

# STATISTICS & ML WITH R

*(special topics)*

- **MetaboAnalyst**
- **Power Analysis**

**2024**

**M. Chiara Mimmi & Luisa M. Mimmi**

# WORKSHOP SCHEDULE

- Modules

- 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
- 6. Extra topics:
  - 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>

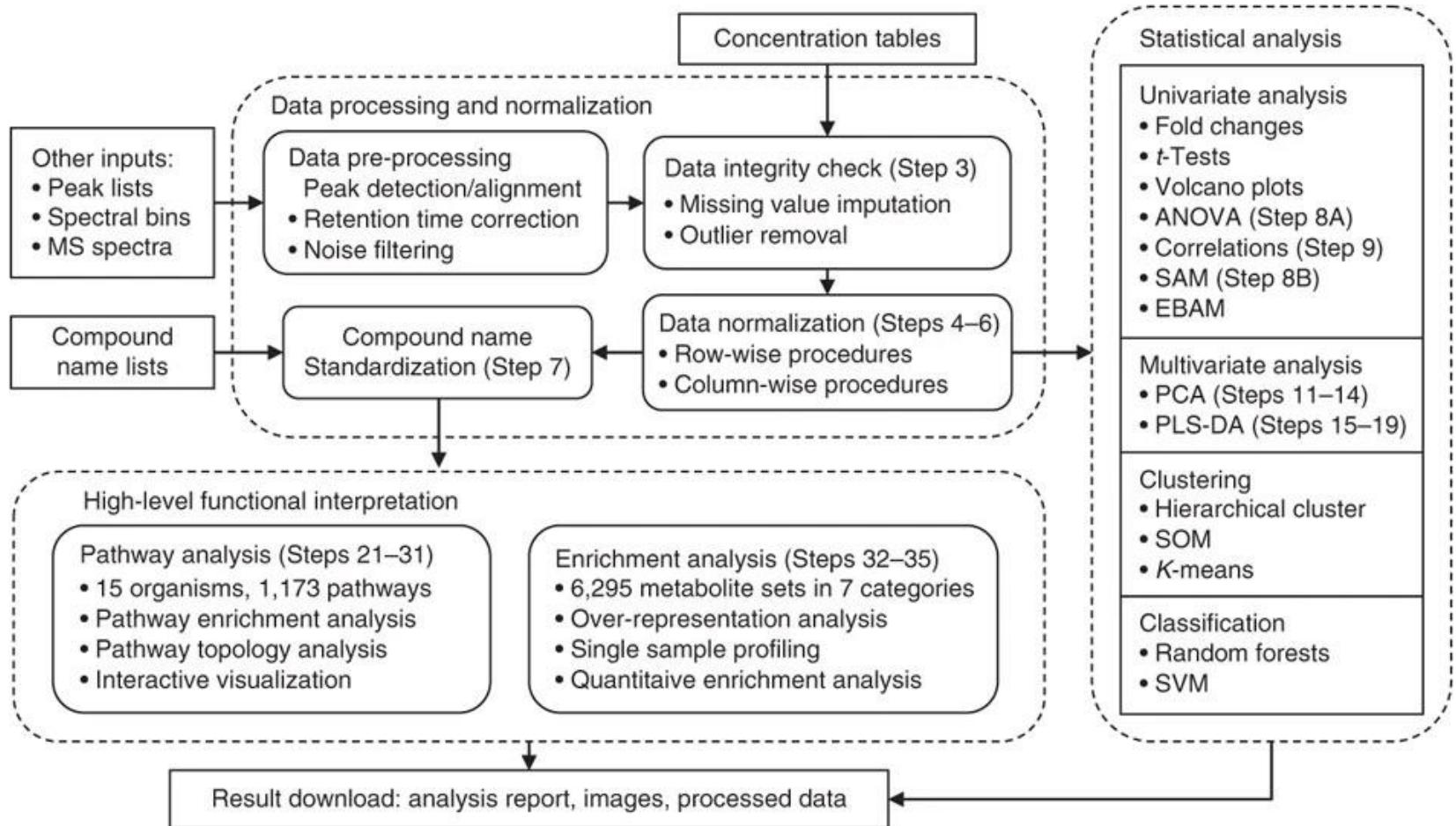
*From raw spectra to biomarkers, patterns, functions and systems biology*

- 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!

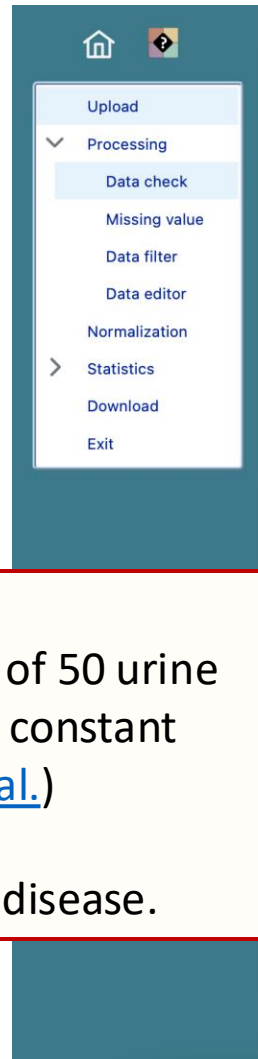
# MetaboAnalyst overview



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

# MetaboAnalyst workflow

## 1) data upload



### 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

### Test data 1:

Binned <sup>1</sup>H 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



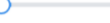


### Data Filtering:

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.](#)

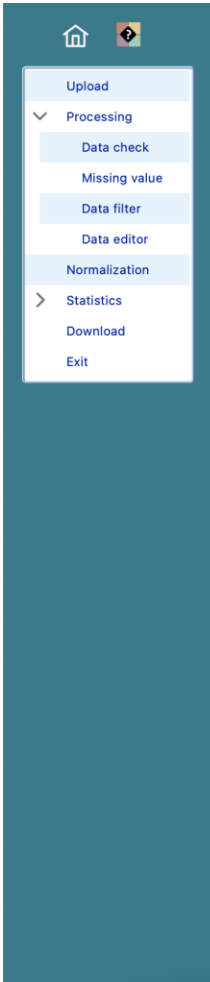
Non-informative variables can be characterized in three groups: 1) variables that show **low repeatability** - this can be measured using QC samples using the relative standard deviation ( $RSD = SD/mean$ ). Features with high percent RSD should be removed from the subsequent analysis (the suggested 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 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.

|                            |  |   |
|----------------------------|--|---|
| <b>Reliability filter:</b> | <input type="checkbox"/> Filtering features based on technical repeatability QC samples  | RSDs greater than:<br> 25%         |
| <b>Variance filter:</b>    | <input checked="" type="radio"/> Interquartile range (IQR)<br><input type="radio"/> Standard deviation (SD)<br><input type="radio"/> Median absolute deviation (MAD)<br><input type="radio"/> Relative standard deviation (RSD = SD/mean)<br><input type="radio"/> Non-parametric relative standard deviation (MAD/median) | Percentage to filter out:<br> 5%   |
| <b>Abundance filter:</b>   | <input checked="" type="radio"/> Mean intensity value<br><input type="radio"/> Median intensity value  | Percentage to filter out:<br> 0% |

# MetaboAnalyst workflow

## 3) data normalization



### Normalization Overview:

The normalization procedures are grouped into three categories. You can use one or combine them to achieve better results.

- Sample normalization is for general-purpose adjustment for systematic differences among samples;
- Data transformation applies a mathematical transformation on individual values themselves. A simple mathematical approach is used to deal with negative values in log and square root Please search OmicsForum using "normalization #metaboanalyst" to find more information.
- Data scaling adjusts each variable/feature by a scaling factor computed based on the dispersion of the variable.

**Sample normalization**

None

Sample-specific normalization (i.e. weight, volume) [Specify](#)

Normalization by sum

Normalization by median

Normalization by a reference sample (PQN) [Specify](#)

Normalization by a pooled sample from group (group PQN) [Specify](#)

Normalization by reference feature [Specify](#)

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)

Normalize

View Result

Proceed

Autoscaling  $\tilde{x}_{ij}$

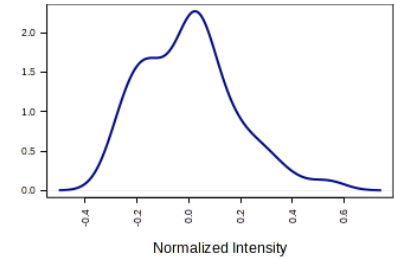
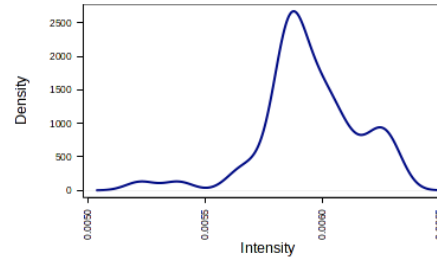
$$= \frac{x_{ij} - \bar{x}_i}{s_i}$$

Pareto scaling  $\tilde{x}_{ij}$

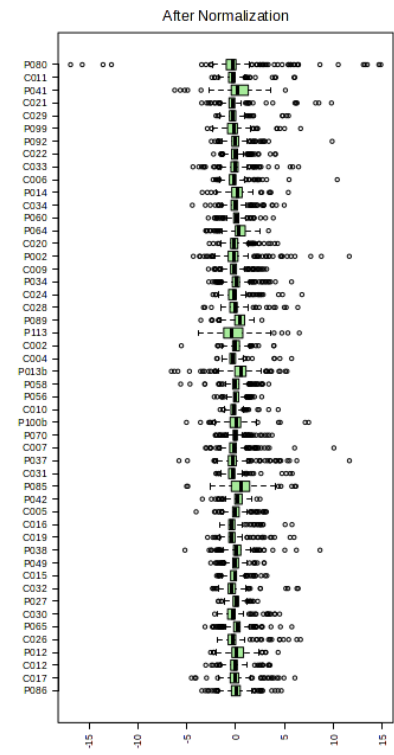
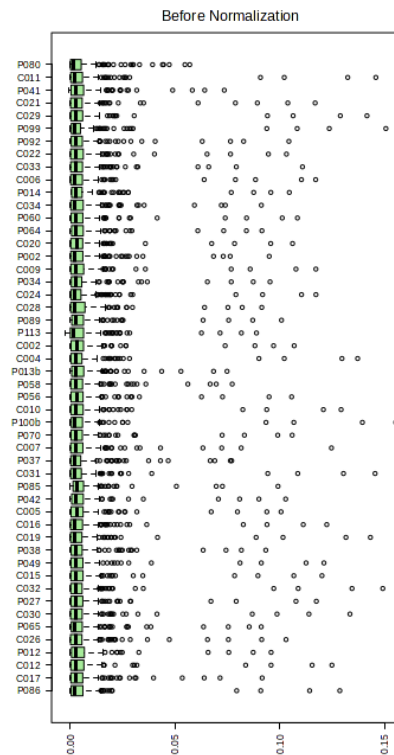
$$= \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$$

