Assignment 1 - Report

Name, Eleni Liarou, Zoë Azra Blei, group 20

23 February 2025

```
install.packages("tinytex", repos = "https://cran.r-project.org")
```

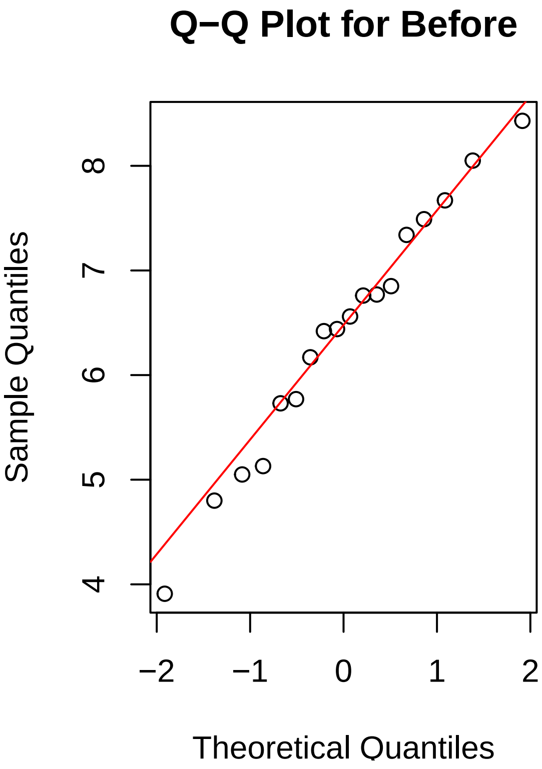
```
##
## The downloaded binary packages are in
##  /var/folders/lt/3tf47cnj2n5f1w0d_h6xmyvr0000gn/T//Rtmprm8PQR/downloaded_packages
```

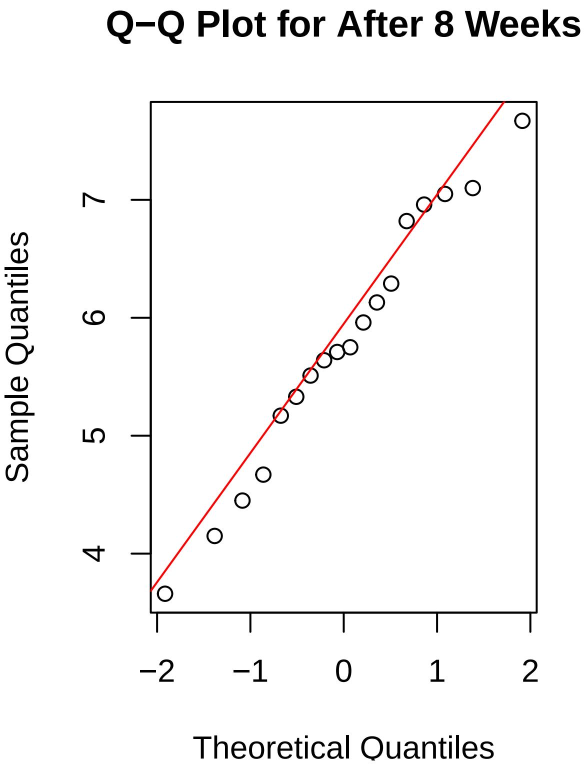
Exercise 1

First we load and read the necessary data set

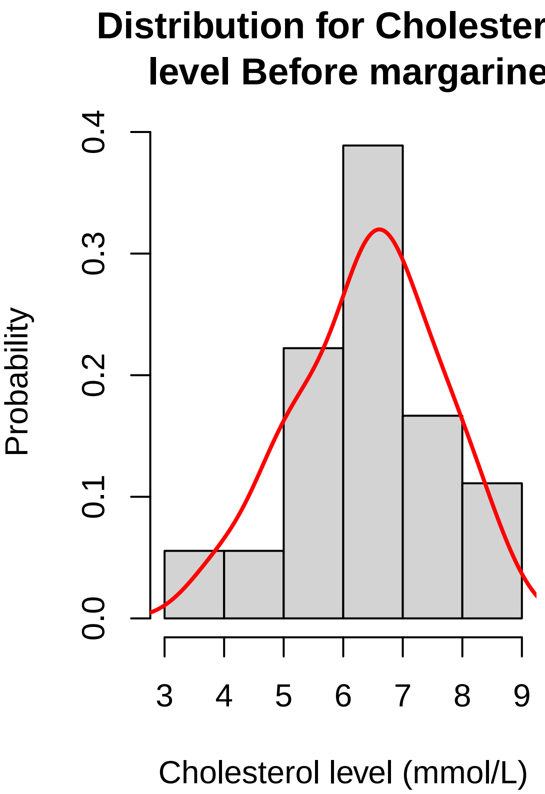
```
data = read.delim("cholesterol.txt", sep=' ')
```

