UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity for Individual Pieces of Work

I am aware that the University defines plagiarism as presenting someone else's work 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 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 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 School's policy on mitigation and procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Student Signature Arafat Hussain Date 07/05/2021

Student Name Arafat Hussain Student Number 201488745

Please note.
When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:
"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.
I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity."

Name: Arafat Hussain
Student ID: 201488745
Title: Course work report for MATH5741: Statistical Theory and Methods.

Introduction-

The coursework entails analysing the dataset based on a sample of killers from the Radford/FGCU Serial Killer database.

The first step involves downloading the killersandmotives.Rdata from Minerva and loading into R. Each student has been assigned a code number as x in order to obtain their data subset from killersandmotives.Rdata. By using the number provided to me which is 14, I created my subset as mysample. The goal from here onwards is to analyse mysample.

Data Cleaning-

Data cleaning involved removing rows with missing values for the variables – AgeFirstKill (missing values are denoted as 99999 for this column), Motive(missing values are denoted as NA for this column). Data cleaning also involved removing any record of killers who first killed before the year 1900.

The following code were run and statistics were collected to meet the objectives of this stage.

- Total number of records in mysample before data cleaning steps were performed is 566.
- Total number of records in mysample for missing values of AgeFirstKill is 9.
- Total number of records in mysample for missing values of Motive is 6.
- Total number of records for killers who first killed before the 1900 is 8.

A total of 23 records meeting the above criteria should be removed from the dataset mysample as a part of the data cleaning activity.

The below code was run to create the dataset mysample_cleaned which contains the cleaned data.

```
mysample_cleaned <- mysample[which(mysample$AgeFirstKill != 99999), ]
mysample_cleaned <- mysample_cleaned[!is.na(mysample_cleaned$Motive), ]
mysample_cleaned <- mysample_cleaned[(mysample_cleaned$YearBorn +
mysample_cleaned$AgeFirstKill) >= 1900, ]
```

Sanity check was performed again running the below code to make sure that the dataset mysample_cleaned does not contain any rows with missing data as described above.

```
nrow(mysample_cleaned[mysample_cleaned$AgeFirstKill == 99999 |
is.na(mysample_cleaned$Motive) | (mysample_cleaned$YearBorn +
mysample_cleaned$AgeFirstKill) < 1900, ])
```

Output – 0

This way we ensured that the dataset mysample_cleaned was a cleaned dataset to be used for analysis and contained 543 records.

The variable for career duration was not readily available in the provided the dataset and was added using the below code.

```
mysample_cleaned[mysample_cleaned$CareerDuration != (mysample_cleaned$AgeLastKill -
mysample_cleaned$AgeFirstKill), ]
```

Data Exploration-

In this stage, numerical and graphical summaries where derived for the distribution of the three variables: AgeFirstKill, AgeLastKill and CareerDuration.

Numerical Summaries -

Moment based summaries are derived for the mentioned variables in R and the stats for the same are captured below.

1. **Mean-**

AgeFirstKill	AgeLastKill	CareerDuration
29.70902	32.78821	3.07919

2. **Standard Deviation-**

AgeFirstKill	AgeLastKill	CareerDuration
9.036828	10.76524	6.15087

Quantile based summaries were derived for the mentioned variables in R and the stats for the same are capture in the below tables

1. **Quantiles**

	Min	Lower Quartile	Median	Upper Quartile	Max
AgeFirstKill	13	23	28	35	75
AgeLastKill	15	25	30	39	77
CareerDuration	0	0	0	3	39

2. **Inter Quartile Range (IQR)**

AgeFirstKill	AgeLastKill	CareerDuration
12	14	13

Graphical Summaries –

Graphical summaries in form of boxplots and histograms have been plotted for the variables AgeFirstKill, AgeLastKill & CareerDuration as shown below.

1. **Boxplots**

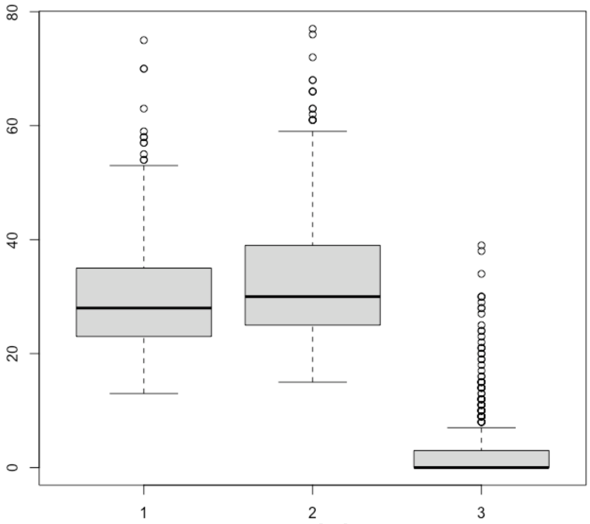


Fig 1.1

The boxplots above depict the distributions of the variables 1) AgeFirstKill 2) AgeLastKill 3) CareerDuration. It can be seen that the median for Age of First Kill and Age of Last Kill lie around the same range with the distribution being skewed towards the right whereas the median for career duration is at 0 with a long tail towards the right showing career duration for most killers in the dataset don’t even last a year.

