Frontespizio

Sommario

[1. Grafici ed indici di posizione centrali 7](#_Toc127015134)

[1.1 Variabili quantitative 7](#_Toc127015135)

[1.2 Grafico a barre delle persone risultate positive 7](#_Toc127015136)

[1.3 Indici di posizione centrali 8](#_Toc127015137)

[1.4 Grafico a barre ordinate 9](#_Toc127015138)

[1.5 Grafico a barre sul numero di persone in terapia intensiva 10](#_Toc127015139)

[1.6 Grafico a barre sul numero totale di persone ospedalizzate 11](#_Toc127015140)

[1.7 Grafico a barre sul numero di tamponi 12](#_Toc127015141)

[1.8 Grafico a barre sul numero di ingressi in terapia intensiva 13](#_Toc127015142)

[2. Frequenze e Distribuzioni 14](#_Toc127015143)

[2.1 Istogramma del numero di risultati positivi 14](#_Toc127015144)

[2.2 Indici di posizione non centrali 16](#_Toc127015145)

[2.3 Boxplot 17](#_Toc127015146)

[2.4 Boxplot del numero di risultati positivi suddivisi per regioni 17](#_Toc127015147)

[2.5 Principio di Pareto 19](#_Toc127015148)

[3. Dispersione e Funzione di Distribuzione 20](#_Toc127015149)

[3.1 Varianza 20](#_Toc127015150)

[3.2 Deviazione Standard 21](#_Toc127015151)

[3.3 Coefficiente di variazione 21](#_Toc127015152)

[3.4 Analisi degli indici di dispersione sulle variabili quantitative 22](#_Toc127015153)

[3.5 Funzione di Distribuzione Empirica 23](#_Toc127015154)

[3.6 Funzione di Distribuzione Empirica Discreta 23](#_Toc127015155)

[3.7 Funzione di Distribuzione Empirica Continua 24](#_Toc127015156)

[3.8 Funzione di Distribuzione Empirica Discreta del Numero di Risultati Positivi 24](#_Toc127015157)

[3.9 Skewness e curtosi campionarie 25](#_Toc127015158)

[4. Regressioni Lineari 26](#_Toc127015159)

[4.1 Studio sulla correlazione tra il numero totale di persone ospedalizzate ed il numero totale di persone in terapia intensiva 26](#_Toc127015160)

[4.2 Covarianza e Correlazione Campionaria 29](#_Toc127015161)

[4.3 Regressione Lineare Semplice 31](#_Toc127015162)

[4.4 Valori Stimati 33](#_Toc127015163)

[4.5 Residui 33](#_Toc127015164)

[4.6 Coefficiente di Determinazione 36](#_Toc127015165)

[4.7 Studio sulla Regressione Lineare Multipla del Medesimo Oggetto 37](#_Toc127015166)

[5. Analisi dei Cluster 41](#_Toc127015167)

[5.1 Distanza e Similarità 41](#_Toc127015168)

[5.2 Misure di Similarità 47](#_Toc127015169)

[5.3 Clustering tramite Metodologie Gerarchiche 48](#_Toc127015170)

[5.4 Analisi del Dendogramma 55](#_Toc127015171)

[5.5 Inserire Individui nel Cluster 56](#_Toc127015172)

[5.6 Misure di sintesi associate ai cluster 56](#_Toc127015173)

[5.7 Misure di Non Omogeneità 58](#_Toc127015174)

[5.8 Metodi non gerarchici 65](#_Toc127015175)

[5.9 Confronto tra Metodi Gerarchici e Metodi Non Gerarchici 68](#_Toc127015176)

**PARTE UNO**

Prefazione

Il presente documento descrive le fasi della attività di analisi statistica condotta sulla base di uno dei tanti dataset pubblicati dal Dipartimento della Protezione Civile. Il tema di quest’analisi consiste nel rapporto dettagliato di tutte le informazioni utili alla comprensione dello stato in cui si trova l’Italia in merito all’emergenza COVID-19 nel giorno 2 febbraio 2023.

Questo documento si suddivide in più capitoli, i quali hanno lo scopo di disquisire riguardo argomenti statistici strettamente legati fra loro. In questo modo si vuole fornire del contenuto nozionistico, riguardante esclusivamente la statistica e dunque non dedotto dallo studio statistico preso in esame, sia inferenziale; quest’ultimo contenuto riguarda le proposizioni statistiche che saranno derivate durante o studio dal dataset (tali informazioni saranno spesso anche descritte graficamente con l’uso di diagrammi diversi).

Il capitolo introduttivo è dedicato alle discipline statistiche ed alle loro suddivisioni. Successivamente inizia lo studio del dataset a partire da un’esposizione iniziale dei dati e da una descrizione delle caratteristiche che ne hanno decretato la scelta, con l’aggiunta di dettagli sui diversi criteri di suddivisione granulare selezionati.

**Introduzione alla statistica**

La statistica è una disciplina che studia in termini quantitativi e qualitativi un particolare fenomeno collettivo che non può essere studiato in termini deterministici. Questa incertezza può essere in sé oppure causata dalla mancanza di conoscenza (sufficiente) per poter fare

previsioni. Secondo il dizionario Palazzi, la statistica è “*la scienza dei fatti sociali espressi con termini numerici*”.

La statistica, avvalendosi della matematica, offre agli studiosi potenti strumenti di comprensione, sintesi e analisi dei fenomeni studiati, dando modo di rilevare eventuali correlazioni tra due o più fenomeni e misurare il grado di precisione con il quale dichiarare legami di causa-effetto. I metodi statistici permettono di analizzare informazioni numeriche - si parla di proprietà quantitative - oppure no - quindi proprietà qualitative.

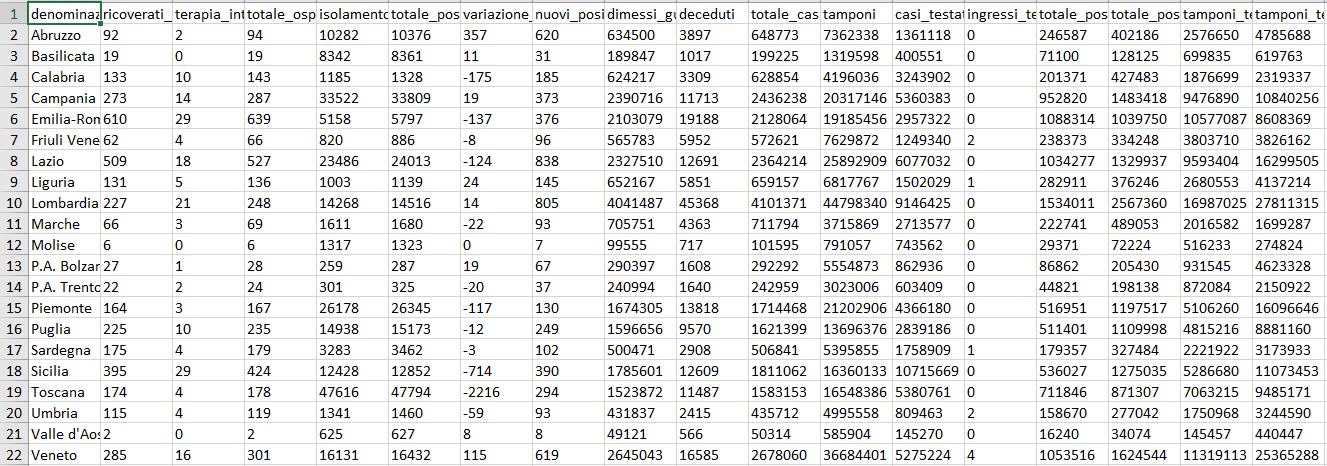
Un altro importante aspetto della statistica è il frequente uso di varie tipologie di diagrammi. Molti aspetti che saranno mostrati matematicamente verranno immediatamente confermati (se non chiariti) tramite rappresentazioni grafiche mediante diagrammi. Per generare i

grafici, così come per elaborare tutti i calcoli che costituiscono lo studio, viene usato il linguaggio di programmazione R.

**Il caso di studio**

Come anticipato, gli argomenti statistici trattati in questo documento saranno prima esposti in maniera “data-agnostic” (indipendenti dai dati in esame) e poi messi in pratica attraverso le informazioni presenti nel dataset in esame. In particolare, questi sono dati riguardanti la situazione dello stato italiano in merito all’emergenza covid in data 2 febbraio 2023. Queste attività sono suddivise geograficamente in regioni italiane, in modo da partire già con un raggruppamento iniziale. In particolare, le informazioni che ci interessano per ogni regione sono: “ricoverati con sintomi”, “terapia intensiva”, “totale ospedalizzati”, “isolamento domiciliare”, “totale positivi”, “variazione totale positivi”, “nuovi positivi”, “dimessi guariti”, “deceduti”, “totale casi”, “tamponi”, “casi testati”, “ingressi terapia intensiva”, “totale positivi test molecolare”, “totale positivi test antigenico rapido”, “tamponi test molecolare”, “tamponi test antigenico rapido”.

Ecco il dataset in questione.



In questo dataset quelle che vengono chiamate regioni in realtà tecnicamente sono delimitazioni del territorio italiano. Infatti, troviamo alcune province autonome.

A questo punto comincia l’analisi dei dati considerati partendo da una vista di alto livello sulla quantità di persone che sono risultate positive il 2 febbraio 2023.

# Grafici ed indici di posizione centrali

# 1.1 Variabili quantitative

Una variabile quantitativa è una collezione di valori numerici. Spesso, i valori di una variabile quantitativa si suddividono nelle classi ordinate in maniera crescente. Quindi si considerano  *= frequenza assoluta in*  e  *=* la frequenza relativa in . Siano inoltre le frequenze relative cumulate. I dati raccolti per le analisi statistiche saranno modellati attraverso variabili quantitative.

# Grafico a barre delle persone risultate positive

Come detto, la statistica offre strumenti grafici molto potenti al fine di analizzare le relazioni quantitative e qualitative che insistono nei dati. La prima rappresentazione grafica di cui ci si avvarrà è quella del grafico a barre.

Un *grafico a barre* è uno strumento di rappresentazione grafica che permette di mostrare le differenze tra i valori presenti in una variabile quantitativa, rappresentando ogni misura tramite una barra (spesso verticale, ma può essere anche orizzontale) posizionata in uno

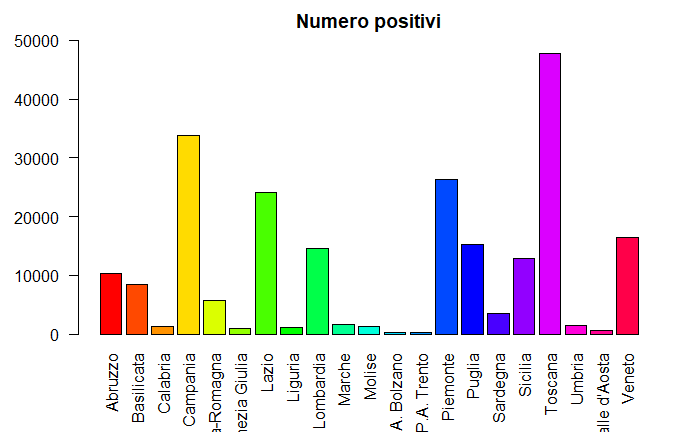
spazio in cui è presente un’asse verticale che scandisce il progressivo aumento di misura man mano che si sale verso l’alto.

Qui verranno mostrati e analizzati una serie di grafici a barre che analizzano i dati estrapolati dal dataset in esame. Il dataset in studio è stato memorizzato in R nel data frame ***dataset***.

Mediante la variabile ***dataset$denominazione\_regione*** è possibile accedere alla colonna del dataset relativa ai nomi delle regioni italiane.

Il seguente codice mostra un grafico a barre dei valori nella colonna relativa al numero di positivi.





Dal grafico risulta evidente che, escluse le provincie autonome, la Valle d’Aosta è la regione che ha presentato il minor numero di positivi, mentre in Toscana, al contrario, è stato rilevato il maggior numero di casi di persone positive al COVID-19.

# Indici di posizione centrali

Gli indici statistici di posizione che saranno descritti in questo documento si dividono in *centrali* e *non centrali*. Essi, in generale, possono fornire indicazioni approssimate circa le grandezze dei valori presenti nella variabile quantitativa considerata.

Dato un campione , si considerano la media campionaria come , la mediana campionaria come se è pari, altrimenti .

