Data Exploration - Introduction

Pietro Franceschi

FEM - UBC

11/02/2020

Section 1

Introduction

Why

Because nowadays almost all biological/natural systems are characterized in high throughput omics experiments

Biosystem Data Analysis

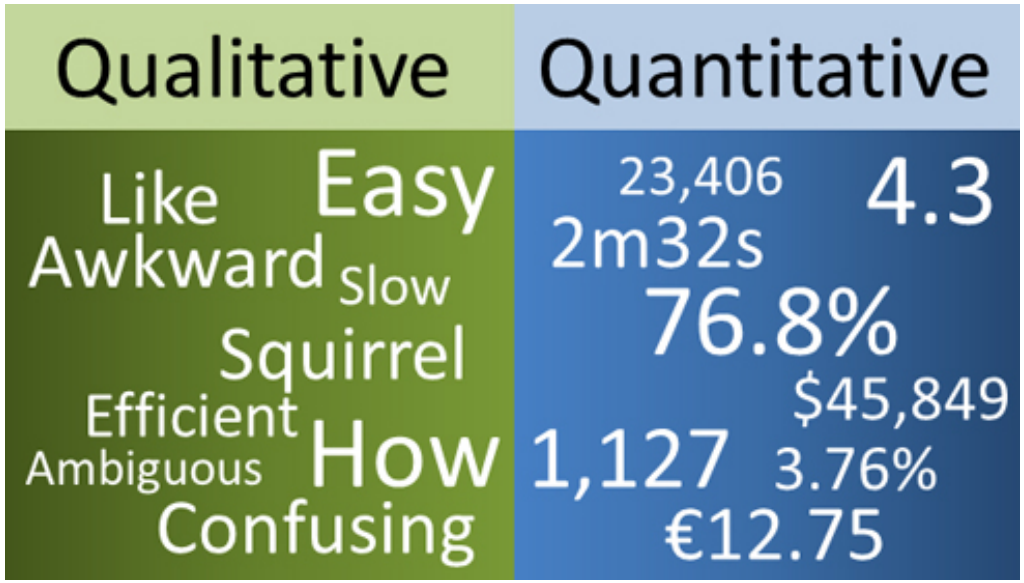
Develop and validate methods for organizing, summarizing and visualizing complex biological data through the integration of bioinformatics and biostatistics

Always more data

- Genomics
- Metabolomics
- Proteomics
- High throughput phenotyping
- Distributed network of sensors

Qualitative and quantitative

Technological advancement (experimental techniques, IT) is transforming the "soft sciences" in quantitative disciplines

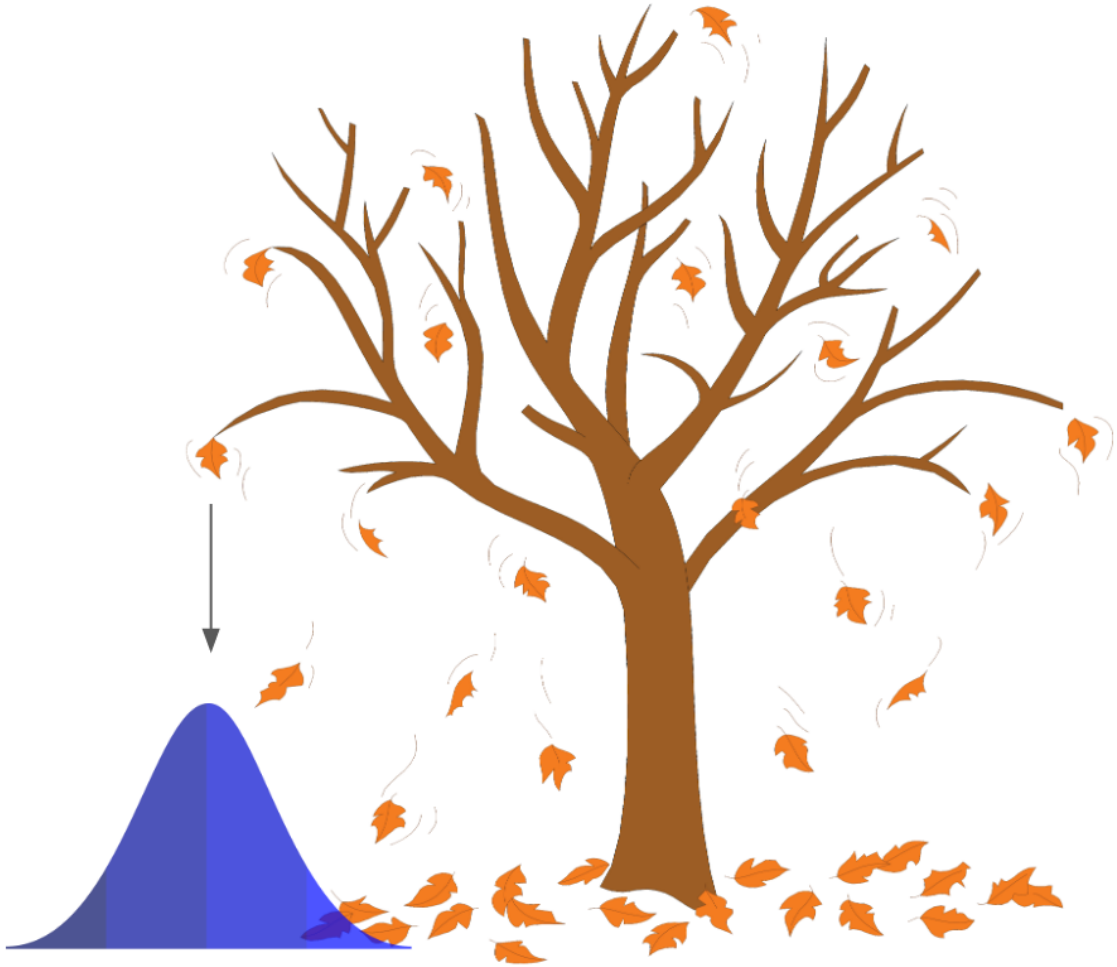


Data, knowledge, data analysis

bioinformatics, statistics, chemometrics, provide the tools to

- Promote the incremental progress of science
- Guarantee the validity and the correctness of the results
- Facilitate the production of "scientific" results (of general validity . . .)
- Be consistent
- Get the maximum from complex data

Why we need statistics . . .



Some Ideas

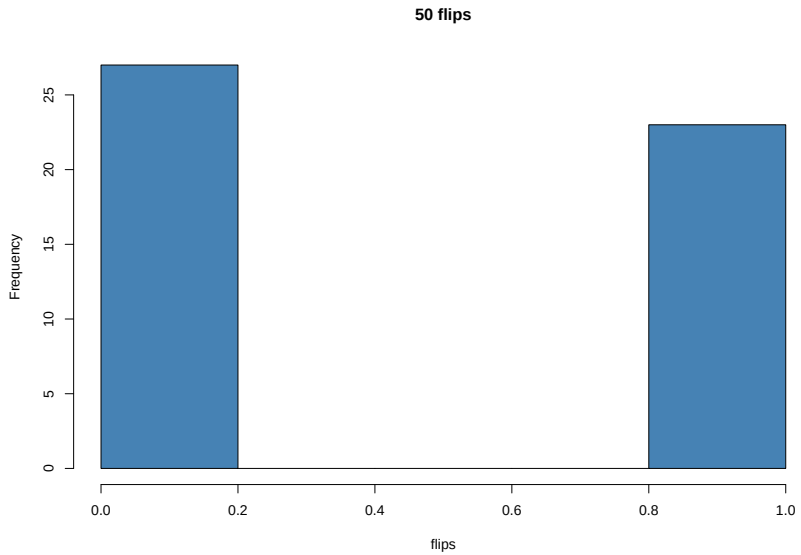
- The leaf always fall
- It is not at all easy to say where
- If it would be an apple the prediction would be easier
- Every leaf is different . . . so measuring once is not sufficient
- I will end up with a probability
- The distribution can be **very** narrow

