Analysis of serial killer's age at First kill with different motives

PhaniKanchan Adabala

ABSTRACT:

A serial killer can be defined as a person who murders people repeatedly, over a period of time, for some known or unknown reason. They murder people with the gaps of time between the murders, ranging from days to months, or years which is known as their career duration. In some cases, the exact number of people murdered by a serial killer is not known for several reasons. There is also a possibility that killer murdered more people even though he is not convicted for all crimes. However, it is important to find the average age at, which killers started their career depending upon the motives for killing people and see how their career duration lies on the scale of average. To investigate this, the samples are addressed using Normal and Exponential distributions and various hypothesis tests are conducted to check if the mean age at first kill differs between motives. These findings are interpreted using the sample mean, confidence interval and p-value for each motive.

DATASET:

Data set based on a sample of serial killers from the Radford/FGCU Serial Killer Database (https://www.fgcu.edu/skdb/). A sample out of the available serial killer dataset is created by function CREATESAMPLE(X) in R, for which the X value for this report is 36. This created a sample of 634 observations with 9 columns, of which only the below columns(measures) are of our major interest and more centric to our research.

- AgeFirstKill, the age of each killer (years) when they committed their first murder.
- AgeLastKill, the age of each killer (years) when they committed their last murder.
- Motive, the motive of each killer, if known (three different motives for sample).

PREPROCESSING:

The sample data was cleaned by removing any missing values in the column AgeFirstKill or motive. 9 such rows(1.4%) with uncertainties were removed from our dataset leaving us with 625 observations. Similarly, for 6 killers(0.9%), motives weren't noted, so by removing them, we now have 619 records. Also, we are not interested in the serial killers whose first kill was before the year 1900, there are 9 (1.4%) such records and after removing them, the sample data set now has 610 observations. Cleaning the dataset removed 24 observations(3.7%). A new column was introduced to the sample data set which denotes the total number of years a killer is active in the killing.

RESULTS:

DATA EXPLORATION:

The summaries of the three quantitative columns are tabled as below:

Parameter	AgeFirstKill	AgeLastKill	CareerDuration
Minimum	13.00	15.00	0.00
1st Quartile	23.00	25.00	0.00
Median	27.00	31.00	0.00

Mean	29.63	33.13	3.50
3rd Quartile	35.00	39.00	4.00
Maximum	75.00	77.00	39.00
Standard Deviation	9.08	10.93	6.61

Figure 1: Summaries of quantitative variables

Summaries can be better visualised with the help of the below boxplots.

