Analysis of Quinoa Seed Traits

Dataset: Lysine.14.mgG

1066809

Table of contents

1 Dataset and libraries used for the analysis

Below is a list of libraries used in conjunction with the R programming language, Quarto, and Jupyter for the creation of this statistical analysis. All datasets and libraries needed were imported at the start of the quarto project. Data was imported into a GitHub Repository to access it directly. Access the repository. The raw data `data` was selected to include only the traits desired. This data is referred to as `selected_data` in the project code. The dataset `data` was modified to treat missing data as described in Section 3.1. This data is referred in this project as `imputed_data`

```r
#Standard library
library(tidyverse)

#Data visualitzation
library(visdat)

#Table creation
library(knitr)

#Mutiple imputation
library(mice)

#MCAR test
library(naniar)

#Raw data analysis
data <- read.csv(
"https://raw.githubusercontent.com/biotechdesigner/quinoa-analysis-coursework/main/SM_Data
)

#Selected data for Lysine content, Yield and TSW
data1 <- data %>% select(Lysine.14.mgG, TSW, Yield_g)
selected_data <- arrange(data1, -Yield_g)

#Raw imputed data
imputed_data <- read.csv(
"https://raw.githubusercontent.com/biotechdesigner/quinoa-analysis-coursework/main/imputed
)
```

2 Data Description

The dataset contains measurements from 360 quinoa accessions, detailing various seed traits retreived from Craine et al. (2023). The data includes continuous measurements such as lysine content (measured in mg/g), yield (measured in g/plant) and Thousand Seed Weight, or TSW (measured in grams).

3 Analysis Description

This report presents an analysis of seed traits from a quinoa dataset consisting of 360 accessions. The analysis focuses specifically on the `Lysine.14.mgG` trait and its relationship with `yield` and thousand seed weight (`TSW`), although there are more phenotypic traits included in the dataset. An analysis of the dataset was made to tidy the data. Then, the missing values of the dataset were reported and the missing values were implanted using Multiple imputation (MI) using the Predictive Mean Matching method (PMM) because the MCAR (Missing Not At Random) test was made and resulted negative to the lysine content,the percentage of missing data on lysine content was above 10% and also to keep the statistical power of the analysis, so it reduces the chance of false-positive or false-negative conclusions. The article of Craine et al. (2023) eliminates the missing values completely of the analysis, but it should be useful to compare the correlation and distribution analysis with and without MI. PPM was used because imputations are based on values observed elsewhere, so they are realistic and is a recomended method when the data is Missing At Random (MAR) (Li, Stuart, and Allison 2015). Furthermore, a distribution plot of the `Lysine.14.mgG` trait was made with a normality test to see if the data is normally distributed using the Shapiro-Wilk test, which is common choice for normality testing due to its power and performance, particularly in moderate sample sizes like this one. Finally, a correlation and regression analysis was made between the `Lysine.14.mgG` trait and the `yield` and (`TSW`) separately using spearman correlation to account for the non-normal distribution of the data (Sarmento, n.d.). Additionaly, logaritmic values of the lysine content were used to try to normalize the data to some extent. These analysis were made to determine wether Lysine content in quinoa seeds is an influential factor to determine the yield and TSW.

3.1 Missing Data Assessment

Given the format of the missing values in the dataset (every missing data cell was filled with NA), the proportion of missing values of the traits of interest was visualized:

```
vis_miss(selected_data, sort_miss = TRUE)
```

