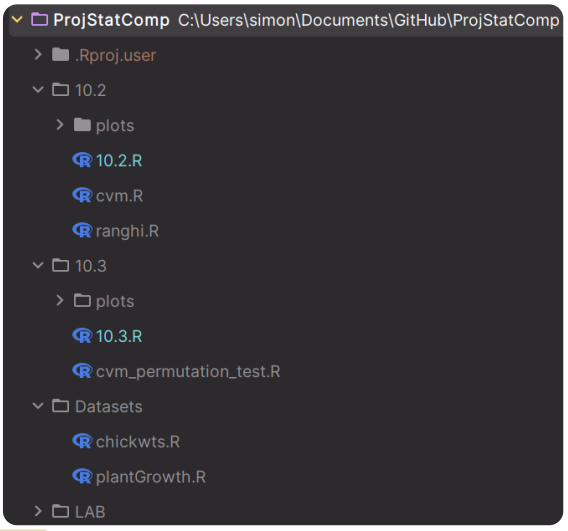
Progetto 10.x

Impostazione Progetto

Siccome la mia scelta di implementare l'esercizio 10.3 comporta inevitabilmente di implementare il 10.2 ho deciso di dividere i lavori in modo da differenziare sia i paragrafi della relazione sia il progetto in R.
Il progetto ha le sue parti divise in modo che con il comando `source()` si possano importare i sorgenti presenti nel progetto. Utile ad esempio per richiamare l'implementazione del 10.2 nella funzione del 10.3.

Struttura Progetto

Il progetto è diviso come segue



Nella cartella *10.2* si trova il file `10.2.R` che contiene il codice main per far andare l'esercizio mentre nel file `cvm.R` c'è la vera e propria implementazione della statistica Cramér-Von Mises. La cartella *plots* contiene i file immagine dei grafici plottati ed eventualmente salvati dei vari esperimenti.
Nella cartella *10.3* si rispetta lo stesso schema.
Nella cartella *Datasets* si trovano gli import dei dataset usati per gli esperimenti.

10.2

Write a function to compute the two-sample Cramér–von Mises statistic.
The Cramér–von Mises distance between distributions is

$$\omega^2 = \int \int \big(F(x) - G(y)\big)^2 dH(x,y)$$

where $H(x,y)$ is the joint CDF of X and Y . For a test of equal distributions,
the corresponding test statistic is based on the joint empirical
distributions, so it is a function of the ranks of the data. First,
compute the ranks r_i of the X sample, $i = 1, \dots, n$ and the ranks s_j of the Y sample $j = 1, \dots, m$ (see rank function).
compute

$$U = n \sum_{i=1}^{n} (r_i - i)^2 + m \sum j = 1^m (s_j - j)^2$$

Note that U can be vectorized and evaluated in one line of R code. Then
the Cramér–von Mises two-sample statistic is

$$W^2 = \frac{U}{nm(n+m)} - \frac{4nm-1}{6(m+n)}$$

Large values of W^2 are significant.

Tradotto

Scrivere una funzione per calcolare la statistica di Cramér-von Mises a due campioni.
La distanza di Cramér-von Mises tra distribuzioni è

$$\omega^2 = \int \int \big(F(x) - G(y)\big)^2 dH(x,y)$$

dove $H(x,y)$ è la CDF congiunta di X e Y. Per un test di uguaglianza delle distribuzioni, la statistica corrispondente si basa sulle distribuzioni empiriche congiunte, quindi è una funzione dei ranghi dei dati. Per prima cosa, calcolare i ranghi r_i del campione X, $i = 1, \dots, n$ e i ranghi s_j del campione Y $j = 1, \dots, m$ (vedi funzione rank). calcolare

$$U = n \sum_{i=1}^{n} (r_i - i)^2 + m \sum j = 1^m (s_j - j)^2$$

Si noti che U può essere vettorializzato e valutato in una sola riga di codice R. Allora
la statistica a due campioni di Cramér-von Mises è

$$W^2 = \frac{U}{nm(n+m)} - \frac{4nm - 1}{6(m+n)}$$

Valori elevati di W^2 sono significativi.

