Webinar 2 - the first things to look at . . .

Pietro Franceschi
pietro.franceschi@fmach.it

FEM - UBC

Recap of key ideas

- Variability is unavoidable: we need to measure more and plan well our experiments

Recap of key ideas

- Variability is unavoidable: we need to measure more and plan well our experiments
- . . . unfortunately ***false discoveries*** are always possible

Recap of key ideas

- Variability is unavoidable: we need to measure more and plan well our experiments
- . . . unfortunately ***false discoveries*** are always possible
- The variable we measure are always associated

Recap of key ideas

- Variability is unavoidable: we need to measure more and plan well our experiments
- ... unfortunately ***false discoveries*** are always possible
- The variable we measure are always associated
- We can do good science with relatively low samples

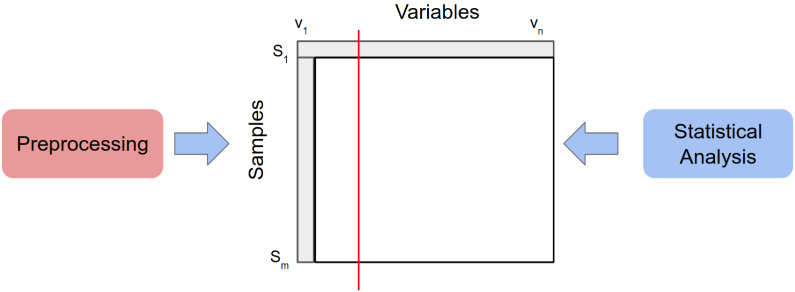
Recap of key ideas

- Variability is unavoidable: we need to measure more and plan well our experiments
- . . . unfortunately ***false discoveries*** are always possible
- The variable we measure are always associated
- We can do good science with relatively low samples
- Data $\leftrightarrow$ (external) knowledge $\leftrightarrow$ complexity of the data analysis strategy

Recap of key ideas

- Variability is unavoidable: we need to measure more and plan well our experiments
- . . . unfortunately ***false discoveries*** are always possible
- The variable we measure are always associated
- We can do good science with relatively low samples
- Data ↔ (external) knowledge ↔ complexity of the data analysis strategy
- There is no such thing as a free lunch . . . **in statistics**

Data Matrix


Preprocessing
Variables
v_1
v_n
S_1
S_m
Samples
Statistical Analysis

IMPORTANT

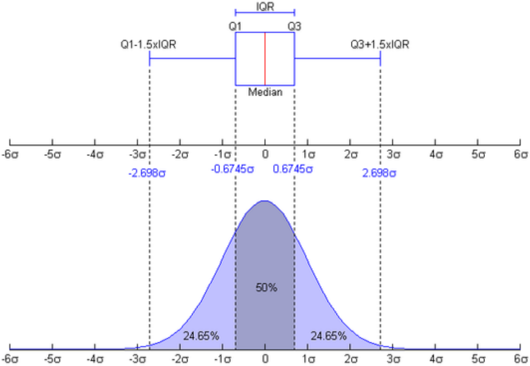
- Variables (columns) are not independent: either for fundamental reasons, or by chance ;-).
- Use domain specific knowledge to asses that!
- In many experimental designs also the samples (rows) are not independent:
  - a person followed over time in a longitudinal study
  - the trees in the same orchard
  - . . .

Multilevel Data Analysis

Choosing the right way to visually inspect the data allows to ....

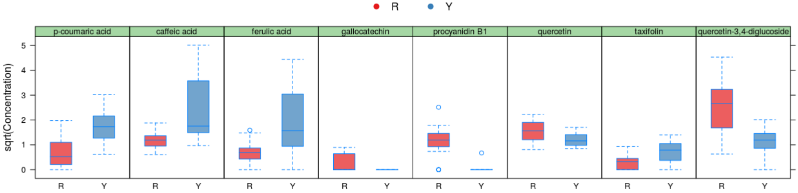
- Check if my experiment is running smoothly
- Identify sub-populations or outliers
- Check the distribution of the data
- Assess the need of variable scaling and sample normalization
- Manage ***missing values***
- Check if my discoveries are there!
- Publish in a better journal ;-)

