ANALYSIS OF SERIAL KILLERS AND THEIR MOTIVES

Rohini Krishna Preetha (201669138)

INTRODUCTION

A serial killer is a person who commits at least two separate murders over a term, ranging from hours to years. To comprehend the patterns and forecast the likelihood of future killings, the motives of serial killers have to be understood. Previous studies have revealed that serial murderers who operated throughout the 1900s had a population mean age of 27 years and a variance of 74. The main motives of serial killers in the given data set are anger, convenience and enjoyment or power. This study aims to determine whether these motives affect the average age at first killing.

DATA CLEANING

Records containing "N/A" and "99999" values in the columns ‘Motive’ and ‘AgeFirstKill’, respectively, were removed during cleaning. The research used the data of killers who only started their careers after 1900. To identify and delete any outliers, such as negative or unreasonably lengthy durations, a new column called 'CareerDuration' was added based on the difference between age at last kill and age at first kill. Table 1 shows the specifics of the records eliminated and Table 2 displays the details of cleaned data set.

Criteria for removal	Value	Number of records removed	Percentage of records removed
Missing value in ‘Motive’	NA	6	6/942=0.64%
Missing value in ‘AgeFirstKill’	99999	9	9/942=0.96%
Started career before 1900	<1900	7	7/942=0.74%
Negative value in ‘CareerDuration’	<0	1	1/920=0.11%
Total		23	23/942=2.44%

Table 1: Details of eliminated records

Total number of rows	919
Total number of columns	10

Table 2: Summary of cleaned dataset

DATA EXPLORATION

Table 3 shows the number of males, and females, motive-based count of killers, and average career duration (estimated using the median because the median is more robust than the mean and less affected by outliers). Figure 1 and Figure 2 show the boxplots of the periods in which the murderers' careers began and terminated, respectively. It can be presumed that most killers began their professions in the 1970s and finished them in the 2000s and that most of them had careers that lasted, on average, roughly two years.

Criteria	Detail	Count	Median of career duration	
Gender	Male	891	2	
	Female	28	2	
Motive	Anger (including mission-oriented killers)	207	3	2
	Convenience (didn’t want children/spouse)	10	3	
	Enjoyment or power	702	2	

Table 3: Gender and motive wise summary of career duration

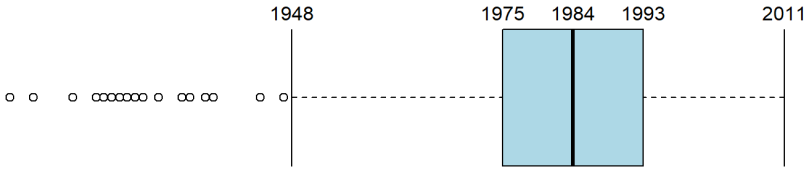

Figure 1: Box plot showing the start of career

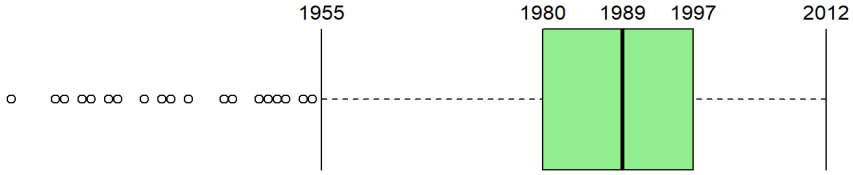

Figure 2: Box plot showing the end of career

Figure 3 is a boxplot that displays the spread of the age at the first kill, according to motive. It was noticed that most of the murderers under each motive began their careers at various ages, such as those motivated by anger starting first, then those who enjoyed or needed power, and finally those motivated by convenience, around the late 20s and early 30s.

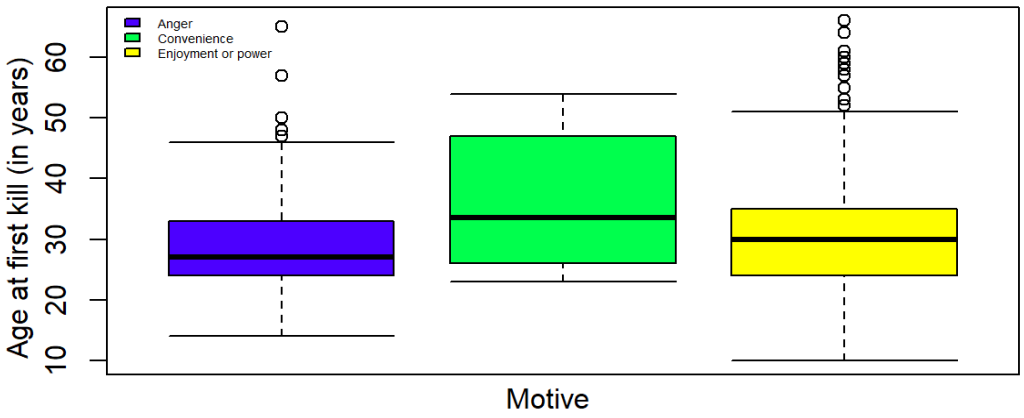

Figure 3: Boxplot depicting the motive-wise spread (quantiles) of age at first kill

