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	Anupam Bose
Student ID	201570198


THE ART OF SERIAL KILLING

INTRODUCTION

The precise definition of serial murder has sparked heated debate among criminologists. In simple words, a serial killer is someone who murders three or more people, usually for abnormal psychological stimulation, over a period of more than a month and with a large interval of time between the murders.

The major goal of our study is to see if the age at which serial killers commit their first murder is related to their purpose for killing. We have a range of data on serial killers giving information like Year killer is Born, Sex, Race etc. But we'll mostly focus on the Killer's Age at first kill/Last kill, & Motive. We have three different motives: robbery or financial gain, convenience (didn't want children or a spouse), and unknown.

RESULTS

Data cleaning:

We had information on 620 killers. The motive was missing in 0.96 percent of the data, while the age at first kill was incorrect in 1.45 percent of the data (99999). We also had eight records of killers who began killing before 1900, which we eliminated because we are researching data on killers who began killing after 1900. Hence, we are left with **597** cleaned records.

Data exploration:

	Mean	*Standard Deviation*	*Maximum Value of the parameter*	*Minimum value of the parameter*	*50% Quantile*	*75% Quantile*	*Skewness*
Age at First kill	29.64	9.14	75	13	27	35	1.21
Age at Last Kill	33.16	11.04	77	15	31	39	1.05
Career Duration (in Years)	3.52	6.16	39	0	0	4	2.61

Figure 0: Numerical summary of Killer's data of Age at First Kill, Age at Last kill & Career Duration

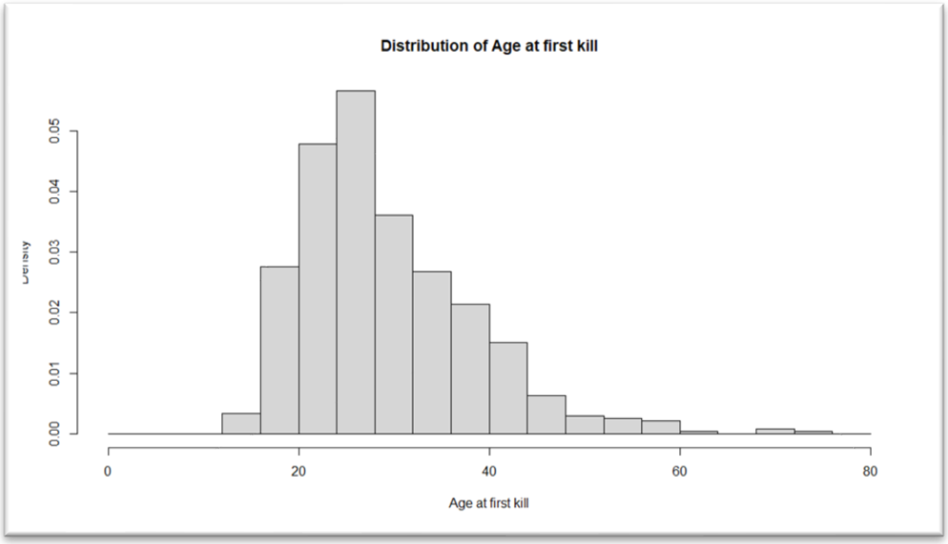

Figure 1: A histogram of our serial killer data, showing densities for age at first kill with intervals of 4 from the age 0 to 80.

The histogram shows that the majority of killers start their careers between the ages of 20 and 30, and relatively few killers starting late as the age increases. 50% of killers start their career by the age of 27.

