# Applied Multivariate Statistical Analysis
# AMSA

I0P16

*Eddie Schrevens*

# Applied Multivariate Statistical Analysis

- Theorethical course in LAND 00.215, recorded and streamed lectures, and/or on-line knowledge clips
  - 13 modules of 2 h per week

- Practical exercises in Ludit PC-Class in 200C (Campus Arenberg III, Celestijnenlaan 200, building C , Heverlee 2001) or on-line R-demo-programs with contact moments in Blackboard Collaborate Ultra extended with learning task on real world data
  - 2 h demo modules in R
  - R-homework
  - Starting date to be announced
  - 5 - 7 sessions

- Assignment: 5 page paper: problem definition, data, analysis, interpretation
  - Send by mail to eddie.schrevens@biw.kuleuven.be at the latest 5 days before examination time (paper example on Toledo)

- Examination: closed book written examination (examination example on Toledo)
  - 1 theoretical question
  - 1 typical listing of results
  - 1 problem solving exercise
  - 1 question about the assignment

# Introduction

- Data
- Scale types of variables
- Random variables
- Information
- Geometrical representation of data
- Missing values and outliers
- Other data structures
- Definitions
  - Statistics
  - Computational sciences
  - Applied Mutivariate Statistical Analysis
  - Exploratory versus confirmatory  methodologies
  - Univariate-multifactorial-multivariate
  - Variable versus observation directed methods
- Outline of the course
- First principles

# Standard Form of Data

n x p data matrix

Data set

$$\begin{vmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{vmatrix}$$

# Data matrix or data set

- Rows: **observations (obs)**, individuals, units, readings, …

- Columns: **variables (var)**, descriptors, attributes, measurements, items, characteristics, responses, features, properties, …

# Data matrix or data set examples

- Clinical data
  - Obs: patients
  - Var: measurements on each patient, blood pressure, weight, …
- Fruit research
  - Obs: apples
  - Var: sugar content, color, weight, Vitamin C, …
- Agricultural field experiments
  - Obs: field, plant, plot in the field, ..
  - Var: yield, fertilization, disease and pest applications, growth, ..

# Data matrix or data set

- Observations are the  smallest units of information on which a set of variables (characteristics, attributes, features, properties) are measured
- Types of variables?
- In life sciences:  random variables
- What is the information we are looking for?
- Geometrical representations

# Scale types of variables

**Nominal:** Variable reflects the category to which an observation belongs.

Unordered categorical variable.

fi gender, variety, color, …

**Ordinal:** Variable reflects the ordered category to which an observation belongs.

Ordered categorical.

fi quality class, infection class, age class, health status, …

**Interval:** Numerical scale with arbitrary zero. Relative differences.

fi temperature

**Ratio:** Numerical scale with absolute zero. Absolute differences.

fi size, weight, volume, …

# Scale types of variables

*Categorical data: nominal or ordinal var, defining categories, classification-, indicator- or qualitative variables*

↳

Segment the data in different categories, fi Male and Female

We can NOT apply linear algebra

*Numeric data: interval or ratio, defining amounts, continuous- or quantity variables*

↳

Main emphasis of this course

We can apply linear algebra

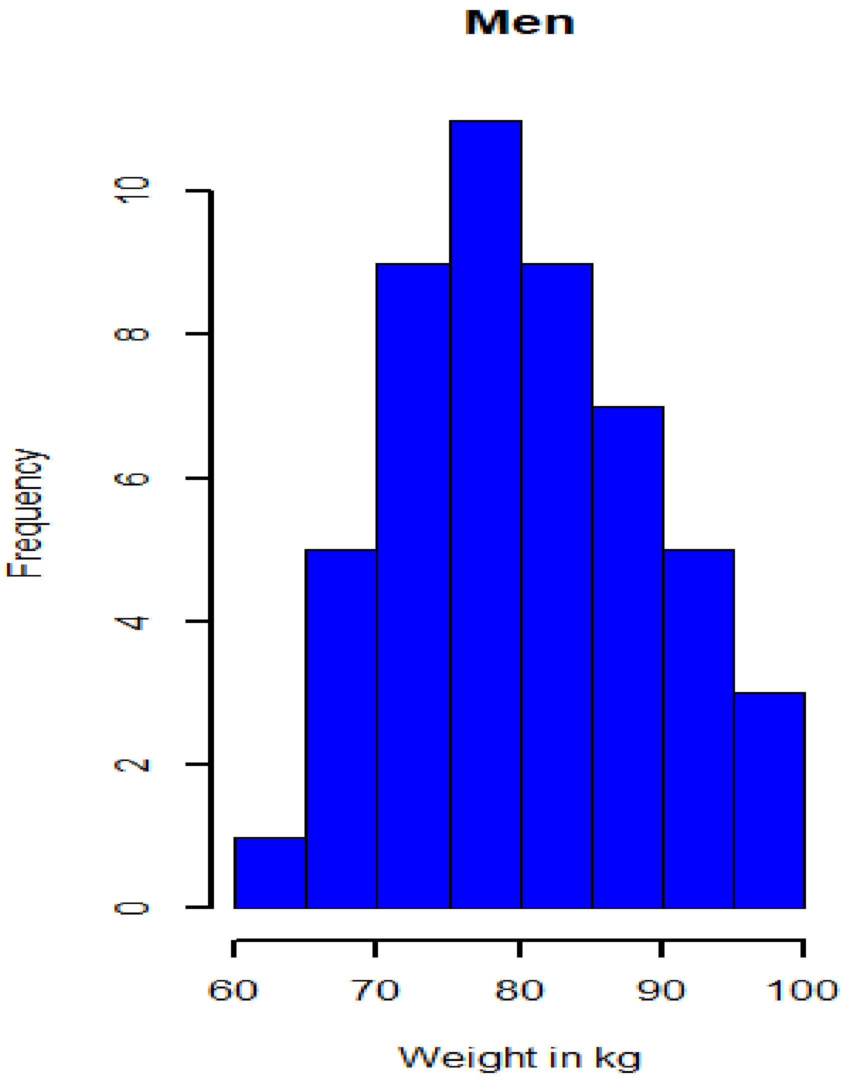
# Random variables in life sciences

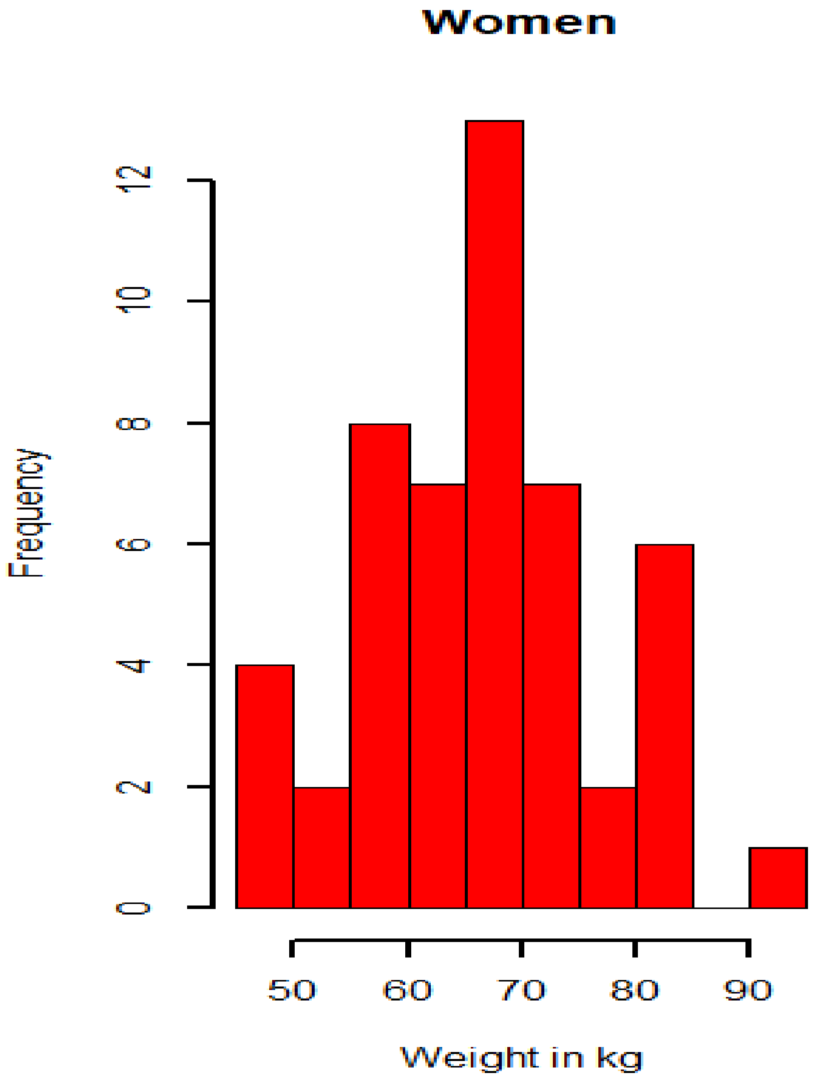
- Random variable
  - Variable that can take on a number of values
    - Categorical, tossing a coin
    - Numeric, weight of  males and females
  - A function that maps from an event (outcome of a probabilistic experiment) into a set of numbers
  - A variable characterised by intrinsic uncertainty

- Assume underlying probability models to tackle the uncertainty problem
  - Probability distributions for categorical data
  - Probability distributions for numeric data
  - For large datasets, construct probability distributions on the data

# Random variables: an example

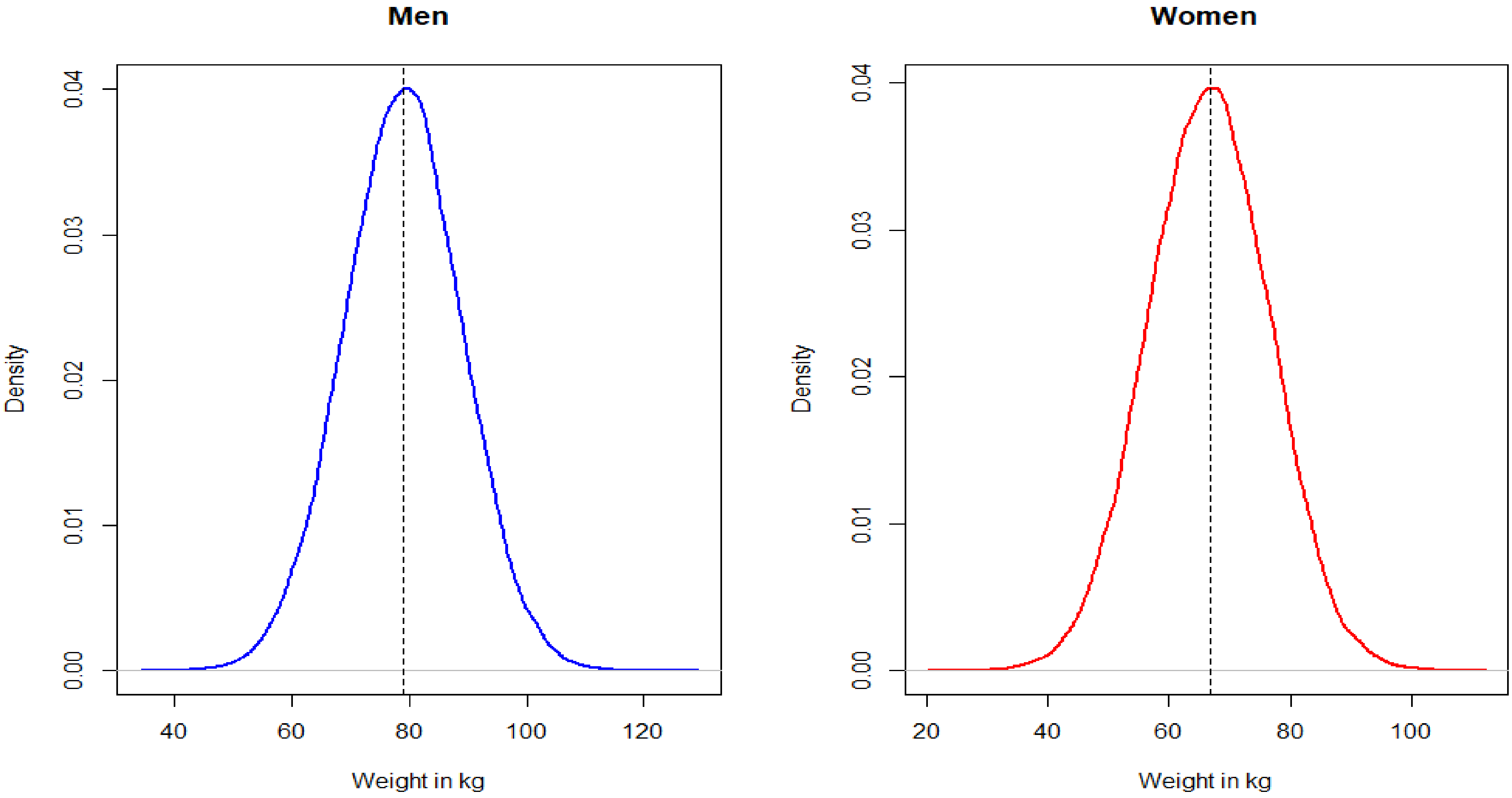
- Weight in kg of female and male Belgian adults (sample size 50 M, 50 F)