Univariate Visualization: boxplots

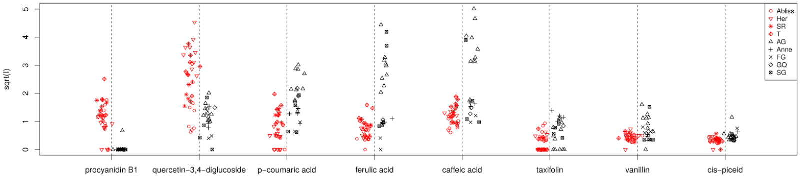


Boxplots nicely summarize the properties of a population, but I need to have a population!

Red and yellow berries



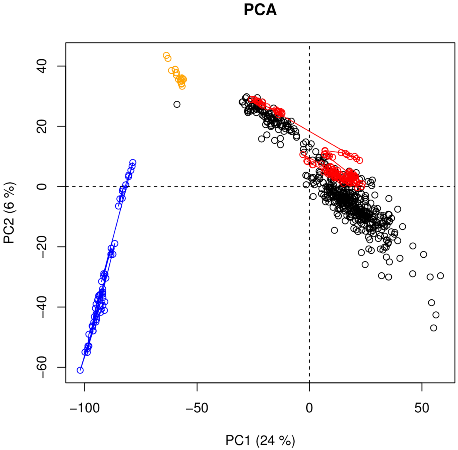
Rubus subpopulations



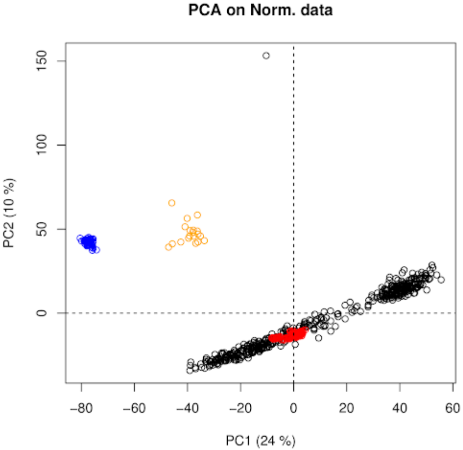
Take home message

- Eventually look to your data ;-)
- In many cases the experimental design is the limiting factor
- Use wisely boxplots and bar plots
- . . . it is unfair to use them to represent three samples . . .

Multivariate visualization: running experiment



Outliers!



Dealing With Outliers

IMPORTANT

- It is unfair to exclude samples which are not in keeping with our theory/hope
- Sometimes outliers are indicators of unexpected and relevant science
- Samples can be excluded if there are **indisputable** evidence that they were “bad” (e.g. analytical errors)
- The best is to keep them in and rely on **robust** data analysis methods (e.g. robust PCA)