Cramér-Von Mises

La statistica di cmv restituisce un valore che misura la distanza fra due cdf. Siccome è una distanza ben definita sappiamo che:

1. $d(F_X(x), F_X(x)) = 0$ dove d è la distanza e X la vc
2. $d(F_X(x), F_Y(y)) = \omega \in [0, +\infty)$ dove d è la distanza e X, Y le vc

Questo test non parametrico torna utilissimo nel confronto di due vc. *Siccome è un test non parametrico significa che io non posso fare assunzioni sulle distribuzioni dei miei campioni e questo ci permette di usarlo in esperimenti di cui non conosco a priori la distribuzione dei dati*. L'adattamento proposto dal libro consente di poter utilizzare questa statistica anche con campioni empirici poiché si utilizzano i *ranghi* nel calcolo della U.

Ranghi

In statistica, il "rango" di un'osservazione in un insieme di dati è la sua posizione nella sequenza ordinata di dati. I ranghi sono assegnati ai dati ordinati dal più piccolo al più grande, assegnando il rango 1 al dato più piccolo, il rango 2 al secondo dato più piccolo, e così via. Quando ci sono dei pareggi (valori identici) o *ties*, si assegna a ogni valore pareggiato la media dei ranghi. Ad esempio, se due valori sono i terzi più piccoli, entrambi ricevono un rango di 3.5, che è la media dei ranghi 3 e 4. La gestione dei ties non è affidata solo alla media ma se serve anche a `max(), min(), first(), last(), random()`.

Esempio

Prendiamo linseed e soybean dal dataset chickwts. Si precisa che hanno rispettivamente lunghezze 12 e 14.

	Linseed	Soybean
1	309	243
2	229	230
3	181	248
4	141	327
5	260	329
6	203	250
7	148	193
8	169	271

	Linseed	Soybean
9	213	316
10	257	267
11	244	199
12	271	171
13	/	158
14	/	248

Eseguendo il seguente codice e dando per scontato l'import del dataset

```
ranghi ← rank(c(sort(linseed), sort(soybean)))
rango_linseed← ranghi[seq_along(linseed)]
rango_soybean←ranghi[(length(linseed)+1):
                                        length(linseed)+length(soybean))]
print(cbind(rango_linseed= rango_linseed, rango_soybean=rango_soybean))
```

otteniamo la seguente tabella che rappresenta appunto la posizione di ogni valore del campione che avrebbe in un ipotetico vettore congiunto ordinato.

	Rango Linseed	Rango Soybean
1	1.0	3.0
2	2.0	5.0
3	4.0	7.0
4	6.0	8.0
5	9.0	12.0
6	10.0	13.0
7	11.0	15.5
8	14.0	15.5
9	18.0	17.0
10	19.0	20.0
11	21.5	21.5
12	23.0	24.0
13	/	25.0
14	/	26.0

Osservazioni su CVM

Cramér-Von Mises viene spesso usato come test di esplorazione prima di procedere ad analisi più precise, se la misura di distanza è piccola ha senso procedere a considerare i campioni come appartenenti alla stessa distribuzione, altrimenti no.
Siccome faccio dei calcoli sui ranghi mi aspetto che campioni molto grandi in dimensione possano avere grandi differenze indicate dal test anche se in realtà questo potrebbe non essere vero. I test è potente nel misurare piccole variazioni e con campioni molto grandi è facile incombere in variazioni numeriche grandi che non sono direttamente correlate a differenze grandi fra le distribuzioni dei campioni.

Implementazione CVM

Premessa implementativa

L'ordinamento dei dati è fondamentale per il calcolo perché la statistica di Cramér-von Mises si basa sulla differenza cumulativa tra le distribuzioni. Senza ordinare i dati, non saremmo in grado di calcolare correttamente le differenze cumulative necessarie per il test.
Togliendo l'ordinamento dei dati si perde la proprietà di identità $d(x, y) = 0 \Leftrightarrow x = y$.

Codice

Questo codice si trova al percorso `10.2/cvm.R`

```
cramer_von_mises <- function(sample_x, sample_y, plotting = FALSE, save_plot = F, name_plot = "") {
  # Siccome non ho garanzia che i campini siano ordinati la prima cosa
  # che faccio è un sort
  sample_x <- sort(as.vector(sample_x))
  sample_y <- sort(as.vector(sample_y))

  # Lunghezza del campione X
  n <- length(sample_x)
  # Lunghezza del campione Y
  m <- length(sample_y)

  # unisco i campioni
  combinati <- c(sample_x, sample_y)
  # faccio un rango congiunto
  ranked_sample <- rank(combinati, ties.method = "average")
  # separo i ranghi
  rank_x <- ranked_sample[1:n]
  rank_y <- ranked_sample[(n + 1):(n + m)]

  # Scommentare se serve parlare dei ranghi
  # print(cbind(sample_x, rank_x,sample_y, rank_y))

  U <- n * sum((rank_x - (1:n))^2) + m * sum((rank_y - (1:m))^2)

  # Distanza delle distribuzioni
  W_squared <- U / (n * m * (n + m)) - ((4 * n * m) - 1) / (6 * (n + m))

  if (plotting == T) {
            plotting_data_cvm(sample_x, sample_y, name_plot = name_plot,
             save_plot = save_plot)
  }
  return(W_squared)
}
```