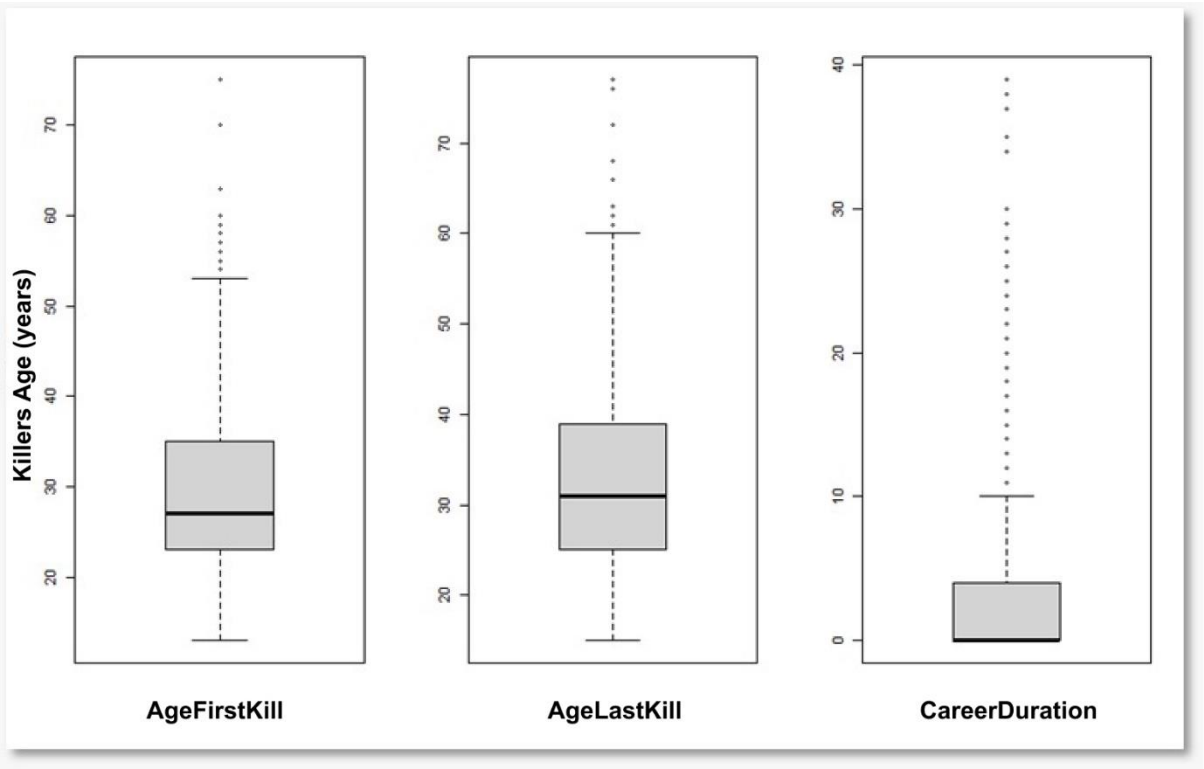

Figure 2: Box plots of the First and Last killing age of killers and their career duration showing some outliers and spread of Killers' Age

With respect to the relationship between these variables, only the age at first murder and last murder seem to have a fairly strong positive linear relationship with a correlation coefficient value of 0.7974.

Figure 3: Scatterplot of killers' age at first kill versus the last kill showing the positive linear relationship

DATA MODELLING :

By plotting the histogram of the three quantitative variables, we can approximately propose the normal distributions for AgeFirstKill, AgeLastKill and an exponential distribution for career Duration.

Age at First kill :

The histogram of the killers' Age at first kill across the density is unimodal and resembles a normal distribution with its peak coinciding with its mean(29.63 years) and has a standard deviation of 9.08. It was a bit skewed to right (mean > median). Its Normal Distribution is checked with the Normal Quantile plots where the approximately most number of points fall on the line.

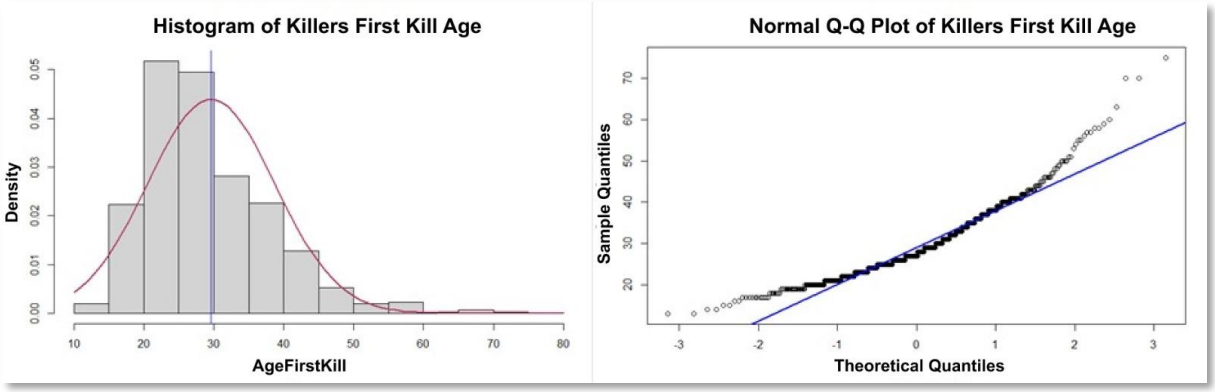

Figure 4: Histogram of AgeFirstKill against its density on left, the QQ plot showing the goodness of fit for its normal distribution

Killers Age at Last kill :

Similar to the Killers' first kill age, their last kill age also follows a normal distribution with a peak at its mean age of 33.13 years. QQ plot can show enough evidence to consider this distribution as a Normal one with a standard deviation of 10.93.