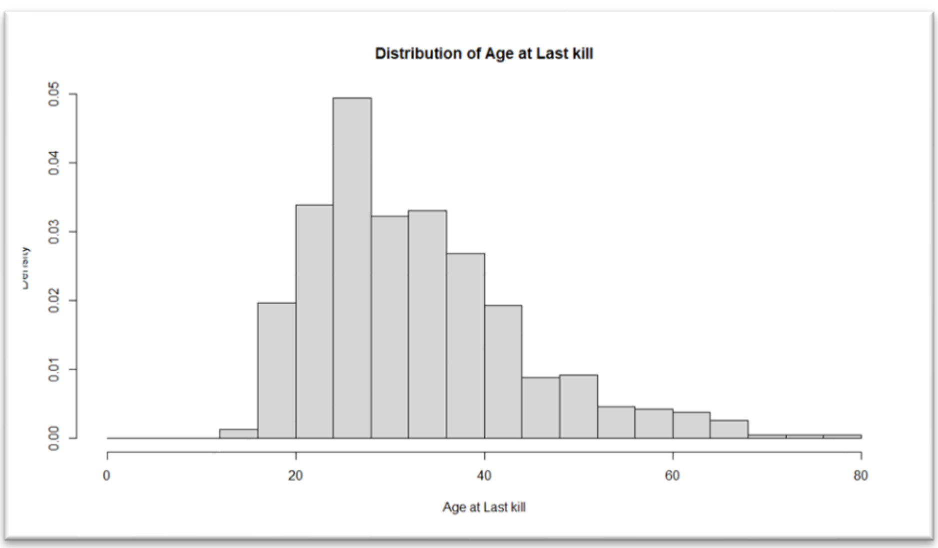

Figure 2: A histogram of our serial killer data, showing densities for age at first kill with intervals of 4 from the age 0 to 80.

The histogram tells us that the bulk of killers finish their careers between the ages of 20 and 35, and fewer killers finishing their professions as they get older. By the age of 31, half of all serial killers have ended their career.

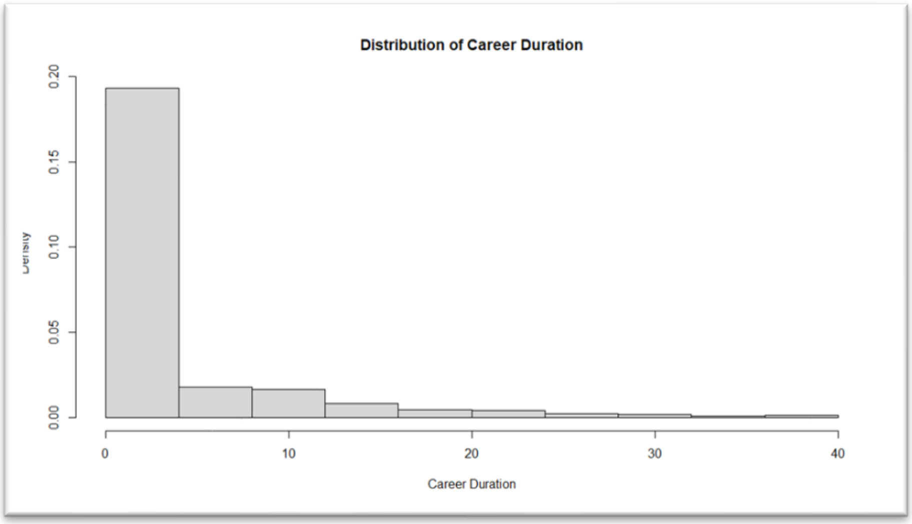

Figure 3: A histogram of our serial murderer data, displaying career duration density with intervals of four from 0 to 40 years.

Figure 3 shows that most killers' careers terminate abruptly, with most of them having a maximum career of 0 to 4 years. For 75% of serial killers, their careers are ended within four years.

