School of Mathematics
FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Academic integrity statement

You must sign this (typing in your details is acceptable) and include it with each piece of work you submit.

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the school's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name Sree Lalitha Gorty

Student ID 201684407

A statistical perspective on serial killers and their motives

Introduction

Statistics is the study of data and navigating through common problems which derives appropriate conclusions. From sports, healthcare, population estimates to space, natural calamities etc, statistics is an important aid that deals with interpretation and aggregation of large complex data into a simpler form of data.

It is interesting that on an average, a human would walk past 36 murderers in his/her lifetime knowingly or unknowingly. Serial killers are those individuals who are likely to murder three or more people, probably attaining a predictable behavioural pattern. A serial killer could often have a motive behind killing a person, three of which include: Angel of Death who purposefully mask themselves as a caregiver who end up killing with a grave intention, Revenge or Vigilante Justice involving killers who take the law into their hands and Convenience (didn't want children/spouse) which is most likely to be seen in unhappy marriages or families. Our analysis begins with the question “Does the average age at first kill differ between killers with different motives?”. The variables that we use to answer this question would be age at first kill, sex, motives, race etc. This question requires us to investigate and analyse each motive individually and obtain findings that lead us to our solution.

Data Cleaning

Initially, our sample consisted of 111 rows that had null values, potential outliers and killers with age at first kill before 1900 which we choose to ignore in our analysis. A new column depicting the career duration of each killer was added and interestingly there were no unusual values as 25 years (which is the highest value) could be a tangible duration for a killer to maintain his killing streak. About 19% of the data has been removed as a result of data cleaning thus giving us the final data sample to work on.

Number of rows with null or garbage values	15
Number of rows with age at first kill <= 1900	6
Number of rows with unusual career duration	0
Percentage data cleaned	19%

Table 1: Summary of Cleaned data

Data Exploration

As our prime interest of this analysis is how the age at first kill of each serial killer affects different motives, it is a wise idea to group the killers based on their motive and draw numerical summaries from them.

Motive	Mean	Median	Coefficient of skewness
Angel of Death	32.6	30	1.35
Revenge	31.08	28	0.95
Convenience	36.5	33.5	0.27

Table 2: Numerical Summaries of each motive

On taking a close look at table 2, we could start our journey on finding a distribution for each of these motives begins. When the median and mean are same or nearly equal, the given sample is likely to be evenly distributed and gives us an estimated if the data is skewed. In this case, they are almost in the same range and not unusual which can be interpreted as nearly symmetric distribution. Further, the coefficient of skewness gives us an estimate if the given sample is likely to follow a normal distribution. As we compare the three motives "Convenience" has a lesser value of skewness compared to that of the other two giving us a peripheral idea that it could be a normal distribution.

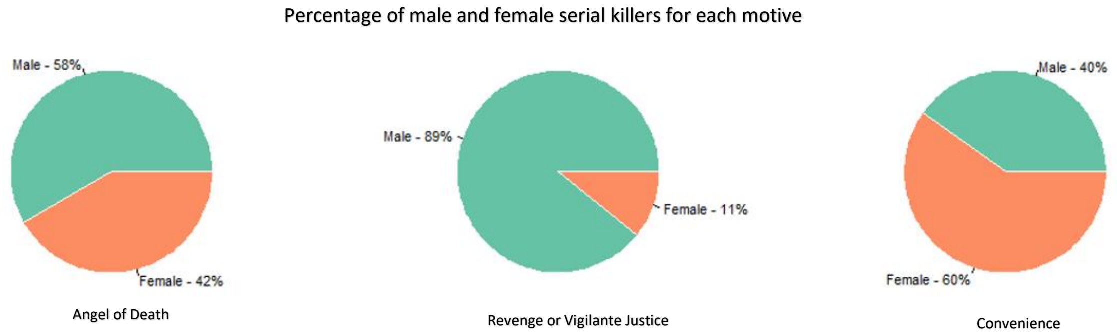

Figure 2: Percentage of male and female serial killers for each motive

Taking genders into consideration, we could see some interesting patterns on our sample like, percentage of females committing a murder for convenience of not wanting a child or a spouse was more when compared to males and for the other two motives men contributed to most murders where revenge was the motive.

