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 Himanshu Jaiswal

Student ID 201577162

INTO THE MINDS OF SERIAL KILLERS

HIMANSHU JAISWAL
Student id: 201577162

INTRODUCTION

In this report we're going to study and analyse the killers dataset in order to understand the behaviour of the killers based on different factors but the main being motive of each of the killer. A killer can be called a serial killer when he/she kills more than one victim in more than one location in a very short period of time.

Based on this understanding and analysis we would be able to answer different behavioural questions like killers who are motivated by enjoyment or power start at a younger age on average to the ones motivated by angel of death or not? Does the average age at first murder differ between killers with different motives? 'Angel of Death,' 'Enjoyment or power,' and 'Escape or escape arrest,' which is the key factor of our study, are the varied motives of the killers in the dataset.

Data Cleaning:

For our analysis to be reliable it's important for our data to be consistent. Removal of rows whose 'AgeFirstKill' is missing (recorded as 99999) are 9 out of total rows (771) which is 1.16% of our data.

Missing values for 'Motive' are recorded as 'NA', therefore 6 rows are eliminated, accounting for 0.77 % of our data. Similarly for the killers whose first kill was before 1900 are removed which is 8 rows (1.03% of our data).

RESULTS & FINDINGS

Data Exploration

Table 1.1: Numerical summaries like IQR (25%), Mean, IQR (75%), Maximum, Standard Deviation for Age of first kill, Age of last kill and Career Duration of the killer.

	Inter Quantile Range (25%)	**Mean**	**Inter Quantile Range (75%)**	**Max**	**Standard Deviation**
Age of first Kill	24	30.66	35	66	8.44
Age of last kill	28	34.99	41	66	9.49
Career Duration	0	4.32	6	34	5.98

For better understanding of the data and distribution of these variables graphical summaries can be done as follows:

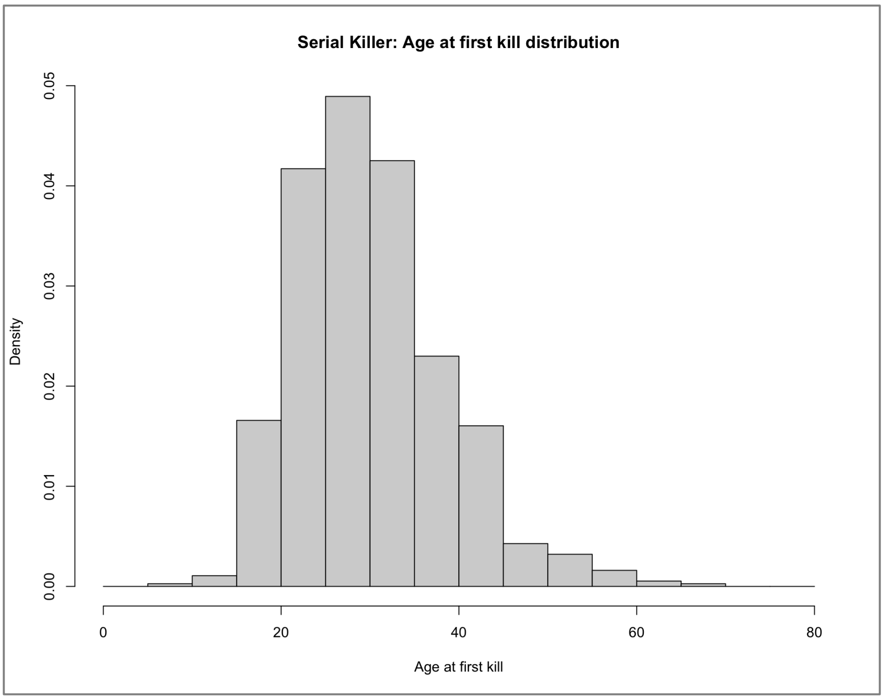

Fig (a): Histogram for the Age of first kill, showing densities for chosen interval.

The distribution of Age at first kill can be seen clearly from the Fig (a). As we have plotted the densities rather than the frequencies, the area of each bar is the proportion of our data contained in the corresponding interval. We can observe that there are many killers whose age at first kill is heavily skewed between age range of 20 to 35 and as we move towards extremes of the age(very low or high), lesser are the killers.