# MetaboAnalyst workflow

## 3) data normalization



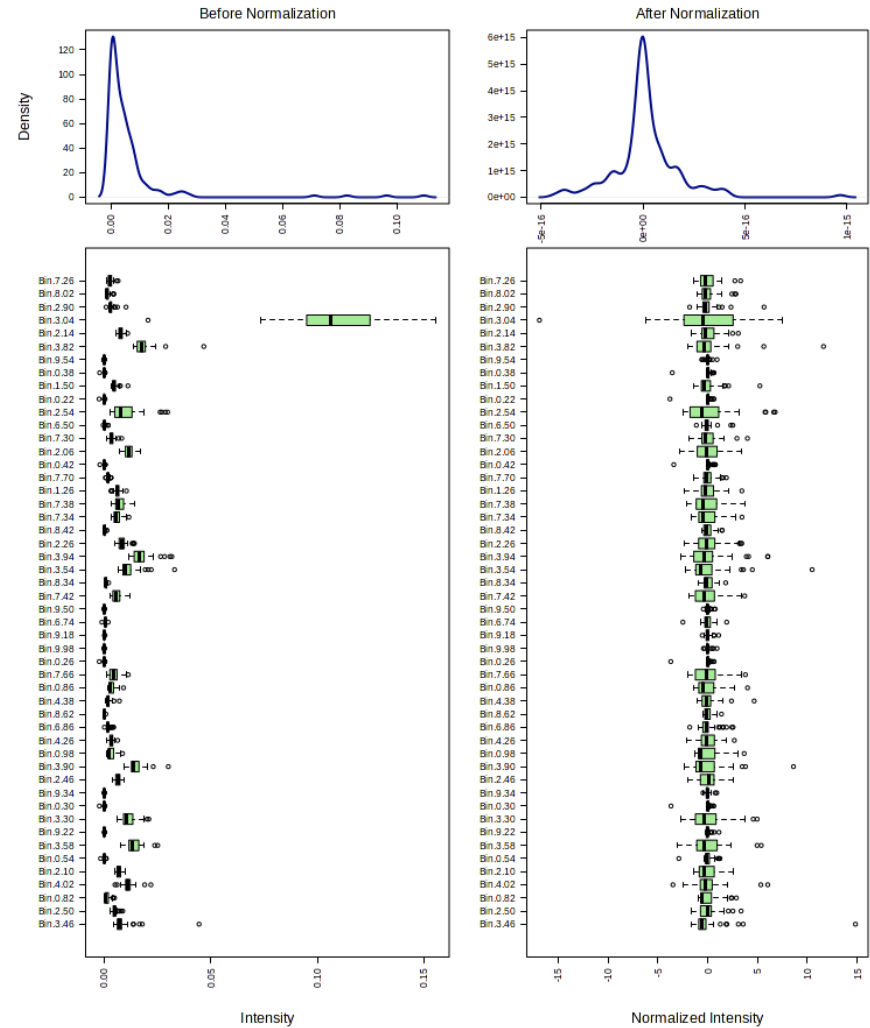
Effect of normalization over sample



# MetaboAnalyst workflow

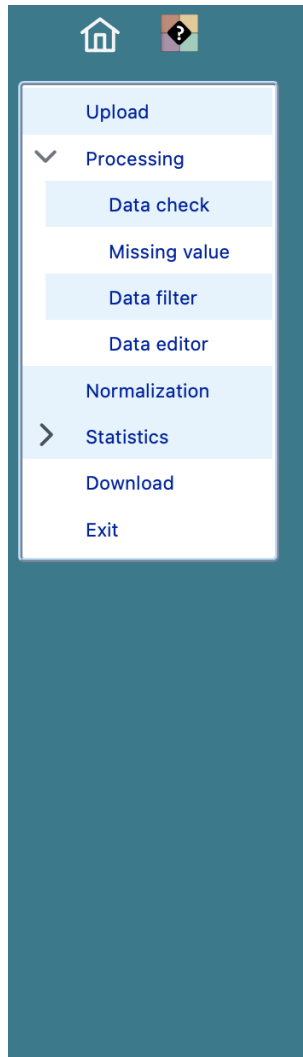
## 3) data normalization

Effect of features/metabolites scaling



# MetaboAnalyst workflow

## 4) statistical analysis



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 \(DSPC\)](#)

### 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 \(SVM\)](#)

«Classical» analysis of variance among groups

Machine learning algorithms

# MetaboAnalyst workflow

## 4) univariate analysis

Show R Commands

- Home
- Upload
- Processing
  - Data check
  - Missing value
  - Data filter
  - Data editor
- Normalization
- Statistics
  - Fold change
  - T-test
  - Volcano plot
  - ANOVA
  - Correlations
  - DSPC network
  - PatternHunter
  - PCA
  - PLSDA
  - sPLSDA
  - OrthoPLSDA
  - SAM
  - EBAM
  - Dendrogram
  - Heatmap
  - SOM
  - K-means
  - RandomForest
  - SVM
- Download
- Exit

### Two-sample t-tests & Wilcoxon rank-sum tests

For large data set (> 1000 variables), both the paired information and the group variance will be ignored, and the default parameters will be used for t-tests to save computational time. If you choose non-parametric tests (Wilcoxon rank-sum test), the group variance will be ignored.

**Analysis type:** Unpaired

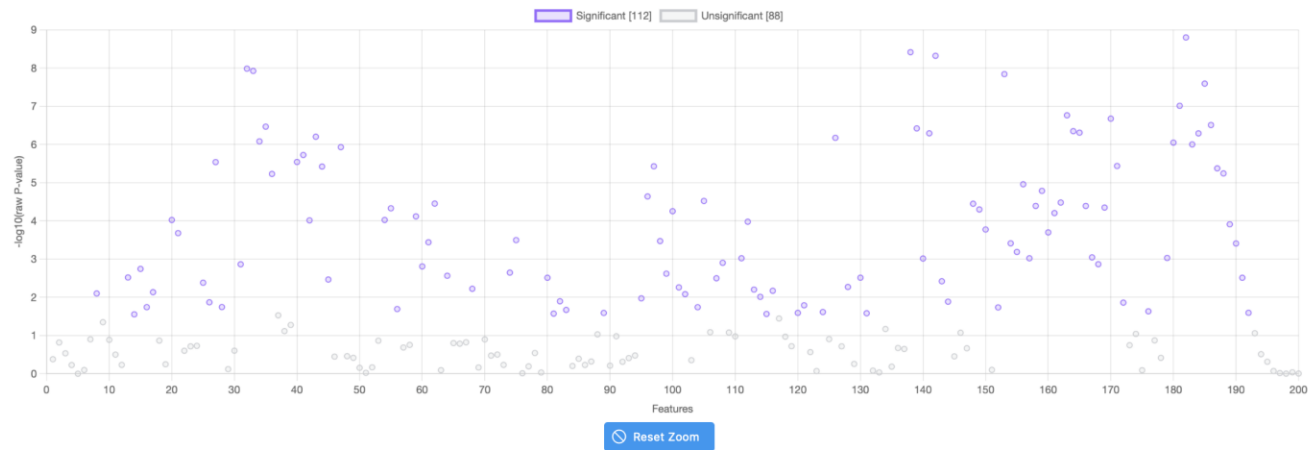
**Group variance:** Equal

**Non-parametric tests:**

**P-value threshold:** 0.05  Raw  FDR

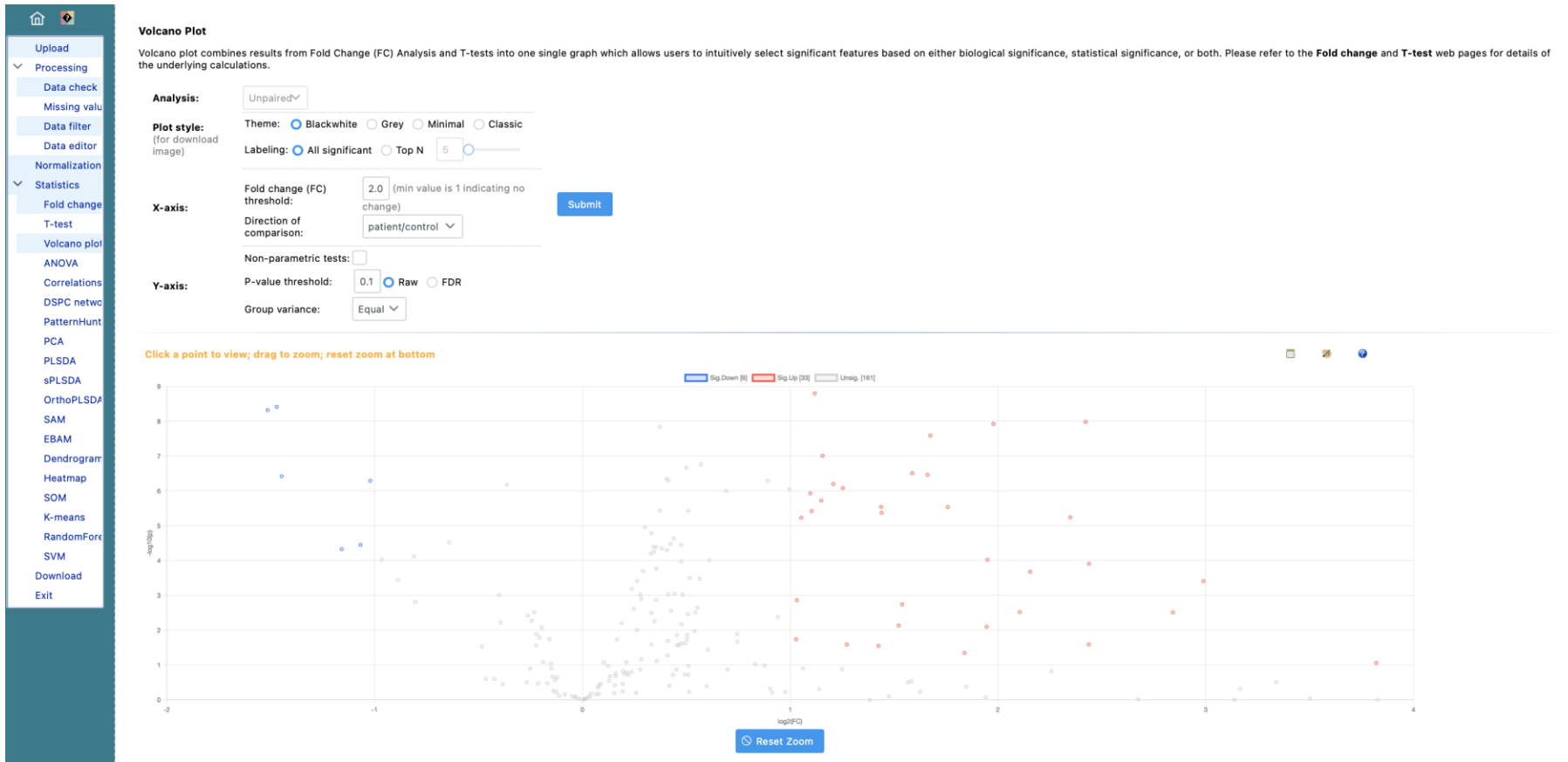
Submit

Click a point to view; drag to zoom; reset zoom at bottom



# MetaboAnalyst workflow

## 4) univariate analysis







# MetaboAnalyst workflow

## 5) chemometric analysis

- Upload
- Processing
  - Data check
  - Missing values
  - Data filter
  - Data editor
- Normalization
- Statistics
  - Fold change
  - T-test
  - Volcano plot
  - ANOVA
  - Correlations
  - DSPC network
  - PatternHunter
  - PCA
  - PLSDA
  - sPLSDA
  - OrthoPLSDA
  - SAM
  - EBAM
  - Dendrogram
  - Heatmap
  - SOM
  - K-means
  - Random Forest
  - SVM
- Download

