Data Analysis of Category 4 and 5 Hurricanes

Josh Levine jl2108
Harsh Patel hkp49
Jaini Patel jp1891
Yifan Liao yl1463
Aayush Shah avs93

Aggregating the Data by Decade

1. We were given two Wikipedia links:
    - Category 4 (https://en.wikipedia.org/wiki/List_of_Category_4_Atlantic_hurricanes)
    - Categeory 5 (https://en.wikipedia.org/wiki/List_of_Category_5_Atlantic_hurricanes)
2. We compiled both data sets into csv files. Seen below is the first 5 rows of both sets of csv files. (Note: we changed the month to a numerical values 1 through 12)

Category 4

Name	Year	Month	Max. sustained winds(mph)	Minimum pressure(hPa=mbar)
Hurricane #3	1853	8	150	924
1856 Last Island Hurricane	1856	8	150	934
Hurricane #6	1866	9	140	938
Hurricane #7	1878	9	145	935
Hurricane #2	1880	8	150	931

Category 5

Name	Year	Month	Duration_as_a_Category_5_in_hours	max_wind_speeds_mph	max_pressure_hPa	Deaths
Cuba	1920	10	12	165	910	90
San Felipe II Okeechobee	1928	9	12	160	929	4000
Bahamas	1932	9	24	160	921	16
Camaguey	1932	11	78	175	915	3103
CubaBrownsville	1933	8	12	160	930	179

3. After compiling the data sets into csv files, we wrote an R script to aggregate the data into averages by decade and then wrote the aggregated data into a csv file for further analysis.

R Script for aggregating the data

```
# code to clean the data of category 4 hurricane.

cat41 <- read.csv("C:/Users/Jaini Patel/Desktop/Category41.csv")
cat41$Month <- gsub(" .*$", "",cat41$Month)
cat41$Month <- gsub(",", "", cat41$Month)
cat41$Month <- match(cat41$Month,month.name)
decades <- seq(1850, 2020, 10)
convert_to_decades <- function(x){
      for (index in 2:length(decades)){
             if (x <= decades[index] && x >= decades[index-1]){
                   return(decades[index])
               }
         }
   }
cat41$decade <- unlist(lapply(cat41$Season, convert_to_decades))

install.packages("plyr")
library(plyr)

# Aggregating and storing the cleaned data for category 4 hurricane (decadewise) in new dataframe.

cat42 <- ddply(cat41, .(cat41$decade), summarize,
                 Max_sustained_winds_mph = paste(mean(Max_sustained_winds_mph),collapse=","),
                 Month= paste(mean(Month),collapse=","))
cat42 <- ddply(cat41, .(decade), summarize,
               Hurricane_Count = paste(table(decade), collapse = ","),
               Max_winds_mph = paste(round(mean(Max_sustained_winds_mph), digits = 2),collapse = ","),
               Month = paste(round(mean(Month),digits=2),collapse = ","),
               Min_pressure_mbar = paste(round(mean(Minimum_pressure_mbar), digits = 2),collapse = ","))
View(cat42)

write.csv(cat42,"C:/Users/Jaini Patel/Desktop/decadewise4.csv")

# code to clean the data of category 5 hurricane.

cat5 <- read.csv("C:/Users/Jaini Patel/Desktop/Category5.csv")
cat5$Month <- gsub(" .*$", "",cat5$Month)
cat5$Month <- gsub(",", "", cat5$Month)
cat5$Month <- match(cat5$Month,month.name)

cat5$decade <- unlist(lapply(cat5$Year, convert_to_decades))

# Aggregating and storing the cleaned data for category 5 hurricane (decadewise) in new dataframe.

cat51 <- ddply(cat5, .(decade), summarize,
               Hurricane_Count = paste(table(decade), collapse = ","),
               Month = paste(round(mean(Month),digits=2),collapse = ","),
               Duration = paste(round(mean(Duration_in_hours), digits = 2), collapse = ","),
               Deaths = paste(round(mean(Deaths),digits = 2), collapse = ","),
               Max_winds_mph = paste(round(mean(Max_sustained_winds_mph), digits = 2),collapse = ","),
               Max_pressure_hPa = paste(round(mean(max_pressure_in_hPa), digits = 2),collapse = ","))

View(cat51)

write.csv(cat51,"C:/Users/Jaini Patel/Desktop/decadewise5.csv")
```

Analyzing the Aggregated Data

1. Displayed below are the Category 4 and 5 Hurricane csv files aggregated by decade

Aggregated Data for Category 4 Hurricanes

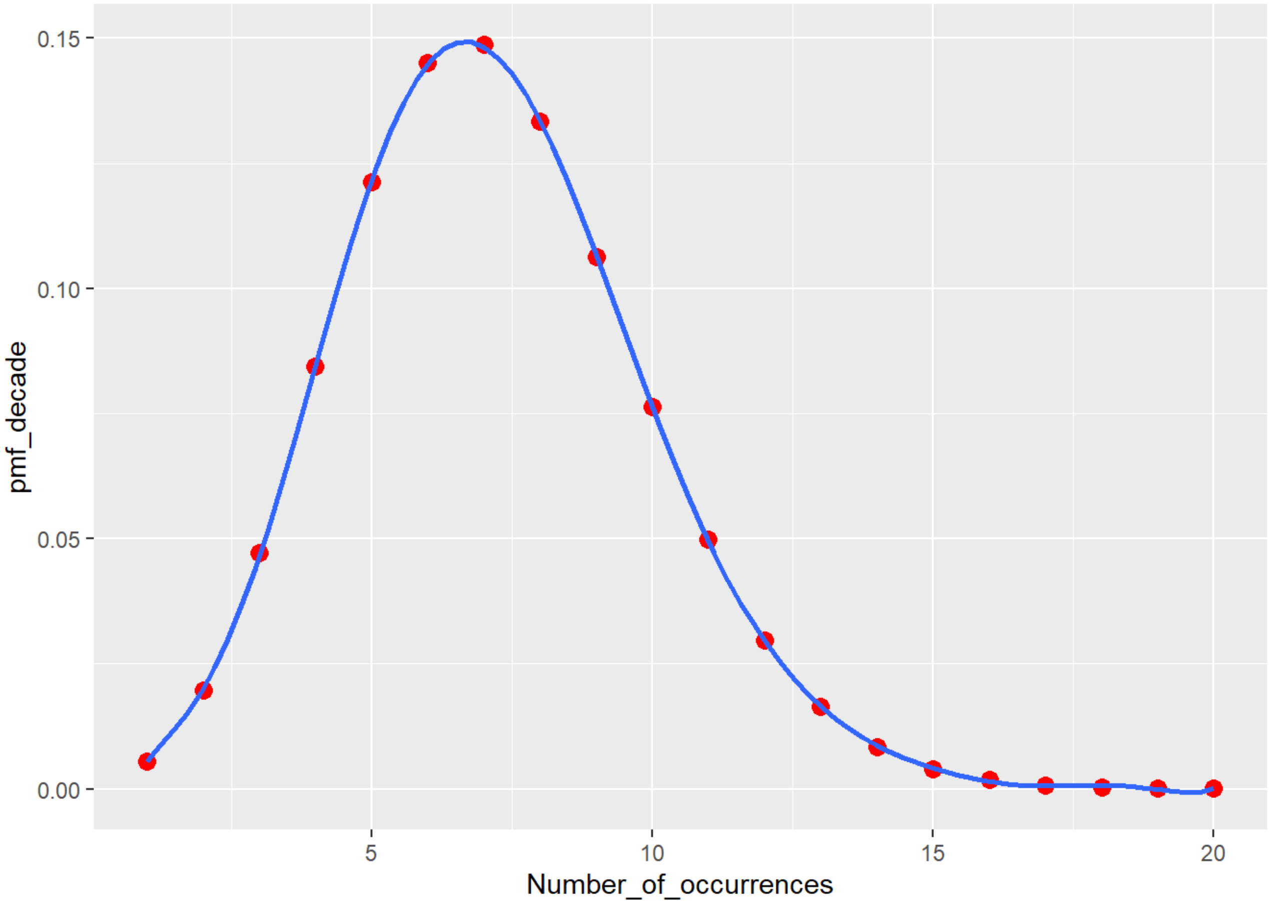
```
##     X decade Hurricane_Count Max_winds_mph Month Min_pressure_mbar
## 1   1   1860               2        150.00  8.00            929.00
## 2   2   1870               1        140.00  9.00            938.00
## 3   3   1880               3        145.00  8.67            931.33
## 4   4   1890               2        145.00  9.00            950.00
## 5   5   1900               5        137.00  8.80            935.00
## 6   6   1910               2        140.00  9.00            937.00
## 7   7   1920               5        144.00  8.60            931.60
## 8   8   1930               7        147.14  8.86            940.86
## 9   9   1940               9        139.44  8.67            944.78
## 10 10   1950              12        136.67  9.00            866.00
## 11 11   1960              12        138.33  8.83            941.83
## 12 12   1970              10        139.50  8.80            939.00
## 13 13   1980               5        138.00  8.40            942.60
## 14 14   1990               7        138.57  9.00            938.43
## 15 15   2000              14        144.64  8.79            934.29
## 16 16   2010              17        143.53  8.88            939.06
## 17 17   2020               9        143.89  9.22            939.33
```