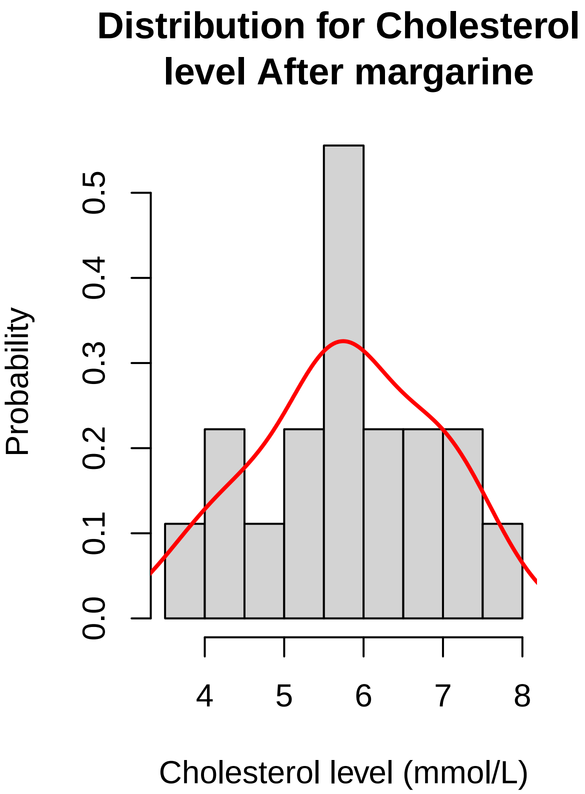
a) Make some relevant plots of this data set, comment on normality. Investigate whether the columns *Before* and *After8weeks* are correlated. In order to investigate the normality of the data set, Q-Q plots are created below for both the *Before* and *After8weeks* columns. As in both plots the data points closely follow the diagonal red line, the data is approximating a normal distribution. While some minor deviations may be present in the tails, the overall pattern suggests that the normality assumption is reasonable.





To further explore the normality assumption, histograms below were plotted for both ‘Before’ and ‘After8weeks’. The histograms exhibit a roughly bell-shaped distribution, which supports the assumption of normality.





However, to address normality more formally, a Shapiro-Wilk test is conducted, as this test is suitable to test on normallity for small data sets. For the test the null hypothesis is as follows:

H0: The data is normally distributed.

The W-statistic measures how closely the data aligns with a normal distribution, ranging from 0 to 1, where values closer to 1 indicate a stronger likelihood of normality. Considering the results for *Before* and *After8weeks*, both W-values are close to 1. Additionally, with a 95% confidence level, both p-values exceed 0.05, meaning that we fail to reject H0. These findings provide strong evidence that the data in both columns can be considered to be normally distributed.

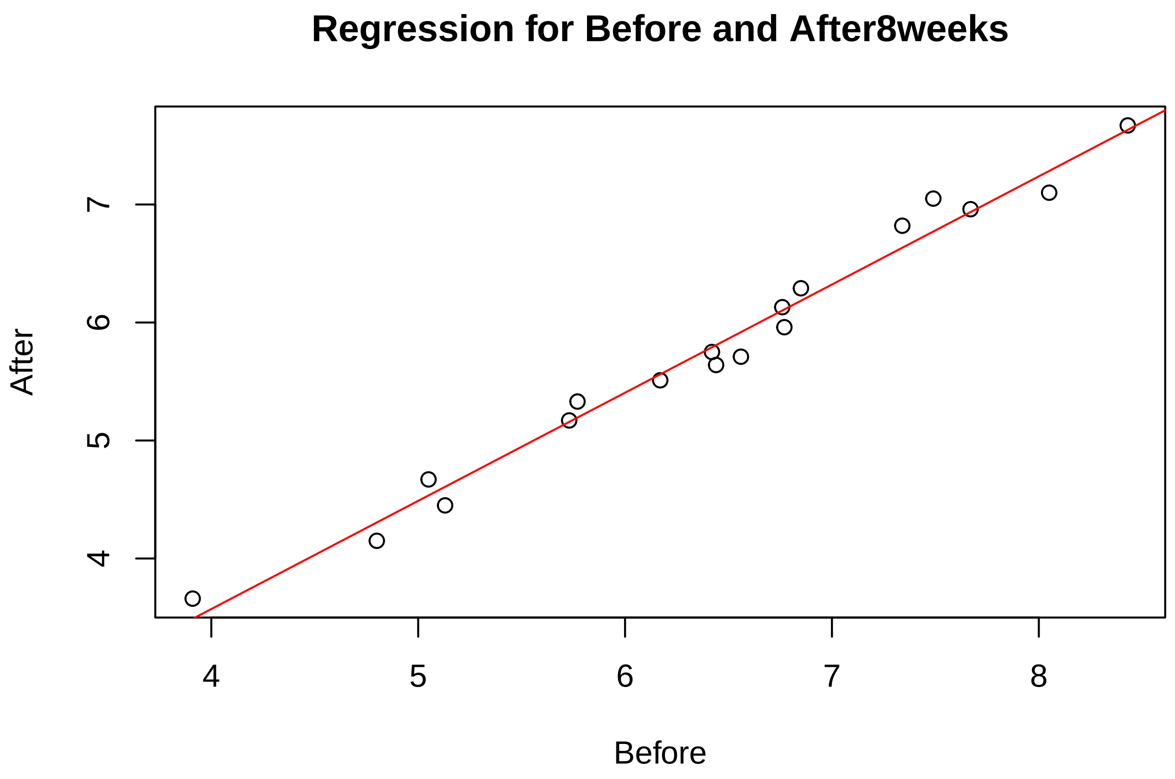
```r
shapiro.test(data$Before)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Before
## W = 0.9819, p-value = 0.9675
```

```r
shapiro.test(data$After8weeks)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$After8weeks
## W = 0.97733, p-value = 0.9183
```

In order to investigate the relationship between the columns of *Before* and *After8weeks* a scatter plot is created below. The scatter plot demonstrates a strong positive correlation between 'Before' and 'After8weeks' cholesterol levels. The data points align closely with the red regression line, suggesting that individuals with higher cholesterol levels before the diet intervention also tend to have higher cholesterol levels after 8 weeks. This indicates that while cholesterol levels may have decreased, there remains a strong relationship between pre- and post-diet measurements.



