Assignment 1

Yichen TONG, group 20

27 February 2023

Exercise 4

a) We can solve this question in this way. To randomize the distribution of soil additives over plots in such a way that each soil additive is received exactly by two plots within each block, the following steps in R can be used:

```
library(MASS)
set.seed(123)  # Set seed for reproducibility

# Create a matrix to store the randomized distribution of soil additives
dist_mat <- matrix(0, nrow = 6, ncol = 4)

# Randomly assign soil additives to each block
for (i in 1:6) {
  dist_mat[i, ] <- sample(c(1,1,0,0), 4)
}

# Verify that each soil additive is received by exactly two plots within each block
colSums(dist_mat)  # The sum of each column should be 2
## [1] 2 4 2 4
rowSums(dist_mat)  # The sum of each row should be 2
## [1] 2 2 2 2 2 2

# Combine the block information and the soil additive distribution
npk_dist <- data.frame(npk, dist_mat)
```

Within this code above, firstly we set the seed for the reproducibility after load the MASS package. Then in order to store the distribution of soil additives which is randomized, a matrix is supposed to be created. After that, we need to make verification to study if every soil additive can be received by the two plots. In the end, the data of the additive distribution information is integrated and combined.

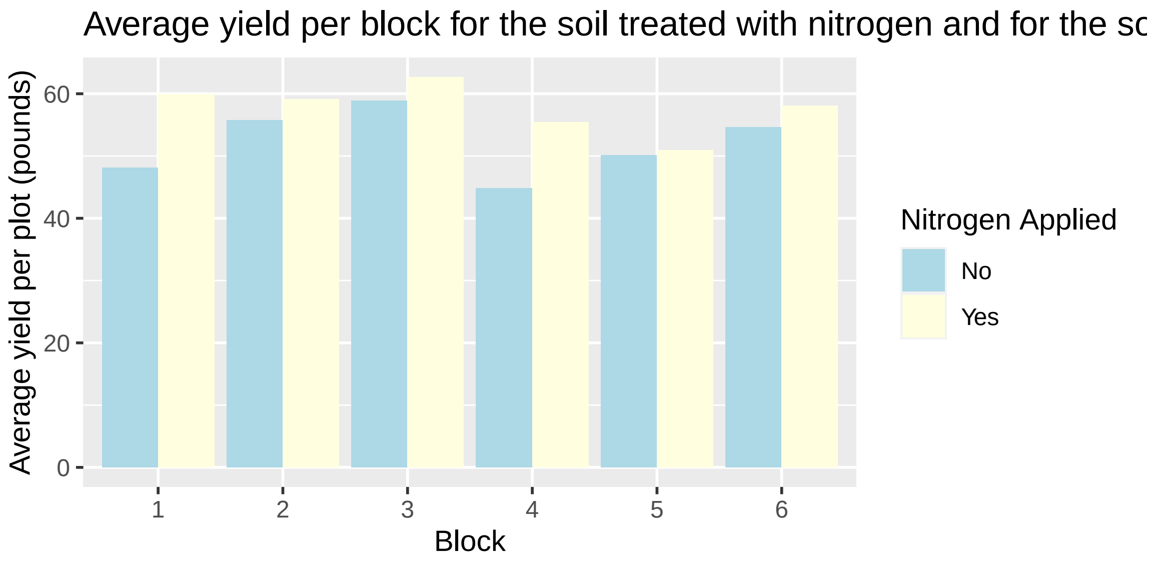
To be more specific, we use the sample function to randomly assign soil additives to each block. We specify the probability of each soil additive with 1 = applied, 0 = not applied, using a vector of length 4 with two 1's and two 0's, which we then randomly sample from using the sample function. We repeat this process for each block, and store the resulting distribution in a matrix called `dist_mat`. We then verify that each soil additive is received by exactly two plots within each block by checking the column and row sums of `dist_mat`. Finally, we combine the block information and the soil additive distribution into a data frame called `npk_dist`.

b) The method of solving this question is to make a plot showing the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen, we can perform the following steps to realize:

```r
library(ggplot2)
npk_dist_agg <- aggregate(yield ~ block + N, data = npk_dist, mean)
```

