ANALYSIS OF THE AVERAGE AGE OF KILLERS WITH DIFFERENT MOTIVES

MATHS 5741

WASIF SHAUKAT

201589942

DECEMBER 16, 2022

UNIVERSITY OF LEEDS
WASIF SHAUKAT

1. INTRODUCTION

Homicide cases and serial killings are tragic and difficult to investigate. This report will only analyse the average age of serial killers with different motives and how it differs from the known population mean age of 27 years old (μ=27 years) and variance of 74 (σ^2=74) which were active in the 1900s. The provided data set will be analysed to investigate which motive results in committing a crime at an early age.

There is no standard definition for a person to be called a serial killer. A serial killer, the rarest form of homicide, is a criminal who commits at least two or more (or some also define it as three or more) murders in a given time (usually within a month) with a cooling down time between murders. The data set contains 71 observations with three different motives i.e. Angel of death, Convenience (didn't want children/spouse) and Escape or avoid arrest. The data contains different variables which will be discussed in detail in the next section. However, the age of the killer when he or she committed the first murder is of particular interest as it will be analysed for different motives.

2. DATA CLEANING AND EXPLORATION

Before performing any analysis, data was cleaned which included missing values (99999 values) and killers that were active before 1900 were also removed as shown in table 1. The data was then explored to find

Table 1 Percentage of data removed for different reasons

Sr. No.	Reason for removing data	Percentage of Data removed
1.	Missing age of the first kill which was recorded as 99999	12.7%
2.	Killers were active before the 1900s	1.4%

out various facts related to different attributes of data. The career duration was then plotted as shown in Figure 1, and most of the cases had career durations of up to 5 years except for a few. However, one had a whooping career duration of 31 years and while looking at the motive of this case, which was an angel of death, it was decided to keep this point. It could be possible because such cases are difficult to catch as the suspect is usually a caregiver. It was also found that 69% of the killers were male and 83% of them were White in terms of an ethnic group.

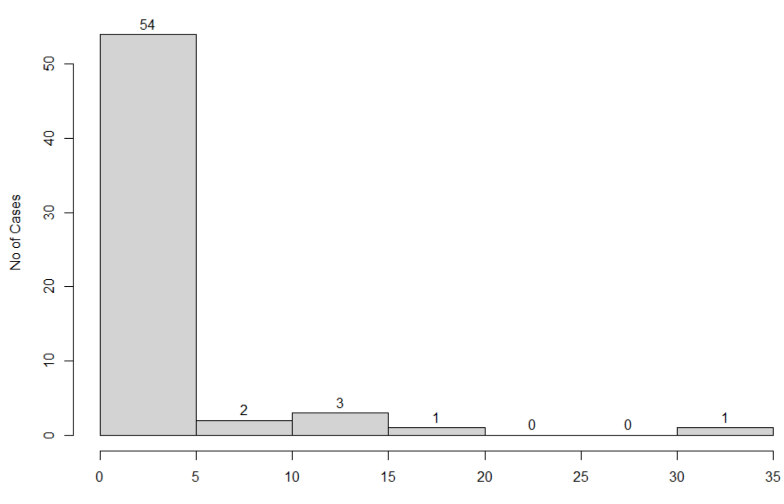

Figure 1 Frequencies of Career duration in years

Figure 3 shows a box plot of age at first murder for three different motives whereas blue dots are actual data points. The data appears to be symmetrical across the median (the median is different and between 30 to 35 years for all motives) however it appears to be positively skewed since the median is near the lower interquartile range. The figure also suggests that most killings are done between the ages of 25 to 40 years for all motives. One point for motive Angel of Death appears to be an outlier where the killer is 58 years old while committing their first crime but in actuality, it is not an outlier. This is only because of the 1.5 interquartile range of R which is not appropriate in our case. It is worth mentioning that no relation

