Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica
Curriculum Data Science e Machine Learning

Elaborato progetto di Statistica e Analisi dei Dati

Candidati

Dott.ssa **Grazia Margarella**
Matricola: 0522501448

Dott. **Nicola Pio Santorsa**
Matricola: 0522501434

Anno Accademico 2022-2023

Sommario

In questo report si andrà a descrivere l'elaborato prodotto nel contesto del corso di Statistica e Analisi dei Dati.
Esso è suddiviso in due parti, la prima relativa allo sviluppo di un elaborato nell'ambito della statistica descrittiva, mentre la seconda nell'ambito della statistica inferenziale.
Per quanto riguarda la statistica descrittiva sono stati analizzati i dati relativi alla produzione energetica dei paesi appartenenti all'Unione Europea suddivisi per varie fonti come il carbon fossile, le rinnovabili, il petrolio e il nucleare. Le analisi sono state svolte con i metodi della statistica descrittiva univariata, bivariata e l'analisi dei cluster, con l'obiettivo di comprendere in che modo si comportano i vari paesi relativamente a tale fenomeno.
Nel secondo caso sono stati descritti i problemi dell'analisi inferenziale tramite la variabile aleatoria di Poisson.
In particolare si è scelta questa variabile aleatoria per descrivere il numero di goal effettuati per partita durante i mondiali di calcio, utilizzando come campione le partite del mondiale del 2022.
Per questa sezione sono state utilizzate le funzioni di R per derivare la funzione di distribuzione e la legge di probabilità, calcolarne i quantili ed effettuare procedure come la stima puntuale, la stima per intervalli di fiducia approssimati, la verifica delle ipotesi e il test del chi-quadrato.

Indice

1 Statistica Descrittiva: Introduzione

1.1 Descrizione Dataset Utilizzato

La versione originale del dataset utilizzato per le analisi dei successivi capitoli di questo elaborato proviene dal sito dell'ente europeo della statistica Eurostat[1].
Questo contiene un insieme di dati che descrivono il bilancio relativo alla produzione energetica di tutti i paesi appartenenti all'Unione Europea (EU) per gli ultimi tre anni. Per effettuare un analisi compatibile con i requisiti del corso si è optato per considerare il solo anno 2020.
Per ogni stato sono state analizzate soltanto le produzioni energetiche relative alle seguenti fonti:

- **Solid Fossils Fuel**: Carburanti a combustibili fossili;
- **Natural Gas**: Gas Naturale;
- **Oil and Petroleum Products**: Prodotti derivanti dal Petrolio;
- **Renewable and Biofuel**: Fonti rinnovabili e Carburanti Bio;
- **Non-Renewable Waste**: Scorie non rinnovabili (come minerali, fossili, derivati animali ecc.);
- **Nuclear Heat**: Energia Nucleare.

Ogni valore qui sopra descritto ha come unità di misura *"tonnellate di litri di petrolio"* che sono direttamente proporzionali ai *Giga Watt-ora (GWh)* con una relazione di 1 a 1.
Nel dataset inoltre sarà inserita un'ulteriore informazione, l'**abbreviazione del paese**, ossia una sigla di 2 lettere definita da una convenzione europea per identificare univocamente ogni paese.
Di seguito è riportata la tabella 1.1 in cui sono descritti i dati utilizzati nel corso dell'analisi.

[1] Eurostat Simplified energy balances

Paese	Abbreviazioni	Solid fossil fuels	Natural gas	Oil and petroleum
Belgio	BE	2.351,567	15.181,631	26.044,676
Bulgaria	BG	4.282,891	2.515,152	4.397,124
Repubblica Ceca	CZ	12.188,634	7.276,103	8.609,930
Danimarca	DK	712,250	2.112,236	6.359,524
Germania	DE	44.595,681	74.599,950	100.784,472
Estonia	EE	0,000	369,776	251,035
Irlanda	IE	445,757	4.554,664	6.372,414
Grecia	GR	1.830,920	4.928,498	11.194,104
Spagna	ES	3.099,844	27.936,527	52.046,121
Francia	FR	5.298,894	34.894,929	65.921,054
Croazia	HR	361,191	2.525,202	2.807,396
Italia	IT	5.094,549	58.285,821	47.350,988
Cipro	CY	14,004	0,000	2.231,709
Lettonia	LV	22,966	910,260	1.632,842
Lituania	LT	134,591	1.971,754	3.065,392
Lussemburgo	LU	38,442	621,269	2.393,290
Ungheria	HU	1.683,004	8.764,381	7.472,440
Malta	MT	0,000	318,243	2.535,858
Paesi Bassi	NL	4.110,084	31.551,702	39.316,758
Austria	AT	2.471,505	7.282,584	11.114,261
Polonia	PL	40.914,707	17.440,338	29.737,980
Portogallo	PT	565,735	5.190,543	9.635,067
Romania	RO	3.481,925	9.681,938	9.675,485
Slovenia	SI	1.018,083	735,587	2.126,39 4
Slovacchia	SK	2.304,985	4.088,253	3.598,446
Finlandia	FI	1.840,438	2.117,289	8.152,937
Svezia	SE	1.453,487	1.272,915	10.920,026

Paese	Renewables	Non-renewable	Nuclear
Belgio	4.928,258	646,716	8.369,961
Bulgaria	2.550,038	66,158	4.334,677
Repubblica Ceca	5.123,327	368,493	7.496,295
Danimarca	6.216,974	447,190	0,000
Germania	46.937,670	4.189,166	16.576,800
Estonia	1.310,241	34,224	0,000
Irlanda	1.766,433	146,752	0,000
Grecia	3.349,997	9,673	0,000
Spagna	19.094,184	539,722	15.174,000
Francia	28.603,321	1.632,975	92.211,000
Croazia	2.195,322	38,939	0,000
Italia	29.344,686	1.190,111	0,000
Cipro	280,847	34,802	0,000
Lettonia	1.809,280	52,620	0,000
Lituania	1.655,641	58,589	0,000
Lussemburgo	396,525	44,739	0,000
Ungheria	2.965,021	211,068	4.053,000
Malta	57,243	0,000	0,000
Paesi Bassi	7.016,951	790,343	955,890
Austria	10.498,759	668,176	0,000
Polonia	12.950,824	1.069,116	0,000
Portogallo	6.369,492	187,125	0,000
Romania	5.989,202	281,648	2.887,000
Slovenia	1.176,231	57,197	1.496,852
Slovacchia	2.147,033	234,833	4.044,000
Finlandia	12.017,328	293,709	5.547,600
Svezia	23.209,504	989,491	12.028,000

1.2 Definizione Dataset In Environment R

Per definire il dataset nell'environment di R utilizziamo un pacchetto open-source chiamato **"readxl"** che consente di creare facilmente un dataframe[2] a partire da un file con un formato *.xlsx* tipico di programmi come Excel.
Installiamo questo pacchetto su R utilizzando il comando **install.packages('readxl')** e, dopo esser terminata l'installazione, carichiamo la libreria con il comando **library('readxl')**.
Dopo aver installato e caricato la libreria ci basterà chiamare il metodo **read_excel()**, passandogli come argomento il nome del file, per avere il dataframe con i dati desiderati.