To begin with, we can calculate the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen by using `npk_dist_agg` with `aggregate()`.

```r
ggplot(npk_dist_agg, aes(x = block, y = yield, fill = factor(N))) +
  geom_col(position = "dodge") +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_discrete(name = "N") +
  labs(x = "Block", y = "Average yield per plot (pounds)", fill = "Nitrogen Applied") +
  scale_fill_manual(values = c("lightblue", "lightyellow"), labels = c("No", "Yes")) +
  ggtitle("Average yield per block for the soil treated with nitrogen and for the soil that did not rece
## Scale for fill is already present.
## Adding another scale for fill, which will replace the existing scale.
```



After calculating the average data, we start to make a plot to realize the same target. This plot is to show the average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen.

Basically, we use the aggregate function to calculate the average yield per block for the soil treated with nitrogen which is `N = 1` and for the soil that did not receive nitrogen which is `N = 0`. We then use the `ggplot2` package to make a grouped bar chart showing the average yield per block for the two groups, with the `x-axis` representing the block and the `y-axis` representing the average yield per plot in pounds. The fill argument is used to group the bars by whether or not nitrogen was applied, and the position argument is used to place the side of the bars by side. We also add appropriate axis labels, a title of "Average yield per block for the soil treated with nitrogen and for the soil that did not receive nitrogen", and a legend. What's more, we define adequate colours for the plot.

The purpose of taking the factor block into account is to control for any variability in yield that may be due to the location of the plots within each block. By comparing the average yield per block between the two groups `N = 1` and `N = 0`, we can see whether there is a difference in yield that is due to the application of nitrogen or to other factors, such as differences in soil conditions or weather.

c) For this question, we conduct a full two-way ANOVA model with the response variable yield and the two factors block and N as follows. To realize this, we use the `aov` function in R. To begin with, we load the required `npk` dataset from the MASS package and then make the structure of the data.

