Coding font

Writing font

<https://www.dropbox.com/sh/qj8cxr9q0c4jhyw/AAAih4jziVqueZsnjX8-80YFa?dl=0> (R files)

Objectives & Topics

* Writing hypotheses for ANOVA
* Assessing data for assumptions of ANOVA
* Conducting one-way and two-way ANOVA with Post-Hoc tests
* Using statistical methods to critique the primary literature
* Developing coding skills in the R language

Introduction to the Functions and Tests

This lab utilizes a number of functions from the tidyverse, MASS, and car packages. Before beginning this lab, it is vital to have a familiarity with the plotting functions within ggplot, and the data manipulation functions in dplyr. Students should have a basic understanding of the R language, including how to create variables, using various base R functions such as print(), and use operators such as # and $. Students should also already have an RStudio account and understand how to view files and run code on the software. This activity is meant to be performed within a single day of class.

If students are not familiar with the ggplot and dplyr packages, they can find more information here:

<https://ggplot2.tidyverse.org/>

<https://dplyr.tidyverse.org/>

A manual for using R version 4.1.0 can be found here:

<https://cran.r-project.org/doc/manuals/r-release/R-lang.pdf>

If you are also performing this lab activity in JMP, the resulting models will be different than in RStudio but the conclusions you should reach will be the same.

Introduction to ANOVA

Analysis of variance, or “ANOVA” for short, is a method for comparing three or more means. By using a single ANOVA rather than many t-tests, there are fewer opportunities for a type I error to occur. Differences between particular means can be investigated after the ANOVA (“post hoc”) with a multiple comparison’s procedure such as Tukey’s Honest Significant Difference.

“Analysis of variance” refers to how this method partitions variation in the response variable according to multiple sources:

* In a one-way ANOVA, there is one factor with three or more groups. The only sources of variation are the groups and error.
* In a randomized block ANOVA, there is one factor of interest plus a blocking variable to account for variability across time, space, or even across different researchers. The sources of variation are the groups, blocks, and error.
* In a two-way ANOVA, there are two factors that are “crossed” so that every combination of levels from the two factors is represented. This configuration allows a test to be performed for interaction between the two factors. If no significant interaction is found, then the factors (or "main effects") can then be tested individually. In this case the factors are additive. If there is a significant interaction, then the factors are not additive, and the means for all the combinations of the factors are investigated. The sources of variation are groups for factor A, groups for factor B, interaction, and error.

Interpreting an interaction plot

Interaction plots are used to determine if there is a relationship between two categorical variables and a continuous response variable. If the lines on the graph are parallel to one another, then no interaction occurs. If the lines are not parallel, then there is an interaction, meaning that the effects of one categorical variable depends on the effects of the other. The greater the slope of the lines, the greater the interaction between the variables.

Interpreting the Shapiro-Wilk test

The Shapiro-Wilk test is a statistical test for normality when examining a continuous response variable. It is important that your data does meet the assumptions of normality, as the models and other parametric tests used on this data depend on normally distributed data in order to properly function. If you do not first assess a dataset’s normality, then you run the risk of your results being inaccurate and creating unusable evidence for your hypotheses.

The basis of the Shapiro-Wilk test is that it is a hypothesis test, with the null hypothesis being that the data comes from a normally distributed population, and the alternate hypothesis being that the data is not normally distributed. When looking at the results of the test, the P-value associated with the test statistic determines whether or not there is evidence of non-normality. There has been much debate among statisticians over what an appropriate significance level of the Shapiro-Wilk test should be, with values ranging from 0.01 to 0.25, though for the purposes of this lab you will use 0.25 as your significance level. If a P-value for a line of data is greater than 0.25, then you can assume that the null hypothesis of this test is correct and there is no evidence of non-normality. However, if the P-value is 0.25 or smaller, then there is evidence of non-normality and more tests must be performed.