To quantify this correlation, the Pearson's correlation coefficient is calculated below. A high Pearson correlation (close to 1) indicates a strong positive relationship between the two columns. The correlation coefficient exhibits a value of approximately 0.99, confirming the strong positive relationship. Additionally, the p-value is smaller than 0.05, indicating that the correlation is statistically significant.

```
cor.test(data$Before, data$After8weeks, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$Before and data$After8weeks
## t = 29.428, df = 16, p-value = 2.321e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9751289 0.9966788
## sample estimates:
##       cor
## 0.9908885
```

b) Apply a couple of relevant tests (at least two tests, see Lectures 2–3) to verify whether the diet with low fat margarine has an effect (argue whether the data are paired or not). Is a permutation test applicable? Is the Mann-Whitney test applicable? As the cholesterol data was measured on the same population at different times, we consider the data to be paired. In this case it is possible to conduct a T-test for paired samples. However, in order to This test assumes that the mean difference of the two populations is normally distributed, thus, a Shapiro-Wilk test is conducted first to investigate the distribution.

```r
difference = data$Before - data$After8weeks
shapiro.test(difference)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference
## W = 0.98501, p-value = 0.9869
```

Explain that the data is paired:

We are doing a T-test

The null hypothesis is as follows:

H0: The margarine diet has no effect, i.e. the mean cholesterol levels *Before* and *After8weeks* are the same.

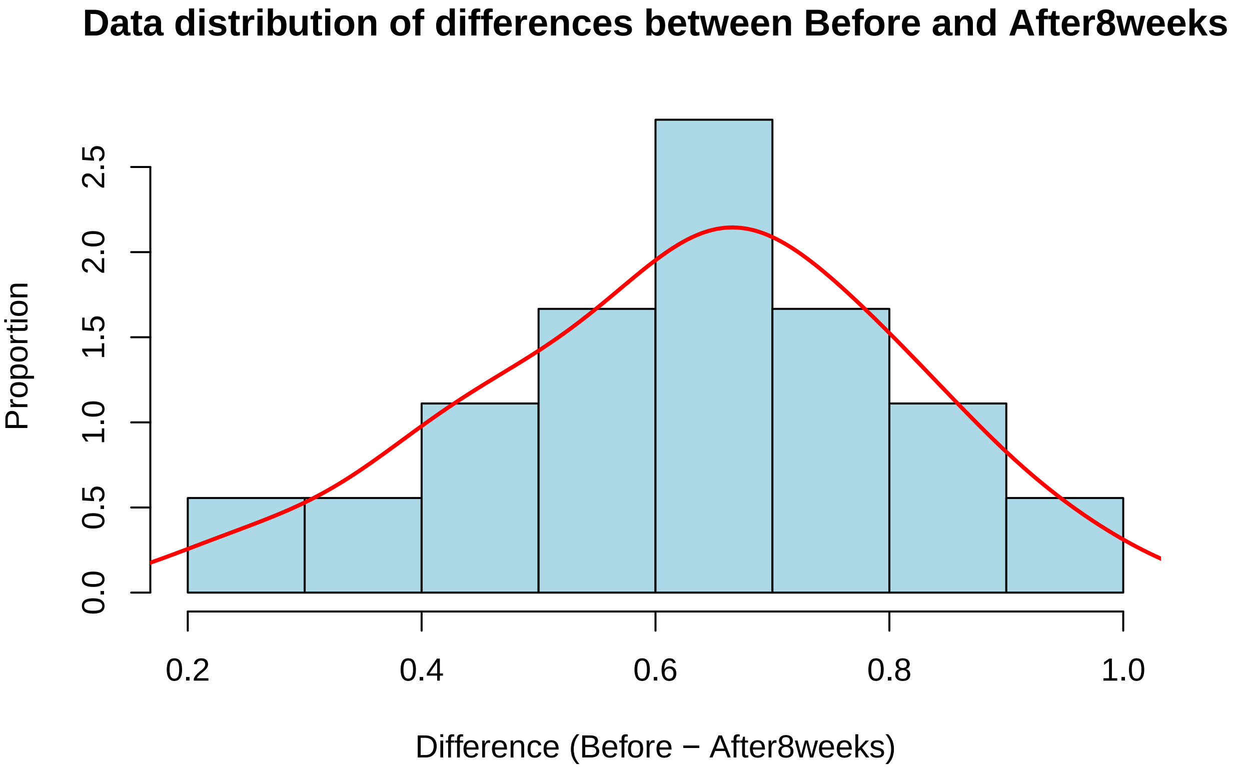
```r
# H0: The margarine diet has no effect, i.e. the mean cholesterol levels Before and After 8 we
# H1: The margarine diet reduces cholesterol levels --> mean_before > mean_after
# Paired t-test
# Assumption: the differences between Before and After should be normally distributed
# Outcome: if p < 0.05, reject H0
t.test(data$Before, data$After8weeks, paired = TRUE, alternative = "greater")
```

```
##
##  Paired t-test
##
## data:  data$Before and data$After8weeks
## t = 14.946, df = 17, p-value = 1.639e-11
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.5556906       Inf
## sample estimates:
## mean difference
##       0.6288889
```

