SERIAL KILLER STATISTICAL ANALYSIS

INTRODUCTION

In this report "KillersandMotive" Dataset will be analyzed. We'll investigate Killer's motivation for becoming a serial killer in this section. A serial killer is someone who conducts a series of killings for no apparent reason. The Radford/FGCU Serial Killer Database is the source of this information. Convenience (didn't want children/spouse), Mental illness (including paranoia, visionary or Munchausen's disease), and Escape or escape arrest are the various causes in our dataset.

DATA CLEANING

Data cleaning needs to be done as we need to be our results as accurate as possible for correct and better understanding. I have removed 9(13.23%) rows from our dataset containing garbage value “99999” in column AgeFirstKill and 6(8.82%) rows containing “NA” values in Motive Column. Even I have tried to remove rows containing the data of killers who first killed before the year 1900 but there was none in my report. After that I have added the column CareerDuration giving the information about the duration of serial Killers. Therefore, now our dataset has 53 rows and 10 columns.

RESULTS

DATA EXPLORATAION

Here we will be seeing the relation between our attributes and will be analyzing what it tells us. We will be focusing mainly on 3 most important attributes affecting the motive of serial killer AgeFirstKill, AgeLastKill and CareerDuration.

Attributes	min	25%(IQR)	mean	75%(IQR)	max	Standard Deviation
AgeFirstKill	18	26	31.66038	35	54	8.708845
AgeLastKill	20	28	35.88679	42	64	10.65122
CareerDuration	0	0	4.226415	3	31	6.957678

Table 1. Data Description

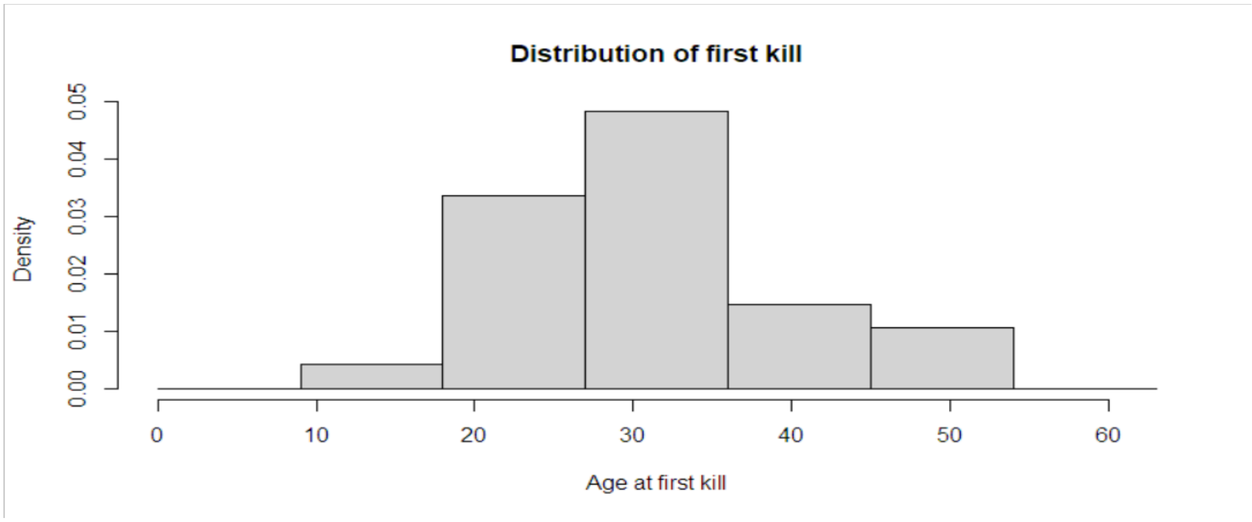

Fig. 1 Histogram of Distribution of First Kill

The above histogram shows that majority of serial killers, started their career between 18 to 36 years of age which is almost 77% of our dataset. Very less no. of serial killers had their first kill after 36 years of age and before 18 years of age.

