

STATISTICS & ML WITH R

(special topics)

- MetaboAnalyst
- Power Analysis

2024

M. Chiara Mimmi & Luisa M. Mimmi

WORKSHOP SCHEDULE

- 4 days
 - 1. Intro to R and data analysis
 - 2. Statistical inference & hypothesis testing
 - 3. Modeling correlation and regression
 - 4 Mapping causal & predictive approaches
 - 5. 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 6 – *Extra* TOPICS

- MetaboAnalyst
 - Overview
 - Workflow
- Power analysis
 - Hypothesis testing
 - Decision errors
 - Statistical power
 - Effect size

DAY 4 – LECTURE OUTLINE

- 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>

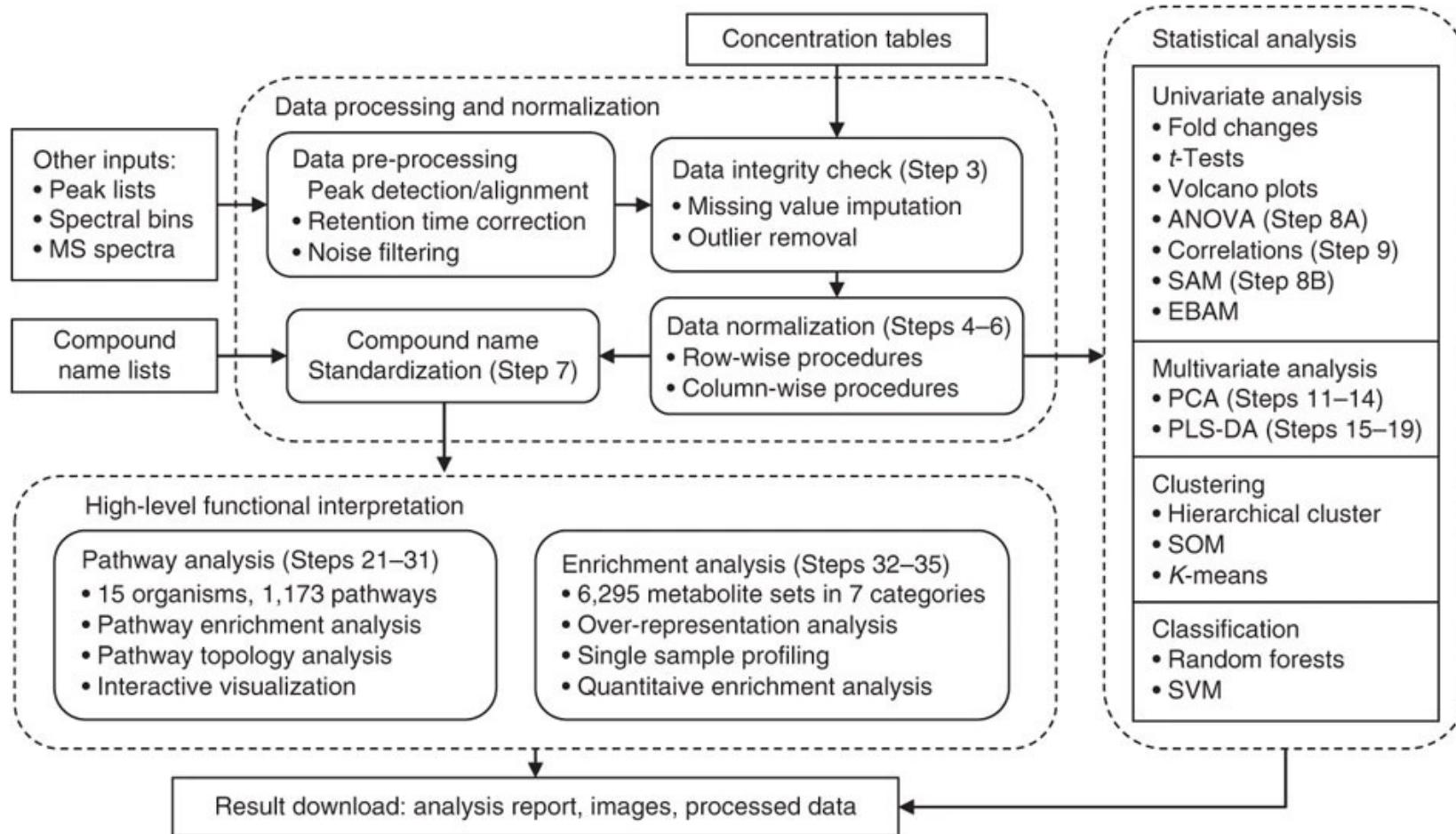
From raw spectra to biomarkers, patterns, functions and systems biology

- 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



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.

Data Integrity Check:

- Checking sample names - spaces will be 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

MetaboAnalyst workflow

2) data filtering

Upload

Processing

Data check

Missing value

Data filter

Data editor

Normalization

Statistics

Download

Exit

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($RSD = SD/\text{mean}$). 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 interquartile 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:	<input type="checkbox"/> Filtering features based on technical repeatability QC samples	RSDs greater than:
Variance filter:	<input checked="" type="radio"/> Interquartile range (IQR) <input type="radio"/> Standard deviation (SD) <input type="radio"/> Median absolute deviation (MAD) <input type="radio"/> Relative standard deviation ($RSD = SD/\text{mean}$) <input type="radio"/> Non-parametric relative standard deviation (MAD/median)	Percentage to filter out:
Abundance filter:	<input checked="" type="radio"/> Mean intensity value <input type="radio"/> Median intensity value	Percentage to filter out:

Submit **Proceed**

MetaboAnalyst workflow

3) data normalization

The screenshot shows the MetaboAnalyst interface for data normalization. The left sidebar has a dark teal background with white icons for home, upload, processing, missing value, data filter, data editor, normalization, statistics, download, and exit. The 'Normalization' section is currently selected. The main content area has a white background with a dark blue header bar. The header bar contains the title 'Normalization Overview:' and a sub-subtitle 'The normalization procedures are grouped into three categories. You can use one or combine them to achieve better results.' Below this is a list of normalization types:

- Sample normalization
 - None
 - Sample-specific normalization (i.e. weight, volume)
 - Normalization by sum
 - Normalization by median
 - Normalization by a reference sample (PQN)
 - Normalization by a pooled sample from group (group PQN)
 - Normalization by reference feature
 - Quantile normalization (suggested only for > 1000 features)
- Data transformation
 - None
 - Log transformation (base 10)
 - Square root transformation (square root of data values)
 - Cube root transformation (cube root of data values)
- Data scaling
 - None
 - Mean centering (mean-centered only)
 - Auto scaling (mean-centered and divided by the standard deviation of each variable)
 - Pareto scaling (mean-centered and divided by the square root of the standard deviation of each variable)
 - Range scaling (mean-centered and divided by the range of each variable)

At the bottom are three buttons: 'Normalize' (blue), 'View Result' (light blue), and 'Proceed' (light blue).

Autoscaling

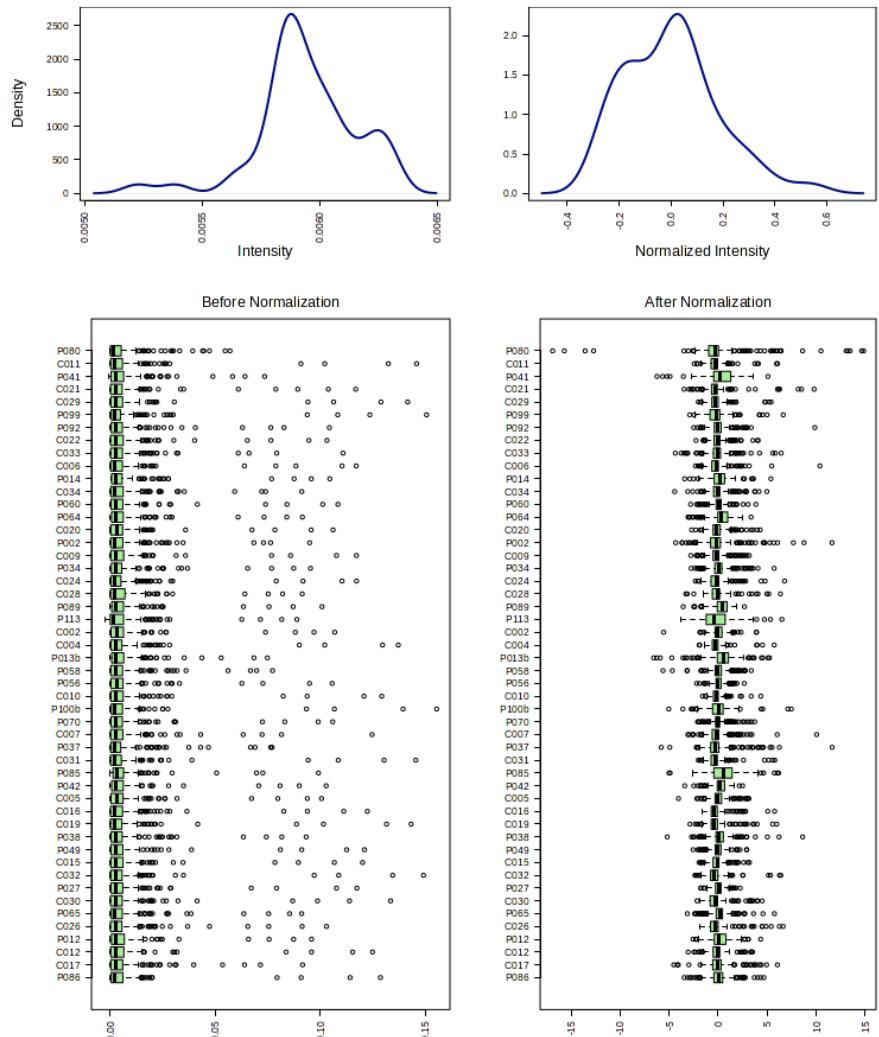
$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$$