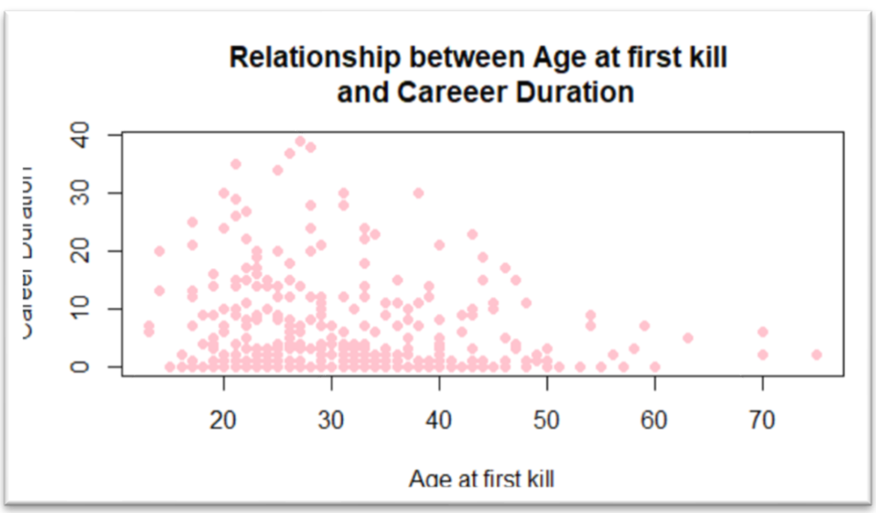

Figure 4: A graphical representation of our serial murderer data, showing how career length varies with age at first murder.

Figure 5: A graphical representation of our serial killer data, demonstrating how the age at first murder differs from the age at last murder.

Figures 4 and 5 both lead to the same conclusion. It shows that serial killers' careers vary linearly with their ages. The figure 4 is positively skewed which tells us that the serial killers who begin their careers when they are young (20-35 years old) have a better probability of having a longer-lasting career. The age at first kill and the age at last kill have a good correlation ($r_{xy} \approx 0.80$) and a positive linear relationship. This is quite expected as killers have more enthusiasm psychologically when they start young with an adrenaline rush. Starting career at old age gives them less opportunity for longer lasting career if they're convicted.

Modelling:

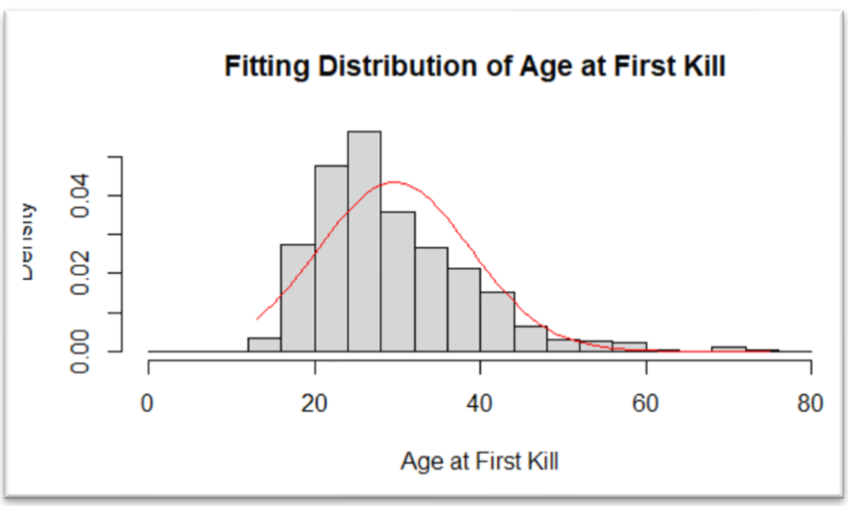

Figure 6: A graphical representation of our serial killer data, demonstrating a roughly bell-shaped distribution which looks similar to normal distribution density function (red) with positive skewness.