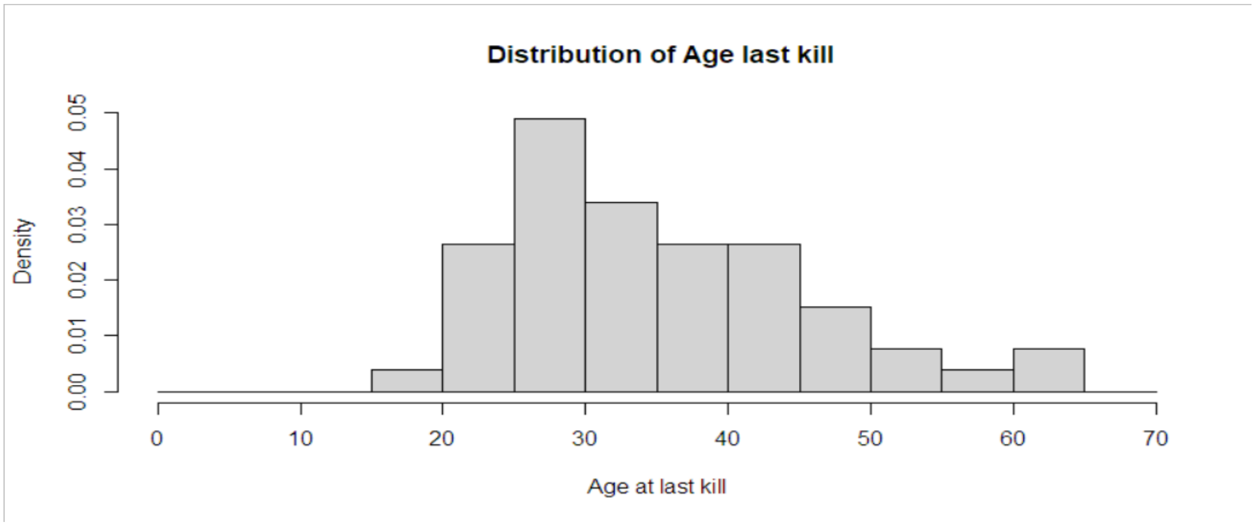

Fig. 2 Histogram of Distribution of Last Kill

The above histogram shows that more than half of the Killers ended their career between 20 to 45 years of age which 57% of total killers. Less no. of killers had their last kill when they were above 45 years of age and the graph at right side is not dense and unequally spread.

The below histogram Fig. 3 shows the career duration of killers. Here, we can see that about 77% of serial killers had career duration of 5 years or lesser than that. The below graph is positively skewed. And on the other end the graph is declining exponentially suggesting that very less killers have career duration of more than 5 years.