Pareto scaling

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

MetaboAnalyst workflow

3) data normalization

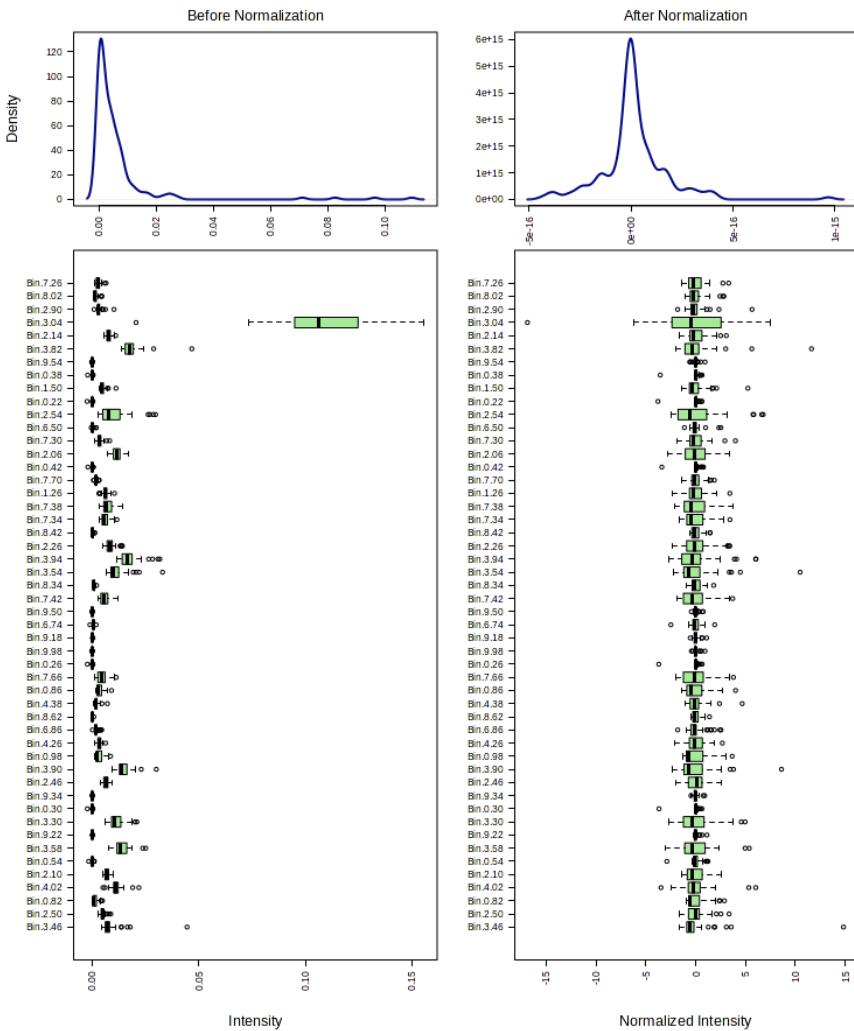


Effect of normalization over sample

MetaboAnalyst workflow

3) data normalization

Effect of features/metabolites scaling



MetaboAnalyst workflow

4) statistical analysis

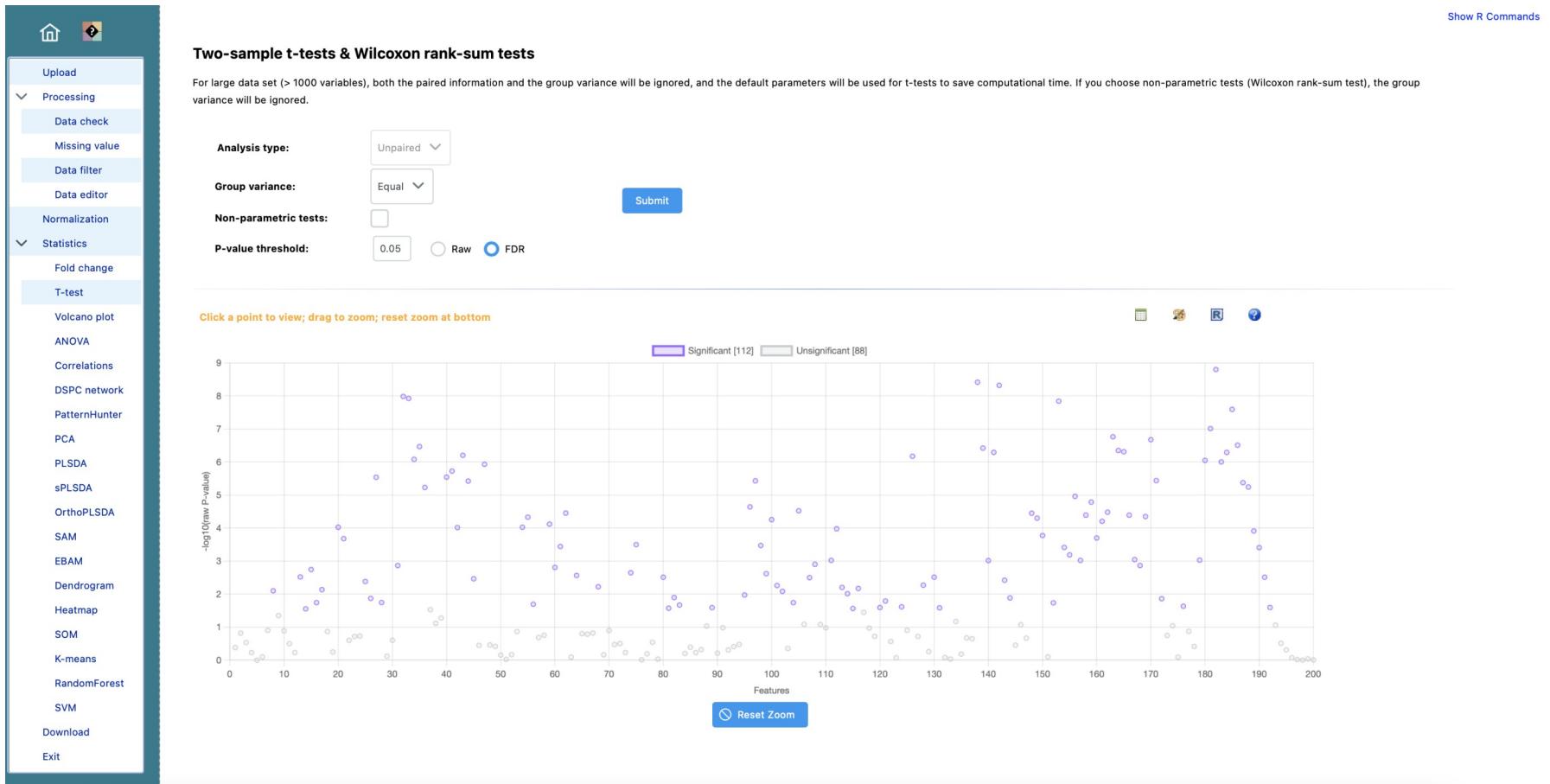
The screenshot shows the MetaboAnalyst software interface. On the left, a sidebar menu lists various analysis paths: Upload, Processing (Data check, Missing value), Data filter, Data editor, Normalization, Statistics (selected), Download, and Exit. The main content area is titled "Select an analysis path to explore:" and contains several sections:

- Univariate Analysis**: Includes [Fold Change Analysis](#), [T-tests](#), [Volcano plot](#), [One-way Analysis of Variance \(ANOVA\)](#), [Correlation Heatmaps](#), [Pattern Search](#), and [Correlation Networks \(DSPC\)](#).
- Advanced Significance Analysis**: Includes [Significance Analysis of Microarray \(and Metabolites\) \(SAM\)](#) and [Empirical Bayesian Analysis of Microarray \(and Metabolites\) \(EBAM\)](#).
- Chemometrics Analysis**: Includes [Principal Component Analysis \(PCA\)](#), [Partial Least Squares - Discriminant Analysis \(PLS-DA\)](#), [Sparse Partial Least Squares - Discriminant Analysis \(sPLS-DA\)](#), and [Orthogonal Partial Least Squares - Discriminant Analysis \(orthoPLS-DA\)](#).
- Cluster Analysis**: Includes [Hierarchical Clustering: Dendrogram](#) and [Heatmaps](#), and [Partitional Clustering: K-means](#) and [Self Organizing Map \(SOM\)](#).
- Classification & Feature Selection**: Includes [Random Forest](#) and [Support Vector Machine \(SVM\)](#).

A large red bracket on the right side groups the first two sections (Univariate Analysis and Advanced Significance Analysis) under the heading "«Classical» analysis of variance among groups". Another large red bracket groups the last three sections (Chemometrics Analysis, Cluster Analysis, and Classification & Feature Selection) under the heading "Machine learning algorithms".