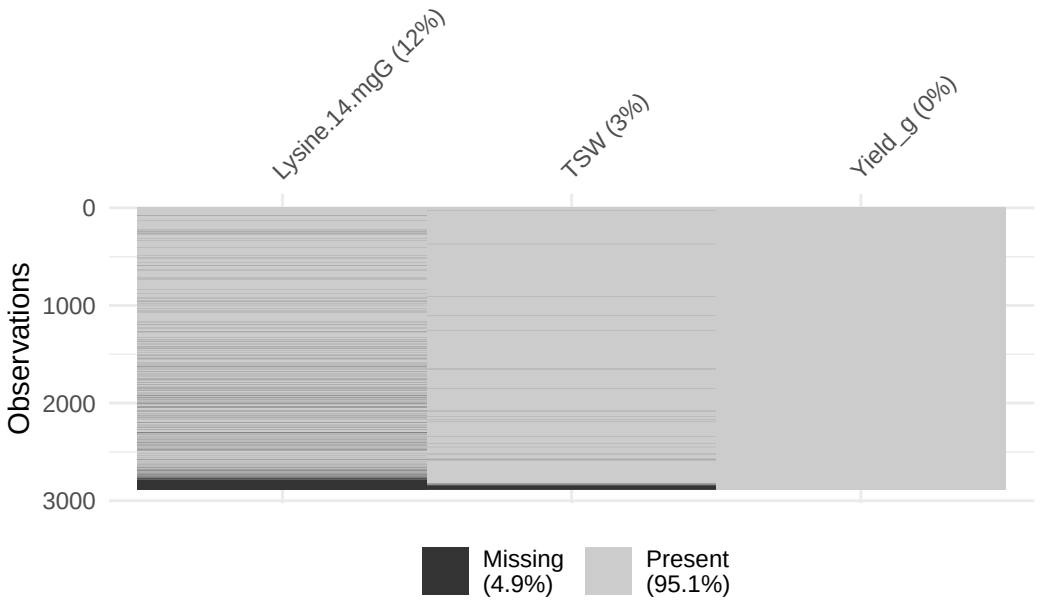

Figure 1: Graphic visualization and percantage of missing data from Yield, TSW and Lysine content

The analysis of missing data from table Figure 1 reveals 12% missing lysine content, 3% in TSW, and no yield data loss. Lee and Huber (2021) suggests using strategies like multiple imputation for increased and varied missing data for precise analyses. An MCAR analysis was performed to assess the feasibility of deleting columns with missing data.

```
#mcar test
result <- mcar_test(selected_data)
kable(result, align = "llcr")
```

Table 1: MCAR test for Yield, TSW and Lysine content

statistic	df	p.value	missing.patterns
428.0913	5	0	4

The MCAR analysis results determined that the p-value is 0, rejecting the null hypothesis and indicating the data are not MCAR. Therefore, as per Bennett (2001), multiple imputation (MI) is recommended for proper data handling. This method will be used in the distribution (Section 3.2), correlation (Section 3.4), and regression (Section 3.5) analyses. Below is the

code for generating the dataset with imputed values in CSV format (See on GitHub). The project now refers to a new table with imputed data as `imputed_data`.

```
#Using MICE library to do the MI using PPM
imputed_data <- mice(data[c("Yield_g", "TSW", "Lysine.14.mgG")], m=5, method='pmm')
completed_data <- complete(imputed_data, 1)

#create csv with the new imputed data:
write.csv(completed_data, "imputed_data.csv")
imputed_data <- read.csv("imputed_data.csv")
```

3.2 Frequency Distribution Analysis

```
#Histogram plot code
ggplot(imputed_data, aes(x = Lysine.14.mgG)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  theme_minimal() +
  labs(title = "Frequency distribution of Lysine Content", x =
"Lysine content (mg/g)", y = "Frequency")
```

