# Introduzione

Questo documento riassume gli step fondamentali dell’analisi statistica per arrivare a trarre conclusioni sui dati.  
Non viene spiegato qui nel dettaglio il codice R utilizzato, che può essere consultato a questo link ([click](https://drive.google.com/file/d/1uYSRP6obmJvkKyHEJeWjYHlkPQV5Vzcj/view?usp=sharing)).   
Oltre ai file .png e al codice R, è possibile scaricare in formato .docx questa relazione ([click](https://docs.google.com/document/d/1Mu6Jc3ekIUDlezDVqe3tjgTboOdWfX31/edit?usp=sharing&ouid=102545256122025424730&rtpof=true&sd=true)).

## Note varie

* In RStudio ho utilizzato il package **readxl** per caricare in memoria il contenuto del file vendite.xls, il package **crayon** per scrivere su console con i colori e il package **ggpubr** per poter utilizzare la funzione ggpaired che permette di plottare dati accoppiati.  
  Tutte le altre funzioni che ho utilizzato non hanno bisogno di ulteriori package.
* Nel caso servisse, la versione di R che ho utilizzato è la **4.1.3**.

## Prime osservazioni sulla traccia

Occorre verificare se una campagna pubblicitaria è stata efficace, avendo a disposizione **due campioni appaiati** poiché ogni “coppia” (prima e dopo la campagna) fa riferimento allo stesso punto vendita, ossia i due campioni non sono indipendenti.  
Inoltre il numero di punti vendita è quindi i campioni sono sufficientemente grandi da permetterci di non richiedere l’ipotesi di normalità sulla popolazione, per utilizzare in seguito il **test t**.

# Statistica descrittiva

## Immagine che contiene testo Descrizione generata automaticamenteIndici

Dopo aver caricato i dati in un dataframe grazie alla funzione **read\_excel**, come prima cosa nella funzione main decido di mostrare gli **indici di posizione** e gli **indici di variabilità** dei due campioni.   
Per fare ciò, ho definito una funzione che utilizza le funzionalità del package crayon per stampare su console con i colori.

A destra uno screenshot della console R dopo l’esecuzione del codice.

Di seguito invece una tabella che riassume i diversi indici sul numero di vendite.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Indice | Tipo | Prima della campagna | Dopo la campagna | Funzione in R |
| Media campionaria | Indice di posizione | 215.93 | 227.41 | **mean**(data) |
| Primo quartile | Indice di posizione | 186.5 | 197.5 | **quantile**(data, 0.25, type = 2) |
| Mediana campionaria | Indice di posizione | 216 | 223 | **median**(data) |
| Terzo quartile | Indice di posizione | 240.5 | 252 | **quantile**(data, 0.75, type = 2) |
| Varianza campionaria | Indice di variabilità | ≈ 1652.873838 | ≈ 1868.628181 | **var**(data) |
| Deviazione standard campionaria | Indice di variabilità | ≈ 40.65555 | ≈ 43.22763 | **sd**(data) |
| Scarto interquartile | Indice di variabilità | 54 | 54.5 | **IQR**(data, type = 2) |
| Range (min e max) | Indice di variabilità | 132 333 | 135 345 | **range**(data) |

Nota: alle funzioni quantile e IQR passo il parametro type = 2 perché esistono definizioni alternative di quartile.   
Di default R utilizza una definizione diversa da quella vista a lezione; quindi, è necessario specificare type = 2.

## Grafici

Continuando l’analisi per quanto riguarda la statistica descrittiva, può essere utile visualizzare i dati con grafici. Ho deciso di utilizzare quattro grafici che riporto:

|  |  |
| --- | --- |
| **Istogramma vendite precedenti** alla campagna e gaussiana di media e varianza uguali a quelle del campione. | **Istogramma vendite successive** alla campagna e gaussiana di media e varianza uguali a quelle del campione. |
| **Plot di paired data prima** e **dopo**. | **Boxplot** prima e dopo la campagna, ho aggiunto i label per indicare esplicitamente il valore numerico per il minimo, il primo quartile, la mediana, il terzo quartile e il massimo (qui si escludono gli outliers, se occorre sapere il massimo anche tra gli outliers è sufficiente andare nella pagina precedente, poiché l’indice range lo indica). |

