ANALYSIS OF SERIAL KILLER'S MOTIVES AND MEAN AGE

1.INTRODUCTION

The term 'serial killer' has become part of modern vocabulary and has also become a dominant theme in films, media, and literature. Serial killers can be defined as people who murder at least three people in at least three separate locations over a span of time. Serial killers usually work alone, kill strangers, and kill for the sake of killing (as opposed to crimes of passion). We will examine some distinct features and facts in this paper about killers age, motives, race, and gender.

We will work on sample data set taken from Radford/FGCU serial killer database. Our focus is on three variables from the data set which are, agefirstkill, agelastkill and career duration between first and last kill. We will use these variables in our report to perform some statistics to understand killers' profile, background, and motives for crimes. Finally, we investigate average age at first kill variations with different motives. After analysis we found that majority of serial killer tends to fall into the white racial group, followed by black racial group who are second majority in serial killer list. It is also observed that for all variables the primary motive behind killings is robbery or financial gain. There is strong correlation between age first and last kill and a weak correlation between age first kill and career duration. In coming sections, we will analyze and explore the data from data cleansing to till conclusion and interpretation.

Our final objective of the report is, does the average at first kill murder differ between killers with different motives or not.

2.Analysis Phases

Before going via each analysis phase lets first check our data set structure. Data set mysample contains a total of 618 records with 9 columns.

Table 1: mysample data set structure

column name	description
KillerID	A unique identification number for each killer
AgeFirstKill	The age of each killer when they committed their first murder
AgeLastKill	The age of each killer when they committed their last murder
YearBorn	The calendar year in which each killer was born
Motive	The motive of each killer
Sex	The sex of each killer
Race	The race of each killer
Sentence	The sentence each killer received
InsanityPlea	Whether each killer made a plea of not guilty

2.1 Data Cleaning

We cleaned some discrepancy and null values data from age at first kill, year born and motive variables. We found 6 null values in motive variable where it constitutes 0.97% of total observations. Similarly, we had some abnormal data in age at last kill where value of the

column is 99999. A total of 9 observations were found and the percentage of such data in data set is 1.45. we also filtered year born column where value is less than 1900. There are total of 35 records of such data and it constitutes 5.66% of total observations. Finally, we added new column called career duration which is difference between age at last kill and age at first kill.

After cleaning and adding new variable, we now have a total of 569 records and 10 columns.

Table 2: Cleaning of mysample data set summary

Column name	Filter value	Total records	% of total data set
AgeFirstKill	99999	9	1.45
AgeFirstKill	<1900	35	5.66
Motive	NA	6	0.97

2.2 Data Exploration

In this topic we will do some statistics on our three variables, AgeFirstKill, AgeLastKill and Career Duration

Table 3: Numerical summary

parameters	variables		
	AgeFirstKill	AgeLastKill	CareerDuration
Standard Dev	8.91	10.47	6.28
Mean	29.41	32.64	3.22
length	569.00	569.00	569.00
Median	27.00	30.00	0.00
Minimum	13.00	15.00	0.00
Maximum	75.00	77.00	39.00
skewness	1.27	1.10	2.61

From numerical summary we can see that 50% of the observations for age at first kill is <= 27 years old Whereas age at last kill is <=30 years old. The maximum age that committed first kill is 75 and for last kill is 77. Another interesting trend we can see that, more than 50% of observations for career duration variable have zero median. This is due to the age at first and last kill is same. We can also observe that the youngest age that committed first kill is 13 and last kill is 15.

