Frontespizio

Sommario

[1. Grafici ed indici di posizione centrali 6](#_Toc126867018)

[1.1 Variabili quantitative 6](#_Toc126867019)

[1.2 Grafico a barre delle persone risultate positive 6](#_Toc126867020)

[1.3 Indici di posizione centrali 8](#_Toc126867021)

[1.4 Grafico a barre ordinate 8](#_Toc126867022)

**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.

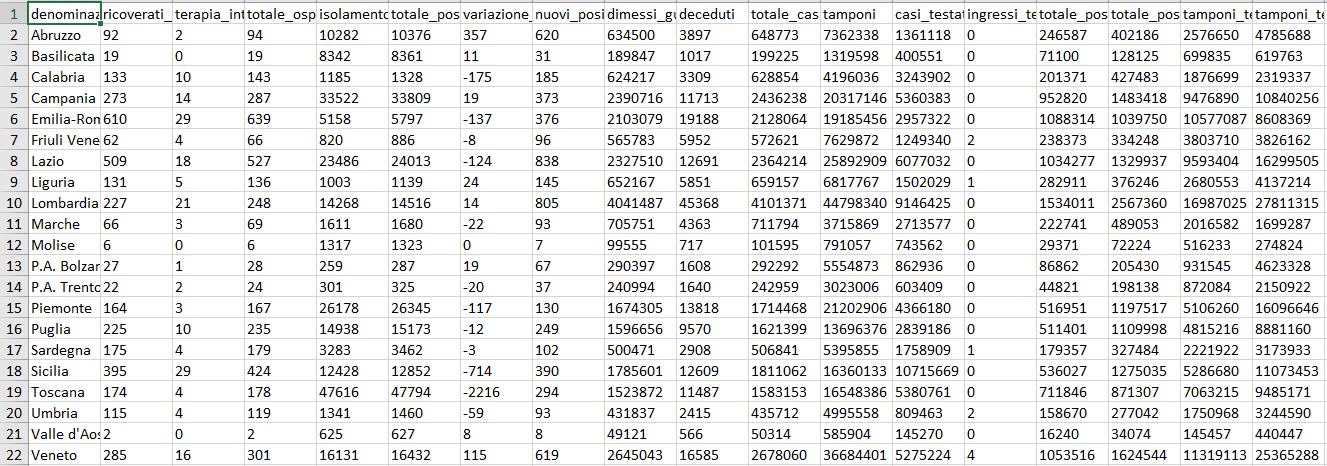
Un altro importante aspetto della statistica è il frequente uso di varie tipologie di diagrammi. Molti aspetti che saranno mostrati matematicamente verranno immediatamente confermati (se non chiariti) tramite rappresentazioni grafiche mediante diagrammi. Per generare i

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

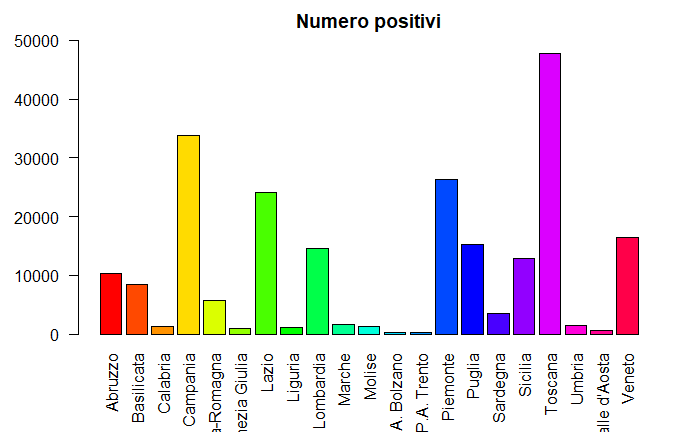
spazio in cui è presente un’asse verticale che scandisce il progressivo aumento di misura man mano che si sale verso l’alto.

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.

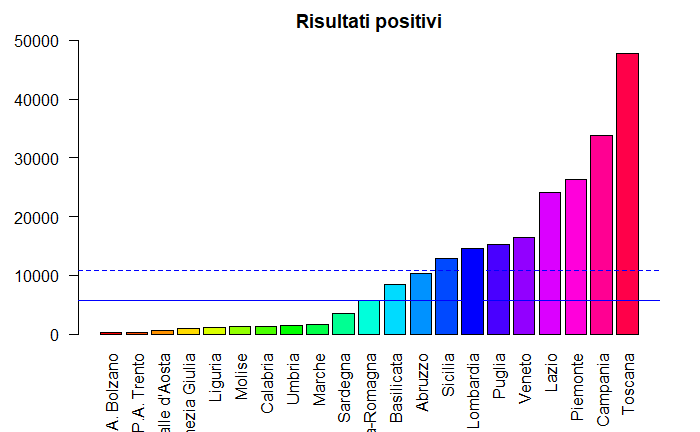
Immagine che contiene testo

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, mentre quella continua rappresenta la mediana.