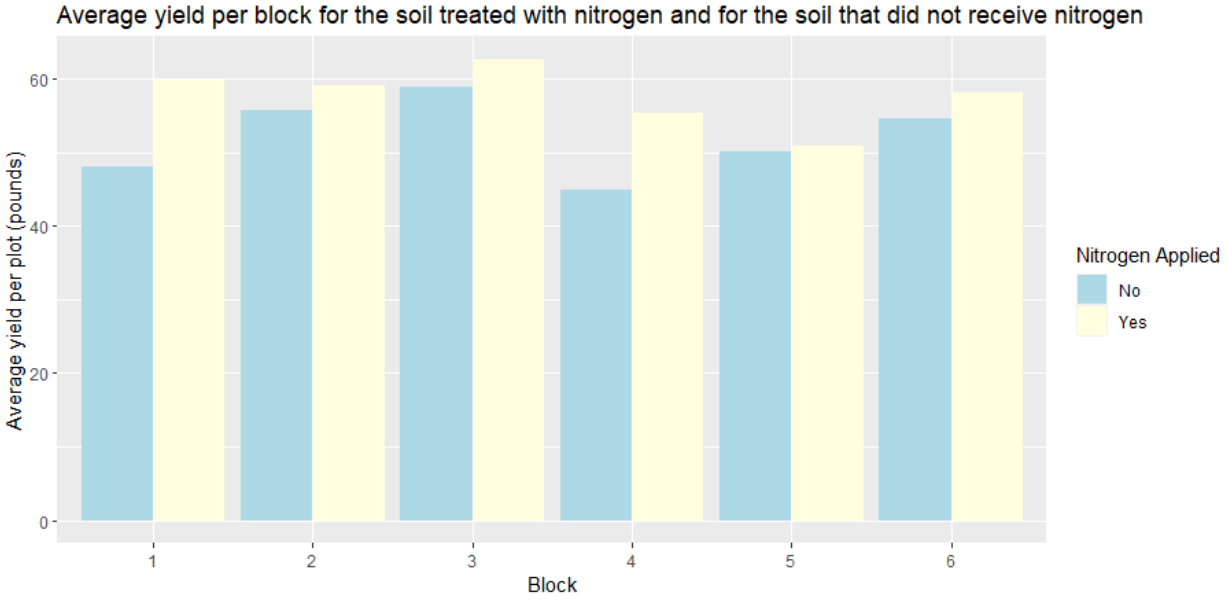


Figure 1: The output plot is as above, displaying the required information in a visualized way.

```
library(MASS)
data(npk)
str(npk)
## 'data.frame':    24 obs. of  5 variables:
##  $ block: Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ N    : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 1 1 2 ...
##  $ P    : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 1 2 2 2 ...
##  $ K    : Factor w/ 2 levels "0","1": 2 1 1 2 1 2 2 1 1 2 ...
##  $ yield: num  49.5 62.8 46.8 57 59.8 58.5 55.5 56 62.8 55.8 ...
```

From the output, we know that the dataset has 24 observations on 4 variables: `block`, `N`, `P`, and `K`. The response variable is yield, which is measured in pounds per plot. The block variable is a factor variable with 6 levels, indicating the block in which the plot was located. The `N`, `P`, and `K` variables are binary variables indicating whether nitrogen, phosphate, or potassium was applied to the plot.

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

The output shows the results of the ANOVA. We can see that both the`block` and `N` variables have significant effects on yield $p < 0.05$. This indicates that both the location of the plot and the application of nitrogen have an impact on the yield of peas. We use the `aov` function to fit a full two-way ANOVA model with the response variable yield and the two factors `block` and `N`. We then use the summary function to view the results of the ANOVA, including the `F-statistic`, `p-value`, and degree of freedom for each factor and their interaction. The Friedman test is not suitable for this situation because the Friedman test is a non-parametric test that is used to compare three or more related groups, and assumes that the data are measured on an ordinal scale. In this case, we have two factors `block` and `N` that are not related, and the response variable `yield` is measured on a continuous scale. Therefore, the assumptions of the Friedman test are not met, and it is not appropriate to use this test for this analysis.

It is sensible and adequate to include the factor block in this model because the block factor represents a grouping variable that is not of primary interest, but that may have an effect on the response variable. By including the block factor in the model, we can control for any variability in yield that is due to the location of the plots within each block, and thus increase the power of the analysis to detect the effect of the `N` factor on yield.

d) This question requires me to give my favorite model and it depends on my choice. To investigate other possible models with all the factors combined, we can use the function to fit a linear regression model. Here are two possible models with one pairwise interaction term:`lm`.

The Model 1 includes the interaction between N and block and the Model 2 includes the interaction between P and block.

```
npk_model1 <- lm(yield ~ N*block + P + K, data = npk_dist)
summary(npk_model1)
##
## Call:
## lm(formula = yield ~ N * block + P + K, data = npk_dist)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.267 -1.608  0.000  1.608  4.267
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.733      2.875  17.648 7.26e-09 ***
## N1            11.750      3.764   3.122   0.0108 *
## block2         7.600      3.764   2.019   0.0711 .
## block3        10.750      3.764   2.856   0.0171 *
## block4        -3.300      3.764  -0.877   0.4012
## block5         2.000      3.764   0.531   0.6068
## block6         6.450      3.764   1.714   0.1174
## P1            -1.183      1.537  -0.770   0.4590
## K1            -3.983      1.537  -2.592   0.0268 *
## N1:block2     -8.350      5.323  -1.569   0.1478
## N1:block3     -8.000      5.323  -1.503   0.1638
## N1:block4     -1.200      5.323  -0.225   0.8262
## N1:block5    -11.000      5.323  -2.067   0.0657 .
## N1:block6     -8.250      5.323  -1.550   0.1522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.764 on 10 degrees of freedom
## Multiple R-squared:  0.8383, Adjusted R-squared:  0.6282
## F-statistic: 3.989 on 13 and 10 DF,  p-value: 0.01731
npk_model2 <- lm(yield ~ P*block + N + K, data = npk_dist)
summary(npk_model2)
##
## Call:
## lm(formula = yield ~ P * block + N + K, data = npk_dist)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.317 -2.038  0.000  2.038  4.317 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   51.083      3.138  16.280 1.59e-08 ***
## P1             4.250      4.108   1.034  0.32528    
## block2         5.750      4.108   1.400  0.19189    
## block3        10.350      4.108   2.519  0.03043 *  
## block4         1.850      4.108   0.450  0.66210    
## block5        -1.250      4.108  -0.304  0.76717    
## block6         4.700      4.108   1.144  0.27927    
## N1             5.617      1.677   3.349  0.00738 ** 
## K1            -3.983      1.677  -2.375  0.03895 *  
## P1:block2     -4.650      5.810  -0.800  0.44212    
## P1:block3     -7.200      5.810  -1.239  0.24356    
## P1:block4    -11.500      5.810  -1.979  0.07596 .  
## P1:block5     -4.500      5.810  -0.775  0.45655    
## P1:block6     -4.750      5.810  -0.818  0.43267    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.108 on 10 degrees of freedom
## Multiple R-squared:  0.8074, Adjusted R-squared:  0.557 
## F-statistic: 3.225 on 13 and 10 DF,  p-value: 0.03535
```

