Final Project, EDDA 2017

Martin de la Riva(11403799) and Kieran O'Driscoll(11426438), Group 23

01 June 2017

Introduction

Air travel is one of the most widely used modes of transport and also one of the fastest. However, once an airline crashes it tends to have a significant impact on consumer's confidence for travelling with that airline. This was the case with Malysian Airline in [1] effectively forcing the airline to reband itself to stay in business.

The question is, are consumers right to feel this way? If an airline crashes, is this an indicator that they are likely to crash in the future? In this paper this topic will be discussed. By exploring this topic additional questions will appear such as if the location of the crash is signicant or if the choice of plane manufacturer has an impact on the number of fatalities.

The datasets used to check this will be drawn from two sources. The first source is taken from the github repository of *Five Thirty Eight* [2] which contains some general airline saftey statistics from 1980s to 2010s. A similar comparison was done by *Five Thirty Eight* but in this paper will use an additional more indept data set and filter by only checking airlines that has an accident to see if they are likely to have one in the future. The second data-set is drawn from the *Socrata* [3] site which contains more detailed data on airline statistics.

Hypotheses

The research question is does the prior number of casualties of an airplane mean that this airplane is more likely to crash in the future. From this one question many subquestions will arise, as mentioned before, such as has air travel become safer with time, how the amount of causalities are affected by the location of crashes, manufacturers and the year of the crash.

Preliminary tests

Firstly, some preliminary tests are conducted on the first dataset. This dataset offers a general overview of two big periods which include the amount of aircrashes per airline from 1985 to 1999 and from 2000 to 2014. From this data set, a general consensus can be obtatined which can be further expanded upon with the second data set.

```
## The following object is masked _by_ .GlobalEnv:
##
##     airline
```

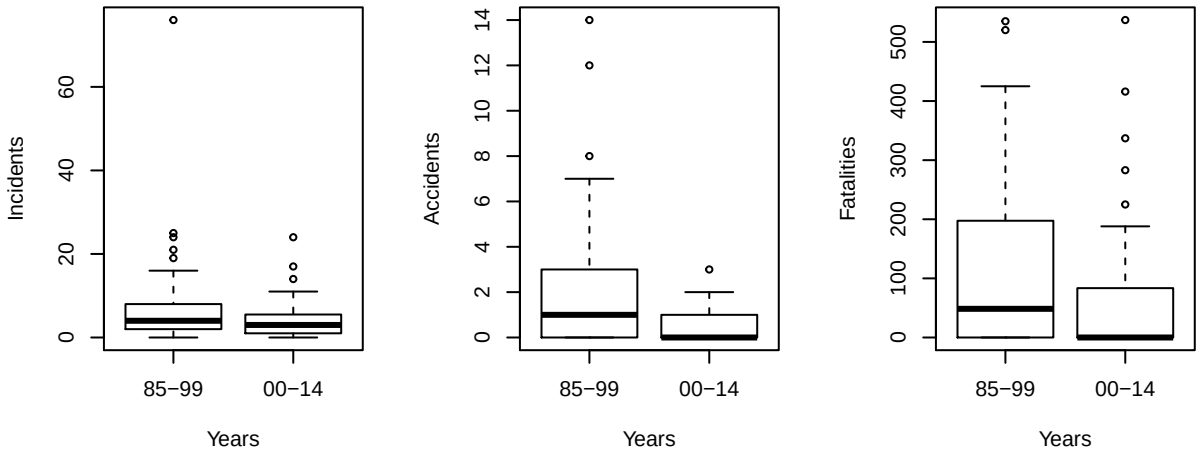
```
## [1] "airline"                "avail_seat_km_per_week"
## [3] "incidents_85_99"        "fatal_accidents_85_99"
## [5] "fatalities_85_99"       "incidents_00_14"
```

```
## [7] "fatal_accidents_00_14"  "fatalities_00_14"
```

The data is in terms of incidents, fatal accidents and fatalities. With this dataset, testing how the flying industry has improved over the years can be used as a starting point for our reasearch. The data can be expressed as an experiment of individuals (in this case airlines) that are tested on two periods of time. There are three experimental units that can be tested to confirm the hypothesis which include: teh amount of incidents, fatal accidents and fatalities. On this dataset the hypothesis that the difference on the mean between the two periods is 0 will be tested. In other words, the test will be the amount of aerial accidents/fatalities that remain unchanged over time without neither improvement nor deterioration.

