Background

Plant Genetic Resources (3502-470)

Karl Schmid, University of Hohenheim

26 May 2025

Table of contents

The purpose of this computer lab is to introduce you to the phenotypic analysis of genebank accessions using a diversity of statistical approaches.

The goal of this analysis is

1. to describe the phenotypic diversity of the genebank material using standard measures of diversity
2. to establish correlations between traits
3. to use multivariate analysis for grouping accessions
4. and to test hypotheses regarding the relationship between traits.

The dataset

For this computer lab, we use a dataset on the characterization of quinoa phenotypic diversity from the Bolivian quinoa genebank.

The dataset is based on 2,815 accessions which were characterized for 55 traits.

The traits are described in a brochure called “Descriptores para Quinua”, which can be downloaded from here: http://www.fao.org/3/aq658s/aq658s.pdf.#Using this brochure for phenotyping, it ensures a standardized description of varieties in different experiments.

The phenotypic dataset can be downloaded from this link: http://181.115.233.146/gringlobal/method.aspx?id=1 as comma-separated file (CSV). We
downloaded the file and it can be found on our PGR website with the file name `web350b.ipsp.uni-hohenheim.de/pgr/data/quinoa-proinpa-2010-phenotypes.csv`.

A description of the phenotype names can be found at: http://germoplasma.iniaf.gob.bo/gringlobal/cropdetail.aspx?type=descriptor&id=439, but the file can also be downloaded from the course website

The dataset has the following features:

- Items are separated by commas. All items are indluded in “ “, e.g. ”105”,“4”, etc.

- Some additional data about the accessions are provided in several columns
- The phenotypic data were generated in a field trial conducted in La Paz/Bolivia in 2010
- There are quantitative and categorical (expressed as numerical scores) phenotypes
- The dataset contains missing data

For the analysis, we first have to prepare the data. For the analysis and the plotting we use the `tidyverse` and `ggplot2` packages.

Data import

We first import all the necessary libraries

```
library(tidyverse) # install if not present.
```

We then import the data into a dataframe.

Important: Adapt the path to a location where your data file is located on the system.

```
df <- read.csv("data/quinoa-proinpa-2010-phenotypes.csv")
```

Determine the amount of missing data.

We first visualize the proportion of missing data for each trait and each accession and filter accordingly to create a dataset with a small proportion of missing data.

> **Tip**
>
> The %>% command is called a *pipe*. You can find details about %>% here: https://r4ds.had.co.nz/pipes.html

```
missing_cols <- df %>% summarize_all(funs(sum(is.na(.)) / length(.)))
barplot(unlist(missing_cols), las=2)
```

First calculate the proportion of NA (missing data) per column)

```
missing_df <- df %>%
  summarise(across(everything(), ~ sum(is.na(.)) / n())) %>%
  pivot_longer(
    cols      = everything(),
    names_to  = "trait",
    values_to = "missing_pct"
  ) %>%
  filter(!trait %in% c("acno", "acp", "acs", "crop", "method_name", "taxon"))
```

Then plot the traits and the missing columns

```
ggplot(missing_df, aes(x = trait, y = missing_pct)) +
  geom_col(fill = "steelblue") +
  scale_x_discrete(limits = rev(sort(unique(missing_df$trait)))) +
  coord_flip() + # make a horizontal barplot
  theme_minimal() +
```

```
  labs(
    x     = "Phenotypic trait",
    y     = "Proportion Missing",
    title = "Proportions of missing values (NA) by trait"
  )
```

The barplot show that 8 traits have a high proportion of missing data. We therefore exclude them from further analysis.

```
drop.cols <- colnames(missing_cols[unlist(missing_cols) > 0.2])
df2 <- df %>% select(-one_of(drop.cols))
```

It is important to know that in the `df2` dataframe the first six columns contain information about the accessions and it is ignored in the following.

List of phenotypes

The phenotypes are grouped into several categories, and two types of data:

- ordinal
- categorical

For both groups, we need different approaches for a statistical analysis.

In the following we want to use the data to identify accessions with a high level of resistance to powdery mildew and use variation in the trait to test the hypothesis whether incidence or severity of mildew infection is related to flowering time, because we expect that early flowering and maturing accessions are more resistant.

Distribution of phenotypic values

Resistance is measures as ordinal values in five groups (see http://germoplasma.iniaf.gob.bo/gringlobal/descriptordetail.aspx?id=44)

Several phenotypes related to flowering time are also ordinal.

We first describe the variation of the four phenological traits related to flowering time to select one that we want to compare with mildew infection.

To make a grid of plots, we use the `GGally` package.

```
library("GGally") # install if not present

# select the phenological phenotypes and put into new data frame
traits.phenological <- c("FIF", "DFL", "FDI", "FMF")
df.phenological <- df %>% select(one_of(traits.phenological))

ggpairs(df.phenological, axisLabels = "none")
```

The resulting grid plot shows that all phenological traits are highly correlated.

> “Exercise
>
> Look at the grid plot and try to explain the results. Which traits are more strongly correlated? Do you have an explanation for this observation?

1 Relationship between mildew infection and flowering time.

In the following use DFL (Days to 50%) flowering as phenotype for testing the hypothesis that severity of mildew is reduced in early flowering accessions.

To test the hypothesis first check whether the distribution of flowering time differs between the clases of mildew infection and then do a formal statistical test.

The plot is a violin plot, which shows the distribution of trait values.

```
# first create a new dataframe with only the two target columns
df3 <- df %>% select(one_of(c("BSM", "DFL"))) %>% drop_na()
df3$BSM <- as.factor(df3$BSM) # convert the mildew persistence as factor

p <- ggplot(df3, aes(x = BSM, y = DFL, fill = BSM)) +
  geom_violin(trim = FALSE)
p
```

The figure does not seem to indicate a difference. To test whether there is a statistical relationship, we test whether the flowering time varies between groups.

```
# Compute the analysis of variance
res.aov <- aov(DFL ~ BSM, data = df3)
# Summary of the analysis
results <- summary(res.aov)
results
```

> Note
>
> 1. Check out whether the trait incidence of mildew (‘BIM’) shows a stronger association with flowering time (days to 50% flowering, DFL)
> 2. Check out whether the trait *days to physiological maturity* (FMF) shows an association with *mildew severity* (BSM)

References