Figure 1 Histogram Plot: Graphical Summary

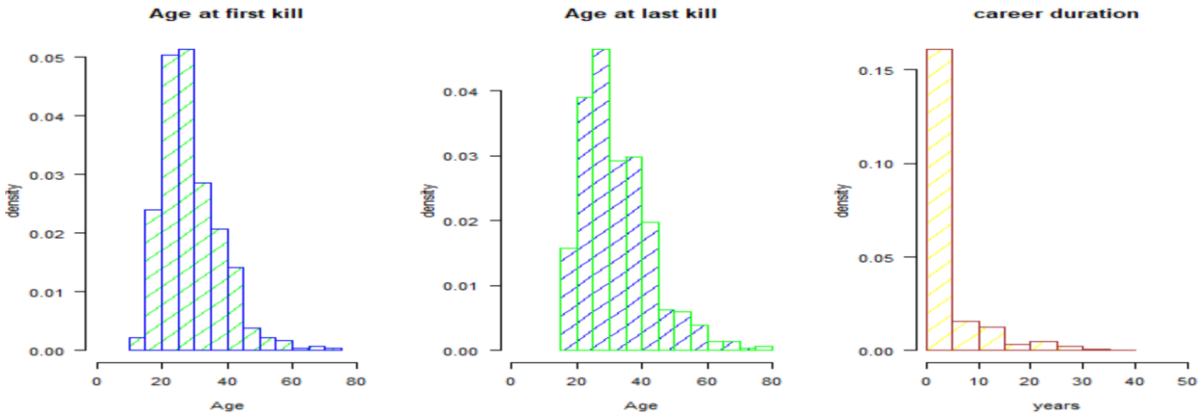


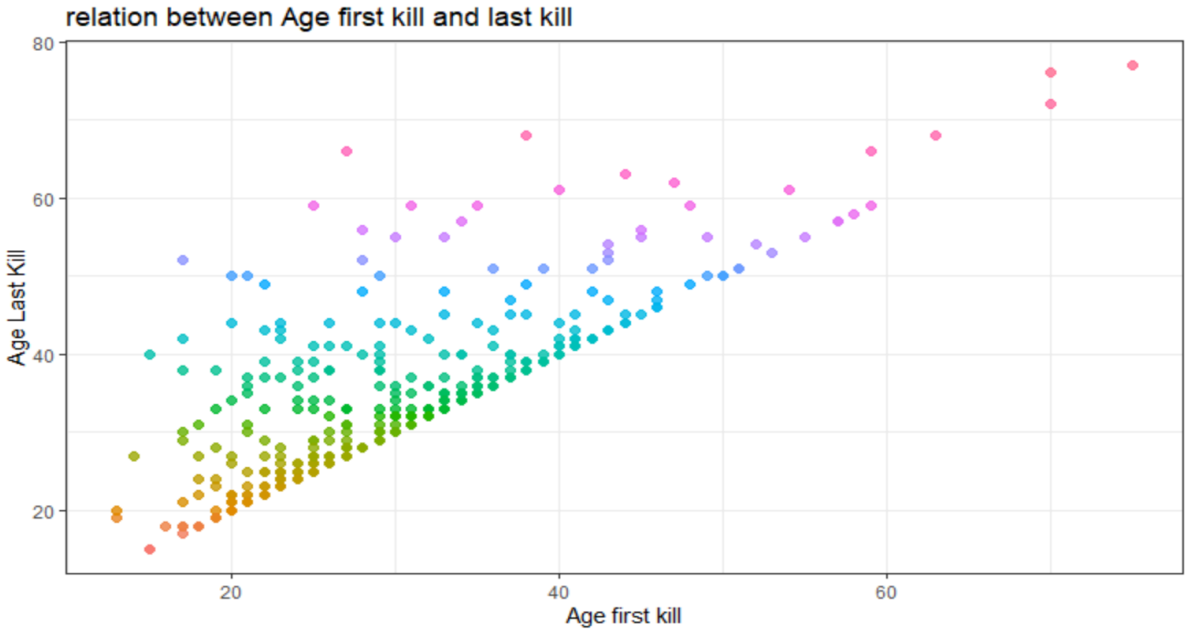
Table 4: Relationship between variables.

Variables	AgeFirstKill	AgeLastKill	CareerDuration
AgeFirstKill	1.0	0.8	-0.1
AgeLastKill	0.8	1.0	0.5
CareerDuration	-0.1	0.5	1.0

The above values are drawn by using cor () function. There is strong correlation between age first and last kill which means number of killers by age is approximately same at first and last kill. Weak correlation can be seen between career duration and Age at first kill.