[2]Un dataframe è una lista di vettori (le variabili), che devono avere tutti la stessa lunghezza (numero di casi), ma possono essere di tipo diverso: variabili nominali (fattori), variabili cardinali (vettori numerici), ecc.

```
DataFramePaesi<-read_excel('nomefile.xlsx')
# OPPURE
DataFramePaesi<-read_excel(file.choose())
```

1.2.1 Ordinamento delle Righe

Il dataset risulta non avere un ordine ben definito al suo stato originale, motivo per cui abbiamo deciso di effettuare un ordinamento alfabetico basato sulla colonna delle abbreviazioni.
La scelta è ricaduta su questa colonna e non su quella del nome dal momento che, come detto già in precedenza, le varie abbreviazioni sono state definite da una convenzione e quindi immutate al cambiamento di lingua, problema riscontrato dal passaggio dal dataset originale alle successive elaborazioni.
L'ordinamento così descritto è stato effettuato tramite questo comando:

```
DataFramePaesi<-DataFramePaesi[order(DataFramePaesi$Abbreviazioni),]}
```

1.2.2 Normalizzazione dei Dati

L'ultima operazione che effettueremo sul dataset, prima di poter iniziare ad effettuare le analisi, sarà la normalizzazione dei dati riducendoli in percentuali.
Questa scelta è stata presa considerando il fatto che per alcune analisi che effettueremo qui in seguito numeri grandi, come quelli che abbiamo in tabella, causerebbero problemi di visualizzazioni e di calcoli approssimati.
Il metodo di normalizzazione che abbiamo utilizzato consiste nel calcolo della percentuale considerando il contributo del singolo valore sulla colonna.
Il codice che abbiamo utilizzato per effettuare questa trasformazione è il seguente:

```
for (i in 4:9){
    somma<-sum(DataFramePaesi[1:27,i])
    DataFramePaesi[1:27,i]<-(DataFramePaesi[1:27,i]/sum)*100
}
```

2 Analisi Esplorativa

2.1 Scelta della Suddivisione In Classi

Quando si fa un analisi di questo tipo la scelta che viene presa da ogni analista è quella di suddividere i dati in classi.
Ma come definiamo queste classi? Con quale criterio?
Ricordandoci che i nostri dati sono percentuali potremmo pensare di cercare il massimo valore nei nostri dati e suddividere in n classi da 0-10% , 10-20% ecc.
Teoricamente non c'è niente di sbagliato in questo ragionamento, però se si vanno a vedere i dati ci balza subito all'occhio che i dati non sarebbero distribuiti uniformemente in tutte le n classi poiché più di $\frac{3}{4}$ dei dati risiedono nella classe 0-10% come mostra anche la seguente funzione :

```
table(cut(DataFramePaesi[ ,3:8],c(0,10,20,30,40,50,60),right=FALSE))
```

[**0,10**)	[**10,20**)	[**20,30**)	[**30,40**)	[**40,50**)	[**50,60**)
148	8	4	1	0	1

Con questi risultati alla mano quindi abbiamo deciso di virare su una suddivisione in classi *asimmetrica*, ovvero una suddivisione dove gli l'intervalli di ogni classe non si equivalgono.
La suddivisione in classi scelta sarà quindi :

$$Classi : \{\ [0,1)\ ,\ [1,5)\ ,[5,10)\ ,[10,20)\ ,[20,30)\ ,[40,60)\ \}$$

Per confrontare questa suddivisione con la precedente, richiamiamo la stessa funzione e confrontiamo i valori per ogni classe :

```
table(cut(DataFramePaesi[ ,3:8],c(0,1,5,10,20,30,40,60),right=FALSE))
```

[**0,1**)	[**1,5**)	[**5,10**)	[**10,20**)	[**20,30**)	[**30,40**)	[**40,60**)
69	60	19	8	4	1	1

2.2 Frequenze Assolute e Relative

2.2.1 Definizione

Sia $X = (x_1, x_2, ..., x_n)$ un campione con n osservazioni, ognuna delle quali può assumere valori $(z_1, z_2, ..., z_k)$, si definisce ***distribuzione di frequenze assolute*** l'insieme:

$$D_a := \{n_i : n_i = numero\ di\ occorrenze \text{ di } z_i \text{ in X } \}$$

mentre si definisce ***distribuzione di frequenze relative*** l'insieme:

$$D_r := \{\ f_i : f_i = n_i/n\ \}$$

Per ottenere i grafici delle frequenze assolute in R si esegue il comando *table* su ogni colonna e si graficano i risultati, mentre per quanto riguarda le relative, si esegue sui risultati precedenti la divisione per la lunghezza della colonna, per ottenere il contributo percentuale delle frequenze.

2.2.2 Produzione energetica Solid Fossils Fuels

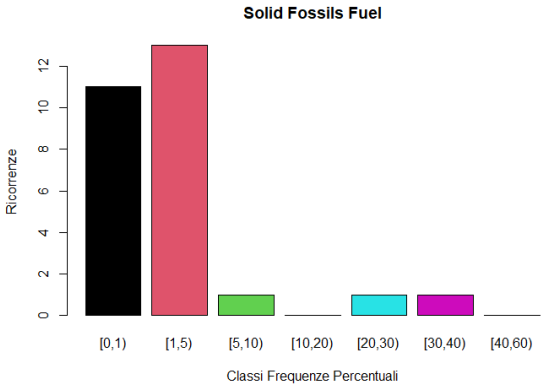


Figura 1: Frequenze assolute

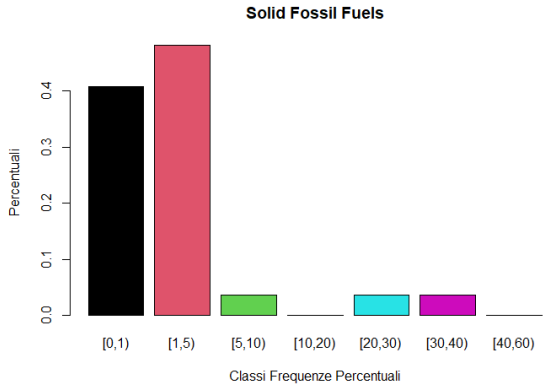


Figura 2: Frequenze relative

Frequenze	[**0,1**)	[**1,5**)	[**5,10**)	[**10,20**)	[**20,30**)	[**30,40**)	[**40,60**)
Assolute	11	13	1	0	1	1	0
Relative	0.41	0.48	0.04	0	0.04	0.04	0

Dai grafici vediamo che :

- Sono principalmente tre i paesi "leader" in Europa della produzione energetica tramite combustibili fossili : La **Germania** e la **Polonia** con le percentuali più alte (rispettivamente 31% e 29%) e la **Repubblica Ceca** che si assesta intorno all'8%;
- Solamente due Paesi in tutta Europa non producono energia tramite i combustibili fossili e sono l'**Estonia** e **Malta**;
- I restanti paesi oscillano tra lo 0 e il 3% .

2.2.3 Produzione energetica Natural Gas

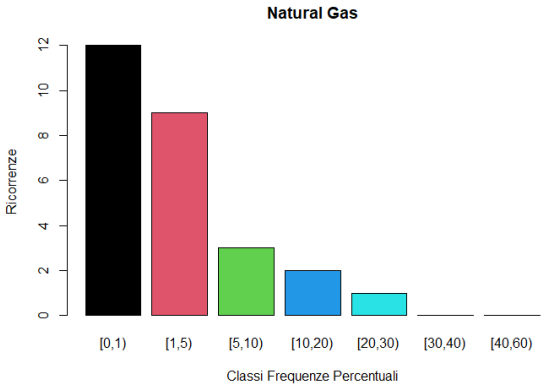


Figura 3: Frequenze assolute

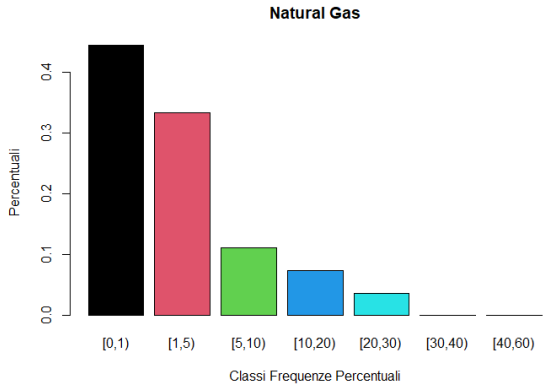


Figura 4: Frequenze relative

Frequenze	[**0,1**)	[**1,5**)	[**5,10**)	[**10,20**)	[**20,30**)	[**30,40**)	[**40,60**)
Assolute	12	9	3	2	1	0	0
Relative	0.44	0.33	0.11	0.07	0.04	0	0

Dai grafici vediamo che :

- I paesi che producono energia tramite il gas naturale in maggior quantità sono la **Germania** al 22% , l'**Italia** al 17%,la **Francia** al 10% e l'**Olanda** quasi al 10% ;
- Solamente un solo paese in tutta l'Europa non produce energia tramite gas naturale e questo è **Cipro**;
- Per i restanti paesi non citati, solamente la **Spagna** e la **Polonia** contribuiscono in maniera corposa con rispettivamente l'8 e il 5% , i restanti paesi contribuiscono in maniera molto minore.

