Analysis of Serial Killers' Age at First Murder

In this work, we analyse serial killer data and focus on the age at which serial killers committed their first murder. The data being used for this analysis is based on a sample of serial killers from the Radford/FGCU Serial Killer Database. The sample data set has nine features: *KillerID, AgeFirstKill, AgeLastKill, YearBorn, Motive, Sex, Race, Sentence* and *InsanityPlea.*

For our analysis, the features we're interested in are *AgeFirstKill, AgeLastKill,* and *Motive. AgeFirstKill* tells us the age (in years) at which each killer committed their first murder. We only consider the data where the killers' first kill was from 1900. *AgeLastKill* contains the age (in years) of the killer when they committed their last murder. *Motive* gives us the reason each killer committed the murders if known. There are three motives in the sample used in this analysis – *Angel of Death, Robbery or Financial Gain,* and *Unknown,* where the motive is not known.

Previous research suggests that the population mean for serial killers who were active before the 1900s is 27. The main aim of this project is to determine whether the average age at first murder differs between killers with different motives. We will also explore other questions like: do killers with different motives start killing at a different age compared to the general population?

1. Data Cleaning:

The data contains some missing values. The rows containing these missing values need to be removed. We also remove the rows containing the data of killers whose first kill was before 1900. Table 2.1 tells us how many observations have been removed and the reasons for removal.

Variable	*Total no. of observations*	*No. of observations (rows) removed*	*Percentage of observations removed*	*Reason for removal*
Age At First Kill	634	9	1.4%	Missing values
Age At First Kill	634	6	0.94%	Killers born before 1900
Motive	634	9	1.4%	Missing values

Table 2.1: Number and percentage of observations removed for Age At First Kill and Motive, and the reasons for removal

We also create a new variable, *Career Duration,* defined as the number of years between the first kill and the last kill of each killer. After cleaning the data set, we are left with 10 columns and 610 rows.

2. Data Exploration:

The following figures (fig 2.2a, 2.2b, and 2.2c) and table2.2 depict the graphical and numerical summary respectively of the distributions of the variables: age at first kill, age at last kill, and career duration.

