Name-SAHELI MAZUMDER

**Behaviour of serial killers in terms of various motives**

**Introduction:** -

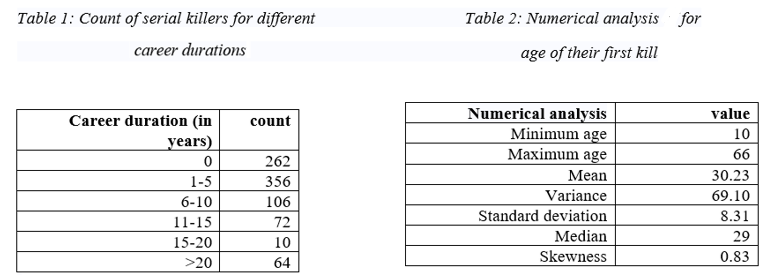
Serial killing is one of the forms of homicide, occurring when an individual has killed three or more people who were previously unknown to him/her, including a significant period of time between them. The variables of interest are age of the first kill by a serial killer and the motive. These variables matter as serial killer classification are mainly done on these two parameters and further analysis of our dataset are also carried forward on basis of them. Dataset “mysample” has 3 below motives-

1. Anger (including mission-oriented killers)
2. Enjoyment or power
3. Escape or avoid arrest

**Results: -**

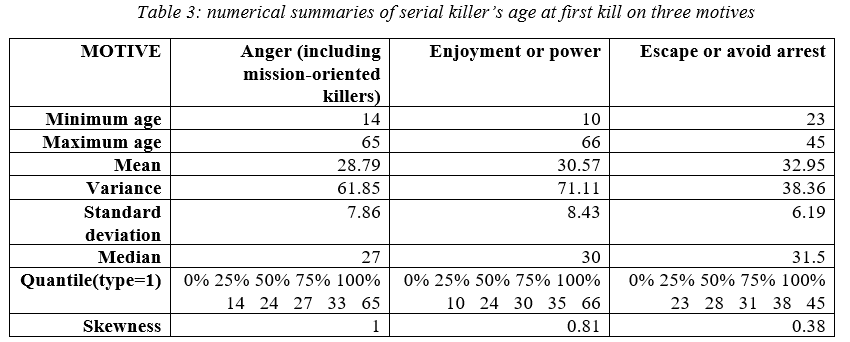
“mysample” had total 954 observations of 9 columns (KillerID, AgeFirstKill, AgeLastKill, YearBorn, Motive,Sex,Race,Sentence,InsanityPlea). There were 78 missing observations for columns “AgeFirstKill” and “Motive”. Also 5 rows contained the data of killers who first killed before year 1900. After removal these values, the sample size got reduced by 8.7%. A new column called “CareerDuration” was added into the dataset which is defined as number of years between the age of each killer (years) when they committed their first and last murder. It had 1 negative value which after removing brought down dataset to size of 870 observations of 10 columns.

**The dataset consists of 847 male and 23 female. Table** 1 **shows the count of serial killers in different range of career duration (in years). Table 2 shows numerical analysis of dataset for age of their first kill.**

