COURSE OF STUDY	**TWO-YEAR MASTER OF SCIENCE PROGRAMME IN MATHEMATICS**
ACADEMIC YEAR	**2023-2024**
ACADEMIC SUBJECT	**MATHEMATICAL STATISTICS**

General information	
Programme year	Second
Term	First semester (September 25, 2023 – December 22, 2023)
European Credit Transfer and Accumulation System credits (ECTS)	7
SSD	MAT/06 – Probability and Mathematical Statistics
Language	Italian
Mode of attendance	Not mandatory

Lecturer	
Name and surname	Marcello De Giosa
E-mail	marcello.degiosa@uniba.it
Telephone	+39 080 544 2707
Department and office	Department of Mathematics, floor 4, room 12
Virtual meeting room	
Web page	https://www.dm.uniba.it/it/members/degiosa
Office hours	Wednesday, from 10 a.m. to 12 a.m., by email appointment.

Work schedule				
	Total	Lectures	Hands-on learning (recitations)	Self-study
Hours	175	40	30	105
ECTS credits	7	5	2	

Learning objectives	
	Acquiring knowledge of the main methods of Mathematical Statistics, their application to solve real world problems and related modern R code.

Course prerequisites	
	Calculus, multivariate calculus, linear algebra, elementary probability, as usually offered during a degree in Mathematics of L-35 class.

Syllabus	
Course contents	**Part I. Modern Computer Intensive Statistics.** **Introduction to data.** Sampling: principles and strategies. Experiments. Observational studies. **Exploratory data analysis.** Exploring categorical data, R code. Exploring numerical data, R code, Numerical summaries. Case study with R. **Regression modelling.** Linear regression with one predictor. Linear regression with multiple predictors. Logistic regression. **Foundations of Inference.** Hypothesis testing with randomization. Confidence intervals with bootstrapping. Inference with mathematical models. Decision errors. Case study: Malaria vaccine. **Statistical Inference.** Inference for a single proportion: bootstrap test and

mathematical model. Inference for comparing proportions: randomization test, bootstrap confidence intervals and mathematical model. Inference for two-way table: randomization test of independence and mathematical model. Inference for a mean: bootstrap confidence interval and mathematical model. Inference for comparing two means: randomization test, bootstrap confidence intervals and mathematical model. Inference for paired means: randomization test, bootstrap confidence intervals, mathematical model. Inference for many means: randomization test and mathematical model. Case study and R tutorials.

Inference for linear regression with a single predictor. Case study. Randomization test and bootstrap confidence intervals for the slope. Mathematical model for hypothesis testing and confidence intervals for the slope. Checking model assumptions.

Inference for linear regression with multiple predictors. Output from software, multicollinearity, cross-validation for prediction error.

Inference for logistic regression. Model diagnostic. The software output. Cross-validation for prediction error.

Part II. Mathematical insights into the modelling approach.

Review of the linear model. Key linear model concepts modelling the age of the universe: least squares point estimate, confidence intervals, hypothesis testing for the slope. The linear model in general. R code. Three useful fundamental theorems: chi-squares, t-student and Fisher distributions.

Theory of linear model. Least squares estimation by QR decomposition. Distributional properties of estimators. Variance estimation. From QR decomposition to hypothesis testing and confidence intervals. From Cholesky decomposition to Fisher hypothesis testing. From QR decomposition to Fisher distribution in the sequential Analysis of Variance (ANOVA) for comparing nested models. From QR decomposition to influence matrix. Distributional properties of fitted values and residuals. The results in terms of model matrix. The R code on a real data example (Galapagos). The Gauss-Markov theorem. Geometric interpretation of least squares, QR decomposition and nested models.

Practical linear models with R code. Basic code, diagnostic plots, R-squared, model selection, Akaike Information Criteria (AIC), predictions, collinearity.

Practical linear model with Factor. Identifiability problem. Multiple factor and interactions.

R specifications for linear models.

Further insights on linear models' theory. Constraints, contrasts and factors. Likelihood. AIC and Mallow's statistics.

Practical linear models with R code. Simple linear model. Real data examples. Mean estimation and prediction of a new observation. The multiple model. Collinearity and variance inflation. Model selection: AIC< hypothesis testing, Fisher test. Coefficients interpretation. Customer satisfaction analysis in an amusement park.

Review of logistic regression. Binomial and Bernoulli distribution. Odds, logit, logistic, odds-ratio. Binary predictor. Continuous predictor. Parameters interpretation. Model selection. Graphical model representation.

Multiple logistic regression. Maximum likelihood and Iteratively reweighted least squares (IRLS). Deviance. AIC. Likelihood ratio test (LRT). Residuals. Dispersion parameter estimation.

Generalized Linear Models (GLM). Two introductory example: e[idemic and predators. The exponential family. Mean and variance of exponential distributions. General formulation of GLM. Parameters estimations and

	asymptotic distribution of estimators. Models comparison with AIC and deviance. Residuals. Practical GLM with R. A binomial model for the probability of heart attack. Poisson regression in epidemic models.
Reference books	S.N. Wood - *Generalized Additive Models* - 2017 - CRC Press M. Cetinkaya-Rundel, J. Hardin - *Introduction to Modern Statistics* - 2023 - OpenIntro R.S. Kenett, S. Zacks, P. Gedeck - *Modern Statistics. A computer-based approach with Python* - 2022 - Birkhauser A. Agresti, C.A. Franklin, B. Klingenberg - *Statistics. The art of learning from data.* - 2023 - Pearson A. Agresti, M. Kateri - *Foundations of Statistics for Data Scientists* - 2021 - CRC Press G. Casella, R.L. Berger – *Statistical Inference* – 2002 - Duxbury
Additional course materials	
Repository	

Expected learning outcomes	
Knowledge and understanding	Acquiring the main concepts, techniques and R code understanding for Mathematical Statistics
Applying knowledge and understanding	Being able to apply the theoretic results and R code to solve real world problems with Mathematical Statistics.
Soft skills	*Making judgements*: Being able to choose the right Mathematical Statistics tools and R code.
	Communication skills: Being able to report and explain the Mathematical Statistics methods and results of statistical analysis.
	Learning skills: Being able to apply learning methods to improve knowledge in the field of Mathematical Statistics.

Teaching methods	
	In person frontal theoretical lessons and R laboratory with use of slides.

Assessment	
Assessment methods	The final exam consists of an oral part and an R laboratory part.
Evaluation criteria	• *Knowledge and understanding*: mastering and deep understanding of the Mathematical Statistics theory and methods and their translation in modern R code. • *Applying knowledge and understanding*: being able to apply the Mathematical Statistics methods and related R code to solve practical real world problems. • *Making judgement*: mastering the ability to choose the right Mathematical Statistics methods and R code to solve real problems. • *Communication skills*: ability to explain the theoretical foundation of applied methods and the statistical analysis results in a clear way. • *Learning skills*: well organizing the acquired knowledge and being able to autonomously find sources of insight.

Grading policy	The exam consists in the presentation of some methods among those proposed during the classes and the discussion of some R code chosen at random from those presented during the classes. The vote, out of thirty, will depend on the mastery of the topics, knowledge of the codes, but also on the ability to clearly explain results.

Further information	