### Partial Least Squares Discriminant Analysis (PLS-DA)

Overview **2D Scores Plot** Loadings Plot Imp. Features Synchronized 3D Plots Cross Validation Permutation

Select component for X-axis:

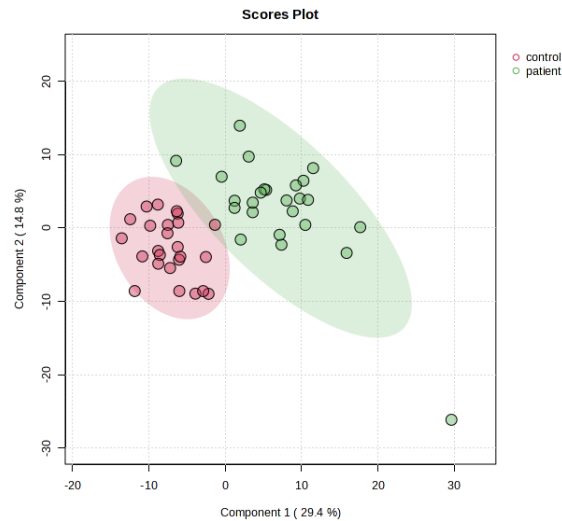
Select component for Y-axis:

Update font/label size:

Display 95% confidence region:

Display sample names:

Use grey-scale colors:



# MetaboAnalyst workflow

## 5) chemometric analysis

- Upload
- Processing
  - Data check
  - Missing value
  - Data filter
  - Data editor
- Normalization
- Statistics
  - Fold change
  - T-test
  - Volcano plot
  - ANOVA
  - Correlations
  - DSPC network
  - PatternHunter
  - PCA
  - PLSDA
  - sPLSDA
  - OrthoPLSDA
  - SAM
  - EBAM
  - Dendrogram
  - Heatmap

### Partial Least Squares Discriminant Analysis (PLS-DA)

Overview 2D Scores Plot Loadings Plot Imp. Features Synchronized 3D Plots **Cross Validation** Permutation

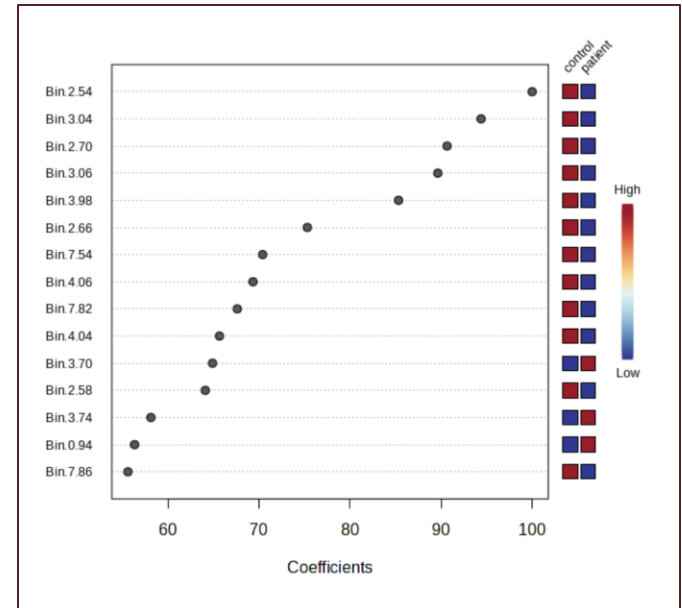
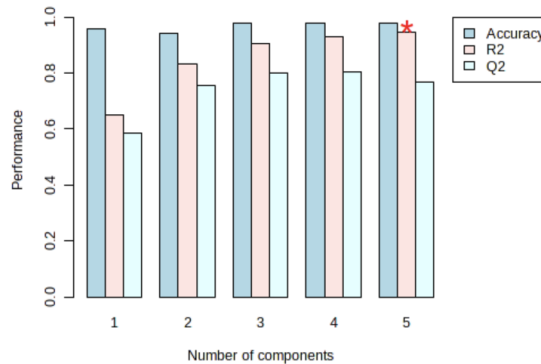
#### Select optimal number of components for classification

Maximum components to search: 5

Cross validation (CV) method: LOOCV

Update

Performance measure: R2



# MetaboAnalyst workflow

## 5) chemometric analysis



### Hierarchical Clustering Heatmaps

A heatmap provides intuitive visualization of a data table. Each colored cell on the map corresponds to a concentration value in your data table, with samples in rows and features/compounds in columns. You can use a heatmap to identify samples/features that are unusually high/low. The maximum number of features can be displayed is **2000** features (selected based on IQR by default). You can use **Select features** for better control

**Data source** Normalized data

**Standardization** Autoscale features

**Distance measure** Euclidean

**Clustering method** Ward

**Color contrast** Default

**Column option** Width: 10  Show names Font size: 8

**Row option** Height: 10  Show names Font size: 8

**Annotation bar** Height: 0.02 % Font size: 10.0

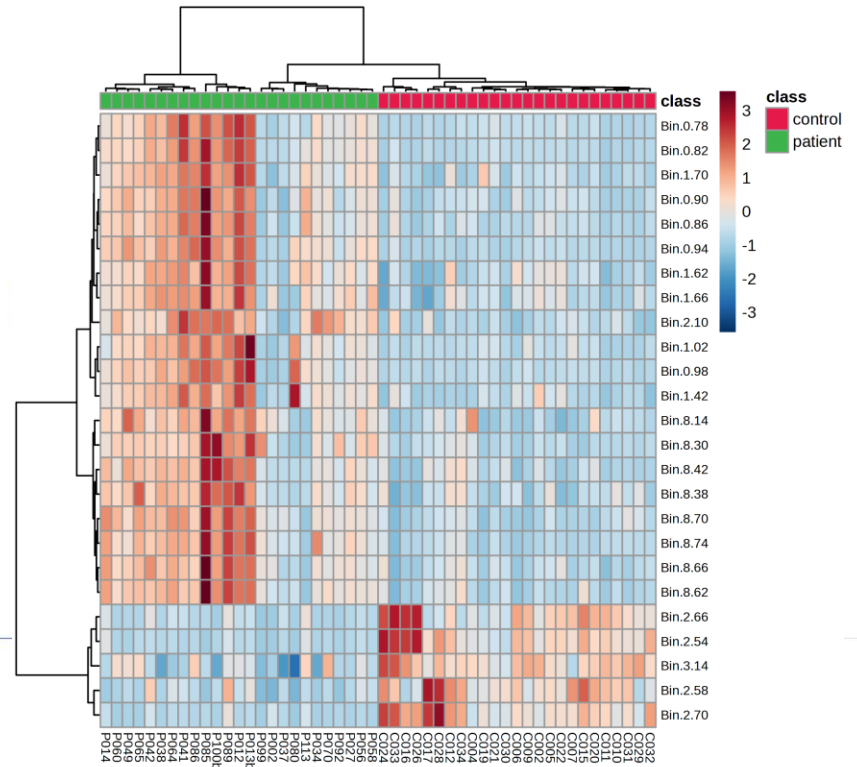
**View mode (only for download)**  Overview  Detail View (< 1000 features)