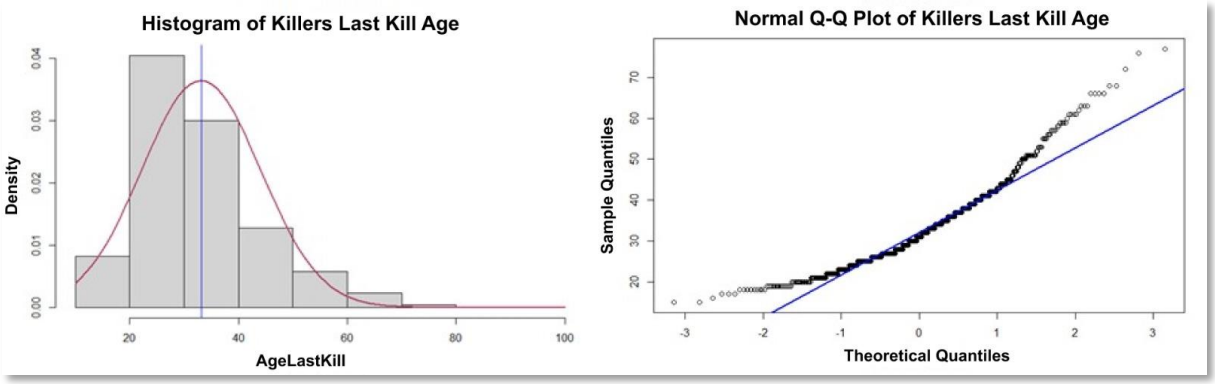

Figure 5: Histogram of AgeLastKill against its density on left, the QQ plot showing the goodness of fit for its normal distribution

Killers Career Duration :

The duration in which the killers are active followed an exponential distribution unlike the above both. 75% of the killers were active only for 4 or fewer years and the rate gradually decreased as the duration increased. This clearly showed an exponential curve on the histogram when plotted across its density. The QQ plot for the exponential distribution of this variable is shown on the right side in the fig6.

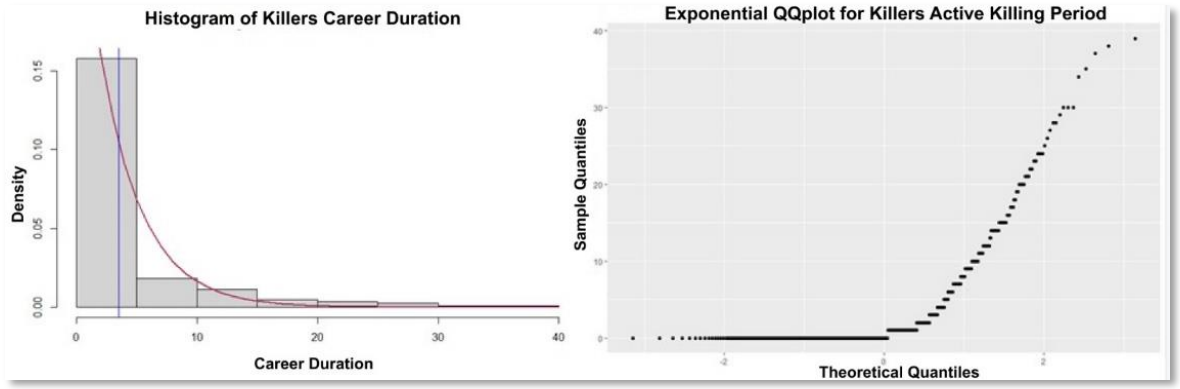

Figure 6: Histogram of Career Duration against its density on left, the QQ plot showing the goodness of fit for its Exponential distribution

It is evident that they don't exactly follow the proposed distributions and show some noise but we assume to be following these distributions approximately.