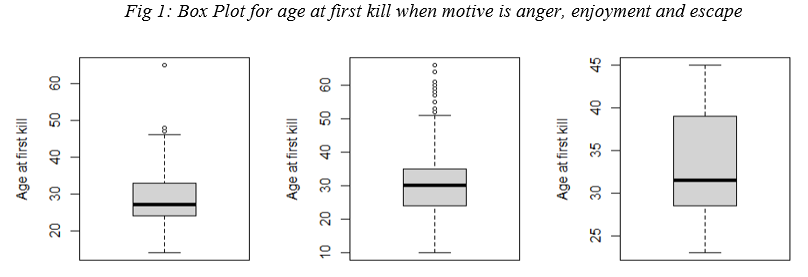


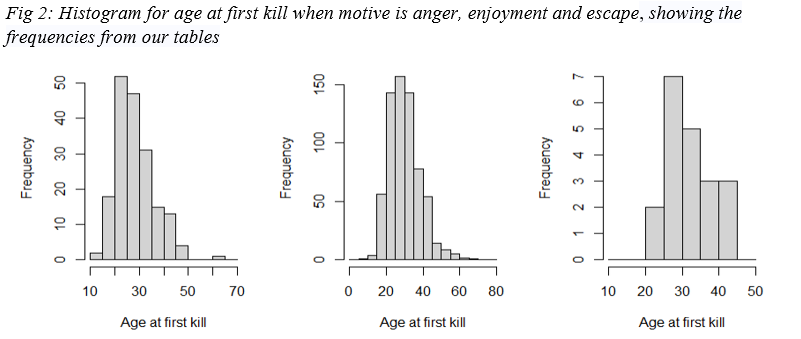
The dataset is split by 3 motives- “Anger (including mission-oriented killers)”,” Enjoyment or power” and “Escape or avoid arrest” and each has size of 183 rows, 667 rows and 20 rows respectively with 10 columns each. Since main research question is based on age at first kill, so below analysis is done on this parameter.

Table 3 suggest that the minimum and maximum age for motive **Escape or avoid arrest** is totally different from the other motives and has a lower variance value due to least number of observations. It means that the values are more consistent and that the data points tend to be very close to the mean, and to each other. For the three motives, data are positively skewed because more people did their first kill at a younger age. But this would not discourage us from using a normal distribution as our model for the motives as the value is not large.

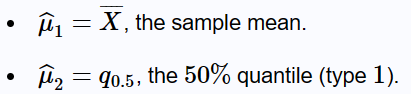


For graphical summaries, box plot and histogram are used.

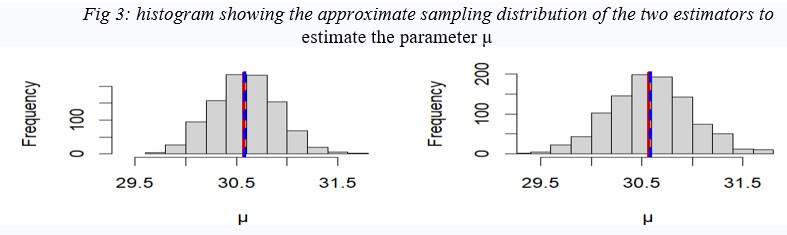
Fig 1 shows that the “maximum” in boxplots is 48,51 and 39 respectively rather than the actual maximum 65, 66 and 45. This is because 65, 66 and 45 lies more than 1.3 IQRs from the box and are therefore judged to be an outlier by R.

**In Fig 2, frequency is the count of serial killers who committed their first murder of different age groups. According to the histograms, maximum count of killers is of the age group 20-30, 25-30 and 25-30 respectively for motives *anger, enjoyment and escape*. For motive anger, distribution is right-skewed. This means the data’s mean value is around the beginning of the data range. For other two, distributions are almost symmetric, which means the mean of the distribution is around the mid-value of the data set. Based on graphical summaries, normal distribution is proposed to model age at first kill for three motives.

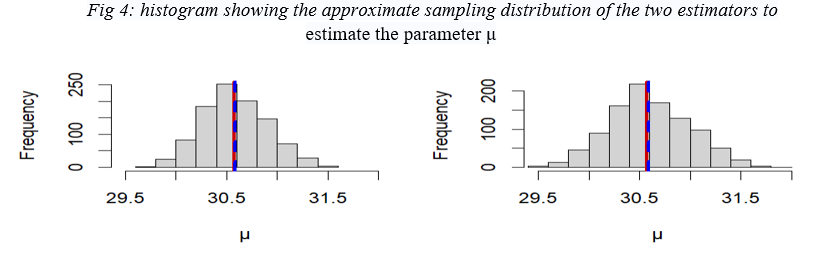
To estimate the parameter μ for age at first kill, it is considered two estimators:



