELS

An anova() function was then applied in order to determine which of the models and to what is extent is best fitting to the data. The operation performs a hypothesis test comparing two models. The null hypothesis (Ho) is that no variables are significant for the model, and the alternative hypothesis (H_A) is that at least one variable is significant for the model.

```
> anova(model2, model1)

Analysis of Variance Table

Model 1: Riders$riders ~ highT + precip + weekday + day + clouds
Model 2: Riders$riders ~ highT
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     79 660583
2     88 969063 -9   -308480 4.0991 0.0002375 ***
```

```
> anova(model2, model3)
Analysis of Variance Table

Model 1: Riders$riders ~ highT + precip + weekday + day + clouds
Model 2: Riders$riders ~ highT + weekday + precip
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     79 660583
2     86 789275 -7   -128692 2.1986 0.04304 *
```

```
> anova(model4, model2)
Analysis of Variance Table

Model 1: Riders$riders ~ highT + clouds + precip * weekday
Model 2: Riders$riders ~ highT + precip + weekday + day + clouds
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     84 679922
2     79 660583  5     19339 0.4626  0.803
```

```
> anova(model3, model1)
Analysis of Variance Table

Model 1: Riders$riders ~ highT + precip + weekday + clouds
Model 2: Riders$riders ~ highT
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     85 726752
2     88 969063 -3   -242311 9.4468 1.874e-05 ***
```

```
> anova(model4, model3)
Analysis of Variance Table

Model 1: Riders$riders ~ highT + clouds + precip * weekday
Model 2: Riders$riders ~ highT + precip + weekday + clouds
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     84 679922
2     85 726752 -1    -46830 5.7855 0.01836 *
```

REPORT ANOVA HYPOTHESIS TESTS

The ANOVA tests reveal there is evidence that model2 and model3 are significantly better at predicting the DV, *riders*, than model 1, F(2, 88) = 4.01, p = 0.0002 and F(2, 88) =9.45, p<0.001, respectively. between model 2 and 3 against 1 in that they both have more predictive power (p=0.0002 and p=1.874e-05, respectively. Thus adding more variables significantly helps with improving the predictive performance of our model.

When the two top models with the highest R-squared values (model2 and model4) were tested compared to each other, the results of the ANOVA showed that the differences between the two were not statistically significant(p=0.803). However, note that the F-stat and significance in difference between model4 and model3, F(2, 85) =5.79 , p=0.02, is larger than it is between model2 and model 3, F(2,86)=2.20, p=0.04. This, coupled alongside it's smaller differences between the multiple R-squared and adjusted R square value than model 2 would lead us to suggest that model4 is a superior fit to the data than model2 and is the superior in predicting the volume of train *riders.*

FINDINGS AND CONCLUSIONS

Therefore, with an R-Squared value of 0.502, the selecting explanatory variables in our final chosen model (model4) can explain 50% of the variability in the response variable. In addition, when we plot the residuals of our final model, we can observe that the data appears homoscedastic. QQ-plots also support that the data in this model meets the assumptions of normality (Please refer to original R script for graph results). This has direct business relevance as it could be applied to determine the numbers of trains required at particular peak times of the week and under certain weathers conditions where we may expect to see high customer demand. With our model we could ensure the availability of transport to meet their needs. However, it must be of note that a model with an R-squared value of above 60% is typically the minimum desirable score for most studies, and thus the variables in our final model may not be particularly be considered as strong explanatory variables for *riders*.

TWO-WAY ANOVA

INTRODUCTION

Whilst motorcyclist deaths and fatalities have been steadily falling over the years, they still represent the most vulnerable road users in Ireland. According to figures reported by the RSA, their risk of dying in a traffic crash is much higher than that of car occupants (Road Safety Authority, 2017). Previous existing studies on road injuries reveal that there are certain trends and behavioural risk factors involved in motorcyclists' crashes, particularly factors of age and gender but few analyses have been performed on their potentially joint effects on the prevalence of these accidents (Chang & Yeh, 2007). Thus, the purpose of research analysis is to explore the effect of two independent variables, *Age* and *Gende*r, as potential contributing factors to motorcycles crashes and deaths and to see if they have interactive effects in order to further define populations within this road user category that are most at risk. To conduct this study, Irish governmental data which classifies all killed and injured motorcyclists casualties by sex and several age groups was retrieved from the CSO for years 2005-2016 (Central Statistics Office). Given the age groups 18-29, 21-24 and 25-34 represent the most frequent users of motorcycle vehicles, this analysis will focus it's investigation on these three levels.

