BIOSTATISTICS WITH R

SUMMER WORKSHOP
(MITGEST network)

24–27 JULY 2024

Maria Chiara Mimmi, PhD

WORKSHOP SCHEDULE

- 4 days
  - 1. Intro to R and data analysis
  - 2. Statistical inference & hypothesis testing
  - 3. Modeling correlation and regression
  - 4. Machine Learning; MetaboAnalyst; Power Analysis

- Each day will include:
  - Frontal class (MORNING)
  - Practical training with R about the topics discussed in the morning. (AFTERNOON)

DAY 4 – LECTURE OUTLINE

- Examples of Machine Learning
  - PCA
  - PLS-DA
- MetaboAnalyst
  - Overview
  - Workflow
- Power analysis
  - Hypothesis testing
  - Decision errors
  - Statistical power
  - Effect size

Principal Component Analysis (PCA)

A type of *unsupervised* learning algorithm for dimensionality reduction

Purpose of PCA

- The goal of PCA is to transform a high-dimensional dataset into a lower-dimensional dataset while retaining as much of the variance in the data as possible.

- Common use cases of PCA:
  1. to reduce the dimensionality of high-dimensional datasets
  2. to visualize the structure of the data
  3. to remove noise and redundant information from the data
  4. as a preprocessing step for other machine learning algorithms

Covariance

Population mean is unknown

$$var(x) = \frac{\sum_i^n (x_i - \overline{x})^2}{N-1}$$

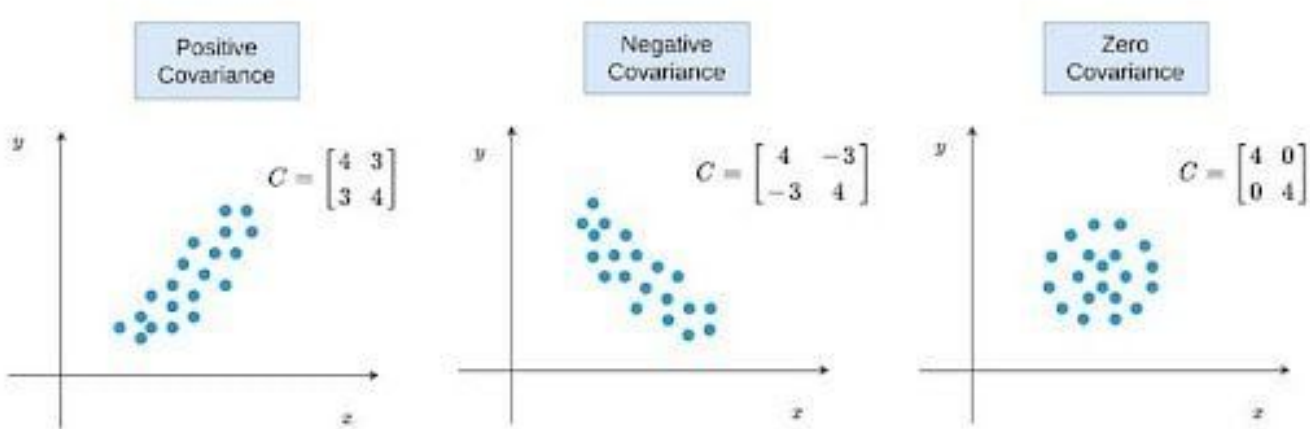
$$\begin{array}{c c} & \begin{array}{ccc} x & y & z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$

Population mean is unknown

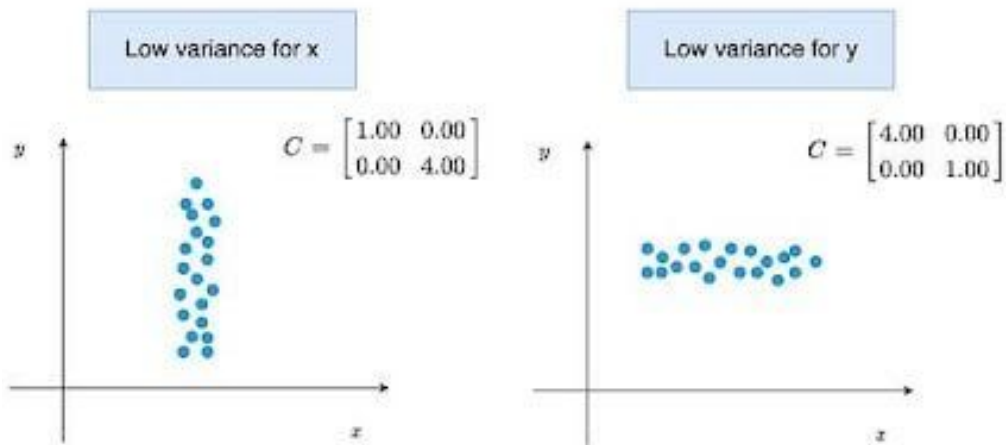
$$cov(x,y) = \frac{\sum_i^n (x_i - \overline{x}) \cdot (y_i - \overline{y})}{N-1}$$

Variance measures how the values vary in a variable. **Covariance** measures how changes in one variable are associated with changes in a second variable.

Covariance



Positive, negative and zero covariance.



Different variances and zero covariance.

Source: https://builtin.com/data-science/covariance-matrix

PCA

PCA originally is a linear algebra operation.

It is a transformation method that creates (weighted linear) combinations of the original variables in a data set, with the intent that the new combinations will capture as much variance in the dataset as possible while eliminating correlations (i.e., redundancy).

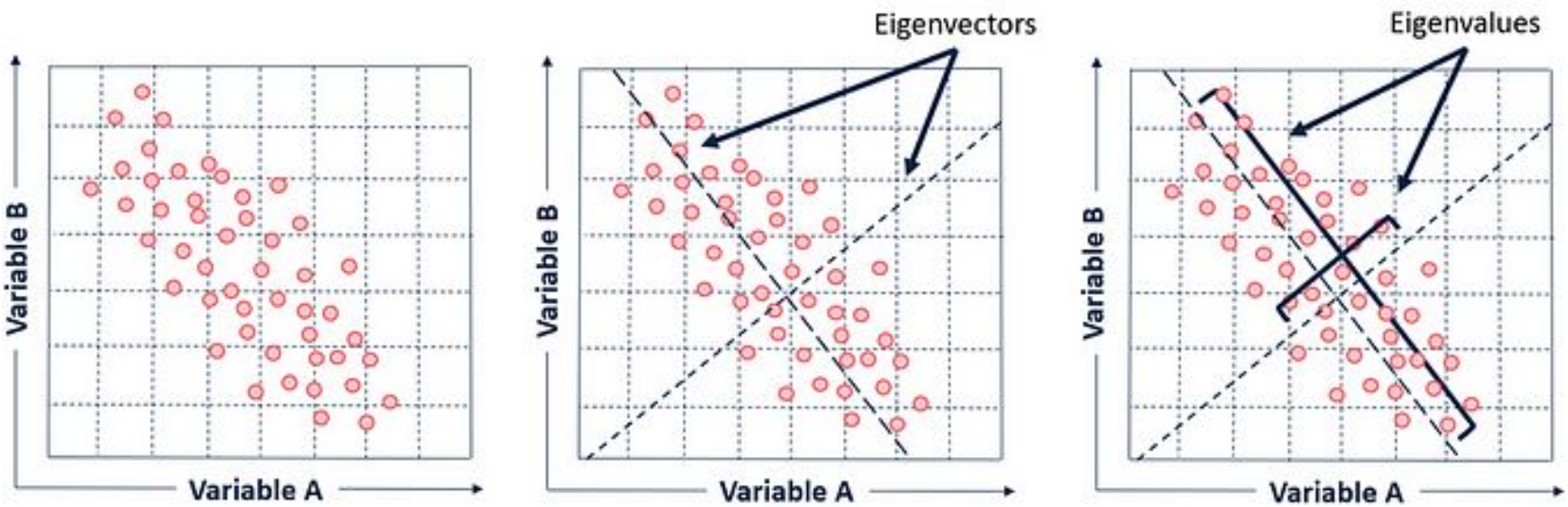
PCA creates the new variables using the eigenvectors and eigenvalues calculated from the covariance matrix of your original variables.

Eigenvectors & Eigenvalues

In the context of PCA

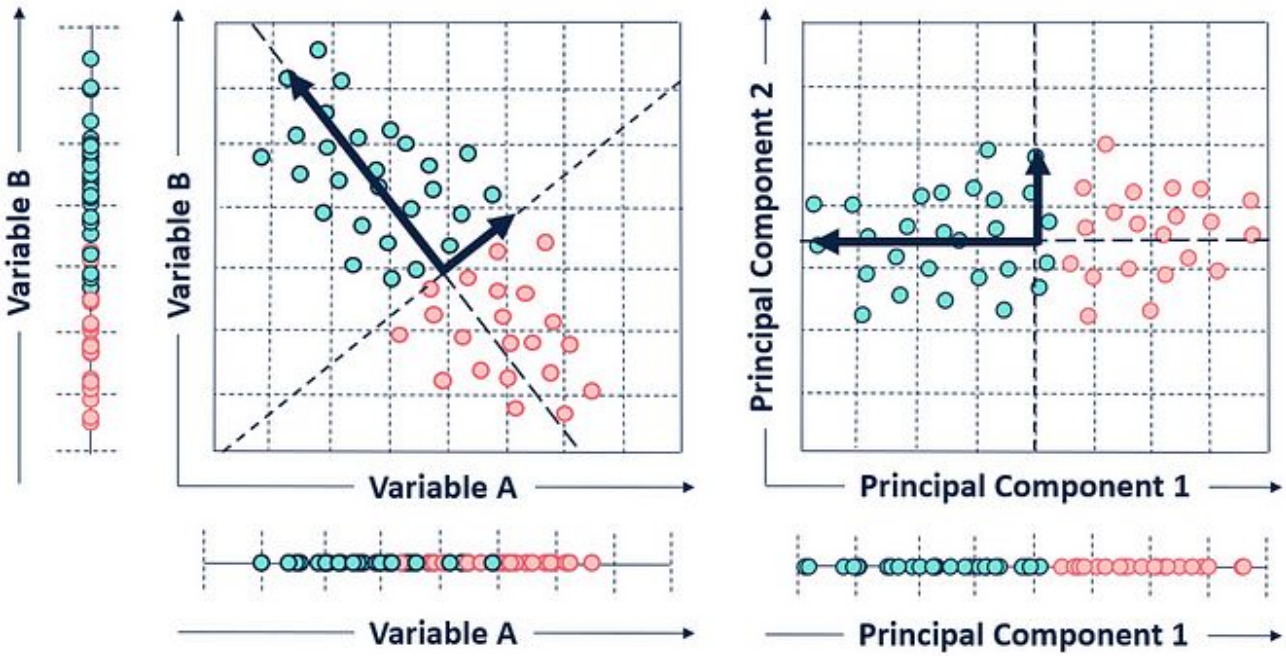
- The **eigenvectors** of the covariance matrix define the directions of the principal components calculated by PCA.

- The **eigenvalues** associated with the eigenvectors describe the variance along the new axis.



Source: https://towardsdatascience.com/tidying-up-with-pca-an-introduction-to-principal-components-analysis-f876599af383

Principal components

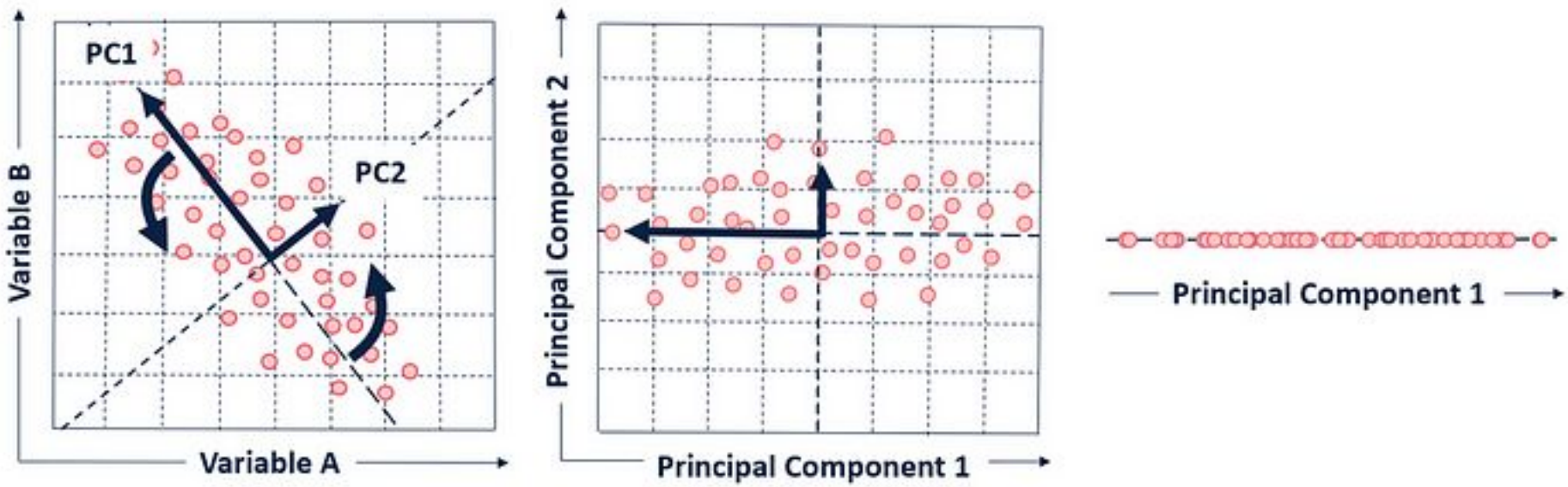


Principal Component 1 accounts for variance from both variables A and B. (dimension reduction)

The principal components (eigenvectors) are sorted by descending eigenvalue. The principal components with the highest eigenvalues are "picked first" as principal components because they account for the most variance in the data.

Source: https://towardsdatascience.com/tidying-up-with-pca-an-introduction-to-principal-components-analysis-f876599af383

Principal components



To convert our original points, we create a projection matrix.
This projection matrix is just the selected eigenvectors concatenated to a matrix. We can then multiply the matrix of our original observations and variables by our projection matrix.
The output of this process is a transformed data set, projected into our new data space — made up of our principal components!

Source: https://towardsdatascience.com/tidying-up-with-pca-an-introduction-to-principal-components-analysis-f876599af383

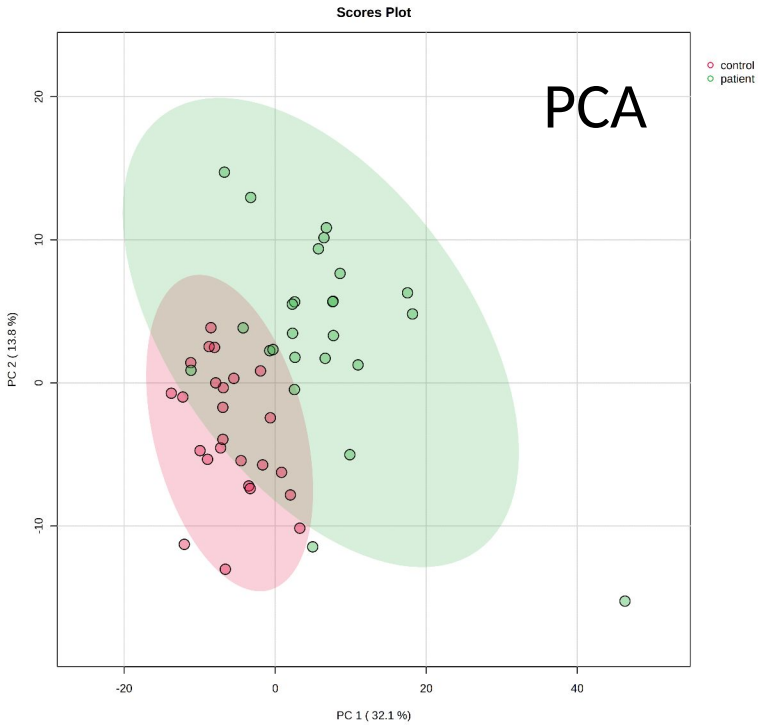
PLS Discriminant Analysis (PLS-DA)

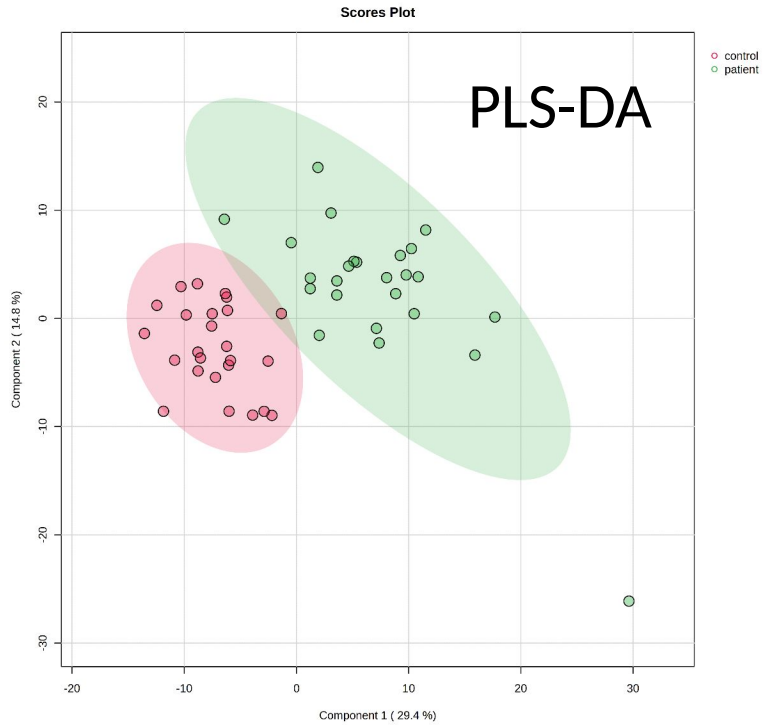
A *supervised* alternative to PCA performing simultaneous dimensionality reduction and classification

Purpose: PLS-DA vs PCA

- PCA is completely unsupervised (i.e. you don't know in advance if there are classes in your dataset)
- In PLS-DA you know how your dataset is divided in classes from the response vector Y. The goal here is then to project the predictors into a space, while maximizing the
- Common scenarios for using PLS-DA: omics sciences.