Do not cluster

Use top: 15

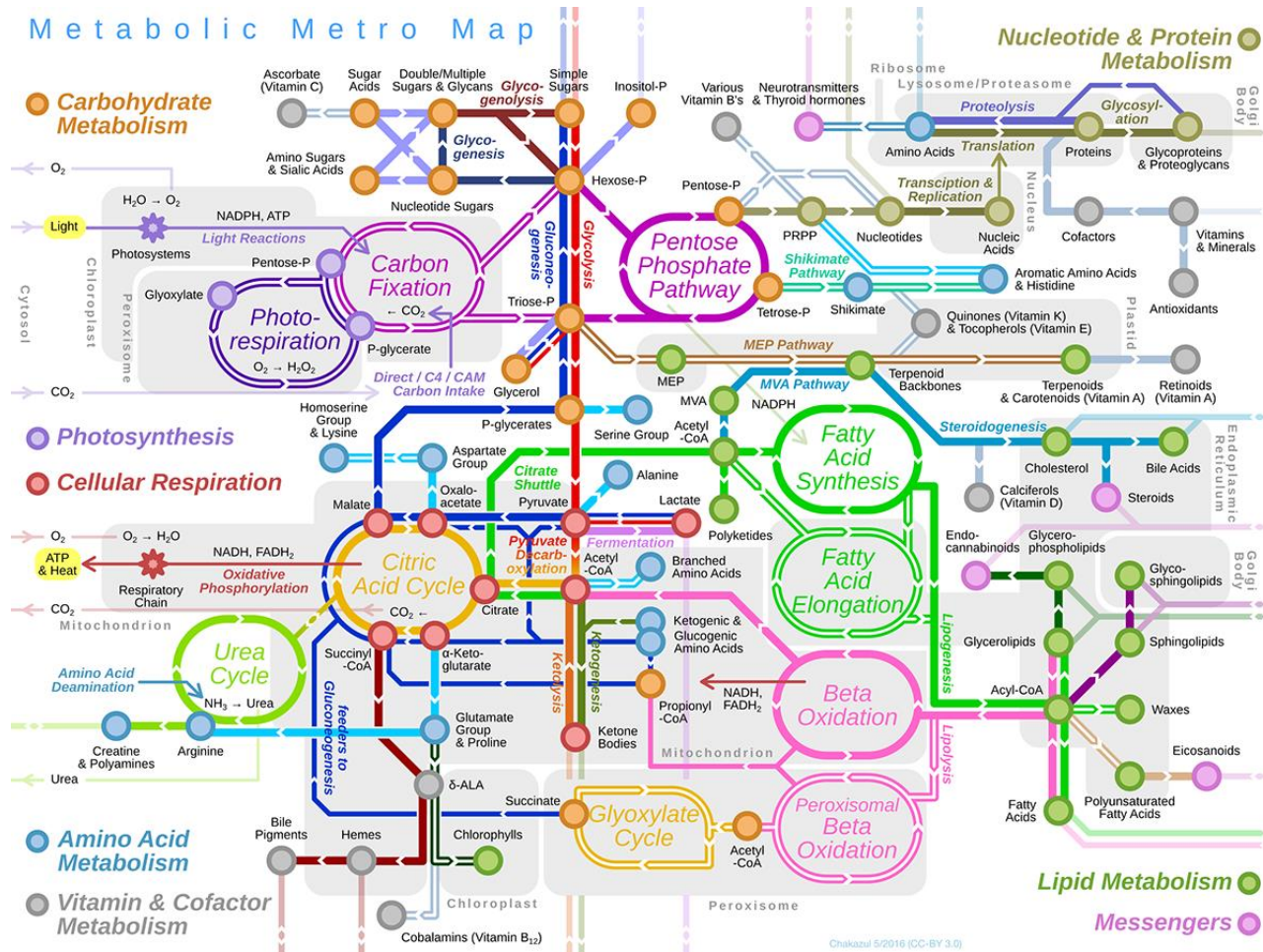
Show group annotation legend

Show only group averages



Heatmap of the top 25 T-test features

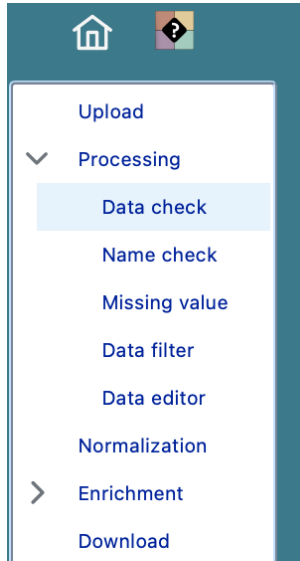
# Identifying the metabolic pathways deregulated by a pathology is finding a target for pharmacological therapy!



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 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 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.

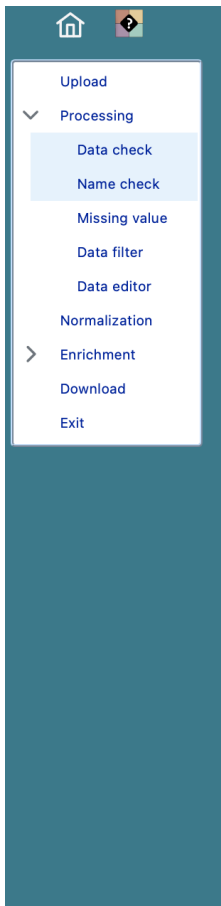
Edit Groups

Missing Values

▶ Proceed

# MetaboAnalyst workflow

## 6) enrichment analysis



Upload

Processing

Data check

**Name check**

Missing value

Data filter

Data editor

Normalization

Enrichment

Download

Exit

### Name/ID Standardization:

- For enrichment analysis, only well-annotated HMDB compounds (i.e. the tool in **Other Utilities** module;
- Greek alphabets are not recognized, they should be replaced by English
- Query names in normal white indicate exact match - marked by "1" in the
- Query names highlighted indicate **no exact or unique match** - marked
- For **compound name**, you should click the **View** link to perform **appro**
- For **KEGG ID**, it is possible to have multiple hits, you should click the **V**

| Query                      | Hit                 |
|----------------------------|---------------------|
| 1,6-Anhydro-beta-D-glucose | Levogluconan        |
| 1-Methylnicotinamide       | 1-Methylnicotinamid |
| 2-Aminobutyrate            | L-alpha-Aminobutyri |
| 2-Hydroxyisobutyrate       | 2-Hydroxyisobutyr   |
| 2-Oxoglutarate             | Oxoglutaric acid    |
| 3-Aminoisobutyrate         | 3-Aminoisobutanoic  |
| <b>3-Hydroxybutyrate</b>   |                     |
| 3-Hydroxyisovalerate       | 3-Hydroxyisovaleric |
| 3-Indoxylsulfate           | Indoxyl sulfate     |
| 4-Hydroxyphenylacetate     | p-Hydroxyphenylac   |
| Acetate                    | Acetic acid         |
| Acetone                    | Acetone             |
| Adipate                    | Adipic acid         |
| Alanine                    | Alanine             |

### Name match

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

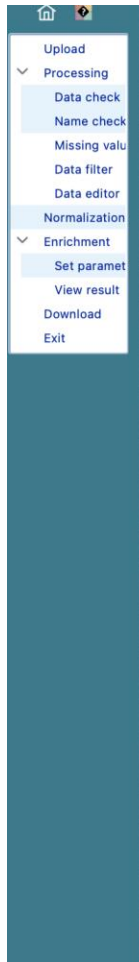
OK

Cancel



# MetaboAnalyst workflow

## 6) enrichment analysis



### Parameter Setting

**Enrichment tests** are based on the well-established [globaltest](#) to test associations between metabolite sets and the outcome. The algorithm uses a generalized linear model to compute a 'Q-stat' for each metabolite set. The Q-stat is calculated as the average of the Q values calculated for the each single metabolites; while the Q value is the squared covariance between the metabolite and the outcome. The globaltest has been shown to exhibit similar or superior performance when tested against several other popular methods.

**Metabolite sets:** Unlike transcriptomics which allows comprehensive gene expression profiling, targeted metabolomics usually covers only a small percentage of metabolome (the actual coverage is platform/protocol specific). This means that metabolites (defined in our current pathways or metabolite sets) do not have equal probabilities of being measured in your studies, and the enriched functions are the results from both platform/protocol specific effects and biological perturbations. Since the primary interest is to detect the latter, we highly recommend **uploading a reference metabolome** containing all measurable metabolites from your platform to eliminate the former effects.

#### Please select a metabolite set library

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

Only use metabolite sets containing at least

#### Please specify a reference metabolome

- Use all the compounds in the selected library  
 Upload a reference metabolome based on your analytical platform