Risultati

Risultati Chickwts

- `cramer_von_mises(linseed, linseed)`
  - $W^2 = 0 \rightarrow$ come detto sopra è atteso siccome sono distribuzioni identiche e sappiamo che la distanza è ben definita
- `cramer_von_mises(linseed, soybean)`
  - $W^2 = 0.1574 \rightarrow$ non è detto sia un buon risultato come sembra ma promette bene essendo molto vicino allo zero.

Applicazione a PlantGrowth

Ho scelto per pura curiosità di tentare con un dataset differente che contiene i risultati di un esperimento di confronto delle rese (misurate in base al peso essiccato delle piante) ottenute in condizioni di controllo e in due diverse condizioni di trattamento.

Risultati PlantGrowth

- `cramer_von_mises(ctrl_group, trt1_group)`
  - $W^2 = 0.2425 \rightarrow$ sembra un risultato promettente
- `cramer_von_mises(ctrl_group, trt2_group)`
  - $W^2 = 0.395 \rightarrow$ sembra un risultato poco promettente
- `cramer_von_mises(trt1_group, trt2_group)`
  - $W^2 = 0.875 \rightarrow$ sembra un risultato pessimo

10.3

Implementare il test di Cramér-von Mises a due campioni per distribuzioni uguali come test di permutazione utilizzando

$$W^2 = \frac{U}{nm(n+m)} - \frac{4nm-1}{6(m+n)}$$

Applicare la statistica a soybean e linseed del dataset chickwts.

Commento

Vediamo le ipotesi che ho a disposizione:

- **Ipotesi Nulla (H_0)**: Le due distribuzioni campionarie sono identiche. Questo significa che qualsiasi differenza osservata tra i campioni può essere attribuita al caso.
- **Ipotesi Alternativa (H_1)**: Esiste una differenza tra le due distribuzioni campionarie. Questo indica che le differenze osservate tra i campioni non sono dovute al caso, ma riflettono piuttosto una differenza reale nelle popolazioni da cui i campioni sono stati estratti.

Si considera un valore standard $\alpha = 0.05$ di significatività.
Andando a lavorare col test di permutazioni posso ottenere un p-value che mi darà la probabilità che la mia ipotesi nulla non sia falsa. Breve cenno:

$$\text{p-value} = \mathbb{P}(|T| > \text{t.oss} | H_0)$$

Quindi se io pongo un livello di significatività $\alpha = 0.05$ e ottengo $\text{p-value} > \alpha$ significa che ho probabilità con significatività alpha di non dover rifiutare l'ipotesi nulla.

Algoritmo

1. dico che $H_0 : CDF(X) = CDF(Y)$
2. Misuro W^2
3. Costruisco un test con permutazioni dove misuro tanti $(W^2)^*$
4. Calcolo il p-value
5. Confronto il p-value con α ho due possibilità:
    1. $\text{p-value} > \alpha \rightarrow \text{accetto } H_0$
    2. $\text{p-value} \leq \alpha \rightarrow \text{rifiuto } H_0$

