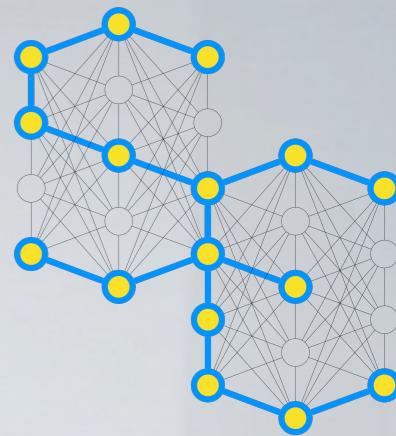


**800**  
A N N I  
1222 • 2022



UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA



# Statistical Hypothesis Testing

## Search Engines

Master Degree in Computer Engineering

Master Degree in Data Science

Academic Year 2023/2024

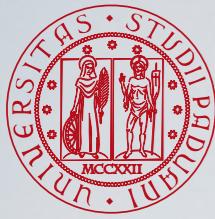


DIPARTIMENTO  
DI INGEGNERIA  
DELL'INFORMAZIONE

Nicola Ferro

Intelligent Interactive Information Access (IIIA) Hub  
Department of Information Engineering  
University of Padua





# Outline

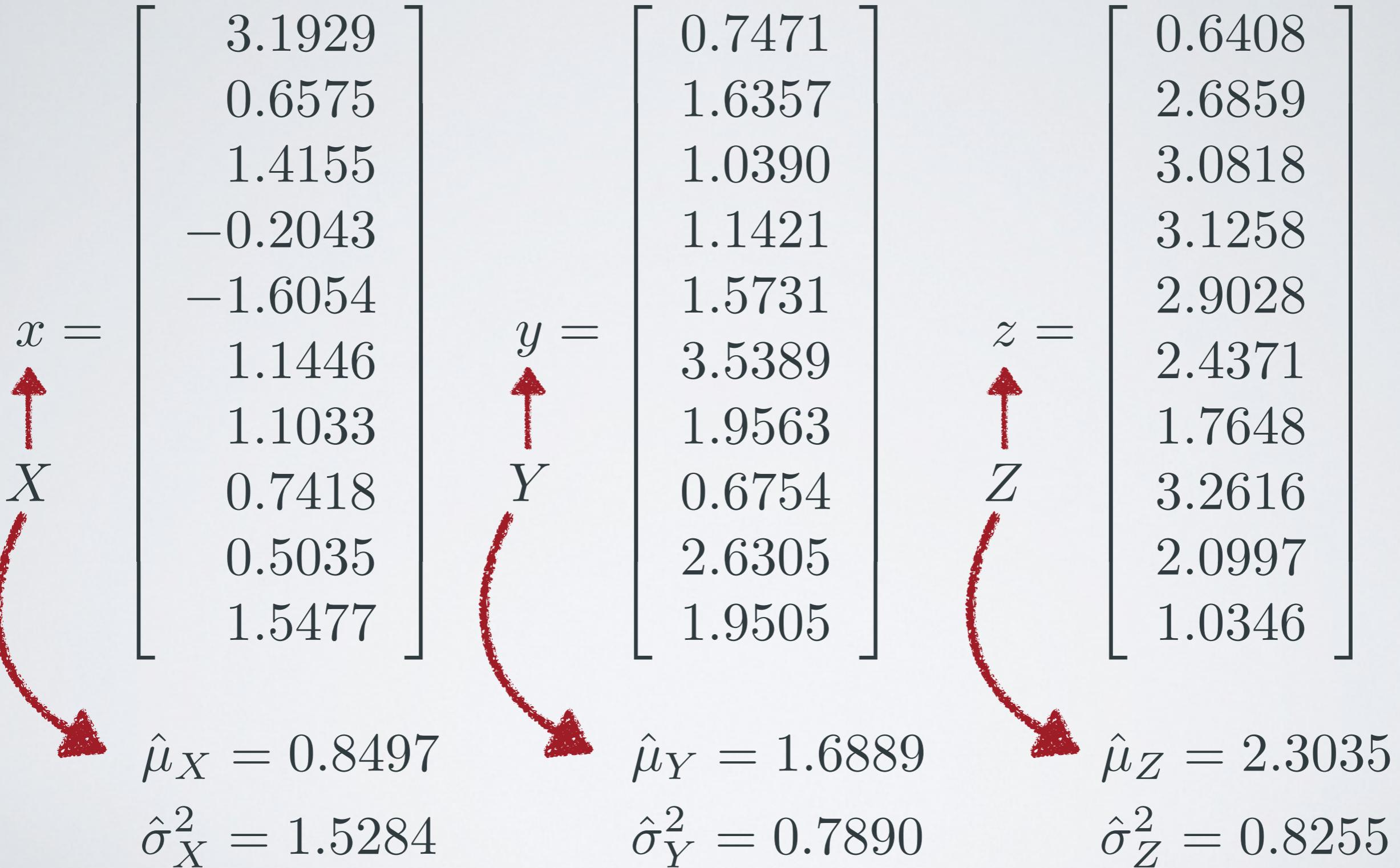
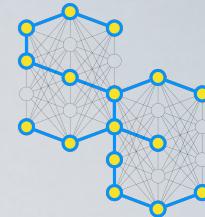


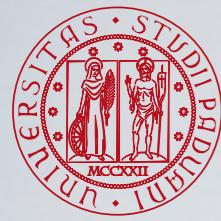
- Hypothesis Testing
- Student's t test
- Analysis of Variance (ANOVA)

# Hypothesis Testing

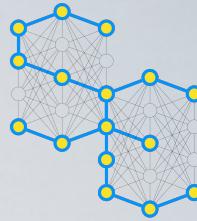


# The Problem: Are They Different?

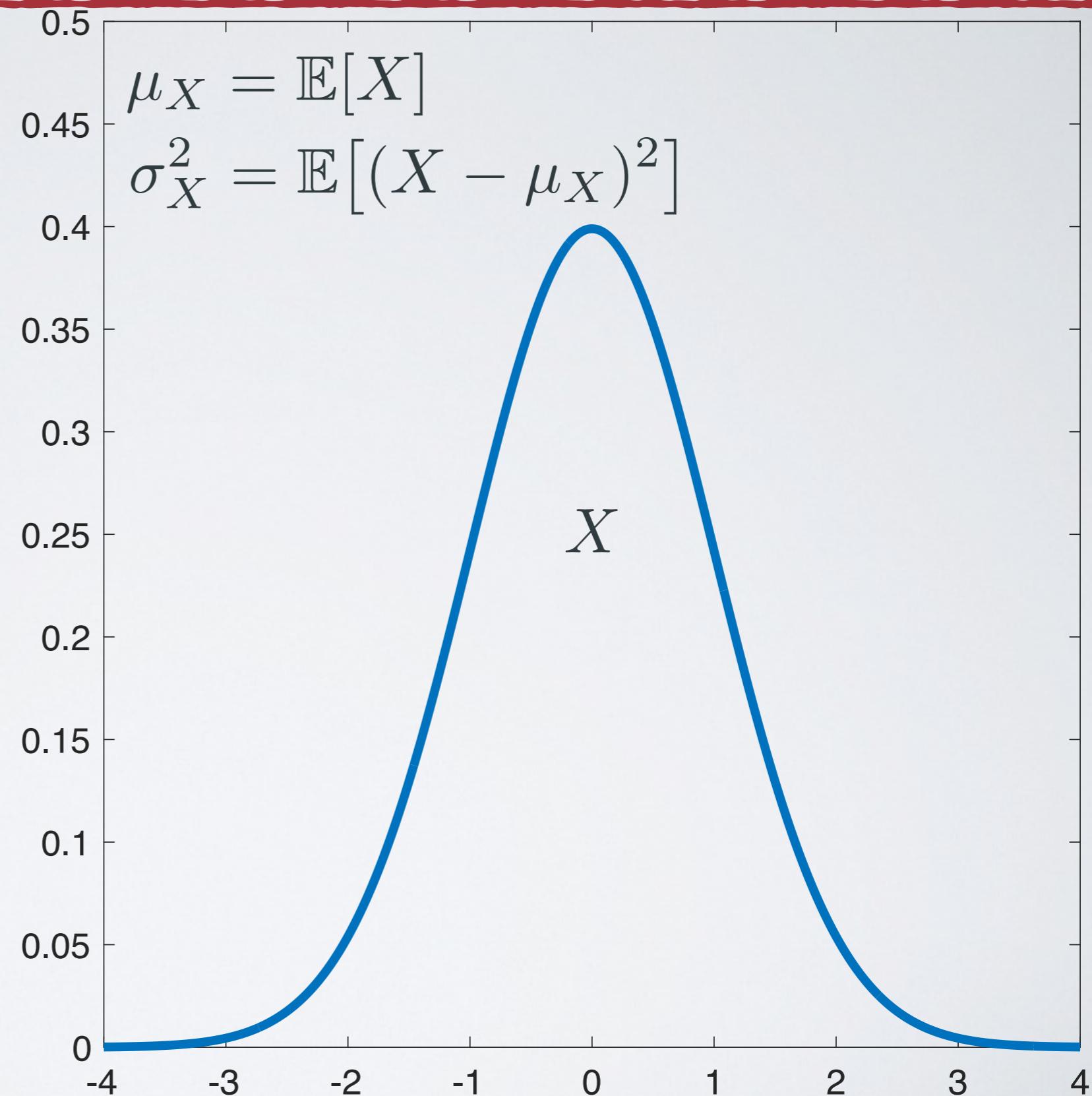




# What Did We Do?



$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix}$$



# What Did We Do?

$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix} \leftarrow \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \\ X_{10} \end{bmatrix}$$

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu}_X)^2$$

- $\{X_1, X_2, \dots, X_n\}$  is a **random sample** of size  $n$ , i.e. a sequence of **independent and identically distributed (i.i.d)** random variables drawn from a distribution
- $x_i$  is an **observation** (realisation) of the random variable  $X_i$ , i.e. the actual value assumed by that random variable in a given trial
- The **sample mean**  $\hat{\mu}_X$  and the **sample variance**  $\hat{\sigma}_X^2$  are **unbiased estimators** of the **population mean**  $\mu_X$  and **population variance**  $\sigma_X^2$

# What Did We Do?

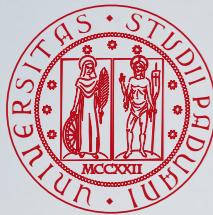
$$x = \begin{bmatrix} 3.1929 \\ 0.6575 \\ 1.4155 \\ -0.2043 \\ -1.6054 \\ 1.1446 \\ 1.1033 \\ 0.7418 \\ 0.5035 \\ 1.5477 \end{bmatrix} \leftarrow \begin{array}{l} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \\ X_{10} \end{array}$$

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i$$

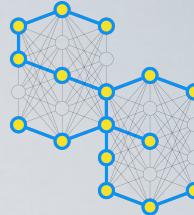
$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu}_X)^2$$

Degrees of Freedom (DF)

- $\{X_1, X_2, \dots, X_n\}$  is a **random sample** of size  $n$ , i.e. a sequence of **independent and identically distributed (i.i.d)** random variables drawn from a distribution
- $x_i$  is an **observation** (realisation) of the random variable  $X_i$ , i.e. the actual value assumed by that random variable in a given trial
- The **sample mean**  $\hat{\mu}_X$  and the **sample variance**  $\hat{\sigma}_X^2$  are **unbiased estimators** of the **population mean**  $\mu_X$  and **population variance**  $\sigma_X^2$



# Law of Large Numbers & Central Limit Theorem



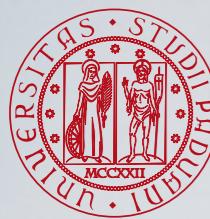
- The (strong) **Law of Large Numbers** states that the sample mean converges almost surely to the population mean

$$\hat{\mu}_X \xrightarrow{\text{a.s.}} \mu_X \quad \text{for } n \rightarrow \infty$$