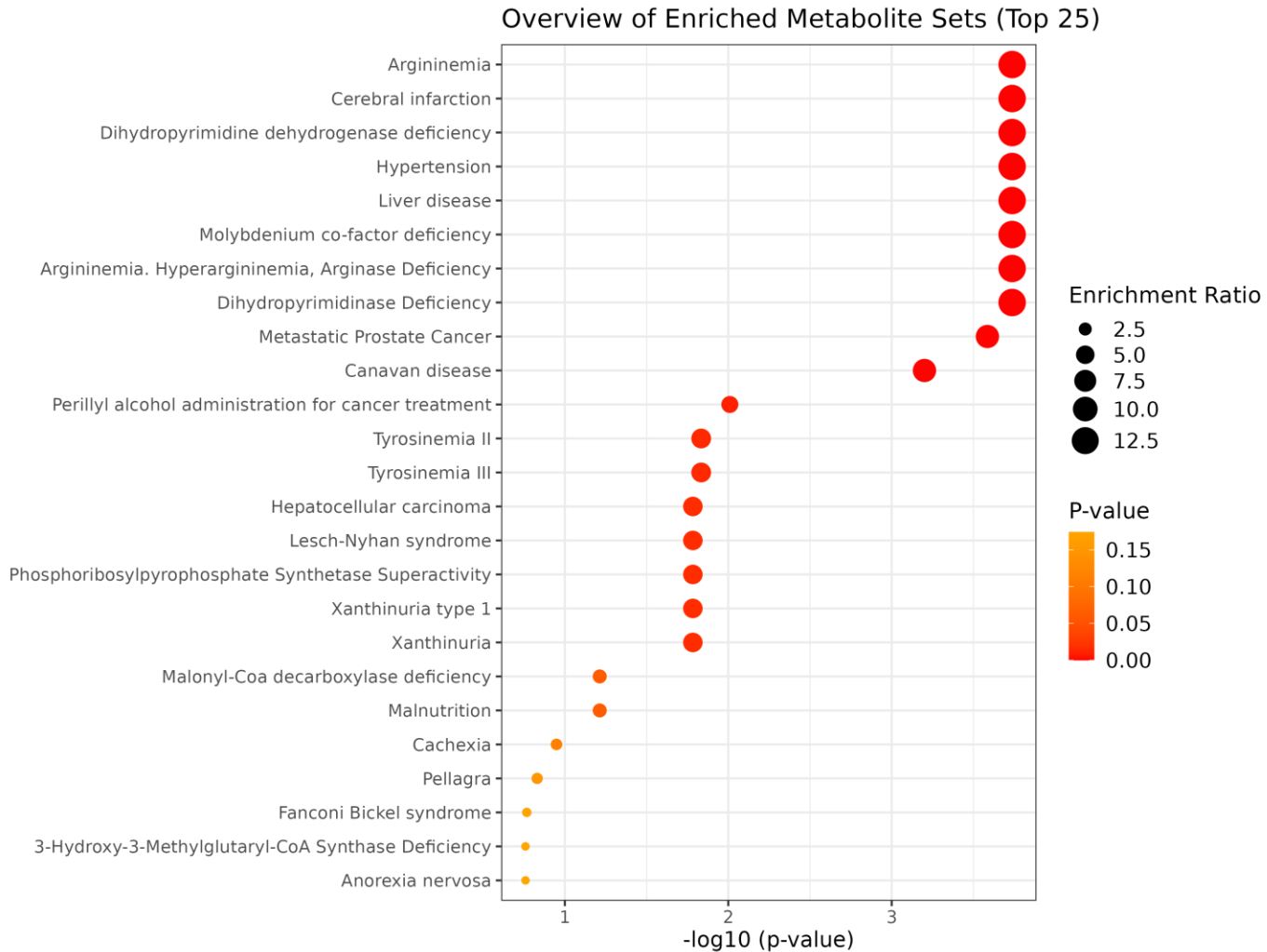
Submit

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

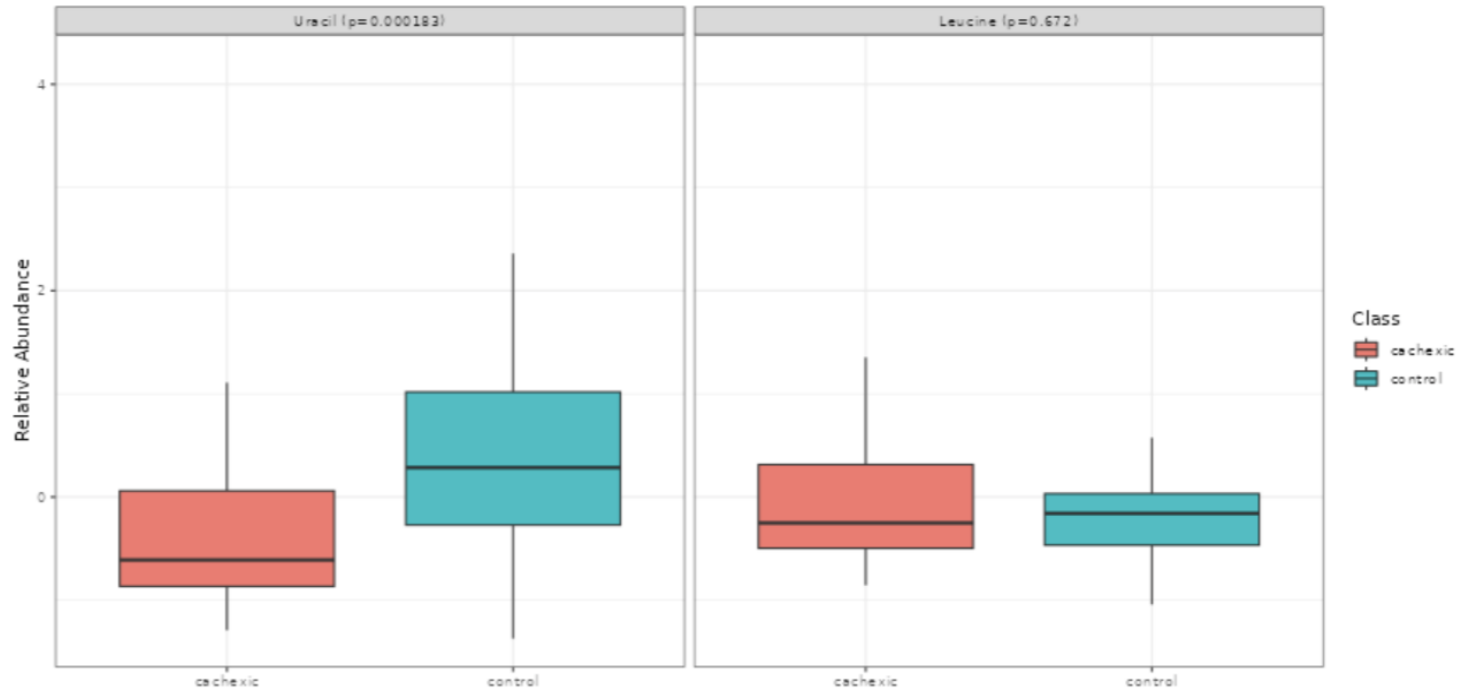




# MetaboAnalyst workflow

## 6) functional interpretation

Current metabolite set:



| Set Name                   | Metabolites  | References             |
|----------------------------|--|------------------------|
| Metastatic Prostate Cancer | Sarcosine; <b>Uracil</b> ; Kynurenine; Glycerol 3-phosphate; <b>Leucine</b> ; DL-Proline | <a href="#">PubMed</a> |

# MetaboAnalyst workflow

## Metabolic pathway analysis and visualization

**Home**

**Steps**

- Upload
- Process
- Statistics
- Enrichment
- Pathway
  - Set param.
  - View result**
- Peak search
- Metabolites
- Download
- Log out

### Result

The **metabolome view** on the left shows all matched pathways according to p values from pathway enrichment analysis and pathway impact values from pathway topology analysis. Place mouse over each pathway node will reveal its pathway name. Click each node will launch the **pathway view** on the right panel.

The pathway can be launched either by clicking the corresponding node on the left image or by clicking the pathway name from the table below. Please note, each node (compound) is clickable. You can zoom in and out using the control buttons below, and then drag the image to the locations of your interest. Place mouse over each metabolite node will reveal its common name. Click the node will trigger **compound view** of the selected compound.

### Overview of Pathway Analysis

**a**

### Alanine, aspartate and glutamate metabolism

**b**

**c**

L-Glutamine  
Importance : 0.20703  
P value : 0.03233106 [Close](#)  
DB Links :

Navigation controls: Zoom in, Zoom out, Fit, Up, Down, Left, Right.

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

$H_0$

- $H_0$  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  $=$ ,  $\leq$ , or  $\geq$
- Test the Null Hypothesis directly: reject  $H_0$  or fail to reject  $H_0$

## Alternative Hypothesis

$H_1$  or  $H_a$

- $H_1$  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  $H_0$  is false (corresponding to  $=$ ,  $\leq$ , or  $\geq$  conditions)
- 'opposite' of Null Hypothesis

# Decision Errors

Two types of errors can result from a hypothesis test.

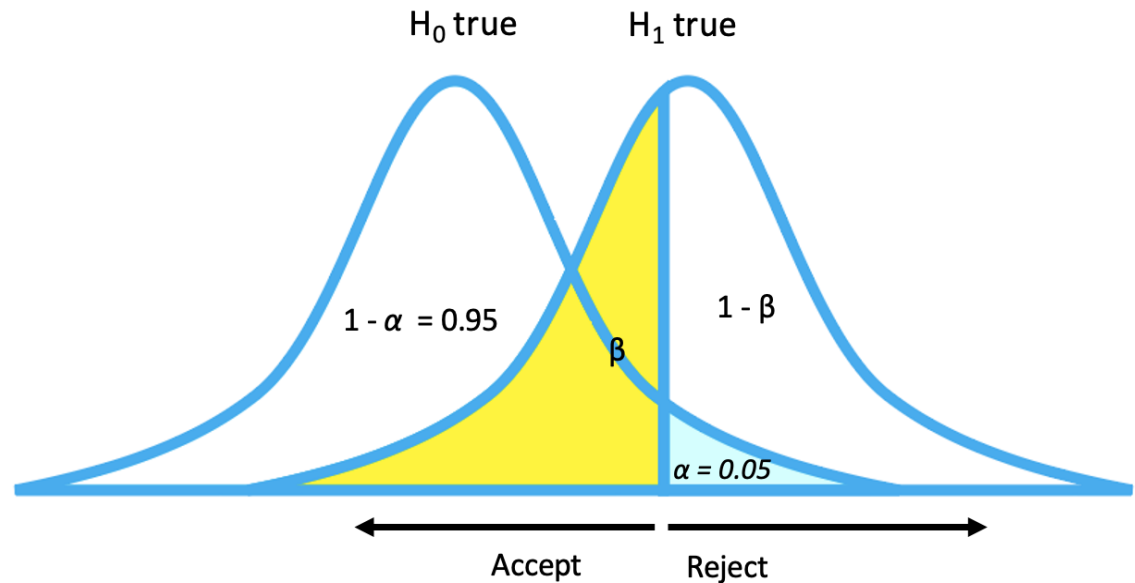
- **Type I** error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called alpha, and is often denoted by  $\alpha$ .
- **Type II** error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by  $\beta$ . 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                          | <b>Type I error<br/>FALSE POSITIVE</b> |
| <i>probability</i> | $1-\alpha$                              | $\alpha$                               |
| H1 is true         | <b>Type II error<br/>FALSE NEGATIVE</b> | Correct action                         |
| <i>probability</i> | $\beta$                                 | <i>power</i> = $1-\beta$               |

$$\alpha = P(H1 | H0)$$

$$\beta = P(H0 | H1)$$



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

|                    | Fail to reject H0     | Reject H0                           |
|--------------------|-----------------------|-------------------------------------|
| H0 is true         | <b>TRUE NEGATIVE</b>  | <b>FALSE POSITIVE</b>               |
| <i>probability</i> | $1-\alpha$            | $\alpha$                            |
| H1 is true         | <b>FALSE NEGATIVE</b> | <b>TRUE POSITIVE</b>                |
| <i>probability</i> | $\beta$               | <i>power = 1-<math>\beta</math></i> |

Example 1. Covid-19 test:

- False you
- your
- False the
- You

Example 2. Quality control in a pharma production company

- False-POSITIVE: The test declared a product faulty. We

Example 3. Disease diagnosis

- False
- False
- False

Example 3. Criminal court

- False disea
- False wron
- False prov
- 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

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

# Controlling Type I and Type II error

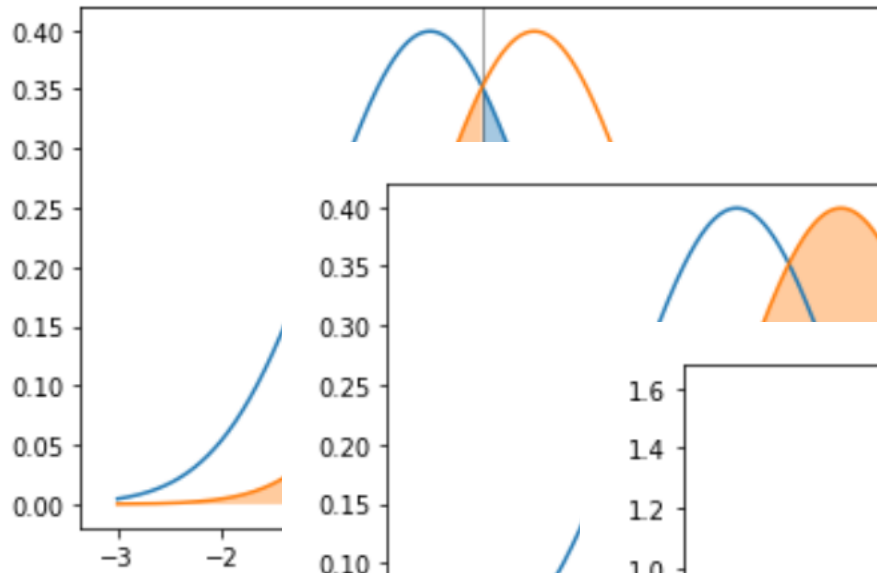


Figure 1: Equal  $\sigma$  false negatives.

