Plane Crash Analysis: Does your seat matter?

Inferential Statistics 2025/2026

Barbato Alberto, Naldoni Martina

2026-01-15

Abstract

Despite the high safety standards of the aviation industry, the question of whether seat selection influences survival chances during an accident remains a popular topic of debate. While media outlets often claim a statistical advantage for a specific section of the aircraft, these claims rarely rely on rigorous inferential statistics. This study investigates the relationship between seating location and mortality rates using data from 48 commercial aircraft accidents. We analyzed mortality rates across different aircraft sections (thirds and halves) and examined survivability factors including fire, restraint integrity, and energy absorption. Although preliminary visual analysis suggested a trend favoring the rear section, Kruskal-Wallis tests revealed no statistically significant difference in mortality rates between sections (p-value>0.05). Further analysis using Generalized Linear Models with the Quasibinomial family demonstrated that specific accident dynamics influence sections differently: the presence of fire and restraint failure impact the front section, while energy absorption mechanisms (landing gear status) impact the middle section. We conclude that survivability is not determined by seat location, but is influenced by accident-specific structural and environmental factors.

Contents

1 Introduction

The Aviation industry is one of the safest in the world. The systems that it has put in place to learn lessons from accidents and make sure they never happen again is well-established and highly regarded. This has made traveling by plane the safest way to travel[1].

[1] [https://flyfright.com/plane-crash-statistics/]

But perhaps due to these high safety standards, when an aircraft accident happens it makes headlines all over the world. Anxious passengers fear it's going to happen to them as well and they can't help but ask themselves if there is something they could do to have a safer flight.

One of the most frequent question that gets asked is : is there a part of the plane that is "safer" than other parts? Does your seat matter?

We are going to try to answer this question using statistics.

1.1 Vocabulary

Before starting the analysis, the meaning of some terms must be clarified

- **Flight** : A flight is a trip made by an aircraft that connects two airports. It is identified by a number. (Example : AirFrance 225 is the name of the regular service from New Deli to Paris)
- **Accident** : An Accident is an occurrence where a person is fatally or seriously injured, the aircraft sustains significant structural damage, or the aircraft goes missing.
- **Incident** : An incident is a dangerous situation where no one is seriously hurt and the plane isn't badly damaged.
- **Crash** : A crash is a type of accident where an aircraft strikes the ground, water, or an obstacle with enough force to cause severe damage or total destruction.
- **Accident flight** : when a serious accident happens on a flight, usually the companies save the flight number to refer to that flight, and change the flight number for new flights. (Example : Air France 447 was the name of the connection from Rio to Paris, after the 2009 accident the flight became Air France 445)
- **Final Report** : The final report is the result of the investigation conducted by an aviation safety board. It is an official document where the results of the investigation into the cause of the accident are reported along with many other informations pertinent to the analysis that the investigators conducted.
- **Serious injury** : Any injury requiring more than 48 hours of hospitalization or involving broken bones (other than fingers/toes)
- **Survivability** : an accident is defined as survivable if the forces are within human limits and if the structure remained substantially intact
- **CREEP factors** : factors that influence the survivability of an accident. They are
  - *Container* : amount of cabin deformation
  - *Restraints* : analysis of the seating structure
  - *Environment* : presence of lethal contact points
  - *Energy Absorption* : G-load mitigation by fuselarge or landing gear
  - *Post Crash* : Other factors (We considered : airport proximity, presence of fire, daytime, phase of flight)

1.2 Literature

While we found no scientific article about this issue, there are many news articles that claim to have conducted a "statistical analysis" on aircraft accident data to find which area is the safest in case of a crash. We now list some of their findings :

- *Time Article*[2] : "Statistics show that the middle seats in the rear of an aircraft historically have the highest survival rates. [...] The analysis found that the seats in the back third of the aircraft had a 32% fatality rate, compared with 39% in the middle third and 38% in the front third.". This article was written in June 2015.

[2] [https://time.com/3934663/safest-seat-airplane/]

- *Reuters Article*[3] : "Data reveals civil aviation's most astonishing, exceptional survivals—and shows no seat is reliably safe."
- *Aeroclass Article*[4] : "An analysis of several accidents reveals that:The seats in the back third of the aircraft have a fatality rate of 32%. The seats in the middle third of the plane have a fatality rate of 39%. The seats in the front third have a fatality rate of 38%. [...] The crash data indicates that the front third and middle third of the plane have higher fatality rates than the back third of the plane." This article was written in November 2017.
- *Wired Article*[5] : "While no part of the plane may generally be the safest [...] Each airline emergency plays out differently, affecting different seats more than others each time."
- *Allianz Report - Popular Mechanics*[6] : Popular Mechanics also examined 20 accidents and calculated the survival rate in each of four sections of the aircraft. Its results found that in 11 of the 20 crashes, passengers in the rear of the aircraft had a better chance of survival. In seven of those 11 crashes they found the rear section was the only section with survivors. In five accidents the first class and business class section fared the best with a 49% survivability rate. In three out of the 20 crashes no location had an advantage.
- *Allianz Report - University of Greenwich*[7] : the University of Greenwich studied 105 airline accidents worldwide, and this study concluded that the safest seat on an aircraft is in the one on the aisle nearest the exit, in the front of the aircraft. This seat has a survivability rate of 65% whereas a passenger seated in the rear section only has a 53% survivability rate. Additionally, any seat in the aisle near an exit offers a greater chance of survivability. When seated more than six rows from an exit "the chances of perishing far outweigh those of surviving"

We can see that they portray the reality in different ways. We wanted to conduct a similar analysis using statistical tools so that we can get statistical guarantees on the results we get.

1.3 Perspective

We are studying if the seating in an airplane has an effect on the survivability of an aircraft accident. To study this, we need to look at all aircraft accidents, then rule out the accidents where every passenger survived, and every accident in which every passenger died. This is an extremely narrow data set of accidents. Of these accidents, we gathered the seatings arrangements and survivor seating maps for 48 crashes.

