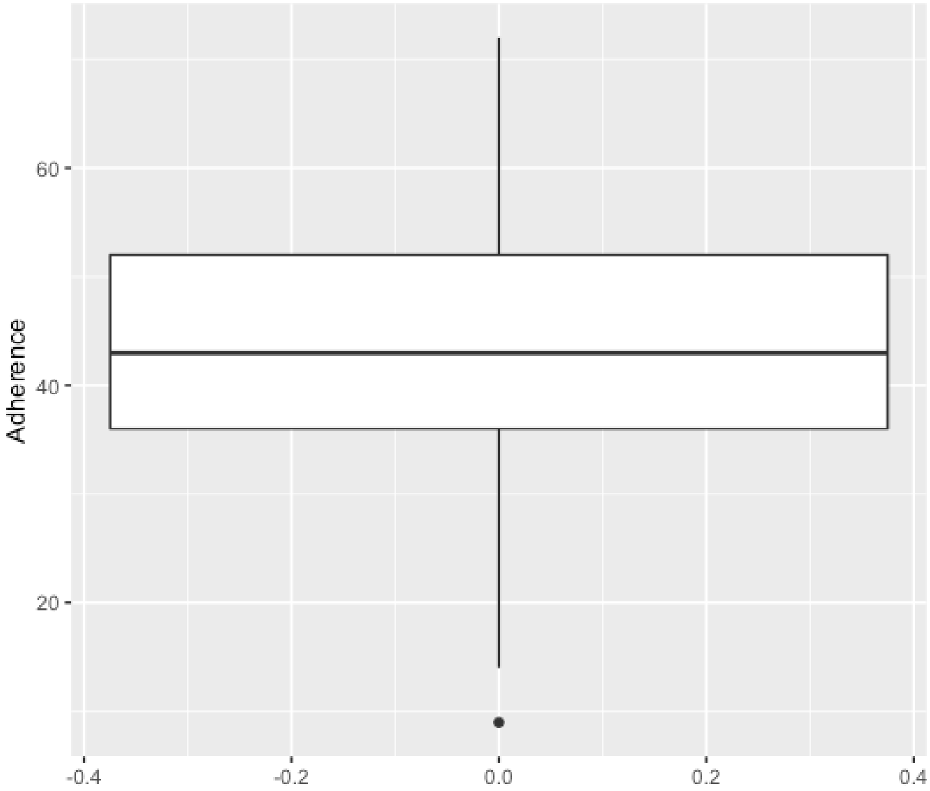
R Assignment

Question 1
A randomised controlled trial was designed to assess the impact of an educational programme on patients' adherence to antihypertensive drugs. A sample of 190 hypertensive patients were randomly allocated to receive either special care (the educational programme) or their usual care. Each patient's adherence to their antihypertensive drugs was recorded for six weeks. An adherence score was then calculated for each patient. This score ranged from a possible minimum of 0 (complete non-adherence) to a possible maximum of 100 (complete adherence). The collected data are stored in the file "adherenceScore.csv". Variable Description Care Group (1 = Special Care, 2 = Usual Care) Age (in years) Gender (1 = Male, 2 = Female) Family Support (1 = Yes, 2 = No) Adherence score from 0 to 100.