2. Histograms

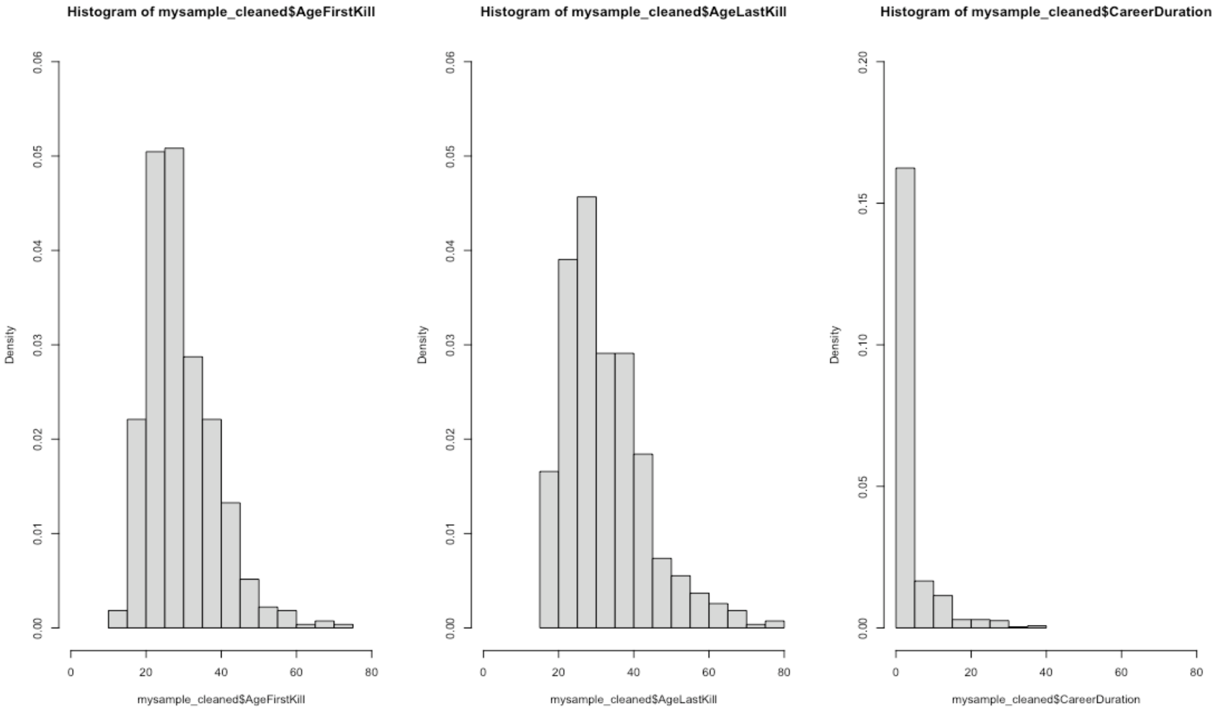


Fig 1.2

The histograms for Age of First Kill and Age of Last Kill clearly show that the data is right skewed. For career duration as well the data show a heavy level of skewness towards the right. The histograms for Age of First Kill and Age of Last Kill points towards a normal distribution for their distributions whereas the histogram for CareerDuration points towards an exponential distribution.

Strong +ve correlation of 0.8209063 observed between Age of First Kill & Age of Last Kill indicating killers starting at young age stop at a young age and killers starting at old age stop at an older age. Very weak -ve correlation -0.03244588 observed between Age of First Kill and Career duration. This can be observed in the below scatter plot as well.

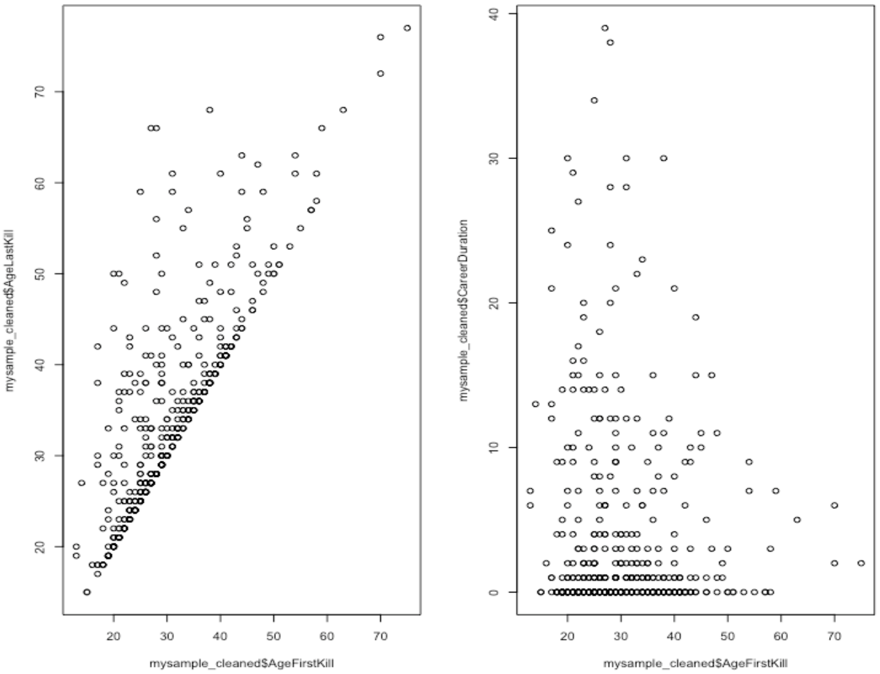
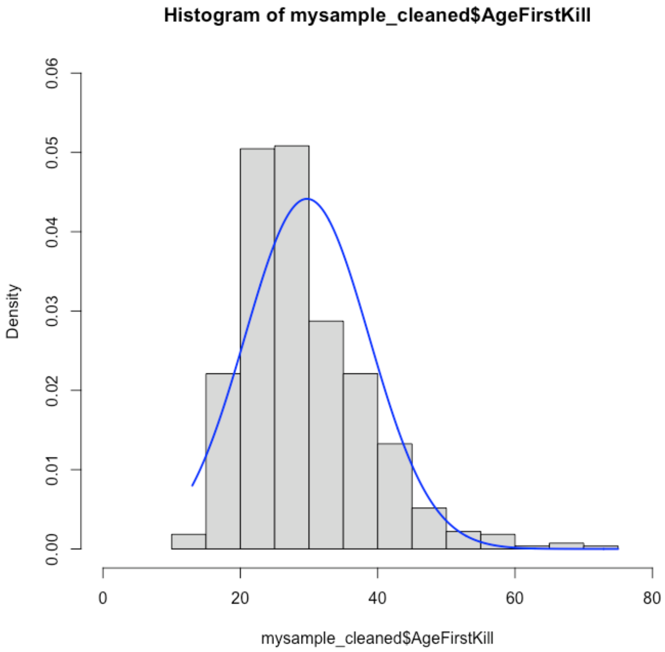