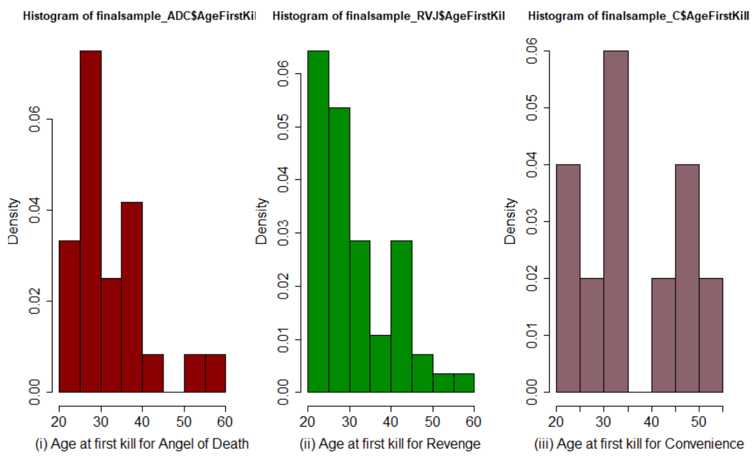

Figure 3: Graphical summary of the three motives using Histogram

Our first interest would be to check if the histogram of age of first kill for each motive follows the features to that of a normal distribution density curve. The results might seem a bit confusing, however for samples with relatively smaller size like ours we could consider that fig (iii) might follow a normal distribution and that figures (i) and (ii) may not be based on the histograms.

Modelling and Estimation

In this part of the analysis, we wish to understand the distribution (if it exists) for each motive to interpret if age at first kill follows a certain distribution between different motives which would ideally direct us to arrive at a solution.

The analysis begins by checking if our samples follow a normal distribution as we generally assume that the given sample follows a normal distribution. For this, we rely on multiple ways namely:

1. a Q-Q Plot
2. Kolomogorov- Smirnov Test (KS Test)
3. Shapiro- Wilk Test
4. Chi- Square Test

Before we perform any of the tests, we would first assume that

H0 (Null Hypothesis)	The sample follows a normal distribution
H1(Alternate Hypothesis)	The sample does not follow a normal distribution

Table 4: Hypothesis test assumptions

This assumption would probably help us understand the distribution of the sample through the outcomes we obtain after each test which would help us summarize our results efficiently.

Q-Q Plot:

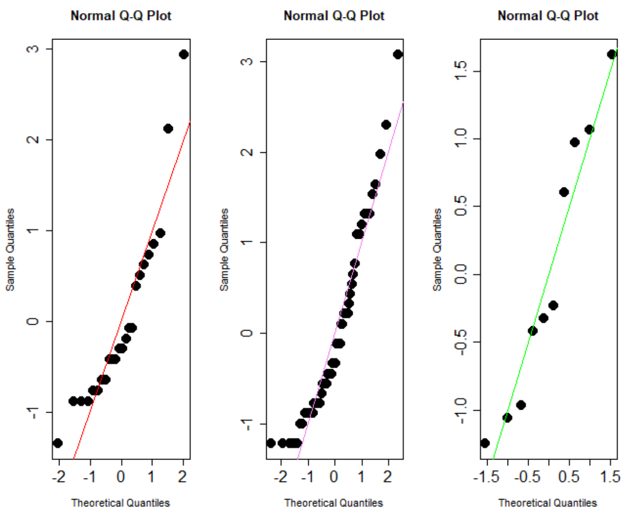

Figure 5: (i) Q-Q plot for Angel of Death (ii) Q-Q plot for Revenge (iii) Q-Q plot for Convenience

The Q-Q plot gives us an estimate whether our sample follows a normal distribution. Our aim here is to see how well the obtained points align with the line that in return tells us the likelihood of it following a normal distribution. According to Figure 5, all the three motives do exhibit normality to some extent which does not allow us to reject the null hypothesis. So, our deduction from this plot would suggest the possibility of the samples being a normal distribution.