ESTIMATES:

The MOM and MLE estimators for the normally distributed variables (AgeFirstKill and AgeLastKill) are the estimated values of mu and sigma of the sample itself whereas it is lambda(inverse of its mean) for an exponential distribution. These values are estimated and visually peak on the graph of the log-likelihood function across the parameters.

Estimates	AgeFirstKill	AgeLastKill	CareerDuration
Distribution	Normal	Normal	Exponential
Mean	29.63	33.13	-
Variance	82.27	119.34	-
Rate	-	-	0.28

Figure 7: Estimates of the quantitative variables according to their distribution

TESTING HYPOTHESES :

The three motives for the sample with X value 36 are Angel of Death, Robbery or Financial Gain and unknown and the total number of killers with these motives are 23, 510 and 77 respectively. Analysing the Age of First Kill in these three datasets, it is evident that they follow an approximately normal distribution from figure 7 which is also validated by QQ Plots.

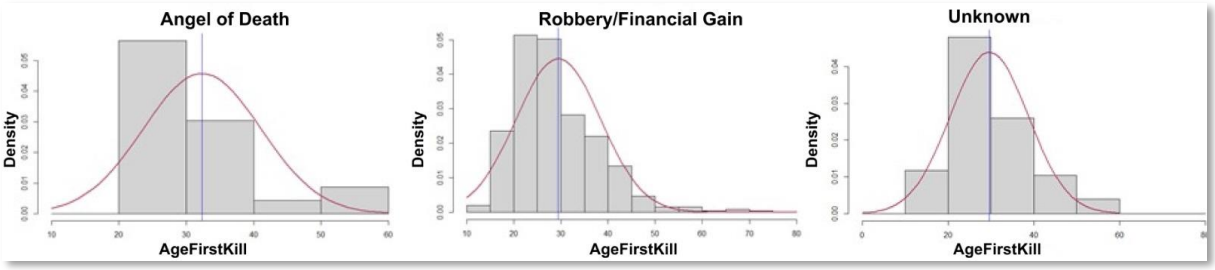

Figure 8: Normal Distribution curves of AgeFirstKill for all the motives

The histogram doesn't look great for Angel of Death, for such a small sample size(23). But the normal quantile plot appears to be very close to a straight line and the hypothesis tests don't reject the normality assumption.

Hypothesis :

Null Hypothesis: **H0: μ = 27** (the average age at first murder between killers with different motives is 27)

Alternative hypothesis: **H1: μ ≠ 27** (the average age at first murder between killers with different motives is not 27)

Assumptions:

- Samples are from Normal distribution (i.e. that X is normally distributed)
- Independent and identically distributed
- Population Variance is unknown

Based on the above assumptions, we opt for a t-test for these motives samples.

Parameters	Angel of Death(n=23)	Robbery/Financial gain(n=510)	Unknown(n=77)
Mean	32.35	29.40	29.63
Standard Deviation	8.70	8.96	9.07
alpha	5%	5%	5%
Confidence Intervals	[28.6, 36.1]	[28.7, 30.2]	[27.6, 31.7]
p-value	0.0074	1.235e-09	0.01011
H0	rejected	rejected	rejected
H1	Accepted	Accepted	Accepted

Figure 9: Summaries of the variable AgeFirstKill across the 3 motives at a 95% confidence level

We can conclude that since the mean age of 27 does not lie in any of the confidence intervals of the samples and also the p values are much lesser than the alpha(5%), we have strong evidence against the null hypothesis.

From the mean age value of the first kill across different motives, it can be said that the killers with the motives of Robbery or Financial gain and unknown start their killing career at a less age when compared to those with the motive of Angel of death.

