Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZIONE 1 - Overview: introduzione al corso, alla materia e al software (R)

Docente: Valentina MINI
valentina.mini@unife.it

RICEVIMENTO: sempre su appuntamento previa mail
Lunedi 14:00-16:00
c/o Dipartimento matematica 3° Piano

docente

Valentina MINI

Professore Aggiunto, Università degli Studi di Ferrara (Dipartimento di Matematica; Dipartimento di Economia e Management)

Già professore Aggiunto, Università della Svizzera Italiana (Lugano, CH)
Già Adjunct Professor, Franklin University (Lugano, CH)

Valutatore esperto di progetti di formazione (For.Te; Fondo Professioni; INTERREG; graduatoria ANVUR)

i discenti


HELLO
MY NAME IS!
HELLO
MY NAME IS!
HELLO
MY NAME IS!
HELLO
MY NAME IS!

q.1 – Hai frequentato corsi di statistica nella tua carriera accademica? Quali?

.....
.....
.....

q.2 – Hai utilizzato software di analisi e di calcolo? Quali?

.....
.....
.....

q.3 – Qual è il tuo piano di studio? Scrivi alcune indicazioni per descriverlo

.....
.....
.....

q.4 – Quali sono i tuoi interessi accademici? (es. ricerche effettuate, corsi particolarmente amati, ecc.)

.....
.....
.....

q.5 – Su quale argomento si è basata la tua tesi triennale?

.....
.....
.....

q.6 – Ho scelto Statistica Multivariata perché ...

INTRODUZIONE AL CORSO

Introduzione al corso

OBIETTIVI DEL CORSO

- Basi **teoriche** alla statistica multivariata e alle tecniche utilizzate **per l'analisi** multivariata
- Indipendenza **nell'impostazione analitica applicata**
- Impostazione e soluzione analitica in **R**

MATERIALE DIDATTICO

Slide presentate durante le lezioni

Libri di testo, articoli o paper consigliati durante le lezioni per approfondimenti (ad es. Lattin J. Et al (2003) Analyzing Multivariate Data. Thomson Brooks/Cole (Canada)).

Software utilizzati: R (free source) già presente nei pc in dotazione

Introduzione al corso

STRUTTURA DELLE LEZIONI

Dove: laboratorio A2 Palazzo Manfredini

Quando: Lunedi: 16-18; Mercoledi: 14-16

Struttura lezioni:

-Fondamenti teorici
-Esemplificazioni
-Esercizi
-Applicazione in R (dove previsto)
-Selezione di letture consigliate

Introduzione al corso

ESAME FINALE E VALUTAZIONE

La valutazione finale si baserà su una **verifica** di apprendimento **DURA 90':**

A) **TEST A RISPOSTA MULTIPLA (15 DOMANDE IN 60')** che includerà una parte teorica e una parte più applicata, con testo dell'esame, modulo delle risposte e fogli protocollo per effettuare eventuali esercizi pratici:
- Risposta corretta +2
- Risposta errata -1
- Risposta non data 0

B) **SCRIPT DI UN'ANALISI IN R (30')** da effettuarsi al pc su R utilizzando un database fornito dalla docente con relativa descrizione. Lo script (originariamente salvato in .txt) verrà salvato in formato PDF su supporto informatico in presenza dello studente e successivamente corretto dalla docente.

Appelli:

a) Gennaio/febbraio
b) Giugno/luglio
c) Settembre

Introduzione al corso

PROGRAMMA

giorno	lezione	ora	argomento	h
07-ott-19	2	16-18	Introduzione al corso, alla materia e all'ambiente R	2
09-ott-19	3	14-16	Vettori e matrici: overview e laboratorio in R	2
14-ott-19	4	16-18	Modello di regressione lineare semplice	2
16-ott-19	5	14-16	Applicazione pratica: MRLS in R	2
21-ott-19	6	16-18	Questioni analitiche e interpretazione di MRLS	2
23-ott-19	7	14-16	Modello di regressione lineare multivariata	2
28-ott-19	8	16-18	Applicazione pratica: MRLM in R	2
30-ott-19	9	14-16	Analisi per componenti principali (PCA)	2
04-nov-19	10	16-18	PCA: Applicazione pratica in R	2
06-nov-19	11	14-16	Analisi fattoriale	2
11-nov-19	12	16-18	AF: applicazione pratica in R	2
13-nov-19	13	14-16	Approfondimento: analisi fattoriale confermativa ed esplorativa	2
18-nov-19	14	16-18	Analisi per gruppi (CA)	2
20-nov-19	15	14-16	Cluster gerarchici e applicazione in R	2
25-nov-19	16	16-18	Cluster non gerarchici e applicazione in R	2
27-nov-19	17	14-16	Cluster gerarchici e non gerarchici: laboratorio in R	2
02-dic-19	18	16-18	Test di permutazione	2
04-dic-19	19	14-16	Analisi di dipendenza e interdipenenza: overview	2
09-dic-19	20	16-18	ESERCITAZIONI (RLS-RLM)	2
11-dic-19	21	14-16	ESERCITAZIONI (FA-PCA)	2
16-dic-19	22	16-18	ESERCITAZIONI (CA)	2

INTRODUZIONE ALLA MATERIA

Introduzione alla materia
LA STATISTICA MULTIVARIATA

Nella Statistica Multivariata si cerca di trovare ***relazioni tra dati in alta dimensione*** *(ossia in spazi Rm per 40 m 100, dove Rm denota il prodotto cartesiano di R fatto m volte, individuando dei vettori riga).*
Una delle difficoltà maggiori è la "Maledizione della Dimensione" (Curse of Dimensionality): essa si manifesta poiché per m abbastanza grande la metrica euclidea "perde" gran parte del suo significato.

Quasi **nessun problema** statistico e' caratterizzato da **una sola** variabile:

I fenomeni oggetto di studio sono spesso il risultato ***di molteplici elementi concomitanti*** *difficilmente controllabili.*

L'esistenza di **molte variabili interagenti l'una con l'altra** complica alquanto l'analisi rispetto all'ideale caso univariato:

le procedure statistiche univariate possono essere ***generalizzate,*** *ma la complessità aumenta sempre più all'aumentare delle dimensioni del problema.*

Introduzione alla materia
LA STATISTICA MULTIVARIATA

Col termine ***analisi multivariata*** *si indica* ***quell'insieme di metodi statistici usati*** per analizzare simultaneamente più caratteri

Cosa si intende per **non multivariato?**
La statistica più semplice: singola variabile.