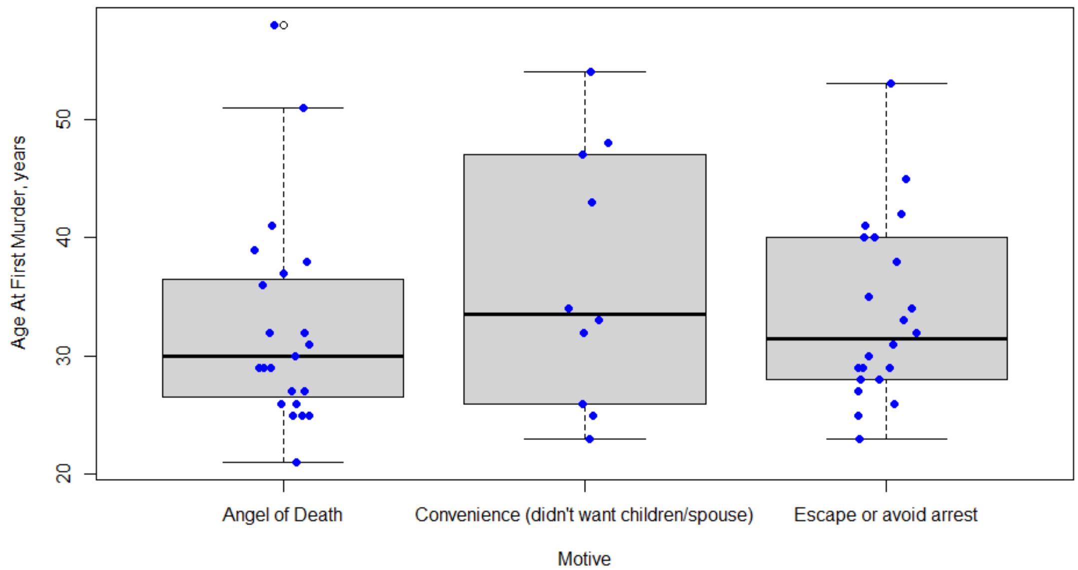

Figure 2 Graphical summary of the age of the first kill for different motives

was found between career duration and age at the first kill during the data exploration.

3. RESULTS

To begin the analysis, the mean, variance, and standard deviation for all motives (table 2) were compared and it was found that the average age of the sample differs according to motive and is different from the population average.

Table 2 Mean, Variance and Standard Deviation of all motives

Motive	**Angel of death n = 23**	**Convenience n = 10**	**Escape or avoid arrest n = 22**
Mean	32.3	33.5	36.5
Variance	75.8	55.8	117.2
Standard Deviation	8.7	7.5	10.8

For the sake of this analysis, our null hypothesis H_0 is equal to 27 years following a normal distribution. To decide frequency distributions for each motive QQ plots and the Shapiro-Wilk test were relied on the most since they do not need to specify the parameters of a particular normal distribution. Figure 3 shows the plots for all motives which clearly shows that the first one is curved and not following the normal distribution whereas the other two suggest normal distribution. The same findings were also confirmed through Shapiro-Wilk tests. Based on these findings, the null hypothesis for the motive Angel of death was rejected and failed to reject for the remaining motives. None of the distributions could fit the motive angel of death as shown in figure 4 because the distribution is bimodal. However, for the sake of