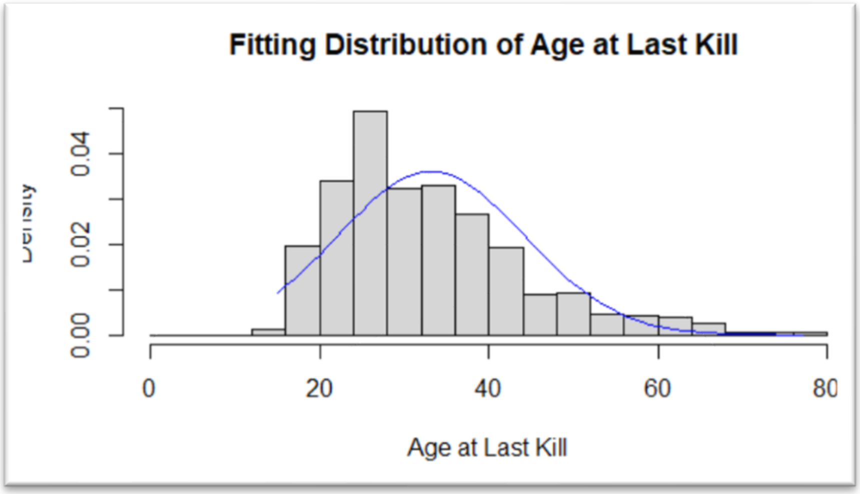

Figure 7: A graphical representation of our serial murderer data, demonstrating a roughly bell-shaped distribution showing how the age at last kill distribution resembles a normal distribution density function(blue) with a positive skewness

Age at First and Last Kill: The age of a person is a continuous variable with values ranging from 0 to infinity. Both distributions are positively skewed & have an approximately bell-shaped shape that resembles a normal distribution, as seen in figures 6 and 7. They also have a comparable spread and position. As a result, using the normal distribution for further investigation makes sense.

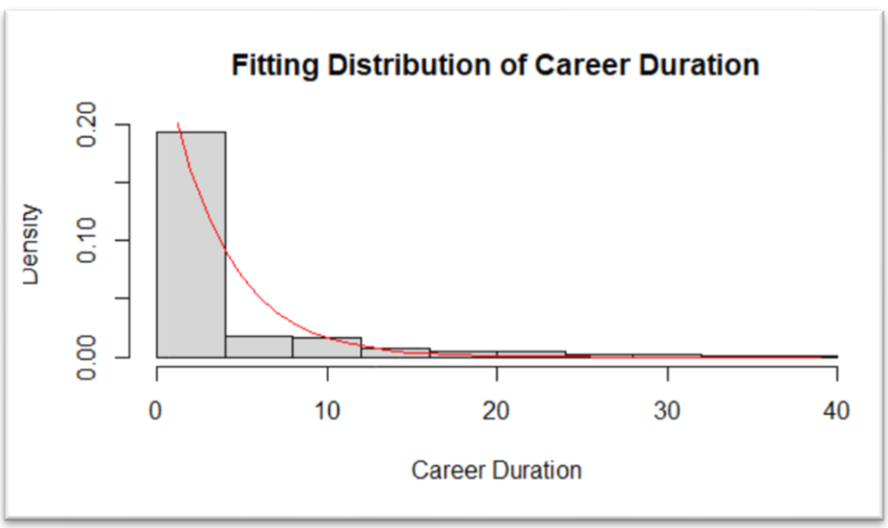

Figure 8: A histogram of career durations shows a non-negative sample with a noticeable positive skew. With an identical form to our histogram, an exponential distribution density curve (red) (non-negative with a positive skew).

Figure 8 shows how the exponential distribution density curve with a long tail to the right and heavy positive skewness fits the distribution of career duration well. Because of this asymmetrical distribution, the exponential distribution is the best option for further research

Estimation of parameters:

Now that we have selected a model for each of the three variables (age at first kill, age at last kill, and career duration), now we must estimate the parameters for each distribution.

Parameter Estimation	Estimating population mean (μ)	Estimating population variance (σ^2)	Justification
Age at first kill	Sample mean ≈ population mean ≈ 29.64	Sample Variance ≈ population variance ≈83.68	Sample is quite large (n=597)
Age at last kill	Sample mean ≈ population mean ≈ 33.16	Sample Variance ≈ population variance ≈122.05	Sample is quite large (n=597)

Table 2: Parameter estimation for variables (age at first kill & last kill) for normal distribution density function

Career Duration:

By method of moments:

For an exponentially distributed variable X ~ Exp(λ) we know that the population mean is 1/λ. Sample mean could be a good estimate of population mean. Hence, 1/λ ≈ sample mean (3.52 years).

λ ≈ 0.283

But we cannot de dependent on this value as this gives a rough idea of the parameter. We must choose the estimator that maximizes the likelihood of the observed data.