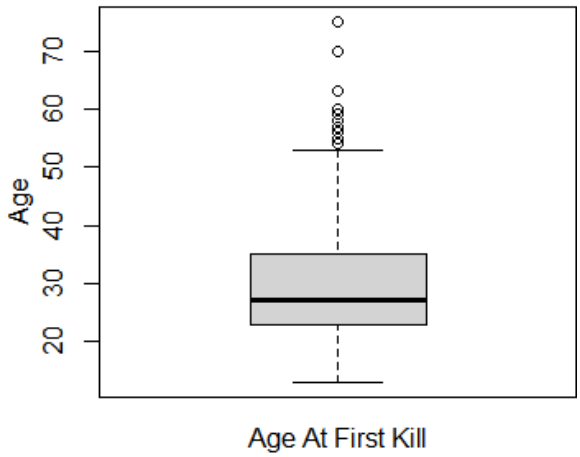

Fig 2.2a: Boxplot of Age At First Kill

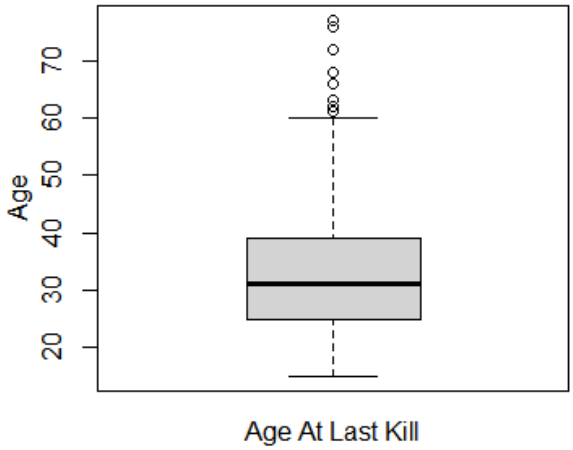

Fig 2.2b: Boxplot of Age At Last Kill

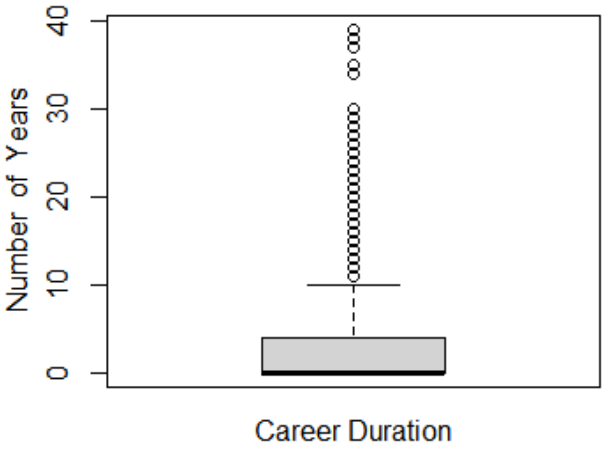

Fig 2.2c: Boxplot of Career Duration

	Age At First Kill (n=610)	*Age At Last Kill (n=610)*	*Career Duration (n=610)*
Mean	29.63	33.13	3.50
Standard deviation	9.08	10.93	6.61
Skew	1.24	1.06	2.60

Table 2.2a: Numerical summary of the distribution of the variables: Age At First Kill, Age At Last Kill, and Career Duration

We visualise the relationship between the career duration & age at first kill and the career duration & the age at last kill using a scatter plot to examine the relationship between the three variables and calculate the correlation coefficient.

The career duration is related to the age at first kill and age at last kill. We expected that serial killers who started killing at an earlier age would have a longer career duration compared to those who started killing at a much older age. Serial killers whose last kill was at an earlier age would have a shorter career duration as opposed to those who stopped killing at an older age.

However, from the scatter plot (fig 2.2d and 2.2e) and correlation coefficient, we can see that this is true for only some cases. Age at first kill has a weak negative linear relationship with career duration, while age at last kill has a positive linear relationship with career duration.

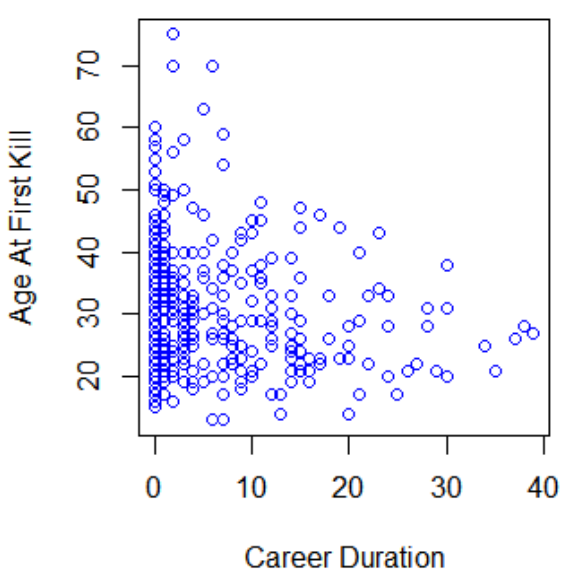

Fig 2.2d: Scatter plot of Career Durations vs Age At First Kill

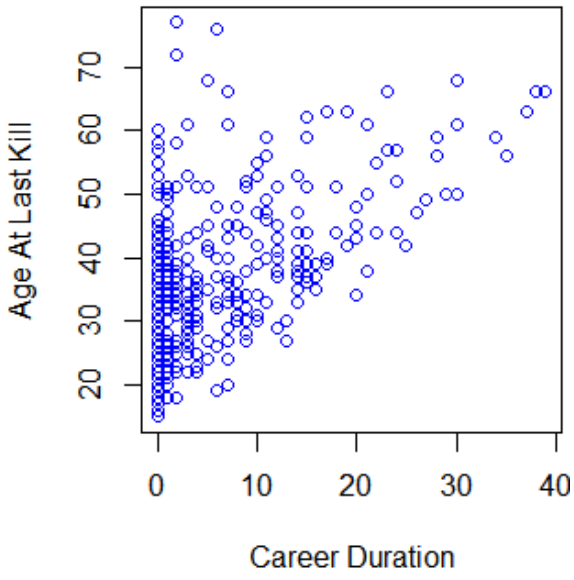

Fig 2.2e: Scatter plot of Career Durations vs Age At First Kill

	Career duration and age at first kill	*Career duration and age at last kill*
Correlation coefficient	-0.05	0.56

Table 2.2b Correlation coefficients for Career duration and age at first kill and Career duration and age at last kill

3. Modelling:

To propose an appropriate distribution for three variables – age at first kill, age at last kill, and career duration, we plot histograms for each of the variables.

