Data processing and *Staphylococcus aureus*

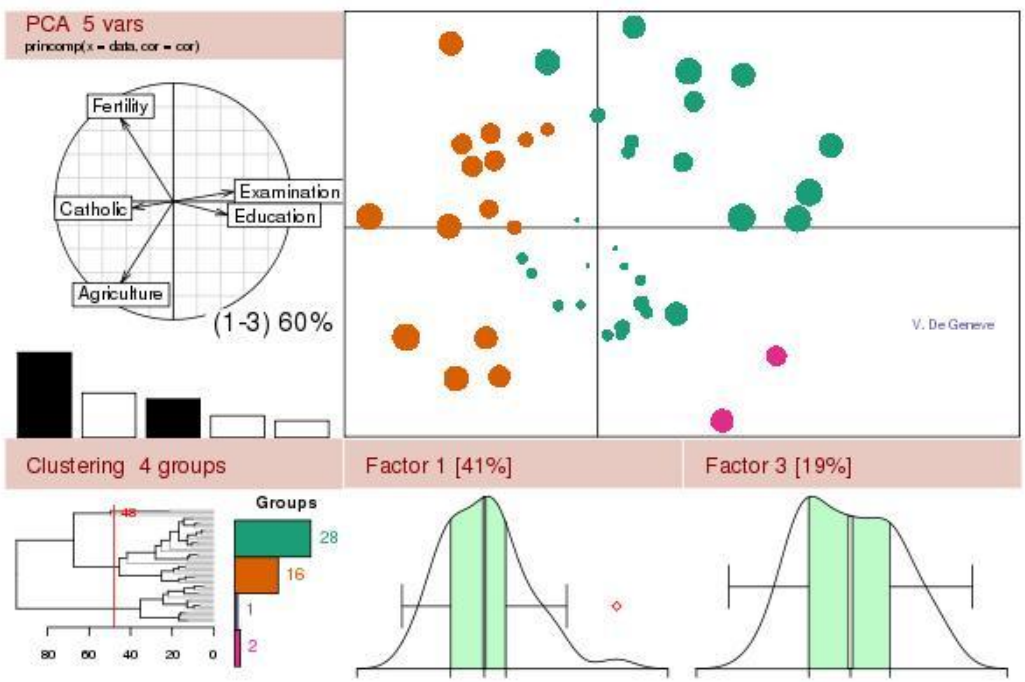
Ruslan Akulenko
Center for Bioinformatics

Outline:

