DEGREE COURSE IN STATISTICS
ACADEMIC YEAR 2023 - 2024
MULTIVARIATE STATISTICS

General information	
Year of the course	III
Academic calendar (starting and ending date)	First term (11/09/2023 – 15/12/2023)
Credits (CFU/ETCS):	10 CFU
SSD	Statistica, SECS-S/01
Language	Italian
Mode of attendance	Strongly suggested

Professor	
Name and Surname	Alessio Pollice
E-mail	alessio.pollice@uniba.it
Telephone	080 504 9267
Department and address	Room n. 3, 5-th floor
Virtual room	MS Teams channel “Prof. A. Pollice - Ricevimento studenti”, code: y7zenm7
Office Hours (and modalities: e.g., by appointment, on line, etc.)	Tuesday 9.30 – 11.30, Friday 9.30 – 11.30, or by appointment.

Work schedule			
Hours			
Total	Lectures	Hands-on (laboratory, workshops, working groups, seminars, field trips)	Out-of-class study hours/Self-study hours
250	46	24	180
CFU/ETCS			
10			

Learning Objectives	Understanding and knowledge of the theoretical foundations and methodological developments of multivariate inference, linear models and multidimensional data analysis. Familiarity and autonomy in the application of the aforementioned methods with the aid of the R software.
Course prerequisites	• Notions of mathematical analysis; • Notions of matrix algebra; • Notions of probability theory; • Notions of statistical inference.

Teaching strategies	• The Multivariate Statistics course provides an introduction to statistical techniques that allow the study of multiple variables, in order to make inferences on their joint distribution, study the relationship between variables, the similarity between statistical units and represent the statistical units and/or variables in a space of reduced dimensionality. The lessons are accompanied by exercises in which the theoretical topics covered are applied, using the statistical software R, to real examples and case studies. • Frontal lessons on theoretical topics (about 46 hours) and exercises on the same topics using the R software (about 24 hours). If necessary, lessons and exercises can be provided in distance learning mode. • Course materials, self-assessment tests and exercises on the e-learning platform.

	• Self-assessment tests (multiple choice questions for each chapter of the course) are provided for the purpose of familiarizing with the procedures for carrying out the (partial) exams and are to be taken individually within the pre-established time limits. The outcome of the self-assessment tests helps to improve the overall evaluation of the commitment of the students enrolled in the course.

Expected learning outcomes in terms of	
Knowledge and understanding	The expected learning outcomes coincide with the acquisition of skills on the various chapters of the detailed program: understanding and knowledge of the theoretical foundations and methodological developments related to multivariate inference, linear models and multidimensional data analysis.
Applying knowledge and understanding	Familiarity in applying multivariate statistics methods to the analysis of economic data or data from other application contexts with the aid of R software.
Soft skills	• *DD3 - Making informed judgments and choices* Autonomy in the choice of multivariate statistics methodologies and in the evaluation of the results of their application with the R software to exercises and case studies referring to the economic context or to other application fields; • *DD4 - Communicating knowledge and understanding* Ability to communicate to specialists and non-specialists the theoretical contents of the discipline, the reasons for the choices to be made for carrying out exercises and examples and the interpretation of the results of the analyses conducted with the R software; • *DD5 - Capacities to continue learning* Autonomy in learning the topics of multivariate statistics and in the practice of using the R software in order to enhance skills and competences in subsequent studies and in work activity.

Content knowledge (Syllabus)	
	First part: Multivariate inference and linear models. • Discrete and continuous multidimensional random variables. Stochastic independence. Expected values. Variance/covariance matrices. Covariance matrices. Moment generating functions. • Multivariate Normal distribution and its parameters. Bivariate Normal distribution. Standardization. Moment generating function. Properties of the multivariate Normal. Wishart's and Hotelling's Distributions. • Inference on the parameters of the multivariate Normal. Maximum likelihood estimation. Sampling distributions of maximum likelihood estimators. Multivariate central limit theorem. Multivariate tests: union-intersection principle, generalized likelihood ratio. Hotelling's test. Hypothesis testing on the variance/covariance matrix. • General linear model. Multiple linear regression. Parameter estimation with the least squares method. Properties of estimates. Parameter estimation with the maximum likelihood method. Sum of squares and R^2. Hypothesis testing and confidence intervals for regression coefficients. Forecasts using the linear model. Removal of assumptions, analysis of residuals, intrinsically linear models, choice of explanatory variables, generalized least squares (heteroskedasticity and 1st order autocorrelation), multicollinearity, ridge estimators. Analysis of variance. Analysis of covariance.. • Generalized linear models. Exponential family, score and total score functions, canonical exponential family. Definition of GLM and generalities. Estimation of GLM parameters (Newton-Raphson and scoring methods), sampling distribution of the estimators. Model assessment. Quasi likelihood (outline).

