Project Report by Asif Hasan (301376671) & Mirrien Liang (301325351)

Section 1. Preamble

Purpose of the study: The purpose of the study is to assess whether it is economically viable to grow two varieties of plant C. sativa, in high humidity greenhouses, compared to low humidity, for full-scale production. It aims to determine the effects of humidity and plant variety on yield.

Treatment structure: The study has two factors: humidity and plant variety; each with two levels. The humidity factor has two levels: high and low. The plant variety factor has two levels: B52 and Northern Light.

Experimental unit structure: The experimental units in this study are the pots within the greenhouses. Each pot can hold up to two seedlings of the same or different varieties, potentially leading to pseudo-replication at the plant level. As we will plant two seeds in a single pot, we will plant seeds of the same plant variety in each pot. Therefore, when analyzing the data, it is important to treat each plant as a separate observation and account for any potential sources of interaction between plants within the same pot to ensure that the observations are independent of each other.

Randomization structure: In this study, the humidity levels will be randomly assigned to the greenhouses, and the plant variety levels will be randomly assigned to the pots within each greenhouse. Additionally, we should also randomly assign pots to different positions within each greenhouse to account for any potential location effects of environmental factors such as sunlight or temperature.

Design considerations: A split-plot design is a natural choice for this study because the humidity level is a whole-plot factor that affects all the pots within a greenhouse, while the plant variety is a sub-plot factor that varies within each greenhouse. Using a split-plot design is efficient because it requires fewer resources than a completely randomized design (CRD) while still allowing for the assessment of the main effects and interactions of the factors of interest.

Section 2. Pre-test

To find the required sample size for the study, we conducted a pilot test that involved 6 greenhouses, with three at high and three at low humidity levels. Within each greenhouse, we used two pots for each plant variety, B52 and NL, and in each pot, we planted two plants of the same variety, which we labeled position '1' and '2' during the harvest.

The model for this study is as follows:

$$y_{i,j,k,l,m} = \mu_i + \beta_j + \gamma_{i,j} + \epsilon^1_{i,k} + \epsilon^2_m + \epsilon^3_{i,j,k,l,m}$$

where:

- $y_{i,j,k,l,m}$ is the yield for each plant
- μ_i is the effect size of the ith level of density; i = {h, l}
- β_j is the effect size of the jth level of plant variety; j = {B52, NL}
- $\gamma_{i,j}$ is the effect size of interaction between factors humidity and plant variety

- $\epsilon^1_{i,k}$ is the error term for the replicates of i; total k replicates of i
- ϵ^2_m is the error term for the number of replicates of plants of the same variety in each pot; m = {1, 2}, 2 replicates of plant in each pot
- $\epsilon^3_{i,j,k,l,m}$ is the error term for each plant; each plant is a unique observation

All three error terms are assumed to be independent and normally distributed, with a mean of zero and a variance of σ^2.

Snippet of ANOVA test:

```
Error: position
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  1 0.01308 0.01308
Error: greenhouse:humidity
          Df Sum Sq Mean Sq F value Pr(>F)
humidity   1 0.8446  0.8446   1.748  0.412
Residuals  1 0.4832  0.4832
Error: Within
               Df Sum Sq Mean Sq F value Pr(>F)
humidity        1   5.93   5.935   4.016 0.0517 .
plant           1   0.63   0.629   0.425 0.5179
humidity:plant  1   0.13   0.128   0.087 0.7699
Residuals      41  60.59   1.478
```

The results of the pilot study provide us with three important statistics:

1. The standard deviation (SD) in yield between greenhouses: $\sqrt{0.4832}$
2. The SD in yield between plants within a pot: $\sqrt{0.01308}$
3. The SD deviation in yield within the same greenhouse: $\sqrt{1.478}$

Since we know that the square root of the residual mean square of ANOVA is the pooled SD, these statistics will be useful in generating the error terms for the power study simulation.