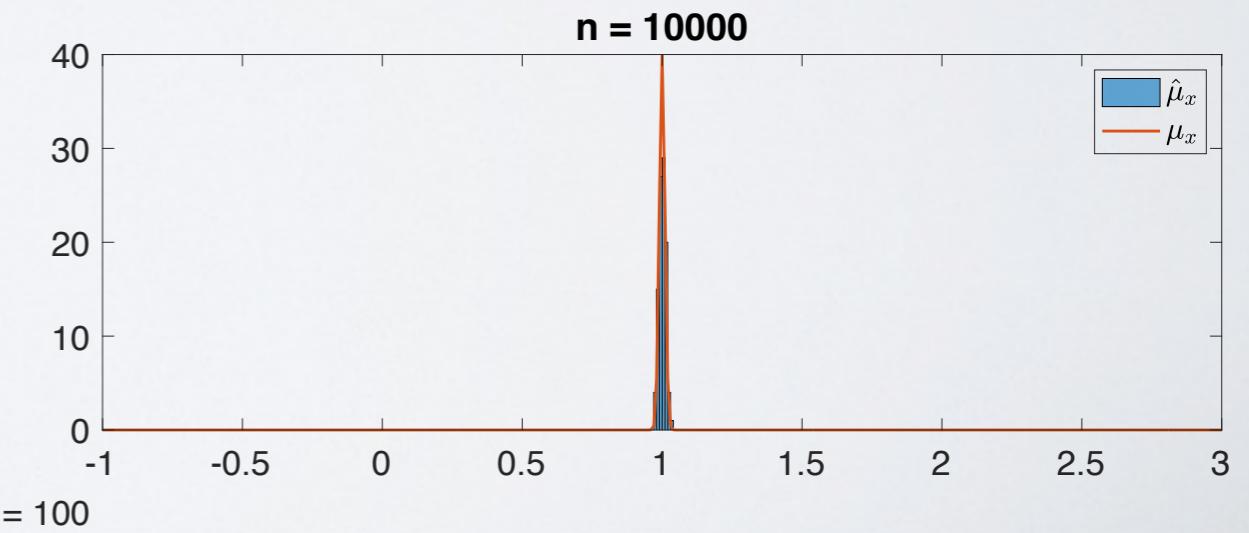
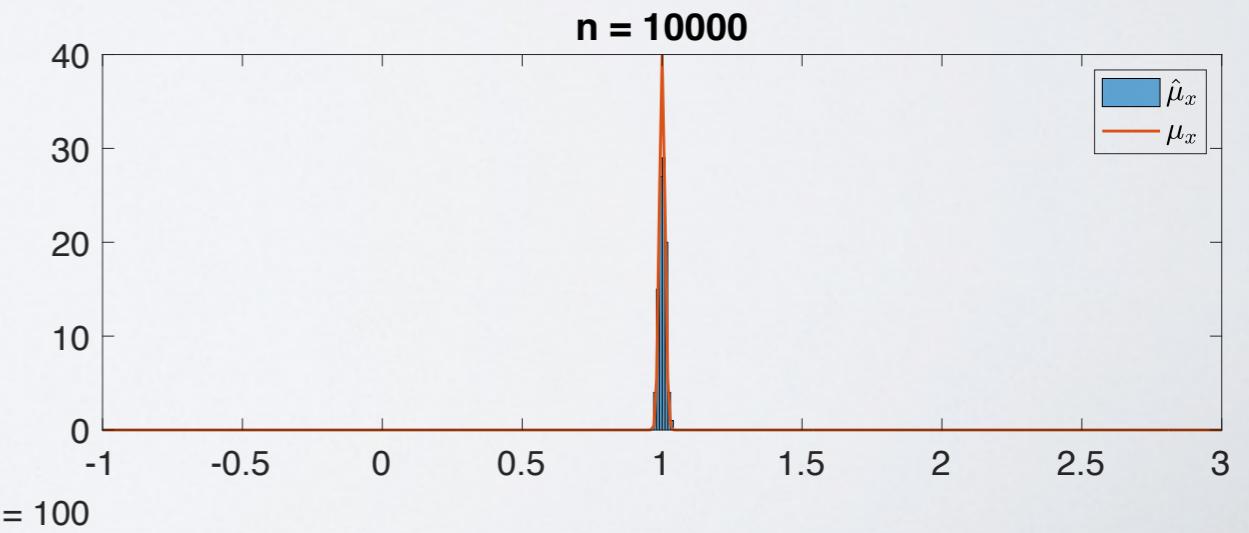
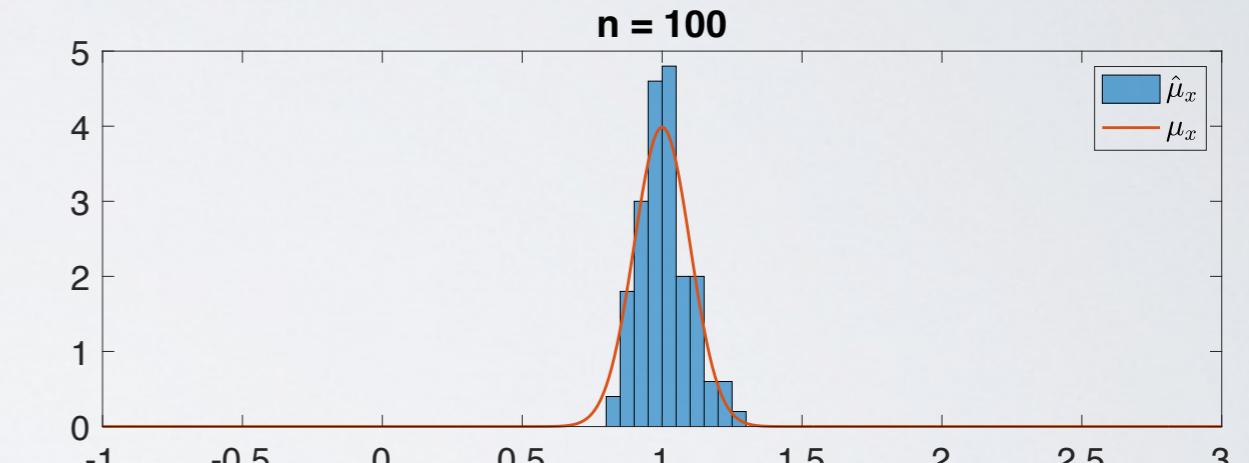
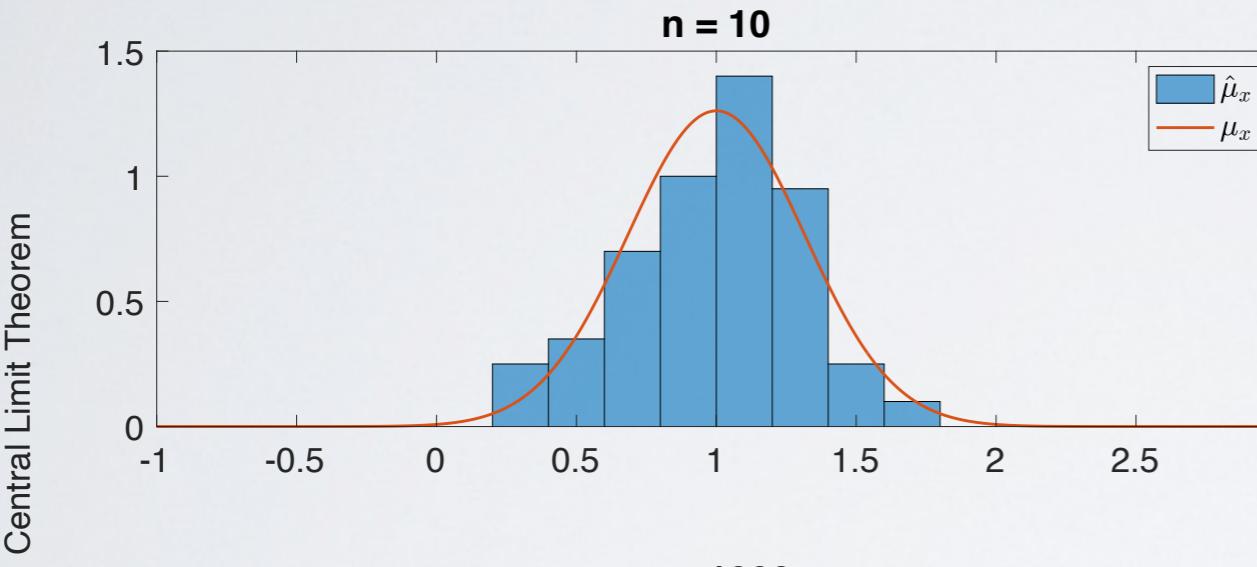
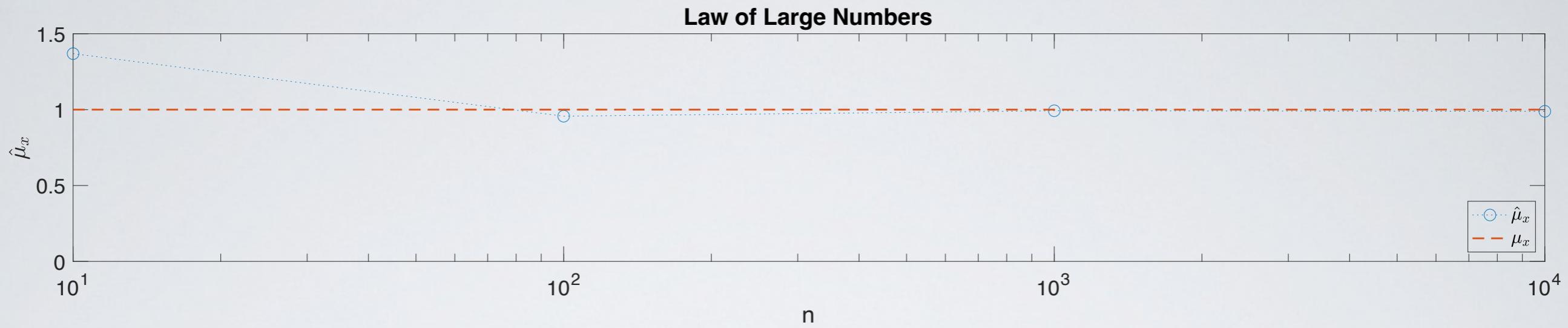
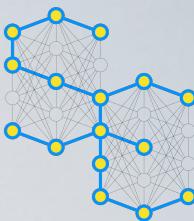
- The **Central Limit Theorem** states that the sample mean follows a normal distribution

$$\hat{\mu}_X \sim \mathcal{N} \left( \mu_X, \frac{\sigma_X^2}{n} \right)$$

$$\frac{\hat{\mu}_X - \mu_X}{\sigma_X / \sqrt{n}} \sim \mathcal{N}(0, 1)$$

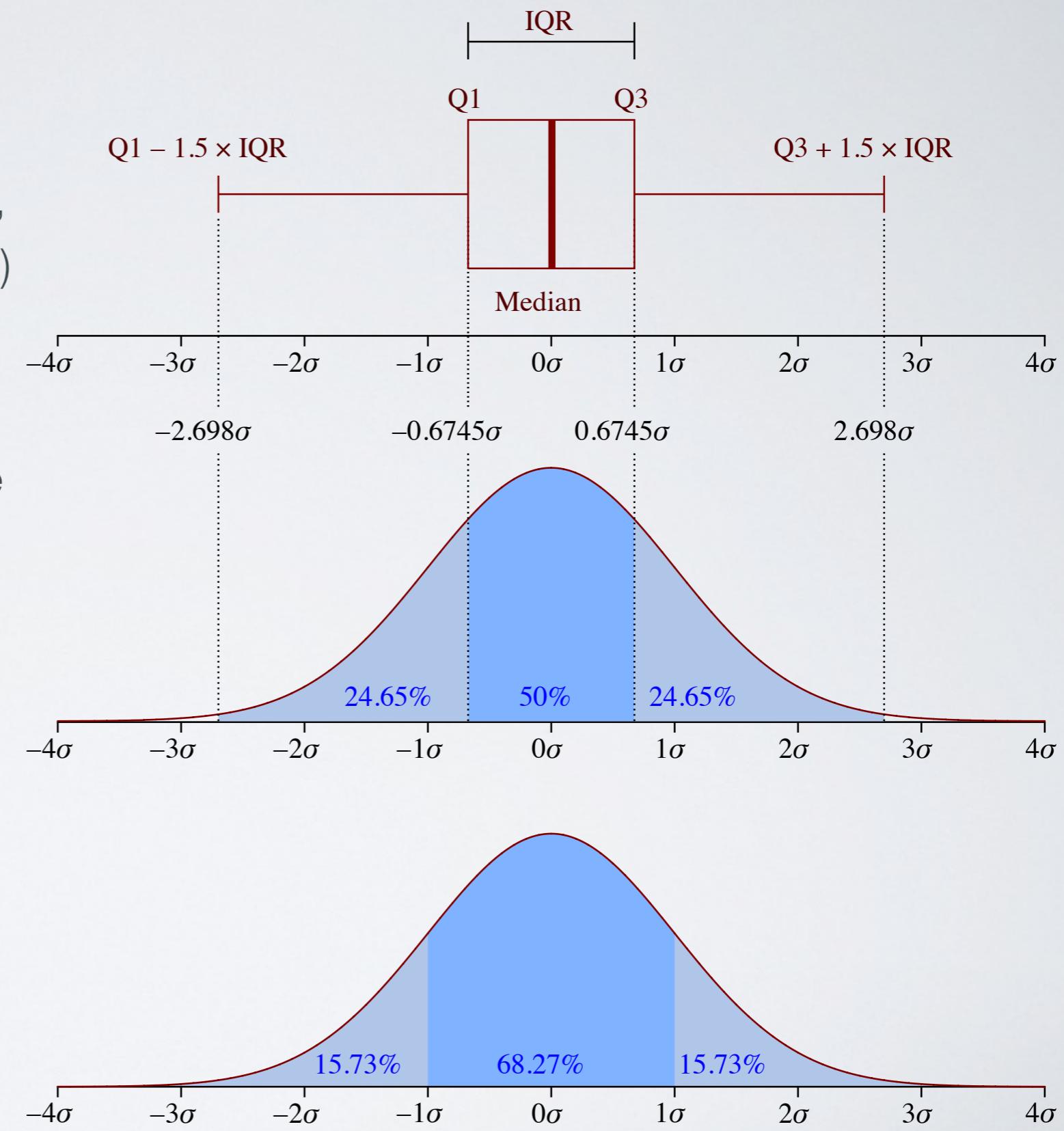


# Law of Large Numbers & Central Limit Theorem

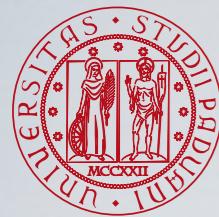


# Box Plot

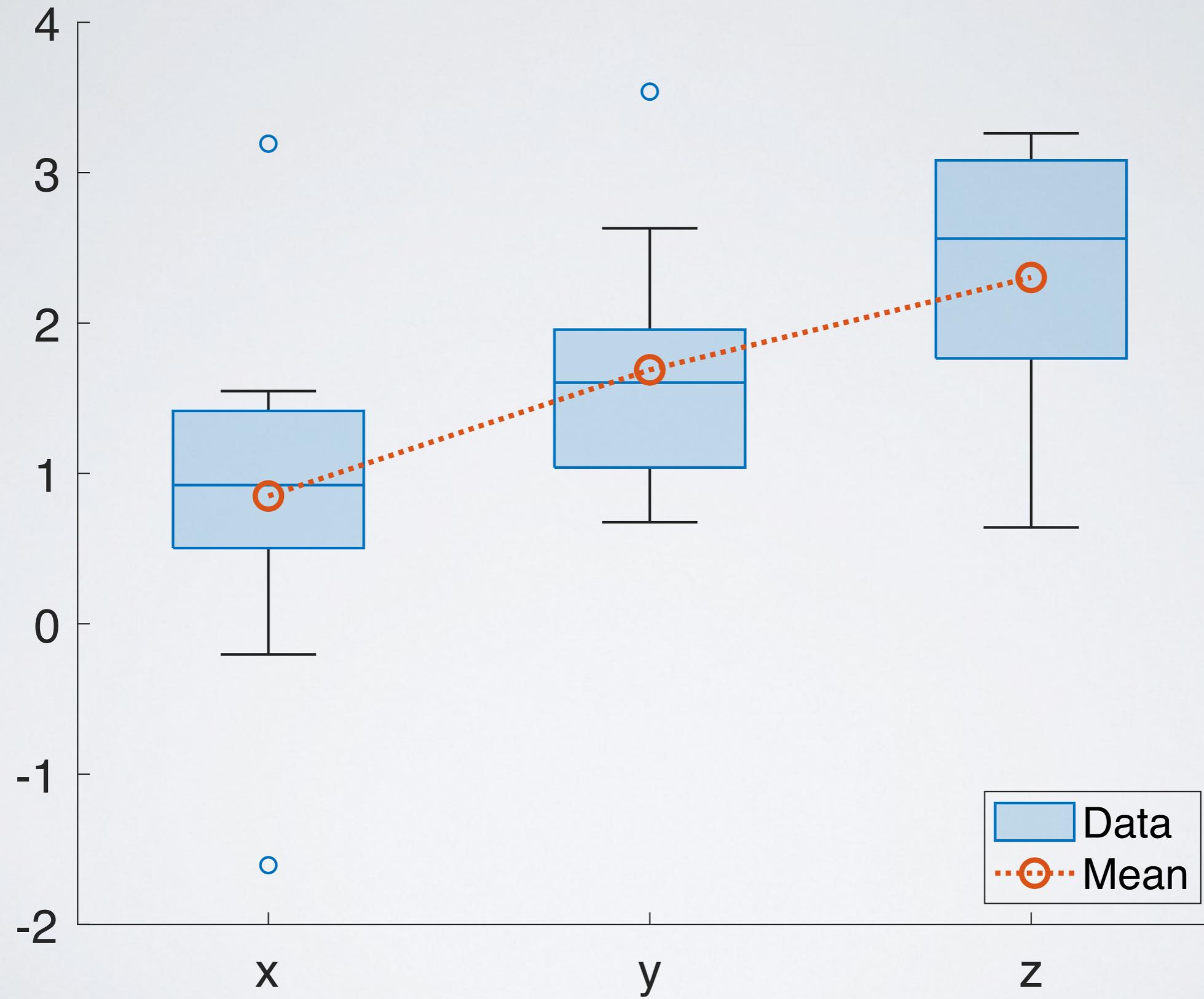
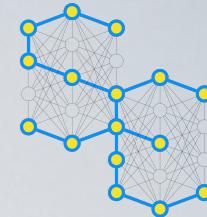
- A **boxplot** is a graphical tool to summarise a distribution of data
- The box shows the **first quartile** (Q1), the **second quartile** (Q2, the **median**) with a line inside the box, and the **third quartile** (Q3)
- The box represents the **Inter-Quartile Range (IQR)**, i.e. the difference Q3-Q1
- The extension of the **whiskers** represents  $1.5 \cdot \text{IQR}$ 
  - they roughly cover ~99% of the data, assuming a normal distribution
- Any data outside the whiskers is considered an **outlier**

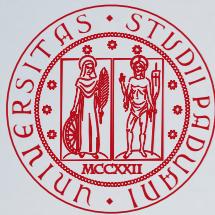


McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician*, 32(1):12–16.