Data

First, the data will be explored to check what insghts can be drawn.



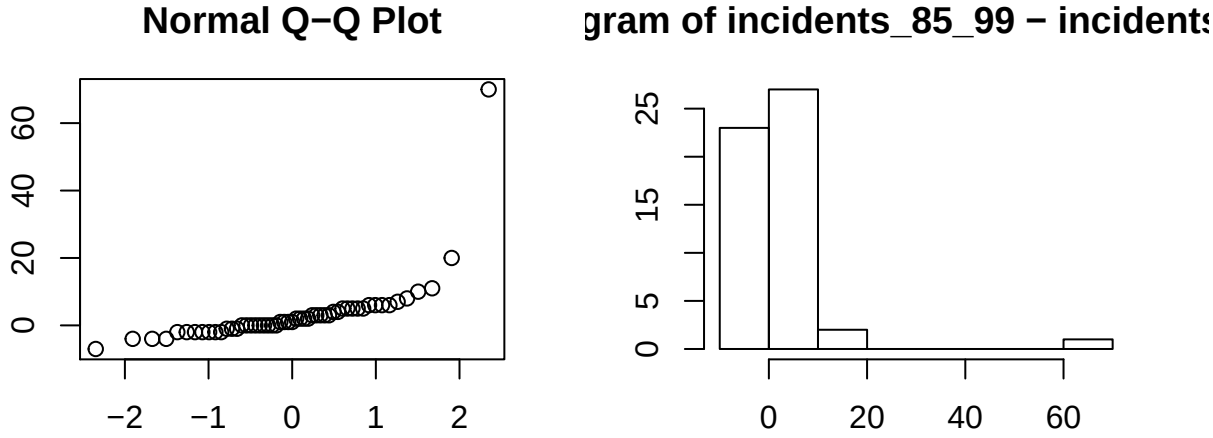
Looking at the boxplots, of the three numerical outcomes, it does seem that the hypothesis is wrong, as it seems that the number of accidents/fatalities decreases on the second period. In the following sections, this will be checked if this is also the case when performing statistical tests.

Incidents

Firstly, to test the hypothesis on the numerical outcome *Incidents*, the Two-Paired test can be applied as the experimental units are individuals however this is assuming that the difference between the periods comes from a normal distribution.

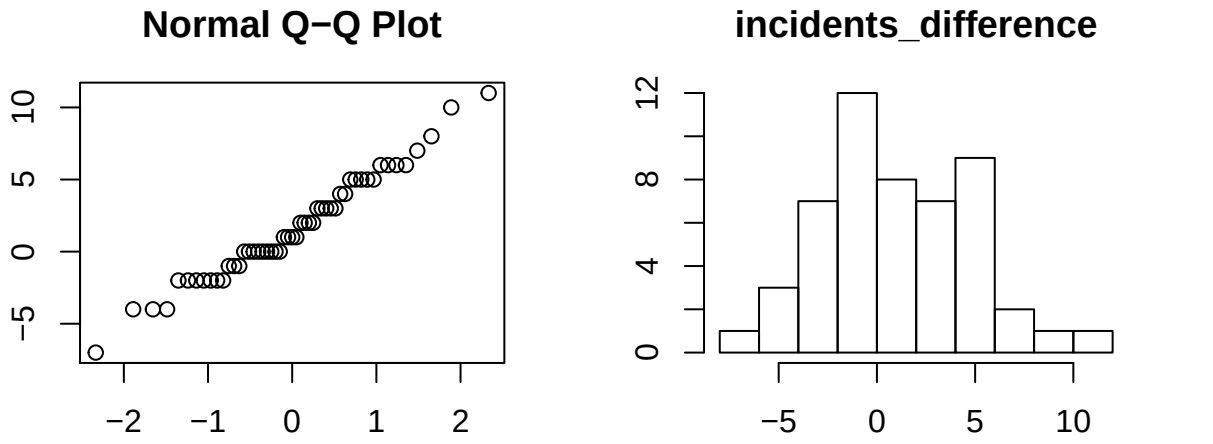
Firstly, the instances in the data that had 0 *Incidents* in the first period will be removed.