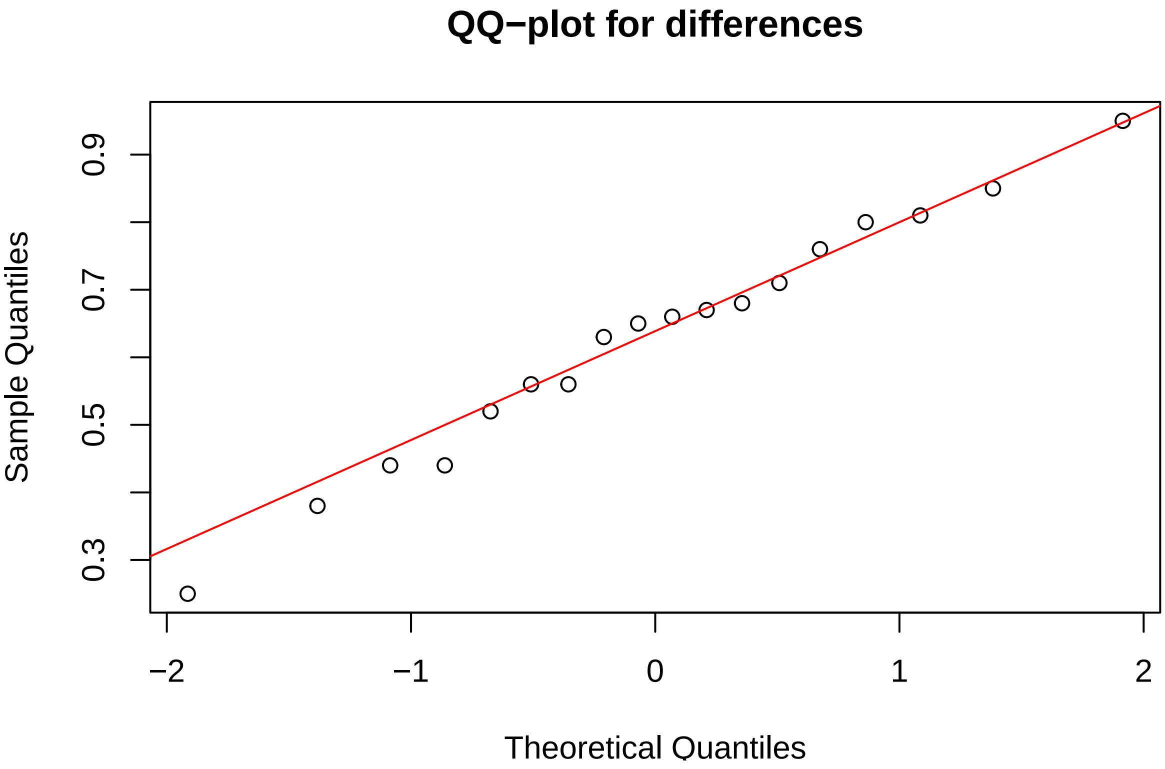
```r
# Visualizing distribution of the differences
difference = data$Before - data$After8weeks

shapiro.test(difference)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  difference
## W = 0.98501, p-value = 0.9869
```

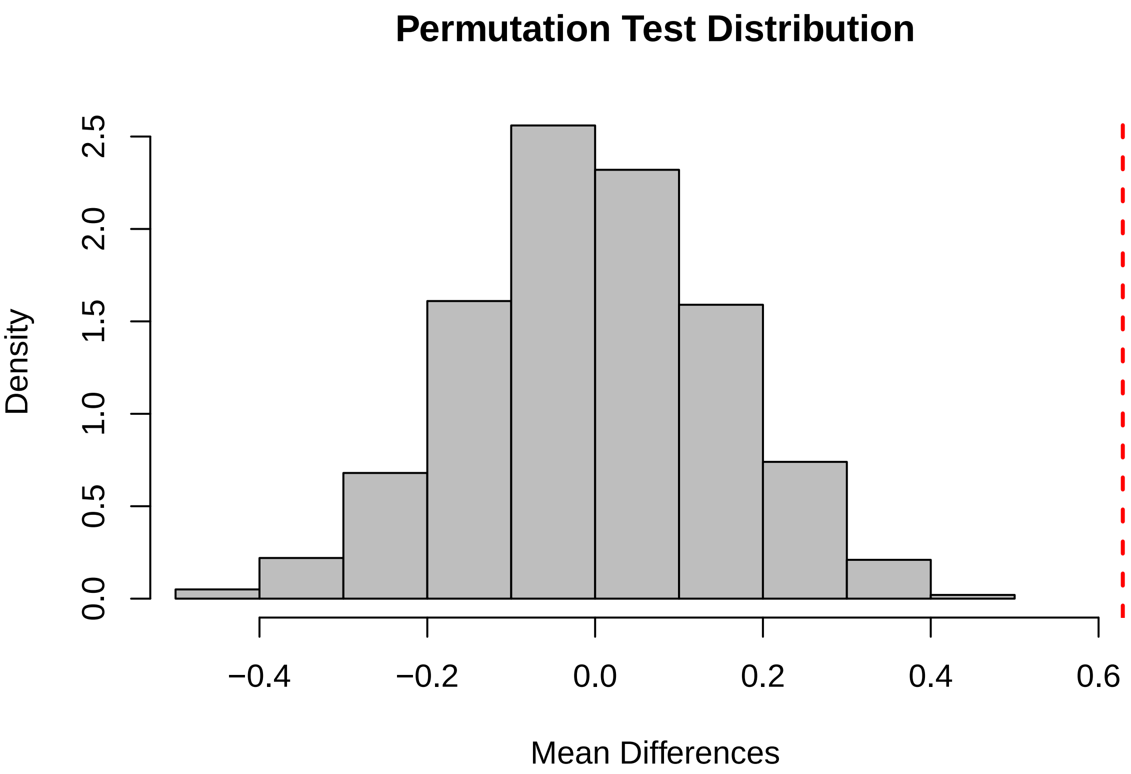
```r
hist(difference, probability = TRUE,
     main = 'Data distribution of differences between Before and After8weeks',
     col = 'lightblue',
     xlab = 'Difference (Before - After8weeks)',
     ylab = 'Proportion')
lines(density(difference), col = 'red', lwd = 2)
```



```
qqnorm(difference,
       main = 'QQ-plot for differences',
       xlab = 'Theoretical Quantiles',
       ylab = 'Sample Quantiles')
qqline(difference, col = "red")
```



```r
# Doing permutation test:
# H0: there is no difference between the Before and After8weeks groups
diff = data$Before - data$After8weeks
n_permutations = 1000
observed_mean = mean(diff)

permute_test = function(diff) {
  permuted_diff <- diff * sample(c(-1, 1), length(diff), replace = TRUE)  # Randomly flip sign
  return(mean(permuted_diff))
}

set.seed(42)
permute_distr = replicate(n_permutations, permute_test(diff))

# Histogram of the permutation distribution
hist(permute_distr, probability = TRUE, col = "gray",
     main = "Permutation Test Distribution", xlab = "Mean Differences",
     xlim = range(c(permute_distr, observed_mean)))  # Adjust x-axis limits

# Add a red line at observed mean difference
abline(v = observed_mean, col = "red", lwd = 2, lty = 2)
```



```r
# Print observed mean to check if it is within range
cat("Observed mean difference:", observed_mean, "\n")
```

```
## Observed mean difference: 0.6288889
```

```r
# Compute p-value
p_value = mean(abs(permute_distr) >= abs(observed_mean))
cat("Permutation test p-value:", p_value, "\n")
```

```
## Permutation test p-value: 0
```

```r
# Doing a Mann-Whitney U-test, as we are testing ordinal data
# H0: There is no difference in mean cholesterol level in the Before and After groups
wilcox.test(data$Before, data$After8weeks, paired = TRUE, alternative = "two.sided")
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Before and data$After8weeks
## V = 171, p-value = 7.629e-06
## alternative hypothesis: true location shift is not equal to 0
```

```r
# H1: cholesterol levels are lower after 8 weeks
wilcox.test(data$Before, data$After8weeks, paired = TRUE, alternative = "greater")
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$Before and data$After8weeks
## V = 171, p-value = 3.815e-06
## alternative hypothesis: true location shift is greater than 0
```

```
# calculating 97% CI for mu using t-score
n = length(data$After8weeks)
sample_mean = mean(data$After8weeks)
sample_sd = sd(data$After8weeks)
critical_value = qt(1-0.015, df=17)
standard_error = sample_sd / sqrt(n)

left_bound = sample_mean - critical_value * standard_error
right_bound = sample_mean + critical_value * standard_error

cat("97% Confidence Interval for mu: [", left_bound, ",", right_bound, "]\n")
```