COMPARISON OF DIFFERENT POPULATIONS:

Since the motives are independent and mutually exclusive, assuming normality as no killer can have two motives, we choose the Independent t-test for the pair of motives.

In order to run an independent t-test,

- One independent categorical variable(Motive) that has groups
- One continuous dependent variable(Age First Kill)

are required, which satisfies the requirement and is an appropriate two-sample hypothesis test to check if mean age difference at first kill exists between motives

Hypothesis :

Null Hypothesis: **H0: $\mu - \mu_0 = 0$** (true mean age difference at first murder between the two motives is zero)

Alternative hypothesis: **H1: $\mu - \mu_0 \neq 0$** (true mean age difference at first murder between the two motives is not zero)

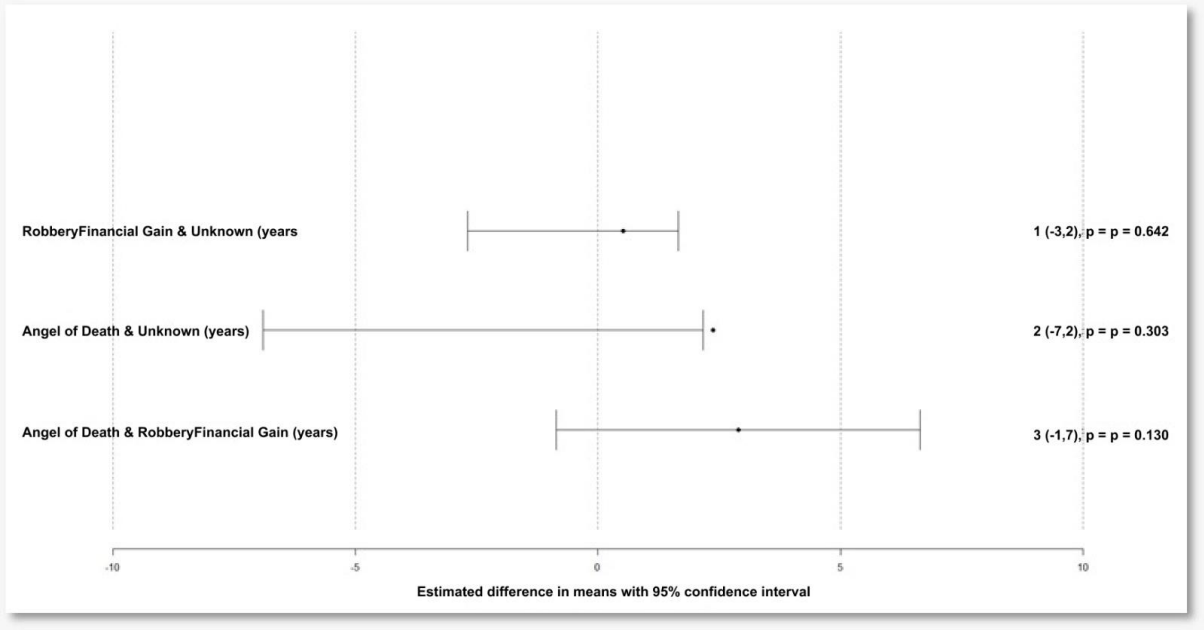

Figure 10: A caterpillar plot showing the estimated mean difference with CI for the pair of motives

We can infer from figure 9 that, the true mean age difference between any pair of motives is not equal to 0 and we reject the null hypothesis in this case accepting the alternate hypothesis.

INTERPRETATIONS:

The data is a bit skewed and we approximately assume the variables First Kill & Last kill to follow a normal distribution and Career Duration an exponential distribution. The career Duration has many outliers which could result in summaries that might be misleading in this case. We performed a t-test based on the assumptions and could come up with the 95% confidence intervals for the three motives. Surprisingly, the mean age differs concerning their motive for killing and is not 27 years. We can also assume the population variance as 74 (for serial killers who were active during the 1900s from previous research) and can do a z-test instead of a t-test, but this gives us the limitation (because the true population is unknown). There are some future research options on how the Age of the first murder killers can also vary with respect to the other attributes along with the motive so that we can build a predictive model to predict the sentence they can receive.