Possiamo:
- raccogliere molti dati (o effettuare delle *realizzazioni* della variabile aleatoria);
- sintetizzare questi dati calcolando media, varianza etc.;
- confrontare due diverse collezioni di dati per la stessa variabile aleatoria (es. gruppi di controllo).

In tutto ciò però **non** ci si pone il problema di **capire**
da dove provengano eventuali differenze tra le collezioni di campioni

MULTIVARIATO→ dati raccolti che presentano più caratteristiche per ogni dato:
- quando ogni esperimento ha per risultato **non un singolo valore** (realizzazione di una singola variabile aleatoria) ma
- un numero $n > 1$ **di risultati** ognuno realizzazione di una diversa variabile aleatoria.

Introduzione alla materia
LA STATISTICA MULTIVARIATA

ES: al momento di una donazione di sangue vengono effettuate analisi sanguigne (quelle che seguono sono le variabili misurate presso il Servizio di Immunoematologia e Medicina trasfusionale dell'Ospedale di Cona)

Immunoematologia:
- Gruppo sanguigno
- Fattore Rh
- Fenotipo Rh
- Kell

Sierologia
- Epatite B: HBsAg
- Epatite C: HCVAc IgG
- Epatite C: HCVAc
- HIV 1-2: HIVAc IgG/IgM
- LUE: Ac IgG/IgM
- Transaminasi

Chimica Clinica:
- Glucosio
- Creatina
- Proteine totali
- Colesterolo totale
- Trigliceridi
- Ferro

Introduzione alla materia
LA STATISTICA MULTIVARIATA

•Per ogni *caso* preso in esame vengono misurate un **numero elevato di variabili.**

•**Per ogni variabile**, ovviamente, è possibile fare **analisi statistiche separate.**

•La presenza di **tante informazioni** : posso esserci altri modi di analisi, per evidenziare la dipendenza reciproca delle variabili.

•Spesso si vogliano utilizzare i dati per **classificare i casi** esaminati in **opportune categorie** (es. rapporto peso/ altezza tra "gravemente sottopeso", "sottopeso", "normopeso", "sovrappeso", "obeso").

•Infine anche nel caso multivariato dobbiamo **distinguere tra la fase *descrittiva* e quella *inferenziale*** (caso monovariato: interesse nell'inferenza; caso multivariato anche la parte descrittiva!)

Introduzione alla materia

NATURA DEI DATI MULTIVARIATI

Sviluppo delle analisi multivariate: ultimi due decenni

2 ragioni principali

1) **Comprensione della COMPLESSITA' del comportamento** (umano, di situazioni ecc.) incluse le caratteristiche multivariate

2) **Sviluppo di IT** che ha permesso:
   - **Analizzare** e processare un ampio ammontare di informazioni
   - **Misurare** fattori che influenzano direttamente o indirettamente il fenomeno di interesse

Il problema non consiste nella carenza di dati o di capacità di processarli;
ma nella carenza di abilità per estrarre informazioni interessanti dai molti dati a disposizione
NB → METODI

Corso:
METODI MULTIVARIATI + LORO APPLICAZIONE + PROBLEMATICHE TIPICHE NELL'ANALISI

→ descrizione della FILOSOFIA DELLE TECNICHE MULTIVARIATE e come l'analisi dei dati rientra nel processo più ampio di ricerca e decision making

Introduzione alla materia
ALCUNE DEFINIZIONI

Analisi multivariata consiste nello studio dell'**associazione tra insiemi di misurazione.**

Metodi Multivariati come l'insieme delle procedure per analizzare l'associazione tra 2 o più insiemi di misurazioni che sono state rilevate su ciascun **oggetto** in 1 o più campioni di oggetti (J. Lattin et al. 2003, p.4).

OGGETTI: possono essere cose, persone, situazioni, organizzazioni, eventi ecc.
Sono le entità sulle quali viene effettuata la misurazione.

OBIETTIVO:
focus su veri dati multidimensionali con 3 o più set di misurazioni

Introduzione alla materia
ALCUNE DEFINIZIONI

Esempio
Consulenza per società di investimento immobiliare USA.

Caratteristiche di **31 proprietà immobiliari** sulla costa di San Francisco in un **arco temporale 10 anni**

In tali dati: ogni proprietà è un oggetto sul quale sono state effettuate varie misurazioni:
- es. la dimensione della proprietà (espressa in ettari)
- es. la altitudine slm (in metri)

OGGETTI (o unità statistiche): Misurati non nella completezza, ma solo rispetto ad alcune variabili di interesse.

VARIABILI (talvolta chiamate caratteristiche o proprietà): sono gli aspetti degli oggetti che vengono misurati. Ad es. la dimensione della proprietà è una variabile che descrive un particolare aspetto di ognuna delle 31 proprietà.

DATI: produciamo un database 31 righe X 2 colonne (+ intestazioni)

Introduzione alla materia
OSSERVAZIONE E DATI

Riprendendo Coombs (1964) ci sono importanti
differenze tra osservazioni e dati

PASSAGGI NEL PROCESSO DA OSSERVAZIONE A DATO
RICHIEDE ALCUNI FONDAMENTI TEORICI

1) Il ricercatore si confronta con il problema "**Quale fenomeno osservare?**"

2) **Livello di misurazione** e **tipologia** di serie **di dati**

Introduzione alla materia

1) OSSERVAZIONE DAI DATI

Quale fenomeno osservare? (in base al focus)

Es. proprietà: il ricercatore era interessato a confrontare il $ di vendita delle proprietà con caratteristiche della stessa

→ sceglie di focalizzarsi su alcune variabili (es. dimensione, altitudine slm, tempo in vendita) che precedentemente (a priori) aveva ritenuto potessero influenzare la variabile di interesse.