c) Let *X1,…,X18* be the column *After8weeks*. Assume *X1,…,X18 ~ N(mu, sigma^2)* *(*irrespective of your conclusion in a)) with unknown and ^2. Construct a 97%-CI for based on normality. Next, construct a bootstrap 97%-CI for and compare it to the above CI.

```
## 97% Confidence Interval for mu: [ 5.16385 , 6.393928 ]
```

```
# calculating 97% CI for mu with bootstrapping
bootstrap_ci = function(x, conf_level = 0.97, B = 10000) {
  alpha = 1 - conf_level
  Bstats = lapply(1:B, FUN = function(i) {
    boot_sample = sample(x, size = length(x), replace = TRUE)
    mean(boot_sample)
  } )
  Bstats = unlist(Bstats)
  quantile(Bstats, prob = c(alpha/2, 1-alpha/2))
}

set.seed(42)
bootstrap_ci(data$After8weeks)
```

```
##     1.5%    98.5%
## 5.229992 6.320008
```

d) Using a bootstrap test with test statistic T=max(X1,…,X18), determine those [3,12] (if there are any) for which H0:X1,…,X18 Unif[3,] is not rejected. Can the Kolmogorov-Smirnov test be also applied for this question? If yes, apply it; if not, explain why not.

```
median(data$After8weeks)
```

e) Using an appropriate test, verify whether the median cholesterol level after 8 weeks of low fat diet is less than 6. Next, design and perform a test to check whether the fraction of the cholesterol levels after 8 weeks of low fat diet less than 4.5 is at most 25%.

```
## [1] 5.73
```

```
wilcox.test(data$After8weeks, mu = 6, alternative = "less")
```

```
## Warning in wilcox.test.default(data$After8weeks, mu = 6, alternative = "less"):
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data$After8weeks
## V = 67.5, p-value = 0.223
## alternative hypothesis: true location is less than 6
```

```
# Count how many values in After8weeks are less than 4.5
count_below_4.5 = sum(data$After8weeks < 4.5)
percentage_below_4.5 = (count_below_4.5 / length(data$After8weeks)) * 100
cat("Percentage of cholesterol levels below 4.5:", percentage_below_4.5, "%\n")
```

```
## Percentage of cholesterol levels below 4.5: 16.66667 %
```

```
# H0: The fraction of cholesterol levels below 4.5 is at most 25%
# H1: The fraction is greater than 25%
# if the p-value is small (<0.05), we reject H0 and conclude that the fraction is significantl
binom.test(count_below_4.5, length(data$After8weeks), p = 0.25, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  count_below_4.5 and length(data$After8weeks)
## number of successes = 3, number of trials = 18, p-value = 0.8647
## alternative hypothesis: true probability of success is greater than 0.25
## 95 percent confidence interval:
##  0.04702488 1.00000000
## sample estimates:
## probability of success
##              0.1666667
```

Exercise 2: Crops

Section a

We want to investigate whether two factors County and Related (and possibly their interaction) influence the crops by performing relevant ANOVA model(s), without taking Size into account. So we create and test 3 separate Null Hypotheses with a two-way ANOVA and a one-way ANOVA on the additive model:

- H_(01): no main effect of factor County
- H_(02): no main effect of factor Related
- H_(03): no interactions between factors County and Related

```
model_a <- lm(Crops ~ County * Related, data = crops_data)
anova(model_a)
```

```
## Analysis of Variance Table
##
## Response: Crops
##                 Df    Sum Sq Mean Sq F value Pr(>F)
## County           2   8841441 4420721  0.7644 0.4766
## Related          1   2378957 2378957  0.4113 0.5274
## County:Related   2   1497573  748786  0.1295 0.8792
## Residuals       24 138805865 5783578
```

From this table we can see the following:

For County: The p-value (0.476) is greater than 0.05, meaning we fail to reject the null hypothesis H_(01), suggesting that there is no significant effect of the County on the Crops variable.