APPENDIX:

```
########################## IMPORTING LIBRARIES #######################################

install.packages("ggplot2")
library("ggplot2")

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

createsample(36)
mysample_rec <- dim(mysample) # total sample records #634

frstkill_cleaned <- mysample[mysample$AgeFirstKill != 99999, ] # (9 rows removed)
motive_cleaned <- frstkill_cleaned[!is.na(frstkill_cleaned$Motive), ] # (6 rows removed)
motive_cleaned$yearFirstKill <- motive_cleaned$YearBorn + motive_cleaned$AgeFirstKill
cleaned_ds <- motive_cleaned[motive_cleaned$yearFirstKill > 1900,] # (9 rows removed)

cleaned_ds$career_duration <- cleaned_ds$AgeLastKill - cleaned_ds$AgeFirstKill # new column
career duration

########################## DATA EXPLORATION #######################################

summary(cleaned_ds$AgeFirstKill,type = 1)
summary(cleaned_ds$AgeLastKill,type = 1)
summary(cleaned_ds$career_duration,type = 1)

par(mfrow = c(1,3))
boxplot(cleaned_ds$AgeFirstKill,xlab = 'AgeFirstKill',ylab = "Killers Age(years)")
boxplot(cleaned_ds$AgeLastKill,xlab = 'AgeLastKill' )
boxplot(cleaned_ds$career_duration,xlab = 'Career Duration')
```

```
plot(cleaned_ds$AgeFirstKill,cleaned_ds$AgeLastKill,xlab = "AgeFirstKill(years)",ylab =
"AgeLastKill(years)",
    main = "Scatter plot of Age at First murder versus Last murder")
cor(cleaned_ds$AgeFirstKill,cleaned_ds$AgeLastKil)


unique(mysample$Motive) # 3 unique motives 1)Angel of Death 2)Robbery or financial gain
3)Unknown


########################## DATA MODELLING ########################################


# AgeFirstKill column
par(mfrow = c(1,1))
hist(cleaned_ds$AgeFirstKill,freq = F ,
    breaks = seq(from = 10, to = 80,by = 5),
    xlab = "AgeFirstKill",
    main = "Histogram of Killers First kill Age")
lines(10:80,dnorm(10:80,mean = mean(cleaned_ds$AgeFirstKill),sd= sd(cleaned_ds$AgeFirstKill)),
     type = "l",lwd = 2, col = "maroon")
abline(v=mean(cleaned_ds$AgeFirstKill),col= "blue",lwd = 1)


qqnorm(cleaned_ds$AgeFirstKill,main = "Normal Q-Q Plot of Killers First Kill Age")
#  points lie on the same line so we assume a normal distribution
qqline(cleaned_ds$AgeFirstKill,col = "blue",lwd = 2)


# AgeLastKill column
hist(cleaned_ds$AgeLastKill,freq = F ,
    breaks = seq(from = 10, to = 100,by = 10),
    xlab = "AgeLastKill",
    main = "Histogram of Killers Last kill Age")
lines(10:100,dnorm(10:100,mean = mean(cleaned_ds$AgeLastKill),sd= sd(cleaned_ds$AgeLastKill)),
     type = "l",lwd = 2,col = "Maroon")
```

```
abline(v=mean(cleaned_ds$AgeLastKill),col= "blue",lwd = 1)

qqnorm(cleaned_ds$AgeLastKill,main = "Normal Q-Q Plot of Killers Last Kill Age")
# #  points lie on same line so we assume normal distribution
qqline(cleaned_ds$AgeLastKill,col = "blue",lwd = 2)

# CareerDuration column
hist(cleaned_ds$career_duration,freq = F,
    xlab = "Active Duration",
    main = "Histogram of Killers Career Duration")