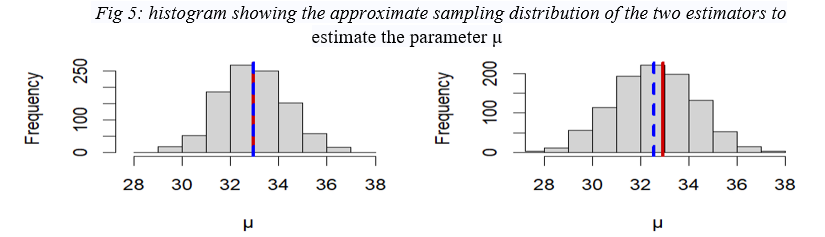
**For motive Anger,** above two estimators are compared by simulating dataset with mean=28.79 and variance = 61.85.

****

**For motive Enjoyment or power**: -above two estimators are compared by simulating dataset with mean=30.57 and variance = 71.11.

****

**For motive Escape or avoid arrest:** above two estimators are compared by simulating dataset with mean=32.95 and variance = 38.36.

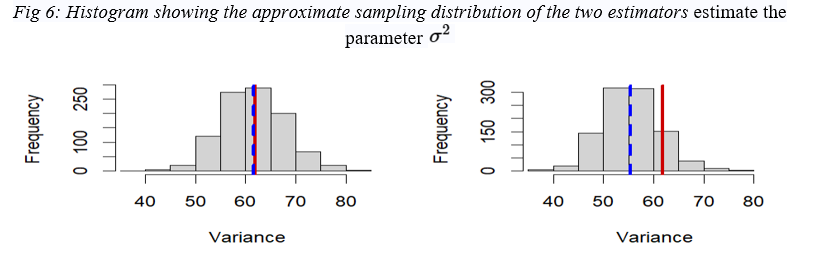


Red line is showing the "true" value ofμ and blue line shows the average of the estimator. In Fig 3 and Fig4, mean is unbiased (as there is no difference between the red and blue lines). In Fig 5 sample mean appears to be biased. The average of this estimator appears to be below the true value of  (blue line is to the left of red line), suggesting a negative bias.

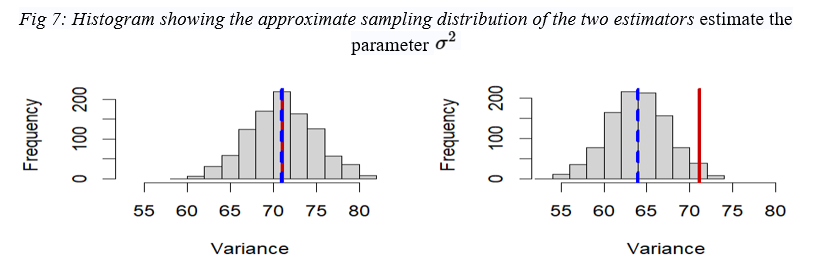
To estimate the parameter for age at first kill, it is considered two estimators: 



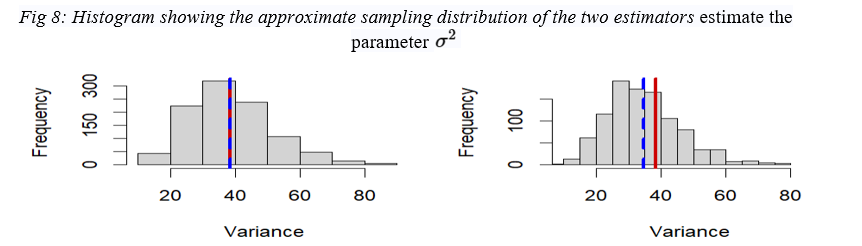
**For motive Anger: -**



**For motive Enjoyment or power: -**

****

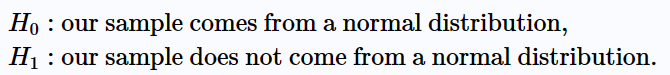
**For motive Escape or avoid arrest: -**



Red line showing the "true" value ofand blue line shows the average of the estimator. In Fig 6,7,8 sample variance appears to be biased for the second estimator with a tendency to underestimate . The average of this estimator appears to be below the true value of  (blue line is to the left of red line), suggesting a negative bias.