Scores plot: PCA vs PLS-DA



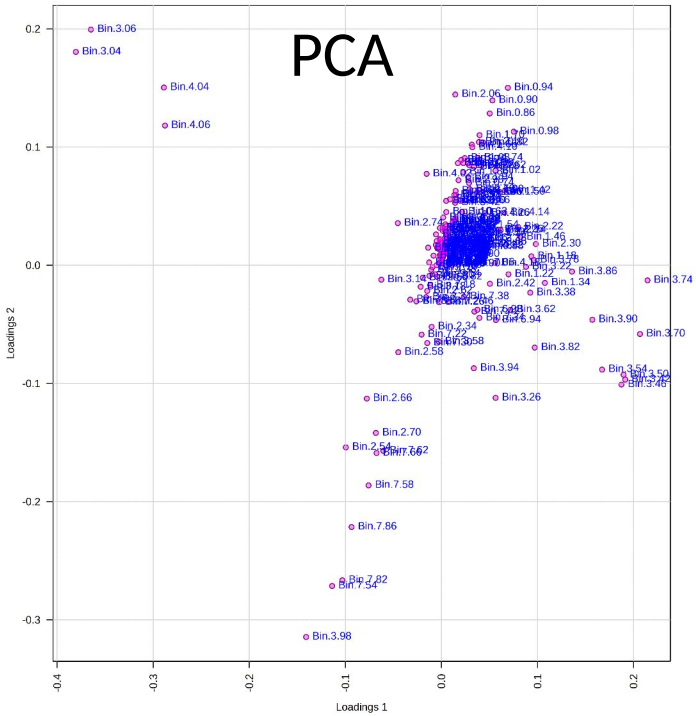


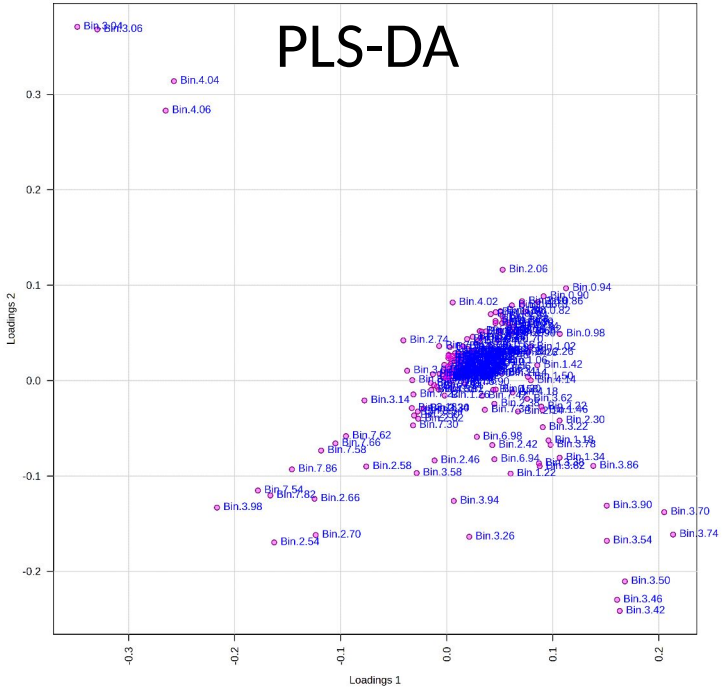
Samples projected in the space of Principal Components

Samples projected in the space of latent variables (components) that maximize the separation between groups

*Source: Test data (**NMR spectral bins**) provided by METABOANALYST platform: https://www.metaboanalyst.ca*

Loadings plot: PCA vs PLS-DA



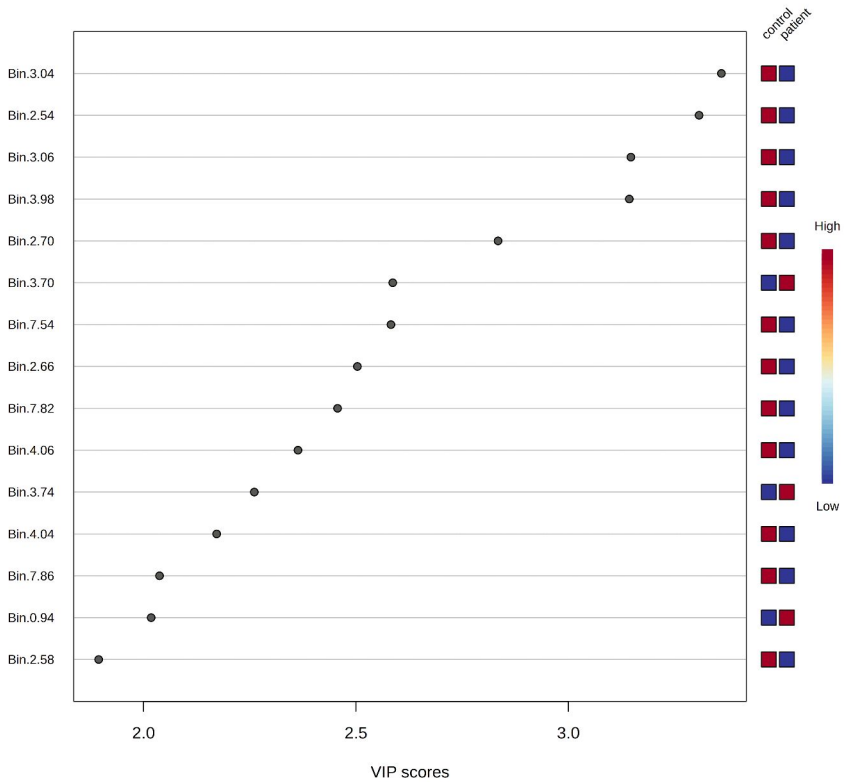


The loading vectors (here shown as points) represent the original variables in the space PCs.

The loading vectors (here shown as points) represent the original variables in the space of latent components retrieved by PLS-DA.

*Source: Test data (**NMR spectral bins**) provided by METABOANALYST platform: https://www.metaboanalyst.ca*

Feature Importance in PLS-DA

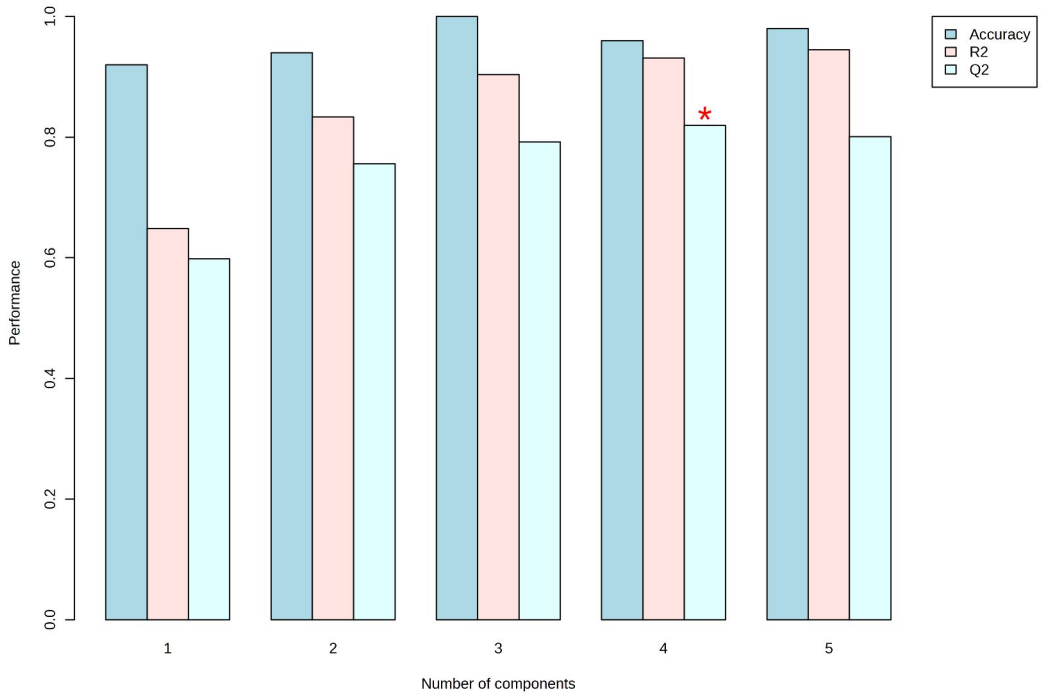


VIP (Variable Importance in Projection) scores, ranking the variables based on their significance in the PLS-DA **model of classification**.

...very useful to select potential biomarkers!

*Source: Test data (**NMR spectral bins**) provided by METABOANALYST platform: https://www.metaboanalyst.ca*

Cross validation in PLS-DA



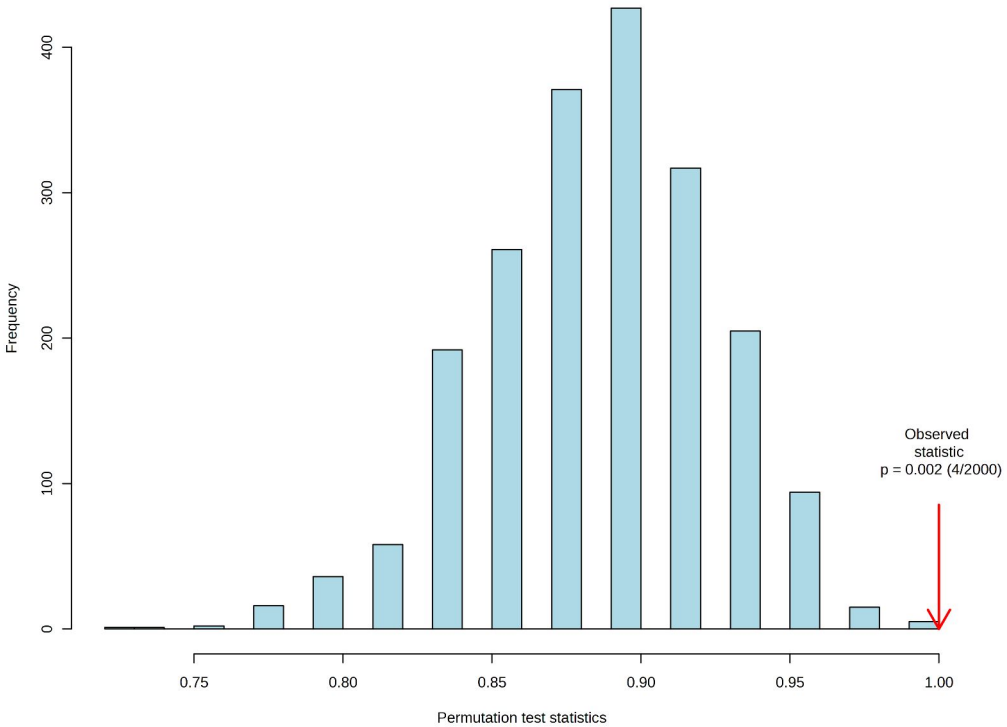
PLS-DA generate a model of classification.

By partitioning the dataset and iteratively testing the model, cross validation estimate the predictive ability of the model.

Q^2 is an analogous of R^2 in regression: the higher the better!

*Source: Test data (**NMR spectral bins**) provided by METABOANALYST platform: https://www.metaboanalyst.ca*

Permutation in PLS-DA



Permutation testing is a non-parametric approach to assess the significance of a model's results.

In the context of PLS-DA, this test helps verify whether the observed classification accuracy is better than what would be expected by chance.

*Test data (**NMR spectral bins**) provided by METABOANALYST platform: https://www.metaboanalyst.ca*

DAY 4 – LECTURE OUTLINE

- Examples of Machine Learning
  1. PCA
  2. PLS-DA
- MetaboAnalyst
  1. Overview
  2. Workflow
- Power analysis
  1. Hypothesis testing
  2. Decision errors
  3. Statistical power
  4. Effect size

MetaboAnalyst

An R-driven Software

Introduction to MetaboAnalyst

https://www.metaboanalyst.ca

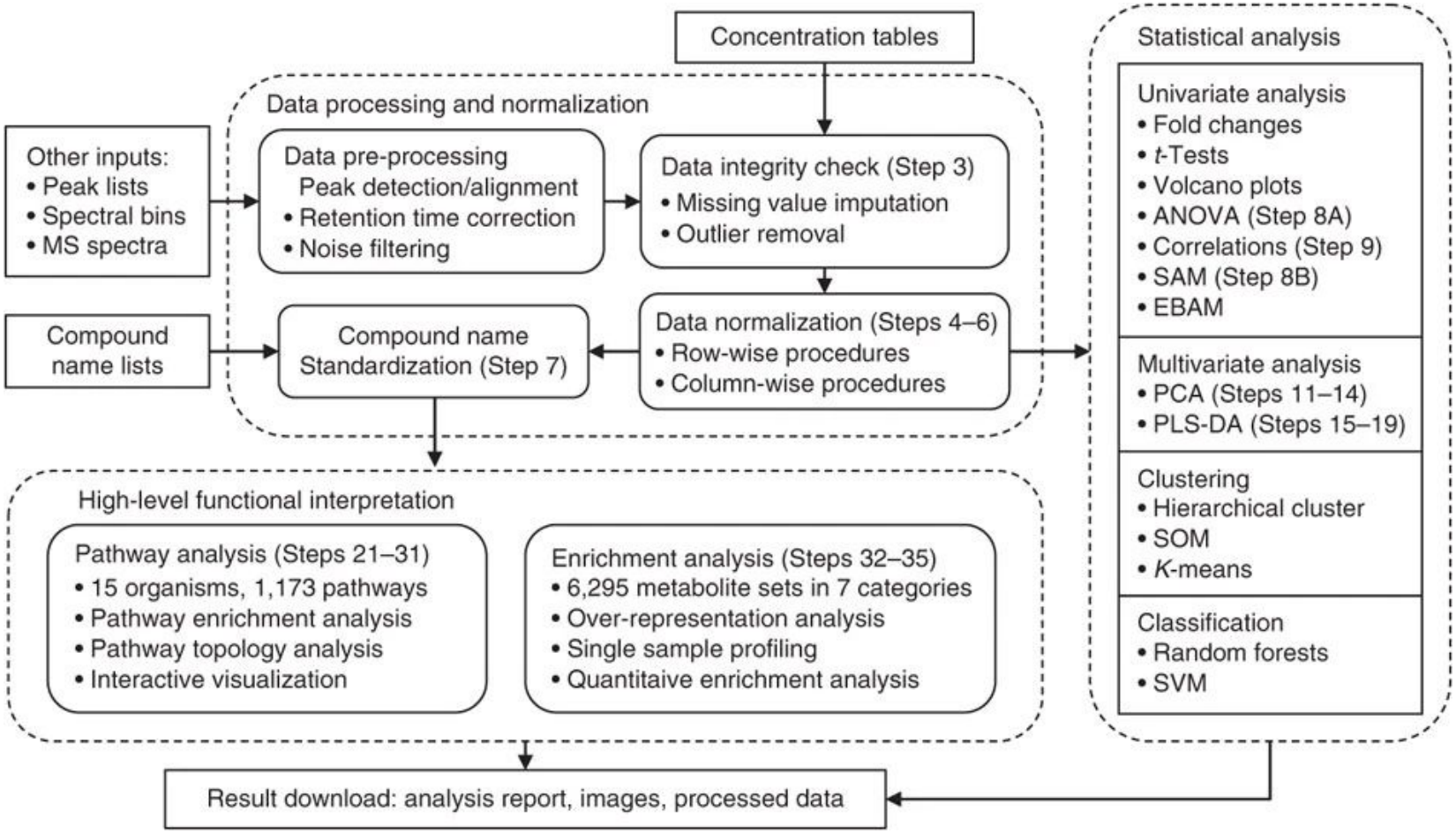
From raw spectra to biomarkers, patterns, functions and systems biology

- it is a free web-based platform
- it works with **R** but it has a *friendlier* GUI: anyone can make metabolomics data analysis, interpretation and integration with other omics data
- the whole metabolomics community uses it!!!