Aggregated Data for Category 5 Hurricanes

```
##     X decade Hurricane_Count Month Duration  Deaths Max_winds_mph
## 1   1   1920               1 10.00    12.00   90.00        165.00
## 2   2   1930               1  9.00    12.00 4000.00        160.00
## 3   3   1940               6  9.17    27.00  762.00        166.67
## 4   4   1960               2  9.00    15.00  514.00        167.50
## 5   5   1970               4  9.00    18.00  161.00        165.00
## 6   6   1980               4  8.50    33.00  596.25        175.00
## 7   7   1990               2  9.00    15.00  212.50        172.50
## 8   8   2000               2  9.00    29.00 9695.00        177.50
## 9   9   2010               8  8.62    27.00  302.25        172.50
## 10 10   2020               6  9.33    24.25  662.50        170.83
##    Max_pressure_hPa
## 1            910.00
## 2            929.00
## 3            921.17
## 4            921.50
## 5            914.00
## 6            923.00
## 7            903.00
## 8            913.50
## 9            908.38
## 10           918.33
```

2. R Code to display the probability of the number of **category 4** hurricanes per decade. Followed by the resulting plot. Also known as the probability mass function.

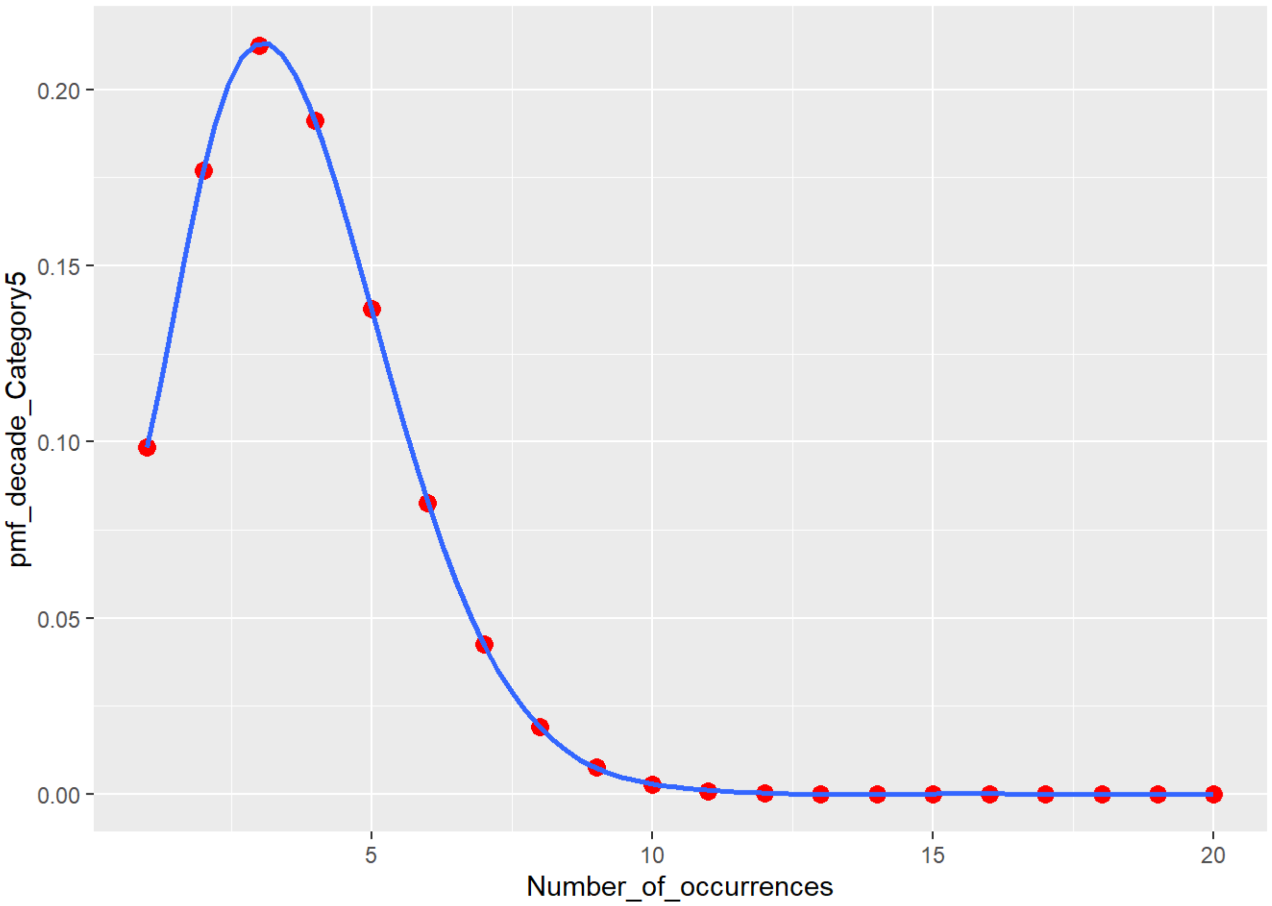
```
x<- data4['Hurricane_Count']
vec1<- data4$Hurricane_Count
# lambda = mean of No. of hurricanes occuring from 1850 to 2020
lambda<- mean(vec1)
pmf_decade<- c()
for(i in 1:20){
  pmf_decade[i] <- dpois(i,lambda)
}
Number_of_Occurrence<- c(1:20)
Hurricane <- data.frame(pmf_decade=pmf_decade , Number_of_occurrences = Number_of_Occurrence)
ggplot(Hurricane, aes(x=Number_of_occurrences, y=pmf_decade)) + geom_point(colour = "red",size=3)+stat_smooth(
 method = lm, formula = y ~ poly(x, 10), se = FALSE)
```



```
## [1] "lambda = 7.17647058823529"
```

3. R Code to display the probability of the number of **category 5** hurricanes per decade. Followed by the resulting plot.

```
x1<- data5['Hurricane_Count']
vec5<- data5$Hurricane_Count

lambda4<- mean(vec5)
pmf_decade_Category5<- c()
for(i in 1:20){
  pmf_decade_Category5[i] <- dpois(i,lambda4)
}
Number_of_Occurrence4<- c(1:20)
Hurricane <- data.frame(pmf_decade_Category5=pmf_decade_Category5 , Number_of_occurrences = Number_of_Occurrence4)
ggplot(Hurricane, aes(x=Number_of_occurrences, y=pmf_decade_Category5)) + geom_point(colour = "red",size=3)+stat_smooth( method = lm, formula = y ~ poly(x, 10), se = FALSE)
```