The Kolmogorov- Smirnov Test, Shapiro-Wilk Test and Chi-square goodness of fit test helps us to estimate the p-value based on which we can either fail to reject or reject null hypothesis at 5% level. Table 6 gives us the summary of the results that we have obtained for each test. It is important to note that we would decide w.r.t hypothesis at 5% level.

Motive	KS Test	Shapiro - Wilk	Chi- square
(i)Angel of Death	0.307	0.0055	0.1231
(ii)Revenge	0.155	0.00057	0.006174
(iii)Convenience	0.792	0.407	0.07855

Table 6: Summary of p-values for the three tests for each motive

According to the results, each test gives us a different insight of the sample's distribution. As per the KS test and Chi-square, we fail to reject the null hypothesis as the p-values > 5% which does not allow us to negate normality, on the other hand, Shapiro- Wilk test rejects null hypothesis for (i) and (ii) whereas it fails to reject the null hypothesis for (iii). A detailed summary of the same could be seen in table 7.

Tests	Angel of Death	Revenge	Convenience
Q-Q plot	Fail to reject H0	Fail to reject H0	Fail to reject H0
Shapiro- Wilko	Reject H0	Reject H0	Fail to reject H0
KS Test	Fail to reject H0	Fail to reject H0	Fail to reject H0
Chi-Square	Fail to reject H0	Fail to reject H0	Fail to reject H0

Table 7: Summary of outcomes of tests

Furthermore, it is safe that we test other forms of distribution for our motives in order to investigate and attain outcomes that could favour our obtained table of hypothesis via method of moments.

Distribution	*(i)Angel of Death*	*(ii)Revenge*	*(iii)Convenience*
Exponential	1.704e-05	2.248e-11	0.01595
Poisson	0.4189	0.02406	0.3665

Table 8: Summary of p-values for Exponential and Poisson distributions for each motive

If we look at Tables 7 and 8, the samples clearly reject the hypotheses of the samples following an exponential distribution as the p-values are less than 5% for all the three motives whereas the p-values for motives (i) and (iii) for Poisson distribution fail to reject the hypothesis that the samples follow a Poisson distribution at 5% level.

Inference about each motive

Our next interest would be to calculate a confidence interval to check if the mean is 27 years. We would now consider either a t-test or a z-test based on the sample size and normality. Now that we have the results from different tests, Q-Q plot and Shapiro- Wilk test give us a better understanding if our sample directly comes from a normal distribution as they do not depend on the parameters. Considering the sample size, value of variance being known, and the normality check based on Q-Q and Shapiro-Wilk tests, we could decide on which test to be used on each motive as depicted in table 9.

Motive	*Z-test/ T-test*	*Comments*
(i)Angel of Death	T- test	Sample size is medium, accepting Normality
(ii)Revenge	Z-test	Sample size is large, rejecting Normality
(iii)Convenience	T-test	Sample size is small, accepting Normality

Table 9: Choice of Z-test/ T-test for each motive

Before calculating our confidence intervals, we first need to define our null hypothesis. In this case,

H0: $\mu = 27$ vs H1: $\mu \neq 27$

As per the assumptions that we made in table 9, we can now obtain the confidence intervals for each motive which is shown in figure 10.

Angel of Death	CI = (29.01153 36.32180)
Revenge or Vigilante Justice	CI = (28.83620 33.34237)
Convenience	CI = (28.75672 44.24328)