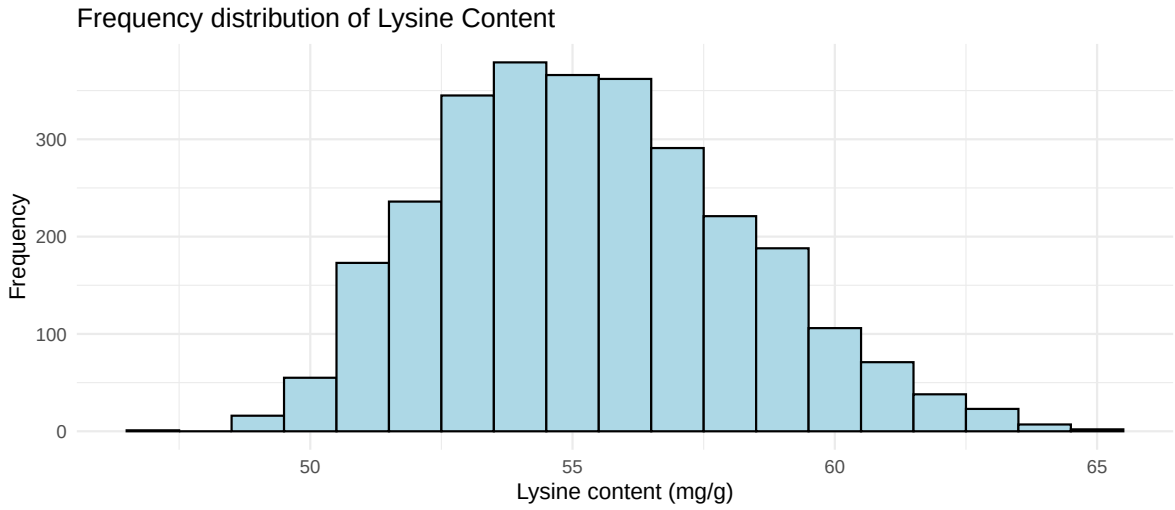

Figure 2: Histogram plot of the frequency distribution of Lysine content

From Figure 2, it is visually inferred that the data seems to be approximately normally distributed with a single peak and symmetric shape, however, it doesnt̓t follow exactly the bell curve, so it is needed to do a nromality test to make sure how the data is distributed.

3.3 Normality Test

```
#Test for normal distribution
normality <- shapiro.test(imputed_data$Lysine.14.mgG)
result <- data.frame(
    W = normality$statistic,
    Pvalue = normality$p.value
)
kable(result, align = "llr")
```

Table 2: Shapiro_Wilk test results for lysine content

	W	Pvalue
W	0.9883829	0

The Shapiro-Wilk normality test has a p-value of almost 0 which is far below any conventional alpha level (e.g., 0.05). despite the high W value, the test finds significant evidence to suggest that the lysine content does not follow a normal distribution. For that reason, this will be needed to take into account the correlation and regression analyses.

3.4 Correlation analysis

```
#Add a column in imputed data with Lysine content transformed to logaritmic values
imputed_data$Lysine.14.mgG_log <- log(imputed_data$Lysine.14.mgG)

#Correlation analysis
cor_value_yield <- cor(imputed_data$Lysine.14.mgG_log, imputed_data$Yield_g,
method = "spearman")

# Correlation plot
ggplot(imputed_data, aes(x = Lysine.14.mgG_log, y = Yield_g)) +
  geom_point(shape = 21, fill = '#0f993d', color = 'white', size = 3) +
  annotate("text", x = Inf, y = Inf, label = paste("Spearman Correlation: ",
    round(cor_value_yield, 5)),
           hjust = 1.1, vjust = 1.1) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Lysine content  log(mg/g)", y = "Yield (g/plant)")
```

