Politecnico di Milano – School of Industrial and Information Engineering

Applied Statistics
052498 - 052742

Academic Year 2024/2025

1 Credits

8 -10 CFU

2 Teachers and tutors

2.1 Course leader

Prof. Piercesare Secchi
MOX - Dipartimento di Matematica
La Nave, III floor
Politecnico di Milano
e.mail: piercesare.secchi@polimi.it
Webex room: https://politecnicomilano.webex.com/meet/piercesare.secchi
Office Hours (by appointment): Friday, 16:00 - 18.00.

2.2 Lab teacher

Ing. Guillaume Koechlin
MOX - Dipartimento di Matematica
La Nave, VI floor
Politecnico di Milano
e.mail: guillaume.koechlin@polimi.it
Webex room: https://politecnicomilano.webex.com/meet/guillaume.koechlin
Tutoring activities: to be announced

3 Timetable

Class & Lab:
MON, Room 3.1.5, 10:15-12:15 (Lab&Lecture),
TUE, Room 3.1.4, 11:15-13:15 (Lecture&Lab),
THU, Room 3.1.12, 11:15-13:15 (Lecture&Lab),
FRI, Room 3.1.3, 10:15-12:15 (Lecture&Lab).

4 Web page

The course web page is here:

https://webeep.polimi.it/course/view.php?id=16791

Short-term notices, source code for the lab sessions, test results, open problems, etc., will be regularly posted on the course web page.

5 Course Program

The topics covered by the 8 CFU versions of the course are the following:

1. *Exploring a multivariate dataset.* Descriptive statistics and graphical displays. The geometry of a multivariate sample. Sample mean, covariance and correlation. Generalized variance and total variance. The metric induced by the covariance matrix.

2. *Data representation and dimensional reduction.* The analysis of the covariance structure, principal component analysis (PCA).

3. *Classification: discrimination and clustering.* Statistical classification: model, misclassification costs and prior probability. Bayesian supervised classification and the Fisher approach to discriminant analysis. Cross-validation for the evaluation of a classifier. Alternative approaches to classification: CART, support vector machines. Similarity measures. Unsupervised classification; hierarchical and nonhierarchical methods. DBSCAN. K-means and K-medoids. Multidimensional scaling.

4. *Inference about mean vectors.* The multivariate normal distribution, the Wishart distribution, the F distribution. Hotelling T^2 test. Confidence regions and simultaneous comparisons of component means. The Bonferroni method for multiple comparisons. Familywise Error Rate and False Discovery Rate. Comparisons of several multivariate means. ANOVA and MANOVA. Inference for Linear Models. Beyond Ordinary Least Squares: ridge regression, lasso, regularized least squares. Random effects and mixed effects linear models.

In the 10 CFU version of the course, the above topics are complemented with the following two modules of Advances in Statistical Learning:

5. *Introduction to Functional Data Analysis.* Data smoothing, dimensional reduction and representation. Functional principal component analysis. Data registration: phase and amplitude variability. Classification of functional data.

6. *Statistics for spatial data.* Random fields, variogram models and variogram fitting. Spatial prediction and Kriging, Functional data with spatial dependence.

6 Lab sessions and data analysis project

Methods and algorithms will be illustrated in the lab sessions through applications to real data sets; analyses will be performed in R, an open-source package for statistics downloadable at

www.r-project.org

Students shall actively participate in the lab sessions. **All students** – those taking the course for 8CFU and those taking the course for 10CFU – **must** work in a team on a data analysis project developed along the course: each team shall show the project work in progress during routinely scheduled meetings with all other teams.

6.1 Data analysis project

Every student taking the course **must** participate in a data analysis project developed by an **independently** formed team of **3-5 members**. The work in progress of the projects will be collectively discussed during a general meeting to be held **Tuesday, the 8th of April, 2025**. Final analyses and results will be presented in a workshop, which will take place a day **in June 2025 to be collectively decided**.

Data sets available for the team projects will be presented in class **Thursday, the 27th of February**.

