ABSTRACT

Car accidents and resulting deaths of car drivers has had significant impact on multiple stakeholders. Data have been collected since 1926, in which year there were 4,886 fatalities in some 124,000 crashes. Between 1951 and 2006 a total of 309,144 people were killed and 17.6 million were injured in accidents on British roads. The highest number of deaths in any one year was 9,169 people in 1941 during World War II. The highest figure during peacetime was 7,985 in 1966. Car accidents have a significant impact on the entire economy.

The police collect details of all incidents which they attend or become aware of within 30 days which occur on the highway in which one or more person is killed or injured and involving one or more vehicles using the STATS19 data collection system. STATS19 is the reference number for the police form used to record incidents. Additional information is gathered from death registrations, coroners' reports and traffic and vehicle registrations. STATS19 data are used in European Union road safety studies.

The report contains statistics about month-wise sum total figures of deaths of car drivers for the given period of fifteen years between 1969 and 1984 (16 years). Having taken into consideration the given statistics, a comprehensive characterization and comparison of data is summarized.

The study is based on the assumption that data is normally distributed, which is reflected in the relevant statistical analysis in the form of a histogram. The report also has tested the hypothesis that proportion of deaths over the years is the same or not. Furthermore, we have tested whether the introduction of seatbelts in 1983 has been effective in decreasing the number of deaths.

INTRODUCTION

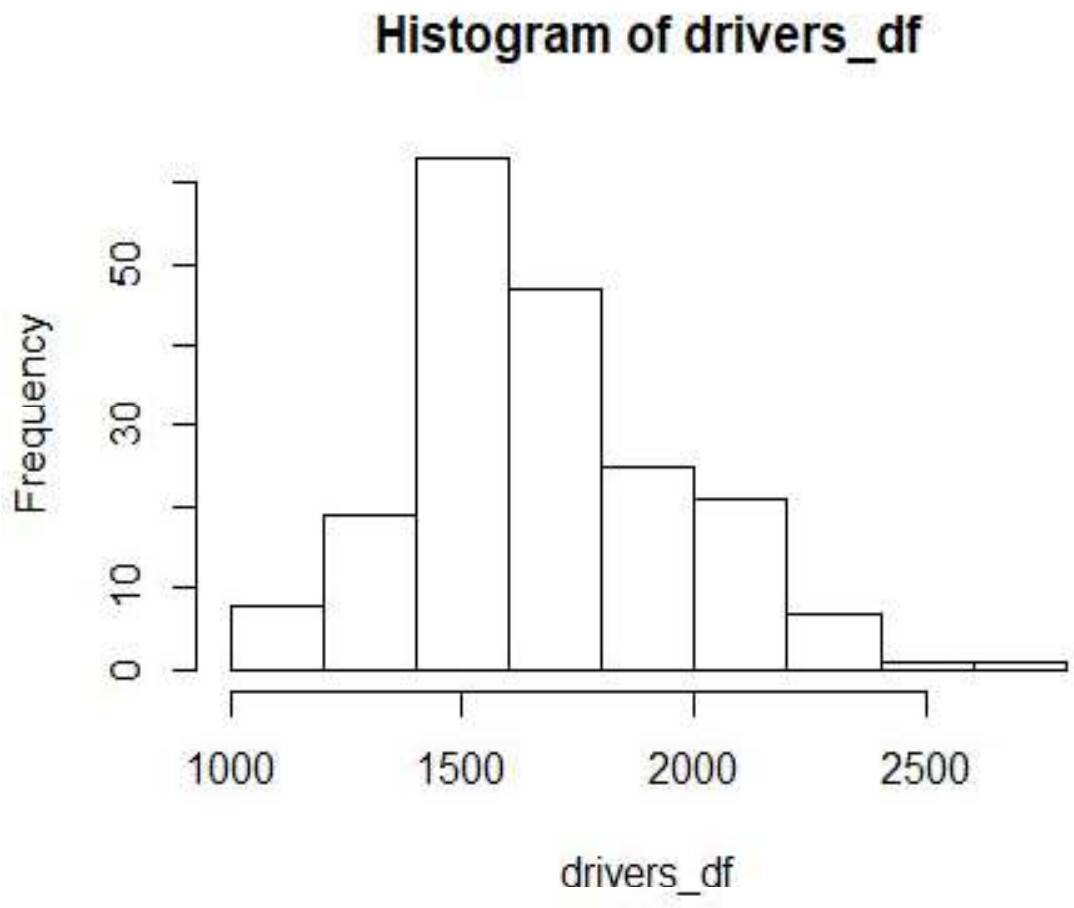
In this report, we have used the data on deaths of car drivers in Great Britain for the period of fifteen years from 1969 to 1984. You can find this data as part of the MASS R library.