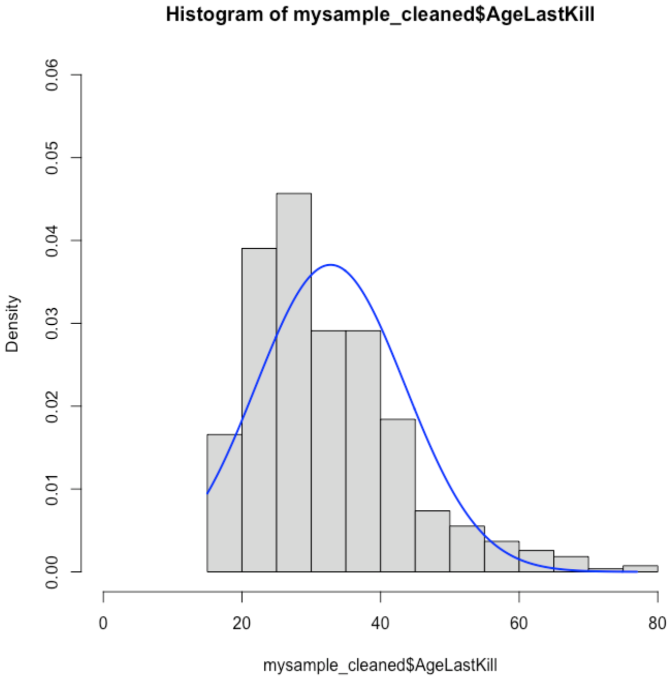
Fig 1.3

Modelling-

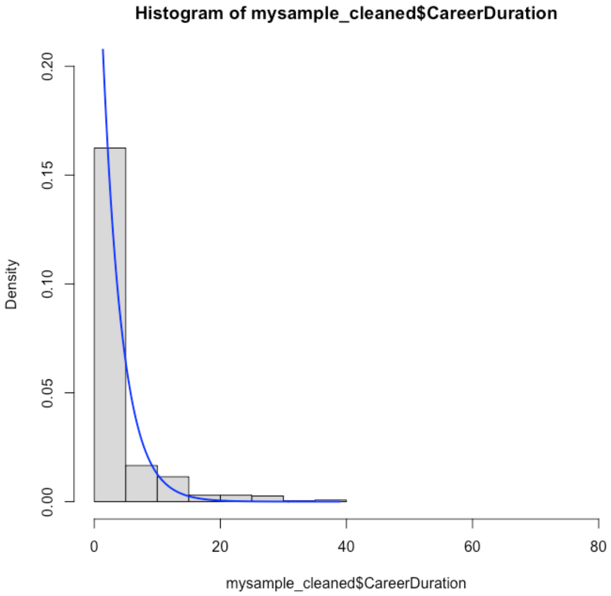
Age of First Kill – The histogram plotted earlier hints that the distribution of this variable is approximately a normal distribution(Baruah, M. and Thorpe, B. 2021) with a positive skew. The nature of the variable tells us that it is time related variable which is continuous. Hence we model it by trying to plot its normal distribution density curve with the parameter values mu = 29.70902, sigma = 9.036828 and get the below graph.



Age of Last Kill - The histogram plotted earlier hints that the distribution of this variable is approximately a normal distribution(Baruah, M. and Thorpe, B. 2021) with a positive skew. The nature of the variable tells us that it is time related variable which is continuous. Hence we model this by trying to plot its normal distribution density curve with the parameter values mu = 32.78821, sigma = 10.76524 and get the below graph.



Career Duration- The histogram plotted earlier for this variable suggests an exponential distribution for this variable (Baruah, M. and Thorpe, B. 2021). The nature of the variable tells us that it is time related variable which is continuous. Hence we try to model this by plotting its exponential distribution density curve but calculating the rate parameter lambda using method of moments estimation and found lambda = 0.3247 there by using this value.