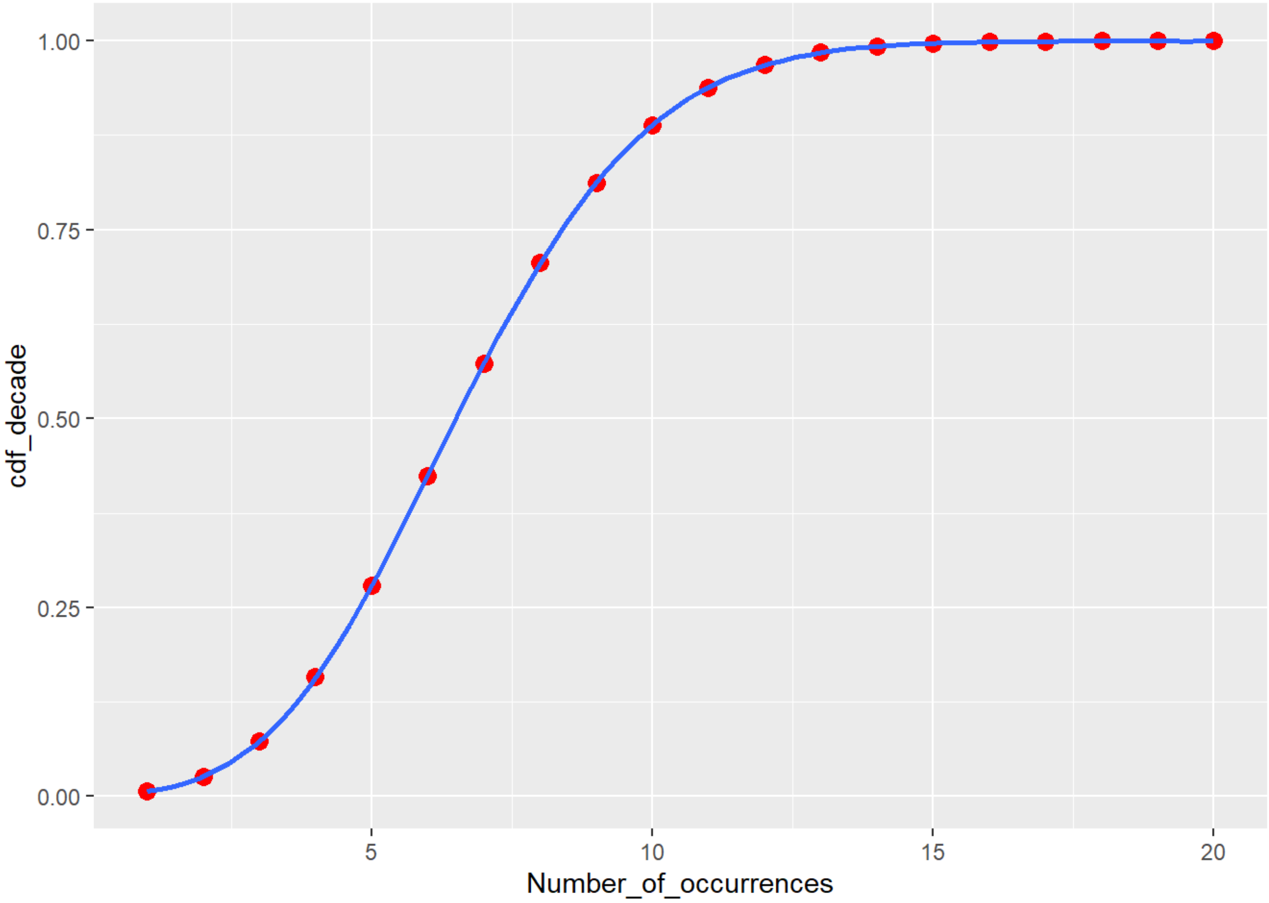


```
## [1] "lambda = 3.6"
```

Conclusion: We can conclude that category 4 hurricanes are more common than category 5 hurricanes on a per decade basis.

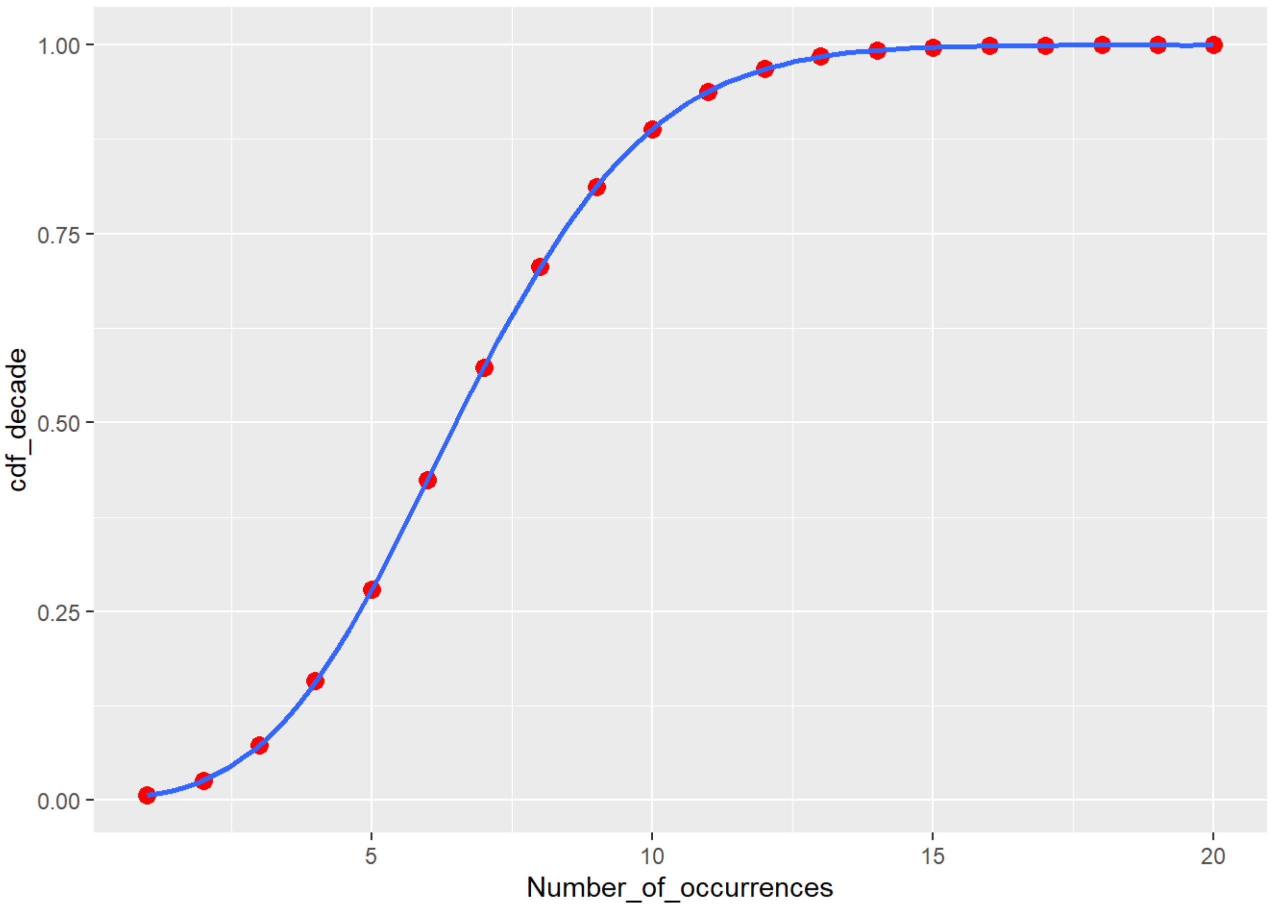
4. R Code to display the cumulative probability of the number of **category 4** hurricanes per decade. Followed by the resulting plot. Also known as the cumulative distribution function. Notice the curve starts at 0 and as we move to infinity hurricanes it ends at 1.

```
cdf_decade<- c()
for (i in 1:20){
  cdf_decade[i] <- ppois(i,lambda)
}
Hurricane_CDF <- data.frame(cdf_decade=cdf_decade , Number_of_occurrences = Number_of_Occurrence)
ggplot(Hurricane_CDF, aes(x=Number_of_occurrences, y=cdf_decade)) + geom_point(colour = "red",size=3)+stat_smoo
th( method = lm, formula = y ~ poly(x, 10), se = FALSE)
```



5. R Code to display the cumulative probability of the number of **category 5** hurricanes per decade. Followed by the resulting plot.

