

**Università degli Studi di Salerno
Laurea in Statistica per i Big Data**

ANALISI E VISUALIZZAZIONE DEI DATI

Prof. Pietro Coretto (Modulo 1, 30h)
pcoretto@unisa.it

Dott.ssa Cecilia Viscardi (Modulo 2, 30h)
cviscardi@unisa.it

SYLLABUS

Sommario. Il corso ha ad oggetto sia gli aspetti concettuali e metodologici dell'analisi esplorativa dei dati (EDA), sia gli strumenti e i metodi di elaborazione dei dati, calcolo numerico, e programmazione (ECP). L'oggetto del corso è lo studio delle caratteristiche della distribuzione empirica dei dati attraverso metodi numerici e grafici. La componente ECP del corso è basata sulla base sull'ambiente di sviluppo R, essa è strettamente funzionale all'EDA. Nella parte iniziale del corso i due aspetti sono trattati separatamente. Successivamente le due componenti (EDA e ECP) saranno presentate in modo integrato. Di seguito si riporta il cronoprogramma dettagliato del corso.

Indice

| | |
|---|----------|
| 1 Organizzazione della didattica | 2 |
| 2 Materiali didattici | 2 |
| 3 Programma dettagliato | 3 |
| 4 Esame e valutazione del profitto | 5 |
| 5 Bibliografia | 5 |

1 Organizzazione della didattica

Lezioni. Il corso si articola su 10cfu (60h). Di seguito si riporta la programmazione settimanale.

| | | |
|-----|-------------|---------------------|
| MAR | 08:30-10:10 | Aula A, Edificio D1 |
| MER | 08:30-10:10 | Aula 2, Edificio C |
| GIO | 15:10-16:50 | Aula D, Edificio D1 |

Avvisi e ricevimento. Per quanto riguarda gli avvisi ed il ricevimento studenti consultare l'[homepage](#) del docente. Il calendario esami sarà disponibile attraverso la [bacheca appelli](#) di ateneo.

2 Materiali didattici

Slides, software, scripts, note, esercitazioni e prove d'esame saranno distribuiti durante tutta la durata del corso. La distribuzione del materiale si basa sulla piattaforma di ateneo Google Workspace. L'accesso alle risorse condivise richiede un account @studenti.unisa.it.

I testi indicati in bibliografia (vedi Sezione 5) insieme al materiale fornito durante le lezioni coprono l'intero programma del corso. La valutazione del profitto riguarderà solo gli argomenti affrontati al corso, di conseguenza i testi consigliati vanno studiati solo per le parti trattate al corso.

Le lezioni si basano sull'uso del linguaggio di programmazione R¹ sviluppato e distribuito con licenza GPL attraverso il *Comprehensive R Archive Network* (CRAN)². Per questo corso è necessario installare R-base-package (comunemente detto R), ed un buon IDE.

Installazione di R-base-package

- **linux:** <https://cran.r-project.org/bin/linux/>
Selezionare la distro appropriata e seguire le istruzioni.
- **macOS:** <https://cran.r-project.org/bin/macosx/>
Nella sezione “Latest Release” sono disponibili diversi file R-x.y.z.pkg, dove x.y.z sono gli identificativi della versione ordinati a partire dalla più recente. Verificare quale è la versione più recente compatibile con la versione di macOS in uso. Scaricare il corrispondente file *.pkg e seguire la procedura di installazione. Si consiglia inoltre l'installazione di XQuartz³.
- **Microsoft Windows:** <https://cran.r-project.org/bin/windows/base/>
Scaricare il file *.exe dal link “Download R x.y.z for Windows”, dove x.y.z sono solitamente gli identificativi della versione corrente. Eseguire il file *.exe scaricato (possibilmente con permessi di amministratore), e seguire la procedura guidata di installazione.

¹Vedi: [https://it.wikipedia.org/wiki/R_\(software\)](https://it.wikipedia.org/wiki/R_(software))

²Vedi: [https://en.wikipedia.org/wiki/R_package#Comprehensive_R_Archive_Network_\(CRAN\)](https://en.wikipedia.org/wiki/R_package#Comprehensive_R_Archive_Network_(CRAN))

³Disponibile al seguente link: <https://www.xquartz.org/>

Installazione dell'IDE

Vi sono tantissime alternative. Durante il corso, e nei laboratori del campus destinati alle attività didattiche, utilizzeremo **RStudio**:

- <https://rstudio.com/products/rstudio/download/#download>

3 Programma dettagliato

Nozioni introduttive (lezione 1). Statistica, informatica e dati – nozioni introduttive: popolazione, campione, campionamento, inferenza, EDA – primi esempi di data set – elementi di un data set: unità statistica, variabili/features – espressione di variabili: livelli e labels – categorie di variabili rispetto alla loro espressione: quantitative, ordinali, nominali – categorie di data sets rispetto all'indicizzazione delle unità: cross-section, serie temporali, serie spaziali, serie spazio-tempoarali, dati longitudinali – esperimenti e studi osservazionali – fasi e componenti dell'EDA.