Variability

- What we observe is the result of a "chain" of processes (e.g. gene -> protein -> metabolite)
- We never observe only one chain (e.g we consider *many* people with similar metabolism)
- The fact that we have noise on the "chain" produces variability in the output
- This variability has a bell shaped profile, which is more often than not **Gaussian**

Coin toss

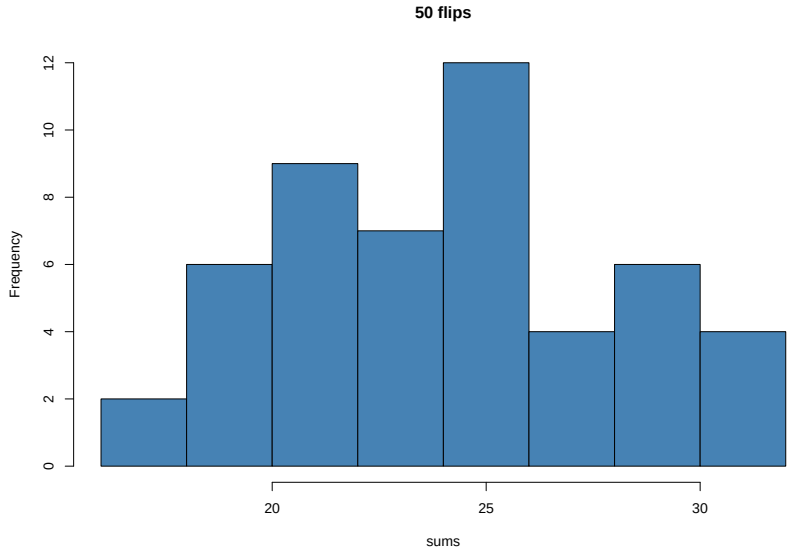
The distribution of the outcomes of 50 tosses of the same coin



This "biological" process is clearly non normally distributed, but has variability

Sum of 50 coin tosses

Suppose that now my "biological" process is the result of the sum of 50 coin tosses . . .



Here the "biological" process yields normally distributed data!

The empirical low of chance

The Law

In a sequence of experiments performed on the same conditions the relative frequency of a phenomenon gets closer to the probability of the phenomenon itself . . . and the goodness of this approximation improves as the number of experiments increases.

Data Deluge

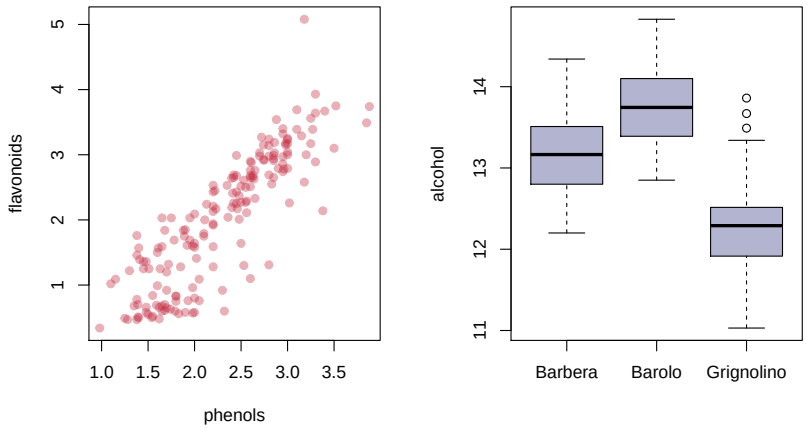
Measuring more does not necessarily mean understanding more

Nate Silver: *The Signal and the Noise: Why So Many Predictions Fail, but Some Don't*

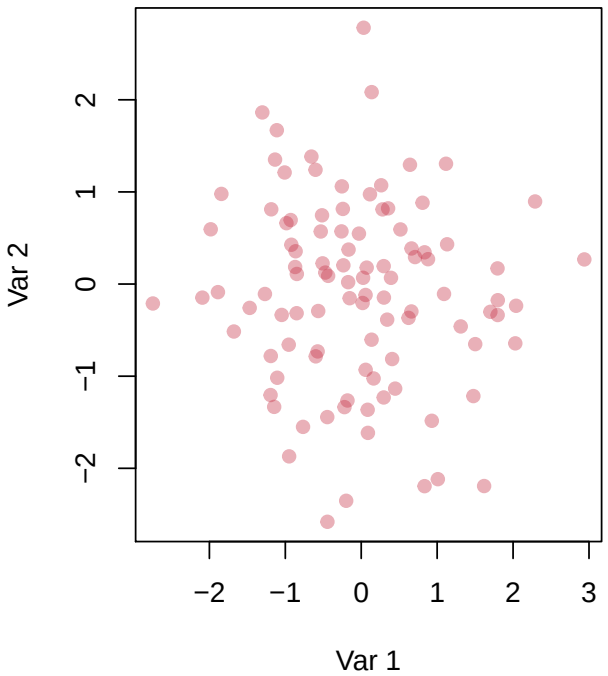
Data Analysis in a Nutshell

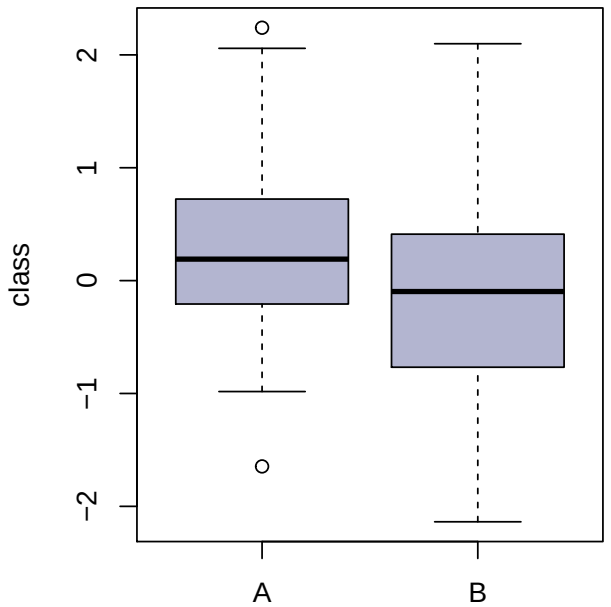
Highlight the presence of **organization** inside complex datasets, trying to measure with which ***confidence*** one can say that this organization is true at the population level

Is what I'm observing ***true beyond my sample***?