Estimation-

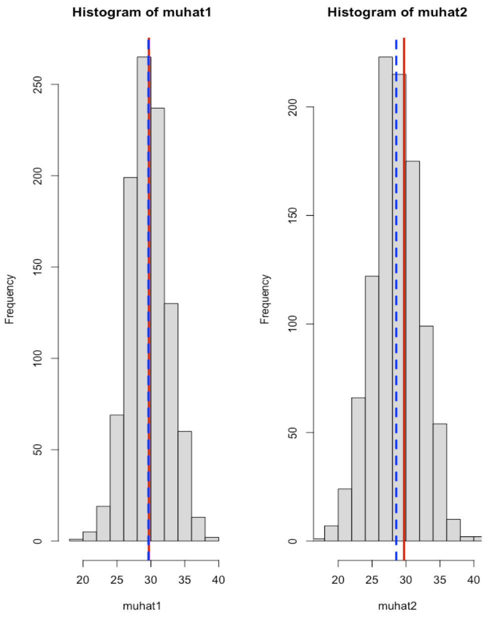
Age of First Kill – We have seen by modelling that distribution of this variable is a Normal distribution. We estimate its parameters **μ** and σ^2 as – the sample mean and the sample variance of this variable, respectively by **method of moments** (Baruah, M. and Thorpe, B. 2021).

μ = 29.70902 , σ^2 = 81.66425

Estimating μ - We then generate 1000 different samples simulating this distribution each of size 10 using these parameter values. We then consider two estimators for **μ**.

muhat1 = The sample mean.
muhat2 = The 50% quantile.

We then calculate the sample mean and the 50% quantile of each of these 1000 samples and plot the below histograms which show the sampling distribution of these two estimators.

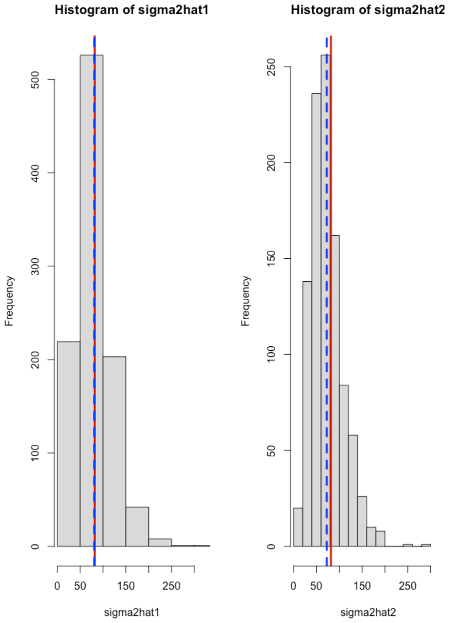


It can be observed from the graphs above that the average of the estimator **muhat1** (represented by the blue line) is unbiased while the average of the estimator **muhat2** (represented by the blue line) has a negative bias underestimating the true value of **μ** (represented by the red line in both the graphs). There appears a little difference in the variance of the two estimator however since **muhat2** has a bias, it suggests that **muhat2** has a greater **MSE** than **muhat1**. Hence we choose **muhat1,** i.e. the sample mean as the better estimator.

Estimating σ^2 - We again generate 1000 different samples simulating the distribution of **Age of First Kill** each of size 10 using the same parameter values as done in the previous section and consider two possible estimators for σ^2.

sigma2hat1 = S^2 which is the sample variance
sigma2hat2 = (n-1/n) S^2 which is the which is the maximum likelihood estimator (MLE) of σ^2

We then calculate the sample variance and the MLE of each of these 1000 samples in R and plot the below histograms which show the sampling distribution of these two estimators.



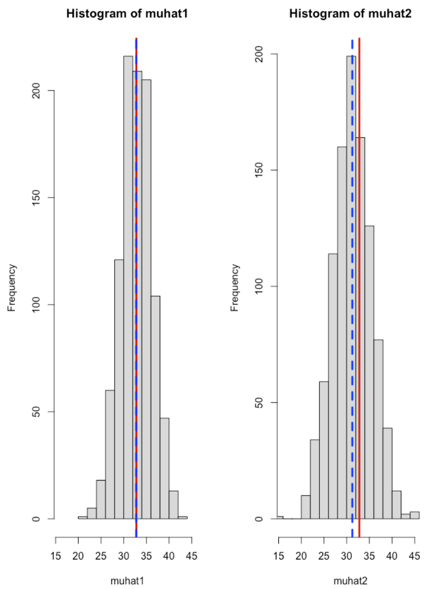
It can be observed from the graphs above that the average of the estimator **sigma2hat1** (represented by the blue line) is unbiased while the average of the estimator **sigma2hat2** (represented by the blue line) has a negative bias underestimating the true value of σ^2 (represented by the red line in both the graphs). There appears a some difference in the variance of the two estimator but since the size of the sample is large the difference is negligible. Hence unbaised estimator **sigma2hat1** is a better estimator for this parameter.

Age of Last Kill – We have seen by modelling that distribution of this variable is a Normal distribution. We estimate its parameters **μ** and σ^2 as – the sample mean and the sample variance of this variable respectively by **method of moments**.

μ = 32.78821 , σ^2 = 115.8905

Estimating μ - We then generate 1000 different samples simulating this distribution each of size 10 using these parameter values.We then consider two estimators for **μ** –

muhat1 = The sample mean.
muhat2 = The 50% quantile.