If 48 crashes look like a lot, we should consider the greater perspective of air travel safety in general.

We can get a clear picture of just how rare an accident is by looking into the US General aviation data between 2015 and 2020[8] :

- $\frac{1}{260256}$: chance of boarding any flight and it being an accident flight
- $\frac{1}{6{,}864{,}250}$: chance of being on a plane involved in an accident that results in at least 1 fatality (possible case study of this study)
- $\frac{1}{816{,}545{,}929}$ chance of you specifically, dying in a plane crash

1.4 Data Gathering Process

For this experiment, we found that there weren't any available datasets to get the accident seating maps from. We decided that we would gather the information from the final reports of the accidents. We also

[3] [https://www.reuters.com/graphics/AVIATION-SAFETY/lgpdaagabvo/]
[4] [https://www.aeroclass.org/the-safest-place-to-sit-on-a-plane/]
[5] [https://www.wired.com/story/whats-the-safest-seat-on-an-airplane/]
[6] [https://www.allianz.com/content/dam/onemarketing/azcom/Allianz_com/migration/media/press/document/other/AGCS-Global-Aviation-Safety-Study-2014.pdf]
[7] [https://www.allianz.com/content/dam/onemarketing/azcom/Allianz_com/migration/media/press/document/other/AGCS-Global-Aviation-Safety-Study-2014.pdf]
[8] [https://flyfright.com/plane-crash-statistics/#tve-jump-18c020d9166]

gathered data from Wikipedia after checking that the source was indeed the final report of the investigation into the accident.

We examined aircraft accidents where there 3 or more deaths, and more than 30 people overall onboard, to be able to divide the aircraft into sections and study them. Of 597 accidents that happened from 1977 to 2019, we found 105 that met this criteria. Of these, we found seat injury maps for 48 of them. This was due to the kind of information contained in the final reports.

The final reports don't always feature a seat injury map built in a way that allows us to conduct a statistical analysis on it. Sometimes, the map is not even present and the only information is a rough idea of the location of the survivors and deaths. This is an extract from the final report of eastern airlines flight 401 :

- *"Most of the survivors were located in the vicinity of the cockpit area, the mid-cabin service area, the overwing area, and the empennage section; these sections were located at the far end of the wreckage path. In contrast, most fatalities were found in the center of the crash path. Crushing injuries to the chest were the predominant causes of death."*[9]

We know that 101 people died in this accident and that 75 survived, but we can't assign a precise mortality rate to each section, so we had to discard this accident from our data sample.

In some cases, only the locations of the deaths were marked, thus making a statistical analysis impossible since the total number of passengers per section was unknown.

Whenever a full seat x injury map was featured in the report, we calculated by hand the total passenger counts. We intentionally left out the crew from our totals, since this analysis is focused on passengers. We decided that to be able to get some insight on the safest part of the aircraft, we should not use single seats but rather chunks of the aircraft, as all aircraft are different and can't be compared seat by seat. Not knowing if dividing into thirds or halves, we inserted both these information in our dataset.

We collected the additional information about the accident details from the aviation safety network database[10] and from Wikipedia[11]

2 Data Description

Let's take a look at the structure and meaning of the data we gathered:

```
accident_data <- read.csv("data/All_Plane_Crash_Data.csv")
str(accident_data)
```

```
## 'data.frame':    48 obs. of  26 variables:
##  $ Airline         : chr  "singapore airlines" "british airtours" "british midland" "china airlines"
##  $ FlightNum       : int  6 28 92 120 123 129 140 148 191 204 ...
##  $ X1.third.minor  : int  17 36 0 0 0 4 0 0 0 3 ...
##  $ X1.third.major  : int  2 0 11 0 0 0 0 0 0 0 ...
##  $ X1.third.dead   : int  15 0 22 0 136 14 18 16 55 33 ...
##  $ X2.third.minor  : int  1 30 4 5 0 5 7 1 0 25 ...
##  $ X2.third.major  : int  15 0 30 0 0 0 0 0 8 0 ...
##  $ X2.third.dead   : int  64 16 13 8 214 60 139 34 51 1 ...
##  $ X3.third.minor  : int  26 10 0 5 36 24 0 8 10 29 ...
##  $ X3.third.major  : int  17 0 27 0 0 0 0 0 7 0 ...
##  $ X3.third.dead   : int  0 36 11 21 109 43 91 30 16 0 ...
##  $ X1.half.minor   : int  17 56 32 3 0 7 7 1 0 7 ...
##  $ X1.half.major   : int  3 0 0 0 0 0 0 0 1 0 ...
##  $ X1.half.dead    : int  41 8 34 6 226 23 95 36 79 35 ...
##  $ X2.half.minor   : int  26 20 39 7 4 26 0 7 10 46 ...
```

[9] [http://libraryonline.erau.edu/online-full-text/ntsb/aircraft-accident-reports/AAR73-14.pdf]
[10] [https://aviation-safety.net/asndb/]
[11] [https://en.wikipedia.org/wiki/List_of_accidents_and_incidents_involving_commercial_aircraft]

```
##  $ X2.half.major   : int  31 0 0 0 0 0 0 0 14 0 ...
##  $ X2.half.dead    : int  34 44 13 23 225 90 145 46 48 0 ...
##  $ DataOrigin      : chr  "w" "w" "w" "w" ...
##  $ PhaseOfFlight   : chr  "takeoff" "takeoff" "landing" "landing" ...
##  $ Time            : chr  "night" "day" "night" "day" ...
##  $ Place           : chr  "airport" "outside" "outside" "airport" ...
##  $ HasFire         : chr  "fire" "fire" "fire" "fire" ...
##  $ CrushedFuselage : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ RestraintIntact : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ Environment     : chr  "dangerous" "dangerous" "dangerous" "clear" ...
##  $ EnergyAbsorption: chr  "nogear" "nogear" "gear" "nogear" ...
```

2.1 Data Points and Variables

The data consists of 48 observations (aircraft accidents) and 26 variables. They are the focus of our analysis.

