Final Project, EDDA 2017

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

01 June 2017

Introduction

Air travel increased by blah% in 2016 [reference] in the USA. However, what is relevant or significant when a crash occurs. In this paper this topic will be discussed. The first question that would be interesting is to check to see if a plane that crashes is more likely to crash in a periods after ther the crash. Also additional intresting questions to consinder are if the location of the crash is signicant and 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 "fivethirtyeight" [1] which contains some general airline saftey statistics from 1980s to 2010s. The second is drawn from the opendata.socrata [2] 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 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 is a good dataset to do so, as it is really general and it stores information on two big periods which include the amount of aircrashes per airline from 1985 to 1999 and from 2000 to 2014.

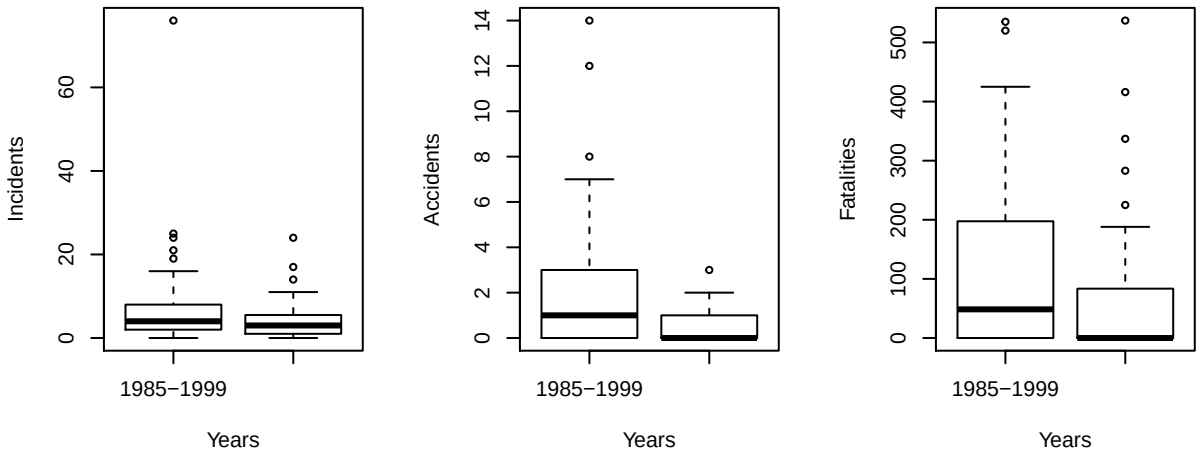
```
## The following object is masked _by_ .GlobalEnv:
##
##     airline
```

```
## [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. Per experimental unit, there are amount of incidents, fatal accidents and fatalities. That is three numerical outcomes over two periods per experimental unit. 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 remains unchanged over time without neither improvement nor deterioration.

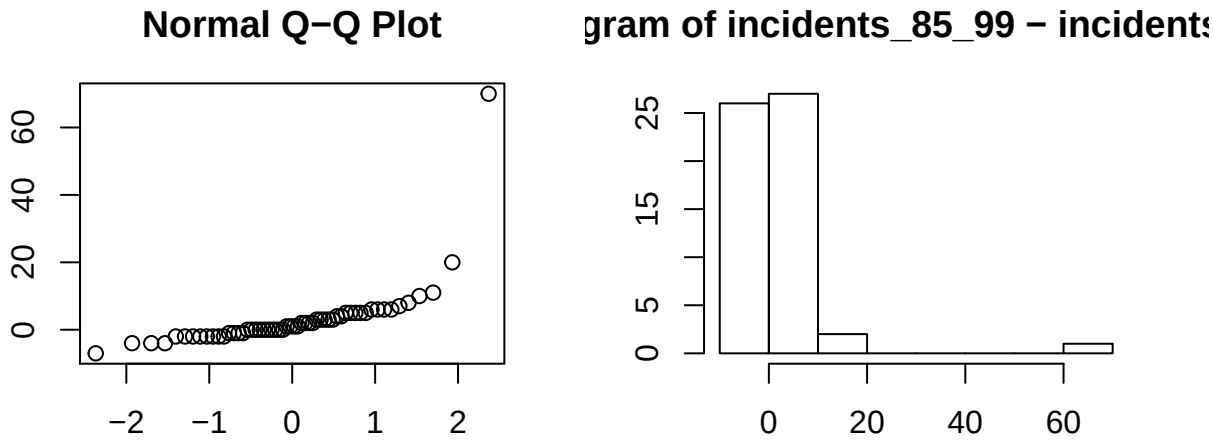
Plots



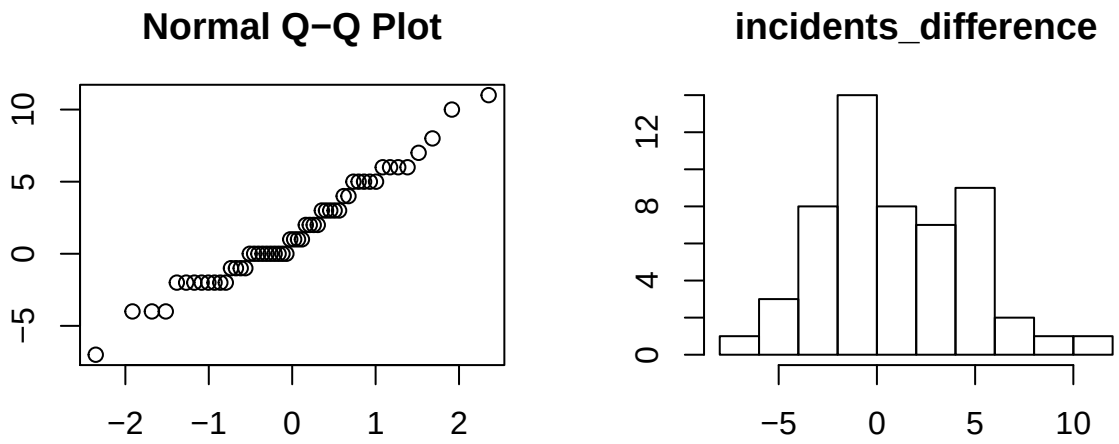
Looking at the boxplots, of the three numerical outcomes, it does seem that the hypothesis is wrong, as it seems that the number of accidents/fatalities decreases on the second period. In the following sections we will prove if this is also the case when performing statistical tests.

Incidents

Firstly, to test the hypothesis on the numerical outcome Incidents. As the experimental units are individuals, we can apply a Two-Paired test, assuming the difference between the periods comes from a normal distribution.



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



For this case, a Two-Paired test on the difference without outliers is perform 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 = 3.0084, df = 53, p-value = 0.004012
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4999342 2.5000658
## sample estimates:
## mean of x
##       1.5
```

