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).
Oltre ai file .png e al codice R, è possibile scaricare in formato .docx questa relazione (click).

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

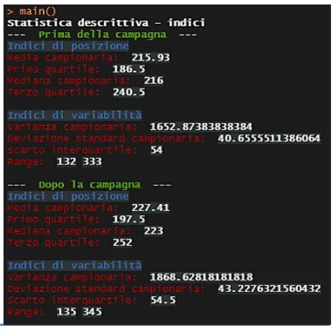
Occorre verificare se una campagna pubblicitaria è stata efficace, avendo a disposizione **due campioni appaiati** poiché ogni "coppia" (prima e dopo la campagna) fa riferimento allo stesso punto vendita, ossia i due campioni non sono indipendenti.
Inoltre il numero di punti vendita è $n = 100 \geq 30$ 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

INDICI

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.



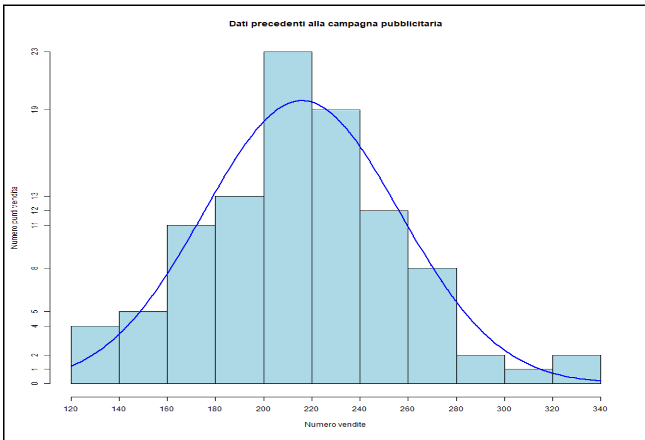
Di seguito invece una tabella che riassume i diversi indici sul numero di vendite.

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)

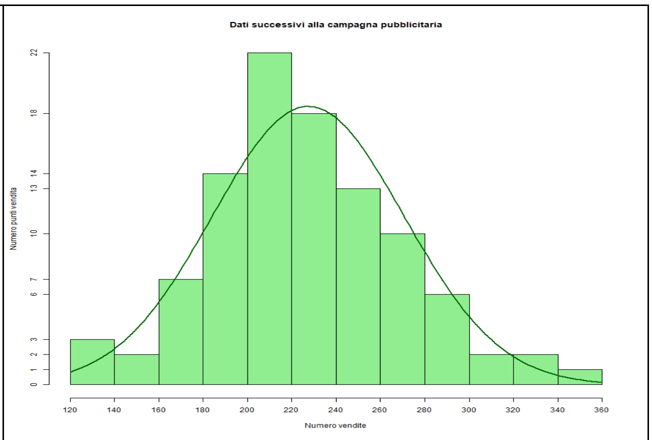
Nota: alle funzioni quantile e IQR passo il parametro type = 2 perché esistono definizioni alternative di quartile.
Di default R utilizza una definizione diversa da quella vista a lezione; quindi, è necessario specificare type = 2.

GRAFICI

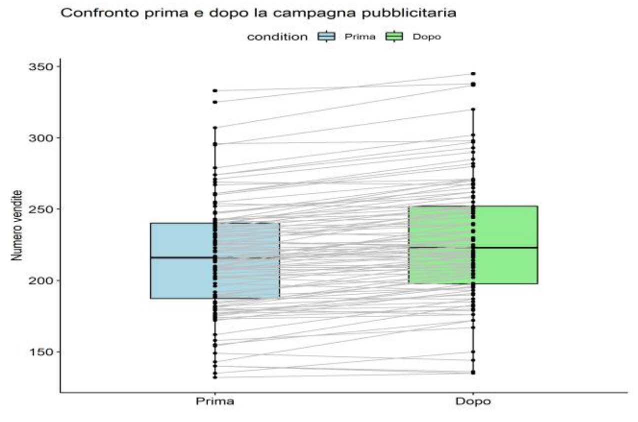
Continuando l'analisi per quanto riguarda la statistica descrittiva, può essere utile visualizzare i dati con grafici. Ho deciso di utilizzare quattro grafici che riporto:



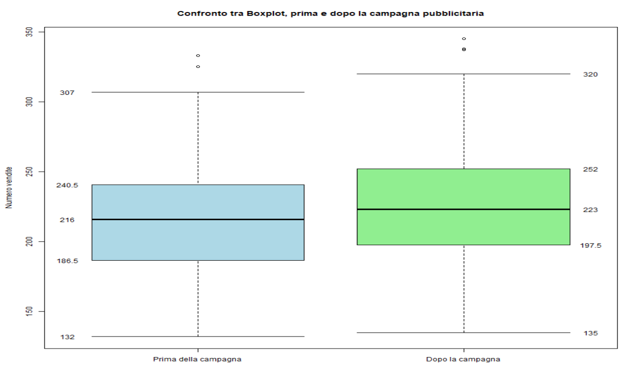
Istogramma vendite precedenti alla campagna e gaussiana di media e varianza uguali a quelle del campione.



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).
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** $x_1, \dots, x_{100}$ (dati registrati durante la settimana **successiva** alla campagna pubblicitaria) di media μ_x e $y_1, \dots, y_{100}$ (dati registrati durante la settimana **precedente**) di media μ_y.

Per svolgere il test, R offre una funzione **t.test**; tuttavia voglio prima svolgere "manualmente" il test a livello di significatività $\alpha = 0.05$ e in seguito utilizzare la funzione t.test, che mi permette anche di avere altri dati importanti come ad esempio il p-value.

Con $\mu_0 = 0$, poiché voglio verificare l'aumento della media nella popolazione successiva alla campagna pubblicitaria, utilizzo come ipotesi alternativa $\mu_x > \mu_y$.

Ipotesi nulla	Ipotesi alternativa	Valore della statistica	Regione critica
$\boldsymbol{H_0 : \mu_x \leq \mu_y}$	$H_1 : \mu_x > \mu_y$	$t = \frac{\overline{d_n}}{s_d}\sqrt{n}$	$t > t_{n-1,\alpha}$

```
[Paired T-Test, alpha = 0.05, mu_0 = 0, alternative = greater]
n:  100
Mean of the differences:  11.48
Variance of the differences:  113.080404040404
Standard deviation of the differences:  10.6339270281681
t:  10.7956354878031
t-student df = 99, alpha = 0.05:  1.66039115601699
```

Inizio calcolando, in una funzione definita da me, il vettore delle differenze, che salvo in una variabile, la media campionaria $\overline{d_n}$, la varianza campionaria s_d^2 e la deviazione standard s_d delle differenze.
Ora per applicare il test serve avere il valore della statistica t (che posso calcolare, avendo tutti i dati necessari e $n = 100$) e il 95% percentile della distribuzione t di Student a 99 gradi di libertà.

Come 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 $t \approx 10.796$ è maggiore del percentile $t_{99,0.05} \approx 1.6604$, **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.

```
        Paired t-test

data:  first and second
t = 10.796, df = 99, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 9.714352      Inf
sample estimates:
mean of the differences
                  11.48
```

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 $\mu_0 = 0$ 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 t che avevo calcolato prima, i gradi di libertà della t di Student,
il p-value che risulta $< 2.2 \times 10^{-16}$, 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 t 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 $\overline{\alpha} < 2.2 \times 10^{-16}$ e, con α livello di significatività, risulta che $\forall \alpha > \overline{\alpha}$ 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 9.714352.
- I dati non ci permettono quindi di escludere che la campagna pubblicitaria sia stata **efficace**.