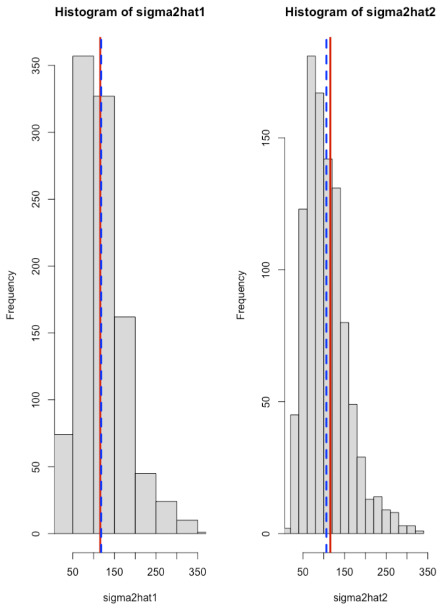
We then calculate the sample mean and the 50% quantile of each of these 1000 samples in R and plot the histograms above which show the sampling distribution of these two estimators.

It can be observed from the graphs above that the average of the estimator **muhat1** (represented by the blue line) is unbiased while the average of the estimator **muhat2** (represented by the blue line) has a negative bias underestimating the true value of **μ** (represented by the red line in both the graphs). There appears a little difference in the variance of the two estimator however since **muhat2** has a bias, it suggests that **muhat2** has a greater **MSE** than **muhat1**. Hence we choose **muhat1,** i.e. the sample mean as the better estimator.

Estimating σ^2 - We again generate 1000 different samples simulating the distribution of **Age of First Kill** each of size 10 using the same parameter values as done in the section above and consider two possible estimators for σ^2.

sigma2hat1 = S^2 which is the sample variance
sigma2hat2 = (n-1/n) S^2 which is the which is the maximum likelihood estimator (MLE) of σ^2

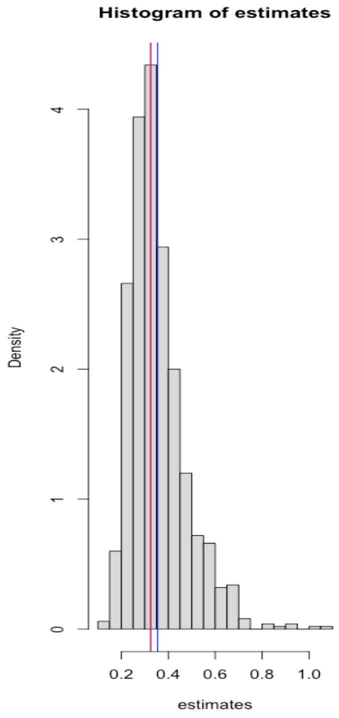
We then calculate the sample variance and the MLE of each of these 1000 samples and plot the below histograms which show the sampling distribution of these two estimators.



It can be observed from the graphs above that the average of the estimator **sigma2hat1** (represented by the blue line) is unbiased while the average of the estimator **sigma2hat2** (represented by the blue line) has a negative bias underestimating the true value of σ^2 (represented by the red line in both the graphs). There appears a some difference in the variance of the two estimator but since the size of the sample is large the difference is negligible. Hence unbaised estimator **sigma2hat1** is a better estimator for this parameter.

Career Duration – For this variable our modelling exercise suggests an exponential distribution. The parameter for this distribution is lambda λ. We estimate this parameter by the **method of moments** as 1 / mean of career duration by which we get λ = 0.3247 (Baruah, M. and Thorpe, B. 2021).

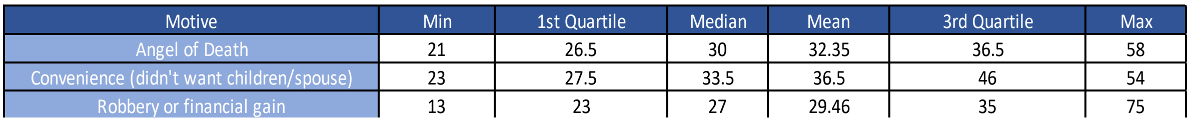
We generate 1000 samples simulating this distribution each of size 10 and plot the sampling distribution by which we get the below graph.



The histogram above shows the that estimates here shown by the blue line tend to overestimate the true value of the parameter represented by the red line.

Hypothesis Testing-

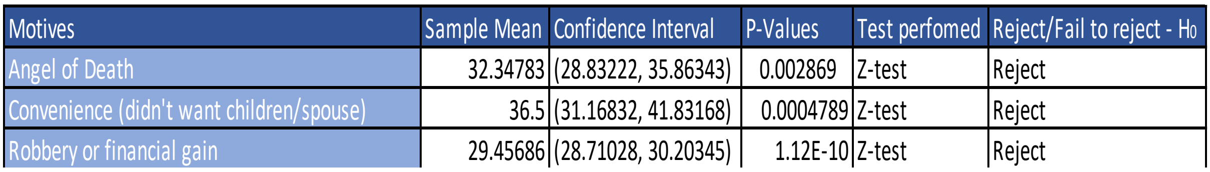
Mentioned below are the numerical summaries of the Age of First Kill divided by motives –



Motive	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Angel of Death	21	26.5	30	32.35	36.5	58
Convenience (didn't want children/spouse)	23	27.5	33.5	36.5	46	54
Robbery or financial gain	13	23	27	29.46	35	75

In order to check if the distribution of all the three motive have normal distribution, we perform the Chi-squared goodness of fit test using the pearson.test( ) in R (Baruah, M. and Thorpe, B. 2021). The p-value of Age of First Kill obtained for Angel of Death, Convenience and Robbery or financial gain are 0.09597, 0.07855 and < 2.2e-16. Since all our tests are at 5% significance level, and the p- value for the motives Angel of Death and Convenience are above the significance level we fail to reject the null hypothesis that they have a normal distribution. However for the motive Robbery and Financial gain, the p-value is below the significance level. Therefore we reject the null hypothesis that it has a normal distribution.