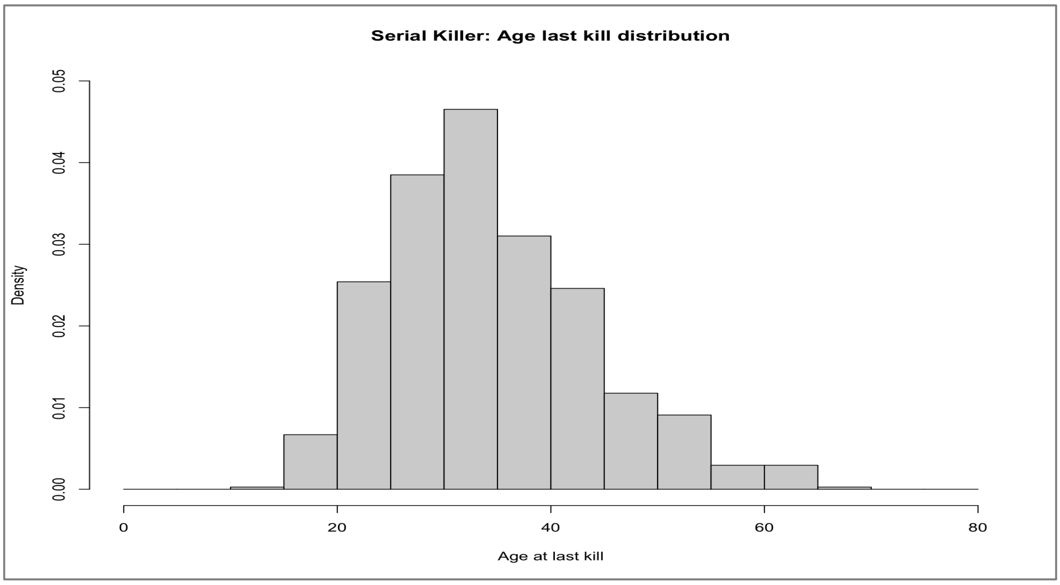

Fig (b): Histogram for the Age at last kill, showing the densities for chosen interval.

Similarly the density distribution of Age at last kill can be seen in Fig (b). The Age at last kill is heavily skewed in the age range of 25 to 40 and the skewness decreases at two extreme age ranges.

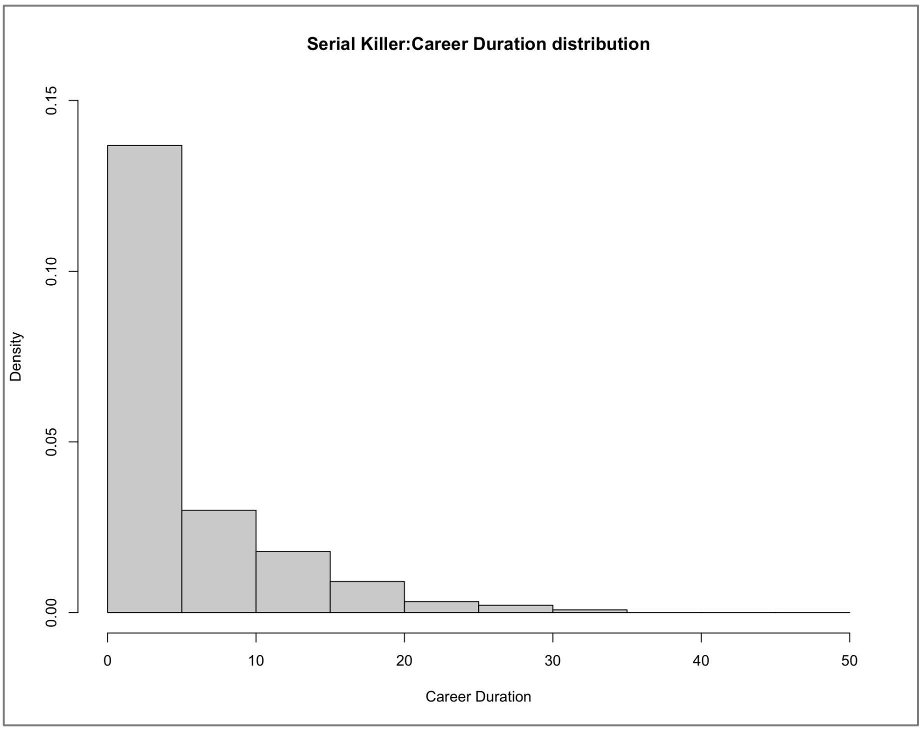

Fig (c): Histogram for the Career duration, showing the densities for chosen interval (in years).

The density distribution for the career duration can be seen in Fig (c). We can infer that the killers career duration mostly ranges from 0 to 5 years as it is highly skewed from 0 to 5, the skewness is seen to be gradually decreasing as the years of the career duration increases. Positive skew is very evident.

