UNIVERSITÀ DEGLI STUDI DI PALERMO

DEPARTMENT	Scienze Economiche, Aziendali e Statistiche
ACADEMIC YEAR	2022/2023
MASTER'S DEGREE (MSC)	STATISTICS AND DATA SCIENCE
INTEGRATED COURSE	CATEGORICAL DATA - INTEGRATED COURSE
CODE	20668
MODULES	Yes
NUMBER OF MODULES	2
SCIENTIFIC SECTOR(S)	SECS-S/01
HEAD PROFESSOR(S)	SCIANDRA MARIANGELA Professore Associato Univ. di PALERMO
OTHER PROFESSOR(S)	ABBRUZZO ANTONINO Professore Associato Univ. di PALERMO SCIANDRA MARIANGELA Professore Associato Univ. di PALERMO
CREDITS	9
PROPAEDEUTICAL SUBJECTS	
MUTUALIZATION	
YEAR	2
TERM (SEMESTER)	1° semester
ATTENDANCE	Not mandatory
EVALUATION	Out of 30
TEACHER OFFICE HOURS	**ABBRUZZO ANTONINO** Monday 15:00 17:00 DSEAS secondo piano stanza 222 **SCIANDRA MARIANGELA** Wednesday 12:00 14:00 DSEAS 2 piano

DOCENTE: Prof.ssa MARIANGELA SCIANDRA

PREREQUISITES	The course requires a deep knowledge of inferential statistics and probability.
LEARNING OUTCOMES	Knowledge of basic methods of probabilistic networks and categorical data models. Acquisition of language and terminology of the discipline. Understanding of derivations, theoretical properties and relations among the presented methods. Ability to deal with concrete problems with the methods acquired during the lectures. Ability to use the statistical environment R to apply the skills students have acquired during the lectures and to check via simulation the theoretical results. Being able to critically understand the characteristics, potential, and limits of the proposed methodologies Being able to discuss the characteristics of a given problem. Being able to use the statistical terminology and the formalization of the problems in writing. Being able to see the scientific literature; ability to learn the patterns of extensions studied in class; learning the ability of specialized statistical software also different from that used in the classroom.
ASSESSMENT METHODS	The final examination will consist of a discussion, and it depends on the fact that the student had passed the written test. The written test will beed in English. The oral test will be held in English. The teacher may, if he/she deems it appropriate, make the candidate present one or more arguments in Italian. Written test The written test strives to establish the knowledge and abilities possessed by the student. The written test will cover probabilistic networks and models for categorical data, acquired during the course and it will be carried out with the support of the statistical software R. The test lasts a maximum of 3 hours. The test regards matters about both modules. The sufficiency (equivalent to a score of 18 on a scale of 18 to 30), which is necessary to pass the test, is reached if the student shows an adequate use of the terms relating to the concepts in question, and i) in case of practical question, by identifying the appropriate statistical methodology even if it is spoiled by the mere computation error; ii) in the case of a theoretical question, in the consistency of the answer, although not completely exhaustive of the topic. Oral examination The oral test is intended to dig up the topic of the written test and to evaluate the knowledge of the students on the subject. This will consist of at least two questions aimed at graduate better knowledge and abilities possessed by the student, and its ability to provide it with a suitable statistical language. The sufficiency of the oral test will be reached when the student has knowledge and understanding of the topics at least in the general terms (definition of the concepts). The more, however, the examination has brilliantly passed the written test and has shown, in the oral test, their capacities, as well as the status of statistical language and the connection with the other subjects, much more the evaluation will be positive. Oral examination regards matters about both modules. FINAL ASSESSMENT The final evaluation of the examination will take into account two aspects: i) mastery of the topics; ii) the property of statistical language, assessed on the whole of the written and oral test. The teacher will also have the opportunity to take into account the context factors of the exam (such as active participation during the lessons and exercises or the presence of some disabilities) for the purpose of determining the outcome of the test. Both tests (written and oral) are evaluated in thirty-five and are considered to be exceeded with a minimum vote of 18/30. The final score is given by the weighted arithmetic mean (weigths: 6 for Categorical Data and 3 for Stochastic Networks).
TEACHING METHODS	The course will be divided into lectures and practicals. All the theoretical arguments developed during the lectures will be addressed in terms of applications, by means of computer-statistical practice, with the use of the program environment R. the entire course will be held in english.