Infine, la moda campionaria corrisponde al valore (oppure modalità) più frequente. Se il campione è costituito da ripetizioni di un solo valore allora esso è priva di moda campionaria. In R, si calcola la media con il comando ***mean*** e la mediata con il comando ***median***.

Il seguente codice calcola alcune informazioni riguardo i valori degli indici di posizione centrali dei dati nella colonna relativa al numero di persone risultate positive il 2 febbraio 2023.

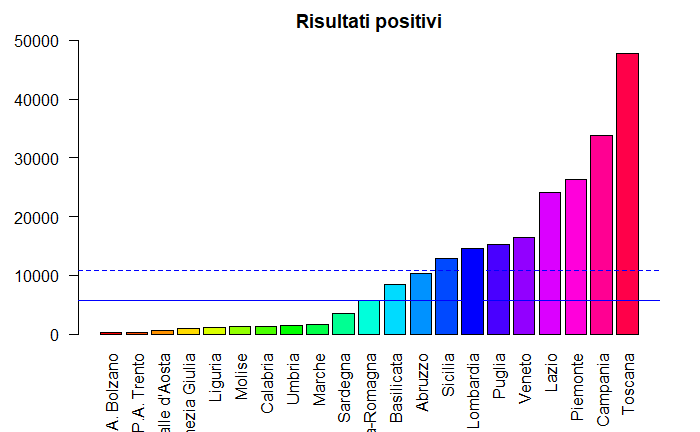
Immagine che contiene testo

Descrizione generata automaticamente

In particolare, il minimo risulta 287, il massimo 47794, la media 10856.43 e la mediana 5797.

# Grafico a barre ordinate

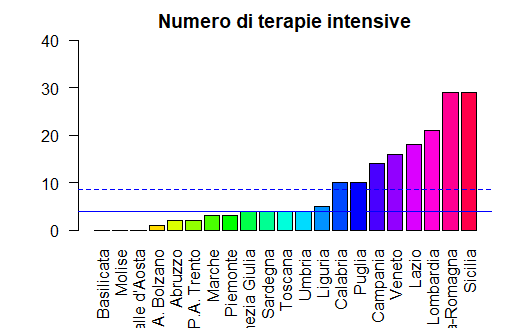
Possiamo infatti renderci conto della veridicità di questi valori dando uno sguardo a un grafico che mostra le stesse informazioni ma con un ordinamento sulle ascisse, grazie al seguente codice.



Risulta evidente il numero più basso è della provincia di Bolzano, il più alto invece è rappresentativo della Toscana. La linea tratteggiata indica la media campionaria di , mentre quella continua rappresenta la mediana.

# Grafico a barre sul numero di persone in terapia intensiva

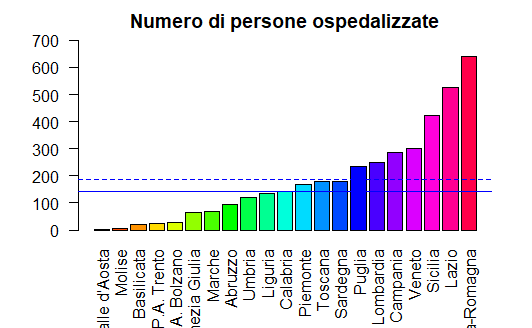
Si analizzeranno ora i dati delle altre colonne nel dataset, partendo da quella relativa al numero di persone in terapia intensiva.



Il grafico appena mostrato indica i numeri di persone, suddivisi per regione, che si trovano in terapia intensiva il 2 febbraio 2023. Mediante un codice R simile a quello visto per la generazione di informazioni riguardo indici di posizione centrali del precedente dataset, otteniamo che la media campionaria è pari a 8.5 mentre la mediana campionaria a 4.

# Grafico a barre sul numero totale di persone ospedalizzate

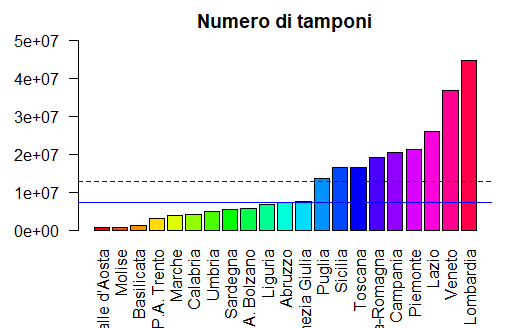
Si analizzeranno ora i dati delle altre colonne nel dataset, partendo da quella relativa al numero di persone ospedalizzate.



Tramite R è stato possibile calcolare che la mediana campionaria risulta equivalente a 639, mentre la mediana campionaria è pari a 143.

# Grafico a barre sul numero di tamponi

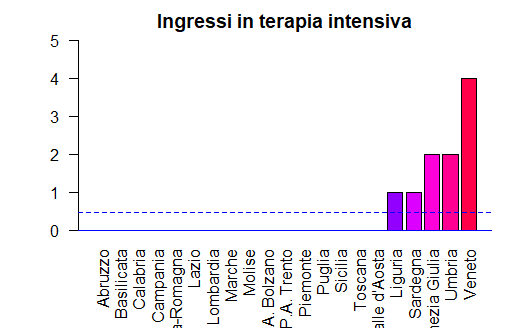
Si analizzeranno ora i dati delle altre colonne nel dataset, partendo da quella relativa al numero di tamponi effettuati.



Tramite l’analisi di tali dati, la media campionaria risulta pari a 12670180, mentre la mediana campionaria è di 7362338.

# Grafico a barre sul numero di ingressi in terapia intensiva

Si analizzeranno ora i dati delle altre colonne nel dataset, partendo da quella relativa al numero di persone che sono entrate in terapia intensiva.



Dall’analisi di questi dati ricaviamo che la media campionaria è pari a circa 0.5, mentre la mediana campionari è pari a 0.

# Frequenze e Distribuzioni

# Istogramma del numero di risultati positivi

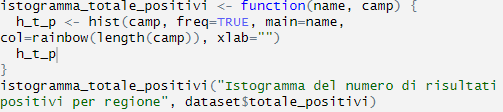
Molto utile è la suddivisione dei valori assunti da una variabile quantitativa in classi per studiarne la distribuzione di frequenze. In generale, le frequenze che non sono troppo dissimili fra loro posso essere accorpate in classi per poter studiare l’andamento di un argomento al fine di catturarne alcune informazioni quantitative. Una delle modalità più usate per l’analisi delle distribuzioni di frequenze consiste nella realizzazione di *istogrammi*.

Un istogramma è una rappresentazione grafica il cui compito è associare ad ogni classe di frequenze un rettangolo che ha per estremi orizzontali i valori estremi della classe, mentre per altezza ha la quantità di istanze nel campione che fanno parte di quella classe. Viene rappresentata un’asse orizzontale che percorre così tutti i valori assunti dal campione, avendo poi come asse verticale i numeri che costituiscono le frequenze (assolute o relative) della distribuzione.

Per studiare la distribuzione risulta determinante eseguire la suddivisione in classi. Critica per ottenere informazioni utili è dunque la scelta del numero di classi. Viene sconsigliata, a causa del probabile oscuramento di dettagli utili, la suddivisione in un numero esiguo di classi di una distribuzione di frequenze molto variata.

Mettendo in pratica gli argomenti appena esposti, adesso si analizzeranno delle informazioni estrapolate dai valori del dataset preso in esame. La prima caratteristica da evidenziare corrisponde alle relazioni dei dati relativi al numero di risultati positivi ai tamponi eseguiti il 2 febbraio 2023. I diversi valori regionali sono raggruppati come mostra l’istogramma di frequenze assolute che ci si accinge a mostrare.

Il seguente codice produrrà l’istogramma richiesto.



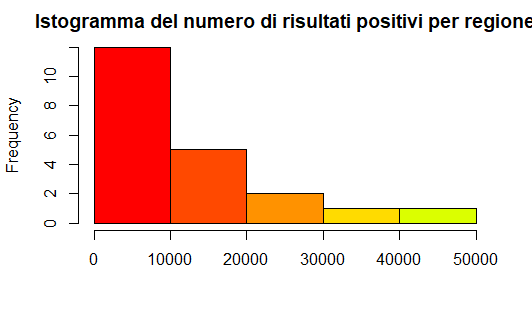
Di seguito è mostrato l’istogramma ottenuto.  


Immagine che contiene testo

Descrizione generata automaticamente

Osservando l’istogramma, appare piuttosto evidente che la gran parte delle regioni sono state raggruppate nelle prime due classi (0-10000 e 10000-20000).

# Indici di posizione non centrali

Tra gli indici di posizione troviamo gli indici di posizione non centrali. Questi si suddividono in quantili, quartili e percentili. Si consideri una variabile quantitativa che abbia dei valori orizzontalmente disposti in un ordine non decrescente, ossia dove il valore più a sinistra costituisce il minimo dei valori di questa. In funzione della variabile stessa, le distanze tra due valori adiacenti possono essere o non essere regolari e omogenee. Un altro strumento statistico per studiare le distribuzioni

di frequenze è quello di usare quantili, quartili e percentili. Un quantile è un valore che divide in due la suddetta sequenza di valori ordinati, in modo tale che a sinistra esistano soltanto valori minori e a destra soltanto valori maggiori. Tre quantili si chiamano quartili se dividono in quattro parti la sequenza ordinata, ovvero il 25% dei dati sono alla sinistra del primo quartile, il 50% alla sinistra del secondo quartile (che coincide con la mediana della variabile quantitativa), il 75% alla sinistra del terzo quartile.

Analogamente, novantanove percentili suddividono i dati ordinati in cento sotto sequenze. In R è possibile ottenere la suddivisione in quartili tramite la funzione ***quantile***.

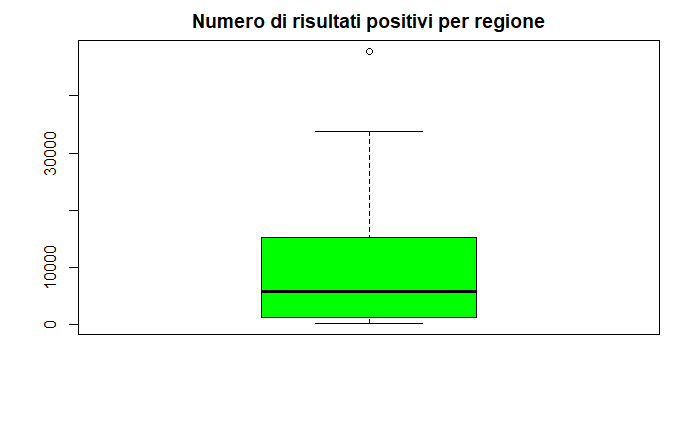
# Boxplot

Il boxplot, o diagramma a scatola e baffi, è una rappresentazione grafica che ha il compito di delimitare i quartili di una variabile quantitativa. In particolare, il primo quartile e il terzo quartile delimitano gli estremi di un rettangolo (la scatola), diviso in due da un segmento all’altezza del secondo quartile. Questa scatola è confinata da ambo i lati da due baffi che hanno con estremità dei valori assunti dalla variabile quantitativa che sono rispettivamente il valore minore e il valore maggiore in un dato intervallo di tolleranza. I valori della variabile quantitativa che oltrepassano questi limiti (in entrambe le direzioni) vengono chiamati valori anomali (anche detti outliner).

Il boxplot ad intaglio è invece una variante del boxplot di base che presenta un intaglio all’altezza della mediana la cui ampiezza esprime la quantità di valori della variabile quantitativa appartenenti all’intervallo di confidenza del 95% per la mediana.

# Boxplot del numero di risultati positivi suddivisi per regioni

Di seguito si mostrano il codice in R per la generazione del boxplot ed il boxplot stesso.

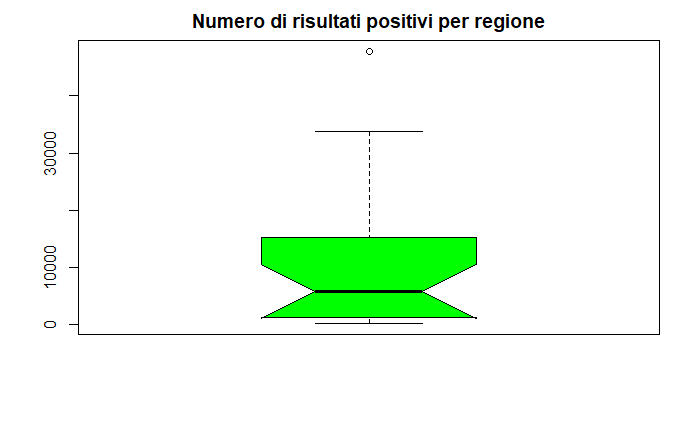


Di seguito si riportano dei dati analitici sul boxplot appena ottenuto.