# Random variables: an example

- Weight in kg of female and male Belgian adults



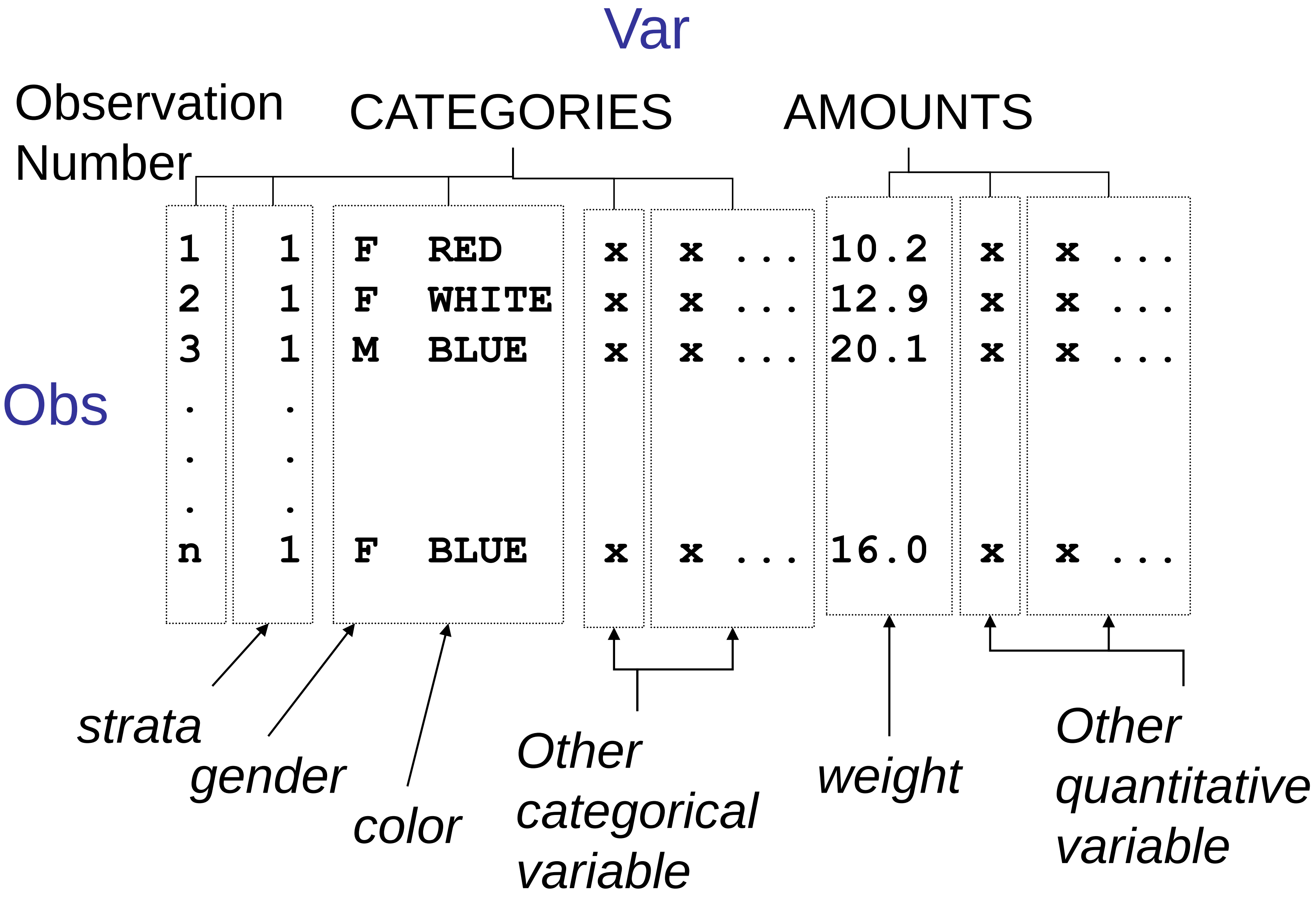
# Random variables: an example

R-Program

```
#  ‘#’ is  the code for program comments
#plot histograms
#Male: sample 50 values from a  univariate normal distribution with mean=79 and std=10
par(mfrow=c(1,2))
Weight_Male<- rnorm(50, mean=79,  sd=10)
hist(Weight_Male,xlab="Weight in kg",main="Men",lwd=2,col="blue")
#Female: sample 50 values from a  univariate normal distribution with mean=66.7 and std=10
Weight_Female<- rnorm(50, mean=66.7,  sd=10)
hist(Weight_Female,xlab="Weight in kg",main="Women",lwd=2,col="red")

#plot densities
#Male: sample 100000 values from a univariate normal distribution with mean=79 and std=10
par(mfrow=c(1,2))
Weight_Male<- rnorm(100000, mean=79,  sd=10)
Dens_Weight_Male <- density(Weight_Male)
plot(Dens_Weight_Male,xlab="Weight in kg",main="Men",lwd=2,col="blue")
abline(v=79,lty="dashed")
#Female: sample 100000 values from a  univariate normal distribution with mean=66.7 and std=10
Weight_Female<- rnorm(100000, mean=66.7,  sd=10)
Dens_Weight_Female <- density(Weight_Female)
plot(Dens_Weight_Female,xlab="Weight in kg",main="Women",lwd=2,col="red")
abline(v=66.7,lty="dashed")
```

# Standard Form for a Data Set



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | F | RED | x | x ... | 10.2 | x | x ... |
| 2 | 1 | F | WHITE | x | x ... | 12.9 | x | x ... |
| 3 | 1 | M | BLUE | x | x ... | 20.1 | x | x ... |
| . | . | | | | | | | |
| . | . | | | | | | | |
| . | . | | | | | | | |
| n | 1 | F | BLUE | x | x ... | 16.0 | x | x ... |

# Information?

DATA ≠ INFORMATION

DATA ANALYSIS → INFORMATION

KNOWLEDGE?

KNOWLEDGE = INTEGRATED INFORMATION

# Information?

Searching for information in data is looking for non-random patterns

# Data matrix

- Row-vector of p column vectors (var)

$$\mathbf{X} = [\mathbf{var}_1\ \mathbf{var}_2\ \mathbf{var}_3\ \ldots\ \mathbf{var}_p]$$

- Column-vector of n row-vectors (obs)

$$\mathbf{X} = \begin{bmatrix} \mathbf{obs}_1 \\ \mathbf{obs}_2 \\ \mathbf{obs}_3 \\ \ldots \\ \mathbf{obs}_n \end{bmatrix}$$

Convention in AMSA

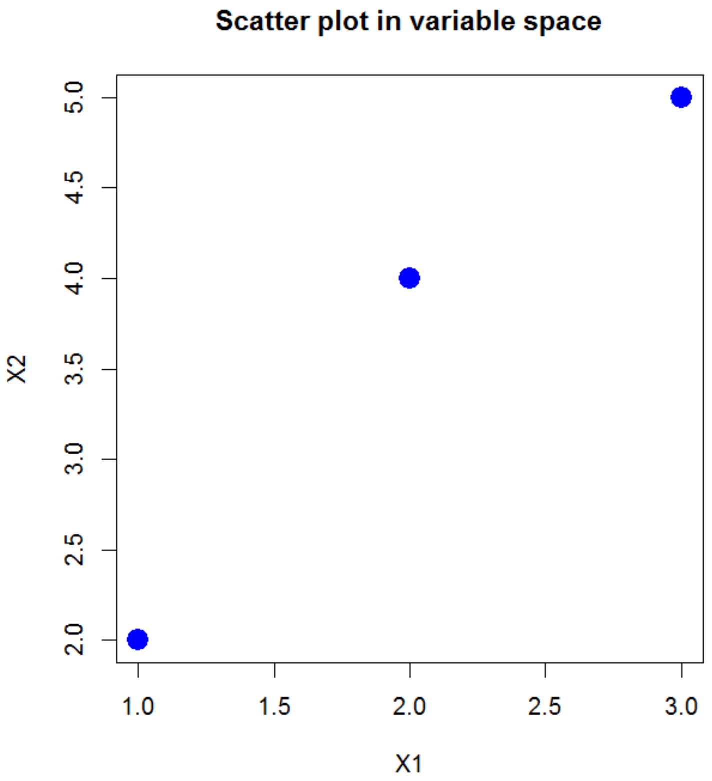
**X: capital bold = matrix**
**x: small caps bold = vector**
x: regular = scalar

# Geometrical representation of numerical data

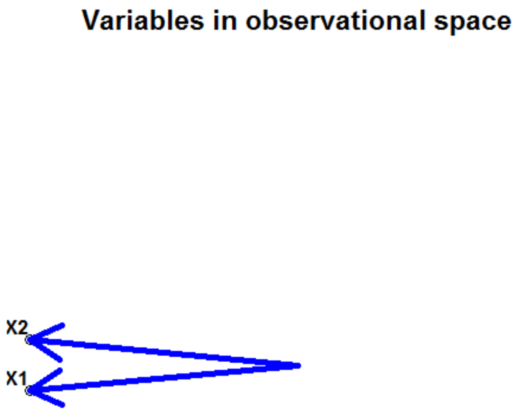
- Point representation in the variable space
  - Scatter plot
- Vector representation in the observational space
  - Vector plot

Geometrical equivalent of data is a ‘cloud’ or ‘swarm’ of points and/or vectors in a multidimensional space (var and/or obs definied)

# Variable space ⇔ observational space



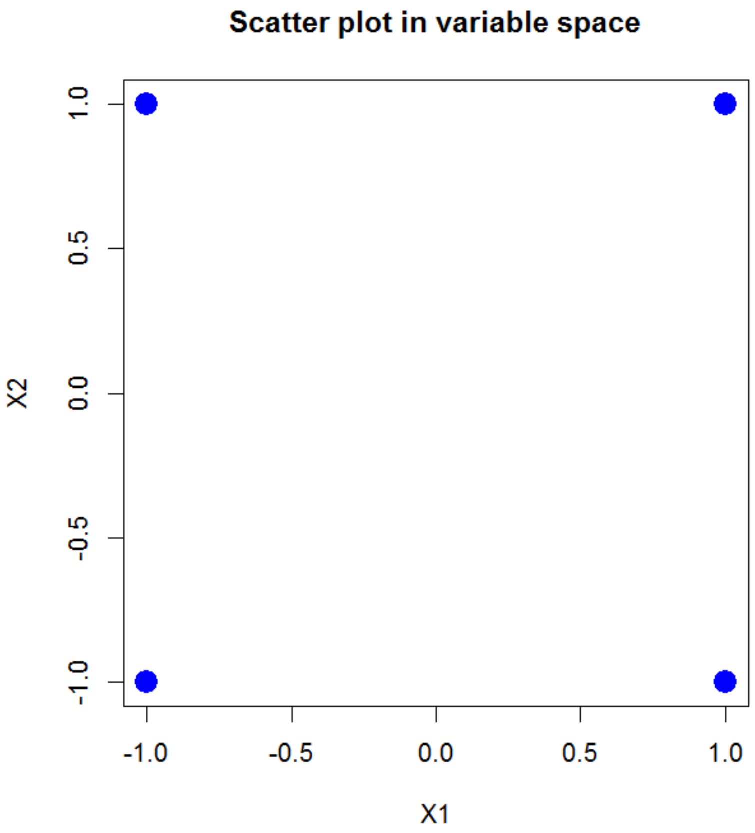
Scatter plot in variable space

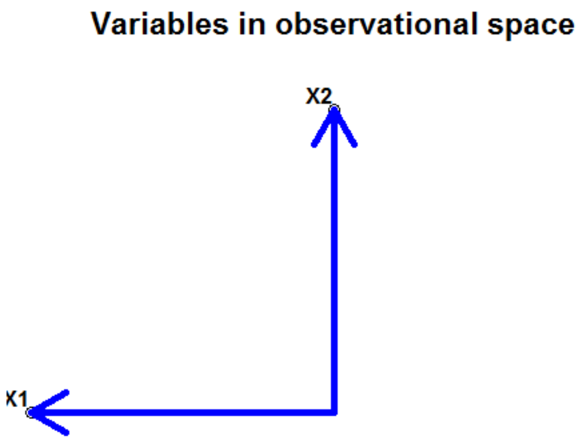


Variables in observational space

$$\begin{array}{cc} & \begin{array}{cc} \mathbf{x1} & \mathbf{x2} \end{array} \\ \mathbf{X} = & \begin{bmatrix} 3 & 5 \\ 1 & 2 \\ 2 & 4 \end{bmatrix} \end{array}$$