**For motive Anger (including mission-oriented killers):**

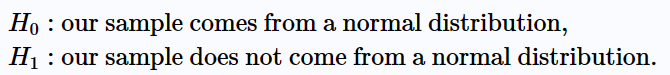
To check the proposed model, it is considered that



Test would have a 5% significance **level (p-value)** that is if   is true, there is less than a 5% risk of **incorrectly**rejecting it. For Quantile-quantile (Q-Q) plots Fig 9 , points almost lie in **a straight line** then. Hence the sample data does have normal distribution.

**For motive Enjoyment or power: -**

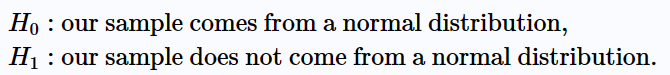
To check the proposed model, it is considered that



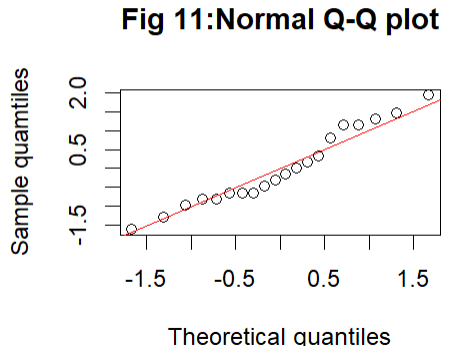
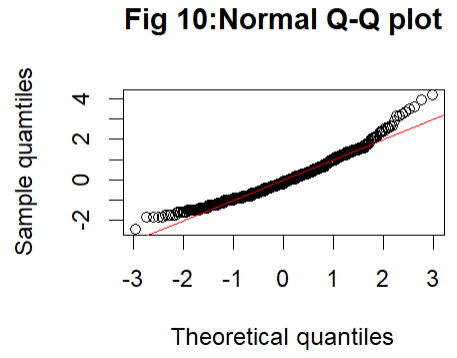
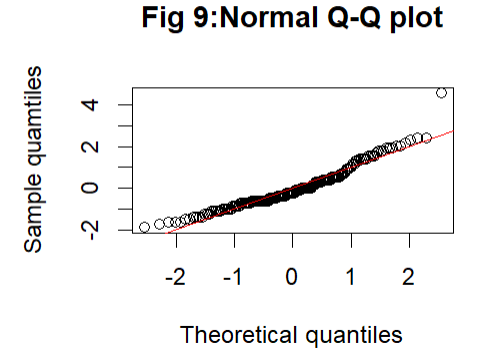
Test would have a 5% significance **level (p-value)** that is if   is true, there is less than a 5% risk of **incorrectly**rejecting it. For Quantile-quantile (Q-Q) plots Fig 10, points almost lie in **a straight line** and the sample size is large. Hence the sample data does have normal distribution.

**For motive Escape or avoid arrest: -**

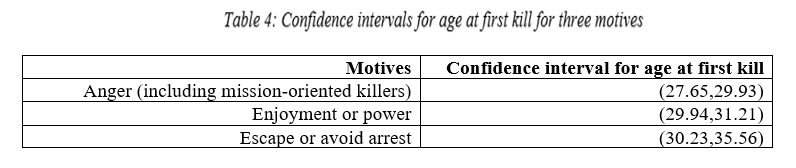
To check the proposed model, it is considered that



Test would have a 5% significance **level (p-value)** that is if   is true, there is less than a 5% risk of **incorrectly**rejecting it. In Kolmogorov-Smirnov test, D = 0.13817, p-value = 0.8397 that is p>5%. For Quantile-quantile (Q-Q) plots Fig 11, points lie in **a straight line**. Hence the sample data have normal distribution.