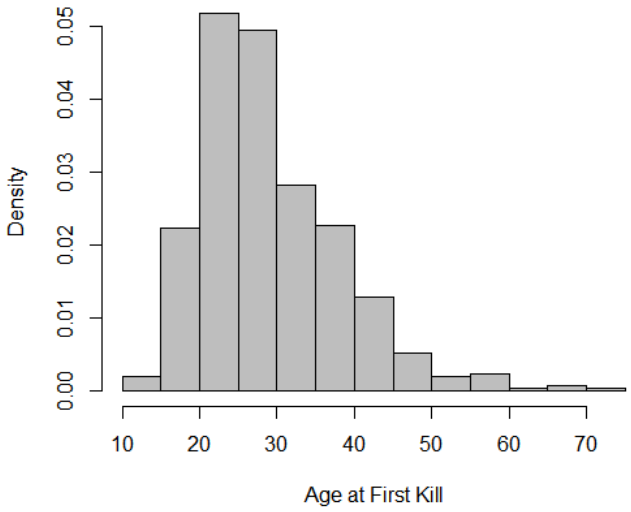

Fig 2.3a: Density and Histogram Plot for Age At First Kill

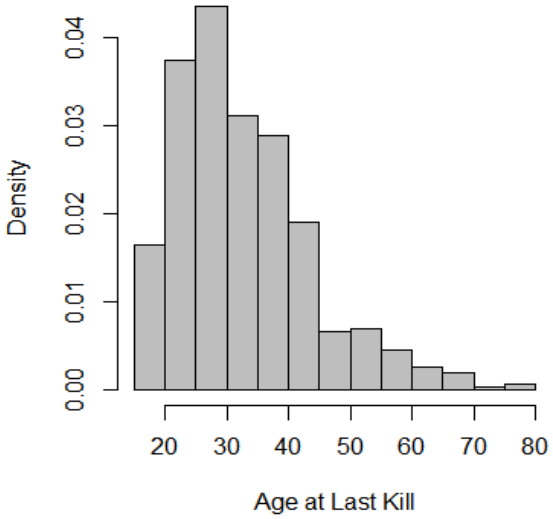

Fig 2.3b: Density and Histogram Plot for Age At Last Kill

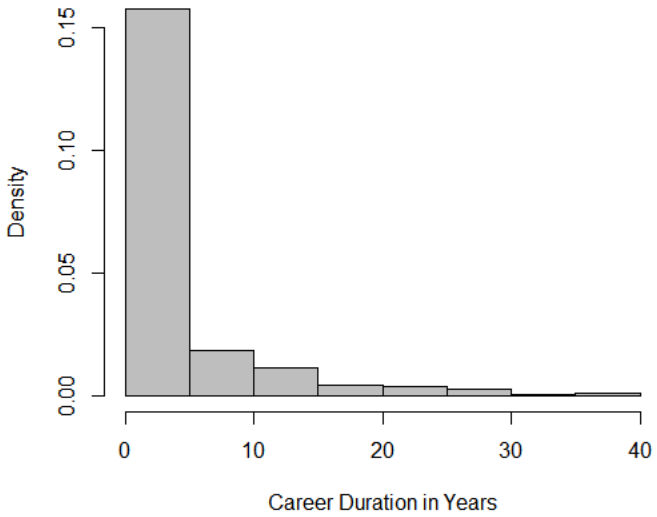

Fig 2.3c: Density and Histogram Plot for Career Duration

Age at first kill and age at last kill are continuous variables. Based on the histograms plotted (fig 2.3a and 2.3b), we can see that both the histograms are roughly bell-shaped and unimodal. We, therefore, propose a normal distribution to model age at first kill and age at last kill. Career duration is also a continuous variable. From the histogram (fig 2.3c) we can see that it is unimodal and extremely skewed to the right and thus use an exponential distribution to model the career duration.

4. Estimation:

The estimators for each distribution have been calculated using method of moments, and the density curves have been plotted based on the estimators.