# Variable space ⇔ observational space





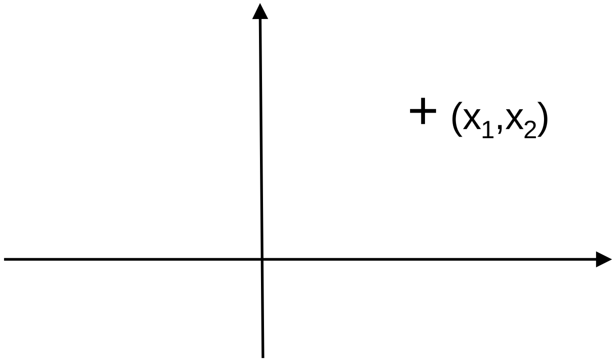
$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{bmatrix}$$

$2^2$ Factorial design

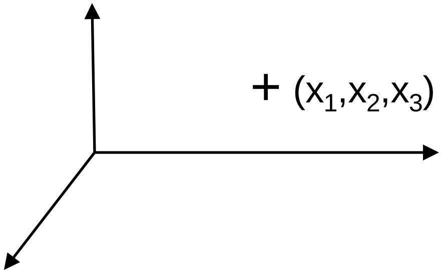
# Geometrical Interpretation

A vector consisting of p elements can be regarded geometrically as a point in an p-dimensional space.

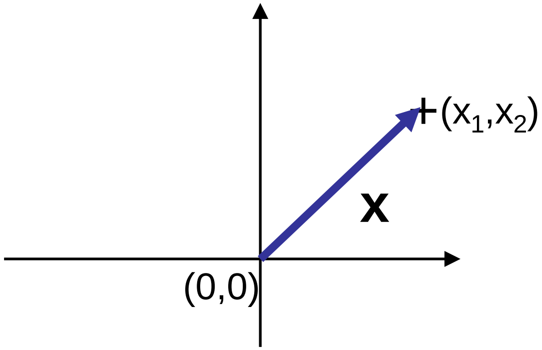
2-dimensional space

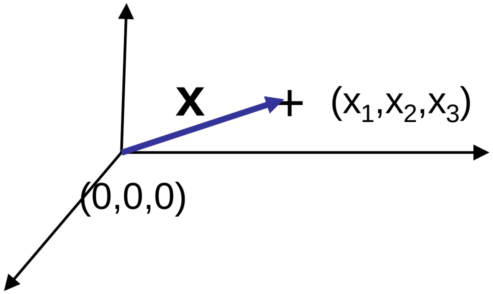


3-dimensional space



A vector consisting of p elements can also be regarded geometrically as a (directed) line from the origin to a point in p-dimensional space.





$$\boxed{\mathbf{x} = \vec{\mathbf{x}}}$$

# Major problems with data matrices

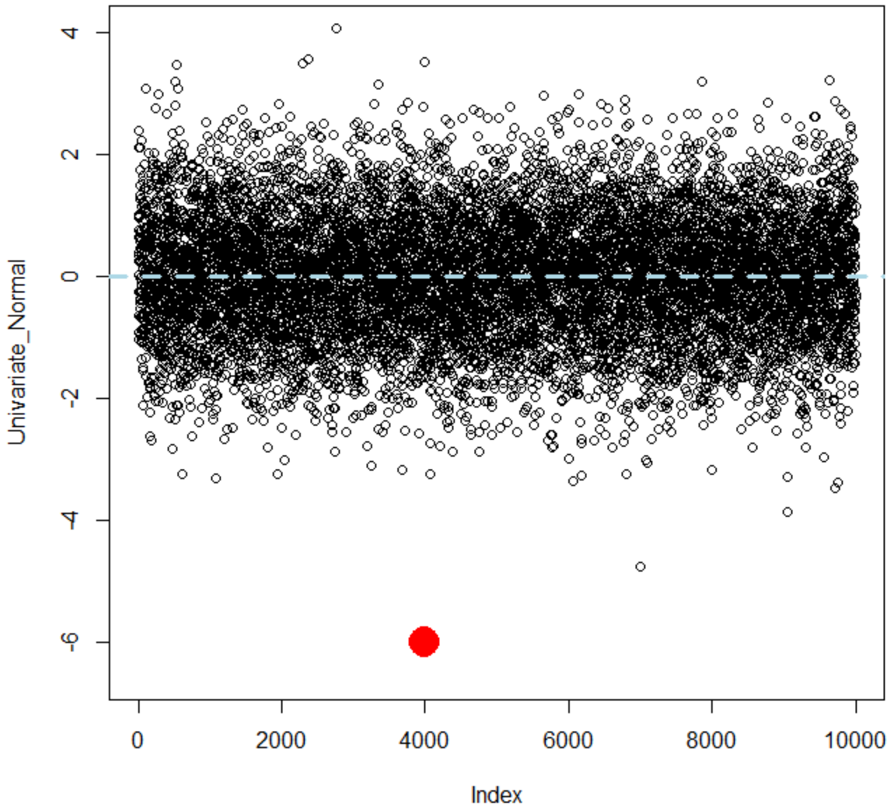
- Missing values:
  - Some obs do not have  values for certain var (holes in the data matrix!)
  - Examples:  soil samples were lost, food samples perished, patients died, accidents in the lab, …
  - Solutions?
    - Delete the obs with missing values (R see exercises)
    - Missing value imputation:  Use a model to fill the holes in the data matrix. Never impute more than 5 % of your data!

- Outliers:
  - Obs that are not belonging to the population under study

# Outliers: univariate case

Obs not belonging to the population under study

10000 obs sampled from a standard normal distribution (mean=0, std=1)



Is the red obs an outlier?

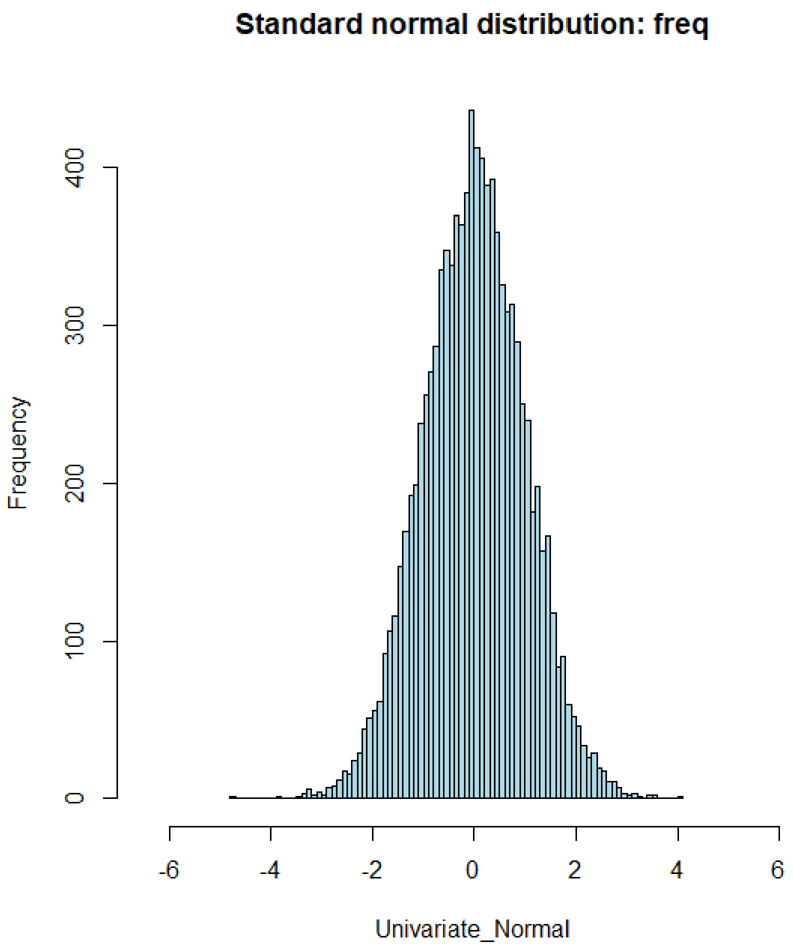
Very “far” away from the mean of the population

Would a distance measure be a good approach to detect outliers?

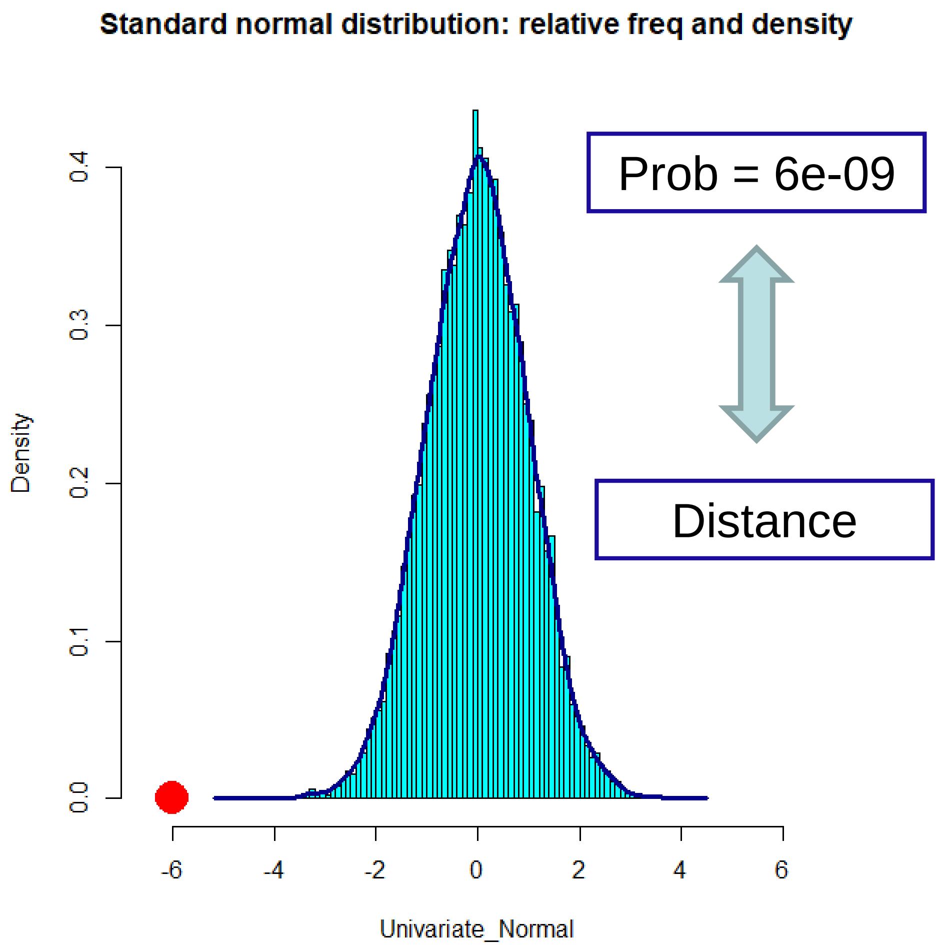
# Outliers: univariate case

## Obs not belonging to the population under study

10000 obs sampled from a standard normal distribution (mean=0, std=1)



Standard normal distribution: freq



Standard normal distribution: relative freq and density

Prob = 6e-09

Distance

# Outliers: univariate case

10000 obs sampled from a standard normal distribution (mean=0, std=1)