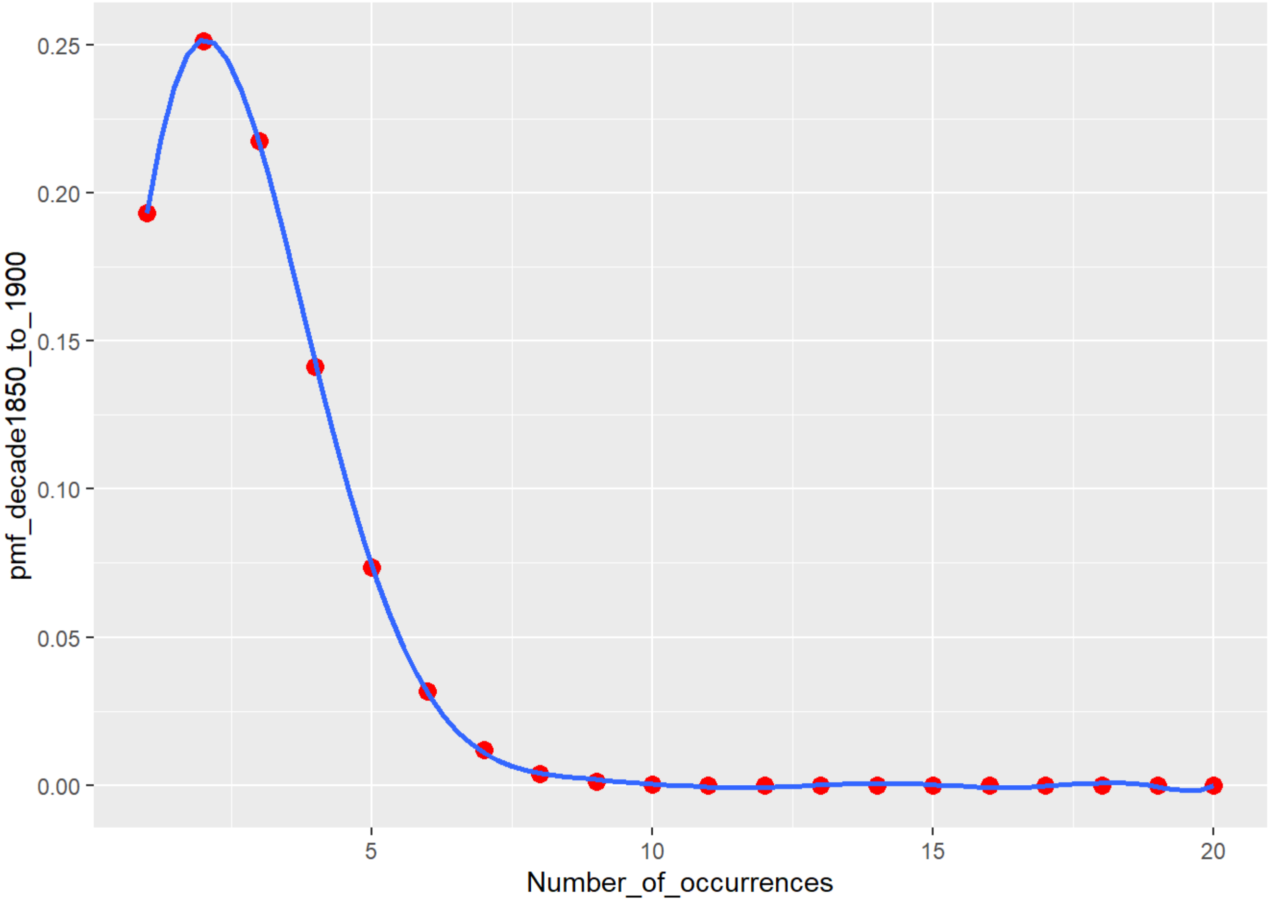
```
cdf_decade1<- c()
for (i in 1:20){
  cdf_decade1[i] <- ppois(i,lambda4)
}
Hurricane_CDF <- data.frame(cdf_decade1=cdf_decade1 , Number_of_occurrences = Number_of_Occurrence4)
ggplot(Hurricane_CDF, aes(x=Number_of_occurrences, y=cdf_decade)) + geom_point(colour = "red",size=3)+stat_smoo
th( method = lm, formula = y ~ poly(x, 10), se = FALSE)
```



Conclusion: We can see that the CDF for category 4 and 5 hurricanes are similar. Both reach a probability of one near 15. This means it is almost certain there will be 15 or less category 4 and 5 hurricanes in any given decade.