Interpreting a QQ Plot

A normal quantile, or QQ, plot is a graph designed to help examine if a dataset has a normal distribution. It is a much more visual way to interpret a dataset’s normality than the Shapiro-Wilk test, which can be easier to understand for some people, though it is more subjective. Each point on a QQ plot is one of the residuals (the difference between a value and the mean of all of the values in a dataset), and in a normally distributed dataset, the residuals should follow the black line on the graph (which is a perfectly normal line for the dataset). Because these are real measurements of the early growth of strawberries, the data will not be perfectly normal and you will likely see some small movement away from the black line. So long as the points seem to be approximately linear, then the data can be called normal. If there seems to be a curved pattern to the points, or if some of the points seem to drift away from the line, then there is evidence of a non-normal distribution.

Lab Activity

In this lab you will utilize a published dataset from De Kort et al., 2020 to assess the experimental methods used in the study. The study by De Kort et al. seeks to understand plasticity in the phenotype (change in the appearance or function) of wild strawberry (*Fragaria vesca*) due to environmental conditions. The authors measured many different plant traits and environmental variables as part of their study and used a complex regression approach to understand the relationship between the variables. In order to conduct such a study, the authors needed to ensure equivalency across their treatment groups and plots, ultimately measuring the initial growth of the plants as a proxy measurement for “maternal effects” like seed quality. While the authors report that they controlled for “maternal effects”, they do not provide any evidence that “maternal effects” are similar across sites or treatments.

The dataset you will be using is a reduced version of the data published by De Kort et al., 2020. Rather than exploring all 16 variables that were measured, you will be focusing on three of them: soil moisture treatment level, gradient (the location where the seeds were collected), and the early growth of the strawberries. The original dataset is still available for you to look at (labeled “original\_data.csv” in the “datasets” folder), but for the purposes of this lab you will be using the “reduced.csv” dataset.

Our goal in this lab activity is to test for the presence of maternal effects across sites and treatments in De Kort et al., 2020 by assessing whether the “Early Growth” variable differs between gradients (locations; see Figure 1 in the study) and treatments (dry, normal, wet). You will present your results as a bug-free R script, as well as in the form of a scientific paper.

Main Publication:

De Kort H, Panis B, Helsen K, Douzet R, Janssens SB, and Honnay O. 2020. Pre-adaptation to climate change through topography-driven phenotypic plasticity. Journal of Ecology, 108:4, 1465-1474. <https://doi.org/10.1111/1365-2745.13365>

Published Data:

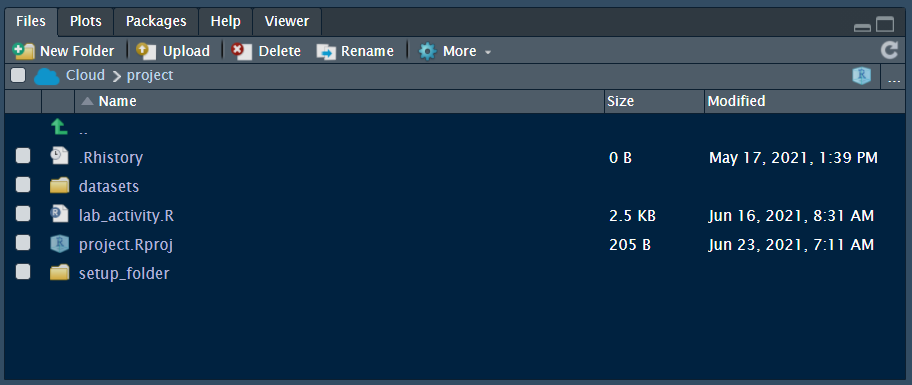
De Kort H, Panis B, Helsen K, Douzet R, Janssens SB, and Honnay O. 2020. Pre-adaptation to climate change through topography-driven phenotypic plasticity. Dryad, Dataset. <https://doi.org/10.5061/dryad.5tb2rbp17>

Summary of Lab Procedure

1. Familiarize yourself with De Kort et al., 2020 and write hypotheses
2. Make figures that visually summarize your hypotheses
3. Test the normality assumption in two-way ANOVA by conducting normality tests
4. Transform the data to achieve multivariate normality
5. Test homogeneity (equality) of variance assumption by conducting Levene tests.
6. Perform the two-way ANOVA
7. Interpret the two-way ANOVA
8. Consider the implications of the two-way ANOVA for De Kort et al., 2020 (this step includes conducting a one-way ANOVA)
9. Develop recommendations on how to treat the data in De Kort et al., 2020 (this step includes post-hoc tests from the one-way ANOVA and a second two-way ANOVA)
10. Write a report about your findings in the format of a scientific paper