Relationships between these variables

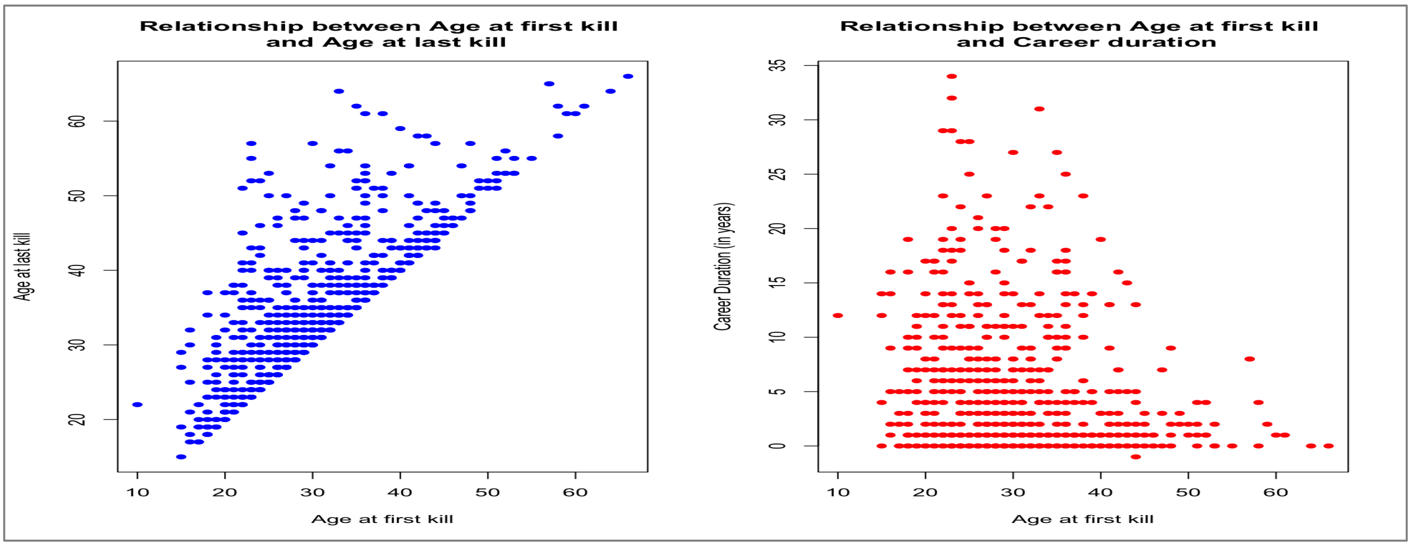

Fig (d): Relationship between Age at First kill, Age at last kill and Career duration (in years)

From the Fig (d) we can say that there is a linear relationship between the Age at first kill and age at last kill and they are highly correlated(0.78). We can infer that if the Age of first kill ranges from approximately 20 to 35 then the career duration of that killer is usually bigger which also makes sense from what I expected. The serial killers who starts to kill at younger age tend to have bigger career duration.

Modelling

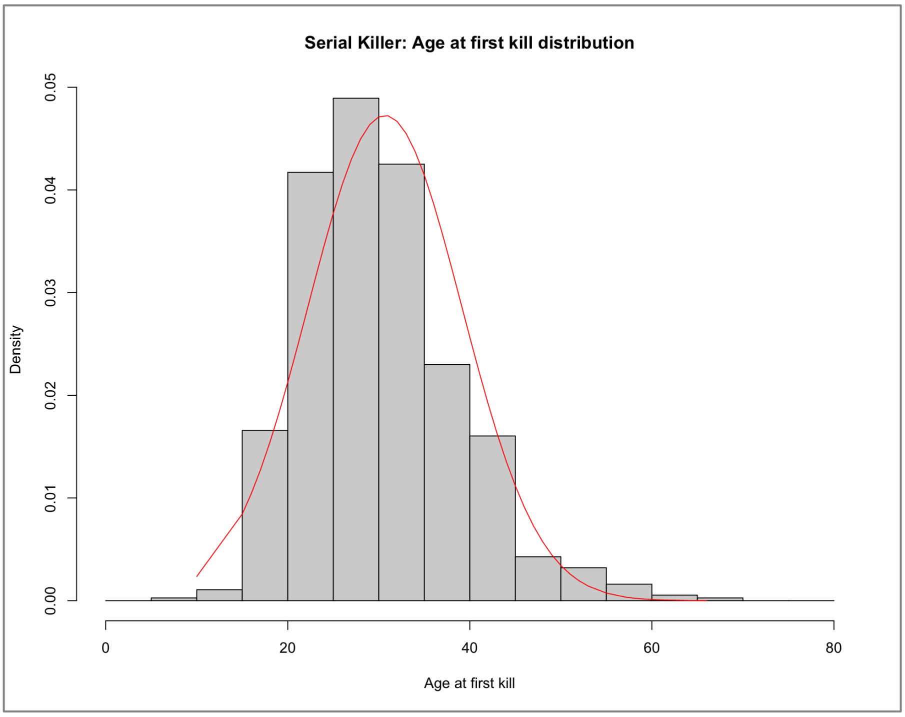

Fig (e): Histogram of Age at first kill, showing a roughly bell shaped distribution within our sample. A normal distribution density function plotted as a smooth curve (red curve) with a similar shape to our histogram (symmetrical with similar location and spread) is fitted.

From the histogram in Fig (e), we can propose the ~ N (μ, σ^2) for Age at first kill as it shows a roughly bell shaped distribution. The red curve which is a normally distributed density function plotted as smooth curve shows that the normal distribution would be a good fit for this model.

Similarly, for the histogram in Fig (f) below, Age at last kill ~ N(μ, σ^2), normally distributed density function (blue curve) approximately fits for the model.