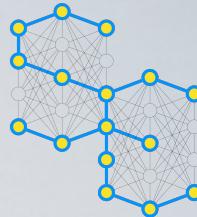


# The Problem: Are They Different?





# Compute it Yourself!



## Basic Statistics, Law of Large Numbers, and Central Limit Theorem

This example shows the computation of some basics statistics on simulated data as well as the application of the Law of Large Numbers and the Central Limit Theorem.

### Basic Statistics and Box Plot

Define the parameters to be used.

```
% set the random number generator for reproducibility
rng(6, 'twister');

% total number of samples to be used
n = 10;
```

Generate the random (simulated) data from normal distributions with same variance as follows:

- $x \sim \mathcal{N}(1, 1)$
- $y \sim \mathcal{N}(1, 1)$
- $z \sim \mathcal{N}(2.3, 1)$

```
x = 1 + randn(n, 1);
y = 1 + randn(n, 1);
z = 2.3 + randn(n, 1);
```

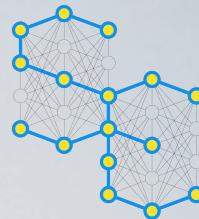
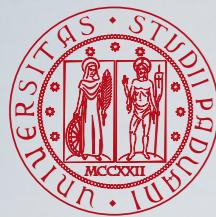
Compute basics statistics:

- sample mean

$$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^n x_i = \text{mean}(x)$$

- sample variance

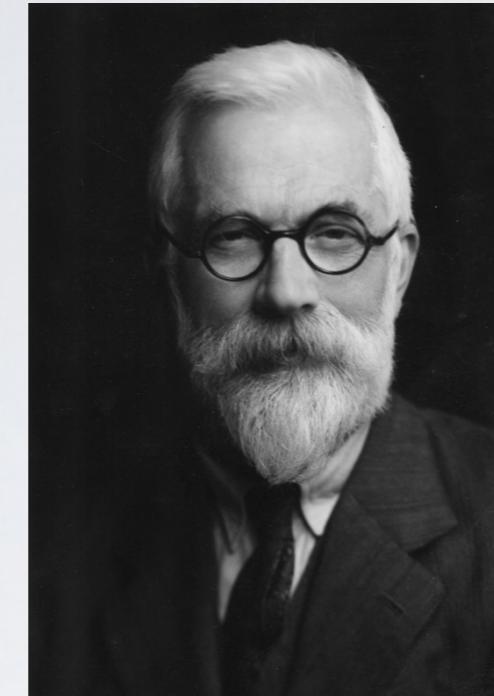
$$\hat{\sigma}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu}_x)^2 = \text{var}(x)$$



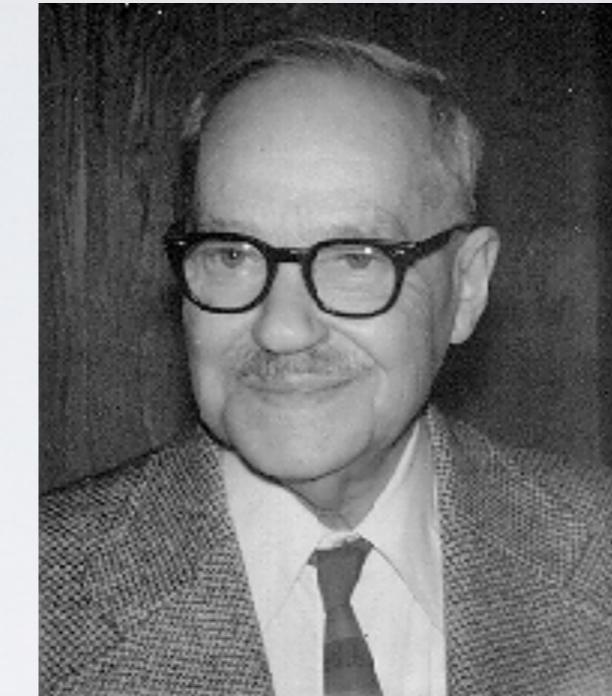
# Statistical Hypothesis Testing



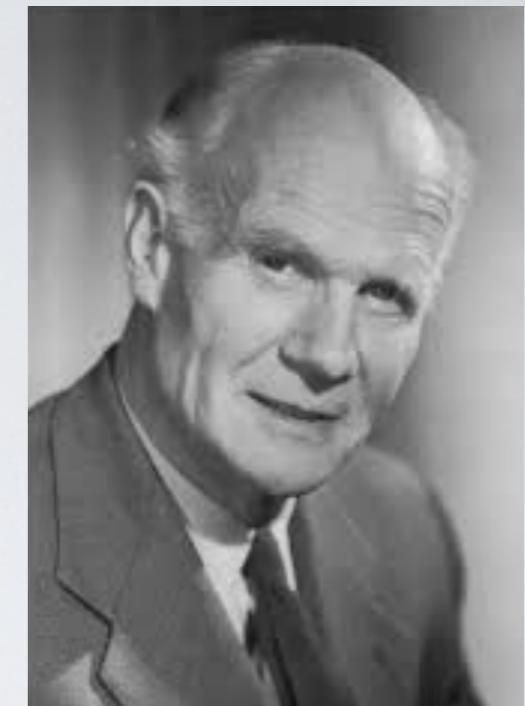
William Sealy Gosset  
“Student”



Ronald Aylmer Fisher



Jerzy Neyman



Egon Sharpe Pearson

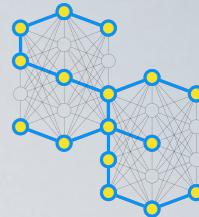
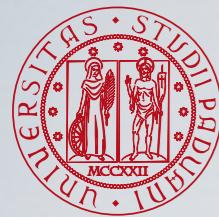
- Statistical hypothesis testing provides us with a mathematical framework to conduct statistical inference from the data
- It compares the so-called null hypothesis  $H_0$  against an alternative hypothesis  $H_1$  or  $H_A$
- The comparison is statistically significant if the data are unlikely to be a realisation of the null hypothesis with respect to a chosen threshold, called significance level  $\alpha$ . In this case we reject the null hypothesis; in the opposite case, we fail to reject the null hypothesis

Student (1908). The Probable Error of a Mean. *Biometrika*, 6(1):1–25.

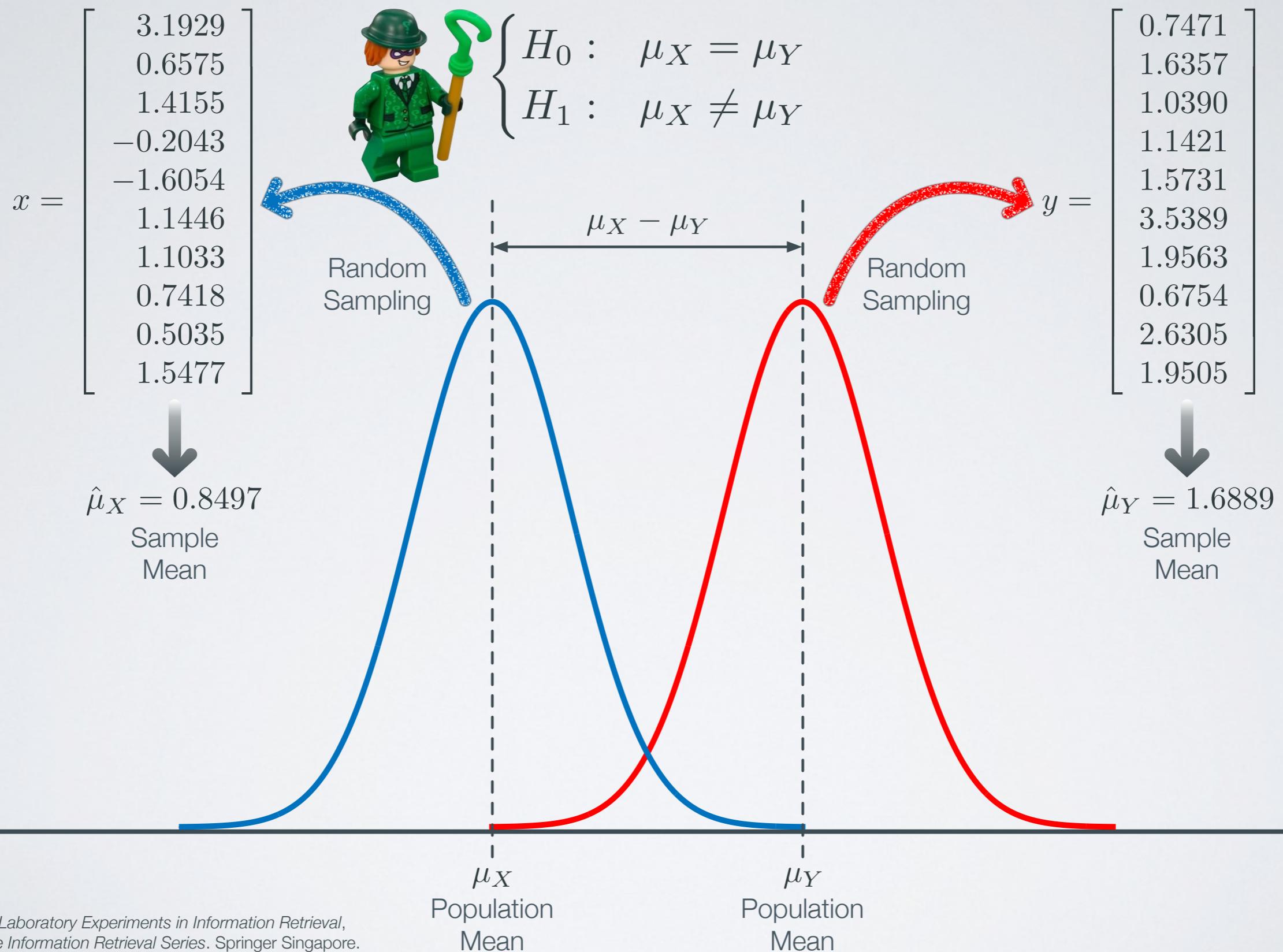
Fisher, R. A. (1925). Statistical Methods for Research Workers. Oliver & Boyd, Edinburgh, UK.

Neyman, J. and Pearson, E. S. (1928). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. *Biometrika*, 20A(1/2):175–240.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American Statistical Association*, 88(424):1242–1249.



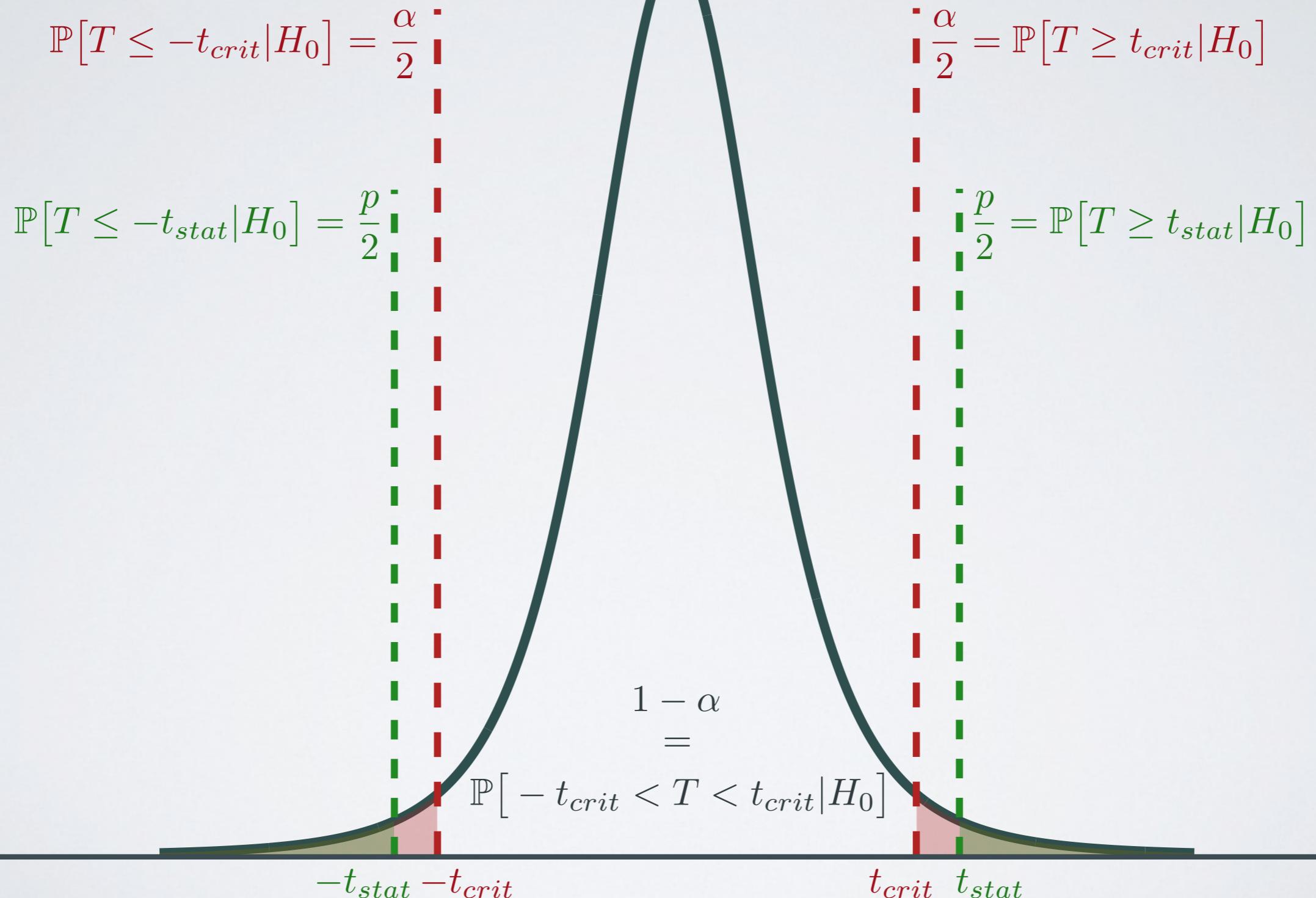
# Formulating the Problem

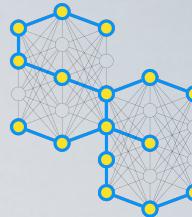




# Test Statistic

Test statistic distribution  $T$  under the null hypothesis  $H_0$





# Test Statistic

Test statistic distribution  $T$  under the null hypothesis  $H_0$

$$\mathbb{P}[T \leq -t_{crit}|H_0] = \frac{\alpha}{2}$$

$$\mathbb{P}[T \leq -t_{stat}|H_0] = \frac{p}{2}$$

$$\frac{\alpha}{2} = \mathbb{P}[T \geq t_{crit}|H_0]$$

$$\frac{p}{2} = \mathbb{P}[T \geq t_{stat}|H_0]$$

Reject the null hypothesis  $H_0$   
when

$$t_{stat} \geq t_{crit}$$

or, equivalently, when

$$p \leq \alpha$$

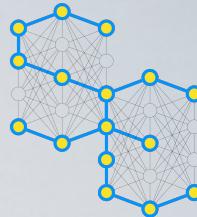
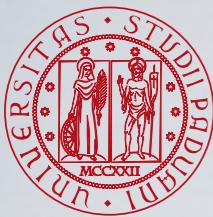


$$1 - \alpha$$
  
 $=$

$$\mathbb{P}[-t_{crit} < T < t_{crit}|H_0]$$

$-t_{stat}$   $-t_{crit}$

$t_{crit}$   $t_{stat}$



# Types of Error



We fail to reject  
 $H_0$   
[not statistically significant]