Linguaggio R: tipi numerici e principali strutture dati (lezione 2–3). L'ambiente di calcolo R – operazioni aritmetiche – operazioni logiche – funzioni matematiche di uso comune – funzioni logiche di uso comune – assegnazione di variabili ed uso della memoria – variabili locali, globali, environments e workspace – tipi di dati – funzioni – strutture dati e ADS (abstract data structures) – vettori (1D-arrays) – attributi e metadati di un vettore – codifica delle variabili non numeriche (la funzione `factor()`) 2D-arrays – 3D-arrays – arrays di dimensione qualsiasi – operazioni su arrays – matrici e algebra lineare – liste – data.frame – subsetting/slicing di un ADS – metadati più comuni di un ADS.

Linguaggio R: inputs/outputs (lezione 4–5). Tecniche elementari di storage e codifica dei dati (ascii, csv, tsv) – lettura di dati in formato ascii – data-import e formattazione dei dati – RData altri formati proprietari (xls, ods, mat, dta, etc) – scrittura dei dati in formato RData e ascii – interazioni elementari con OS – tecniche elementari per manipolare stringhe – gestione dell'output della console.

Linguaggio R: device grafico (lezione 5–6). introduzione al device grafico – primi esempi di rappresentazioni grafiche – parametri grafici generici – rappresentazione dei colori e palettes – overlays e grafici multipli – aspect ratio, scaling, e risoluzione – grafica raster (bitmat), vettoriale, lossy, lossless – scrittura di grafici in formati comuni (pdf, jpg, png, etc) – ottimizzazione dei formati grafici in funzione dell'utilizzo.

Linguaggio R: programmazione (lezione 7–9). Strutture di controllo (if, if-else, if-else, if-esle if-else, while, for, repeat) – breakers (stop, break, etc.) – scripts – stile di programmazione (commenti descrittivi, buona sintassi, etc) – approfondimenti sulla rappresentazione numerica (elementi della lista `.Machine`, effetti del “rumore di troncamento”) – programmazione funzionale – costruzione di una funzione – output complessi e passaggio di argomenti – classe di un oggetto, metodi e programmazione ad oggetti – costruzione di “*print method*” e “*plot method*” – *environment* globale vs *function environment* – funzioni nascoste e funzioni primitive – introduzione al debugging.

Distribuzione, funzione quantile e densità (lezione 10–11). Raw data e visualizzazione mediante stripchart – distribuzione di dati univariata (DU) – caratteristiche rilevanti di DU – frequenze relative e assolute – windowing dei dati continui – costruzione di DU

aggregate per frequenze – stripchart, barplot, e piechart – massa e densità dei dati – misure di densità – misura di densità mediante istogramma – rappresentazione grafica dell’istogramma – frequenze cumulate – funzione di distribuzione empirica cumulata (ECDF) – quantile di livello α – approssimazione dei quantili basata sull’ECDF – ordinamento dei dati e ranks – principali tecniche di ordinamento in R – ordinamento gerarchico di un data set – approssimazione dei quantili basata sullo smoothing dei dati ordinati.

Centralità e dispersione di una dispersione (lezione 11–12). Weighting dei dati e somme pesate – weighting basato sulla distribuzione empirica – media empirica, mediana e moda – serie di quantili – quartili e Tukey 5-numbers – decili e percentili – trasformazioni dei dati – trasformazioni lineari come operazioni di shift e scaling – trasformazioni lineari e unità di misura – effetti delle trasformazioni lineari sulle misure di dispersione (scatter, spread) – deviazioni rispetto ad un punto – deviazione totale quadratica e assoluta – centralità della media rispetto alle deviazioni quadratiche – centralità della mediana rispetto alle deviazioni assolute – variabilità – varianza campionaria, varianza non distorta, deviazione standard, IQR, ampiezza del range – decomposizione della varianza nei momenti – proprietà delle misure di variabilità – effetti delle trasformazioni lineari sulle misure di dispersione – standardizzazione dei dati.