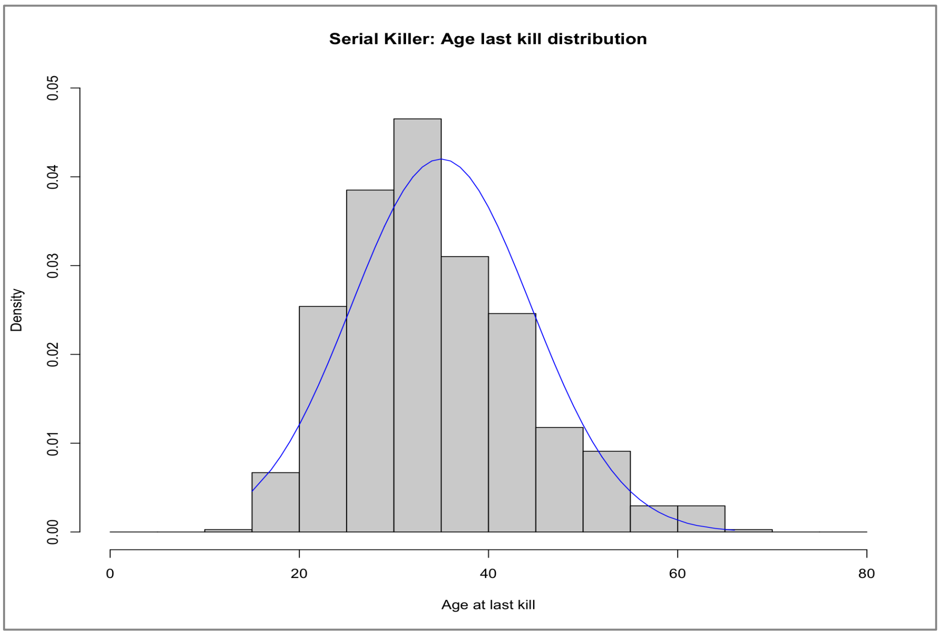

Fig (f): Histogram of Age at last kill, showing a roughly bell shaped distribution within our sample. A normal distribution density function plotted as a smooth curve (blue curve) with a similar shape to our histogram (symmetrical with similar location and spread) is fitted.

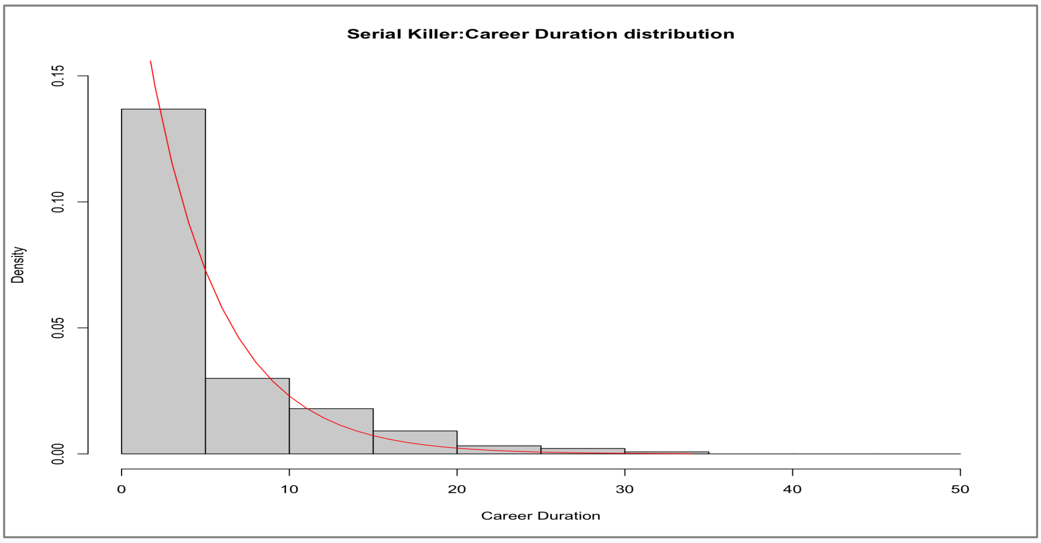

Fig (g): A histogram of career durations, showing non-negative sample with clear positive skew. An exponential distribution density curve (red curve) with a similar shape to our histogram (non-negative with a positive skew) fits.

From fig (g) we can see that the data for career durations is very skewed, with a long tail towards right. To model this we would need a distribution that reflects such asymmetry. Exponential distribution would be appropriate (red density curve fits) as it is very often used to model time-to-event variables.

Estimation

After we have chosen the sensible models for the distribution we now need to estimate the parameters as in real life we don't know their true values.

Table 1.2: Parameter estimation - Population mean (E(x)) & Population variance (σ^2) for Age at First kill Fig (e) & Age at last kill Fig (f) both having normal distribution with n=748 (very large).

Parameter Estimation	Estimating E(x) Population Mean	Estimating (σ^2) Population Variance	Reasoning for the estimation
Age of first kill	Sample mean (x_bar) ≈ Population mean (E(x)) ≈ 30.66	Sample Variance (s^2) ≈ Population variance (σ^2) ≈ 71.23	As n=748 is very large by Method of Moments
Age of last kill	Sample mean (x_bar) ≈ Population mean (E(x)) ≈ 34.99	Sample Variance (s^2) ≈ Population variance (σ^2) ≈ 90.18	As n=748 is very large by Method of Moments

Parameter estimation of rate parameter (λ) for the Career duration ~ Exp(λ) (Fig (g)) by **Method of Moments** is: x_bar ≈ 1/λ so λ ≈ 1/x_bar(sample mean) ≈ 0.2309.