We reject  
 $H_0$   
[statistically significant]

$H_0$   
is true  
[e.g. systems are equivalent]

Correct conclusion  
[true negative]

Probability  
 $1 - \alpha$

Type I error  
[false positive]

Probability  
 $\alpha$

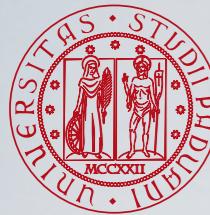
$H_0$   
is false  
[e.g. systems are not equivalent]

Type II Error  
[false negative]

Probability  
 $\beta$

Power (Correct conclusion)  
[true positive]

Probability  
 $1 - \beta$



# Multiple Comparisons



- Type I errors concern the comparison of 2 alternatives
- What happens when you need to compare  $c$  alternative?
- They originate  $m = \binom{c}{2}$  possible pairs to be compared, i.e.  $m$  hypotheses to be tested simultaneously
- Performing multiple comparisons increases the Type I error probability, i.e. it is easier to reject the null hypothesis when you should not, since the  $m$  pairwise comparisons are **independent**

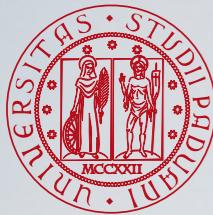
$$P(\text{No Type I Error}) = 1 - \alpha$$

$$P(\text{No Type I Errors}) = \prod_{i=1}^m (1 - \alpha) = (1 - \alpha)^m$$

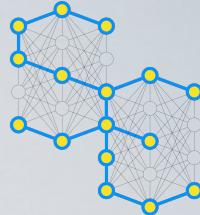
$$P(\text{At Least One Type I Error}) = 1 - (1 - \alpha)^m$$

Family-wise  
Error Rate (FWER)

# **Student's t Test**



# (Paired) Student's t test



- Let us assume to have a random sample of size  $n$  (**independence**) from two **normally-distributed** random variables with **same variance** (homoskedasticity)

$$X \sim \mathcal{N}(\mu_X, \sigma^2)$$

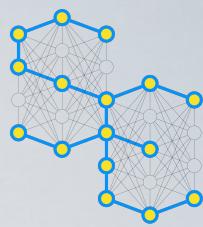
$$Y \sim \mathcal{N}(\mu_Y, \sigma^2)$$

- Let us define the difference among these two random variables

$$D_i = X_i - Y_i \sim \mathcal{N}(\mu_X - \mu_Y, 2\sigma^2)$$

$$\hat{\mu}_D = \frac{1}{n} \sum_{i=1}^n d_i$$

$$\hat{\sigma}_D^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \hat{\mu}_D)^2$$



# (Paired) Student's t test

- The test statistic

$$T = \frac{\hat{\mu}_D - \mu_D}{\sqrt{\hat{\sigma}_D^2/n}} \sim t(n-1)$$

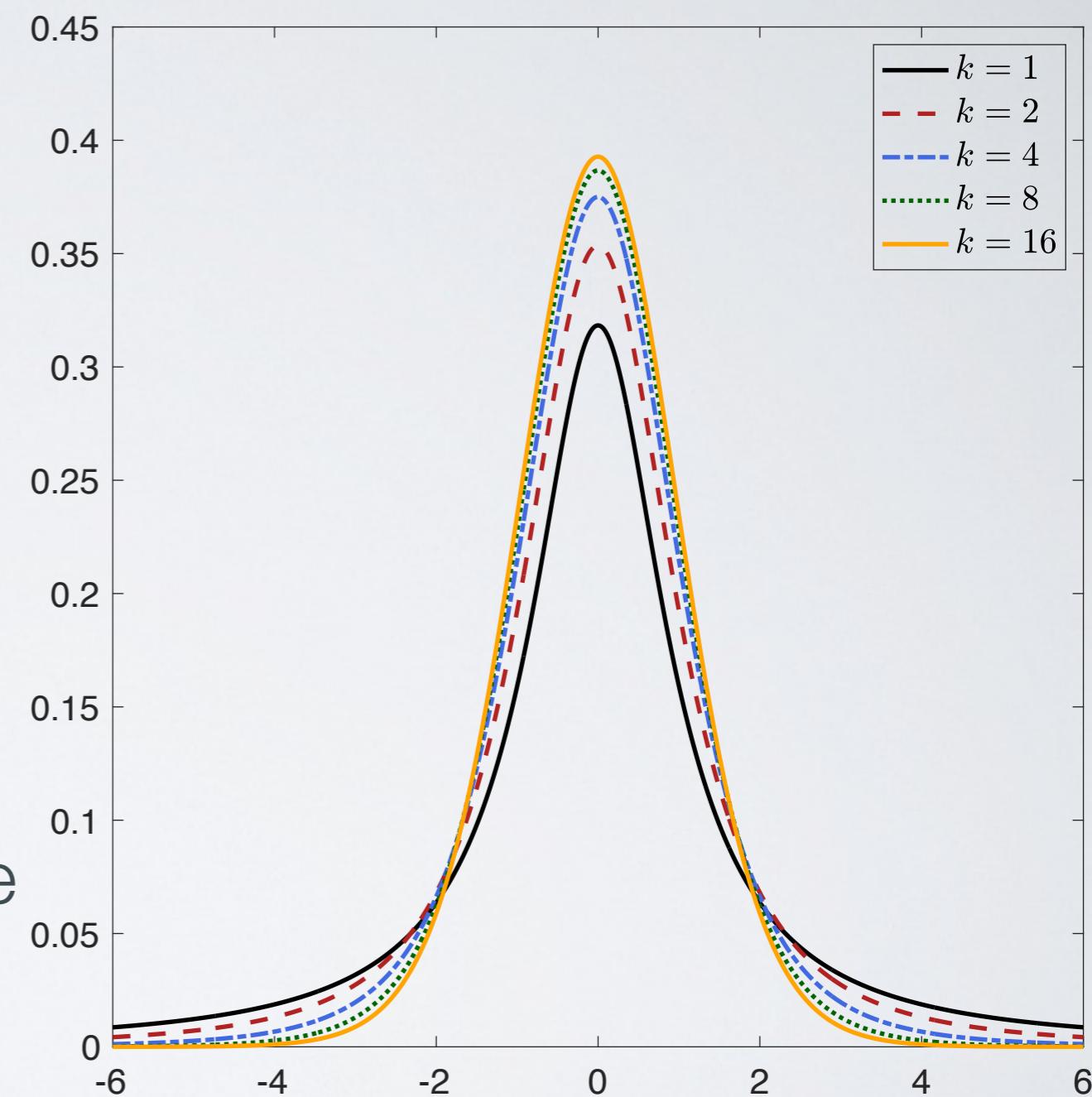
follows a **Student's t distribution** with  $k = n - 1$  degrees of freedom

- Under the null hypothesis

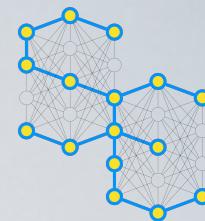
$$H_0 : \mu_X = \mu_Y \Leftrightarrow \mu_D = 0$$

we obtain the test statistic value

$$t_{stat} = \frac{\hat{\mu}_D}{\sqrt{\hat{\sigma}_D^2/n}}$$



# (Paired) Student's t test



- The test statistic

$$T = \frac{\hat{\mu}_D - \mu_D}{\sqrt{\hat{\sigma}_D^2/n}} \sim t(n-1)$$

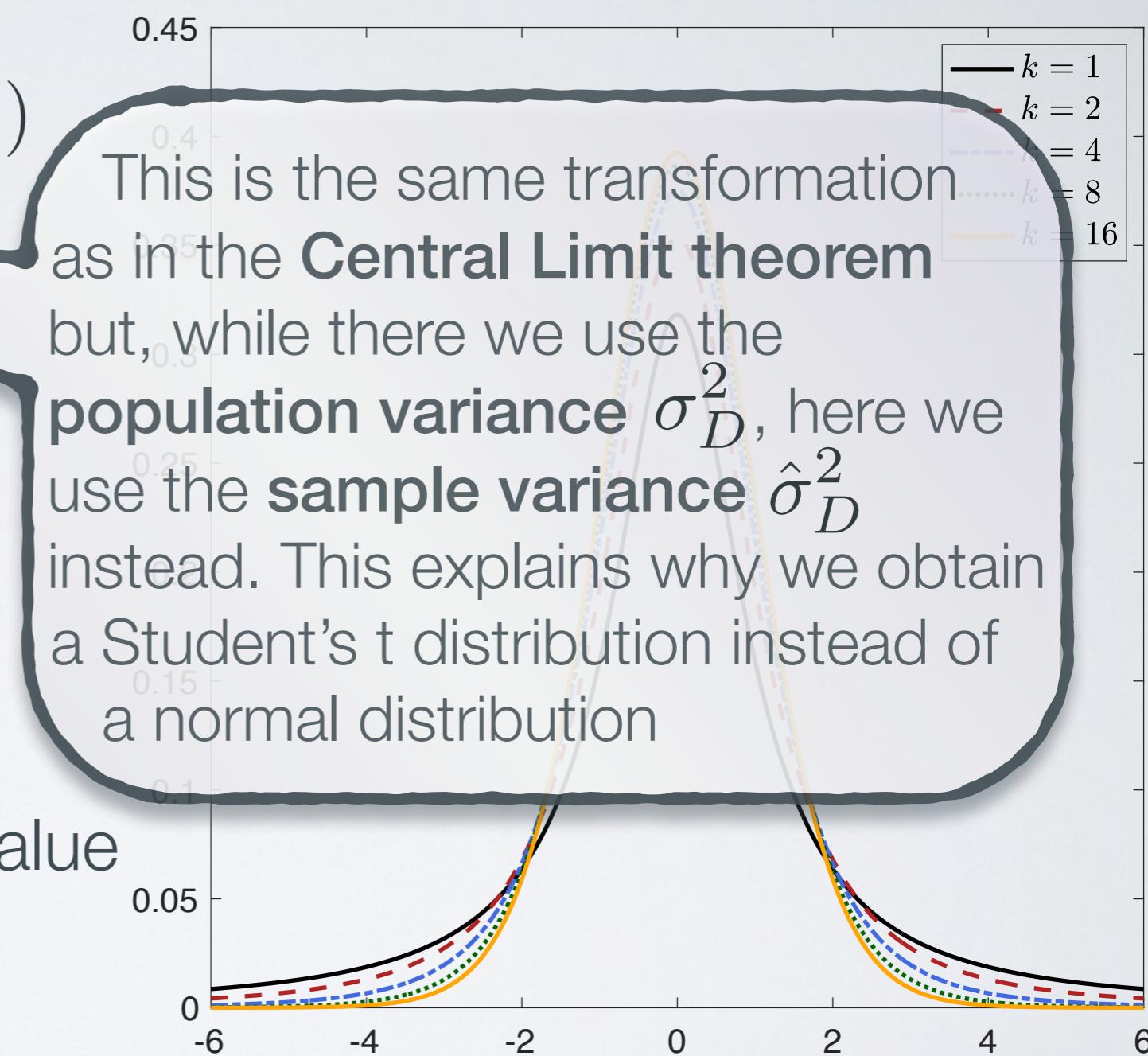
follows a **Student's t distribution** with  $k = n - 1$  degrees of freedom

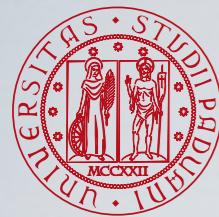
- Under the null hypothesis

$$H_0 : \mu_X = \mu_Y \Leftrightarrow \mu_D = 0$$

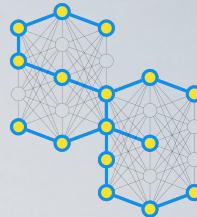
we obtain the test statistic value

$$t_{stat} = \frac{\hat{\mu}_D}{\sqrt{\hat{\sigma}_D^2/n}}$$





# Compute it Yourself!



## Computation of (Paired) Student's t test on simulated data

This example shows step-by-step computation of the Student's t test on simulated data.

### Setup

Define the parameters to be used.

```
% set the random number generator for reproducibility  
rng(6, 'twister');  
  
% total number of samples to be used  
n = 10;  
  
% the significance level  
alpha = 0.05;
```

Generate the random (simulated) data from normal distributions with same variance as follows:

- $x \sim \mathcal{N}(1, 1)$
- $y \sim \mathcal{N}(1, 1)$
- $z \sim \mathcal{N}(2.3, 1)$

```
x = 1 + randn(n, 1);  
y = 1 + randn(n, 1);  
z = 2.3 + randn(n, 1);
```

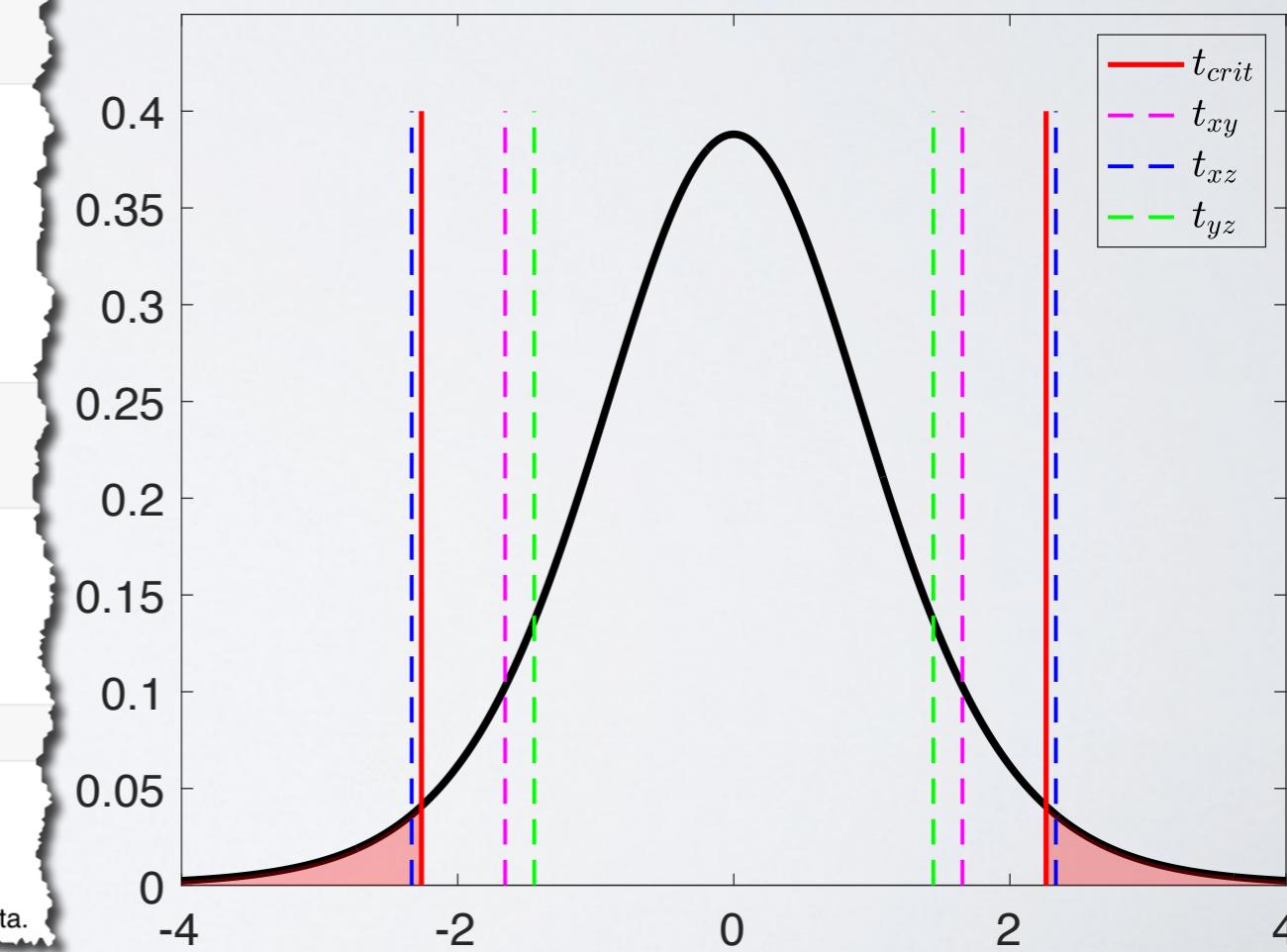
Compute the degrees of freedom of the Student's t test.