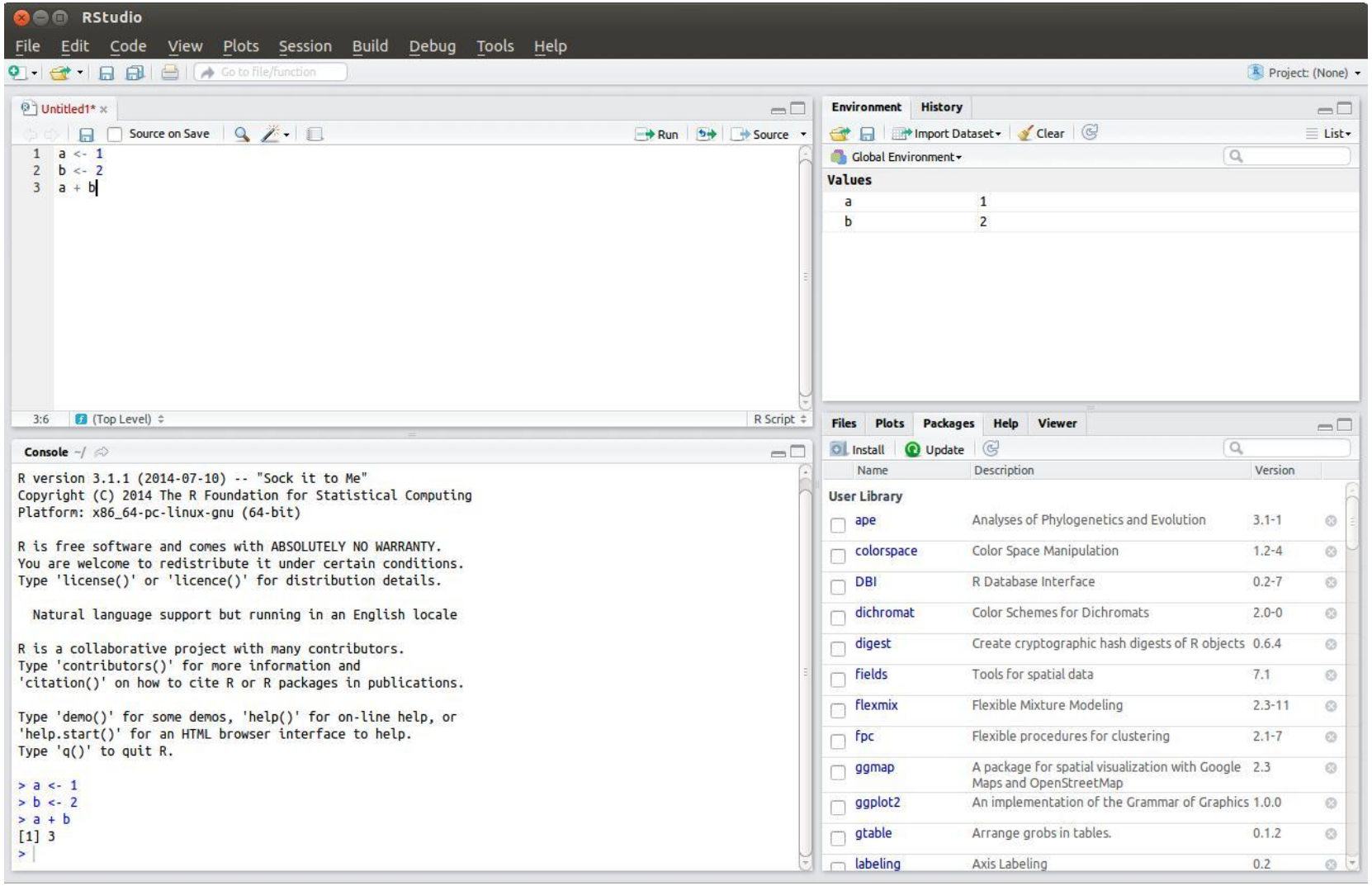
- Introduction:
  - R – free software environment for statistical computing and graphics;
  - RStudio - Integrated Development Environment for R;
  - R packages and basic functions;
- Data preprocessing:
  - Data and it's peculiarities;
  - Completing missing values;
- Data analysis:
  - T-test;
  - Kolmogorov – Smirnov test;
  - Wilcoxon signed – rank test.

R - simplifies your data analysis

- Performs data manipulation;
- All types of calculations;
- Optimize or automatize calculations using loops;
- Wide range of statistical tests;
- Data visualization;
- R can answer the question – what do I find in my data?



RStudio – makes use of R easier



R basics. Packages

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) #logical vector

d <- c(1,2,3,4)
e <- c("red", "white", "red", NA)
f <- c(TRUE,TRUE,TRUE,FALSE)
mydata <- data.frame(d,e,f)
names(mydata) <- c("ID","Color","Passed") # variable names
```

```
value1 <- 2
value2 <- 3
if (value1 < value2)
{
  print('value1 is less than value2')
} else {
  print('value1 is greater than value2')
}
```

```
for (i in 1:10)
{
  print(i)
}
```

```
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9
[1] 10
```

	ID	Color	Passed
1	1	red	TRUE
2	2	white	TRUE
3	3	red	TRUE
4	4	NA	FALSE