We, now perform hypothesis tests for each of the motives to determine if the mean Age of First Kill is 27 years. The null hypothesis here is that the mean Age of First Kill is 27 years for each of the motives. We perform Z-test on the motives Angel of Death and Convenience even though there sample size is small but their distribution is assumed to be a normal distribution from the results of the Chi-squared goodness of fit test performed above. For the motive Robbery and financial gain, we select a Z-test as well, since the sample size in this case is large i.e. 510 samples. (Baruah, M. and Thorpe, B. 2021). Mentioned below are the results of the Z- test.

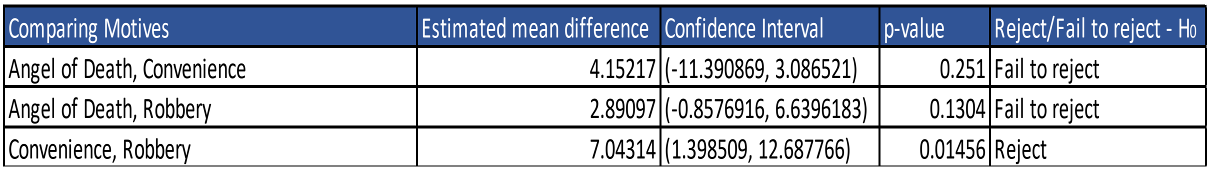


Motives	Sample Mean	Confidence Interval	P-Values	Test perfomed	Reject/Fail to reject - Ho
Angel of Death	32.34783	(28.83222, 35.86343)	0.002869	Z-test	Reject
Convenience (didn't want children/spouse)	36.5	(31.16832, 41.83168)	0.0004789	Z-test	Reject
Robbery or financial gain	29.45686	(28.71028, 30.20345)	1.12E-10	Z-test	Reject

On the basis of the results obtained from the test above we can see that the p-value of the three motives are well below the significance level of 5%. Also the mean value of 27 proposed in the null hypothesis fall outside the confidence intervals of all the three motives. These evidences against the null hypothesis lead us to reject the same at 5% significance level for all the three motives.

Comparison of populations-

Here we perform a hypothesis test to check if the mean Age of First Kill differ by motives. Our null hypothesis is that there is no difference in the true means of Age of First Kill across the motives while the alternative hypothesis is that the true difference in the means is not equal to 0.
For each comparison we perform t-test (Baruah, M. and Thorpe, B. 2021) which yields the below results –



Comparing Motives	Estimated mean difference	Confidence Interval	p-value	Reject/Fail to reject - Ho
Angel of Death, Convenience	4.15217	(-11.390869, 3.086521)	0.251	Fail to reject
Angel of Death, Robbery	2.89097	(-0.8576916, 6.6396183)	0.1304	Fail to reject
Convenience, Robbery	7.04314	(1.398509, 12.687766)	0.01456	Reject

The p-values from these tests for the first two comparisons in the table above are above the 5% significance level as well as the confidence intervals for those includes the 0 difference in the mean values which is evidence is in favour of the null hypothesis hence we fail to reject the null hypothesis for those two comparisons at 5% significance level.
For the last comparison in the table above, the p-value is well below the significance level and the confidence interval for the same excludes the 0 difference in mean for that comparison. This is evidence against the null hypothesis. Hence we reject the null hypothesis at 5% significance level in this case.

Interpretation-

In this course work we have primarily analysed three variables from the dataset which are Age of First Kill, Age of Last Kill and Career Duration. Based on the nature of the variables and its data, we have proposed appropriate models for each of them.
A strong positive correlation was observed between the variables Age of First Kill and Age of Last Kill.
The Z-tests performed for all the motives to check the null hypothesis that the mean of Age of First Kill is 27 for each motive is rejected at 5% significance level based on the results of the test.
The results from the t-tests performed to check the null hypothesis, that there is no difference in the average Age of First Kill between the motives, fail to reject the hypothesis at 5% significance level for comparisons between- Angel of Death & Convenience as well as Angel of Death & Robbery or financial gain. However, the results from the t-test performed, for comparison between the motives - Convenience & Robbery or financial gain, lead us to reject the null hypothesis at 5% significance level.

Reference

Baruah, M. and Thorpe, B. 2021. *Estimation.* [PDF document accessed through Minerva]. MATH5741 School of Mathematics. University of Leeds.

Baruah, M. and Thorpe, B. 2021. *Inference.* [PDF document accessed through Minerva]. MATH5741 School of Mathematics. University of Leeds.

Baruah, M. and Thorpe, B. 2021. *Statistical Models.* [PDF document accessed through Minerva]. MATH5741 School of Mathematics. University of Leeds.

Baruah, M. and Thorpe, B. 2021. *t-tests.* [PDF document accessed through Minerva]. MATH5741 School of Mathematics. University of Leeds.