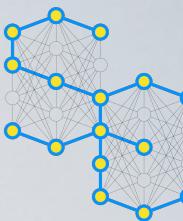
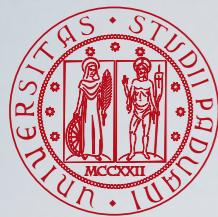
Note that they depend on the number of samples  $n$  but not on the actual simulated data.

```
df = n - 1;
```

Compute the critical value of the Student's t test under the null hypothesis  $H_0$ .

Note that it depends on the significance level  $\alpha$  and the degrees of freedom  $df$  but not on the actual simulated data.





# Compute it Yourself!

Computation of (Paired) Student's t test on TREC 08, 1999, AdHoc

[~, p] = ttest(x, y)

This example shows how to compare runs from TREC 08, 1999, AdHoc using the Student's t test.

computes the Student's t test among two vectors x and y

## Setup

Load the performance score about the runs. The file contains:

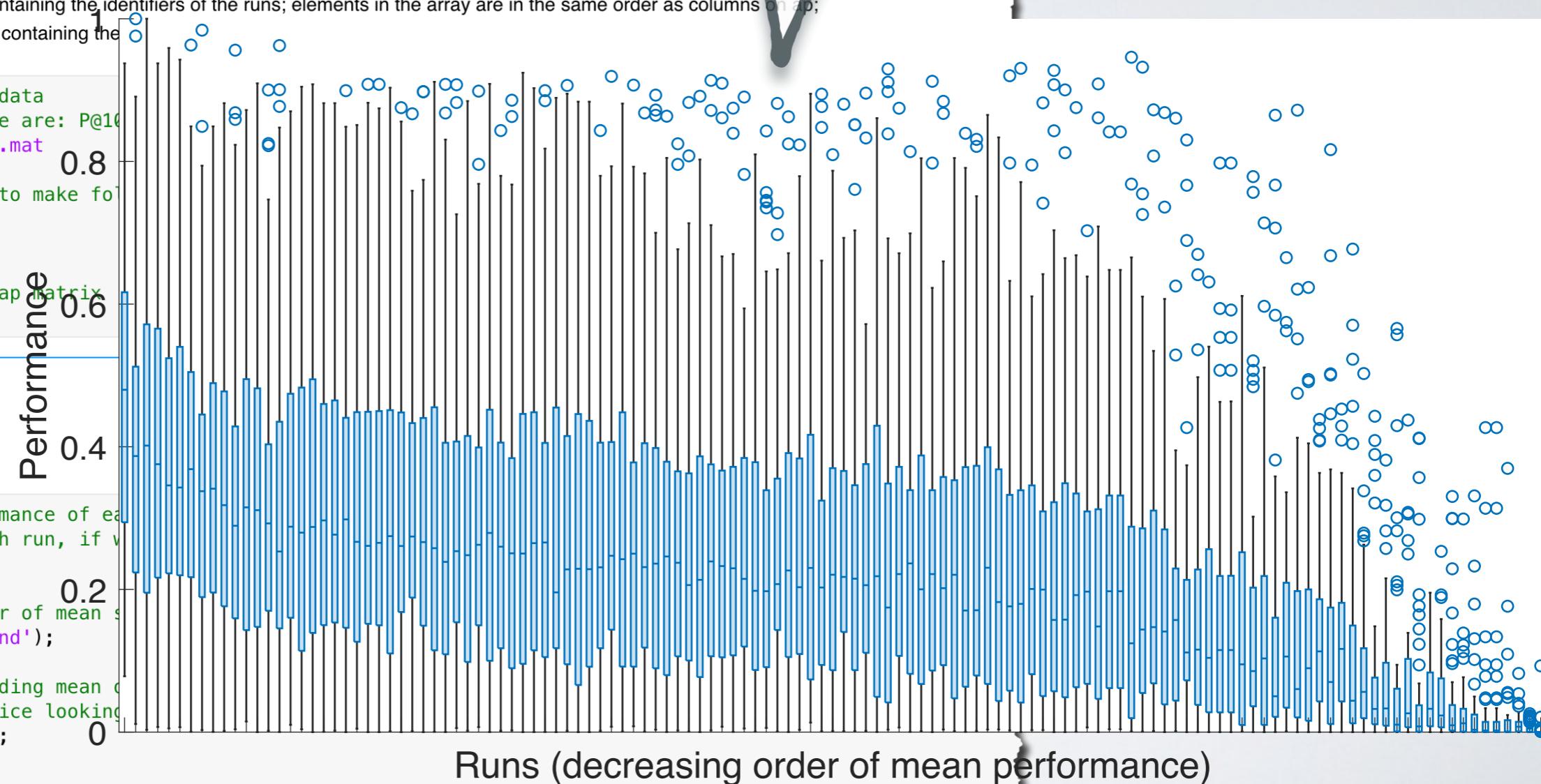
- ap (or another measure name): a matrix where rows are topics and columns are runs; each cell contains the AP score for a run on a given topic;
- runs: an array of strings containing the identifiers of the runs; elements in the array are in the same order as columns on ap;
- topics: an array of strings containing the

```
% Load Average Precision data  
% Other measures available are: P@10  
load ../../data/T08/ap.mat  
  
% rename ap into measure to make fo  
% measure we load  
measure = ap;  
  
% remove the now useless ap matrix  
clear ap;
```

## Box Plot

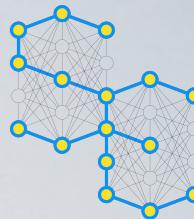
Show the box plot of the runs.

```
% Compute the mean performance of ea  
% This is the MAP for each run, if w  
m = mean(measure);  
  
% sort in descending order of mean s  
[~, idx] = sort(m, 'descend');  
  
% re-order runs by descending mean o  
% needed to have a more nice looking  
measure = measure(:, idx);  
runs = runs(idx);
```





# Type I and Type II Errors



## Simulation of Type I and Type II Errors

This example simulates Type I and Type II errors in the case of Student's t test.

### Setup

Define the parameters to be used.

```
% set the random number generator for reproducibility
rng(6, 'twister');

% total number of samples to be used
n = 10;

% the significance levels
alpha = [0.0001 0.001 0.01 0.05 0.10];

% the number of trials
trials = 1000;
```

### Type I Errors

Generate the random (simulated) data from the same normal distributions as follows:

- $x \sim \mathcal{N}(1, 1)$
- $y \sim \mathcal{N}(1, 1)$

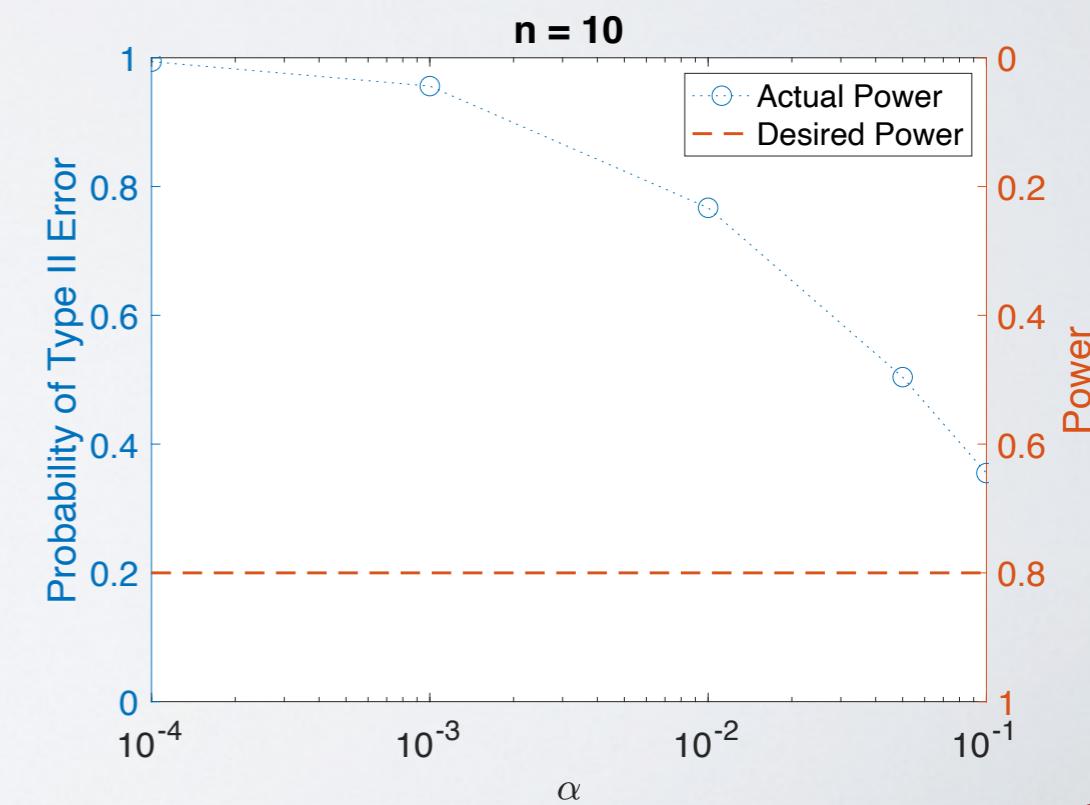
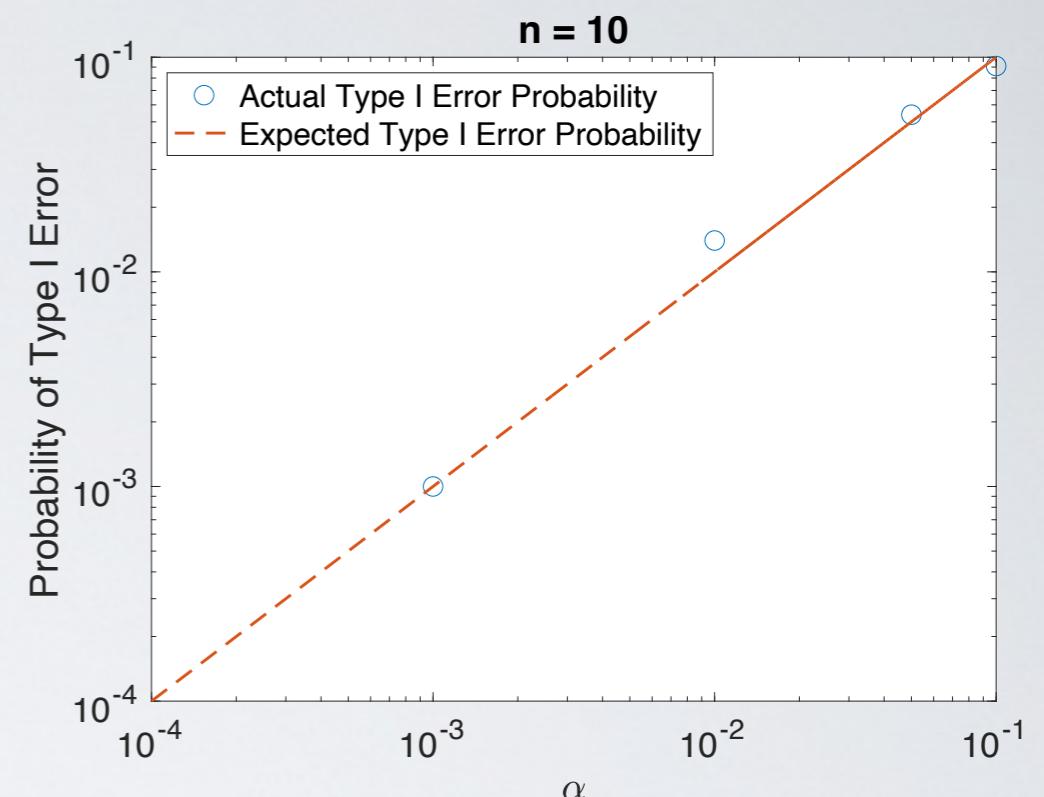
By construction,  $x$  and  $y$  are samples drawn from the same distribution, so the null hypothesis  $H_0$  holds in their case.

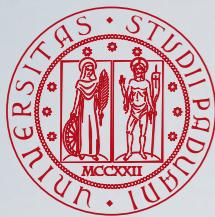
```
% the first random sample of size n, repeated trials times (columns)
x = 1 + randn(n, trials);

% the second random sample of size n, repeated trials times (columns)
y = 1 + randn(n, trials);
```

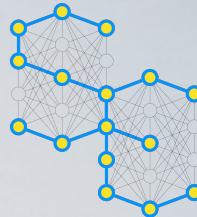
Compute trials Student's t test between each pair of  $x$  and  $y$  samples

```
% the p-values computed in the t tests
p = NaN(1, trials);
```





# Bonferroni Correction



- Test each individual hypothesis at a significance level

$$\alpha' = \frac{\alpha}{m}$$

- Criticisms

- too conservative
- increases the likelihood of Type II errors

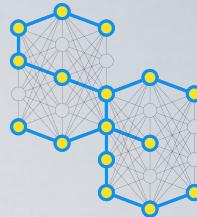


Carlo Emilio Bonferroni

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Number 8 in Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze. Libreria internazionale Seeber, Firenze, Italia.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, 316(7139):1236–1238.

# Student's t Test Assumptions?

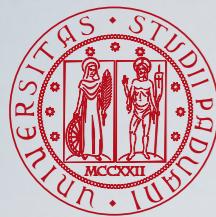


- **Independence:** topics and systems can be considered (reasonably) independent
- **Normality:** typical IR measures are bounded in  $[0, 1]$  so they cannot be normal
  - the normal distribution is unbounded
- **Homoskedasticity:** variance changes across systems
- The Student's t test is considered robust to violations of normality and also homoskedasticity when the sample sizes are equal (our typical case)

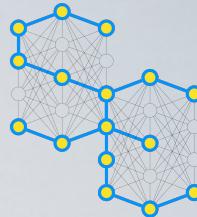


Markowski, C. A. and Markowski, E. P. (1990). Conditions for the Effectiveness of a Preliminary Test of Variance. *The American Statistician*, 44(4):322–326.