```
> value1 <- 2
> value2 <- 3
> if (value1 < value2)
+ {
+   print('value1 is less than value2')
+ } else {
+   print('value1 is greater than value2')
+ }
[1] "value1 is less than value2"
```

User Library

☐	apcluster	Affinity Propagation Clustering	1.3.5
☐	ape	Analyses of Phylogenetics and Evolution	3.1-1
☐	colorspace	Color Space Manipulation	1.2-4
☐	DBI	R Database Interface	0.2-7
☐	dichromat	Color Schemes for Dichromats	2.0-0
☐	digest	Create cryptographic hash digests of R objects	0.6.4
☐	fields	Tools for spatial data	7.1
☐	flexmix	Flexible Mixture Modeling	2.3-11
☐	fpc	Flexible procedures for clustering	2.1-7

Outline:

- Introduction:
  - R – free software environment for statistical computing and graphics;
  - RStudio - Integrated Development Environment for R;
  - R packages and basic functions;
- Data preprocessing:
  - Data and it's peculiarities;
  - Completing missing values;
- Data analysis:
  - T-test;
  - Kolmogorov – Smirnov test;
  - Wilcoxon signed – rank test.

Input data

StaphyType Test Report

Operator	
Sample ID	2192119
Experiment ID	2192119 - {4083AD2C-7D42-4FB9-82D5-E50CC0FD6206}
Date of Result	Thu Apr 14 10:46:01 2011
Assay Name	StaphyType
Assay ID	10248
Well Position	01 (01-A)
Software Version	2009-07-09
Device	04a0022

Internal Controls

Data Quality	passed

Genetic markers for S. aureus / MRSA / PVL

Taxonomy	Species Marker (*S.aureus*) **positive**
MRSA (mecA)	**positive**
PVL	negative

Resistance Genotype

Hybridisation (Gene)	Result	Expected Resistance
mecA	**positive**	Methicillin, Oxacillin and all Beta-Lactams, defining **MRSA**
blaZ	negative	Beta-Laktamase
ermA	**positive**	Macrolide, Lincosamide, Streptogramin
ermB	negative	Macrolide, Lincosamide, Streptogramin
ermC	negative	Macrolide, Lincosamide, Streptogramin
linA	negative	Lincosamides

	11	46	10	33	28
MRSA (mecA)	0	0	0	0	0
PVL	0	0	0	0	0
23S-rRNA	1	1	1	1	1
gapA	1	1	1	1	1
katA	1	1	1	1	1
coA	1	0	1	1	1
Protein A	1	1	1	1	1
sbi	1	1	1	1	1
