

B4M36SAN
BE4M36SAN
Seminar/Tutorial

Alikhan Anuarbekov

Ali

anuarali@fel.cvut.cz

cw.fel.cvut.cz

Total: 100 points



Seminar: 50 points + 5 bonus



During 13 weeks of semester,
programming, theory, project

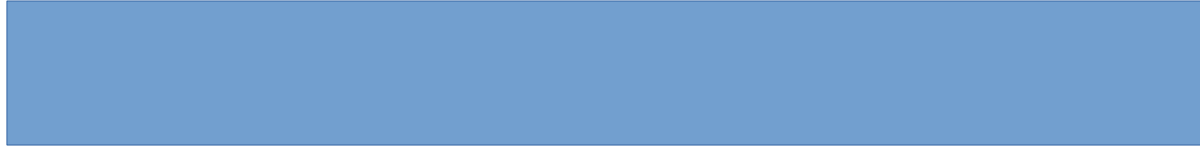


Exam: 50 points

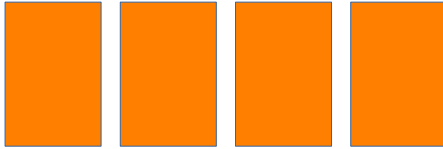


At the end of the course, 1 big exam
theory from lectures only

Seminar: 50 points + 5 bonus



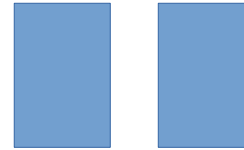
10 points for theory test (midterm)
8.week



4 homeworks, each for 5 points



15 points for big team project



5 + 5 bonus points
for seminar activity

25 points to pass



4 / 5 assignments for more than zero points

Every seminar (a total of 10 seminars):

15-25 minutes theoretical review -----> not covers all of lecture!

5-15 minutes assignment presentation

50-70 minutes for activity assignment solution

Attendance is obligatory, max. 2 missing tutorials

BRUTE -> solved and submitted till the end of day, earn 1 point (0.5 point if partially)

Homework for 5 points – parallel to activity assignment, typically 2-3 weeks to solve
<1 week after deadline = 50% of points, 1+ week after deadline = 0 points

Big team project (Final assignment section in courseware)

- Start looking for 4 people team **NOW**
- Try to find a question/dataset that is interesting for you
- Look for previous year best projects for inspiration
- Ask question before the deadline for team creation (3.11)
- If no team on 3.11 (midterm week), then I will forcefully assign a team for you

Tips and recommendations:

- Do not focus on single question, try to combine two or more questions
 - Example: Not only look for medical question, but try to add socio-economy
- Try to find more than one dataset
 - Example: UNICEF + World Bank
- Start working early, at least to formulate a question

Questions?

Rstudio vs Python

- In this tutorials I will go with R/Rstudio programming language
 - Has all math and formulas inside, just call functions
 - Syntax is the similar to Matlab/Javascript/....
 - If you want, you could go with Python
 - I will not check you homework, Jan Blaha will, later feedback
 - I cannot help you with python questions
 - But python is much more easier and familiar :)
 - R is used only by statisticians, python by everyone
- Type in google: **Rstudio download** and **R download**

<https://cran.r-project.org/>

R (programming lang)
e.g. same as Python

<https://posit.co/download/rstudio-desktop/>

RStudio (editor, IDE)
e.g. same as IntelliJ/PyCharm

R_version - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins R_version

workflow.Rmd pvalue.cpp dfs.in\$circInter list.in\$`circInter`

Knit on Save Knit Run

Source Visual Outline

```
512 return(list(pvalue=list.pval.out,
513             tpr=pvals1.tpr, fpr=pvals1.fpr,
514             precision=pvals1.prec, recall=pvals1.recall,
515             f1=pvals1.f1))
516 }
517
518 list.to.pr <- function(pvals.list.in){
519
520   pvals.list.in$pvalue[is.na(pvals.list.in$pvalue)] <- 1
521   as.data.frame(pvals.list.in) %>%
522     mutate(correct=as.factor(correct), score=-log(pvalue)/50) %>%
523     pr_curve(correct, score)
524 }
```

575:1 Chunk 9 R Markdown

Console Terminal Background Jobs

```
R 4.1.2 · ~/Desktop/work/R_version/
+ }
> rcPlot(list(
+   "GEO"=all.geo.pvals,
+   "circInter"=all.inter.pvals,
+   "rand"=all.rand.pvals
+ ))
Called from: rcPlot(list(GEO = all.geo.pvals, circInter = all.inter.pvals,
+   rand = all.rand.pvals))
Browse[1]> c
> install.packages("BSDA")
```