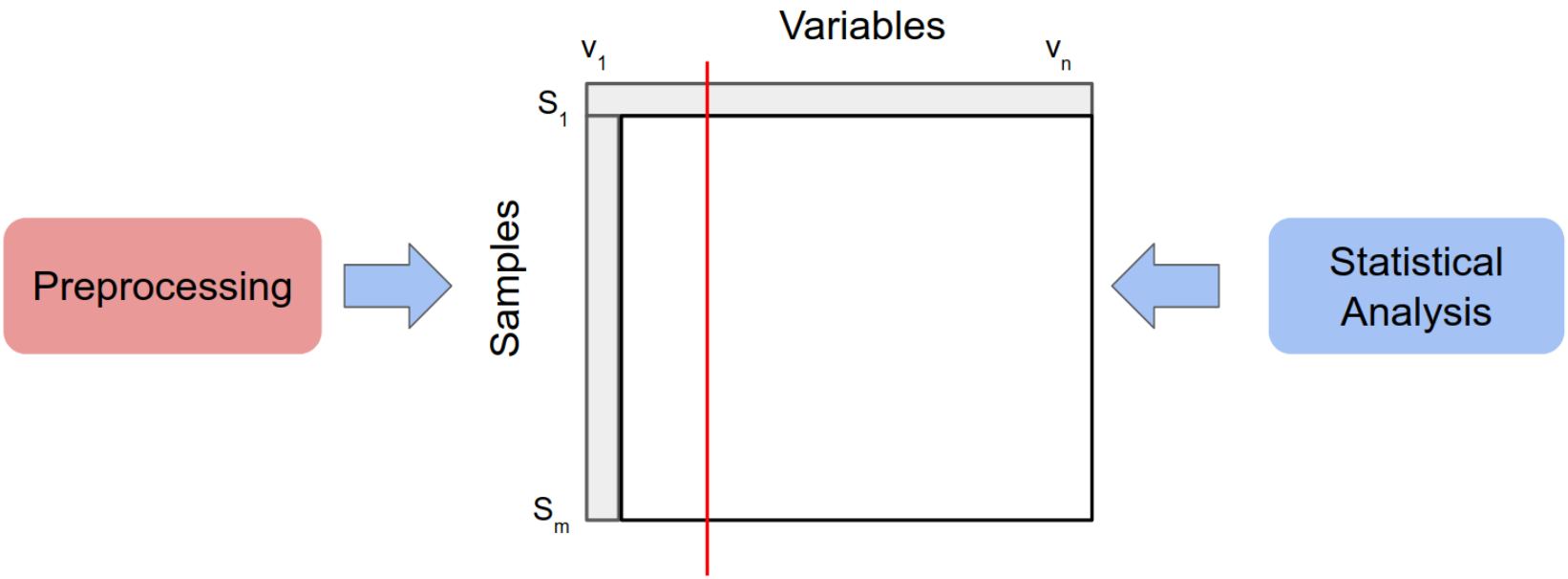


No organization ...





Data Matrix


Variables
v_1
v_n
s_1
s_m
Samples
Preprocessing
Statistical Analysis

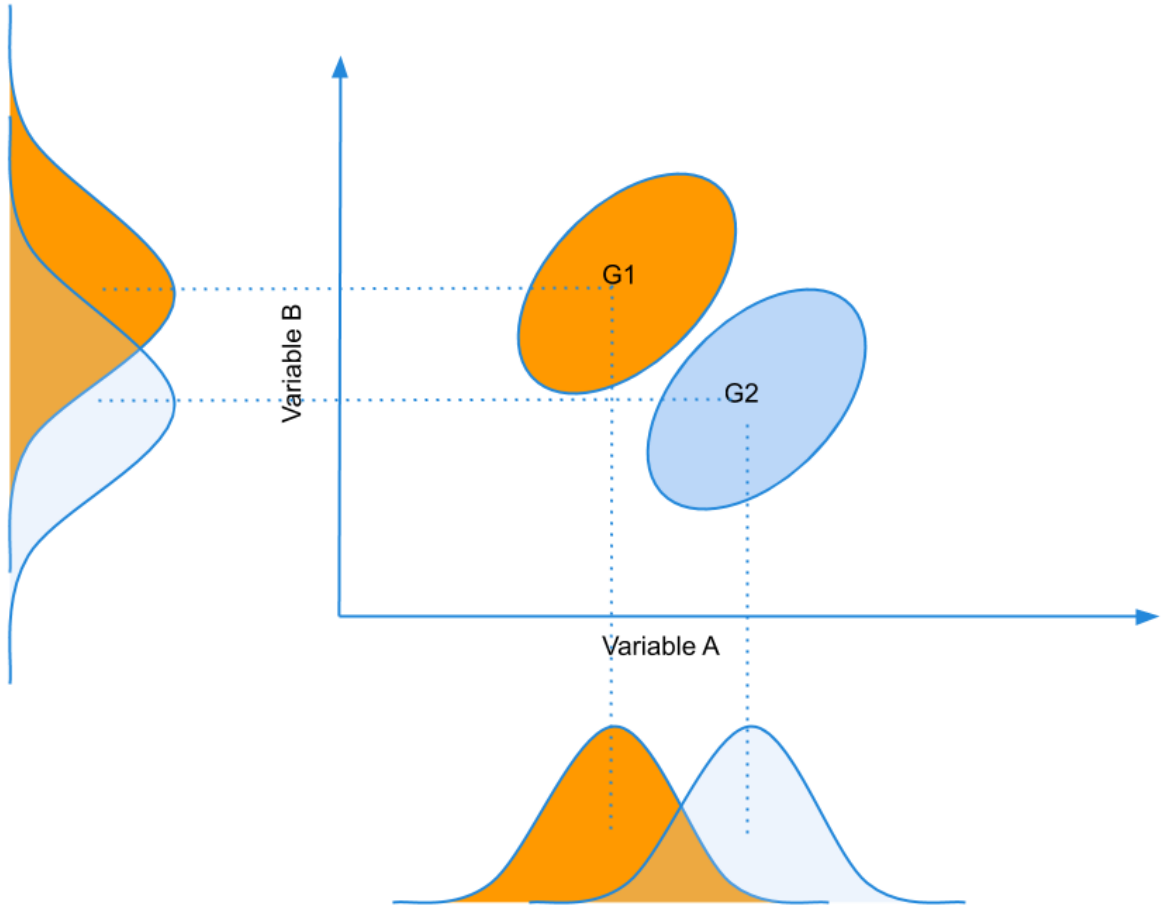
For the audience

- What are the preprocessing steps if you have to analyze a gene expression experiment?
- What are the variables in a targeted metabolomics experiment?
- What are the variables in a next generation sequencing experiment?
- What are the variables in a metagenomic investigation?

Multivariate vs Univariate

- **Univariate Approach**: each experimental variable is analyzed/visualized autonomously
- **Multivariate Approach**: Each sample (observation) is a point in the n dimensional space of the variables. E.g If we measure three properties the space is three dimensional.

Why multivariate . . .



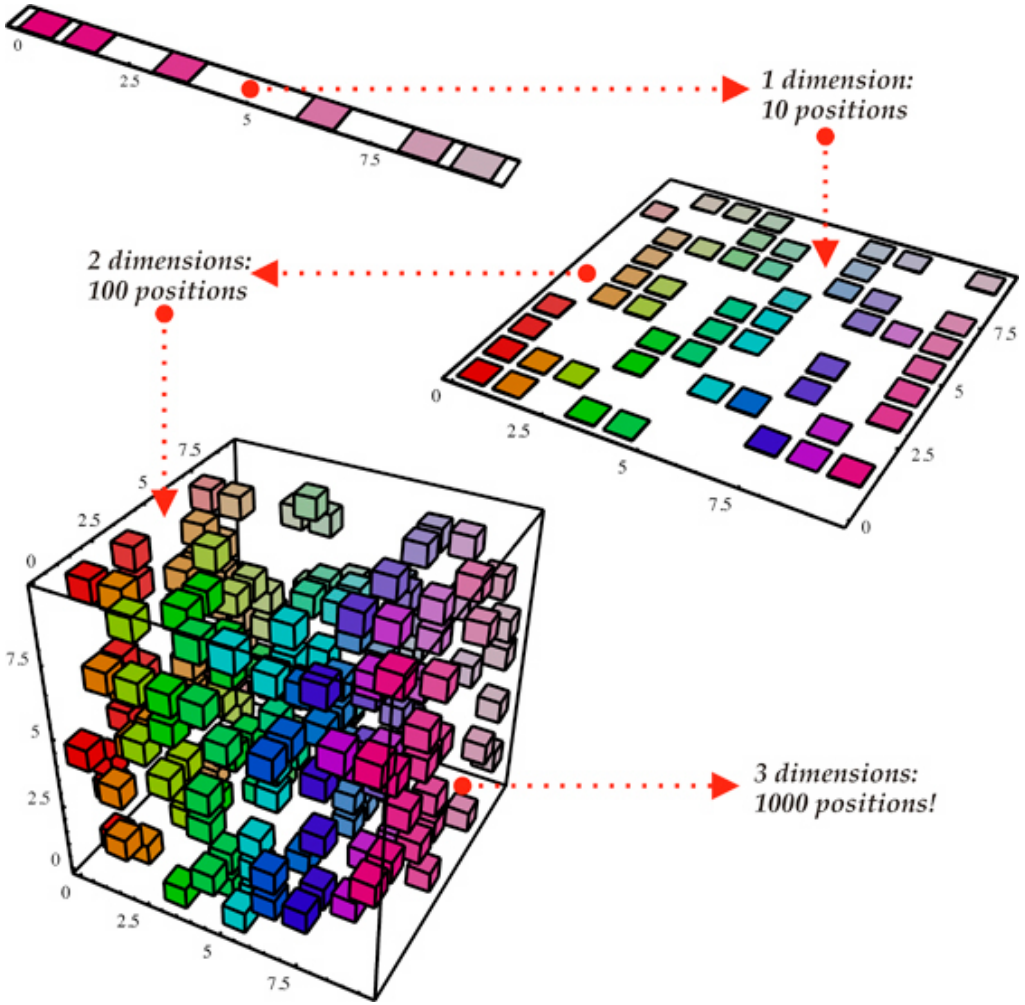
How Big is the space ?