Setting up the Data and Exporting Your Script

1. When first opening the project in RStudio, you should see a number of folders and files as shown below.

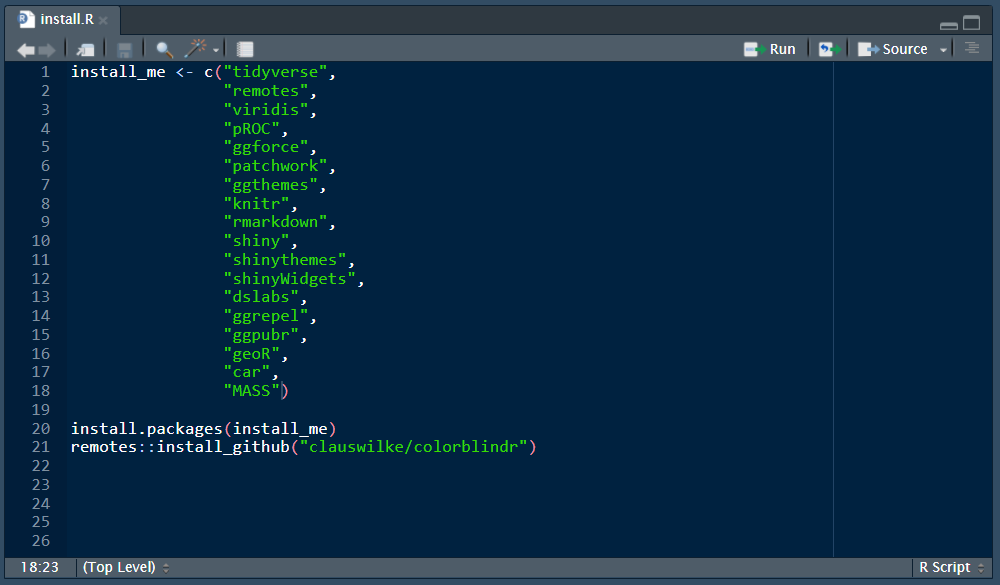


1. If the packages have already been installed for this project, then simply skip down to step 4. If they haven’t, then click on “setup\_folder” and open the “install.R” script.

Text

Description automatically generated

1. Click the “Source” button to run the script, then close it once the packages have finished installing.



1. Return to the main project repository and now open the “lab\_activity.R” script. Regardless of if you did the previous setup step, press the “Run” button four times to run the four lines at the top of the page. This will load the necessary packages and dataset which you will be working with.

library(tidyverse)

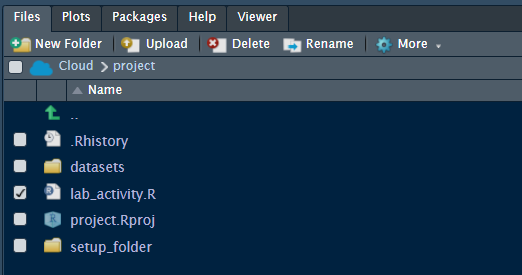
library(MASS)

library(car)

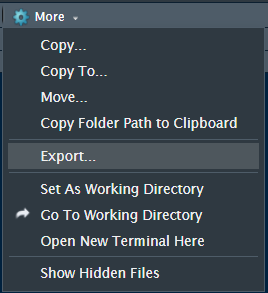
read\_csv("datasets/reduced.csv") -> reduced

Skip to Part 1 after performing step 4. Do not complete these next two steps until after your lab activity has been finished.

1. After your lab activity is complete, you will need to export your script and submit it alongside your paper. To do this, you will first need to click on the white box next to the script, then press “More.”



1. Click on “Export…”, then “Download” when the popup window appears, and save it as an R file to a folder on your computer.