Environment History Connections Tutorial

Import Dataset 1.38 GiB List

R Global Environment

Files Plots Packages Help Viewer Presentation

Home > Desktop > work > R_version > source > GPACDA

	Name	Size	Modified
☐	downloadMiRNAGO.R	801 B	Nov 10, 202
☐	downloadMultiMiR.R	1.3 KB	Nov 10, 202
☐	gene2disease.csv	118.4 KB	Nov 10, 202
☐	goterms.R	906 B	Nov 10, 202
☐	goterms.RData	4.3 MB	Nov 10, 202
☐	mirna-go-terms.csv	84.3 KB	Nov 10, 202
☐	mirna-mrna-interactions.csv	15 MB	Nov 10, 202
☐	mirnaGO.R	1.4 KB	Nov 10, 202
☐	pvalue.cpp	5.9 KB	Nov 10, 202
☐	README.txt	5.2 KB	Nov 10, 202
☐	run.R	1.2 KB	Nov 10, 202
☐	run.sh	207 B	Nov 10, 202
☐	runDisease.sh	154 B	Nov 10, 202
☐	diseasedata		
☐	diseases-small		
☐	GPACDA-paper-experiments		
☐	result2024		

Rstudio/R basic programming

LOGICAL -> INTEGER -> NUMERIC -> CHARACTER

“boolean” “integer” “double” “string”

a = 1

class(a) = type of variable

vector = array of same type

....

https://www.w3schools.com/r/r_get_started.asp

<https://www.datacamp.com/courses/free-introduction-to-r>

Download today's tutorial script (just click, no programming):

Tutorials

T	Date	Teacher	Contents	Materials
1	23.9.	JB, AA, JK	Statistical testing, t-test, significance, power of the test.	 san_intro.zip

Rstudio/R library installation + problems

> I want to install a library that will compute z-test



R how to do z test



<https://www.geeksforgeeks.org/z-test-in-r/>

```
library(BSDA)
```

```
> install.packages("BSDA")
```

* DONE (BSDA)

Successful!

```
Warning message:
```

```
In install.packages("curl") :
```

```
installation of package 'curl' had non-zero exit status
```

Error!

```
----- [ERROR MESSAGE] -----  
<stdin>:1:10: fatal error: curl/curl.h: No such file or directory  
compilation terminated.  
-----
```

Reason (read above)!

```
Using PKG_CFLAGS=
Using PKG_LIBS=-lcurl
----- [ANTICONF] -----
Configuration failed because libcurl was not found. Try installing:
 * deb: libcurl4-openssl-dev (Debian, Ubuntu, etc)
 * rpm: libcurl-devel (Fedora, CentOS, RHEL)
If libcurl is already installed, check that 'pkg-config' is in your
PATH and PKG_CONFIG_PATH contains a libcurl.pc file. If pkg-config
is unavailable you can set INCLUDE_DIR and LIB_DIR manually via:
R CMD INSTALL --configure-vars='INCLUDE_DIR=... LIB_DIR=...'
----- [ERROR MESSAGE] -----
<stdin>:1:10: fatal error: curl/curl.h: No such file or directory
compilation terminated.
-----
ERROR: configuration failed for package 'curl'
* removing '/usr/local/lib/R/site-library/curl'

The downloaded source packages are in
      '/tmp/RtmpNNghX1/downloaded_packages'
Warning message:
In install.packages("curl") :
  installation of package 'curl' had non-zero exit status
  
```

Rstudio/R library installation solution:



R library install fatal error: curl/curl.h: No such file or directory



9 Answers

Sorted by: Hi



~~sudo apt-get install curl-devel~~

214

```
sudo apt-get install libcurl-dev
```



(will install the default alternative)



OR



```
sudo apt-get install libcurl4-openssl-dev
```



Questions?

Statistics 101 (basics)

Prerequisites are in courseware (**Exam** section):

6. Sample questions pertaining to prerequisites:

-  [stat_min.pdf](#).

Today we will start with basics and go over it fast

If you don't understand anything from PDF, do a self study!

What is statistics?

- A typical scientific approach was to find the cause with the effect
- If we find all causes, we can 100% predict the effect
- A topic for math, algebra, physics, chemistry,

Cause

Why something happens.

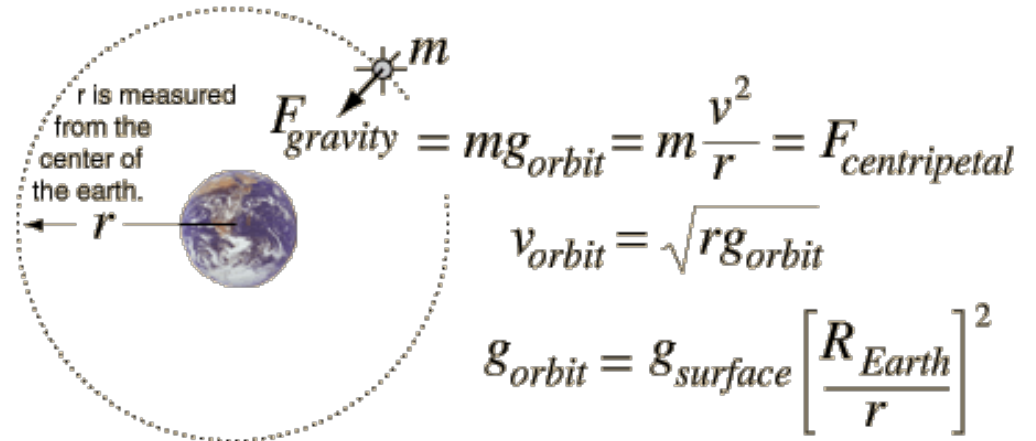
For example:
John was unwell, so he went to the hospital.