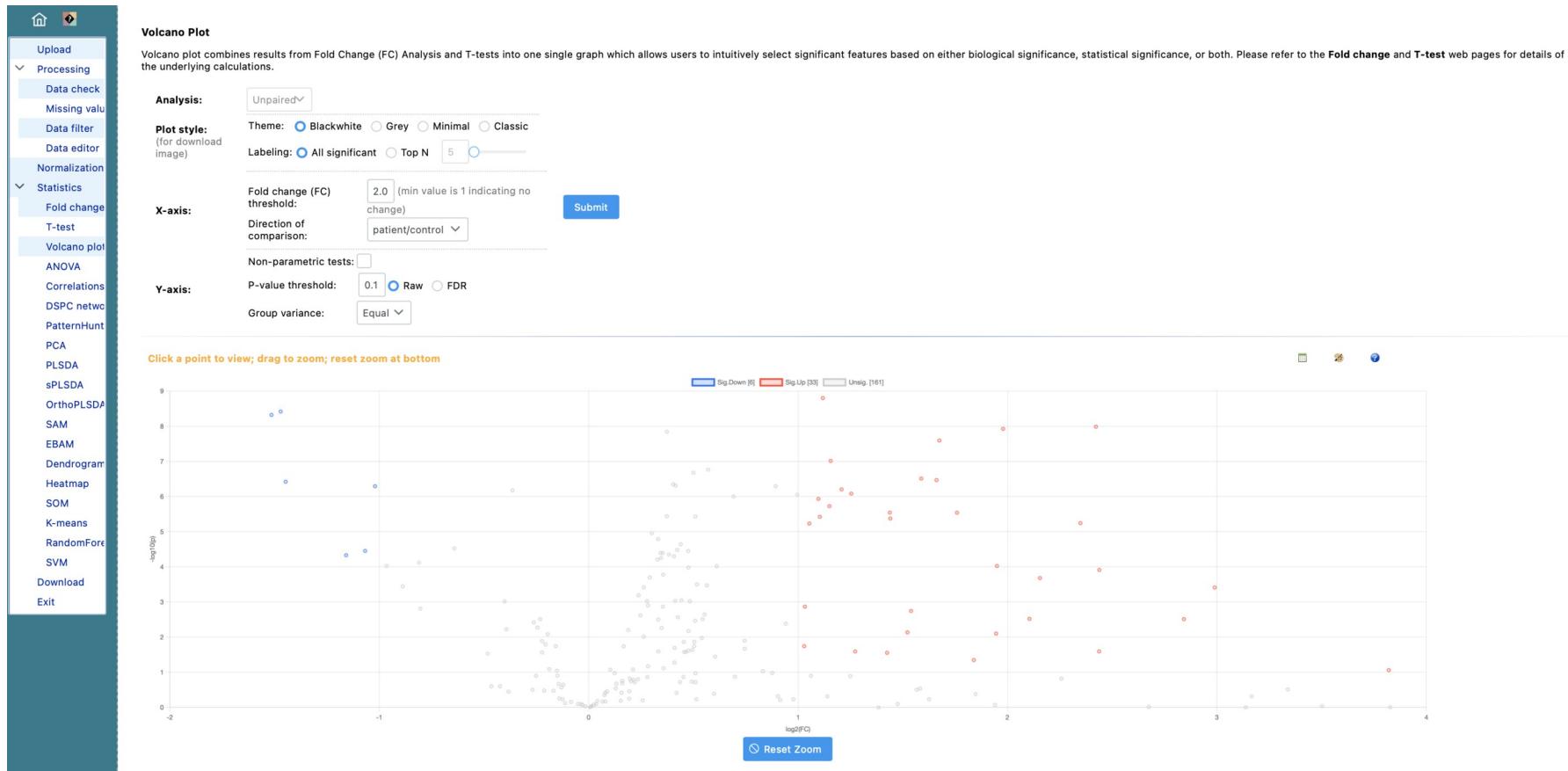
MetaboAnalyst workflow

4) univariate analysis



MetaboAnalyst workflow

4) univariate analysis



MetaboAnalyst workflow

5) chemometric analysis



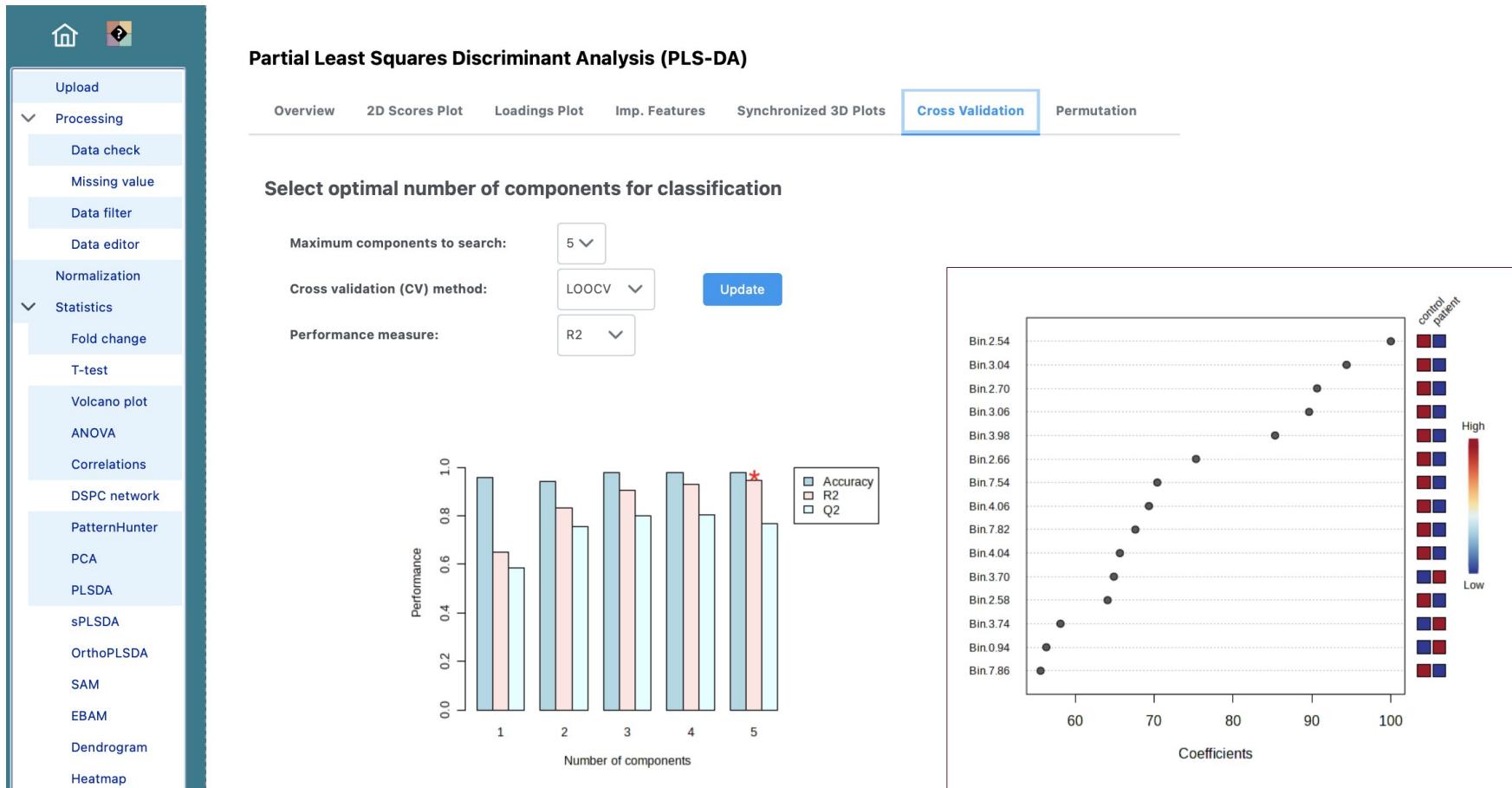
MetaboAnalyst workflow

5) chemometric analysis



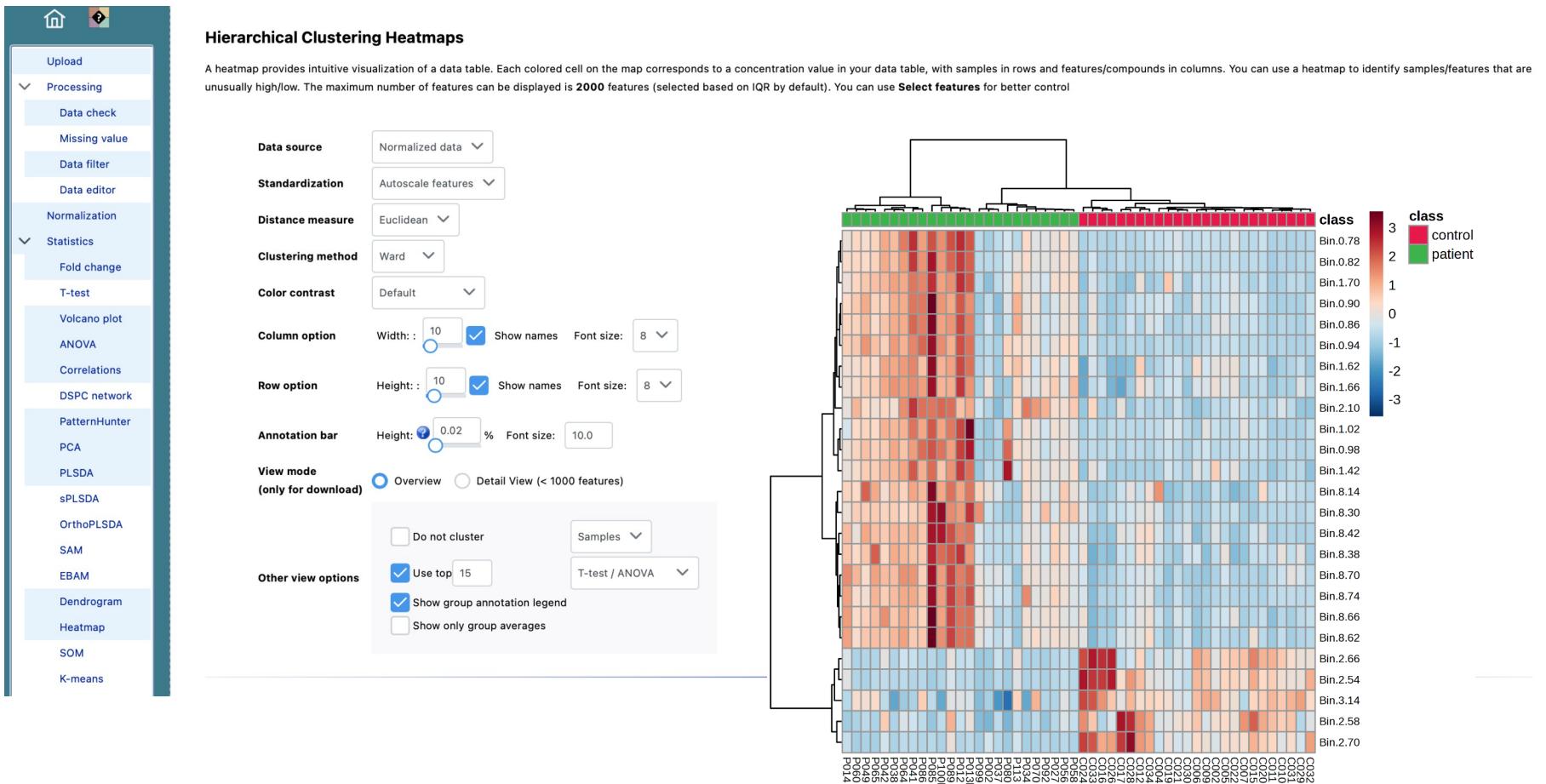
MetaboAnalyst workflow

5) chemometric analysis



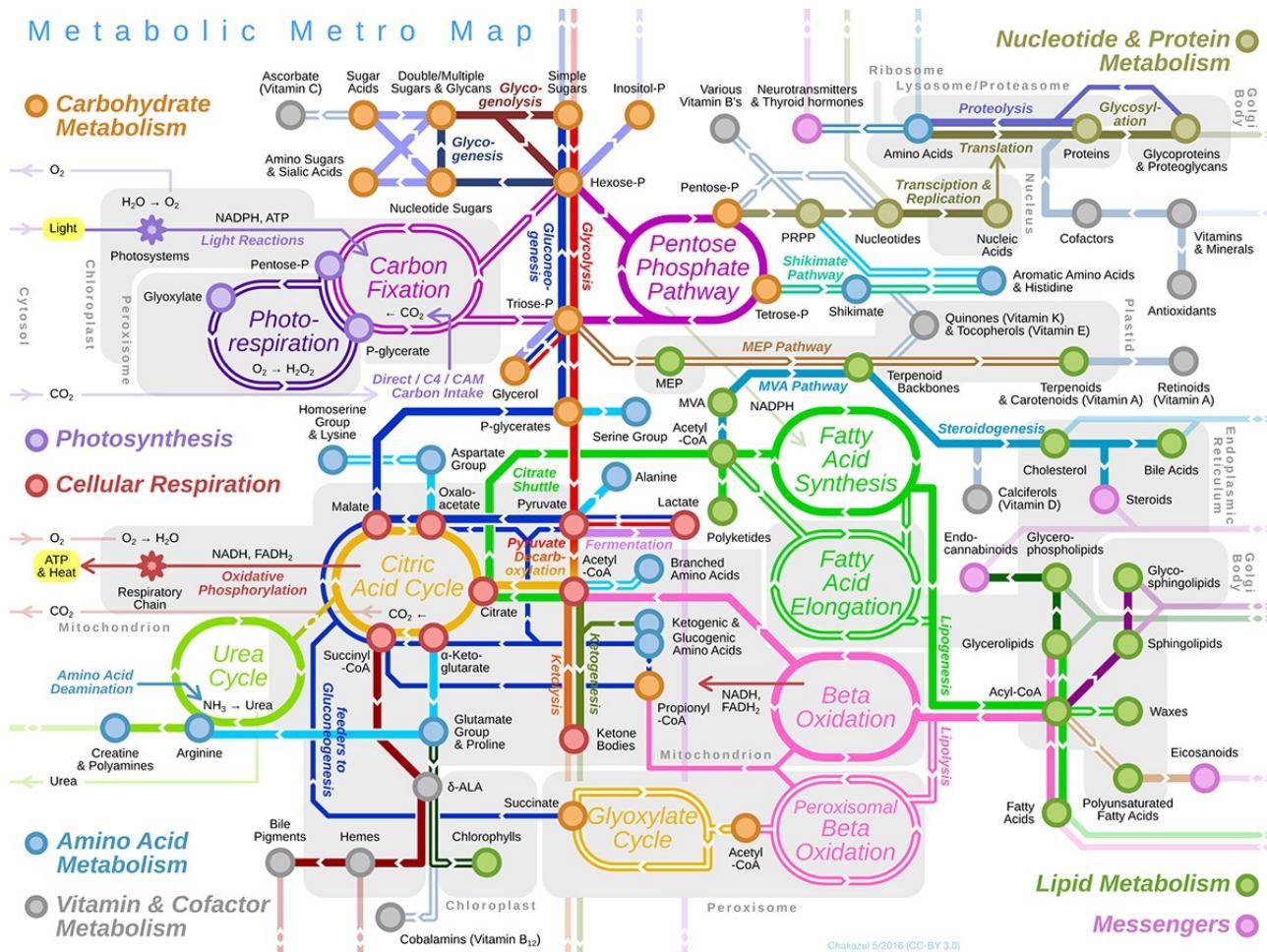
MetaboAnalyst workflow

5) chemometric analysis



Heatmap of the top 25 T-test features

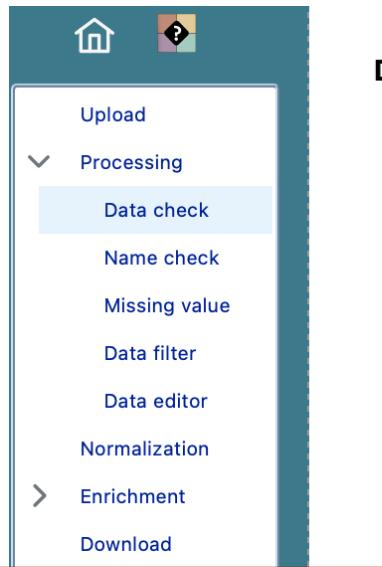
Identifying the metabolic pathways deregulated by a pathology is finding a target for pharmacological therapy!



