

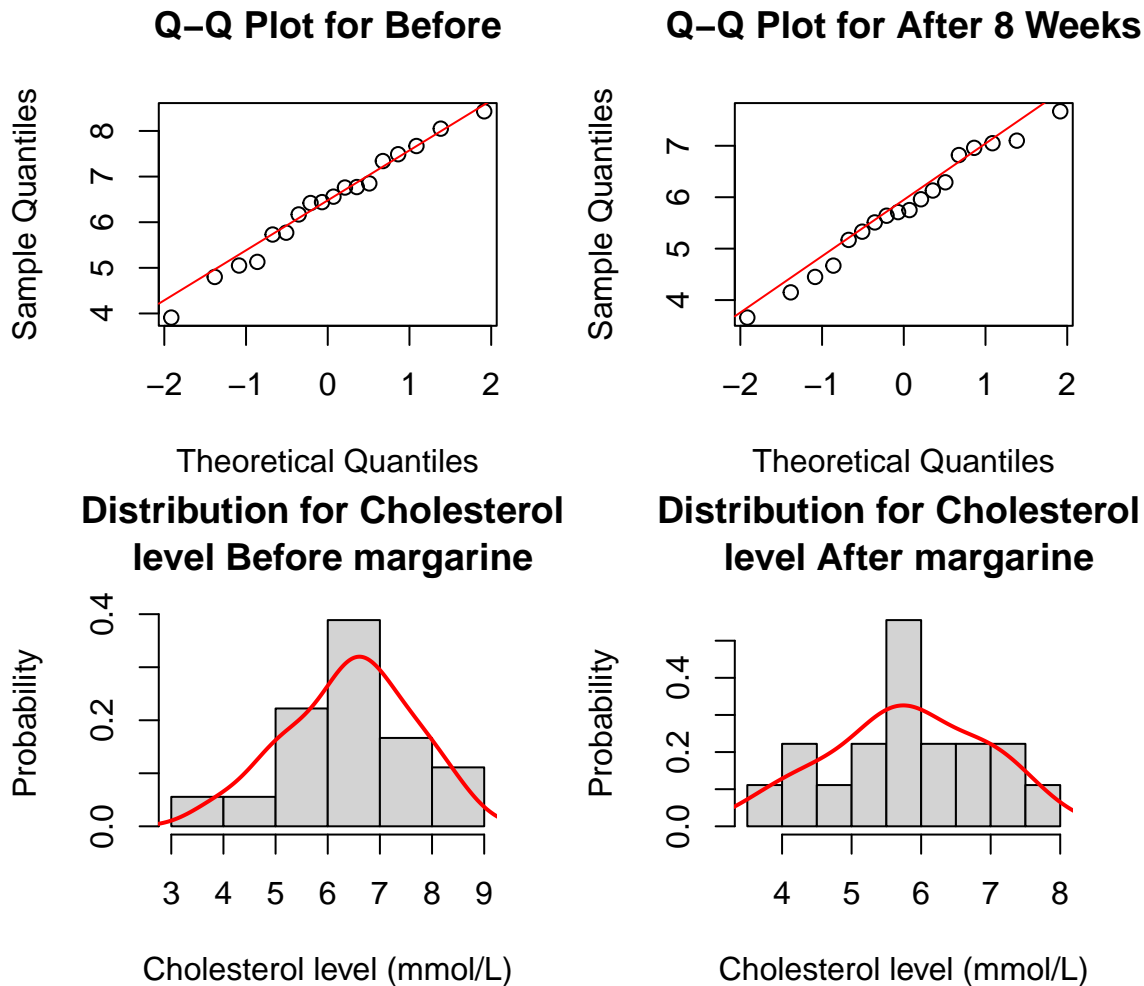
# Assignment 1 - Report

Eleni Liarou, Zoë Azra Blei, Frederieke Loth, group 20

23 February 2025

## Exercise 1: Cholesterol

**Section a** 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, above histograms are plotted for both *Before* and *After8weeks*. The histograms exhibit a roughly bell-shaped distribution, which supports the as-

sumption of normality.

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

H0: The data is normally distributed.