Effect

What happens as a result.

For example:
I am very hungry because I didn't eat breakfast.



What is statistics?

In complex systems/objects, we cannot physically predict the effect with 100% accuracy

Laplace's demon principle

*"If you were to measure all atom's position and speed,
you can predict everything in future and
know everything in past"*

> but no one can measure all atoms in universe



Practical example:

Can you with 100% accuracy predict whether the drug will work on patient?

Human body: 3.2 billion DNA bases, 30 trillion cells -> any DNA problem, cancer cells, ...

Surroundings: around 20 – 30 people interacts with you daily, you measure their bodies as well

+ Human's diet, behavior (smoking, sport,...), location (tropical disease?),.....

Bayesian statistics

Instead of 100% accurate answer give the most probable one

> yes, if you have a 1 / 1,000,00 chance rare genetic disease, our drug will harm you,
but in 999,999 / 1,000,000 cases it will help

Based on that principle:

Reality: complicated mechanism/equation
that is impossible to calculate

$$\begin{aligned} & \rho_t C_{p_t} \frac{\delta T_t}{\delta t} \\ = k_t & \left[\left(\frac{\delta^2 T_t}{\delta x^2} \right) + \left(\frac{\delta^2 T_t}{\delta y^2} \right) + \left(\frac{\delta^2 T_t}{\delta z^2} \right) \right] \\ & + \rho_b C_{p_b} \dot{V} (T_A - T_V) + \dot{Q}_m \end{aligned}$$

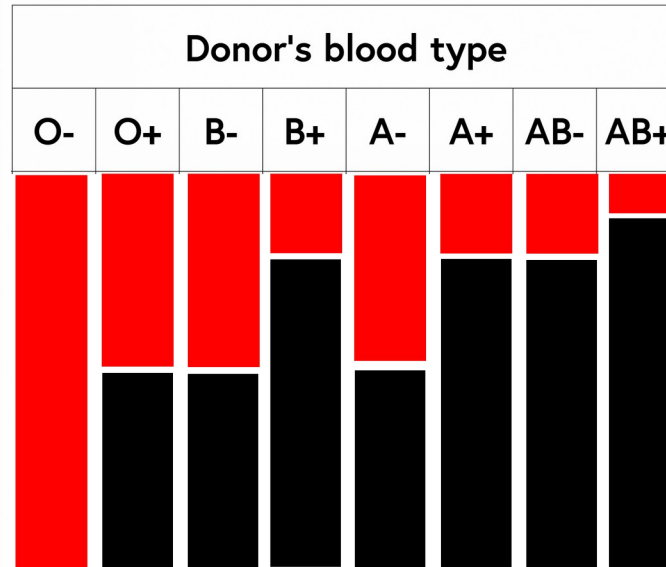


$$\begin{aligned} \text{disease chance} & \approx \text{age} * 0.5 \\ \text{disease chance} & = \text{age} * 0.5 + \Delta_{\text{error}} \end{aligned}$$

Example (2D discrete space -> 1D distribution):