Before March 10, each team should send an email to the students

(1) Sara Auletta (*sara.auletta@mail.polimi.it*),

(2) Alisa Pesotskaia (*alisa.pesotskaia@mail.polimi.it*)

containing the following information:

(a) name and email address of the team leader;

(b) the title of the project;

(c) the list of the team members, their names and personal code;

(d) max 5 lines of abstract with a short description of the data set analyzed and the temporary goals of the project; these could always be updated and modified while the project is under development.

This information will be made public on the course web page as soon as the students (1) and (2) will organize it in a file to be sent to Ing. Guillaume Koechlin.

6.2 Open workshop for project presentations

Teams will show the final results and analyses of their projects during an open workshop that will take place after the end of classes (June 2025).

The workshop will consist of a speedy pitch session (2 minutes per team), during which each team very briefly presents its project. This will be followed by a poster session, during which each team will illustrate an A0 poster reporting the results of their analysis to all participants, students, and teachers.

The projects presented at the final workshop will be collectively evaluated **by the course students participating – in presence – to the entire workshop and by the course teachers**.

The teams students (1) and (2) belong to is in charge of the organization of the work-in-progress meeting and of the final workshop (program schedule of the event, chairing the pitch session, presiding over food and drinks for coffee breaks etc.).

7 Exam

The exam consists of two parts:

(a) A written exam. The written exam will feature multiple-choice questions, that assess both theoretical knowledge and practical skills, along with several data analysis problems to be individually solved with R; two problems for the students following the 8 CFU version of the course, three problems - with extra time - for those registered in the 10 CFU version. For the students taking the course for 10 CFU, the extra problem will be related to the two topics treated in the modules characterizing their additional 2 CFU; working on this problem is **mandatory** to pass the exam of the 10 CFU version of the course. For all students, the use of a personal computer is allowed, as well as that of books, personal notes, etc. (it's an open-book written exam).

In the written exam the student must show the ability to conduct a stylized data analysis, by selecting the appropriate methods and algorithms - among those introduced in the course - for solving the problems, by running the algorithms with R, by identifying the significant results and by reporting them with the precision and property of language which characterize the technical and scientific communication.

(b) Team project evaluation. Projects will be collectively evaluated by the teachers of the course and by the students participating – in presence – in a final workshop at the end of the course. **The grade of the project expires after the 5th exam session, that is at the beginning of the second semester of the Academic Year 2025/2026.**

During the presentation of their projects, teams must prove their ability to conduct and report a real life statistical data analysis, showing knowledge and understanding of the problem at hand and the nature of the

data, proper judgment for the selection of the appropriate methods and algorithms, sensible interpretations of the generated results - grasping not only their strengths but also their weaknesses - and, finally, communication skills when informing an audience of peers.

To pass the exam, students must pass **both part (a) and part (b) with a score greater than or equal to 18/30**; their final grade is then obtained as the weighted average of the two scores, with weights respectively equal to 0.6 for the written exam and 0.4 for the project evaluation.

8 Course bibliography

required Johnson, R.A. and Wichern, D.W., (2007). *Applied Multivariate Statistical Analysis (sixth edition)*, Prentice Hall, Upper Saddle River

required James G., Witten D., Hastie T. and Tibshirani R. (2013). *An introduction to statistical learning, with application to R*, Springer, New York (`http://www-bcf.usc.edu/~gareth/ISL/`)

required Hastie T., Tibshirani R. *Statistical Learning MOOC*, `https://www.edx.org/course/statistical-learning`.

suggested Ramsay, J.O. e Silverman, B.W., (2005). *Functional Data Analysis (second edition)*, Springer Series in Statistics, Springer, New York

additional Galecki A. e Burzykowski T. (2013). *Linear Mixed-Effects Models using.* Springer Texts in Statistics, Springer, New York

additional Cressie, N. (1993). *Statistics for Spatial Data (Revised Edition)*, John Wiley & Sons

additional Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: data mining, inference and prediction. (Second Edition)*, Springer-Verlag, New York.

additional Ramsay, J.O. e Silverman, B.W., (2002). *Applied Functional Data Analysis: methods and case studies*, Springer Series in Statistics, Springer, New York