è importante **la scelta del focus** nell'universo delle variabili che possono essere osservate

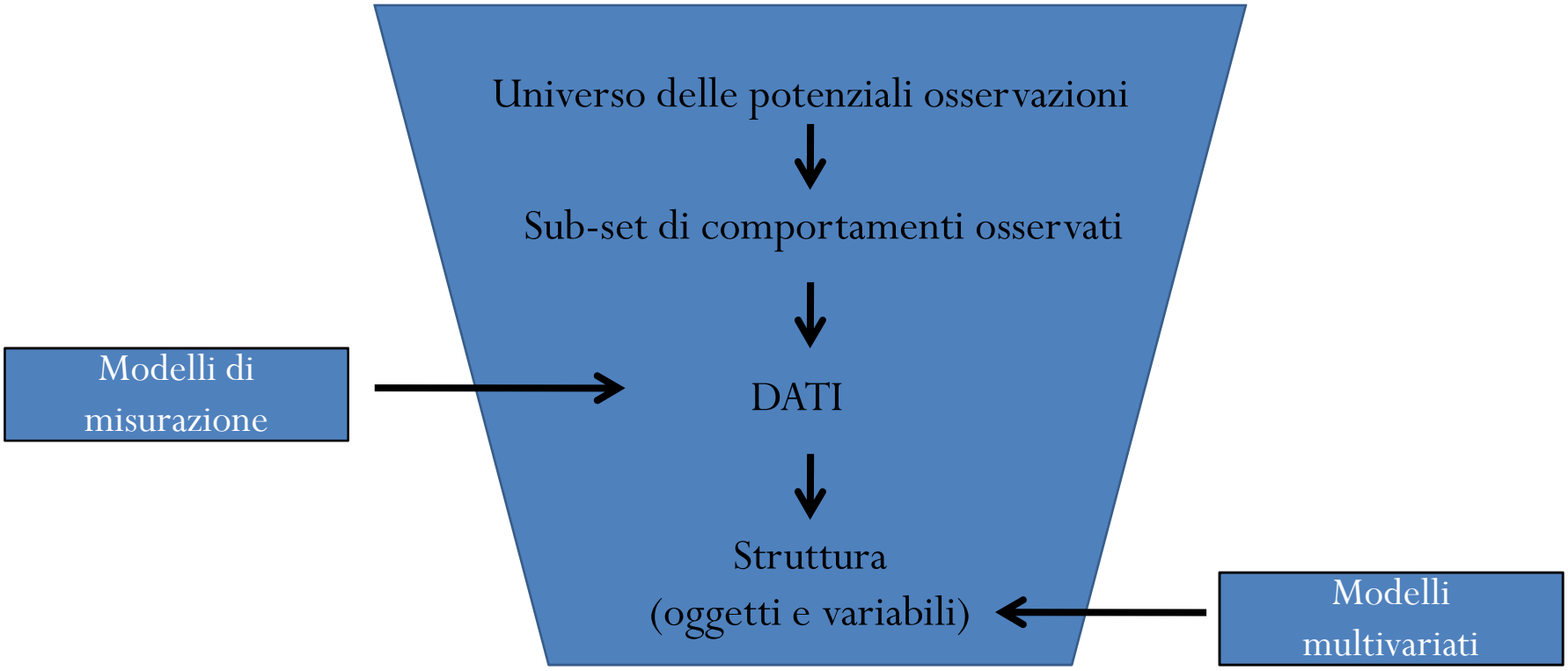
COME SCEGLIERE LE VARIABILI?

- teoria
- pratica
- precedenti esperimenti da provare o confutare

Introduzione alla materia

1) OSSERVAZIONE DEI DATI

Processo di ricerca (Coombs, 1964)



Introduzione alla materia

1) OSSERVAZIONE DEI DATI

Figura:
Disegno di ricerca del ricercatore CONDIZIONA l'universo delle osservazioni potenziali
Anche i comportamenti osservati devono essere codificati (insieme di regole)

→l'interazione tra ricercatore e ambiente delle potenziali osservazioni → guida i dati

•Ultimo passo = guardare la **struttura di associazioni** nei dati per far calzare i modelli multivariati
•Ogni scelta effettuata = effetti restrittivi (selezione)
•Il ricercatore procede per **cercare pattern di regolarità** nel comportamento di 2 o più variabili
• Successivamente alla scoperta del pattern l'obiettivo è quello di individuare il **modello che meglio corri sponde ai dati**

NB **regole nella raccolta dei dati e strutturazione del database**

La validità dell'inferenza statistica può essere molto influenzata **dalle scelte** effettuate, dal **processo della ricerca** e dai cicli di **selezione del modello** per scoprire e codificare regolarità nei dati o testare prime ipotesi sui pattern di osservazione

Introduzione alla materia

2) LIVELLI DI MISURAZIONE

La **MISURAZIONE** è
il processo attraverso il quale **i numeri sono attribuiti alle caratteristiche**
o proprietà degli oggetti.

Es. regole per assegnare un numero corrispondente alla dimensione di una proprietà

Una scala di misurazione identifica:

-Quanta informazione è contenuta nella misura

- Misura e relazione tra due oggetti

Stevens (1946): 4 tipologie di misurazione (e relative scale)

- Nominale
- Ordinale
- Intervallare
- Razionale

Ogni livello: diversa quantità di informazione e diversi tipi di confronto tra oggetti esaminati

Introduzione alla materia

2) LIVELLI DI MISURAZIONE

DATI (e scale) NOMINALI (o categorici)

Unica informazione su oggetto: rientra in gruppi
mutualmente esclusivi e
collettivamente esaustivi

Gruppi: **non ordine**

Es. ultima marca di jeans scelta

A – B – C – D : possiamo arbitrariamente assegnare 1 ad un individuo che nell'ultimo acquisto ha scelto A; assegnare 2 ad un individuo che nell'ultimo acquisto ha scelto B

Con scale nominali: NUMERI = info solo su ETICHETTA della categoria
SBAGLIATO: confronto numeri o operazioni matematiche

Utilizzati spesso nei metodi multivariati, ma codificati in modo appropriato →

Introduzione alla materia

2) LIVELLI DI MISURAZIONE

DATI (e scale) NOMINALI (o categorici)

CODIFICARE = creare una variabile separata (ad es. che assuma solo valori 0 o 1) per ogni categoria definita attraverso i dati nominali

ES. individui che, durante l'ultimo acquisto, hanno effettuato la loro scelta fra 4 brand di jeans.

