Task1

Data and research question:

The provided data contained measurements of mushroom yield (kg of fresh mushrooms per bag) of 3 different mushroom strains (H35, L357, P70). Each strains have relatively large sample size (40 replicate/strain). The question was is there any yield specific differences between the 3 strains.

Interpreting data:

The data analysed by R (R 4.1.3). After loading the data, the groups are visualized on boxplot (*Figure1*).

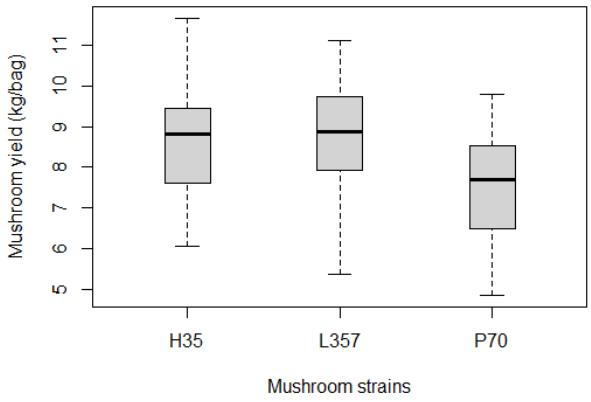

Figure1: Show the 3 strains of mushroom by their mushroom yields.

For comparing the three groups ANOVA testing was used. All the assumption for the test was meet. All the groups had normal distribution (checked by shapiro.test (p-value>0.05) and all the data had homogeneity of variance (checked by barlett.test (p-value = 0.9449 > 0.05) and leveneTest (Pr(>F) = 0.9801 > 0.05)). The null hypothesis (H0) was that all the group means were equal which was rejected by the test result (p-value < 0.05). The alternative hypothesis(H1) was accepted, which means at least one of the group is differ from the others. For checking which groups are differ, posthoc analysis was applied (Tukey's posthoc, conf.level=0.95).

Results:

For the analysis of variance in ANOVA (*Table1*), the Df was 2 for variables and 117 for residuals. The Sum Sq for variables was 33.917, showing the variation between groups, while Sum Sq of the residuals was 215.135, which shows the variation attributed to the error. The Mean Sq (sum sq/df) was 16.9585 for variables and 1.8388 for the residuals, these shows the unbiased estimate of the variation. The F value was 9.2228, which shows the variation within the group (smaller value means smaller variation) and this F statistics showed statistically significant (Pr(>F) is 0.001908<0.05) result which means the variance of the group means are different. Therefore, we do not assume equal variances.

```
Analysis of Variance Table

Response: Mushroom_yield
                Df  Sum Sq Mean Sq F value    Pr(>F)
Mushroom_strain   2  33.917 16.9585  9.2228 0.0001908 ***
Residuals       117 215.135  1.8388
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table1: Result table of analysis of variance of ANOVA (R)

In ANOVA one-way analysis of means (not assuming equal variances) (*Table2*) resulted statistically significant difference (p-value is 0.0002635<0.05) at least one in the groups mean. (F = 9.1774, num df = 2.000, denom df = 77.963)

```
One-way analysis of means (not assuming equal variances)

data:  Mushroom_yield and Mushroom_strain
F = 9.1774, num df = 2.000, denom df = 77.963, p-value = 0.0002635
```

Table2: One-way analysis of means of ANOVA (R)

The posthoc test (*Table3*) suggested that the P70 strain statistically significantly differ from the other two (H35, L357) group (p adj is 0.0016860 < 0.05 and 0.0005049 <0.05), and no statistically significantly difference between the H35, L357 groups (p adj is 0.9363689 > 0.05).

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Mushroom_yield ~ Mushroom_strain, data = mushrooms)

$Mushroom_strain
             diff        lwr        upr     p adj
L357-H35  0.10475 -0.6150511  0.8245511 0.9363689
P70-H35  -1.07175 -1.7915511 -0.3519489 0.0016860
P70-L357 -1.17650 -1.8963011 -0.4566989 0.0005049
```

Table3: Tukey posthoc test results

Task2

Data and research question:

The provided data contained the stature (cm) and femoral length(mm) of 100 individuals from an osteological collection. The question was what the degree of linear association between the length of the femur and stature is.

Interpreting data:

The data was analyzed by R (R 4.1.3). After loading the data, the groups are visualized on a scatterplot (*Figure2*).

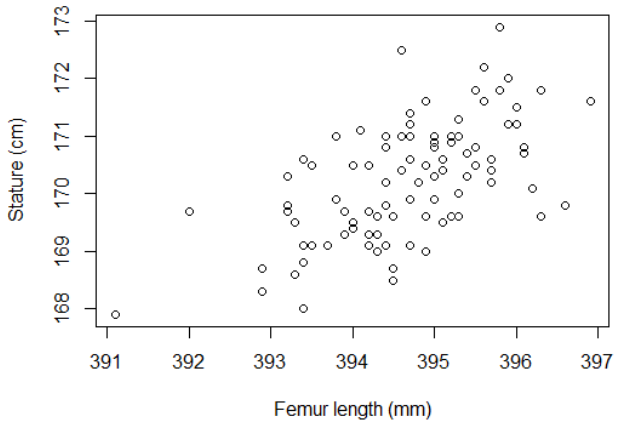

Figure2: Plot of correlation of Femur length (mm) and Stature (cm).