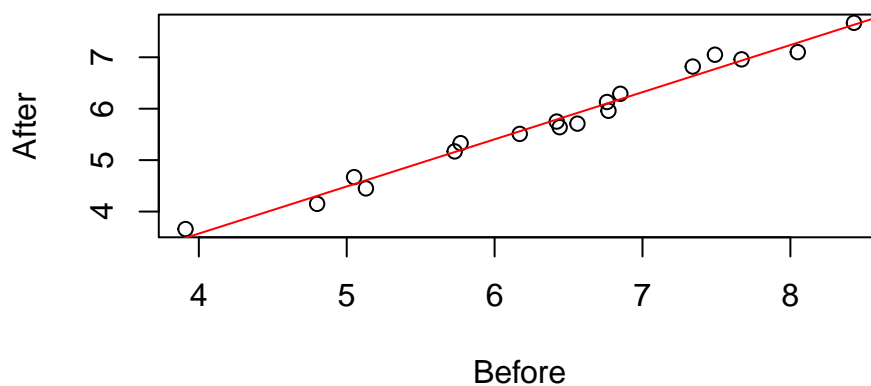
The W-statistic measures how closely the data aligns with a normal distribution, ranging from 0 to 1, where values closer to 1 indicate a stronger likelihood of normality. Considering the results for *Before* and *After8weeks*, both W-values are close to 1. Additionally, with a 95% confidence level, 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.

```
shapiro_before = shapiro.test(data$Before)
shapiro_after = shapiro.test(data$After8weeks)
```

**Shapiro-Wilk Test for Before:** W-statistic = 0.982 , p-value = 0.967

**Shapiro-Wilk Test for After 8 Weeks:** W-statistic = 0.977 , p-value = 0.918

### Regression for Before and After8weeks



In order to investigate the relationship between the columns of *Before* and *After8weeks* a scatter plot is created above. 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_result = cor.test(data$Before, data$After8weeks, method = "pearson")
```

**Pearson Correlation Test:**

Correlation coefficient ( $r$ ) = 0.991 , p-value = 2.321e-15 , 95% CI: [ 0.975 , 0.997 ]

**Section b** Since the cholesterol levels were measured on the same individuals at different times, the data is paired, making a paired t-test appropriate. However, since the t-test assumes normality of the mean difference, we first assess this assumption by applying a Shapiro-Wilk test. H0: The data is normally distributed.

The W-statistic (0.99) is close to 1, and the p-value exceeds 0.05, meaning we fail to reject H0, providing strong evidence that the mean difference follows a normal distribution.

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

**Shapiro-Wilk Test for Mean Difference:** W-statistic = 0.985 , p-value = 0.987

The paired t-test shows strong evidence against the null hypothesis ( $t = 14.946$ ,  $p = 1.639e-11$ ), indicating a significant cholesterol reduction after 8 weeks on the margarine diet. The mean difference of 0.629 mmol/L supports this, with a 95% confidence interval [0.556, Inf] confirming that the reduction is at least 0.556 mmol/L. These results suggest a clear effect of the diet on lowering cholesterol levels.

```
t_test_result = t.test(data$Before, data$After8weeks, paired = TRUE, alternative = "greater")
```

#### Paired T-Test Results:

T-statistic = 14.946 , p-value = 1.639e-11 , 95% Confidence Interval: [ 0.556 , Inf ]

The permutation test ( $p = 1.0e-05$ ) provides strong evidence to reject the null hypothesis, confirming a significant cholesterol reduction of 0.629 mmol/L after 8 weeks. This supports the effectiveness of the margarine diet. A Mann-Whitney U-test is not applicable as it requires independent samples, while this data is paired.

```
diff = data$Before - data$After8weeks
n_permutations = 100000
observed_mean = mean(diff)

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

set.seed(42)
permute_distr = replicate(n_permutations, permute_test(diff))
rounded_observed_mean = round(observed_mean, 3)
rounded_p_value = formatC(mean(abs(permute_distr) >= abs(observed_mean)), format = "e", digits
```

**Observed mean difference:** 0.629 **Permutation test p-value:** 1.000e-05

**Section c** To estimate the mean cholesterol level after 8 weeks, we compute a 97% confidence interval using both the t-test and bootstrapping. The t-test CI, assuming normality, is [5.164, 6.394], while the bootstrap CI, based on resampling, is [5.230, 6.320]. The substantial overlap suggests that both methods provide similar estimates. The slight differences arise because the t-test relies on the sample standard deviation, whereas bootstrapping does not assume normality.

Since both approaches yield consistent results, we can confidently conclude that cholesterol levels remain moderate after 8 weeks on the diet.

```
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

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)
  ci = round(quantile(Bstats, prob = c(alpha/2, 1-alpha/2)), 3)
  return(ci)
}
```

### Results:

Sample mean: 5.779, Sample standard deviation: 1.102, Critical value: 2.368, 97% CI for  $\mu$ : [ 5.164 , 6.394 ]

Bootstrap test with 97% CI: [ 5.23 , 6.32 ]

**Section d** For the bootstrap test, we consider the following null hypothesis:

H0: The data follows a uniform distribution with a minimum of 3 and a maximum of theta.