We want to remark that the data consists only of aircraft accidents where there were at least 3 fatalities and more than 30 passengers onboard, because the final goal was to compare mortality rates. In this section we give more detail on the additional variables gathered for each accident

The variables are:

- **Airline & FlightNum** : identifier of the aircraft accident. Useful to retrieve more information about the accident if needed.
- **X variables** : these variables represent the number and the type of injury (minor-none / serious / fatal) in each section of the airplane (Each third and each half). The format is the following : `X{section}.{part}.{type of injury}`, where section is 1,2,3 for the thirds and 1,2 for the halves, part is "third" or "half", and type of injury is "minor", "major" or "dead".
- **PhaseOfFlight** : phase of flight when the accident happened, The possible values are "takeoff" or "landing", landing includes emergency landings.
- **Time** : time of the day when the accident happened, possible values are "day" or "night". This is intended to signal if natural light was present or not.
- **Place** : location where the accident happened. Possible values are "airport" or "outside". "Airport" means that the accident happened within airport boundaries (where the airport fire rescue service arrive in less that 3 minutes), "outside" means that the accident happened outside an airport.
- **HasFire** : indicates if there was an uncontained fire after the crash that spread to the cabin. Possible values are "fire" or "no-fire".
- **CrushedFuselage** : indicates if the fuselage was crushed, so deformed with significant loss of volume, during the accident. Possible values are 1 (yes) or 0 (no).
- **RestraintIntact** : if the seating structure remained intact and the seat belt system functioned correctly, the Restraint system is considered to have functioned correctly. Possible values are 1 (yes) or 0 (no).
- **Environment** : indicates the state of the cabin after the crash. Possible values are "clear" or "dangerous". It is considered dangerous if it was determined that parts of the cabin like overhead bins failed and debris was scattered in the cabin.
- **EnergyAbsorption** : indicates if the landing gear was extended and if it absorbed a significant amount of impact forces before collapsing. If not, the airplane either hit the ground wit the gear retracted or hit the ground nose-first or tail-first, causing the impact forces to be absorbed by the passenger cabin. Possible values are "gear" or "nogear"

2.2 Mortality Rates

To start the analysis, let's preprocess the data by creating the mortality rates for each section. We first proceed to add the totals for each section, and then use it to generate the mortality rates. We also transform the categorical variables into factors.

```
# add a colum of # of passengers for each airplain section
accident_data <- within(accident_data, {
  X2.half.total  <- X2.half.minor  + X2.half.major  + X2.half.dead
  X1.half.total  <- X1.half.minor  + X1.half.major  + X1.half.dead
  X3.third.total <- X3.third.minor + X3.third.major + X3.third.dead
  X2.third.total <- X2.third.minor + X2.third.major + X2.third.dead
  X1.third.total <- X1.third.minor + X1.third.major + X1.third.dead
})

# now we transform the variables that are categorical into factors
factor_cols <- c("PhaseOfFlight", "Time", "Place", "HasFire", "Environment",
                 "EnergyAbsorption", "CrushedFuselage", "RestraintIntact")

accident_data[factor_cols] <- lapply(accident_data[factor_cols], as.factor)
```

Now we can add a new variable that will be the center of our analysis : the mortality rate. It is measured for each section of the airplane. The mortality rate is defined as the number of deaths divided by the total number of passengers in the section, for each section. Note that aircraft don't always fly on a full load of passengers, so there might be sections with very few passengers or possibly empty. That's why we are considering rates per total passengers instead of per number of seats.

```
accident_data <- within(accident_data, {
  X1.third.mortality.rate <- X1.third.dead / X1.third.total
  X2.third.mortality.rate <- X2.third.dead / X2.third.total
  X3.third.mortality.rate <- X3.third.dead / X3.third.total
  X1.half.mortality.rate  <- X1.half.dead  / X1.half.total
  X2.half.mortality.rate  <- X2.half.dead  / X2.half.total
  Total.mortality.rate <- (X1.third.dead + X2.third.dead + X3.third.dead) /
                          (X1.third.total + X2.third.total + X3.third.total)
})
```