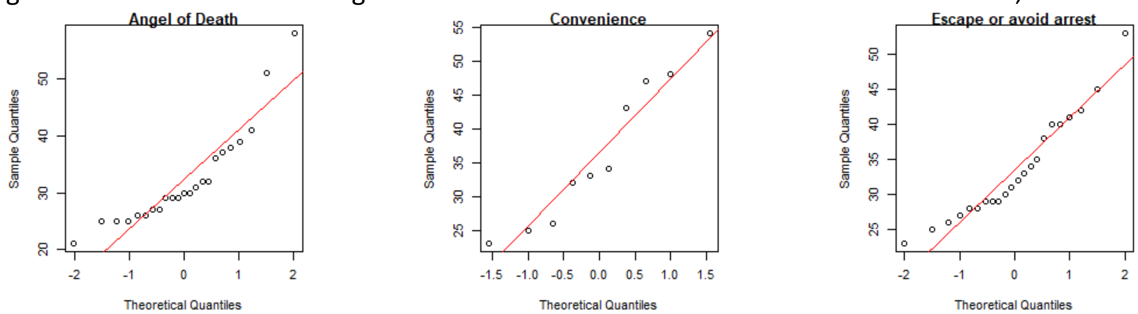

Figure 3 QQ plots of different motives

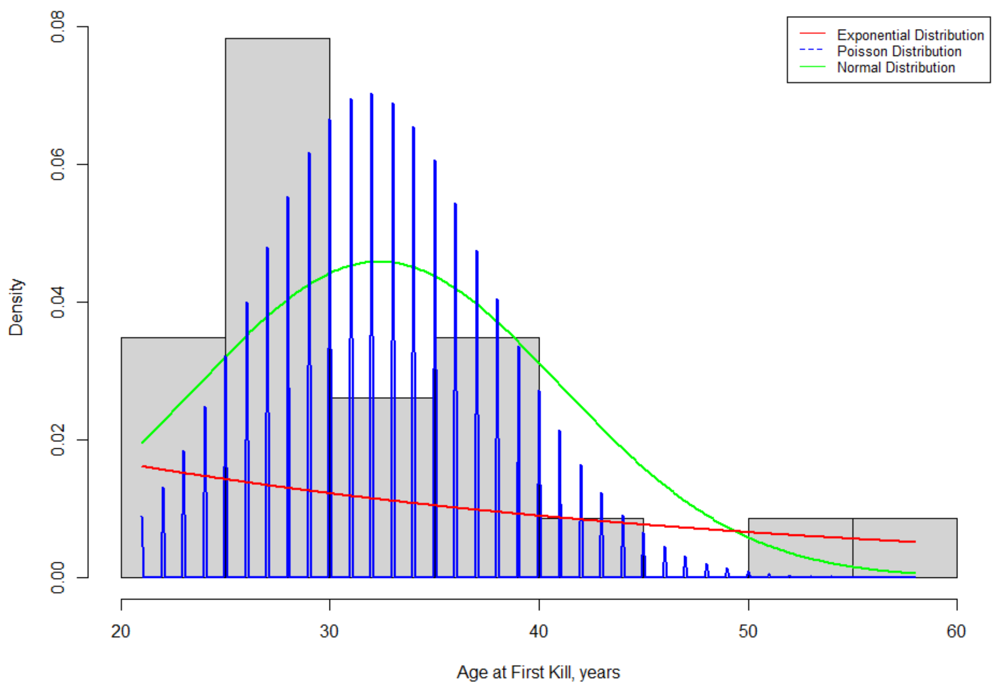

Figure 4 Density plot of the motive angel of death with different distributions

calculating confidence intervals, the data for this motive will be assumed as normal.

For Z-test, it will be misleading to assume that sample variance is equal to population variance since data points for all motives were less than 30 (a normal practice in industry). Therefore, in this scenario t-tests are more appropriate. Figure 5 clearly shows that the null hypothesis (shown with a red dotted line) must be rejected for all motives since it does not fall in the confidence interval of any motive. Although the average age at first kill for different motives is different, confidence intervals are overlapping mostly suggesting that the average does not differ significantly between different motives. For instance, the average age from 30 to 35 years cannot be rejected for all motives. The confidence interval for motive convenience (didn't want children/spouse) is large because of fewer data points. More data points may reduce the interval size and may even suggest strongly that the average does not differ for motives under discussion. This data also suggests that age at the first crime is not affected significantly by motive.

