SummeR School-DATA MINING AND DATA ANALYSIS WITH R

PROPONENTE

Dott. Davide Gentilini Prof.ssa Luisa Bernardinelli

OBIETTIVI FORMATIVI

La gestione e l'analisi dei dati rappresenta uno dei fattori più importanti e simultaneamente più critici in molti ambiti lavorativi; spesso ci si avvale di strumenti e software proprietari che sono soventemente costosi o risultano essere in definitiva limitati nella loro applicazione.
Il corso ha l'obiettivo di fornire ai partecipanti le conoscenze necessarie e sufficienti per introdurre ed utilizzare il linguaggio e le potenzialità di R nel proprio lavoro.
R può essere definito come un sistema di analisi statistica e contemporaneamente un linguaggio ed un software. E' uno strumento open source potentissimo e largamente utilizzato, per l'analisi statistica dei dati, inoltre, essendo un vero e proprio linguaggio di programmazione, racchiude in se la potenzialità per creare e sviluppare in autonomia svariate applicazioni utili alla manipolazione, gestione ed analisi di ogni tipo di dato. Le sue caratteristiche principali comprendono infatti la semplicità nella gestione e manipolazione dei dati, la disponibilità di una suite di strumenti per calcoli su vettori, matrici ed altre operazioni complesse, l' accesso ad un vasto insieme di strumenti integrati e funzioni sviluppate da altri e resi disponibili per l'analisi statistica, la produzione di numerose potenzialità grafiche particolarmente flessibili, possibilità di adoperare un vero e proprio linguaggio di programmazione orientato ad oggetti che consente l'uso di strutture condizionali e cicliche, nonché di funzioni create dall'utente.
Offrendo un'elaborata introduzione alla programmazione con R questo corso ha lo scopo di intercettare le esigenze dei partecipanti focalizzandosi in particolar modo su alcuni aspetti fondamentali come la manipolazione e gestione dei dati, la loro analisi attraverso l'identificazione del test statistico più appropriato e la visualizzazione di dati e risultati utilizzando le potenzialità grafiche messe a disposizione da R.
Il corso prevede l'impiego di numerosi "dataset " ed esempi che possano essere familiari alle varie aree di interesse in modo da agevolare i partecipanti nella comprensione ed applicazione delle nozioni acquisite.

NUMERO DI ORE(CFU)/LEZIONI

Il corso avrà una durata di almeno 20 ore (5 CFU), e sarà suddiviso in 5 incontri.

Possono essere frequentati anche singoli incontri.

PERIODO DI SVOLGIMENTO

Il periodo di svolgimento del corso sarà dal 21 al 25 Giugno 2021 (summeR school) Il calendario potrà subire variazioni

DOCENTI

Dott. Davide Gentilini

COMITATO SCIENTIFICO

Prof Luisa Bernardinelli, Dr Davide Gentilini

COMITATO ORGANIZZATORE

Dipartimento di Scienze del Sistema Nervoso e del Comportamento e Residenza Biomedica – Fondazione Collegio Universitario S. Caterina da Siena, Pavia

SEDE DEL CORSO

Residenza Biomedica –Fondazione Collegio Universitario S. Caterina da Siena, Pavia

POTENZIALI DOTTORATI INTERESSATI

Il corso ha l'intento di essere trasversale e di fornire competenze utili in ogni ambito che abbia a che fare con dati e con la necessità di elaborarli e gestirli. Per tale ragione il corso ha l'obiettivo di essere utile ad ogni tipologia di dottorato, in particolare macroarea di Scienze della vita, Bioingengneria e Bionformatica, Economics, Political Studies, DREAMT. Nell'attività pratica verranno utilizzati esempi e dati nei vari ambiti applicativi di interesse per i diversi Dottorati.

MODALITA' DI VERIFICA DELL'APPRENDIMENTO

Il grado di apprendimento verrà testato al termine di ogni lezione sottoponendo i partecipanti ad un test.
Il test sarà composto da una serie di 10 domande a scelta multipla riguardanti la parte teorica a cui verrà aggiunto un esercizio inerente le argomentazioni trattate.