DATA MODELLING AND ESTIMATION

The aforementioned computed numbers simply displayed a summary of the sample that was provided. Finding the probability distribution was necessary to identify trends in the population as a whole.

		Mean	Standard deviation	Variance	Min	$Q_{0.25}$	$Q_{0.5}$ (Median)	$Q_{0.75}$	Max
Entire sample		30.23	8.425375	70.98694	10	24	29	35	66
Motive	Anger	28.99	8.070677	65.13583	14	24	27	33	65
	Convenience	36.5	10.82436	117.1667	23	26	33	47	54
	Enjoy or power	30.5	8.439555	71.22609	10	24	30	35	66

Table 4: Age at first murder for killers with various motives, summarised numerically

The numerical summaries such as mean, median, variance, standard deviation and quantiles of age at first kill are shown in Table 4. Quantiles, instead of moments, were considered for analysis purpose due to the presence of potential outliers.

From the distributions of age at first kill for each motive in the sample, shown in Figure 4, it can be assumed that all three motives—enjoy or power, convenience, and anger—follow normal distribution because they all generated bell-shaped curves. Their normality was examined using Q-Q plots, depicted in Figure 5. For the motives, anger and enjoyment or power, the majority of the points fell along the reference lines with the slope as the mean and intercept as the standard deviation of age at the first kill, supporting their normality. Since the sample size of convenience is too small, it is challenging to fit the whole set of points in a straight line.

Figure 4: Histograms displaying the distribution of age at first kill for the motives of enjoyment or power, convenience, and anger, respectively.

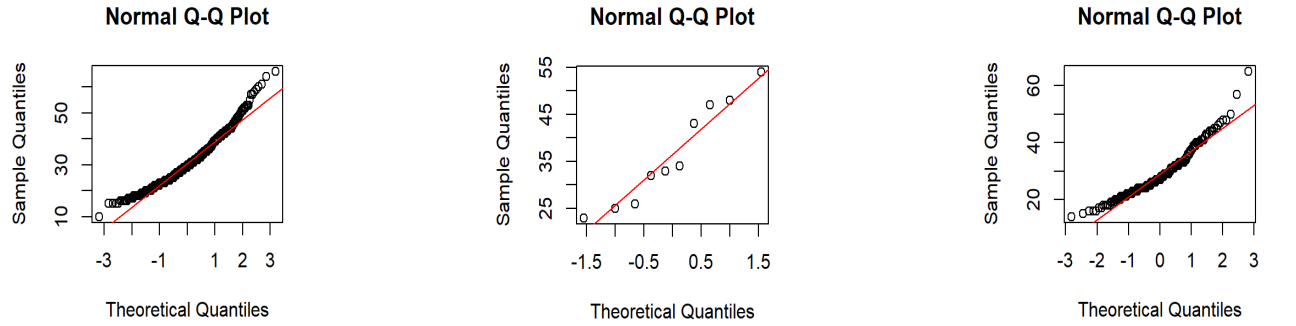

Figure 5: Q-Q plots of age at first kill for the motives of enjoyment or power, convenience, and anger, respectively.

To estimate the unknown parameters in the suggested model, the method of moments (MOM) estimation could be employed as follows.

As per the assumption, all the three motives follow normal distribution. For normal distribution, sample mean and variance are the population mean and variance.

Mean

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \qquad (1)$$

$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (2)$$

Variance

$$\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad (3)$$

Using equations (1), (2) and (3), the population mean and variance for each motive were calculated as shown in Table 5.

Motive	$\hat{\mu}$	$\hat{\sigma}^2$
Anger (n=207)	28.99	65.13583
Enjoyment or power (n=702)	30.5	117.1667
Convenience (n=10)	36.5	71.22609

Table 5: MOM estimate of mean and variance for each motive

RESULTS

According to a popular research, serial killers have a mean age of μ=27 years and a variance of σ^2=74. Consequently, it leads to the hypothesis that the age at first kill is normally distributed, with X ~ N (27,74). Apparently, the observed sample mean for each of the motives differed as can be seen from Table 4. Therefore, a two-sided hypothesis test was used to verify this contradiction for all three motives, with:

The null hypothesis, H_0: μ_0=27 (the mean age of the killers is equal to 27)

The alternative hypothesis, H_1: $\mu_0 \neq 27$ (the mean age of the killer is not equal to 27)

Due to the larger sample size for the motives of anger and enjoyment or power, a z-test was used, whereas a t-test was used for convenience due to its small size. The results of the z-tests are shown in Table 6 and t-test in Table 7 respectively.