...but

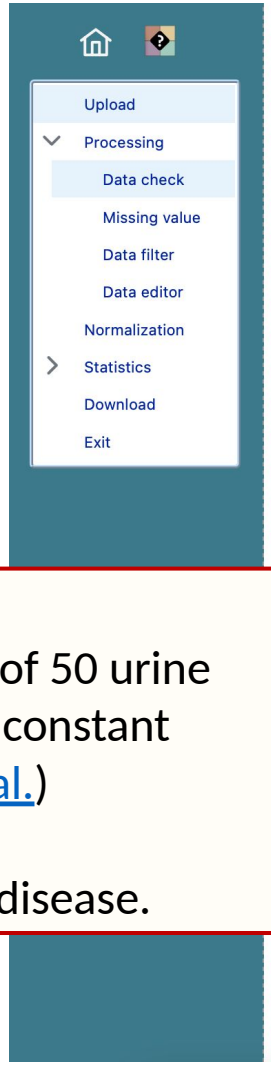
- you need a statistical background to interpret the **MetaboAnalyst** outputs and to get the most of it!

MetaboAnalyst overview



*Source: Xia, J., Wishart, D. Nat Protoc **6**, 743–760 (2011).*

MetaboAnalyst workflow
1) data upload



Data Integrity Check:

- Checking sample names - spaces will replaced with underscore, and special characters will be removed;
- Checking the class labels - at least three replicates are required in each class.
- The data (except class labels) must not contain non-numeric values.
- If the samples are paired, the pair labels must conform to the specified format.
- The presence of missing values or features with constant values (i.e. all zeros).

Data processing information:

Checking data content ...passed.

Samples are in rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 50 (samples) by 200 (spectra bins) data matrix.

Samples are not paired.

2 groups were detected in samples.

Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.

Other special characters or punctuations (if any) will be stripped off.

All data values are numeric.

A total of 0 (0%) missing values were detected.

By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables

Click the **Proceed** button if you accept the default practice;

Or click the **Missing Values** button to use other methods.

Edit Groups | Missing Values | Proceed

Test data 1:
Binned 1H NMR spectra of 50 urine samples using 0.04 ppm constant width (Psihogios NG, et al.)
Group 1- control;
Group 2 - severe kidney disease.

MetaboAnalyst workflow
2) data filtering

Data Filtering:

The purpose of the data filtering is to identify and remove variables that are unlikely to be of use when modeling the data. No phenotype information are used in the filtering process, so the result can be used with any downstream analysis. This step is strongly recommended for untargeted metabolomics datasets (i.e. spectral binning data, peak lists) with large number of variables, many of them are from baseline noises. Filtering can usually improve the results. For details, please refer to the paper by Hackstadt, et al.

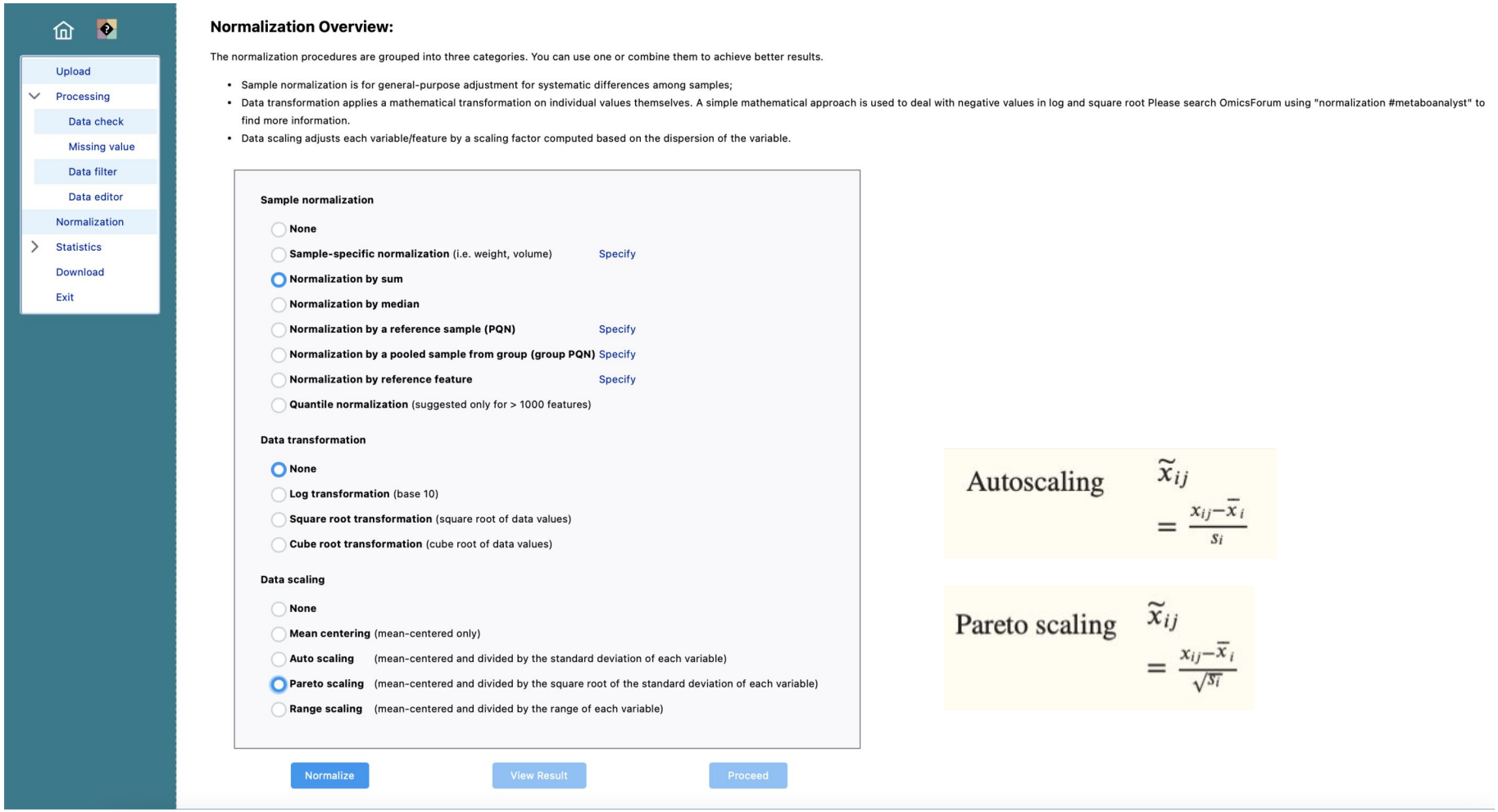
Non-informative variables can be characterized in three groups: 1) variables that show **low repeatability** - this can be measured using QC samples using the relative standard deviation(RSD = SD/mean). Features with high percent RSD should be removed from the subsequent analysis (the suggested threshold is 20% for LC-MS and 30% for GC-MS); 2) variables that are **near-constant** throughout the experiment conditions - these variables can be detected using standard deviation (SD); or the robust estimate such as interquantile range (IQR); and 3) variables of **very small values** (close to baseline or detection limit) - these variables can be detected using mean or median.

For data filtering based on the last two categories, the default parameters follow the empirical rules: 1) Less than 250 variables: 5% will be filtered; 2) Between 250 - 500 variables: 10% will be filtered; 3) Between 500 - 1000 variables: 25% will be filtered; and 4) Over 1000 variables: 40% will be filtered. You can turn off data filtering by dragging the slider to adjust the percentage to filter out to be 0, when your data contain less than 5000 features (or 2500 for power analysis) to control computing time on our server.

Reliability filter:	☐ Filtering features based on technical repeatability QC samples	RSDs greater than: 25%
Variance filter:	◉ Interquantile range (IQR) ○ Standard deviation (SD) ○ Median absolute deviation (MAD) ○ Relative standard deviation (RSD = SD/mean) ○ Non-parametric relative standard deviation (MAD/median)	Percentage to filter out: 5%
Abundance filter:	◉ Mean intensity value ○ Median intensity value	Percentage to filter out: 0%

Submit Proceed

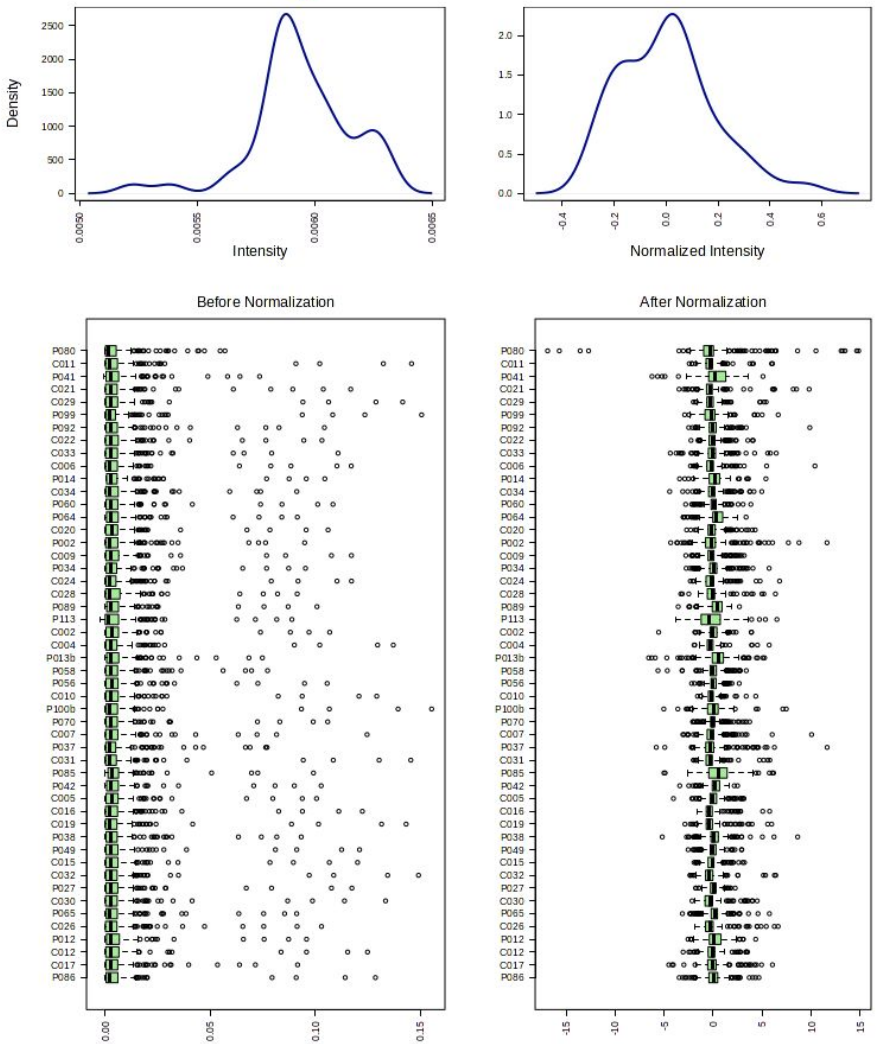
MetaboAnalyst workflow
3) data normalization



Autoscaling $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$

Pareto scaling $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$

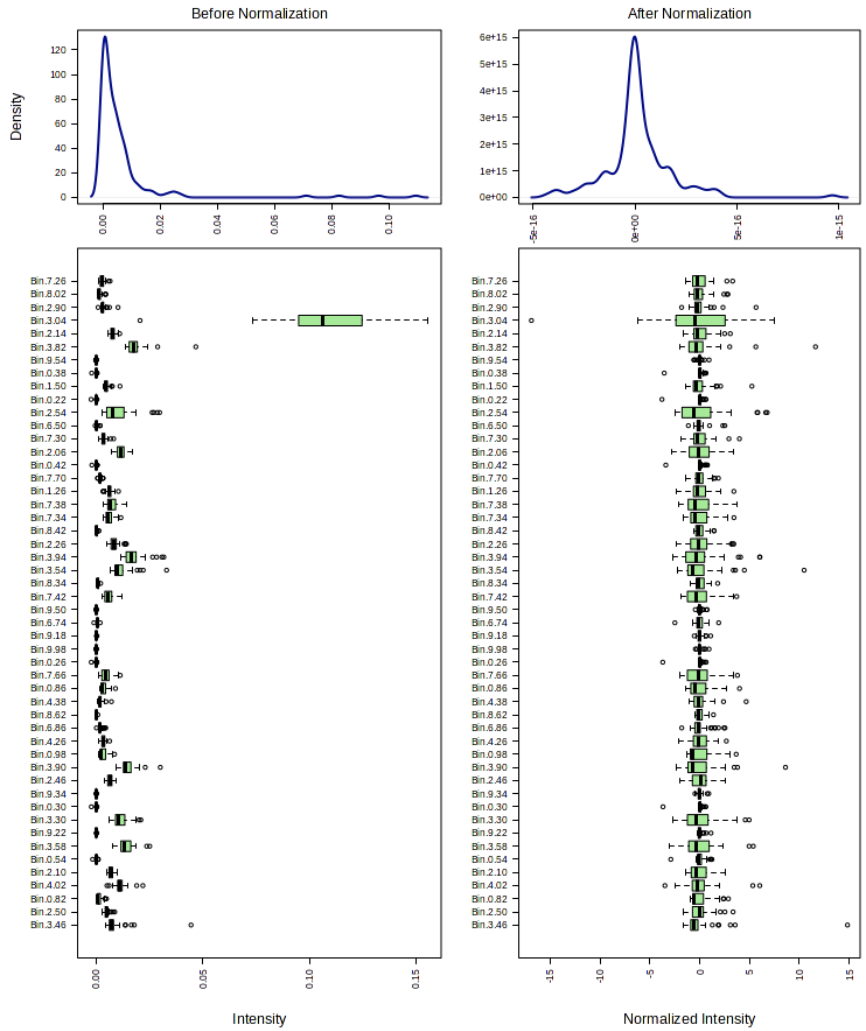
MetaboAnalyst workflow
3) data normalization



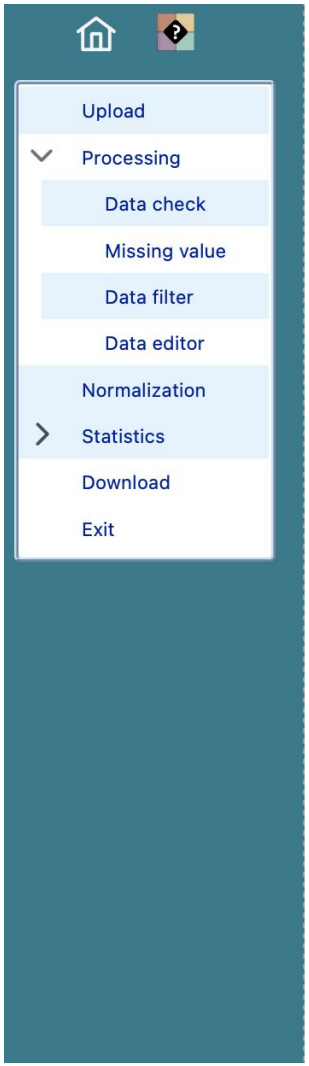
Effect of normalization over sample

MetaboAnalyst workflow
3) data normalization

Effect of features/metabo scaling



MetaboAnalyst workflow
4) statistical analysis



Select an analysis path to explore :