For Related: The p-value (0.527) is also greater than 0.05, meaning we fail to reject the null hypothesis H_(02), which suggests that there is no significant effect of whether the landlord and tenant are related on the Crops variable.

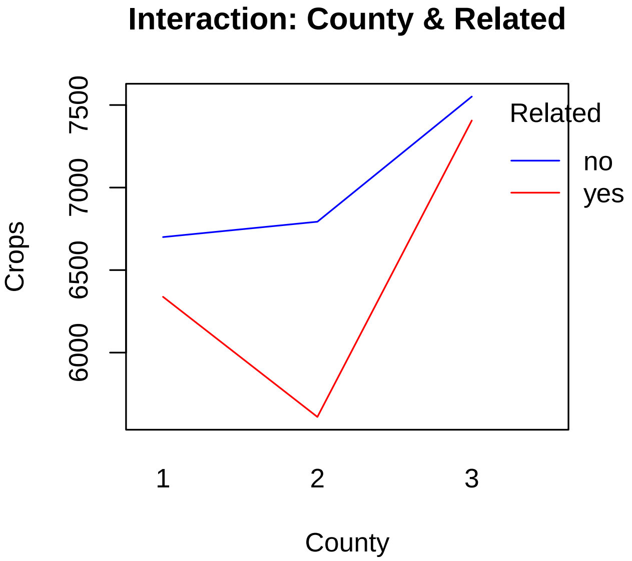
For both: The p-value (0.879) is much greater than 0.05, meaning we fail to reject the null hypothesis H_(03), implying there is no significant interaction between County and Related on the Crops variable.

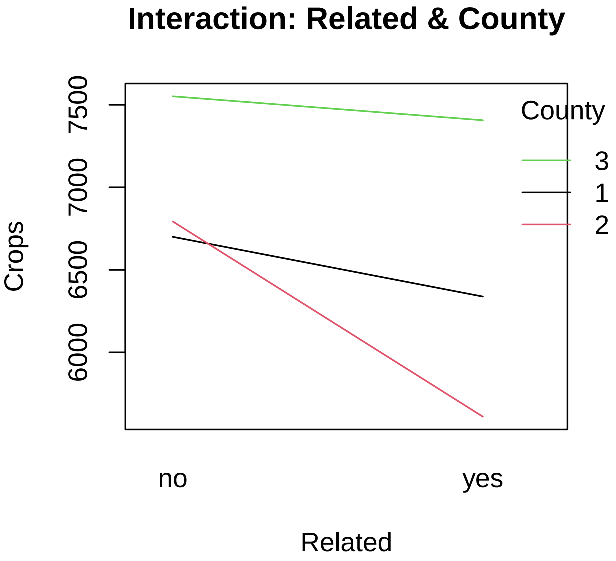
```
summary(model_a)
```

```
##
## Call:
## lm(formula = Crops ~ County * Related, data = crops_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3120.4 -1744.7  -176.9  2064.2  4806.6
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6700.0     1075.5   6.230 1.94e-06 ***
## County2                93.0     1521.0   0.061    0.952
## County3               851.2     1521.0   0.560    0.581
## Relatedyes           -362.0     1521.0  -0.238    0.814
## County2:Relatedyes   -820.6     2151.0  -0.381    0.706
## County3:Relatedyes    217.0     2151.0   0.101    0.920
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

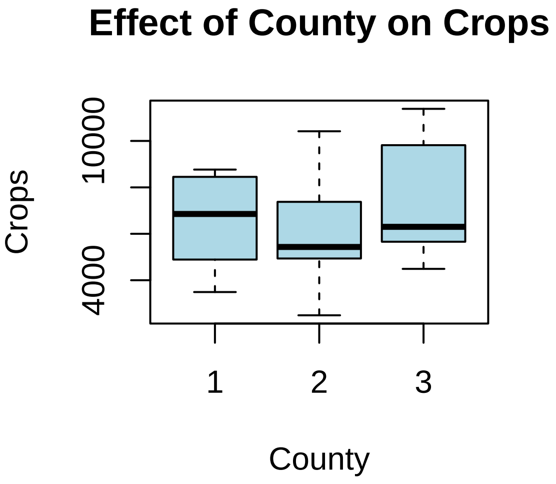
```
## Residual standard error: 2405 on 24 degrees of freedom
## Multiple R-squared:  0.08393,    Adjusted R-squared:  -0.1069
## F-statistic: 0.4398 on 5 and 24 DF,  p-value: 0.8163
```

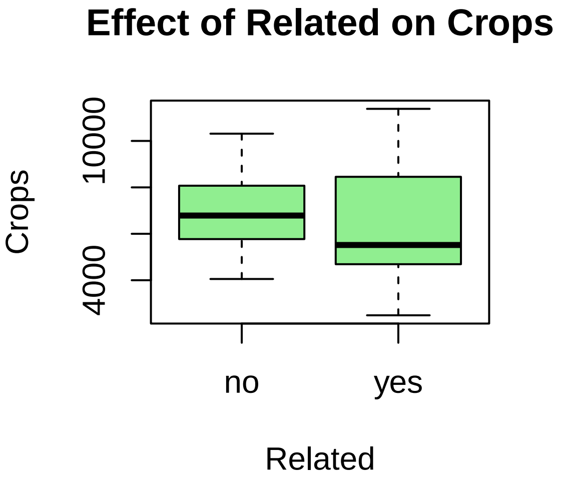
The above model summary table aligns with the ANOVA p-values as both show that none of the predictors(County, Related, or their interaction) are significant in either table.





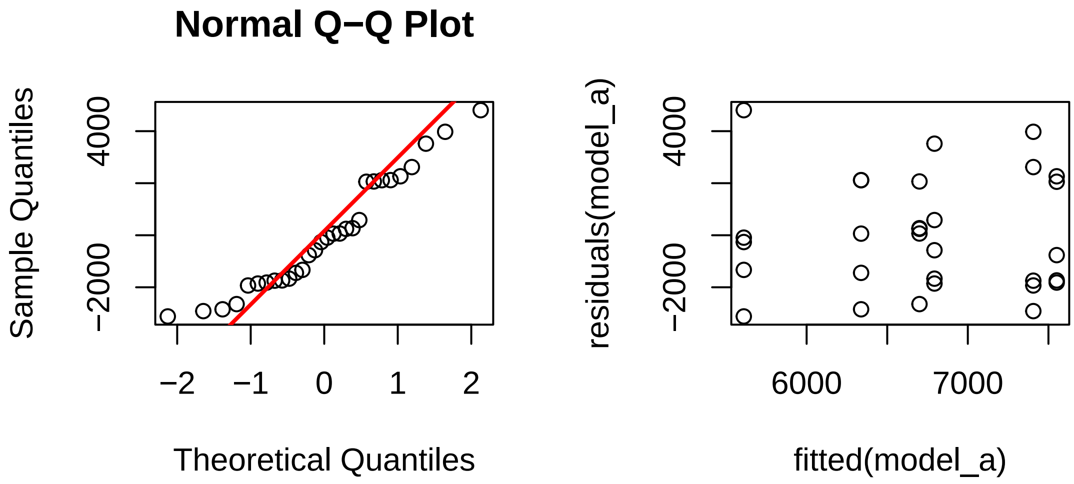
In the above interaction plots the lines seem parallel, therefore interaction seems to not be present, verifying the two-way anova results.





ANOVA shows no significant effect of County or Related on crop yield. High p-values suggest no strong differences, consistent with the boxplot, where distributions overlap, medians are close, and no outliers appear.

Finally, we have to check the model assumptions