As the experimental units are individuals, we can apply a Two-Paired test, assuming the difference between the periods comes from a normal distribution.



Unforunately, this assumption cannot be applied to this data due to two big outliers. If the outliers are eliminated, one can clearly see an underlying Normal Distribution.

```
#@Martin Check if this relevant
incidents_difference=incidents_85_99-incidents_00_14
par(mfrow=c(1,2), mar=c(2.1, 3.1, 3.1, 2.1))
qqnorm(incidents_difference[incidents_difference<15], xlab="", ylab="")
hist(incidents_difference[incidents_difference<15], main="incidents_difference", xlab="", ylab=
```



For this case, a Two-Paired test on the difference without outliers is performed and will be backed up with a permutation test due to the fact that the data might not not truly have an underling Normal Distribution.

```
t.test(incidents_difference[incidents_difference<15])
```

```
##
##  One Sample t-test
##
## data:  incidents_difference[incidents_difference < 15]
## t = 3.1671, df = 50, p-value = 0.002624
## alternative hypothesis: true mean is not equal to 0
```

```
## 95 percent confidence interval:
##  0.6025119 2.6916058
## sample estimates:
## mean of x
##  1.647059
```

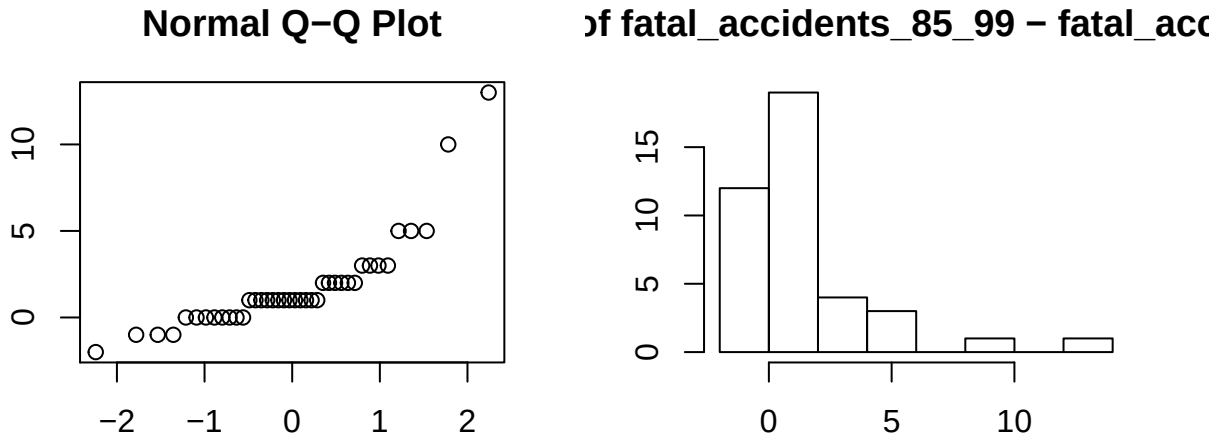
The Two-Paired test clearly shows that the difference between the population has to be different from 0, estimating it in a confidence interval between 0.5 and 2.5, which is not that high.

```
permutation_test = function(mystat, col1, col2){
  B=1000
  tstar=numeric(B)
  for (i in 1:B){
    temp=t(apply(cbind(col1,col2),1,sample))
    tstar[i]=mystat(temp[,1],temp[,2])
  }
  myt=mystat(col1,col2)
  #print(myt)
  pl=sum(tstar<myt)/B
  pr=sum(tstar>myt)/B
  p=2*min(pl,pr)
  print(paste("Mean Diff: ", myt ,"P-value:", p))
  #print(p)
}
mystat=function(x,y) {mean(x-y)}
permutation_test(mystat, incidents_85_99, incidents_00_14)
```

