How to perform statistical analysis?

Exploratory data analysis

Etienne Thévenot
Mélanie Pétéra
Jean-François Martin
Gildas Le Corguillé

24/09/2015
Objectives

• Visualize your data
• Detect potential clusters of samples
• Detect potential sample outliers

=> To be performed before any statistical analysis
The "Quality Metrics" tool

- Includes:
  - graphics for data visualization
  - metrics for outlier detection and quality control
    - $p$-values for samples (columns added in `sampleMetadata`)
      - Hotelling's T2 (Mason et al, 1997)
      - intensity distribution (Alonso et al, 2011)
      - proportion of missing values (Alonso et al, 2011)
    - metrics for variables (columns added in `variableMetadata`)
      - coefficient of variation

- Depending on the results:
  - intensities can be log transformed
  - outliers can be discarded
Sacurine dataset

- **Objective:** influence of age, body mass index and gender on metabolite concentrations in urine
- **Cohort:** 183 employees from the CEA institute
- **Analytics:** LTQ-Orbitrap (negative ionization mode)
- **Annotation:** 109 metabolites were identified or annotated at the MSI level 1 or 2.
- **Pre-processing:**
  - XCMS followed by Quan Browser
  - Signal drift and batch effect correction
  - Normalization to the osmolality
  - log10 transformation


http://dx.doi.org/10.1021/acs.jproteome.5b00354
The "Quality Metrics" tool

Summary of the intensities in the dataMatrix

Quality Metrics
- NAs: 0%
- 0 values: 0%
- min: -0.3
- median: 4.3
- mean: 4.2
- max: 6

Thresholds used in plots:
- p-value = 0.001
Exploratory data analysis

Quality Metrics

NA: 0%
0 values: 0%
min: -0.3
median: 4.3
mean: 4.2
max: 6

Thresholds used in plots:
p-value = 0.001

View of the intensities in the dataMatrix
View the PCA scores (check for clusters, outliers)
Check the absence of signal drift (LC-MS)
Check the absence of samples with outlier intensity distribution or outlier proportion of missing values.
Exploratory data analysis

Quality Metrics

- NaNs: 0%
- 0 values: 0%
- min: -0.3
- median: 4.3
- mean: 4.2
- max: 6

Thresholds used in plots:

$p$-value = 0.001

Check the absence of correlation between mean and standard deviation
Using the metrics generated

Filter your data according to the metrics calculated by Quality Metrics module

example for variables:
CV calculation

variable metadata

example for sample metadata
Hotelling P-values
Other tools for unsupervised analysis

• Principal Component Analysis

ACP  Multivariate

• Clustering

Hierarchical Clustering  Heatmap