Statistics for data science
Project A.Y. 2024/25

Motivation

- As a data scientist, you must be able to **understand and reproduce** innovations in the field, which are typically described in scientific papers
- **Understanding** means being able to comprehend and communicate to others theoretical and methodological innovation, *using appropriate formal statistical language*
- **Reproducing** means being able to re-implement portions of methods and algorithms, or to re-implement portions of the tests and experimental validation, *using the R programming language*
- The project of Statistics for Data Science will focus on practicing with those skills, applied to topics reported in selected papers
  - Selected papers available in Teams under the Files/Project tab
  - The list of topic papers may expand up to the end of the course

Group formation and paper assignment

Students autonomously cluster in groups of up to 3 members

- They select *at least* 5 papers and send a ranked list to salvatore.ruggieri@unipi.it
- The teacher assigns to the group one of the selected paper
- Deadline: **26 May 2025 h. 7:59 at the latest, but the sooner the better** (papers already assigned become unavailble for other)

Project delivery

The group studies the assigned paper and deliver to salvatore.ruggieri@unipi.it

- A presentation of up to 15 slides
- An R script with re-implementation of some methods and/or experiments
  - Both presentation and R script may focus on a portion of the paper – *no need to cover everything*

At the time of delivery, each student must have registered to an exam date (appello), only to ensure filling the student's questionnaire

Deadlines (no extension will be granted!):

Batch	Delivery deadline
1st	30 May h. 23:59
2nd	23 June h. 23:59
3rd	12 July h. 23:59
4th	5 Sept h. 23:59

Project and oral discussions

The group gives a 20 min talk with Q&A using the presentation slides and R script

- All members of the group must be present, no split discussions
- Q&A will regard the project only

Each member of the group takes the individual oral discussion, which may include Q&A on the project as well as Q&A on the course theory/R programming

- Students must demonstrate to be able to summarize both the theory and the software related to any of the lessons using the slides and R scripts of the lessons.

Batch	Delivery deadline	Group presentation	Individual oral
1st	30 May h. 23:59	3 June h. 9:00	3 June h. 14:00
2nd	23 June h. 23:59	25 June h. 9:00	25 June h. 14:00
3rd	12 July h. 23:59	16 July h. 9:00	16 July h. 14:00
4th	5 Sept h. 23:59	9 Sept h. 9:00	9 Sept h. 14:00

Grade composition

Grade composition

- up to 5 points: slides writing, organization, clarity
- up to 5 points: R script quality, organization, documentation
- up to 5 points: project presentation (overall)
- up to 5 points: project discussion (individual)
- up to 10 points: individual oral discussion

This means

- 15 points assigned collectively to the group + 15 points assigned individually to each member of the group
- 10 points assigned to the offline work + 10 point assigned to the reporting of such work + 10 points assigned to the knowledge of the topics of the course
- 20 points assigned to the project + 10 points assigned to the oral discussion

Q&A

Q: What is the language of slides/presentation?

A: Slides must be in English, presentation can be in English or in Italian

Q: Can the project be discussed after September 2025?

A: No, students with assigned project that do not discuss within September 2025 will have to do the normal written+oral test.

Q: Can the project be discussed remotely?

A: No, only in presence. The room for the discussion will be communicated after the delivery of the project.