For Escape, sample size is small so performing t-tests. For other two motives , sample size is large and variance of age at first kill is known so Z-test is used to get the confidence intervals.

**

This means that there is a 95% probability the above-mentioned confidence interval will contain the true population mean for each of the motives.

To test whether the mean is 27 years, it can be stated from the above table that it does not lie in confidence intervals in any of the motives.

***Fig 12: A caterpillar/forest plot of three 95%-confidence intervals for three different motives comparing mean of age of first kill***

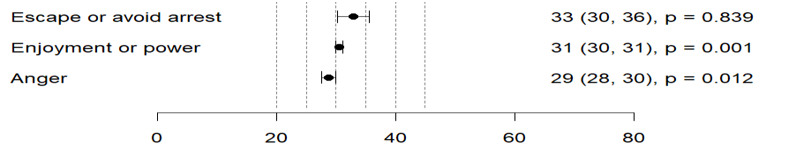


Fig 12 presents the results of motives comparing estimated difference in means of age at first kill with 95% confidence interval. For each motive on the right side of figure, mean of the sample is mentioned along with confidence intervals and p-value.

**Conclusion: -**

The above analysis suggests that **the average age at first murder differ between killers with different motives. Serial killers did their first murder at a very young age for purpose of enjoyment than those motivated by anger or escape. Age at first kill follows normal distribution for all motives. The limitation is that there is large difference of dataset sizes among the motives which may** mislead to summaries.

## APPENDIX

R codes that are used: -

load("killersandmotives.Rdata")

createsample(201678825)

#to get which row numbers has agefirstkill as 99999

which(mysample$AgeFirstKill=="99999")

#to get which row numbers has motive (missing motives are recorded as NA)

mysample<-na.omit(mysample)

#Removed the above rows

mysample <- mysample[mysample$AgeFirstKill != "99999", ]

#rows containing the data of killers who first killed before the year 1900

mysample[(mysample$YearBorn + mysample$AgeFirstKill) < 1900 ,]

#removed any rows containing the data of killers who first killed before the year 1900 and kept the remaining rows

mysample<-mysample[(mysample$YearBorn + mysample$AgeFirstKill) >= 1900 ,]

names(mysample)

#CareerDuration variable is created

mysample$CareerDuration <- (mysample$AgeLastKill - mysample$AgeFirstKill)

#careerduration is negative

mysample[mysample$CareerDuration < 0 ,]

#careerduration all positive

mysample<-mysample[mysample$CareerDuration >= 0 ,]

#male and female count

sum(mysample$Sex == "Female")

sum(mysample$Sex == "Male")

#Numerical analysis of mysample data

quantile(mysample$CareerDuration,type=1)

skewness(mysample$AgeFirstKill)

#Splitting into 3 different dataset as per motives

Escape <- mysample[mysample$Motive == 'Escape or avoid arrest', ]

Enjoyment\_power <- mysample[mysample$Motive == 'Enjoyment or power', ]

Anger <- mysample[mysample$Motive == 'Anger (including mission-oriented killers)', ]

#Numerical analysis of data for motive Anger

min(Anger$AgeFirstKill)

max(Anger$AgeFirstKill)

mu1 <- mean(Anger$AgeFirstKill)

var(Anger$AgeFirstKill)

sigma1 <- sd(Anger$AgeFirstKill)

quantile(Anger$AgeFirstKill, type=1)

median(Anger$AgeFirstKill)

skewness(Anger$AgeFirstKill)

#Numerical analysis of data for motive Enjoyment\_power

min(Enjoyment\_power$AgeFirstKill)

max(Enjoyment\_power$AgeFirstKill)

var(Enjoyment\_power$AgeFirstKill)

quantile(Enjoyment\_power$AgeFirstKill, type=1)

median(Enjoyment\_power$AgeFirstKill)

skewness(Enjoyment\_power$AgeFirstKill)

mu2 <- mean(Enjoyment\_power$AgeFirstKill)

sigma2 <- sd(Enjoyment\_power$AgeFirstKill)