Simmetria, modello normale standard e code (lezione 13–14). Distribuzione dei dati limite ($n \rightarrow \infty$) – continuità al limite della densità e dell’ECDF – modello normale standard – proprietà di forma del modello normale standard – simmetria ed asimmetria – simmetria e misure di centralità – simmetria e quantili – misure di asimmetria – code di una distribuzione – decadimento esponenziale delle code nel modello normale – Normal Q-Q-plot e confronto con le code normali – Q-Q-plot per il confronto tra distribuzioni qualsiasi – eccesso di curtosì.

Contaminazione e robustezza (lezione 15). Contaminazione dei dati – processi di contaminazione vs code pesanti/lunghe – nozione intuitiva di robustezza – *finite sample replacement breakdown point* – breakdown point asintotico – calcolo del breakdown point per media e mediana – misure di centralità e dispersione robuste (trimean, medie troncate, MAD, etc) – *Tukey fences* e classificazione dei punti regolari, outliers, e gross-outliers. – boxplot – Tukey fences e boxplot.

Kernel density e smoothing (lezione 17). Stima kernel della densità per DU – bandwidth e funzione kernel – scelta della bandwidth e metodi data-driven – kernel rettangolare, Gaussiano, di Epanechnikov – undersmoothing e oversmoothing.

Strumenti per il data-wrangling (lezione 18–19). Introduzione agli strumenti di data-wrangling in R – data frame e tibble – formato *long* vs *wide* – introduzione all’uso di Tidyverse – importare dati con **Readr** – gestione delle date con **Lubridate** – Operatore “pipe” e costruzione di “pipeline” – **Dplyr** per selezionare/modificare righe e/o colonne di un dataset.

Introduzione alla visualizzazione con ggplot (lezione 20). Introduzione all’uso di **ggplot** e confronto con gli strumenti grafici di base – specifica dei “layers” – boxplot, diagramma a dispersione, diagramma a barre, istogramma e densità – uso dei parametri **aes** per individuare un fattore “condizionante” – uso delle funzioni **facet**.

Data reporting (lezione 21–22). Introduzione a Rmarkdown – linguaggi implicati – formato dell’output – linguaggio Markdown – “chunks” di codice – formulazione matematica

e linguaggio **LaTeX**.

Distribuzioni Multivariate (lezione 23–24). Introduzione all’analisi multivariata – caso bivariato – distribuzioni di frequenze assolute congiunte e marginali – distribuzioni di frequenze relative e distribuzioni condizionate – distribuzioni e dipendenza – EDA nel caso multivariato – ECDF nel caso multivariato – indici di centralità e dispersione condizionata.

Dipendenza e connessione: Associazione (Lezione 25–27). Dipendenza e legami associativi – condizione di indipendenza basata su ECDF – misure di associazione – indice Chi-quadro di Pearson – indice di Cramer – dipendenza e confronti tra proporzioni: differenza tra proporzioni, rischio relativo, odds e odds ratio.

Dipendenza e connessione: Correlazione (lezione 28–30). Dipendenza tra variabili quantitative – diagramma a dispersione – dipendenza lineare – covarianza: definizione, calcolo, interpretazione e proprietà – indice di correlazione: definizione, calcolo, interpretazione e proprietà – correlazione vs dipendenza – correlazione spuria e correlazione a blocchi – matrici di covarianza e di correlazione e loro proprietà – dipendenza lineare in media – smoothing lineare – smoothing e relazioni non lineari.

4 Esame e valutazione del profitto

L’esame prevede una prova scritta in laboratorio e un colloquio orale. Lo scritto, superato con un punteggio ≥ 18 pt, valuta i diversi quesiti con un massimo di 30 punti. Il colloquio è solitamente focalizzato sulla discussione dello scritto; tuttavia, in alcuni casi potrà essere esteso ad un approfondimento sui contenuti del corso per verificare l’autonomia di giudizio e le capacità critiche ed espositive dello studente.

5 Bibliografia

Iacus, S. M., e G. Masarotto (2007). *Laboratorio di statistica con R*. McGraw-Hill.

Muggeo, Vito M. R., e Giancarlo Ferrara. 2005. *Il linguaggio R: concetti introduttivi ed esempi*. II edizione. <https://cran.gedik.edu.tr/doc/contrib/nozioniR.pdf>.

Wickham, H., e G. Grolemund (2016). *R for data science: import, tidy, transform, visualize, and model data*. O’Reilly Media, Inc. (Anche disponibile in versione elettronica: <https://it.r4ds.hadley.nz/index.html>)