- Untargeted metabolomics $\sim$ 1000/10000 variables/dimensions
- Targeted proteomics $\sim$ 1000/3000 variables/dimensions
- Targeted metabolomics $\sim$ 100/200 variables/dim
- NGS, Metagenomics ...

How many samples

Most experiments are performed on 10-100 samples. This is the number of points in the multidimensional space

The course of dimensionality

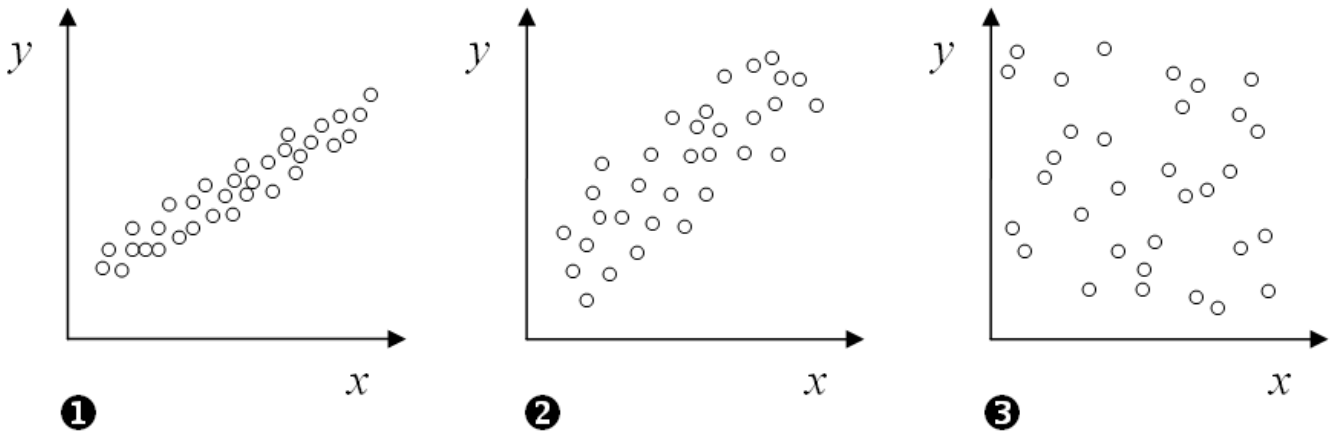


Variable Dependence

Since the variable we measure are **not independent** (e.g. network of genes, associated proteins, ecc . . .).

The effective size of the space occupied by the samples is smaller than the number of variables.

This is equivalent to say that the cloud of samples can only populate a limited part of the available multivariate space



In presence of dependence between the variables (**1** and **2**), the samples are occupying a smaller part of the available space.

Type of variable dependence

- "Analytical"/chemical
- Biological
- ...

Let's look for the source of dependence in your research areas ...

Dependence has the following positive implications

- multivariate approaches are potentially more "effective" since they use explicitly variable correlation
- we can make reasonably **good science** even from experiments with **relatively small number of samples**

- With small number of samples we'll always find **spurious organization**.
- This will always lead to FALSE DISCOVERIES


LIVE
I am
here
now
being
becoming
sensing

False Discoveries

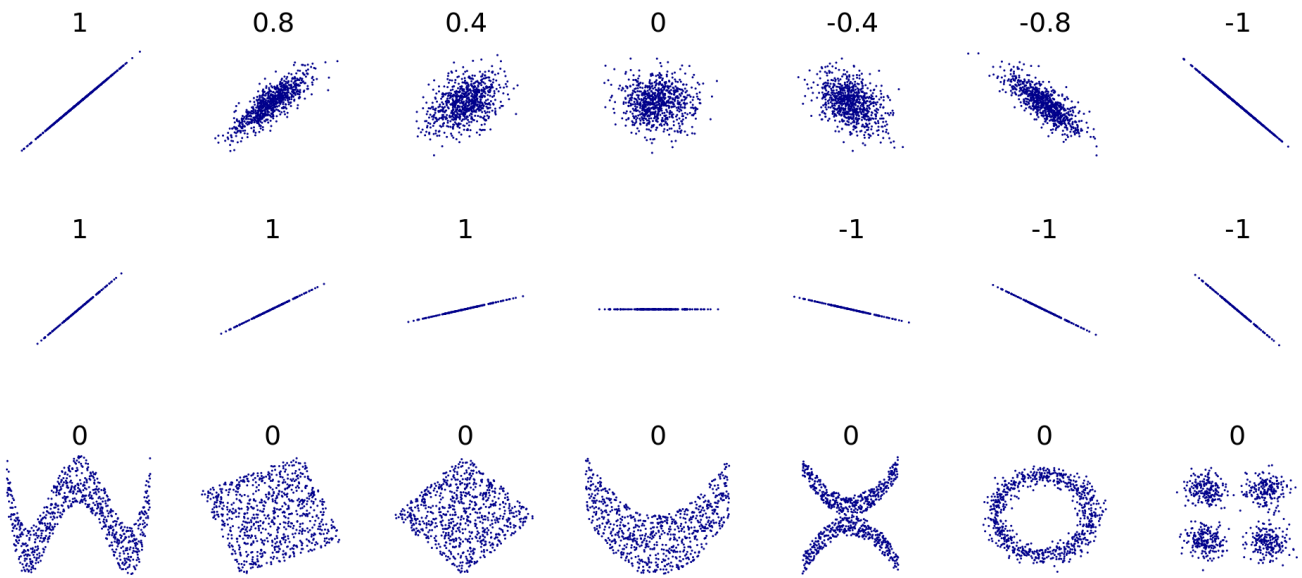
False Discovery

- any for of organization which is visible in my dataset, but cannot be generalized at the population level.
- is **not an error**, but is an inherent result of chance during **sampling** . . . we can see it a sort of "bad luck".

Section 2

Interlude: Correlation and Causation

Correlation



- Correlation (Pearson), measures the noisiness and direction of a linear relationship, but not the slope of that relationship.
- Two variables linked by causal relation are correlated
- The reverse is not true ...