Immagine che contiene testo

Descrizione generata automaticamente

Per evidenziare ulteriori informazioni mostriamo il seguente boxplot a intaglio (come detto nella sezione teorica, con intervallo di confidenza al 95% per la mediana).



Ciò che si nota (già dal primo boxplot) è che la distribuzione di frequenze studiata non è soggetta a valori anomali.

# Principio di Pareto

Il principio di Pareto è un principio che prevede che circa l’80% dei risultati derivino da circa il 20% delle cause. Questo è un principio empirico che si verifica spesso in sistemi complessi basati su strutture causa-effetto. Può essere ulteriormente sintetizzato nel seguente modo: “la maggior parte degli effetti è dovuta a un numero ristretto di cause”. Esso è poi ulteriormente supportato dalla distribuzione paretiana.

Altri esempi possono essere:

● il 20% dei venditori esegue l’80% delle vendite;

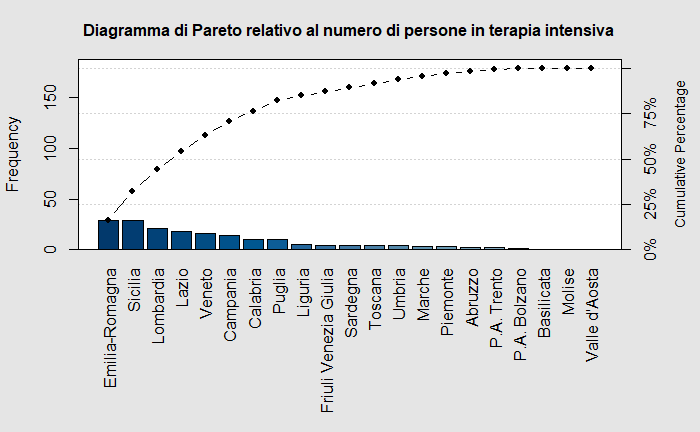
● il 20% delle istruzioni di un programma viene eseguito l’80% del tempo (a supporto del principio di località temporale);

● l’80% dei ricavi di treni e aerei deriva dal 80% di rotte;

● l’80% del successo di una persona deriva dal 20% delle sue imprese.

Prende il nome da Vilfredo Pareto che studiando la distribuzione dei redditi, nel 1897 evinse che solo pochi individui possedevano la maggior parte delle ricchezze in una data regione.

Ecco, infine il diagramma di Pareto relativo al numero di terapie intensive.



# Dispersione e Funzione di Distribuzione

# Varianza

Il primo indice di dispersione statistico che si evidenzia immediatamente è la varianza campionaria. Essa fornisce una misura di dispersione statistica, in particolare della variabilità dei valori assunti da una variabile statistica o aleatoria . Dato un campione , la varianza è pari a:

Come è facile notare, l’unità di misura della varianza è il quadrato dell’unità di misura dei dati del relativo campione.

Nell’ambiente R è possibile calcolare la varianza di un *dataframe df* usando direttamente il comando ***var(df)***.

# Deviazione Standard

Un indicatore statistico strettamente legato alla varianza è la deviazione standard. Essa è un indice di dispersione statistico che descrive quanto i dati di un certo campione si discostano dalla *media campionaria* (indice di posizione centrale che verrà descritto nel prossimo capitolo). Dato un campione di dati , la deviazione standard è pari a , ovvero la radice quadrata del campione . Si può facilmente notare che l’unità di misura della deviazione standard è la stessa di quella dei dati del campione di partenza.

Nell’ambiente R è possibile calcolare la deviazione standard di un *dataframe df* usando direttamente il comando ***sd(df)***.

# Coefficiente di variazione

Infine, il coefficiente di variazione è un altro indicatore statistico che descrive il valore di scostamento dei dati in un campione rispetto alla sua media campionaria. Esso si calcola con la seguente formula:

Questo valore è il primo indicatore statistico adimensionale esposto in questo documento. Esso è infatti il rapporto tra la deviazione standard e la media campionaria (in valore assoluto) di un campione dato, ossia un rapporto tra valori della stessa unità di misura, che quindi si elide. Questo indice di dispersione può essere considerato soltanto per campioni aventi media campionaria non nulla. Nell’ambiente R è possibile calcolare il coefficiente di variazione di un *dataframe df* usando direttamente il comando ***cv(df)***.

# Analisi degli indici di dispersione sulle variabili quantitative

Qui verranno mostrati e analizzati i dati estrapolati dal dataset per tutti i dati numerici relativo allo stato in cui si trova l’Italia relativamente al problema COVID-19.

I dati relativi al numero di nuovi risultati positivi sono caratterizzati da una varianza pari a 166665183, una deviazione standard pari a 12909.89 ed un coefficiente di variazione pari a 118.92.

I dati relativi al numero di terapie intensive sono caratterizzati da una varianza pari a 84.4619, una deviazione standard pari a 9.19 ed un coefficiente di variazione pari a 107.81.

I dati relativi al numero di persone ospedalizzate sono caratterizzati da una varianza pari a 29950.61, una deviazione standard pari a 173.06 ed un coefficiente di variazione pari a 93.40.

I dati relativi al numero di tamponi eseguiti sono caratterizzati da una varianza pari a 1.449247e+14, una deviazione standard pari a 12038466 ed un coefficiente di variazione pari a 95.01.

I dati relativi al numero di ingressi in terapia intensiva sono caratterizzati da una varianza pari a 1.06, una deviazione standard pari a 1.03 ed un coefficiente di variazione pari a 216.40.

Come si può notare dai risultati, il coefficiente di variazione è l’unico dei tre che non varia così tanto rispetto agli altri due. Infatti, questo è l’unico numero puro tra i tre e può essere considerato un indice di informazioni assolute non dipendente dalle grandezze complessive dei dati né ovviamente dall’unità di misura.

Si evidenzia inoltre che il coefficiente di variazione è più basso nel campione relativo al numero di persone ospedalizzate, ossia quello il cui

grafico a barre risulta contenere le altezze più simili (vedi capitolo 1).

# Funzione di Distribuzione Empirica

In probabilità spesso le informazioni che vengono studiate insistono nelle diverse variabili aleatorie - discrete o continue - di cui si considera la funzione di distribuzione. Nel caso dell’analisi dei dati, pur non avendo una variabile aleatoria si ha comunque la necessità di comporre in qualche modo una funzione di distribuzione.

Ecco che si crea una funzione di distribuzione empirica continua o discreta, in base al tipo di dati numerici (discreta se per esempio si considera il numero di studenti laureati con la lode nelle varie facoltà, continua per esempio la media dei voti di laurea di un gruppo di studenti universitari).

# Funzione di Distribuzione Empirica Discreta

Quando si raggruppano le frequenze in classi, e non si considerano valori intermedi, allora una funzione di distribuzione empirica discreta può essere di aiuto. Questa è così definita:

Si verifica che .

La funzione di distribuzione empirica discreta gode delle seguenti proprietà:

* è una funzione a gradini
* è una funzione non decrescente
* e

Nell’ambiente R, il grafico di una funzione di distribuzione empirica discreta può essere mostrato usando la seguente funzione:

Immagine che contiene testo

Descrizione generata automaticamente

# Funzione di Distribuzione Empirica Continua

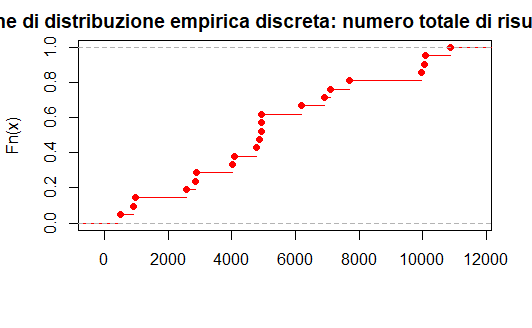
Quando bisogna classificare i dati di una variabile quantitativa in classi , modellare in qualche modo tutti i valori intermedi che insistono negli estremi di una classe, allora è bene considerare una funzione di distribuzione empirica continua. Questa è costruita come segue:

Si verifica che .

# Funzione di Distribuzione Empirica Discreta del Numero di Risultati Positivi

Si analizza adesso la funzione di distribuzione empirica discreta ottenuta mediante i dati della colonna nel dataset corrispondente al numero di nuovi risultati positivi (ai tamponi) per regione. In seguito, è mostrato il grafico della suddetta funzione.

Il seguente codice produce il prossimo grafico nel documento.



In seguito, saranno illustrati metodi più analitici per ricavare informazioni riguardo caratteristiche grafiche di una distribuzione di frequenze.

# Skewness e curtosi campionarie

La statistica offre anche strumenti più analitici per caratterizzare la simmetria e la piccatezza di una distribuzione di frequenze. La skewness campionaria è un indice di simmetria della distribuzione di frequenze, e si calcola tramite i momenti centrati campionari di ordine due e tre. La curtosi campionaria, invece, indica la piccatezza di una distribuzione di frequenze e viene calcolata tramite l’indice di Pearson (che si basa sui momenti centrali campionari di ordine due e quattro).

La skewness campionaria e la curtosi campionaria del campione corrispondente alla colonna del numero totale di risultati posotivi ai tamponi nel dataset possono essere calcolate in R tramite il seguente programma.

Immagine che contiene testo

Descrizione generata automaticamente

In particolare, la skewness campionaria della distribuzione di frequenze appena ottenuta è pari a 0.271 poiché infatti la distribuzione presenta una asimmetria negativa, ossia ha una coda sinistra più lunga come tendenza

La curtosi campionaria della distribuzione di frequenze appena ottenuta è pari -0.858, poiché infatti la distribuzione presenta una platicurtosi, ossia la distribuzione è eccessivamente appiattita (con code troppo corte).

# Regressioni Lineari

In questo nuovo capitolo si presenteranno vari argomenti riguardanti la statistica descrittiva multivariata. In particolare, ci si soffermerà sugli strumenti statistici per la rilevazione di relazioni tra due variabili, come la covarianza, la correlazione e le regressioni. Nell’applicare questi strumenti verrà ristretto il campo alla statistica descrittiva bivariata.

La statistica descrittiva bivariata è un ramo della statistica che si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili.

Le modalità di rappresentazione grafica consistono nella composizione di diagrammi di dispersione (anche detti scatter plot) composti da un piano euclideo e da una serie di punti nel piano. Ogni punto è associato a una coppia di osservazioni appartenenti alle due variabili di cui si ha interesse a mostrare il grado di relazione reciproco.

# Studio sulla correlazione tra il numero totale di persone ospedalizzate ed il numero totale di persone in terapia intensiva

La statistica descrittiva bivariata fornisce strumenti matematici che permettono di descrivere le relazioni che intercorrono tra due variabili: nella sezione di questo documento si mostreranno gli esiti degli studi effettuati sulla esistenza di relazioni tra il numero totale di persone ospedalizzate ed il numero totale di persone in terapia intensiva. Si descriveranno prima graficamente le correlazioni tra i vari campioni del dataset e si tireranno le somme riguardo a correlazioni più marcate rispetto ad altre.

In generale, le relazioni tra due variabili possono avere diversa natura, nel nostro caso ci focalizzeremo su quelle di tipo lineari.

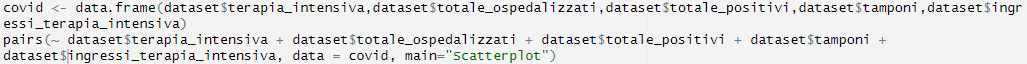
Se tra due variabili intercorre una relazione lineare, questa può essere:

● positiva, se l’aumento di valore di una variabile implica l’aumento di valore dell’altra;

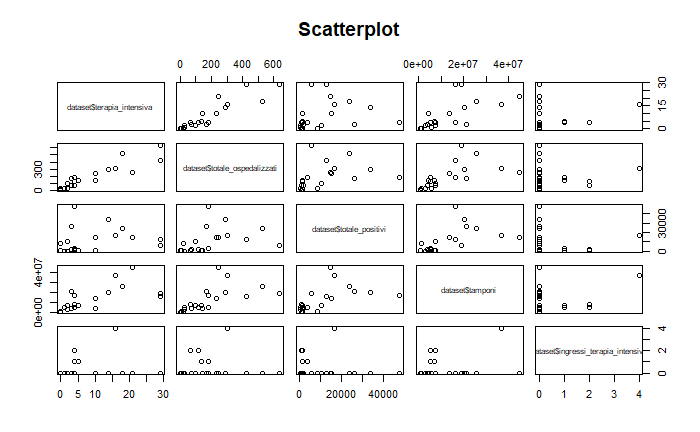
● negativa, se l’aumento di valore di una variabile implica la diminuzione di valore dell’altra.