Assumptions used for z-test:

Age at first kill for the motives anger and enjoyment or power, follows a normal distribution, and the sample size is very large. Although the population variance is unknown, it is estimated to be equal to the sample variance. A 5% significance level and a 95% confidence interval are used.

Motive	Z-statistic	p-value	Confidence Interval	Mean	Result
Anger	3.329	0.0008716	27.81851-30.16216	28.99034	Rejected H_0
Enjoy or power	10.794	2.2e-16	29.86795-31.14060	30.50427	Rejected H_0

Table 6: z-test results of the motives anger and enjoyment or power

Assumptions used for t-test:

Age at first kill for the motive convenience follows normal distribution, even though its sample size is very small. A 5% significance level and a 95% confidence interval are used and variance is unknown.

Motive	T-statistic	Degrees of freedom	p-value	Confidence Interval	Mean	Result
Convenience	2.7754	9	0.02156	28.75672-44.24328	36.5	Rejected H_0

Table 7: t-test results of the motive convenience

DISCUSSION

The p-values for all of the motives were found to be less than 5%, consequently, the results rejected the null hypothesis of the killers being 27 years old on average. The lower limits of the confidence intervals were all greater than 27. Also, it was noticed that the average age of killers varies for different motives. However, because the two-sample test was not carried out, this was not tested and cannot be assumed to be so.

The main advantage of the sample for the motive ‘Enjoyment or power and anger had 702 and 207 observations respectively, which strengthens the analysis due the large sample size.

Limitations:

The histogram obtained for the motive 'Convenience' appeared to be of uniform distribution, although its curve was bell-shaped. The assumption that q-q plots follow a normal distribution was made. Because of these assumptions, the analysis has a flaw. Furthermore, due to the lack of sufficient data, the t-test resulted in a wider confidence interval spanning from 28.76 to 44.24. The result rejected the null hypothesis solely because the bottom limit of the confidence interval was greater than 27. If the population mean was considered to be some value between 28 and 44, suppose 44, the null hypothesis would have been rejected. The median, on the contrary, suggests that it is far lower than 44. As a result, the hypothesis tests cannot yield conclusive results.

Further research can be undertaken in unexplored areas such as sex, race, insanity plea, and sentence to determine if these provide a basis for the killer to begin or continue his profession.

ACKNOWLEDGEMENT

I want to express my gratitude to Dr. Benjamin, Adia, and Harsha for their unwavering assistance during this analysis task.