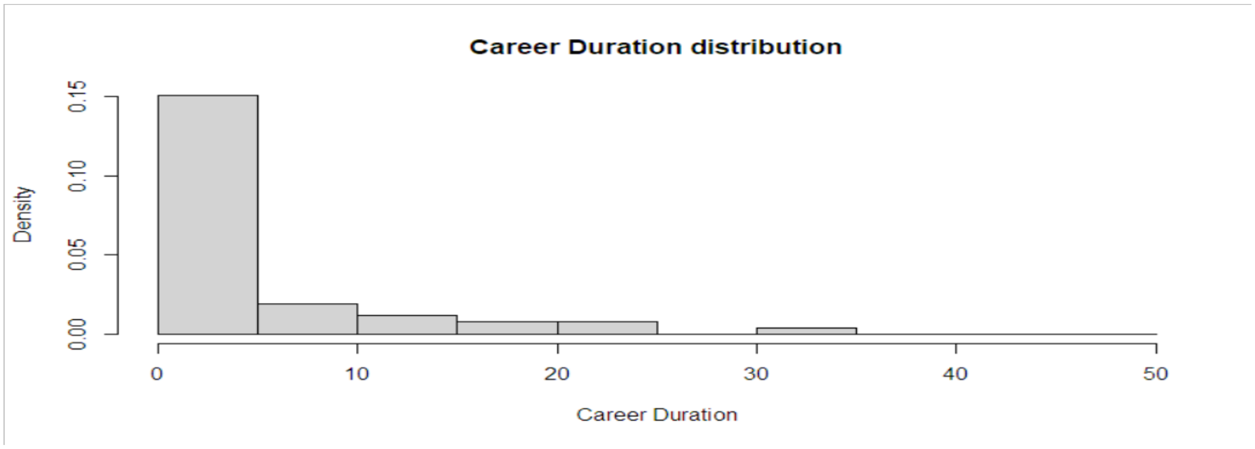

Fig. 3 Histogram of Career Duration distribution

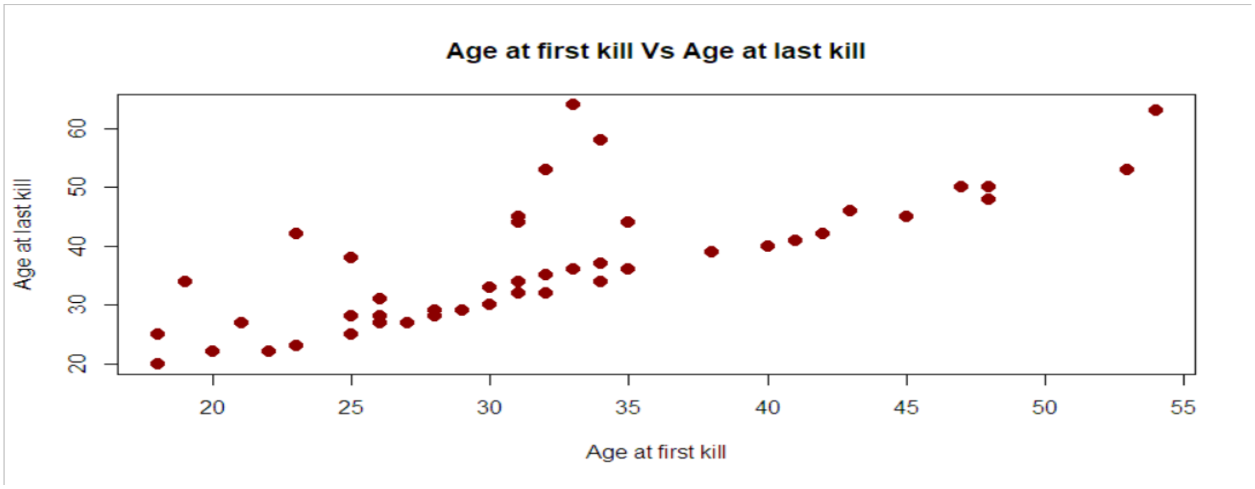

Fig. 4 Correlation between Age at first kill and Age at Last Kill

We can see that from fig that correlation between Age at first kill vs Age at last Kill is highly correlated with correlation value of 0.76 where its stating that most of the Killers has very small Career Duration as I have mentioned above Killers having Career Duration <=3 comprises of 75% of our Killers population. This same thing is evident in below Fig Age at Last Kill vs Career Duration which is moderately corelated with correlation value of 0.58.