```
## [1] "These are the p-values of the Shapiro-Wilk test for model_a: p = 0.099, W = 0.941"
```

The Q-Q plot shows deviations from normality, particularly in the tails, but the overall trend follows the theoretical quantiles. The residuals vs. fitted plot suggests no strong patterns, indicating an approximately random distribution of residuals. The Shapiro-Wilk test (W = 0.941, p = 0.099) fails to reject the null hypothesis of normality at the 0.05 level. Given the small sample size (n = 30), results should be interpreted with caution, as minor departures from normality can impact statistical inference.

To estimate crop yields for County 3 when the landlord and tenant are unrelated, we use the emmeans function to calculate the adjusted mean yield. This estimation is based on model_a, which incorporates the County-Size interaction. The emmeans function provides the estimated marginal means, accounting for the effects of County and Related while adjusting for interactions.

```
## Estimated crops for County 3 (Landlord and Tenant NOT related): 7551.2
```

Section b

We define 3 different models:

1. Model_county_size: This model examines how crop yields are influenced by County, Related, and Size, with an additional focus on the interaction between County and Size. It does not include an interaction term for Related.

```
model_county_size <- lm(Crops ~ Size * County + Related, data = crops_data)
anova(model_county_size)
```

```
## Analysis of Variance Table
##
## Response: Crops
##             Df    Sum Sq   Mean Sq  F value   Pr(>F)
## Size         1 119569344 119569344 135.6241 4.01e-11 ***
## County       2    767179    383589   0.4351  0.65242
## Related      1   1381334   1381334   1.5668  0.22325
## Size:County  2   9528654   4764327   5.4040  0.01192 *
## Residuals   23  20277325    881623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Model_related_size: This model evaluates how County, Related, and Size affect crop yields, and it adds an interaction term for Related and Size to test if the effect of Size on crop yields depends on whether the landlord and tenant are related.

```
model_related_size <- lm(Crops ~ Size * Related + County, data = crops_data)
anova(model_related_size)
```

```
## Analysis of Variance Table
##
## Response: Crops
##              Df    Sum Sq   Mean Sq  F value    Pr(>F)
## Size          1 119569344 119569344 100.8587 4.521e-10 ***
## Related       1   1380585   1380585   1.1645    0.2913
## County        2    767927    383964   0.3239    0.7264
## Size:Related  1   1353666   1353666   1.1418    0.2959
## Residuals    24  28452313   1185513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

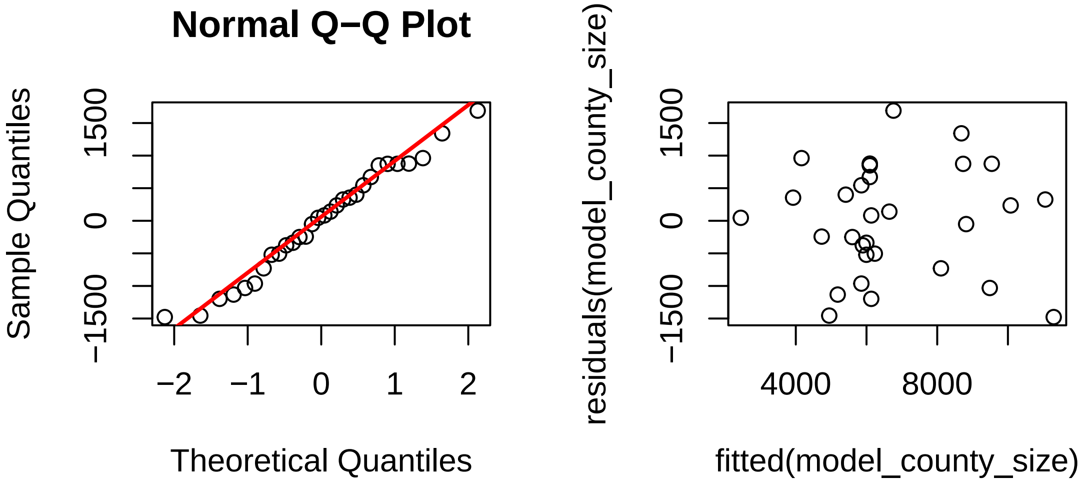
3. Model_additive: This model assumes that the effect of each factor on crop yield is independent of the others. It does not test for any interaction effects, only the individual contributions of County, Related, and Size to crop yields.