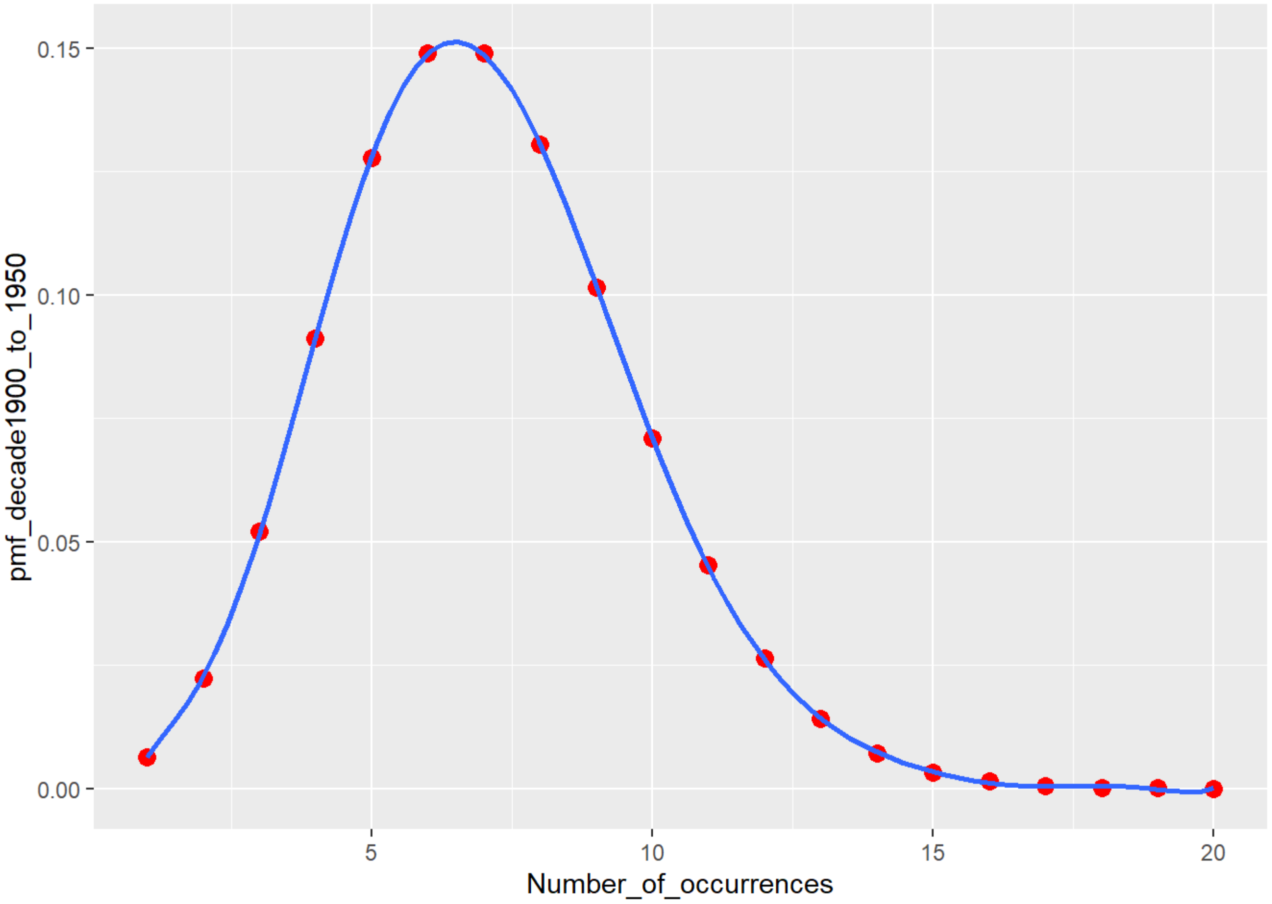
6. Next we are going to show the probability mass function of category 4 hurricanes in 50 year periods. First we will show from 1850 to 1900, second we will show from 1900 to 1950 and third we will show from 1950 to 2020. Lets see what we can conclude by graphing the 3 pmfs.

```
# Taking mean of decades ranging from 1850 to 1900
# Slicing the vector for 5 decades 1850 to 1900
vec2 <- vec1[1:5]
pmf_decade1850_to_1900<- c()
lambda1<- mean(vec2)
for(i in 1:20){
  pmf_decade1850_to_1900[i] <- dpois(i,lambda1)
}
Number_of_Occurrence1<- c(1:20)
Hurricane <- data.frame(pmf_decade1850_to_1900=pmf_decade1850_to_1900 , Number_of_occurrences = Number_of_Occurrence1)
ggplot(Hurricane, aes(x=Number_of_occurrences, y=pmf_decade1850_to_1900)) + geom_point(colour = "red",size=3)+stat_smooth( method = lm, formula = y ~ poly(x, 10), se = FALSE)
```



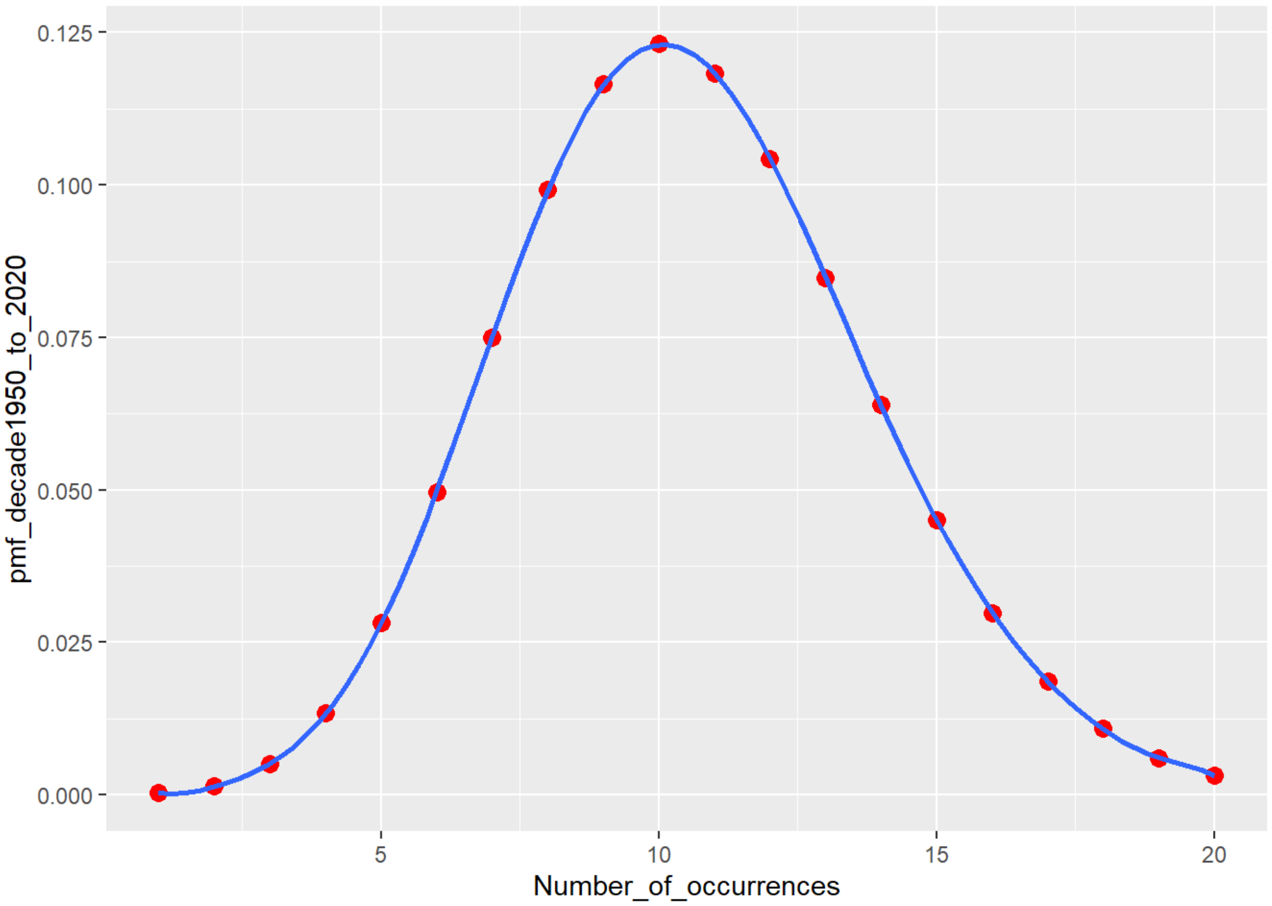
```
## [1] "lambda = 2.6"
```

```
# Taking mean of decades ranging from 1900 to 1950
# Slicing the vector for 5 decades 1900 to 1950

vec3 <- vec1[6:10]
pmf_decade1900_to_1950<- c()
lambda2<- mean(vec3)
for(i in 1:20){
  pmf_decade1900_to_1950[i] <- dpois(i,lambda2)
}
Number_of_Occurrence2<- c(1:20)
Hurricane <- data.frame(pmf_decade1900_to_1950=pmf_decade1900_to_1950 , Number_of_occurrences = Number_of_Occur
rence2)
ggplot(Hurricane, aes(x=Number_of_occurrences, y=pmf_decade1900_to_1950)) + geom_point(colour = "red",size=3)+s
tat_smooth( method = lm, formula = y ~ poly(x, 10), se = FALSE)
```



```
## [1] "lambda = 7"
```

```
# Taking mean of decades ranging from 1950 to 2020
# Slicing the vector for 5 decades 1950 to 2020

vec4 <- vec1[11:17]
pmf_decade1950_to_2020<- c()
lambda3<- mean(vec4)
for(i in 1:20){
  pmf_decade1950_to_2020[i] <- dpois(i,lambda3)
}
Number_of_Occurrence3<- c(1:20)
Hurricane <- data.frame(pmf_decade1950_to_2020=pmf_decade1950_to_2020 , Number_of_occurrences = Number_of_Occurrence3)
ggplot(Hurricane, aes(x=Number_of_occurrences, y=pmf_decade1950_to_2020)) + geom_point(colour = "red",size=3)+stat_smooth( method = lm, formula = y ~ poly(x, 10), se = FALSE)
```