Implementazione

Nel progetto ci sarà un sorgente nel percorso `10.3/cvm_permutation_test.R` leggermente differente che avrà la sola aggiunta di codice per il plotting che qui non ho ritenuto necessario ma anzi confusionario.

```r
# Definizione della funzione Cramér-von Mises con simulazione di permutazione
cvm_permutation_test <- function(sample_x, sample_y, n_permutations = 999) {

  # Calcolo della statistica W^2 osservata
  W_squared_observed <- cramer_von_mises(sample_x, sample_y)

  # Preparazione per la simulazione di permutazione
  combined_samples <- c(sample_x, sample_y)
  n_x <- length(sample_x)
  n_y <- length(sample_y)
  W_squared_permutations <- NULL

  # Esecuzione delle permutazioni e calcolo di W^2 per ciascuna
  for (i in 1:n_permutations) {
    permuted_samples <- sample(combined_samples, replace = FALSE)
    permuted_x <- permuted_samples[1:n_x]
    permuted_y <- permuted_samples[(n_x + 1):(n_x + n_y)]
```

```
    W_squared_permutations[i] <- cramer_von_mises(permuted_x, permuted_y)
  }

  # Calcolo del p-value

  W_squared_tot <- c(W_squared_observed, W_squared_permutations)
  # non ho usato mean perché mi piace vedere
  # una rappresentazione simile a quella
  # matematica vista a lezione
  p_value <- sum(W_squared_tot >= W_squared_observed)/(n_permutations+1)

  # Output dei risultati
  list(W_squared_observed = W_squared_observed, p_value = p_value)
}
```

Risultati

Applicazione a chickvts

Sono stati presi in esame come richiesto *linseed* e *soybean*.

```
# Caricamento dataset dei pulcini
attach(chickwts)
boxplot(formula(chickwts))
soybean <- sort(as.vector(weight[feed == "soybean"]))
casein <- sort(as.vector(weight[feed == "casein"]))
linseed <- sort(as.vector(weight[feed == "linseed"]))
sunflower <- sort(as.vector(weight[feed == "sunflower"]))
detach(chickwts)
# Esecuzione del test di permutazione Cramér-von Mises
result <- cvm_permutation_test(linseed, linseed)
result <- cvm_permutation_test(linseed, soybean)
# Stampa del risultato
cat("W^2 osservato:", result$W_squared_observed, "\n")
cat("p-value:", result$p_value, "\n")
```

Risultati Chickwts

- `cvm_permutation_test(linseed, linseed)`
  - $W^2 = 0 \rightarrow$ come detto sopra è atteso siccome sono distribuzioni identiche e sappiamo che la distanza è ben definita
  - $\text{p-value} = 1 > \alpha \rightarrow$ che conferma l'accettazione dell'ipotesi nulla con un valore schiacciante
- `cvm_permutation_test(linseed, soybean)`
  - $W^2 = 0.1574 \rightarrow$ non è detto sia un buon risultato come sembra ma promette bene essendo molto vicino allo zero.
  - $\text{p-value} = 0.398 > \alpha \rightarrow$ altissimo, quindi non posso rifiutare H_0.

Commento Chickwts

La distanza fra la crescita dei pulcini con semi di lino e con fagioli di soia non è significativa quindi il criterio per scegliere uno rispetto all'altro dovrà arrivare da ulteriori analisi: costi mangimi, facilità di reperibilità, conservazione, benessere pulcini, etc.
Questa ipotesi è rafforzata dai risultati visti a lezione e riportati sul libro dove altre statistiche hanno confermato l'impossibilità di non accettare l'ipotesi nulla nel confronto fra i campioni `linseed` e `soybean`.

Risultati PlantGrowth

- `cvm_permutation_test(ctrl_group, trt1_group, 999)`
  - $W^2 = 0.2425 \rightarrow$ sembra un risultato poco promettente
  - $\text{p-value} = 0.211 > \alpha \rightarrow$ accetto l'ipotesi nulla
- `cvm_permutation_test(ctrl_group, trt2_group, 999)`
  - $W^2 = 0.395 \rightarrow$ sembra un risultato poco promettente

    - p-value $= 0.095 > \alpha \rightarrow$ accetto l'ipotesi nulla
- `cvm_permutation_test(trt1_group, trt2_group, 999)`
    - $W^2 = 0.875 \rightarrow$ sembra un risultato poco promettente
    - p-value $= 0.004 \leq \alpha \rightarrow$ rifiuto l'ipotesi nulla e accetto H_1.

Commento

La crescita delle piante in situazione di controllo risulta simile a quella ottenuta con il trattamento 1 e col trattamento 2 anche se con evidente maggiore distanza da parte del secondo trattamento. Essendo il trattamento di controllo il riferimento per una buona crescita delle piante io mi sentirei di dire che il primo trattamento è quello che conviene indagare meglio perché più promettente avendo una distanza minore. I risultati di `trt2_group` però non sono da buttare e non sarebbe sciocco pensare di indagare anche quel trattamento ma va sottolineata la vicinanza al rifiutare l'ipotesi nulla.
La misura della distanza fra i due trattamenti suggerisce che il modo in cui deviano rispetto al campione di controllo è differente.