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

MetaboAnalyst workflow

6) enrichment analysis



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

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

MetaboAnalyst workflow

6) enrichment analysis

The screenshot shows the MetaboAnalyst interface for 'Name/ID Standardization'. On the left, a sidebar lists various tools: Upload, Processing, Data check (selected), Name check, Missing value, Data filter, Data editor, Normalization, Enrichment, Download, and Exit. The main area displays a table comparing 'Query' compound names with their corresponding 'Hit' names.

Query	Hit
1,6-Anhydro-beta-D-glucose	Levoglucosan
1-Methylnicotinamide	1-Methylnicotinamic acid
2-Aminobutyrate	L-alpha-Aminobutyrate
2-Hydroxyisobutyrate	2-Hydroxyisobutyric acid
2-Oxoglutarate	Oxoglutaric acid
3-Aminoisobutyrate	3-Aminoisobutanoid
3-Hydroxybutyrate	
3-Hydroxyisovalerate	3-Hydroxyisovaleric acid
3-Indoxylsulfate	Indoxyl sulfate
4-Hydroxyphenylacetate	p-Hydroxyphenylacetic acid
Acetate	Acetic acid
Acetone	Acetone
Adipate	Adipic acid
Alanine	Alanine

A red box highlights the row for '3-Hydroxybutyrate' in the 'Query' column. A modal dialog titled 'Name match' is displayed, listing potential matches:

Matched Name	HMDB	PubChem	KEGG
3-Hydroxyisovaleric acid	HMDB0000754	69362	C20827
<input checked="" type="checkbox"/> 3-Hydroxybutyric acid	HMDB0000011	441	C01089
<input type="checkbox"/> (S)-3-Hydroxybutyric acid	HMDB0000442	94318	C03197
<input type="checkbox"/> Ethyl (±)-3-hydroxybutyrate	HMDB0040409	62572	NA
<input type="checkbox"/> Methyl 3-hydroxybutyrate	HMDB0041603	15146	NA
<input type="checkbox"/> L-Threonine	HMDB0000167	6288	C00188
<input type="checkbox"/> 4-Amino-3-hydroxybutyrate	HMDB0061877	2149	C03678
<input type="checkbox"/> 2-Methyl-3-hydroxybutyric acid	HMDB0000354	160471	NA
<input type="checkbox"/> None of the above			

At the bottom of the dialog are 'OK' and 'Cancel' buttons.

MetaboAnalyst workflow

6) enrichment analysis

The screenshot shows the 'Parameter Setting' page of the MetaboAnalyst workflow. The left sidebar lists various steps: Upload, Processing, Data check, Name check, Missing value, Data filter, Data editor, Normalization, Enrichment (which is selected), Set paramet, View result, Download, and Exit. The main content area is titled 'Please select a metabolite set library'. It contains a table with five rows, each representing a different type of library:

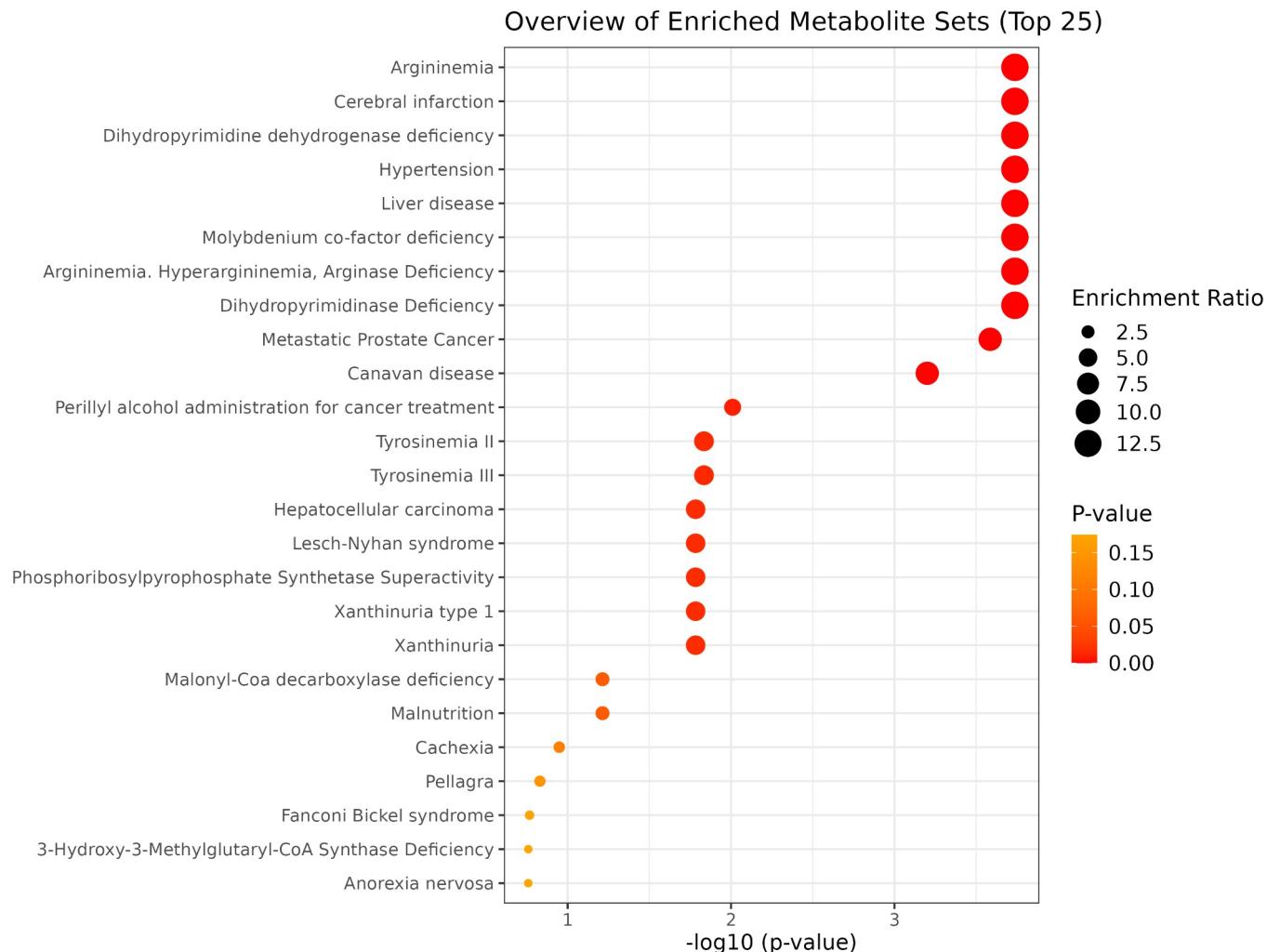
Pathway based	<input type="radio"/> SMPDB <input type="radio"/> KEGG <input checked="" type="radio"/> Drug related <input type="radio"/> RaMP-DB	99 metabolite sets based on normal human metabolic pathways. 80 metabolite sets based on KEGG human metabolic pathways (Dec. 2023). 461 metabolite sets based on drug pathways from SMPDB. 3694 metabolite and lipid pathways from RaMP-DB (integrating KEGG via HMDB, Reactome, WikiPathways).
Disease signatures	<input type="radio"/> Blood <input checked="" type="radio"/> Urine <input type="radio"/> CSF <input type="radio"/> Feces	480 metabolite sets reported in human blood. 385 metabolite sets reported in human urine. 174 metabolite sets reported in human cerebral spinal fluid (CSF). 67 metabolite sets reported in human feces.
Chemical structures	<input type="radio"/> Super-class <input type="radio"/> Main-class <input type="radio"/> Sub-class	39 super chemical class metabolite sets or lipid sets 617 main chemical class metabolite sets or lipid sets 1250 sub chemical class metabolite sets or lipid sets
Other types	<input type="radio"/> SNPs <input type="radio"/> Predicted <input type="radio"/> Locations <input type="radio"/> Exposure	4,598 metabolite sets based on their associations with SNPs loci. 912 metabolic sets predicted to change in the case of dysfunctional enzymes. 78 metabolite and lipid sets based on organ, tissue, and subcellular localizations. 62 metabolite sets based on dietary and chemical exposures.
Self defined	<input type="radio"/> Upload here	define your own customized metabolite sets

Below the table, there is a checkbox: Only use metabolite sets containing at least 2 entries. A note says 'Please specify a reference metabolome'. Two radio button options are shown: Use all the compounds in the selected library and Upload a reference metabolome based on your analytical platform. A 'Submit' button is at the bottom.

Enrichment analysis, based on the glo baltest, 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