```
set.seed(42) # Ensure reproducibility
T_max_observed = max(data$After8weeks)
n = length(data$After8weeks)
theta_values = 3:12
p_values = numeric(length(theta_values))
Bootstraps = 100000
for (i in seq_along(theta_values)) {
  theta = theta_values[i]
  t_star = numeric(Bootstraps)
  for (j in 1:Bootstraps) {
    x_star = runif(n, min = 3, max = theta)
    t_star[j] = max(x_star)
  }
  pl = sum(t_star < T_max_observed) / Bootstraps
  pr = sum(t_star > T_max_observed) / Bootstraps
  p_values[i] = 2 * min(pl, pr)
}
```

**Results:** Table 1 Theta values and their p-values

	3	4	5	6	7	8	9	10	11	12
P-Value	0	0	0	0	0	0.58	0.021	0.001	0	0

To generate bootstrap samples from a  $\text{Uniform}(3, \theta)$  distribution, we iterate over  $\theta$  values from 3 to 12 in steps of 1, generating 100,000 bootstrap samples for each. The maximum value from each re-sample is recorded, and the p-value is computed by comparing the observed maximum ( $T_{\max\_observed}$ ) with the bootstrap distribution ( $t_{\star}$ ).

```
accepted_theta <- c(8, 9, 10, 11)
ks_results <- list()

for (theta in accepted_theta) {
  ks_test_result <- ks.test(data$After8weeks, "punif", min = 3, max = theta)
  ks_results[[as.character(theta)]] <- list(
    theta = theta,
    D_statistic = round(ks_test_result$statistic, 3),
    p_value = formatC(ks_test_result$p.value, format = "e", digits = 3)
  )
}
```

**Kolmogorov-Smirnov Test for Uniform(3, 8 ):** D-statistic: 0.212 , P-value: 3.451e-01

**Kolmogorov-Smirnov Test for Uniform(3, 9 ):** D-statistic: 0.261 , P-value: 1.429e-01

**Kolmogorov-Smirnov Test for Uniform(3, 10 ):** D-statistic: 0.359 , P-value: 1.401e-02

**Kolmogorov-Smirnov Test for Uniform(3, 11 ):** D-statistic: 0.432 , P-value: 1.438e-03

For  $\theta = 9$ ,  $\theta = 10$ , and  $\theta = 11$  we get a p-value  $< 0.05$ , thus we reject  $H_0$  indicating that the observed maximum for these values is significantly different from the expected maximum under  $U(3, \theta)$ . For  $\theta = 8$ , we fail to reject  $H_0$ , meaning that the data could be from a uniform distribution.

**Section e** To test whether the median of the cholesterol level is less than 6, we will use a Wilcoxon signed-rank test, because it is a non-parametric test that does not assume normality, making it suitable for small sample sizes and skewed data. The null hypothesis is:

$H_0$ : The median cholesterol level is 6 or greater.

The test yields a p-value of 0.223, meaning we fail to reject  $H_0$ . This suggests that the median cholesterol level after 8 weeks is not significantly less than 6 at the 5% significance level.

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

**Wilcoxon Signed-Rank Test:** V-statistic: 67.5 , p-value: 2.230e-01

To determine whether the fraction of cholesterol levels below 4.5 after 8 weeks exceeds 25%, a binomial test is conducted. The null hypothesis is:

$H_0$ : The fraction of cholesterol levels below 4.5 is at most 25%.

The test yields a p-value of 0.865, meaning we fail to reject  $H_0$  at the 5% significance level. This suggests that there is no significant evidence that the proportion of cholesterol levels below 4.5 is greater than 25%.

```
count_below_4.5 = sum(data$After8weeks < 4.5)
binom_result = binom.test(count_below_4.5, length(data$After8weeks), p = 0.25, alternative = ">")
```

### Exact binomial test

Number of successes: 3 , Number of trials: 18 , p-value: 0.865 , 95% CI: [ 0.047 1 ], Estimated probability of success: 0.167

## 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 and  $H_{(03)}$ : no interactions between factors County and Related

```
## 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 County p-value  $> 0.05$ , meaning we fail to reject the null hypothesis  $H_{(01)}$ , suggesting that there is no significant effect of the County on the Crops variable.