Table 9: Confidence interval for each motive

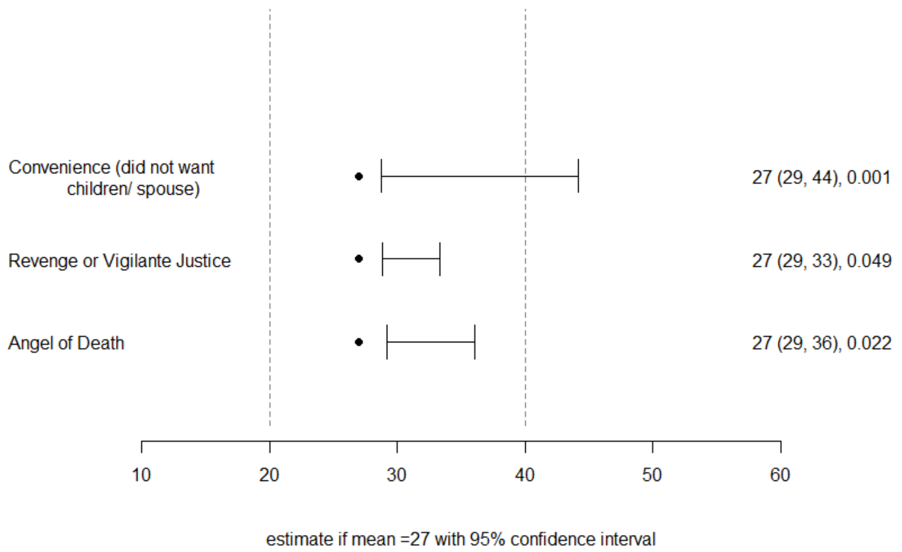

Figure 10: A plot of confidence intervals for different motives

From figure 10, we can decipher that the test has rejected null hypothesis as the value of μ does not lie in the confidence interval for any of our motives. This particular result helps us answer our research question which is- yes, the average age at first kill differs between killers with different motives. As our results reject the null hypothesis, it clearly indicates that the average mean is not same across different motives.

Discussion

Over the course of analysing and interpreting results in order to answer our research question, we have made a few assumptions to proceed with our findings.

Assumptions:

- For our motive, Angel of death, we failed to reject that it comes from a Poisson distribution but still went ahead with the assumption that it comes from a normal distribution.

- In addition to it, the Shapiro- Wilko test which we considered to be advantageous over other tests has rejected the hypothesis that it comes from a normal distribution.

Limitation:

- Further the usage of Z test is based on the sample size and is most accurate for samples with large sizes but in our case, we performed z-test on a sample with n= 64 which may not be a very large sample.

From our analysis, we were able to deduce that our sample with motives “Angel of death” and “Convenience (didn’t want children or spouse)” follow a normal distribution whereas for our sample with motive “Revenge or Vigilante justice”, we were unable to reckon a suitable model among the distributions that we have.