Brand ultima scelta	Variabile nominale
A	1
B	2
C	3
D	4

Codifichiamo l'informazione: creiamo altre 4 variabili con modalità 1-4

VLORI BINOMIALI (DUMMY):
1 ⇔ brand scelto è il brand di interesse
0 altrimenti
NE BASTANO 2!

Introduzione alla materia

2) LIVELLI DI MISURAZIONE

DATI (E SCALE) ORDINALI
Variabili misurate in modo **ordinale**
Su 2 oggetti i e j → INFORMAZIONE = oggetto J ha +,-,= **ammontare di caratteristiche** rispetto a oggetto i

ES. preferenze ordinate in ranking
Chiediamo a individuo di ordinare i 4 brand di jeans
dal "PIU' PREFERITO" (assegniamo valore 1) →al "MENO PREFERITO" (valore 4)

Se preferenza individuo A>B>C>D => la scala ordinale assumerà gli stessi valori della scala nominale presentata in tabella (1,2,3,4)

POSSIBILE INFERENZA: la **preferenza** per il brand A> preferenza per il brand B
NON possiamo esprimerci su QUANTO A sia preferito rispetto a B ecc.

NON HA SENSO PROCEDERE CON OPERAZIONI ARITMETICHE

Introduzione alla materia

2) LIVELLI DI MISURAZIONE

DATI (E SCALE) INTERVALLARI

Permettono di dire QUANTO della caratteristica misurata è posseduta da A rispetto a C
MA
Non permettono un confronto assoluto rispetto ad uno 0 significativo

Es. Temperatura misurata in gradi Centigradi: in questo caso
-0 = è definito arbitrariamente al punto in cui l'acqua ghiaccia.
L'origine è arbitraria perché non ha il significato di "totale assenza di temperatura"
-100 = definito sulla situazione in cui l'acqua bolle.

→Non è strettamente corretto dire che un oggetto con temperatura rilevata di 100°C è "caldo il doppio" rispetto ad un oggetto con temperatura rilevata pari a 50°C.

POSSIBILE INFERENZA: paragonare le differenze su scala intervallare
NON HA SENSO il confronto rispetto ad un'origine arbitraria

Introduzione alla materia

2) LIVELLI DI MISURAZIONE

DATI (E SCALE) INTERVALLARI

Es. Rilevata temperatura in °C di 4 oggetti:

A) 80°C ; B) 100°C ; C)80°C ; Z)70°C
Possiamo dire che la differenza [B-A]> differenza [C-Z]

Notiamo che **questo confronto non è qualitativamente differente da ogni trasformazione lineare positiva (X=a+bx con a>0)**
Quindi stessa conclusione qualitativa se usiamo °F (dove F= 32+ 9/5 °C)

POSSIAMO EFFETTUARE OPERAZIONI ARITMETICHE assumendo che la scala di valori abbia almeno le proprietà intervallari

ES. PREF. MEDIA ⇔ DIFFERENZA COSTANTE

Diverse preferenze brand	Valore scalare
MOLTO ALTA	5
ALTA	4
MODERATA	3
BASSA	2
MOLTO BASSA	1

Introduzione alla materia

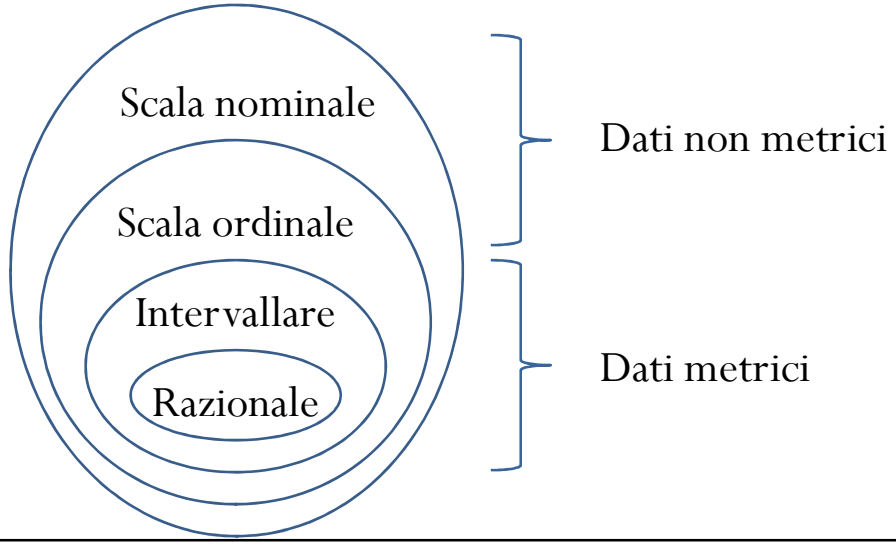
2) LIVELLI DI MISURAZIONE

DATI (E SCALE) RAZIONALI

La scala di valori razionali possiede le stesse proprietà di quella intervallare e in più **possiede un'origine significativa**

Es. Variabili come ETA' e PREZZI hanno proprietà delle scale di dati razionali
Confronto tra individui con diversa età [A)10 e B)20 => B ha il doppio di età di A]

Passando da scala NOMINALE a scala RAZIONALE la quantità di informazioni aumenta



Introduzione alla materia

METODI MULTIVARIATI

I **dati multivariati** (metrici o non-metrici) sono organizzati in **strutture dette MATRICI**

	VARIABILI				
OGGETTI		1	2	…	P
	1	X_{11}	X_{12}		X_{1p}
	2				
	3				
	n	X_{n1}			X_{np}

n= numero di oggetti (uno per ogni riga)

Ogni colonna della matrice = variabili (p)
Caratteristiche misurate dell'oggetto nel campione

Introduzione alla materia

METODI MULTIVARIATI

OVEVRVIEW DEI METODI MULTIVARIATI

DISTINZIONI:

- TECNICA UTILIZZATA PER ANALISI DI **INTERDIPENDENZA O DI DIPENDENZA**
- TECNICA UTILIZZATA PER **OBIETTIVO ESPLORATIVO O CONFERMATIVO**
-TECNICA DISEGNATA PER ESSERE UTILIZZATA CON **DATI METRICI O NON METRICI**

Questa distinzione segue la filosofia del corso.