With this in mind, we created two balanced designs similar to the pilot test but involving 8 and 10 greenhouses, respectively. Then we simulated the data for the independent variables, where we chose the effect sizes of both humidity and plant variety to be 1 kg or more, as we decided that 1 kg would be the indicator of a biologically significant difference in yield. Using the model: yield = humidity.effect + plant.effect + error1 + error2 + error3 (without an interaction), we simulated the data for the response variables. We chose not to include an interaction term in the simulation because it would make the simulation much more complicated; after all, we are just estimating the required sample size.

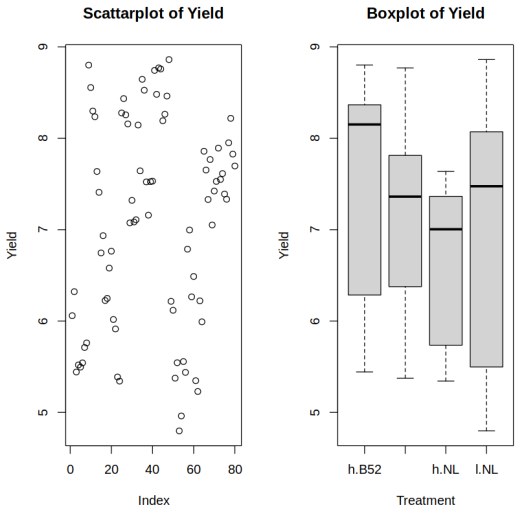
We performed the simulation 10,000 times for each design to calculate the power of the designs using a significance level of $\alpha = 0.05$. Power is the probability of detecting an effect when there is an effect, i.e., when the null hypothesis is false. With 8 greenhouses, we saw a power of 0.30 for detecting the effect of humidity and 0.85 for detecting the effect of plant variety. With 10 greenhouses, we again saw a power of 0.30 for detecting the effect of humidity but 0.93 for detecting the effect of plant variety. *Code for Section. 2 is provided in Appendix B.*

As we did not see any difference in the power of detecting the effect of humidity in the two different models, we assumed that there was no effect of humidity and continued to the full experiment to see the effect of plant variety on yield, as the pilot study suggested it to be significant. Finally, we chose to use the balanced experimental design involving 10 greenhouses, as it provides more power for the experiment, and we have a sufficient budget. *Budget Memo provided in Appendix A.*

Section 3. Full experiment

For the full experiment, we used a balanced design involving 10 greenhouses. We randomly assigned five greenhouses to high humidity levels and the other five to low humidity levels. Within each greenhouse, we randomly assigned two plant varieties to the four pots, with two pots for each plant variety. We planted two seeds of the same plant variety in each pot. Then, we also randomly assigned four pots to different positions within each greenhouse. Finally, when the plants were ready to be harvested, we randomly labeled the two plants (of the same plant variety) within each pot as '1' and '2' and measured their yield. This labeling was done to account for any variation in yield between the two plants planted in each of the 40 pots.

Checking central tendency, variability, and outliers



Comments: Using the scatterplot, we can identify up to two potential outliers at the lower end of the yield values. Then, using the boxplots, it appears that both outliers are in the treatment group "l.NL". Based on the boxplots, it appears that the "h.B52" and "l.NL" groups have greater variability than the "l.B52" and "h.NL" groups, and the pairs appear to have similar variability. Additionally, it indicates that there are overlaps between the values in all four groups. The median yield in the "h.B52" group is the highest, followed by "l.NL", "l.B52", and "h.NL". To check for equal variance across all treatment groups, we will examine the standard deviation of the raw data.

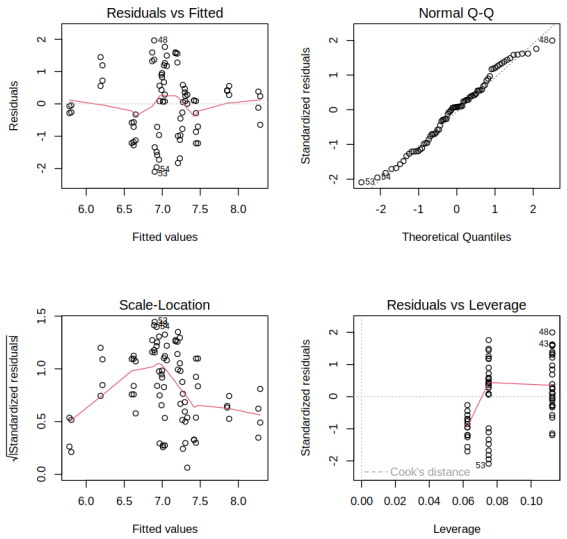
Examining the standard deviation of raw data (checking equal variances)

```
humidity    plant    sd
```

```
h           B52     1.1571228
h           NL      0.8485688
l           B52     1.0259299
l           NL      1.3454974
```

