# “Data” project

This is the final project for the subject “Data Acquisition Analysis and Scientific Methods for Life Sciences” shorted as “Data” from the first course of the Master Plant Health at the Universitat Politècnica de València.

In this project the students chose a scientific article and redo all possible statistic analyses and graphics from it, trying different approaches to see if they can improve them. They will use the data included in the article and follow the Material and methods chapter as a guide.

## Group members

Group E:

Arfa Ghias, Ana Belem Garcia, Cann Doan, Fatima Tahir.

## Chosen article

[Insect repellent and chemical agronomic treatments to reduce seed number in ‘Afourer’ mandarin: Effect on yield and fruit diameter](https://www.sciencedirect.com/science/article/pii/S0304423818308069)

The data is based on the research presented in Scientia Horticulturae (Volume 246, 27 February 2019, Pages 437-447), titled Insect repellent and chemical agronomic treatments to reduce seed number in ‘Afourer’ mandarin: Effect on yield and fruit diameter. The dataset includes measurements related to: - Application of insect repellents and chemical treatments. - Seed numbers in ‘Afourer’ mandarin. - Yield and fruit diameter metrics.

## Proposal

This document serves as a preliminary research plan for reanalyzing the effects of agronomic treatments on ‘Afourer’ mandarin. The analysis will help clarify the efficacy and broader implications of the treatments studied.

Objective of Article:

Focuses on studying methods to manage seed production in Afourer mandarins, a variety known for its desirable seedless characteristics, but it can develop seeds under certain conditions due to cross-pollination.

Objective of Our Project:

A. Replicate statistical analyses from the research article using available data to exercise R skills learned in class. B. Verify results presented in the article for seed-related data to confirm accuracy, consistency, and explore potential alternatives in processing and presenting data. C. To extract data independently from an open resource as opposed to a pre-defined datasets.

The analyses you propose are right.

I would add some repeated measures analyses, as you have number of seeds and aphids at three different dates.

### Main Hypotheses

If the seed data analysis is replicated, the results will align with the article’s findings, while revealing opportunities for improved data management and visualisation.

### Analyses Intended

1. Descriptive Statistics
   1. Average and Standard Error: Summarising the central tendency and variability of numerical variables.
   2. Frequency Distribution and Density Curves: Assessing how data is spread across different values. Density curves offer a smoothed visualisation of data distribution.
2. Inferential Statistics
   1. ANOVA (Analysis of Variance): To compare the mean values of variables across different treatments and blocks.
   2. Normality Testing (Shapiro-Wilk Test): To determine if data meets the normality assumption for parametric tests.
   3. Non-Parametric Tests (Kruskal-Wallis Rank Sum Test): To determine the average seed number per fruit achieved in each treatment.
3. Correlation and Regression Analysis
   1. Linear Models: Analyse the effect of independent numerical variables on yield.
4. Categorical Data Analysis
   1. Chi-Squared Test: The Chi-square test showed significant differences in the percentages of the seeded and seedless fruits among treatments
5. Visualization Techniques
   1. Box Plots: Compare seed count distributions across treatment groups.
   2. Bar Charts with Error Bars: Show mean seed count per treatment with confidence intervals.
   3. Scatterplots: Explore relationships between seed counts and continuous variables .
   4. Violin Plots: Visualize both distribution and density of seed counts for each treatment.

## Files description

1. DATA: this file contains the documents of raw data needed for analysis. Mainly excel files.
2. Analysis Inferential.qmd: in this document you will find the inferential analysis made for the number of seeds experiment. There were conducted the following tests:

* ANOVA
* Tukey Test
* HSD test
* Shapiro-Wilk
* Kruskal-Wallis

1. TwoSeedsAnalysis: contains the analysis for the data considering <= 2 seeds as seedless fruits.Is a percentage bar plot.
2. Arfa Data file: It contains the descriptive analysis, Correlation and Regression Analysisand Visualization Techniques which are explained above.
3. freqdist.chisq.qmd : It contains the Categorical data analysis.
4. Results: In this folder you will find the figures/images of plots generated by the codes.

## Protocol

1. The first thing to run is the descriptive analysis, Correlation and Regression Analysisand Visualization Techniques.

quarto::quarto\_render("Arfa Data", output\_format = "html")

1. The second thing is to execute the categorical

quarto::quarto\_render("freqdist.chisq.qmd", output\_format = "html")

1. After, the inferential analysis follows.

quarto::quarto\_render("Analysis Inferential.qmd", output\_format = "html")

1. Run the document for two seeds analysis

quarto::quarto\_render("TwoSeedsAnalysis.qmd", output = "html")

## Methods

The data used in this project originates from the research article “Insect repellent and chemical agronomic treatments to reduce seed number in ‘Afourer’ mandarin: Effect on yield and fruit diameter” published in Scientia Horticulturae (2019). The original data was collected from experiments investigating the effects of different insect repellents and chemical agronomic treatments on seed production, yield, and fruit diameter in ‘Afourer’ mandarin.