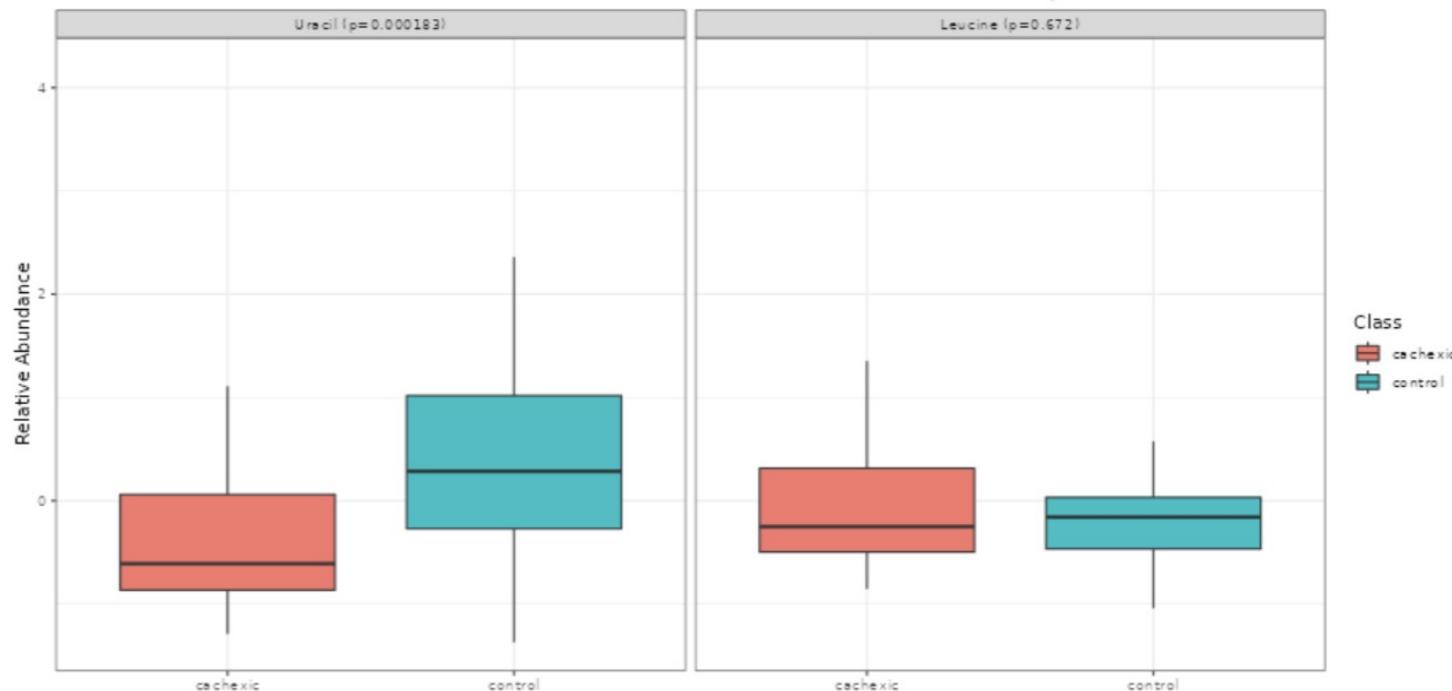
6) enrichment analysis



MetaboAnalyst workflow

6) functional interpretation

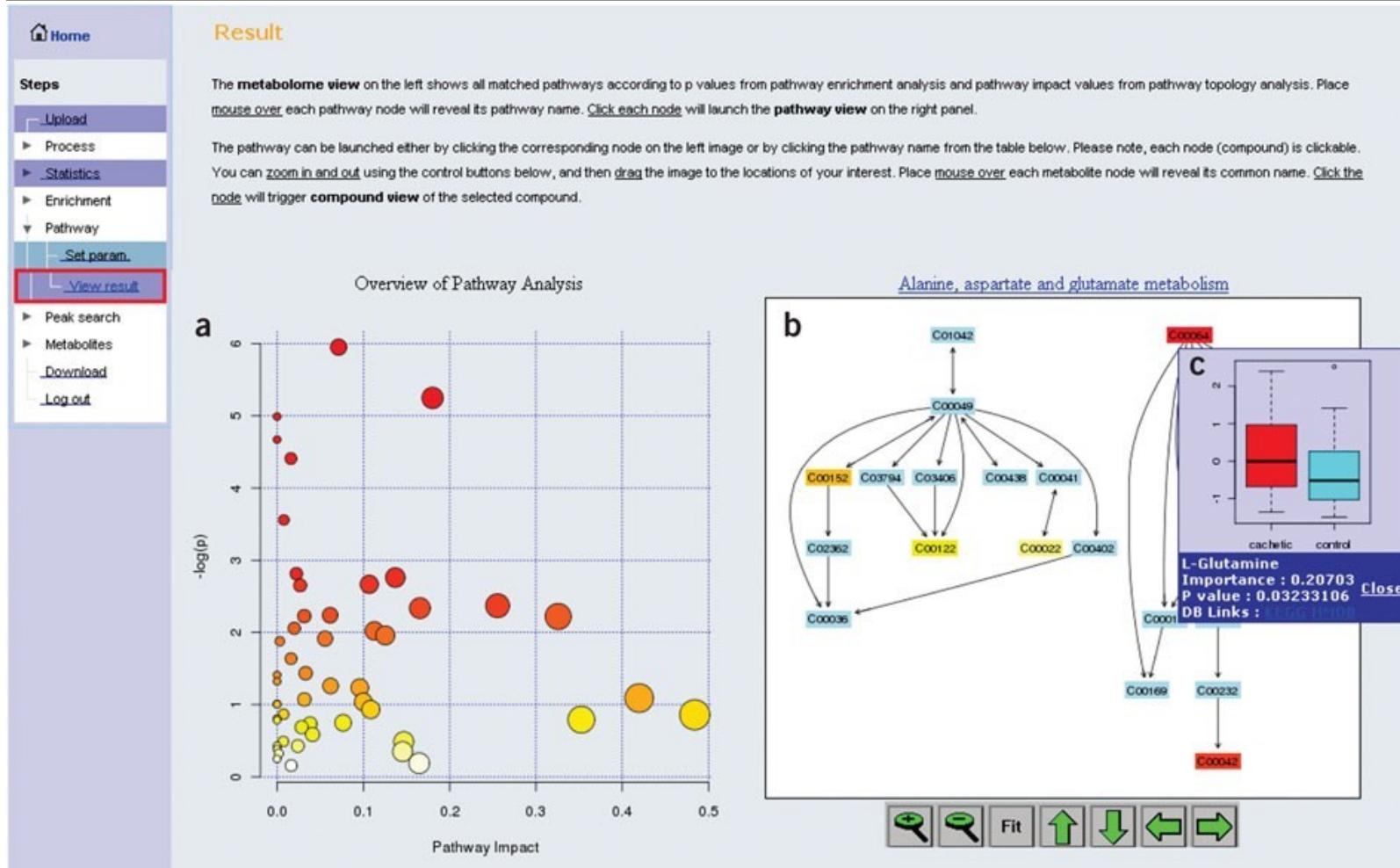
Current metabolite set:



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

- 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
<ul style="list-style-type: none">• is the hypothesis that a sample data statistic occurs purely from chance<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• is the hypothesis that a sample data statistic is influenced by some non-random cause<ul style="list-style-type: none">• 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.

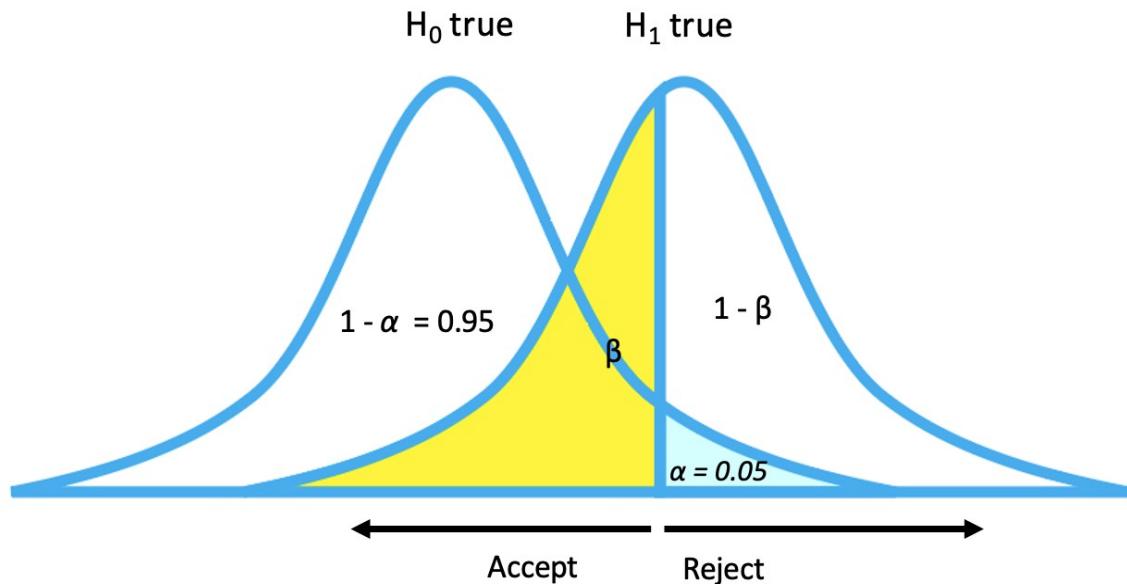
- Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called alpha, and is often denoted by α .
- Type II error occurs when the researcher fails to reject a null 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 H_0	Reject H_0
H_0 is true	Correct action	Type I error FALSE POSITIVE
probability	$1-\alpha$	α
H_1 is true	Type II error FALSE NEGATIVE	Correct action
probability	β	$power = 1-\beta$

$$\alpha = P(H_1 | H_0)$$

$$\beta = P(H_0 | H_1)$$



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:

- False positive: although the quality control has been certified
 - False negative: presence of disease was as a defendant in the criminal trial
- Example 2. Quality control in a pharma production company
- False positive: We rejected the hypothesis of disease as a defendant in the criminal trial
 - False negative: was found guilty and is sent to prison or receives the death penalty
- Example 3. Disease diagnosis
- False positive: a criminal is declared innocent and escapes punishment
 - False negative: an innocent citizen is found guilty and is sent to prison or receives the death penalty
- Example 3. Criminal court