Appendix

R- code:

```
setwd("C:/Users/Lalitha Gorthi/OneDrive/Documents/Stats Coursework")

load(file = "killersandmotives.Rdata")

#selected my Data Frame.

createsample(201684407)

save(mysample, file = "mysample.RData")

mysample

dim(mysample)

mysample[1:10, ]

table(mysample$Motive)

# determining the time period of each killer in years

for(i in 1:nrow(mysample))

  mysample$Year_of_first_kill <- c(mysample$YearBorn + mysample$AgeFirstKill)

mysample

for(i in 1:nrow(mysample))

  mysample$Year_of_last_kill <- c(mysample$YearBorn + mysample$AgeLastKill )

mysample


#added a new column career duration.

for(i in 1:nrow(mysample))

  mysample$careerduration=c( mysample$AgeLastKill - mysample$AgeFirstKill )

mysample


#removed NA values from the Data Set.

mysample=mysample[!is.na(mysample$Motive),]

mysample

dim(mysample)


#removed rubbish values from the Data Set.

mysample <-mysample[!(mysample$AgeFirstKill)=="99999",]

mysample

dim(mysample)

#removed killers born before 1990.
```

```
mysample <-mysample[!(mysample$AgeFirstKill)<="1900",]

finalsample=mysample

dim(finalsample)

finalsample

mean(finalsample$AgeFirstKill)

median(finalsample$AgeFirstKill)


#checking for unusual Outliers.

boxplot(finalsample$careerduration,

    main = "career duration of killers",

    ylab = "years active")

#grouping based on Motive

finalsample_ADC <-subset(finalsample, Motive== 'Angel of Death')

finalsample_ADC

dim(finalsample_ADC)

sum(finalsample_ADC$Sex=="Female") #----> 14 male, 10 Female


finalsample_RVJ <-subset(finalsample, Motive== 'Revenge or vigilante justice')

dim(finalsample_RVJ)

sum(finalsample_RVJ$Sex=="Female")# --------> 50 male,6 Female

 #count of male is more for Revenge or vigilante justice


finalsample_C <- subset(finalsample, Motive == "Convenience (didn't want
children/spouse)")dim(finalsample_C)

sum(finalsample_C$Sex=="Male") #--------> 4 male, 6 female


par(mfrow = c(1,3))

Prop1 <- c(14, 10)

library(RColorBrewer)

myPalette <- brewer.pal(5, "Set2")

pie(Prop1 , labels = c("Male - 58%", "Female - 42%"), border="white", col=myPalette, title= "percentage of
male and female ")

Prop2 <- c(50, 6)

myPalette <- brewer.pal(5, "Set2")

pie(Prop2 , labels = c("Male - 89%", "Female - 11%"), border="white", col=myPalette)
```

```
Prop3 <- c(4,6)
myPalette <- brewer.pal(5, "Set2")
pie(Prop3 , labels = c("Male - 40%", "Female - 60%"), border="white", col=myPalette)
par(mfrow = c(1,3))
boxplot(finalsample_ADC$careerduration, col = 'red', title = 'Angel of Death')
boxplot(finalsample_RVJ$careerduration, col = 'green3', title= 'Revenge or vigilante justice')
boxplot(finalsample_C$careerduration, col = 'blue', title = 'Convenience' )
#numerical summaries
mu_ADC <- mean(finalsample_ADC$AgeFirstKill) #---> 32.67
sd_ADC <- sd(finalsample_ADC$AgeFirstKill)
sort(finalsample_ADC$AgeFirstKill)
median(finalsample_ADC$AgeFirstKill) #----> 30
library(moments)
skewness(finalsample_ADC$AgeFirstKill) #----> 1.3579

mu_RVJ <- mean(finalsample_RVJ$AgeFirstKill) #----> 31.08
sd_RVJ <- sd(finalsample_RVJ$AgeFirstKill)
median(finalsample_RVJ$AgeFirstKill)#-----> 28
skewness(finalsample_RVJ$AgeFirstKill)#------> 0.9502869

mu_C <- mean(finalsample_C$AgeFirstKill) #---------> 36.5
sd_C <- sd(finalsample_C$AgeFirstKill)
median(finalsample_C$AgeFirstKill) #----> 33.5
a <- c(mu_ADC,mu_RVJ,mu_C)
skewness(finalsample_C$AgeFirstKill)#-----> 0.27927

library(ggplot2)
library(dplyr)
# Hole size
hsize <- 3.5
df <- df %>%
  mutate(x = hsize)
ggplot(df, aes(x = hsize, y = value, fill = group)) +
  geom_col() +
```

```