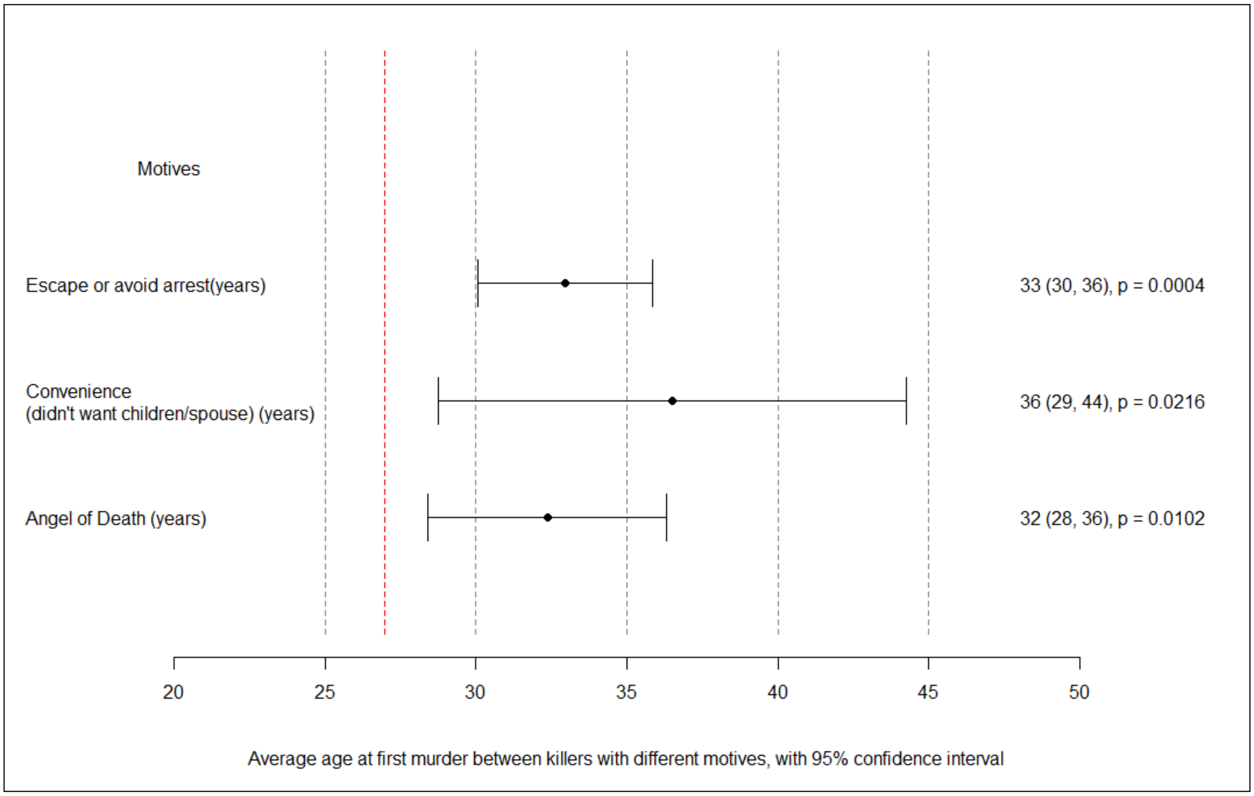

Figure 5 Forest plot of 95% confidence intervals for different motives having mean (lower level, upper level), p values for different motives.

4. DISCUSSION

The data suggests that there is no relation between age at first crime/career duration and the average age at first kill between different motives. However, having more points could possibly unearth findings which might not be consistent with these results as the data points were fewer, especially for our second motive which is Convenience. Furthermore, the confidence interval for the motive angel of death is based on assumption that data is normal which is not the case. The distribution of the log of age at first kill for the motive angel of death (or any other transformation) could be checked and then attempt calculating confidence interval if the distribution becomes normal to be more certain about the confidence intervals.

Commenting on why most killers are white males could be misleading as we don't know the circumstances. This could either mean that the investigation process is prejudiced against male or white males, or it could mean that women/other ethnicities tend to get away with crime. Also, it is possible that the samples were taken from the white male majority society. This needs to be further investigated or normalised before reaching conclusions.