3 Data Analysis

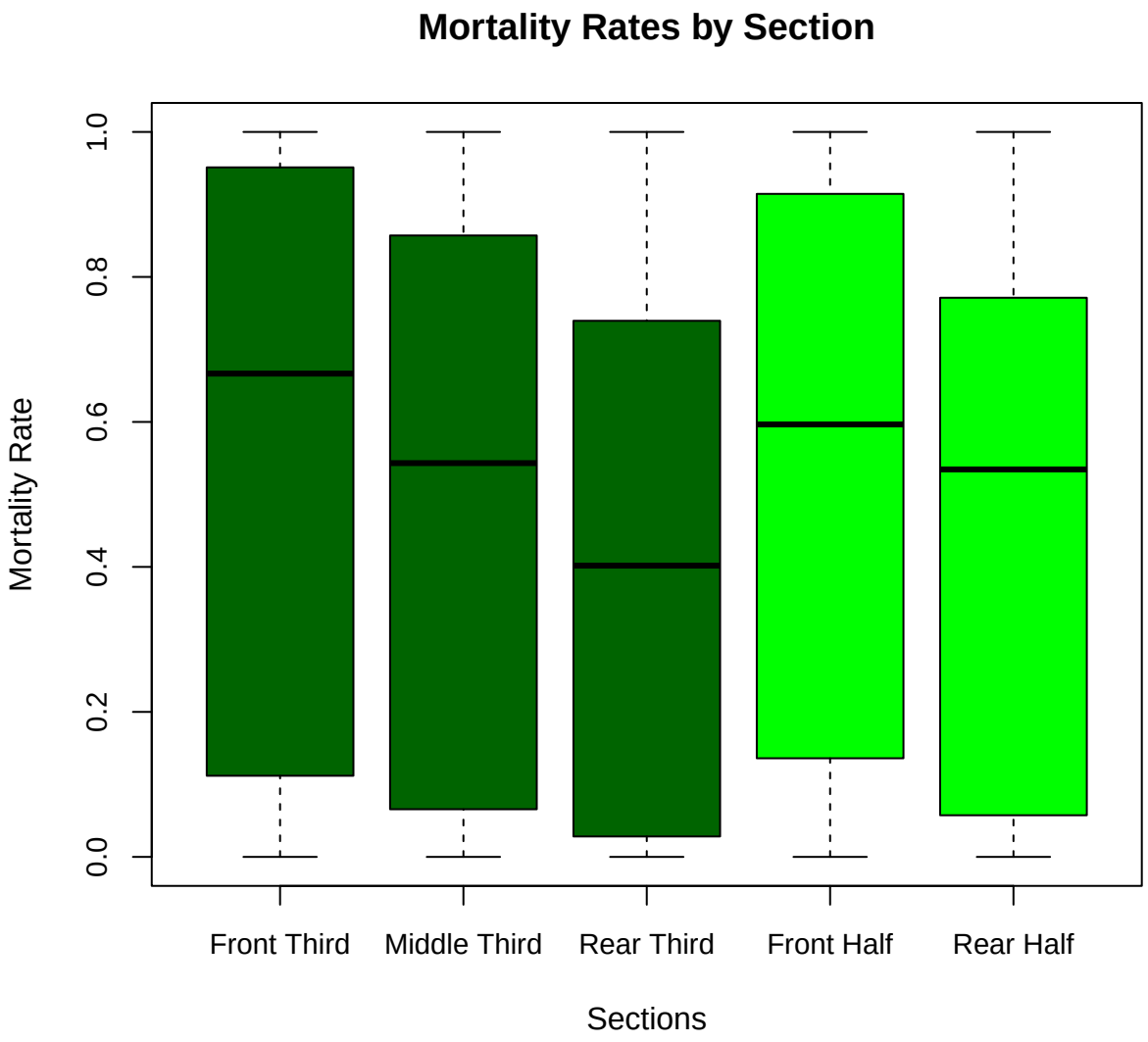
3.1 Preliminary analisys

3.1.1 Plotting

The first step we can do is plotting some features of the data to have a visual understanding of it. We do this to get some useful insights and also check for possible outliers or anomalies.

```
with(accident_data, {
  boxplot(X1.third.mortality.rate,
          X2.third.mortality.rate,
          X3.third.mortality.rate,
          X1.half.mortality.rate,
          X2.half.mortality.rate,
          names = c("Front Third", "Middle Third", "Rear Third", "Front Half", "Rear Half"),
          main = "Mortality Rates by Section",
          ylab = "Mortality Rate",
          xlab = "Sections",
```

```
          col = c("darkgreen", "darkgreen", "darkgreen", "green", "green"),
          cex.axis = 0.9)
})
```



Intestingly, there seems to be a general trend in the mortality rates : the more rear the section is, the lower the mortality rate. This is seen in both the third and half divisions of the airplane.

3.1.2 Distribution

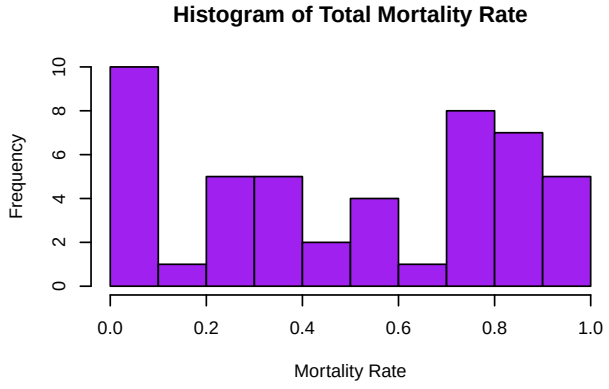
Now we can try to find the kind of population distribution of the mortality rates is. This will help us choose the right statistical tests for our analysis. We test to see if the data follows a normal distribution with the Shapiro-Wilk test.

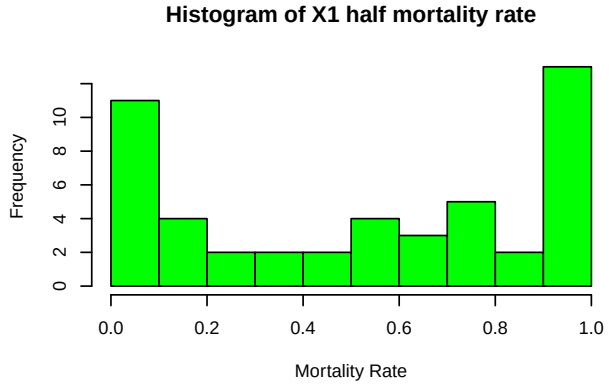
```
# shapiro test for normality, p-value extrated and labeled
shapiro_results <- data.frame(
  Section = c("Front Third", "Middle Third", "Rear Third", "Front Half", "Rear Half"),
  P_Value = c(
    shapiro.test(accident_data$X1.third.mortality.rate)$p.value,
    shapiro.test(accident_data$X2.third.mortality.rate)$p.value,
    shapiro.test(accident_data$X3.third.mortality.rate)$p.value,
    shapiro.test(accident_data$X1.half.mortality.rate)$p.value,
```