GRAPHIC EXLPLORATIONS

Boxplots were employed to identify potential trends from the distribution of motorcyclist incidents across the categories of age groups and sex. There appear to be differences between the means of the different age and gender groups. However, their does appear to be overlap between the different age groups with considerable different variances. Thus, these graphics were also used to assess their levels of normality. Interestingly, the boxplots indicate that amongst the population of motorcyclists, those that appear most at risk of road injuries and deaths are males belonging to the age category of 24-35 years (Figure 9, Figure 10)

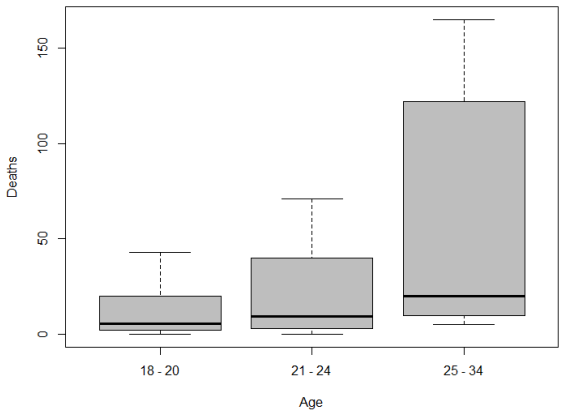


Figure 9: Boxplots showing the number of motorcyclist deaths and injuries across three different age groups

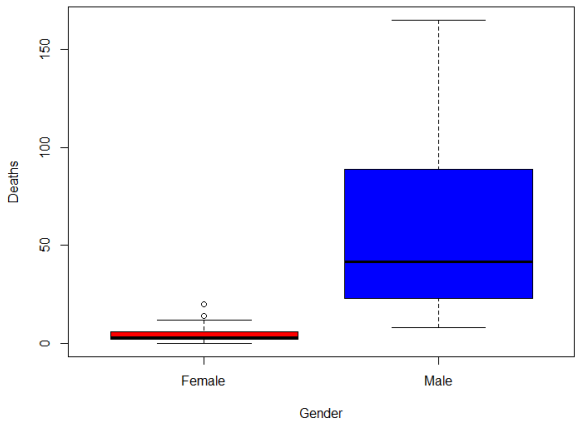
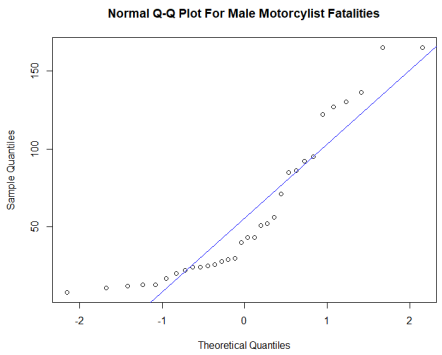
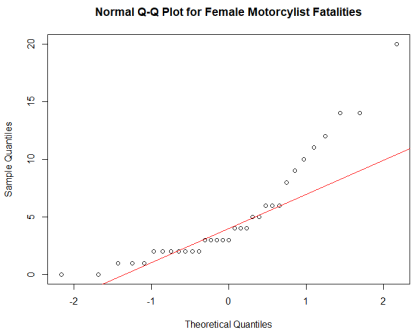


Figure 10: Boxplots showing the number of motorcyclist deaths and injuries acrosst the genders of females and males

NORMALITY ASSESSMENTS

The distributions of road fatalities across males and females were further explored in order to validate if the assumptions of normality were met, which is a requirement for running this test. Visually the QQ plots indicate that both populations are not normally distributed (figure 11) which was quantitatively confirmed with Shapiro-Wilk test as both resulted with a p value<0.05. While all visuals and test for the normality show the data is not normal, ANOVAs tests have been shown to be adequately robust to withstand against violations of normality (Blanc et al, 2017).






```
> shapiro.test(Deaths_Male)

        Shapiro-Wilk normality test

data:  Deaths_Male
W = 0.85541, p-value = 0.0005446


# some normality test for feamles
> shapiro.test(Deaths_Female)

        Shapiro-Wilk normality test

data:  Deaths_Female
W = 0.8419, p-value = 0.000227
```

Figure 11: QQ-plot to assess the normal distributions of males and female motorcyclist causalities

TWO WAY ANOVA TEST

Therefore, the three null hypothesis in this 2x3 factorial ANOVA test design are as follows:
H_0: *Gender* has no significant effect on the mean number of motorcyclist casualties
H_0: *Age* has no significant effect on the mean number of motorcyclist casualties
H_0: *Gender* and *Age* interaction has no significant effect on the mean number of motorcyclist casualtiess

Without Interactions

```
# Two-way ANOVA (without interactions)

> m1 = aov(Deaths ~ Gender + Age)
> summary(m1)
            Df Sum Sq Mean Sq F value   Pr(>F)
Gender       1  45644   45644   74.75 3.43e-12 ***
Age          2  33799   16899   27.68 2.80e-09 ***
Residuals   61  37248     611
```