Spurious correlations . . . ;-P

The NEW ENGLAND JOURNAL of MEDICINE

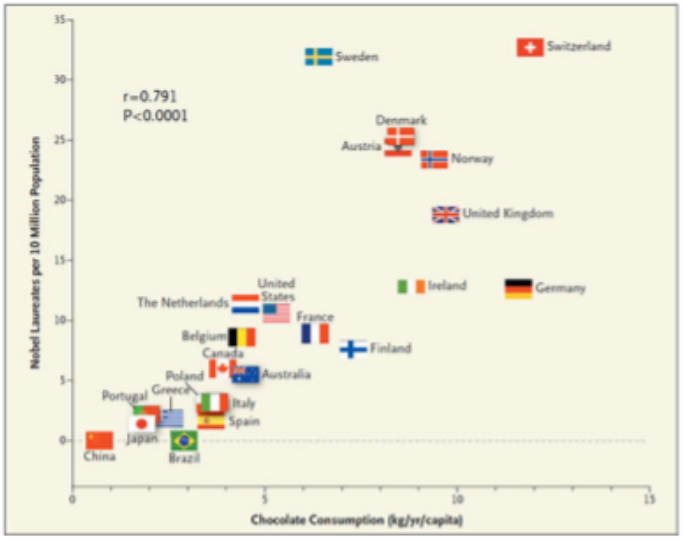
Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

N Engl J Med 2012; 367:1562-1564 October 18, 2012 DOI: 10.1056/NEJMon1211064

Chocolate consumption could hypothetically improve cognitive function not only in individuals but in whole populations. Could there be a correlation between a country's level of chocolate consumption and its total number of Nobel laureates per capita?

There was a close, significant linear correlation (r=0.791, P<0.0001) between chocolate consumption per capita and the number of Nobel laureates per 10 million persons in a total of 23 countries (Fig. 1)



Section 3

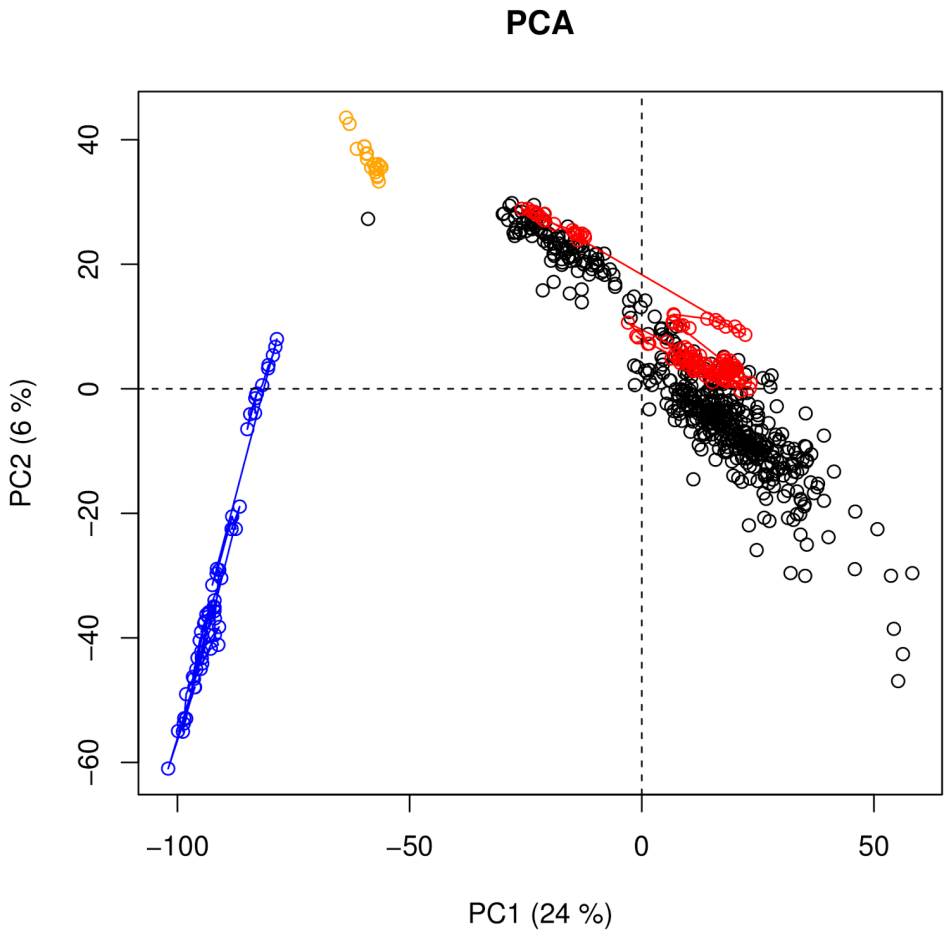
Data visualization

Data Visualization

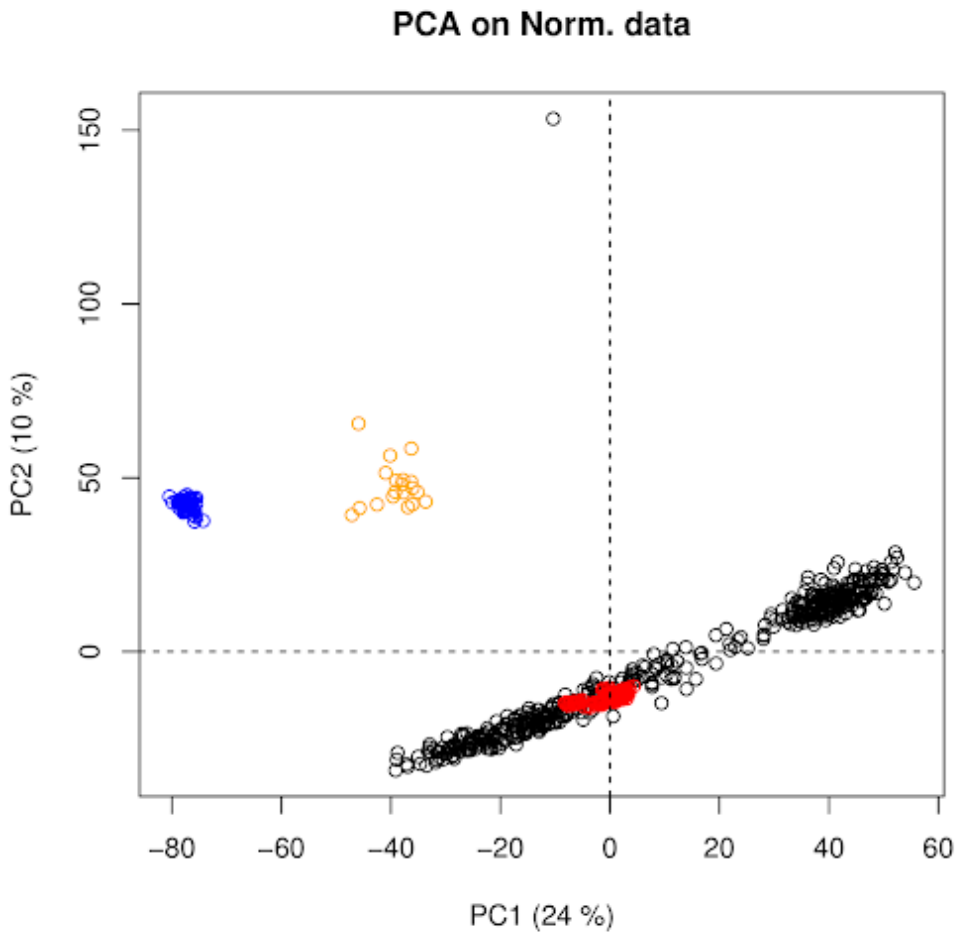
Choosing the right way to visually inspect the data allows to ....