APPENDIX – R CODE

```
######### CREATING UNIQUE DATA SET FOR COURSEWORK############

# unique sample was create and saved for later use

setwd("D:/MS Data Science/MATH 5741 STAT AND METHODS/Coursework")
#load(file = "killersandmotives.Rdata")
#createsample(201589942)
#save(mysample, file = "mysample.RData")
load(file = "mysample.Rdata")


############# CHECKING THE MOTIVES FOR THIS DATA SET ###########################

table(mysample$Motive)
# There are three different motives in this data set
# 1. Angel of Death
# 2. Convenience (Did not want children spouse)
# 3. Escape or avoid arrest
 ########## PREVIOUS RESEARCH #########
# Sigma square = variance = 74
# Meu = mean = 27

############################### DATA CLEANING ###################################

str(mysample)
head(mysample)
mysample[,]
rows = nrow(mysample)
rows
# There are total 71 observations in this Data Set stored in variables rows

# Three data quality issues were spotted,
# 99999 values in AgeFirstKill column
# NA values in motive
# NA values in sentence

mysample1 <- mysample[!(mysample$AgeFirstKill == 99999),]
str(mysample1)

#Entries that had 99999 in them
missing_agefirstkill <- sum(mysample$AgeFirstKill == 99999)
missing_agefirstkill
```

```
# so there are 9 enteries that needs to be removed because of missing value
perc_agefirstkill = missing_agefirstkill/rows * 100
perc_agefirstkill
# In terms of percentage there are 12.67606% values that are missing


# Creating Calculated column for removing killers that killed before 1900

mysample1$firstkillyear <- mysample1$AgeFirstKill + mysample1$YearBorn

str(mysample1)

#No. of Entries that are to be removed because the kills were before 1900
kills_before_1900 <- sum(mysample1$firstkillyear< 1900)
kills_before_1900
perc_kills_before_1900 = kills_before_1900/rows *100
perc_kills_before_1900
# In terms of percentage there 1.408451% of values that killed before 1900


mysample2 <- mysample1[!(mysample1$firstkillyear < 1900),]
str(mysample2)

# No of Entries that have no values in column motive
missing_motive <- sum(is.na(mysample2$Motive) == TRUE)
missing_motive

mysample2
mysample5 <- mysample2
# No of Entries that have no values in column Sentence
missing_sentence <- sum(is.na(mysample2$Sentence))
missing_sentence
perc_missing_sentence = missing_sentence/rows * 100
perc_missing_sentence
# In terms of percentage there 2.816901% of values that are missing in Sentence column

mysample3 <- na.omit(mysample2)  #(TO BE DISCUSSED)
mysample3$CareerDuration <- mysample3$AgeLastKill - mysample3$AgeFirstKill
str(mysample3)

mysample4 <- mysample2[!is.na(mysample2$Motive),]
mysample4

mysample5$CareerDuration <- mysample2$AgeLastKill - mysample2$AgeFirstKill
str(mysample5)
```

```
mysample5
# The career duration appears to be normal with most of the killers having career less than 5 years
# The career of one individual is 31 years which is surprising.
##########################################
#CHECK ITS IMPACT ON
max(mysample5$CareerDuration)

View(table(mysample5$Sex,mysample5$Race))

View(table(mysample5$CareerDuration,mysample5$Race))

View(table(mysample5$Sex,mysample5$CareerDuration))

View(table(mysample5$Motive,mysample5$AgeFirstKill))


mean(mysample5$AgeFirstKill)
  #### comparing all three motives#####################

avg_ageoffirstkill_motiveangelofdeath <- mean (mysample5$AgeFirstKill[which(mysample5$Motive ==
"Angel of Death")])
var_ageoffirstkill_motiveangelofdeath <- var(mysample5$AgeFirstKill[which(mysample5$Motive ==
"Angel of Death")])
sd_ageoffirstkill_motiveangelofdeath <- sd(mysample5$AgeFirstKill[which(mysample5$Motive ==
"Angel of Death")])
avg_ageoffirstkill_motiveangelofdeath
var_ageoffirstkill_motiveangelofdeath
sd_ageoffirstkill_motiveangelofdeath

avg_ageoffirstkill_convenience <- mean (mysample5$AgeFirstKill[which(mysample5$Motive ==