Part 1: Developing Hypotheses

1. Familiarize yourself with the De Kort et al. (2020) study and data. The study can be found here (<https://doi.org/10.1111/1365-2745.13365>). To view the data you will be using, go to the “datasets” folder, then click on “reduced.csv” and “View File.” It will open in a new tab in the Source Pane where you can see the variables and values you will be working with.

Graphical user interface, application

Description automatically generated

1. Moving back to the “lab\_activity.R” script, you need to develop your hypotheses regarding the Early Growth variable. You will write out your null and alternate hypotheses for this experiment in the spaces indicated in the script and save them to variables (there must be 6 hypotheses and variables total; 3 null and 3 alternate). Remember, your goal is to test for the presence of maternal effects across sites and treatments. Hint: we will be using a two-way ANOVA so one set of hypotheses must address the interaction between the predictor, or independent, variables.
2. Once you have finished creating your six variables, use the paste() and print() functions to combine all of your hypotheses into the same string, or line of code with values that are made up of text instead of numbers. An example of how to write the code with only four hypotheses, as well as what it would look like once ran, is shown below (though remember to include your interaction hypotheses in your final code).

HO1 <- "sample null hypothesis 1,"

HA1 <- "sample alternate hypothesis 1,"

HO2 <- "sample null hypothesis 2."

HA2 <- "sample alternate hypothesis 2."

print(paste("Our null hypotheses are that", HO1, “as well as”, HO2, "Our alternate hypotheses are that", HA1, “as well as”, HA2))

--------------------------------------------------------------

[1] "Our null hypotheses are that sample null hypothesis 1, as well as sample null hypothesis 2. Our alternate hypotheses are that sample alternate hypothesis 1, as well as sample alternate hypothesis 2."

Part 2: Exploring the Data

1. After constructing your hypotheses, you should explore the data and create figures that help you visualize the hypotheses you are planning to test. Use glimpse(), summary(), str(), or any other summary function you know to explore the data and print the results.

glimpse(reduced)

print(summary(reduced))

1. Using the ggplot2 package, create two boxplots to visualize the data. The boxplots should be related to your hypotheses and show your response variable (y) plotted against your predictor variables (x). You can add as many details as you want to the plots, but they should be colorful and easy to read. Once you have finished creating your boxplots, save them to variables. An example of how to set up a basic boxplot and save it is shown below.

ggplot(reduced) +

geom\_boxplot() +

aes(x = predictor, y = response) -> figure\_1

1. Use ggsave() to save your figures as pictures. You should save them to the main project repository where the lab activity is. Remember to specify the height and width of your figures to ensure that they will not be stretched out or too small to read. An example is shown below.

ggsave("Figure 1.png", figure\_1, height = 4, width = 7)

1. You should also create and save an interaction plot to visualize the third hypothesis. Run the code shown below with your own text labels added in. You can also add in colors using the argument col = c("color") in the interaction plot function, and specify the colors for each variable in the legend.

png("interaction\_plot.png", width = 7, height = 4, units = "in", res = 100)

interaction.plot(x.factor = reduced$Treatment,

trace.factor = reduced$Gradient,

response = reduced$EarlyGrowth,

fun = mean, type = "b", leg.bty = "o", fixed = TRUE,

trace.label = "Legend label", xlab = "axis label",

ylab = "axis label")

dev.off()

Part 3: Conducting Normality Tests

1. You must first conduct a normality test of the data using a Shapiro-Wilk test. The example code below shows how to perform the test, save it to a variable, and print it. To get full credit, however, you must add an extra column to the dataframe you create from this example code. This column should state whether a given P-value is significant or not in a row (remember that the significance level for the Shapiro-Wilk test is 0.25). Hint: you have to use functions from the dplyr package within the Tidyverse.

reduced %>%

group\_by(Treatment, Gradient) %>%

summarise\_all(.funs = funs(statistic =

shapiro.test(.)$statistic,

p.value = shapiro.test(.)$p.value)) -> SWtest

print(SWtest)