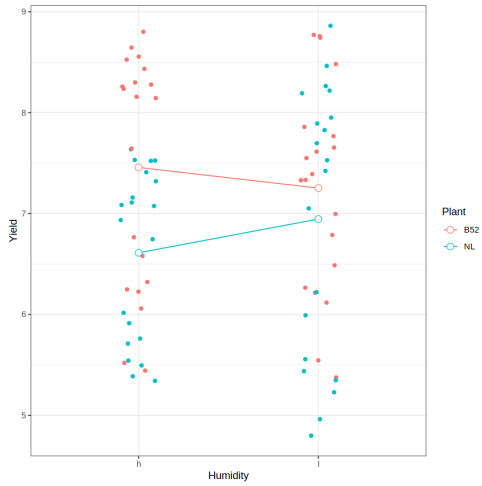
Comments: We know that if $sd_{max}/sd_{min} < 2$, which is true in our case, we can assume that the standard deviation is roughly equal in each treatment group and thus the variance.

Checking assumptions for ANOVA



Comments: (1) Using the Normal Q-Q plot, we can see that the points are roughly aligned along the dotted line. Thus, we can conclude that the error terms are from a normal distribution. (2) Using the Residuals vs Fitted plot, we can see that the points are roughly equally spread on the upper and lower parts of the plot. Thus, we can conclude that the error terms are independent, with a mean of zero. (3) Using the Scale-Location Plot, we do not see any significant pattern. Thus, we can conclude that the error terms have equal variance. (4) Using the Residual vs Leverage plot, we do not see any influential points. This means that the results wouldn't be much different if we either include or exclude them from the analysis. Thus, we can conclude that the errors are also identically distributed, satisfying all assumptions for ANOVA.

Checking for interaction between the factors: Profile plot



Comments: The plot indicates that there might be an interaction between humidity levels and plant variety, as plant 'B52' appears to have a larger mean yield at high humidity levels and a smaller mean yield at low humidity levels, while plant 'NL' shows the opposite trend. However, it is not clear if this interaction is statistically significant. Therefore, we can perform an ANOVA test to examine the effect of each factor (humidity and plant variety) and their interaction on the yield. This will help us determine if the interaction is significant and to what extent it affects the yield.

ANOVA

```
Error: position
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  1 0.01276 0.01276
Error: greenhouse:humidity
          Df Sum Sq Mean Sq F value Pr(>F)
humidity   1  0.003   0.003       0  0.991
Residuals  1 12.492  12.492
Error: Within
               Df Sum Sq Mean Sq F value Pr(>F)
humidity        1   1.11   1.113   1.016 0.3168
plant           1   6.64   6.637   6.058 0.0162 *
humidity:plant  1   1.46   1.455   1.328 0.2528
Residuals      73  79.98   1.096
```

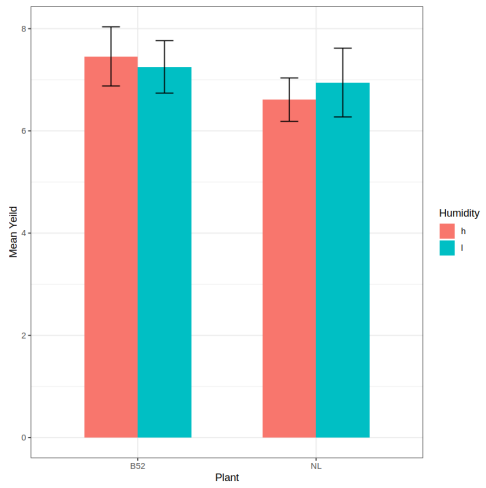
Comments: The ANOVA table shows that only the plant variety has a significant main effect ($p < 0.05$). However, there is no significant main effect for humidity and no significant interaction effect between the two factors ($p > 0.05$). We considered a difference of 1 kg or more to be a biologically important difference, which is approximately 15% of the mean yield under low humidity levels.