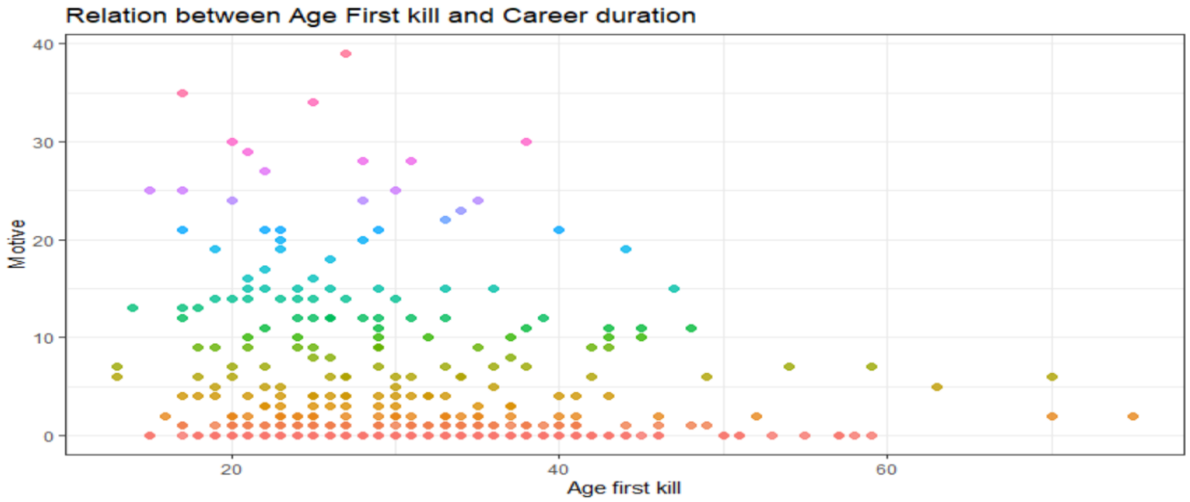
It is also observed that majority of the killings is driven by robbery or financial gain and majority of killers tend to fall in white racial group, followed by black race.

Figure 2 Scatter plot: Relationship between Age first and Last Kill



As you see small circle points correlated between x-axis and y-axis values.

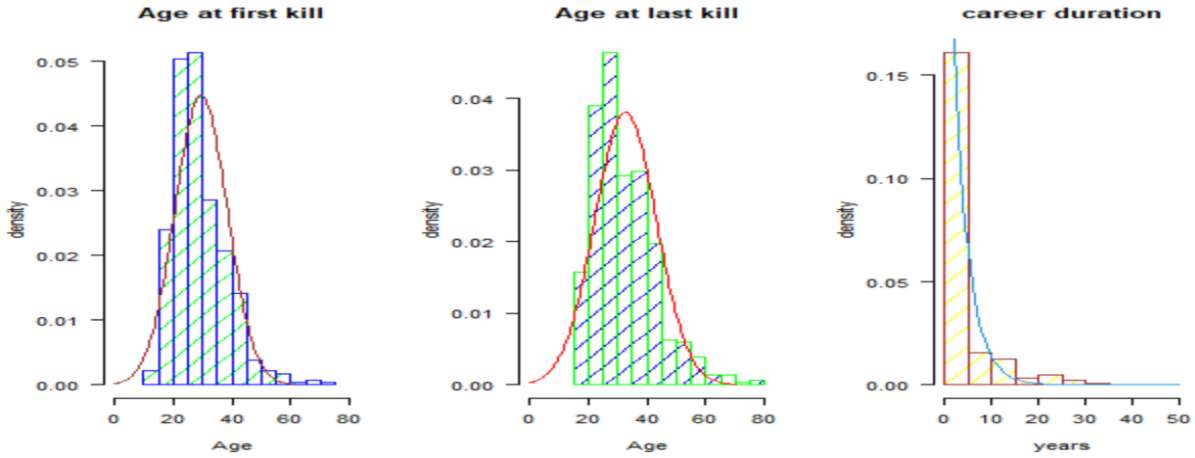
Figure 3 Scatter plot: Relationship between Age first kill and Motive.



We can clearly see from above plot x-axis and y-axis values are poorly correlated.

2.3 Modelling

Figure 4 Histogram plot: density curve on three variables



from above graph it is obvious that most of the histogram bins in age at first kill and last kill fits in the bell-shaped curve. Based on empirical rule (68-95-99), almost all observed data is within three standard deviation as per analysis we done. And there are not many outliners in data as well. So, for variables age at first and last kill we propose normal distribution.

For career duration we have seen that in table 3, the skewness is high to the right side of the distribution with rate parameter lambda=1/mean. Which is 0.3 for this variable. Thus, we recommend exponential distribution for career duration.

2.4 Estimation

Figure 5 Histogram plot: Estimating MLE for mu and variance from mysample data set for variables age at first kill and age at last kill.