2.2.4 Produzione energetica Oil and petroleum

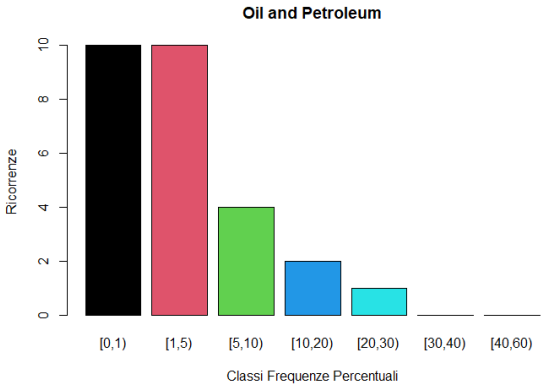


Figura 5: Frequenze assolute

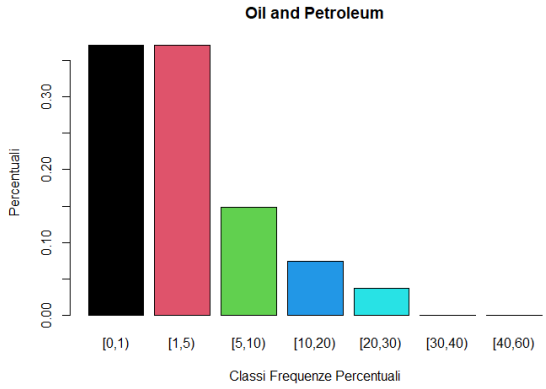


Figura 6: Frequenze relative

Frequenze	[**0,1**)	[**1,5**)	[**5,10**)	[**10,20**)	[**20,30**)	[**30,40**)	[**40,60**)
Assolute	10	10	4	2	1	0	0
Relative	0.37	0.37	0.15	0.07	0.04	0	0

Dai grafici vediamo che :

- Il paese leader nella produzione di energia dal petrolio e derivati è la **Germania** con il 21%, le seguono la **Francia** con l'13% e la **Spagna** con l' 11%;
- Non esistono paesi che non producono energia tramite il petrolio, però quelli che ne contribuiscono dallo 0 al 5% al bilancio energetico europeo tramite questa fonte sono 20;
- I paesi che contribuiscono mediamente con una percentuale compresa tra il 5 e il 10% sono 4 e sono **Belgio** con il 5% , l'**Italia** con il quasi 10% , l'**Olanda** con l'8% e infine la **Polonia** con il 6%.

2.2.5 Produzione energetica Renewable and biofuel

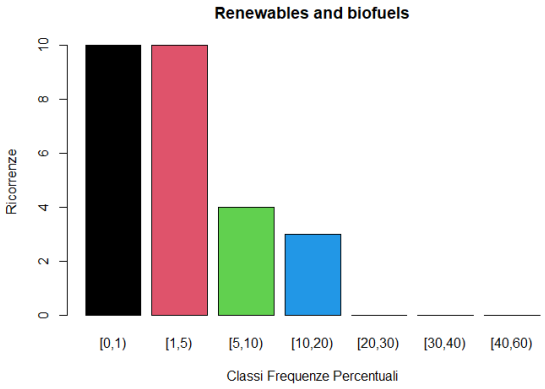


Figura 7: Frequenze assolute

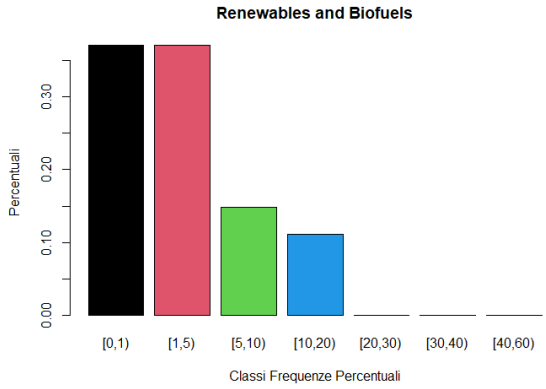


Figura 8: Frequenze relative

Frequenze	**[0,1)**	**[1,5)**	**[5,10)**	**[10,20)**	**[20,30)**	**[30,40)**	**[40,60)**
Assolute	10	10	4	3	0	0	0
Relative	0.37	0.37	0.15	0.11	0	0	0

Dai grafici vediamo che :

- Ci sono principalmente 3 paesi che contribuiscono maggiormente all'energia prodotta tramite fonti rinnovabili e biofuel e questi sono : La **Germania** con una percentuale del 19% , l'**Italia** con una percentuale del 12% e infine la **Francia** con una percentuale dell'11%;

- Non ci sono paesi che non producono energia tramite fonti rinnovabili però , allo stesso modo del caso precedente , sono 20 i paesi che contribuiscono tra lo 0 e il 5% ;

- Sono 4 i paesi che contribuiscono mediamente al bilancio energetico europeo per le fonti rinnovabili e questi paesi sono : La **Svezia** con quasi il 10%, la **Spagna** con l'8% e infine **Polonia** e **Finlandia** con il 5%.

2.2.6 Produzione energetica Non-Renewable waste

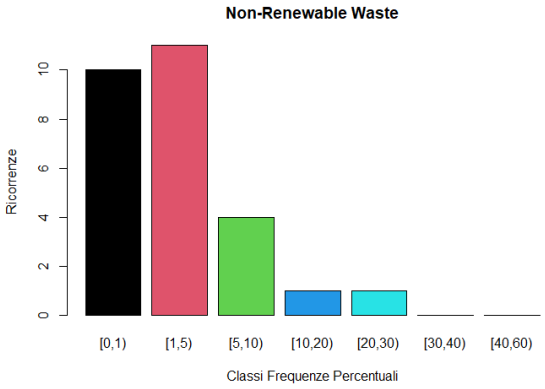


Figura 9: Frequenze assolute

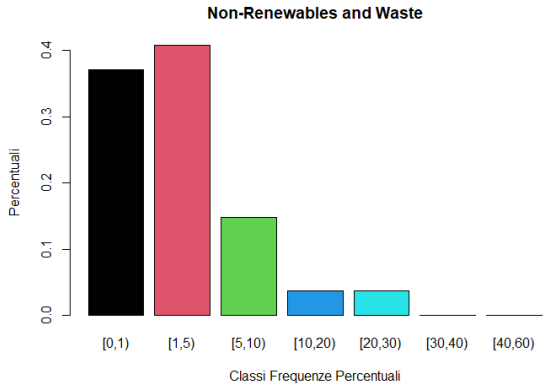


Figura 10: Frequenze relative

Frequenze	**[0,1)**	**[1,5)**	**[5,10)**	**[10,20)**	**[20,30)**	**[30,40)**	**[40,60)**
Assolute	10	11	4	1	1	0	0
Relative	0.37	0.41	0.15	0.04	0.04	0	0

Dai grafici vediamo che :

- Sono principalmente due i paesi "leader" nella generazione di energia tramite questa fonte e sono la **Germania** che contribuisce al 29% e la **Francia** che contribuisce 11%;
- Un solo paese non contribuisce al bilancio energetico europeo con questa fonte energetica ed è **Malta** , oltre questo paese , ci sono altri 20 che contribuiscono minormente con una percentuale compresa tra lo 0 e 5%;
- Sono 4 i paesi che contribuiscono mediamente al bilancio energetico e sono : l'**Italia** con l' 8% , la **Polonia** e la **Svezia** con il 7% e infine l'**Olanda** con il 5%.