R-Program

```
#  "#" this characters marks comments in the code
#sample 10000 values of the standart normal distribution
#put them in an object called Univariate_Normal
Univariate_Normal <- rnorm(10000, mean=0,  sd=1)
#plot the values from 1 tot 10000 and set the scale of the x-axis
plot(Univariate_Normal,ylim=c(-6.5,4))
#put an extra obs at (4000,-6)
points(4000,-6,col=2,pch=16,cex=3)
#draw a horizontal line at 0
abline(h=0,col="lightblue",lty="dashed",lwd=3)

#make plotting window for 1 row and 2 columns
par(mfrow=c(1,2))
#plot a histogram of the absolute freq
hist(Univariate_Normal,breaks=100,col="lightblue",main="Standard normal distribution: freq",xlim=c(-6.5,6.5))
#plot a histogram of the relative freq
hist(Univariate_Normal,breaks=100,col="cyan",main="Standard normal distribution: relative freq and
density",cex=0.7,freq=F,xlim=c(-6.5,6.5))
#calculate the probability density of these data
Dens_Univ_Norm <- density(Univariate_Normal)
#overlay the density on the last plot
lines(Dens_Univ_Norm,col="darkblue",lwd=3)
#add the outlier
points(-6,0,col=2,pch=16,cex=3)

#calculate the probabilityy of the outlier
dnorm(-6,mean=0,sd=1)
```

# Outliers: bivariate case

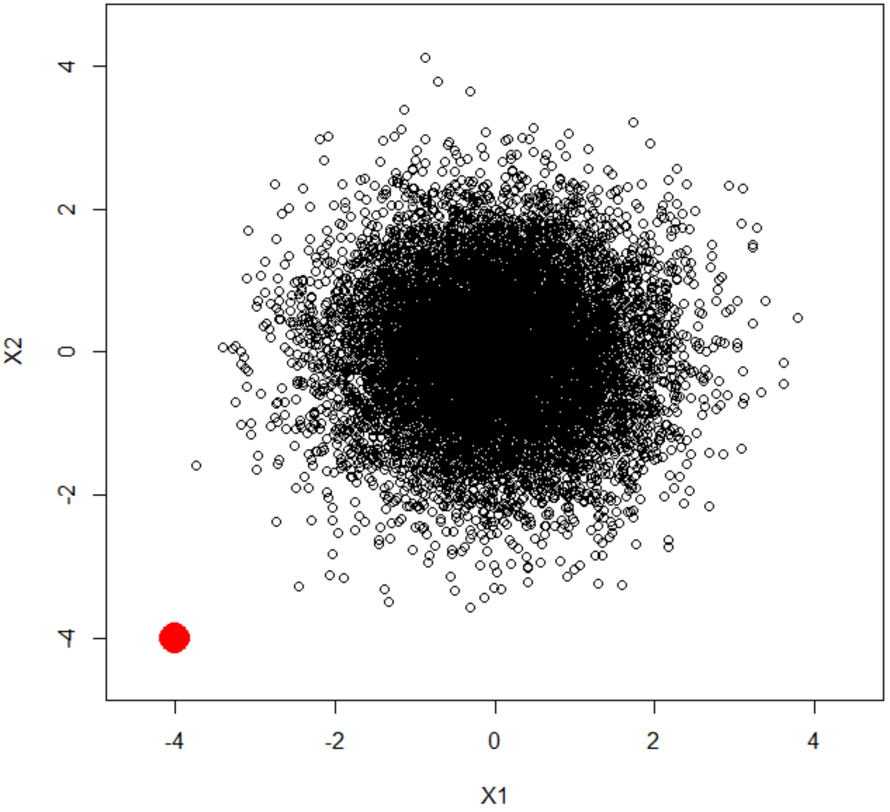
Obs not belonging to the population under study

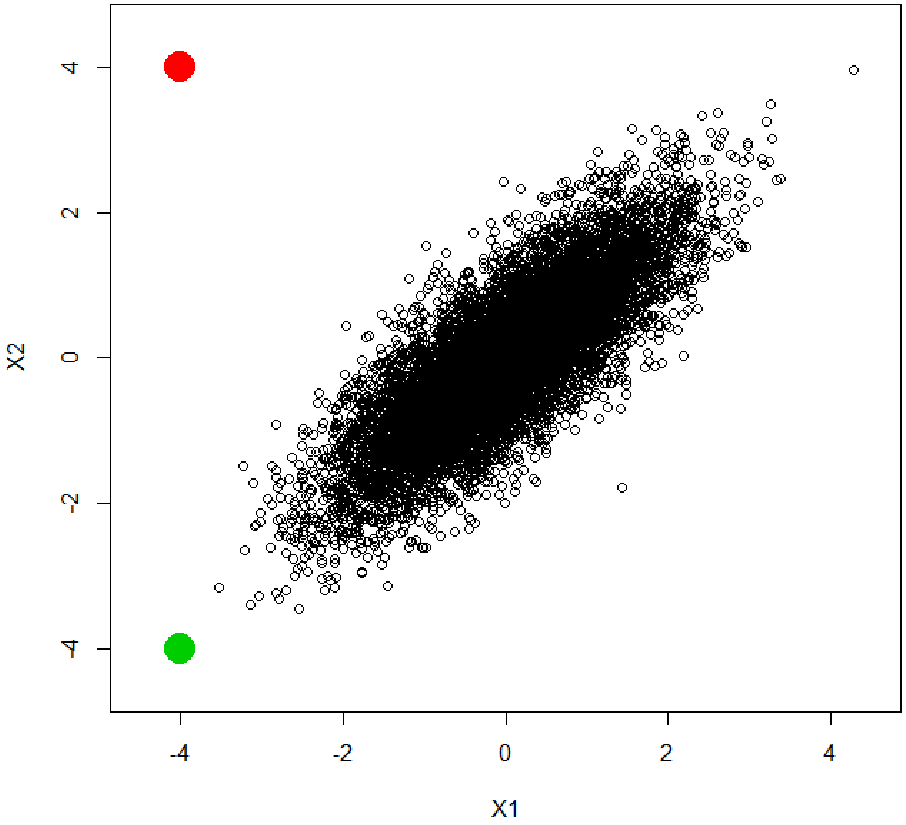
Bivariate point clouds of 1000 obs

Prob = 1.7e-08

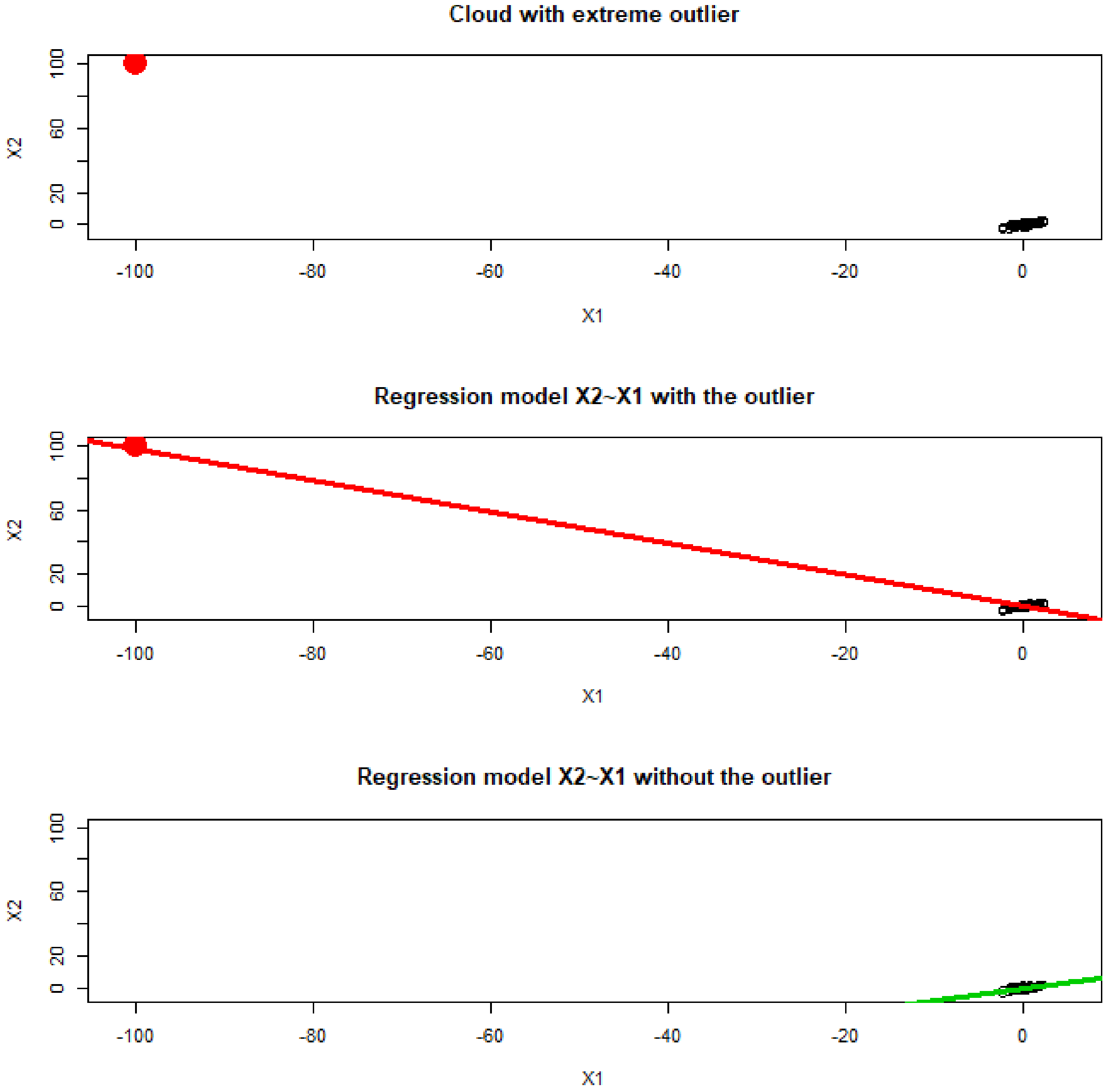
Prob = 4.7e-36

Prob = 3.6e-05





# Outliers: effect on stats models



Cloud with extreme outlier

Regression model X2~X1 with the outlier

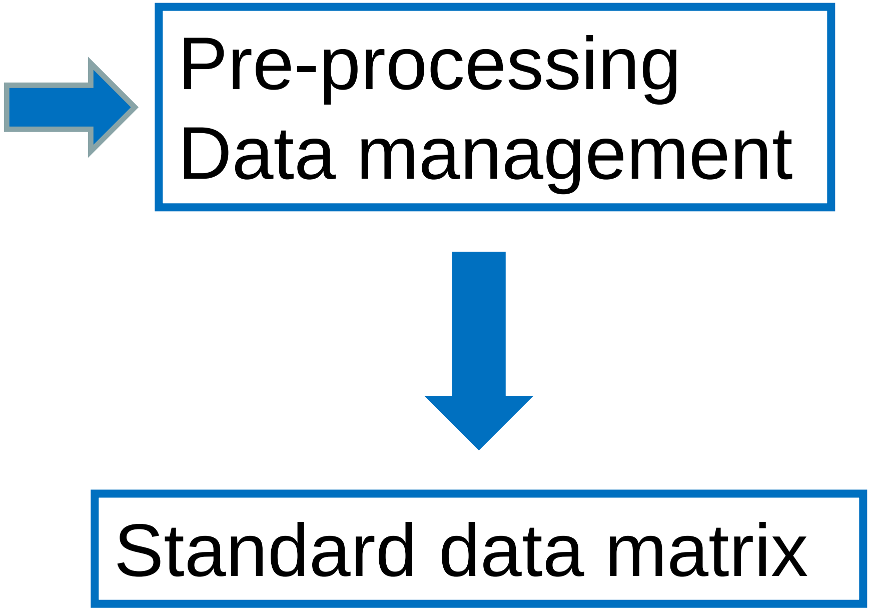
Regression model X2~X1 without the outlier

Many standard multivariate methodologies are also based on linear model theory!