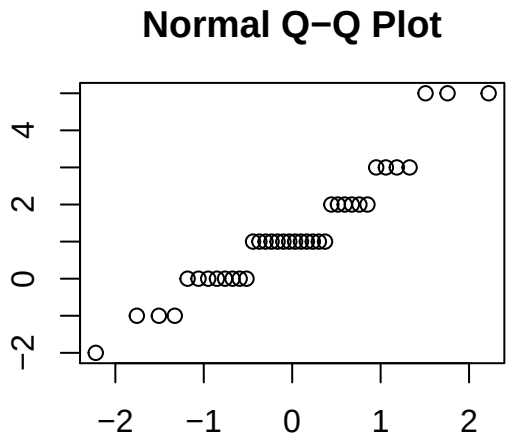
```
## [1] "Mean Diff:  3.28301886792453 P-value: 0.002"
```

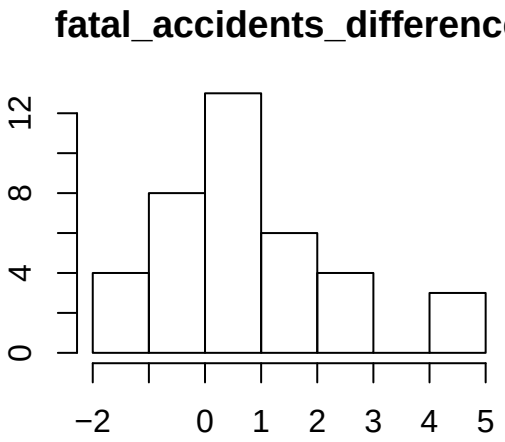
The permutation test clearly backs what was previously seen in the Two-Paired test. Therefore, it can be concluded that there is a significant difference between both periods in terms of the number of *Incidents*.

Fatal Accidents



In terms of *Fatal Accidents*, there is a similar situation as before. The plots do not seem to resemble a Normal Distribution, although it may be caused by the clearly detectable outliers. Therefore, the distribution will be checked after getting rid of these outliers:





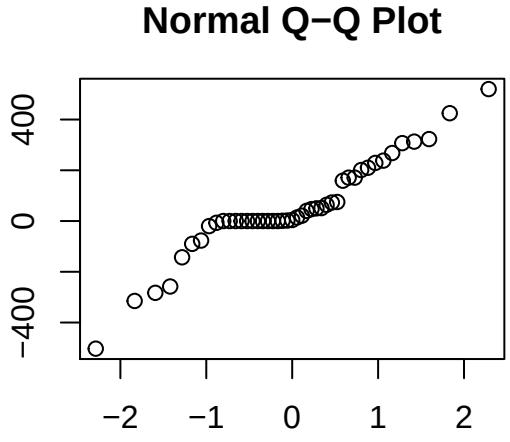
In this case, even after removing the outliers, a Normal Distribution cannot be definitively assumed. Therefore, only a Permutation Test will be carried out for this case:

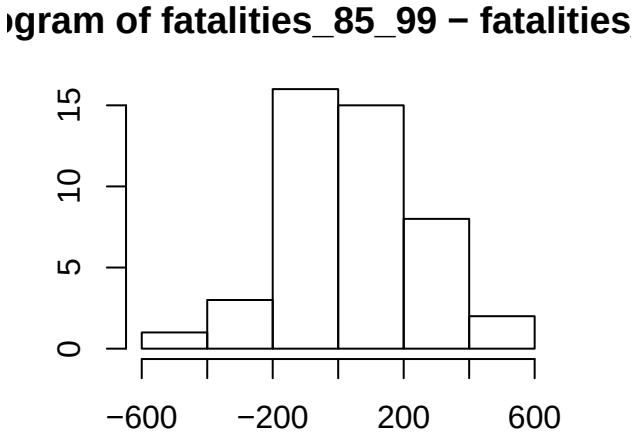
```
permutation_test(mystat, fatal_accidents_85_99, fatal_accidents_00_14)
```

```
## [1] "Mean Diff:  1.75 P-value: 0"
```

Again, the test on *Fatal Accidents* clearly shows that there is a significant difference between both periods.