Distribution	*Parameter*	*Estimator*	*Age At First Kill (n=610)*	*Age At Last Kill (n=610)*
Normal distribution	Mean	$\hat{\mu}_{MOM}$ = sample mean	29.63	33.13
	Variance	$\hat{\sigma}^2_{MOM}$ = sample variance	82.41	119.54

Table 2.4a: Estimated mean and variance for Age At First Kill and Age At Last Kill

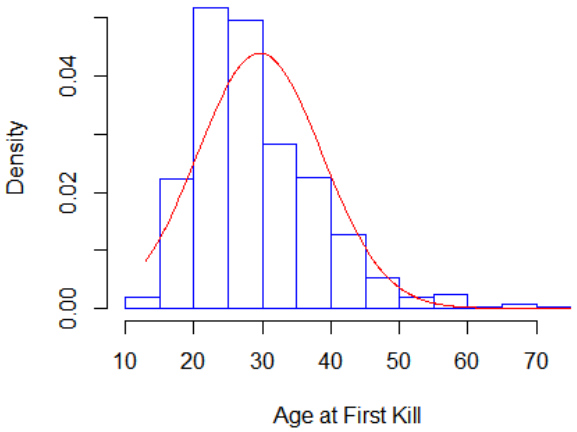

Fig 2.4a: Histogram and Normal Distribution Density Curve for Age At First Kill

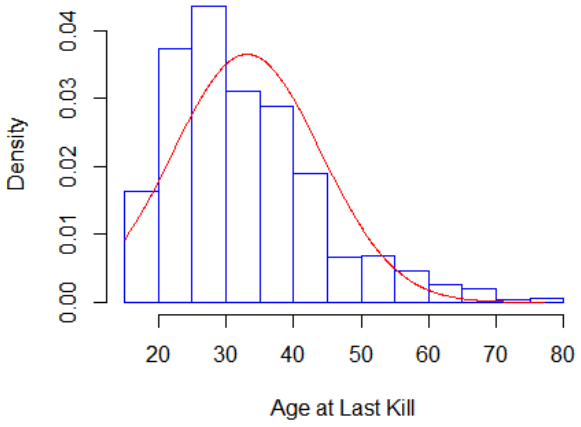

Fig 2.4b: Histogram and Normal Distribution Density Curve for Age At Last Kill

Method of moments has been used to estimate the parameters for the distributions in this work because it is easier to use. However, it is not based on any particular criterion which would give us the 'best' estimate like the maximum likelihood estimator might. The estimates for the mean of age at first kill, and age at last kill (table 2.4a) seem reasonable enough. The estimated variances are quite large, which tells us that there is a lot of variability in the age at which serial killers start and stop killing.

Distribution	*Parameter*	*Estimator*	*Career Duration (n=610)*
Exponential distribution	Rate	λ_{MOM} = 1/ sample mean	0.29

Table 2.4b: Estimated rate parameter for Career Duration

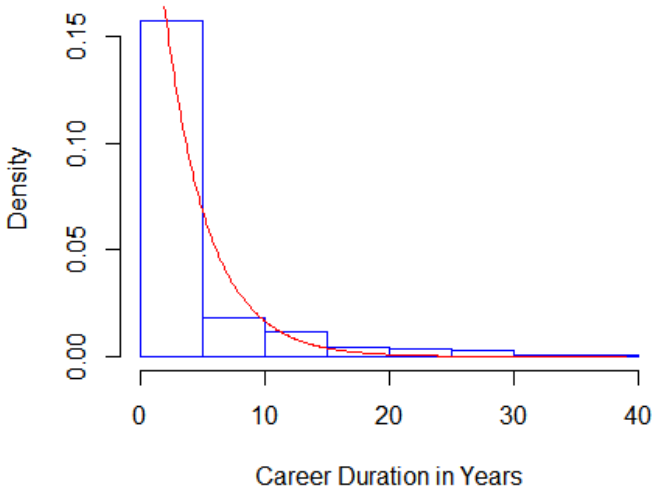

Fig 2.4c: Histogram and Exponential Distribution Density Curve for Career Duration

5. Testing Hypotheses:

We perform hypotheses tests for each of the three motives to determine if the mean age at first kill is 27 years. Our null hypothesis is that the mean age at first kill for all three motives is 27 years.