PROGRAMMA

Calendario Attività e Programma

21 Giugno 2021 - ore 9.30-13.30

Titolo: “L’ambiente di programmazione R generalità e i principali oggetti”
Obiettivo della lezione è introdurre i partecipanti all’ambiente R, illustrandone le principali funzionalità, l’architettura e presentando i principali oggetti.
• Introduzione del Corso e Generalità
• Installazione di R e configurazione
• Nozioni preliminari sulla sua architettura sui pacchetti e sulle funzioni
• Utilizzo dei I pacchetti
• Utilizzo dell’Help
• Gli oggetti principali
• I vettori
o Assegnazione Vettori (Vettori aritmetici, Logici, di Caratteri)
o Operazioni e funzioni per lavorare con i vettori
• Matrici ed Array
o Operatori e funzioni per il calcolo Matriciale
• Liste e Data Frames
o Operatori e funzioni per lavorare con Data Frames e Liste
• Esempi
• Esercitazione Pratica

22 Giugno 2021 - ore 9.30-13.30

Titolo: “Data import e data mining con R”

Obiettivo della lezione è insegnare ai partecipanti ad importare dati di varia tipologia e formato nell’ambiente R. La lezione ha inoltre lo scopo di insegnare “data mining” ovvero ad utilizzare R per visualizzare gestire, estrarre e manipolare i dati. La lezione tratterà le funzioni grafiche di R in modo che i partecipanti al corso acquisiscano le conoscenze base per poter visualizzare i propri dati e i propri risultati in modo autonomo. Esempi ed esercizi guidati avranno lo scopo di verificare e consolidare le nozioni.
• Importazione e pulizia di data-set
• Funzioni di importazione dei dati in base alla loro natura e formato
• Funzioni di visualizzazione e manipolazione dei dati
• Quality Control dei dati
• Trattamento dei dati mancanti
• Imputazione
• Metodi di riclassificazione delle variabili
• Funzioni per la manipolazione dei dati

23 Giugno 2021 - ore 9.30-13.30

Titolo: “La Grafica con R”

• L’ambiente grafico
• Pacchetto Base
• Pacchetto ggplot2
• I parametri grafici
• Tipi di grafici e funzioni
• Classificazione del tipo di dati e discussione sulla modalità di visualizzazione
• Esempi
• Esercitazione Pratica

24 Giugno 2021 - ore 9.30-13.30

Titolo: “Statistica di base con R”

Obiettivo della lezione è entrare nel merito della statistica descrittiva/inferenziale e fornire gli strumenti essenziali per poter scegliere e applicare il metodo statistico più appropriato

- Principi di disegno dello studio e tipi di dati
- Misure descrittive univariate/bivariate

• Test statistici
o Test parametrici su media, differenza tra medie, proporzione, differenza tra proporzioni e correlazione
o Test non parametrici su media, differenza tra medie, differenza fra mediane, proporzione, differenza tra proporzioni e correlazione
o Test di associazione per tabelle 2x2 e JxK
• Esempi
• Esercitazione Pratica

25 Giugno 2021 - ore 9.30-13.30

Titolo: “Statistica di base con R”

Obiettivo della lezione è entrare nel merito della statistica descrittiva/inferenziale e fornire gli strumenti essenziali per poter scegliere e applicare il metodo statistico più appropriato
• Analisi della varianza
• Regressione lineare semplice e multipla
• Regressione logistica
• Analisi delle componenti principali
• Esempi
• Esercitazione Pratica

PROPOSER

Dr. Davide Gentilini Prof. Luisa Bernardinelli

EDUCATIONAL OBJECTIVES

Data management and analysis is one of the most important and simultaneously the most critical factors in many work environments; proprietary tools and software are often used which are frequently used or are ultimately limited in their application.
The course aims to provide participants with the necessary and sufficient knowledge to introduce and use the language and potential of R in their work.
R can be defined as a system of statistical analysis and at the same time a language and a software. It is a powerful and widely used open source tool for statistical data analysis, furthermore, being a real programming language, it contains the potential to independently create and develop various applications useful for manipulation, management and analysis. of any type of data. Its main features are in fact the simplicity in the management and manipulation of data, the availability of a suite of tools for calculations on vectors, matrices and other operations, access to a vast set of integrated tools and functions developed by others and made available. for statistical analysis, the production of numerous particularly flexible graphic potentials, the possibility of using a real object-oriented programming language that uses the use of conditional and cyclic structures, as well as functions created by the user.
Offering an elaborate introduction to programming with R, this course aims to intercept the needs of the participants by focusing in particular on some fundamental aspects such as data manipulation and management, their analysis through the identification of the most appropriate statistical test and the visualization of data and results using the graphic potential made available by R.
The course involves the use of numerous "data sets" and examples that may be familiar to the various areas of interest in order to facilitate the participants in reading and applying the knowledge acquired.