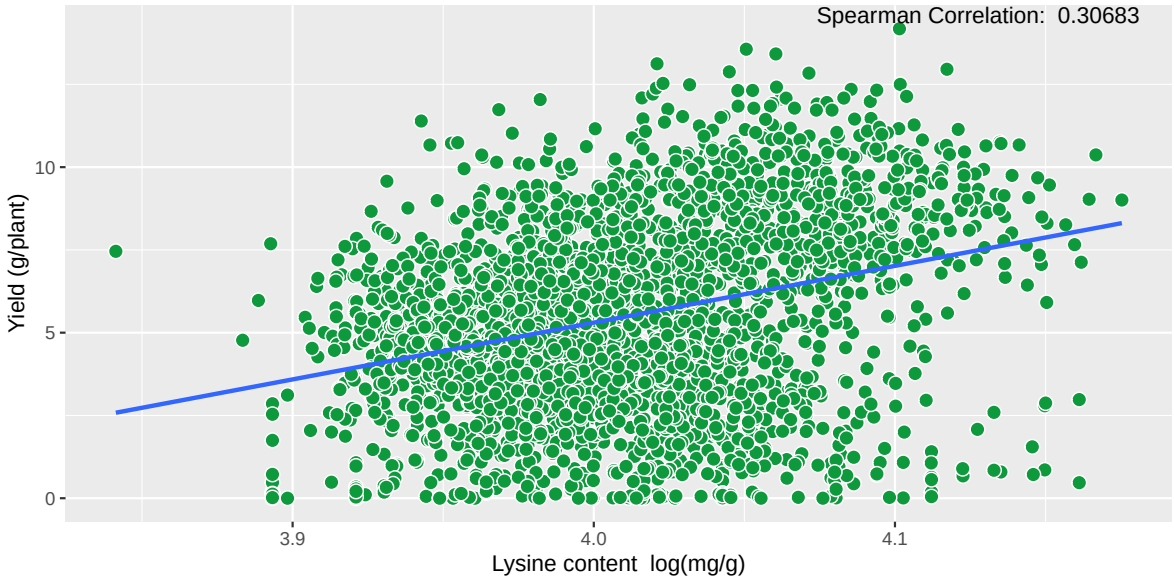

Figure 3: Correlation analysis between Lysine content and Yield traits

3.4.1 Correlation of Lysine content and TSW

```
#Correlation analysis
cor_value_tsw <- cor(imputed_data$Lysine.14.mgG_log, imputed_data$TSW,
method = "spearman")

# Correlation plot
ggplot(
  imputed_data, aes(x = Lysine.14.mgG_log, y = TSW)) +
  geom_point(shape = 21, fill = '#0f993d', color = 'white', size = 3) +
  annotate("text", x = Inf, y = Inf, label = paste("Spearman Correlation: ",
  round(cor_value_tsw, 5)), hjust = 1.1, vjust = 1.1) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Lysine content log(mg/g)", y = "TSW (g)"
)
```

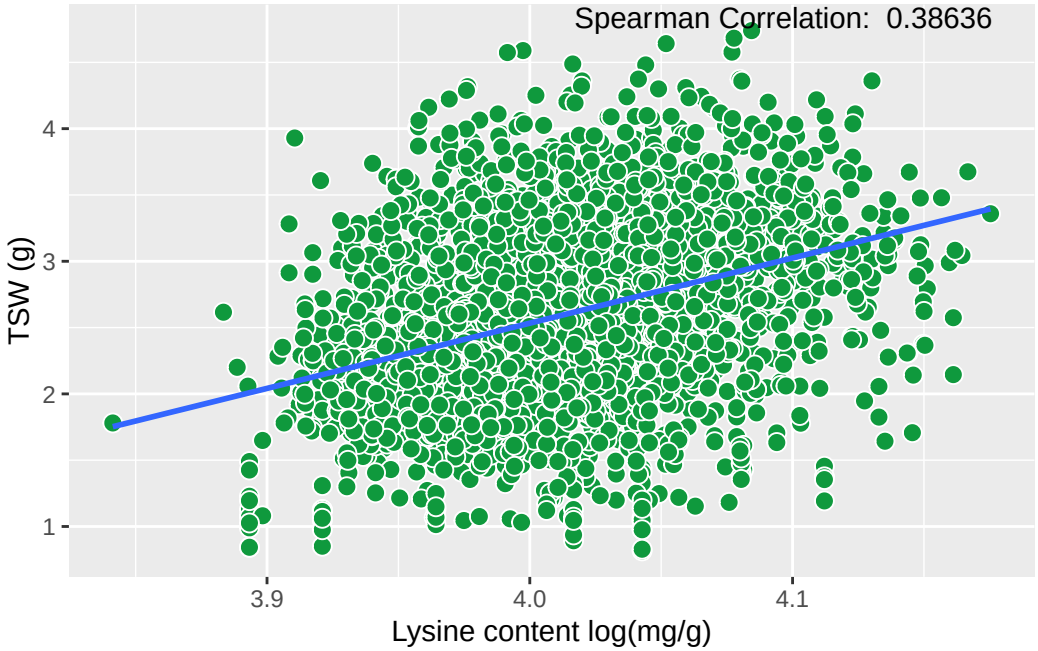

Figure 4: Correlation analysis between Lysine content and Thousand Seed Weight (TSW)

As can be seen, the correlation level between yield variables and lysine content is moderately positive (0.30 with yield and 0.39 with TSW) similar as reported in Craine et al. (2023), indicating that when the lysine content in quinoa seeds is higher, the yield and TSW will also be higher.

3.5 Regression analysis

3.5.1 Regression analysis and ecuation of Lysine content and Yield

This analysis is crucial for understanding the relationship between these two variables, indicating whether higher lysine content correlates with higher yield. In Table 3, it is observed that the p-value is less than 0.05, leading to the rejection of the null hypothesis and the conclusion that there is a significant relationship between lysine content and yield. Furthermore, the R^2 value indicates that approximately 10% of the variability in Yield_g is explained by the model. This suggests that there are other factors influencing the yield that are not accounted for in this analysis. Additionally, this analysis is based on a single value derived from many, so it was expected that lysine content would not explain the entire model. For this same reason, the reported regression equation will not be as precise for determining yield values.

```r
#Regression analysis
reg1_yield <- lm(Yield_g ~ Lysine.14.mgG_log, data = imputed_data)
sum_reg1_yield <- summary(reg1_yield)
broom_yield_summary <- broom::glance(sum_reg1_yield)
knitr::kable(broom_yield_summary, align = "lllllllrr")
```