With Interaction

```
# Two-way ANOVA
> m2 = aov(Deaths ~ Gender + Age + Gender:Age)
> summary(m2)
            Df Sum Sq Mean Sq F value   Pr(>F)
Gender       1  45644   45644  220.13  < 2e-16 ***
Age          2  33799   16899   81.50  < 2e-16 ***
Gender:Age   2  25014   12507   60.32 5.44e-15 ***
Residuals   59  12234     207
```

REPORT TWO WAY ANOVA TEST RESULTS

A 3x2 ANOVA with age groups (18-20, 21-24, 25-34), high) and gender (female, male) as between-subject factors revealed that the population means of *Age* and *Gender* are different, indicating that both factors have a significant effect on the number of motorcyclist causalities. To briefly summarize these findings, our test revealed a main effects of Gender, $F(1,59) = 220.13$, $p<0.001$ and *Age*, $F(2, 59) = 81.50$, $p<0.001$. Analysis of two-way interactions of age and sex show that sex is the primary cause of these casualties Furthermore, these main effects were qualified by a significant interaction between *Gender* and *Age*, $F(2, 59) = 60.32$, $p<0.001$.

POST HOC TEST

Post hoc analyses using Tukey's HSD test was performed in order to identify between which group means do these previously established differences lie? All significant main effects of each factor and their simple main effects from the results of the test are highlighted in the R output on the right.

Here we see that, at p value of >0.05, all group means are significantly different between the two groups in *Gender* and between the three groups in *Age*. With regards to the simple main effects in *Gender: Age*, we see that, at p value of >0.05, all of the group means that are significantly different between *Genders* at each level *Age*, bar the unhighlighted ones.

The output results of the Tukey test was graphically displayed in (figure 12) in order to visually compare the estimated different differences between all possible multiple pair level combinations in *Gender* and *Age* at a 95% probability level (figure 12). Our post hoc analysis show that, at a $p>0.001$, the most large and significant differences lie between all *Males :25-34* group pair combinations.

```
> TukeyHSD(m2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Deaths ~ Gender + Age + Gender:Age)

$Gender
                  diff      lwr      upr p adj
Male-Female 53.00473 45.85613 60.15334     0

$Age
                      diff       lwr      upr     p adj
21 - 24-18 - 20 10.77273  0.3343494 21.21111 0.0416344
25 - 34-18 - 20 53.20574 42.6438272 63.76765 0.0000000
25 - 34-21 - 24 42.43301 31.8710999 52.99493 0.0000000

$`Gender:Age`
                                          diff         lwr         upr     p adj
Male:18 - 20-Female:18 - 20        18.363636   0.2786984  36.448574 0.0445578
Female:21 - 24-Female:18 - 20       1.727273 -16.3576653  19.812211 0.9997497
Male:21 - 24-Female:18 - 20        38.181818  20.0968802  56.266756 0.0000008
Female:25 - 34-Female:18 - 20       8.818182  -9.2667562  26.903120 0.7050840
Male:25 - 34-Female:18 - 20       118.663636 100.1320895 137.195183 0.0000000
Female:21 - 24-Male:18 - 20       -16.636364 -34.7213016   1.448574 0.0884806
Male:21 - 24-Male:18 - 20          19.818182   1.7332438  37.903120 0.0237850
Female:25 - 34-Male:18 - 20        -9.545455 -27.6303925   8.539483 0.6309594
Male:25 - 34-Male:18 - 20         100.300000  81.7684531 118.831547 0.0000000
Male:21 - 24-Female:21 - 24        36.454545  18.3696075  54.539483 0.0000024
Female:25 - 34-Female:21 - 24       7.090909 -10.9940289  25.175847 0.8560177
Male:25 - 34-Female:21 - 24       116.936364  98.4048167 135.467911 0.0000000
Female:25 - 34-Male:21 - 24       -29.363636 -47.4485743 -11.278698 0.0001681
Male:25 - 34-Male:21 - 24          80.481818  61.9502713  99.013365 0.0000000
Male:25 - 34-Female:25 - 34       109.845455  91.3139076 128.377001 0.0000000
```