Appendix

R – Code

Data cleaning –

```
nrow(mysample)
nrow(mysample[mysample$AgeFirstKill == 99999, ])
nrow(mysample[is.na(mysample$Motive), ])
nrow(mysample[(mysample$YearBorn + mysample$AgeFirstKill) < 1900, ])
```

Sanity check performed using the below code to check the counts pertaining to the above criteria which should add up to 23.

```
nrow(mysample[mysample$AgeFirstKill == 99999 | is.na(mysample$Motive) | (mysample$YearBorn +
mysample$AgeFirstKill) < 1900,  ])
```

Data Exploration-

Mean –

```
mean(mysample_cleaned$AgeFirstKill)
mean(mysample_cleaned$AgeLastKill)
mean(mysample_cleaned$CareerDuration)
```

Standard Deviation –

```
sd(mysample_cleaned$AgeFirstKill)
sd(mysample_cleaned$AgeLastKill)
sd(mysample_cleaned$CareerDuration)
```

Quantiles –

```
quantile(mysample_cleaned$AgeFirstKill, type = 1)
quantile(mysample_cleaned$AgeLastKill, type = 1)
quantile(mysample_cleaned$CareerDuration, type = 1)
```

Inter quantile range –

```
IQR(mysample_cleaned$AgeFirstKill, type = 1)
IQR(mysample_cleaned$AgeLastKill, type = 1)
IQR(mysample_cleaned$CareerDuration, type = 1)
```

Graphical Summaries-

```
boxplot(mysample_cleaned$AgeFirstKill, mysample_cleaned$AgeLastKill, mysample_cleaned$CareerDuration)

par(mfrow = c(1,3))
hist(mysample_cleaned$AgeFirstKill, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.06))
hist(mysample_cleaned$AgeLastKill, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.06))
hist(mysample_cleaned$CareerDuration, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.2))

cor(mysample_cleaned$AgeFirstKill, mysample_cleaned$AgeLastKill)
cor(mysample_cleaned$AgeFirstKill, mysample_cleaned$CareerDuration)
cor(mysample_cleaned$AgeLastKill, mysample_cleaned$CareerDuration)

plot(mysample_cleaned$AgeFirstKill, mysample_cleaned$AgeLastKill)
plot(mysample_cleaned$AgeFirstKill, mysample_cleaned$CareerDuration)
```

Modelling-

Age of First Kill -

```
hist(mysample_cleaned$AgeFirstKill, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.06))

mean(mysample_cleaned$AgeFirstKill) # 29.70902
sd(mysample_cleaned$AgeFirstKill) # 9.036828
```

```
x <- seq(from = min(mysample_cleaned$AgeFirstKill), to = max(mysample_cleaned$AgeFirstKill), by = 0.1)

lines(x, dnorm(x, mean = 29.70902, sd = 9.036828), lwd = 2, col = "blue")
```

Age of last Kill

```
hist(mysample_cleaned$AgeLastKill, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.06))
mean(mysample_cleaned$AgeLastKill)
sd(mysample_cleaned$AgeLastKill)
x <- seq(from = min(mysample_cleaned$AgeLastKill), to = max(mysample_cleaned$AgeLastKill), by = 0.1)
lines(x, dnorm(x, mean = 32.78821, sd = 10.76524), lwd = 2, col = "blue")
```

Career Duration

```
hist(mysample_cleaned$CareerDuration, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.2))
x <- seq(from = min(mysample_cleaned$CareerDuration), to = max(mysample_cleaned$CareerDuration), by = 0.1)
mean(mysample_cleaned$CareerDuration)
sd(mysample_cleaned$CareerDuration)

lines(x, dexp(x, rate = 0.3247), lwd = 2, col = "blue")
```

Estimation-

Age of first kill

```
hist(mysample_cleaned$AgeFirstKill, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.06))
mean(mysample_cleaned$AgeFirstKill)
sd(mysample_cleaned$AgeFirstKill)
sd(mysample_cleaned$AgeFirstKill)^2
```

Estimating the mu - mean

```
mu      <- mean(mysample_cleaned$AgeFirstKill)
sigma   <- sd(mysample_cleaned$AgeFirstKill)

muhat1  <- rep(NA, 1000)
muhat2  <- rep(NA, 1000)

for(i in 1:1000){

  x <- rnorm(n = 10, mean = mu, sd = sigma)
  muhat1[i] <- mean(x)
  muhat2[i] <- quantile(x, type = 1)[3]

}

par(mfrow = c(1, 1))

hist(muhat1, xlim = range(c(muhat1, muhat2)))
abline(v = mu, col = "red3", lwd = 3)
abline(v = mean(muhat1), col = "blue", lty = 2, lwd = 3)

hist(muhat2, xlim = range(c(muhat1, muhat2)))
abline(v = mu, col = "red3", lwd = 3)
abline(v = mean(muhat2), col = "blue", lty = 2, lwd = 3)
```

Estimating the sigma

```
sigma2hat1 <- rep(NA, 1000)
sigma2hat2 <- rep(NA, 1000)

for(i in 1:1000){

  x <- rnorm(n = 10, mean = mu, sd = sigma)

  sigma2hat1[i] <- sd(x)^2
  sigma2hat2[i] <- (9/10)*sd(x)^2
}
```

```
par(mfrow = c(1, 2))

hist(sigma2hat1, xlim = range(c(sigma2hat1, sigma2hat2)))
abline(v = sigma^2, col = "red3", lwd = 3)
abline(v = mean(sigma2hat1), col = "blue", lty = 2, lwd = 3)

hist(sigma2hat2, xlim = range(c(sigma2hat1, sigma2hat2)))
abline(v = sigma^2, col = "red3", lwd = 3)
abline(v = mean(sigma2hat2), col = "blue", lty = 2, lwd = 3)
```