nuc	1	1	1	1	1
fnbA	1	1	1	1	1
vraS	1	1	1	1	1
sarA	1	1	1	1	1
eno	1	1	1	1	1
saeS	1	1	1	1	1
mecA	0	0	0	0	0
blaZ	0	1	0	0	0
blaI	0	1	0	0	0
blaR	0	1	0	0	0
ermA	0	0	0	0	0
ermB	0	0	0	0	0
ermC	0	0	0	0	0
linA	0	0	0	0	0

Completing missing values

Task – predict ambiguous values. Methods tested:
Baseline prediction using average values

total average

$$\mu = \frac{1}{N} \sum_{(i,j)\in\Omega} D_{ij}$$

sample average

$$b_i = \frac{1}{N_i} \sum_{(i,j)\in\Omega} D_{ij} - \mu$$

gene average

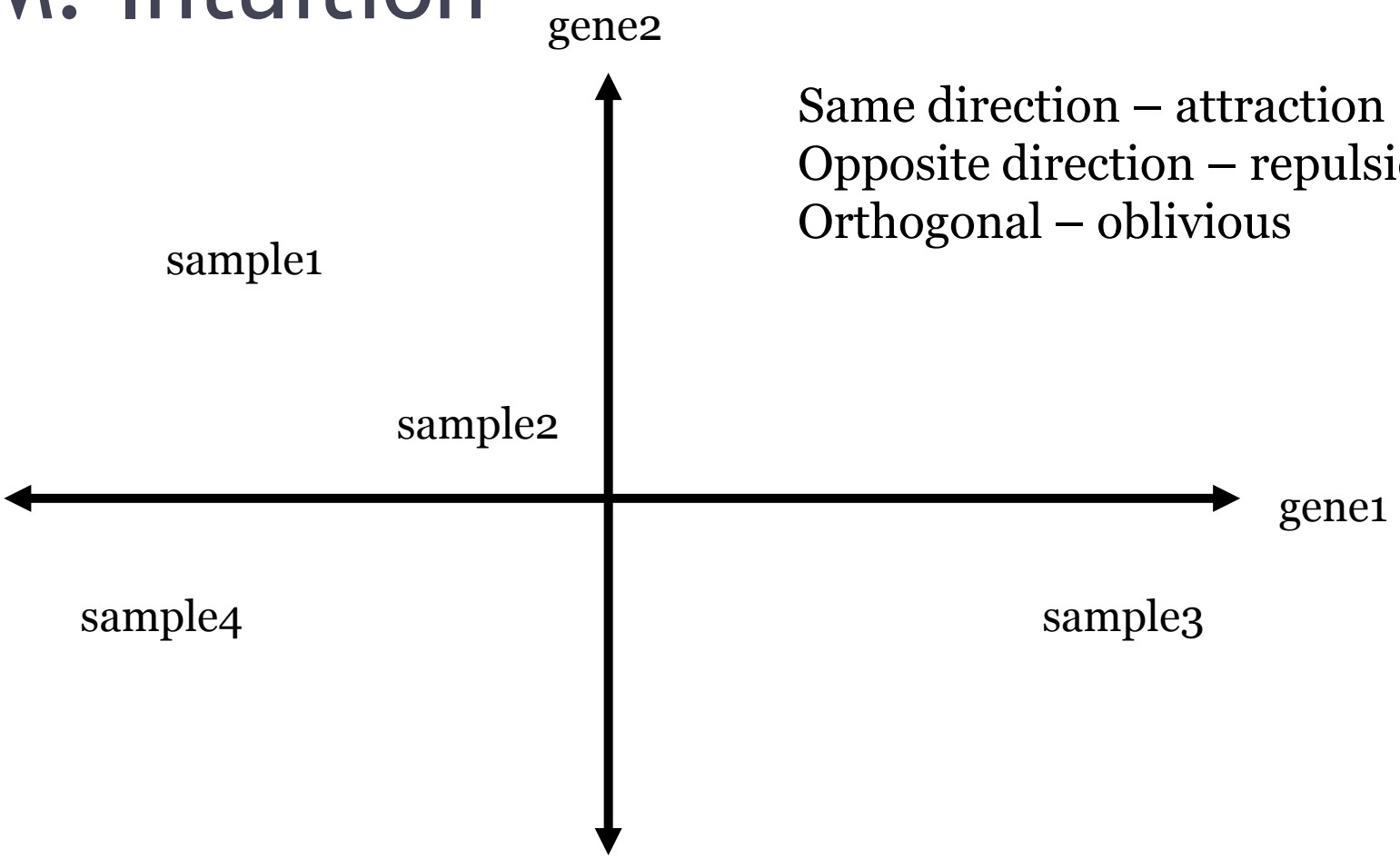
$$b_j = \frac{1}{N_j} \sum_{(i,j)\in\Omega} D_{ij} - \mu$$

$$b_{prediced_{ij}} = \mu + b_i + b_j$$

~85% Correct predictions

Latent Factor Models (LFM)
~95% Correct predictions

LFM: Intuition



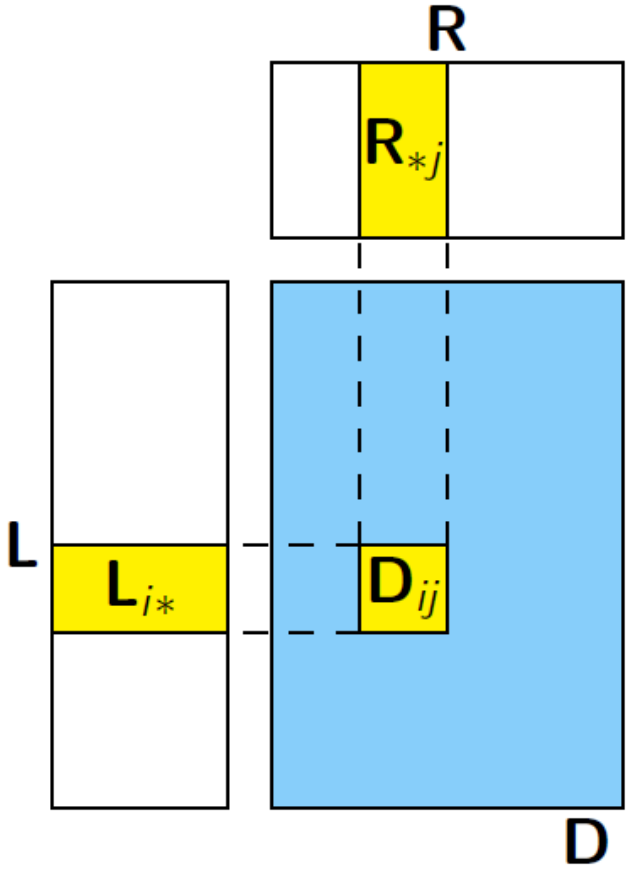
Same direction – attraction
Opposite direction – repulsion
Orthogonal – oblivious

LFM: mathematical background

$$L = \sum_{(i,j)\in\Omega} \left(D_{ij} - [LR]_{ij}\right)^2 + \lambda\left(\|L\|_F^2 + \|R\|_F^2\right)$$