To compute the estimated effect sizes, we used the 'lsr' library in R and passed the model: "yield = humidity * plant" as an argument to the 'etaSquared' function. Note that the function does not allow for the addition of error terms, as adding them would result in an incorrectly computed F-stat value

for the model. The results returned by the function are 0.00082, 0.0653, and 0.0143, which are the effect sizes of humidity, plant, and their interaction (humidity:plant), respectively. This is in agreement with the mean square or p-values we saw in the ANOVA table.

Since we have only found a significant main effect for plant variety, we looked at the difference in mean yield between the two plant varieties, which is 0.576. This is lower than our biologically important difference of 1 kg. Note that the mean yield under low humidity levels is larger than the mean yield under high humidity levels.

Summary Plot



Comments: The summary plot displays the mean yield for each combination of humidity and variety, along with error bars that show the standard error of the mean. The plot indicates that plant variety has a significant impact on yield, as B52 yields more than NL in both humidity treatments. The error bars suggest that yield variability is relatively low within each treatment group. Similarly, based on the ANOVA results, there are significant main effects of plant variety on yield, while there are no significant main effects of humidity and no significant interaction between the two factors. *Code for Section. 3 is provided in Appendix C.*

Section 4. Conclusion

Based on the statistical analysis and the results of the experiment, it is evident that plant variety has a significant effect on yield, while humidity does not have a significant effect on yield. There was also no significant interaction between humidity and variety. Therefore, for full-scale production, it may not be economically viable to grow both varieties of plant C. sativa in high humidity greenhouses compared to low humidity greenhouses, as the yield does not seem to be significantly affected by humidity. So, we recommend that future experiments focus on selecting and harvesting high-yield plant varieties to examine which varieties produce higher yields rather than manipulating humidity levels in the greenhouse environment.

Appendices

Appendix A

```
Budget Memo:

    Pre-test Equipment:
        # of greenhouse: 6
        # of pot: 24
        # of plant: 48

    Full Experiment Equipment:
        # of greenhouse: 10
        # of pot: 40
        # of plant: 80

    Cost:
        greenhouse: (10*100)=1,000
        pot: ((40+24)*10)=640
        plant: ((80+48)*1)=128

    Total Cost: (1,000+640+128)=1,768

    Remaining Budget: (2,500-1,768)=732
```

