Final Project, EDDA 2017

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

01 June 2017

Introduction

%Reserach questions %Why is it important %Whats is in the data

%Is a place a

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.

provide referenece

Setting Up Experiment

The datasets used to check this will be drawn from two sources. The first source is taken from the github repository of "fivethirtyeight" which contain airline saftey statistics. The second is drawn from the opendata.socrata site.

%https://github.com/fivethirtyeight/data/blob/master/airline-safety/airline-safety.csv %https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq

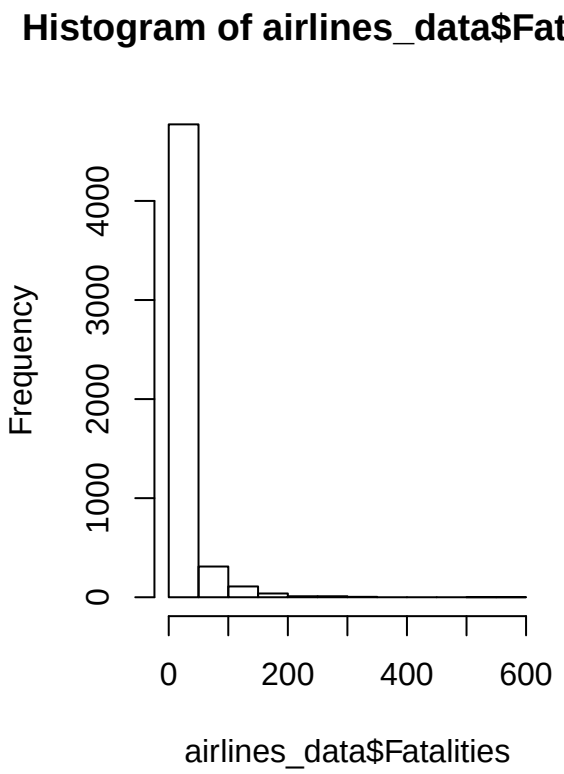
Data

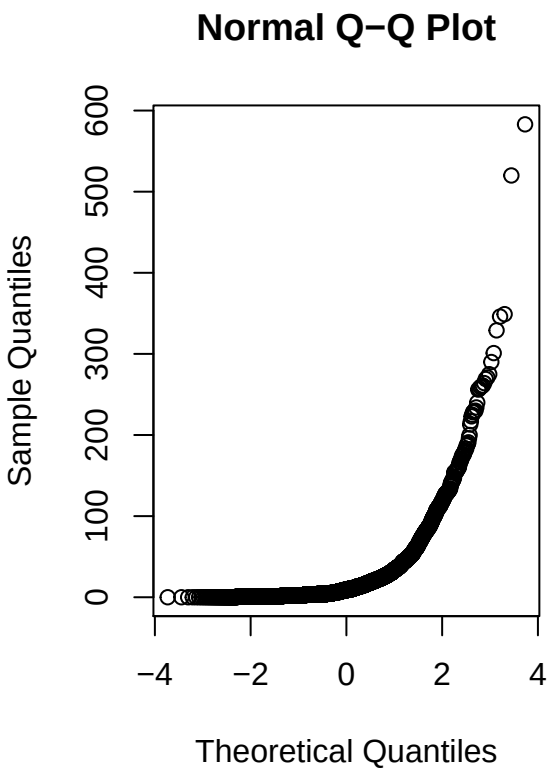
%for Second data set

Data Distributions

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

```
par(mfrow=c(1,2))
hist(airlines_data$Fatalities)
qqnorm(airlines_data$Fatalities)
```





The data is clearly not normally distrubuted which will filter out any statistical test that is based on this assumption.

Data Cleaning

Due to the format of some of the data, The column "Date" needs to be cleaned as the data is agg

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

Setting up experiment

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.

```
aggregate_data = setup_data(1965, 5, airlines_selection_data)

#boxplot(airlines_selection_data$Fat, airline$fatal_accidents_00_14)
#boxplot(airlines_selection_data$fatalities_85_99, airline$fatalities_00_14)
```

References