Correlation testing was applied (Pearson correlation). The hypothesis was: H0, there isn't linear correlation and H1, there is linear correlation The H1 was accepted, the data had normal distribution (checked by shapiro.test, p-value >0.05).

Result:

The correlation coefficient (degree of linear association) was 0.5966, which suggested a moderate positive correlation between femoral length and stature. (t = 7.3604, df = 98, p-value = 5.718e-11)

Evaluate linear model:

With the data, a linear model was built for predicting the stature by femur length, by using the lm() function. The model equation: E(Stature)=E(0.59580)Femur-E(64.93339). "E" is stands for estimated. The summary of the model is on *Table4*. The visualized model is on *Figure3*. The Femur has statistical significance to the model (Pr(>|t|) = 5.72e-11), therefore, can be used as a predictor. The result proposed that the model has statistical significance and performs better than expected by chance (p-value: 5.718e-11 < 0.05 and F-statistic: 54.17 on 1 and 98 DF). The regression model accounts for 35.6% of the variability of outcome measures (Multiple R-squared: 0.356).

Table4: Summary of the linear model (R)

```
Call:
lm(formula = Stature ~ Femur, data = ost_collection)

Residuals:
     Min       1Q   Median       3Q      Max
-1.60799 -0.51383 -0.05799  0.51642  2.33243

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.93339   31.94494  -2.033   0.0448 *
Femur         0.59580    0.08095   7.360 5.72e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8074 on 98 degrees of freedom
Multiple R-squared:  0.356,    Adjusted R-squared:  0.3494
F-statistic: 54.17 on 1 and 98 DF,  p-value: 5.718e-11
```

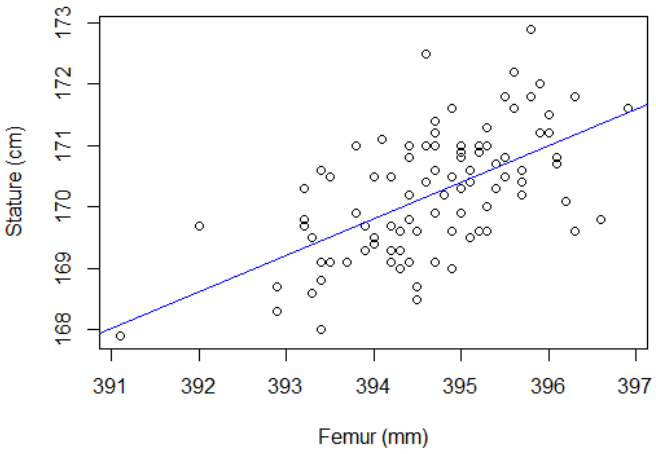

Figure3: Linear model for predict stature from femoral length.

Residual analysis:

The residuals analysed for linearity (*Figure4*), homoscedasticity *(Figue5*) and normality (*Figure6*). For normality shapiro.test also applied (p-value> 0.05).

On *Figure4* we can see that the spots scattered around the residual=0 in random distribution (y-axis: residuals, x-axis: fitted values). The red line shows how close the residuals fit to the horizontal line (residual=0). This shows that it has linear distribution.

On *Figure5* we can see the spots has a similar pattern as on Figure4. Here the plot shows the absolute value of the residuals (y-axis) and make all the y-values positive. The random distribution we can see, suggest there is no relationship between the predicted y values and residuals. The red line shows the equality of distribution (randomness), which close to fit to the horizontal line, which means the homoscedasticity is not violated.

On *Figure6* we can see the QQ plot for normality. On the y-axis: actual residuals, on the x-axis: predicted residuals. We can see that the spots fit on the straight line which means the equation of the axis's values. This means the data has normal distribution.

Outcome:

All the assumptions were met; the model can be used for predicting stature with 35.6% explanation of the variability of outcome measure. For testing prediction, a function was made, which estimates the stature of the individual in centimetres by adding femoral length in millimetres as input. As an example, 390mm of femur length was added as input, which resulted an estimated 167.4cm as stature.

Additional figures for Task2:

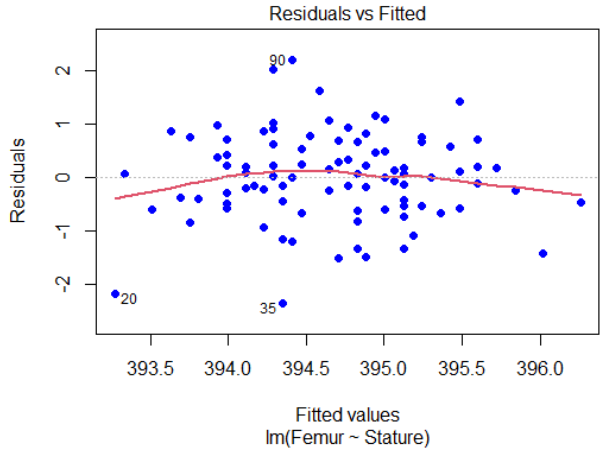

Figure4: Linearity of residuals

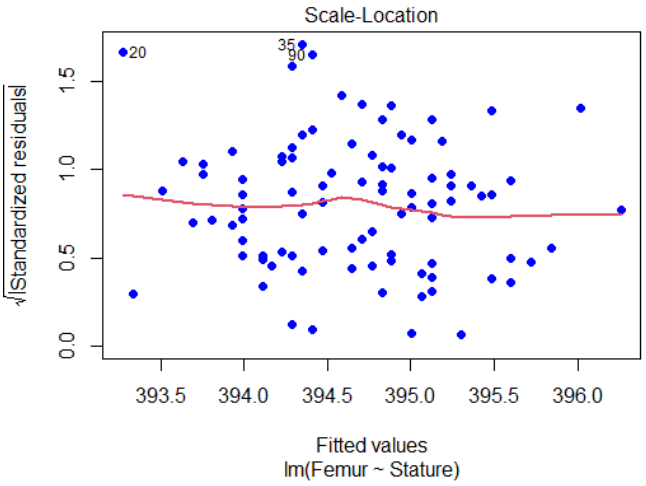

Figure5: Homoscedasticity of residuals

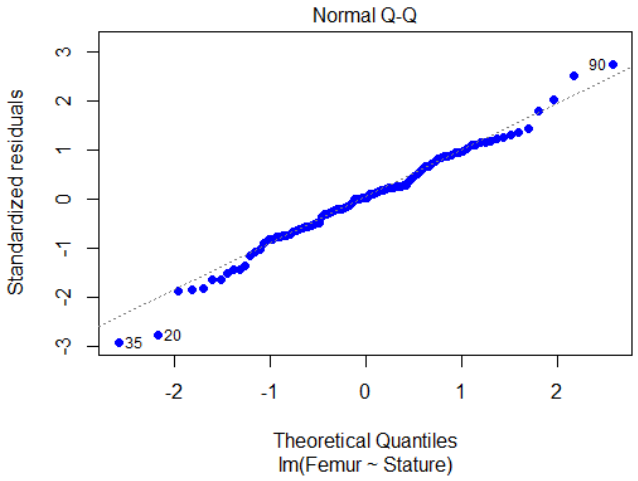

Figure6: Normality of residuals

Task3

Data and research question:

The provided data contained samples of GGT (gamma-glutamyl transpeptidase) levels (IU/L) from two group. Group1 samples was from no liver-related issues, Group2 was from individuals with liver-related issue. The data provided 20 sample for each group.

Interpreting data:

The data analysed by R (R 4.1.3). After loading the data, the groups are visualized on boxplot (Figure7).

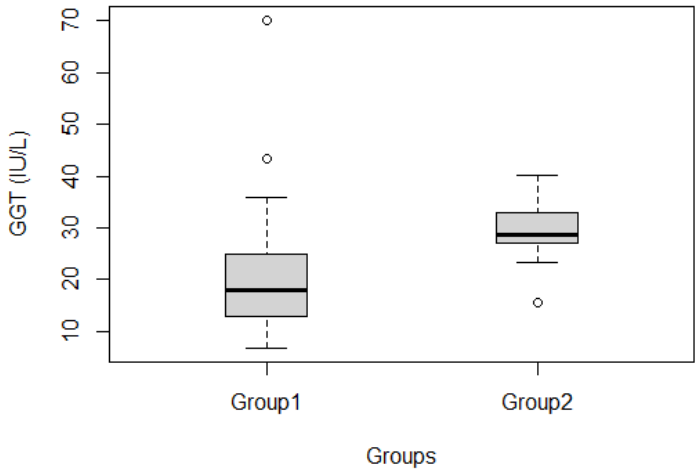

Figure7: Distribution of the level of GGT by group

The initial test was two sample t-test. For this test the null (H0) hypothesis was that the mean level of GGT in Group1 (no liver disease) is the same as the mean level of GGT in Group2 (liver disease) and the alternative (H1) was that the mean of the two group's GGT level is different. The visualised data showed outliers, also the Group1 had no normal distribution, it was failed for normality test (p-value<0.05). Therefore, instead of two sample t-test, Wilcoxon rank sum test (Mann-Whitney U test) was applied, and new hypothesis was made. The null hypothesis (H0) was that the distribution of the groups is identical, the alternative (H1) was that there is not identical distribution.

Result:

The Wilcoxon rank sum test's result (*Table5*) suggested that there is statistically significant difference between the distribution of the two groups (p-value 0.001435 <0.05), the alternative hypothesis was accepted. The W-value was 85 indicates the sum of positive ranks. In conclusion, there is statistically significant difference between the distribution of the two groups.

```
        Wilcoxon rank sum exact test

data:  blood_samples$Group1 and blood_samples$Group2
W = 85, p-value = 0.001435
alternative hypothesis: true location shift is not equal to 0
```

Table5: Wilcoxon rank sum exact test result