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

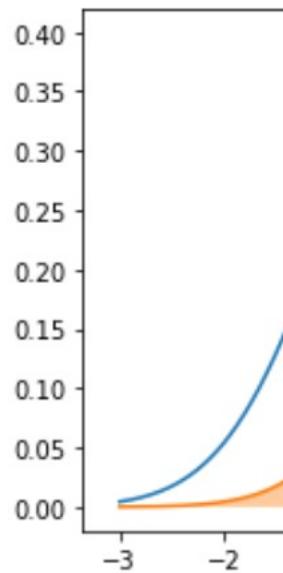


Figure 1: Equally informative features
and false negative rate

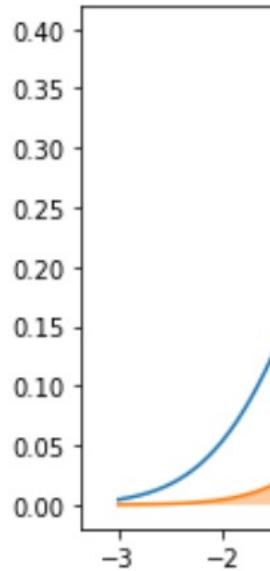


Figure 2: Greater uncertainty in true class
than false negative rate

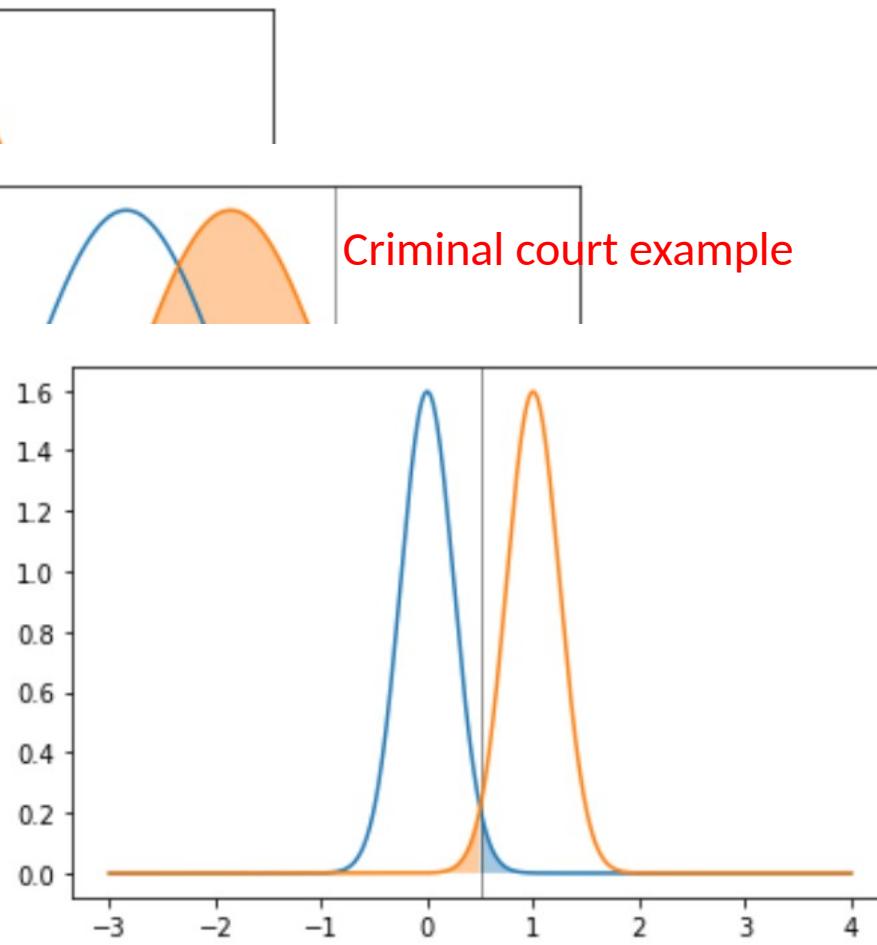
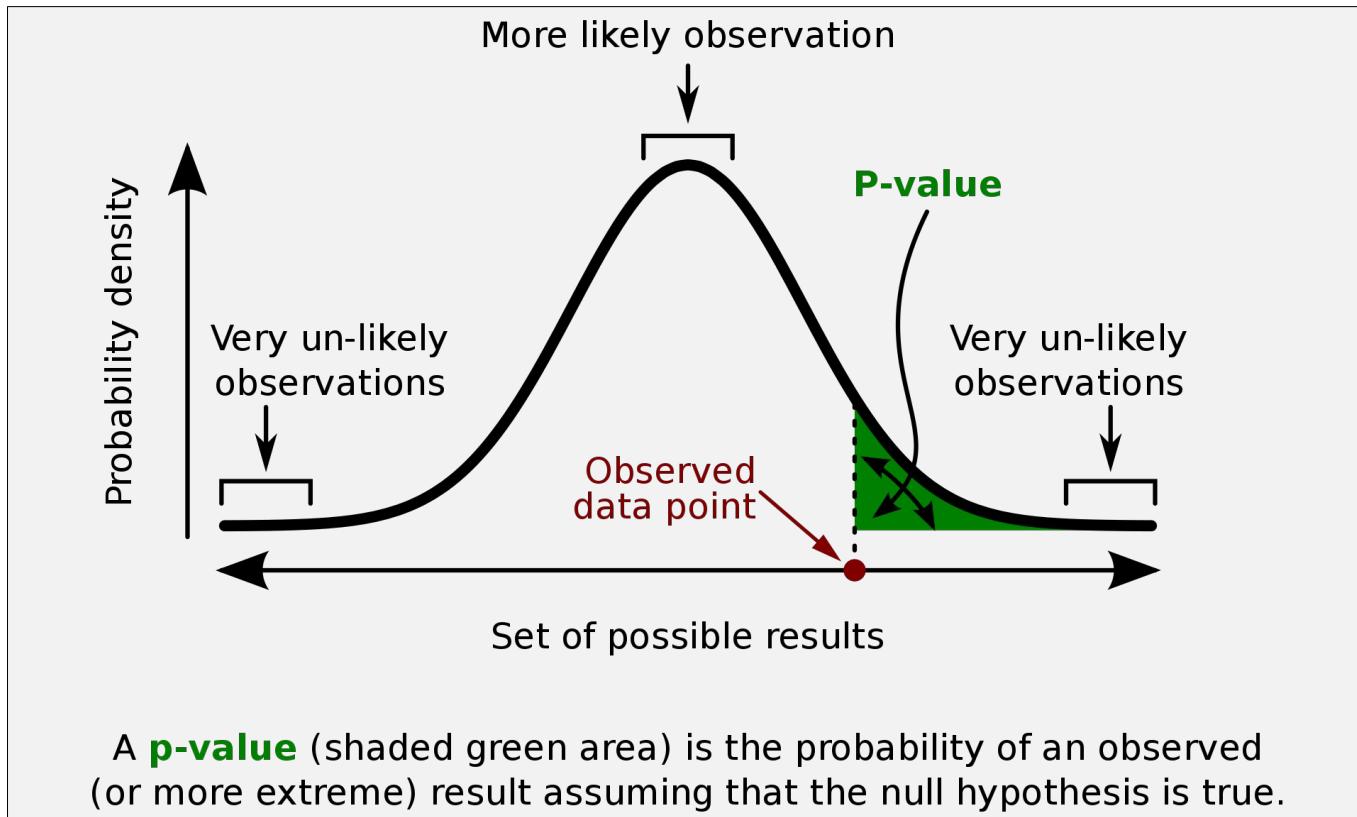


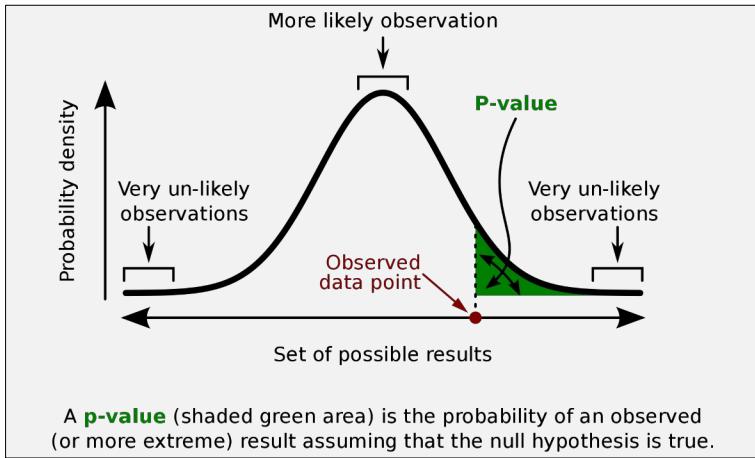
Figure 3: Lowered uncertainty through
more informative features.

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?



p-value interpretation



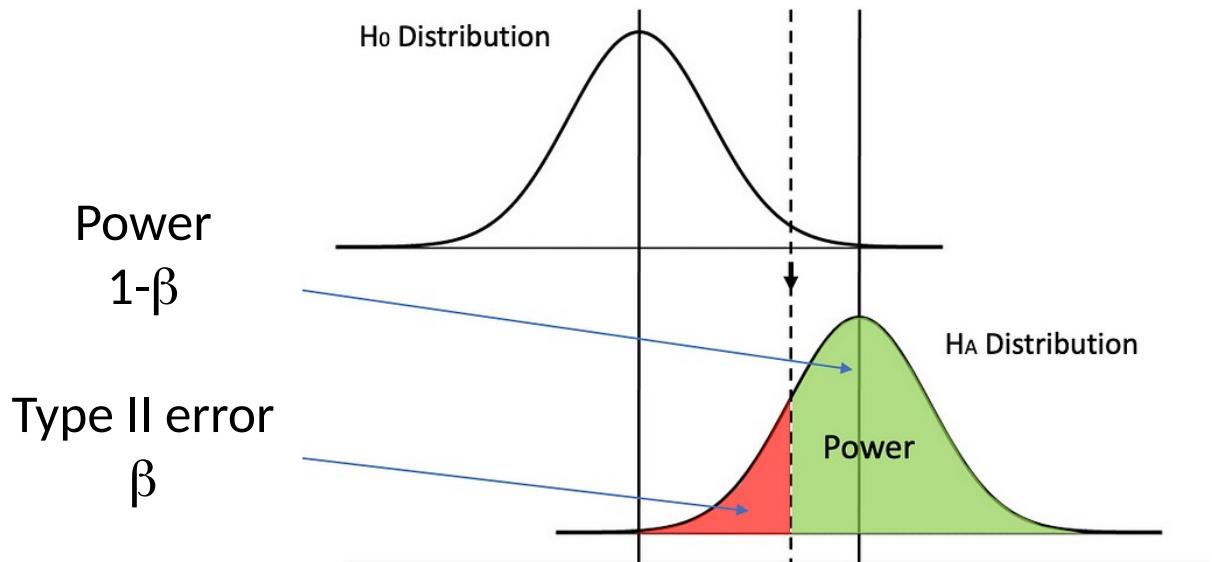
- A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.
- Usually the researcher fix α before experiment and then compare the p-value and takes a decision.

Conventions

$P > 0.10$	\Rightarrow	<i>non-significant evidence against H_0</i>
$0.05 < P \leq 0.10$	\Rightarrow	<i>marginally significant evidence against H_0</i>
$0.01 < P \leq 0.05$	\Rightarrow	<i>significant evidence against H_0</i>
$P \leq 0.01$	\Rightarrow	<i>highly significant evidence against H_0</i>

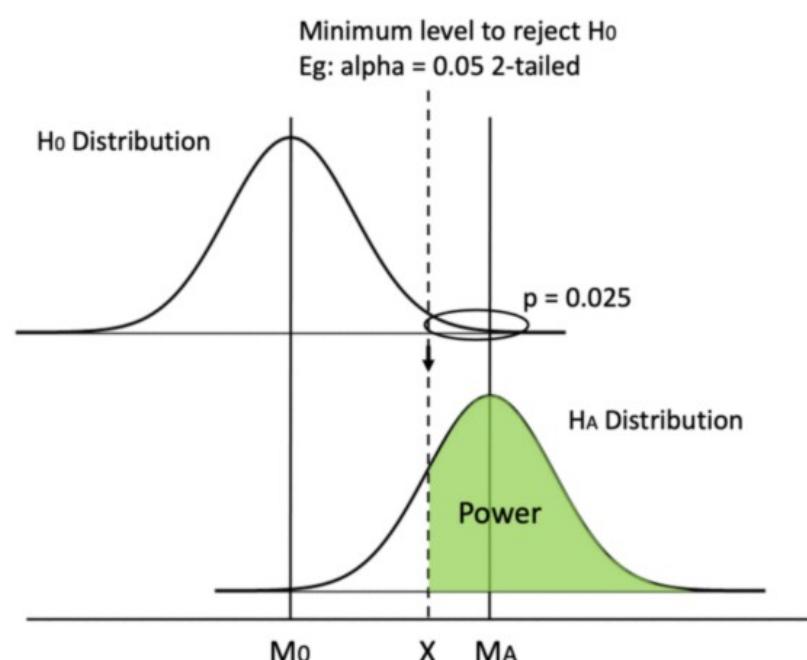
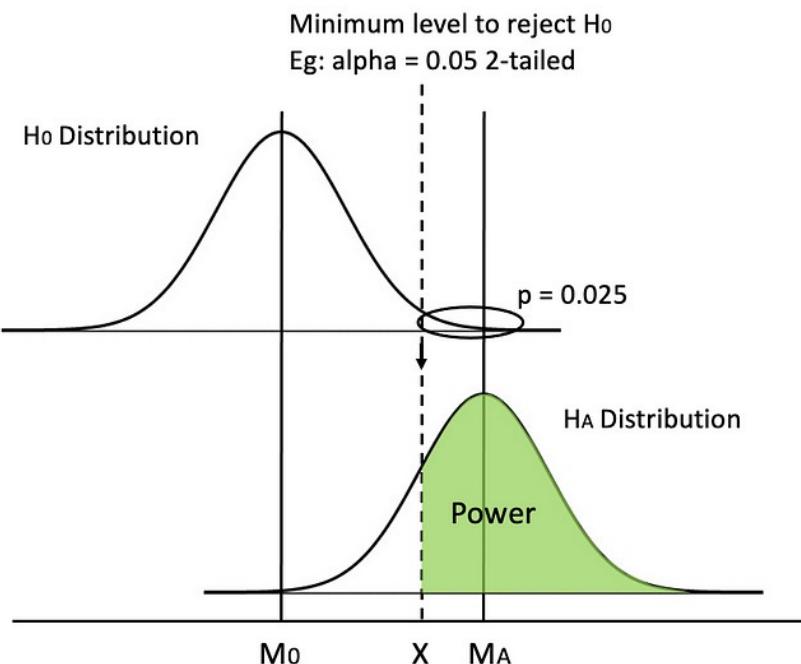
How to increase statistical power