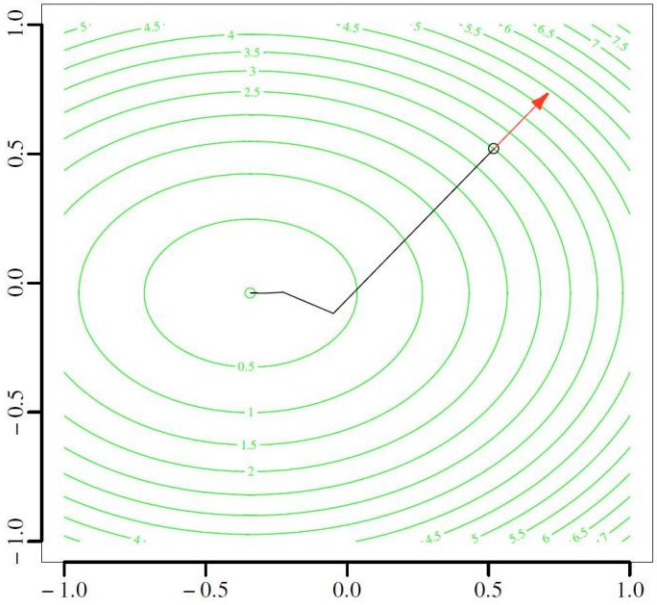
L (m × r) and R (r × n) are sought matrices of rank r

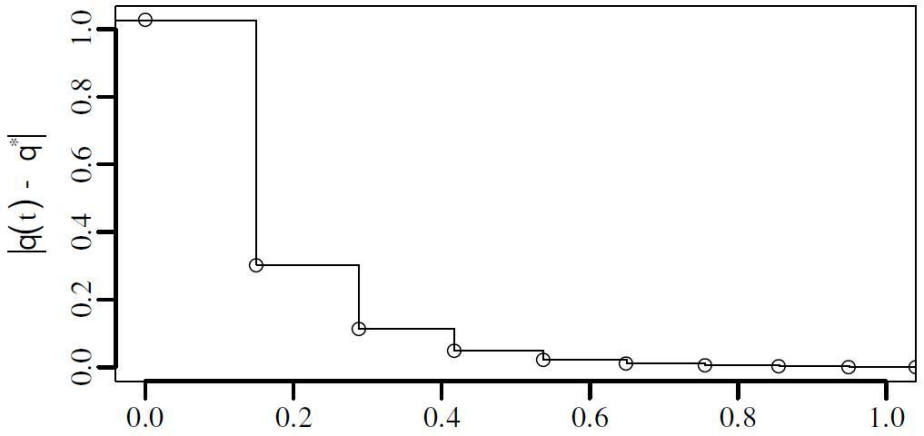
D (m × n) is a given matrix



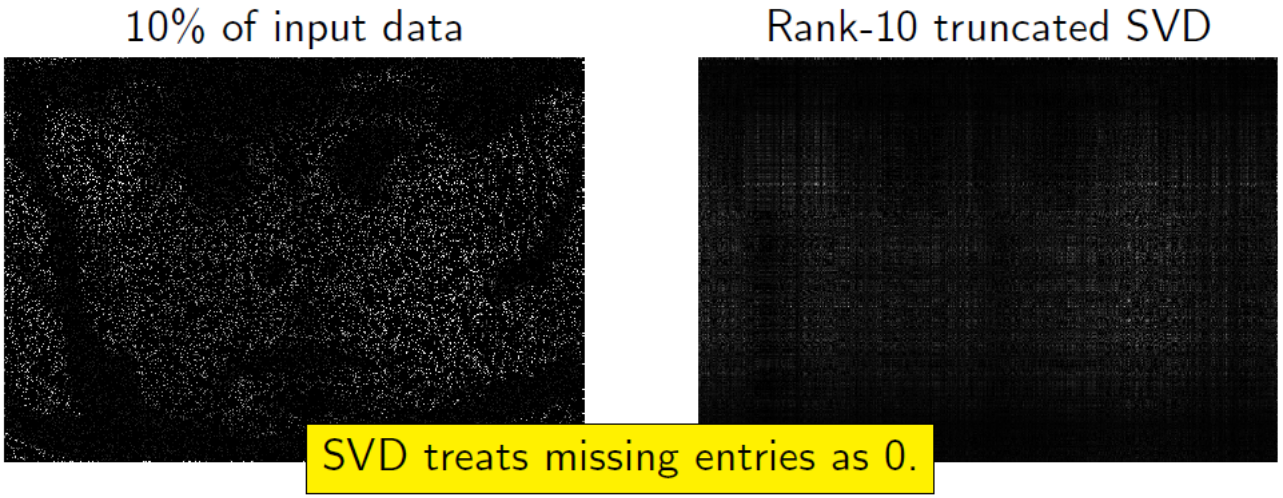
LFM: Discrete Gradient Descent

- Pick a starting point;
- Compute gradient;
- Update parameters L and R
- Repeat N times.





LFM: one more benefit

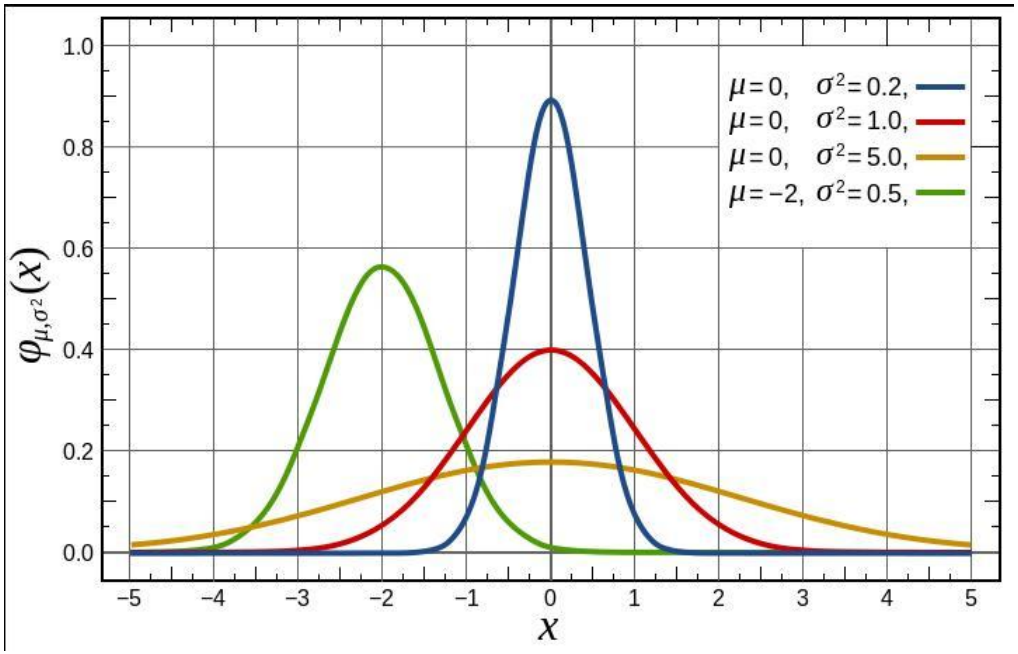


Outline:

- Introduction:
  - R – free software environment for statistical computing and graphics;
  - RStudio - Integrated Development Environment for R;
  - R packages and basic functions;
- Data preprocessing:
  - Data and it's peculiarities;
  - Completing missing values;
- Data analysis:
  - T-test;
  - Kolmogorov – Smirnov test;
  - Wilcoxon signed – rank test.

Student's t-test

Is statistical hypothesis test which is used to determine if two sets of data are significantly different from each other.

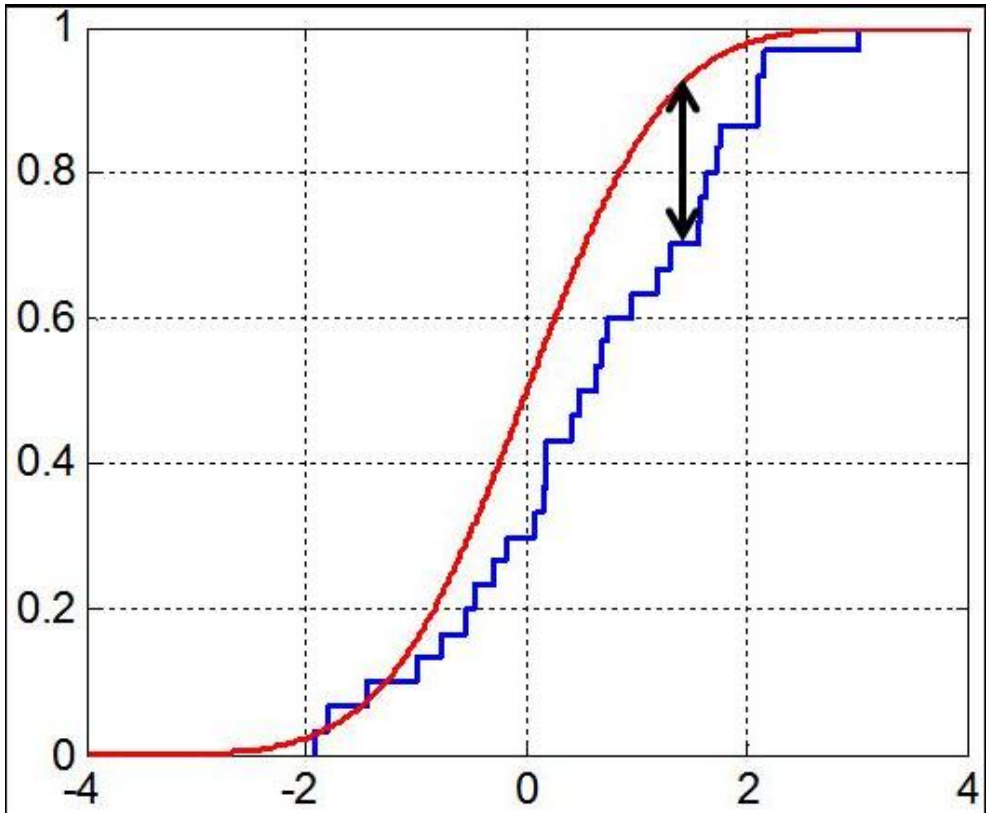