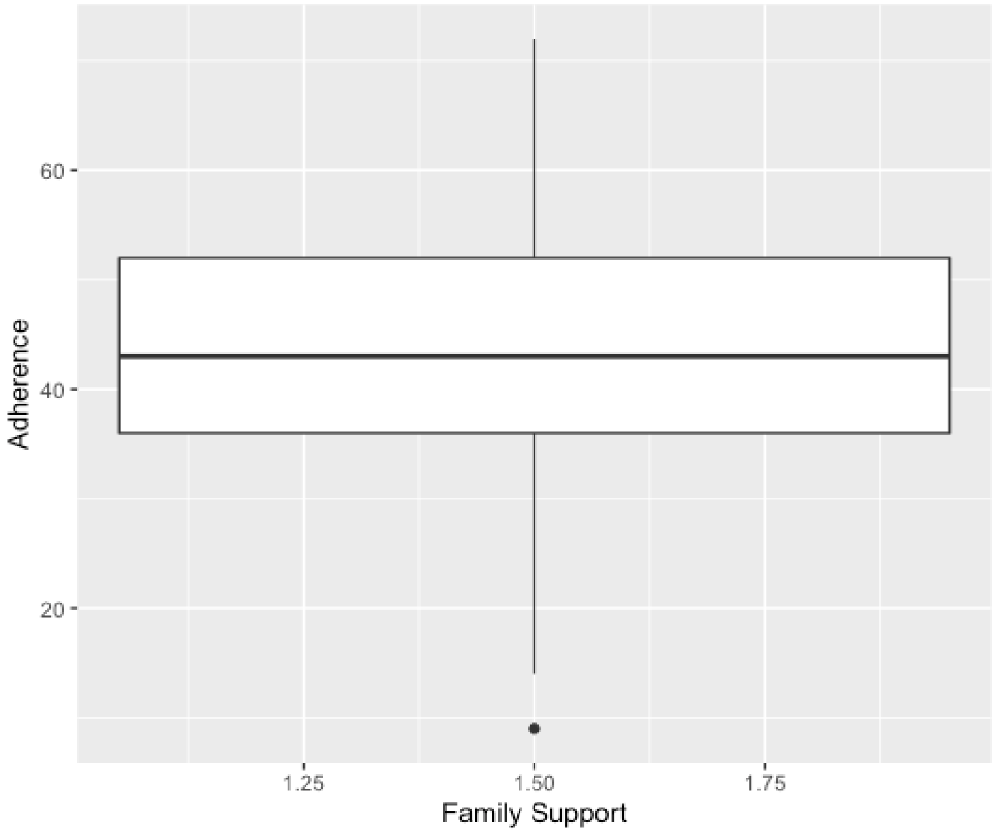
(a) Present the boxplot for adherence score.



(b) Describe the shape of adherence score.
The boxplot of adherence score is Right Skewed (skewedness is positive) since the median is located at the lower end of the box rather than the middle and the upper end is longer than the shorter end.

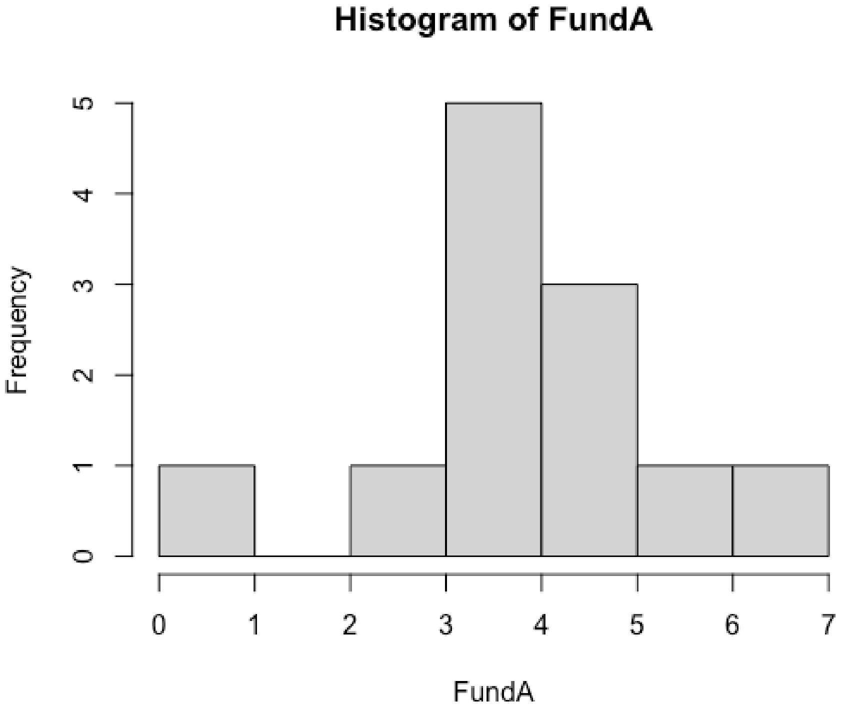
(c) Is there any outlier for adherence score?
Yes there is and the outlier for adherence score is shown by a small black dot at the end of the lower part.