Introduzione alla materia

METODI MULTIVARIATI

METODI PER L'ANALISI DELLA DIPENDENZA TRA VARIABILI

-Necessità di comprendere, spiegare il legame di dipendenza tra una variabile esplicativa (x) e una variabile dipendente (y) → **REGRESSIONE LINEARE SEMPLICE**

-Necessità di comprendere, spiegare il legame di dipendenza molteplici variabili esplicative (x1, x2, x3…, xn) e una variabile dipendente (y) → **REGRESSIONE LINEARE MULTIPLA**

-Necessità di prevedere valori ignoti attraverso il legame di dipendenza lineare tra due o più variabili (**MODELLO DI REGRESSIONE**)

Introduzione alla materia

METODI MULTIVARIATI

PCA = riduzione della dimensione di dataset multivariati

permette di ri-esprimere i dati (attraverso una combinazione lineare delle variabili originarie) così che le prime poche variabili risultanti (dette componenti) contano per la maggior parte dell'informazione.

riduzione di dimensione = 1) visualizzazione più semplice

2) analisi più gestibile

TRADE-OFF fra semplicità e completezza (il ricercatore deve decidere quante componenti considerare)

-FACTOR ANALYSIS

-MULTIDIMENSIONAL SCALING

-CLUSTER ANALYSIS

Introduzione alla materia

METODI MULTIVARIATI

FACTOR ANALYSIS = riduzione dei dati multivariati, ma # da PCA perché modelli sottostanti diversi

Analisi fattoriale: **individuazione di fonti di varianza comune a 2 o + variabili** (chiamate fattori comuni). Tuttavia non c'è un modello sottostante di misurazione delle componenti (PCA: ogni componente è una combinazione lineare esatta delle variabili originarie).

-PUO' ESSERE UTILIZZATA PER **SCOPO ESPLORATIVO O CONFERMATIVO**

-ESPLORATIVA → inferenza sulla struttura dei fattori a partire dai pattern di correlazione nei dati
-CONFERMATIVA → testare una nozione iniziale forte per vedere se è consistente con i nostri dati
NB per queste tecniche esistono test di bontà del modello (per testare il livello di significatività di diversi modelli)

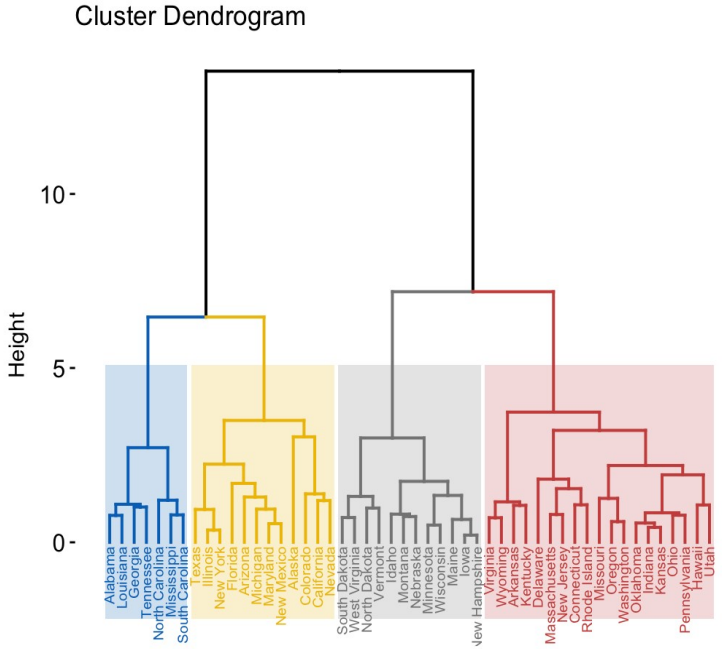
Introduzione alla materia

METODI MULTIVARIATI

CLUSTER ANALYSIS O ANALISI PER GRUPPI :

Le osservazioni in ogni gruppo sono simili;

Le osservazioni in gruppi diversi sono diverse (distanza)



Si fornisce una variabile nominale che indica l'appartenenza di ogni oggetto ad un cluster.

DIVERSI METODI:

a) **Gerarchici** → bottom-up (aggregativi) o top-down (divisivi)