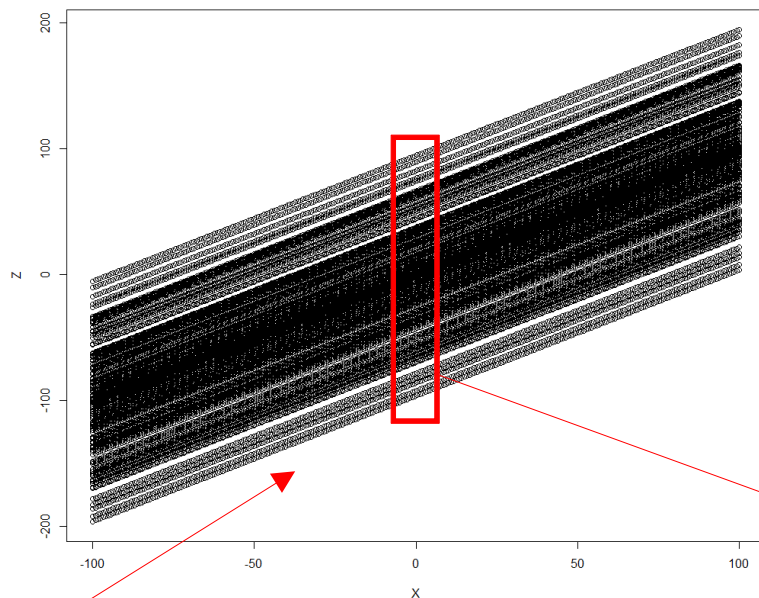
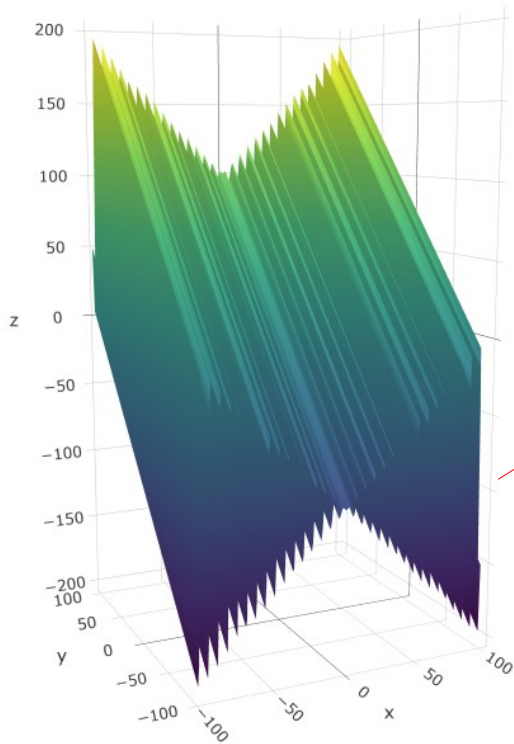
You need to performed a blood transfusion, but you don't have time to measure patient blood type. You only know donor's blood type. Then you have an uncertainty about transfusion success

		Donor's blood type							
		O-	O+	B-	B+	A-	A+	AB-	AB+
Recipient's blood type	AB+	✓	✓	✓	✓	✓	✓	✓	✓
	AB-	✓		✓		✓		✓	
	A+	✓	✓			✓	✓		
	A-	✓				✓			
	B+	✓	✓	✓	✓				
	B-	✓		✓					
	O+	✓	✓						
	O-	✓							



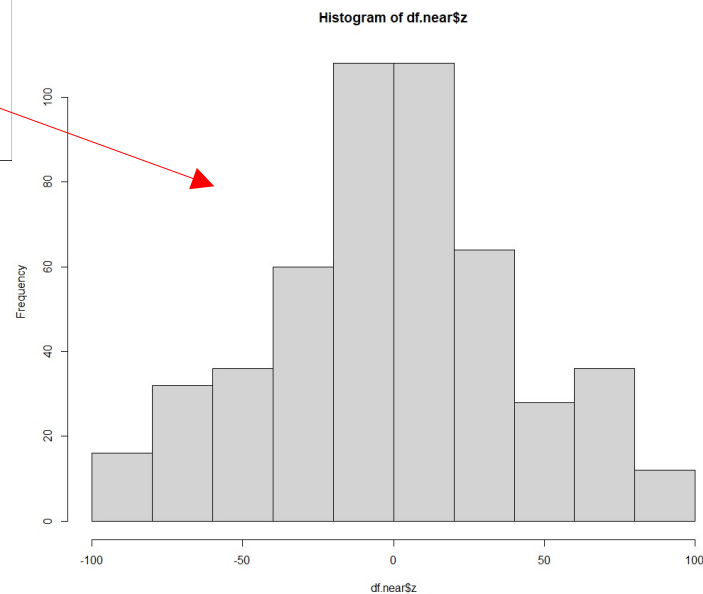
Example (2D continuous space -> 1D distribution):

$$Z = X + Y * \sin(Y)$$



Hide Y axis

Look at some X-Z range:

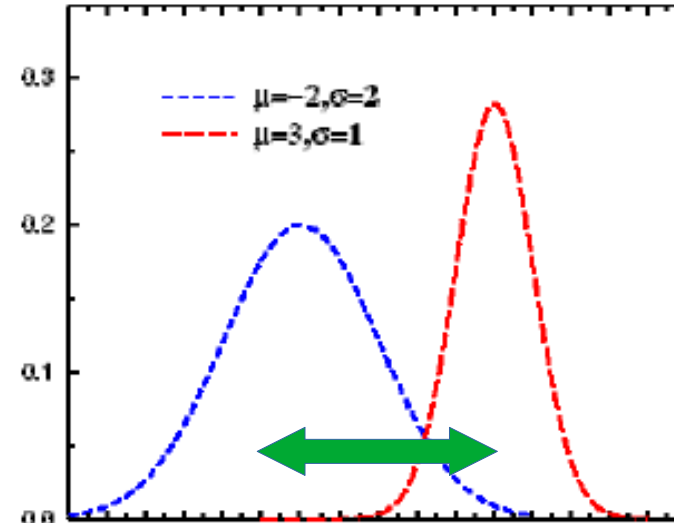
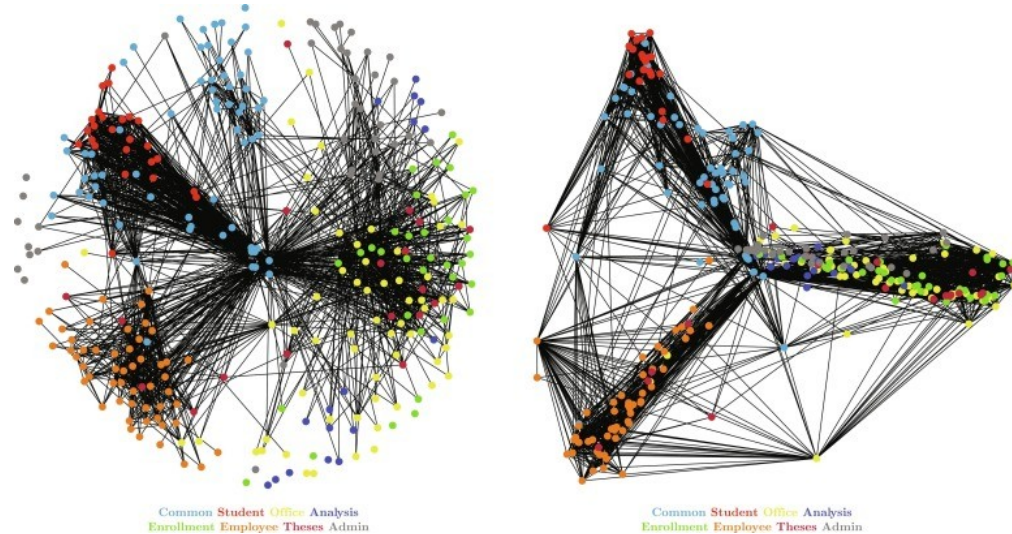


Bayesian statistics

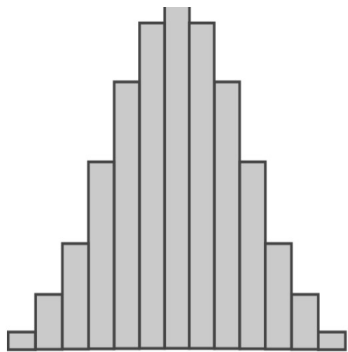
Distribution -> if we cut parts of complicated reality equation for subset of causes,
lets say we want to predict if the drug will work based on age only,
then **sometimes it works, sometimes not**.

100% accurate complicated formula is some
complicated curve in 1,000,000D
(or even infinitely complex, **intractable**)

Statistical distribution if only age is considered
(blue are age values when drug works,
red are when drug does not work)
72,5% accurate

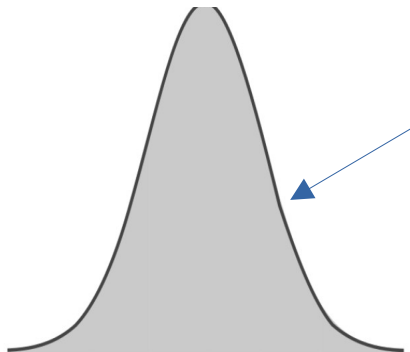


Distribution can have different shapes, formulas and properties



Discrete

$\sum \text{value} = 1$



Continuous

$\int \text{values} = 1$

Gaussian normal:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Exponential:

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$

Poisson:

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

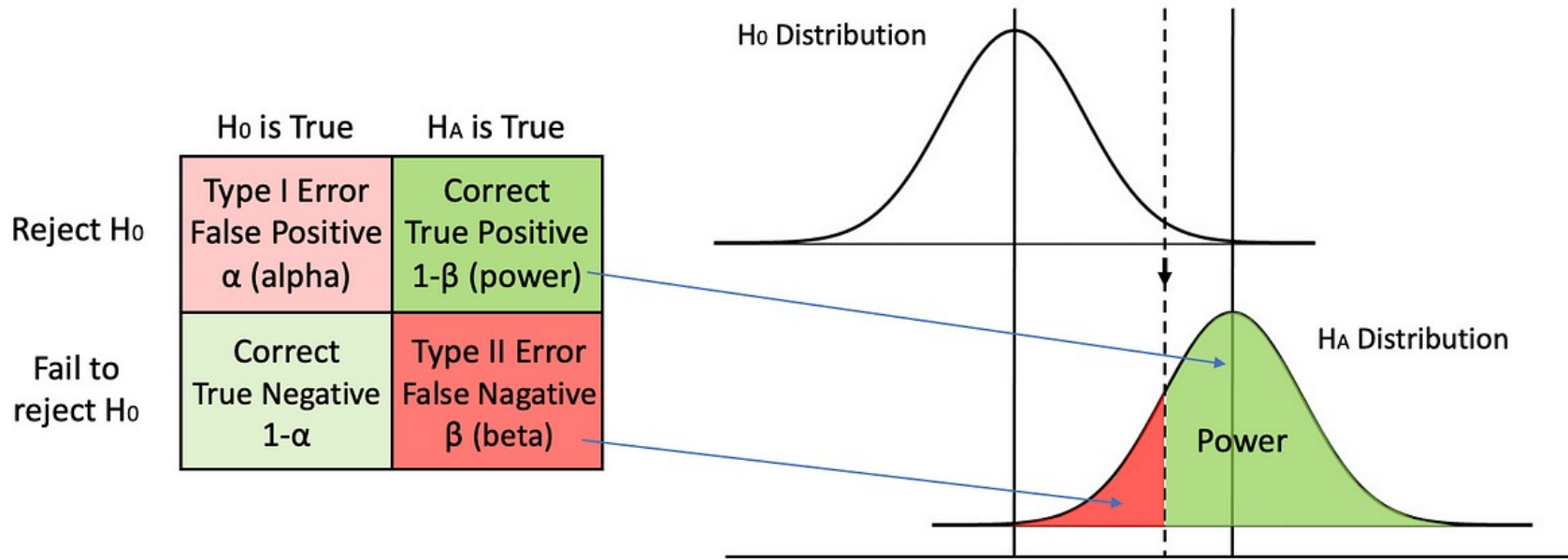
All of them have some **parameters** that we need to **MEASURE / ESTIMATE**