Fatalities





In the case of *Fatalities*, it can be assumed that the difference on Fatalities between the two periods comes from a Normal Distribution. Therefore, a Two-Paired test can be used.

```
t.test(fatalities_85_99, fatalities_00_14, paired=TRUE)
```

```
##
```

```
##  Paired t-test
##
## data:  fatalities_85_99 and fatalities_00_14
## t = 1.8178, df = 44, p-value = 0.07591
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -5.508918 106.886696
## sample estimates:
## mean of the differences
##                50.68889
```

There seems to be a significant difference between the *Fatalities* in the 20th century and the 21st century with the first being significantly higher (with a confidence interval between 9 and 105 instances bigger), following the results on the previous tests.

Conclusion on preliminary experiments

From these preliminary experiments, the conclusion is that there has been a significant improvement in flight saftey from the 20th to the 21st century and that an airline that has crashed in not a indication that they are likely to crash again. However, this experiment has used data only from two periods where the period, in this case 10 years, might be two small or two big. Therefore in the next section, different year periods will be used along with additional information such as manufatuer and location of crash. With this improvement, a deeper experiment can be carried out.

Deeper Investigation

The second data set contains 21 columns with over 500 entries. The data-set contains data from 1908 to 2009 with airplane crashes from commercial airlines and miltary planes. For this experiment, only commercial airlines will be consindered and only the columns *Date, Operator, Manufacturer* and *Location* will be used. *Operator* in this data-set is the Airline of the plane that crashed.