We would estimate the parameter using MLE (maximum likelihood) in R and plot the log likelihood function as a curve for to confirm whether it is a good estimate, which returns λ ≈ 0.23 from fig (h).

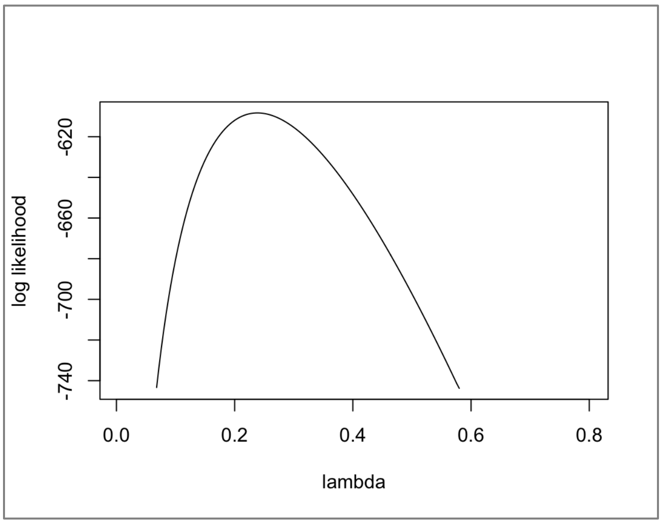

Fig (h): log likelihood function as a curve for λ which confirms λ ≈ 0.23 is a good estimation by MLE.

Table 1.3: Numerical summaries for Age of first kill of killers with three different motives considering **m1: ‘Angel of Death’, m2: ‘Enjoyment or power ‘**, **m3: ‘Escape or avoid arrest’**

	Minimum	Mean	Maximum	Standard Deviation
m1 (n=23)	21	32.34	58	8.70
m2 (n=703)	10	30.52	66	8.44
m3 (n=22)	23	33.54	53	7.46

Hypothesis Testing

Before carrying out Hypothesis testing it is important to check the normality each of m1, m2 & m3:

