ANALYSIS OF SERIAL KILLERS WITH DIFFERENT MOTIVES

Introduction:

Serial Killers are the people who commit murders in a pattern. The dataset is taken from Radford/FGCU Serial Killer Database - killersandmotives.Rdata. The analysis conducted on this dataset is intended to reveal the underlying patterns and trends between them. The report includes cleaning, exploration, modelling, estimation, hypothesis testing, comparison and interpretation of the dataset. The dataset provided for the analysis consists of 253 rows and nine attributes of the killer as columns. There is a motivation behind every killer. The motives of the killers provided in this dataset are anger, Mental illness and Convenience. The given average age at a first kill is 27. This report summarizes the analysis performed using this dataset.

Objective:

Does the average age at the first kill for each motive differ from the other?

Results:

The data cleaning is the foremost step to detect and remove or impute the outliers or missing values that would cause misleadingness in our analysis. The best way to detect outliers is by using a boxplot [Fig.1] where the plot is not visible due to the outliers.

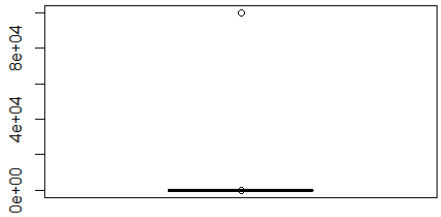

Fig.1 Outliers detection using boxplot

Sometimes the outlier values may be true in reality but the outlier we got at the age at First Kill is 99999 which is not possible. The rows were removed instead of imputing because imputing the values of the age of the first kill will contradict the age of last kill values.

```
KillerID AgeFirstKill  AgeLastKill     YearBorn       Motive          Sex
       0            0            0            0            6            0
    Race     Sentence InsanityPlea
       0           22           10
```

Fig.2 Sum of missing values in each column

The number of missing values are calculated [Fig.2] and the rows are removed as the missing values are in the data type of string. It will be tough to impute the values. The final rows obtained after cleaning the dataset are 213.

There are several ways to explore the data. The easiest way to read and understand the data is to use a graphical summary. Different types of graphs can be created but the best way to determine the shape of an attribute is to plot it in a histogram. The representation of the attributes in the histogram is based on density rather than frequency because the uneven distribution of the widths can affect the shape of the curve.

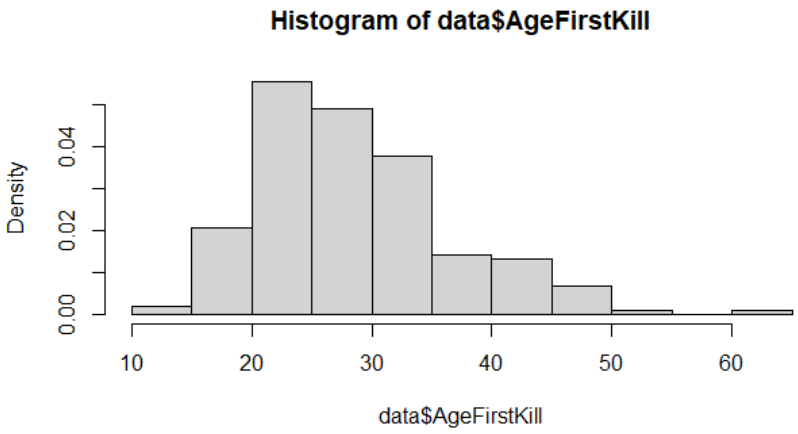

Fig.3 Age of First Kill histogram

The age at first kill between 20 to 25 has the highest density as seen in Fig.3 so most of the killers' first kill is between 20 to 25. The shapes of the age at the first kill histogram plot[Fig.3] to the density looks like a normal distribution with little skewness. The age at the first kill is assumed to be normally distributed.

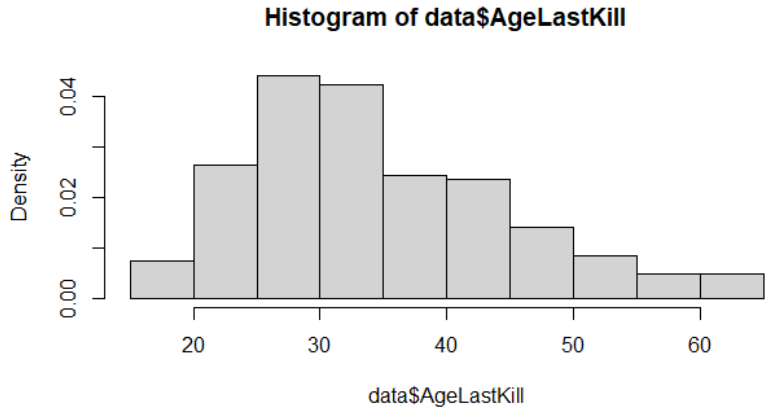

Fig.4 Age of Last Kill histogram

The age at last kill distribution[Fig.4] between 25 to 30 has the highest density so most of the killers' last kill is between 25 to 30. The shape of the age at the last kill histogram[Fig.4] to the density looks like a normal distribution with little skewness.