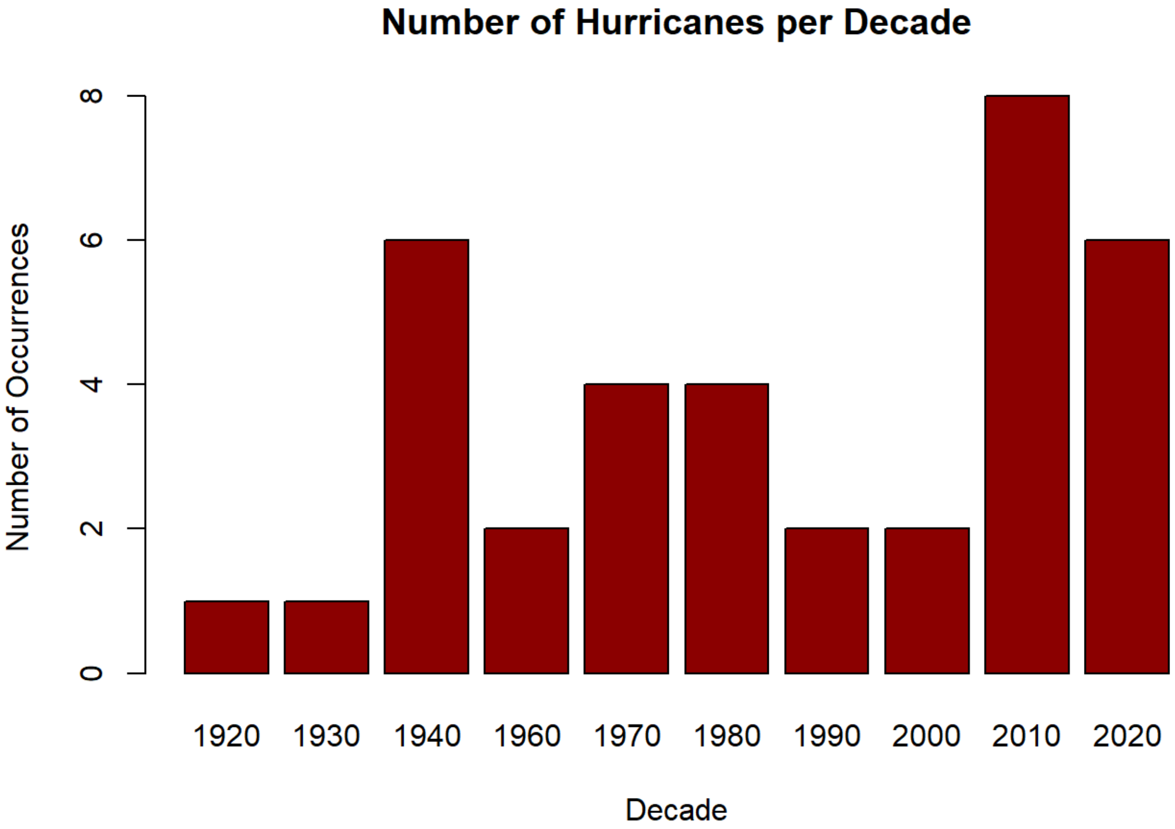


```
## [1] "lambda = 10.5714285714286"
```

Conclusion: We can clearly see that the probability of a higher number of category 4 hurricanes expected per decade has increased since 1850.

7. R Code to generate the number of category 5 hurricane occurrences per decade. Followed by the resulting plot.

```
barplot(data5$Hurricane_Count, main = "Number of Hurricanes per Decade",
        xlab = "Decade",
        ylab = "Number of Occurrences",
        names.arg = c("1920" , "1930", "1940", "1960", "1970", "1980", "1990", "2000", "2010", "2020"),
        col = "darkred")
```

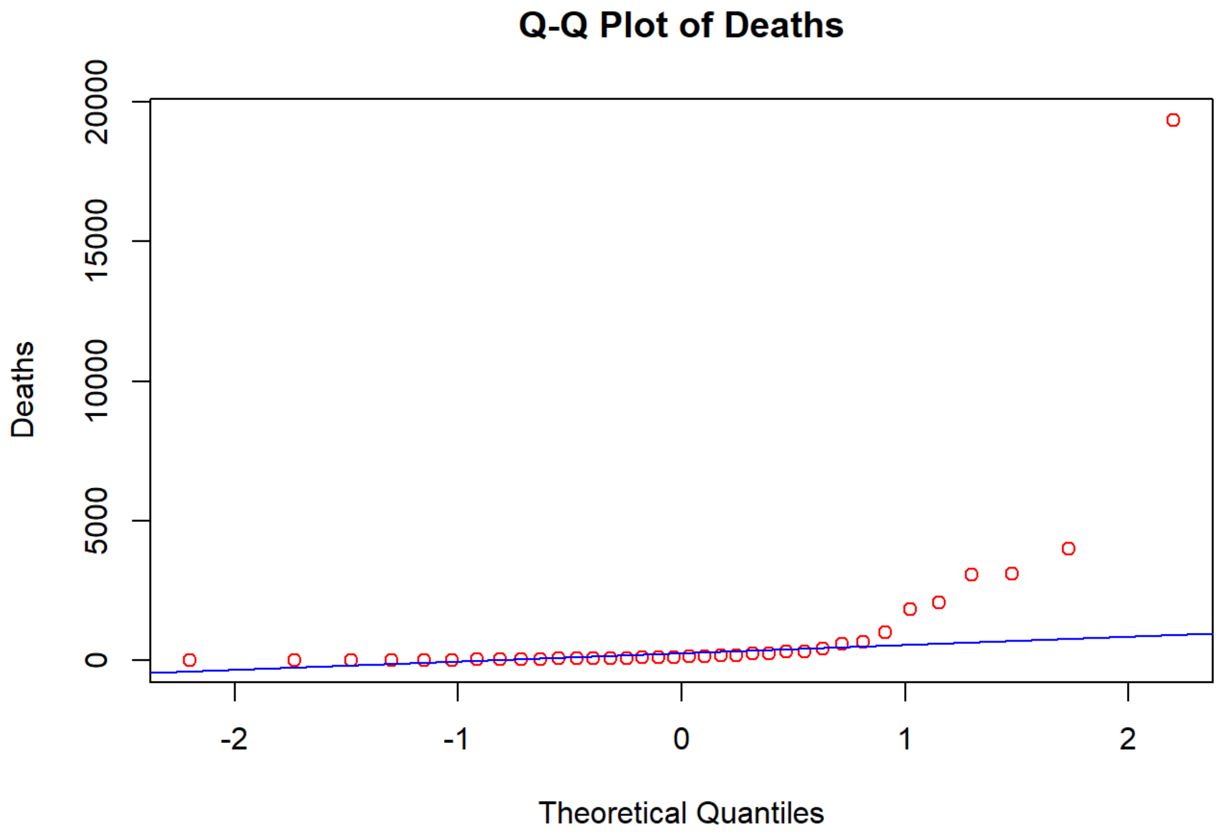


Conclusion: It does appear that Category 5 Hurricanes are occurring more often than in decades past.