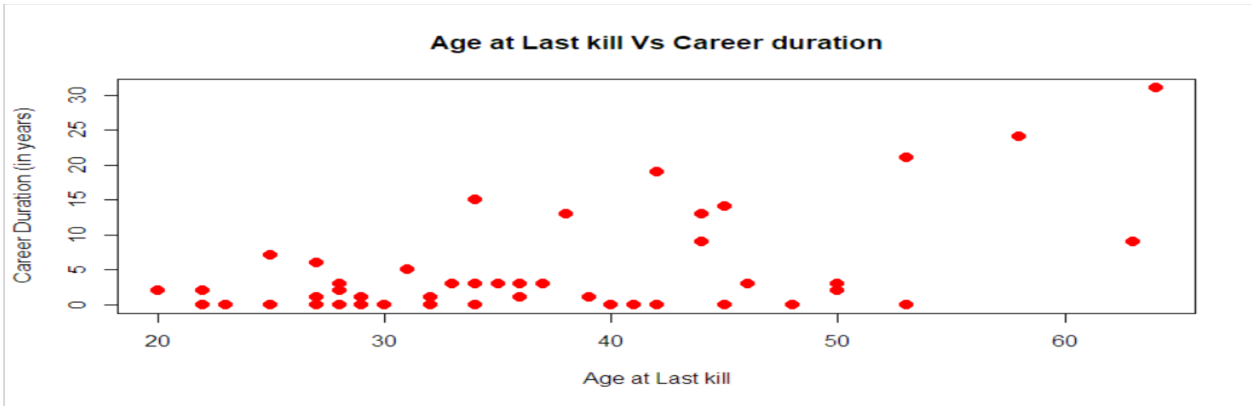

Fig. 5 Correlation between Age at Last Kill and Career Duration

Modelling

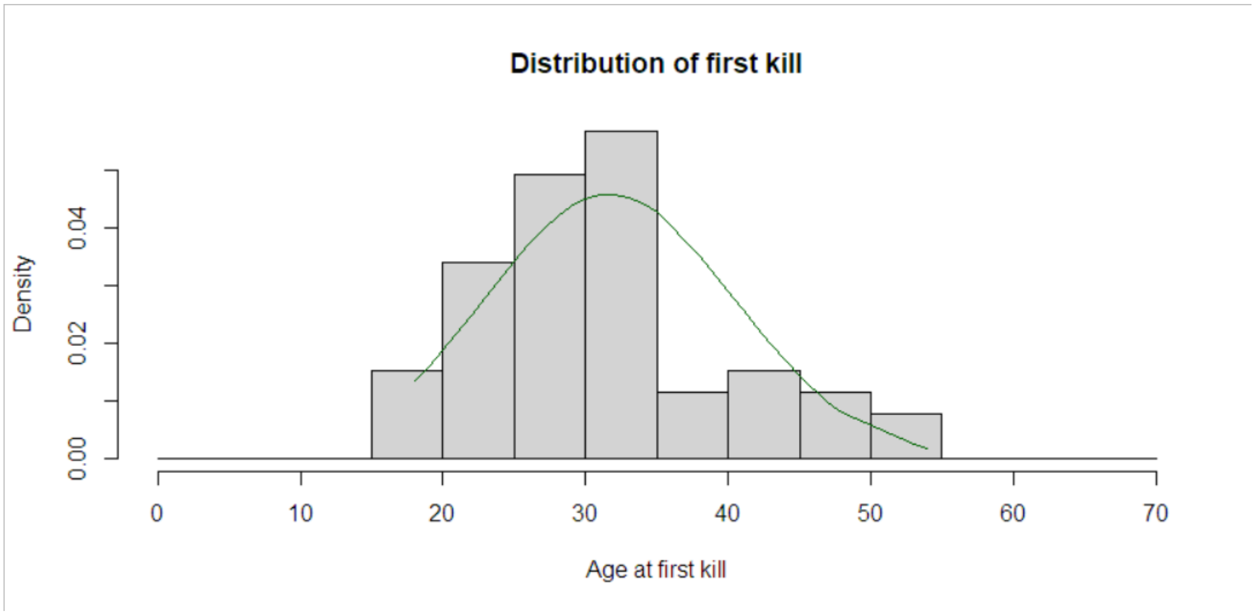

Fig.6 Fitting of Normal Density Curve over Distribution of First Kill

From the Fig. 6 histogram we can say that normal distribution will be a good fit for Age at first kill distribution as it shows a roughly bell-shaped distribution. The normal distribution would be an excellent fit for this model, as shown by the green curve, which is a normally distributed density function shown as a curve.

Same thing we can say that for Fig. 7 a roughly bell-shaped distribution can be a good fit for the Distribution for Age at Last Kill so that we can take normal distribution density function (green curve) for Age at First Kill going further for our research.