Questo tipo di relazione può essere immediatamente riconosciuto grazie a uno scatter plot delle due variabili. Se la nuvola di punti mostrata è poco dispersa e si raggruppa attorno a una retta crescente, allora esiste una relazione lineare positiva. Se si raggruppa attorno a una retta decrescente allora esiste una relazione lineare negativa. Il grado di intensità della relazione dipende anche da quanto la nuvola di punti è fitta intorno alla retta. Chiameremo d’ora in poi retta di regressione la retta attorno alla quale si attornia la nuvola di punti.

Di seguito si ha il codice utilizzato per generare la matrice di scatter plot in R:



Ecco invece la matrice dei diagrammi di dispersione:



In questa immagine è evidente la presenza di alcune correlazioni abbastanza marcate tra variabile quantitative. Secondo lo scopo del presente documento, si è deciso di focalizzarsi sullo studio delle relazioni tra il numero totale di persone ospedalizzate ed il numero totale di persone in terapia intensiva.

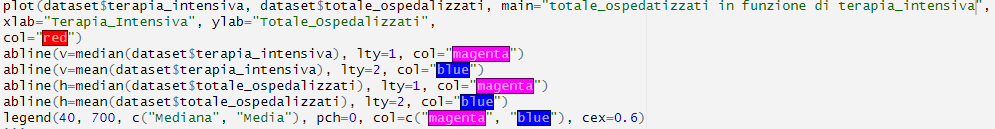
Prima di procedere allo studio vero e proprio, si inseriscono di seguito i dati numerici presi dal dataset.

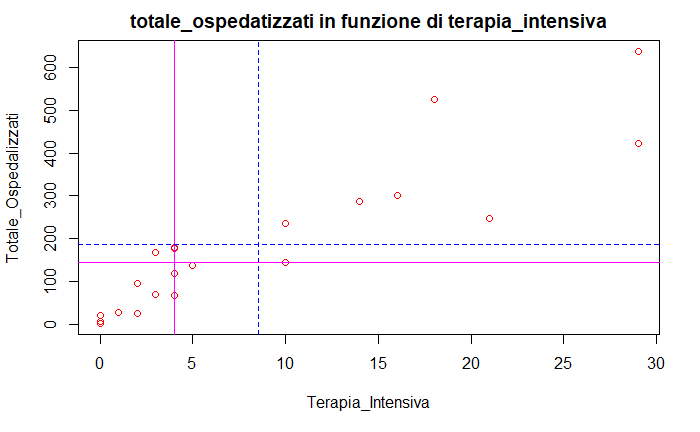
Immagine che contiene tavolo

Descrizione generata automaticamente

Si ribadisce che la media del campione “terapia\_intensiva” è 8.5 e la mediana è 4. La media del campione “totale\_ospedalizzati” è 639 e la mediana 143.

Di seguito viene mostrato un plot in cui vengono relazionati i dati tra il numero totale di persone ospedalizzate ed il numero totale di persone in terapia intensiva, contenente anche media e mediana dei dati. Questo viene generato dal seguente codice R:





Si può notare che i dati sembrano essere posizionati intorno ad una retta scendente e ciò induce a pensare che esista una correlazione lineare positiva tra le variabili.

# Covarianza e Correlazione Campionaria

Per essere certi della reale presenza di una relazione tra le variabili 𝑋 e 𝑌, e per ottenere con precisione informazioni dettagliate rispetto la natura (positiva o negativa) della relazione, si è previsto il calcolo di covarianza e coefficiente di correlazione.

La covarianza è un numero che misura il grado di dipendenza tra la variabile indipendente 𝑋 e la variabile dipendente 𝑌.

Si possono verificare i seguenti casi:

● covarianza positiva, ossia la variabile dipendente è correlata positivamente alla variabile indipendente;

● covarianza negativa, ossia la variabile dipendente è correlata negativamente alla variabile indipendente;

● covarianza nulla, ossia le variabili non sono correlate.

Il coefficiente di correlazione permette di ottenere ancora più informazioni rispetto la dipendenza di una variabile dipendente da una variabile 𝑌 indipendente 𝑋.

Il coefficiente di correlazione assume un valore appartenente all’intervallo [− 1, 1]. Si possono verificare i seguenti casi:

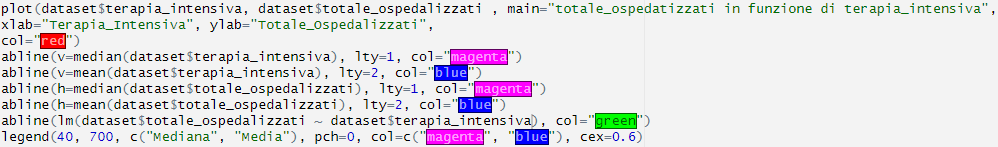
* ossia la nuvola di punti giace esclusivamente su una retta ascendente 𝑥𝑦 = 1 (correlazione positiva perfetta);
* , ossia la nuvola di punti è addensata a una retta interpolante ascendente 𝑥𝑦 ∈ (0, 1) (correlazione positiva);
* ossia non esiste correlazione tra le variabili considerate;
* , ovvero la nuvola di punti è addensata attorno una retta interpolante discendente (correlazione negativa);
* ossia la nuvola di punti giace esclusivamente su una retta discendente (correlazione negativa perfetta).

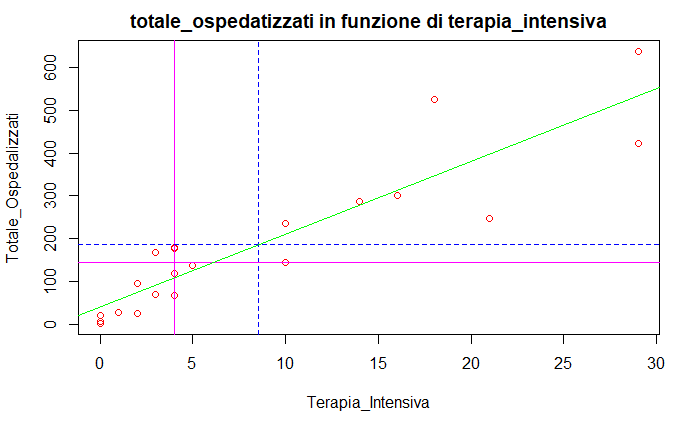
La covarianza e la correlazione campionaria possono essere calcolate in R rispettivamente tramite le formule **cov(x, y)** e **cor(x, y)**.Calcoliamo la covarianza e la correlazione tramite il seguente codice in R:



Osserviamo dunque che la covarianza assume un valore molto alto, dunque il numero totale di ospedalizzati ha una forte correlazione positiva con il numero totale di terapie intensive. A testimonianza di ciò, si nota come infatti il coefficiente di correlazione abbia quindi un valore molto vicino all’1.

Il passo successivo consiste nel realizzare il grafico contenente la retta di regressione che interpola i punti mediante il seguente codice in R:





Si nota subito che il punto della media si trova sulla retta di regressione discendente: ciò sta a significare che la retta rappresenta al meglio il legame tra 𝑋 e 𝑌 a testimonianza di ciò, difatti, si ha il coefficiente di correlazione campionario uguale a 0. 90, molto vicino a 1: ciò indica una correlazione negativa molto forte.

# Regressione Lineare Semplice

Dopo aver calcolato covarianza e correlazione, si passa al calcolo di α e β per scrivere l’equazione della retta:

Dove:

* Abbiamo α che è l’*intercetta*: ;
* Mente β è il *coefficiente angolare* della retta: ;
* è la variabile dipendente;
* 𝑋 è la variabile indipendente.

Tramite il seguente codice in R è possibile calcolare l’intercetta ed il coefficiente angolare:

Immagine che contiene testo

Descrizione generata automaticamente

Output:



I coefficienti possono essere alternativamente calcolati in R tramite la funzione lm(y~x). Il nome lm della funzione rappresenta l’acronimo di linear model. L’argomento y˜x passato alla funzione lm() indica che y dipende da x, ossia che x è la variabile indipendente e y la variabile dipendente.

Quindi, alpha e beta possono essere calcolati in maniera più rapida utilizzando quest’altro codice in R:



Output:

Immagine che contiene testo

Descrizione generata automaticamente

In conclusione, la retta di regressione avrà come equazione:

Y = 40.02 +17.04 X.

# Valori Stimati

Una volta calcolati i valori dei coefficienti α e β e disegnata la retta di regressione che interpola la nuvola dei punti nel corrispondente scatter plot, è possibile osservare quanto questa retta si adatta ai punti che individuano le osservazioni.

In generale, esisteranno degli scostamenti (detti residui) tra le ordinate dei punti (ovvero i valori osservati) e i corrispondenti valori stimati nella retta di regressione, dove:

Data la retta di regressione lineare dei punti costituiti da osservazioni di 𝑋 e 𝑌, si verifica che la media campionaria dei valori osservati coincide con la media campionaria dei valori stimati ella suddetta retta di regressione; possiamo verificare quanto detto dal seguente codice in R:



Output:

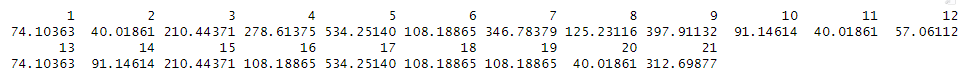


Come si vede, la media è esattamente uguale.

In R, per il calcolo dei valori stimati si utilizzano le seguenti righe di codice:



Output



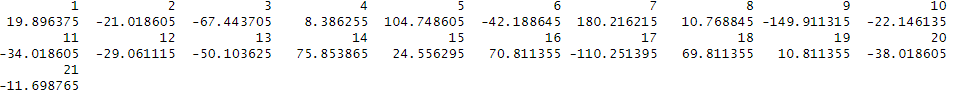
# Residui

I valori residui mostrano gli scostamenti dei valori osservati dai valori stimati nella retta di regressione. Questi si indicano nel seguente modo:

Data la retta di regressione lineare dei punti costituiti da osservazioni di 𝑋 e 𝑌, si verifica che la media dei residui è nulla. Questo perché gli scostamenti si compensano tra loro. Per calcolare i residui in R si usa la funzione ***resid(lm(y~x))***:



Output:

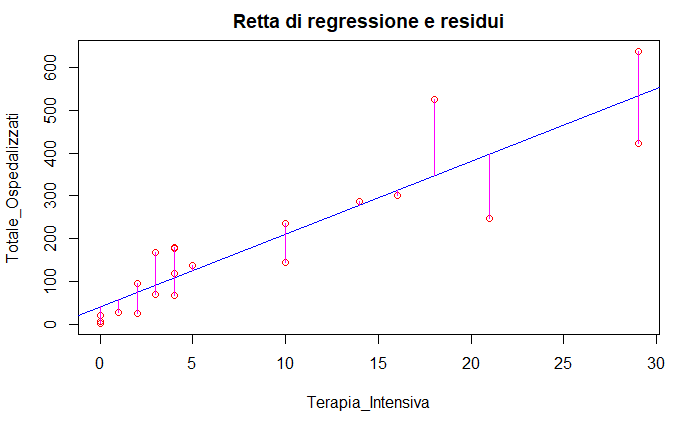


I residui possono anche essere rappresentati graficamente:

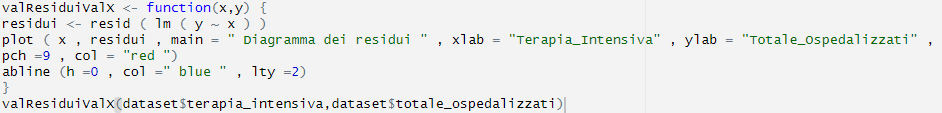
* Tracciando dei segmenti verticali che congiungono i valori stimati (sulla retta di regressione) e i valori osservati , usando:

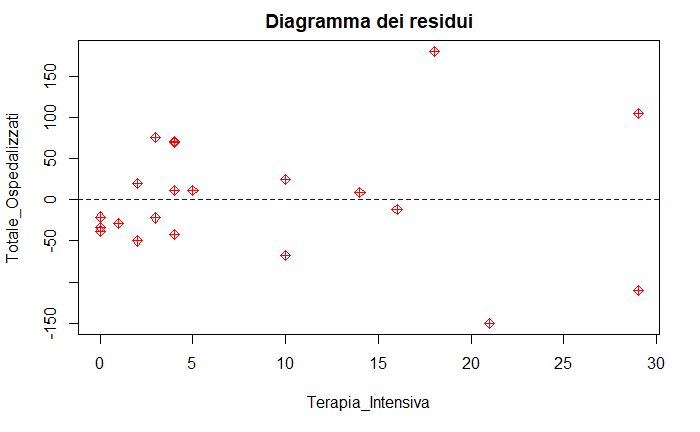
Immagine che contiene testo

Descrizione generata automaticamente



* Rappresentando i valori dei residui rispetto alle osservazioni , usando

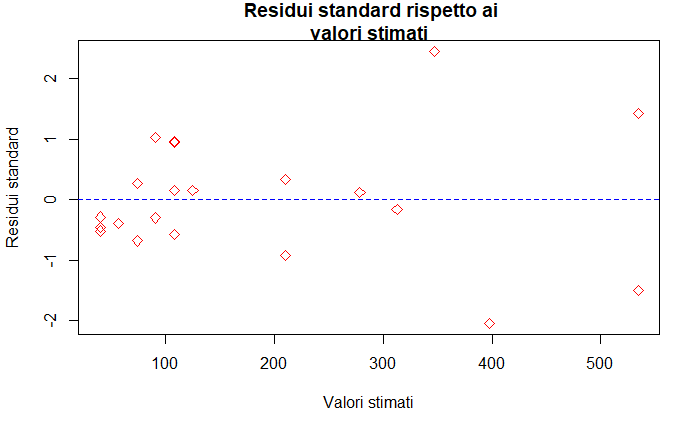




* Rappresentando i residui standardizzati rispetto ai valori stimati, usando

Immagine che contiene testo

Descrizione generata automaticamente



# Coefficiente di Determinazione

Poiché si è interessati a vedere quanto la retta si adatta ai dati, si stabilisce il coefficiente di determinazione.

Il coefficiente di determinazione è un indicatore di adattamento della retta di regressione ai dati, ed è il rapporto tra la varianza dei valori stimati nella retta di regressione e la varianza dei valori osservati. Questo coefficiente appartiene sempre all’intervallo [0, 1]: quanto più è maggiore tanto più i dati stimati sono rappresentativi dei dati osservati.

Il coefficiente di determinazione può essere calcolato in due modi:

Immagine che contiene testo

Descrizione generata automaticamente

Output:



Con le verifiche fatte finora, quindi, possiamo affermare che è presente una forte dipendenza tra il numero totale di persone ospedalizzate ed il numero totale di persone in terapia intensiva.

# Studio sulla Regressione Lineare Multipla del Medesimo Oggetto

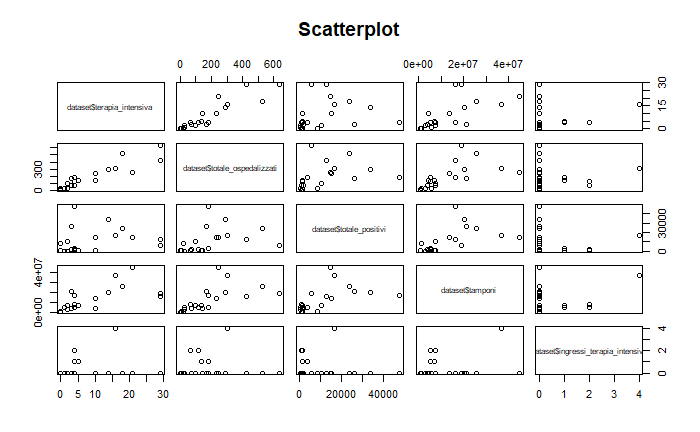
Nella sezione precedente si è evidenziata la regressione lineare semplice, mentre nella corrente si andrà ad approfondire la regressione lineare multipla.

Questo nuovo modello viene utilizzato per spiegare la relazione tra una variabile quantitativa dipendente e una serie di variabili quantitative 𝑌 indipendenti .

Una regressione lineare multipla si esprime nel seguente modo:

Dove è l’intercetta, mentre le sono i regressori.

Innanzitutto ci generiamo tramite la funzione **pairs()** di R, un’unica finestra grafica contenente scatter plot multipli:



Successivamente viene calcolata la covarianza e correlazione di un data frame (contenente i dati del grafico soprastante) contenente n osservazioni:

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che la matrice delle covarianze contiene sulla diagonale principale la varianza delle singole colonne del data frame.

Immagine che contiene testo

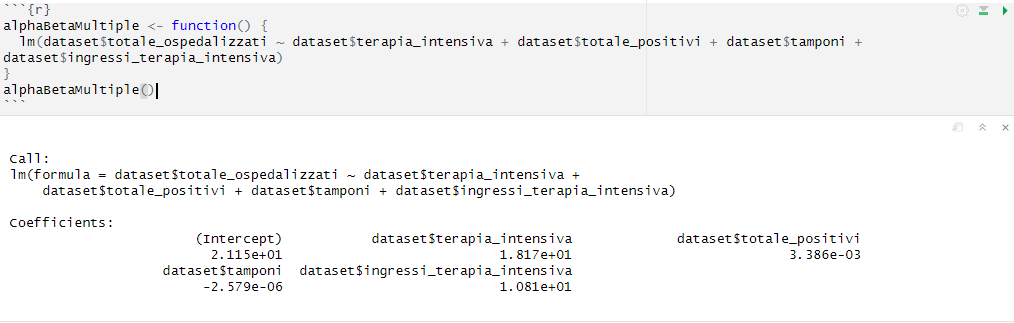
Descrizione generata automaticamente

Riguardo la matrice delle correlazioni, invece, essa contiene il numero 1 sulla diagonale principale. Inoltre, possiamo notare che esiste una forte correlazione negativa tra i ricoverati in terapia intensiva e le persone ospedalizzate.

Nel nostro caso, utilizzeremo gli ospedalizzati totali come dipendente, mentre tutte le altre saranno variabili indipendenti.

Ricordiamo che Il modello di regressione lineare multipla con p variabili indipendenti è esprimibile attraverso l’equazione:

Ricaviamo, tramite il seguente codice, l’intercetta ed i regressori:



Pertanto, il modello di regressione multipla stimato è:

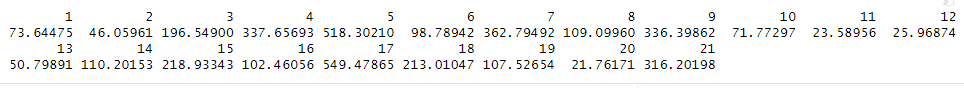
Riguardo i residui, invece, è necessario ricavare, come prima, i valori stimati ottenuti mediante la regressione lineare multipla:

Successivamente, possono essere calcolati i residui:

Tramite la seguente funzione in R è possibile calcolare i valori stimati:



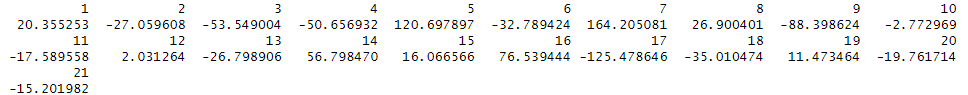
Ecco il risultato:



Successivamente, possono essere calcolati i residui:



Ecco l’output:



Si conclude osservando il valore del coefficiente di determinazione:



Output:



Esso è pari a 0.8567136.

# Analisi dei Cluster

L’analisi dei cluster è una metodologia che permette di raggruppare in sottoinsiemi (detti cluster) entità appartenenti a un insieme più ampio. L’insieme originario delle entità su cui si svolge l’analisi per ricavare i cluster non è vincolato ad alcuna restrizione; infatti, esso può contenere variabili, individui, osservazioni, dati, misure, ecc... I vari metodi attraverso cui si attua l’analisi dei cluster hanno in comune uno scopo generale: ottenere raggruppamenti in base alla somiglianza in modo che gli elementi di uno stesso gruppo siano tra loro il più possibile simili e gli elementi appartenenti a gruppi distinti siano tra loro il più possibile diversi.

Ciò significa che questa analisi ha lo scopo di suddividere in gruppi le osservazioni in modo da garantire due proprietà:

● alta omogeneità all’interno dei gruppi, ossia bisogna che sia presente un alto grado di associazione tra membri dello stesso gruppo;

● alta eterogeneità tra gruppi diversi, ossia avere un basso grado di associazione tra membri di gruppi differenti.

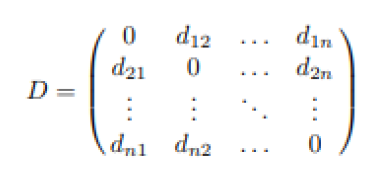
L’analisi dei cluster si può applicare in tutti i casi in cui è richiesto riunire in maniera non intuitiva, con il sostegno di metodi quantitativi, dati di qualsiasi natura. Lo scopo è quello di organizzare in una struttura un insieme di dati e di scoprire quali siano i legami esistenti tra essi.

# Distanza e Similarità

Le misure metriche di somiglianza sono soprattutto basate sulle funzioni distanza tra i vettori delle caratteristiche. Una funzione a valori reali , è detta funzione distanza se e soltanto se essa soddisfa le seguenti condizioni:

* , se e solo se , con (Questa proprietà implica che è a distanza zero da sé stesso e che ogni due punti a distanza nulla devono essere identici);
* (La seconda proprietà afferma che la funzione distanza è non negativa);
* (La terza proprietà impone la simmetria richiedendo che la distanza tra e deve essere la stessa della distanza tra e );
* (L’ultima proprietà, nota come disuguaglianza triangolare, richiede che la distanza tra e debba essere sempre minore o uguale della somma delle distanze di ognuno dei due vettori considerati da qualunque altro terzo vettore.

Generalmente, tutte le distanze vengono inserite in una matrice simmetrica D:



Dove .

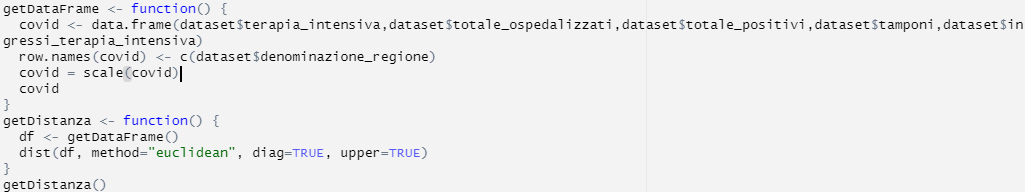
La matrice delle distanze D può essere calcolata in R tramite la funzione