2.2.7 Produzione energetica Nuclear Heat

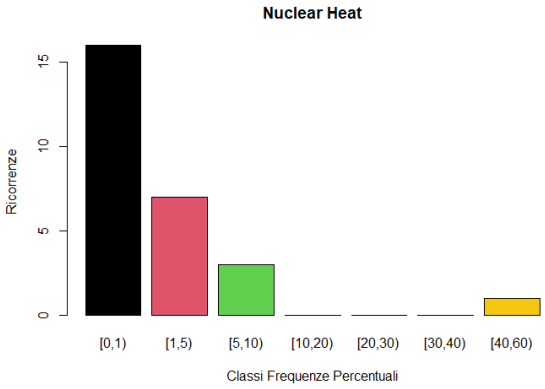


Figura 11: Frequenze assolute

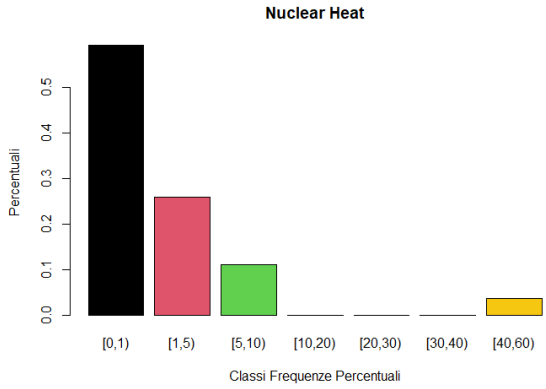


Figura 12: Frequenze relative

Frequenze	[0,1)	[1,5)	[5,10)	[10,20)	[20,30)	[30,40)	[40,60)
Assolute	16	7	3	0	0	0	1
Relative	0.59	0.26	0.11	0	0	0	0.04

Dai grafici vediamo che :

- Sono ben 14 i paesi che non producono energia tramite il nucleare e ben 9 i paesi che contribuiscono minormente con una percentuale compresa tra lo 0 e 5%;
- Sono 3 i paesi che contribuiscono mediamente al bilancio energetico europeo con questa fonte di energia e sono : **Germania** con lo 9% , **Spagna** con l'8% e infine la **Svezia** con il 7%;
- Unico paese leader nell'ambito nucleare è la **Francia** con oltre il 52% !.

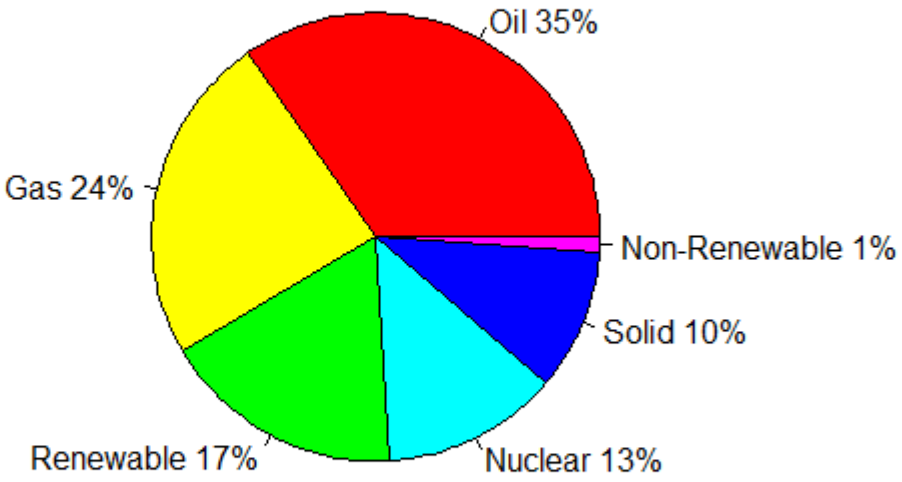
Curiosità, la **Francia** è l'unico paese al mondo che modula i propri reattori nucleari per diminuire e aumentare l'energia prodotta poiché secondo una statistica l'energia prodotta da tutti i suoi 58 reattori nucleari riesce a coprire circa il 70% del fabbisogno energetico di tutto il paese, motivo per cui anche la Francia è il primo esportatore in Europa di energia elettrica.[3]

2.3 Contributi Fonti Energetiche sul Bilancio Energetico

Per comprendere al meglio il contributo delle singole fonti energetiche all'interno della produzione energetica europea è stato realizzato un diagramma a torta.

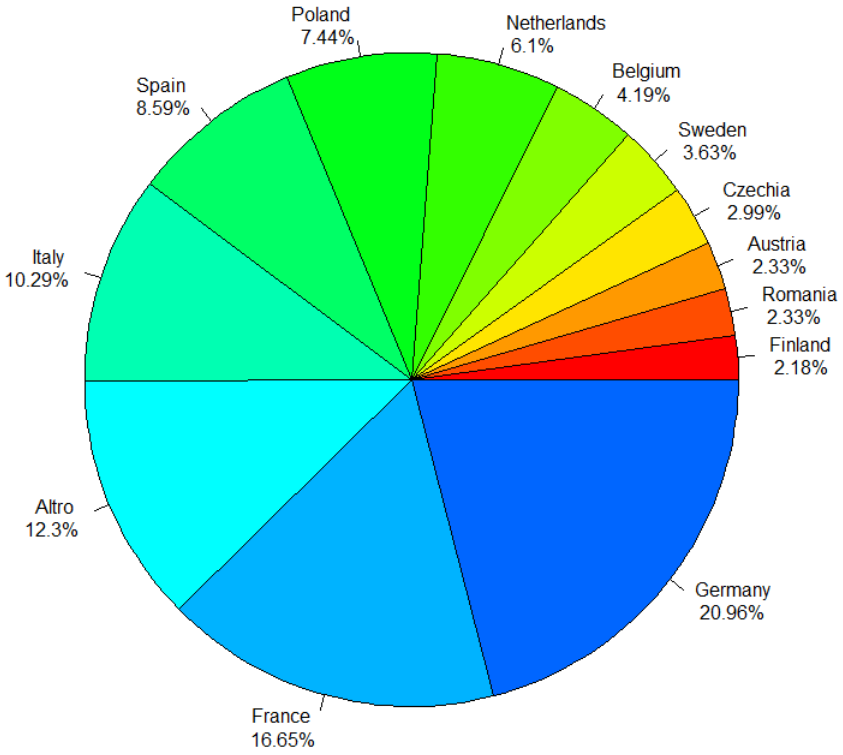
[3]Fonte: Geopop, canale youtube che si occupa di divulgazione scientifica.
Qui per il video in questione

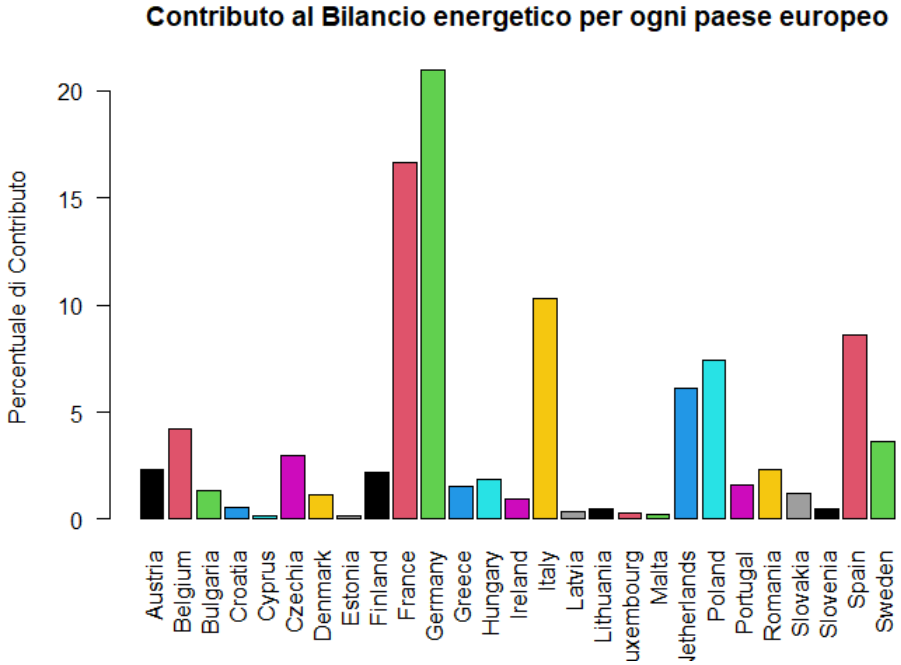
I dati sono il risultato della somma per colonne della matrice i cui valori sono stati divisi per il totale della produzione.



2.4 Contributi Singoli Paesi sul Bilancio Energetico

Altro tipo di dato da voler analizzare è il contributo per paese alla produzione totale. Una prima visualizzazione è stata effettuata con un diagramma a torta, ma questo tipo di diagrammi non è estremamente leggibile.
In questo caso risulta più efficace la costruzione di un barplot. Per semplicità mostreremo entrambi.