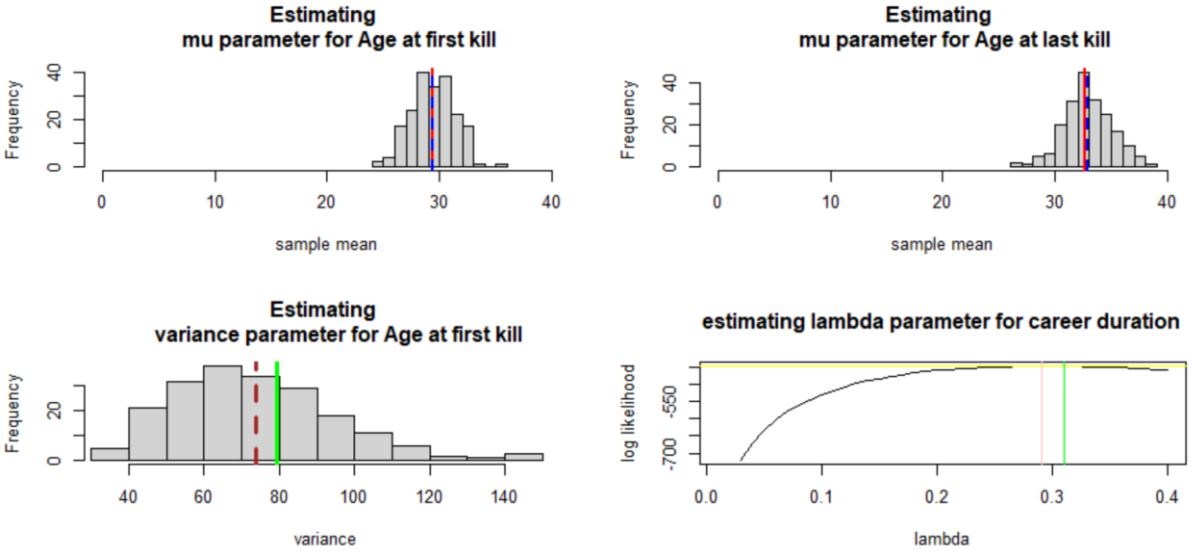


Figure 1: Mean of age at first kill mu =29.41, average estimator =29.55
Figure 2: Mean of age at last kill mu =32.64, average estimator =32.58.
Figure 3: variance of age at first kill sigma^2 = 79.26, average estimator = 73.74
Figure 4: true lambda of career duration = 0.30, MLE of lambda=0.33

Generated 200 samples with size n=25 by using mysample data set mean and standard deviation. We plotted 200 samples mean using histogram, we found that the average estimator (color blue) appears to be unbiased. Both true mean (color red) and average estimator is almost overlapping each other, suggesting it has very little difference. So MLE of mu for age at first kill and last kill is 29.55 and 32.58, respectively. The MLE for variance sigma^2 in figure 3 is 73.74. As you can see the MLE has smaller spread, which makes unbiased estimator.

From figure 4 we clearly see, the MLE of lambda (color pink) is directly below the peak of the curve where log likelihood is maximized at 1/mean. The horizontal line (color yellow) appears to confirm this. The green vertical line is true lambda.