(d) Present the boxplot of adherence score by family support.



Question 2
An analyst records a random sample of the annual percentage returns on Fund A and Fund B as follows: Fund A (%): {0.5, 3.9, 6.3, 3.5, 4.2, 4.1, 4.1, 5.8, 3.2, 2.8, 3.3, 3.2}
Fund B (%): {6.5, 4.9, 4.3, 6.0, 5.6, 4.8, 6.0, 5.9, 6.4, 6.7}

(a) Use R to construct a histogram of the first dataset and copy this into a Word file. Based on the histogram do you think this dataset is drawn from a Normal population? Explain.



Yes, I do think that the dataset is derived from the normal population as the normally distributed variables have bell-shaped histograms with a single peak usually at the median values. It is possible to say that the above histogram is drawn from a normal distribution as it is close to having a bell-shape, but a histogram looks different depending on the number of points and bars, so it is more difficult to interpret when the sample size is small. If there are fewer than 20 data points, the bars on the histogram are inadequate for showing the distribution.

(b) Use R to calculate the sample mean, median, standard deviation and interquartile range for the second dataset. Then, compare the mean to the median. Is there evidence from these values that the dataset is drawn from a Normal population? Use a 5% level of significance and carry out the Shapiro test of Normality in R to investigate this point further.

For Funds B:-
Inserting Fund B into a dataframe:-
FundB <- c(6.5,4.9,4.3,6.0,5.6,4.8,6.0,5.9,6.4,6.7)
new_datb <- data_frame(FundB=FundB)
view(new_datb)

Mean for Funds B =
mean(new_datb$FundB)
= 5.71
Median for Funds B=
Median(new_datab$FundB)
= 5.95
Standard Deviation for Funds B= sd(FundB)

0.8006248

Interquartile range for Funds B= IQR(FundB)

1.225

Hence, we can say that the dataset is drawn from a normal population since the mean and median are approximately equal.

Shapiro Normality test-

Significance level – 0.05

```
shapiro.test(new_datb$FundB)
```

```
Shapiro-Wilk normality test data:  new_datb$FundB
W = 0.92561, p-value = 0.4061
```

As the p-value (0.4061) >significance level (0.05) we can conclude that dataset is drawn from normal population.

(c) Use R to carry out a 1 sample t-test to determine whether the annual percentage return from Fund A exceeds 3%. Use a 5% level of significance and state the appropriate null and alternative hypotheses applying.

Null Hypothesis: $H_0 : \mu =3$

Alternative Hypothesis $H_A : \mu>3$

Significance Level – 5%

T-test code :-

```
t.test(FundA , mu=3 , alternative = "greater", conf.level = 0.95,)
```

Result:-

```
One Sample t-test

data:  FundA
t = 1.7574, df = 11, p-value = 0.0533
alternative hypothesis: true mean is greater than 3
95 percent confidence interval:
 2.98377     Inf
sample estimates:
mean of x
 3.741667
```

Conclusion :- We do not have enough evidence to reject null hypothesis because P-value is larger than 0.05.

(d) Use R to carry out a 2-sample unpaired t-test to determine whether the annual percentage return from Fund A is more than 1 percentage point lower from that of Fund B. Use a 5% level of significance and state the appropriate null and alternative hypotheses applying.