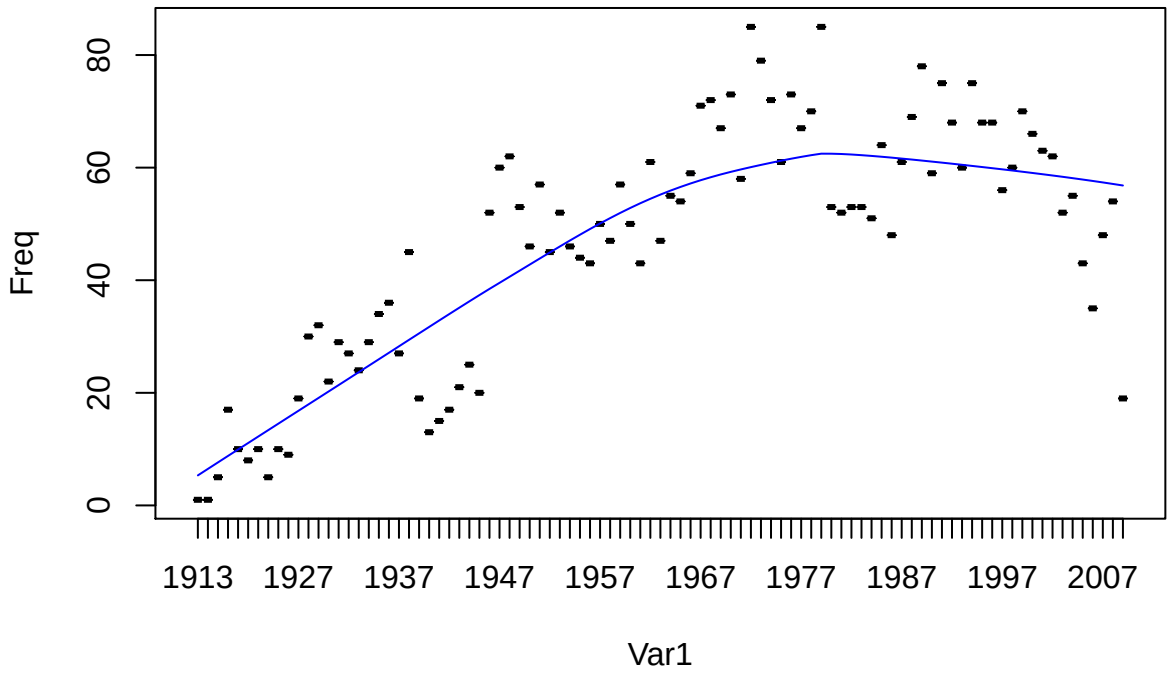
Data Cleaning

Due to the format of some of the data, some small data cleaning had to be performed. The column *Date* needs to be cleaned as the data needs to be aggregated by year for the statiscal test. Also invalid entires need to be filtered out along with any miltary crashes.

```
#transform Location column into countries
library(stringr)
temp=str_split_fixed(airlines_data$Location, ", ", 2)
countries=temp[,2]
airlines_data$Location=countries
#delete Empty Locations
airlines_data = airlines_data[!airlines_data$Location=="",]
#delete Military data
```

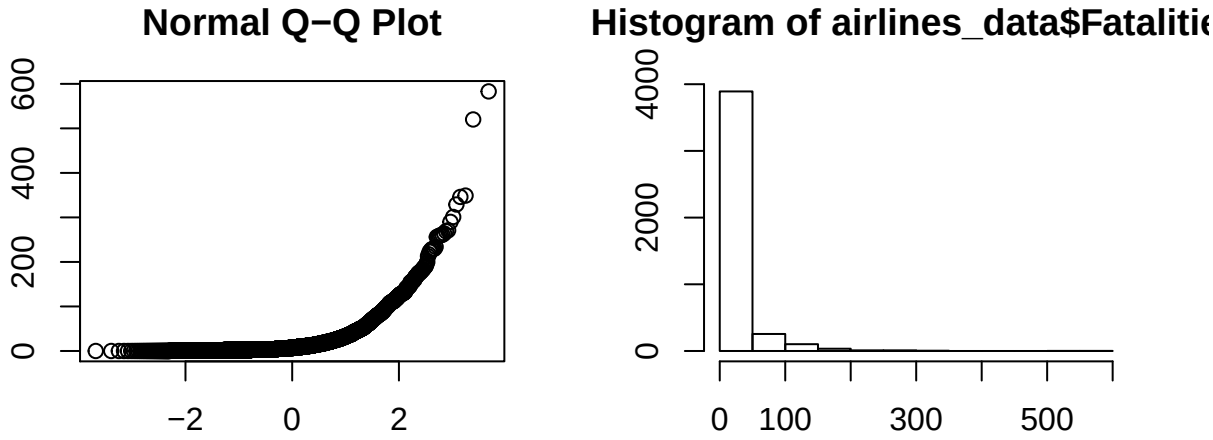
```r
airlines_data = airlines_data[airlines_data$Classification=="Non Military",]
#transform Date to Years
temp = as.Date(airlines_data$Date,'%m/%d/%Y')
airlines_data$Date = format(temp,'%Y')
```

Data



The plot above shows the amount of accidents per year. It reveals the improvement on airflight technology over time leading to more flights occuring but also the lack of saftey standards in the early 90s gives a growth in number of accidents as air travel becomes more popular. The amount of accidents seems to stabalize at the around the 1970s in which from this point onwards only a slight decrease in accidents can be seen.

To check what test can be applied, the distrubution of the data is checked.



The data is clearly not normally distrubuted which will filter out any statistical test that is based on this assumption. In this case a permuation test will be used, however to do this the data needs to be formatted in such away in which it has two groups of experimental units.

Setting up Experiments

To set up the experiment, a function was created to format data into two groups of experimental units, *before* and *after*. The function will also allow for experimentation with different ranges years thus altering the time period in which the *before* and *after* columns would cover. The function is given below:

```
setup_data <- function(date, range, data) {
    airlines = unique(data$Operator[data$Date == date])
    before = rep(0, length(airlines))
    after = rep(0, length(airlines))
    lower_bound = date - range - 1
    upper_bound = date + range
    df <- data.frame(airlines, before, after)
    for(i in seq(from=1, to=length(airlines), by=1)){
      df$before[i] = sum(data$Fatalities[data$Date >= lower_bound & data$Date <= date &
                                         data$Operator == df$airlines[i]])
      df$after[i] = sum(data$Fatalities[data$Date <= upper_bound & data$Date > date &
                                         data$Operator == df$airlines[i]])
    }
    return(df)
}
```

This function is created to aggregate the data for a year. This gives for each airline a *before* and *after* figure. This function is fed a year and range. The range alters the *before* and *after* columns. The aim is to provide a data-set where it can be determined that if a plane crashes(the *before* column) are they more likely to crash in the near future(the *after* column).