Poland
7.44%
Netherlands
6.1%
Belgium
4.19%
Sweden
3.63%
Czechia
2.99%
Austria
2.33%
Romania
2.33%
Finland
2.18%
Germany
20.96%
France
16.65%
Altro
12.3%
Italy
10.29%
Spain
8.59%


Contributo al Bilancio energetico per ogni paese europeo
Percentuale di Contributo
0
5
10
15
20
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Ireland
Italy
Latvia
Lithuania
uxembourg
Malta
letherlands
Poland
Portugal
Romania
Slovakia
Slovenia
Spain
Sweden

3 Analisi descrittiva univariata

Procediamo dunque nell'analisi descrittiva dei dati precedentemente introdotti. Attraverso i metodi della statistica descrittiva univariata vengono indagati i comportamenti dei fenomeni che i dati rappresentano. Questi dati possono essere di tipo qualitativo o quantitativo ed in particolare discreti o continui. Nel caso preso in esame tratteremo dati quantitativi continui. Procediamo quindi nell'indagare le proprietà delle diverse colonne del nostro dataset.

3.1 Solid Fossil Fuels

Una delle prime analisi effettuabile è quella data dalla definizione della funzione di distribuzione empirica. Essa può essere di due tipi a seconda del tipo di dato considerato.

3.1.1 Funzione Distribuzione Empirica Discreta

Data una variabile X che assume k valori distinti $z_1, z_2, \ldots, z_k$ ordinati in ordine crescente e un campione $(x_1, x_2, \ldots, x_n)$ di n osservazioni di X. Denotiamo con n_i la frequenza assoluta per ciascun valore di z_k e con $f_i = \frac{n_i}{n}$ la frequenza relativa. Le frequenze relative cumulate sono definite come

$$F_i = f_1 + f_2 + \cdots + f_i = \frac{n_1 + n_2 + \cdots + n_i}{n}$$

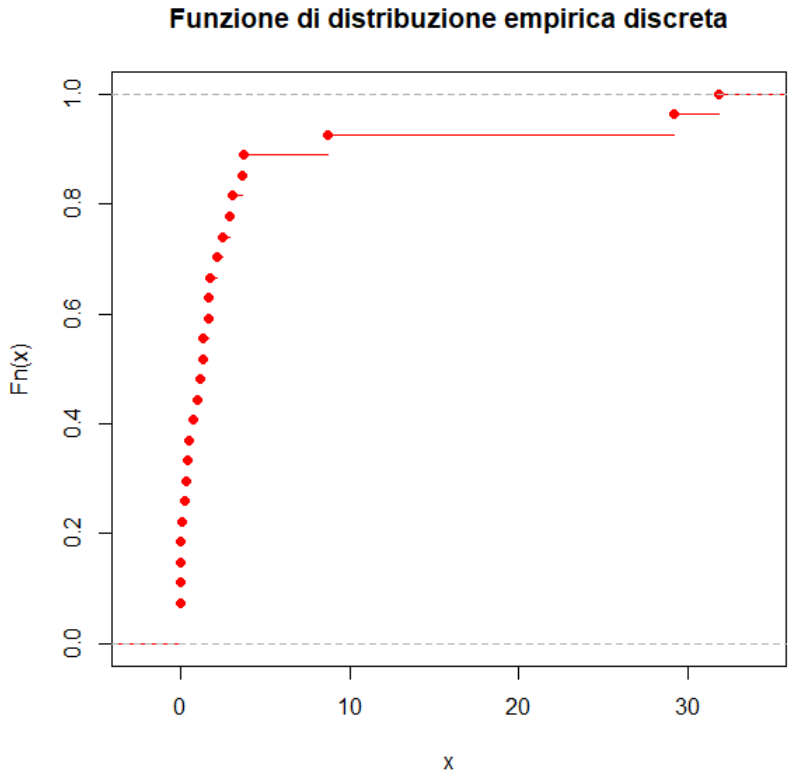
$$(i = 1, 2, \ldots, k)$$

dove F_i rappresenta la proporzione dei dati del campione minori o uguali a z_i.

Supponendo che i k valori distinti assunti dalla variabile X siano ordinati in ordine crescente, la funzione di distribuzione empirica $F(X)$ è definita come segue:

$$F(x) = \frac{\#x_i \leq x, i = 1, 2, \ldots, n}{n} = \begin{cases} 0 & x < z_1 \\ F_1 & z_1 \leq x < z_2 \\ \ldots & \\ F_i & z_i \leq x < z_{i+1} \\ \ldots & \\ 1 & x \geq z_k \end{cases}$$

Il grafico risultante da questa funzione è crescente a gradini dove la funzione assume il valore a sinistra in corrispondenza ad ogni punto di salto e nel caso il valore sia minore o uguale dell'osservazione minima è uguale a 0 e 1 nel caso sia maggiore o uguale dell'osservazione massima.
In R è derivabile tramite la funzione **ecdf()** (*empirical cumulative distribution function*). Nel caso del Solid Fossil Fuel la funzione risulta essere la seguente.



3.1.2 Funzione Distribuzione Empirica Continua

Per quanto riguarda i fenomeni quantitativi continui occorre considerare la funzione di distribuzione empirica continua. Questo tipo di funzione è strutturata in k classi $C_1 = [z_0, z_1), C_2 = [z_1, z_2), \ldots, C_k = [z_{k-1}, z_k]$ con $z_0 < z_1 < \cdots < z_{k-1} < z_k$ dove z_0 corrisponde al minimo delle osservazioni e z_k al massimo delle osservazioni.
La **funzione di distribuzione empirica continua** è così definita:

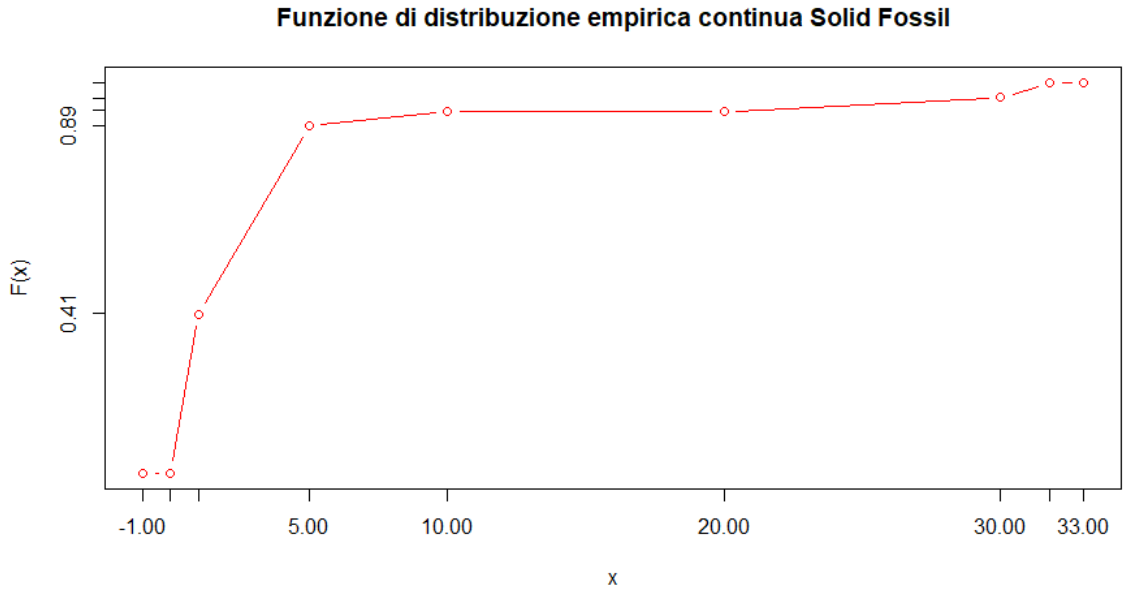
$$F(x) := \begin{cases} 0 & \text{x} < z_0 \\ \ldots & \\ F_{i-1} & \text{x} = z_{i-1} \\ \dfrac{F_i - F_{i-1}}{z_i - z_{i-1}} x + \dfrac{z_i F_{i-1} - z_{i-1} F_i}{z_i - z_{i-1}} & z_{i-1} < x < z_i \\ F_i & x = z_i \\ \ldots & \\ 1 & x \geq z_k \end{cases}$$

Essa infatti coincide con il segmento che passa per i punti (z_{i-1}, F_{i-1}) e (z_i, F_i).