MODULE
STOCHASTIC NETWORKS

Prof. ANTONINO ABBRUZZO

SUGGESTED BIBLIOGRAPHY

Introduction to graphical modelling, Dempster. Capitoli 1, 2, 3, 5, 6, 7

Graphical Models with R, Søren Højsgaard, David Edwards. Steffen Lauritzen, Springer, 2012. Capitoli 1, 2, 3, 4, 6, 7.

Dispense del docente.

AMBIT	21031-Attività formative affini o integrative
INDIVIDUAL STUDY (Hrs)	54
COURSE ACTIVITY (Hrs)	21

EDUCATIONAL OBJECTIVES OF THE MODULE

The course offers the student the opportunity to acquire knowledge on graphical models for the analysis of categorical data and to apply these methodologies to several datasets. At the end of the course, the student should be able to recognize the strengths and weaknesses of the graphical models for categorical data and to describe complex real datasets through the use of the techniques learned.

The course guides the student to the knowledge of the basic methods of probabilistic networks and the acquisition of the ability to apply these methodologies to real datasets. Students should be able: i) to understand both the positive and negative aspects of probabilistic networks; ii) to use these techniques to investigate real datasets.

SYLLABUS

Hrs	Frontal teaching
4	Log-linear graphical models
4	Bayesian networks for categorical data analyis
4	Continuous graphical models

Hrs	Practice
3	Data analysis with log-linear graphical models
3	Data analysis with Bayesian Networks
3	Data analysis with Gaussian graphical models

MODULE
MODELS FOR CATEGORICAL DATA

Prof.ssa MARIANGELA SCIANDRA

SUGGESTED BIBLIOGRAPHY

Agresti A. (2002) The analysis of categorical data (2nd ed.), Academic Press, London. (Chs. 1 to 9) Disponibile presso la biblioteca del DSEAS

Johnson, Valen E., Albert, James H. (1999), Ordinal Data Modeling,SpringerVerlag New York (Chs. 3 and 4) Acquistabile online o presso la libreria universitaria.

AMBIT	21031-Attività formative affini o integrative
INDIVIDUAL STUDY (Hrs)	108
COURSE ACTIVITY (Hrs)	42

EDUCATIONAL OBJECTIVES OF THE MODULE

This course aims to provide students with a statistical background and practical skills to apply more advanced modeling techniques specific for categorical data problems. Students must be able to identify the best statistical tool to investigate a problem related to categorical data (binary, ordered). In presence of a multiway categorical problem, students must be able to understand if the problem asks for an associative or dependence structure and also to identify the most parsimonious way to describe the data generating process. In the end, students must be able to represent categorical data problems and results using specific graphical tools Ability to discuss the characteristics of a practical problem and comment the obtained results and interpret results to non statisticians.

SYLLABUS

Hrs	Frontal teaching
4	1.1 Basic concepts and definitions: categorical variables, categorical data matrices, analysis of directed and undirected relationships, approaches with and without probabilistic formalisation 1.2 Recall of discrete multivariate distributions.
6	2.1 The 2x2 contingency table, Measures of association and dependence , Logit-linear and log-linear models 2.2 Extensions to the IxJ contingency table 2.3 Polytomous response models 2.4 Ordinal categorical variables models
8	3.1 Measures and models of association and dependence in multiway contingency tables 3.2 Model selection procedures
4	Models for zero inflated data: ZIP and Hurdle models
2	Quasi and complete separability in categorical data modelling: the Firth regression

Hrs	Practice
4	Introduction: laboratory tutorials in R
8	Two-way contingency tables: laboratory tutorials in R
6	Multiway contingency tables: laboratory tutorials in R