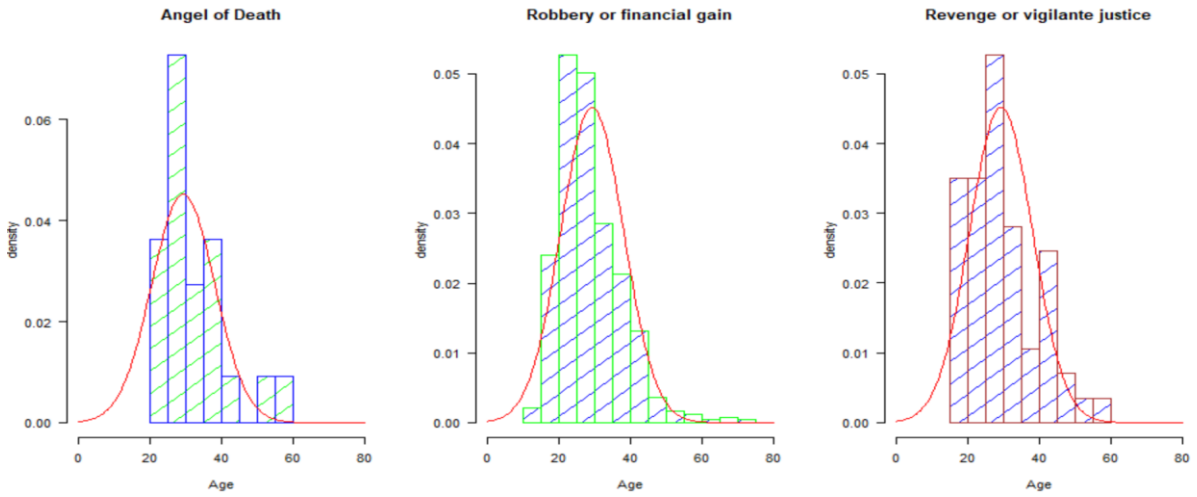
2.5 Testing Hypothesis

Table 5: Numerical summary of age at first kill for different types of motives.

motive	Summary						
	length	mean	median	Std dev	min	max	skewness
Angel of death	22	32.59	30	8.83	21	58	1.32
robbery or financial gain	490	29.20	27	8.82	13	75	1.34
Revenge or vigilante justice	57	30.09	28	9.64	15	59	0.79

Figure 6 Histogram plot: Distribution of age at first kill for different type of motives.

ANALYSIS OF SERIAL KILLER'S MOTIVES AND MEAN AGE



Based on density curve all motives does not seem to be fit in standard normal distribution. However, they are approximately can fit in normal distribution.

We will choose Z-test for all three motives because sample size is greater than 20 and we know population variance and mean.

Using below Z test statistic, p-value and confidence interval formulas, table 6 is generated.
We want to test null hypothesis H0: μ=27 against H1: μ≠27

$$Z = (\bar{X} - \mu)/\sqrt{(\frac{\delta^2}{n})}$$

p-value = $P(Z \epsilon R) \leq \alpha = 0.05$
Where Z is approxiametely N(0,1) by CLT and R is rejection region

$$CI=(\bar{X} - 1.96/\sqrt{(\frac{\delta^2}{n})} , \bar{X} + 1.96/\sqrt{(\frac{\delta^2}{n})})$$

Table 6: Hypothesis testing summary

Motives	Summary				
	size	Sample mean	Z-score	CI	p-value
Angel of death	22	32.59	3.05	(28.44,29.96)	0.002288414
robbery or financial gain	490	29.20	5.67	(28.44,29.96)	0.000000014
Revenge or vigilante justice	57	30.09	2.71	(27.86,32.32)	0.006728321

For all motives we can reject null hypothesis at the 5% significance level since p-value is less than α(0.05). From 95% Confidence interval for all motives, we can see that the parameter μ is outside the interval so we can reject null hypothesis H0: μ=27. For motive "Revenge or vigilante justice" μ is close to the interval but missing by margin. It is approximately matching average mean age 27 at first kill.

2.6 Comparison of Populations

Since sample size is different for all three motives and big, we will choose independent sample testing for all motives for Age at first kill.

We will use below formula for pair 1, pair 2 and pair 3 because standard deviation for all motives between each other is in the ration between 1/3 and 3. So we assume population variation is same for all samples.

$$\left(\overline{X}_1 - \overline{X}_2\right) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Where S_p is the pooled estimate of the variance.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Table 7: Independent samples testing

Pair number	Motives	size	Estimated mean difference	Confidence interval	p-value
1	Angel of death and robbery or financial gain	22 and 490	3.39	(-0.38,7.16)	0.05054
2	Angel of death and Revenge or vigilante justice	22 and 57	2.5	(-2.21, 7.21)	0.049792
3	robbery or financial gain and Revenge or vigilante justice	490 and 57	0.89	(-1.59, 3.37)	0.050674

For all above pairs zero is within confidence interval. So, we cannot reject the hypothesis that the difference between two samples is zero. For pair 1 and pair 3 p-value is slightly greater than α(0.05) which means we cannot reject the mean difference hypothesis. Interestingly pair 2 is less than α(0.05), in this case we should reject our hypothesis as per p-value, but pair 2 p-value is almost closer to 0.05. We cannot judge confidence interval and p-value is contradicting hypothesis result because p-value approximately equal to 0.05.