APPENDIX

```
#Loading and saving the dataset
#createsample(201669138)
save(mysample, file = "mysample.RData")

#Counting the number of invalid/missing values
nrow(mysample)
sum(is.na(mysample$Motive))
sum(mysample$AgeFirstKill==99999)
sum(rowSums(mysample[,c(2,4)])<1900)

#Removing unwanted records
mysample<-subset(mysample,mysample$AgeFirstKill!=99999)
mysample<-mysample[!is.na(mysample$Motive),]
mysample<-mysample[rowSums(mysample[,c(2,4)])>=1900,]

#Creating CareerDuration column,checking and removing invalid values
mysample$CareerDuration<-mysample$AgeLastKill-mysample$AgeFirstKill
mysample$CareerDuration>=100
sum(mysample$CareerDuration<0)
mysample<-subset(mysample,mysample$CareerDuration>=0)

#Finding the unique motives in the dataset
unique(mysample$Motive)

#Data exploration
n=nrow(mysample)
sum(mysample$Sex=='Male')
sum(mysample$Sex=='Female')
quantile(subset(mysample$CareerDuration,mysample$Sex=='Male'))
```

```
quantile(subset(mysample$CareerDuration,mysample$Sex=='Female'))
quantile(mysample$CareerDuration,type=1)
anger<-subset(mysample,mysample$Motive=='Anger (including mission-oriented killers)')
quantile(anger$CareerDuration)
convenience<-subset(mysample,mysample$Motive=='Convenience (didn\'t want children/spouse)')
quantile(convenience$CareerDuration)
enjoy<-subset(mysample,mysample$Motive=='Enjoyment or power')
quantile(enjoy$CareerDuration)

timeperiodstart<-(mysample$AgeFirstKill+mysample$YearBorn)
timeperiodend<-(mysample$AgeLastKill+mysample$YearBorn)
boxplot(timeperiodstart,horizontal = TRUE, axes = FALSE, staplewex = 1, col="lightblue")
text(boxplot.stats(timeperiodstart)$stats, labels = boxplot.stats(timeperiodstart)$stats, y=1.25)
boxplot(timeperiodend,horizontal = TRUE, axes = FALSE, staplewex = 1, col="lightgreen")
text(boxplot.stats(timeperiodend)$stats, labels = boxplot.stats(timeperiodend)$stats, y=1.25)

#Finding the numerical summaries of age at first kill for variance motives
mean(mysample$AgeFirstKill)
variance=sd(mysample$AgeFirstKill)^2
quantile(mysample$AgeFirstKill)

mean(anger$AgeFirstKill)
sd(anger$AgeFirstKill)
variance_a=(sd(anger$AgeFirstKill))^2
quantile(anger$AgeFirstKill,type=1)

mean(convenience$AgeFirstKill)#36.5
sd(convenience$AgeFirstKill)#10.82436
variance_c=(sd(convenience$AgeFirstKill))^2
quantile(convenience$AgeFirstKill,type=1)
```

```
mean(enjoy$AgeFirstKill)

sd(enjoy$AgeFirstKill)

variance_e=(sd(enjoy$AgeFirstKill))^2

quantile(enjoy$AgeFirstKill,type=1)



#Plotting the graphical summary of age at first kill based on motives

boxplot(mysample$AgeFirstKill ~ mysample$Motive,ylab="Age at first kill (in
years)",xlab="Motive",

      col = topo.colors(3),xaxt="n",staplewex = 1)

legend("topleft",bty="n", fill = topo.colors(3),c("Anger","Convenience","Enjoyment or
power"),cex=0.45)



install.packages("tidyverse")

library(tidyverse)



# Plotting histogram for the motive anger

mu_anger<-mean(anger$AgeFirstKill)

sigma_anger<-sd(anger$AgeFirstKill)

v_a <- seq(from = min(anger$AgeFirstKill), to = max(anger$AgeFirstKill), by = 1)

hist(anger$AgeFirstKill,freq=F,breaks=v_a,col="lightgreen",ylim=c(0,0.1),xlab="Age at first kill (in
years)",main = "Anger")

lines(v_a,dnorm(v_a,mu_anger,sigma_anger), col = "red", lwd = 2,type='l')

#Plotting q-q plot for the motive anger

qqnorm(anger$AgeFirstKill)

abline(a = mu_anger, b = sigma_anger, col = "red")
```

```
# Plotting histogram for the motive convenience

mu_convenience<-mean(convenience$AgeFirstKill)

sigma_convenience<-sd(convenience$AgeFirstKill)

v_c <- seq(from = min(convenience$AgeFirstKill), to = max(convenience$AgeLastKill), by = 1)

hist(convenience$AgeFirstKill,breaks=v_c,freq=F,col="lightblue",ylim=c(0,0.15),xlab="Age at first kill (in years)",main = "Convenience")

lines(v_c,dnorm(v_c,mu_convenience,sigma_convenience),col="red",lwd=2,type='l')

#Plotting q-q plot for the motive convenience

qqnorm(convenience$AgeFirstKill)

abline(a = mu_convenience, b = sigma_convenience, col = "red")

# Plotting histogram for the motive enjoy

mu_enjoy<-mean(enjoy$AgeFirstKill)

sigma_enjoy<-sd(enjoy$AgeFirstKill)

v_e <- seq(from = min(enjoy$AgeFirstKill),to = max(enjoy$AgeLastKill), by = 1)

hist(enjoy$AgeFirstKill,breaks=v_e,freq=F,col="lightyellow",xlab="Age at first kill (in years)",ylim=c(0,0.08),main="Enjoyment or power")

lines(v_e,dnorm(v_e,mu_enjoy,sigma_enjoy), col = "red", lwd = 2,type='l')

#Plotting q-q plot for the motive enjoy

qqnorm(enjoy$AgeFirstKill)

abline(a = mu_enjoy,b=sigma_enjoy, col = "red")

#Null hypothesis testing using z and t tests.

library("BSDA")

z.test(mysample$AgeFirstKill,mu=27,sigma.x=8.602,alternative='two.sided',conf.level=.95)
```

```
z.test(anger$AgeFirstKill,mu=27,sigma.x=8.602,conf.level=.95)
z.test(enjoy$AgeFirstKill,mu=27,sigma.x=8.602,conf.level=.95)
t.test(convenience$AgeFirstKill,alternative="two.sided",mu=27,conf.level=0.95)
```