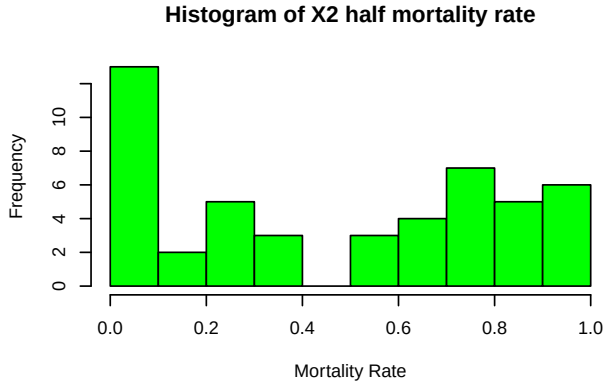
```
    shapiro.test(accident_data$X2.half.mortality.rate)$p.value
  )
)
shapiro_results
```

```
##        Section      P_Value
## 1  Front Third 1.629503e-05
## 2 Middle Third 4.224501e-05
## 3   Rear Third 1.354496e-04
## 4   Front Half 1.100865e-04
## 5    Rear Half 1.712003e-04
```

All of the p-values are far below the standard threshold of 0.05, indicating that we can reject the null hypothesis of normality, for all sections. Let's try to plot a histogram for all of the sections to visualize an approximation of the distribution.







It seems that there is no clear distribution pattern. What we can see is that the data is more concentrated at both ends of the mortality rate values, especially for the front third section. It seams like in most accidents, either most people die or most people survive, with few accidents in the middle ground.

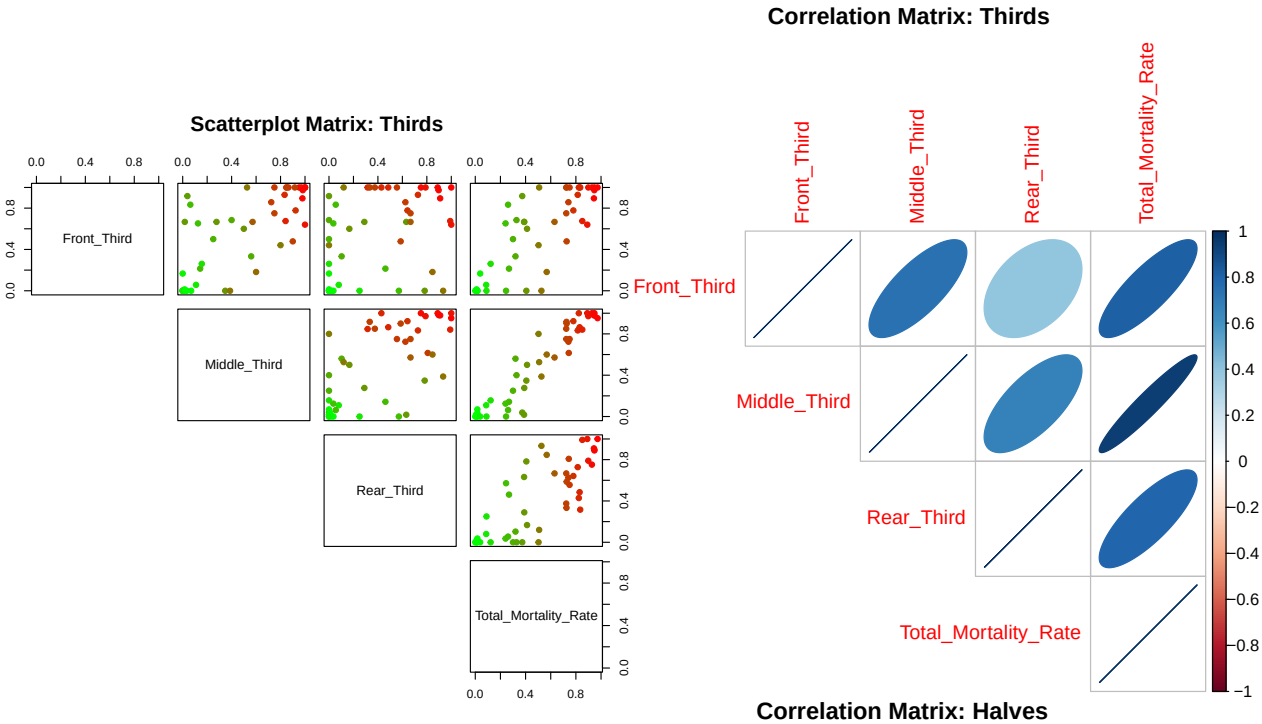
3.2 Correlation between sections

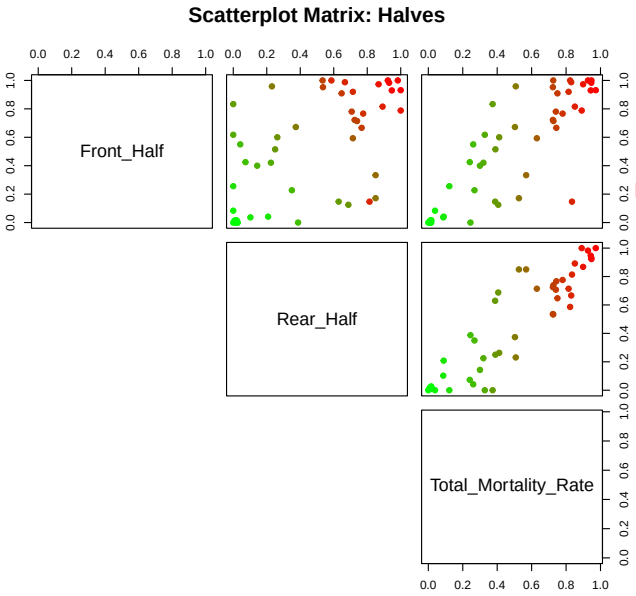
Since the mortality rates are from the same accidents, it is possible that there is some correlation between them. To visualize this, we can create a scatterplot matrix of the mortality rates of the different sections. We will also color the points based on the total mortality rate of the accident, to se if there is some pattern.