By Maximum Likelihood:

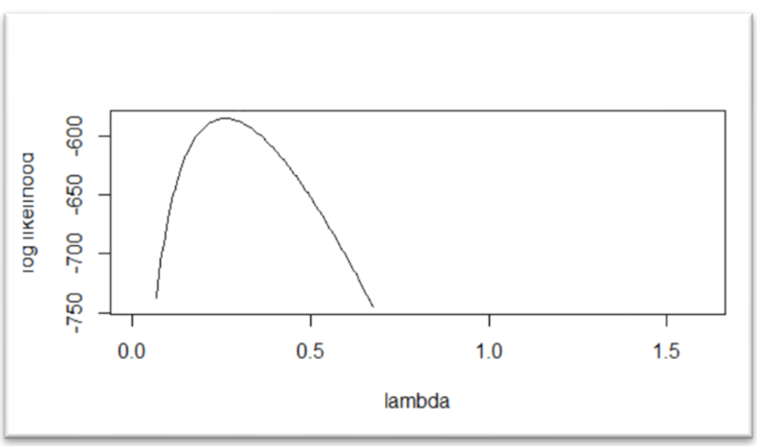

Figure 9: The likelihood of lamda's value changing with different values is shown by a log likelihood curve

Figure 9 shows that the peak of the curve (likelihood) is maximum around $\lambda = 0.3$, which is substantially identical to what we calculated using the method of moment method. The slope of the log likelihood curve is zero at $\lambda \approx 1/(\text{sample mean})$. However, because we have limited sample of data, we cannot guarantee a perfect evaluation of lamda's value as maximum likelihood performs better for larger samples.

Testing Hypothesis:

Motives	Mean (Age a first kill)	Standard deviation	skewness	Max Age	Min Age	Total killers
Unknown (Motive 1)	**29.97**	**9.89**	**0.93**	**60**	**14**	**77**
Robbery or financial gain (Motive 2)	**29.45**	**8.96**	**1.29**	**75**	**13**	**510**
Convenience (didn't want children/spouse) (Motive 3)	**36.5**	**10.82**	**0.27**	**54**	**23**	**10**

Table 2: The table numerically summarizes the data of the age at first kill of killers with 3 different motives