**dist(X, method=”euclidean”, diag=FALSE, upper=FALSE)**

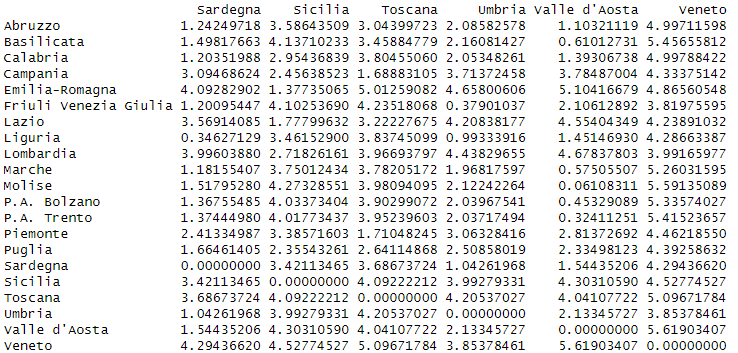
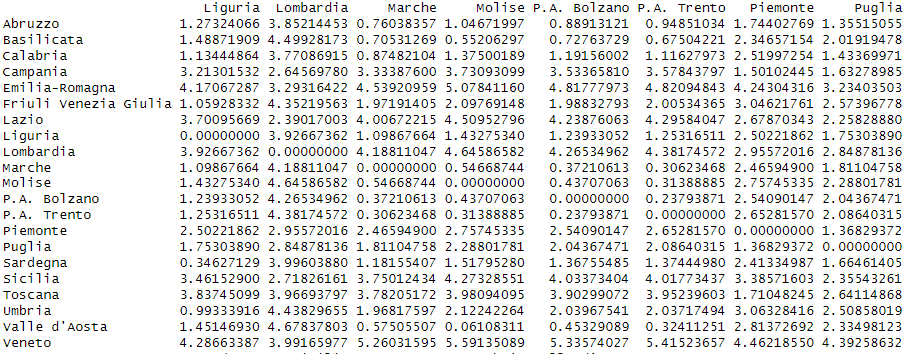
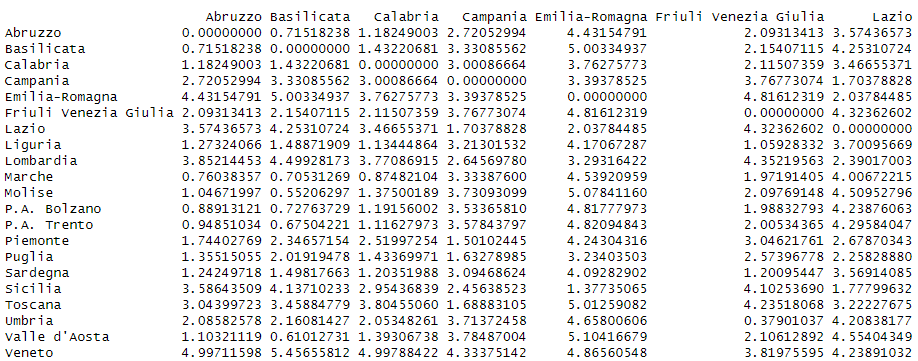
dove:

* rappresenta una matrice numerica o un data frame;
* **method** seleziona la misura di distanza da utilizzare (che di default è già euclidean);
* **diag** è posta uguale a TRUE se si desidera che la matrice delle distanze contenga anche i valori nulli sulla diagonale (di default è FALSE);
* **upper** è posta uguale a TRUE se si desidera che la matrice delle distanze contenga anche i valori al di sopra della diagonale principale (di default `e FALSE).

Tale matrice in R può essere ottenuta eseguendo il seguente codice:



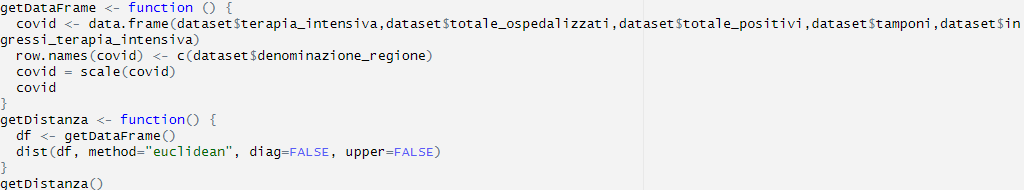
Output:

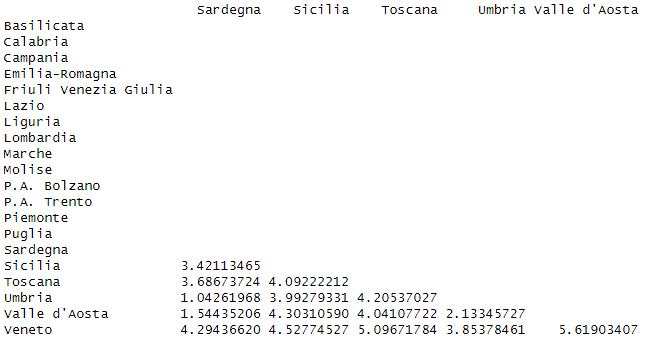
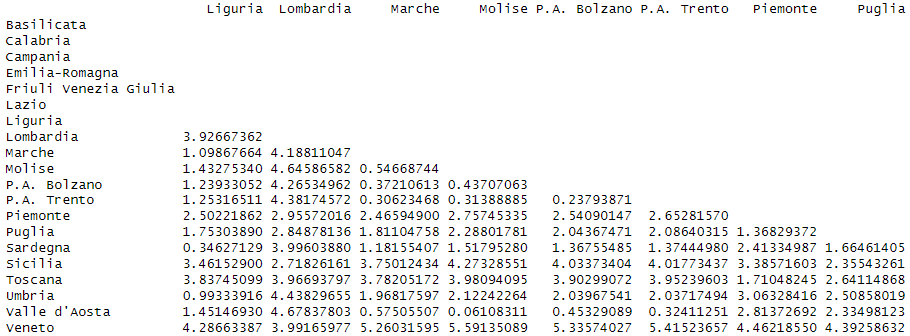
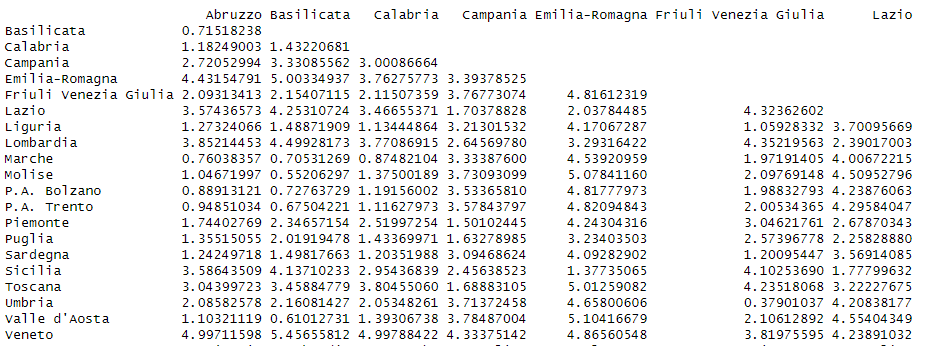


Il cluster con le distanze è stato calcolato applicando il metodo euclideo.

Come possiamo notare, i valori sulla diagonale sono tutti nulli, ovvero pari a zero, mentre quelli al di sotto e quelli al di sopra risultano essere equivalenti. Quindi, il numero di distanze che è necessario conoscere affinché sia definita la posizione di ciascuna delle n variabili rispetto alle rimanenti n − 1 è, dunque, n(n − 1)/2. Ciò significa che è sufficiente considerare la matrice triangolare al di sopra o al di sotto della diagonale principale di D.

Infatti, potremmo ottimizzare ulteriormente l’output precedente applicando questa modifica al codice precedente:



Output:

Questo risultato è stato ottenuto modificando:

* diag, è stato impostato su FALSE, in questo modo non appaiono i valori nulli sulla diagonale;
* upper, è stato impostato su FALSE, in questo modo, non compaiono i valori al di sopra della diagonale, perché risulterebbero duplicati dato che già sono presenti al di sotto della diagonale.

Quindi, come possiamo notare, adesso compaiono solamente i valori non nulli e non duplicati.

Uno dei problemi che si presenta nell’analisi dei cluster riguarda la standardizzazione o meno delle variabili, poiché attribuire un peso diverso a ciascuna caratteristica potrebbe condurre a risultati differenti a seconda delle tecniche di clustering utilizzate. In molti metodi di clustering si raccomanda la standardizzazione di ogni variabile (caratteristica) usando la media campionaria e la deviazione standard campionaria entrambe derivate dall’insieme completo di individui della popolazione.

Ciò è stato eseguito grazie all’uso della funzione **scale(covid)**, dove **covid** identifica il dataset in esame.

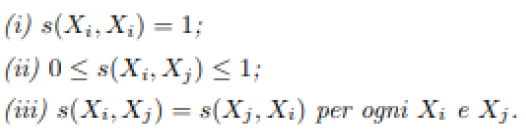
# Misure di Similarità

Nella maggior parte delle tecniche di clustering occorre inizialmente calcolare la matrice D delle distanze oppure una matrice S delle similarità. Una misura di similarità permette di definire in modo quantitativo la somiglianza o differenza tra due individui e .

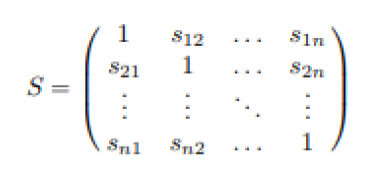
Viene fornito come risultato un valore numerico compreso tra 0 e 1, dove:

* Se 0, si intende l’assoluta assenza di similarità;
* Se 1, si intende l’assoluta presenza di somiglianza.

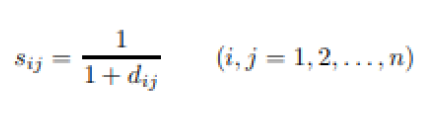
Per definizione una funzione a valori reali è detta misura di similarità se e soltanto se essa soddisfa le seguenti condizioni



La proprietà (i) implica che la misura di similarità è unitaria se i due punti sono identici. La proprietà (ii) afferma che la misura di similarità è compresa tra 0 e 1. La proprietà (iii) impone la simmetria richiedendo che la misura di similarità tra e deve essere la stessa della misura di similarità tra e . La quantità è chiamata semplicemente coefficiente di similarità e risulta essere l’elemento nella riga i-esima e colonna j-esima della matrice di similarità S, così definita:



La più ovvia differenza tra una misura di similarità e una misura di distanza è che la prima misura assume valori tra 0 e 1 mentre la seconda misura assume valori non negativi. È importante comunque sottolineare che è sempre possibile trasformare una misura di distanza in una misura di similarità. Supponiamo che sia la distanza tra e , allora mediante la seguente formula



si può effettuare la trasformazione e quindi ottenere la misura di similarità tra e .

Non sempre, però, è possibile effettuare il contrario, ovvero passare da una misura di similarità ad una distanza, in quanto quella di distanza deve soddisfare anche la disuguaglianza triangolare.

# Clustering tramite Metodologie Gerarchiche

Gli algoritmi di clustering che trattano la divisione gerarchica si suddividono in due categorie: gli algoritmi di tipo agglomerativi partono dalle foglie del *dendrogramma* per poi arrivare alla radice, quindi seguono in bottom-up; gli algoritmi di tipo divisivo partono dalla radice per poi arrivare alle foglie, lavorando quindi in maniera opposta rispetto agli agglomerativi, quindi in top-down.

L’obiettivo finale dei metodi gerarchici non è quello di ottenere una singola partizione degli n individui di partenza, ma di ottenere una sequenza di partizioni che possono essere rappresentate graficamente mediante una struttura ad albero, detta dendrogramma.

Nel dendrogramma, sulle ascisse sono riportati i livelli di distanza, mentre sulle ordinate sono riportati i singoli individui. Molti metodi di analisi gerarchica hanno un algoritmo comune che è formato da cinque step:

1. Si considera la matrice delle distanze D o la matrice di similarità S tra gli individui considerati come singoli cluster contenenti un solo individuo;

2. Si individua la coppia di cluster meno distanti (o più somiglianti) e si raggruppano in un unico cluster;

3. Si costruisce una nuova matrice di distanza (o di similarità) che risulterà ridotta di una riga e di una colonna rispetto a quella che la precede;

4. Effettuare n-1 iterazioni a partire dallo step 2, in modo da arrivare alla fine ad avere una matrice di cardinalità 2x2;

5. Rappresentare graficamente il processo di agglomerazione attraverso un dendrogramma.

I cinque passi sopra delineati, come già detto, costituiscono lo scheletro comune di molti metodi gerarchici. Le uniche differenze esistenti si riscontreranno nel passo 1 e nel passo 2.

L’analisi gerarchica di tipo agglomerativo può essere effettuata in R attraverso la funzione:



Alcune delle opzioni disponibili per ***method*** sono (ricordiamo che tutti questi metodi, sono stati calcolati standardizzando i dati, la variabile dataset rappresenta il dataset standardizzato):

1. Metodo del legame singolo (single)

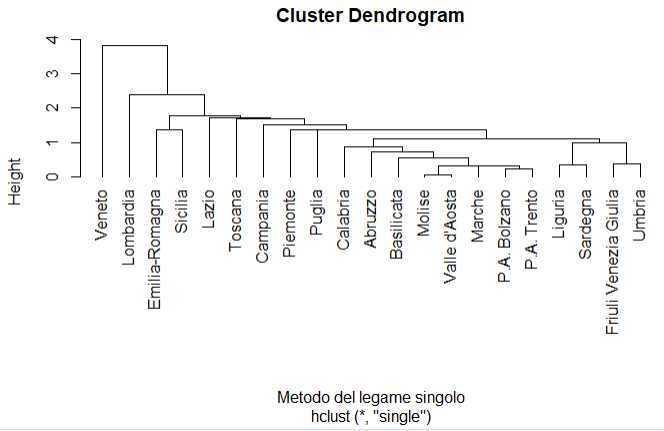
Il primo metodo è quello del legame singolo, il quale accorpa i cluster basandosi sulla distanza più piccola. Una volta fatto l’accorpamento, si ricalcolano le distanze e si riconsidera la nuova distanza minima. Si continua fino ad avere un unico cluster.