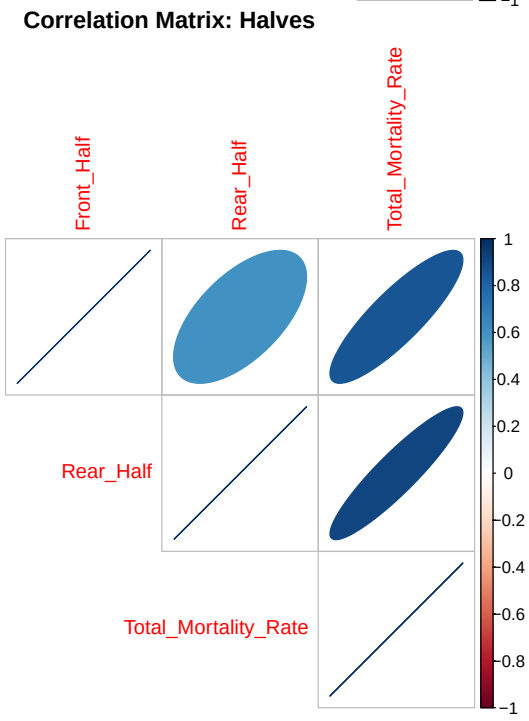
```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
data_thirds <- data.frame(
  Front_Third = accident_data$X1.third.mortality.rate,
  Middle_Third = accident_data$X2.third.mortality.rate,
  Rear_Third = accident_data$X3.third.mortality.rate,
  Total_Mortality_Rate = accident_data$Total.mortality.rate
)

data_halves <- data.frame(
  Front_Half = accident_data$X1.half.mortality.rate,
  Rear_Half = accident_data$X2.half.mortality.rate,
  Total_Mortality_Rate = accident_data$Total.mortality.rate
)

# Create a color palette from green to red
rbPal <- colorRampPalette(c("green", "red"))

# Generate colors based on Total_Mortality_Rate
# We can use the Total from accident_data directly to ensure consistent coloring
point_colors <- rbPal(100)[as.numeric(cut(accident_data$Total.mortality.rate, breaks = 100))]


# Helper function for Scatterplots
plot_my_scatter <- function(data, title) {
  pairs(data,
        main = title,
        pch = 19,
        col = point_colors,
        lower.panel = NULL)
}

# Helper function for Correlation Plots
plot_my_corr <- function(data, title) {
  corr_res <- cor(data, use = "complete.obs")
  corrplot(corr_res, method = "ellipse",
           title = title,
           mar = c(0,0,2,0),
           type = "upper")
}

plot_my_scatter(data_thirds, "Scatterplot Matrix: Thirds")
plot_my_corr(data_thirds, "Correlation Matrix: Thirds")

plot_my_scatter(data_halves, "Scatterplot Matrix: Halves")
plot_my_corr(data_halves, "Correlation Matrix: Halves")
```







As we can see from both the scatterplot matrix and the correlation plot, there is a strong positive correlation between the mortality rates of the different sections and the total mortality rate. Also there are no evident asymmetries on the color distribution. The correlation with the total mortality rate is very high for all sections, indicating that when the total mortality rate is high, all sections tend to have high mortality rates, which is expected, as the increase of mortality rate in a section also leads to the increase of the total mortality rate.

We notice that the correlation is weakest between the front third and the rear third: in fact, the more 2 sections are distant physically, the less correlated they are in this data; also similar distance leads to similar correlation values. This is a sign that probably there is "a safer zone", since it looks like the mortality rates

for distant sections are not as correlated as they are for closer sections.

This, and the decreasing trend in the boxplots seems like the reason why most articles halted their analysis, since they thought that a difference in mortality rates and correlation meant a significant difference in survivability.

We want to base our conclusion on statistical evidence, so we now conduct further statistical analysis to determine if these results are actually significant.

3.3 Statistical Analysis of differences in mortality rates between sections

Since the distribution of the mortality rates is not normal, we will not use tests that assume normality, like t or ANOVA tests. We will instead use the Kruskal-Wallis test, which is a non-parametric test to compare the distribution of two or more groups. The null hypothesis of the Kruskal-Wallis test is that all the populations have the same distribution.

```
# Kruskal-Wallis test for differences in mortality rates between sections (only thirds)
mortality_data_third <- data.frame(
  SectionThird = rep(c("Front Third", "Middle Third", "Rear Third"), each = nrow(accident_data)),
  MortalityRateThird = c(accident_data$X1.third.mortality.rate,
                   accident_data$X2.third.mortality.rate,
                   accident_data$X3.third.mortality.rate)
)

kruskal_result_third <- kruskal.test(MortalityRateThird ~ SectionThird, data = mortality_data_third)
kruskal_result_third
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  MortalityRateThird by SectionThird
## Kruskal-Wallis chi-squared = 3.4175, df = 2, p-value = 0.1811
```

```
# Kruskal-Wallis test for differences in mortality rates between sections (only halves)
mortality_data_half <- data.frame(
  SectionHalf = rep(c("Front Half", "Rear Half"), each = nrow(accident_data)),
  MortalityRateHalf = c(accident_data$X1.half.mortality.rate,
                   accident_data$X2.half.mortality.rate)
)

kruskal_result_half <- kruskal.test(MortalityRateHalf ~ SectionHalf, data = mortality_data_half)
kruskal_result_half
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  MortalityRateHalf by SectionHalf
## Kruskal-Wallis chi-squared = 0.80032, df = 1, p-value = 0.371
```

The Kruskal-Wallis tests returned p-values of 0.1811 (for thirds) and 0.371 (for halves). In both cases, the p-value is above the standard significance level of 0.05. Therefore, we cannot reject the null hypothesis : despite the visual trends observed in the boxplots, our statistical analysis indicates that there is no significant difference in mortality rates between the different sections of the aircraft, no matter if is divided into thirds or halves. The trend that we observed in the boxplots could be caused just by random variation in the data, and not by a real effect of the seating location on mortality rates. There could also be a problem of low statistical power due to the small sample size and/or the high variability of the data observed in the histograms.