Andiamo a produrre dunque con il seguente codice il grafico relativo alla colonna Solid Fossil Fuel.

```
> freqrel <- table(X$`Solid fossil fuels`)/length(X$`Solid fossil
    fuels`)
> classi <- c(0,1,5,10,20,30,32)
> frelclassi <- table(cut(X$`Solid fossil fuels`, breaks = classi,
    right=FALSE))/length(X$`Solid fossil fuels`)
> Fcum <- cumsum(frelclassi)
```

I valori delle classi sono dunque selezionati come ascisse e i valori delle frequenze cumulate (descritte nella tabella successiva) come ordinate per poi proseguire con il plot del grafico risultante.



[0,1)	**[1,5)**	**[5,10)**	**[10,20)**	**[20,30)**	**[30,31.78]**
0.41	0.89	0.93	0.93	0.96	1

Come possiamo notare nel grafico i valori sono maggiormente distribuiti nelle prime due classi, ovvero quasi il 90% dei valori è compreso tra 0 e 5, come già evidenziato dalle precedenti analisi. Nelle classi successive, il grafico evidenzia una crescita minore dovuta alla ridotta quantità di dati in quegli intervalli. Nel caso in cui la linea è orizzontale notiamo che non ci sono valori che contribuiscono alla somma cumulata delle frequenze.

3.1.3 Indici di sintesi

Queste statistiche sono utili a descrivere i dati numerici. In particolare verranno descritti i seguenti indici con le relative definizioni:

- *indici di posizione centrali* come media campionaria, mediana campionaria e moda campionaria;
- *indici di posizione non centrali* come i quantili e quartili;
- *indici di dispersione* quali la varianza campionaria, deviazione standard campionaria e coefficiente di variazione;
- *indici di simmetria* come la skewness campionaria e la curtosi campionaria.

3.1.4 Media campionaria, mediana campionaria e moda campionaria

Definiamo **media campionaria** di $X = (x_1, x_2, ..., x_n)$ la media aritmetica degli elementi del campione x_i con i che varia tra 1 e n:

$$\overline{x} := \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Supponendo che il campione $x = (x_1, x_2, ..., x_n)$ sia ordinato in modo crescente, definiamo **mediana campionaria** il valore centrale dell'ordinamento:

$$Me := \begin{cases} x_{(n+1)/2} & \text{se n è dispari} \\ (x_{n/2} + x_{n/2+1})/2 & \text{se n è pari} \end{cases}$$

Definiamo **moda campionaria** la modalità a cui è associata la frequenza assoluta (o relativa) più alta. Se esistono più modalità con frequenza massima ciascuna è un valore modale. Nel caso di variabili quantitative continue si parla di *classe modale*, ossia la classe con la massima densità di frequenza.

3.1.5 Varianza, deviazione standard e coefficiente di variazione campionari

Definiamo dunque gli indici di dispersione che descrivono quanto i dati si differenziano dagli indici di centralità come la media. In dettaglio sono definiti **varianza campionaria, deviazione standard campionaria e coefficiente di variazione campionario** che sono rispettivamente uguali ai seguenti valori:

$$s_x^2 := \frac{1}{n-1} \cdot \sum_{i}^{n} (x_i - (\overline{x}))^2 \quad \textit{Varianza Campionaria}$$

$$s_x := \sqrt{s_x^2} \quad \textit{Deviazione Standard Campionaria}$$

$$cv_x := \frac{s_x}{|\overline{x}|} \quad \textit{Coefficiente di Variazione Campionario}$$

Essi sono calcolati in R tramite le funzioni var() per la varianza, sd() per la deviazione standard e il coefficiente di variazione tramite la seguente funzione.

```
> cv <- function (x){sd(x)/abs ( mean(x))}
```

La varianza e la deviazione standard dipendono dall'unità di misura dei dati, mentre il coefficiente di variazione è adimensionale. La statistica più utilizzata su campioni con la stessa unità di misura è la deviazione standard per la sua più semplice interpretabilità.

3.1.6 Skewness

Un indice che permette di misurare la simmetria di una distribuzione di frequenze è la **skewness campionaria (coefficiente di simmetria)**.
Assegnato un insieme di dati numerici $x_1, x_2, ..., x_n$, si definisce **skewness** campionaria il valore:

$$\gamma_1 = \frac{m_3}{m_2^{3/2}}$$

con m_3 viene denotato il momento centrato campionario di ordine 3.
Il generico momento di ordine j è definibile tramite la seguente formula:

$$m_j = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^j$$

In base al valore della skewness possiamo definire che:

- con $\gamma_1 = 0$, la distribuzione di frequenza è detta **simmetrica**
- con $\gamma_1 > 0$, la distribuzione di frequenza è detta **simmetrica positiva** (distribuzione ha coda di destra più allungata)
- con $\gamma_1 < 0$, la distribuzione di frequenza è detta **simmetrica negativa** (distribuzione ha coda di sinistra più allungata)

3.1.7 Curtosi

Un indice che permette di misurare la densità dei dati intorno alla media è la **curtosi campionaria**.
Assegnato un insieme di dati numerici $x_1, x_2, ..., x_n$, si definisce **curtosi campionaria** il valore:

$$\gamma_2 = \beta_2 - 3$$

con

$$\beta_2 = \frac{m_4}{m_2^2}$$

detto *indice di Pearson*.
Gli indici γ_2 e β_2 permettono di confrontare la distribuzione di frequenze dei dati con una densità di probabilità normale standard, caratterizzata da $\beta_2 = 3$ e indice di curtosi $\gamma_2 = 0$.
Se risulta :

- $\beta_2 < 3(\gamma_2 < 0)$: la distribuzione di frequenze si definisce **platicurtica**, ossia la distribuzione di frequenze è più piatta di una normale;
- $\beta_2 > 3(\gamma_2 > 0)$: la distribuzione di frequenze si definisce **leptocurtica**, ossia la distribuzione di frequenze è più piccata di una normale;
- $\beta_2 = 3(\gamma_2 = 0)$: la distribuzione di frequenze si definisce **normocurtica** (mesocurtica), ossia piatta come una normale.

Il calcolo della curtosi campionaria ha significato soltanto per distribuzioni di frequenze unimodali, dato che tale indice è confrontato con quello di una normale standard.
Nel caso del Solid Fossil Fuel otteniamo i seguenti risultati:

Indice di Sintesi	**Valore**
Media Campionaria	3.7037
Mediana Campionaria	1.3048
Classe Modale	[1,5)
Varianza Campionaria	62.9804
Deviazione Standard Campionaria	7.9360
Coefficiente di Variazione	2.1427
Skewness	2.9947
Curtosi Campionaria	7.5120