Null Hypothesis : $H_0 : \mu_A - \mu_B =1$

Alternative Hypothesis : $H_{A\,:}$ $\mu_A - \mu_B > 1$

T-test code :-

```
t.test(FundA,FundB , mu=1 , alternative = "greater", conf.level = 0.95)
```

Result:-

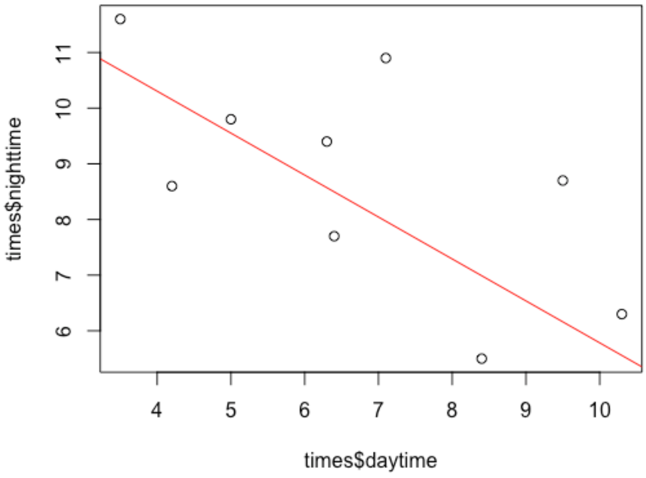
```
Welch Two Sample t-test

data:  FundA and FundB
t = -6.0315, df = 17.562, p-value = 1
alternative hypothesis: true difference in means is greater than 1
95 percent confidence interval:
 -2.822885       Inf
sample estimates:
mean of x mean of y
 3.741667  5.710000
```

Conclusion :- We do not have enough evidence to reject null hypothesis because P-value is larger than 0.05.

Question 3

A sample of dolphins was selected at random. For each dolphin, the number of hours spent in a unihemispheric sleep state during a day (8:00am to 8:00pm) and during the following night (8:00 pm to 8:00am) was recorded. The data is presented below:
Daytime (x) 6.3 4.2 10.3 7.1 8.4 9.5 5.0 3.5 6.4
Night-time (y) 9.4 8.6 6.3 10.9 5.5 8.7 9.8 11.6 7.7

(a) Draw a scatter plot of night-time against daytime. Interpret the plot.



According to the above scatterplot, there is a negative correlation between daytime values and night-time values.

(b) Calculate the correlation coefficient and interpret its value.

```
cor(times$daytime,times$nighttime) =   -0.6463796
```

An inverse relationship exists between two independent variables when there is a negative correlation coefficient value, for example, when the value of daytime increases, the value of nighttime decreases. The closer the value of correlation coefficient to -1, stronger the inverse relationship. The correlation coefficient value of r = -0.6463796 indicates a strong relationship.

(c) We want to see the effect of daytime sleep on the night-time sleep. Find the equation of the linear regression line. Interpret the slope and the intercept of this line.

Equation of linear regression line is :-

$$y_i = \beta_0 - \beta_1 x_i + \varepsilon$$

where β_1 = Slope of the line

β_0 = Intercept when x=0

ε = Residual Error

Calculation of coefficients to obtain linear regression line:-

```
lmtime= lm(formula = nighttime~daytime,data=times)
summary(lmheight)
```

Output:-

```
  Call:
lm(formula = nighttime ~ daytime, data = times)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3046 -1.2131  0.1109  1.0796  2.3748 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.4603     1.7541   7.103 0.000193 ***
daytime      -0.5542     0.2473  -2.241 0.059961 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.63 on 7 degrees of freedom
Multiple R-squared:  0.4178, Adjusted R-squared:  0.3346 
F-statistic: 5.023 on 1 and 7 DF,  p-value: 0.05996
```

Hence, $\beta_0 = 12.4603$ and $\beta_1 = -0.5542$

Regression line : $y_i = 12.4603 - (-0.5542)x_i + 1.63$

(d) Calculate the coefficient of determination. Interpret this value.
Coefficient of Determination (Multiple R-squared) = 0.4178
This indicates that 41.78% of the data fits into the regression model/line or 41.78% of variance in dependent variable night-time(y) can be explained by the independent variable daytime(x) with help of the regression line.

(e) Using the regression line from (iii) estimate the number of nighttime hours in a unihemispheric sleep state for a dolphin with 6 daytime hours.

```
predict(lmtime,data.frame(daytime=6))
```

Linear Regression Equation :- yi= 12.4603 – (-0.5542)xi
= 9.13483hrs

(f) Calculate the residual for X = 8.4

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.3046 -1.2131  0.1109  1.0796  2.3748 
```