Tramite il seguente codice in R:

Immagine che contiene testo

Descrizione generata automaticamente

È possibile ottenere il dendrogramma generato dalle iterazioni del metodo del legame singolo:



Osservando attentamente il dendogramma, risultano esserci quattro cluster principali: il primo contiene solo il Veneto, il secondo va dalla Lombardia alla Puglia, il terzo dalla Calabria alla P. A. di Trento ed il quarto dalla Liguria all’Umbria.

È bene ribadire, però, che la suddivisione in cluster, dipende anche dal metodo utilizzato; quindi, vediamo anche gli altri metodi che cluster producono.

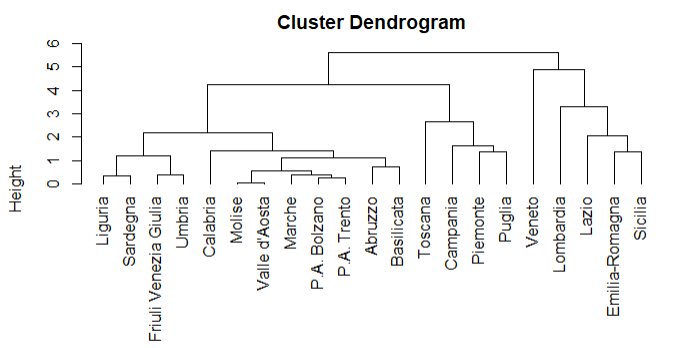
2. Metodo del legame completo (complete)

Il secondo metodo è quello del legame completo, che, a differenza del singolo, effettua un’analisi opposta, considerando le distanze massime tra gli elementi esterni al cluster.

Di seguito è possibile trovare il codice in R per tale metodo:

Immagine che contiene testo

Descrizione generata automaticamente

Output:  


Dal seguente dendrogramma si può dedurre innanzitutto che adesso la suddivisione in cluster è cambiata rispetto al metodo singolo utilizzato precedentemente, infatti con l’utilizzo del metodo del legame completo si notano subito grandi differenze riguardo la divisione in cluster.

Quindi, tramite il metodo del legame completo, in risalto si hanno cinque cluster principali: il primo che va dalla Liguria all’Umbria, il secondo che va dalla Calabria alla Basilicata, il terzo che va dalla Toscana alla Puglia, Il quarto che contiene solo il Veneto ed il quinto che va dalla Lombardia alla Sicilia.

3. Metodo del legame medio (average)

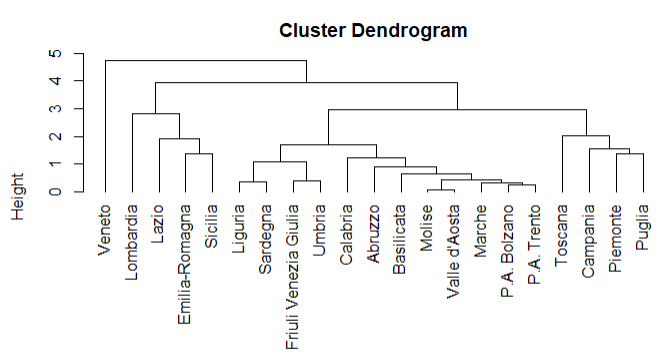
Il terzo metodo pone il calcolo della nuova distanza effettuando una media tra tutte le possibili distanze tra i valori interni nel cluster appena creato e quelli all’esterno. È importante sapere che il numero di individui in ogni cluster incide molto sul calcolo delle nuove distanze, difatti cluster più grandi avranno un peso maggiore rispetto agli altri.

Di seguito è possibile trovare il codice in R per tale metodo:

Immagine che contiene testo

Descrizione generata automaticamente

Output:

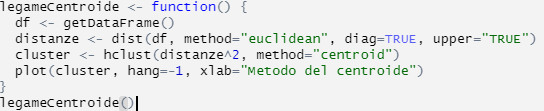


Quest’ultimo dendrogramma è molto simile a quello riprodotto dal metodo del legame completo: i principali cluster individuati sono gli stessi e i raggruppamenti sono molto simili.

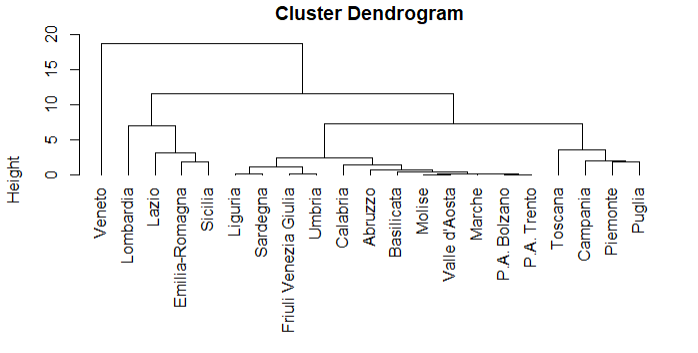
4. Metodo del centroide (centroid)

Il metodo del centroide utilizza una metrica basata sulle distanze dei centroidi dei gruppi, dove il centroide è la media campionaria tra tutti gli elementi dei due cluster in considerazione. Come per il metodo precedente, anche in questo caso è rilevante il numero degli elementi appartenenti ad un singolo cluster, siccome la metrica perde l’appellativo di distanza dato che in questo nuovo caso si considera la distanza quadratica.

Di seguito è possibile trovare il codice in R per tale metodo:



Output:



A parte le distanze quadratiche, anche questo dendrogramma risulta essere molto simile a quello del metodo del legame completo, siccome i cluster sono molto simili.

5. Metodo della mediana

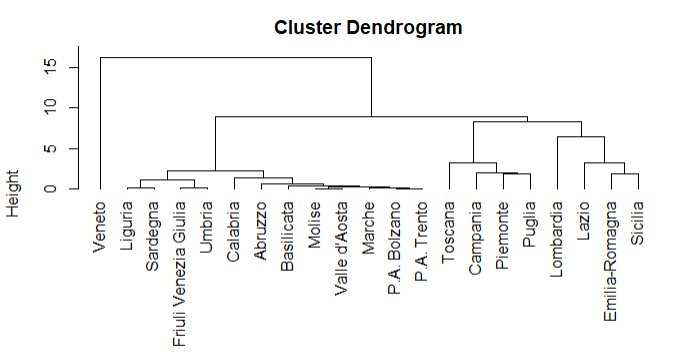
Oltre al metodo del centroide, anche il metodo della mediana si basa sulle distanze elevate al quadrato. Tale metodo è molto simile a quello del centroide, con la differenza che non si basa sulla numerosità dei cluster.

Di seguito è possibile trovare il codice in R per tale metodo:

Immagine che contiene testo

Descrizione generata automaticamente

Output:



Anche il metodo della mediana sembra riportare gli stessi cluster.

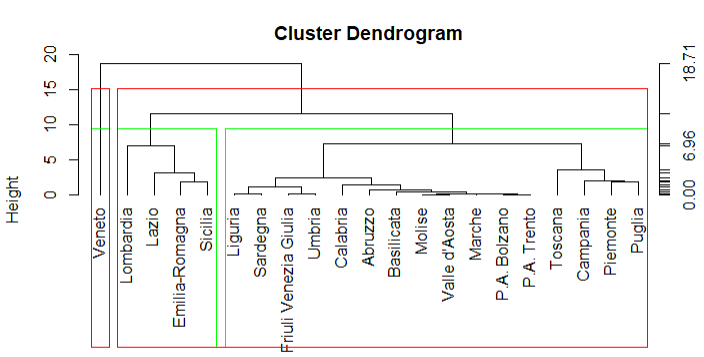
# Analisi del Dendogramma

Procediamo adesso ad analizzare il dendrogramma ottenuto tramite il metodo del centroide, andando a calcolare le misure di non omogeneità della partizione trovata. Utilizziamo la funzione R ***rect.hclust()*** ed in particolare gli passiamo il parametro *k* che ci permette di dividere in base al numero di cluster che si vogliono ottenere (nel nostro caso con il verde partizioniamo in 3 e in rosso partizioniamo in 2):

Immagine che contiene testo

Descrizione generata automaticamente

Output:



Il dendrogramma risulta diviso in 3 cluster (colore verde) ed in 2 cluster (colore rosso).

# Inserire Individui nel Cluster

Per ottenere una suddivisione degli individui in cluster in corrispondenza di un determinato livello di distanza oppure in corrispondenza di un prefissato numero di cluster, R utilizza anche la funzione cutree():

Immagine che contiene testo

Descrizione generata automaticamente

Output:

Immagine che contiene testo

Descrizione generata automaticamente

L’output della funzione cutree() è un vettore contenente numeri interi positivi associati ai cluster in cui sono stati inseriti i vari individui.

# Misure di sintesi associate ai cluster

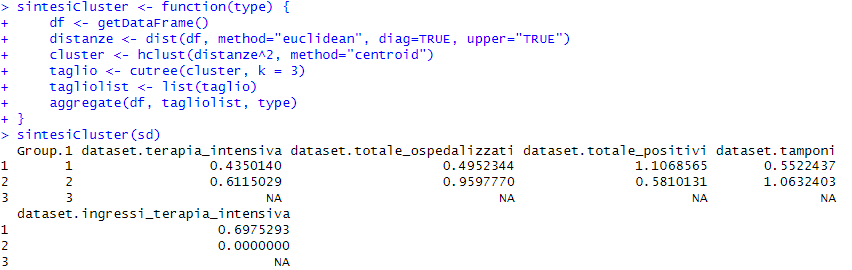
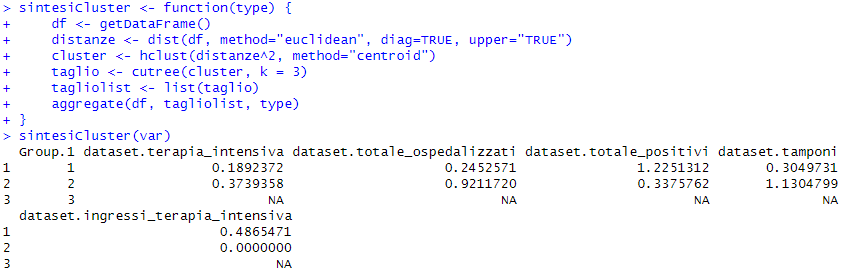
In R è inoltre possibile ricavare misure di sintesi come la media campionaria, la varianza campionaria, la deviazione standard, etc, sui singoli cluster, ottenuti tagliando il dendrogramma tramite la funzione **cutree()** e successivamente utilizzando la funzione **aggregate(X, by, FUN)**

dove:

* X rappresenta una matrice numerica o un data frame;
* by è una lista di indici sulla base dei quali le colonne di X vanno aggregate;
* FUN è la funzione da applicare alle colonne di X, separatamente per i vari gruppi individuati in base a by.

Segue il codice in R usato:

Immagine che contiene testo

Descrizione generata automaticamente

# Misure di Non Omogeneità

A questo punto è possibile calcolare le misure di non omogeneità totale, uguale per tutti i metodi, ed interna ad ognuno dei tre cluster per ciascun metodo in modo da verificare quale metodo dia una suddivisione migliore.