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

Running Experiment



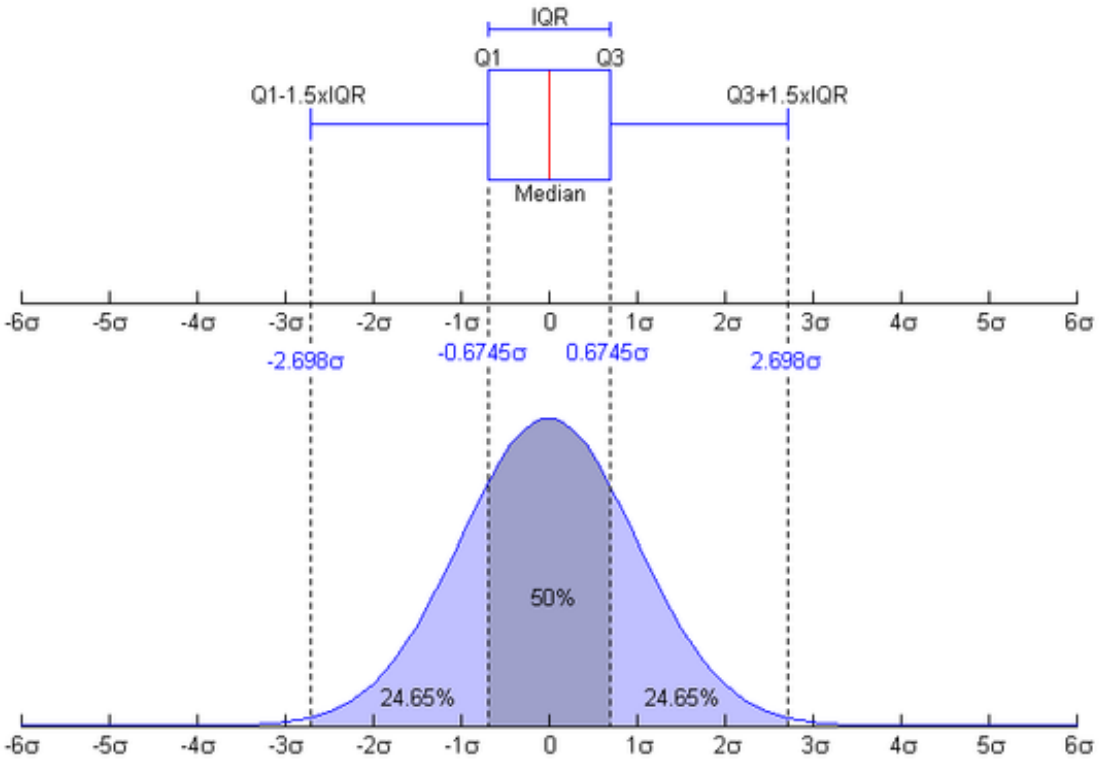
Outliers!



Dealing With outliers

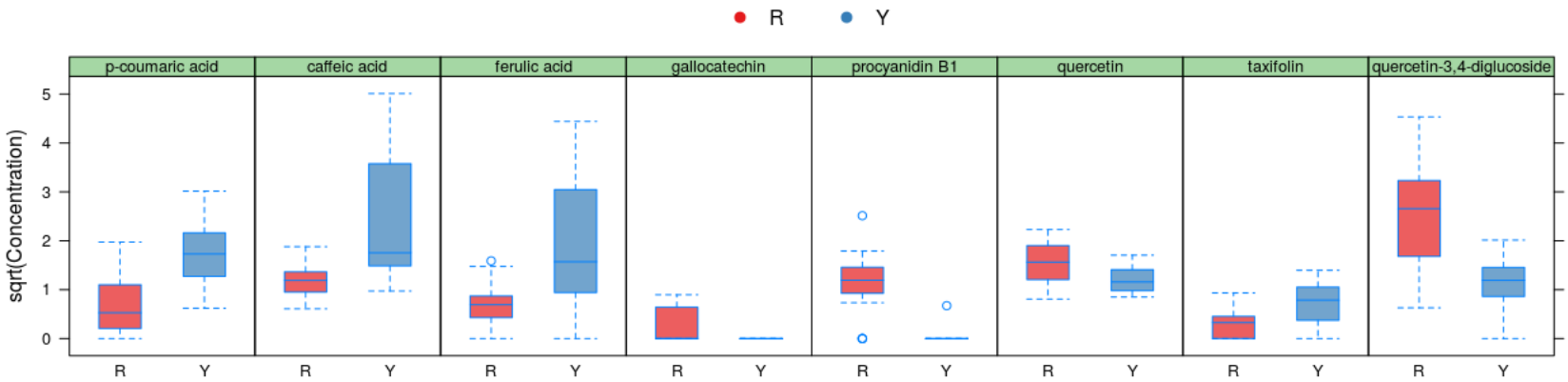
- It is unfair to exclude samples which are not in keeping with our theory/hope
- Sometimes outliers are indicators of unexpected and relevant science
- Samples can be excluded if there are **indisputable** evidence that they were "bad" (e.g. analytical errors)
- The best is to keep them in and rely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

Interlude: Boxplots

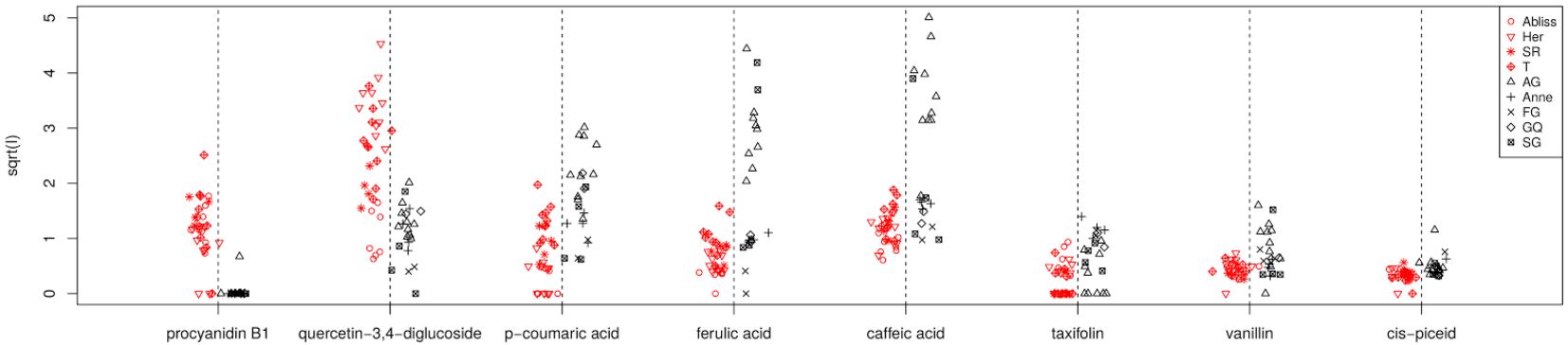


Boxplots nicely summarize the properties of a population, but I need to have a population !

Red and yellow berries



Rubus subpopulations



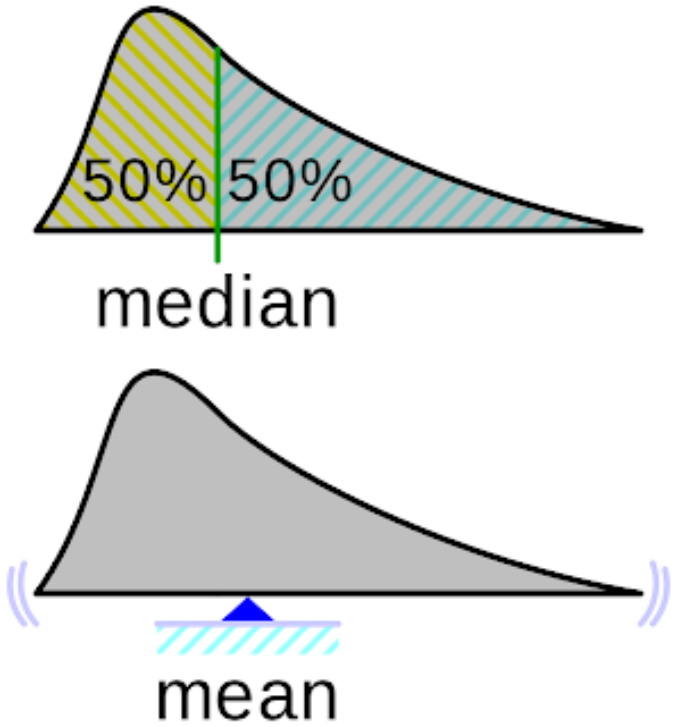
Take home messages

- Use boxplots wisely
- Plot the raw data points!
- Think in advance making a well planned experimental design