Analisi degli indici

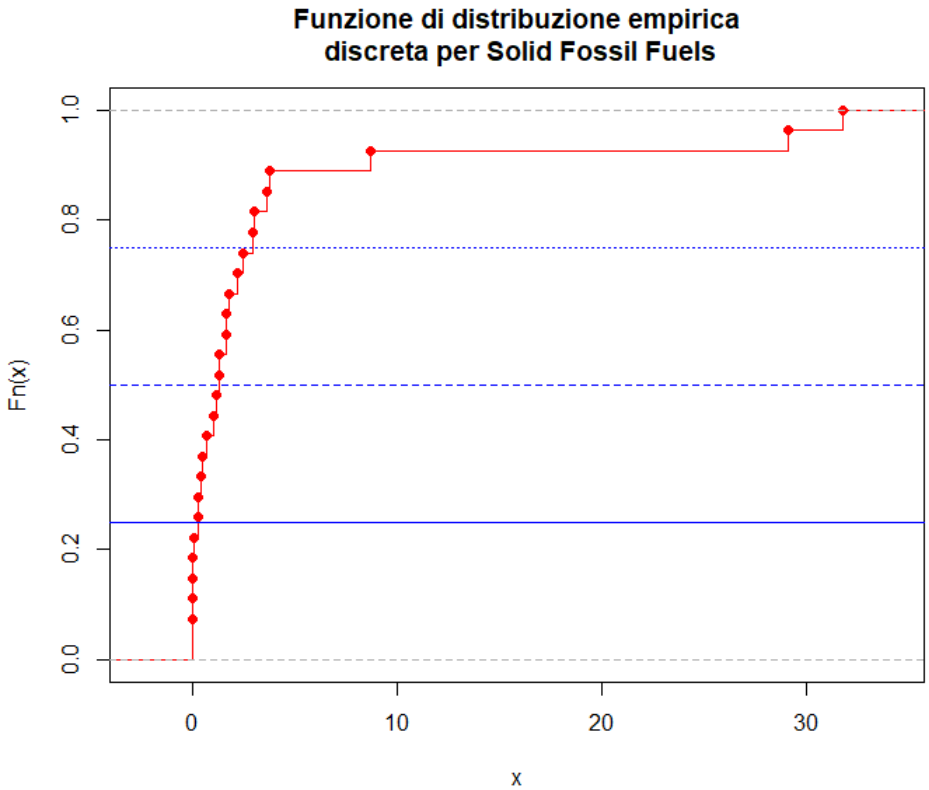
- **Confronto media e mediana** : Come possiamo notare il valore della media risulta essere discostato positivamente dalla mediana, il che ci porta a ipotizzare un'asimmetria positiva;
- **Varianza, deviazione standard e coefficiente di variazione**: Dalla deviazione standard vediamo che i valori si discostano circa dell'8% dalla media campionaria, informazione che è facile estrapolare anche dagli altri 2 valori;
- **Skewness** : Il valore della skewness risulta essere > 0 , il che va a confermare la nostra ipotesi di *asimmetria positiva* ;
- **Curtosi** : dal momento che il valore della curtosi è positiva, possiamo dire che la distribuzione di frequenze sia *leptocurtica*, ossia è più piccata di una normale.

3.1.8 Quantili

Sia X una variabile quantitativa e sia $x_1, x_2, ..., x_n$ un campione di n osservazioni disposte in ordine crescente.
Supponiamo di suddividere i dati ordinati in α gruppi, ognuno dei quali contenga (circa) lo stesso numero di osservazioni; gli α - 1 numeri che consentono tale suddivisione sono i **quantili** di ordine α. Nel caso in cui α sia uguale a 4 otteniamo i quartili. In R viene utilizzata la funzione **quantile(v, probs=, type=)** dove v è il vettore dei dati, probs il vettore delle probabilità da calcolare e type il tipo di algoritmo da utilizzare per calcolare i quantili.

L'algoritmo di default è denotato dal valore 7 ed è basato sull'interpolazione tra punti e non restituisce sempre valori compresi nel campione.

Altro algoritmo che non ha questo problema è dato dall'algoritmo di tipo 1 basato sulla distribuzione di frequenze. Esso si ottiene ordinando le frequenze relative e considerando il valore per cui $F_{i-1} < p$ e $F_i \geq p$. Ciò è facilmente deducibile dal grafico della distribuzione empirica discreta riportato di seguito utilizzando la colonna Solid Fossil Fuel nel quale sono stati evidenziati i valori dei quartili.



Nel caso del Solid Fossil Fuel i quartili sono nella seguente tabella e sono rappresentati nella prossima sezione tramite boxplot.

Quartili	*0%*	*25%*	*50%*	*75%*	*100%*
Valori	0.0000	0.2574	1.3048	2.9291	31.7822

3.1.9 BoxPlot

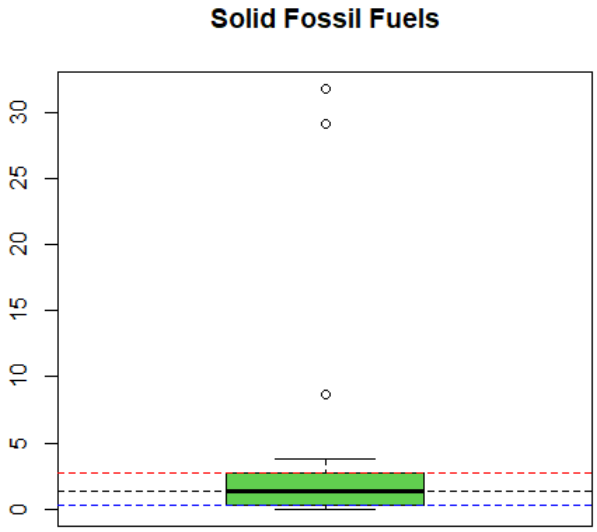
Un boxplot è costituito da una scatola i cui **estremi sono in corrispondenza del valore di Q1 e Q3**, ed è tagliata da una linea orizzontale in corrispondenza di Q2.
Un boxplot presenta anche altre due linee orizzontali, dette **baffi**: per tale motivo il boxplot viene definito anche "**diagramma a scatola e baffi**".
Il baffo inferiore corrisponde al valore più piccolo tra le osservazioni che risulta maggiore o uguale di:

$$a = Q1 - 1.5 * (Q3 - Q1)$$

mentre il baffo superiore corrisponde al valore più grande delle osservazioni che risulta minore o uguale a:

$$b = Q3 + 1.5 * (Q3 - Q1)$$

La distanza tra il primo e il terzo quartile è detta **intervallo interquartile** o **scarto interquartile**.
Nel caso in cui tutti i valori rientrano nell'intervallo (a,b) allora i baffi sono posizionati in corrispondenza del minimo e del massimo del campione.
Se ci sono invece valori che si trovano fuori dall'intervallo (a,b), questi valori vengono evidenziati come punti nel grafico e sono detti **valori anomali** o **outlier**: sono infatti un'anomalia rispetto alla maggioranza delle altre osservazioni del campione e sono pertanto molto interessanti da analizzare.
Attraverso il boxplot ricaviamo molte informazioni: oltre ai già descritti valori anomali, osserviamo la **centralità**, la **dispersione** e la **forma**.
La centralità si deduce dalla mediana.
Dalla forma possiamo verificare se i dati sono simmetrici o meno, questo può essere osservato tramite lo scarto interquantile e la mediana: **se Q3 e Q1 hanno una distanza simile da Q2 allora il nostro set di dati è simmetrico**.
I baffi invece ci danno informazioni sulla **dispersione e la distribuzione dei dati**.
Nel nostro caso studio è stato prodotto per il Solid Fossil Fuel il seguente boxplot.

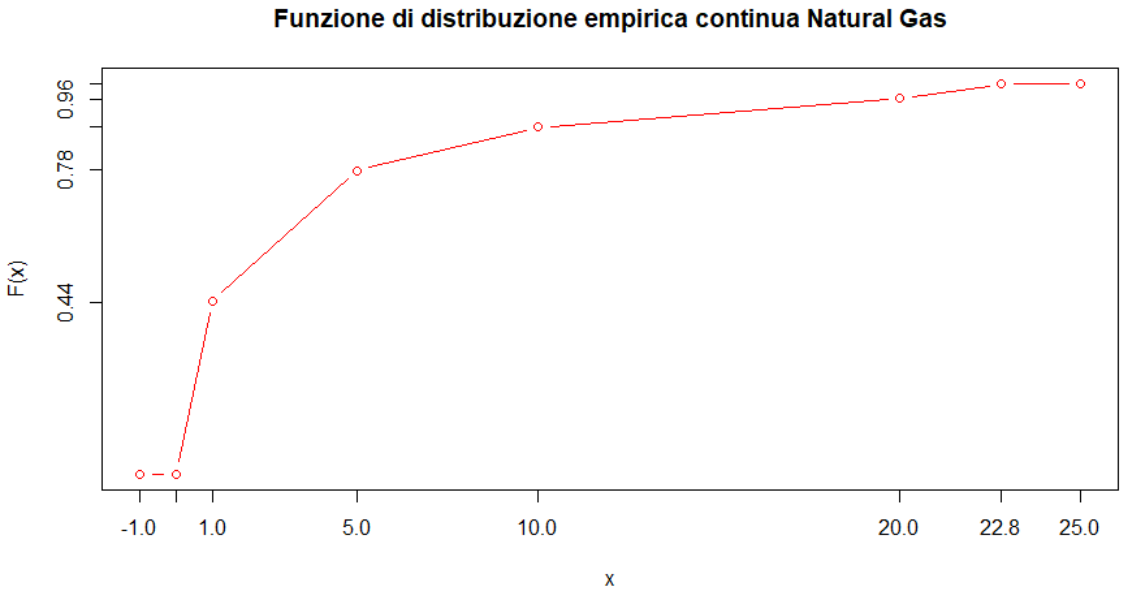


Min	1st Qu.	Median	Mean	3rd Qu.	Max
0	0.29	1.30	3.70	2.70	31.78

Tra tutti i paesi solamente 3 sono outlier (valori anomali) poiché hanno una produzione "fuori dalla media" e questi sono, in ordine dal basso verso l'alto, **Repubblica Ceca** , **Polonia** e **Germania** a conferma di quello che avevamo già analizzato nella distribuzione delle frequenze.