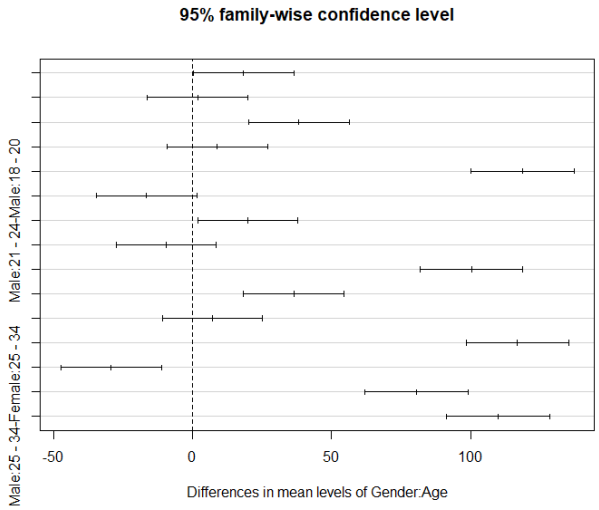


Figure 12: Grpahical disaply of Tukey HSD results

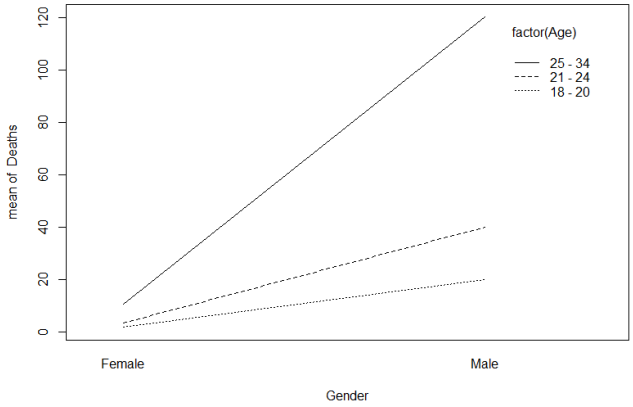


Figure 5 Interaction plot of mean number of motorcyclist deaths/injuries by gender of each age group

Thus, males belonging to 25-34-year age groups are more at risk of suffering from road causalities. We can also provide a good summarised illustration of these differences and the identified interaction effects between factors by graphing the mean number of motorcycle casualties by gender and age in an interaction plot (figure 13). This graphic further supports that the mean number of road casualties for female motorcyclists differ from the mean number for males. Similarly, we see that the mean number of casualties differ between all three age groups. The non-parallel alignment between all of the lines in the plot also displays that there is an obvious interaction between the factor of *Gender* and *Age*.

Furthermore, amongst females, there is very little difference in the number of motorcycle fatalities between the different age groups. However, as previously shown, these differences are much starker amongst males and even more so between the 25-34 year old male group and all age levels of females.

FINDINGS AND CONCLUSIONS

Therefore, the factors of gender and age play a huge role in determining the risk of succumbing to motorcyclists' accidents on the road. This study has produced a age and sex profile of the motorcyclist population in Ireland which could be used for providing actionable insights into the characteristics of motorcyclist that are most at risk of suffering from road casualties.

The result of this analysis concludes that the most dominant sex group at risk of motorcyclist deaths are male riders and was highest for those aged 25-34. years. With these insights, the RSA could launch more road safety awareness interventions and initiative campaigns on this specific subpopulation of motorcyclists. By targeting this demographic with more road safety advertisement, we could hopefully reduce the number of reported casualties in the next year and reduce their risk profile.

REFERENCES

Research Articles

Klein and Moeschberger (1997) *Survival Analysis Techniques for Censored and truncated data*, Springer. Freireich et al. (1963) *Blood* 21: 699-716.

Blanca, M., Alarcón, R., Arnau, J., Bono, R. and Bendayan, R., 2017. Non-normal data: Is ANOVA still a valid option?. *Psicothema*, *29*(4), pp.552-557.

Chang, H.L. and Yeh, T.H., 2007. Motorcyclist accident involvement by age, gender, and risky behaviors in Taipei, Taiwan. *Transportation Research Part F: Traffic Psychology and Behaviour*, *10*(2), pp.109-122.

Road Safety Authority, Road Safety Annual Report, 2017 . Available at : https://www.tispol.org/sites/default/files/article_files/RSAR2017_0.pdf (Accessed: 22 June 2019).

Original Dataset Sources

WILCOXON'S TEST

http://vincentarelbundock.github.io/Rdatasets/doc/KMsurv/drug6mp.html (Accessed: 12 June 2019).

.

KRUSKAL WALLIS TEST

Central Statistics Office, **StatBank** . Available at: https://www.cso.ie/px/pxeirestat/Statire/SelectVarVal/Define.asp?maintable=roa16&ProductID=DB_RSA&PLanguage=0 (Accessed: 10 June 2019).

MULTIPLE LINEAR REGRESSION

http://vincentarelbundock.github.io/Rdatasets/doc/mosaicData/Riders.html (Accessed: 12 June 2019).

TWO-WAY ANOVA

Central Statistics Office, **StatBank** . Available at: https://www.cso.ie/px/pxeirestat/Statire/SelectVarVal/Define.asp?maintable=roa19&ProductID=DB_RSA&PLanguage=0 (Accessed: 2 June 2019).