A new attribute should be created to know the number of years the killers have been killing or it can be said as the career duration of the killer. The career duration of the killer is the difference in age at the first kill and age at the last kill.

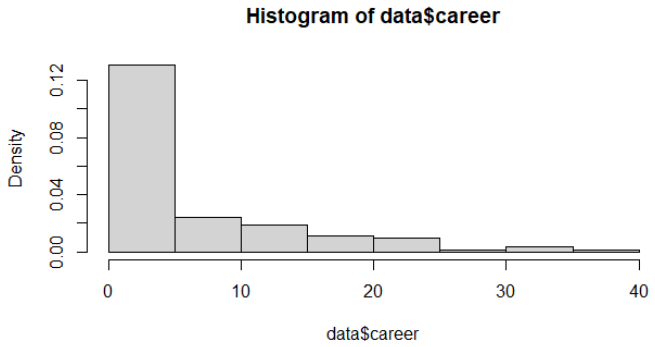

Fig.5 Career duration of the killer

Career duration between 0 to 5 has the highest density as seen in Fig.5 so most of the killers' career duration is less than five years. The shape of the career duration histogram plot looks like an exponential distribution rather than a normal distribution.

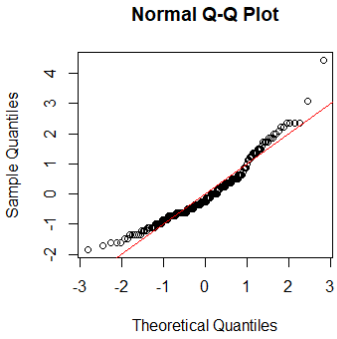

Fig.6a quantile plot age first kill

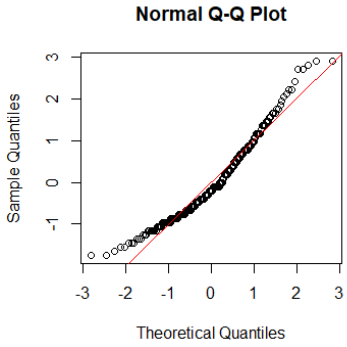

Fig.6b quantile plot age last kill

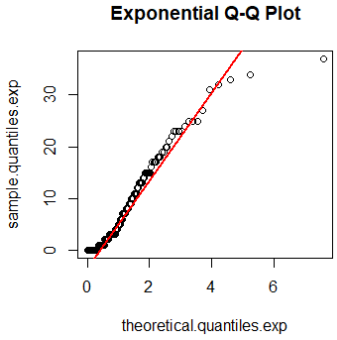

Fig.6c quantile plot career

The quantile plot is a type of graph used to validate our claim that the distribution is either normal or exponential or not. If the plot between theoretical and sample quantiles is a straight line, it is a normal distribution or an exponential distribution.

To validate the claim that the age at first and last kill is normally distributed, the normal quantile plots are used to check the attribute and the plot between theoretical and sample quantiles is assumed to be a straight line for the age at the first kill[Fig.6a] and the age at the last kill[Fig.6b]. The age at the first kill and the last kill is inferred to be normally distributed.

Based on the assumption that career duration is exponentially distributed, the exponential quantile plot[Fig.6c] is used and the plot between theoretical and sample quantiles is assumed to be a straight line. Career duration is inferred to be exponentially distributed.

The next step is to use estimation to calculate the rate of the career duration which is exponentially distributed as the expectation and variance of an exponentially distributed curve depends on the rate. The expectation of career duration is equal to the inverse of the rate. So,

$$E(X)=1/\lambda$$

$$\lambda= 1/E(X)$$

μ= Mean of the career duration

λ=rate of the exponential distribution

E(X)= Expectation of the career duration

The Expectation of the Career duration is assumed to be equal to the mean of the career duration using the method of moments. So,

$$E(X)=\mu$$

$$\lambda=1/\mu$$

Similarly, the estimation for the mean and variance for the normally distributed age at the first kill and the last kill were found.

The next step is hypothesis testing. It is done to estimate the strength of the evidence to support our assumption. The assumption we took here is that the true mean is equal to 27. This assumption is called a null hypothesis(H0). If the null hypothesis is false then we have to reject it and go with an alternative hypothesis(H1).