  geom_text(aes(label = value),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  xlim(c(0.2, hsize + 0.5))+
  scale_fill_discrete(labels = c("Angel of Death",  "Convenience", "Revenge or Vigilante Justice"))


#Modelling and estimation
#plotting histograms
par(mfrow = c(1,3))
hist(finalsample_ADC$AgeFirstKill, freq=FALSE, col= 'dark red', xlab='(i) Age at first kill for Angel of Death', cex.axis = 1.5, cex.lab = 1.5, cex.title = 1.5, cex.title = 1.5)#-----> expected Poisson Distribution
hist(finalsample_RVJ$AgeFirstKill, freq=FALSE, col='green4', xlab='(ii) Age at first kill for Revenge', cex.axis = 1.5, cex.lab = 1.5, cex.title = 1.5, cex.title = 1.5) #------> expected exponential Distribution
hist(finalsample_C$AgeFirstKill, freq=FALSE, col='pink4', xlab='(iii) Age at first kill for Convenience', cex.axis = 1.5, cex.lab = 1.5, cex.title = 1.5, cex.title = 1.5)   #-------> expected (maybe Normal)
#Calculating CDFs
#ADC Angel of Death
Fn <- ecdf(finalsample_ADC$AgeFirstKill)
Fn(200)
G <- function(x){

  return(pnorm(x, mean = mu_ADC, sd = sd_ADC))
}
G(200)
plot(Fn, verticals = TRUE, pch = NA)
x <- 1:500
lines(x, G(x), col = "violet")
#RVJ Revenge or Vigilante justice
Fn <- ecdf(finalsample_RVJ$AgeFirstKill)
Fn(200)
G <- function(x){
  return(pnorm(x, mean = mu_RVJ, sd = sd_RVJ))
}
G(200)
plot(Fn, verticals = TRUE, pch = NA)
x <- 1:500
```

```
lines(x, G(x), col = "Blue1")
#C Convenience
Fn <- ecdf(finalsample_C$AgeFirstKill)
Fn(200)
G <- function(x){

  return(pnorm(x, mean = mu_C, sd = sd_C))
}
G(200)
plot(Fn, verticals = TRUE, pch = NA)
x <- 1:500
lines(x, G(x), col = "Red2")
#Kolmogorov-Smirnov test
#H0:F=G(our sample is from this normal distribution),
#H1:F≠G(our sample is not from this normal distribution).
ks.test(x =  finalsample_ADC$AgeFirstKill,
        y = "pnorm",
        mean = mu_ADC, sd = sd_ADC)
# here p is 0.3072
ks.test(x =  finalsample_RVJ$AgeFirstKill,
        y = "pnorm",
        mean = mu_RVJ, sd = sd_RVJ)
# here p is 0.1557
ks.test(x =  finalsample_C$AgeFirstKill,
        y = "pnorm",
        mean = mu_C, sd = sd_C)
# Q-Q plots
sort(finalsample_ADC$AgeFirstKill)
z <- (sort(finalsample_ADC$AgeFirstKill) - mu_ADC)/sd_ADC
qqnorm(z, cex.axis=1.5, pch = 19, col= 'black', cex =2)
abline(a = 0, b = 1, col = "red")
sort(finalsample_RVJ$AgeFirstKill)
z <- (sort(finalsample_RVJ$AgeFirstKill) - mu_RVJ)/sd_RVJ
qqnorm(z, cex.axis=1.5, pch = 19, col= 'black', cex =2)
```

```
abline(a = 0, b = 1, col = "violet")


sort(finalsample_C$AgeFirstKill)
z <- (sort(finalsample_C$AgeFirstKill) - mu_C)/sd_C
qqnorm(z, cex.axis=1.5, pch = 19, col= 'black', cex =2)
abline(a = 0, b = 1, col = "green")
#Shapiro-Wilk test
shapiro.test(finalsample_ADC$AgeFirstKill)
shapiro.test(finalsample_RVJ$AgeFirstKill)
shapiro.test(finalsample_C$AgeFirstKill)
#Chi-square test
library(nortest)
pearson.test(finalsample_ADC$AgeFirstKill)
pearson.test(finalsample_RVJ$AgeFirstKill)
pearson.test(finalsample_C$AgeFirstKill)
#exponential distribution
#Method of moments
ks.test(x =  finalsample_ADC$AgeFirstKill, y = "pexp", rate = 1/mu_ADC)
ks.test(x =  finalsample_RVJ$AgeFirstKill, y = "pexp", rate = 1/mu_RVJ)
ks.test(x =  finalsample_C$AgeFirstKill, y = "pexp", rate = 1/mu_C)
#SHOWING THAT POISSON DIST. IS ALSO NOT SATISFYING VIA METHOD OF MOMENTS
ks.test(x =  finalsample_ADC$AgeFirstKill, y = "ppois", lambda = mu_ADC)
ks.test(x =  finalsample_RVJ$AgeFirstKill, y = "ppois", lambda = mu_RVJ)
ks.test(x =  finalsample_C$AgeFirstKill, y = "ppois", lambda = mu_C)
#z or t tests
sigma_Square <- 74
#Z-test at 95% CI for mu = 27-- > ADC
n_Con <- nrow(finalsample_ADC)
t_Value_ADC<- qt(p = 0.975, df = n_ADC -1 )
samplemean_ADC <-mu_ADC