Age of last kill

```
hist(mysample_cleaned$AgeLastKill, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.06))

mean(mysample_cleaned$AgeLastKill) #32.78821
sd(mysample_cleaned$AgeLastKill) #10.76524
sd(mysample_cleaned$AgeLastKill)^2 #115.8905

Estimating the mu - mean

mu      <- mean(mysample_cleaned$AgeLastKill)
sigma   <- sd(mysample_cleaned$AgeLastKill)

muhat1  <- rep(NA, 1000)
muhat2  <- rep(NA, 1000)

for(i in 1:1000){

  x <- rnorm(n = 10, mean = mu, sd = sigma)
  muhat1[i] <- mean(x)
  muhat2[i] <- quantile(x, type = 1)[3]

}

par(mfrow = c(1, 2))

hist(muhat1, xlim = range(c(muhat1, muhat2)))
abline(v = mu, col = "red3", lwd = 3)
abline(v = mean(muhat1), col = "blue", lty = 2, lwd = 3)

hist(muhat2, xlim = range(c(muhat1, muhat2)))
abline(v = mu, col = "red3", lwd = 3)
abline(v = mean(muhat2), col = "blue", lty = 2, lwd = 3)

# Estimating the sigma

sigma2hat1 <- rep(NA, 1000)
sigma2hat2 <- rep(NA, 1000)

for(i in 1:1000){

  x <- rnorm(n = 10, mean = mu, sd = sigma)

  sigma2hat1[i] <- sd(x)^2
  sigma2hat2[i] <- (9/10)*sd(x)^2
}


par(mfrow = c(1, 2))

hist(sigma2hat1, xlim = range(c(sigma2hat1, sigma2hat2)))
abline(v = sigma^2, col = "red3", lwd = 3)
abline(v = mean(sigma2hat1), col = "blue", lty = 2, lwd = 3)

hist(sigma2hat2, xlim = range(c(sigma2hat1, sigma2hat2)))
abline(v = sigma^2, col = "red3", lwd = 3)
abline(v = mean(sigma2hat2), col = "blue", lty = 2, lwd = 3)

# Career duration
```

```
hist(mysample_cleaned$CareerDuration, freq = FALSE, xlim = c(0, 80), ylim = c(0, 0.2))

x <- seq(from = min(mysample_cleaned$CareerDuration), to = max(mysample_cleaned$CareerDuration), by = 0.1)

mean(mysample_cleaned$CareerDuration) #3.07919
sd(mysample_cleaned$CareerDuration) #6.15087
1/mean(mysample_cleaned$CareerDuration)

# estimating lambda(rate) by method of moments rate = 1/3.07919 = 0.3247
n <- 10

lambda <- 1/mean(mysample_cleaned$CareerDuration)

estimates <- rep(NA, 1000)   # Empty vector to store estimates.

for(i in 1:1000){

  x <- rexp(n, lambda)       # Sample in the i-th experiment.

  lambdahat <- 1/mean(x)     # Estimate in the i-th experiment.

  estimates[i] <- lambdahat  # Store the lambdahat in our vector.
}

hist(estimates, breaks = "FD", freq = FALSE)

abline(v = mean(estimates), col = "blue")

abline(v = lambda, col = "red")
```

Hypothesis Testing

```
unique(mysample_cleaned$Motive)

Motives_AngelofDeath <- mysample_cleaned[mysample_cleaned$Motive == "Angel of Death", ]
Motives_Convenience <- mysample_cleaned[mysample_cleaned$Motive == "Convenience (didn't want
children/spouse)", ]
Motives_Robbery <- mysample_cleaned[mysample_cleaned$Motive == "Robbery or financial gain", ]


nrow(Motives_AngelofDeath) # 23
nrow(Motives_Convenience) # 10
nrow(Motives_Robbery) # 510

summary(Motives_AngelofDeath$AgeFirstKill)
summary(Motives_Convenience$AgeFirstKill)
summary(Motives_Robbery$AgeFirstKill)

### Normality check using Chi-squared goodness of fit test

pearson.test(Motives_AngelofDeath$AgeFirstKill)
pearson.test(Motives_Convenience$AgeFirstKill)
pearson.test(Motives_Robbery$AgeFirstKill)

### Z test for hypothesis testing of the null hypothesis that the average age of first kill accross motives is 27

Z_AOD <- z.test(x = Motives_AngelofDeath$AgeFirstKill, mu = 27, sigma.x = 74^(1/2), conf.level = 0.95)
Z_CNV <- z.test(x = Motives_Convenience$AgeFirstKill, mu = 27, sigma.x = 74^(1/2), conf.level = 0.95)
Z_ROB <- z.test(x = Motives_Robbery$AgeFirstKill, mu = 27, sigma.x = 74^(1/2), conf.level = 0.95)
```

Comparison of populations-

Two sample hypothesis test

```
t.test(x = Motives_AngelofDeath$AgeFirstKill, y = Motives_Convenience$AgeFirstKill, mu = 0, paired = FALSE,
var.equal = TRUE, conf.level = 0.95)

t.test(x = Motives_AngelofDeath$AgeFirstKill, y = Motives_Robbery$AgeFirstKill, mu = 0, paired = FALSE, var.equal
= TRUE, conf.level = 0.95)

t.test(x = Motives_Convenience$AgeFirstKill, y = Motives_Robbery$AgeFirstKill, mu = 0, paired = FALSE, var.equal
= TRUE, conf.level = 0.95)
```