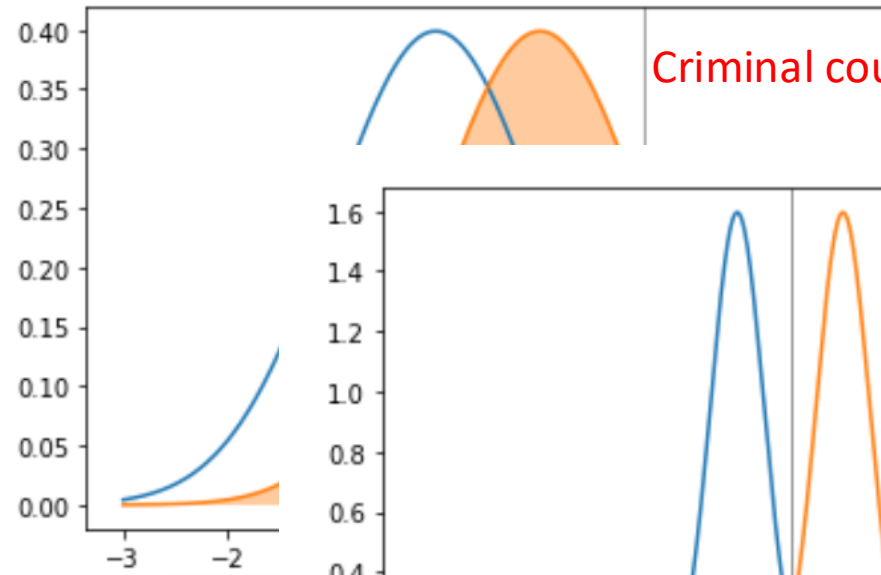


Figure 2: Greater  $\sigma$  false negatives

Criminal court example

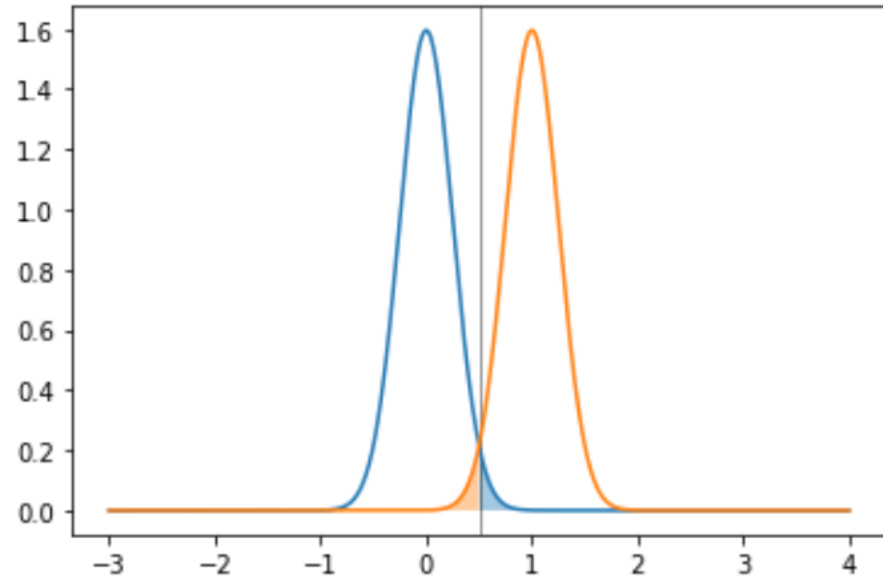
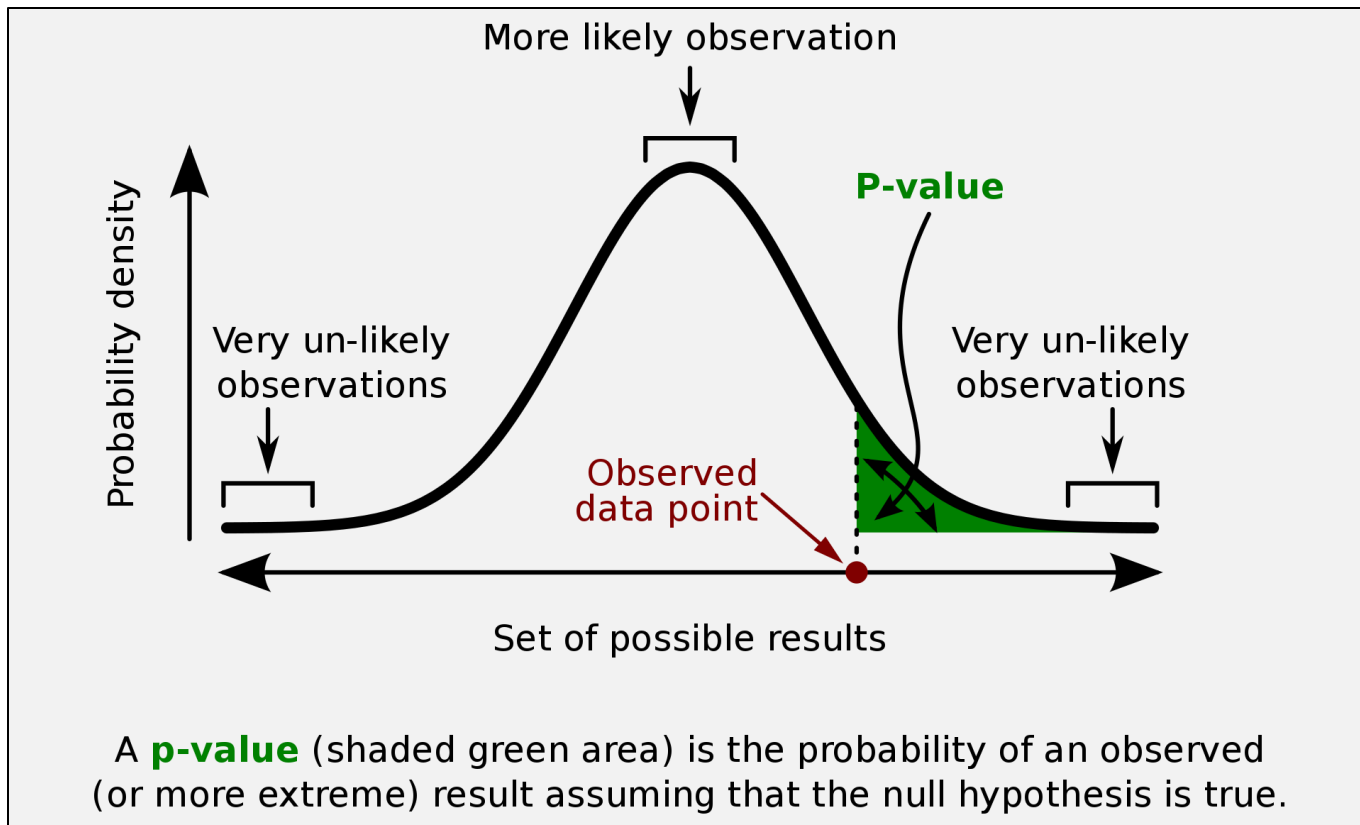


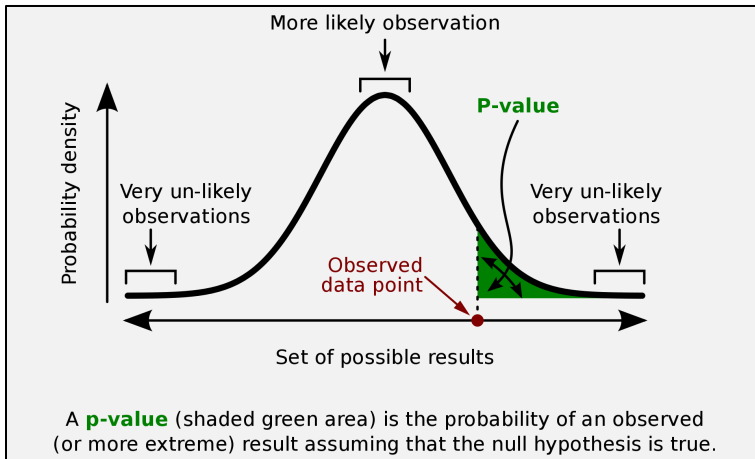
Figure 3: Lowered uncertainty through more informative features.

# 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  $H_0$  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  $\alpha$  before experiment and then compare the p-value and takes a decision.

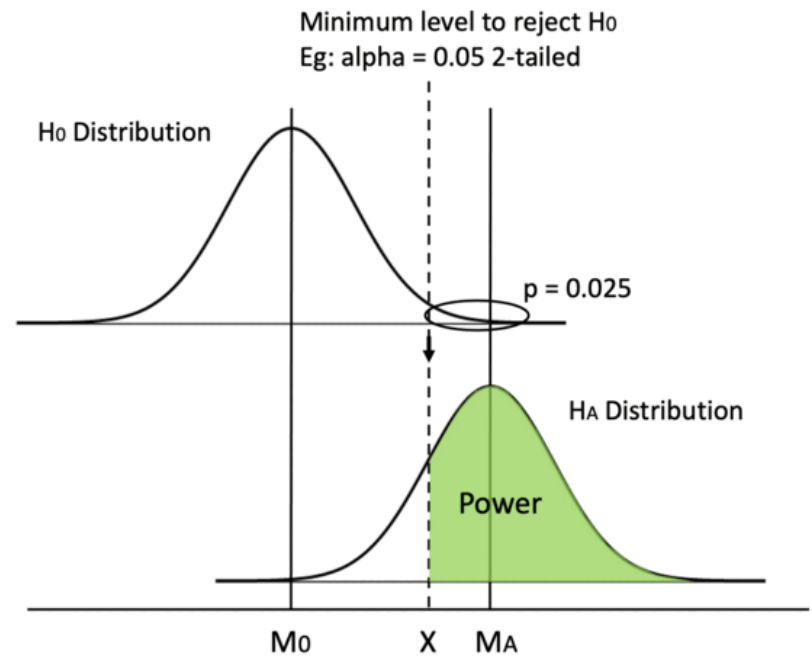
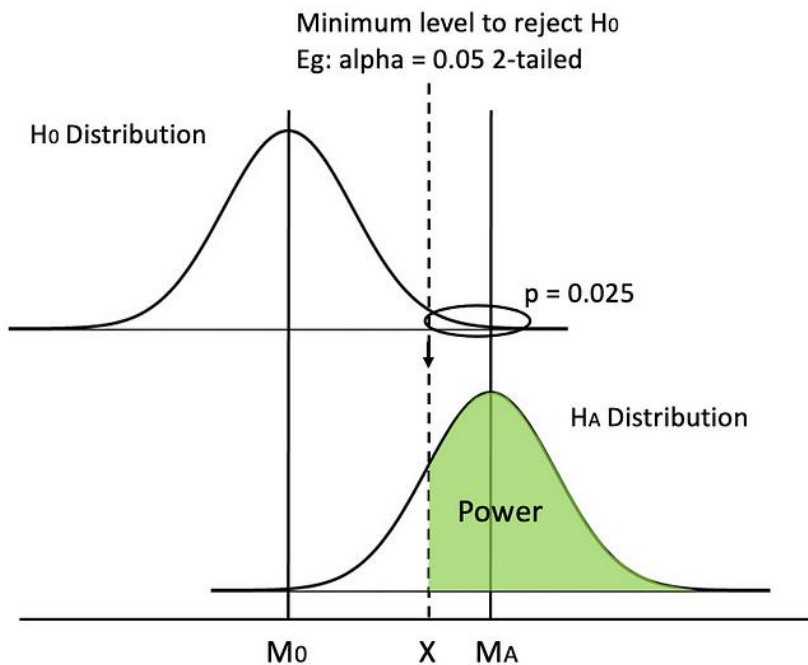
## Conventions