# **Analysis of Variance (ANOVA)**



# General Linear Models (GLM)



Data = Model + Error

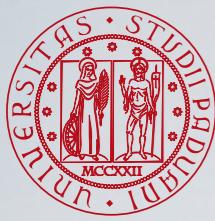
- A GLM explains the variation of a **dependent variable (Data)** in terms of a controlled variation of **independent variables (Model)** in addition to a **residual** uncontrolled variation (**Error**)
- Regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

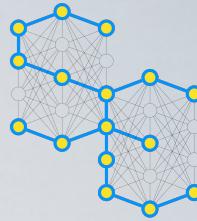
- ANalysis Of VAriance (ANOVA)

$$y_{ij} = \mu_{..} + \alpha_j + \varepsilon_{ij}$$

- the above regression model corresponds to the ANOVA one once you add as many  $x_{ij}$  predictors as many levels there are in the experimental condition  $\alpha_j$ , e.g., by using dummy coding

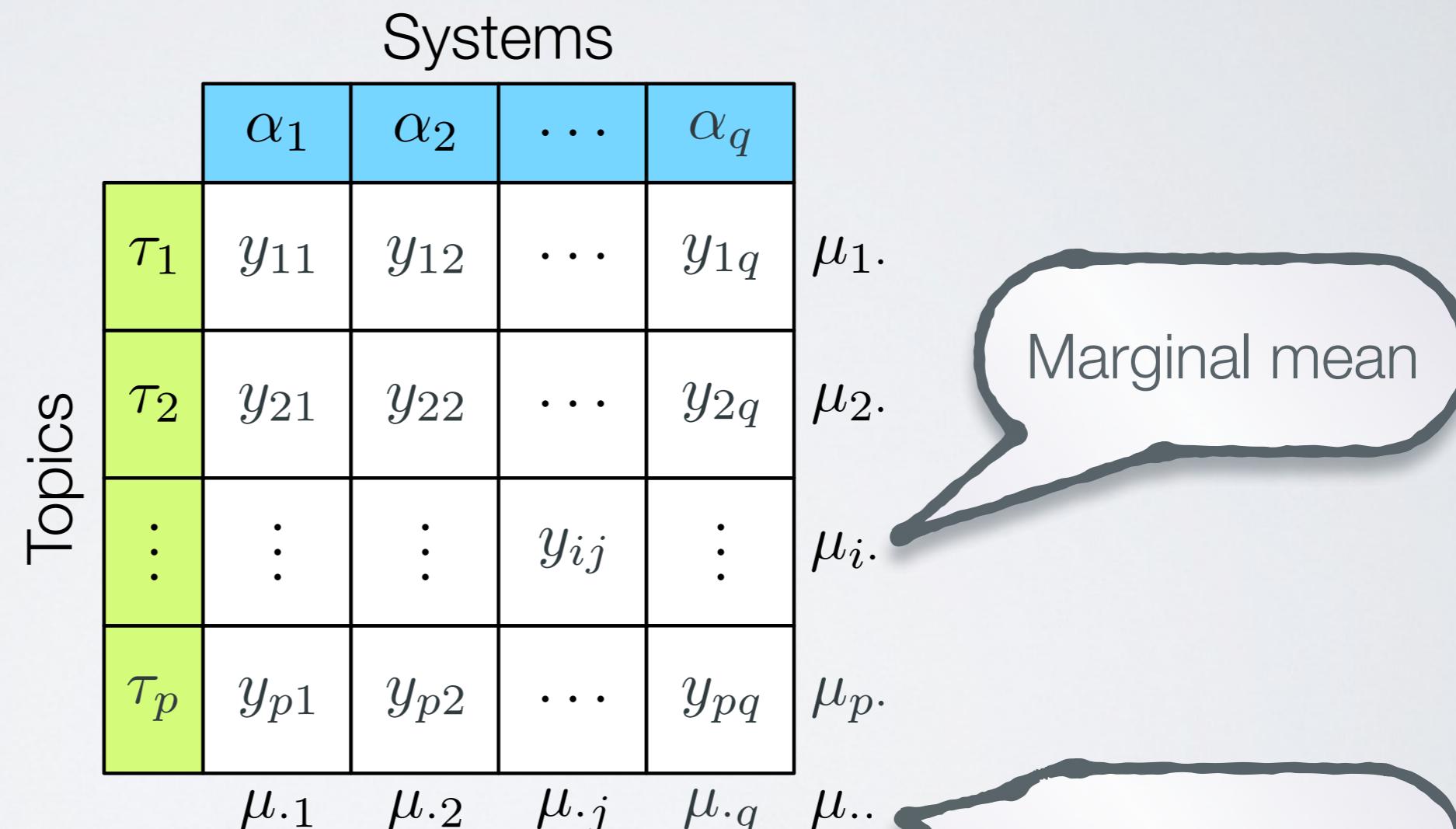


# Modelling System Effects (one-way ANOVA)



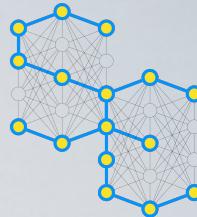
ANOVA

$$y_{ij} = \underbrace{\mu_{..} + \alpha_j}_{\text{Model}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}$$





# Modelling System Effects (one-way ANOVA)

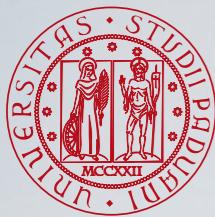


ANOVA

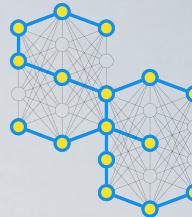
$$y_{ij} = \underbrace{\mu_{..} + \alpha_j}_{\text{Model}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}$$

Regression  $y_{ij} = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_q x_{iq}}_{\text{Model}} + \underbrace{\varepsilon_{ij}}_{\text{Error}}$

		Systems				
		$\alpha_1$	$\alpha_2$	$\dots$	$\alpha_q$	
Topics	$\tau_1$	$y_{11}$	$y_{12}$	$\dots$	$y_{1q}$	$\mu_{1..}$
	$\tau_2$	$y_{21}$	$y_{22}$	$\dots$	$y_{2q}$	$\mu_{2..}$
	$\vdots$	$\vdots$	$\vdots$	$y_{ij}$	$\vdots$	$\mu_{i..}$
	$\tau_p$	$y_{p1}$	$y_{p2}$	$\dots$	$y_{pq}$	$\mu_{p..}$
		$\mu_{.1}$	$\mu_{.2}$	$\mu_{.j}$	$\mu_{.q}$	$\mu_{..}$



# Estimators



- Grand mean

$$\hat{\mu}_{..} = \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q y_{ij}$$

- Marginal mean of the  $j$ -th system

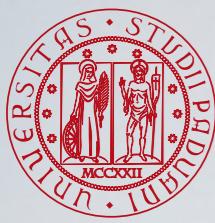
$$\hat{\mu}_{\cdot j} = \frac{1}{p} \sum_{i=1}^p y_{ij} \quad \hat{\alpha}_{\cdot j} = \hat{\mu}_{\cdot j} - \hat{\mu}_{..}$$

- Predicted score

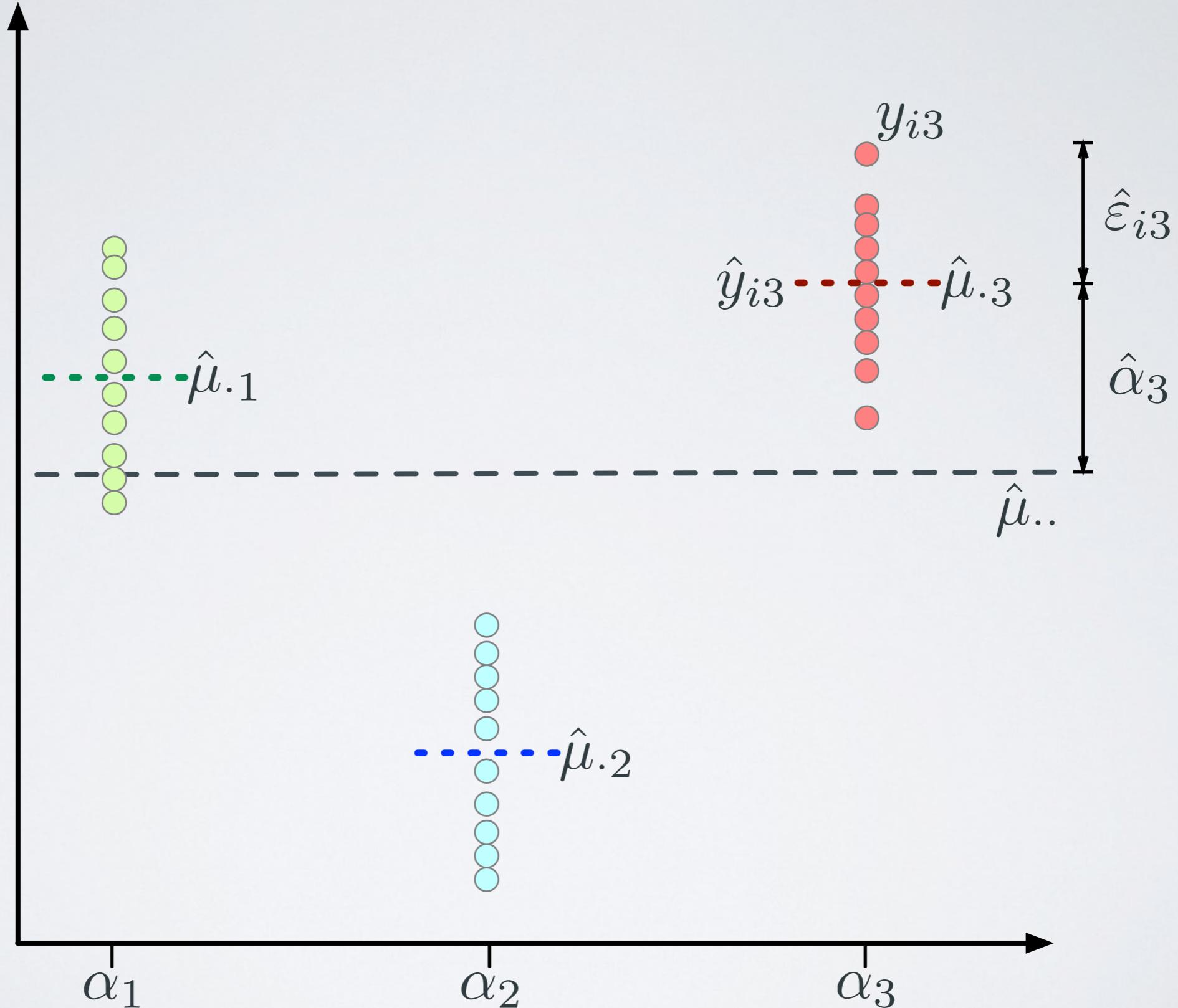
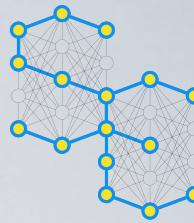
$$\hat{y}_{ij} = \hat{\mu}_{..} + \hat{\alpha}_{\cdot j} = \hat{\mu}_{\cdot j}$$

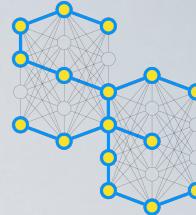
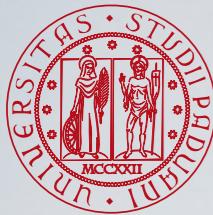
- Prediction error

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \hat{\mu}_{\cdot j}$$



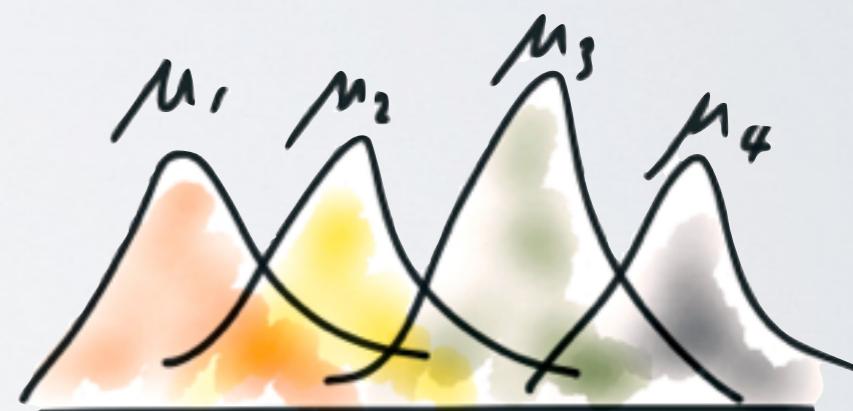
# Estimators





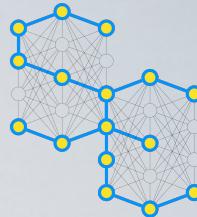
# Assessment: ANOVA

- Analysis of Variance (ANOVA) was developed by statistician and evolutionary biologist Ronald Fisher (1890-1962)
- It provides a statistical test of whether or not the means of several groups are equal
- $H_0$  is the **null hypothesis** that all the means are equal
- It partitions the observed variance in a particular variable into components attributable to different sources of variation



ANOVA

$\mu_1 = \mu_2 = \mu_3 = \mu_4 ?$



# Assessment: Sum of Squares

$$y_{ij} - \hat{\mu}_{..} = \underbrace{\hat{\mu}_{.j} - \hat{\mu}_{..}}_{\text{System Effect } \hat{\alpha}_j} + \underbrace{y_{ij} - \hat{\mu}_{.j}}_{\text{Error Effect } \hat{\varepsilon}_{ij}}$$

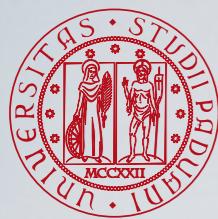
Total Effect      System Effect  $\hat{\alpha}_j$       Error Effect  $\hat{\varepsilon}_{ij}$

- Sum of squares (SS)

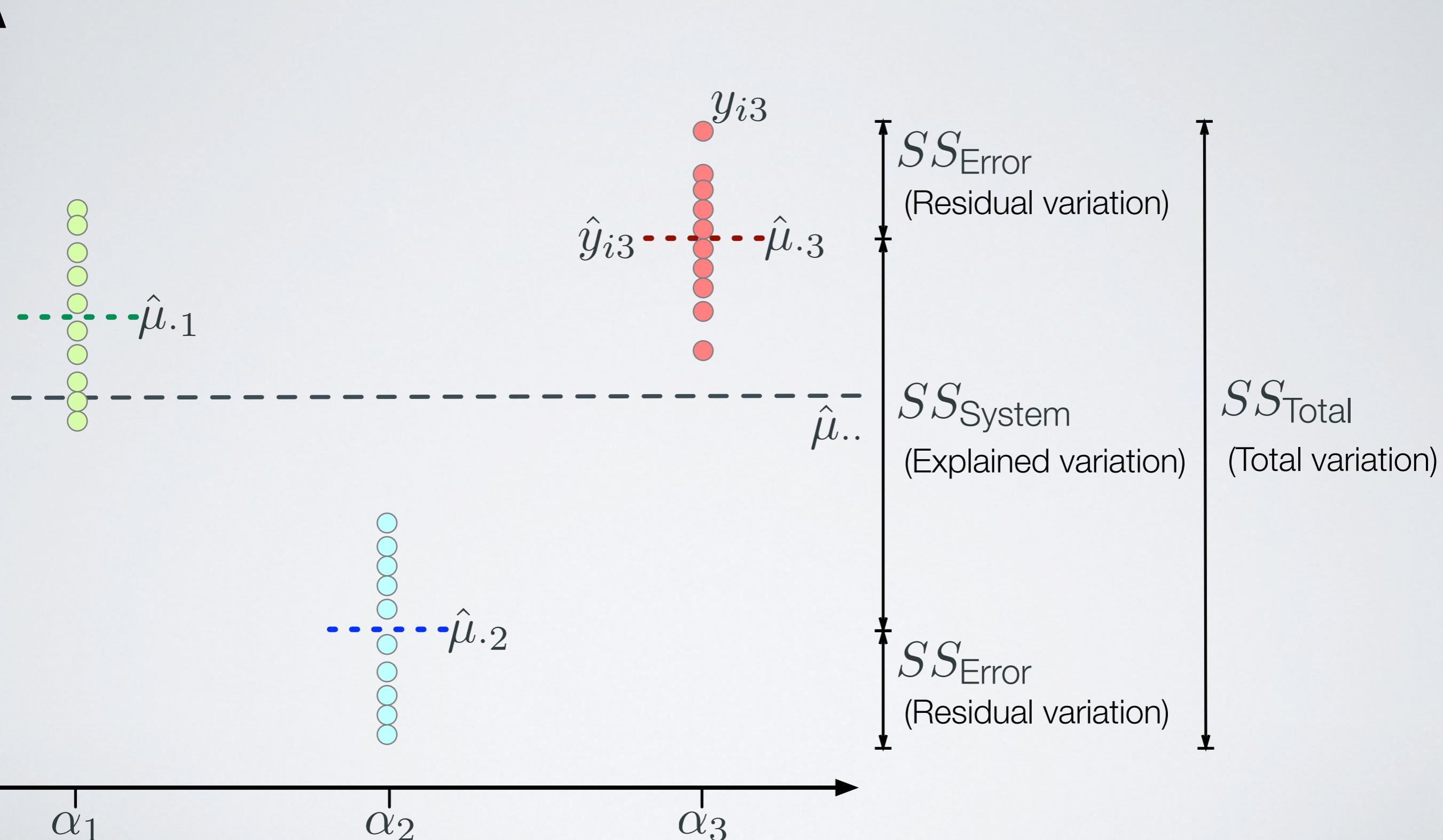
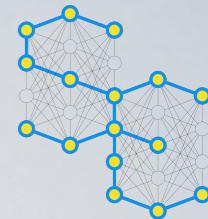
$$SS_{fact} = \sum_{i=1}^p \sum_{j=1}^q ([\text{total} | \text{system} | \text{error}] \text{ effect})^2$$