1. You should also create a QQ plot to check whether the residuals of a model are normally distributed. To do this, you must first fit the data to a linear model using the lm() function and save it to a variable. We want to do a full factorial analysis to test the interaction between the predictor variables, so we will multiply the predictors together within the function. An example of how to create and save the model to a variable is shown below.

lm(response ~ Predictor\_1 \* Predictor\_2, data = reduced) -> model

1. Now you can use the residuals from the model to create a QQ plot, and use the print() function to type your interpretation of the plot. An example of how to create the QQ plot is shown below.

qqnorm(model$residuals); qqline(model$residuals)

Part 4: Transforming the Data to Achieve Multivariate Normality

1. You should have discovered in Part 3 that some of the subsets of data are normal while others are not. The best course of action is to conduct a Box-Cox transformation on the data to try meeting the assumptions of an ANOVA. Perform the Box-Cox transformation using the boxcox() function on your model, and save it to a variable, as shown below in the first line of code. From that transformed data, you then run the code shown below in the second line so that you can find the best fit for the transformed data and save that value to a variable (fit). This “best fit” value is the lambda value from the Box-Cox function which is used to specify what power the data should be raised to.

boxcox(model) -> bc

fit <- bc$x[which.max(bc$y)]

1. Create a new model at the best lambda value from the Box-Cox transformation. To do this, instead of only needing to type response ~ predictor into the model function, the response variable should be written as((EarlyGrowth^fit-1)/fit) to account for the transformed data.

lm(((EarlyGrowth^fit-1)/fit) ~ Predictor\_1 \* Predictor\_2, data = reduced) -> new\_model

1. You must also create a new dataset to use moving forward that uses the values from the Box-Cox power transformation for the response variable. To do this, create a new column in the reduced dataset using the new transformed values for the response variable, as shown in the example below, and save this new, modified dataset. You can name the transformed response variable whatever you like, but for the purposes of this lab activity it will be called BC\_Growth here. Before moving on, you should also remove the untransformed EarlyGrowth variable from your new dataset. Hint: You will have to use a function from the dplyr package to remove it, and must type dplyr:: before the function you use to ensure that RStudio knows where the function you want comes from, otherwise it will not work.

reduced %>%

mutate(BC\_Growth = ((EarlyGrowth^lmda-1)/lmda)) -> new

1. After creating your new dataset with the Box Cox transformed variable, perform the steps from Part 3 again. Use the print() function to type a comparison of your results. Did the Box-Cox Y transformation improve the multivariate normality of the data?

Part 5: Testing for Homogeneity

1. Now you will assess the transformed response variable for homogeneity, or equality of variance, using a Levene Test. The Levene Test is used to test the null hypothesis that the distributions being compared have equal variance. The significance level of the Levene test is 0.05, so values lower than 0.05 will cause you to reject the null and you will have failed to meet the assumption of homogeneity of variance. If your values are greater than 0.05, then you may proceed with the planned ANOVA since no compelling evidence of unequal variance was found. To perform the Levene test, use the leveneTest() function with the same response and predictor variables as you used to create your linear models (though remember to use the new transformed responsible variable), then save it to a variable, as shown in the example below.

leveneTest(response ~ Predictor\_1 \* Predictor\_2, data = new) -> levene

1. Using the print() function, extract and print the P-value from the test (hint: you’ll need to use $ as well). Then on a separate line, say whether or not the P-value is significant and if you have met the assumption of homogeneity.

Part 6: Conducting the Two-Way ANOVA

1. Having passed both the normality tests and the test of homogeneity of variance, you can now proceed with building the ANOVA. First create a linear model with the Box Cox transformed response variable and save the model to a variable, as shown below.

lm(response ~ Predictor\_1 \* Predictor\_2, data = new) -> new\_model

1. The function aov() creates an ANOVA from a linear model as shown below. Print the results of this function along with a summary of the ANOVA.

aov(new\_model)

Part 7: Interpreting the ANOVA

1. Interpret your hypotheses from the summary of the ANOVA using the P-value and a significance level of 0.05. In three separate lines, state whether you can reject or fail to reject the three null hypotheses and include what the P-value is for each of the predictor variables.