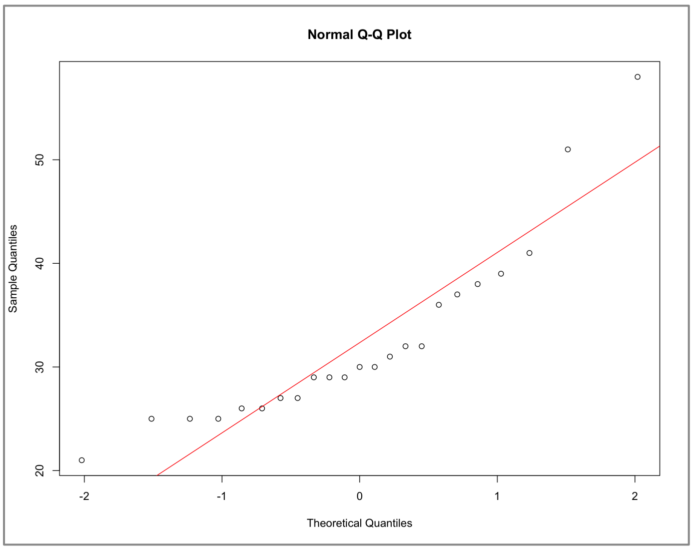

Fig (i): Normal Q-Q plot for m1 motive - Angel of Death

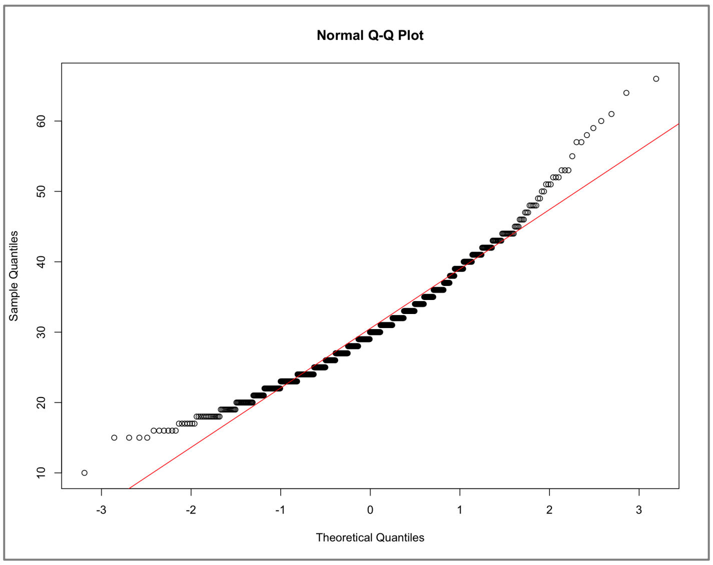

Fig (j): Normal Q-Q plot for m2 motive - Enjoyment or power

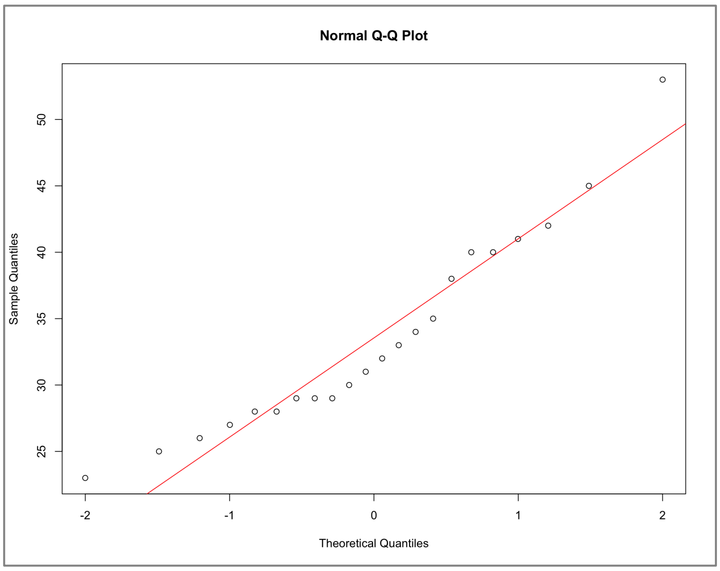

Fig (k): Normal Q-Q plot for m3 - Escape or avoid arrest

From fig (i), fig (j) & fig (k) we can say m1, m2 & m3 ≈ N(μ,σ^2) as the points lie roughly on or close to the straight red line with slight deviation in it, so now we can carry out hypothesis test on them.

Table 1.4: Results of hypothesis testing on each of m1, m2 & m3 where Null Hypothesis is H_0 : μ_0= 27 is the proposed average Age of first kill for each of three motives m1, m2 & m3. Alternative Hypothesis being H_1 : $\mu_0 \neq$ 27 at 95 % significance level.

Motives	Type of test performed	Sample Mean (years)	Result: Status of H_0	95% Confidence Interval	p-value	Z or T test statistic value
m1 (n=23)	t-test	32.34	Failed: rejection of H_0	28.55 to 36.11	0.007	2.946
m2 (n=703)	z-test	30.52	Failed: rejection of H_0	29.89 to 31.14	< 2.2e-16 ≈ 0	11.069
m3 (n=22)	z-test	33.54	Failed: rejection of H_0	30.01 to 37.07	0.0002	3.637

The population variance for **m1 motive 'Angel of Death'** is unknown as well as n=23 is small so we can not consider sample variance as population variance. Based on the normality assumption from fig (i) and these points, t-test seems to be an appropriate choice.

As H_0 fails, it acts as a evidence in favour of H_1 : $\mu_0 \neq$ 27 so we can infer that the average age of first kill for serial killers with motive 'Angel of Death' : m1 is not 27. Our C.I. suggests that 95 % this μ would be in range of {28.55 to 36.11} years, also as the p-value is less than 0.05, it justifies the rejection of H_0.

The sample variance ≈ population variance for **m2 motive 'Enjoyment or power'** as n=703 is large. Also based on the normality assumption from Fig (j) z-test is chosen. As H_0 fails, we can infer that the average age of first kill for killers with motive m2 is not 27. Our C.I suggests that there's a 95 % chance μ would be more here {29.89 to 31.14}.

Similarly for **m3 motive 'Escape or avoid arrest'**, even though n=22 is small, based on the normality assumption from Fig (k) we can apply z-test with help of C.L.T as all we need is Z-test statistic to be normal. As H_0 fails, we can infer that the average age of first kill for killers with motive m3 is not 27.

Comparison of killers with different motives

The killers with various motives are two distinct commodities that are completely independent to one another. Assumption of normality can be taken from Fig (i), Fig (j) & fig (k) for all m1, m2 & m3. Also the variance for all of our sample is different so based on these points 'Two sample t-test with independence' would be a good choice for carrying out hypothesis for comparison of different population.

Table 1.5: Results of two sample t-test with independence hypothesis testing on each pair of motives: {m1,m2}, {m2,m3}, {m1,m3} where Null Hypothesis is H_0 : μ_1- μ_2= δ = 0 i.e. for each pair of motives the μ would be same. Alternative Hypothesis being H_1 : δ ≠ 0 at 95 % significance level.

Pair of motives	Estimated mean difference of Age first kill	Result: Status of H_0	95% Confidence Interval	p-value
{m1,m2}	1.82	H_0 is true	-1.98 to 5.63	0.3324
{m2,m3}	3.02	H_0 is true	-6.38 to 0.33	0.0757
{m1,m3}	1.20	H_0 is true	-6.06 to 3.67	0.6225

For the pair {m1,m2} of motives the H_0 is true, so on an average there is no significant difference in age of first kill for the serial killers with motives **'Angel of Death'** & **'Enjoyment or power'** as μ_1- μ_2= δ = 0 lies in our 95 % CI, p-value is greater than 0.05 justifies that H_0 is true.