NUMBER OF HOURS (CFU) / LESSONS

The course will last at least 20 hours (5 CFU), and will be divided into 5 meetings. Individual meetings can also be attended.

PERIOD OF CONDUCT

The course period will be from 21 to 25 June 2021 (summeR school) The calendar may be subject to variations

TEACHERS

Dr. Davide Gentilini

SCIENTIFIC COMMITTEE

Prof Luisa Bernardinelli, Dr Davide Gentilini

ORGANIZING COMMITTEE

Department of Nervous System and Behavioral Sciences and Biomedical Residence - University College Foundation S. Caterina da Siena, Pavia

Location

On line

POTENTIAL INTERESTED DOCTORATES

The course aims to be transversal and to provide useful skills in every area that has to do with data and the need to process and manage them. For this reason, the course aims to be useful to any type of doctorate, in particular the macro-area of Life Sciences, Bioengineering and Bionformatics, Economics, Political Studies, DREAMT. In the practical activity, examples and data will be used in the various application areas of interest to the various PhDs.

LEARNING VERIFICATION METHOD

The degree of learning will be tested at the end of each lesson by subjecting the participants to a test.

The test will consist of a series of 10 multiple choice questions regarding the theoretical part to which an exercise relating to the topics covered will be added.

PROGRAM

Activities Calendar and Program

21 June 2021 - 9.30-13.30

Title: "The general programming environment R and the main objects"
The objective of the lesson is to introduce the participants to the R environment, illustrating its main functions, architecture and presenting the main objects.
• Introduction of the Course and Generalities
• R installation and configuration
• Preliminary notions on its architecture on packages and functions
• Use of packages
• Use of the Help
• The main objects
• The vectors
o Vector Assignment (Arithmetic, Logic, Character Vectors)
o Operations and functions for working with vectors
• Matrices and Arrays
o Operators and functions for the Matricial calculation
• Lists and Data Frames
o Operators and functions for working with Data Frames and Lists
• Examples
• Practical exercise

June 22, 2021 - 9.30-13.30

Title: "Data import and data mining with R"

The aim of the lesson is to teach participants to import data of various types and formats into the R environment. The lesson also aims to teach "data mining" or to use R to view, manage, extract and manipulate data. The lesson will cover the graphical functions of R so that course participants acquire the basic knowledge to be able to view their data and results independently. Examples and guided exercises will aim to verify and consolidate the notions.
• Import and cleanup of data sets
• Data import functions based on their nature and format
• Data display and manipulation functions
• Data Quality Control
• Treatment of missing data
• Imputation
• Methods of reclassification of variables

• Functions for data
Manipulation
23 June 2021 - 9.30-13.30

Title: "The Graphics with R"

• The graphic environment
• Basic Package
• ggplot2 package
• The graphic parameters
• Types of graphs and functions
• Classification of data type and discussion on how to display
• Examples
• Practical exercise

24 June 2021 - 9.30-13.30

Title: "Basic statistics with R"

The objective of the lesson is to go into the merits of descriptive / inferential statistics and provide the essential tools to be able to choose and apply the most appropriate statistical method
• Study design principles and data types
• Univariate / bivariate descriptive measures
• Statistical tests
o Parametric tests on mean, difference between means, proportion, difference between proportions and correlation
o Nonparametric tests on mean, difference between means, difference between medians, proportion, difference between proportions and correlation
o Association test for 2x2 and JxK tables
• Examples
• Practical exercise

25 June 2021 - 9.30-13.30

Title: "Basic statistics with R"

The objective of the lesson is to go into the merits of descriptive / inferential statistics and provide the essential tools to be able to choose and apply the most appropriate statistical method
• Analysis of variance
• Simple and multiple linear regression
• Logistic regression
• Principal component analysis
• Examples
• Practical exercise