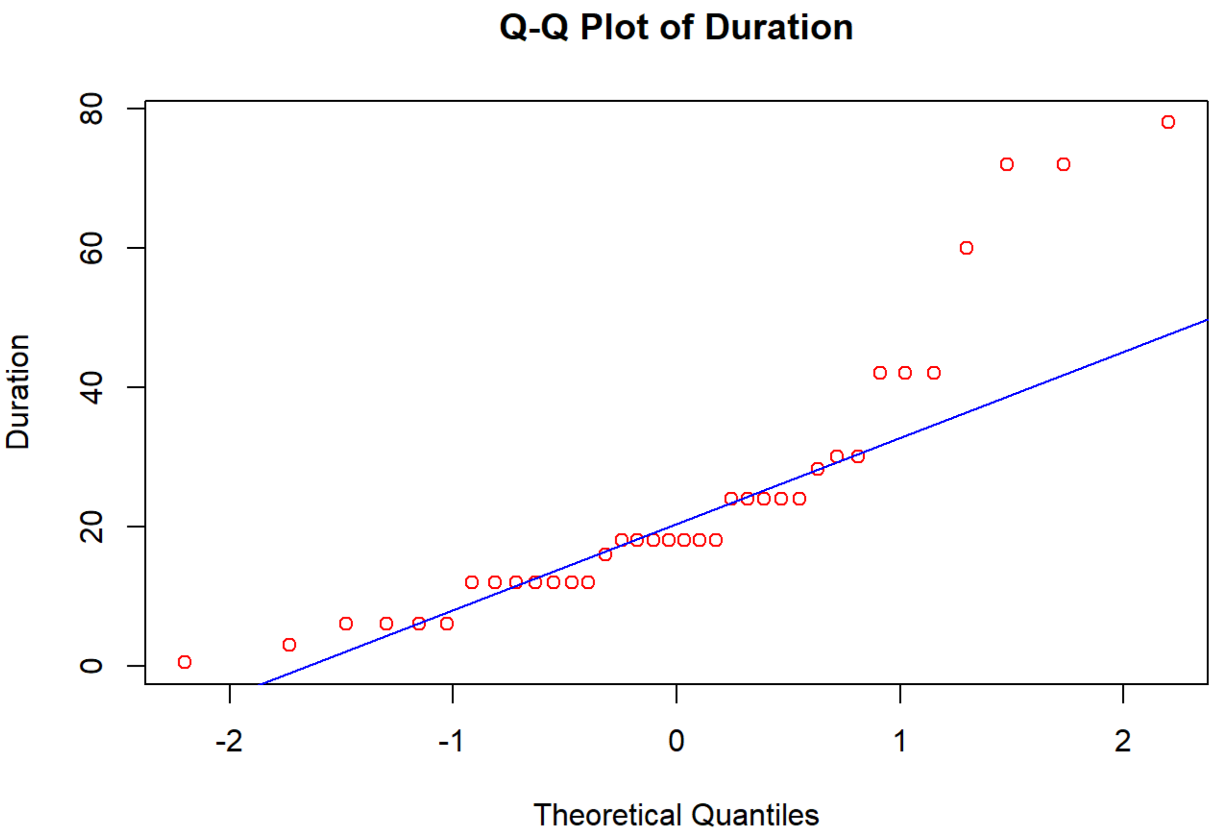
QQ-Plots

1. We will assess whether or not a distribution is normal or not using QQ plots. QQ Norm will plot the theoretical quantiles of a normal distribution along the x axis and the actual quantiles of the data on the y axis. And QQ Line will draw a theoretical line along which the values should be if the distribution was normal. Below is R code to plot Deaths, Duration and Log of Duration of all category 5 hurricanes.

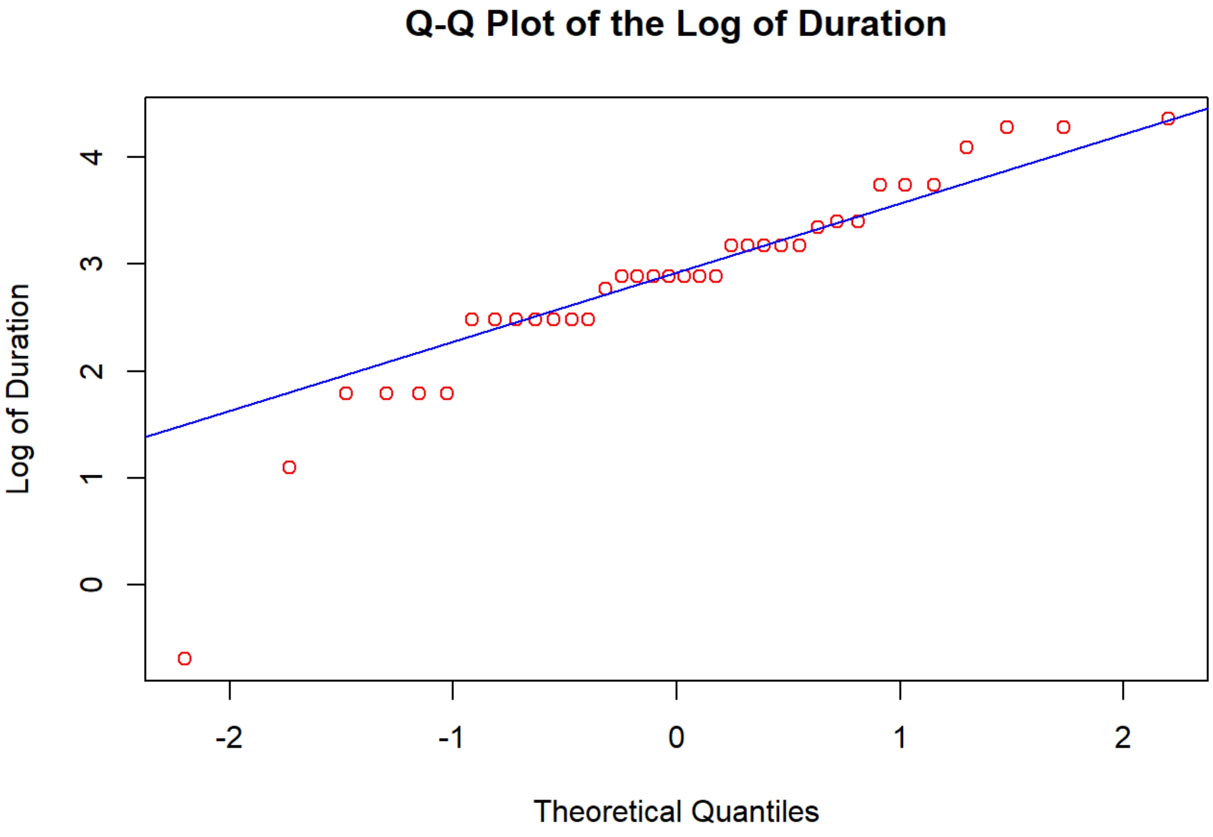
```
qqnorm(data_all_5$Deaths, col='red', main='Q-Q Plot of Deaths', ylab = "Deaths")
qqline(data_all_5$Deaths ,  col='blue')
```