lines(0:40,dexp(0:40,rate = (1/mean(cleaned_ds$career_duration)),log = F),
     type = "l",lwd = 2 ,col= "Maroon")
abline(v=mean(cleaned_ds$career_duration),col= "blue",lwd = 1)

qplot(sample = career_duration, data = cleaned_ds,
     main = "Exponential QQplot for Killers Active Killing period",
     xlab="Theoretical Quantiles",ylab="Sample Quantiles")

#########################    ESTIMATES   ########################################

n <- nrow(cleaned_ds) #total rows

# Agelastkill & AgeLastKill is a normal distribution and MOM and MLE are same in this case
mu_hat_agefirstkill <- mean(cleaned_ds$AgeFirstKill)
variance_hat_agefirstkill <- (sd(cleaned_ds$AgeFirstKill)^2) * ((n-1)/n)

mu_hat_agelastkill <- mean(cleaned_ds$AgeLastKill)
variance_hat_agelastkill <- (sd(cleaned_ds$AgeLastKill)^2) * ((n-1)/n)

# Career duration is following an exponential dist.
# so we estimate the parameter lambda as the inverse of its sample mean
```

```
lambda_hat_career_duration <- 1/mean(cleaned_ds$career_duration)
career_duration_mean <- mean(cleaned_ds$career_duration)



########################## TESTING HYPOTHESIS ####################################

# H0 : mu0 = 27 null hypothesis
# H1 : mu1 <> 27 alternate hypothesis

# Angel of death dataset
AOD_ds <-cleaned_ds[cleaned_ds$Motive == "Angel of Death",]
AOD_ds_mean <- mean(AOD_ds$AgeFirstKill)
AOD_ds_sd <- sd(AOD_ds$AgeFirstKill)

hist(AOD_ds$AgeFirstKill,freq = F,
    breaks = seq(from = 10, to = 60, by=10),
    xlab = "AgeFirstKill",
    main = "Angel of Death")
lines(10:60,dnorm(10:60,mean = AOD_ds_mean, sd = AOD_ds_sd ),
     type = "l",col = "maroon",lwd = 2)
abline(v = AOD_ds_mean, col = "blue",lwd = 1)

qqnorm(AOD_ds$AgeFirstKill)
qqline(AOD_ds$AgeFirstKill,col = "blue",lwd = 2)

# Robbery/ Financial Gain
ROFG_ds <- cleaned_ds[cleaned_ds$Motive == "Robbery or financial gain",]
ROFG_ds_mean = mean(ROFG_ds$AgeFirstKill)
ROFG_ds_sd = sd(ROFG_ds$AgeFirstKill)

hist((ROFG_ds$AgeFirstKill),freq=F,breaks = seq(from = 10, to = 80, by=5),
```

```
    xlab = "AgeFirstKill", main = "Robbery / Financial Gain")
lines(10:80,dnorm(10:80,mean = ROFG_ds_mean , sd = ROFG_ds_sd),type = "l",col = "maroon",lwd = 2)
abline(v = ROFG_ds_mean, col = "blue",lwd = 1)

qqnorm(ROFG_ds$AgeFirstKill)
qqline(ROFG_ds$AgeFirstKill,col = "blue",lwd = 2)

# unknown
UK_ds <- cleaned_ds[cleaned_ds$Motive == "Unknown",]
UK_ds_mean <- mean(UK_ds$AgeFirstKill)
UK_ds_sd <- sd(UK_ds$AgeFirstKill)

hist(UK_ds$AgeFirstKill,freq = F,breaks = seq(from = 0, to = 80, by=10),
    xlab = "AgeFirstKill", main = "Unknown")
lines(0:60,dnorm(0:60,mean = UK_ds_mean,sd = UK_ds_sd) ,type = "l",col = "maroon",lwd = 2)
abline(v = UK_ds_mean, col = "blue",lwd = 1)

qqnorm(UK_ds$AgeFirstKill)
qqline(UK_ds$AgeFirstKill,col = "blue",lwd = 2)



#all of the 3 motives for the first kill follow normal dist.