#Numerical analysis of data for motive Escape

min(Escape$AgeFirstKill)

max(Escape$AgeFirstKill)

var(Escape$AgeFirstKill)

quantile(Escape$AgeFirstKill, type=1)

median(Escape$AgeFirstKill)

skewness(Escape$AgeFirstKill)

mu3 <- mean(Escape$AgeFirstKill)

sigma3 <- sd(Escape$AgeFirstKill)

#Boxplots

par(mfrow = c(1, 3))

boxplot(Anger$AgeFirstKill,ylab = "Age at first kill",main = "Box Plot for age at first kill when motive is anger")

boxplot(Enjoyment\_power$AgeFirstKill,ylab = "Age at first kill",main = "Box Plot for age at first kill when motive is Enjoyment\_power")

boxplot(Escape$AgeFirstKill,ylab = "Age at first kill",main = "Box Plot for age at first kill when motive is Enjoyment\_power")

#Histograms

par(mfrow = c(1, 3))

hist(Anger$AgeFirstKill,xlab="Age at first kill",main="Histogram of age at first kill for motive 'Anger (including mission-oriented killers)'",breaks = seq(from = 10, to = 70, by = 5))

hist(Enjoyment\_power$AgeFirstKill,xlab="Age at first kill",main="Histogram of age at first kill for motive 'Enjoyment or power'",breaks = seq(from = 0, to = 80, by = 5))

hist(Escape$AgeFirstKill,xlab="Age at first kill",main="Histogram of age at first kill for motive 'Escape or avoid arrest'",breaks = seq(from = 10, to = 50, by = 5))

##Q-Q plot of data for age at first kill for motive anger

z <- (sort(Anger$AgeFirstKill) - mu1)/sigma1

n1 <- length(Anger$AgeFirstKill)

r <- (1:n1)

q <- qnorm(p = r/(n1 + 1), mean = 0, sd = 1)

plot(q, z,ylab = "Sample quamtiles",xlab="Theoretical quantiles",main = "Fig 9:Normal Q-Q plot")

abline(a = 0, b = 1, col = "red")

xbar1 <- mean(Anger$AgeFirstKill)

# Calculate the confidence interval:

CI1 = xbar1 + c(-1, 1)\*1.96\*sqrt(sigma1^2/183)

CI1

##Q-Q plot of data for age at first kill for motive Enjoyment\_power

z <- (sort(Enjoyment\_power$AgeFirstKill) - mu2)/sigma2

n2 <- length(Enjoyment\_power$AgeFirstKill)

r <- (1:n2)

q <- qnorm(p = r/(n2 + 1), mean = 0, sd = 1)

plot(q, z,ylab = "Sample quamtiles",xlab="Theoretical quantiles",main = "Fig 10:Normal Q-Q plot")

abline(a = 0, b = 1, col = "red")

xbar2 <- mean(Enjoyment\_power$AgeFirstKill)

# Calculate the confidence interval:

CI2 = xbar2 + c(-1, 1)\*1.96\*sqrt(sigma2^2/667)

CI2

#K-s test for age at first kill for motive Escape

ks.test(x = Escape$AgeFirstKill, y = "pnorm", mean = mu3, sd = sigma3)

##Q-Q plot of data for age at first kill for motive Escape

z <- (sort(Escape$AgeFirstKill) - mu3)/sigma3

n3 <- length(Escape$AgeFirstKill)

r <- (1:n3)

q <- qnorm(p = r/(n3 + 1), mean = 0, sd = 1)

plot(q, z,ylab = "Sample quamtiles",xlab="Theoretical quantiles",main = "Fig 11:Normal Q-Q plot")

abline(a = 0, b = 1, col = "red")

t.test(Escape$AgeFirstKill, alternative = "two.sided", mu = mu3, conf.level = 0.95)