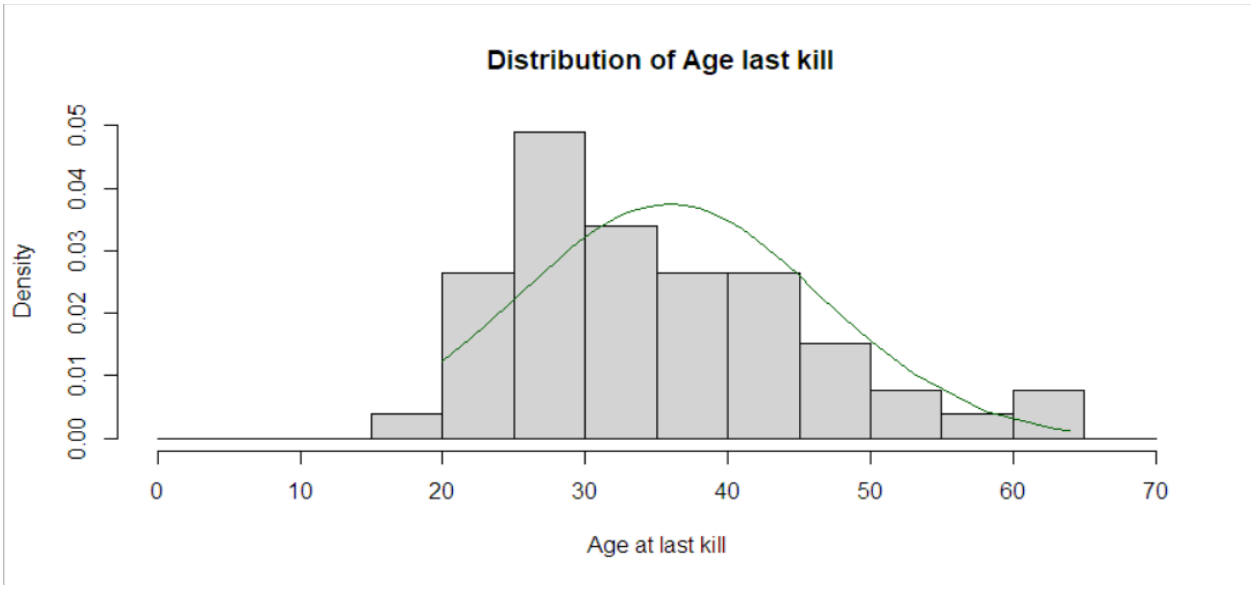

Fig. 7 Fitting of Normal Density Curve over Distribution of Last Kill

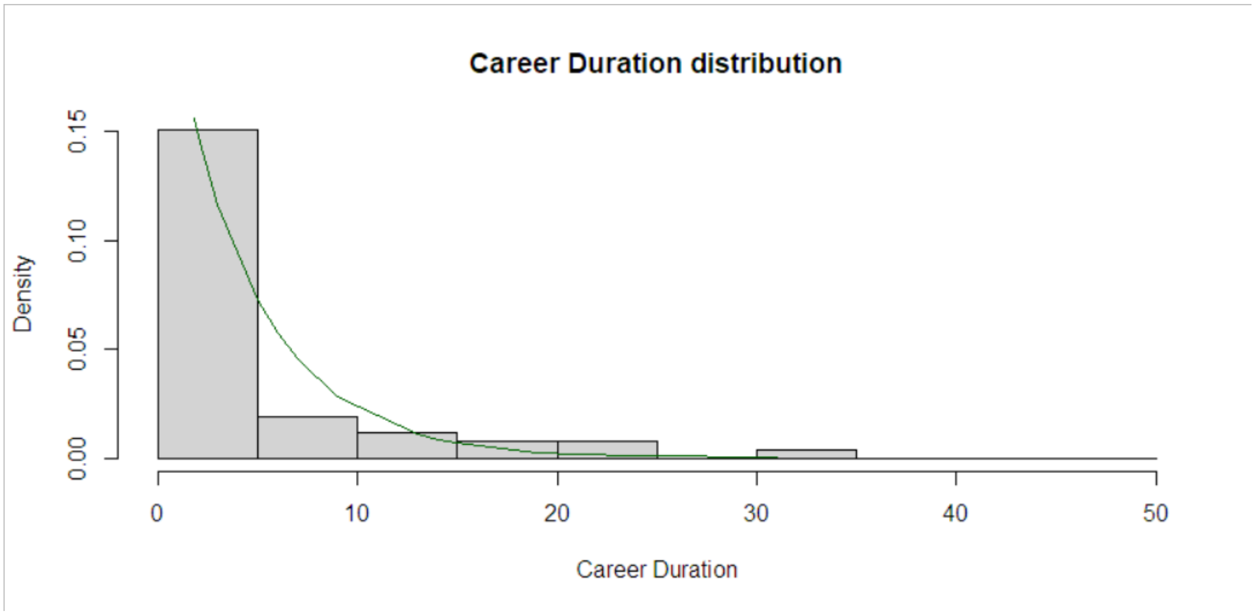

Fig. 8 Fitting of Exponential Density Curve over Distribution of Last Kill

We can observe that the data for career durations is positively skewed, with a lengthy tail going to a right extreme. We'd need a distribution that reflects such skew to model this. Because it is frequently used to model time-to-event variables, an exponential distribution would be acceptable going further for our research (green density curve).

Estimation

After selecting the appropriate distribution models, we must now estimate the parameters because we do not know their true values in real life.