Outliers can completely distort model efforts

# Other data structures

- Functional relations
  - Time dynamics, fi growth curves
  - Spectral data
  - Spatial data
  - Gene expression data
- Images
  - Digital camera images (RGB)
  - Multi-spectral images
- Models
  - Crop growth and development
  - Corona SIR model scenario’s



Pre-processing
Data management

Standard data matrix

# Short versus long data format

## Spectral, temporal and spatial data

Long format

| Obs nr | Time or Sensor nr or Location | Var |
|---|---|---|
| 1 | 1 | 2.3 |
| 1 | 2 | 4.6 |
| 1 | 3 | 6.2 |
| 1 | 4 | 8.4 |
| 1 | 5 | 11.6 |
| 1 | 6 | 14.0 |
| 1 | 7 | 17.8 |

Short format

| Obs nr | Var T1 | Var T2 | Var T3 | Var T4 | Var T5 | Var T6 | Var T7 |
|---|---|---|---|---|---|---|---|
| 1 | 2.3 | 4.6 | 6.2 | 8.4 | 11.6 | 14.0 | 17.8 |
| 2 | | | | | | | |
| 3 | | | | | | | |

Each cell in the short format is a mix of different variables (fi var and time)
A variable should be uniquely defined!

Always use long format

# Definitions

Statistics

A set of methods for obtaining, organizing, summarizing, presenting and analyzing measurable facts (data) in order to help make wise decisions in the face of uncertainty

- Originated from agricultural sciences
  - Intrinsic uncertainty of biological processes and systems
- Problem solving oriented
- Assume underlying probability distributions to tackle the intrinsic uncertainty of life science data

# Definitions

Computational data analysis methods

Set of methods for automated data(-base) analysis

- Originated from computer sciences
- Arose from theoretical considerations, fi AI, ANN, SVM, Machine learning, …

# Definitions

Information processing, multivariate data analysis, data mining, data science

Application of methods on data, consisting of large numbers of variables, measured on each observation of one or more groups, aiming at investigating relations between variables and/or between observations or group structures

Integration of statistics and computational methods

Best of both worlds

# Information?

Searching for non-random patterns/structures in data

- Investigating relations between variables and/or between observations
- Investigating grouping structures of observations and/or variables
- Investigating differences in relations between variables and/or between observations in different grouping structures

# Components of information processing, data science

1. Data acquisition
   - How to get data?
     - Measure theory-experimental design-sampling plans
2. Data management
   - How to organize and understand the data?
     - Database set-up
     - Preprocessing, filtering, transformation, MV, outliers, exploratory graphical analysis, …
3. Data analysis
   - What to do with data?
     - Confirmatory ⇔ exploratory data analysis (MVA)
     - Extract information

# Related disciplines and application fields

Biometrics, psychrometrics, econometrics, chemometrics, environmetrics, agro-metrics, …

Data analysis, IDA, knowledge discovery, KDD, data mining, business intelligence, pattern recognition, machine learning,  bio-informatics, computational molecular biology, informative pattern discovery, multipattern discovery, modelling subjective uncertainty, inductive logic programming, association rules discovery, nugget searching, …

AMSA  =  Data science discipline

# Exploratory ⇔ confirmatory methods

Confirmatory methods, Inferential methods

Formulation and verification of hypotheses

MANOVA

- Compare two or more populations on a large number of numeric variables

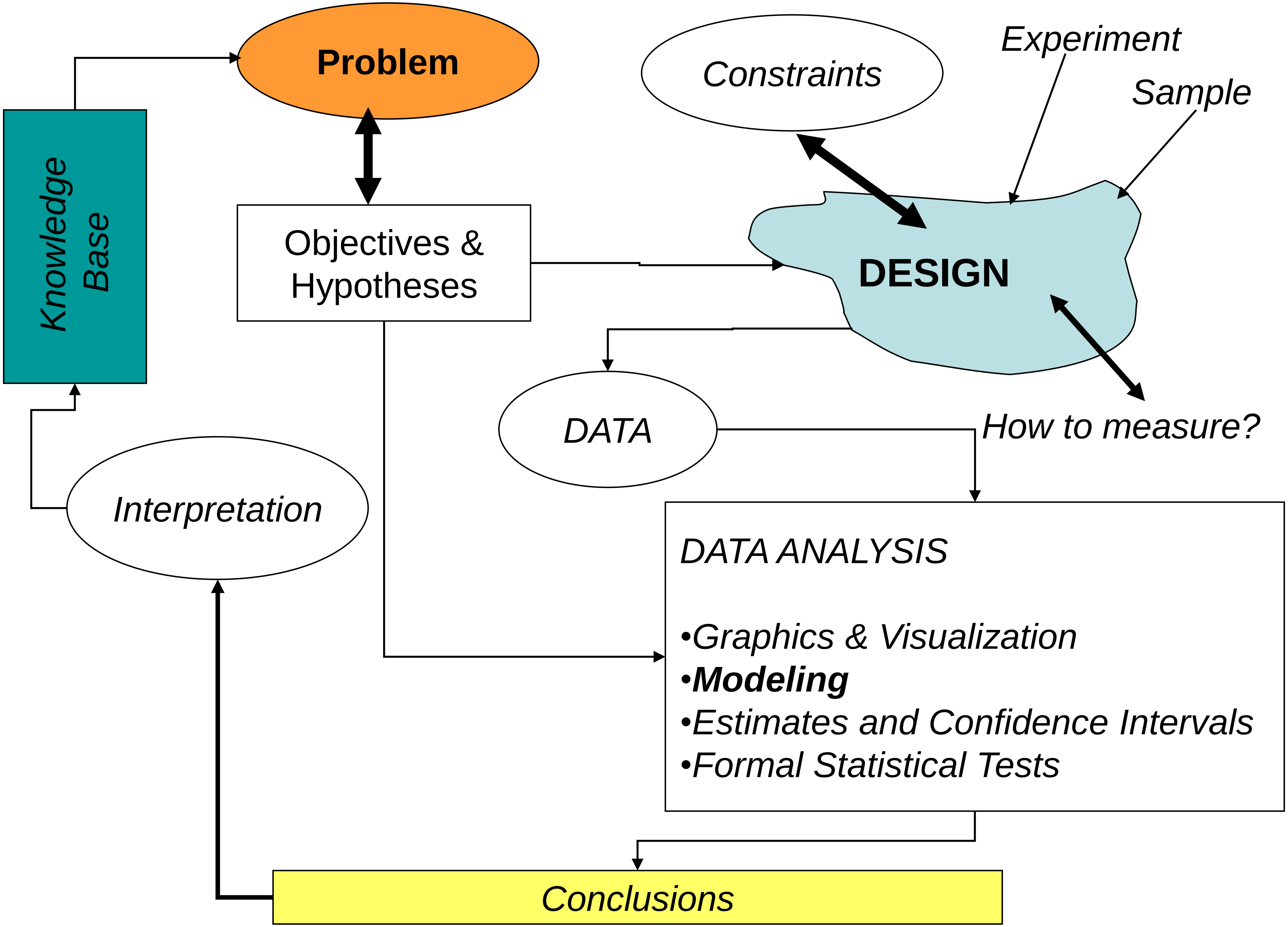
Multiple regression

- Examine the relations between a collection of dependent variables and independent variables

Inferential methods use the (Multivariate) General (Generalised) Linear Model

Aplied Linear Statistical Models, Neter, Kutner, Nachtsheim and Wasserman

## Confirmatory methodology by experiment



Problem

Knowledge Base

Objectives & Hypotheses

Constraints

Experiment

Sample

DESIGN

DATA

How to measure?

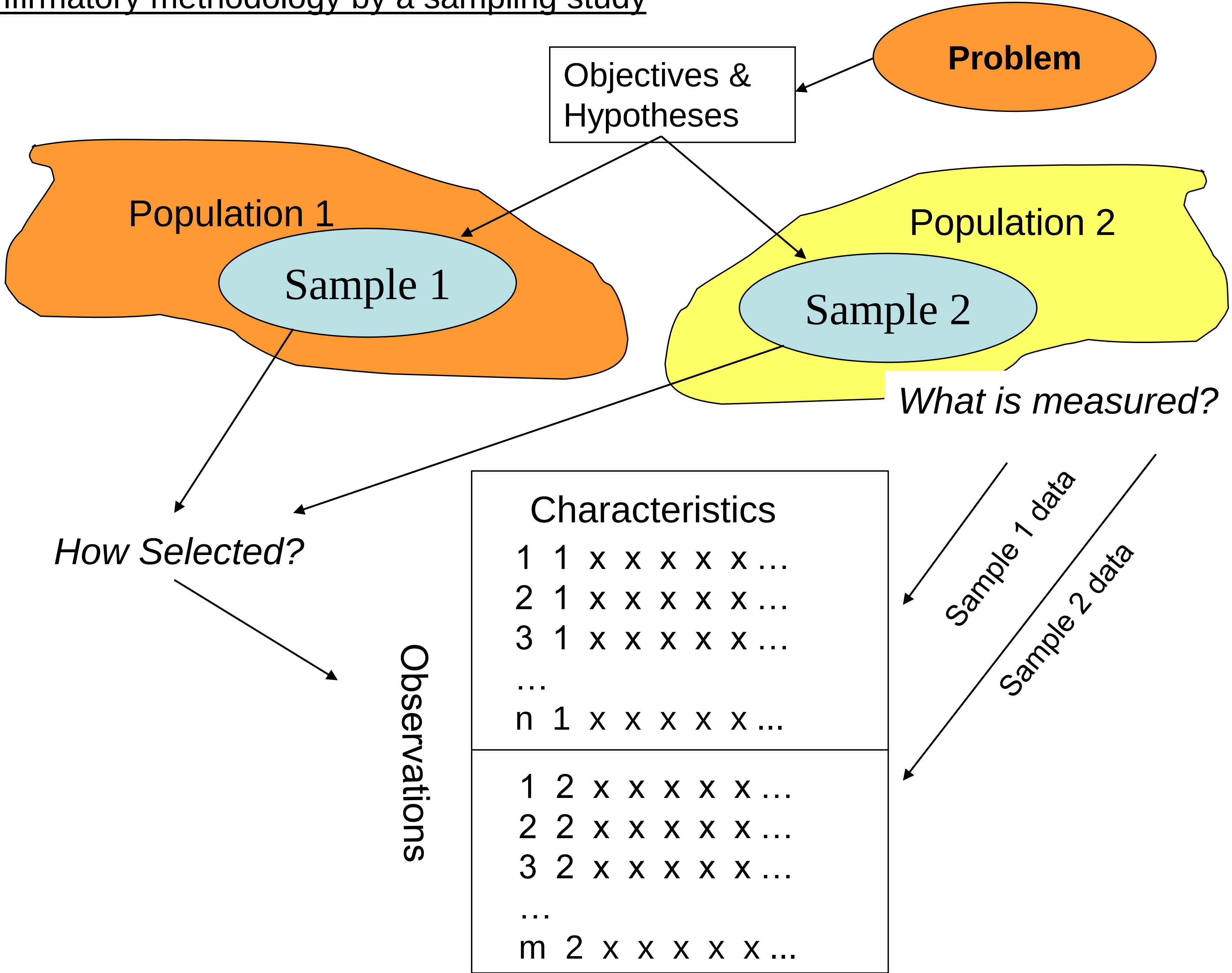
Interpretation

*DATA ANALYSIS*

- *Graphics & Visualization*
- ***Modeling***
- *Estimates and Confidence Intervals*
- *Formal Statistical Tests*

*Conclusions*

# Confirmatory methodology by a sampling study



**Problem**

Objectives & Hypotheses

Population 1

Sample 1

Population 2

Sample 2

*What is measured?*

*How Selected?*

Sample 1 data

Sample 2 data

Observations

Characteristics

1 1 x x x x x …
2 1 x x x x x …
3 1 x x x x x …
…
n 1 x x x x x ...

1 2 x x x x x …
2 2 x x x x x …
3 2 x x x x x …
…
m 2 x x x x x ...

# Exploratory ⇔ confirmatory methods

Exploratory methods ⇔ multivariate analysis

- Graphical exploratory data analysis
  - Relations, grouping structures
- Finding common (latent) factors , ordination
  - Relations, grouping structures
- Discrimination, supervised classification
  - Grouping structures
- Grouping, clustering, unsupervised classification
  - Grouping structures

In this course we will use the original, authentic historical names for different methods

# Exploratory ⇔ confirmatory methods

## Exploratory methods ⇔ multivariate analysis

### Finding common factors , ordination

- Condensing the information contained in a number of original variables into a smaller set of dimensions with minimum loss of information
- Analyze the interrelationships among a large number of variables and explain these variables in terms of their common underlying dimensions (called factors or latent variables).

### Discrimination

- Given the data are divided into predefined groups (2 or more), use the multivariate measurements to predict the likelihood of group membership for a new observation.
- Equivalent to multiple regression with a non-metric or categorical (dichotomous [yes-no] or multi-chotomous [red-blue-green]) response variable.