	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	$power = 1-\beta$



How to increase statistical power

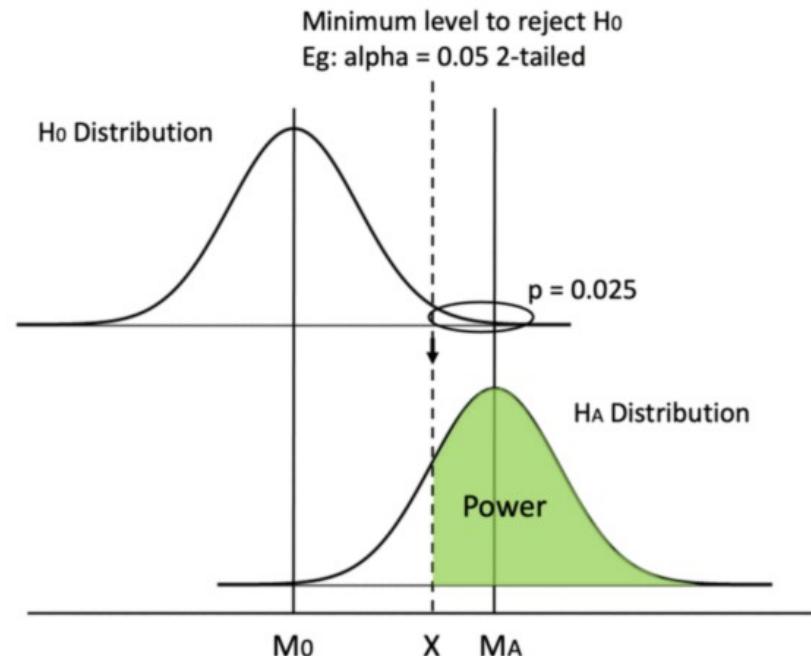
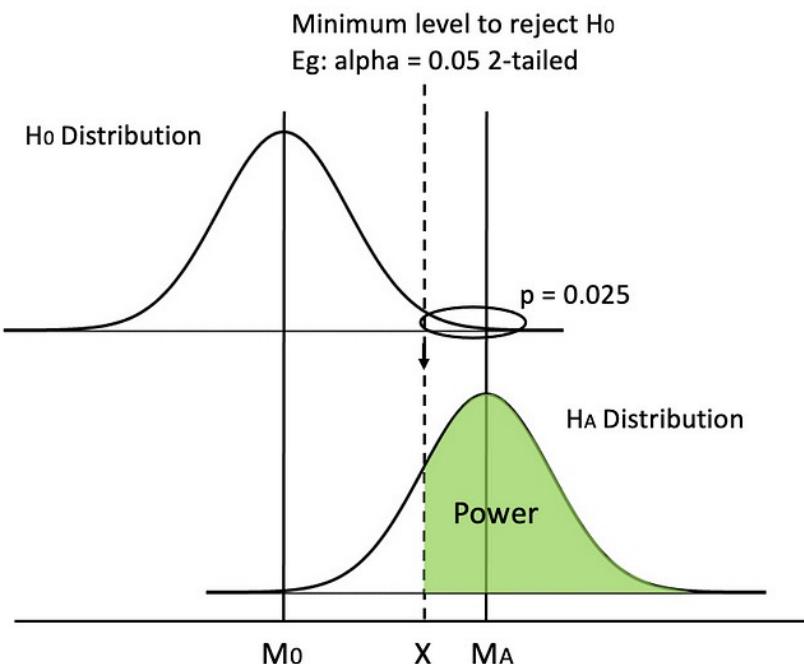
1) Raise significance level alpha (the **WRONG** way)



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

How to increase statistical power

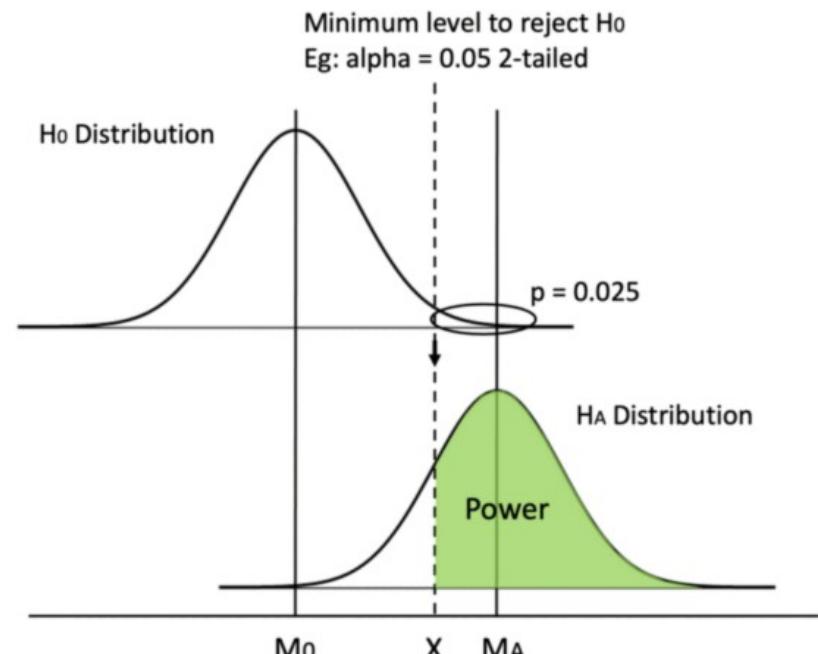
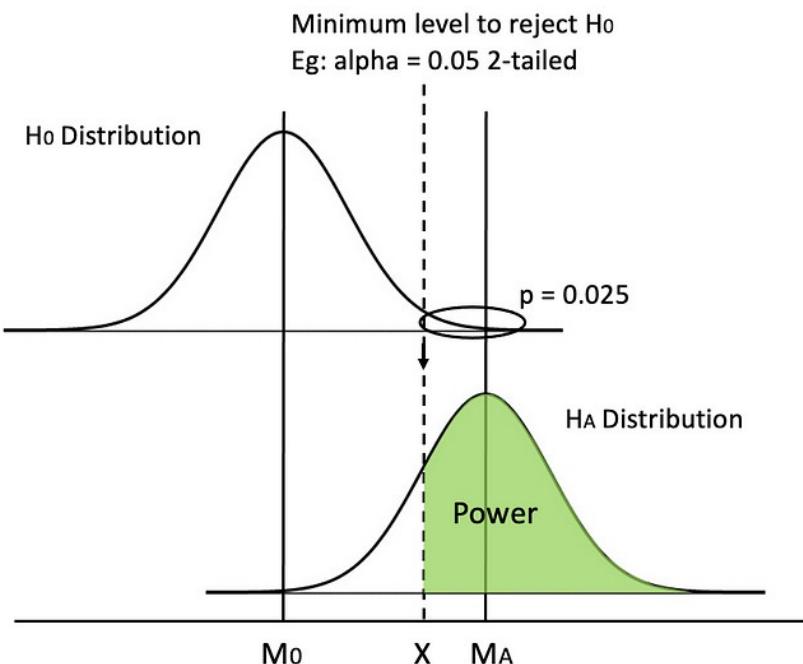
2) Switch from a 2-tailed test to a 1-tailed test (**CORRECT** if possible)



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

How to increase statistical power

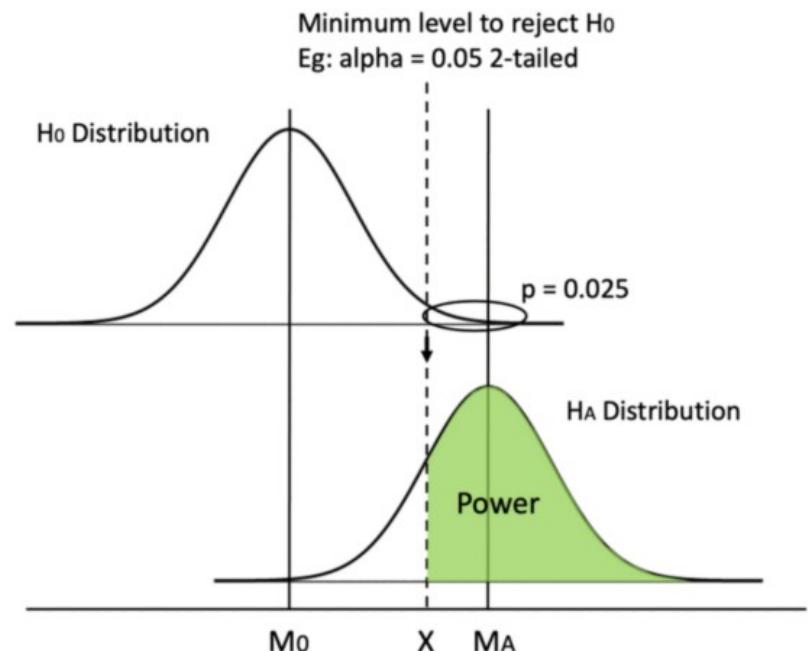
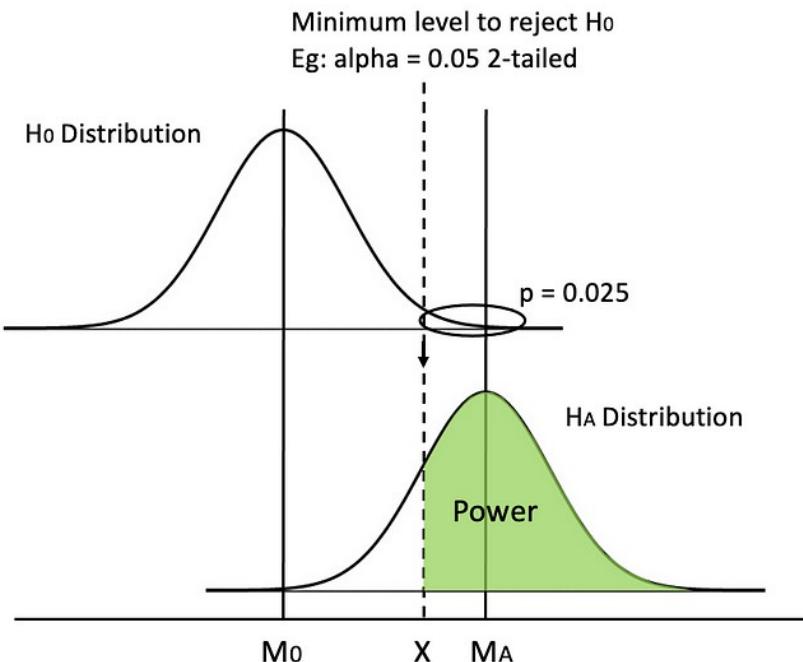
3) Increase mean difference (or increase the effect size)