The information used to create these statistics are collected by police forces, either through officers attending the scene of accidents or from members of the public reporting the accident in police stations after the incident. There is no obligation for people to report all personal injury accidents to the police (although there is an obligation under certain conditions, as outlined in the Road Traffic Act). These figures, therefore, do not represent the full range of all accidents or casualties in the country. Also, from the description we know that there has been a rule of compulsory wearing of seat belts introduced on 31 Jan 1983

In the dataset, we have month-wise figures on total number of deaths of car drivers. In this project, we have considered population to be the data corresponding to be total number of deaths of car drivers in the given time frame. We have used this sample to estimate cumulative probability distribution. Inference of statistical metrics like median of the population are done using sample point estimates and MLE. The confidence intervals for these estimates are obtained using parametric and non-parametric bootstrap. We can see in this report that since sample size is large, confidence intervals for parametric and non- parametric comes out to be close. Further, we have tested the hypothesis that proportion of deaths over the years remains the same from using Chi-squared test. We have also tested an important hypothesis whether the introduction of seatbelts in 1983 has been effective in decreasing the number of deaths.

DISTRIBUTION

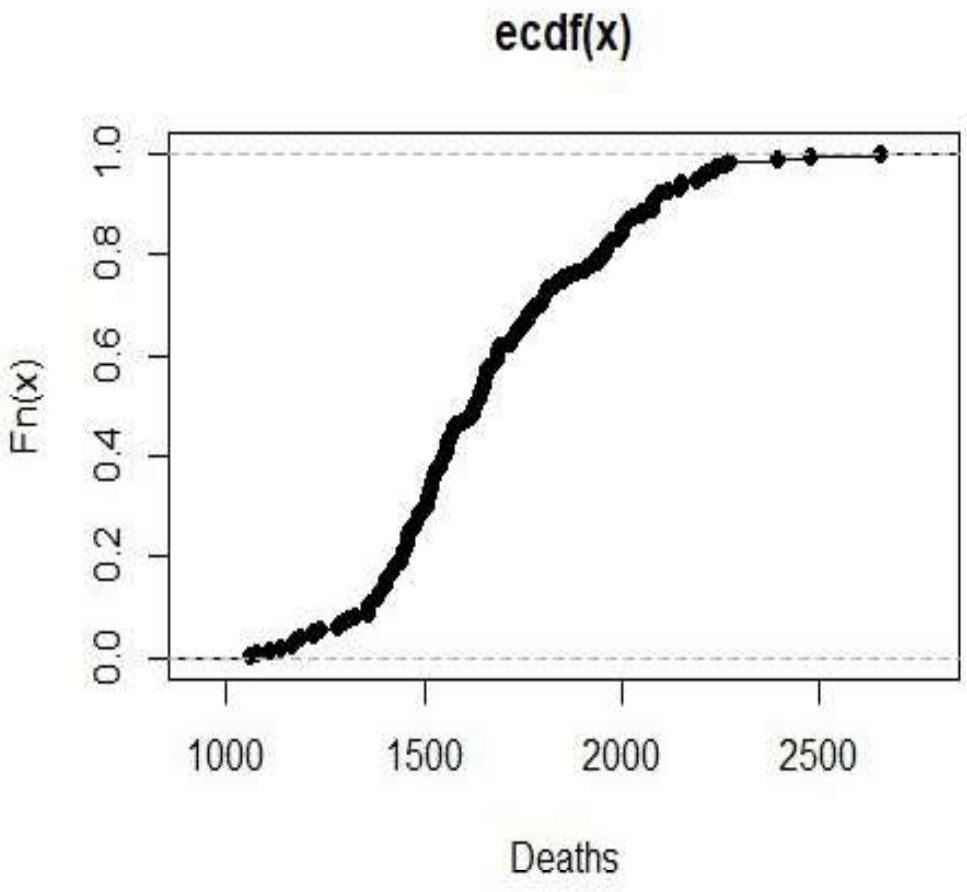
```
library(MASS)
drivers_df <- drivers
hist(drivers_df)
```