Segue il codice usato in R per il calcolo della misura di non omogeneità totale:

Immagine che contiene testo

Descrizione generata automaticamente

Output:



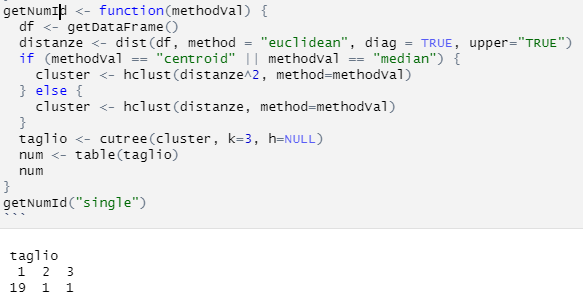
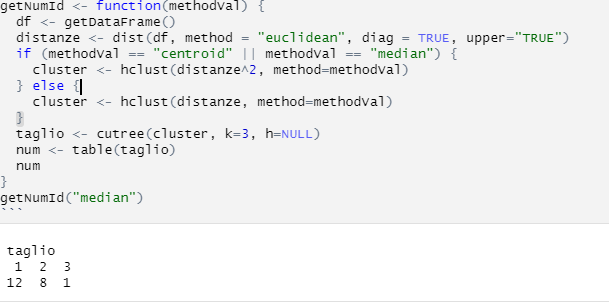
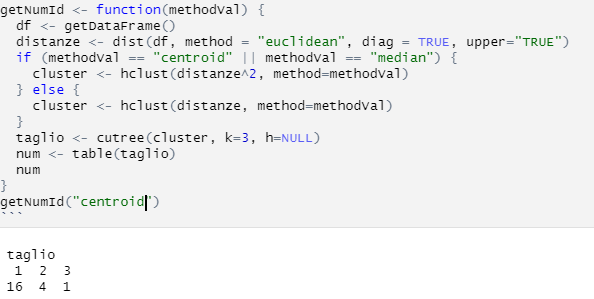
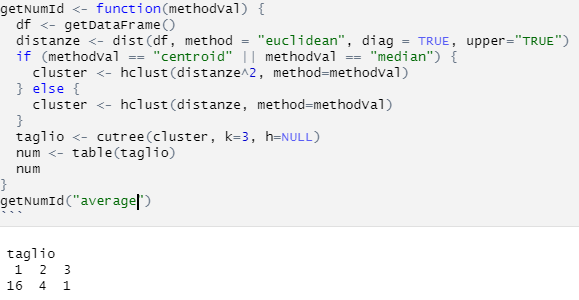
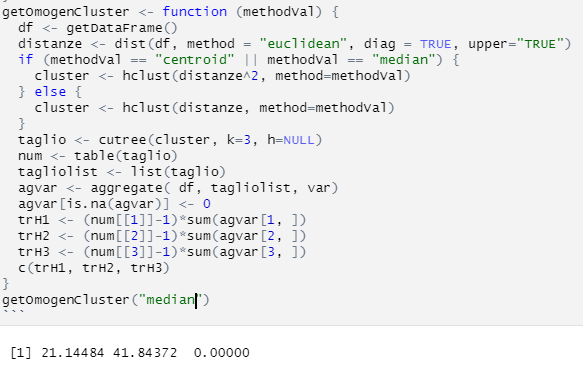
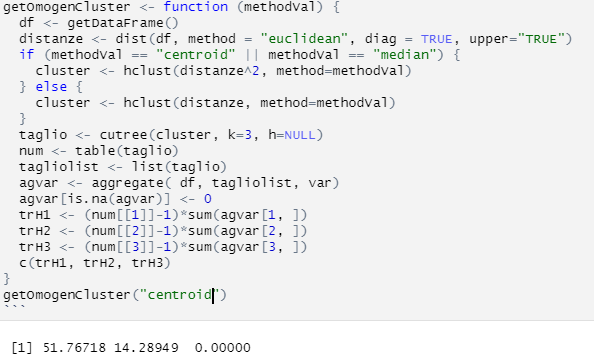
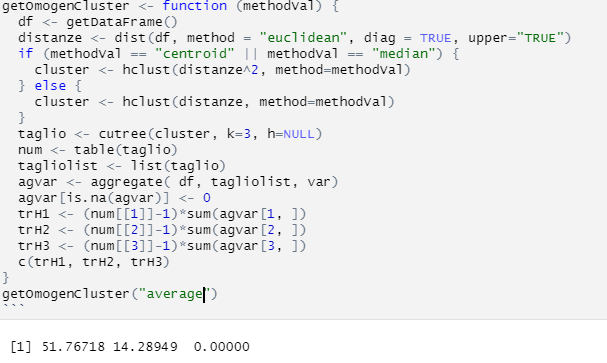
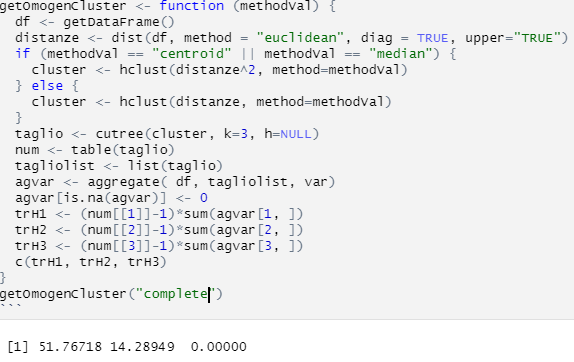
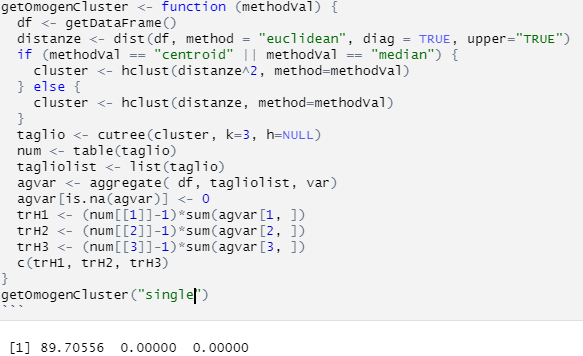
Successivamente, tramite la funzione cutree(), andiamo ad effettuare una suddivisione degli individui in cluster. Il risultato della funzione cutree() verrà passato come parametro a table() per calcolare il numero di individui in ogni cluster. Questa operazione viene ripetuta per ogni metodo gerarchico, tramite la seguente funzione implementata in R:

Immagine che contiene testo

Descrizione generata automaticamente

I risultati ottenuti ci confermano in parte quello che avevamo accennato nelle sezioni precedenti, ovvero che il metodo singolo si discosta da tutti gli altri; infatti, possiamo notare che il numero degli individui nei cluster è uguale per tutti gli altri metodi. In più notiamo anche una leggera differenza anche rispetto al metodo della mediana, che riporta un numero di elementi diverso.

Con il prossimo codice andremo a calcolare le misure di non omogeneità dei gruppi:



Nel metodo del legame completo, della media, del centroide e della mediana abbiamo ottenuto gli stessi valori per ogni gruppo proprio perché, come visto precedentemente, si avevano gli stessi numeri di regioni nei tre gruppi.

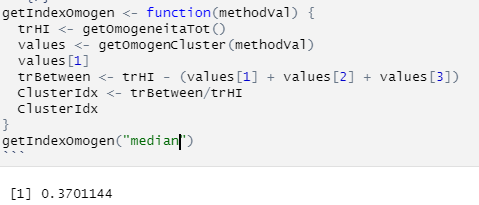
Una volta calcolate le misure di non omogeneità di ogni gruppo, possiamo finalmente passare al calcolo dell’indice di non omogeneità tra i cluster:

Immagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamente

Come già detto precedentemente, anche in questo caso, l’unico valore che si discosta molto dagli altri è quello del metodo singolo. Inoltre, si può notare che il metodo di classificazione migliore è quello della mediana, con indice 0.3701144.

# Metodi non gerarchici

Fino ad ora abbiamo visto i cluster ottenuti da metodi gerarchici, quindi adesso ci rigeneriamo i cluster ma questa volta tramite i non gerarchici.

L’obiettivo dei metodi non gerarchici è quello di ottenere un’unica partizione degli n individui di partenza in cluster. A differenza dei metodi gerarchici, in tali tecniche è consentito riallocare gli individui già classificati ad un livello precedente dell’analisi.

In molti metodi non gerarchici di clustering si assume che il numero di cluster in cui suddividere l’insieme totale degli n individui sia fissato a priori dal ricercatore, mentre in altri tale numero è determinato nel corso dell’analisi.

Il metodo più utilizzato prende il nome di k-means. Tale metodo richiede che il numero di cluster sia specificato a priori e fornisce in output un’unica partizione. Esso è composto dai seguenti step:

1. Fissare a priori il numero k di cluster specificando m punti di riferimento iniziali;

2. Considerare tutti gli individui e attribuire ciascuno di essi al cluster individuato dal punto di riferimento da cui ha distanza minore;

3. Calcolare il baricentro (il centroide) di ognuno dei k gruppi così ottenuti. Tali centroidi costituiscono i punti di riferimento per i nuovi cluster;

4. Valutare la distanza di ogni unità da ogni centroide ottenuto al passo precedente. Se la distanza minima non è ottenuta in corrispondenza del centroide del gruppo di appartenenza, allora si procede a spostare l’individuo presso il cluster che ha il centroide più vicino;

5. Si ricalcolano i centroidi dei k gruppi così ottenuti;

6. Ripetere il procedimento a partire dal punto (4) fino a che i centroidi non subiscono ulteriori modifiche rispetto all’iterazione precedente. Si procede così iterativamente a spostamenti successivi fino a raggiungere una configurazione stabile, ossia gli individui all’interno di ogni cluster non cambiano al ripetersi del procedimento.

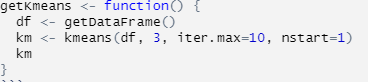
I vantaggi del metodo k-means sono la velocità di esecuzione dei calcoli e l’estrema libertà che viene lasciata agli individui di raggrupparsi e allontanarsi.

L’analisi con il metodo k-means si effettua in R mediante la funzione:

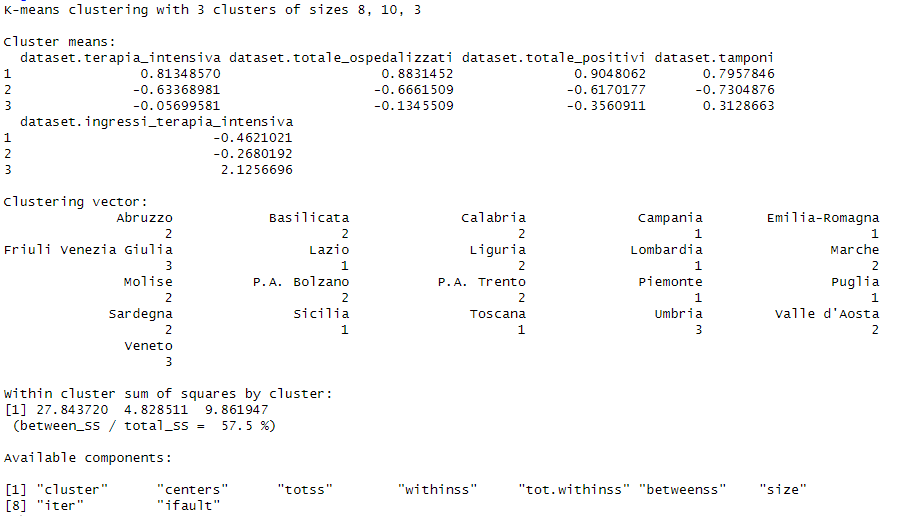


A differenza dei gerarchici, nei non gerarchici non è necessario calcolare prima le distanze.

Snippet di codice usato in R per il calcolo del kmeans:



Output:



Nella funzione kmeans() di R abbiamo impostato come numero massimo di iterazioni 10 e scelta dei punti casuali 1, passando un dataframe con valori standardizzati tramite la funzione scale().

La misura di non omogeneità tra i cluster diviso la misura di non omogeneità totale è pari al 57.5% quindi maggiore del 50%.

Applicando la funzione str() possiamo ottenere maggiori info, come ad esempio:

* $cluster, un vettore di interi che indica l’allocazione di ogni individuo;
* $center, matrice contenente i centroidi dei cluster;
* $withinss, un vettore contenente le misure di non omogeneità statistica calcolate all’interno di ognuno dei cluster;
* $size, dimensione dei gruppi formati.

Segue il codice R utilizzato:

Immagine che contiene testo

Descrizione generata automaticamente

Output:

Immagine che contiene testo

Descrizione generata automaticamente

# Confronto tra Metodi Gerarchici e Metodi Non Gerarchici

In conclusione, i metodi gerarchici presentano un evidente vantaggio dal punto di vista computazionale rispetto ai metodi di ottimizzazione; tuttavia, risultano sensibili a vettori delle caratteristiche anomali e non consentono di modificare la configurazione raggiunta, ossia una volta che un individuo è stato attribuito ad un cluster permane al suo interno per sempre. I metodi non gerarchici non presentano questo problema poiché consentono di riallocare gli individui precedentemente classificati, ma richiedono una scelta opportuna della configurazione iniziale.