Similarly, for the pairs {m2,m3} & {m1,m3} H_0 is true, so we can infer that on an average the age of first kill does not differ significantly based on different motives but for one of the pair of motives {m2,m3} which are **'Enjoyment or power'** & '**Escape or avoid arrest'** we can't say this strongly as the estimated mean difference of age first kill here is 3.02 also p-value is close to 0.05 but as δ = 0 lies in our 95 % C.I. so the H_0 passed.

DISCUSSION

Despite the results obtained in Table 1.5 for pair of motives ‘Enjoyment or power’ & ‘Escape or avoid arrest’ the estimated mean difference of age kill was a bit significant also the p-value was close to 0.05 so we could not strongly infer that no matter what the motives are of killers, the average of first kill would be same. It might slightly be dependent on the motive of the killer as well which can be a potential point of further discussion.

The main objective of our analysis was **does the average age at first murder differ between killers with different motives?** From our findings we can say that on an average age of first murder does not differ significantly with different motives from Table 1.5 which were carried out based on certain assumptions. There were certain assumptions made which from my point of view were in doubt such as assumption of normality for m1 motive ‘Angel of Death’ Fig (i) as we can see it’s not entirely normally distributed but as the sample size for m1 was small (n = 23) & the population variance was unknown, we had to make the assumption for normality to carry out t-test as z-test was not a good option here.