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



# How to increase statistical power

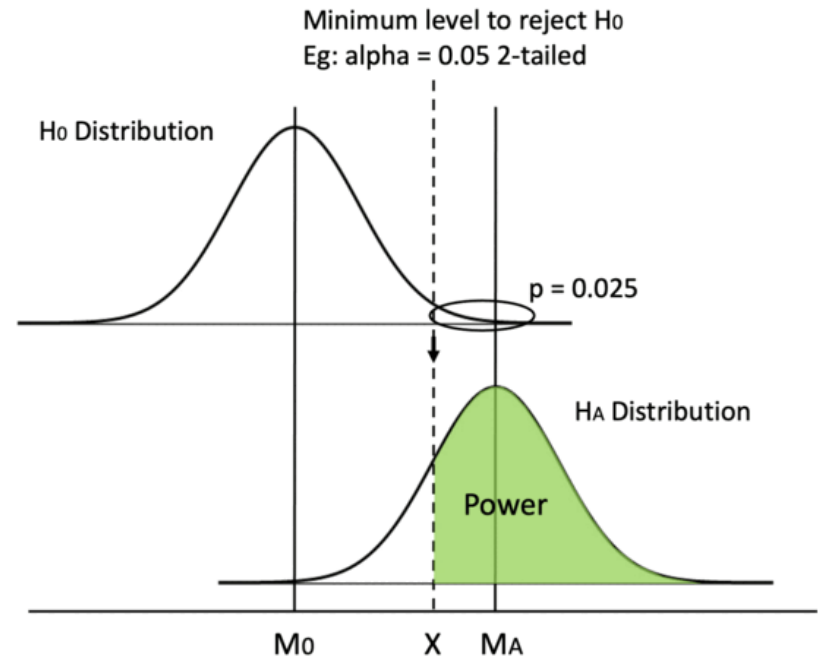
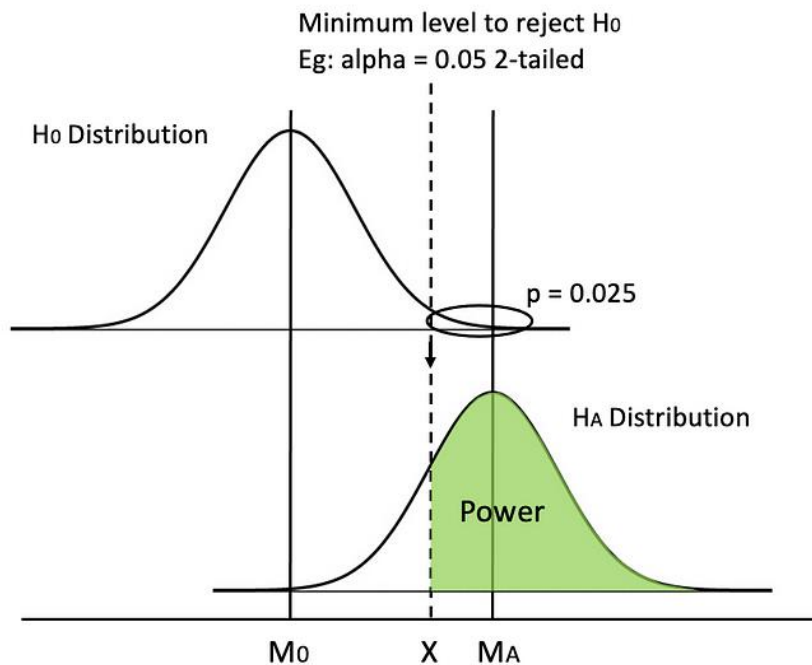
## 1) Raise significance level alpha (the **WRONG** way)



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

# How to increase statistical power

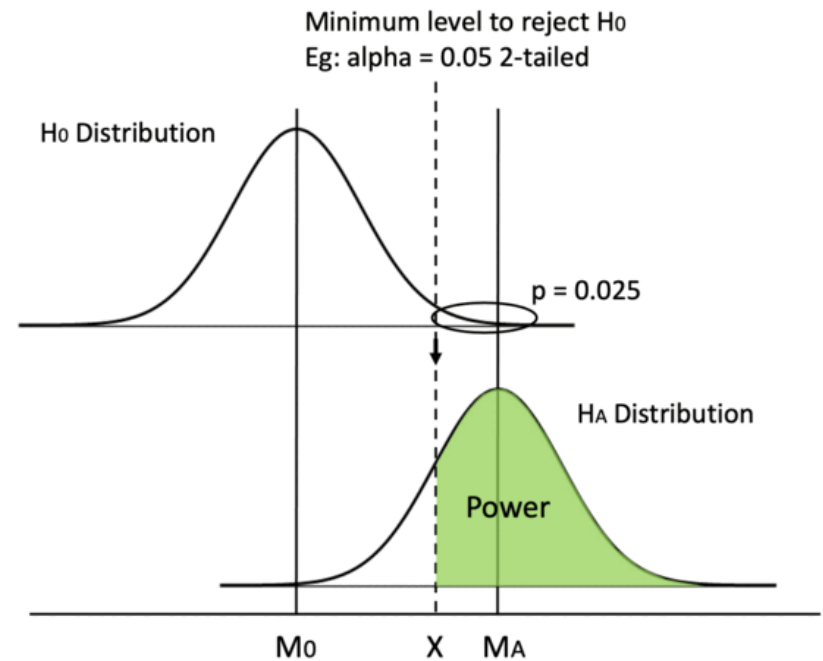
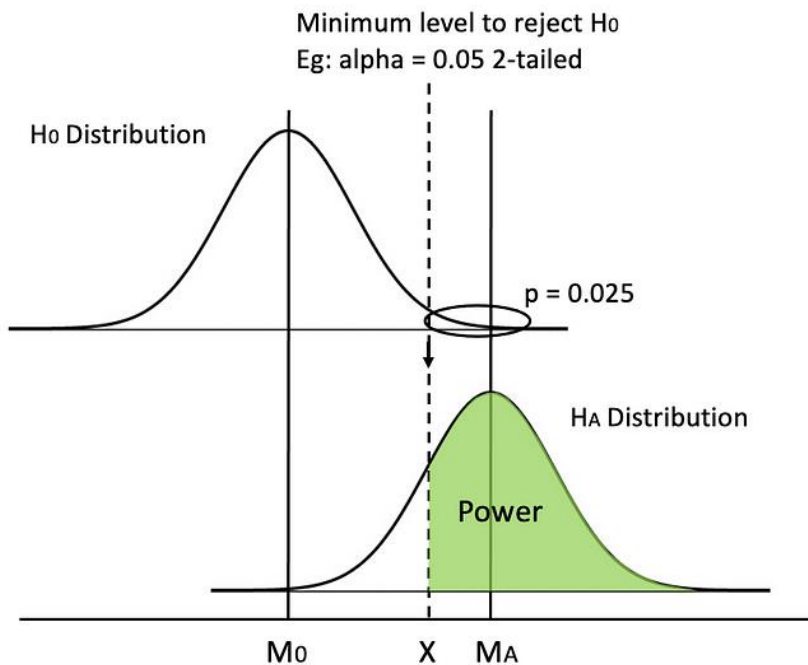
2) Switch from a 2-tailed test to a 1-tailed test (CORRECT if possible)



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)

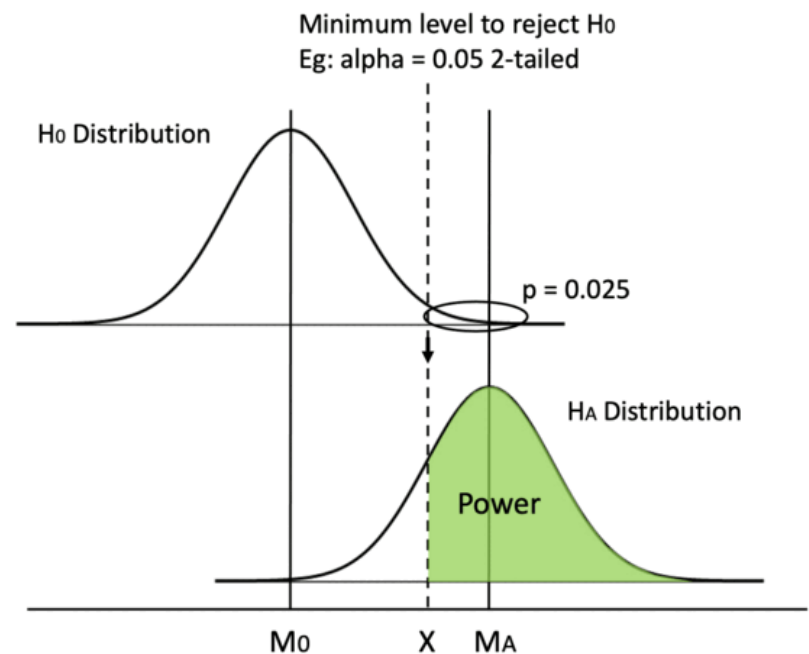
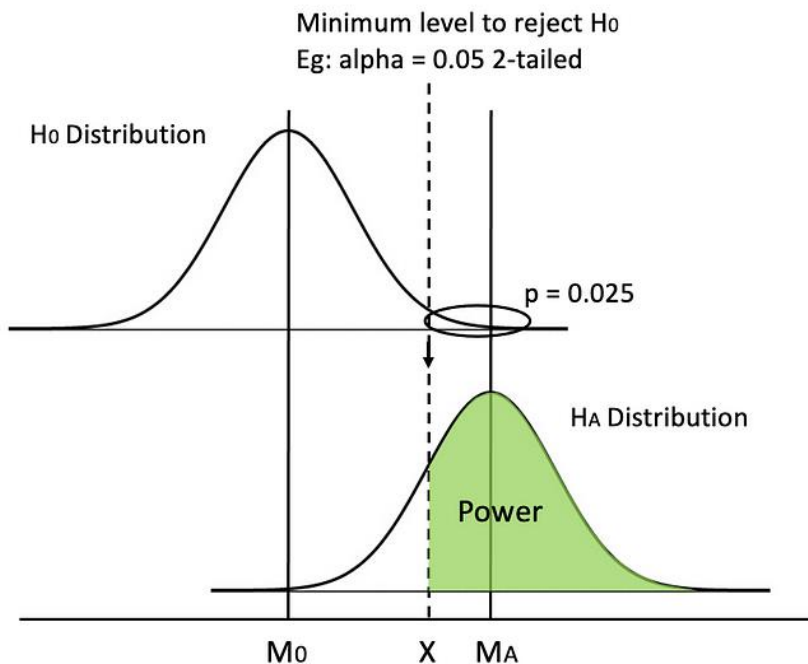


Source: <https://towardsdatascience.com/5-ways-to-increase-statistical-power-377c00dd0214>