### Grouping, clustering

- Techniques for developing meaningful subgroups of individuals or objects.
- Classify observations or variables into a small number of mutually exclusive groups based on similarity among the entities.
- Groups are not predefined.

# Exploratory methods

- Data exploration in an attempt to recognize any nonrandom pattern or structure requiring explanation (data mining)
- Finding the question is more important than seeking subsequent answers
- Formal models designed to yield specific answers to rigidly defined questions are in many situations not required (opposite to confirmatory methods)

- Generate possible hypotheses for further research

Exploratory ⇒ confirmatory ⇒ knowledge increase

# Univariate-multifactorial-multivariate

- Univariate: system description with 1 independent variable
  - Simple regression, 1-way ANOVA
- Multi-factorial: system description with many independent variables
  - Multiple regresssion, n-way ANOVA
- Multivariate: system description with many dependent variables
  - No distinction between dep or indep

# Var directed⬄ Obs directed

- Var directed
  - Ordination methods
  - Dimension reduction
  - Relations between variables
  - Variable grouping
- Obs directed
  - Investigating grouping structure of the obs
  - Relations between obs
- Sometimes Var and Obs are interchangable

# AMSA-outline

- Part 1
  - Introduction
  - Linear algebra: vectors-matrices-data analysis
  - Graphical exploratory data analysis
- Part 2: Ordination methods, in search for latent factors
  - Principal components analysis
  - Factor analysis – structural equation modelling
  - Biplotting
  - Partial least squares
  - Multidimensional scaling
- Part 3: Classification methods
  - Cluster analysis, hierarchical, non-hierarchical, HICUPP
  - Discriminant analysis, logistic regression, Tree based modelling, Neural networks

# Example data matrix

On 150 Iris flowers, belonging to 3 species, length and width of sepals and petals was measured in cm (Fisher's Iris Data)

First 10 obs

| Sepal L. | Sepal W. | Petal L. | Petal W. | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Se |
| 4.9 | 3.0 | 1.4 | 0.2 | Se |
| 4.7 | 3.2 | 1.3 | 0.2 | Se |
| 4.6 | 3.1 | 1.5 | 0.2 | Ve |
| 5.0 | 3.6 | 1.4 | 0.2 | Vi |
| 5.4 | 3.9 | 1.7 | 0.4 | Se |
| 4.6 | 3.4 | 1.4 | 0.3 | Vi |
| 5.0 | 3.4 | 1.5 | 0.2 | Se |
| 4.4 | 2.9 | 1.4 | 0.2 | Ve |
| 4.9 | 3.1 | 1.5 | 0.1 | Vi |

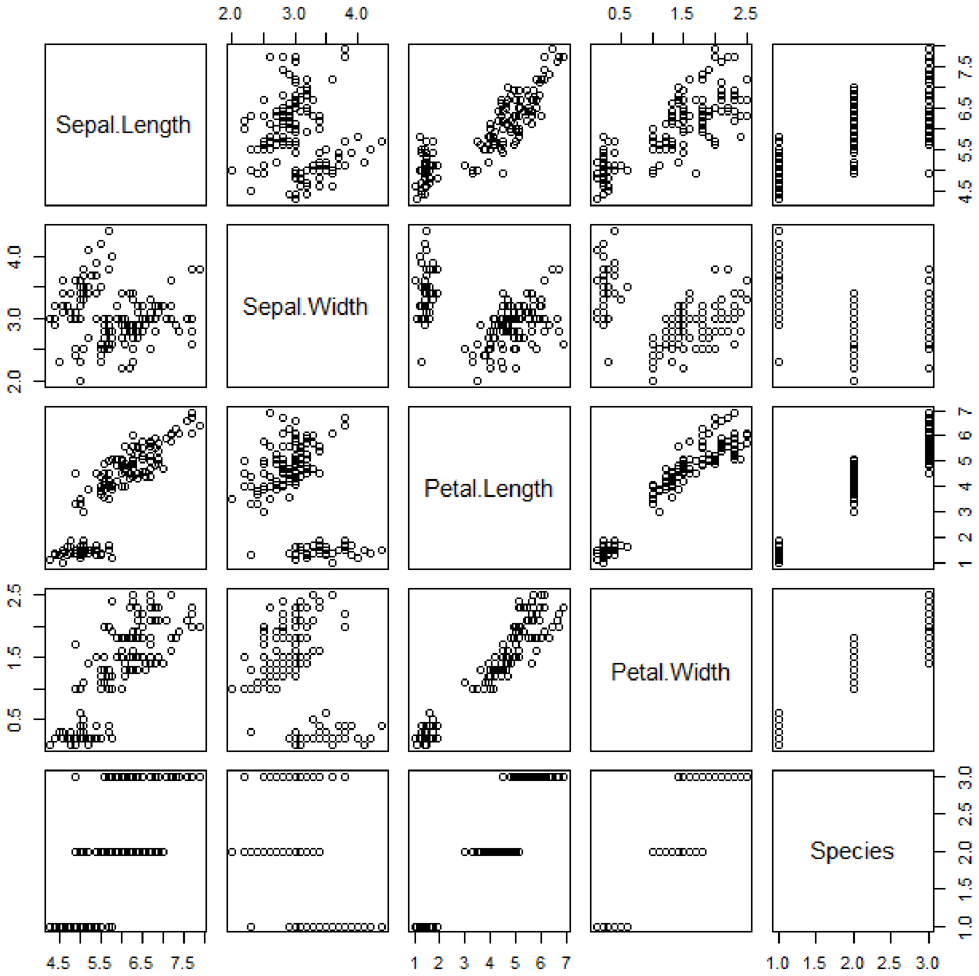
# Ordination

## Graphical exploratory analysis: pairs plot

## Objective: investigating interrelations between variables

**Ordination**

Searching for relations between var and between obs



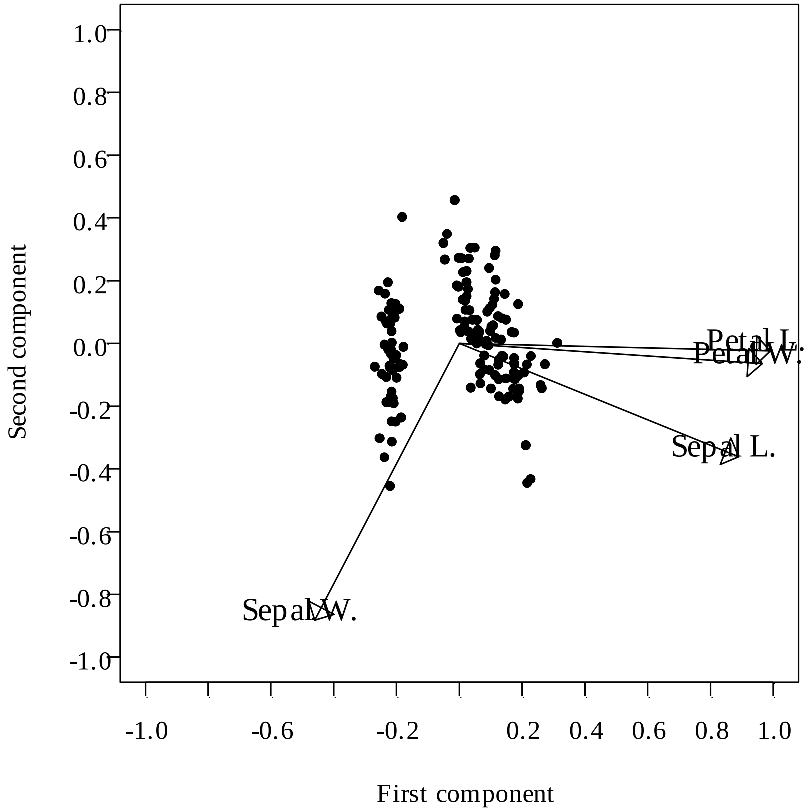
# Ordination

## Graphical exploratory analysis: biplot

## Objective: investigating interrelations between variables

### Ordination

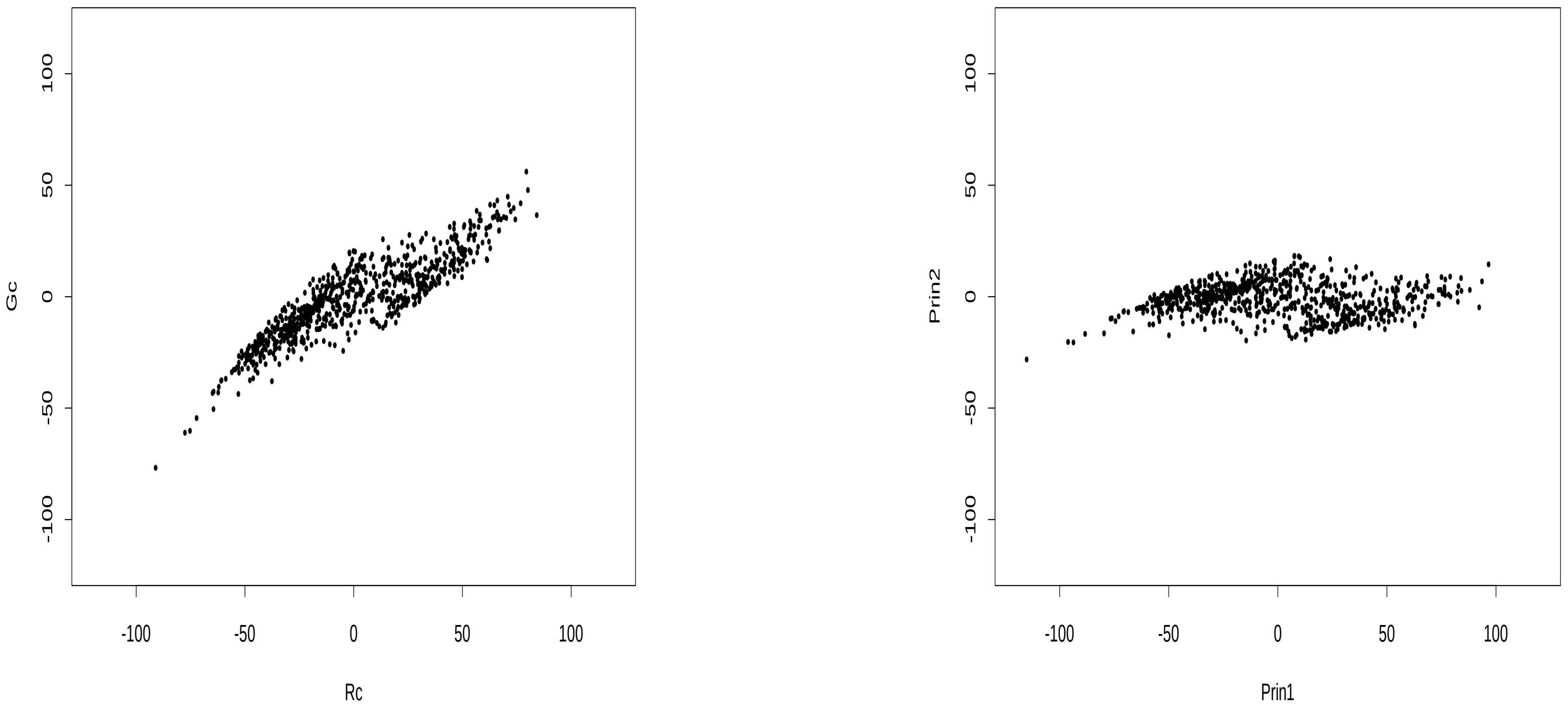
A procedure for adapting a multidimensional swarm of data points in such a way that when it is projected onto a lower-dimensional surface any **intrinsic pattern** possessed by the swarm will become apparent.