3.4 Modelling the mortality rates

Now we will try to fit a model to possibly discover the effect of the other variables on the mortality rates of the different sections. The focus is not specifically on prediction, but rather on trying to find a significant difference between the sections in light of their correlation with the additional variables. Now the questions are:

- Are there variables that significantly affect the mortality rate?
- If so, do they affect differently the various sections of the airplane?

We will try to fit a Generalized Linear Model (GLM) with binomial family, since the mortality rate is a proportion (number of deaths / total number of passengers). Unfortunately, because of the nature of the data, the assumptions of the binomial distribution are not met, since the passenger's deaths are not independent events.

To account for this, we will use the quasibinomial family, which is a more flexible version of the binomial family that allows for overdispersion and corrects the standard errors and p-values accordingly.

3.4.1 Technical Note: The Quasibinomial Family

The standard binomial model assumes that every observation is independent and it assumes a relationship between the mean and the variance. For a probability p and n trials, the variance is expected to be:

$$Var(Y) = n \cdot p(1-p)$$

In aircraft accidents, mortality data is clustered. Often, either everyone survives or everyone dies. This causes the observed variance in the data to be much higher than what the binomial formula predicts. This phenomenon is called **Overdispersion**. If we ignored this, our model would be overconfident (p-values artificially low), meaning we will find "significance" where there is none.

The quasibinomial model does not define a specific probability distribution. Instead, it defines a relationship between the mean and the variance using a Dispersion Parameter ϕ.

The new variance formula becomes:

$$Var(Y) = \phi \cdot n \cdot p(1-p)$$

- If $\phi = 1$: It behaves exactly like a standard Binomial model.
- If $\phi > 1$: It indicates overdispersion. The model "penalizes" our results by scaling up the standard errors, and thus lowers the pvalues, by a factor of $\sqrt{\phi}$.

The estimated coefficients remain the same as a standard logistic regression, but the standard errors are larger, resulting in wider confidence intervals and more conservative p-values.