Table 3: Regression analysis between Lysine content and Yield

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual	nobs
0.0986395	0.0983263	2.678733	314.9512	0	1	2878	2880

```r
#Regression and coefficient analysis
coefs_yield <- coef(reg1_yield)
paste("Y =", coefs_yield[1], "+", coefs_yield[2], "* X")
```

```
[1] "Y = -63.2745685002425 + 17.1445031520236 * X"
```

3.5.2 Regression analysis and ecuation of Lysine content and Thousand Seed Weight TSW

The outcome of this analysis delineates the potential relationship between lysine content and Thousand Seed Weight (TSW). Similar to the previous regression analysis, the null hypothesis can be rejected, leading to the conclusion that there is a highly significant relationship between lysine content and TSW. However, the R^2 value is quite low (13%), indicating that there are other factors affecting TSW that are not considered in this analysis, which was to be expected.

```r
#Regression analysis
reg1_tsw <- lm(TSW ~ Lysine.14.mgG_log, data = imputed_data)
sum_reg1_tsw <- summary(reg1_tsw)
broom_tsw_summary <- broom::glance(sum_reg1_tsw)
knitr::kable(broom_tsw_summary, align = "lllllllrr")
```

Table 4: Regression analysis between Lysine content and Thousand Seed Weight (TSW)

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual	nobs
0.1353708	0.1350704	0.6424025	450.5944	0	1	2878	2880

```r
#Regression and coefficient analysis
coefs_yield <- coef(reg1_tsw)
paste("Y =", coefs_yield[1], "+", coefs_yield[2], "* X")
```

```
[1] "Y = -17.1382340262179 + 4.91783458003395 * X"
```

4 Discussion

The analysis of quinoa seed traits, particularly focusing on lysine content, yield, and thousand seed weight (TSW), provides some insights into quinoa's genetic diversity and agricultural potential. In the initial phase of the analysis, a significant proportion of data pertaining to lysine content was found to be missing in Figure 1, presenting a substantial challenge in the data evaluation process. This issue could arise from a multitude of factors, each with varying implications. Consequently, a decision was made to retain the data rows and employ multiple imputation techniques. This approach aimed to provide an alternative perspective to the analyses previously conducted by Craine et al. (2023). Following the implementation of multiple imputation, the deviation of the dataset from a normal distribution suggested potential specific influences on the yield and Thousand Seed Weight (TSW) of the quinoa seeds as it can be seen in Table 2. Subsequent regression and correlation analyses, focusing on the lysine content in relation to yield and TSW, confirmed the existence of a moderate positive correlation, as seen in Figure 3 and Figure 4. It was observed that lysine content in quinoa seeds positively impacts both yield and TSW. Nonetheless, it is important to recognize that this is not the only influential factor in the model construction, which was confirmed in Table 3 and Table 4 with the R^2 value. In conclusion, a comprehensive analysis incorporating other characteristics within the dataset is essential to fully understand the myriad factors influencing the yield of quinoa seeds.

5 References

Bennett, Derrick. 2001. "How can I deal with missing data in my study?" *Australian and New Zealand Journal of Public Health* 25 (5): 464–69. https://doi.org/10.1111/j.1467-842x.2001.tb00294.x.

Craine, Evan B., Alathea Davies, Daniel Packer, Nathan D. Miller, Sandra M. Schmöckel, Edgar P. Spalding, Mark Tester, and Kevin M. Murphy. 2023. "A comprehensive characterization of agronomic and end-use quality phenotypes across a quinoa world core collection." *Frontiers in Plant Science* 14 (February). https://doi.org/10.3389/fpls.2023.1101547.

Lee, Jin Hyuk, and James Huber. 2021. "Evaluation of Multiple Imputation with Large Proportions of Missing Data: How Much Is Too Much?" *Iranian Journal of Public Health*, July. https://doi.org/10.18502/ijph.v50i7.6626.

Li, Peng, Elizabeth A. Stuart, and David B. Allison. 2015. "Multiple imputation." *JAMA* 314 (18): 1966. https://doi.org/10.1001/jama.2015.15281.

Sarmento, David. n.d. "Chapter 22: Correlation Types and when to use them." https://shorturl.at/lwIZ2 .