Univariate Analysis

Fold Change Analysis T-tests Volcano plot

One-way Analysis of Variance (ANOVA)

Correlation Heatmaps Pattern Search Correlation Networks (DSPC)

«Classical» analysis of variance among groups

Advanced Significance Analysis

Significance Analysis of Microarray (and Metabolites) (SAM)

Empirical Bayesian Analysis of Microarray (and Metabolites) (EBAM)

Chemometrics Analysis

Principal Component Analysis (PCA)

Partial Least Squares - Discriminant Analysis (PLS-DA)

Sparse Partial Least Squares - Discriminant Analysis (sPLS-DA)

Orthogonal Partial Least Squares - Discriminant Analysis (orthoPLS-DA)

Cluster Analysis

Hierarchical Clustering: Dendrogram Heatmaps

Partitional Clustering: K-means Self Organizing Map (SOM)

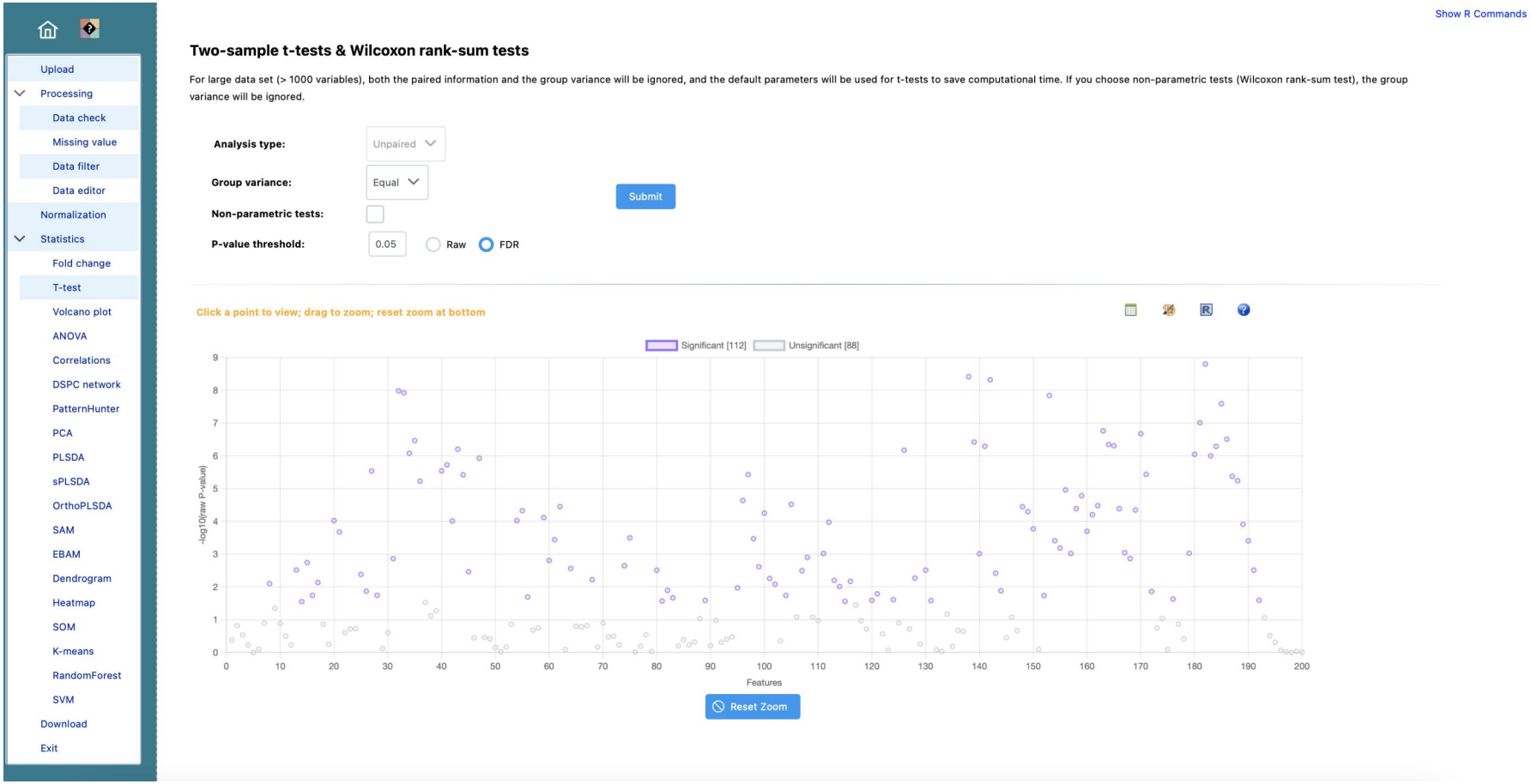
Classification & Feature Selection

Random Forest

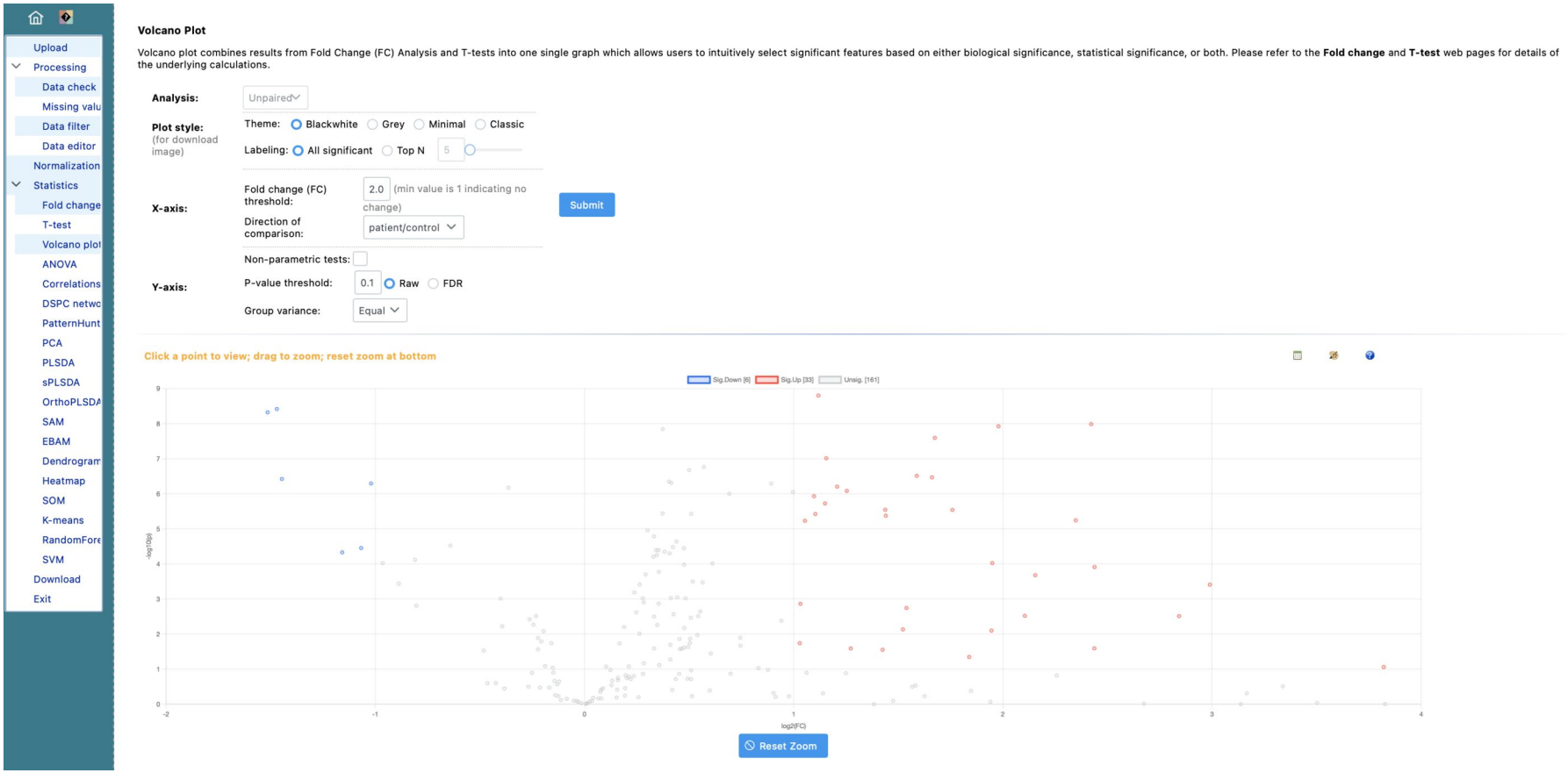
Support Vector Machine (SVM)