We can see from the histogram that the monthly road accidents follow a normal or **approximately normal distribution**

EMPIRICAL CDF

```
library(MASS)
drivers_df <- drivers
drivers_df.ecdf <- ecdf(drivers_df)
plot.ecdf(drivers_df,xlab = "Deaths")
```



Above plot gives us the **empirical distribution from the sample data**

NON-PARAMETRIC ESTIMATION

We try to estimate the **median of our data**.
Following can be done using non-parametric approach of estimation:

- Point Estimation of Median
- CI of the Median using bootstrap

```
th.hat <- median(drivers_df)
print(th.hat)
```

```
## [1] 1631
```

Point estimate of median: **1631**

Confidence Interval of this estimate can be found using three methods after bootstrapping:

1. Normal Confidence Interval
2. Pivotal Confidence Interval
3. Quantile Confidence Interval

```
library(bootstrap)

theta <- function(x, xdata)
{
  median(xdata[x])
}
n = length(drivers_df)

drivers_df.boot <- bootstrap(1:n,3200,theta, drivers_df)

Dboot<-drivers_df.boot$thetastar #vector of bootstrap results
se.boot <- sqrt(var(Dboot))

Normal_CI <- c(th.hat - 2*se.boot, th.hat + 2*se.boot)
print("Normal CI = ")
```

```
## [1] "Normal CI = "
```

```
print(Normal_CI)
```

```
## [1] 1575.959 1686.041
```

```
Pivotal_CI <- c(2*th.hat-quantile(Dboot,.975),2*th.hat-quantile(Dboot, .025))
print("Pivotal CI = ")
```

```
## [1] "Pivotal CI = "
```

```
print(Pivotal_CI)
```

```
##    97.5%     2.5%
## 1601.487 1700.025
```

```
Quantile_CI <- c(quantile(Dboot,.025), quantile(Dboot,.975))
print("Quantile CI = ")
```

```
## [1] "Quantile CI = "
```

```
print(Quantile_CI)
```

```
##     2.5%    97.5%
## 1561.975 1660.513
```

Therefore, following are our results for the confidence interval of the estimate:

1. Normal CI : **[1575.96, 1686.04]**
2. Pivotal CI: **[1601.49, 1700.02]**
3. Quantile : **[1561.97,1660.513]**

PARAMETRIC ESTIMATION

Since the distribution is normal (known), we can also estimate the confidence interval by parametric estimation using MLE. MLE of Mu and Sigma are **Mu_hat** and **Sigma_hat** of the sample

Also, since MLE is equivariant we can estimate the median of the data and it's confidence interval using parametric bootstrapping

```
mu_hat <- mean(drivers_df)
sigma_hat = sd(drivers_df)*sqrt(n/(n-1))
tau_hat <- mu_hat + qnorm(0.5) * sigma_hat
print(tau_hat)
```

```
## [1] 1670.307
```

Point estimate of median: **1670.31**

Confidence Interval of this estimate can be found using parametric bootstrapping using mu_hat and sigma_hat of MLE

```
##parametric bootstrap
tau_bootstrap <- vector()

for(i in 1:1000)
  {
  X = rnorm(n,mu_hat,sigma_hat)
  X_mu = mean(X)
  X_sigma = sd(X)*(sqrt(n/(n-1)))
  tau_bootstrap[i] = X_mu + qnorm(0.5)*X_sigma
  }
tau_bootstrap_se <- sd(tau_bootstrap)
tau_bootstrap_se
```

```
## [1] 20.66982
```

```
#CI
normal_ci <- c(mu_hat - 2 * tau_bootstrap_se, mu_hat + 2 * tau_bootstrap_se)
print(normal_ci)
```

```
## [1] 1628.968 1711.647
```

Normal Confidence Interval after parametric bootstrap is : **[1628.97,1712]**

HYPOTHESIS TESTING

Monthly death tolls over the years due to driving accidents is a cause concern. We try to estimate whether the yearly death tolls have been same or different. This gives us an important insight whether there has been any effort to decrease the yearly death rates.