#estimating of mu:-

muhat1 <- rep(NA, 1000)

muhat2 <- rep(NA, 1000)

for(i in 1:1000){

x <- rnorm(n = 20, mean = mu3, sd = sigma3)

muhat1[i] <- mean(x)

muhat2[i] <- quantile(x, type = 1)[3]}

par(mfrow = c(1, 2))

# A histogram showing the approximate sampling distribution of the first estimator:

hist(muhat1, xlim = range(c(muhat1, muhat2)),xlab="μ")

# A line showing the "true" value of mu:

abline(v = mu3, col = "red3", lwd = 3)

# A dashed (lty = 2) line showing the average of the first estimator for mu:

abline(v = mean(muhat1), col = "blue", lty = 2, lwd = 3)

# the second estimator of mu:

hist(muhat2, xlim = range(c(muhat1, muhat2)),xlab="μ")

abline(v = mu3, col = "red3", lwd = 3)

abline(v = mean(muhat2), col = "blue", lty = 2, lwd = 3)

#Estimating of sigma^2

sigma2hat1 <- rep(NA, 1000)

sigma2hat2 <- rep(NA, 1000)

for(i in 1:1000){

x <- rnorm(n = 20, mean = mu3, sd = sigma3)

sigma2hat1[i] <- sd(x)^2

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

}

par(mfrow = c(1, 2))

hist(sigma2hat1, xlim = range(c(sigma2hat1, sigma2hat2)),xlab="Variance")

abline(v = sigma3^2, col = "red3", lwd = 3)

abline(v = mean(sigma2hat1), col = "blue", lty = 2, lwd = 3)

hist(sigma2hat2, xlim = range(c(sigma2hat1, sigma2hat2)),xlab="Variance")

abline(v = sigma3^2, col = "red3", lwd = 3)

abline(v = mean(sigma2hat2), col = "blue", lty = 2, lwd = 3)

##CI plot long with Estimated difference in means, with 95% confidence interval

analysis = c("Anger", "Enjoyment or power", "Escape or avoid arrest" )

estimate = c(28.79, 30.57, 32.95)

upper = c(29.93, 31.21, 35.66)

lower = c(27.65, 29.93, 30.23)

pval = c(0.012,0.001,0.839)

par(mar = c(6,6,1,6))

plot(x = 0, xlim = c(0, 60), ylim=c(0, 4),type = "n", xaxt = "n", yaxt="n", xlab = NULL, ylab= NULL, ann = FALSE, bty="n")

axis(side = 1, cex.axis = 1)

mtext("Estimated difference in means, with 95% confidence interval", side = 1, line = 4)

for(i in c(20,25,30,35,40,45)){

lines(c(i, i), c(0, 5), lty = 2, col = "gray53") }

verticalpos = 1:3

mtext(text = analysis, at = verticalpos,

side = 2, line = 5, outer = FALSE, las = 1, adj = 0)

points(estimate, verticalpos, pch = 16)

for(i in 1:3 ){

lines(c(lower[i], upper[i]), c(verticalpos[i], verticalpos[i]))

lines(c(lower[i], lower[i]), c(verticalpos[i] + 0.2, verticalpos[i] - 0.2))

lines(c(upper[i], upper[i]), c(verticalpos[i] + 0.2, verticalpos[i] - 0.2))}

est <- formatC(estimate, format='f', digits = 0)

P <- formatC(pval , format = 'f', digits = 3)

pval <- paste("p =", P)

L <- formatC(lower, format = 'f', digits = 0)

U <- formatC(upper, format = 'f', digits = 0)

interval <- paste("(", L, ", ", U, "),", sep = "")

results <- paste(est, interval, pval)

mtext(text = results, at = verticalpos, side = 4, line = 4, outer = FALSE, las = 1, adj = 1)

box("inner")