CI_ADC=  samplemean_ADC + c(-1, 1)*t_Value_ADC*sqrt(sd_ADC^2/n_ADC)
CI_ADC
# does not lie in the confidence interval
#Z- test for RVJ
```

```
n_RVJ <- nrow(finalsample_RVJ)
samplemean_RVJ <-mu_RVJ
CI_of_RVJ <- samplemean_RVJ + c(-1,1)*1.96*(sqrt(sigma_Square/n_RVJ))
CI_of_RVJ
# does not lie in the confidence interval
#T- test for C
n_Con <- nrow(finalsample_C)
t_Value_C<- qt(p = 0.975, df = n_Con -1 )
samplemean_C <-mu_C
CI_C =  samplemean_C + c(-1, 1)*t_Value_C*sqrt(sd_C^2/n_Con)
CI_C
# does not lie in the confidence interval

analysis = c("Angel of Death",
          "Revenge or Vigilante Justice",
          "Convenience (did not want
          children/ spouse)"
          )
# Results of each test (estimated mean,
# upper CI limit, lower CI limit, p-value):
estimate  =  c(27, 27, 27)
upper     =  c(36.3, 33.3,44.2)
lower     =  c(29.0,28.8 ,28.7)

# Note that the order of the results in each vector
# must match the order of the labels in the
# vector "analysis".
# Set the margin widths:
par(mar = c(6,6,1,6))
# Create an empty plot of a suitable
# size (considering the width of your
# confidence intervals):
plot(x = 0,                          # One point at (0,0).
    xlim = c(10, 60), ylim=c(0, 5),       # Axis limits.
```

```
   type = "n", xaxt = "n", yaxt="n",      # No points, no axes drawn.
   xlab = NULL, ylab= NULL, ann = FALSE,   # No axis labels or numbers.
   bty="n")                          # No box.
# Add a horizontal (side = 1) axis:
axis(side = 1, cex.axis = 1)
# Add an axis label 4 lines below the axis:
mtext("estimate if mean =27 with 95% confidence interval",
    side = 1, line = 4)
# Add some grid lines, preferably lined up
# with the numbers on the horizontal axis:
for(i in c(-40, -20, 0, 20, 40)){
  lines(c(i, i), c(0, 5), lty = 2, col = "gray53")
}
# Add labels for each analysis on the left (side = 2)
# at vertical heights of 1, 2, 3 and 4:
verticalpos = 1:3
mtext(text = analysis,  at = verticalpos,
    side = 2, line = 5, outer = FALSE, las = 1, adj = 0)
# Try changing the "line" option to move these closer to
# or away from the plotted intervals.
# Plot the four point estimates (centres
# of the CIs for each analysis):
points(estimate, verticalpos, pch = 16)
# Plot the four interval estimates:
for(i in 1:3 ){
 lines(c(lower[i], upper[i]), c(verticalpos[i], verticalpos[i]))
 lines(c(lower[i], lower[i]), c(verticalpos[i] + 0.2, verticalpos[i] - 0.2))
 lines(c(upper[i], upper[i]), c(verticalpos[i] + 0.2, verticalpos[i] - 0.2))
  }
# Now we add numerical results on the right (side = 4), but we
# need to put them into a nice form first. Note that
# paste() merges words and numbers, and formatC()
# allows us to control the number of decimal places.
est <- formatC(estimate, format='f', digits = 0)
```

```
# Type pval to see what this does.
L <- formatC(lower, format = 'f', digits = 0)
U <- formatC(upper, format = 'f', digits = 0)
interval <- paste("(", L, ", ", U, "),", sep = "")   # Type interval to check.
# Putting it all together:
results <- paste(est, interval, pval)
# Add to the plot:
mtext(text = results, at = verticalpos,
    side = 4, line = 4, outer = FALSE, las = 1, adj = 1)
# Like a Christmas present, an R
# plot belongs in a box:
box("inner")
#reason for choosing Q-Q and shapiro over poisson through cdf
x =  finalsample_C$AgeFirstKill
lambda1 = mu_C
cdf_userDefined_Poisson <- function(x,value){return(ppois(x, lambda1))}
cdf_userDefined_Poisson(600)
K <- ecdf(finalsample_C$AgeFirstKill)
K(600)
G <- function(x){
  return(pnorm(x, mean = mu_C, sd = sd_C))}
G(600)
library('fitdistrplus')
plot(fitdist(finalsample_C$AgeFirstKill,"pois"))
```