H0: True mean is equal to 27

H1:True mean is not equal to 27

The dataset is divided into three, based on the motives: anger, mental illness and convenience. The hypothesis testing is done for age at first kill on all three samples of the data.

1) The conditions and the test for motive anger based on the assumptions
    i. The sample size(n) is 183(n>30)
    ii. It is assumed to be normally distributed
    iii. The standard deviation and the population mean are known

The test that was chosen is Z-test due to the above three conditions

2) The conditions and the test for motive mental illness based on the assumptions
    i. The sample size(n) is 20(n<30)
    ii. The standard deviation and the population mean are known
    iii. It is assumed to be normally distributed

The test that was chosen is Z-test though the sample size is less than 30 because of the conditions(ii, iii) and the Z-test is powerful.

3) The conditions and the test for motive convenience based on the assumptions
    i. The sample size(n) is 10(n<30)
    ii. It is assumed to be normally distributed
    iii. The standard deviation and the population mean are known

The test that was chosen is Z-test though the sample size is less than 30 because of the conditions(ii, iii) and the Z-test is powerful.

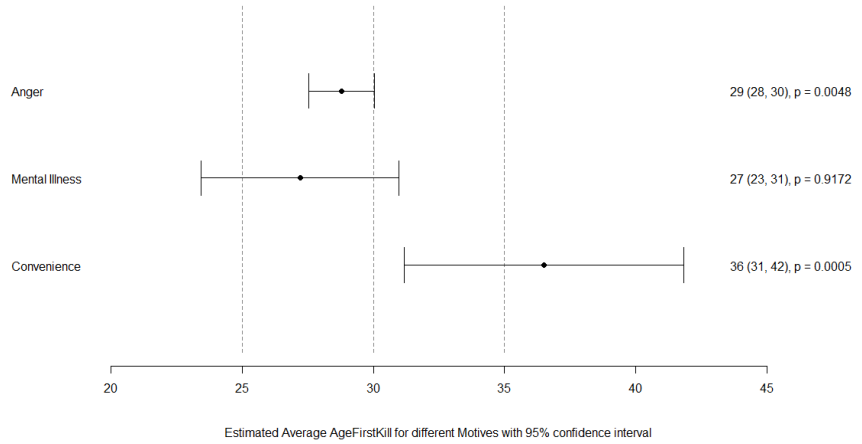

Fig.7 Estimated Average Age at First Kill for different Motives with 95% confidence interval

- The null hypothesis for the motive anger is rejected as the mean value(27) is not in the confidence Interval[Fig.7] and the pvalue< 0.05[Fig.7].
- The null hypothesis for the motive Convenience is rejected as the mean value(27) is not in the confidence Interval[Fig.7] and the pvalue< 0.05[Fig.7].
- We fail to reject the null hypothesis for the motive Mental Illness as the mean value(27) is in the confidence Interval[Fig.7] and the pvalue>0.05[Fig.7].

The next step is the comparison of different motives by finding the difference in their mean values by using the two-sample t-test. The null hypothesis is that the difference in the mean value is zero. We assume that the variances of the two samples are unequal.

$$H0: \mu_1 - \mu_2 = 0$$

$$H1: \mu_1 - \mu_2 = !0$$

μ_1= Mean of motive1

μ_2= Mean of motive2

Motive1	Motive2	P-value
Mental Illness	Convenience	P-value= 7.775e-13<0.05
Anger	Convenience	P-value= 2.2e-16<0.05
Anger	Mental Illness	P-value=0.3661>0.05

Table.1 Two sample t-test with unequal variance

- The P-value of the motives mental illness and convenience is less than 0.05, so we reject the null hypothesis
- The P-value of the motives anger and convenience is less than 0.05, so we reject the null hypothesis
- The P-value of the motives anger and mental illness is greater than 0.05, so we fail to reject the null hypothesis. The is no difference between the mean of the age at the first kill two motives.

The distributions of the attributes are inferred from the graphical. The Z-test concludes that the population mean is not in the confidence interval of the sample mean with the motives anger and convenience but it is in the interval for the motive mental illness. The two-sample t-test concludes that the average age at the first kill differs between the convenience and other the two motives and there is no difference between the motives anger and mental illness.

Discussions:

The hypothesis testing is based on the assumptions that the attribute is normally distributed, sample variance is known. The variances of the two motives are assumed to be unequal in the comparison test with the results as no difference between anger and mental illness. The inferences can vary if the assumptions are different. We can use this dataset to figure out the most common motives of the killers. In future, we can analyse the behaviour of different killers and treat people who show similar behaviour to prevent them from committing a crime.