"Convenience (didn't want children/spouse)")])
var_ageoffirstkill_convenience <- var (mysample5$AgeFirstKill[which(mysample5$Motive ==
"Convenience (didn't want children/spouse)")])
sd_ageoffirstkill_convenience <- sd(mysample5$AgeFirstKill[which(mysample5$Motive == "Convenience
(didn't want children/spouse)")])
avg_ageoffirstkill_convenience
var_ageoffirstkill_convenience
sd_ageoffirstkill_convenience

avg_ageoffirstkill_escape <- mean (mysample5$AgeFirstKill[which(mysample5$Motive == "Escape or
avoid arrest")])
var_ageoffirstkill_escape <- var(mysample5$AgeFirstKill[which(mysample5$Motive == "Escape or avoid
arrest")])
sd_ageoffirstkill_escape <- sd(mysample5$AgeFirstKill[which(mysample5$Motive == "Escape or avoid
arrest")])
avg_ageoffirstkill_escape
```

```
var_ageoffirstkill_escape
sd_ageoffirstkill_escape


motive_firstkill_var <- c(var_ageoffirstkill_motiveangelofdeath, var_ageoffirstkill_escape,
var_ageoffirstkill_convenience)
motive_firstkill_mean <- c(avg_ageoffirstkill_motiveangelofdeath,avg_ageoffirstkill_escape,
avg_ageoffirstkill_convenience)
motive_firstkill_sd <- c(sd_ageoffirstkill_motiveangelofdeath,sd_ageoffirstkill_escape,
sd_ageoffirstkill_convenience)
motive <- c("angel of Death", "Escape", "Convenience")
df <- data.frame(motive, motive_firstkill_mean, motive_firstkill_var, motive_firstkill_sd)
df

Firstmurderaverage <- mean(mysample5$AgeFirstKill)
sd_firstmurder <- sd(mysample5$AgeFirstKill)
variance_firstmurder <- var(mysample5$AgeFirstKill)
Firstmurderaverage


# splitting the data according to motive

data_angeldeath <- mysample5[mysample5$Motive == "Angel of Death",]
data_escape <- mysample5[mysample5$Motive == "Escape or avoid arrest",]
data_convenience <- mysample5[mysample5$Motive == "Convenience (didn't want children/spouse)",]

# only dropping rows if all the columns are NA
data_angeldeath <- data_angeldeath[rowSums(is.na(data_angeldeath)) != ncol(data_angeldeath), ]
data_convenience <- data_convenience[rowSums(is.na(data_convenience)) != ncol(data_convenience),
]
data_escape <- data_escape[rowSums(is.na(data_escape)) != ncol(data_escape), ]

#checking data frames
data_angeldeath
data_escape
data_convenience


max(data_angeldeath$AgeFirstKill)

# numerical summaries

table(data_angeldeath$Sex)
table(data_convenience$Sex)
table(data_escape$Sex)
count(data_angeldeath$Sex)
```

```
table(mysample5$Sex)
table(mysample5$Motive)
table(mysample5$Race)
table(mysample5$Race, mysample5$Sex)
hist(mysample5$CareerDuration, freq = TRUE, xlab = "Career Duration, Years", ylab = " No of Cases",
labels = TRUE, main = "", cex = 1.5)
View(summary((mysample5$CareerDuration)))
hist(mysample5$AgeFirstKill, freq = TRUE, xlab = "Career Duration, Years", ylab = " No of Cases", labels =
TRUE, main = "", cex = 1.5)

table(data_angeldeath$Race)
table(data_convenience$Race)
table(data_escape$Race)

boxplot(AgeFirstKill ~ Motive, data = mysample5, ylab = "Age At First Murder, years")
stripchart(mysample5$AgeFirstKill ~ mysample5$Motive, vertical = TRUE, method = "jitter", pch = 19,
add = TRUE, col = "blue")

boxplot(AgeFirstKill ~ Race, data = data_angeldeath)
boxplot(AgeFirstKill ~ Race, data = data_convenience)
boxplot(AgeFirstKill ~ Race, data = data_escape)
?boxplot

plot (data_angeldeath$AgeFirstKill,data_angeldeath$CareerDuration)
plot (mysample5$AgeFirstKill,mysample5$CareerDuration)

quantile(data_angeldeath$AgeFirstKill, type = 1)

## box plot for all motives with Race
boxplot(AgeFirstKill ~ Race, data = data_angeldeath)
boxplot(AgeFirstKill ~ Race, data = data_escape)
boxplot(AgeFirstKill ~ Race, data = data_convenience)

boxplot(AgeFirstKill ~ Sex, data = data_angeldeath)
boxplot(AgeFirstKill ~ Race, data = data_escape)
boxplot(AgeFirstKill ~ Race, data = data_convenience)


################################################ MODELLING AND ESTIMATION