Machine learning algorithms

MetaboAnalyst workflow
4) univariate analysis

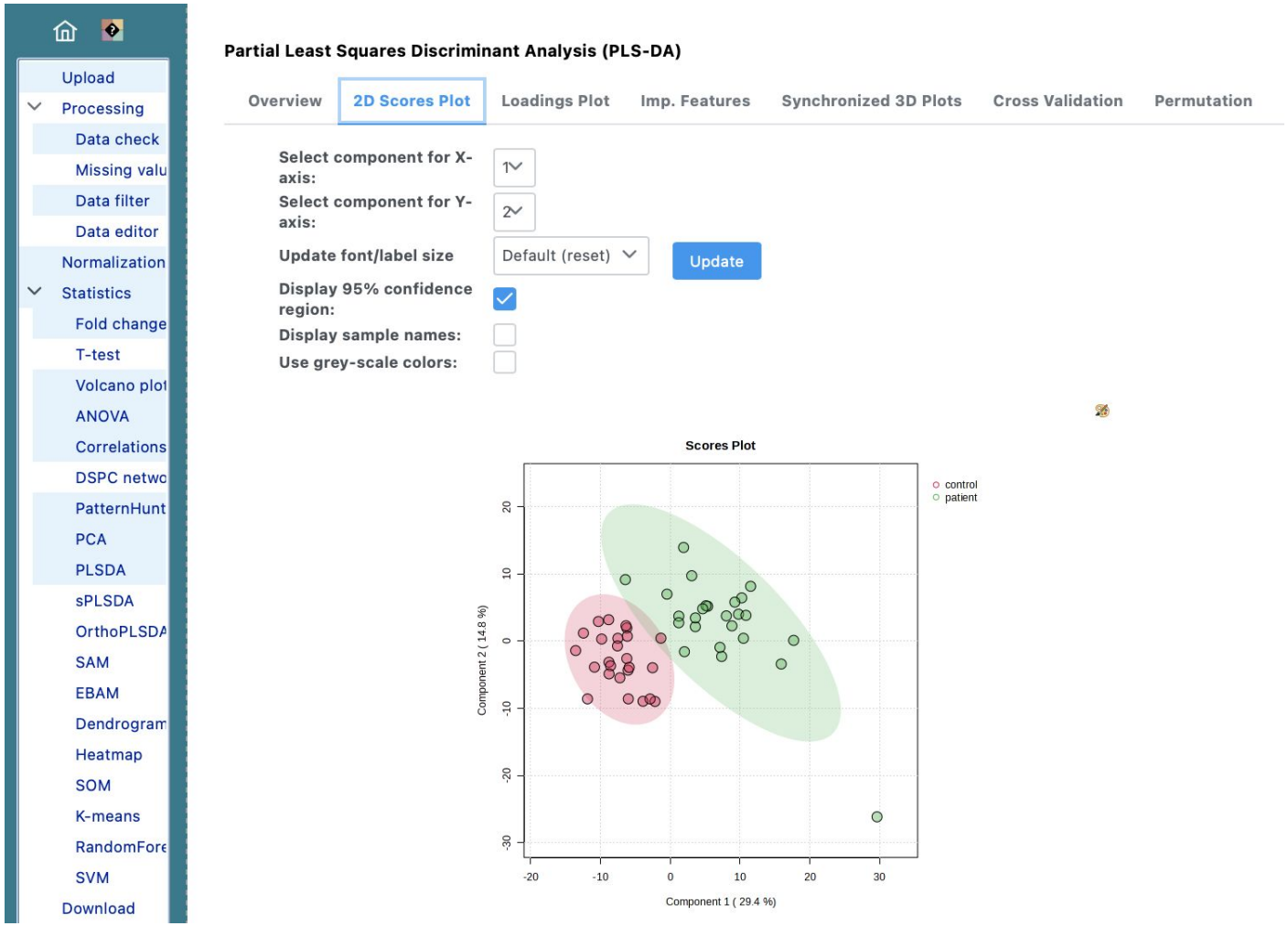


MetaboAnalyst workflow
4) univariate analysis

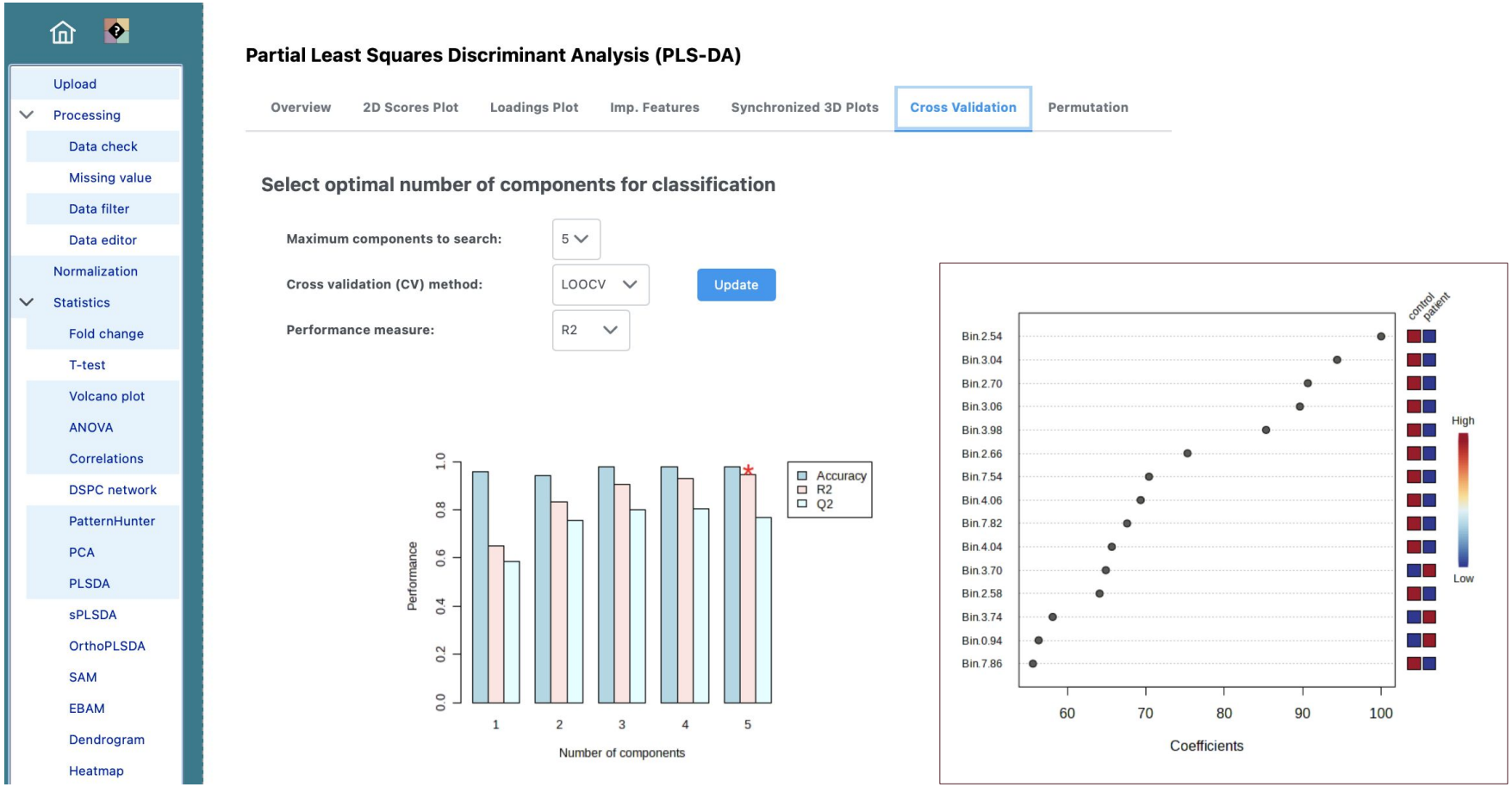


MetaboAnalyst workflow
5) chemometric analysis

MetaboAnalyst workflow
5) chemometric analysis



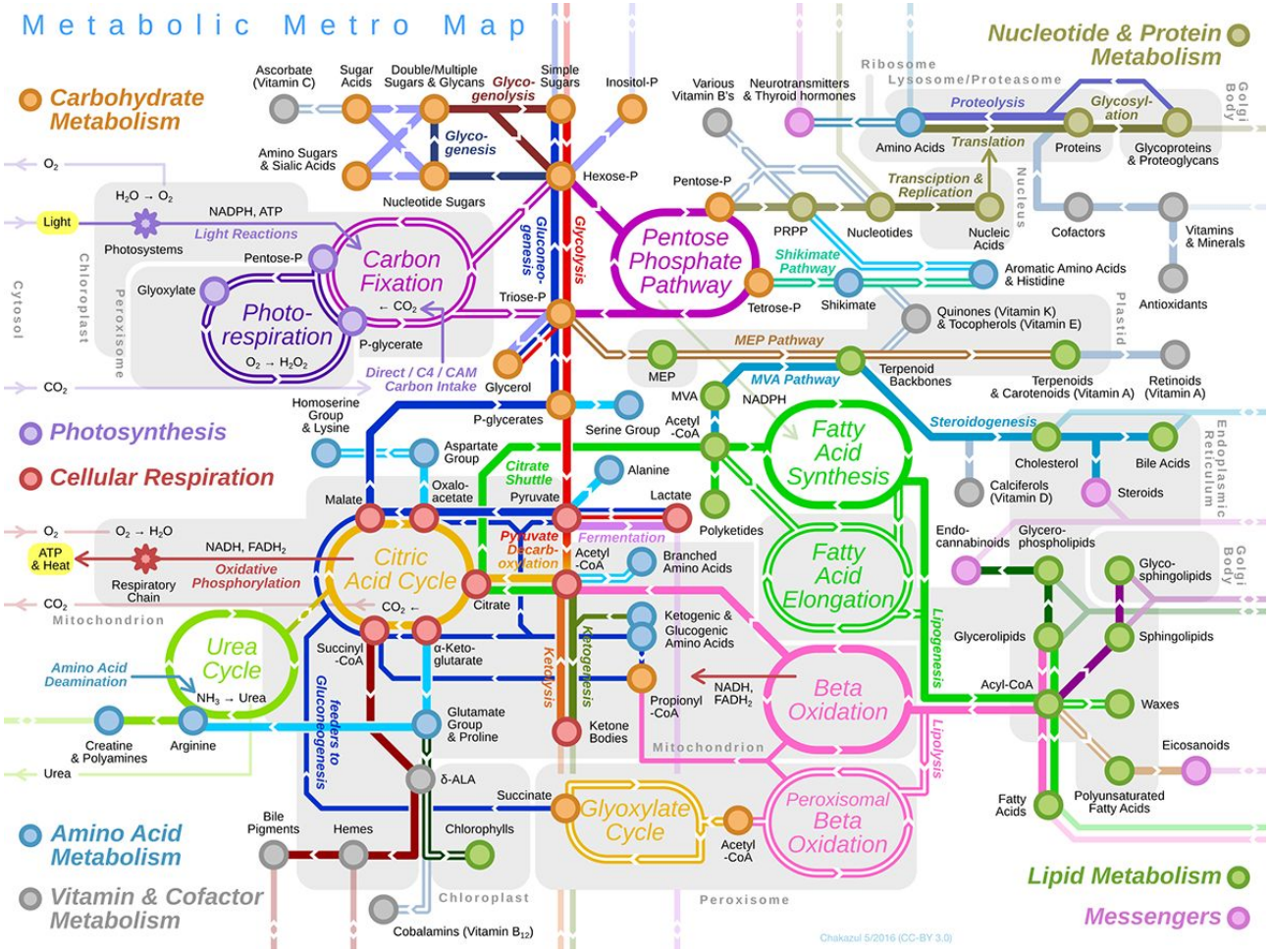
MetaboAnalyst workflow
5) chemometric analysis



MetaboAnalyst workflow
5) chemometric analysis

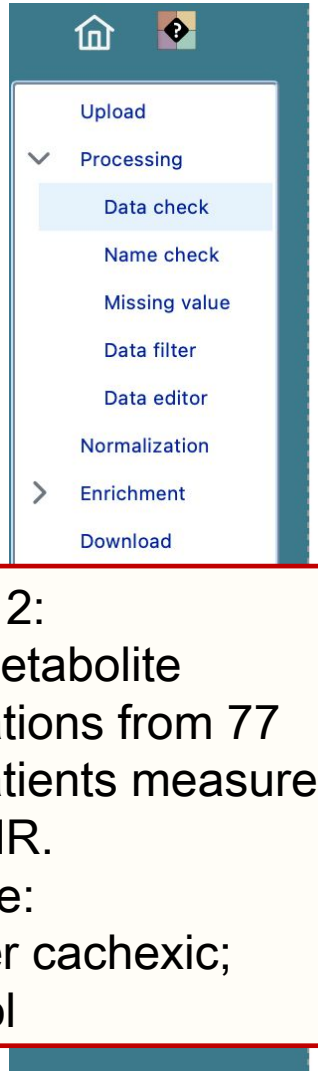
Heatmap of the top 25 T-test features

Identifying the metabolic pathways deregulated by a pathology is finding a target for pharmacological therapy!



Source: https://www.behance.net/gallery/38270165/Metro-Map-of-Metabolism-The-Overview

MetaboAnalyst workflow
6) enrichment analysis



Data Integrity Check:

- Checking sample names - spaces will replaced with underscore, and special characters will be removed;
- Checking the class labels - at least three replicates are required in each class.
- The data (except class labels) must not contain non-numeric values.
- If the samples are paired, the pair labels must conform to the specified format.
- The presence of missing values or features with constant values (i.e. all zeros).

Data processing information:

Checking data content ...passed.

Samples are in rows and features in columns

The uploaded file is in comma separated values (.csv) format.

The uploaded data file contains 77 (samples) by 63 (compounds) data matrix.

Samples are not paired.

2 groups were detected in samples.

Only English letters, numbers, underscore, hyphen and forward slash (/) are allowed.

Other special characters or punctuations (if any) will be stripped off.

All data values are numeric.

A total of 0 (0%) missing values were detected.

By default, missing values will be replaced by 1/5 of min positive values of their corresponding variables

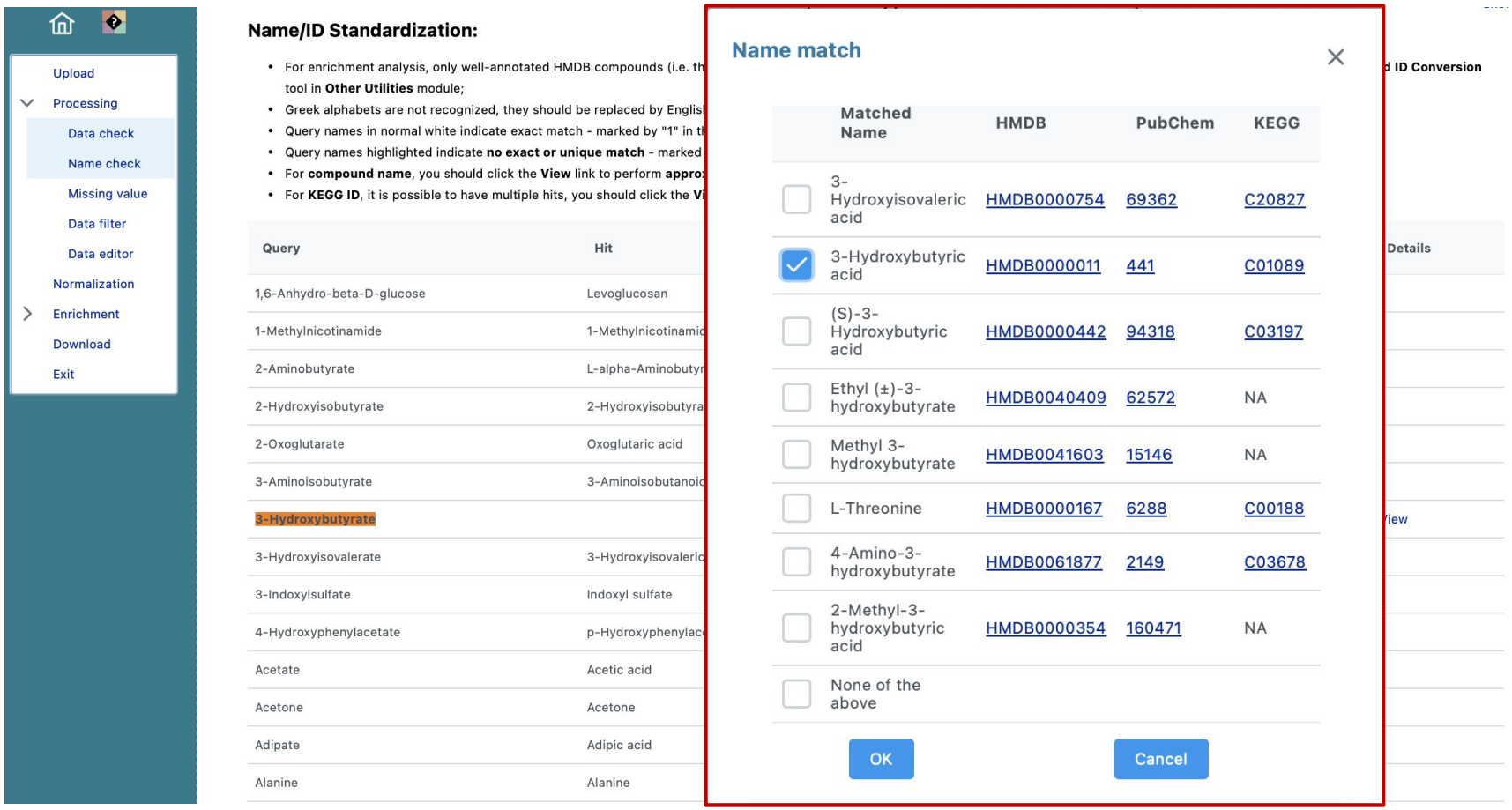
Click the **Proceed** button if you accept the default practice;

Or click the **Missing Values** button to use other methods.

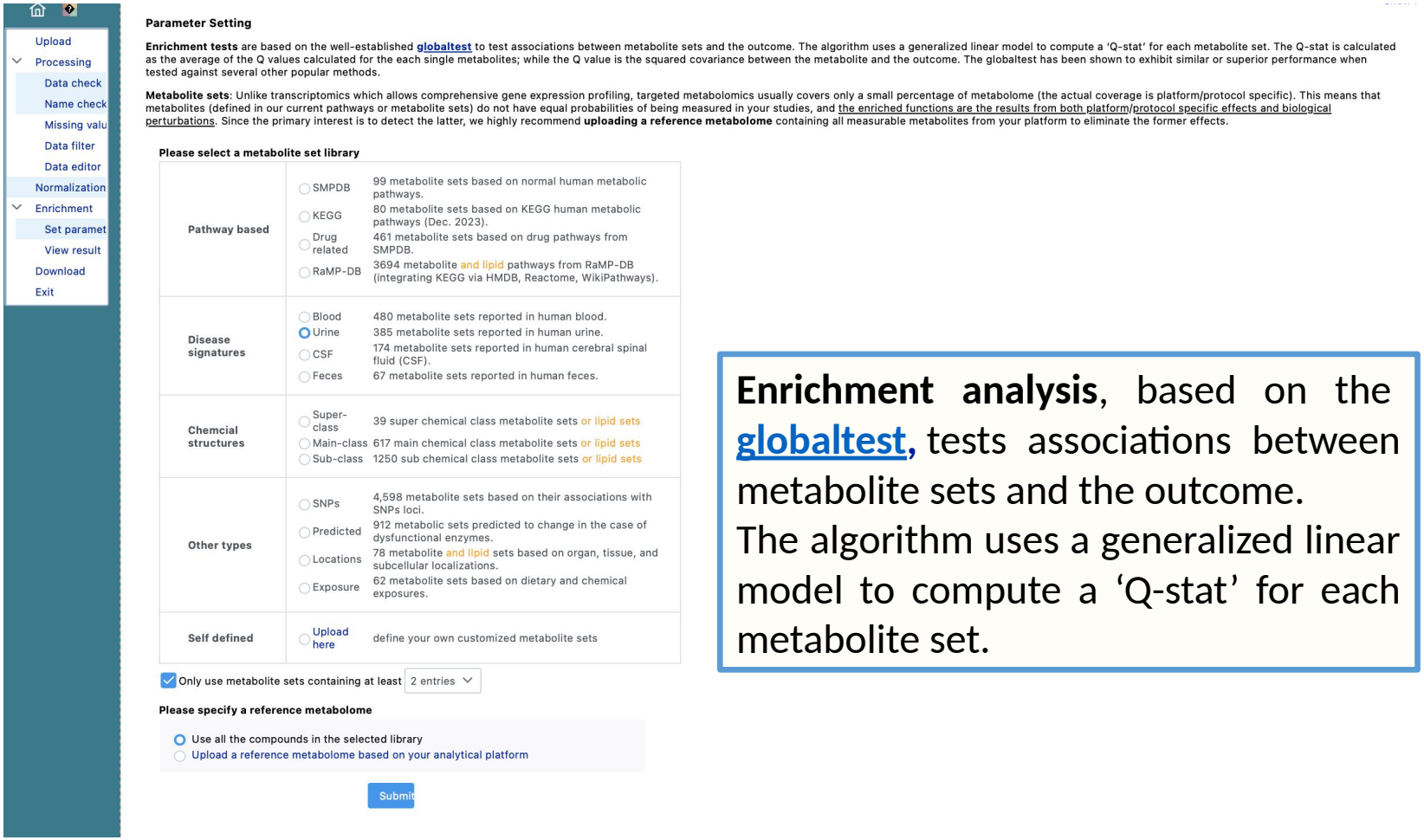
Edit Groups | Missing Values | Proceed

Test data 2:
Urinary metabolite concentrations from 77 cancer patients measured by 1H NMR.
Phenotype:
N - cancer cachexic;
Y - control

MetaboAnalyst workflow
6) enrichment analysis



MetaboAnalyst workflow
6) enrichment analysis

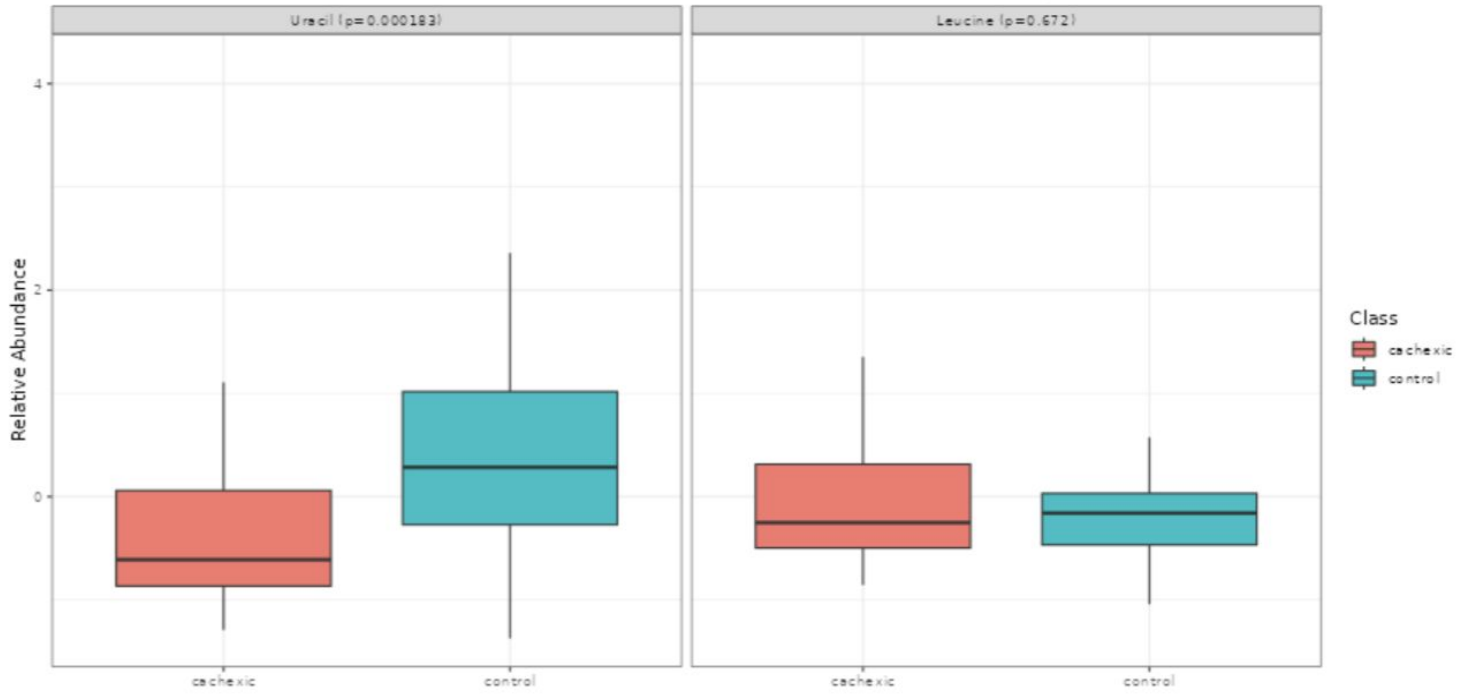


Enrichment analysis, based on the globaltest, tests associations between metabolite sets and the outcome.
The algorithm uses a generalized linear model to compute a 'Q-stat' for each metabolite set.