2.7 Interpretation

Based on analysis we had done so far; we know that mean age for all motives for first kill is different between each other. The highest mean age motive is angel of death standing at 32.59. it is having 1.32 skewness due to inclusion of old age killers. The maximum age killer for motive of angel is 58. The least skewed motive for age at first kill is revenge or vigilante justice

standing at 0.79. The graph for this motive is somehow normally distributed as show in figure 6. We have done some experiments whether mean age will be same or not by combination of race and motive. Except for the motive angel of death, mean age for the rest of motives is same for first kill age. Our limitation is just focusing on motives, we may perhaps require testing with different combinations of attributes like race, sex, motive etc.
Another limitation is sample size is not enough and cleaning records also plays role in deciding mean age. Since we cleaned few records, it can affect the average age at first kill.

Regarding Hypothesis testing and independent testing, the former rejected null hypotheses for all motives whereas later did not reject the claim that mean difference between population samples is zero. Which means independent testing accepts all samples mean average is 27 for age at first kill but hypothesis testing in table 6 contradicting with independent testing. If you check closely in table 6, the CI interval almost contains the proposed mean age value 27 with just 1 to 1.5 points gap. So basically, we can say that the average age at first kill does not much differ between killers and motives. However, by checking accurate results we may to go ahead with analysis results irrespective of assumptions.

References

[1] Benjamin Thorpe and Monita Baruah, 2021. Practical 6 Estimators from Statistical Theory and Methods. https://minerva.leeds.ac.uk/bbcswebdav/pid-8418885-dt-content-rid-18483044_2/courses/202021_37414_MATH5741M/Practical--6--Solutions.html

[2] Benjamin Thorpe and Monita Baruah, 2021. Inference from Statistical Theory and Methods.https://minerva.leeds.ac.uk/webapps/blackboard/execute/content/file?cmd=view&content_id=_8405391_1&course_id=_521621_1&framesetWrapped=true

[3] Benjamin Thorpe and Monita Baruah, 2021. T-tests from Statistical Theory and Methods.https://minerva.leeds.ac.uk/webapps/blackboard/execute/content/file?cmd=view&content_id=_8405391_1&course_id=_521621_1&framesetWrapped=true

[4] Hypothesis testing from Laerd Statistics. https://statistics.laerd.com/statistical-guides/hypothesis-testing-2.php

[5] Benjamin Thorpe and Monita Baruah, 2021.Chapter 3 Models from Statistical Theory and Methods.https://minerva.leeds.ac.uk/webapps/blackboard/execute/content/file?cmd=view&content_id=_8368084_1&course_id=_521621_1&framesetWrapped=true

Appendices

#R code

```
load(file="C:/praveen/leeds university/statistical theory and methods/coursework/killersandmotives.Rdata")
createsample(34)
####packages
install.packages("dplyr")
```

ANALYSIS OF SERIAL KILLER'S MOTIVES AND MEAN AGE

```
library("dplyr")
install.packages("psych")
library(psych)
install.packages("moments")
library(moments)
install.packages("ggplot2")
library(ggplot2)

##percentage of cleaning records
clean_percentage1=(nrow(mysample[mysample$AgeFirstKill=="99999",])/nrow(mysample))*100
clean_percentage2=(nrow(mysample[is.na(mysample$Motive),])/nrow(mysample))*100
clean_percentage3=(nrow(mysample[mysample$YearBorn<"1900",])/nrow(mysample))*100

######cleaning mysample
mysample=mysample[!(is.na(mysample$Motive)),]
mysample=mysample[!(mysample$YearBorn<"1900"),]
mysample=mysample[!(mysample$AgeFirstKill=="99999"),]
mysample["CareerDuration"]=mysample[,3]-mysample[,2]

#######summarize numerically
a_flc=fortest_mysample[,c("AgeFirstKill","AgeLastKill","CareerDuration")]
sapply(a_flc, function(a_flc) c("Stand dev" = sd(a_flc),
                                "Mean"= mean(a_flc),
                                "length" = length(a_flc),
                                "Median" = median(a_flc),
                                "skewness" = skewness(a_flc),
                                "Minimum" = min(a_flc),
                                "Maximum" = max(a_flc)
)

#######summarize graphically#######
hist(age_firstkill,
    main="Age at first kill",
    xlab="Age",
    ylab="density",
    border="blue",
    col="green",
    xlim=c(0,80),
    las=1,
    density=20,
    prob=TRUE)

hist(age_lastkill,
    main="Age at last kill",
    xlab="Age",
    ylab="density",
    border="green",
    col="blue",
    xlim=c(0,80),
    las=1,
    density=20,
    prob=TRUE)
```

