Laboratorio R per le scienze sociali

Federico Vegetti
federico.vegetti@unito.it

Università degli Studi di Torino

Obiettivi del laboratorio

- Familiarizzare con la logica e il linguaggio di R
- Imparare a usare R per:
  - fare data management in modo efficiente
  - esplorare e visualizzare i dati
  - fare alcune analisi statistiche di base
  - scrivere report
- *Imparare a risolvere in autonomia i problemi che inevitabilmente sorgeranno usando R*

Struttura

- 9 sessioni, ~1-1.5 ore circa
- Principalmente **sessioni pratiche**
- I materiali (tutorial `html`, slide, dati) saranno caricati su Moodle prima dell'inizio della lezione
- Il corso è in modalità mista (non registrato)
- Essendo un laoratorio pratico, la frequenza è caldamente consigliata

Risorse

Se vi doveste trovare in difficoltà, la cosa migliore da fare è cercare su Google il vostro problema, perchè con ogni probabilità qualcun altro lo ha avuto prima di voi. Nella maggior parte dei casi la soluzione arriverà da uno di questi siti:

- CrossValidated (principalmente statistica/analisi dati): `http://stats.stackexchange.com/`
- Stack Overflow (principalmente programmazione): `http://stackoverflow.com/`
- R-Bloggers (alcune applicazioni utili): `http://www.r-bloggers.com/`
- Centinaia di tutorial su vari blog di programmatori/analisti dati/nerd generici

Crediti

- Per ottenere i crediti dovrete completare
  - Due “compiti a casa” durante il corso del laboratorio
  - Un esercizio di analisi dati/reportistica a fine laboratorio
- Gli esercizi sarnno gli stessi che abbiate frequentato o meno

Testi

- Essendo un laboratorio, non ci si basa su nessun libro
- Tuttavia, una risorsa utile è `R for Data Science` di Grolemund & Wickham (cliccate sul titolo per accedere al sito del libro, dove è consultabile gratuitamente)

Iniziamo dall'inizio

Che cosa è l'analisi dati?

- Una serie di attività che svolgiamo per **imparare** qualcosa dai dati
- Si può imparare in modo

- **Deduttivo**: abbiamo una teoria, ci serve evidenza per confermarla o falsificarla
  - Vogliamo capire se la nostra lettura di un fenomeno generale è corretta o no
- **Induttivo**: cerchiamo *pattern* (o "schemi ricorrenti") nei dati
  - Vogliamo scoprire l'esistenza di fenomeni che ignoravamo in precedenza

Analisi dati quantitativa

- Il fenomeno a cui siamo interessati è **misurabile**
- Lo misuriamo tante volte, in diverse occasioni
- Utilizziamo metodi **statistici** o **grafici** per sintetizzare tutte le informazioni che abbiamo ottenuto dalle nostre misurazioni
- Abbiamo bisogno dell'aiuto di un software per fare questo

I dati

- Le nostre misurazioni ripetute sono quello che chiamiamo “i dati”
- Nella stragrande maggioranza dei casi sono disposti lungo 2 dimensioni:
  - **Osservazioni**: le “occasioni” di misurazione, i “casi” su cui la nostra teoria si focalizza (ad esempio, individui)
  - **Variabili**: espressioni del fenomeno a cui siamo interessati (ad esempio erà, reddito, scelta di voto)
- Questa struttura (osservazioni X variabili) si chiama “matrice”

Matrice dati

A	B	C	D	E	F	G	H	I	J
region	**female**	**age**	**vote2013**	**vote2013str**	**lrscale**	**netusoft**	**netustm**	**happy**	**health**
Calabria	1	68	8	Other	5	4	60	4	2
Sardegna	1	58	1	PD	88	5	120	8	1
Puglia	0	28	2	M5S	5	5	270	8	4
Lombardia	0	31	9	Did not vote	77	5	120	0	3
Sicilia	1	46	10	No answer	5	5	30	0	3
Lazio	0	30	1	PD	4	4	60	8	4
Emilia-Romagna	1	73	9	Did not vote	1	1		4	1
Lombardia	1	73	1	PD	5	5	120	7	3
Emilia-Romagna	1	39	2	M5S	88	5	120	8	4
Sardegna	0	20	9	Did not vote	3	5	300	9	4
Sardegna	0	52	3	PDL	7	5	180	6	3
Sardegna	1	78	10	No answer	7	1		3	2
Emilia-Romagna	0	45	2	M5S	5	5	180	7	3
Lombardia	1	33	10	No answer	6	5	180	9	4
Calabria	0	68	1	PD	5	2		10	2
Valle d'Aosta	0	31	9	Did not vote	8	3		7	4
Emilia-Romagna	0	52	1	PD	5	5	120	9	4
Sicilia	0	33	9	Did not vote	88	4	60	10	3
Veneto	0	63	10	No answer	10	4	60	6	3
Puglia	1	73	10	No answer	6	4	60	10	2
Campania	0	58	9	Did not vote	5	4	30	7	2
Sicilia	0	62	9	Did not vote	3	4	90	9	2
Friuli-Venezia Giulia	0	75	10	No answer	6	1		8	2
Basilicata	0	48	1	PD	0	2		3	2