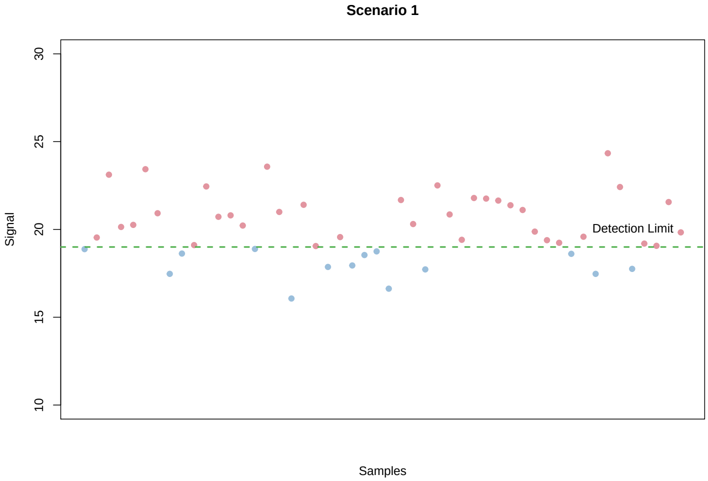
Missing Data

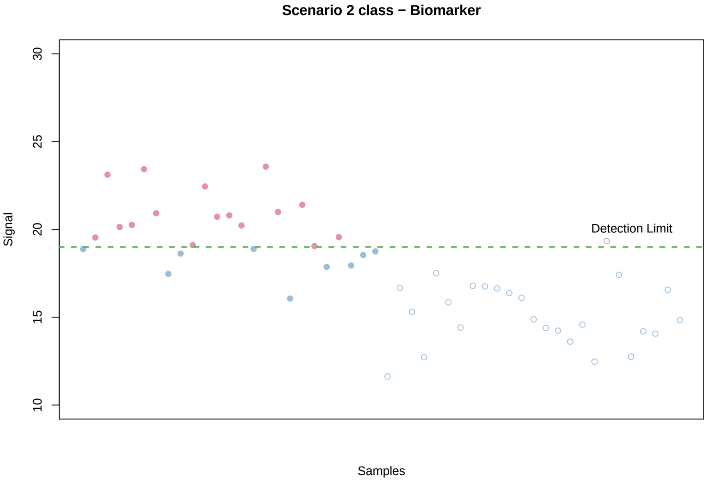
Missing Data

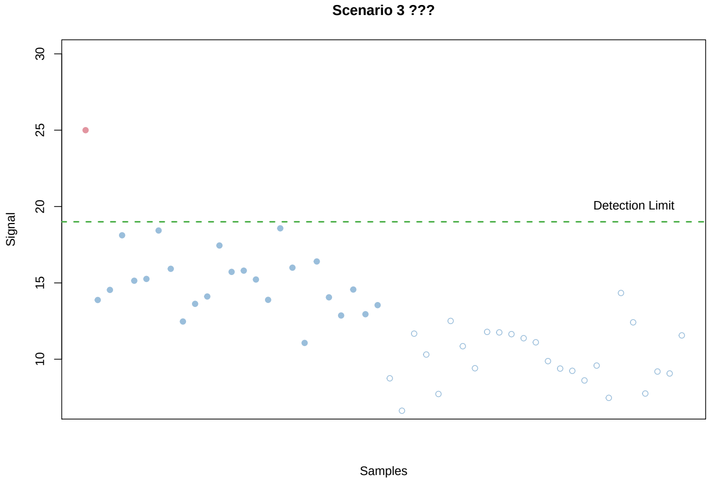
In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. (Wikipedia)

In our *quantitative* scenario:

- Holes in the data matrix
- Often inappropriately filled with **0**
- Preprocessing issue
- Low signal


Scenario 1
Signal
Samples
Detection Limit
10
15
20
25
30


Scenario 2 class – Biomarker
Signal
30
25
20
15
10
Detection Limit
Samples


Scenario 3 ???
Signal
30
25
20
15
10
Detection Limit
Samples

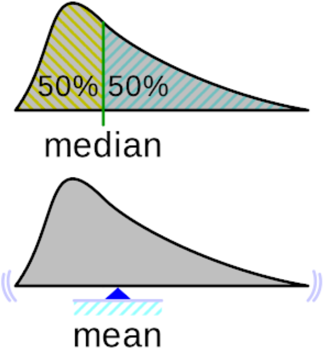
Take home message

- Discard features with too many NAs, but keep in mind the experimental design!
- Use statistical methods able to work with NAs (mainly univariate)
- If you have to **impute** them put a reasonable number with variability (random)
- Use **domain specific knowledge** (e.g random number between zero and the minimum reliable signal)
- **Missing completely at random** (Scenario 1) is easier to handle (also in a multivariate context)
- Try different forms of imputation: are the outcomes sensitive to that?
- To avoid overfitting imputation should be independent from the study factors (e.g control/treatment groups)