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

How to increase statistical power

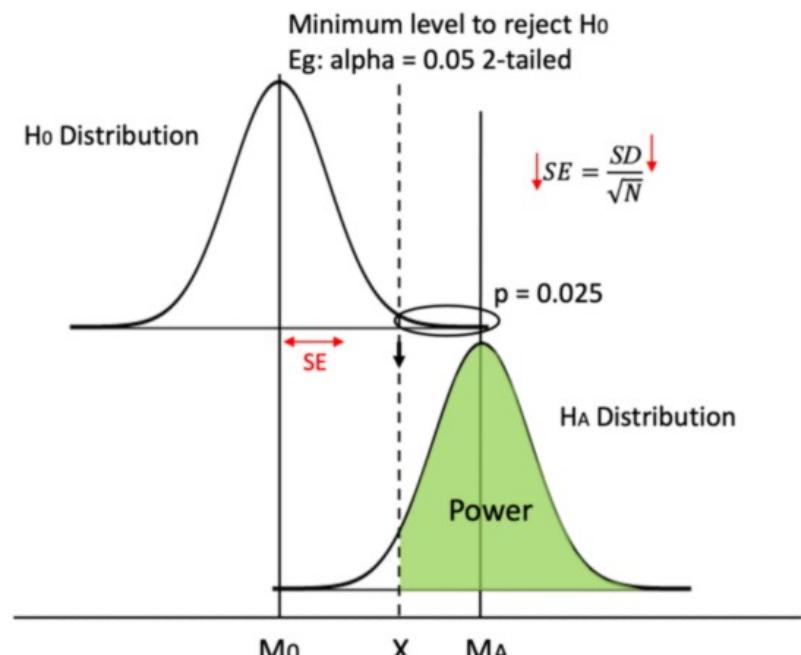
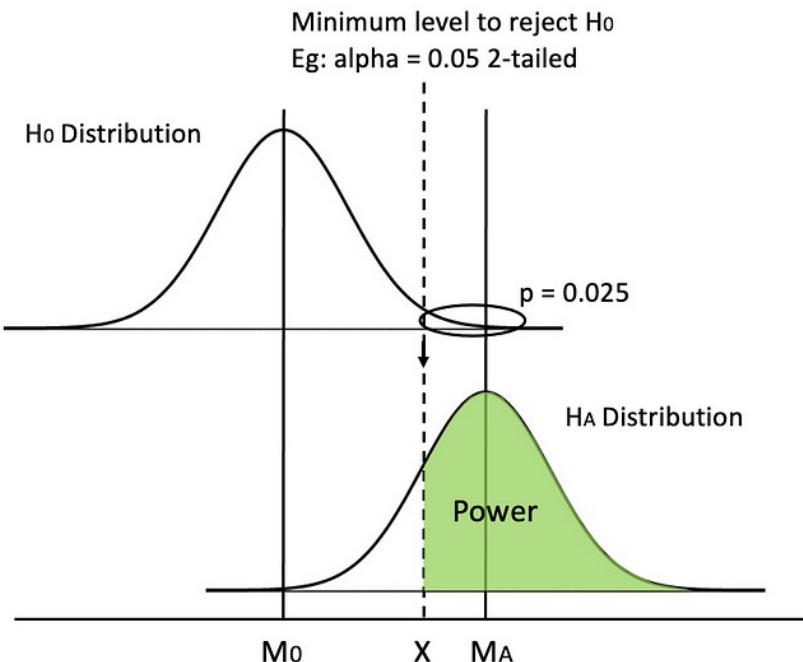
4) Use z distribution instead of t distribution (appropriate when we know the population mean)



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

How to increase statistical power

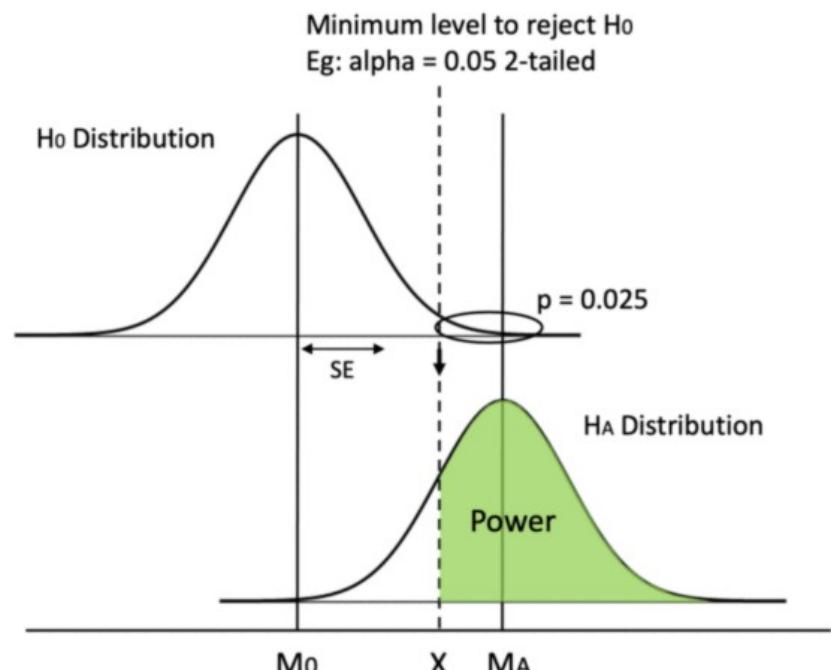
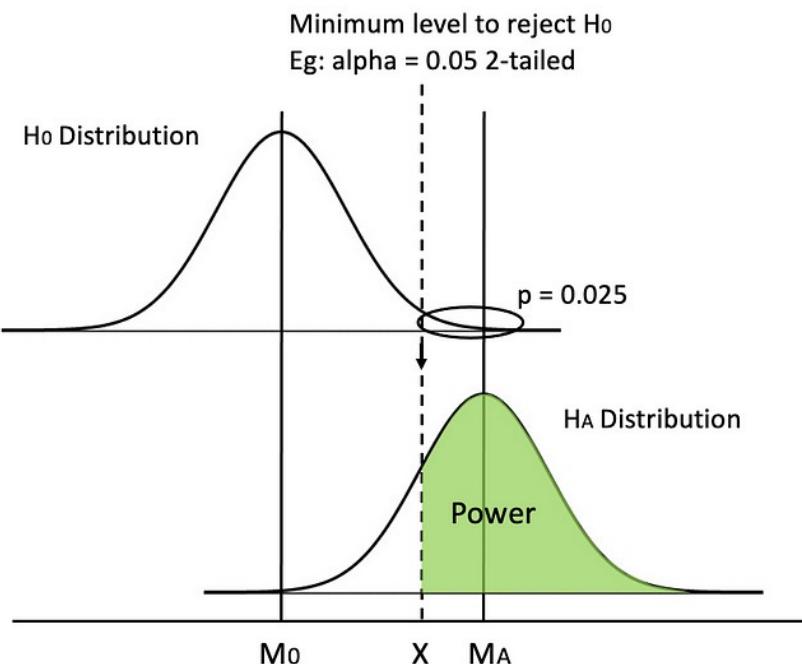
5) Decrease standard deviation (using more precise measurements to have less error and less noise)



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 you are doing:

1. Studying the mean difference between two groups

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

Effect size

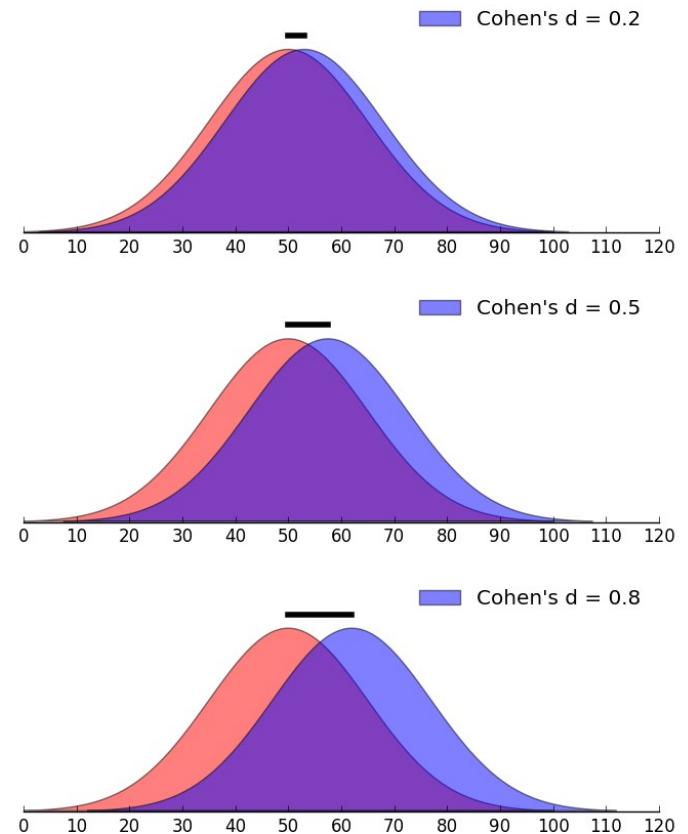
$$Cohen's\ d = \frac{\mu_1 - \mu_2}{\sigma}$$

Mean value of the population 1 Mean value of the population 2
Standard deviation of the population

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

Mean value of sample 1 Mean value of sample 2
Estimated standard deviation of the population from the sample

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



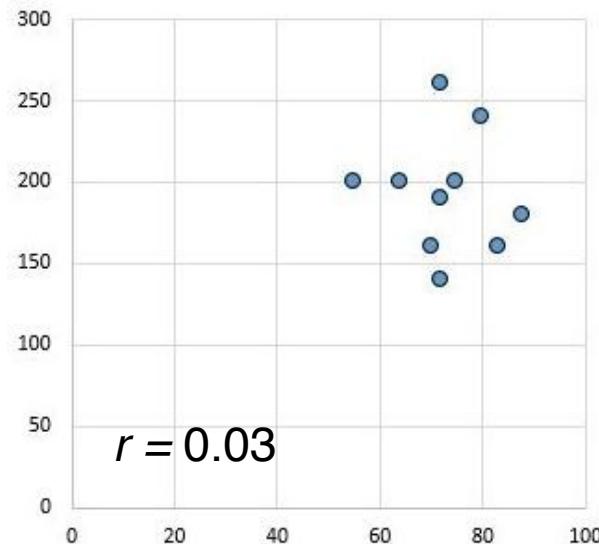
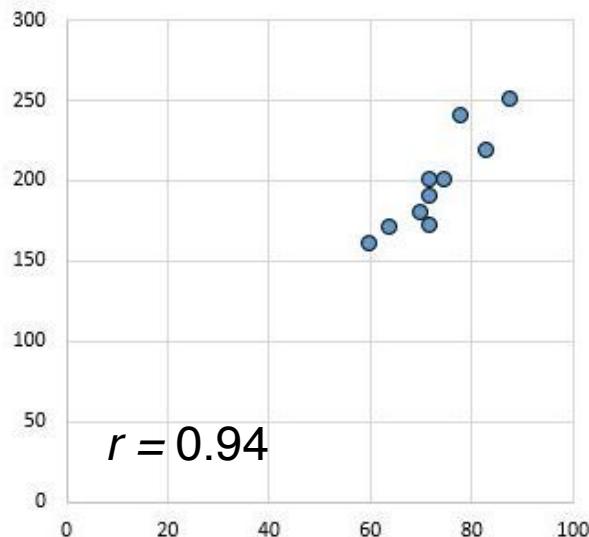
Effect size

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 $\omega =$	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