MetaboAnalyst workflow
6) enrichment analysis

MetaboAnalyst workflow
6) functional interpretation

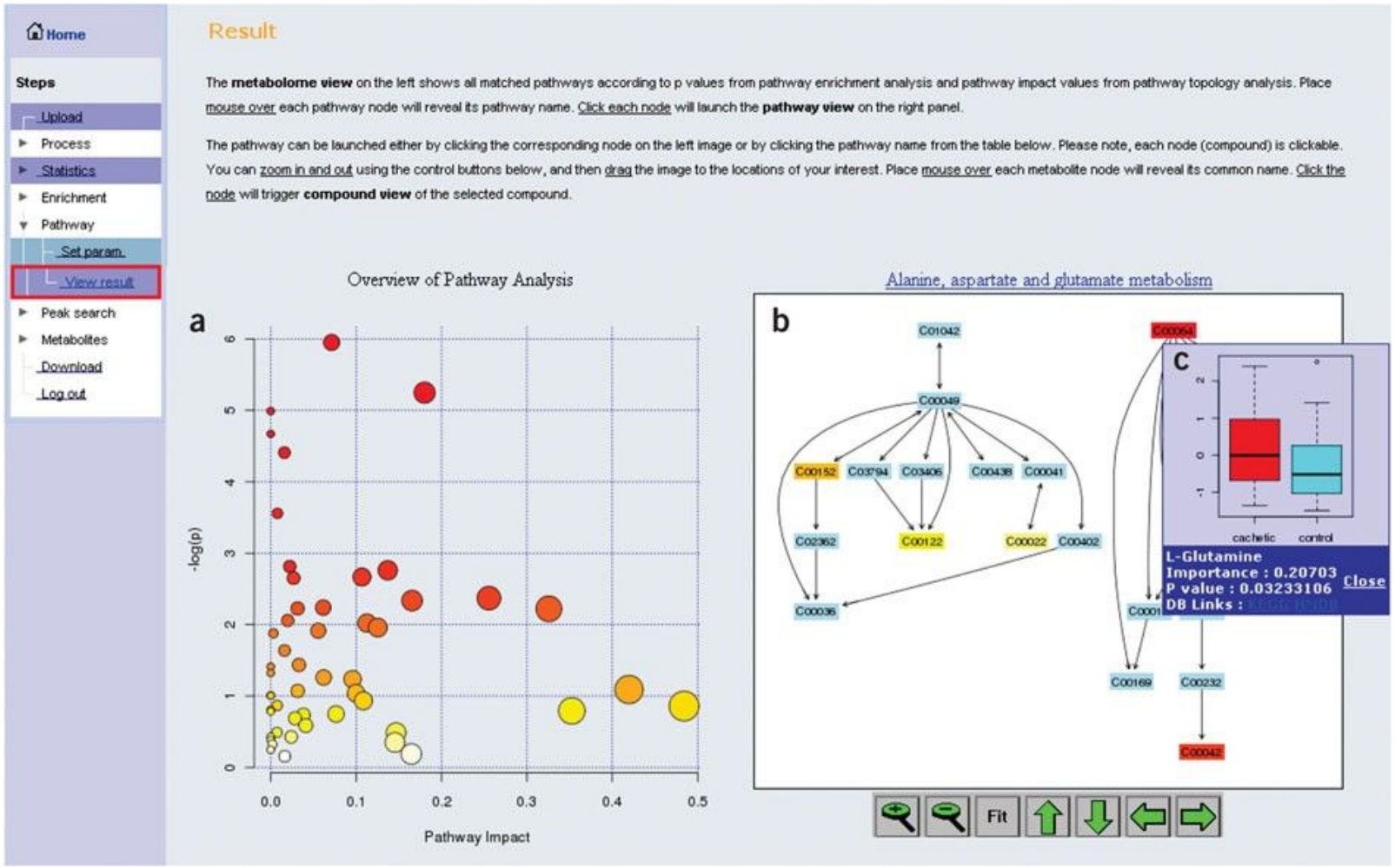
Current metabolite set:



Set Name	Metabolites	References
Metastatic Prostate Cancer	Sarcosine; **Uracil**; Kynurenine; Glycerol 3-phosphate; **Leucine**; DL-Proline	PubMed

MetaboAnalyst workflow

Metabolic pathway analysis and visualization



*Source: Xia, J., Wishart, D. Nat Protoc **6**, 743–760 (2011).*

DAY 4 – LECTURE OUTLINE

- Examples of Machine Learning
  1. PCA
  2. PLS-DA
- MetaboAnalyst
  1. Overview
  2. Workflow
- Power analysis
  1. Hypothesis testing
  2. Decision errors
  3. Statistical power
  4. Effect size

Hypothesis testing steps

1. State the hypotheses (the null hypothesis and an alternative hypothesis)
2. Design the analysis (*e.g.* the significance level is 0.05, the test method one-sample z-test)
3. Analyze sample data
4. Interpret result and make decision

What are the Null and Alternative hypotheses?

Null Hypothesis	Alternative Hypothesis or

• is the hypothesis that a sample data statistic occurs purely from chance
  • e.g. there is no difference between the mean pulse rate for people doing physical exercise and the normal pulse rate
• Must contain condition of equality ,
• Test the Null Hypothesis directly: reject or fail to reject
	• is the hypothesis that a sample data statistic is influenced by some non-random cause
	  • e.g. the mean pulse rate for persons doing the physical exercise is higher than the normal
	• Must be true if is false (corresponding to , conditions)
	• `opposite' of Null Hypothesis

Decision Errors

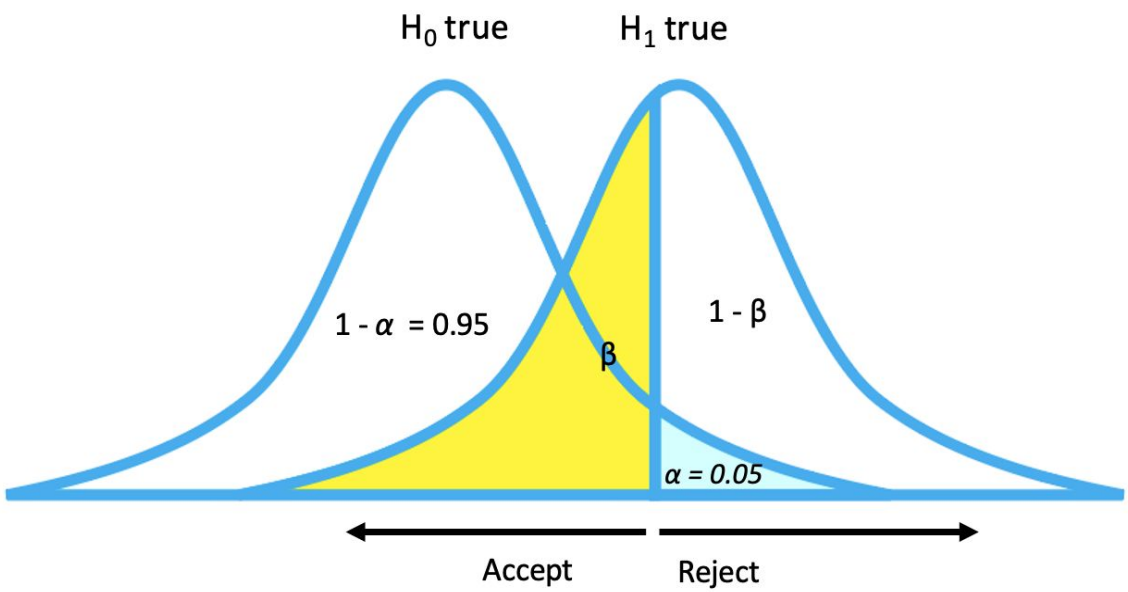
Two types of errors can result from a hypothesis test.

- Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the significance level. This probability is also called alpha, and is often denoted by α.
- Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by β. The probability of not committing a Type II error is called the Power of the test.

Summarizing Type I and Type II Errors

	Fail to reject H0	Reject H0
H0 is true	**Correct action**	**Type I error** **FALSE POSITIVE**
probability	*1-α*	*α*
H1 is true	**Type II error** **FALSE NEGATIVE**	**Correct action**
probability	*β*	*power = 1-β*

α = P(H1|H0)
β = P(H0|H1)



Which is worse: false-positive or false-negative?

	Fail to reject H0	Reject H0
H0 is true	TRUE NEGATIVE	FALSE POSITIVE
probability	$1-\alpha$	α
H1 is true	FALSE NEGATIVE	TRUE POSITIVE
probability	β	*power* = $1-\beta$

Example 1. Covid-19 test:

- *Fals*
  alth
  qua
- *Fals*
  pres
  have
  cent

Example 2. Quality control in a pharma production company

- *Fals*
  We
  shel
- *Fals*
  as a
  defe
  the

Example 3. Disease diagnosis

- *Fals*
  of d
  was
- *Fals*
  prov
  fatal

Example 3. Criminal court

- *False-POSITIVE: an innocent citizen is found guilty and is sent to prison or receives the death penalty*
- *False-NEGATIVE: a criminal is declared innocent and escapes punishment*

Controlling Type I and Type II Errors

- α, β, and n are related
- when two of the three are chosen, the third is determined
- usually the researcher fix the type I error (α) he can tolerate before experiment and then compare the p-value and takes a decision

Controlling Type I and Type II error

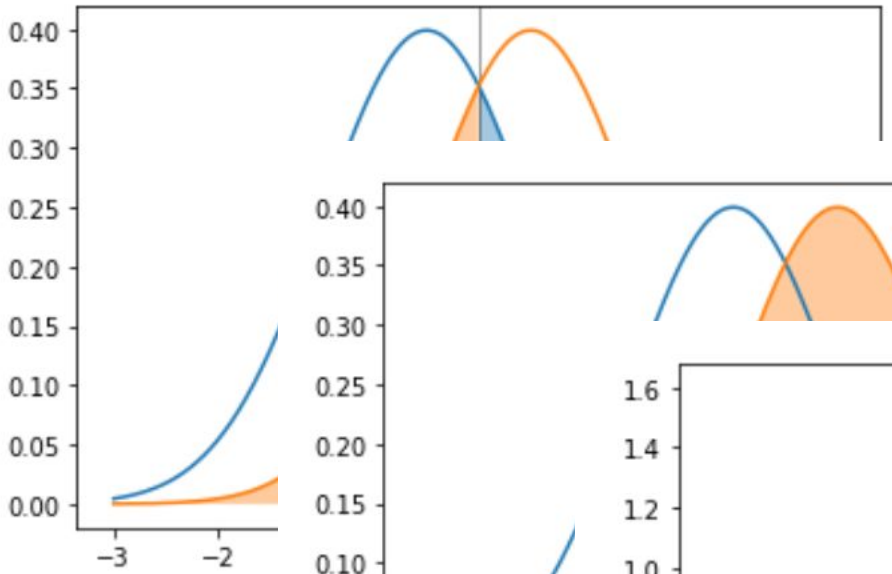


Figure 1: Equa
and false nega

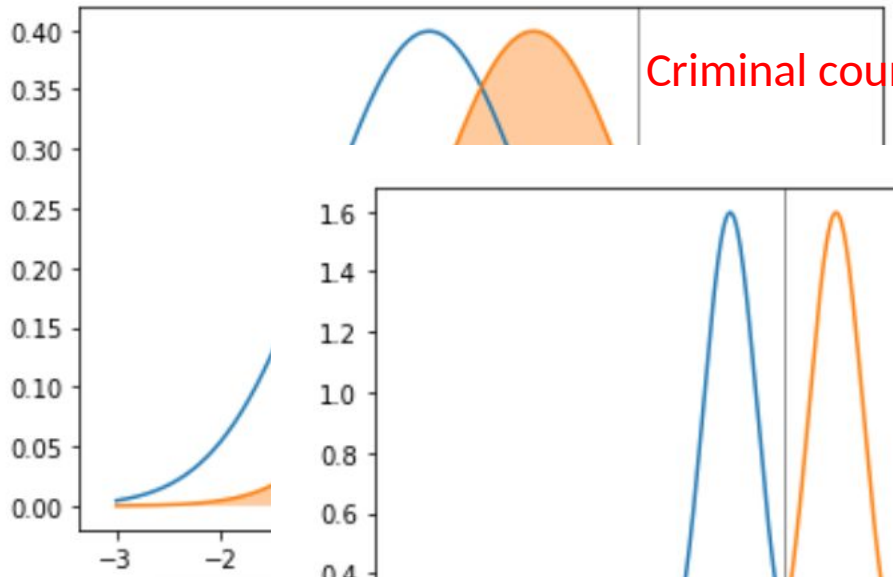


Figure 2: Grea
than false neg

Criminal court example

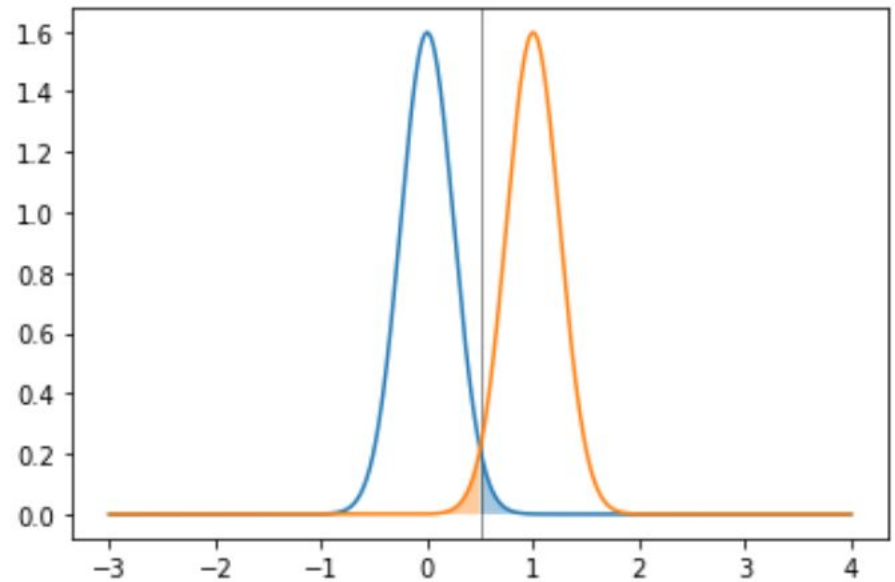
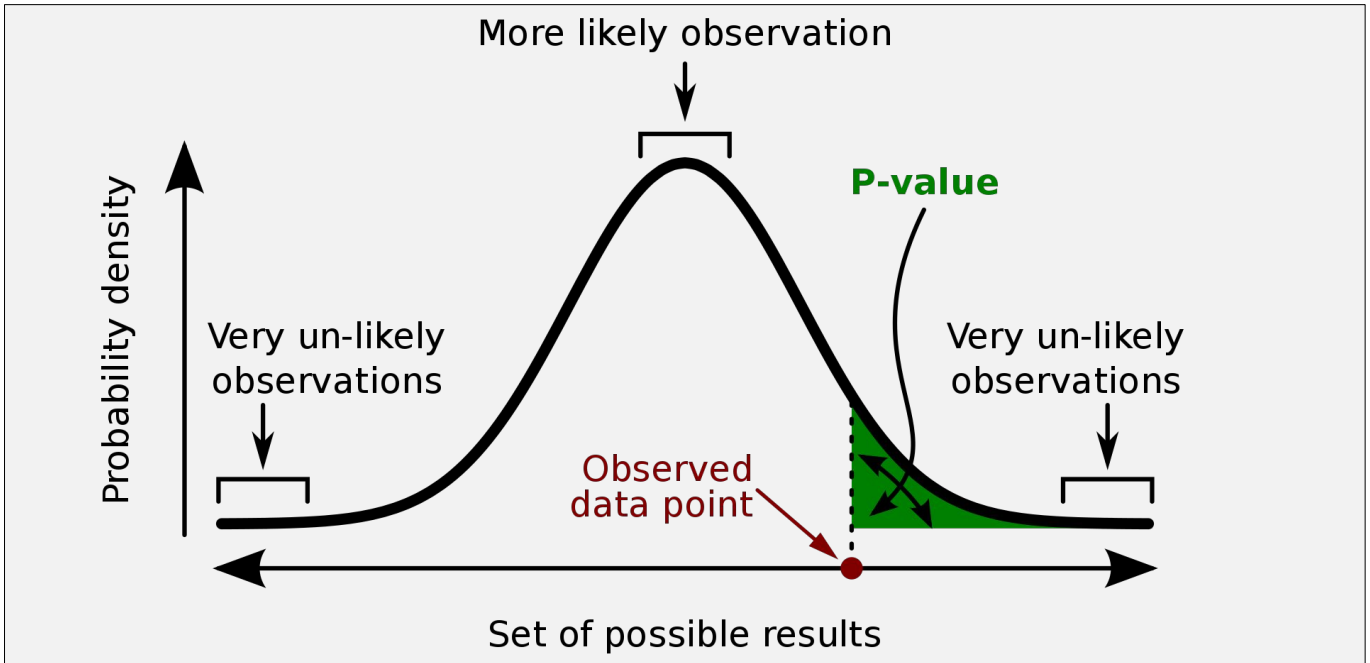


Figure 3: Lowered uncertainty through more informative features.