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.

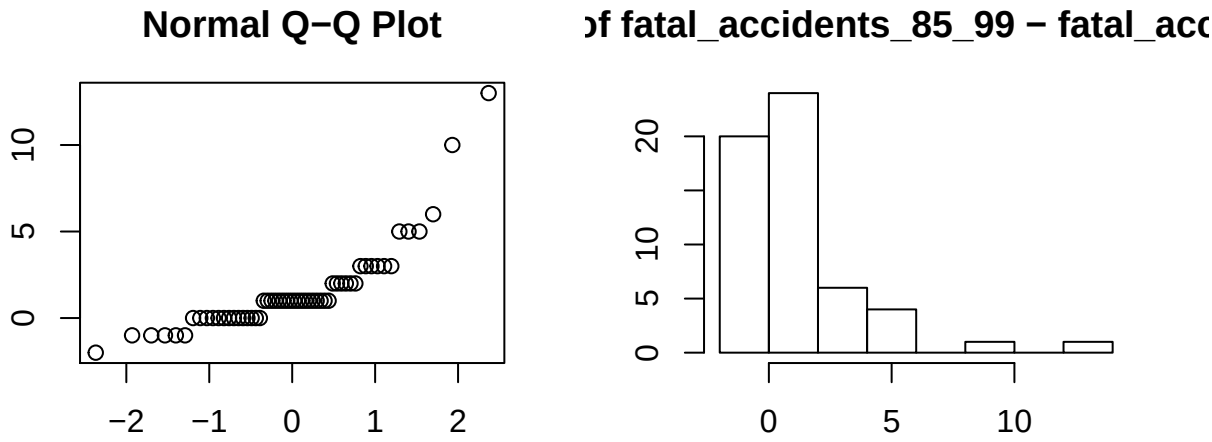
```
### Permutation Tests - No Normal distribution
permutation_test = function(mystat, col1, col2){
  B=1000
  tstar=numeric(B)
  for (i in 1:B){
    temp=t(apply(cbind(col1,col2),1,sample))
    tstar[i]=mystat(temp[,1],temp[,2])
  }
  myt=mystat(col1,col2)
  #print(myt)
  pl=sum(tstar<myt)/B
  pr=sum(tstar>myt)/B
```

```
  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)
```