Data Distribution

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

Mean - Median



Mean and median

```
## my measuremnts
x <- c(1,2,3,4,5)
mean(x)
```

```
## [1] 3
```

```
median(x)
```

```
## [1] 3
```

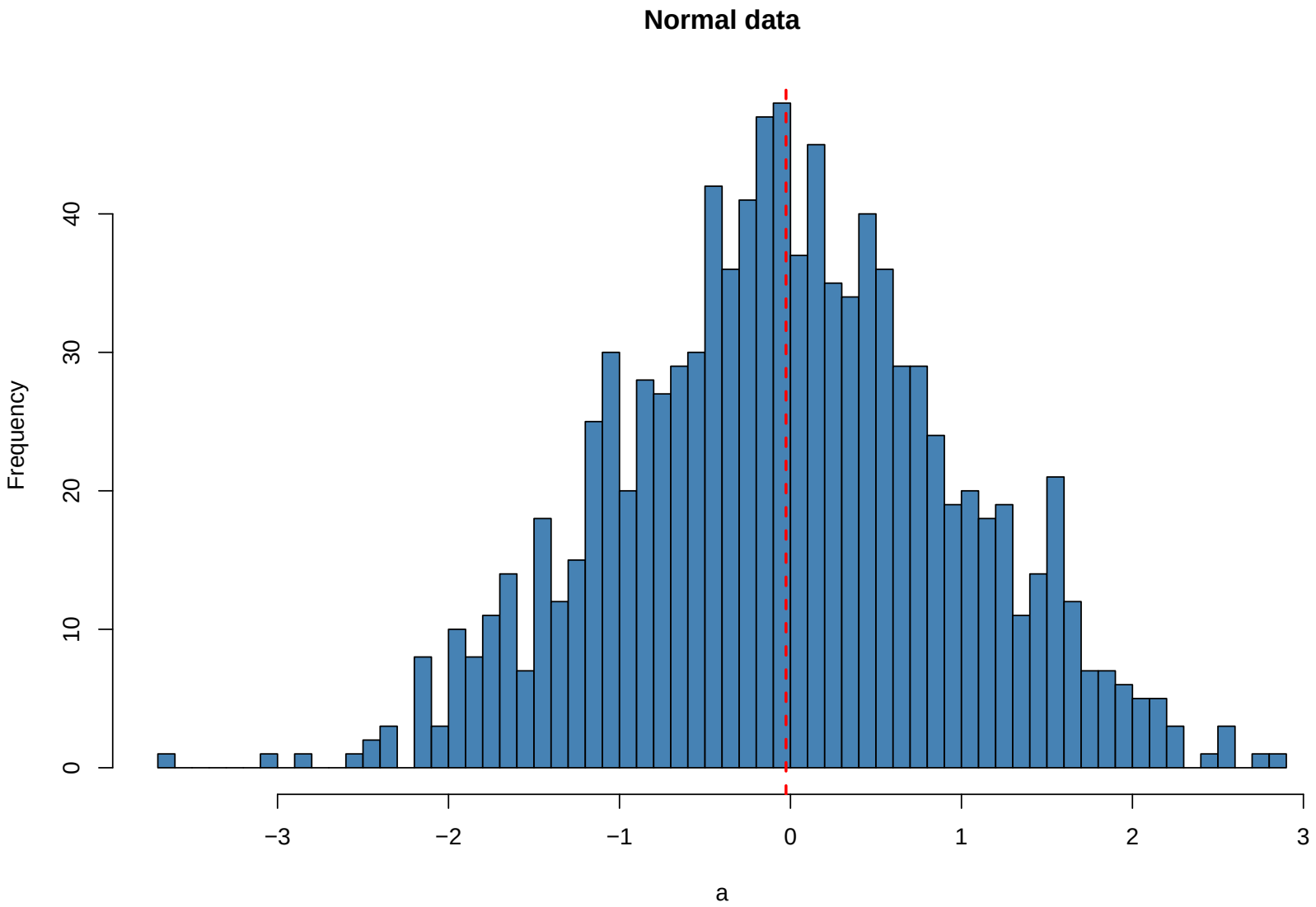
```
## my measuremnts with one outlier
y <- c(1,2,3,4,50)
mean(y)
```

```
## [1] 12
```

```
median(y)
```

```
## [1] 3
```

On the mean: normal data



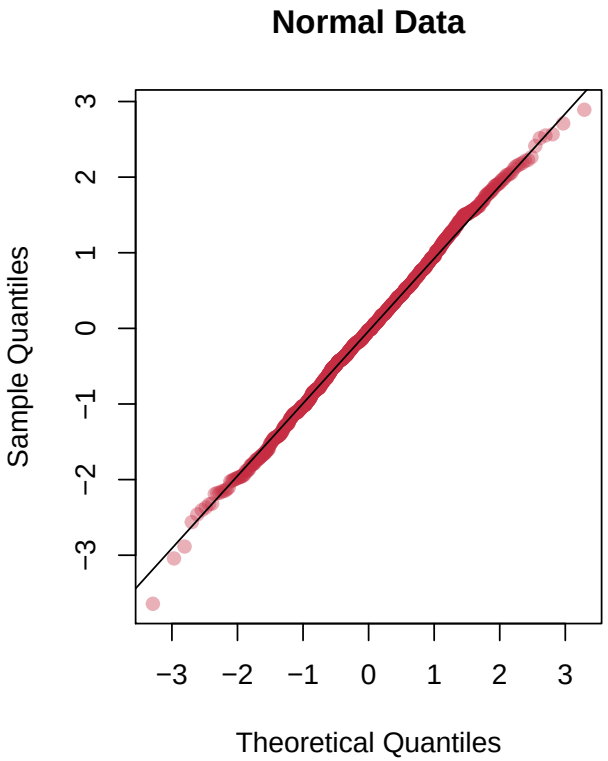
Take home messages

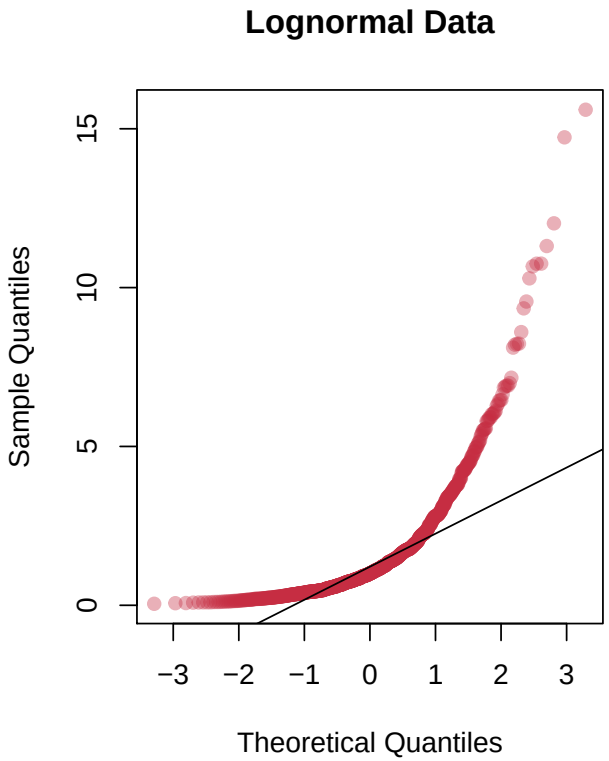
- With normally distributed data the mean is the most probable value
- Normality is a prerequisite for many statistical tools
- Statistics like the mean
- . . . but the mean is not robust
- . . . and many variable we measure are not normally distributed (e.g. counts)