# How to increase statistical power

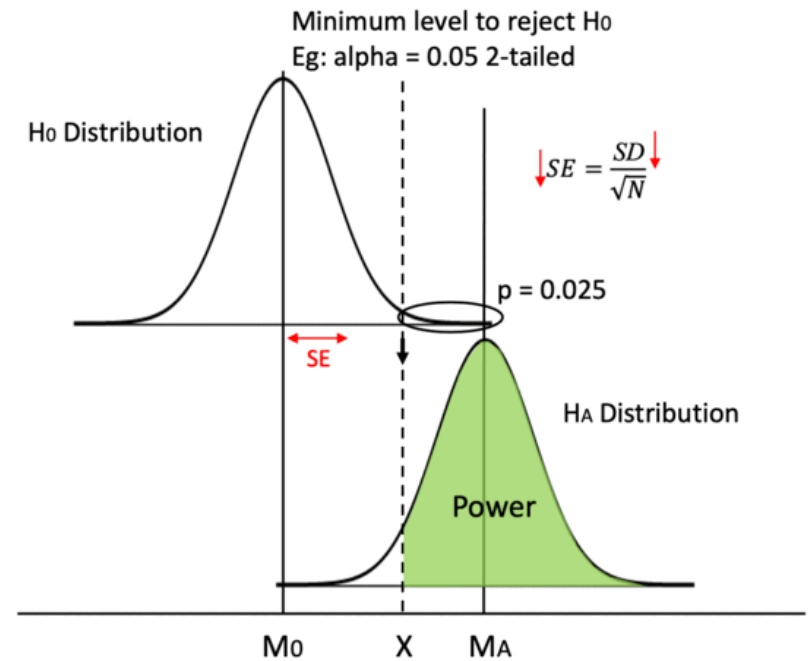
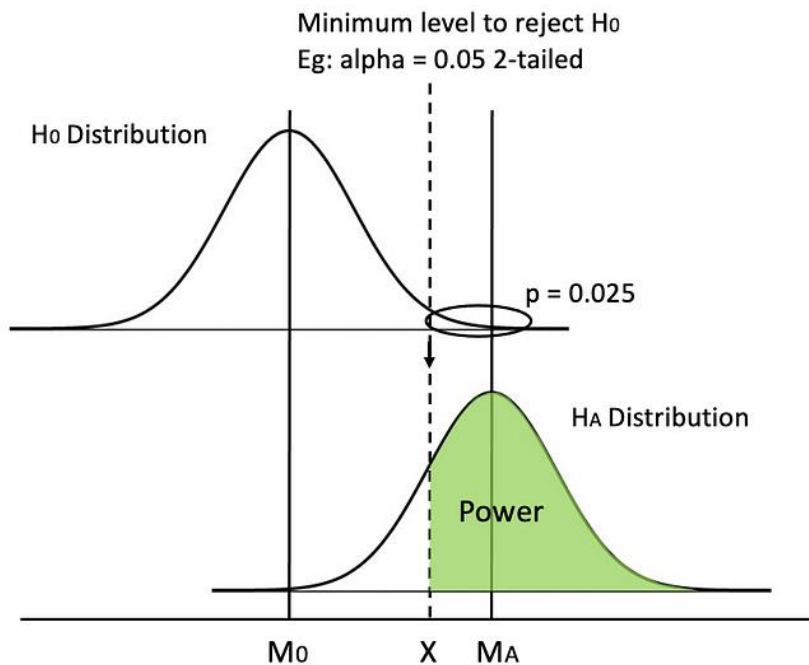
4) Use z distribution instead of t distribution (appropriate when we know the population mean)



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

# How to increase statistical power

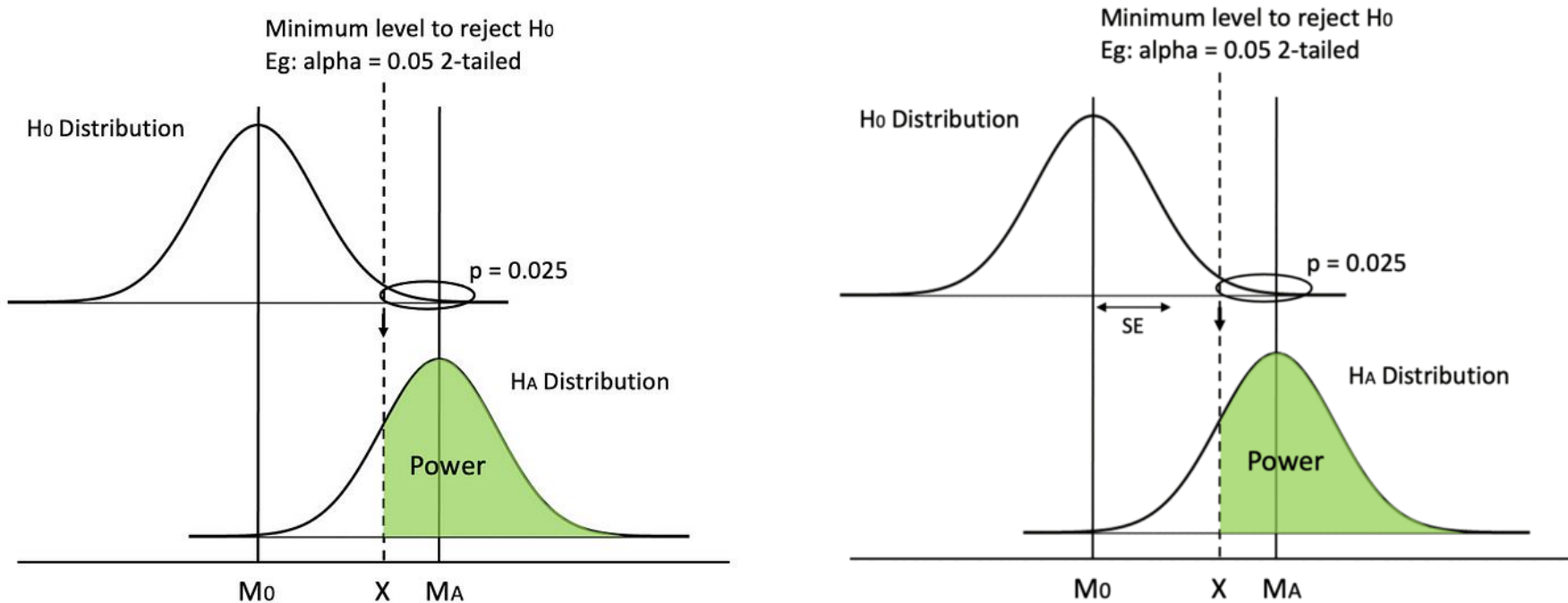
5) Decrease standard deviation (using more precise measurements to have less error and less noise)



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

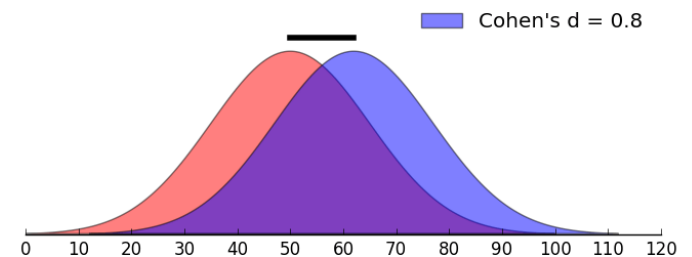
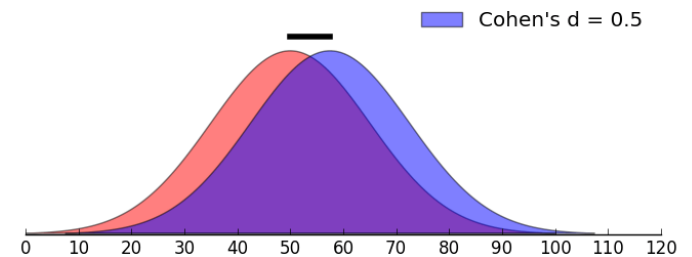
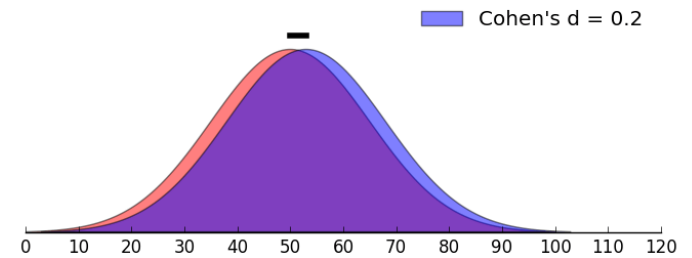
$$Cohen's\ d = \frac{\mu_1 - \mu_2}{\sigma}$$

Mean value of the population 1 ( $\mu_1$ )  
 Mean value of the population 2 ( $\mu_2$ )  
 Standard deviation of the population ( $\sigma$ )

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

Mean value of sample 1 ( $\bar{x}_1$ )  
 Mean value of sample 2 ( $\bar{x}_2$ )  
 Estimated standard deviation of the population from the sample ( $\hat{\sigma}$ )

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



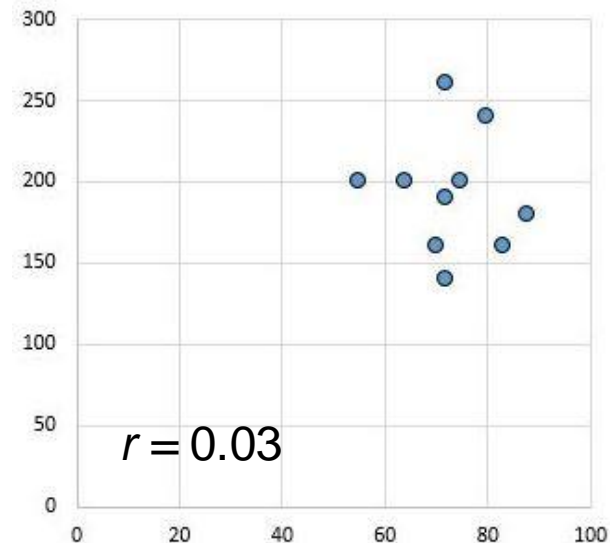
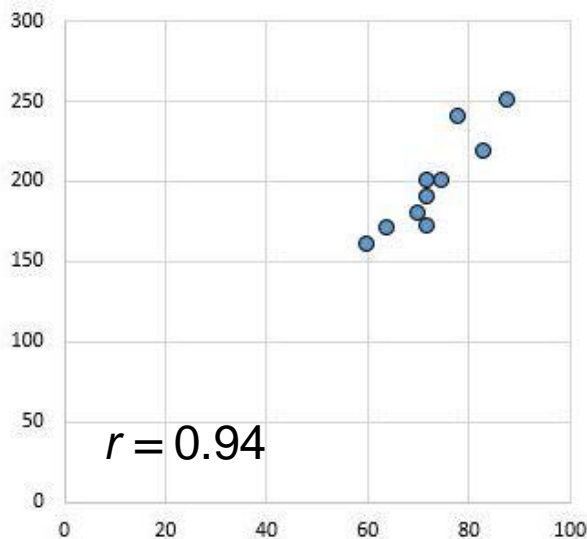
# Effect size

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: <ul style="list-style-type: none"> <li>one-sample t-test</li> <li>independent samples t-test</li> <li>paired samples t-test</li> </ul> | Cohen's $d$<br>$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$                                  | 0.20  | 0.50   | 0.80  |
| Difference between many means (ANOVA)   | Cohen's $f$<br>$f = \sqrt{\frac{\eta^2}{1-\eta^2}}$                                   | 0.10  | 0.25   | 0.40  |
| Chi-squared test  | Cohen's $\omega$<br>$\omega = \sqrt{\sum_{i=1}^m \frac{(p_{1i} - p_{0i})^2}{p_{0i}}}$ | 0.10  | 0.30   | 0.50  |
| Pearson's correlation coefficient   | Pearson's $r$   | 0.10  | 0.30   | 0.50  |
| Linear Regression (entire model)  | Cohen's $f^2$<br>$f^2 = \frac{R^2}{1 - R^2}$  | 0.02  | 0.15   | 0.35  |

Source: [https://en.wikipedia.org/wiki/Effect\\_size#Overview](https://en.wikipedia.org/wiki/Effect_size#Overview)