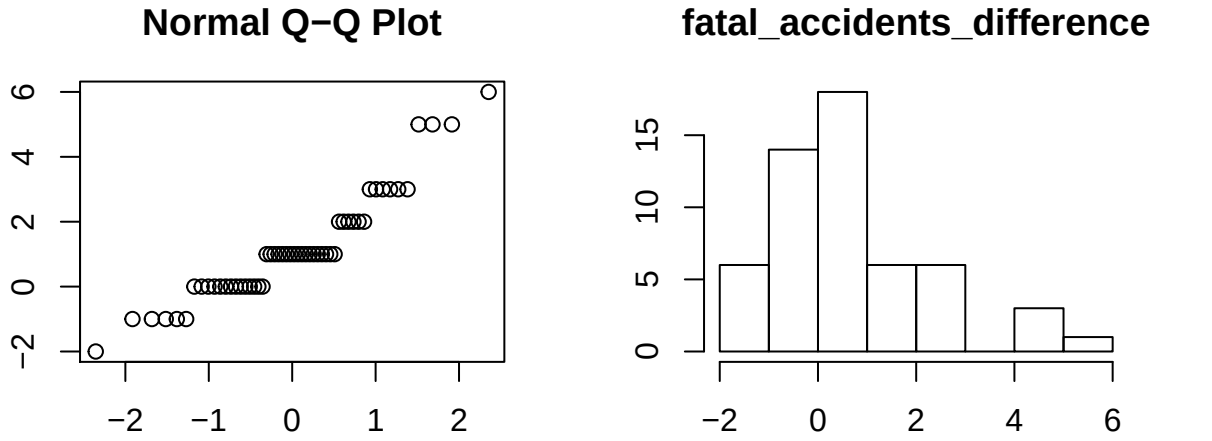
```
## [1] "Mean Diff:  3.05357142857143 P-value: 0"
```

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 state 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:



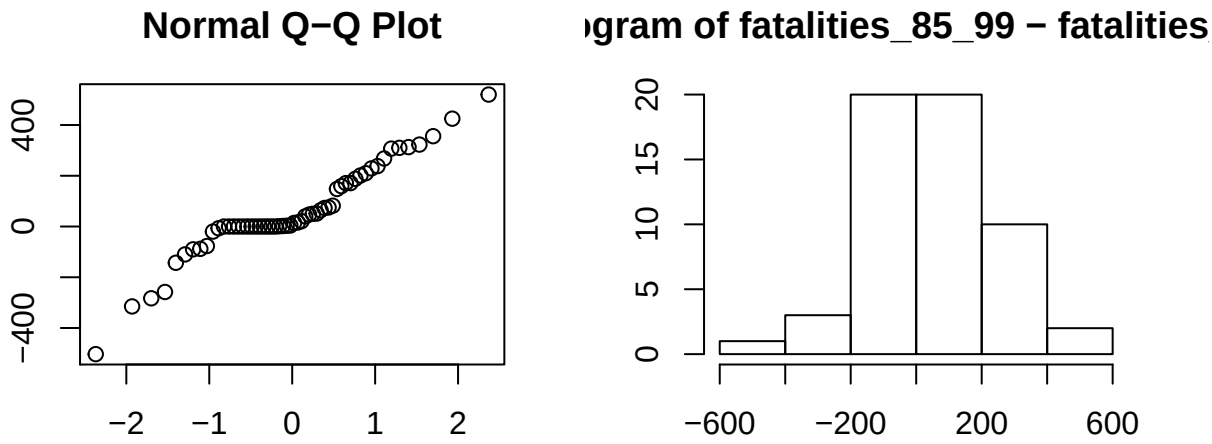
In this case, even after removing the outliers, a Normal Distribution cannot be definitively assumed. Therefore, only a Permutation Test will be carried out for this case:

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

```
## [1] "Mean Diff:  1.51785714285714 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 = 2.366, df = 55, p-value = 0.02153
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    8.704324 105.081390
## sample estimates:
## mean of the differences
##                56.89286
```

There seems to be a significant difference between fatalities in the 20th century and the 21st century, being the first significantly higher (with a confidence interval between 9 and 105 instances bigger), following the results on our 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. This has been a very general experiment, being not enough to get a deeper conclusion. Therefor in the next section, a dataset which stores more information in terms of number of experimental units and attributes will be used. With this improvement, a deeper experiment can be carried outS.

Deeper Investigation

```
airlines_data = read.csv("data/airline_data.csv", header=TRUE)
```

Data Cleaning

Due to the format of some of the data, The column "Date" needs to be cleaned as the data needs to be aggregated by year. Also invalid entires need to be filtered out.