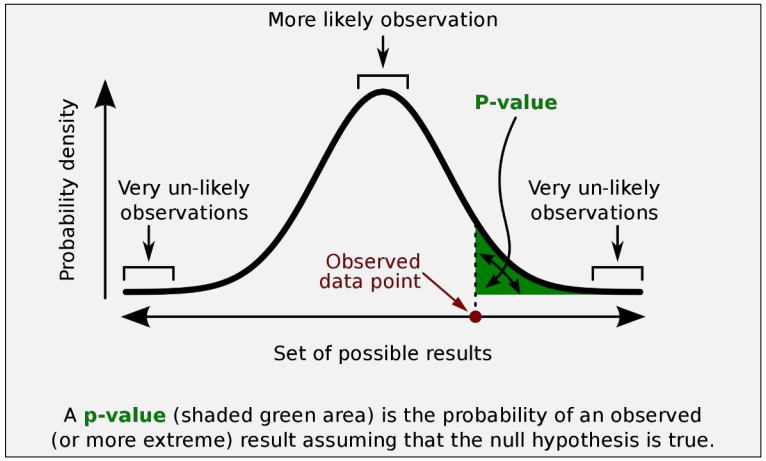
p-value

The p-value corresponds to the answer the question: what is the probability of the observed test statistic or one more extreme when H0 is true?



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

p-value interpretation



- A very small p-value means that such an extreme observed **outcome** would be very unlikely under the null hypothesis.
- Usually the researcher fix α before experiment and then compare the p-value and takes a decision.

Conventions

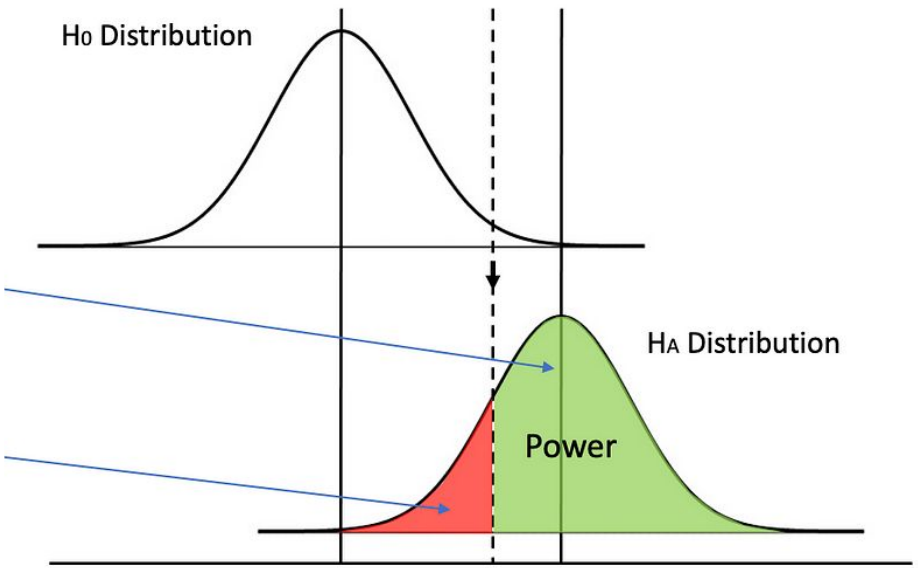
$P > 0.10$	$\Rightarrow$	*non-significant evidence against H0*
$0.05 < P \leq 0.10$	$\Rightarrow$	*marginally significant evidence against H0*
$0.01 < P \leq 0.05$	$\Rightarrow$	*significant evidence against H0*
$P \leq 0.01$	$\Rightarrow$	*highly significant evidence against H0*

How to increase statistical power

	Fail to reject H0	Reject H0
H0 is true	**Correct action**	**Type I error FALSE POSITIVE**
probability	$1-\alpha$	α
H1 is true	**Type II error FALSE NEGATIVE**	**Correct action**
probability	β	*power* = $1-\beta$

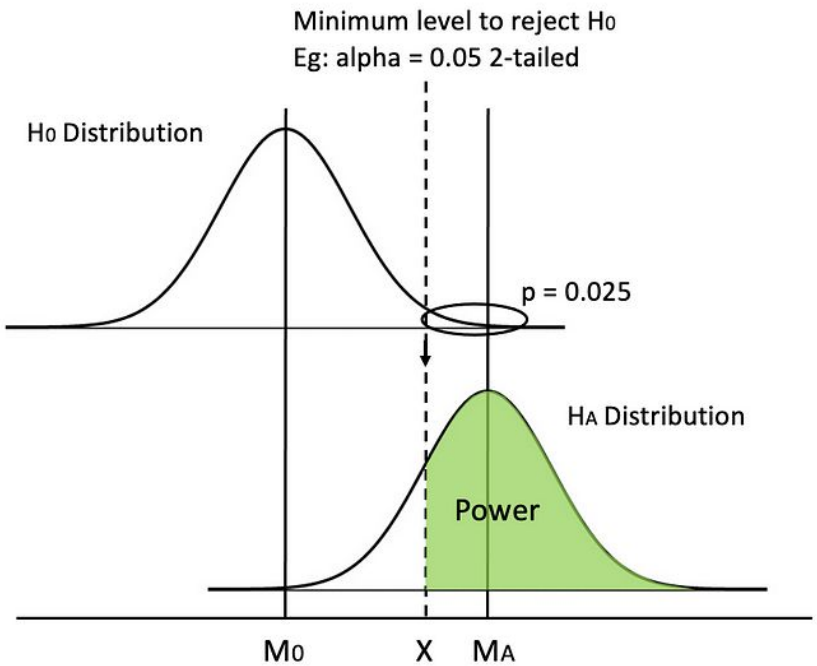
Power
$1-\beta$

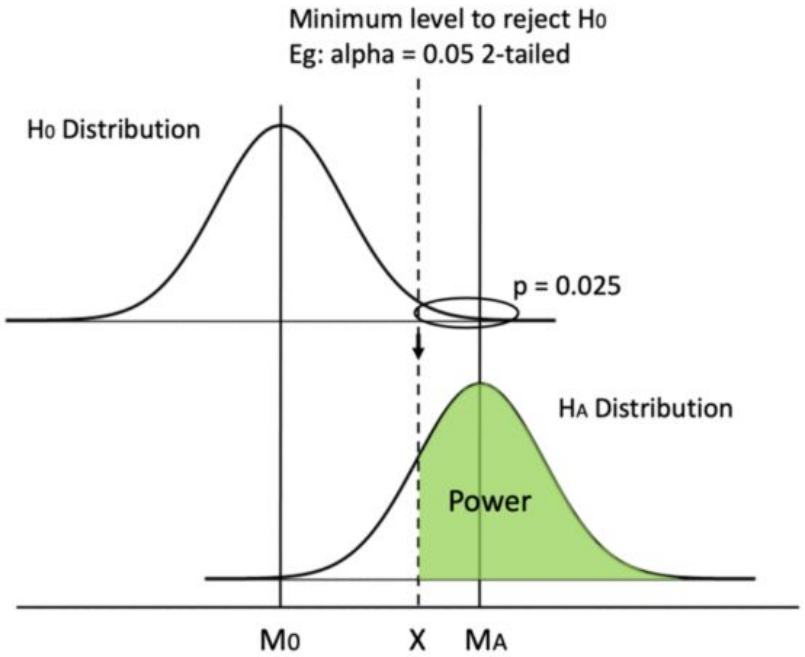
Type II error
β



How to increase statistical power

1) Raise significance level alpha (the WRONG way)

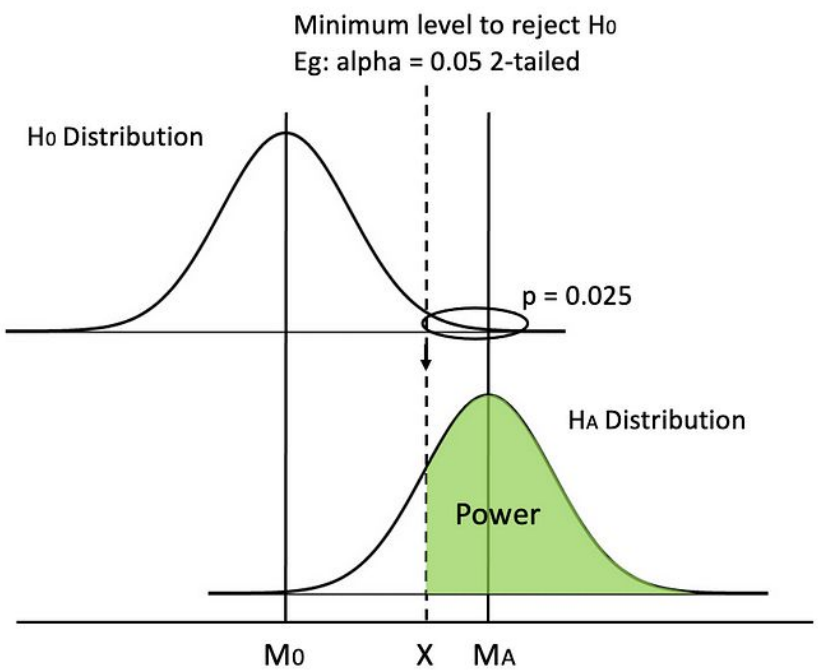


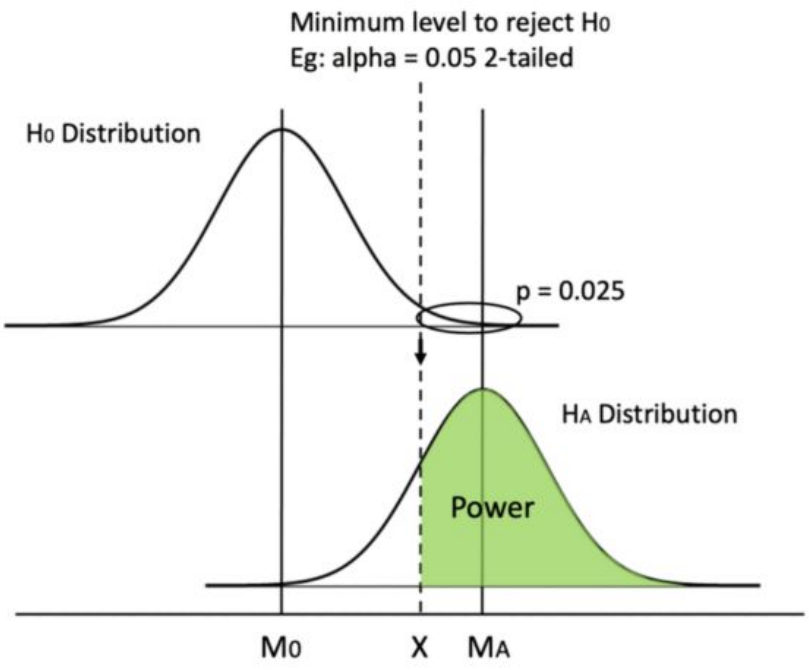


Source: https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to increase statistical power

2) Switch from a 2-tailed test to a 1-tailed test (CORRECT if possible)

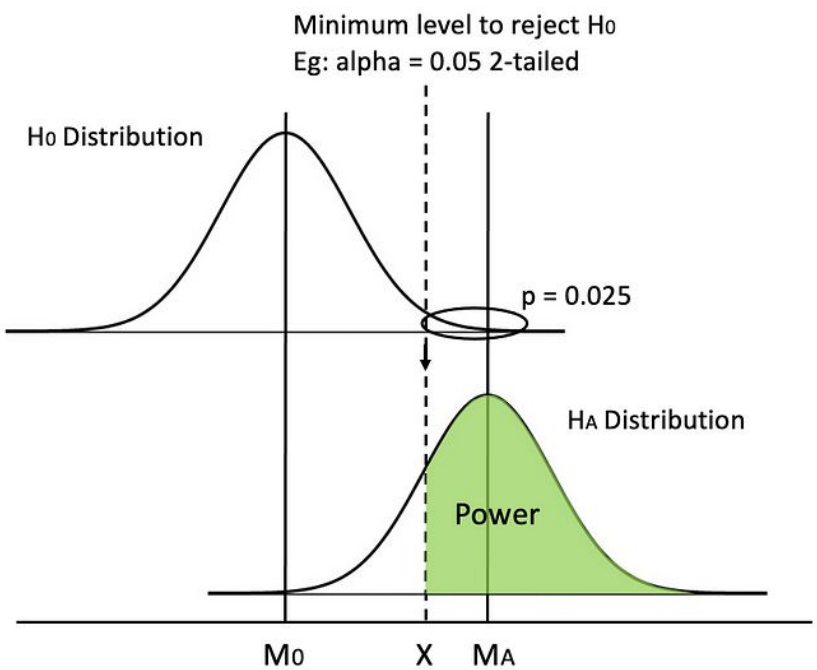


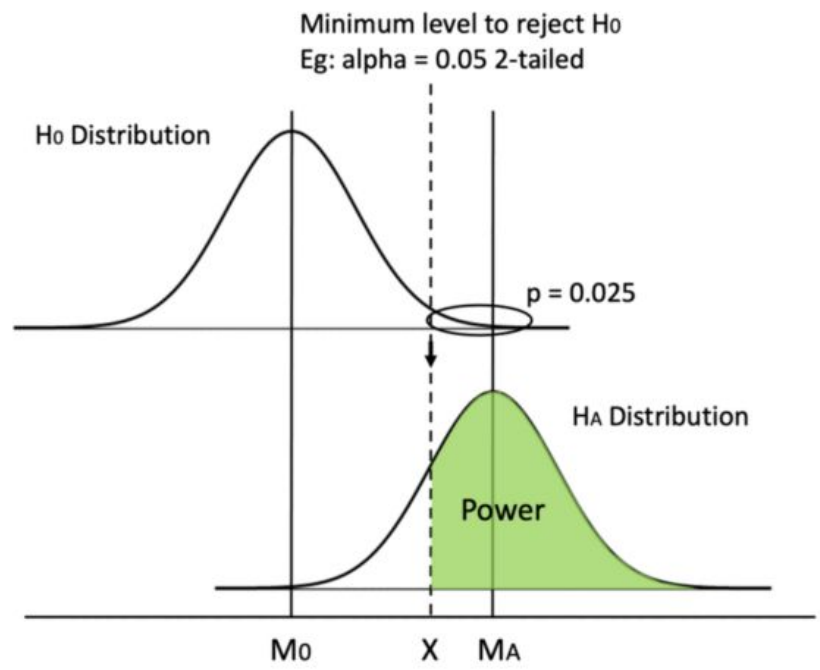


Source: https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to increase statistical power

