

Elementi di analisi statistica con R

Dicembre 2012

Docente: Christian A. Mongeau Ospina

Luogo: Aula XI (piano -2)

Data: 6, 7, 13, 14 dicembre 2012, dalle 9:00 alle 11:00

Contatti: christian.mongeau@uniroma3.it

Descrizione e obiettivi

Scopo di questo mini-corso è quello di fornire agli studenti la preparazione per effettuare l'analisi statistica di un insieme di dati attraverso l'uso di R.

R a è un programma per l'analisi statistica molto usato nella comunità scientifica. Per menzionare alcuni dei suoi vantaggi, rispetto ad altri programmi, si possono citare: alto grado di flessibilità che il suo linguaggio offre; ampia gamma di funzioni per l'analisi dei dati; estensibilità delle funzionalità del programma attraverso l'installazione di pacchetti.

Il corso vale 1 CFU, per il conseguimento del quale è necessario essere presenti ad almeno 3 lezioni (sulle 4 previste) e svolgere un esame finale (solo EPOS).

Requisiti

Si richiede una conoscenza minima di statistica descrittiva. Non è invece necessaria alcuna esperienza con R o altri programmi statistici.

Riferimenti bibliografici

Non è previsto un libro di testo per il corso: per quanto riguarda l'implementazione in R degli argomenti trattati è sufficiente il materiale che verrà distribuito durante il corso. Tuttavia, per chi volesse approfondire la conoscenza di R ci sono delle ottime dispense on-line:

- Muggeo e Ferrara (2005), “Il linguaggio R: concetti introduttivi ed esempi”. <http://cran.r-project.org/doc/contrib/nozioniR.pdf>
- Tancredi (2005), “Materiale didattico per il ciclo di lezioni Laboratorio di Statistica con R”. <http://geostasto.eco.uniroma1.it/utenti/tancredi/dispensaR.pdf>
- Massari, “Introduzione all’analisi statistica con R”, <http://www.dis.uniroma1.it/~statistica/dispensaR.pdf>

Chi, invece, volesse un libro di testo può far riferimento a:

Iacus e Masarotto, “Laboratorio di statistica con R”, McGraw-Hill, 2007 (ISBN: 9788838663697).

D'altra parte, se ci sono lacune riguardo i concetti statistici trattati si raccomanda la consultazione di un qualsiasi manuale di base di statistica (sebbene nel corso verranno fatti degli opportuni richiami).

Struttura del corso

Il corso è articolato in otto ore, secondo il seguente schema:

1. Installazione di R, oggetti, prime operazioni e *workflow*.
 - (a) Installazione: verranno indicati i passi necessari per l'installazione del programma;
 - (b) Oggetti: saranno introdotti gli oggetti più semplici (scalari, vettori, matrici) e quelli più complessi (array, data.frame, liste). Si parlerà dei tipi di dato che possono essere impiegati negli oggetti di R (numerici, alfanumerici, logici);

- (c) Prime operazioni: si effettueranno delle prime operazioni con gli oggetti creati (operazioni algebriche e booleane);
- (d) *Workflow*: si indicherà il tipico flusso di lavoro in R (apertura, importazione e esportazioni di dati, esecuzione di operazioni, chiusura con salvataggio di sessione e di storia delle operazioni).

2. Indicazione di elementi degli oggetti e prime statistiche

- (a) Indicazione di elementi: accesso agli elementi in forma esplicita (numerica) e tramite condizioni logiche. Questo argomento è di fondamentale importanza per il lavoro su R, motivo per cui si dedicherà molto tempo a vedere come si indicano gli elementi dei vari oggetti.
- (b) Prime statistiche: calcolo di massimo, minimo, *range*, media, mediana, deviazione standard. Le statistiche verranno calcolate sia su singoli vettori sia su oggetti a più dimensioni attraverso la famiglia di funzioni *apply*.

3. Analisi statistica (I)

- (a) Quantili (quartili e percentili), media interquartile, coefficiente di variazione, correlazione.
- (b) Distribuzioni di frequenza: frequenze assolute, relative, cumulate e marginali. Sivedrà come si trattano le diverse tipologie di dati (qualitativi e quantitativi) e si vedrà come sia possibile sintetizzare le informazioni dei dati quantitativi attraverso al ricorso a classi.

4. Estensione delle funzionalità di R, scrittura di script e funzioni, e ripasso

- (a) Pacchetti: installazione di pacchetti, e approfondimento del sistema di aiuto (documentazione) di R.
- (b) Scrittura di script e funzioni: si vedrà come è possibile facilitare lo svolgimento di operazioni ripetitive, oppure di semplificare alcuni passaggi, attraverso la creazione di script e/o funzioni.
- (c) Ripasso: arrivati a metà corso, lo studente deve essere in grado di replicare alcune delle analisi fatte in maniera completamente autonoma. Per questo motivo, si presenterà una nuova base di dati e verrà chiesto allo studente di calcolare alcune statistiche. Si approfitterà di quest'occasione per rivedere i concetti sui quali gli studenti presentano maggiori difficoltà.

5. Analisi statistica (II)

- (a) Forma delle distribuzioni: simmetria e curtosi (verranno costruite funzioni *ad hoc*).
- (b) Concentrazione: indice di Gini e curva di Lorenz.
- (c) Analisi di dipendenza: si vedrà come è possibile stabilire se esistono dei legami tra osservazioni e situazioni condizionanti o tra gruppi diversi.

6. Grafici

- (a) Scatter, istogrammi, diagrammi circolari, boxplot, funzione di ripartizione, stima della densità.
- (b) Esportazione attraverso *device* grafici (ad es. PNG).

7. Serie storiche

- (a) Creazione di oggetti di class *ts*; estrazione di informazioni strutturali (inizio osservazioni, fine, frequenza, ecc.) e di sottocampioni.
- (b) Medie mobili; estrazione di stagionalità ciclo e trend.

8. Interfacce grafiche e ripasso finale

- (a) Interfacce grafiche: verrà fatto cenno all'esistenza di interfacce grafiche alternative a quella nativa di R (ad es., R Commander e RStudio).
- (b) Ripasso: si affronteranno di nuovo gli argomenti che possono generare maggiori difficoltà.

Lo schema sopra proposto può subire leggere variazioni in relazione all'effettivo apprendimento dei concetti illustrati.

Si consiglia agli studenti di avere con se un computer portatile.

Verifica

Durante il corso verranno dati alcuni esercizi la cui soluzione non comporta un voto, ma sono essenziali per l'assimilazione degli argomenti trattati.

A fine corso verrà chiesto di effettuare un'analisi statistica di un insieme di dati impiegando R e i concetti insegnati nel corso. Solo per gli studenti EPOS questa verifica ha il carattere di un esame obbligatorio.