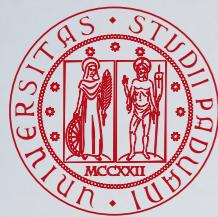
- Variance break-down

$$SS_{\text{Total}} = SS_{\text{System}} + SS_{\text{Error}}$$

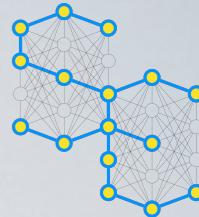


# Assessment: Sum of Squares





# Assessment: Degrees of Freedom and Mean Squares



## ● Degrees of Freedom (DF)

$$DF_{\text{Total}} = pq - 1$$

$$DF_{\text{System}} = q - 1$$

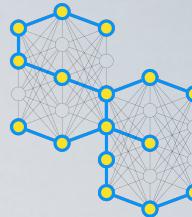
$$DF_{\text{Error}} = (pq - 1) - (q - 1) = q(p - 1)$$

## ● Mean Squares (MS)

$$MS_{fact} = \frac{SS_{fact}}{DF_{fact}}$$

Walker, H. M. (1940). Degrees of freedom. *Journal of Educational Psychology*, 31(4):253–269.

# Assessment: F-test



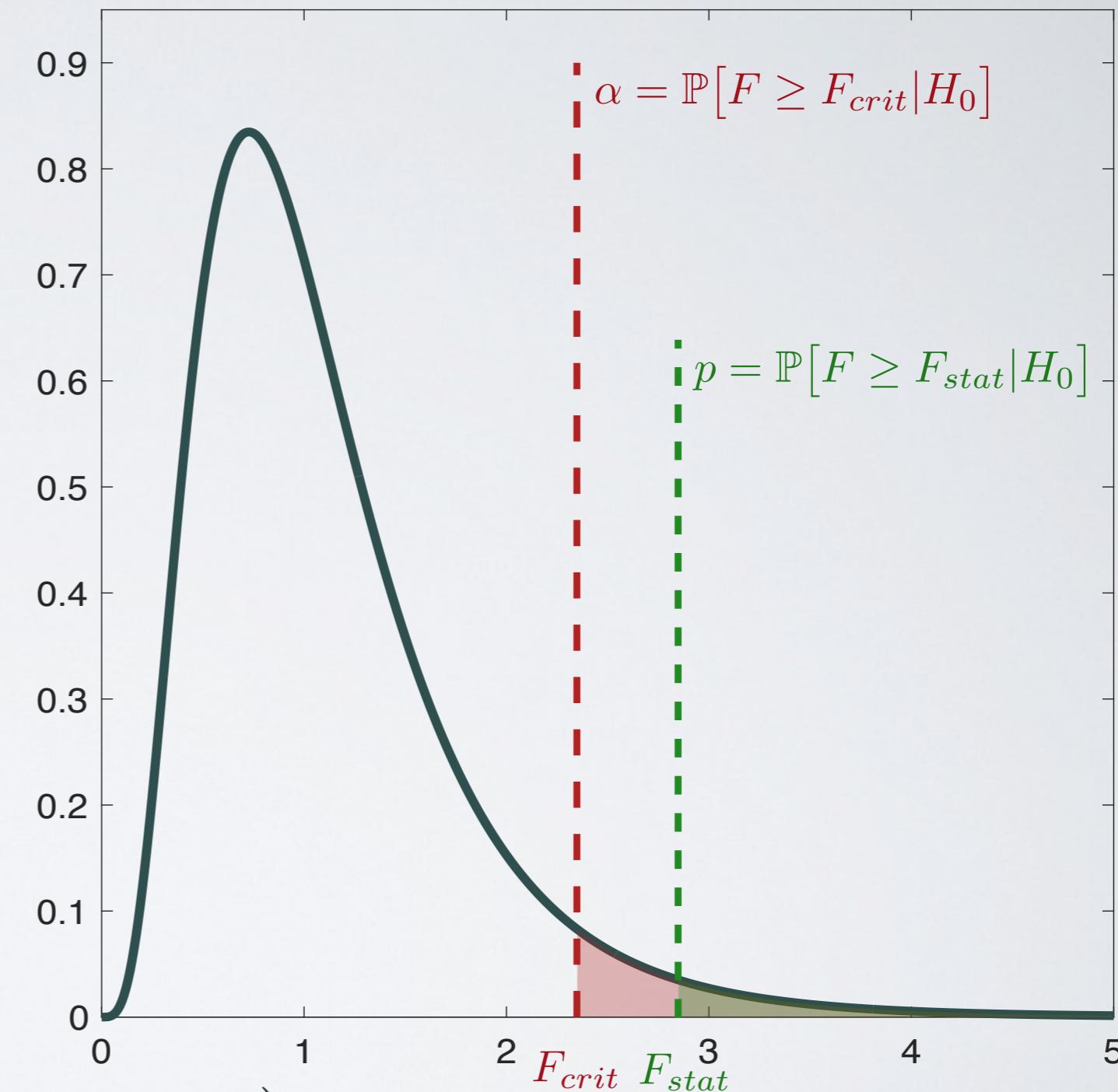
- Let us assume to have a random sample of size  $n = pq$  (**independence**) from  $q$  normally-distributed random variables with **same variance** (homoskedasticity)

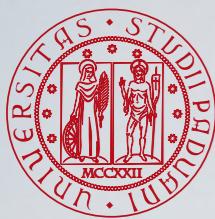
$$Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$

- Under the null hypothesis  $H_0 : \mu_1 = \mu_2 = \dots = \mu_q$

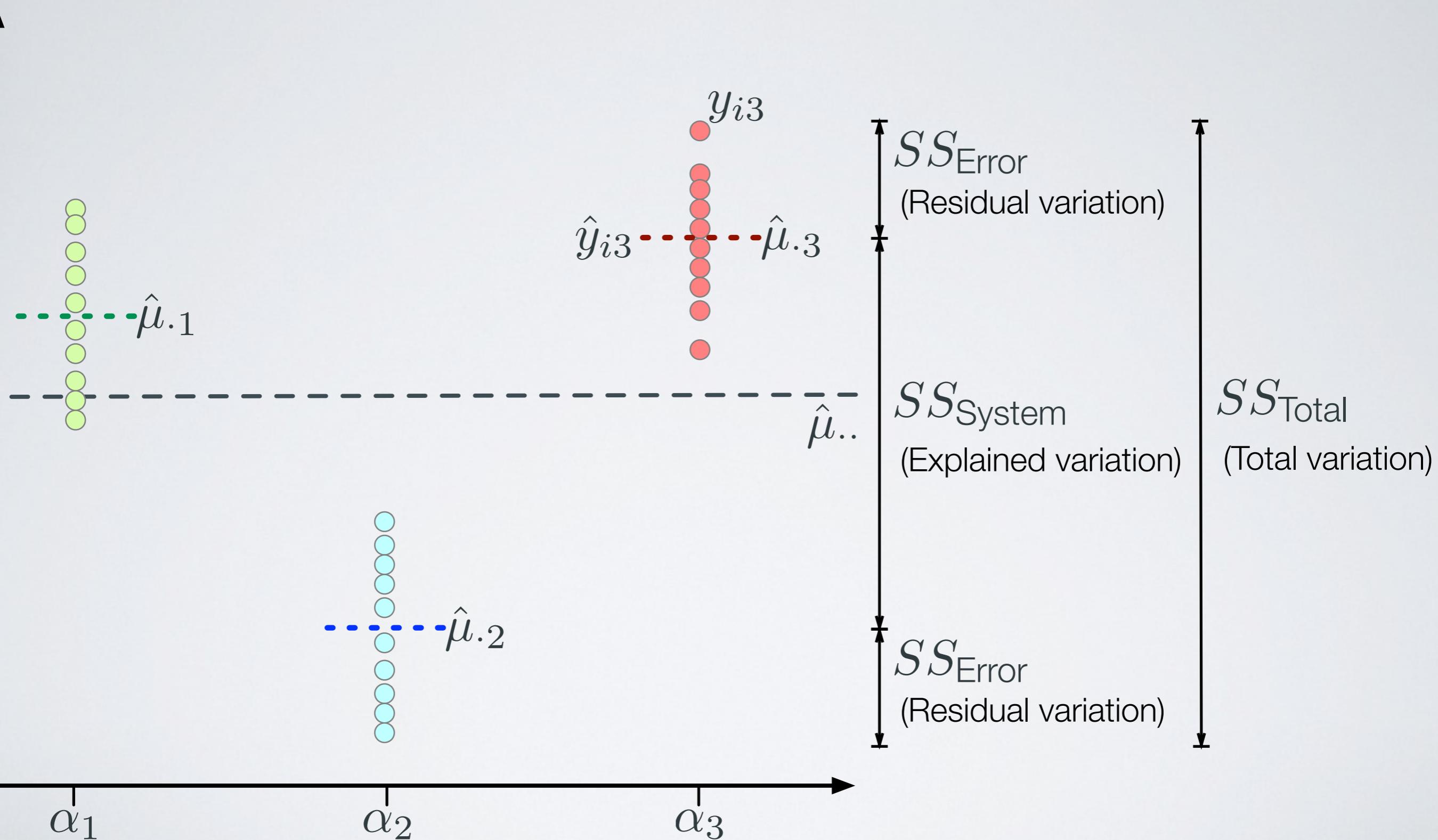
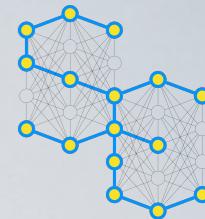
the tests statistic is

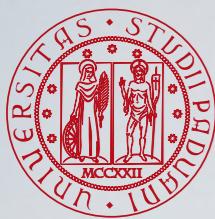
$$F_{stat} = \frac{MS_{System}}{MS_{Error}} \sim F_{(DF_{System}, DF_{Error})}$$



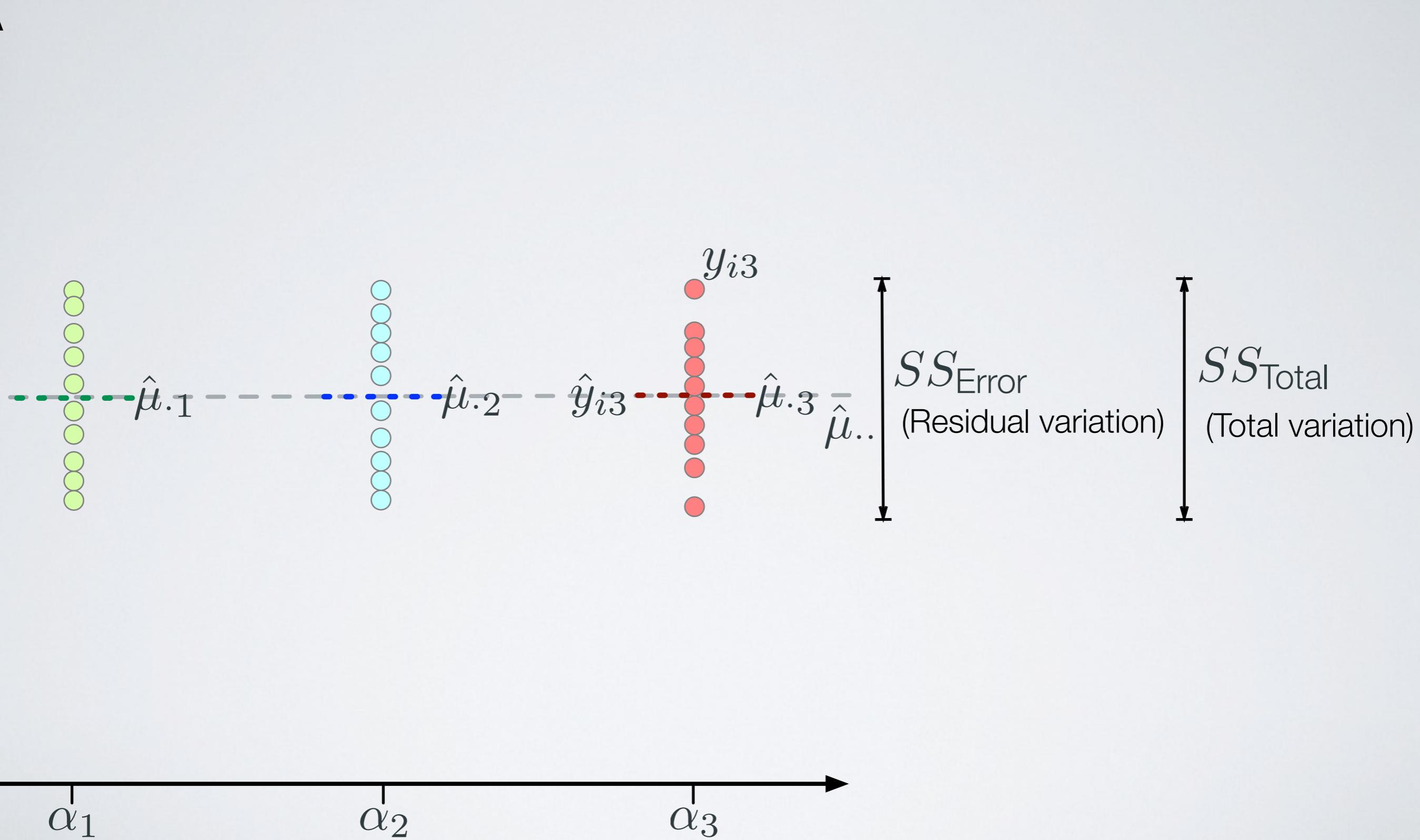
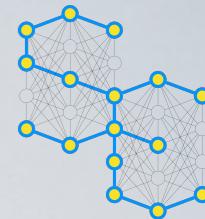


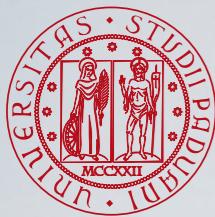
# Assessment: F-test under $H_0$



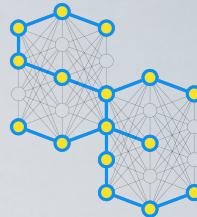


# Assessment: F-test under $H_0$





# ANOVA Assumptions?



- **Independence:** topics and systems can be considered (reasonably) independent
- **Normality:** typical IR measures are bounded in  $[0, 1]$  so they cannot be normal
  - the normal distribution is unbounded
- **Homoskedasticity:** variance changes across systems
- ANOVA is considered robust to violations of normality and also homoskedasticity when the sample sizes are equal and large (our typical case)

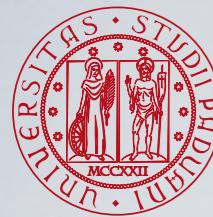


Eisenhart, C. (1947). The Assumptions Underlying the Analysis of Variance. *Biometrika*, 3(1):1–21.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In Krishnaiah, P. R., editor, *Handbook of Statistics – Analysis of Variance*, volume 1, pages 199–236. Elsevier, The Netherlands.