$$\mu = \frac{\sum x}{N}$$

What do these formulas say about distributions?

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Bayesian statistics and Hypothesis testing

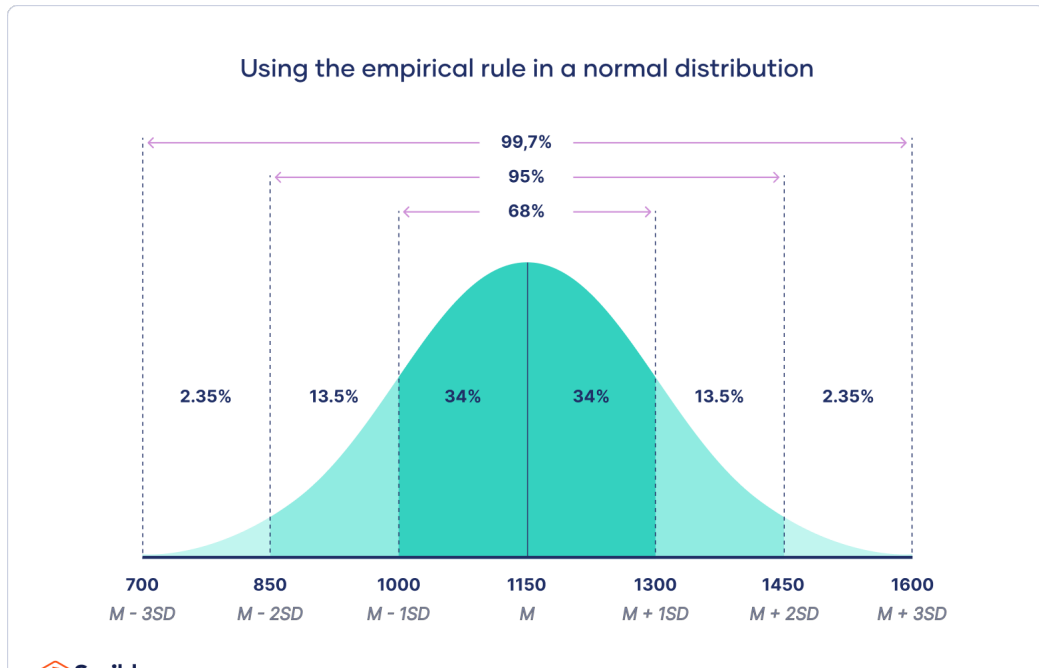


Also – when choosing between two distributions (as for drugs), it **DOES** matter what is the distribution of interest.

Alternative Hypothesis (H_A) = some new fact that I want to proof (discovery)

Null Hypothesis (H_0) = a typical assumption/belief that is held before your discovery

Název vyšetření	17.06.24	Jednotka	Meze	Hodnocení
<u>Základní biochemická vyšetření</u>				
* S_Bilirubin celkový	24,2	umol/L	0,0 - 21,0	*
* S_ALT	0,36	ukat/L	0,10 - 0,68	*
* S_AST	0,42	ukat/L	0,10 - 0,67	*
* S_GGT	0,27	ukat/L	0,17 - 1,19	*
* S_Glukóza v séru	5,4	mmol/L	3,9 - 5,6	*
<u>Hematologie</u>				



- I went to a doctor and they took a sample of my blood
- They measured multiple biomolecules
- If you have too many or too low of any molecules, something is wrong with you
- Scientists decided that **alpha=5%** is the border between “its okay” and “it is not okay”
- This is called **Z-test**
- It needs scientists to **measure** a lot of people to **estimate** (construct) a distribution
e.g. → measure mean and variance:

$$\mu = \frac{\sum x}{N} \quad \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

- > Do you know what is parametric and non-parametric test?
- > Do you know what is p-value?
- > What is confidence interval?
- > What is quantile? One-sided vs two-sided?