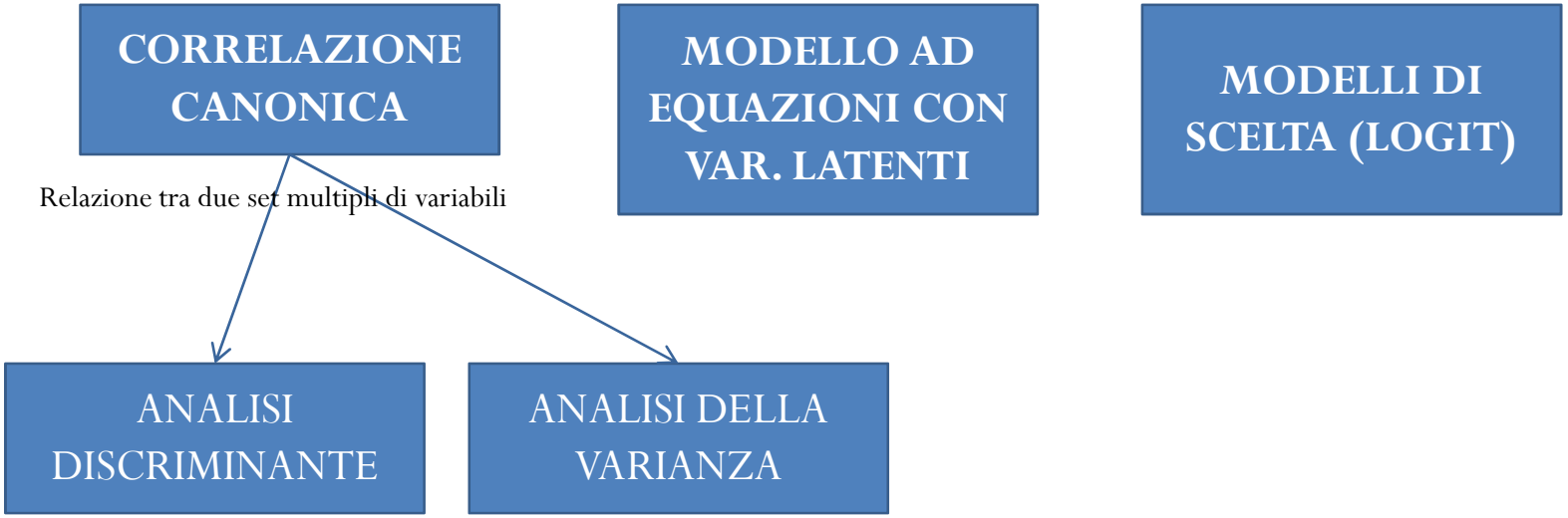
b) **Non gerarchici o a partizione**

SOLUZIONE ALLA ETEROGENEITA' NEI DATI, ma creazione di cluster non è come trovare cluster naturali.

Introduzione alla materia

METODI MULTIVARIATI

ALTRI METODI ATTUALI DI ANALISI MULTIVARIATA



INTRODUZIONE AI DATABASE

Passito

A marketing survey on the demand of the wine «Passito» has been performed.

A sample of n=386 people has been interviewed. The questionnaire includes several questions about their preferences and behaviors related to drinking wine.

Dataset variables:

Label	Description	Coding
ID	Personal ID of the interviewed	Increasing integer number
AgeClass	Age of the person	Age (years)
AGE_CLASS	Age class of the person	1-6
SEX	Sex of the person	M or F
PROV	Province where the interviewed lives	Province code
LIKE_WINE	How much do you like drinking wine?	Integrer number from 1 to 7
FREQ_HOME	How often do you drink wine at home with meals?	Integrer number from 1 to 5
FREQ_BAR	How often do you drink wine in bars/pubs?	Integrer number from 1 to 5
FREQ_REST	How often do you drink wine at restaurants with meals?	Integrer number from 1 to 5
KNOW_PAS	Do you know the wine Passito?	Integrer number from 1 to 7
FREQ_PAS	How often do you drink Passito?	Integrer number from 1 to 5
FREQ_P_HOL	How often do you drink Passito on holidays and celebrations?	Integrer number from 1 to 5
FREQ_P_ALO	How often do you drink Passito when you are alone?	Integrer number from 1 to 5
FREQ_P_MEA	How often do you drink Passito at the end of meals?	Integrer number from 1 to 5
FREQ_P_OFF	How often do you drink Passito offered by someone?	Integrer number from 1 to 5
HOW_MUCH	How much wine do you drink in one year?	Integrer number from 1 to 4
LIKE_PAS	How much do you like drinking Passito?	Integrer number from 1 to 7
LIKE_AROMA	How much do you like aroma and smell of Passito?	Integrer number from 1 to 7
LIKE_SWEET	How much do you like the sweetness of Passito?	Integrer number from 1 to 7
LIKE_ALCOHOL	How much do you like the alcohol content of Passito?	Integrer number from 1 to 7
LIKE_TASTE	How much do you like the intensity of taste of Passito?	Integrer number from 1 to 7
PRICE	How much could you pay for one bottle of Passito? (0.5 litre)	Integrer number from 1 to 5

Heating Habits

Official data by Food and Agricultural Organization (FAO) about per capita food consumption by type of food.

The set of 126 countries with a population greater than 3 milions of people have been considered.

Dataset variables:

- Alcoholic Beverages
- Cereals
- Fruits
- Starchy Roots
- Sugar
- Veg Oils
- Animal Fats
- Meat
- Eggs
- Fish
- Veg_pulses
- Milk
- Population

Hotel

A customer satisfaction survey where four hotels have been evaluated by 40 customers (10 for each hotel) with respect to k=3 variables: cleanliness, courtesy and price.

The data consist of rates from 0 (minimum satisfaction) to 100 (maximum satisfaction).

Dataset variables:

Name	***Type***
Hotel	Categorical
Cleanliness	Numeric
Courtesy	Numeric
Price	Numeric

Students

Let us consider an example of teaching evaluation of k=3 university programs (undergraduate degree in Economics) evaluated by n=20 students with a rate from 0 to 100.

Dataset variables:

Statistics
Mathematics
Econometrics

Mall

A customer satisfaction survey about a a recently opened shopping center.

A sample of n=29 customers was asked to evaluate k=5 different aspects of the shopping center, such as the environmental temperature, the brightness, the presence of sales assistents, the range of products, the background music volume.

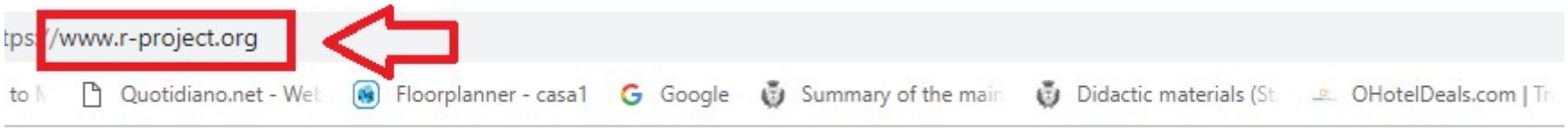
Evaluations are expressed on a scale from -100 («too little») to +100 («too much»), where 0 corresponds to «just right».

Dataset variables:

Temp_Level
Brightness
Salesman
Product_assortmant
Music_volume

INTRODUZIONE ALL'AMBIENTE R

IL SITO DI RIFERIMENTO



[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Conferences
Search
Get Involved: Mailing Lists
Developer Pages
R Blog

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- The R Foundation Conference Committee has released a call for proposals to host useR! 2020 in North America.
- You can now support the R Foundation with a renewable subscription as a supporting member
- **R version 3.5.1 (Feather Spray)** has been released on 2018-07-02.
- The R Foundation has been awarded the Personality/Organization of the year 2018 award by the professional association of German market and social researchers.