3.4.2 GLM models and results

We fit GLMs using the mortality rate as response and our additional variables as covariates, one for each section. We also modeled the total mortality rate, to see if there are variables that affect the overall survivability of the accident, regardless of seating location. We choose to study only the thirds, since they give a more granular view of the aircraft sections, and they where the most promising in the preliminary analysis.

```
predictors <- ~ PhaseOfFlight + Time + Place + HasFire + Environment +
                EnergyAbsorption + CrushedFuselage + RestraintIntact


glm_third_1 <- glm(update(predictors,
                          cbind(X1.third.dead, X1.third.total - X1.third.dead) ~ .),
                   family = quasibinomial(link = "logit"),
```

```
                  data = accident_data)

glm_third_2 <- glm(update(predictors,
                          cbind(X2.third.dead, X2.third.total - X2.third.dead) ~ .),
                  family = quasibinomial(link = "logit"),
                  data = accident_data)

glm_third_3 <- glm(update(predictors,
                          cbind(X3.third.dead, X3.third.total - X3.third.dead) ~ .),
                  family = quasibinomial(link = "logit"),
                  data = accident_data)

glm_total <- glm(update( predictors, cbind(X1.third.dead + X2.third.dead + X3.third.dead,
                       X1.third.total + X2.third.total + X3.third.total -
                         (X1.third.dead + X2.third.dead + X3.third.dead)) ~ .),
                  family = quasibinomial(link = "logit"),
                  data = accident_data)
```

```
summary(glm_third_1)
```

```
## Warning in summary.glm(glm_third_1): observations with zero weight not used for
## calculating dispersion
```

```
##
## Call:
## glm(formula = update(predictors, cbind(X1.third.dead, X1.third.total -
##     X1.third.dead) ~ .), family = quasibinomial(link = "logit"),
##     data = accident_data)
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -0.8762     3.1426  -0.279   0.7819
## PhaseOfFlighttakeoff     -0.9204     0.6463  -1.424   0.1626
## Timenight                 0.7204     0.6901   1.044   0.3031
## Placeoutside              0.5515     0.7305   0.755   0.4550
## HasFireno-fire           -1.5879     0.7779  -2.041   0.0482 *
## Environmentdangerous     -0.8939     1.4475  -0.618   0.5405
## EnergyAbsorptionnogear    0.7601     0.8436   0.901   0.3733
## CrushedFuselage1          2.2791     2.7051   0.843   0.4048
## RestraintIntact1         -2.6441     1.2698  -2.082   0.0441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 17.31082)
##
##     Null deviance: 1359.75  on 46  degrees of freedom
## Residual deviance:  771.33  on 38  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm_third_2)
```

```
##
## Call:
```

```
## glm(formula = update(predictors, cbind(X2.third.dead, X2.third.total -
##     X2.third.dead) ~ .), family = quasibinomial(link = "logit"),
##     data = accident_data)
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.8506     2.6939  -1.058   0.2965
## PhaseOfFlighttakeoff     -0.4293     0.6955  -0.617   0.5407
## Timenight                 0.4988     0.7534   0.662   0.5118
## Placeoutside             -0.6477     0.7995  -0.810   0.4228
## HasFireno-fire           -1.1185     0.8292  -1.349   0.1852
## Environmentdangerous      1.1840     1.1083   1.068   0.2919
## EnergyAbsorptionnogear    2.2160     0.9191   2.411   0.0207 *
## CrushedFuselage1          1.9408     2.4099   0.805   0.4255
## RestraintIntact1         -2.1980     1.2674  -1.734   0.0908 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 36.18404)
##
##     Null deviance: 2373.6  on 47  degrees of freedom
## Residual deviance: 1531.2  on 39  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
summary(glm_third_3)
```

```
##
## Call:
## glm(formula = update(predictors, cbind(X3.third.dead, X3.third.total -
##     X3.third.dead) ~ .), family = quasibinomial(link = "logit"),
##     data = accident_data)
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.1141     2.6286  -1.185   0.2433
## PhaseOfFlighttakeoff     -0.6745     0.5801  -1.163   0.2520
## Timenight                -0.3653     0.6512  -0.561   0.5781
## Placeoutside             -0.6051     0.6643  -0.911   0.3680
## HasFireno-fire           -1.0076     0.7776  -1.296   0.2026
## Environmentdangerous      1.4433     1.0758   1.342   0.1875
## EnergyAbsorptionnogear    1.3123     0.7144   1.837   0.0738 .
## CrushedFuselage1          2.0651     2.3224   0.889   0.3793
## RestraintIntact1         -0.8132     1.0045  -0.810   0.4231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 27.59401)
##
##     Null deviance: 1587.8  on 47  degrees of freedom
## Residual deviance: 1261.4  on 39  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(glm_total)
```

```
##
## Call:
## glm(formula = update(predictors, cbind(X1.third.dead + X2.third.dead +
##     X3.third.dead, X1.third.total + X2.third.total + X3.third.total -
##     (X1.third.dead + X2.third.dead + X3.third.dead)) ~ .), family = quasibinomial(link = "logit"),
##     data = accident_data)
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.6730     2.3407  -1.142   0.2604
## PhaseOfFlighttakeoff     -0.6196     0.5460  -1.135   0.2633
## Timenight                 0.2134     0.5855   0.364   0.7175
## Placeoutside             -0.3342     0.6106  -0.547   0.5873
## HasFireno-fire           -1.1611     0.6869  -1.691   0.0989 .
## Environmentdangerous      0.8498     0.9128   0.931   0.3576
## EnergyAbsorptionnogear    1.5616     0.6912   2.259   0.0295 *
## CrushedFuselage1          2.1600     2.1024   1.027   0.3106
## RestraintIntact1         -1.6927     0.9841  -1.720   0.0933 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 62.06273)
##
##     Null deviance: 4194.5  on 47  degrees of freedom
## Residual deviance: 2673.4  on 39  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

We first need to check wether our assumptions on family choice were correct. To do this, we check the dispersion parameter calculated for each model. We can see that it is significantly higher than 1 in all cases (front : 17.31; middle: 27.59; rear: 36.18; total: 62.06), confirming that overdispersion is present, and needs to be corrected. Therefore we can say that our choice of family was apt.

Next, we analize the statistically significant results present in the models in detail:

- **Front Third** : The variables "HasFire" and "RestraintIntact" are statistically significant (p-values : 0.0482, 0.0441). In particular, "HasFire = no-fire" and "RestraintIntact = 1" are associated with negative coefficients. This suggests that the absence of fire and the presence of intact restraints are associated with lower mortality rates in the front third of the aircraft.
- **Middle Third** : The variable "EnergyAbsorption" is statistically significant (p-value = 0.0207). The "nogear" level has a positive coefficient, meaning that the absence of energy absorption mechanisms is associated with higher mortality rates in the middle third. Interesingly, "HasFire" is not significant here and "RestraintIntact" is only marginally significant (p-value = 0.0908). This is different from the front third.
- **Rear Third** : No variables are statistically significant at the 0.05 level, but "EnergyAbsorption" is marginally significant (p-value = 0.0738) with a positive coefficient for "nogear". This suggests a similar trend as in the middle third, but with less statistical confidence. It seems that the rear third is less affected by the other variables overall, this might be due to the fact that none of the variables in the dataset directly relate to rear third's survivability, unlike the front and middle thirds.

Total Mortality Rate : "EnergyAbsorption" is statistically significant (p-value = 0.0295) with a positive coefficient for "nogear". This means that the absence of a clean, energy-absorbing landing, is highly correlated

with higher overall mortality rates. "HasFire" and "RestraintIntact" are marginally significant (p-values = 0.0989 and 0.0933 respectively), suggesting that they may have some effect on overall survivability, but the evidence is not strong enough to be conclusive.

4 Conclusion

During this analysis, we focused on determining if the seating location in an aircraft has a significant effect on the mortality rates in the event of an accident. The Preliminary analysis was promising, with seemingly clear visual trends in the boxplot and different correlation patterns between the sections. However, using a statistical test coherent with the data distribution (Kruskal-Wallis test), we found no significant difference in mortality rates between the different sections of the aircraft, whether divided into thirds or halves.

On the other hand, our modeling efforts using Generalized Linear Models with a quasibinomial family revealed that certain factors significantly influence mortality rates, and they are different for each section of the aircraft.

Section	Significant Variables	p-values
Front Third	HasFire=no-fire, RestraintIntact=1	0.0482, 0.0441
Middle Third	EnergyAbsorption=nogear	0.0207
Rear Third	None	
Total	EnergyAbsorption=nogear	0.0295

These findings suggest that while seating location alone does not significantly impact mortality rates, other factors related to the accident do play an important role in survivability. Unfortunately, may of these factors are outside the control of passengers, and their significance is just an indication for improvements in aircraft safety.

So, ***Does your seat matter***?

Not significantly. The real things that matter are the the many layers of redundancy, safety procedures and checklists that the industry is so well known for implementing and upholding with attention. It is also the design of the aircraft, the training of the cabin crew and pilots, the effectiveness of the emergency procedures and the work of first responders. These are the reasons why we see so few accidents in the first place, and even in the event of an accident, these are the factors that will most likely determine your chances of survival.