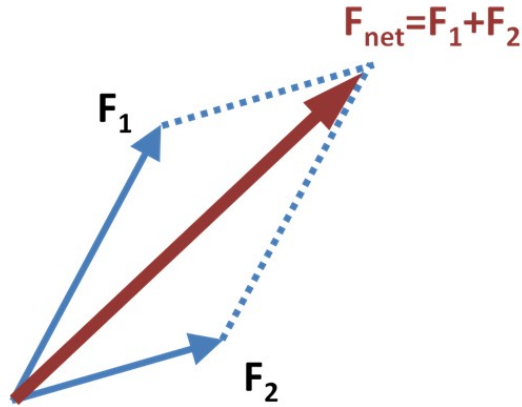
Show them in figure on previous slide

6. Sample questions pertaining to prerequisites:

-  [stat_min.pdf](#).

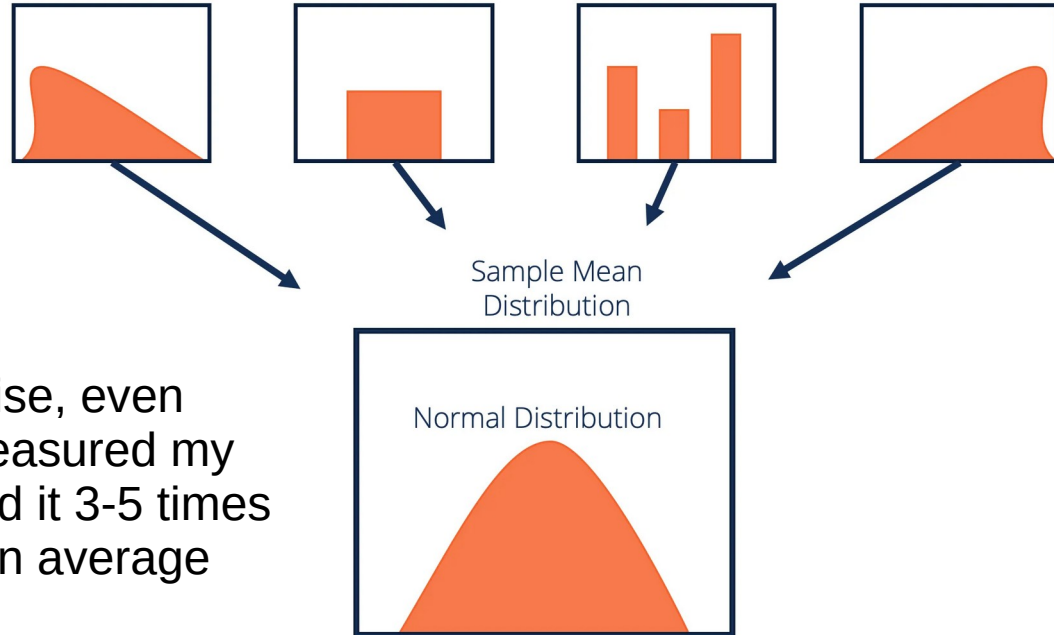
Why Gaussian normal distribution?

- Why we use a Gaussian normal? What if there is another distribution?
- Maybe, but typically when you have multiple factors/causes, they are typically summed



Example: Physics
sum of multiple forces

- Then, when we compute their mean (sum / number), it is mathematically Gaussian! (Central limit theorem)



- **Consequence** – to be more precise, even in previous test they didn't just measured my biomolecules level, they measured it 3-5 times from big blood sample and took an average

Problem with Gaussian for small sample number

$$\mu = \frac{\sum X}{N}$$

Mean is okay, it is calculated same way

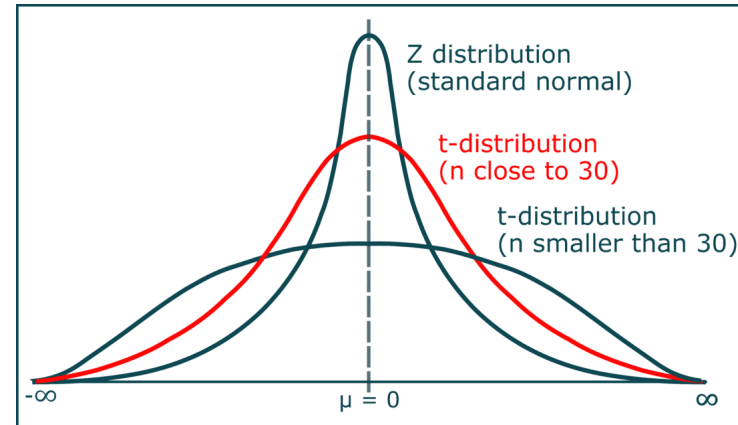
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

But if scientists do not have large enough number of people,
variance will be very different from real one
-> very big impact on correctness