Parameter	Estimation of E(x) Population Mean	Estimation of (σ^2) Population Variance	Reasoning for the estimation
Age at first kill	Sample mean (x_bar) ≈ Population mean (E(x)) ≈ 31.66	Sample Variance (s^2) ≈ Population variance (σ^2) ≈	As n=53 is moderately large by Method of Moments
Age at last kill	Sample mean (x_bar) ≈ Population mean (E(x)) ≈ 35.88	Sample Variance (s^2) ≈ Population variance (σ^2) ≈	As n=53 is moderately large by Method of Moments

Table 2. Parameter estimation

For Career Duration

Method of Moments

We know that the population mean for an exponentially distributed variable X ~ Exp(λ) is 1/ λ. The sample mean may be a reliable predictor of the population mean. As a result, 1/ λ sample mean (4.22 years).

Parameter estimation of rate parameter (λ) for the Career duration which has exponential distribution (Fig (g)) by **Method of Moments** is: x_bar ≈ 1/λ so λ ≈ 1/x_bar(sample mean) ≅ 0.237.

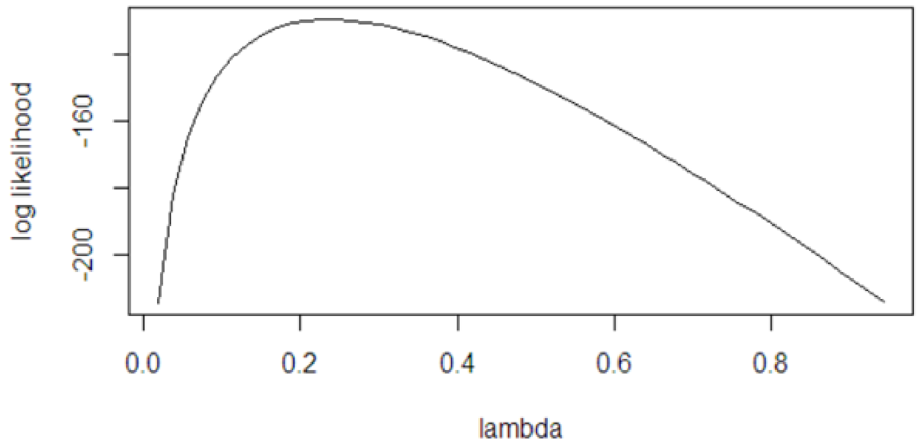

Fig. 9 Log likelihood function as a curve for λ which confirms λ ≈ 0.23

Maximum Likelihood Estimate (MLE)

We would estimate the parameter using MLE (maximum likelihood) in R and plot the log likelihood function as a curve for to validate it is a good guess, as shown below by MLE: λ ≈ 0.23 from fig 8.

	Minimum	Mean	Maximum	Standard Deviation
Motive 1 (n=22)	23	33.54	53	7.46
Motive 2 (n=10)	23	36.5	54	10.82
Motive 3 (n=21)	18	27.38	48	7.14

Table 3. Motive Description

Hypothesis Testing

It is necessary to confirm the normality of M1, M2, and M3 before doing hypothesis testing:

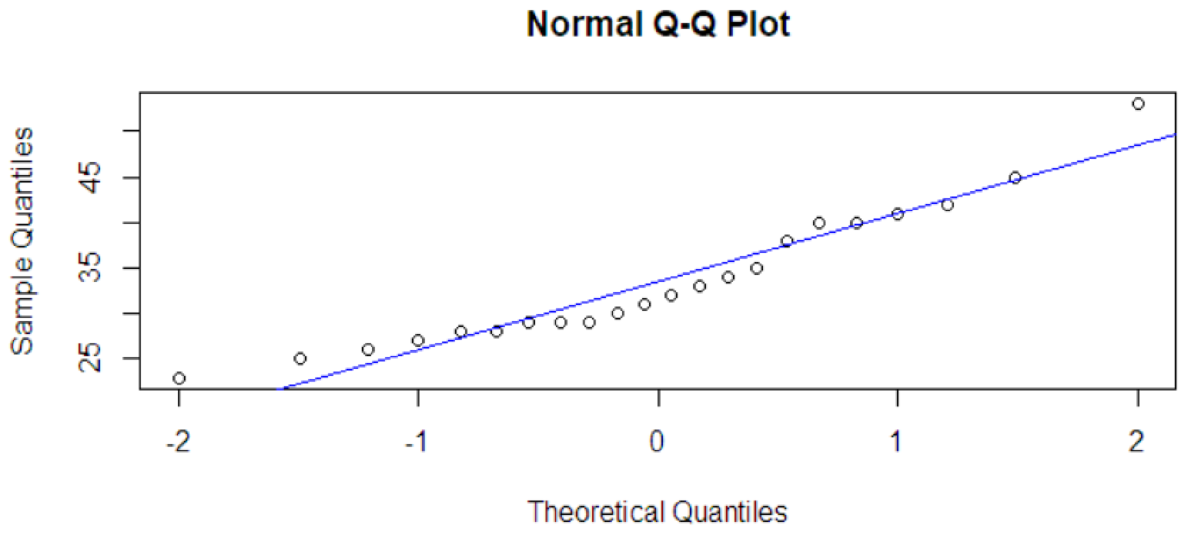

Fig. 10 Normal Q-Q plot for Escape or avoid arrest(M1)