Appendix B

In [ ]:

```
library(tidyverse)
library(ggplot2)
library(lsr)
```

In [ ]:

```
# read pre-test data
data = read.csv("pre_test_data.csv", header=TRUE, sep=',')
data$humidity = factor(data$humidity)
data$plant = factor(data$plant)
data$position =  rep(c(1,2), times=24)
data
# perform anova test
result = aov(yield~humidity * plant +
             Error(greenhouse:humidity + position), data=data)
summary(result)
## residual mean square of ANOVA. Its square root is the pooled SD
## Simulation: showing for experimental design with 10 greenhouses
# set counter
counter1 = 0
counter2 = 0
# set iteration; 10,000
iter = 10000
for(i in 1: iter) {
    # read design
    d = read.csv('design2.csv', header=TRUE, sep=',')
    d$humidity = factor(d$humidity)
    d$plant = factor(d$plant)
    # count number of unique pots
    num.pot = unique(d$pot)
    num.pot = length(num.pot)
    # assign positions {1,2} to plants within each pot
    d$position =  rep(c(1,2), times=num.pot)
    # count number of unique greenhouses
    num.gh = unique(d$greenhouse)
    num.gh = length(num.gh)
    # generate the error terms for between greenhouses
    sd_ = sqrt(0.4832)
    error1 = rnorm(num.gh, mean=0, sd=sd_)
    d$error1 = rep(error1, each=length(d$greenhouse)/num.gh)
    # generate the error terms for between plants within pot
    sd_ = sqrt(0.01308)
    error2 = rnorm(2, mean=0, sd=sd_)
    d$error2 = rep(error2, times=num.pot)
    # generate the error terms for within greenhouses (each plant)
    sd_ = sqrt(1.478)
    num.plants = length(d$plant)
    d$error3 = rnorm(num.plants, mean = 0, sd=sd_)
    # generate effects for humidity, difference is 1 kg or more
    # first half of df contains humidity=h
    # and last half contains humidity=l
    d$humidity.effect[1:num.pot] = 2
    d$humidity.effect[(num.pot+1):length(d$pot)] = 1
    # generate effects for plant variety, difference is 1 kg or more
    # df has 4 B52s followed by 4 NLs and is repeated num.gh times
    plant.effect = c(2,2,2,2,1,1,1,1)
    d$plant.effect = rep(plant.effect, times=num.gh)
```

```
    # generate the yeilds
    # yeild = humidity.effect + plant.effect + error1 + error2 +
    # error3 (no interaction)
    d$yeild = d$humidity.effect + d$plant.effect + d$error1 +
              d$error2 + d$error3
    #analyze result
    result = aov(yeild~humidity * plant +
             Error(greenhouse:humidity + position), data=d)
    x=summary(result)
    x = x$`Error: Within`
    z = unlist(x)
    # check p-val for humidity effect and plant variety efecct
    pval1 = z[17]
    pval2 = z[18]
    # update counter
    counter1 = counter1 + (pval1<0.05)
    counter2 = counter2 + (pval2<0.05)
}
# compute proportions
prop1 = counter1/iter
prop2 = counter2/iter
prop1;prop2
```

Appendix C

In [ ]:

```
# read full-experiment dataset
data = read.csv("full_exp_data.csv", header=TRUE, sep=',')
data$humidity = factor(data$humidity)
data$plant = factor(data$plant)
data$position =  rep(c(1,2), times=40)
# check for outliers and the general spread of data
par(mfrow=c(1,2))
plot(data$yield, main="Scattarplot of Yield", xlab="Index", ylab="Yield")
boxplot(data$yield ~ data$humidity*data$plant, main="Boxplot of Yield",
        xlab="Treatment", ylab="Yield")
# sd across all treatment groups
data %>% group_by(humidity, plant) %>%
summarise(sd=sd(yield), .groups = "keep")
# check for model assumptions
model = lm(yield~ humidity*plant + greenhouse:humidity +
           position, data=data)
par(mfrow=c(2,2))
plot(model)
# check for interactions
plot(data)
interaction.plot(data$humidity, data$plant, data$yield)
# more detailed visualization for interactions
ggplot(data, aes(x=humidity, y=yield, color=plant)) +
       geom_point(position=position_jitter(width=0.1)) +
       stat_summary(fun=mean, geom="line", aes(group=plant)) +
       stat_summary(fun=mean, geom="point", shape=21,
                    fill="white", size=3, aes(group=plant)) +
       labs(x="Humidity", y="Yield") +
       scale_color_discrete(name="Plant") +
       theme_bw()
# analyze data
result = aov(yield~humidity * plant +
             Error(greenhouse:humidity + position), data=data)
summary(result)
# calculate effect size
## ref: https://www.r-bloggers.com/2022/01/how-to-perform-
##      eta-squared-in-r/
model = lm(yield~ humidity*plant, data=data)
etaSquared(model)
# compute difference in mean yeiel between two plant varieties
temp = aggregate(yield ~ plant, data = data, FUN = mean)
abs(temp[1,2]-temp[2,2])
# construct a summary plot
summary_data <- aggregate(yield ~ humidity + plant, data = data,
                          FUN = mean)
summary_data$se <- aggregate(yield ~ humidity + plant, data = data,
                             FUN = sd)$yield / sqrt(4)
ggplot(summary_data, aes(x = plant, y = yield, fill= humidity)) +
  geom_bar(position = position_dodge(), stat = "identity", width = 0.6) +
  geom_errorbar(aes(ymin = yield - se, ymax = yield + se), width = 0.2,
                position = position_dodge(0.6)) +
```

```
  labs(x = "Plant", y = "Mean Yeild", fill = "Humidity") +
  theme_bw()
```