Motive	*Angel of Death (n=23)*	*Robbery or Financial Gain (n=510)*	*Unknown (n=77)*
Min	21.00	13.00	14.00
1st Quartile	26.50	23.00	23.00
Median	30.00	27.00	26.00
Mean	32.35	29.46	29.97
Variance	75.78	80.33	97.87
3rd Quartile	36.50	35.00	36.00
Max	58.00	75.00	60.00

Table 2.5a: Numerical summary for the three motives: Angel of Death, Robbery or Financial Gain, and Unknown

QQ plots have been used to check for normality. However, we should note that QQ plots are subjective and not everyone may interpret them in the same way. To some, a QQ plot may look normal, but to others, it may not. For future work, we can do the Kolmogorov-Smirnov test along with the QQ plot to check the normality assumption.

The following QQ plots show whether or not the three motives are normally distributed.

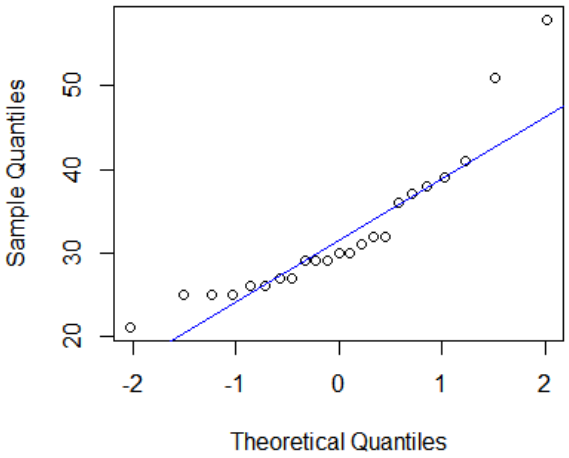

Fig 2.5a: QQ Plot to check for normality of the Angel of Death motive

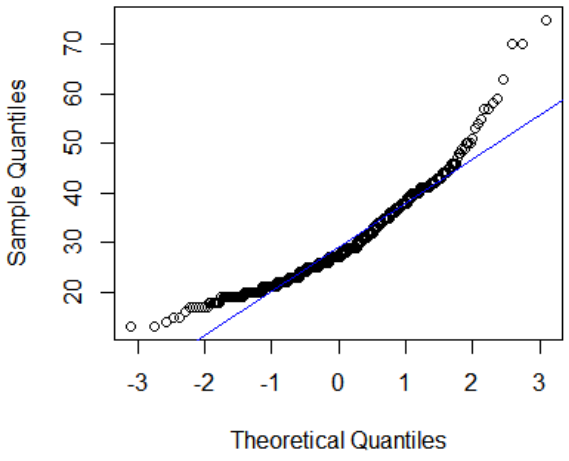

Fig 2.5b: QQ Plot to check for normality of the Robber or Financial Gain motive

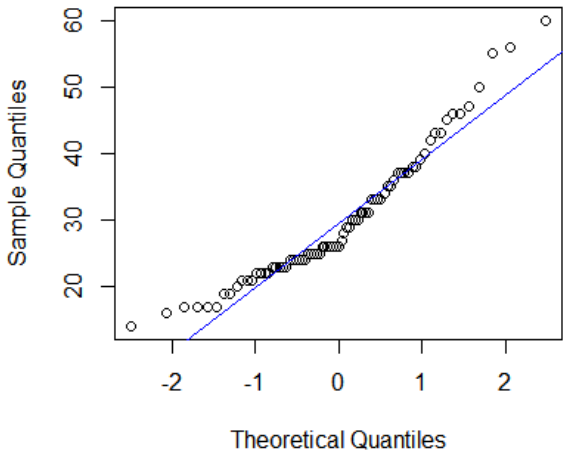

Fig 2.5c: QQ Plot to check for normality of the Unknown motive

From fig 2.5a, we can see that the QQ plot supports normality for the motive Angel of Death. This motive also has a small sample size. Therefore, we use a t-test. The QQ plot does not support the normality assumption for robbery or financial gain (fig 2.5b) and, the sample size is very large. We can, therefore, use a z-test since the z-test does not need the distribution to be normal. For the unknown motive, we use a t-test since the QQ plot supports the normality assumption (fig 2.5c). Z-test cannot be used here since the variance of the unknown motive (97.87) is very different from the population variance (75).