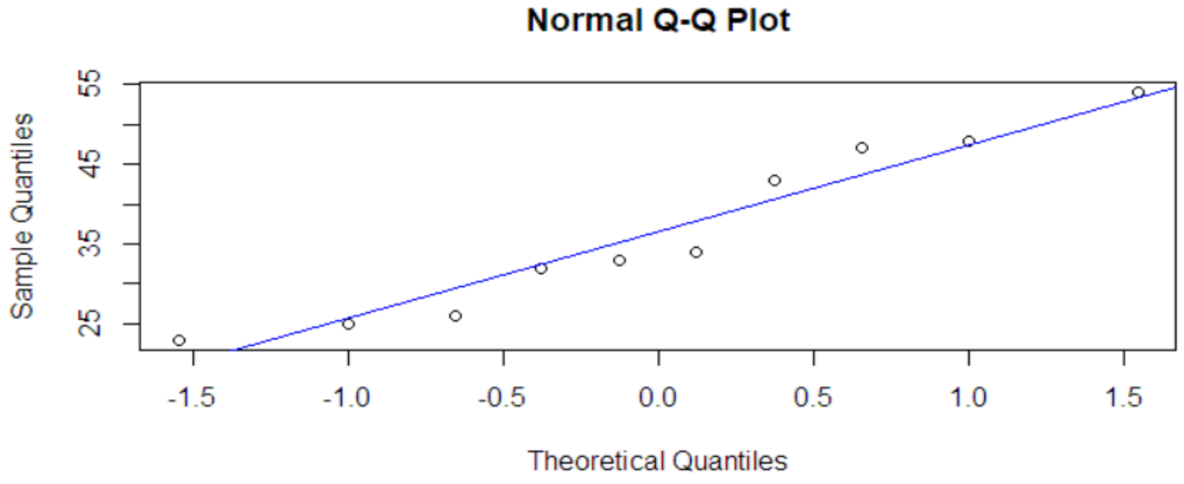

Fig. 11 Normal Q-Q plot for Convenience (didn't want children/spouse) (M2)

We can now perform a hypothesis test M1, M2, M3 because from above fig. 9 and 10 and & below fig 11 we can tell that M1, M2, M3 ≈ N(μ,σ^2) as the data points lie roughly near the blue straight line with a minor deviation.

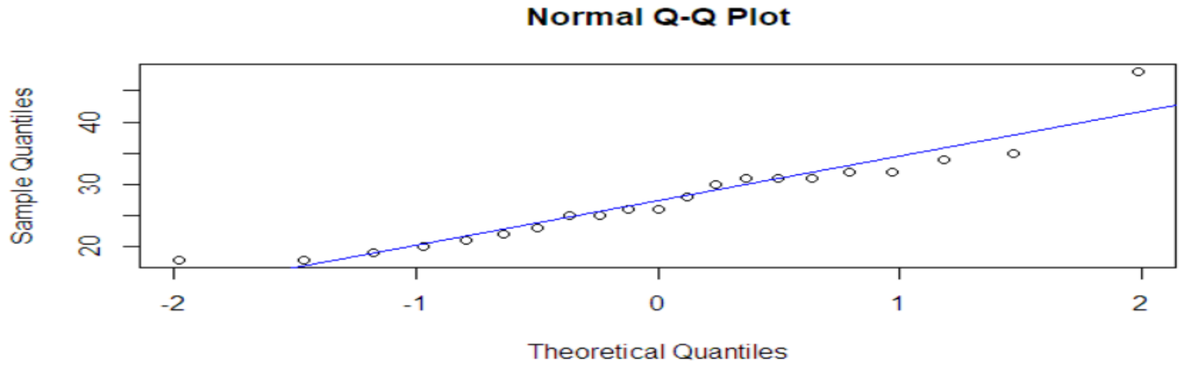

Fig. 12 Normal Q-Q plot for Mental illness (including paranoia, visionary or Munchausen's syndrome) (M3)