Checking normality

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

QQ plots


Normal Data
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3


Lognormal Data
Sample Quantiles
Theoretical Quantiles
0
5
10
15
−3
−2
−1
0
1
2
3

Data Transformations

Promoting Normality

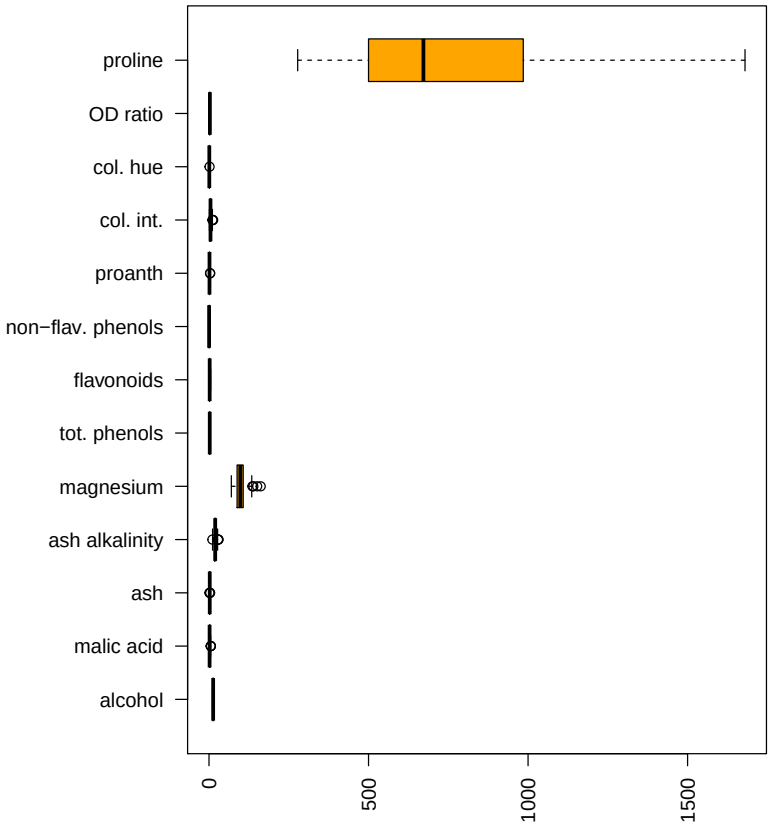
Non normally distributed data can be transformed into almost normal data prior to statistical analysis in order to avoid biased results.

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

Variable Scaling

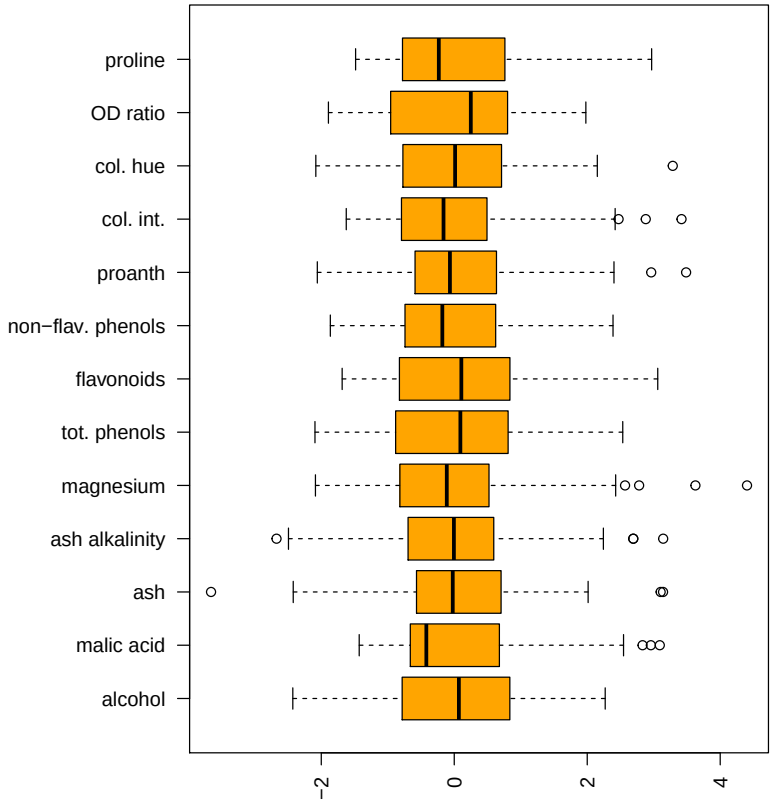
More often than not, the set of variables measured on the samples are highly variable in magnitude. High intensity variables will then determine the shape of the sample cloud in the multidimensional space.

Scaling is the process used to compensate for that



Autoscaling

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



Scaling Alternatives

- `log`: compresses high intensity values, but has problems with zeroes
- `sqrt`: here the compression is less severe, but zeroes are allowed
- Do you see issues with autoscaling?

Normalization

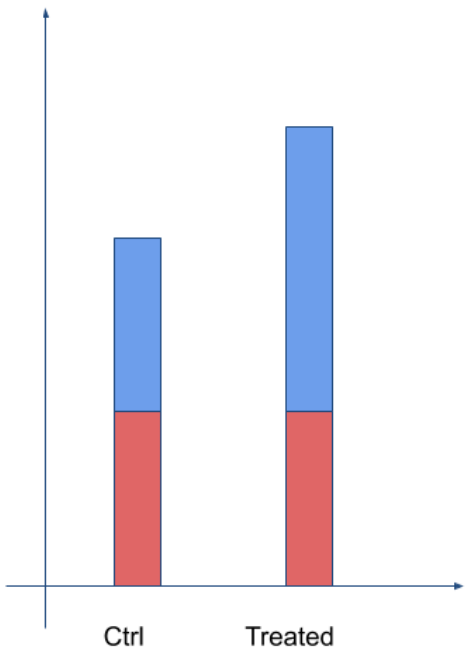
With **normalization** we indicate the process of transforming the intensities of the signal measured in each sample in order to make the samples directly comparable.

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

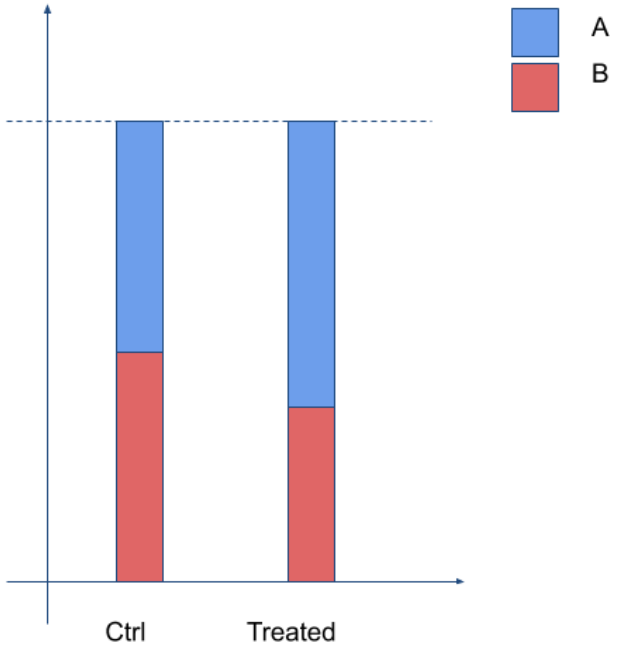
Methods

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

Compositional Data



Raw Intensities



Relative Intensities

Normalization has created a new biomarker!


LIVE
now
being
I am
here
becoming
sensing