ANALYSIS OF SERIAL KILLER'S MOTIVES AND MEAN AGE

```
hist(career_dur,
    main="career duration",
    xlab="years",
    ylab="density",
    border="brown",
    col="yellow",
    xlim=c(0,50),
    las=1,
    density=20,
    prob=TRUE)

###correlation between age_Firstkill,age_lastkill and career_duration#####
cordata = mysample[,c(2,3,10)]
head(cordata,n=5)
corr_variables <- round(cor(cordata), 1)
corr_variables

#correlation by ggplot
ggplot(data=fortest_mysample)+geom_point(aes(x=age_firstkill,
                                y=age_lastkill,color=as.factor(age_lastkill)),size=2,alpha=0.8)+
  #scale_x_continuous(name=age_firstkill,breaks=seq(0,100,5))+
  #scale_y_continuous(name=age_lastkill,breaks=seq(0,300,5))+
  ggtitle(label="relation between Age first kill and last kill")+
  labs(x="Age first kill",y="Age Last Kill")+
  theme_bw()

ggplot(data=fortest_mysample)+geom_point(aes(x=age_firstkill,
                                y=career_dur,color=as.factor(career_dur)),size=2,alpha=0.8)+
  #scale_x_continuous(name=age_lastkill,breaks=seq(0,100,5))+
  #scale_y_continuous(name=motive,breaks=seq(0,300,5))+
  ggtitle(label="Relation between Age First kill and Career duration")+
  labs(x="Age first kill",y="Motive")+
  theme_bw()



par(mfrow = c(1, 2))
###############estimation for mu and sigma for age first and last kill########
mu_firstkill=mean(age_firstkill)
mu_lastkill=mean(age_lastkill)
sigma_firstkill=sd(age_firstkill)
sigma_lastkill=sd(age_lastkill)

avg_mean_firstkill  <- rep(NA, 200)
avg_mean_lastkill  <- rep(NA, 200)

###########first kill MLE mean#######
for(i in 1:200){

  x <- rnorm(n = 25, mean = mu_firstkill, sd = sigma_firstkill)
  avg_mean_firstkill[i] <- mean(x)
  }
```

```
hist(avg_mean_firstkill, xlim = c(0,40),xlab="sample mean",main="Estimating
    mu parameter for Age at first kill")
abline(v = mu_firstkill, col = "red", lwd = 2)
abline(v = mean(avg_mean_firstkill), col = "blue", lty = 2, lwd = 2)

#lastkill MLE mean
for(i in 1:200){

  x <- rnorm(n = 25, mean = mu_lastkill, sd = sigma_lastkill)
  avg_mean_lastkill[i] <- mean(x)
}
hist(avg_mean_lastkill, xlim = c(0,40),xlab="sample mean",main="Estimating
    mu parameter for Age at last kill")
abline(v = mu_lastkill, col = "red", lwd = 2)
abline(v = mean(avg_mean_lastkill), col = "blue", lty = 2, lwd = 2)

#firstkill MLE of sd
mu      <- mean(age_firstkill)
sigma   <- sd(age_firstkill)


sigma2hat1 <- rep(NA, 200)
sigma2hat2 <- rep(NA, 200)

for(i in 1:200){

  x <- rnorm(n = 25, mean = mu, sd = sigma)

  sigma2hat1[i] <- sd(x)^2
  sigma2hat2[i] <- (24/25)*sd(x)^2
}
hist(sigma2hat1, xlim = range(c(sigma2hat1, sigma2hat2)),xlab="variance",main="Estimating
    variance parameter for Age at first kill")
abline(v = sigma^2, col = "green", lwd = 2)
abline(v = mean(sigma2hat1), col = "brown", lty = 2, lwd = 2)

hist(sigma2hat2, xlim = range(c(sigma2hat1, sigma2hat2)),xlab="variance",main="Estimating
    variance parameter for Age at first kill")
abline(v = sigma^2, col = "green", lwd = 3)
abline(v = mean(sigma2hat2), col = "brown", lty = 2, lwd = 3)


##################career_duration mle of lambda
c_dur_mean=mean(career_dur)
c_dur_lamda_true=1/c_dur_mean
x    <-  rexp(n = 200, rate = c_dur_lamda_true)
xbar=mean(x)

loglik <- function(lambda){

  L <- (lambda^200)*exp(-lambda*200*xbar)
  return(log(L))
```

```