####################################
######################################################################################
#######################

# # Plotting histograms of all the motives
```

```
hist(mysample5$AgeFirstKill, freq = FALSE)

hist(data_angeldeath$AgeFirstKill, freq = FALSE, main = "", xlab = "Age at First Kill, years")
x <- seq(from = min(data_angeldeath$AgeFirstKill), to = max(data_angeldeath$AgeFirstKill), by = 0.1)
lines(x, dnorm(x, mean = avg_ageoffirstkill_motiveangelofdeath, sd =
sd_ageoffirstkill_motiveangelofdeath), lwd = 2, col = "green")
lines(x,dpois(x, lambda = avg_ageoffirstkill_motiveangelofdeath), lwd = 2, col = "blue")
lines(x,dexp(x, rate = 1/avg_ageoffirstkill_motiveangelofdeath), lwd = 2, col = "red")
legend( x = "topright", legend=c("Exponential Distribution", "Poisson Distribution", "Normal
Distribution"),col=c("red", "blue","green"), lty=1:2, cex=0.8)

hist(data_escape$AgeFirstKill, freq = FALSE)
x <- seq(from = min(data_escape$AgeFirstKill), to = max(data_escape$AgeFirstKill), by = 0.1)
lines(x, dnorm(x, mean = avg_ageoffirstkill_escape, sd = sd_ageoffirstkill_escape), lwd = 2, col = "dark
green")


hist(data_convenience$AgeFirstKill, freq = FALSE)
x <- seq(from = min(data_convenience$AgeFirstKill), to = max(data_convenience$AgeFirstKill), by = 0.1)
lines(x, dnorm(x, mean = avg_ageoffirstkill_convenience, sd = sd_ageoffirstkill_convenience), lwd = 2,
col = "red")

# not matching any frequencies. Lets generate CDFs for each motive

###### CDF Angel of Death ###########
#####################################

par(mfrow = c(1,1))

Fn_angeldeath <- ecdf(data_angeldeath$AgeFirstKill)

plot(Fn_angeldeath, verticals = TRUE, pch = NA)
x <- 1:500
G <- function(x){return(pnorm(x, mean = avg_ageoffirstkill_motiveangelofdeath, sd =
sd_ageoffirstkill_motiveangelofdeath))}

lines(x, G(x), col = "red3") # less vertical separation between Fx and Gx suggest normal distribution

# Perform the Kolmogorov-Smirnov test

ks.test(x = data_angeldeath$AgeFirstKill ,
    y = "pnorm",
    mean = avg_ageoffirstkill_motiveangelofdeath, sd = sd_ageoffirstkill_motiveangelofdeath)

# we fail to reject null hypothesis since P-value is 0.3 which is greater than 5% suggests normal
distribution, consistent with previous finding
```

```
# Q-Q plots

qqnorm(data_angeldeath$AgeFirstKill, main = "Angel of Death")

abline(a = avg_ageoffirstkill_motiveangelofdeath, b = sd_ageoffirstkill_motiveangelofdeath, col = "red")
# This does suggest to significant extent that distribution is normal

# SHAPIRO WILK TEST
shapiro.test(data_angeldeath$AgeFirstKill)

# P value is less than 5% suggesting that it is not normal distribution, rejects the null hypothesis

#Chi-squared goodness test
library(nortest)

pearson.test(data_angeldeath$AgeFirstKill)
# fails to reject null hypothese

###### CDF convenience ###########
##################################

Fn_conv <- ecdf(data_convenience$AgeFirstKill)

plot(Fn_conv, verticals = TRUE, pch = NA)

G <- function(x){return(pnorm(x, mean = avg_ageoffirstkill_convenience, sd =
sd_ageoffirstkill_convenience))}
lines(x, G(x), col = "red3")

# Perform the Kolmogorov-Smirnov test

ks.test(x = data_convenience$AgeFirstKill ,
      y = "pnorm",
      mean = avg_ageoffirstkill_convenience, sd = sd_ageoffirstkill_convenience)

# we fail to reject null hypothesis since P-value is 0.7 which is greater than 5% suggests normal
distribution, consistent with previous finding

# Q-Q plots

qqnorm(data_convenience$AgeFirstKill, main = "Convenience")

abline(a = avg_ageoffirstkill_convenience, b = sd_ageoffirstkill_convenience, col = 'red')
# QQ suggesting normality
```

```
# SHAPIRO WILK TEST
shapiro.test(data_convenience$AgeFirstKill)

# P value is greater than 5% suggesting that it is distribution, fail to reject the null hypothesis

#Chi-squared goodness test