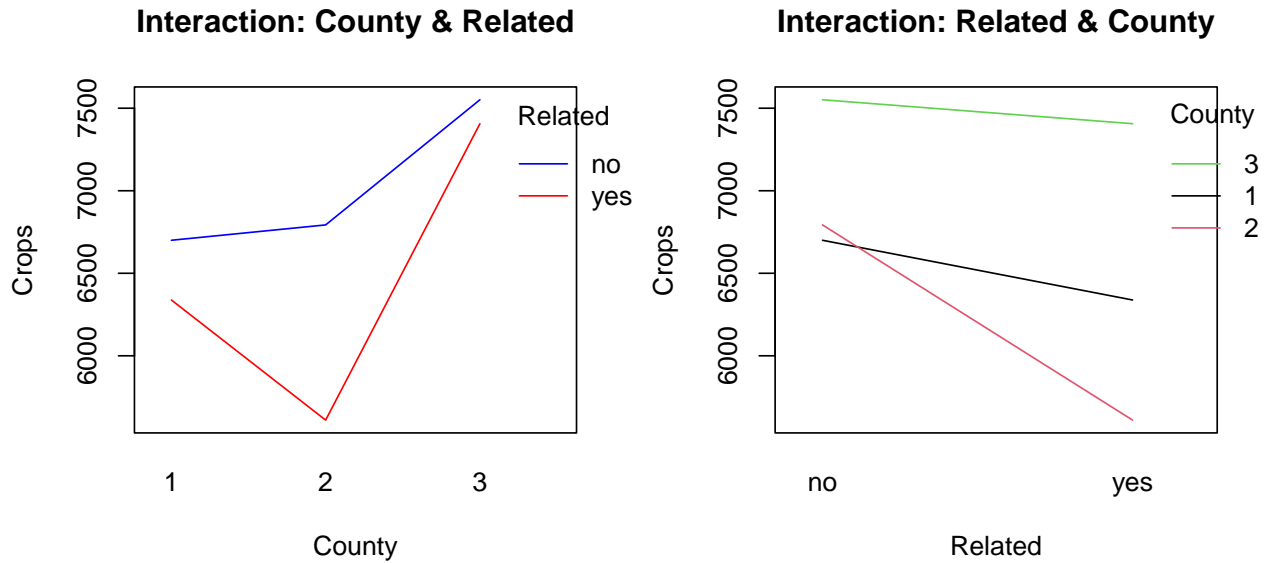
For Related: The Related p-value  $> 0.05$ , meaning we fail to reject the null hypothesis  $H_{(02)}$ , which suggests that there is no significant effect of whether the landlord and tenant are related on the Crops variable.

For both: The Interaction p-value  $> 0.05$ , meaning we fail to reject the null hypothesis  $H_{(03)}$ , implying there is no significant interaction between County and Related on the Crops variable.

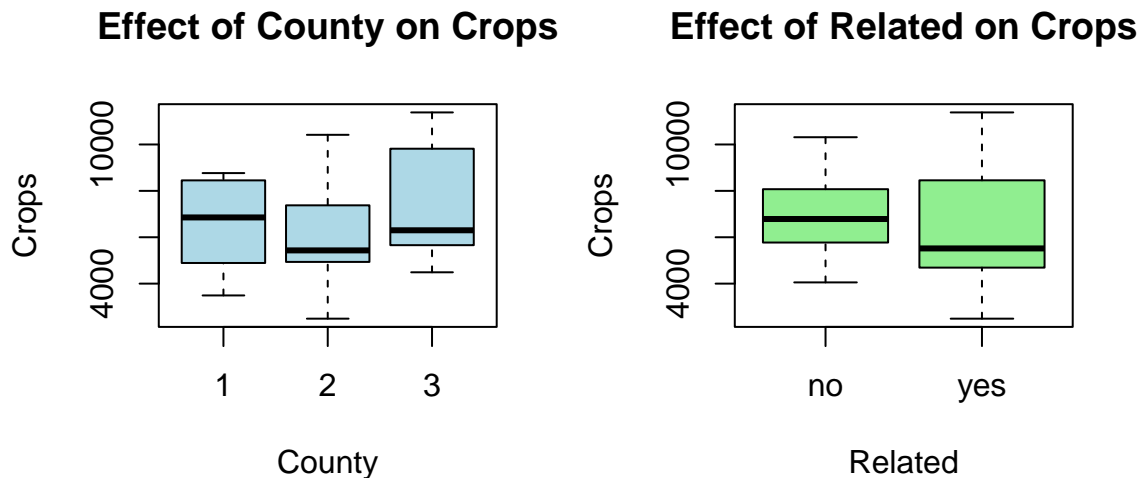
## These are the coefficients:

```
##              Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)      6700.0   1075.507  6.22961945 1.942596e-06
## County2           93.0   1520.997  0.06114412 9.517508e-01
## County3          851.2   1520.997  0.55963302 5.809157e-01
## Relatedyes       -362.0   1520.997 -0.23800183 8.138999e-01
## County2:Relatedyes -820.6   2151.014 -0.38149446 7.061930e-01
## County3:Relatedyes  217.0   2151.014  0.10088264 9.204817e-01
```

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.