Checking Normal Distribution for each motive

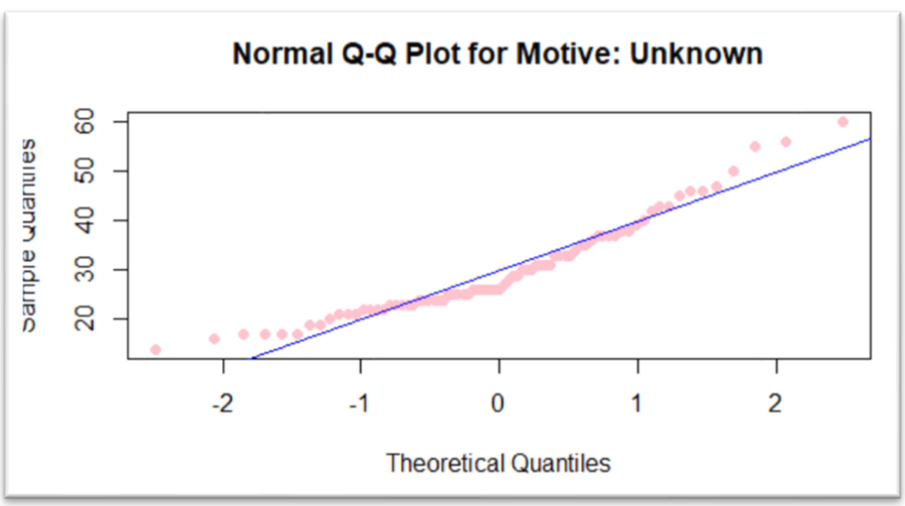

Figure 10: Normal Q-Q Plot for motive 1

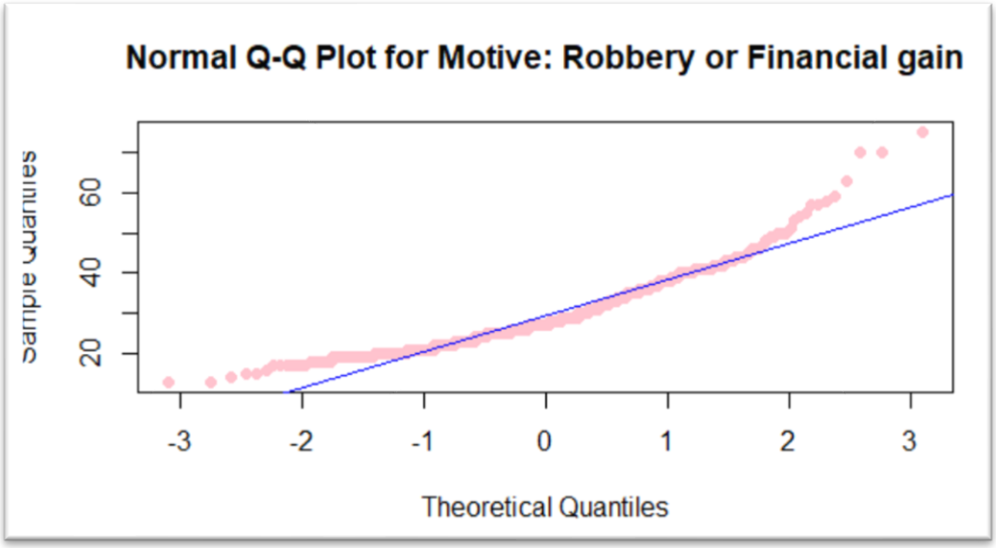

Figure 11: Normal Q-Q Plot for motive 2

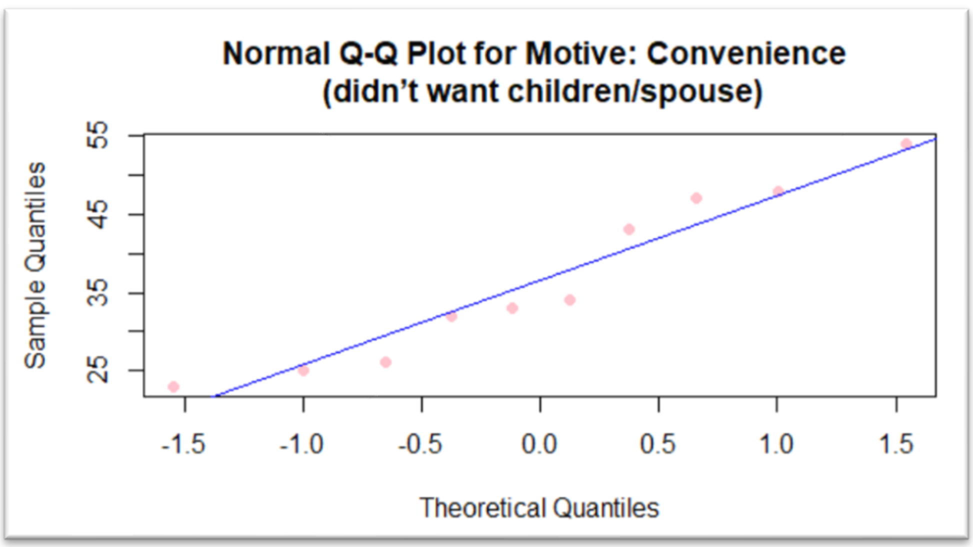

Figure 12: Normal Q-Q Plot for motive 3

A standard Q-Q plot of data of killers with various intentions is shown in Figures 10, 11, and 12. In each of the three figures, the points are basically in a straight line with slight deviation from the line, it's likely that our sample is drawn from a normal distribution. Hence, we can z-test on them. But in figure 12, Because of the limited sample size (10) and most of the data points are not falling on the straight line. I'm not convinced to perform a z-test on it.

Overall, I'll do a z-test for motives 1 and 2. I'll use a t-test for motive 3.