Experiment

The experiments are carried out on three years 1955, 1975, 2004 with ranges of 15, 10 and 5 applied repectively. This allows for experimentation with years and experimentation with the ranges.

```
## [1] "Testing on year: 1955  Using range: 15"

## [1] "Mean Diff:  -43.1315789473684 P-value: 0.354"

## [1] "Testing on year: 1975  Using range: 10"

## [1] "Mean Diff:  55.1132075471698 P-value: 0"

## [1] "Testing on year: 2004  Using range: 5"

## [1] "Mean Diff:  13.3636363636364 P-value: 0"
```

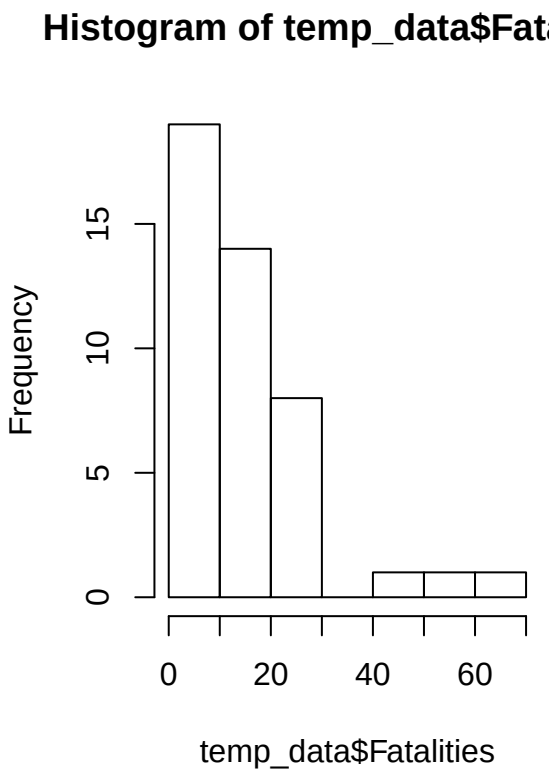
The p-value returns zero for both years 1975 and 2004 using ranges 10 and 5 respectively. However, 1955 with a big range of 15 years returns a p-value of 0.322. This could be because of the date 1955, air transport could have been becoming a more affordable options which in turn could have led to more crashes.

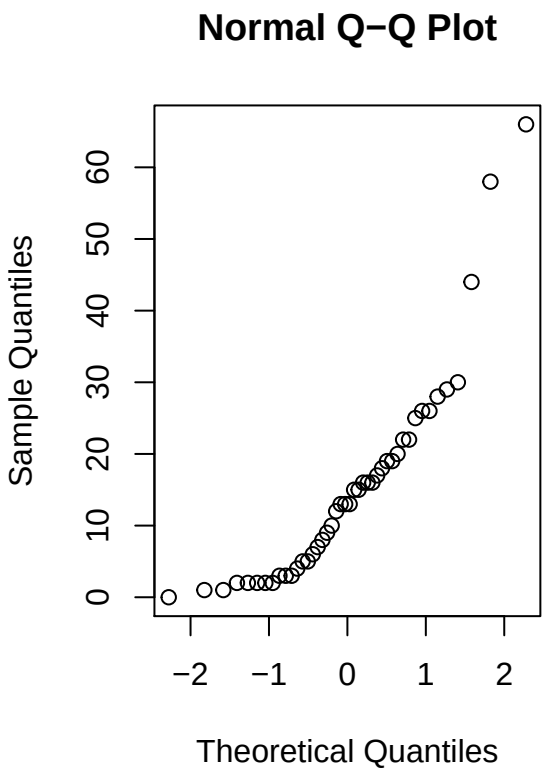
To test this, the test is run again and the range is shortened for the year 1955 to both 10 and 5. Also the year 1975 is taken with a range of 15 to check the impact that the range of years has on the result.

```
## [1] "Testing on year: 1955  Using range: 10"

## [1] "Mean Diff:  -7.28947368421053 P-value: 0.838"

## [1] "Testing on year: 1955  Using range: 5"

## [1] "Mean Diff:  16.4210526315789 P-value: 0.132"

## [1] "Testing on year: 1955  Using range: 2"

## [1] "Mean Diff:  26.4736842105263 P-value: 0"

## [1] "Testing on year: 1975  Using range: 15"

## [1] "Mean Diff:  56.9245283018868 P-value: 0.002"
```