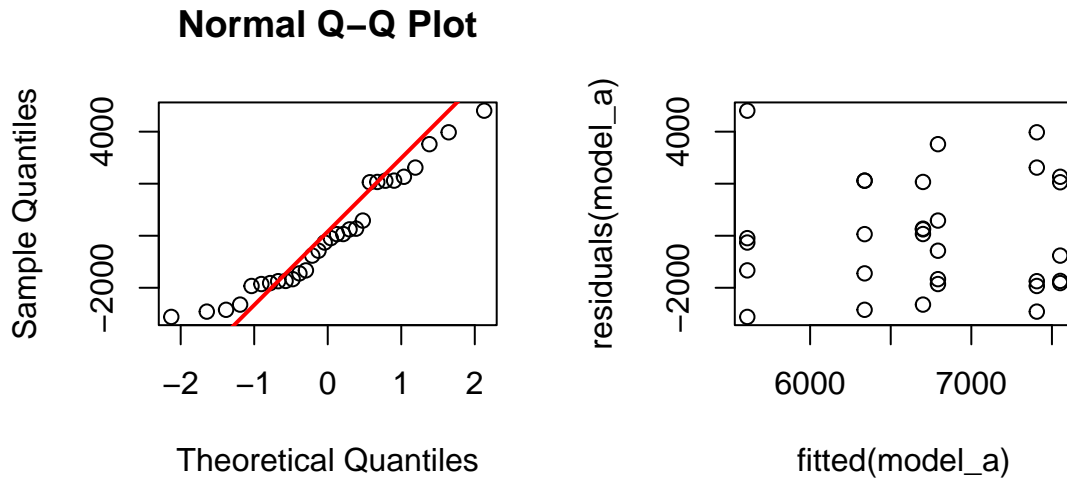


ANOVA shows no significant effect of County or Related on crop yield. High p-values suggest no strong differences, consistent with the boxplot, where distributions overlap, medians are close, and no outliers appear.

Finally, we have to check the model assumptions for normality

```
## [1] "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 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`: Examines the effects of County, Related, and Size on crop yield, focusing on the County  $\times$  Size interaction but excluding Related interactions.

```
## 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`: Evaluates the effects of County, Related, and Size on crop yield, adding a Related  $\times$  Size interaction to test if Size's effect depends on landlord-tenant relation.

```
## 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
```



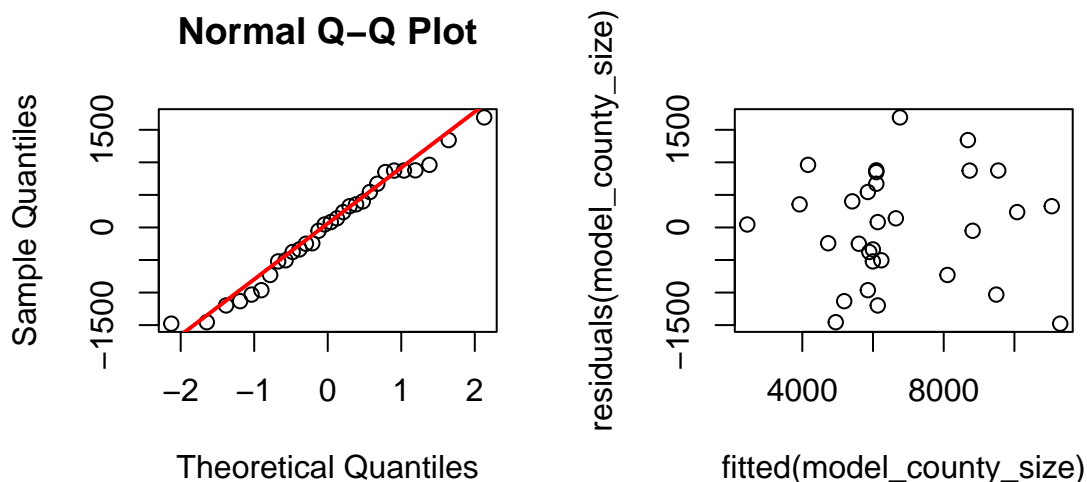
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Model\_additive: Assumes each factor affects crop yield independently, without testing interactions—only the individual effects of County, Related, and Size.

```
## 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( $p$ -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] "Shapiro-Wilk test for model_county_size: p = 0.733, W = 0.977"
```

The Shapiro-Wilk test for model\_county\_size residuals suggests no significant deviation from normality, supported by the QQ-plot's linear pattern.