# Assumptions:
# here this is a normal distribution, with unknown population variance,
# so here we are going for a t-test.

# CI for Angel of death sample n = 23, alpha = 5%
t_AOD = qt(p = 0.975, df = 22)
CI_AODs = AOD_ds_mean + c(-1,1)*t_AOD*(sqrt((AOD_ds_sd^2)/(length(AOD_ds$AgeFirstKill))))# [28.6,36.1 ]
```

```
t.test(AOD_ds$AgeFirstKill, alternative = "two.sided", mu = 27, conf.level = 0.95)


# CI for Robbery / financial gain n= 510, alpha = 5%

t_ROFG = qt(p = 0.975,df = 509)

CI_ROFGs = ROFG_ds_mean + c(-
1,1)*t_ROFG*(sqrt((ROFG_ds_sd^2)/(length((ROFG_ds$AgeFirstKill))))) # [28.7,30.2]

t.test(ROFG_ds$AgeFirstKill, alternative = "two.sided", mu = 27, conf.level = 0.95)


# CI for Unknown n = 77,alpha = 5%

t_UK = qt(0.975,df = 76)

CI_UKs = UK_ds_mean + c(-1,1)*t_UK*(sqrt((UK_ds_sd^2)/(length(UK_ds$AgeFirstKill)))) #
[27.7,32.2]

t.test(UK_ds$AgeFirstKill, alternative = "two.sided", mu = 27, conf.level = 0.95)


################### COMPARISION OF DIFFERENT POPULATIONS ######################


# Angel of Death + Robbery or financial gain


t.test(x= AOD_ds$AgeFirstKill , y = ROFG_ds$AgeFirstKill ,

      paired = FALSE, var.equal = TRUE , conf.level = 0.95)


# Robbery or financial gain + Unknown


t.test(x= ROFG_ds$AgeFirstKill , y = UK_ds$AgeFirstKill ,

      paired = FALSE, var.equal = TRUE , conf.level = 0.95)


# Angel of Death + Unknown


t.test(x= UK_ds$AgeFirstKill , y = AOD_ds$AgeFirstKill ,

      paired = FALSE, var.equal = TRUE , conf.level = 0.95)
```

```
################ Forest plot of the above comparisons ############################

analysis = c("Angel of Death & Robbery/Financial Gain (years)",
             "Angel of Death & Unknown (years)",
             "Robbery/Financial Gain & Unknown (years)")

par(mfrow = c(1,1))
estimate  =  c(2.89, 2.37, 0.52)
lower     =  c(-0.86 , -6.91 , -2.69)
upper     =  c(6.64, 2.17, 1.66)
pval      =  c(0.1304,0.3026,0.6418)

par(mar = c(6,6,1,6))

plot(x = 0,                          # One point at (0,0).
    xlim = c(-10, 10), 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.

axis(side = 1, cex.axis = 1)
mtext("Estimated difference in means with 95% confidence interval",
     side = 1, line = 4)



for(i in c(-10,-5, 0, 5, 10)){

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

}
```

```
verticalpos = 1:3

mtext(text = analysis,  at = verticalpos,
     side = 2, line = 4, outer = FALSE, las = 1, adj = 0)
points(estimate, verticalpos, pch = 16)

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))

}

est <- formatC(estimate, format='f', digits = 0)
P <- formatC(pval , format = 'f', digits = 3)
pval <- paste("p =", P)
L <- formatC(lower, format = 'f', digits = 0)
U <- formatC(upper, format = 'f', digits = 0)

interval <- paste("(", L, ", ", U, "),", sep = "")
results <- paste(est, interval, pval)
mtext(text = results, at = verticalpos,
     side = 4, line = 4, outer = FALSE, las = 1, adj = 1)

box("inner")
```

R FILE :

Rfile_201574463_kanchan.R