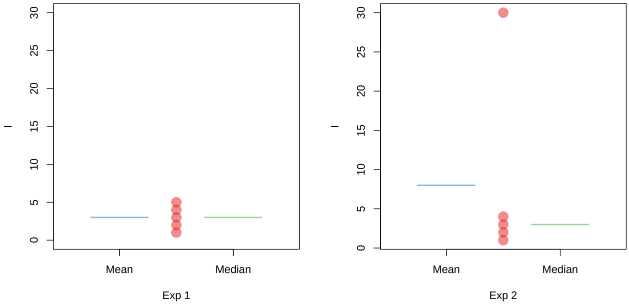
Data Distribution

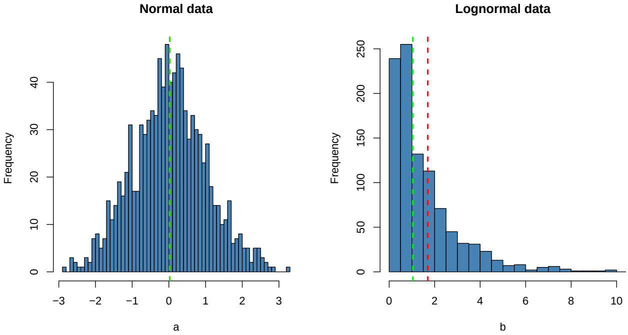
- We measure more than one sample (hopefully)
- **Variability** will make the sample slightly different
- Each property we measure will show some sort of “distribution”
- We assume that the distribution we measure on the sample is related to the distribution of the property in the population

Summarize a distribution: Mean - Median



Mean and median: robustness





. . . would you use the mean in the right case?

Take home messages

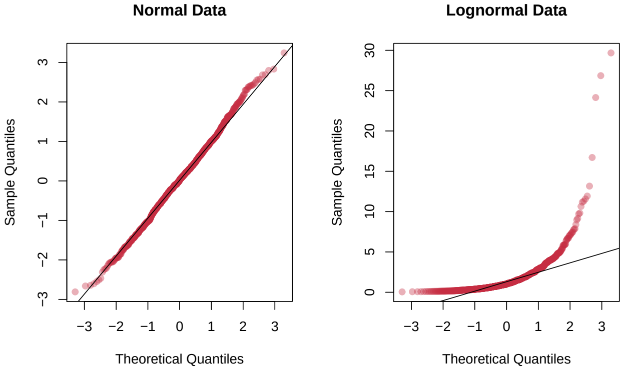
IMPORTANT

- With normally distributed data the mean is the most probable value
- Normality is a prerequisite for many statistical tools
- Statistics likes the mean
- . . . but the mean is not robust
- . . . and many variable we measure are not normally distributed (e.g. counts)

Checking normality

- **statistical tests**: in general unreliable with the typical number of samples we are dealing with
- **quantile-quantile plots**: these graphical tools are really handy to evaluate the distribution of my data. Remember that I need anyway a reasonable amount of samples: 3,5,10 are not sufficient!
- knowledge about data (e.g. are we dealing with counts?) . . .

QQ plots


Normal Data
Sample Quantiles
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Lognormal Data
Sample Quantiles
0
5
10
15
20
25
30
−3
−2
−1
0
1
2
3
Theoretical Quantiles