### Section c

```
## These are the coefficients:
##           Estimate Std. Error  t value    Pr(>|t|)
## (Intercept)  2461.014399  929.763733  2.6469245 1.440848e-02
## Size         22.703995    4.765462  4.7642804 8.377593e-05
## County2      -4214.049646 1447.241530 -2.9117805 7.853645e-03
## County3      -1284.813014 1302.577520 -0.9863620 3.342189e-01
## Relatedyes    -239.098860  347.915613 -0.6872323 4.988081e-01
```

```
## Size:County2      26.589563      8.090696      3.2864369 3.234050e-03
## Size:County3       8.916134      6.397943      1.3935939 1.767616e-01
```

```
##
```

```
## This is the R-squared value: 0.8661773
```

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.

```
##              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 above 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

```
## 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: Yield of peas

**Section a** To optimize possible future instances of a random plot distribution process we create a data frame similar to that of the npk data such that it may be utilized similarly.

```
# Set block and plot dimensions
n_blocks <- 6
n_plots <- 4

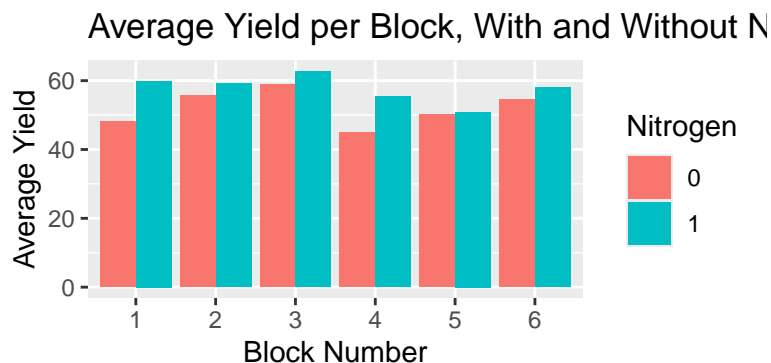
# Create a data frame to randomly distribute over blocks
random_distributed <- data.frame(block = rep(1:n_blocks, each = n_plots),
                                N = rep(0, n_plots*n_blocks),
                                P = rep(0, n_plots*n_blocks),
                                K = rep(0, n_plots*n_blocks))

# Iterate over blocks for index and sampling
for (block in 1:n_blocks) {

  idx <- (n_plots * (block-1) + 1): (n_plots * block)

  random_distributed[sample(idx, 2), "N"] <- 1
  random_distributed[sample(idx, 2), "P"] <- 1
  random_distributed[sample(idx, 2), "K"] <- 1
}
```

**Section b** From the npk data a bar graph is generated to report the average *yield* per *block* in the presence (1) and absence (0) of *N*.



From the graph we can note that there seems to be a correlation between a higher *yield* and the presence of the additive *N* in the plot soil. We will further investigate this hypothesis with different tests such as a full two-way ANOVA. The *block* factor is not of interest but is applied to group plot tests and may introduce variability within the samples.

**Section c** A full two-way ANOVA is conducted with the response variable *yield* and the two factors *block* and *N*. We start with a main effects model.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5  343.3    68.66   3.395 0.0262 *
```

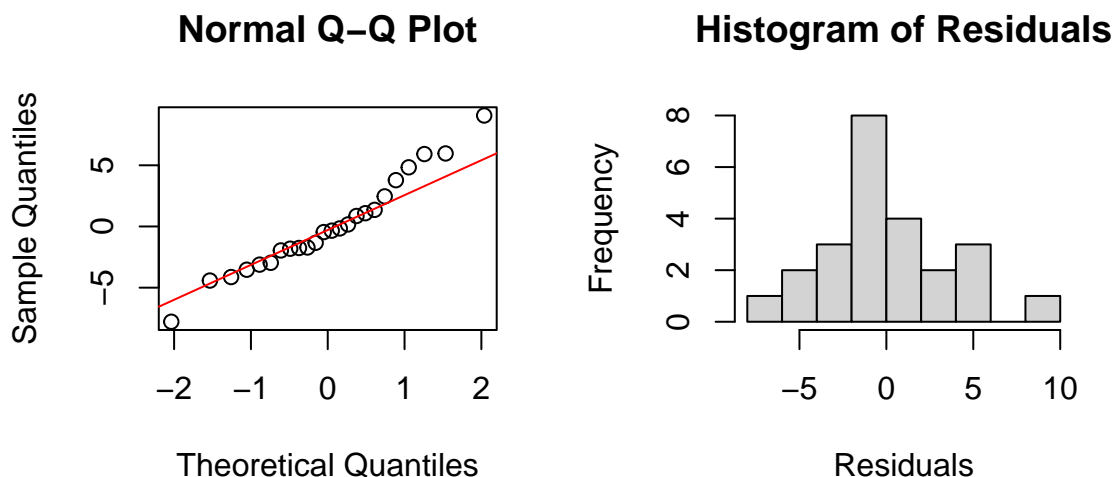
```
## N          1  189.3  189.28   9.360 0.0071 **
## Residuals  17  343.8   20.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We then compute an interaction model.