```
model_additive <- lm(Crops ~ Size + County + Related, data = crops_data)
anova(model_additive)
```

```
## Analysis of Variance Table
##
## Response: Crops
##           Df    Sum Sq   Mean Sq  F value    Pr(>F)
## Size       1 119569344 119569344 100.2897 3.114e-10 ***
## County     2    767179    383589   0.3217    0.7278
## Related    1   1381334   1381334   1.1586    0.2920
## Residuals 25  29805979   1192239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have tested interaction models as well as purely additive. The interaction Size-Related and the individual effect of County and Related are insignificant(all 3 have p-values>0.5). Therefore, the best model is model_county_size, since it shows the significance of Size and of the interaction Size-County.

Finally, we can check this model's assumptions.



```
## [1] "These are the p-values of the Shapiro-Wilk test for model_county_size: p = 0.733, W =
```

The Shapiro-Wilk test for model_county_size residuals (p = 0.733) suggests no significant deviation from normality, supported by the QQ-plot's linear pattern.

Section c

```
summary(model_county_size)
```

```
##
## Call:
## lm(formula = Crops ~ Size * County + Related, data = crops_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1477.64  -517.10    63.89   639.62  1690.29
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2461.014    929.764   2.647  0.01441 *
## Size           22.704      4.765   4.764 8.38e-05 ***
## County2     -4214.050   1447.242  -2.912  0.00785 **
## County3     -1284.813   1302.578  -0.986  0.33422
## Relatedyes   -239.099    347.916  -0.687  0.49881
## Size:County2   26.590      8.091   3.286  0.00323 **
## Size:County3    8.916      6.398   1.394  0.17676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 938.9 on 23 degrees of freedom
## Multiple R-squared:  0.8662, Adjusted R-squared:  0.8313
## F-statistic: 24.81 on 6 and 23 DF,  p-value: 5.851e-09
```

The coefficient for Size is 22.704 (p<0.001), meaning that in County 1 (the reference level), each unit increase in Size leads to an expected 22.7-unit increase in Crops. This effect is statistically significant, confirming a strong positive influence. County 2 has a negative coefficient of -4214.050 (p<0.01), meaning crop yields there are 4214 units lower than in County 1 for the same Size. County 3's coefficient (-1284.813, p=0.334) is not statistically significant, so we cannot conclude a strong difference from County 1. Related has a coefficient of -239.099 (p=0.499), indicating no significant impact on Crops. The Size:County2 interaction is 26.590 (p=0.003), meaning that the effect of Size on Crops in County 2 is stronger than in County 1, with a total increase of 49.3 units per Size unit. The Size:County3 interaction (8.916, p=0.176) is not statistically significant, so we cannot confidently conclude a difference from County 1. The model explains 86.6% of the variation in Crops (R^2=0.866), indicating strong explanatory power.

```
confint(model_county_size)
```

```
##                     2.5 %      97.5 %
## (Intercept)    537.651576  4384.37722
## Size            12.845887    32.56210
## County2      -7207.896850 -1220.20244
## County3      -3979.399914  1409.77389
## Relatedyes    -958.817141   480.61942
## Size:County2     9.852682    43.32644
## Size:County3    -4.319019    22.15129
```

The confidence intervals confirm that Size has a strong, statistically significant positive effect, and that County 2 and the Size:County2 interaction also have significant effects. In contrast, the confidence intervals for County 3, Related, and Size:County3 include zero, meaning there is no strong evidence that these factors significantly influence Crops.

Section d

```
pred <- emmeans::emmeans(model_county_size, specs = ~ County * Related * Size,
                at = list(County = "2", Size = 165, Related = "yes"))

summary(pred)
```

```
##  County Related Size emmean  SE df lower.CL upper.CL
##  2      yes      165   6141 345 23     5428     6855
##
## Confidence level used: 0.95
```

The predicted yield crops for a farm from County 2 of size 165, with related landlord and tenant is 6141, with a 95% CI: (5428, 6855)

```
## Prediction Variance: 118896.9
```

```
## Residual Variance (sigma^2): 881622.8
```

```
## Total Variance: 1000520
```

The fact that the prediction variance is much smaller than the residual variance suggests that most of the uncertainty is due to the residual variation (random noise or factors not captured by the model), rather than the instability of the model's coefficient estimates.

Exercise 3

a) Present an R-code for the randomization process to distribute soil additives over plots in such a way that each soil additive is received exactly by two plots within each block.

b) Make a plot to show the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen, and comment.

c) Conduct a full two-way ANOVA with the response variable *yield* and the two factors *block* and *N*. Was it sensible to include factor *block* into this model? Can we also apply the Friedman test for this situation?

d) Investigate other possible models with all the factors combined, restricting to only one (pair-wise) interaction term of factors *N*, *P* and *K* with block in one model (no need to check the model assumptions for all the models). Test for the presence of main effects of *N*, *P* and *K*, possibly taking into account factor *block*. Give your favorite model and motivate your choice.

e) For the resulting model from d), investigate how the involved factors influence *yield*. Which combination of the levels of the factors in the model leads to the largest yield?

f) Recall the main question of interest. In this light, for the resulting model from d) perform a mixed effects analysis, modeling the block variable as a random effect by using the function *lmer*. Compare your results to the results found by using the fixed effects model. (You will need to install the R-package *lme4*, which is not included in the standard distribution of R.)