	Second part: Multivariate data analysis. • Discriminant analysis. Fisher's linear discriminant function. Maximum likelihood discriminant function. Bayesian discriminant analysis. Least cost of misclassification. Estimation of the misclassification probability (outline). • Principal component analysis. Definition of principal components, sampling properties. Application problems and interpretation of the principal components. Choosing the number of principal components.. • Canonical correlation analysis. Definition of canonical components, sampling properties. Hypothesis testing. • Factor Analysis. Factor model. Model estimation: principal axis factorization, maximum likelihood method. Factor rotation. Factor scores: Bartlett and Thompson estimators. • Cluster analysis. Dissimilarity matrix. Hierarchical techniques. Non-hierarchical techniques. Finding the number of groups. • Correspondence analysis. Graphical representation. Introduction to multiple correspondence analysis.
Texts and readings	A. Pollice, Course handouts, available in e-learning mode. G. James, D. Witten, T. Hastie, R. Tibshirani (2021) An Introduction to Statistical Learning: with Applications in R Second edition. Springer Editor.
Notes, additional materials	The use of slides or personal notes is STRONGLY NOT RECOMMENDED and considered insufficient for the preparation.
Repository	In e-learning mode it is possible to carry out the self-assessment tests and download the course handouts, additional teaching materials, traces and data useful for carrying out the exercises with the R software. The address and password of the Multivariate Statistics course in e-learning mode are shared at the beginning of the course.

Assessment	
Assessment methods	Per i frequentanti la valutazione delle attività formative è distribuita nell'arco del semestre e si conclude con la fine del corso. L'esito dei test di autovalutazione e delle attività di laboratorio contribuisce ad integrare la valutazione complessiva del profitto. Concorrono alla valutazione un esonero intermedio ed uno finale riferiti a parti distinte del programma ed entrambi basati su un test con domande a risposta multipla ed un'esercitazione in laboratorio con il software R. Alla fine di ciascun esonero segue un breve colloquio orale. I non frequentanti devono sostenere una prova riferita all'intero programma del corso e basata su un test con domande a risposta multipla ed un'esercitazione in laboratorio con il software R. Segue un colloquio orale.

Assessment criteria	• *Conoscenza e capacità di comprensione*
	La prima parte degli esoneri riferiti alle due parti del corso consiste in un test con 20 domande a risposta multipla da svolgere in 30 minuti. Per i non frequentanti il test riguarda l'intero programma del corso e comprende 40 domande a risposta multipla da svolgere in 50 minuti.
	• *Conoscenza e capacità di comprensione applicate*
	La seconda parte degli esoneri prevede lo svolgimento di un'esercitazione in cui si richiede di sviluppare in 2 ore una traccia riferita all'analisi di un insieme di dati con il software R. Analogamente, si richiede ai non frequentanti di svolgere in 3 ore l'analisi di un insieme di dati con il software R con riferimento alle metodologie dell'intero programma del corso.
	• *Autonomia di giudizio*
	Allo scopo di valutare l'autonomia di giudizio dei/lle candidati/e verranno valutati i commenti contenuti nello svolgimento delle esercitazioni e riferiti
	  • alle motivazioni delle scelte effettuate;
	  • all'interpretazione dei risultati dell'applicazione delle metodologie oggetto del programma del corso con il software R.

	• *Abilità comunicative* Dopo aver effettuato la consegna dello svolgimento dell'esercizio con R, i/le candidati/e verranno chiamati/e individualmente per una breve discussione pubblica orale basata sulle risposte alle domande a risposta multipla. Allo scopo di valutare le abilità comunicative dei/lle candidati/e verranno inoltre considerati i commenti contenuti nello svolgimento delle esercitazioni e riferiti • alle motivazioni delle scelte effettuate; • all'interpretazione dei risultati dell'applicazione delle metodologie oggetto del programma del corso con il software R. • *Capacità di apprendere in modo autonomo* Allo scopo di valutare se i/le candidati/e hanno sviluppato le capacità di apprendimento necessarie per intraprendere studi successivi con un alto grado di autonomia gli/le stessi/e sono interrogati/e individualmente per una breve discussione pubblica orale basata sulle risposte alle domande a risposta multipla.
Final exam and grading criteria	La correzione e la valutazione degli svolgimenti dell'esercitazione con R sono effettuate nei giorni successivi allo svolgimento, prestando molta attenzione all'autenticità dei contenuti. Sono annullati automaticamente gli esoneri dei/delle candidati/e i cui svolgimenti riportano frasi o espressioni identiche nell'interpretazione dei risultati o gli stessi errori nei comandi di R. L'annullamento dell'esonero comporta l'obbligo di sostenere l'esame sull'intero programma del corso. Ciascun test dà luogo ad una valutazione in 100esimi. Anche le esercitazioni con R danno luogo a valutazioni in 100esimi. Il risultato del colloquio orale va ad migliorare o peggiorare il punteggio conseguito nelle due prove precedenti. La valutazione dell'esonero intermedio e di quello finale danno luogo ad una proposta di voto in trentesimi ottenuta come sintesi delle due valutazioni. Per i non frequentanti la proposta di voto è formulata sulla base delle valutazioni del test, dell'esercitazione con R e del colloquio orale riferiti all'intero programma del corso.

Further information	