Part 8: Considering the Big Picture and Next Steps

1. The analysis revealed that there is a significant interaction between the treatment and gradient for the maternal effects variable. In this case, as you are testing whether the authors have successfully controlled for maternal effects, you *want* P-values that are greater than the significance level. The significant result for the gradient is not that problematic taken individually, but the interaction is problematic because it interferes with the authors’ ability to learn about the effect of their soil moisture level treatments on the plants.
2. There is evidence that De Kort et al. (2020) might have problems with their analysis. The next step is to see if these problems can be fixed using post-hoc tests to guide you. Because the treatment variable was not significant, you can focus on the gradient variable and conduct a One-Way ANOVA. First, create a model using the same Box Cox transformed response variable, but only use the gradient predictor variable, and save that model to a variable, as shown below.

lm(response ~ Predictor, data = new) -> gradient\_model

1. Perform the same steps as in Q. Create the One-Way ANOVA from your new model, then print the results and the summary.

Part 9: Removing Data and Testing an Alternative Dataset

1. As expected, the One-Way ANOVA is significant, so a post-hoc test must be conducted to determine where the significance lies. Conduct a Tukey HSD test using the HSD.test() function with the One-Way ANOVA you made. Be sure to specify that you have an alpha level of 0.05 in the function, as shown below. Print the results of this post-hoc test.

HSD.test(gradient\_model, "Gradient", alpha = 0.05, group = TRUE, main = "Please help", unbalanced=TRUE, console=TRUE)

1. The results from the Tukey HSD test suggest four options:
   1. Remove Grave (173 rows of data removed)
   2. Remove Axat (103 rows of data removed)
   3. Remove Grave and Montoulieu (576 rows of data removed)
   4. Remove Sourroque and Axat (679 rows of data removed)

Choose one of these options and print your choice in the script.

1. In the rest of this script, you will be exploring your choice and report on the effectiveness of it. Use what functions you need from the dplyr package to remove the necessary rows of data. You must conduct a new Box Cox transformation, reassess the ANOVA assumptions of normality, and conduct a new Two-Way ANOVA. For each of these tests, you should print an interpretation of your results.

Part 10: Writing your Paper

1. Based on what you found after choosing one of the four options in W, write a formal scientific paper explaining why this study and your work matters, and what your results were. Be sure to give your reasoning for why you chose the option you did in part W, and an analysis of your results from part X. The length of the paper is not important, but your paper should be detailed and clear.
   1. For the introduction:
      1. You should briefly discuss the study by De Kort et al., 2020 and pay particular attention to the maternal effects in question.
      2. There should be a clear statement on how/why a two-way ANOVA will help you understand the maternal effects in De Kort et al., 2020.
      3. Include your hypotheses in the introduction.
   2. In the methods section:
      1. Describe how the data you plan to test was collected by De Kort et al. 2020 and where we obtained the data. You do not need to describe the entire study; focus on the parts of the study that are relevant to you.
      2. Describe our statistical methods (do not write your methods in a cook-book fashion; avoid step by step directions).
   3. In the results section:
      1. Include the tables that you made above.
      2. Include the figures that you made above.
      3. The results section contains text explaining your analysis and the result of your hypothesis tests.
      4. Do not explain WHY you got a particular result, but do report WHAT your results are.
      5. This is acceptable: “The Shapiro-Wilk test indicated that it was reasonable to assume normality (p=0.54, α= 0.25)”
      6. This is not acceptable: “The Shapiro-Wilk test indicated the data was normal (p=0.54, α=0.25), which was surprising given that the data was collected in France.”
      7. Use correct statistical terminology when describing your results.
   4. In the discussion:
      1. Broadly summarize the result of your hypothesis tests and explain why you think you obtained these particular results. (That is, what could have caused these results?)
      2. Put our study into a broader context (the “big picture”). Can we trust the results of De Kort et al., 2020? What should De Kort et al., 2020 do with our results?
   5. In the appendix:
      1. Include the lab\_activity.R script as an appendix for your paper. Export your file as instructed in steps 5 and 6 of “Setting up the Data and Exporting Your Script” and submit it as a bug-free R script.