Hypothesis Test with 95% Confidence Interval:

Null Hypothesis(H_0) : true mean is equal to 27

Alternative hypothesis: true mean is not equal to 27

Motive 1: Unknown

Motive 2: Robbery or financial gain

Motive 3: Convenience (didn't want children/spouse)

Z-Test:

	p-value (probability)	Confidence Interval (Age)	Sample mean (Age)	Sample size (count)
Motive 1	**0.00834**	**28 - 32**	**29.97**	**77**
Motive 2	**5.993e-10 ≈ 0**	**29 - 30**	**29.45**	**510**

Table 3: Numerical summary of the z-test performed on age at first kill with two different motives.

For motive 1 & 2, we assumed the average age of first kill to be 27 years, but this does not lie within our confidence interval. Furthermore, the probability of the null hypothesis being true is close to zero. As a result, our hypothesis test for motive 1 was unsuccessful.

T -TEST:

	p - value (probability)	Confidence Interval (Age)	Sample mean (Age)	Sample size (count)
Motive 3	**0.022**	**29 - 44**	**36.5**	**10**

Table 4: Numerical summary of the T-test performed on age at first kill with motive 3

We used the t-test for motive 3. However, our test fails here as well, because the average age of 27 does not fall inside the confidence interval, and the p-value is very near to zero.

Comparison of different population:

Killers with different motives are different kind of people. Hence, we have 3 independent samples of data with different motives. Also, the variance for each motive is different as summarized in Table 2. Hence, it makes more sense to perform t-test for two-sample hypothesis test.

Pair 1: Motive 1 & 2 (**Unknown & Robbery or financial gain)**

Pair 2: Motive 2 & 3 (**Robbery or financial gain & Convenience)**

Pair 3: Motive 1 & 3 (**Convenience & Unknown)**

Hypothesis Test with 95% Confidence Interval:

Null Hypothesis(H_0) : true difference in mean is equal to 0

Alternative hypothesis: true difference in mean is not equal to 0

	P-value	Confidence Interval	Difference in mean of first kill	Test Result
Pair 1	**0.67**	**-1.8 to 2.88**	**0.52**	**H_0 is true**
Pair 2	**0.07**	**-14.8 to 0.72**	**7.05**	**H_0 is true**
Pair 3	**0.097**	**-14.45 to 1.4**	**6.53**	**H_0 is true**

Table 5: Numerical summary of the T-test performed on 3 different pairs of motives

The purpose of this test is to answer the study's core question. Because all of the tests were passed, the average age at first murder does not differ between the killers with different motives, as it falls within the confidence interval. However, for pairs 2 and 3 with a significant mean difference, also the probability of the null hypothesis being true is close to zero. As a result, the killers who kill for the sake of convenience and financial gain have a slight difference in the mean of their first kill. The same can be said about pair 3. However, there is little difference in the average age of first kill between killers whose purpose is unknown and those who do it for financial gain.

DISCUSSION

- Despite the fact that all tests were passed, pairs 2 and 3 had a lot of skewness. There was a significant difference in the mean of first kill between two pairs of motives (financial gain & convenience) & (unknown & convenience) as shown in table 5. Furthermore, the probability of the null hypothesis being true is close to zero for pair 2 & 3. However, it was still inside the confidence interval, and the test was passed. This could be very misleading & ineffective.
- There were numerous flaws in the analysis. To begin with, there were numerous outliers in the killer data (age at last kill, age at first kill, and career duration) that influenced the age mean value.
- The sample size for killers with motive 3 (convenience (didn't want children/spouse)) was exceedingly small, which affects the accuracy of results. The points in the QQ plot were close to the straight line, but I wasn't persuaded it was a normal distribution.
- There’s a possibility of type 1 error as our z-test & t-test has failed & we’ve rejected that the average age of first kill is 27 years for all three motives when it might be true.

ACKNOWLEDGEMENT

Despite the fact that I am repeating the obvious, it is a fact. Prof. Benjamin Thorpe, his informative videos, lecture notes, and practical topics have been tremendously helpful to me. I couldn't have done any of it without it. Last but not least, I'd like to express my gratitude to my Indian professor, Dr. M. Selva Balan whose statistics notes helped me understand the mathematical foundations of each density function.