}
lambda <- (1:40)/100

plot(lambda, loglik(lambda), xlab = "lambda", ylab = "log likelihood",
main="estimating lambda parameter for career duration",type = "l")
abline(v = 1/xbar, col = "pink")
abline(h = loglik(1/xbar), col = "yellow")
abline(v=c_dur_lamda_true,col="green")


######Numerical summary of age at first kill for different types of motives.
motive_angel=c(fortest_mysample[fortest_mysample$Motive == "Angel of Death", "AgeFirstKill"])
motive_revenge=c(fortest_mysample[fortest_mysample$Motive == "Revenge or vigilante justice",
"AgeFirstKill"])
motive_robbery=c(fortest_mysample[fortest_mysample$Motive == "Robbery or financial gain",
"AgeFirstKill"])

describe(motive_angel)
describe(motive_revenge)
describe(motive_robbery)

############histogram plot with density curve for age at first kill for different type of motives
par(mfrow = c(2, 2))
hist(motive_angel,
    main="Angel of Death",
    xlab="Age",
    ylab="density",
    border="blue",
    col="green",
    xlim=c(0,80),
    las=1,
    density=20,
    prob=TRUE)
curve(dnorm(x, mean=age_fk_mean, sd=age_fk_sd),
     col="darkblue", lwd=2, add=TRUE, yaxt="n")


hist(motive_robbery,
    main="Robbery or financial gain",
    xlab="Age",
    ylab="density",
    border="green",
    col="blue",
    xlim=c(0,80),
    las=1,
    density=20,
    prob=TRUE)
curve(dnorm(x, mean=age_fk_mean, sd=age_fk_sd),
     col="darkblue", lwd=2, add=TRUE, yaxt="n")

hist(motive_revenge,
    main="Revenge or vigilante justice",
```

```
    xlab="Age",
    ylab="density",
    border="brown",
    col="blue",
    xlim=c(0,80),
    las=1,
    density=20,
    prob=TRUE)
curve(dnorm(x, mean=age_fk_mean, sd=age_fk_sd),
     col="darkblue", lwd=2, add=TRUE, yaxt="n")

################Hypothesis testing calculation ########################################
mean_angel=32.59
n_angel=22

mean_robbery=29.20
n_robbery=490

mean_revenge=30.09
n_revenge=57

mu=27
variance=74

### Z score
Z_angel = (mean_angel - mu)/sqrt(variance/n_angel)
Z_robbery = (mean_robbery - mu)/sqrt(variance/n_robbery)
Z_revenge = (mean_revenge - mu)/sqrt(variance/n_revenge)

####CI interval
CI_angel =  mean_angel + c(-1, 1)*1.96*sqrt(variance/n_angel)
CI_robbery =  mean_robbery + c(-1, 1)*1.96*sqrt(variance/n_robbery)
CI_revenge =  mean_revenge + c(-1, 1)*1.96*sqrt(variance/n_revenge)

#####P-value
P_angel=2*pnorm(-abs(Z_angel))
P_robbery=2*pnorm(-abs(Z_robbery))
P_revenge=2*pnorm(-abs(Z_revenge))


##################Independent two samples' calculations ##################################
#####pair1 angel of death and robbery motive
meandiff_pair1=3.39
s1=sd(motive_angel)
s2=sd(motive_robbery)

t=qt(p = 0.975, df = 22 + 490 - 2)
sp=sqrt( ((22 - 1)*s1^2 + (490 - 1)*s2^2)/(490 + 22 - 2)  )
CI_pair1=meandiff_pair1 + c(-1, 1) * t * sp * sqrt(1/22 + 1/490)

#####pair2 angel death and revenge
meandiff_pair2=2.5
s1=sd(motive_angel)
```

```
s2=sd(motive_robbery)

t=qt(p = 0.975, df = 22 + 57 - 2)
sp=sqrt( ((22 - 1)*s1^2 + (57 - 1)*s2^2)/(57 + 22 - 2)  )
CI_pair2=meandiff_pair1 + c(-1, 1) * t * sp * sqrt(1/22 + 1/57)
```

#####pair3 d robbery and revenge motive

```
meandiff_pair3=0.89
s1=sd(motive_robbery)
s2=sd(motive_revenge)

t=qt(p = 0.975, df = 22 + 490 - 2)
sp=sqrt( ((490 - 1)*s1^2 + (57 - 1)*s2^2)/(490 + 57 - 2)  )
CI_pair3=meandiff_pair1 + c(-1, 1) * t * sp * sqrt(1/490 + 1/57)
```