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      5  343.3   68.66   3.359 0.0397 *
## N          1  189.3   189.28   9.261 0.0102 *
## block:N     5   98.5   19.70   0.964 0.4769
## Residuals  12  245.3   20.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0.0071 and 0.0262 respectively in the main effects model, the ANOVA signifies that both the *N* additive and *block* factor have statistically significant effects on the yield of peas with nitrogen having a larger effect compared to the block. However, we might say that the *block* factor merely introduces variability in the samples and do not represent fixed system effects. Including the *block* factor is, however, important to understand the amount of variability it may cause. In the interaction model both factors again have a p-value below 0.05. Additionally, we notice that there is no statistically significant effect on the *yield* from an interaction between *N* and *block*.

To solidify our findings we perform an analysis on the residuals of the main effects model with a normality test and a plot of the distribution with a histogram.



We note that there is a slight deviation at the tail of the normality test which we investigate this further with a Shapiro Wilk test.

```
##
## Shapiro-Wilk normality test
##
## data:  residuals(two_way_anova_main)
## W = 0.96937, p-value = 0.6514
```

From the Shapiro-Wilk test we receive a p-value greater than 0.05, combined with the Q-Q and histogram it is likely normally distributed.

Tests indicate a normal distribution of the residuals and the data is continuous. We can therefore say that a non-parametric designed test such as a Friedman test is not applicable. Additionally, We have limited in-block variation due to a randomized block design with continuous data which is more applicable to the ANOVA model design.

**Section d** To investigate the effects of other variables and any other possible interactions we explore more models. We first test for the presence of additional main effects from all additives and the block factor.

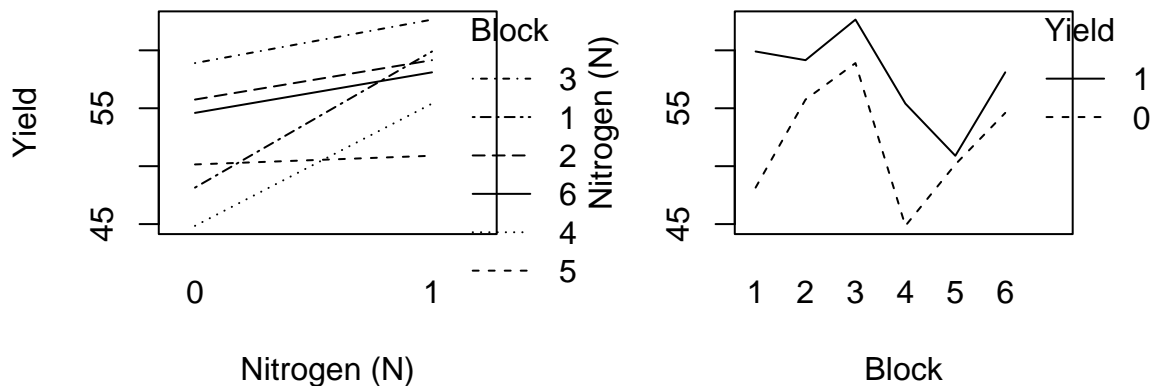
```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## N           1  189.3   189.28   11.821 0.00366 **
## P           1    8.4     8.40    0.525 0.47999
## K           1   95.2    95.20    5.946 0.02767 *
## block       5  343.3    68.66    4.288 0.01272 *
## Residuals   15  240.2    16.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We now learn that potassium has a statistically significant effect on the yield with a p-value of 0.02767. From this main effects model we therefore learn that only *P* does not contribute significantly, we will therefore omit *P* in subsequent models.

Next, we explore the interaction between potassium and the other remaining main effect variables *N* and *block*.

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## N           1  189.3   189.28   13.037 0.00476 **
## K           1   95.2    95.20    6.557 0.02834 *
## block       5  343.3    68.66    4.729 0.01777 *
## N:K         1   33.1    33.14    2.282 0.16179
## K:block     5   70.3    14.05    0.968 0.48123
## Residuals   10  145.2    14.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the data we learn that there is no significant interaction between the different factors. We further explore this with an interaction plot.



From this plot we only observe a minor possible interaction between *block* and *N*. However, due to

the overall linearity we suspect the interaction to be mostly negligible.

We will therefore approach a main effects model where we leave out phosphor to omit arbitrary data and prevent overfitting. The main effect model will include factors *N*, *K* and *block*