```
> t.test(1:10, y = c(7:20))

 Welch Two Sample t-test

data:  1:10 and c(7:20)
t = -5.4349, df = 21.982, p-value = 1.855e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.052802  -4.947198
sample estimates:
mean of x mean of y
      5.5      13.5
```

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

Kolmogorov – Smirnov test

Compares one – dimensional probability distribution of sample(s).




```
> ks.test(rnorm(50), runif(30))

 Two-sample Kolmogorov-Smirnov test

data:  rnorm(50) and runif(30)
D = 0.4467, p-value = 0.0007174
alternative hypothesis: two-sided
```

```
ks.test(x, y, ...,
        alternative = c("two.sided", "less", "greater"),
        exact = NULL)
```

Wilcoxon signed – rank Test

Is alternative to the paired Student's t-test. Population does not need to be normally distributed.

Assumptions:

1. Data are paired and come from the same population.
2. Each pair is chosen randomly and independently.
3. The data are measured at least on an ordinal scale.

```
> wilcox.test(x, y, paired = TRUE, alternative = "greater")

 Wilcoxon signed rank test

data:  x and y
V = 40, p-value = 0.01953
alternative hypothesis: true location shift is greater than 0
```

```
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95, ...)
```

Summary

- R and RStudio contain many features to analyze data.
- LFM was able to predict successfully 95% of missing values.
- Wilcoxon, T-test, KS-test can answer the question – Are these groups of samples different?

Thank you for your attention!

Questions?