Data Structure:

The data includes several variables, such as: Seed count per fruit under various treatments. Yield and fruit diameter measurements. The application of insect repellents and chemical treatments. Data from different experimental blocks and treatment groups. The dataset is organized in Excel format, containing raw data and variables related to each of the above parameters.Our main focus was on number of seeds or seed count per fruit under various treatments.

## Data Analysis:

A. Descriptive Statistics:

We first performed descriptive statistics to summarize the data. This included calculating the average, standard deviation, and standard error for seed counts and other relevant variables. We also created frequency distributions and density plots to visualize how seed counts were distributed across different treatments.

B. Inferential Statistics:

ANOVA (Analysis of Variance) was performed to determine whether there were significant differences in seed counts between treatment groups. Post-hoc tests (Tukey and HSD) were conducted to compare specific treatment pairs. Normality testing (Shapiro-Wilk) was done to assess if the data followed a normal distribution, which is required for parametric tests. The Kruskal-Wallis test was used for non-parametric data analysis, where appropriate.

C. Correlation and Regression Analysis: Linear models were applied to analyze the relationship between numerical variables, such as seed count and fruit diameter.

D. Visualization: We created various plots for data visualization, including box plots, bar charts, scatterplots, and violin plots to illustrate the distribution of seed counts and their relationships with other variables.

E. Categorical Data Analysis: A chi-squared test was performed to determine if there were significant differences in the proportion of seeded vs. seedless fruits across different treatments.

## Tools Used:

The analysis was carried out using R programming language, along with the Quarto framework to generate reports and visualizations. Libraries such as ggplot2, dplyr, and stats were used for data manipulation, visualization, and statistical tests.

## Results

Th results are aligned to our hypothesis as the data obtained after analysis is similar to the article’s findings. The following figures and tables illustrate the results of our analysis:

Figure 1: Box Plot of Seed Counts by Treatment

This box plot shows the distribution of seed counts for each treatment group. The box represents the interquartile range (IQR), with the line inside the box indicating the median seed count. Outliers are marked as points outside the whiskers. Interpretation: This plot highlights any treatment groups that have a notably higher or lower seed count. We can observe whether treatments effectively reduce seed production in the mandarin fruits. Figure 2: Bar Chart of Mean Seed Count with Error Bars

The bar chart represents the average seed count for each treatment group, with error bars showing the standard error. Interpretation: The comparison of the treatment means helps identify whether any treatment leads to a statistically significant reduction in seed count compared to others. Figure 3: Scatterplot of Seed Count vs. Fruit Diameter

A scatterplot showing the relationship between seed count and fruit diameter. Interpretation: This scatterplot visualizes if there is any correlation between the size of the fruit and the number of seeds it contains. Table 1: ANOVA Results for Seed Count by Treatment

This table shows the results of the ANOVA test, including p-values and F-statistics for testing the null hypothesis that all treatment means are equal. Interpretation: If the p-value is less than the significance level (typically 0.05), it suggests that there are significant differences between the treatments in terms of seed count. Figure 4: Chi-Squared Test for Seedless vs. Seeded Fruits

A bar chart showing the percentage of seedless and seeded fruits across treatments, with a chi-squared test result.

## Discussion:

This analysis shows the effectiveness of the treatments in reducing seed formation, with statistical significance indicating whether differences in seedless fruit proportions are due to the treatments.

The primary objective of this project was to replicate the statistical analyses from the selected article and verify the results using our own approach to data handling and analysis. By following the methods outlined in the article, we were able to successfully reproduce the findings related to the effects of insect repellents and chemical treatments on seed production in ‘Afourer’ mandarin.

The descriptive statistics provided a comprehensive summary of the data, and the visualization techniques helped identify trends and outliers. The inferential tests, including ANOVA and Kruskal-Wallis, confirmed that certain treatments had significant effects on seed count, particularly in the reduction of seed numbers in some cases. The correlation analysis revealed interesting relationships between fruit diameter and seed count, suggesting that fruit size may play a role in seed production.

The chi-squared test of categorical data showed clear differences in the proportion of seedless fruits between treatments, supporting the hypothesis that agronomic treatments can significantly reduce seed formation in mandarin fruits.

Overall, the analysis confirmed that the treatments used in the study had a measurable impact on seed production, but there may be potential for improving the treatment methods or applying more targeted treatments to further reduce seed count. Future work could involve experimenting with different concentrations or combinations of treatments, or examining the long-term effects of these interventions on fruit quality and yield.