Per generare i grafici ho utilizzato le funzioni **hist**, **ggpaired** e **boxplot** (come già detto, ggpaired appartiene al package ggpubr).  
Il mio codice non plotta direttamente i grafici su RStudio ma li salva su file con estensione .png; le immagini originali si possono scaricare da qui ([click](https://drive.google.com/drive/folders/160n0uHVFCn1eB0_rRyayNq4nv2WDPMDy?usp=sharing)).  
Gli istogrammi singolarmente presentano anche una gaussiana di media e varianza uguali a quelle del campione, per mostrare che ha senso che non sia necessaria l’ipotesi di normalità quando andremo a svolgere il test t.  
Osservando i grafici (si può notare soprattutto dagli ultimi due grafici), si potrebbe pensare che le vendite siano aumentate dopo la campagna pubblicitaria.  
Questo è necessario verificarlo svolgendo un test di ipotesi.

# Verifica di ipotesi

Per vedere se i dati ci consentono di concludere che non si può escludere che la campagna pubblicitaria risulta efficace, svolgo un **test t sulla differenza delle medie** di due campioni normali **accoppiati** (dati registrati durante la settimana **successiva** alla campagna pubblicitaria) di media e (dati registrati durante la settimana **precedente**) di media .  
  
Per svolgere il test, R offre una funzione **t.test**; tuttavia voglio prima svolgere “manualmente” il test a livello di significatività e in seguito utilizzare la funzione t.test, che mi permette anche di avere altri dati importanti come ad esempio il p-value.

Con , poiché voglio verificare l’aumento della media nella popolazione successiva alla campagna pubblicitaria, utilizzo come ipotesi alternativa .

|  |  |  |  |
| --- | --- | --- | --- |
| Ipotesi nulla | Ipotesi alternativa | Valore della statistica | Regione critica |
|  |  |  |  |

Immagine che contiene testo

Descrizione generata automaticamenteInizio calcolando, in una funzione definita da me, il vettore delle differenze, che salvo in una variabile, la media campionaria , la varianza campionaria e la deviazione standard delle differenze.   
Ora per applicare il test serve avere il valore della statistica (che posso calcolare, avendo tutti i dati necessari e ) e il 95% percentile della distribuzione t di Student a 99 gradi di libertà.

Immagine che contiene testo

Descrizione generata automaticamenteCome mostrato qui sopra, li faccio calcolare da R; in particolare per la t di Student si utilizza la funzione **qt** con i parametri 0.95 e 99.  
Possiamo concludere che a livello di significatività 5%, poiché il valore della statistica è maggiore del percentile , **i dati mi permettono di rifiutare l’ipotesi nulla** e quindi non si può escludere che ci sia un aumento delle vendite dopo la campagna pubblicitaria.

Ora per avere altre informazioni (e come conferma) utilizzo la funzione **t.test** che prende come parametri i due vettori (**dopo e prima** la campagna), **paired = TRUE** e **alternative = “greater”**, per utilizzare l’ipotesi alternativa riportata sopra; di default quindi non è necessario specificare altri parametri.  
Questa funzione che rende disponibile R (non ho utilizzato package aggiuntivi) stampa su console quanto riportato a destra.  
Possiamo vedere (ignorando first e second che sono i nomi dei parametri all’interno della mia funzione che alla fine richiama t.test) il valore della statistica che avevo calcolato prima, i gradi di libertà della t di Student,   
**il p-value** che risulta , l’ipotesi alternativa scelta in precedenza, **un intervallo di confidenza al 95%** e la media delle differenze, anch’essa già calcolata.

# Conclusioni

* Dopo aver utilizzato la funzione **t.test**, ho avuto la conferma sulla correttezza del valore della statistica e della media delle differenze, ma soprattutto ora si conosce una stima del p-value ed è stato costruito l’intervallo di confidenza al 95%.
* Il p-value risulta molto piccolo, quindi vi è **forte evidenza statistica** che ci sia un **aumento delle vendite** dopo la campagna pubblicitaria.
* Infatti, il p-value risulta e, con livello di significatività, risulta che i dati ci permettono di rifiutare l’ipotesi nulla.
* Inoltre, con l’intervallo di confidenza costruito, sappiamo che a livello 95% l’aumento della media è stimato sopra .
* I dati non ci permettono quindi di escludere che la campagna pubblicitaria sia stata **efficace**.