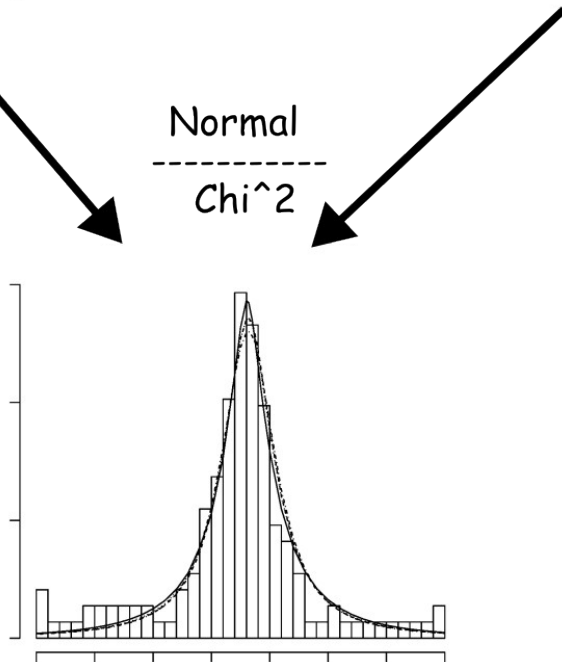
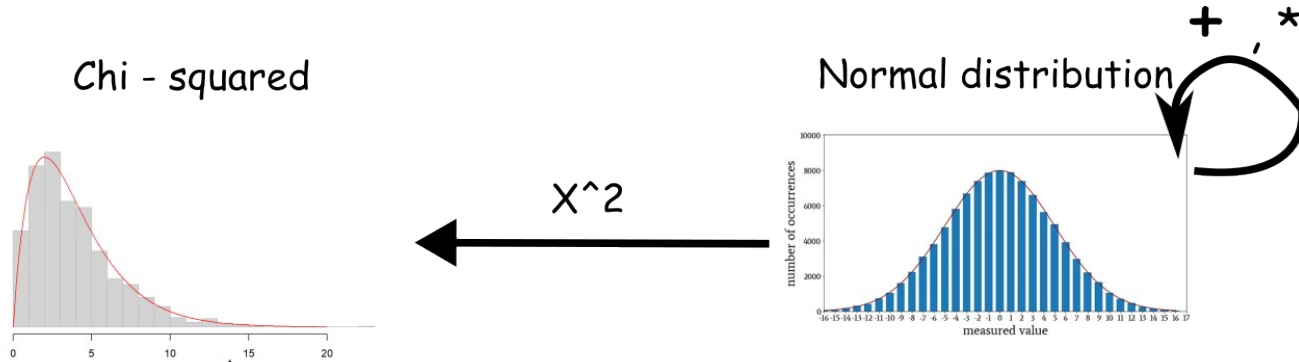
t-student test (t-test)

$$\frac{X - \mu}{\frac{1}{\sqrt{N}} S_H} \in \text{not normal anymore} \in t\text{-student} (df = N - 1)$$

Instead of σ^2 use $S_H^2 = \frac{\sum (X_i - \mu)^2}{N - 1}$



Where does t-distribution and other test distributions come from?



$$t = \frac{\frac{1}{n} \sum_{i=1}^n x_i - \mu}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} / \sqrt{n}}$$

Statistics 101 (basics)

Prerequisites are in courseware (**Exam** section):

6. Sample questions pertaining to prerequisites:

-  [stat_min.pdf](#).

Today we will start with basics and go over it fast

If you don't understand anything from PDF, do a self study!

Today's homework (voluntary) for 1 point

Tutorials

T	Date	Teacher	Contents	Materials
1	22.9.	JB, AA, JK	Statistical testing, t-test, significance, power of the test.	 san_intro.zip ,  r_setup.zip ,  pres-1.pdf

Further questions and tasks:

1. the power of the test that evaluates whether “the mean blood pressure after treatment is above a certain value”. To increase this power, report when you reach 80%.
2. see what happens with the test when violating assumptions with outliers, inject outliers
3. violate other assumptions to see what happens (a distribution different from normal)
4. show the setting in which the central limit theorem can be applied, gradually change sample means

Goals:

- *) Install Rstudio/R and run some code by Ctrl + Enter
- *) Answer some theoretical questions from basic statistics
- *) Or you could look into HTML file with plots/figures if you do not want to run R code
- *) Submit the modified Rmd or plain text file with answers

You want to ask an anonymous question? You have some problem with the seminar?

<https://askany.me/session/acb96d8c-86d5-4763-86da-a2c88724>

[← Back to Home](#) **Live Q&A Session (Owner) - Mode: normal**

Mode: Normal



[Copy Session Link](#)

Session Owner

Your Question

I do not like the way you teach. I want to solve more examples, not more theory!

Your Name (optional)

Enter your name

☒ Ask Anonymously

Submit Question

End of statistics 101
Good luck!