Tague-Sutcliffe, J. M. and Blustein, J. (1994). A Statistical Analysis of the TREC-3 Data. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.

Carterette, B. A. (2012). Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)*, 30(1):4:1–4:34.



# Tukey HSD Test

- The Tukey Honestly Significant Difference (HSD) test creates confidence intervals  $|t|$  for all pairwise differences between factor levels, while controlling the family error rate

$$|t| = \frac{|\hat{\mu}_{\cdot u} - \hat{\mu}_{\cdot v}|}{\sqrt{\frac{2MS_{\text{Error}}}{p}}} > \frac{1}{\sqrt{2}} q_{\alpha, q, q(p-1)}$$

- $\hat{\mu}_{\cdot u}$  and  $\hat{\mu}_{\cdot v}$  are the marginal means of the two factor levels, i.e. the two systems to be compared
- $q_{\alpha, q, q(p-1)}$  is the upper  $100 * (1 - \alpha)$ -th percentile of the **studentized range distribution**, i.e. the distribution of the range of samples drawn from a normal distribution, considering  $q$  systems to compare using  $p$  topics

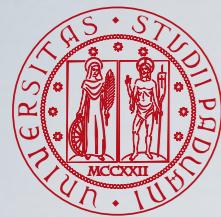


John Wilder Tukey

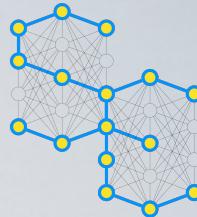
Tukey, J. W. (1949). Comparing Individual Means in the Analysis of Vari- ance. *Biometrics*, 5(2):99–114.

Newman, D. (1939). The Distribution of Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. *Biometrika*, 31(2):20–30.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, USA.



# Compute it Yourself!



## Computation of one-way ANOVA on simulated data

This example shows step-by-step computation of one-way ANOVA on simulated data.

### Setup

Define the parameters to be used.

```
% set the random number generator for reproducibility  
rng(6, 'twister');  
  
% total number of samples to be used  
n = 10;  
  
% the significance level  
alpha = 0.05;
```

Generate the random (simulated) data from normal distributions with same variance as follows:

- $x \sim \mathcal{N}(1, 1)$
- $y \sim \mathcal{N}(1, 1)$
- $z \sim \mathcal{N}(2.3, 1)$

```
x = 1 + randn(n, 1);  
y = 1 + randn(n, 1);  
z = 2.3 + randn(n, 1);
```

Put everything together in a single matrix

```
data = [x y z];  
  
% the number of rows and columns  
[p, q] = size(data);
```

### Compute the one-way ANOVA

We consider the following one-way ANOVA model

$$y_{ij} = \mu_{..} + \alpha_j + \varepsilon_{ij}$$

where  $y_{ij}$  is an element of the data matrix.

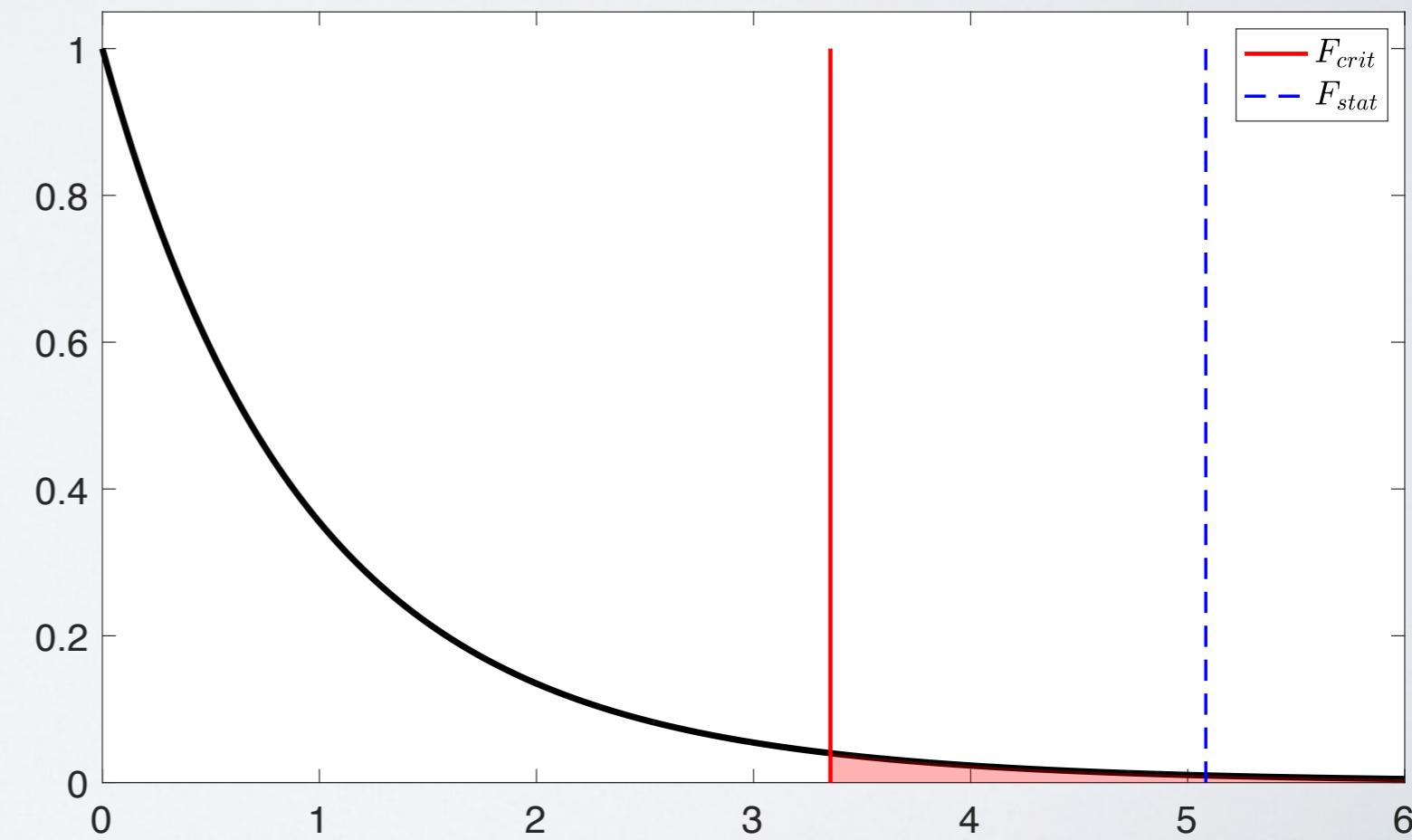
### Fit the ANOVA Model

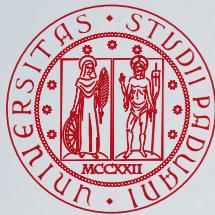
- Grand mean:

$$\hat{\mu}_{..} = \frac{1}{pq} \sum_{i=1}^p \sum_{j=1}^q y_{ij}$$

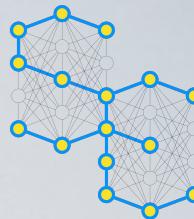
	1	2	3	4	5	6
1	'Source'	'SS'	'DF'	'MS'	'F'	'Prob>F'
2	'Columns'	10.6516	2	5.3258	5.0836	0.0134
3	'Error'	28.2863	27	1.0476	[]	[]
4	'Total'	38.9379	29	[]	[]	[]

ANOVA Table





# Compute it Yourself!



## Computation of one-way ANOVA on TREC 08, 1999, AdHoc

This example shows how to compare runs from TREC 08, 1999, AdHoc using one-way ANOVA.

### Setup

Load the performance score about the runs. The file contains:

- ap (or another measure name): a matrix where rows are topics and columns are runs; each cell contains the AP score for a run on a given topic;
- runs: an array of strings containing the identifiers of the runs; elements in the array are in the same order as columns on ap;
- topics: an array of strings containing the identifiers of the topics; elements in the array are in the same order as rows on ap;

```
% Load Average Precision data
% Other measures available are: P@10, nDCG, RBP
load ../../data/T08/ap.mat

% rename ap into measure to make follow-up processing the same for whatever
% measure we load
measure = ap;

% remove the now useless ap matrix
clear ap;

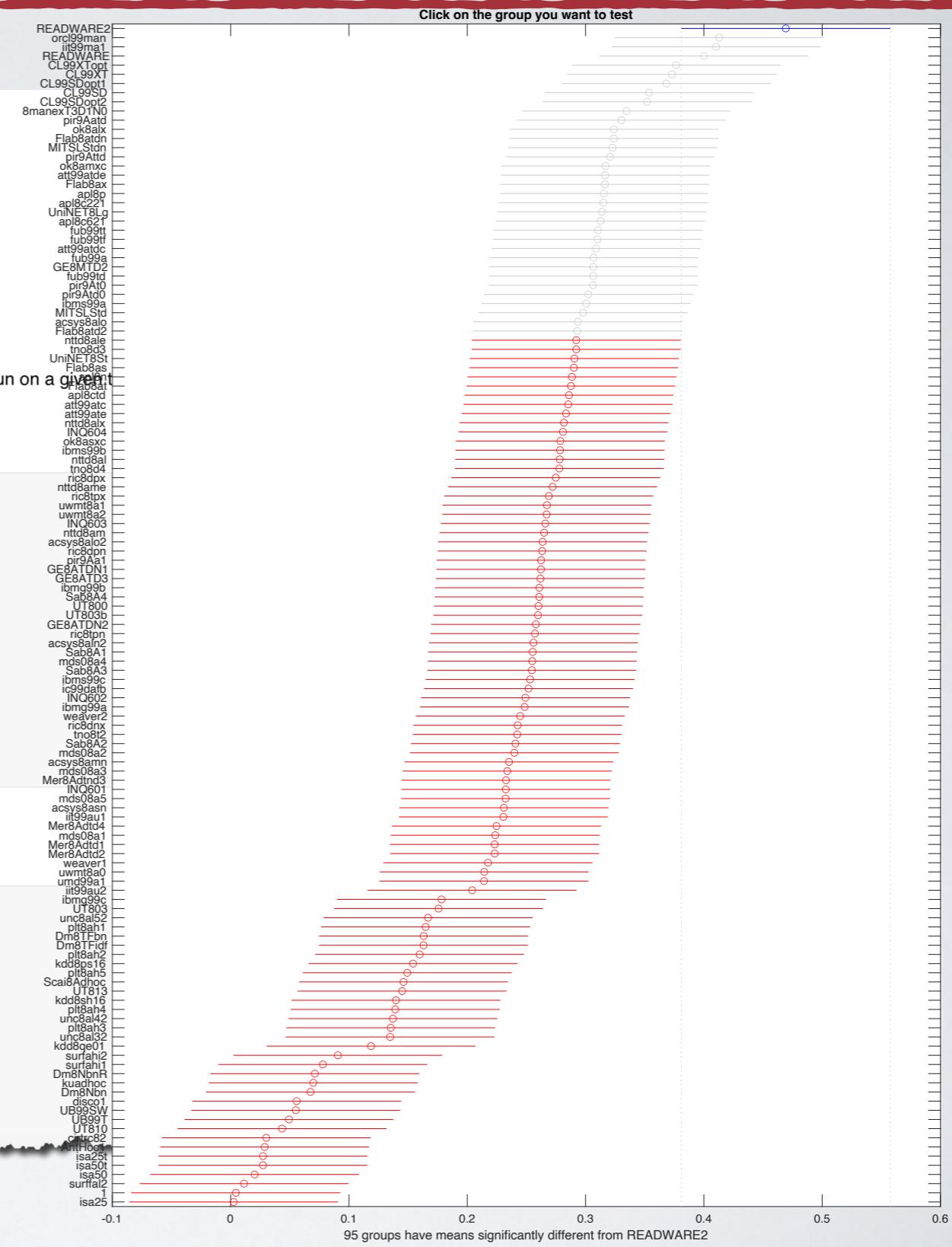
% the significance level
alpha = 0.05;
```

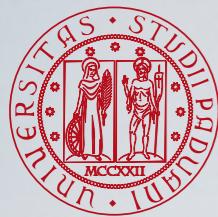
Reorder the measure by mean performance

```
% the mean for each run across the topics
% Note that if the measure is AP (Average Precision),
% this is exactly MAP (Mean Average Precision) for each run
m = mean(measure);

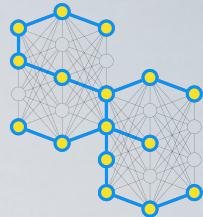
% sort in descending order of mean score
[~, idx] = sort(m, 'descend');

% re-order runs by descending mean of the measure
measure = measure(:, idx);
runs = runs(idx);
```





# What About Two-Way ANOVA?



$$y_{ij} = \mu_{..} + \tau_i + \alpha_j + \varepsilon_{ij}$$

Topic Effect

System Effect

- Check-out `anova2` in Matlab
- When runs on two systems it is basically equivalent to a (paired) Student's t test



Tague-Sutcliffe, J. M. and Blustein, J. (1995). A Statistical Analysis of the TREC-3 Data. In Harman, D. K., editor, *The Third Text REtrieval Conference (TREC-3)*, pages 385–398. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA.

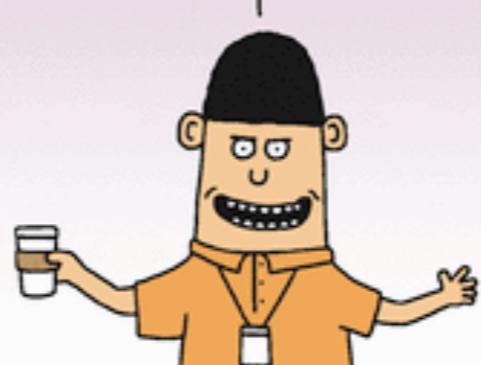
# questions?

YOU KNOW HOW STUDIES  
ALWAYS SAY ONE OUT  
OF TEN PEOPLE HAVE A  
PARTICULAR PROBLEM?



Dilbert.com DilbertCartoonist@gmail.com

I'M ALWAYS THAT  
GUY. STATISTICALLY  
SPEAKING, I KEEP NINE  
PEOPLE SAFE JUST BY  
EXISTING.



THAT'S  
NOT HOW  
STATIS-  
TICS WORK.  
AND...  
EVERYONE  
ELSE IN THE  
DEPARTMENT  
KNOWS THAT?