```
# P could be omitted due to statistical insignificance
```

```
model_3 <- aov(yield ~ N + K + block, data = npk)
summary(model_3)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## N              1  189.3   189.28   12.183 0.00302 **
## K              1   95.2    95.20    6.128 0.02487 *
## block          5  343.3    68.66    4.419 0.01017 *
## Residuals     16  248.6    15.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Section e** To investigate the influence of the factors on the *yield*, we will perform a Tukey's Honest Significant Difference (HSD) test.

```
# Apply TukeyHSD
```

```
tukey_results <- TukeyHSD(model_3)
tukey_results
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = yield ~ N + K + block, data = npk)
##
## $N
##           diff           lwr           upr           p adj
## 1-0  5.616667  2.205368  9.027965  0.0030242
##
## $K
##           diff           lwr           upr           p adj
## 1-0 -3.983333 -7.394632 -0.5720346  0.0248744
##
## $block
##           diff           lwr           upr           p adj
## 2-1    3.425   -5.555691  12.405691  0.8168210
## 3-1    6.750   -2.230691  15.730691  0.2062520
## 4-1   -3.900  -12.880691   5.080691  0.7270343
## 5-1   -3.500  -12.480691   5.480691  0.8035575
## 6-1    2.325   -6.655691  11.305691  0.9565038
## 3-2    3.325   -5.655691  12.305691  0.8338711
## 4-2   -7.325  -16.305691   1.655691  0.1465878
## 5-2   -6.925  -15.905691   2.055691  0.1863101
## 6-2   -1.100  -10.080691   7.880691  0.9985112
## 4-3  -10.650  -19.630691  -1.669309  0.0156223
## 5-3  -10.250  -19.230691  -1.269309  0.0207261
```

```
## 6-3 -4.425 -13.405691 4.555691 0.6173347
## 5-4 0.400 -8.580691 9.380691 0.9999896
## 6-4 6.225 -2.755691 15.205691 0.2761166
## 6-5 5.825 -3.155691 14.805691 0.3396777
```

From the TukeyHSD results we see that  $K$  has a significant negative impact on yield. Therefore we can say that the best combination for model 3 (where we do not consider  $P$ ) is  $N = 1$  and  $K = 0$ . Additionally, the *block* factor introduces variability with *block* 3 having the overall biggest positive impact and *block* 1 a relatively lower *yield*.

**Section f** Finally, we create a mixed effects analysis for our chosen main effect model with factors  $N$ ,  $K$  and *block* where we model *block* as a random effect.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: yield ~ N + K + (1 | block)
## Data: npk
##
## REML criterion at convergence: 131.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.67459 -0.51552  0.09225  0.51006  1.52224
##
## Random effects:
## Groups   Name                Variance Std.Dev.
## block    (Intercept) 13.28      3.644
## Residual                  15.54      3.942
## Number of obs: 24, groups: block, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   54.058      2.039  26.519
## N1              5.617      1.609   3.490
## K1             -3.983      1.609  -2.475
##
## Correlation of Fixed Effects:
##      (Intr) N1
## N1 -0.395
## K1 -0.395  0.000
```

Similarly to the additive model, we can conclude that  $N$  has a significant (5.617) positive effect on the *yield* with a p-value of 0.00302, whereas  $K$  has a statistically significant negative effect on the *yield* (-3.983) with a p-value of 0.02487. This analysis shows that *blocks* have significant variability in data, with a random effects variance of 13.16. This mixed effects model further solidifies the stance that  $N$  has a significant impact on *yield*. We further note that the mixed effects model may provide a better representation, since *blocks* introduce random variance and not fixed systemic effects such as  $N$ ,  $P$  and  $K$ .