	Type of test performed	Test Result	95% Confidence Interval	p- value	Z /T test statistic value
M1 (n=22)	**z-test**	**Failed**	**29.90 to 37.18**	**0.00042**	**3.52**
M2 (n=10)	**t-test**	**Failed**	**28.75 to 44.24**	**0.02156**	**2.77**

M3 (n=21)	**z-test**	**Failed**	**23.65 to 31.10**	**0.8411**	**0.20**

Table 4: Hypothesis testing results for m1, m2, and m3 when the Null Hypothesis is H_0 : μ_0= 27 is average Age of first kill for each of three motives m1, m2 & m3. Alternative Hypothesis is H_1 : $\mu_0 \neq$ 27.

Though my n=22 is small for motive M1, still I have applied z-test as we can use C.L.T to apply the z-test on the normality assumption from Fig. 10, as all we require is the Z-test statistical value to be normal. We may deduce that the average age of first kill for killers with motive m1 is not equal to 27 because H_0 fails.

I have applied z-test for M3 motive for same reasons as M1 motive, as here also we can see normal distribution on M3 from Fig. 12 and with the help of C.L.T. I have chosen t-test. Here H_0 fails, therefore we can say that average age of first kill for serial killers with motive Escape or avoid arrest cannot be 27.

For M2 motive, I have applied t-test. As the population variance is unknown here and due to n=10 being so small we cannot assume population variance ≈ sample variance. Because of normality distribution which we can infer from Fig. 11, I think t-test is the best possible option to go ahead. Here also H_0 fails as it concludes that average age of first kill for serial killers with motive Mental illness is not 27.

Comparison of Killers with Different Motives

By comparing the killers with different motives will answer the main question of our research as we are not able to conclude anything evidently on basis of hypothesis t-test or z-test. As killers with varied intentions are two separate entities with no ties to one another. Figures 10, 11 and 12 for all M1, M2, and M3 can be used to make a normality assumption. And even the variance in our sample is different, a 'two sample t-test with independence' would be a appropriate choice for testing hypotheses for comparing various populations based on these considerations.

Pair of motives	Estimated mean difference	Test Result	95% Confidence Interval	p- value
(M1, M2)	2.96	*Pass*	-11.10 to 5.19	0.4478
(M2, M3)	9.11	*Failed*	0.98 to 17.25	0.0308
(M1, M3)	6.16	*Failed*	1.66 to 10.66	0.0084

Table 5: On each pair of motives, the results of a two-sample t-test with independence hypothesis testing where Null Hypothesis is H_0 : μ_1- μ_2= δ = 0. Alternative Hypothesis being H_1 : $\delta \neq 0$ at 95 % significance level.

For the pair (M1, M2)of motives of H_0 is true, so this suggests that there is no difference between the age of first kill for the killers with motives Escape or avoid arrest and Convenience (didn't want children/spouse) as $\mu_1 - \mu_2 = \delta = 0$ lies in our 95 % CI, p-value is greater than 0.05 which justifies that H_0 is true.

For the pair (M2, M3) and (M1, M3) of motives H_0 is False which suggest that there can be age difference between first kill for killers with motives Convenience (didn't want children/spouse) and Mental illness (including paranoia, visionary or Munchausen's syndrome) and same applies for pair (M1, M3).

Discussion

The main question of our research is **does the average age at first murder differ between killers with different motives?**

The biggest problem which I think is that I have very a smaller number of observations from which It becomes difficult to strongly suggest any finding basis on the data provided. There is so much variance in my dataset because 75% of killers have career duration <=3 years and have many outliers of career duration where values vary from 0 to 31.

As pair (m1, m2) testing is true from this we can say that though age at first kill is same doesn’t mean that their motives will be same. And on other hand for pair (m2, m3) and (m1, m3) test fails. Here I think that may be my assumptions would have been incorrect and the motives will not be having normal distribution as it can happen due to small number of sample data.

Acknowledgements

I'd want to express my gratitude to Professor Benjamin Thorpe, whose direction and lecture notes assisted me in completing my research and who provided me with the opportunity to do so. At the same time, I'd like to express my gratitude to my entire team for their assistance with this study.