PCA Biplot on IRIS data

# Ordination model: PCA

## Objective: Dimension reduction



Rotation of the axis system in function of maximal explained variance in the new coordinate system

# Classification (clustering) model

## Objective: Discovering group structures

First 10 obs of Fischer's Iris data

| Sepal L. | Sepal W. | Petal L. | Petal W. |
|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 |
| 4.9 | 3.0 | 1.4 | 0.2 |
| 4.7 | 3.2 | 1.3 | 0.2 |
| 4.6 | 3.1 | 1.5 | 0.2 |
| 5.0 | 3.6 | 1.4 | 0.2 |
| 5.4 | 3.9 | 1.7 | 0.4 |
| 4.6 | 3.4 | 1.4 | 0.3 |
| 5.0 | 3.4 | 1.5 | 0.2 |
| 4.4 | 2.9 | 1.4 | 0.2 |
| 4.9 | 3.1 | 1.5 | 0.1 |

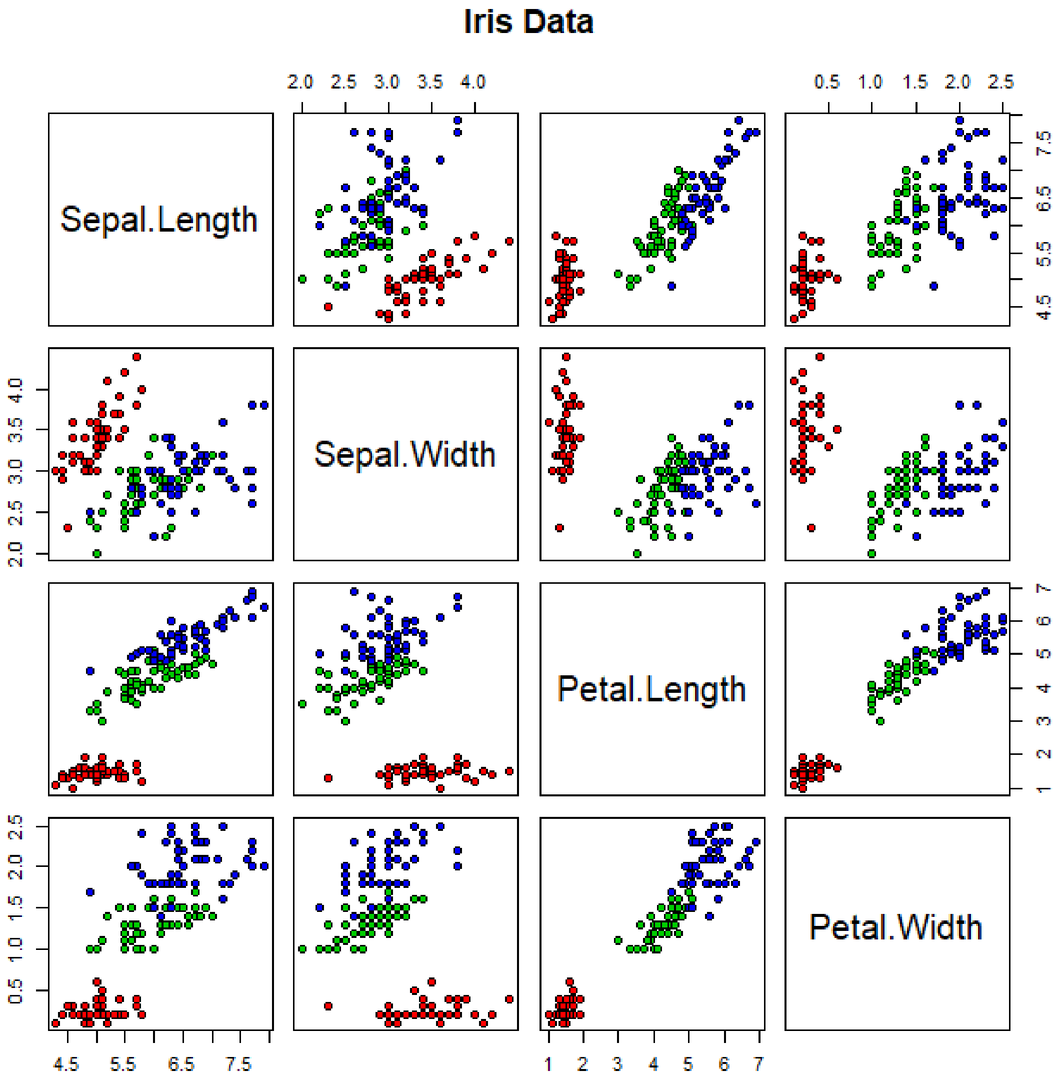
Group structure is NOT known

Investigate if an underlying group structure exists and can be revealed by the variables

How?
Calculate the matrix of inter-obs distances
Group obs that are 'close'

# Classification (clustering) model

## Objective: Discovering group structures

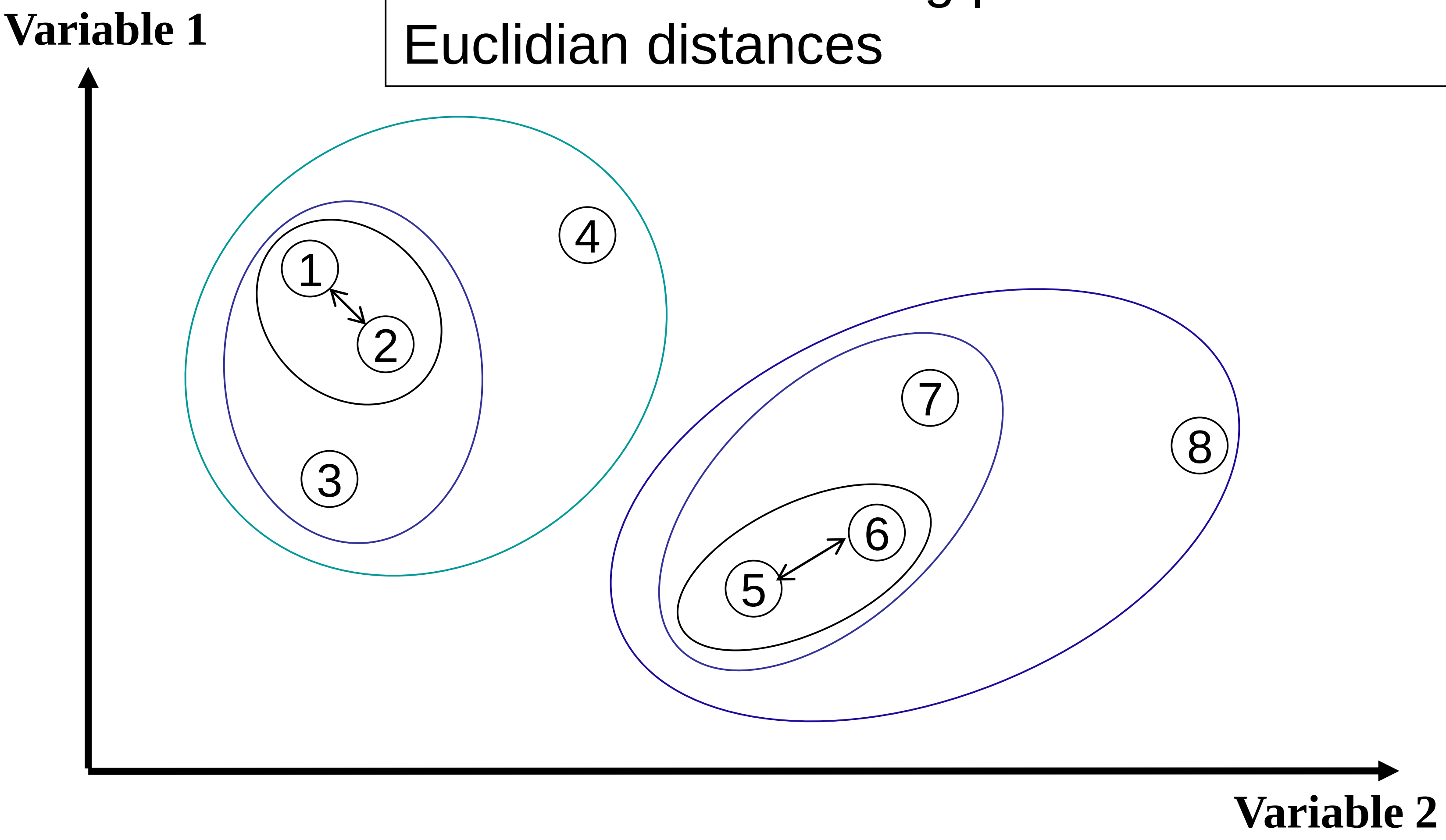


Group structure is NOT known

Investigate if an underlying group structure exists and can be revealed by the variables

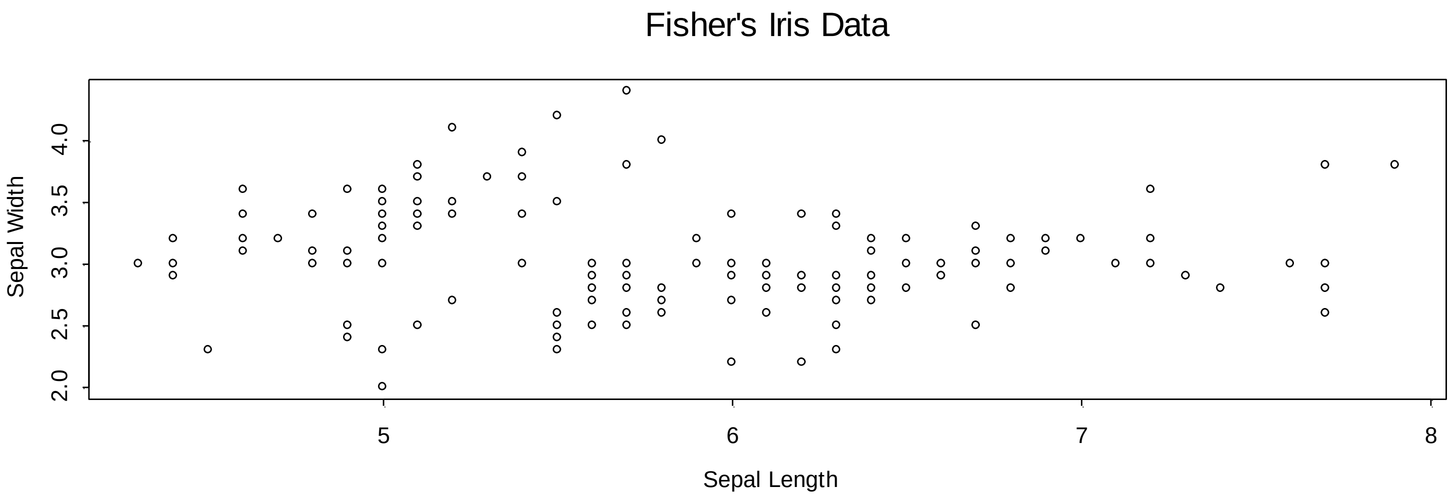
# Grouping or clustering

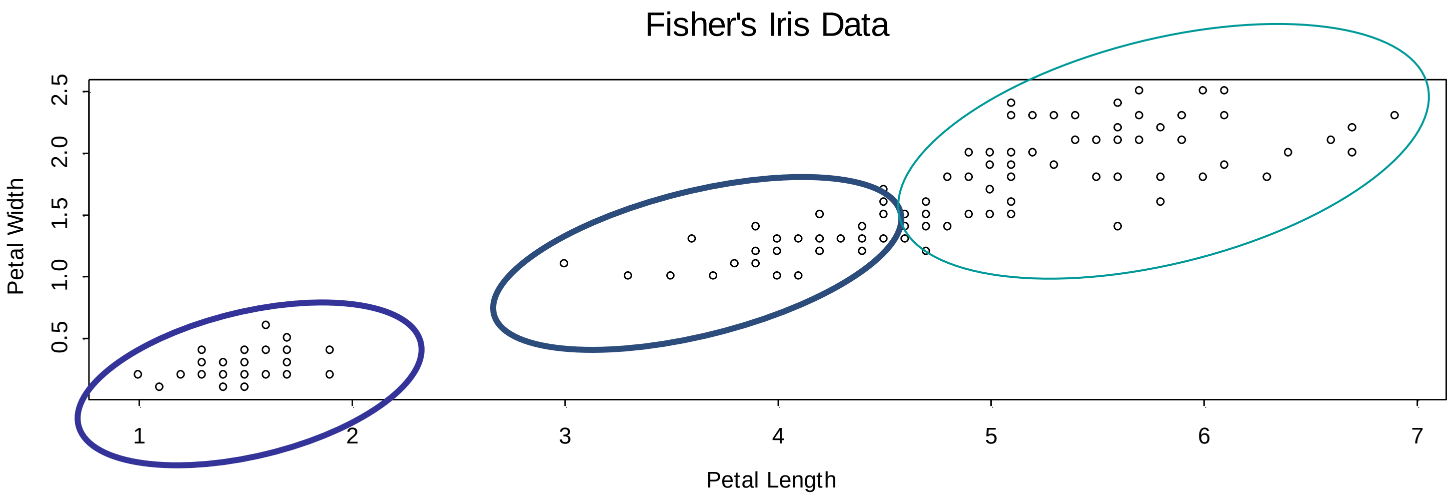
Hierarchical clustering procedure based on Euclidian distances