From this, two conclusions can be drawn. One that the range used has a impact on the p-value, this is seeb in 1975 has the p-value increase but only by a small amount. Secondly, the date 1955 seems significant. After some research it was found that the period from 1950s and 1960s is known as the golden age in aviation [4]. This reminiscent of the graph produced in the Data section as air travel became a much more affordable option for people to persue but lacked many of the saftey standards implemeted today lead to increase in the number of accidents. This would in theroy make the risk of crashing higher however this would have to be investigated further.

Seeing as the data-set is quite big, one year will be used to see if a particular manufacturer or location alter the probablity of having a crash. In this case, 1955 will be checked. The data for this is displayed below to see if the normal assunmption is true. Also only tests that are checked agianst *Fatalties* will be used.





Data is not normally distrubuted so therefoe cannot use anova, therefore Will use a a non-parametric test specifically the Kruskal-Wallis test.

```
print("Manufacturer")
```

```
## [1] "Manufacturer"
```

```
print(kruskal.test(temp_data$Aircraft.Manufacturer,temp_data$Fatalities))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  temp_data$Aircraft.Manufacturer and temp_data$Fatalities
## Kruskal-Wallis chi-squared = 28.628, df = 27, p-value = 0.3792
```

```
print("Location")
```

```
## [1] "Location"
```

```
print(kruskal.test(temp_data$Location,temp_data$Fatalities))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  temp_data$Location and temp_data$Fatalities
## Kruskal-Wallis chi-squared = 24.899, df = 27, p-value = 0.5801
```

For the year 1955, neither the *Location* nor the *Manufacturer* turned out to be significant when compared against *Fatalaties* with both p-values being above 0.05.

Conclusions

Although it would have been nice to apply more advanced data analysis tests (such as 2-way ANOVA), the type of tests to be applied on the experiments is highly dependent on the data you collect and the design of the experiment. After using these datasets, it was found that some tests were not applicable, or could be computationally extensive if applied as multiple columns had over 2000 factors. An exampel of this is if 2-way ANOVA was used. In this case if the relevance of Manufacturer and Location was tested as the same time, more than one hundred unique elements (levels) per factor would be encountered. Some additional, faults with the test performed include the fact that they

In saying this, it is the belief that hypoyhesis provied is largely proved to be correct but it is highly depenedent on the time period choosen. From the research, it is also shown that air accidents are less likely to occur over time thus making a plane that crashed less likely to crash in the future. Future work would include the investigation of weather patterns in different locations and their significance on the liklihood a plane crashes.

References

[1] The Week. *The Week, Malaysia Airlines*. http://theweek.com/speedreads/449761/passengers-are-already-avoiding-malaysia-airlines

[2] Five Thirty Eight. *Five Thirty Eight, Airline Saftey*. https://github.com/fivethirtyeight/data/blob/master/airline safety/airline-safety.csv

[3] Socrata. *OpenData.Socrata, Airplane-Crashes-and-Fatalities-Since-1908* https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq

[4] Air And Space. `Airandspace Heyday` https://airandspace.si.edu/exhibitions/america-by-air/online/heyday/heyday11.cfm