Routine dell'analisi dati

Nella maggior parte dei casi, fare "analisi dati" significa fare una o più di queste cose:

- Caricare ed esplorare i dati
- Data management: preparare i dati per le analisi
  - Ricodifiche
  - Trasformazioni
  - Aggregazioni
- Estrarre informazioni dai dati
  - Graficamente
  - Con analisi statistiche
- Riportare i risultati delle analisi

Cosa è R

- Uno strumento per fare **tutti** i passi nella routine per l'analisi dei dati, dal data management alla reportistica
- Ma anche
  - Un **linguaggio** di programmazione
  - Un "interpreter" che esegue il codice scritto nel linguaggio R
  - Un motore grafico
  - Una applicazione che include l'interpreter, il motore grafico, librerie, e un'interfaccia utente

Perchè R?

- R è gratuito
- R rende possibile implementare quasi ogni tecnica di analisi dati (potete fare *quasi* tutto con R)
- R è flessibile
- Imparando R acquisirete la conoscenza di un linguaggio di programmazione più generale di una sintassi *ad-hoc* che può essere utilizzata solo per analizzare i dati (come quella di Stata)
- La conoscenza di R è sempre più richiesta dalle aziende per posizioni di analista dati

Caratteristiche di R

- R è un linguaggio "**object oriented**"
- Gli "oggetti" sono entità identificate da un nome e da un contenuto
- Potete mettere diverse cose all'interno di un oggetto: numeri, parole e frasi, dataset, funzioni, grafici, ecc.
- R può leggere dati scritti da Stata, SPSS, Excel, e qualsiasi altro formato dati
- R può essere utilizzato con diverse interfacce grafiche ("*graphical user interface*", GUI). In questo laboratorio useremo **RStudio**

R vs. RStudio

- RStudio è un'interfaccia per R
- Vantaggi
  - Semplifica il flusso di lavoro (*workflow*)
  - Permette di implementare diverse funzionalità molto utili:
  - Version control (git, SVN)
  - Creare documenti direttamente con R (Word, PDF, slide)
- Svantaggi
  - Può causare dipendenza (seriamente!)

Installare R

`https://www.r-project.org/`

The CRAN – Comprehensive R Archive Network

- `https://cran.r-project.org/mirrors.html`
- Da qui potete scaricare il programma di installazione e la maggior parte delle librerie aggiuntive
- Potrete dover scegliere un "*mirror*", un server da cui scaricare il programma e le librerie

Il sito include anche un manuale ufficiale e diverse risorse:

`https://cran.r-project.org/manuals.html`

Installare RStudio

`https://rstudio.com/products/rstudio/download/`

- Useremo **RStudio Desktop – Open Source License**
- L'installazione può avvenire separatamente da quella di R

Interagire con R

- Si interagisce con R utilizzando la **sintassi**
- Cosa significa scrivere "sintassi"?
  - Scrivere i comandi in un linguaggio che R può capire
- Con R l'utilizzo della sintassi è inevitabile (al contrario di SPSS o Stata)
- Le risorse online mostrate prima servono a darvi una mano

R e tipi di file

- `.r`: un file di sintassi (nulla di più che un file di testo grezzo)
- `.rds`: oggetti R (possono essere dataset o qualciasi altra cosa)
- `.RData`: oggetti R, ma anche l'intero "*workspace*" (l'insieme di dati, funzioni e altri oggetti che avete creato)
- `.Rhistory`: la "storia" della vostra sessione (tutto quello che avete fatto all'interno della sessione corrente)

Le prossime sessioni