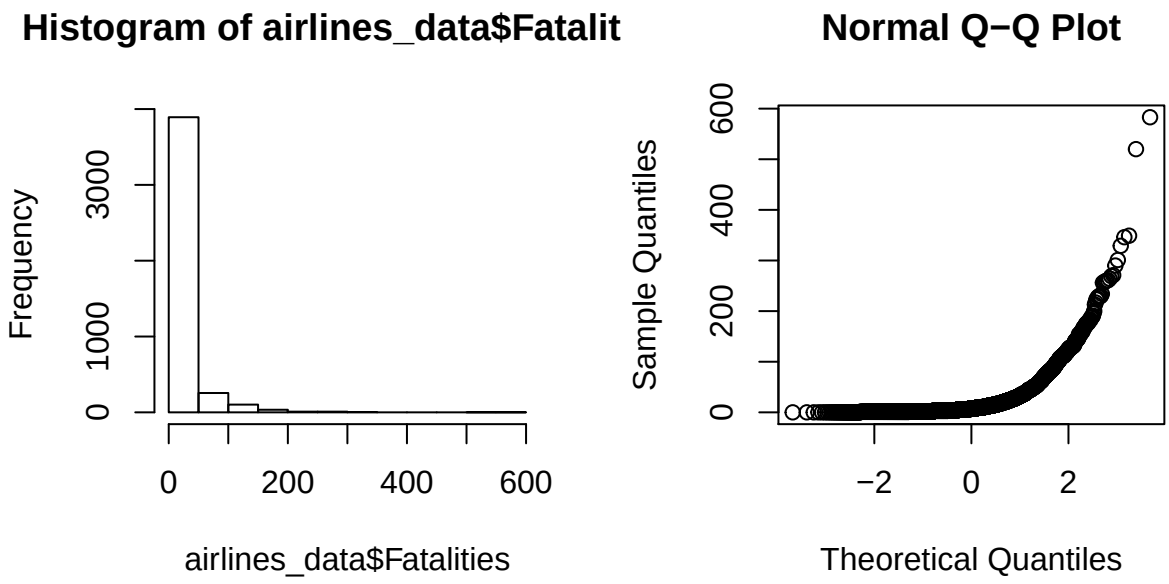
```
#delete Summary column
airlines_data=airlines_data[ , !(names(airlines_data)=='Summary')]
#transform Location column into countries
library(stringr)
temp=str_split_fixed(airlines_data$Location, ", ", 2)
countries=temp[,2]
airlines_data$Location=countries
#delete Empty Locations
airlines_data = airlines_data[!airlines_data$Location=="",]
#delete Military data
airlines_data = airlines_data[airlines_data$Classification=="Non Military",]
#transform Date to Years
temp = as.Date(airlines_data$Date,'%m/%d/%Y')
airlines_data$Date = format(temp,'%Y')
```

Data

%for Second data set

Data Distributions

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.

Setting up Experiments

To set up the experiment, a function was created to give data in the format where before and after columns could be created and used to test the research question. The function is given below:

```
setup_data <- function(date, range, data) {
    airlines = unique(data$Operator[data$Date == date])#remove filter to include airlines that
    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 selects the range plus and minus around the year given. The aim is to provide a data set where it can be determined if a plane crashes(the before column) are they more likely to crash in the near future(the after column). Serveal dates are used 1965, 1975, 1985 , 1995 and 2004. The range used will be 5 years.

Experiment

The expeiments are carries out on three year ranges 1955, 1975, 2004 with a range of 15, 10 and 5 applied repectively.

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

```
## [1] "Mean Diff:  -43.1315789473684 P-value: 0.318"
```

```
## [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 15 years after the date 1955, air transport could have becoming more popular which could have led to more crashes.

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 chekc the impact of the range of years.

```
year = 1955
range = 10
print(paste("Testing on year:", year, " Using range:", range))
```

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

```
aggregate_data = setup_data(year, range, airlines_data)
permutation_test(mystat=function(x,y) {mean(x-y)}, aggregate_data$before, aggregate_data$after
```

```
## [1] "Mean Diff:  -7.28947368421053 P-value: 0.768"
```

```
year = 1955
range = 5
print(paste("Testing on year:", year, " Using range:", range))
```

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

```
aggregate_data = setup_data(year, range, airlines_data)
permutation_test(mystat=function(x,y) {mean(x-y)}, aggregate_data$before, aggregate_data$after
```

```
## [1] "Mean Diff:  16.4210526315789 P-value: 0.154"
```

```
year = 1955
range = 2
print(paste("Testing on year:", year, " Using range:", range))
```

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

```
aggregate_data = setup_data(year, range, airlines_data)
permutation_test(mystat=function(x,y) {mean(x-y)}, aggregate_data$before, aggregate_data$after)
```

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

```
year = 1975
range = 15
print(paste("Testing on year:", year, " Using range:", range))
```

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

```
aggregate_data = setup_data(year, range, airlines_data)
permutation_test(mystat=function(x,y) {mean(x-y)}, aggregate_data$before, aggregate_data$after)
```

```
## [1] "Mean Diff:  56.9245283018868 P-value: 0.004"
```

From this we can draw two conclusions, one that the range used has a impact on the p-value. Also because the date with 1975 that the period 1950s and 1960s which is known as a golden age in aviation due to teh rise in use of comercial airlines [3]. This could be because that the technology is new making the risk of crashing higher.

Conclusions

References

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

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

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