We include the interaction between `N` and `block` in the first model, while in the second model, we include the interaction between `P` and `block`. We also include the main effects of all three soil additives `N`, `P`, `K` and the `block` factor. To test for the presence of main effects of `N`, `P`, and `K`, we can use an ANOVA table to compare the two models with and without the corresponding main effect.

```
npk_model_noN <- lm(yield ~ block + P + K, data = npk_dist)
anova(npk_model_noN, npk_model1)
## Analysis of Variance Table
## 
## Model 1: yield ~ block + P + K
## Model 2: yield ~ N * block + P + K
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)  
## 1     16 429.47                             
## 2     10 141.67  6     287.8 3.3859 0.0433 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To begin with, we make the model without the main effect of N and after that we compare models with and without the main effect of `N`. Similarly, we can test for the main effects of `P` and `K` by comparing the models with and without the corresponding main effect.

Model 1 would be my favorite choice for this given situation, which includes the interaction between `N` and `block`. This model assumes that the effect of `N` on yield may depend on the block where the plot is located. The main effect of `N` is also included, allowing us to test whether `N` has a significant effect on yield, regardless of the `block`. The main effects of `P` and `K` are also included, controlling for any potential effects of these soil additives on yield. I prefer this model because it allows for more flexibility in modeling the relationship between the response variable and the soil additive `N`, which is the factor of primary interest.

e) This question mains to perform a mixed effects analysis, we will use the `lmer` function in the R package `lme4`. This function allows us to model the `block` variable as a random effect. Firstly we load the lme4 package. and get a mixed affects model with `block` as a random affect.

```
library(lme4)
##      Matrix
npk_mixed <- lmer(yield ~ N + P + K + (1|block), data = npk_dist)
summary(npk_mixed)
## Linear mixed model fit by REML ['lmerMod']
## Formula: yield ~ N + P + K + (1 | block)
##    Data: npk_dist
##
## REML criterion at convergence: 128.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.79887 -0.62358  0.05054  0.65032  1.34337
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  block    (Intercept) 13.16    3.628
##  Residual             16.01    4.002
## Number of obs: 24, groups:  block, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   54.650      2.205  24.784
## N1             5.617      1.634   3.438
## P1            -1.183      1.634  -0.724
## K1            -3.983      1.634  -2.438
##
## Correlation of Fixed Effects:
##    (Intr) N1     P1
## N1 -0.370
## P1 -0.370  0.000
## K1 -0.370  0.000  0.000
```

It is obvious the we treat block as a random effect in this code above, allowing for differences in the intercept among the blocks. We include the main effects of all three soil additives: `N`, `P`, and `K` as fixed effects.

Comparing the results from the mixed effects model to the results from the full two-way ANOVA model in the third question, we can see that the estimates of the fixed effects including `N`, `P`, and `K` are similar in both models, but the standard errors and `p-values` are slightly different. The mixed effects model assumes that the variance of the yield response may differ among the blocks, which leads to a larger estimate of the residual variance and smaller p-values for the fixed effects.

Overall, both models suggest that the soil additive has a significant effect on yield. However, the mixed effects model provides a more appropriate way to model the data, by taking into account the variation in yield among the blocks.