	Angel of Death (n=23)	*Robbery or financial gain (n=510)*	*Unknown (n=77)*
Type of test	t-test	z-test	t-test
Sample mean	32.35	29.46	29.97
Confidence Interval	(28.58, 36.11)	(28.68, 30.23)	(27.76, 32.18)
p-value	0.007	5.99e-10	0.008

Table 2.5b: Hypothesis test results for the three motives: Angel of Death, Robbery or Financial Gain, and Unknown

Based on the hypothesis tests performed, we reject the claim that the mean age at first kill is 27 years for all three motives as none of the confidence intervals contain 27.

It is important to note that the QQ plot does not support the normality assumption for the robbery or financial gain motive. The variance assumptions may also be questionable since they have been chosen based on how close the ratios of the variances are to 1.

6. Comparison of Different Populations

We use the two-sample t-test to determine if the mean age at first kill is different between each pair of motives. Our null hypothesis is that there is no difference in the mean of the age at first kill for the three pairs of motives.

Since the motives are different for different killers, we assume that all three motives are independent and normally distributed. For the pair of motives, Angel of Death and robbery or financial gain, we assume they have equal variances (the ratio of the variances is 0.94). We assume that Angel of Death and unknown have unequal variances (the ratio of the variances is 0.77). For robbery or financial gain and unknown, we assume that they have unequal variances (the ratio of the variances is 0.82).

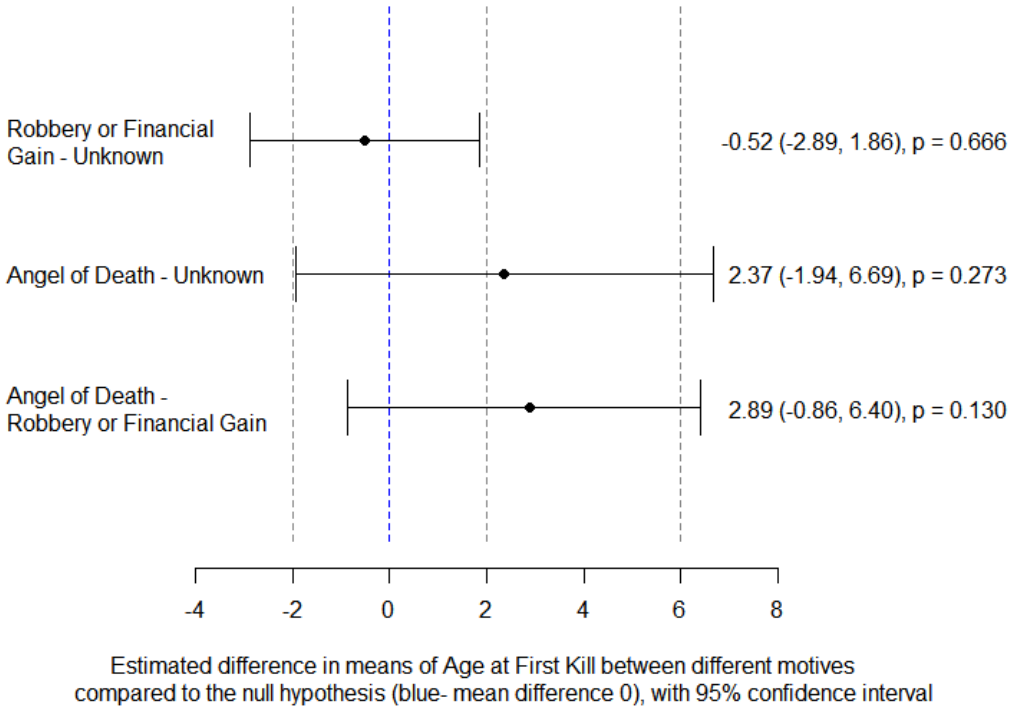

Fig 2.6: Forest plot of three 95% confidence intervals for each pair of motives

From the forest plot (fig 2.6) we see that the confidence intervals for all three pairs of motives contain 0, so we fail to reject the claim that there is no difference between the three pairs of motives.

However, the three confidence intervals suggest that killers whose motive is robbery or financial gain and killers whose motive is unknown start killing at an earlier age than those whose motive is Angel of Death. Our interpretation of the hypothesis test performed could lead to a type 2 error. We might fail to reject the claim that there is no difference in the average age at which killers with different means start killing, when in fact there is a difference.