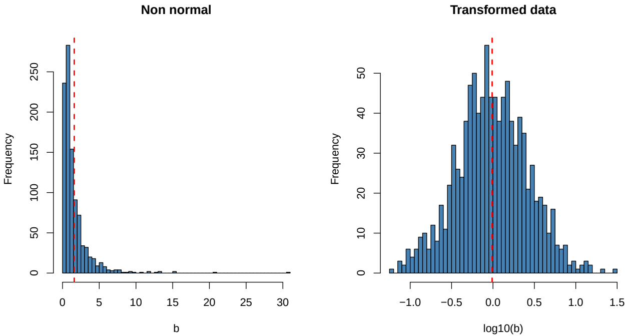
Data Transformations

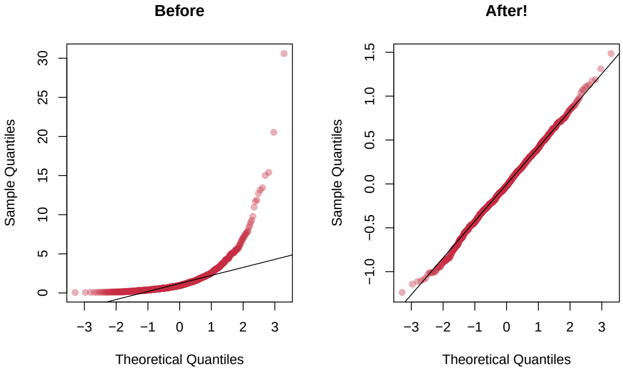
Promoting Normality

Non normally distributed data can be transformed into almost normal data prior to statistical analysis in order to avoid biased results.

- `log` transformation for counts or concentrations
- `arcsin` transformation for percentages
- Box-Cox transformation
- ...

Promoting normality!




Before
After!
Sample Quantiles
Theoretical Quantiles
0
5
10
15
20
25
30
−1.0
−0.5
0.0
0.5
1.0
1.5
−3
−2
−1
0
1
2
3

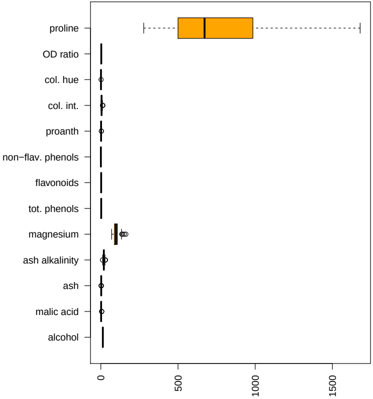
Take home message

- Log has problems with zero!
- Log transformed values can be difficult to digest for the audience
- Generalized Models!
- . . . I almost always log transform . . .

Variable Scaling

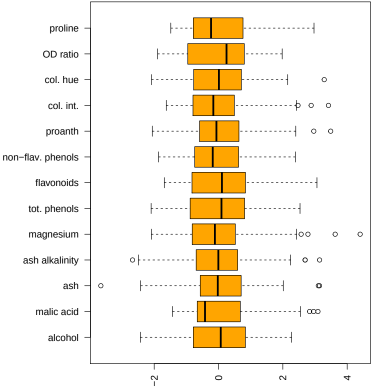
More often than not, the set of variables measured on the samples are highly variable in magnitude. High intensity variables will then determine the shape of the sample cloud in the multidimensional space.

Scaling is the process used to compensate for that



Autoscaling

This is an exceptionally common preprocessing method which uses mean-centering followed by division of each column (variable) by the standard deviation of that column.



Scaling Alternatives

- `log`: compresses high intensity values, but has problems with zeroes
- `sqrt`: here the compression is less severe, but zeroes are allowed
- Remember that autoscaling blows up the noise!

Normalization

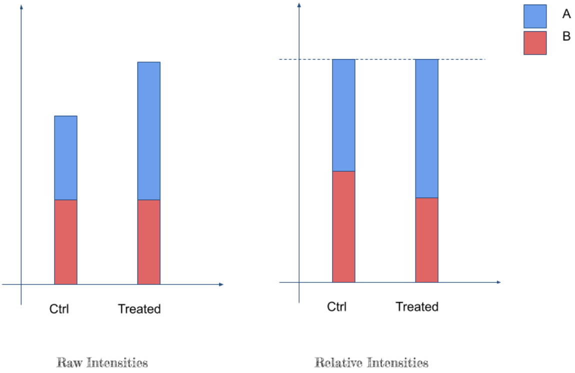
With **normalization** we indicate the process of transforming the intensities of the signal measured in each sample in order to make the samples directly comparable.

- dilution of the sample (e.g. urine)
- different biomass (e.g number of cells)
- amount of DNA in amplification
- ...

Methods

- Matrix specific strategies (e.g. creatinine in urine)
- Housekeeping genes
- Internal standards (proteomics or metabolomics)
- Relative intensities (metagenomics)
- Probabilistic Quotient Normalization

Compositional Data



Normalization has created a new biomarker!