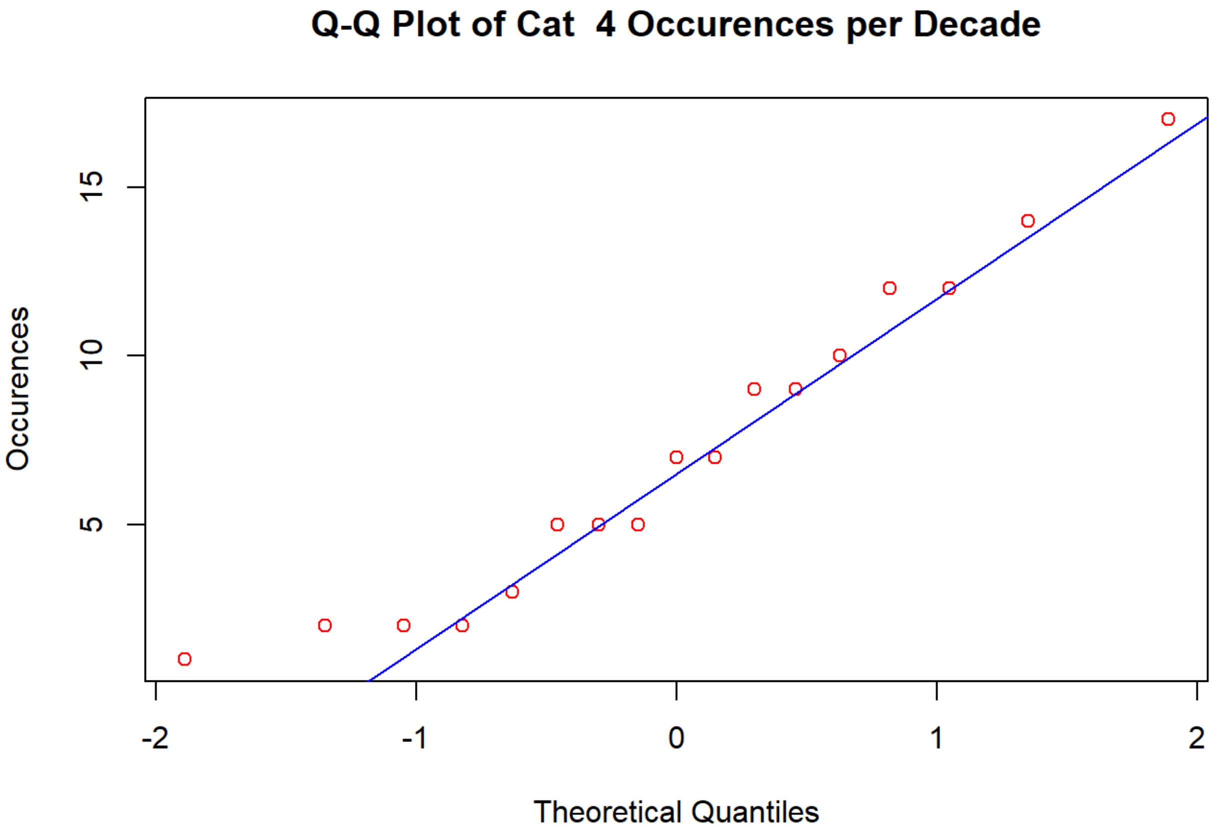
```
qqnorm(data_all_5$Duration, col='red', main='Q-Q Plot of Duration', ylab = "Duration")
qqline(data_all_5$Duration  , col='blue')
```



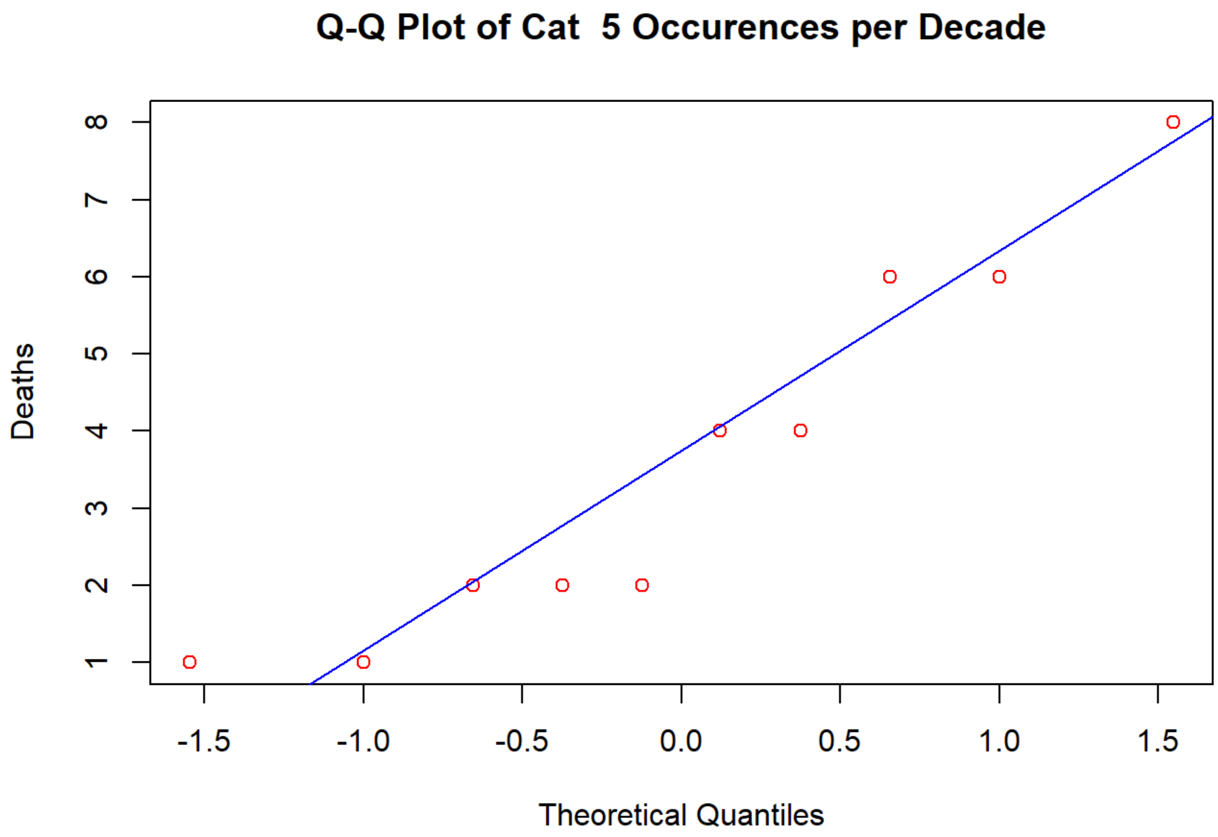
```
log.duration<-log(data_all_5$Duration)
qqnorm(log.duration, col='red', main='Q-Q Plot of the Log of Duration', ylab = "Log of Duration")
qqline(log.duration  , col='blue')
```



```
qqnorm(data_decade_4$Hurricane_Count, col='red', main='Q-Q Plot of Cat  4 Occurences per Decade', ylab = "Occurences")
qqline(data_decade_4$Hurricane_Count  , col='blue')
```



```
qqnorm(data_decade_5$Hurricane_Count, col='red', main='Q-Q Plot of Cat  5 Occurences per Decade', ylab = "Deaths")
qqline(data_decade_5$Hurricane_Count  , col='blue')
```



Conclusion: We can see that the distributions for deaths is normal because most of the points are on the theoretically normal line. We can observe, however, that the distribution of deaths is becoming not normal in more recent years. Clearly, we can see that the distribution of the duration of category 5 hurricanes are not normal and are becoming even less normal in more recent years. We can observe when we perform a log transformation on the duration that the values become more normal but arent quite normal. Finally, the distribution for the count per decade is also not normal.

Final Conclusion

Based on the data it does appear that stronger hurricanes are becoming more frequent and more destructive. Further research on the topic does show a trend over the last 20-40 years of hurricanes worsening. However, there are studies that also show that this may be a cycle called AMO (Atlantic Multidecadal Oscillation). Further details of AMO are below:

What is the AMO?

Amo is a series of changes leading to the temperature seesaw of the ocean surface mainly happening in the north Atlantic Ocean. It often takes a 20-40 year's long period of time to transform from a cool phase to a warm one, vice versa. The scientists believe this phenomenon has been occurring for over 1000 years.

What are the impacts of the AMO?

AMO is able to change the frequency of drought by affecting air temperatures and rainfall. The scope of influence covers much part of North America and Europe. It can also affect the frequency of severe Atlantic hurricanes. We have evidence that AMO becomes exaggerated since global warming.

How important is the AMO when it comes to hurricanes?

During the warm phases of AMO, the number of storms that become hurricanes eventually rises twice as many as the cool one.

Could the data reflect information on change in technology?

We believe the data shows that climate change is affecting the number of hurricanes occuring per decade. We think technology may affect the accuracy of wind speeds and pressure.