# Classification (clustering) model

## Discovering group structures





# Discriminant analysis: FLDA

## Objective: investigating existing group structures

| Sepal L. | Sepal W. | Petal L. | Petal W. | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Se |
| 4.9 | 3.0 | 1.4 | 0.2 | Se |
| 4.7 | 3.2 | 1.3 | 0.2 | Se |
| 4.6 | 3.1 | 1.5 | 0.2 | Ve |
| 5.0 | 3.6 | 1.4 | 0.2 | Vi |
| 5.4 | 3.9 | 1.7 | 0.4 | Se |
| 4.6 | 3.4 | 1.4 | 0.3 | Vi |
| 5.0 | 3.4 | 1.5 | 0.2 | Se |
| 4.4 | 2.9 | 1.4 | 0.2 | Ve |
| 4.9 | 3.1 | 1.5 | 0.1 | Vi |

FDA assumes underlying probability distribution

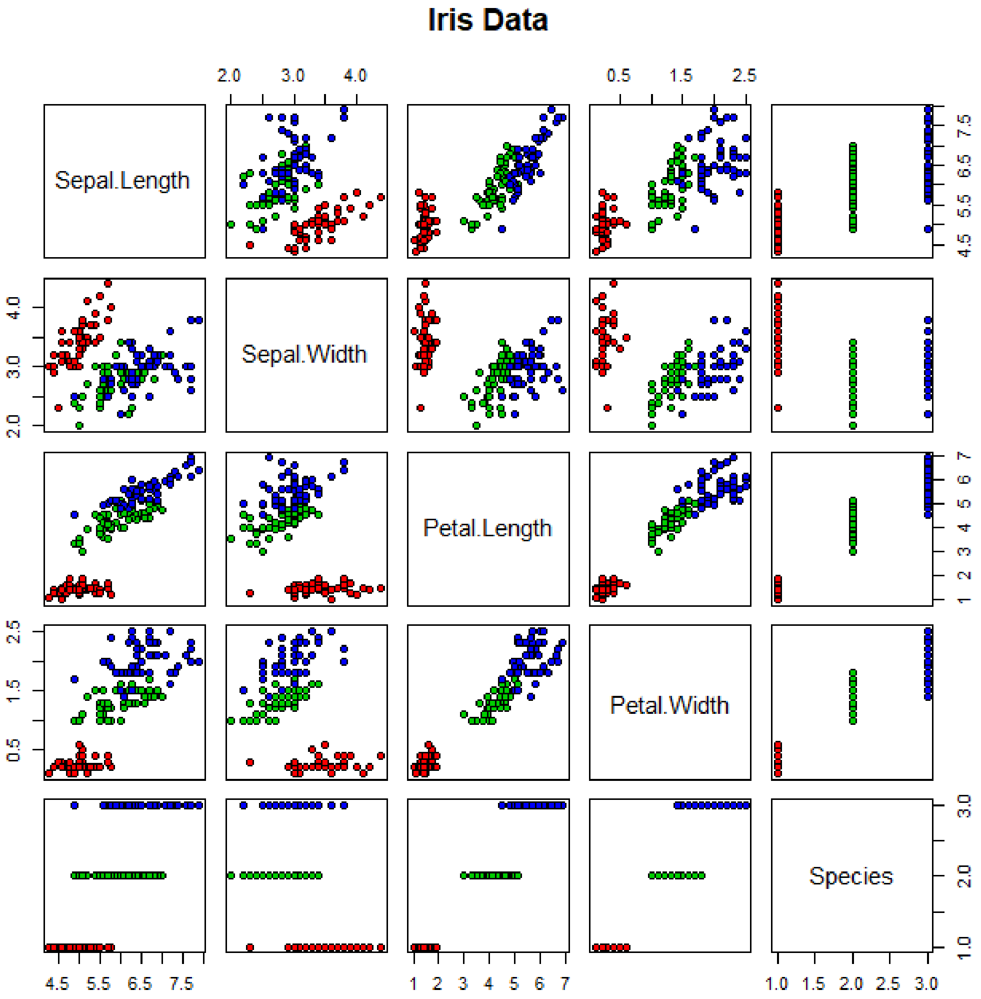
Parametric

Group structure is known

Investigate if the existing and known group structure can be revealed by some discriminating function of the variables

# Discriminant analysis: FLDA

## Objective: investigating existing group structures



Group structure is known

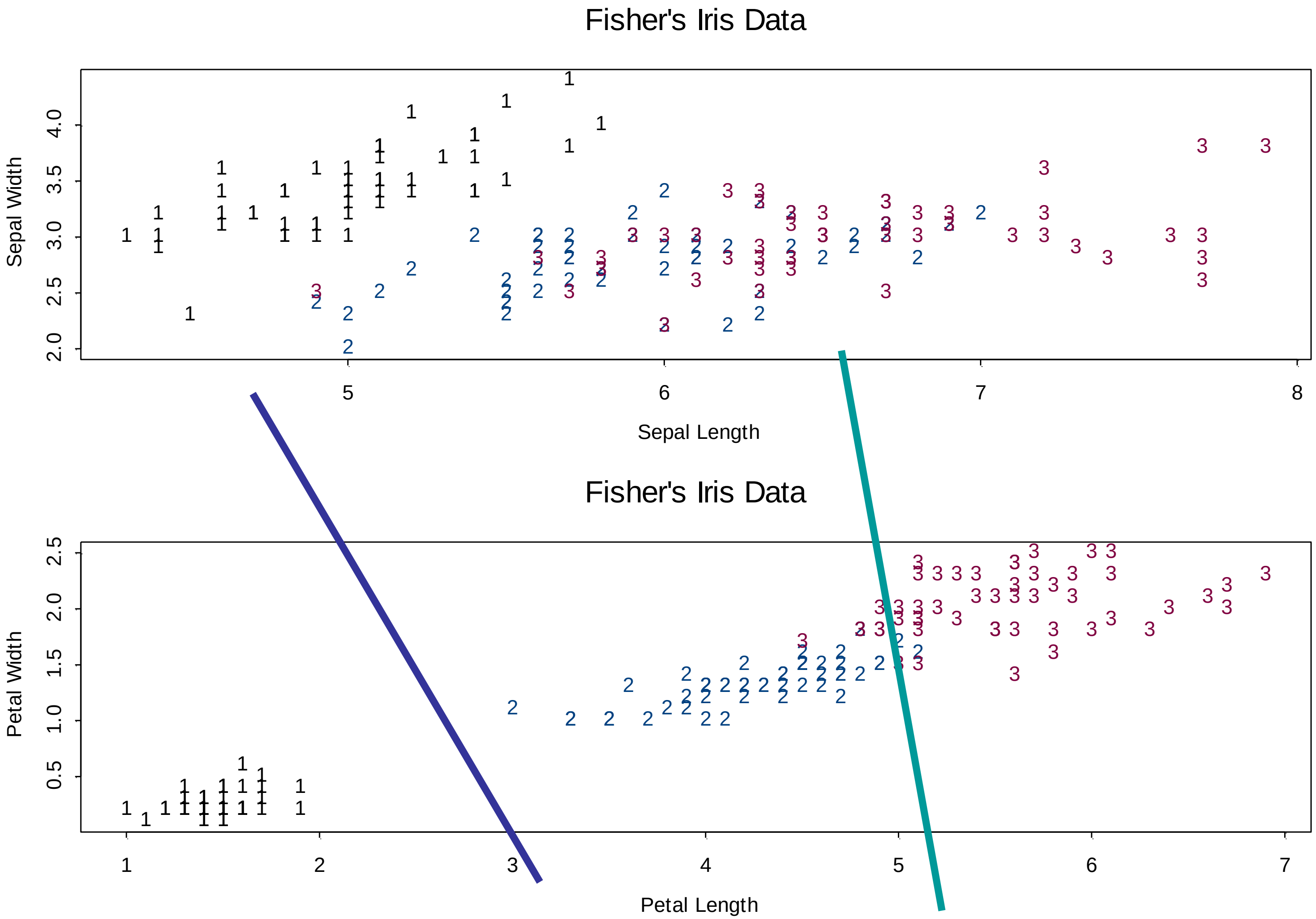
Investigate if the existing and known group structure can be revealed by some discriminating function of the variables

FDA assumes underlying probability distribution

Parametric

# Discriminant analysis

## Objective: investigating existing group structures



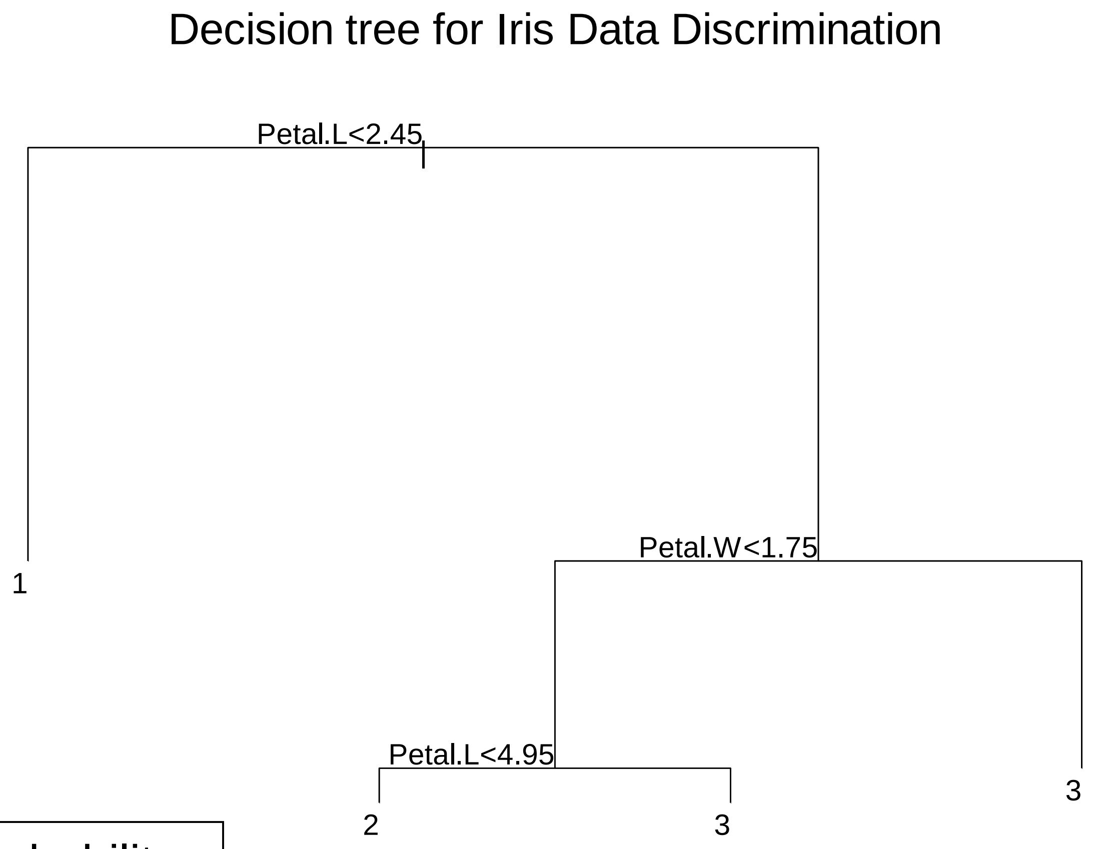
Fisher's Iris Data

Fisher's Iris Data

# Discriminant analysis

## Objective: investigating existing group structures

## Tree beased modelling (TBM)



No underlying probability distribution is assumed

Non-parametric

# Concluding remark

This introductory course concentrates on the first principles of applied exploratory multivariate data analysis:

- ordination (data reduction, relations between var and between obs)
- discrimination, supervised classification (predict group membership)
- clustering, unsupervised classification (search grouping structures)

For each of these first principles several methodologies, based on different algorithmic approaches, are elaborated

The emphasis lays on understanding the underlying mathematics as far as necessary to understand the methodologies

The general idea is not to give an overview of all possible data science methodologies

Being able to develop/adapt your own R-code for the different methodologies is most essential