APPENDIX

```
setwd("/Users/himanshoo/Course_Work/Statistical Theory & Methods/Practical_R")

load(file = "killersandmotives.Rdata")
createsample(16)

mysample

dim(mysample)

#unique(mysample$Race)
#unique(mysample$InsanityPlea)
#unique(mysample$Sentence)
unique(mysample$Motive)


#which(mysample$AgeFirstKill == '99999')
#mysample[mysample$AgeFirstKill=='99999',c('KillerID','AgeFirstKill')]

mysample <- mysample[mysample$AgeFirstKill!='99999',]
dim(mysample)    #771-9 rows = 762 rows remaining

mysample <- mysample[!is.na(mysample$Motive),]
dim(mysample)  # 762- 6 = 756

mysample <- mysample[ (mysample$AgeFirstKill + mysample$YearBorn ) > 1900, ]
dim(mysample) #756 - 8 = 748
#mysample$CareerDuration <- mysample$AgeLastKill - mysample$AgeFirstKill

mysample["CareerDuration"]<- mysample$AgeLastKill - mysample$AgeFirstKill

#data cleaning part done!

mys <- mysample

#Analysing AgeFirstKill

mean(mys$AgeFirstKill)    #30.66845

sd(mys$AgeFirstKill)      # 8.440022

var(mys$AgeFirstKill)
quantile(mys$AgeFirstKill)
max(mys$AgeFirstKill)

boxplot(mys$AgeFirstKill,
        ylab = "Age first kill (in years)",
       main = "Serial killer: Age first kill")

# Histogram to analyze variables
#FIRST VARIABLE : Age First Kill

#Normal Distribution of the age first kill


v <- seq(from = 0, to = 80, by = 5)

hist(mysample$AgeFirstKill ,
     xlab = "Age at first kill",
     ylab = "Density",
     main = "Serial Killer: Age at first kill distribution",
     breaks = v,
     freq = FALSE)

afk <- mysample$AgeFirstKill
afkm <- mean(mysample$AgeFirstKill)
afksd <- sd(mysample$AgeFirstKill)
afkdnorm <- dnorm(sort(afk), mean= afkm , sd = afksd)

lines( sort(afk), afkdnorm , type = "l" , col = "red" )


#Second VARIABLE : Age last Kill

mean(mysample$AgeLastKill)    #34.99198
max(mysample$AgeLastKill)    #66
sd(mysample$AgeLastKill)      #9.496561
quantile(mysample$AgeLastKill)

var(mys$AgeLastKill)

boxplot(mysample$AgeLastKill,
       ylab = "Age last kill (in years)",
       main = "Serial killer: Age last kill")

# Histogram to analyze variables

v <- seq(from = 0, to = 80, by = 5)

hist(mysample$AgeLastKill ,
     xlab = "Age at last kill",
     ylab = "Density",
     main = "Serial Killer: Age last kill distribution",
     breaks = v,
     freq = FALSE,
     ylim = c( 0 , 0.05))
```

```
alk <- mysample$AgeLastKill
alkm <- mean(mysample$AgeLastKill)
alksd <- sd(mysample$AgeLastKill)
alkdnorm <- dnorm(sort(alk), mean= alkm , sd = alksd)

lines( sort(alk), alkdnorm , type = "l" , col = "blue" )


# third Variable - Career Duration

mean(mysample$CareerDuration)    #4.323529

sd(mysample$CareerDuration)     #5.988801

quantile(mysample$CareerDuration)
max(mysample$CareerDuration)

boxplot(mysample$CareerDuration,
        ylab = "Career duration (in years)",
        main = "Serial killer: Career Duration")

# Histogram to analyze variables


v_cd <- seq(from = 0, to = 50, by = 5)

#length(mysample$CareerDuration)  #748
cd_data <- mysample[mysample$CareerDuration >=0,]

hist(cd_data$CareerDuration,
     xlab = "Career Duration",
     ylab = "Density",
     main = "Serial Killer:Career Duration distribution",
     breaks = v_cd,
     freq = FALSE,
     ylim = c( 0 , 0.15),
     right = FALSE)

#Exponential distribution of Career Duration

cd <- cd_data$CareerDuration
cdm <- mean(cd_data$CareerDuration)
cd_lambda = 1/cdm
cdsd <- sd(cd_data$CareerDuration)
cddnorm <- dexp(sort(cd) , rate = cd_l)
points( sort(cd), cddnorm , col = "red" , type = "l" )

#Estimating parameter lambda using maximum likelihood (prac 6)

#x1 <-sample(cd_data$CareerDuration,250)
```

```
#xbar <-mean(x1)

xbar <- cdm

loglik <- function(lambda){
  L <- (lambda^250)*exp(-lambda*250*xbar)
  return(log(L))
}

lambda <- (1:200)/250 # 4000 equally spaced points between 0 and 40.

plot(lambda, loglik(lambda), type = "l",
     xlab = "lambda", ylab = "log likelihood")


(1/xbar)   # by Mom value of our parameter lambda is 0.2309

cdm
cd_lambda #0.2309 in my case!


par(mfrow = c(1, 1))

plot(mysample$AgeFirstKill,mysample$AgeLastKill,
     pch = 16, cex = 1, col = "blue",
     main = "Relationship between Age at first kill \n and Age at last kill",
     xlab = "Age at first kill",
     ylab = "Age at last kill")


plot(mysample$AgeFirstKill,mysample$CareerDuration,
     pch = 16, cex = 1, col = "red",
     main = "Relationship between Age at first kill \n and Career duration",
     xlab = "Age at first kill",
     ylab = "Career Duration (in years)")

cor(mysample$AgeFirstKill,mysample$AgeLastKill)


dim(mys)

m1 <- mys[mys$Motive=='Angel of Death',c('AgeFirstKill')]
length(m1)  #23

m2 <- mys[mys$Motive=='Enjoyment or power',c('AgeFirstKill')]
length(m2) #703

m3 <- mys[mys$Motive=='Escape or avoid arrest',c('AgeFirstKill')]
length(m3)  #22
```

```
mean(m1)  #32.34
sd(m1)    #8.70
min(m1)   #21
max(m1)   #58

qqnorm(m1)
abline(mean(m1),sd(m1), col = "red")     #not normality qqnorm normality
#hist(m1)
t.test(m1, alternative = "two.sided", mu = 27, conf.level = 0.95)

#One Sample t-test

#data:  m1
#t = 2.9462, df = 22, p-value = 0.007469
#alternative hypothesis: true mean is not equal to 27
#95 percent confidence interval:
 # 28.58336 36.11229
#sample estimates:
 # mean of x
#32.34783


mean(m2)  #30.52
sd(m2)    #8.44
min(m2)   #10
max(m2)   #66

qqnorm(m2)
abline(mean(m2),sd(m2), col = "red")  #Normality check done by qnorm
#hist(m2)
z.test(m2,alternative = "two.sided",mu = 27,sigma.x = afksd,conf.level = 0.95)

#One-sample z-Test

#data:  m2
#z = 11.069, p-value < 2.2e-16
#alternative hypothesis: true mean is not equal to 27
#95 percent confidence interval:
# 29.89957 31.14737
#sample estimates:
#  mean of x
#30.52347


mean(m3)  #33.54
sd(m3)    #7.46
min(m3)   #23
max(m3)   #53
```

```
qqnorm(m3)
abline(mean(m3),sd(m3), col = "red")    #normality check done by qnorm
#hist(m3)

z.test(m3,alternative = "two.sided",mu = 27,sigma.x = afksd,conf.level = 0.95)

#One-sample z-Test

#data:  m3
#z = 3.6375, p-value = 0.0002753
#alternative hypothesis: true mean is not equal to 27
#95 percent confidence interval:
  #30.01866 37.07225
#sample estimates:
 # mean of x
#33.54545



#Two sample t- test

t.test(x = m1, y = m2, mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

#Welch Two Sample t-test

#data:  m1 and m2
#t = 0.98991, df = 23.376, p-value = 0.3324
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
  #-1.984670  5.633381
#sample estimates:
 # mean of x mean of y
#32.34783  30.52347


t.test(x = m2, y = m3, mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

#Welch Two Sample t-test

#data:  m2 and m3
#t = -1.8609, df = 22.715, p-value = 0.07575
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
  #-6.3836822  0.3397148
#sample estimates:
 # mean of x mean of y
#30.52347  33.54545
```

```
t.test(x = m1, y = m3, mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

#Welch Two Sample t-test

#data:  m1 and m3
#t = -0.49599, df = 42.513, p-value = 0.6225
#alternative hypothesis: true difference in means is not equal to 0
#95 percent confidence interval:
 # -6.068838  3.673581
#sample estimates:
 # mean of x mean of y
#32.34783  33.54545
```