Formally we define this by chi squared test:

Ho: p1=p2=……p16

Ha: p1 ≠ p2 ≠……p16

where, p1, p2… p16 are the proportions of deaths in 16 years of our data

```
deaths_yearly <- vector()
prob = vector()
#yearly deaths for 16 years
for(i in 1:16)
{
  p = (i-1)*12 + 1
  q = i*12
  deaths_yearly[i] = sum(drivers_df[p:q])
  prob[i] = (1/16)
}
chisq.test(x= deaths_yearly,p=prob)

##
##  Chi-squared test for given probabilities
##
## data:  deaths_yearly
## X-squared = 3783.3, df = 15, p-value < 2.2e-16
```

For alpha = 0.05, **we can reject the null hypothesis** that the proportion of deaths are same yearly. Hence, we can conclude the death tolls are not in equal proportions giving us an insight that there has been an intervention.

The data description states**: Compulsory wearing of seat belts was introduced on 31st Jan 1983**

We can test this hypothesis to see if this is true.

We divide the distribution into 2:

- Data of accidents till 3st Jan 1983 (drivers_before_sb)
- Data of accidents after 31st Jan 1983 (drivers_after_sb)

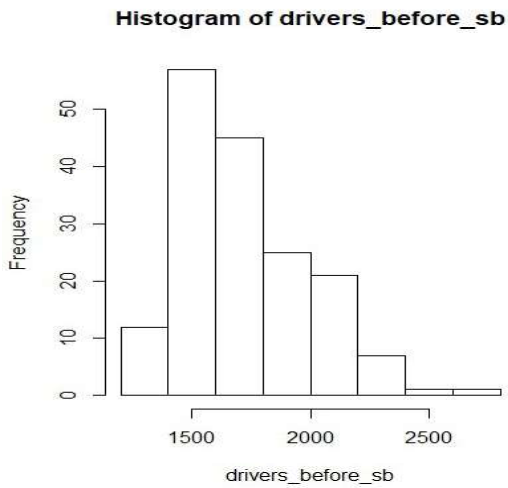
We conduct the following test:

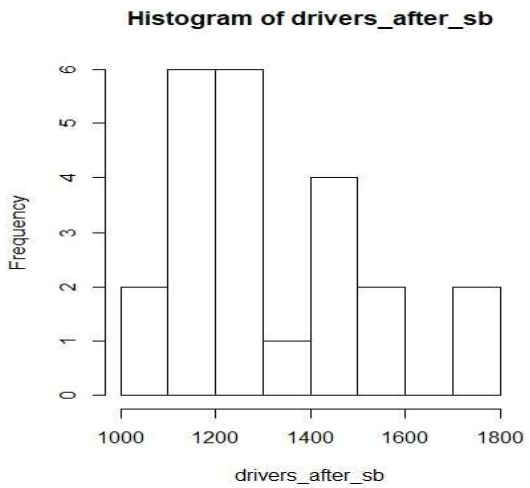
X = mean(divers_before_sb), Y = mean(drivers_after_sb)

Ho: X <= Y

Ha: X > Y

```
drivers_before_sb <- drivers_df[1:169]
drivers_after_sb <- drivers_df[170:192]
par(mfrow=c(1,2))
hist(drivers_before_sb)
hist(drivers_after_sb)
```






```
X = mean(drivers_before_sb)
Y = mean(drivers_after_sb)
Sx = sd(drivers_before_sb)*168/169
Sy = sd(drivers_after_sb)*22/23
Z = (X - Y)/((Sx^2/169)+(Sy^2/23))
p_value = 1-pnorm(abs(Z))
print(p_value)
```

```
## [1] 0.4216394
```

Since the p-value is greater than alpha (0.05) we reject the null hypothesis. Therefore, we conclude the **mean of normal distribution after the introduction of seatbelt compulsion rule is more.**

CONCLUSION

We draw the following important conclusions from this analysis:

- Proportions of yearly deaths over the 16 years are not the same
- Mean of the distribution of monthly deaths after the rule making seatbelts compulsory has decreased

Hence we can conclude, the rule of **making the seatbelts compulsory for drivers has been effective in decreasing the death toll**.