Introduzione al software

Scopo: illustrare i fondamenti logici ed applicativi di R.
R = ambiente, ovvero un insieme di macro, librerie, oggetti che possono essere utilizzati per la gestione, l'analisi dei dati e la produzione di grafici

"**E possibile** in R implementare….?",
deve essere sostituita da
"**Quanto è difficile** in R implementare…?"

R e basato sul linguaggio S a cui è strettamente legato un altro 'ambiente' commerciale probabilmente pi`u conosciuto, S-Plus.
R, a differenza di S-Plus, è un GNU-Sofware, ovvero disponibile gratuitamente sotto i vincoli della GPL (General Public Licence).

R Reference Card

by Tom Short, EPRI PEAC, tshort@epri-peac.com 2004-11-07

Getting help

Most R functions have online documentation.

help(topic) documentation on topic
?topic id.
help.search("topic") search the help system
apropos("topic") the names of all objects in the search list matching the regular expression "topic"
help.start() start the HTML version of help
str(a) display the internal *str*ucture of an R object
summary(a) gives a "summary" of a, usually a statistical summary but it is *generic* meaning it has different operations for different classes of a
ls() show objects in the search path; specify pat="pat" to search on a pattern
ls.str() str() for each variable in the search path
dir() show files in the current directory
methods(a) shows S3 methods of a
methods(class=class(a)) lists all the methods to handle objects of class a

Input and output

load() load the datasets written with save
data(x) loads specified data sets
library(x) load add-on packages
read.table(file) reads a file in table format and creates a data frame from it; the default separator sep="" is any whitespace; use header=TRUE to read the first line as a header of column names; use as.is=TRUE to prevent character vectors from being converted to factors; use comment.char="" to prevent "#" from being interpreted as a comment; use skip=n to skip n lines before reading data; see the help for options on row naming, NA treatment, and others
read.csv("filename",header=TRUE) id. but with defaults set for reading comma-delimited files
character or factor columns are surrounded by quotes ("); sep is the field separator; eol is the end-of-line separator; na is the string for missing values; use col.names=NA to add a blank column header to get the column headers aligned correctly for spreadsheet input
sink(file) output to file, until sink()

Most of the I/O functions have a file argument. This can often be a character string naming a file or a connection. file="" means the standard input or output. Connections can include files, pipes, zipped files, and R variables.

On windows, the file connection can also be used with description = "clipboard". To read a table copied from Excel, use
x <- read.delim("clipboard")
To write a table to the clipboard for Excel, use
write.table(x,"clipboard",sep="\t",col.names=NA)
For database interaction, see packages RODBC, DBI, RMySQL, RPgSQL, and ROracle. See packages XML, hdf5, netCDF for reading other file formats.

Data creation

c(...) generic function to combine arguments with the default forming a vector; with recursive=TRUE descends through lists combining all elements into one vector
from:to generates a sequence; ":" has operator priority; 1:4 + 1 is "2,3,4,5"
seq(from,to) generates a sequence by= specifies increment; length= specifies desired length
seq(along=x) generates 1, 2, ..., length(along); useful for for loops
rep(x,times) replicate x times; use each= to repeat "each" element of x each times; rep(c(1,2,3),2) is 1 2 3 1 2 3; rep(c(1,2,3),each=2) is 1 1 2 2 3 3
data.frame(...) create a data frame of the named or unnamed arguments; data.frame(v=1:4,ch=c("a","B","c","d"),n=10); shorter vectors are recycled to the length of the longest
list(...) create a list of the named or unnamed arguments; list(a=c(1,2),b="hi",c=3i);
array(x,dim=) array with data x; specify dimensions like dim=c(3,4,2); elements of x recycle if x is not long enough
matrix(x,nrow=,ncol=) matrix; elements of x recycle
factor(x,levels=) encodes a vector x as a factor
gl(n,k,length=n*k,labels=1:n) generate levels (factors) by specifying the pattern of their levels; k is the number of levels, and n is the number of replications
expand.grid() a data frame from all combinations of the supplied vectors or factors

Indexing lists

x[n]	list with elements n
x[[n]]	n^{th} element of the list
x[["name"]]	element of the list named "name"
x$name	id.

Indexing matrices

x[i,j]	element at row i, column j
x[i,]	row i
x[,j]	column j
x[,c(1,3)]	columns 1 and 3
x["name",]	row named "name"

Indexing data frames (matrix indexing plus the following)

x[["name"]]	column named "name"
x$name	id.

Variable conversion

as.array(x), as.data.frame(x), as.numeric(x), as.logical(x), as.complex(x), as.character(x), ... convert type; for a complete list, use methods(as)