3) Increase mean difference (or increase the effect size)

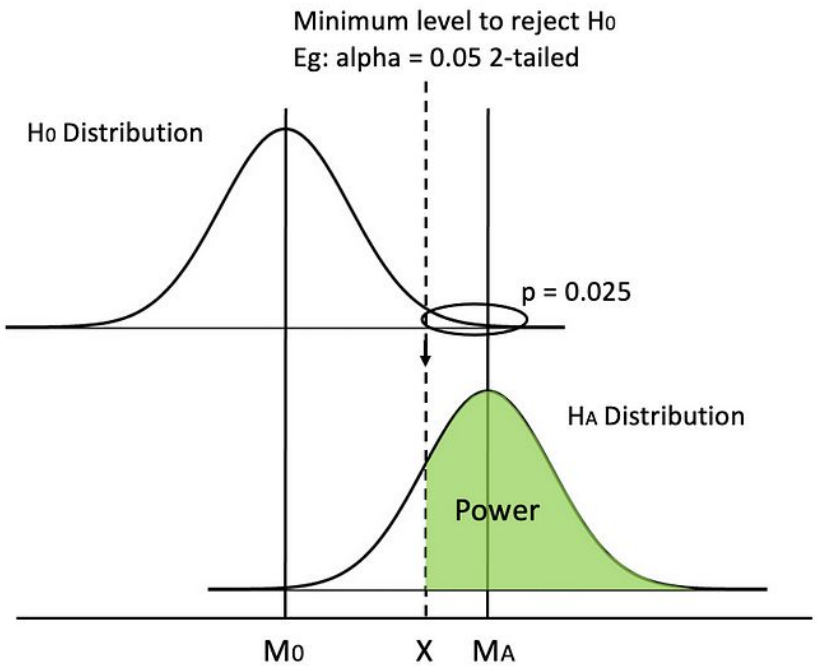


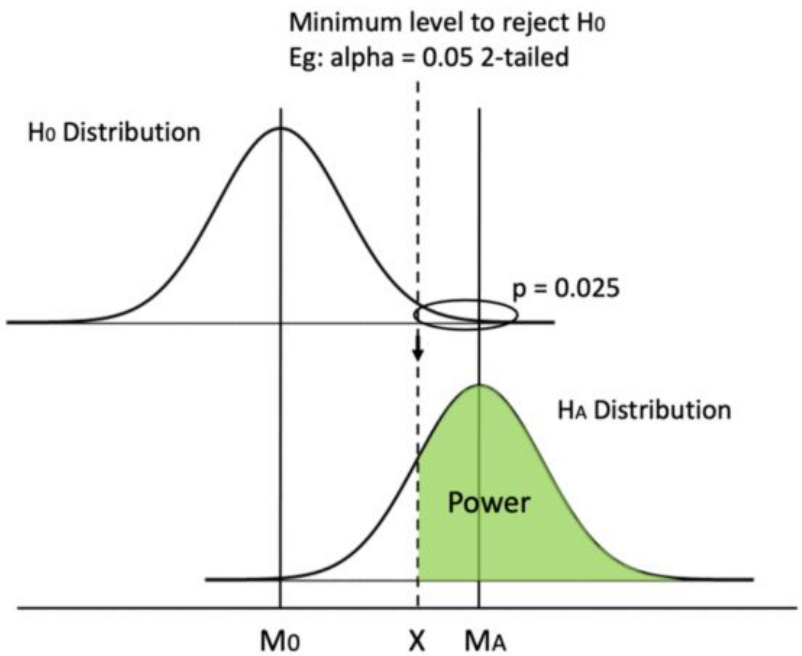


Source: https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to increase statistical power

4) Use z distribution instead of t distribution (appropriate when we know the population mean)

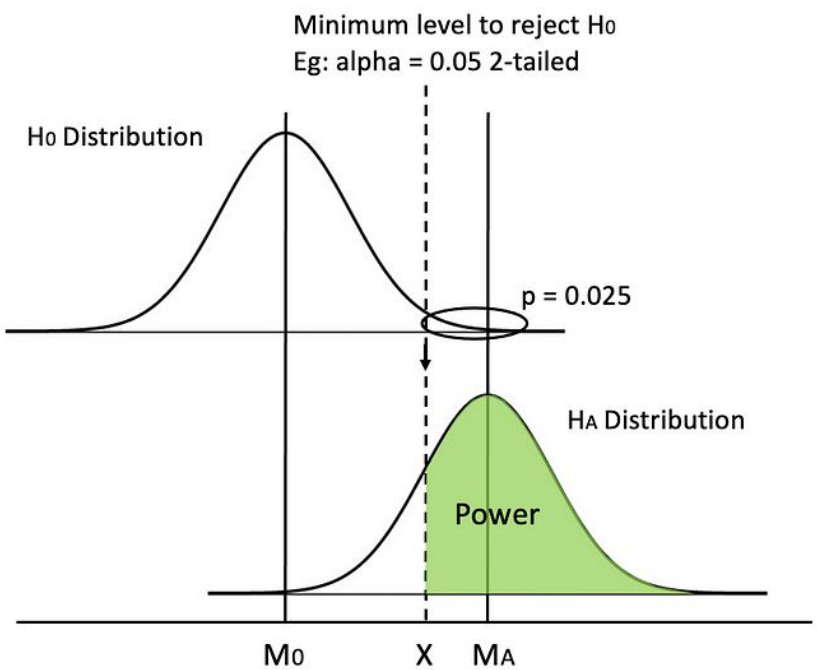


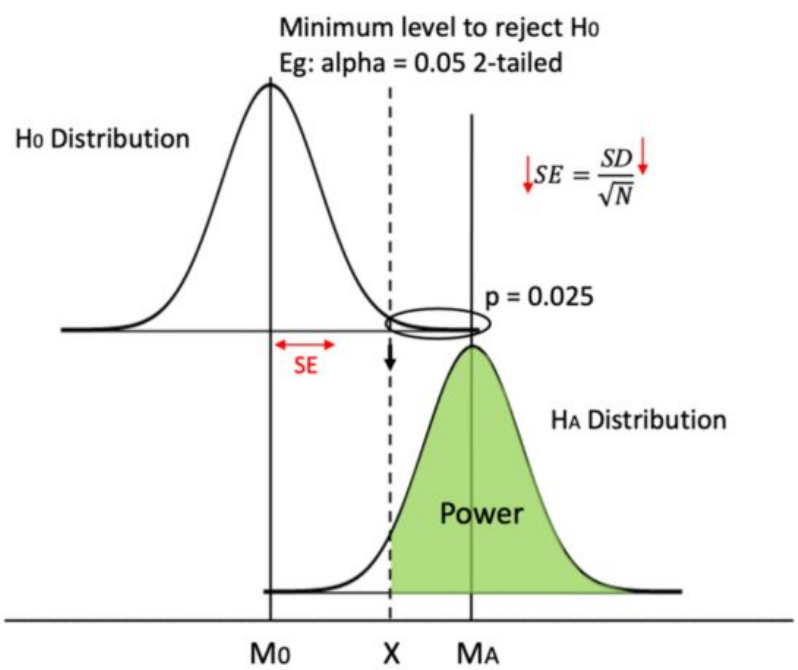


Source: https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to increase statistical power

5) Decrease standard deviation (using more precise measurements to have less error and less noise)

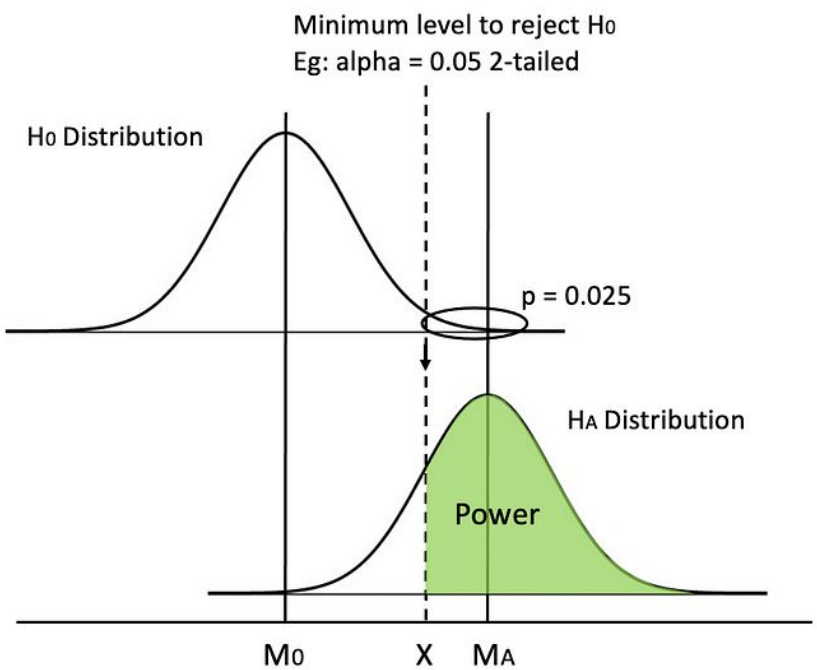


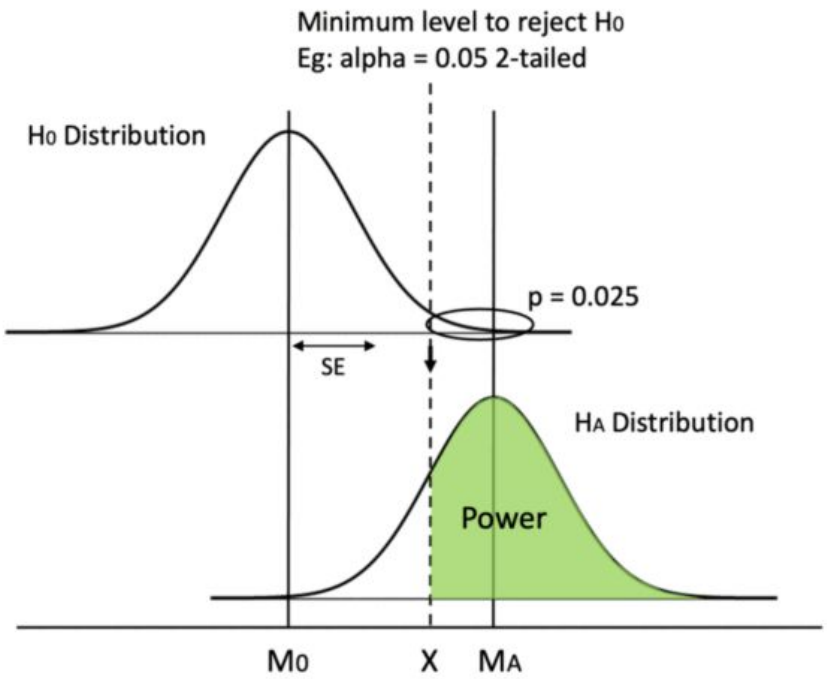


Source: https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

How to increase statistical power

6) Increase sample size (the most practical way)





Source: https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214

Effect size

The effect size is an estimate of the difference between two or more groups.

The measurement of the effect size depends on the type of analysis your are doing:

1. Studying the mean difference between two groups

In this case you use a standardized mean difference (Cohen's *d*)

Effect size

Mean value of the population 1

Mean value of the population 2

$$Cohen's\ d = \frac{\mu_1 - \mu_2}{\sigma}$$

Standard deviation of the population

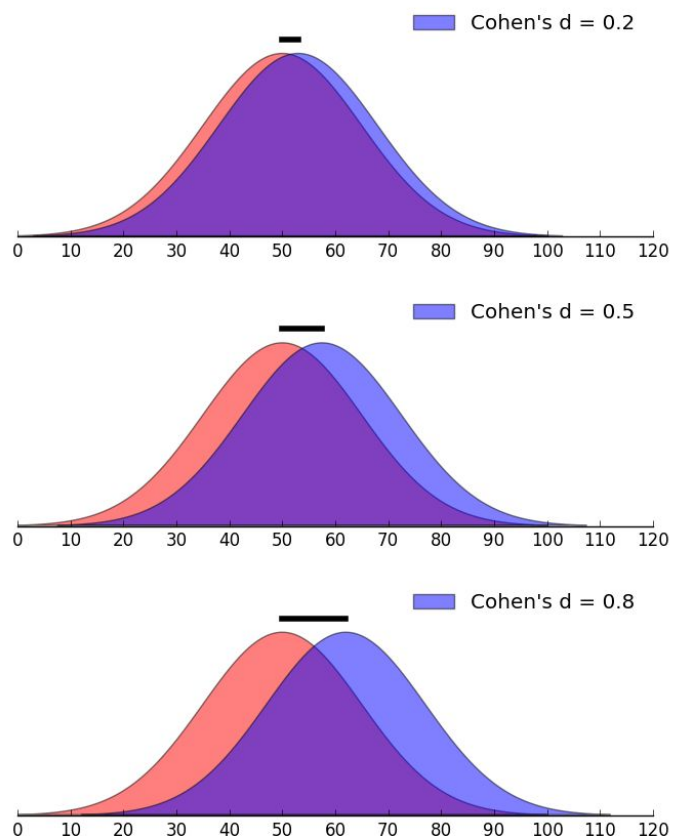
Mean value of sample 1

Mean value of sample 2

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}}$$

Estimated standard deviation of the population from the sample

Cohen's d	Effect size
0.20	Small
0.5	Medium
0.8	Strong



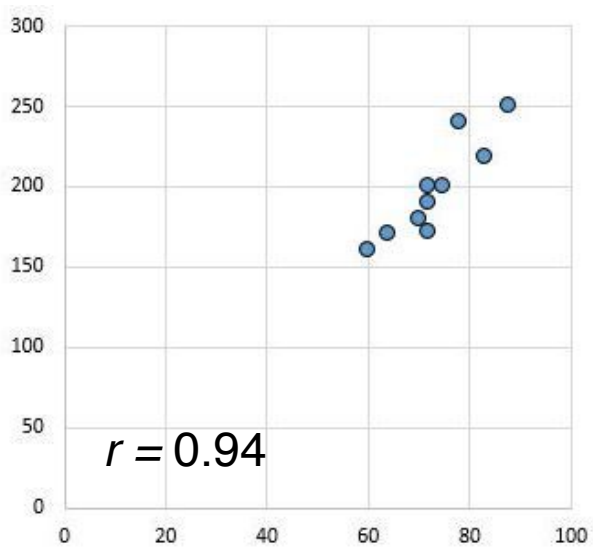
Effect size

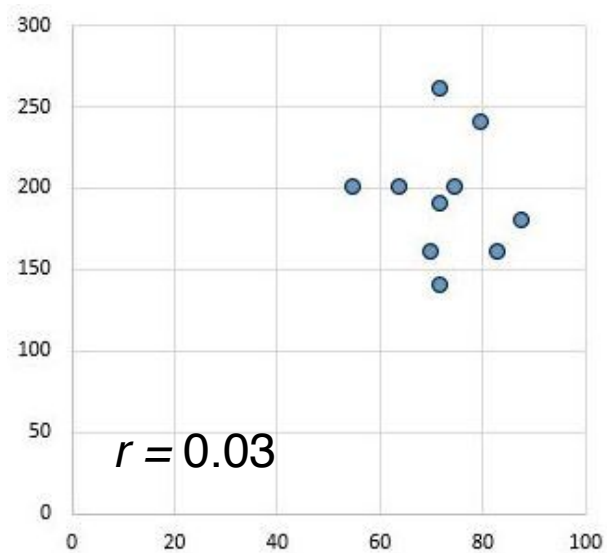
2) Pearson Correlation Coefficient: measuring the linear association between two variables X and Y.

-1 = perfectly negative linear correlation between two variables

0 = no linear correlation between two variables

1 = perfectly positive linear correlation between two variables





Source: https://www.statology.org/effect-size/

Effect size

Pearson Correlation Coefficient

r	Effect size
0.1	small
0.3	medium
>0.5	large

Effect size in different scenarios

Test	Effect Size	Small	Medium	Large
All t-tests: • one-sample t-test • independent samples t-test • paired samples t-test	Cohen's d d =	0.20	0.50	0.80
Difference between many means (ANOVA)	Cohen's f f =	0.10	0.25	0.40
Chi-squared test	Cohen's ω ω=	0.10	0.30	0.50
Pearson's correlation coefficient	Pearson's	0.10	0.30	0.50
Linear Regression (entire model)	Cohen's	0.02	0.15	0.35

Source: https://en.wikipedia.org/wiki/Effect_size#Overview