library(nortest)

pearson.test(data_convenience$AgeFirstKill)

# fails to reject null hypotheses

###### CDF escape ###########
##############################


Fn_escp <- ecdf(data_escape$AgeFirstKill)

plot(Fn_escp, verticals = TRUE, pch = NA)

G <- function(x){return(pnorm(x, mean = avg_ageoffirstkill_escape, sd = sd_ageoffirstkill_escape))}
lines(x, G(x), col = "red3")

# suggests normality

# Perform the Kolmogorov-Smirnov test

ks.test(x = data_escape$AgeFirstKill ,
       y = "pnorm",
       mean = avg_ageoffirstkill_escape, sd = sd_ageoffirstkill_escape)
# P value is greater than 5% suggesting that it is distribution, fail to reject the null hypothesis

# Q-Q plots

qqnorm(data_escape$AgeFirstKill, main = "Escape or avoid arrest")

abline(a = avg_ageoffirstkill_escape, b = sd_ageoffirstkill_escape, col = "red")
# QQ suggesting normality

# SHAPIRO WILK TEST
shapiro.test(data_escape$AgeFirstKill)
#suggesting normality

#Chi-squared goodness test
library(nortest)
```

```
pearson.test(data_escape$AgeFirstKill)
#suggesting normality

df
c_results <- c("Test","Histogram", "CDF", "K-s test", "Q-Q plots", "Shapiro-Wilk", "Chi-squared")
c_results_angelofdeath <- c("Angel of Death","Not Normal","Normal","Normal","Not Normal","Not
Normal","Normal")
c_results_conv <- c("Convenience", "Normal", "Normal","Normal", "Normal","Normal", "Normal" )
c_results_escp <- c("Escape or Avoid arrest","Not Normal", "Normal", "Normal",
"Normal","Normal","Normal")
df.results <- data.frame(c_results,c_results_angelofdeath, c_results_conv,c_results_escp)
View(df.results)


##################################################INFERENECE#########################
###############
#####################################################################################
###############
#####################################################################################
###############
length(data_angeldeath$AgeFirstKill)
length(data_convenience$AgeFirstKill)
length(data_escape$AgeFirstKill)

# since my data set is less than 30 i would use T test
# it would be wrong to assume that sample variance is equal to population variance

############ t-tests############
#############################################
x1 <- data_angeldeath$AgeFirstKill
t.test(x1, alternative = "two.sided", mu = 27, conf.level = 0.95)

x2 <- data_convenience$AgeFirstKill
t.test(x2, alternative = "two.sided", mu = 27, conf.level = 0.95)

x3 <- data_escape$AgeFirstKill
t.test(x3, alternative = "two.sided", mu = 27, conf.level = 0.95)

# performing z test for population mean of 27


##############################################################################

# Analysis labels for the left side:
```

```
analysis = c("Angel of Death (years)",
          "Convenience
(didn't want children/spouse) (years)",
          "Escape or avoid arrest(years)")

# Results of each test (estimated mean,
# upper CI limit, lower CI limit, p-value):

estimate  =  c(32.36364, 36.5 , 32.95 )
upper     =  c(36.31404,44.24328, 35.84889)
lower     =  c(28.41323,28.75672, 30.05111)
pval      =  c(0.01018,0.02156, 0.00039 )


# 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(20, 50), 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("Average age at first murder between killers with different motives, 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(25, 30, 35, 40,45)){

  lines(c(i, i), c(0, 5), lty = 2, col = "gray53")

}

lines(c(27, 27), c(0, 5), lty = 20, col = "red1")

# 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)
mtext(text = "Motives",
      side = 2 , at= 4:4, 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 1:4 ){

  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)
```

```
P <- formatC(pval , format = 'f', digits = 4)

pval <- paste("p =", P)    # 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")
```