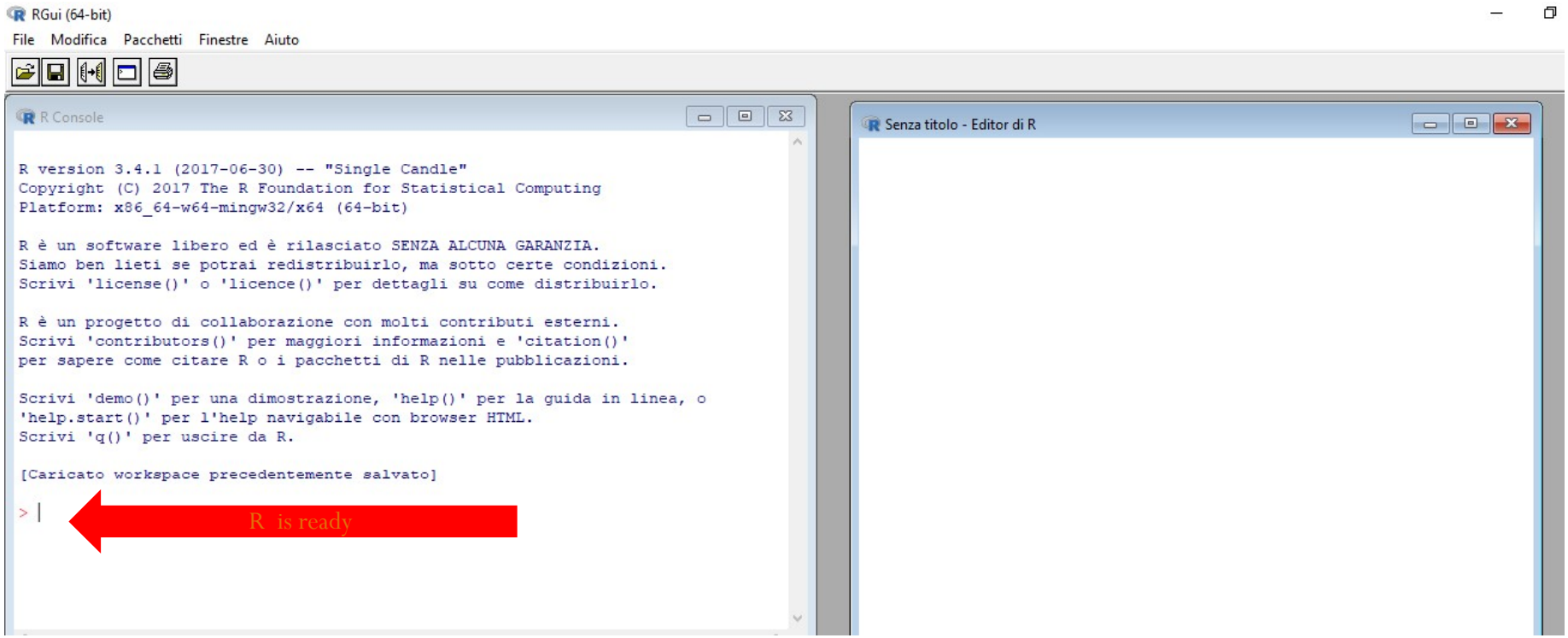
Variable information

is.na(x), is.null(x), is.array(x), is.data.frame(x), is.numeric(x), is.complex(x), is.character(x), ... test for type; for a complete list, use methods(is)
length(x) number of elements in x
dim(x) Retrieve or set the dimension of an object; dim(x) <- c(3,2)
dimnames(x) Retrieve or set the dimension names of an object
nrow(x) number of rows; NROW(x) is the same but treats a vector as a one-row matrix
ncol(x) and **NCOL(x)** id. for columns
class(x) get or set the class of x; class(x) <- "myclass"
unclass(x) remove the class attribute of x
attr(x,which) get or set the attribute which of x
attributes(obj) get or set the list of attributes of obj

Data selection and manipulation

which.max(x) returns the index of the greatest element of **x**
which.min(x) returns the index of the smallest element of **x**
rev(x) reverses the elements of x
sort(x) sorts the elements of x in increasing order; to sort in decreasing

INTERFACCIA DI R



Console: COMANDI E RISULTATI

EDITOR: scrivere e salvare i comandi (script)

ALCUNI ELEMENTI DI BASE

In R si utilizza il simbolo # per aggiungere commenti: la stringa di testo preceduta da # non viene considerato comando, ma commento

Just like Twitter # !!! I commenti non inficiano i vostri risultati

esempi:

calcolare 3 + 4 → è un commento

R PUO' FUNZIONARE COME CALCOLATRICE:

esercizi di base:

1) Using the console write the following comment: INTRODUCTION TO THE SOFTWARE
2) Calculate 3+6 and insert a comment explaining the computation
3) Calculate 3*3 and insert a comment explaining the computation
4) Calculate 3/3 and insert a comment explaining the computation
5) Calculate 6-3 and insert a comment explaining the computation
6) Calculate (3+3)*2 and insert a comment explaining the computation
7) Calculate 3^2 and insert a comment explaining the computation

ALCUNI ELEMENTI DI BASE

R COMUNICA CON VOI:

> R PRONTO PER RICEVERE COMANDI
[1] R STA COMUNICANDO IL RISULTATO
Error: R VI COMUNICA CHE AVETE COMMESSO UN ERRORE

ESERCIZIO:
APRITE R E IMPOSTATE I SEGUENTI COMANDI

```
#introduction to the software
# R can work as a simple calculator
3+6

# we compute the addition 3+6 and we obtain the result = 9
```

Pratica

Using the console write the following title (as a **comment)**:
INTRODUCTION TO THE SOFTWARE

Perform the following commands,
trying to understand the results and commenting them

1) Calculate 3+6 and insert a comment explaining the computation
2) Calculate 3*3 and insert a comment explaining the computation
3) Calculate 3/3 and insert a comment explaining the computation
4) Calculate 6-3 and insert a comment explaining the computation
5) Calculate (3+3)*2 and insert a comment explaining the computation
6) Calculate 3^2 and insert a comment explaining the computation

ALCUNI ELEMENTI DI BASE

SOME BASIC RULES:

-R is key sensitive (be careful !!! Capital and small letters are different!)

-If you don't close your command, R will waits for it

(ex. Write 3+4- and tape return key … please observe the result)

-How to save your work:
The command you want to save must be taped in the **EDITOR window**. - The **Editor should be saved using the** extension **.txt**

BE CAREFUL:
Data can be picked from an Excel dataset: in this case we must previously save it using the extension **.csv**

ALCUNI ELEMENTI DI BASE

2 operazioni separate da punto-e-virgola";"

Ex: 3+5*(3.5/15)+5-(2/6*4); 3+2

Operazioni che utilizzano la radice quadrata:

10+(7-2)*4-8/2+sqrt(9)

Crea una variabile

x

Assegna un valore ad una variabile (you may use = or direct arrow →)

x=6 # R registers the assignment

x and tape return key # R visualizes the content of the object x

Crea una serie di valori(o vectore):

v= c(9,5,4)

v

PRIMO LABORATORIO PRATICO IN R

APRIAMO R DA PC ED
ESEGUIAMO ALCUNI ESERCIZI INIZIALI
UTILIZZANDO COMANDI BASE