3.2 Natural Gas

3.2.1 Funzione Distribuzione Empirica



[0,1)	[1,5)	[5,10)	[10,20)	[20,22.80]
0.44	0.78	0.89	0.96	1

Come possiamo notare nel seguente grafico i valori sono maggiormente distribuiti nelle prime tre classi, ovvero quasi il 90% dei valori è compreso tra 0 e 10, come già evidenziato dalle precedenti analisi.
Nelle classi successive, il grafico evidenzia una crescita minore dovuta alla ridotta quantità di dati in quegli intervalli.

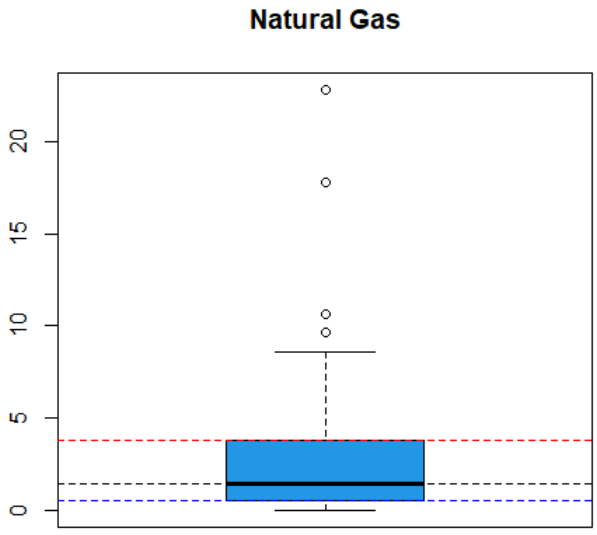
3.2.2 Indici Di Sintesi

Indice di Sintesi	Valore
Media Campionaria	3.7037
Mediana Campionaria	1.3923
Classe Modale	[0,1)
Varianza Campionaria	32.0652
Deviazione Standard Campionaria	5.6626
Coefficiente di Variazione	1.5289
Skewness	2.1471
Curtosi Campionaria	3.9362

Quartili	*0%*	*25%*	*50%*	*75%*	*100%*
Valori	0.0000	0.3891	1.3923	4.6408	22.8045

- **Confronto media e mediana**: Come possiamo notare il valore della media risulta essere discostato positivamente, il che ci porta a ipotizzare un'asimmetria positiva;
- **Varianza, deviazione standard e coefficiente di variazione**: Dalla deviazione standard vediamo che i valori si discostano circa del 5% dalla media campionaria il che indica una grande variazione dei dati, informazione che è facile estrapolare anche dagli altri 2 valori;
- **Skewness** : Il valore della skewness risulta essere > 0 , il che va a confermare la nostra ipotesi di *asimmetria positiva*;
- **Curtosi** : dal momento che il valore della curtosi è positiva, possiamo dire che la distribuzione di frequenze sia *leptocurtica*, ossia è più piccata di una normale.

3.2.3 BoxPlot

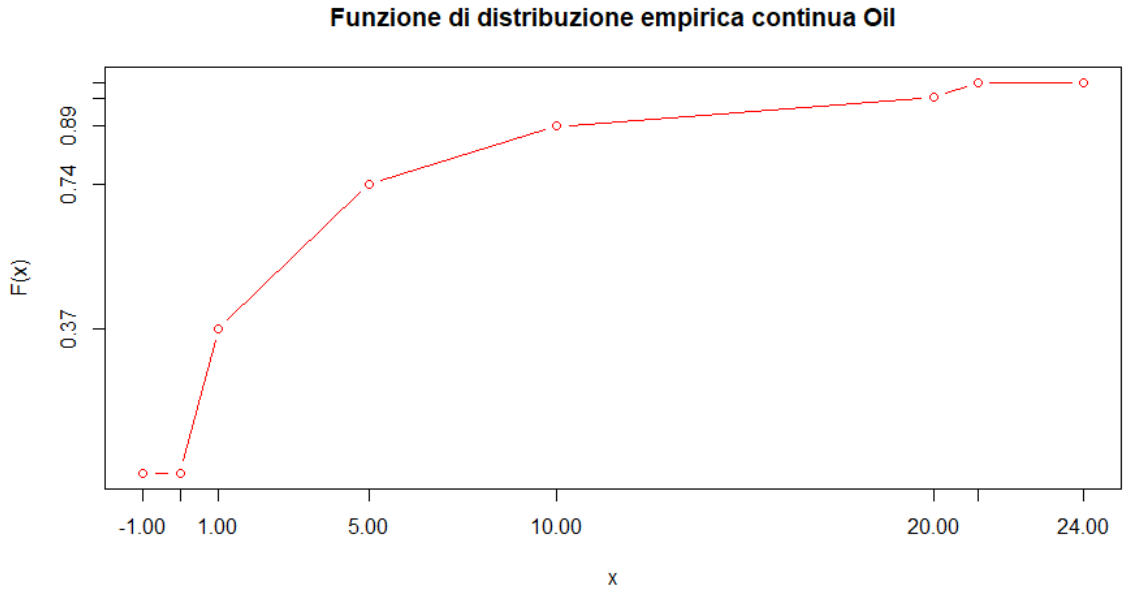


Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
0	0.50	1.39	3.70	3.80	22.80

Tra tutti i paesi solamente 4 sono outlier (valori anomali) poiché hanno una produzione "fuori dalla media" e questi sono , in ordine dal basso verso l'alto , **Olanda** , **Francia** , **Italia** e **Germania** a conferma di quello che avevamo già analizzato nella distribuzione delle frequenze.

3.3 Oil and Petroleum

3.3.1 Funzione Distribuzione Empirica



[**0.05,1**)	[**1,5**)	[**5,10**)	[**10,20**)	[**20,21.18**]
0.37	0.74	0.89	0.96	1

Come possiamo notare nel seguente grafico i valori sono maggiormente distribuiti nelle prime tre classi, ovvero quasi il 90% dei valori è compreso tra 0 e 10, come evidenziato già dalle precedenti analisi.
Nelle classi successive il grafico evidenzia una crescita minore dovuta alla ridotta quantità di dati in quell'intervallo.

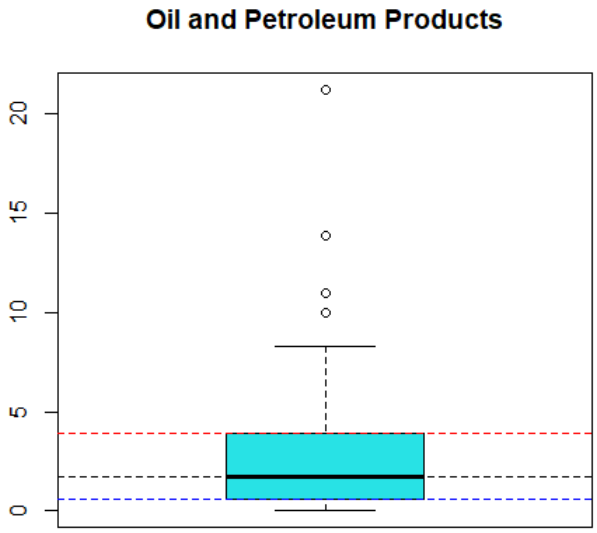
3.3.2 Indici Di Sintesi

Indice di Sintesi	**Valore**
Media Campionaria	3.7037
Mediana Campionaria	1.7137
Classi Modali	[0,1) e [1,5)
Varianza Campionaria	25.6352
Deviazione Standard Campionaria	5.0631
Coefficiente di Variazione	1.3670
Skewness	2.0336
Curtosi Campionaria	3.6819

Quartili	*0%*	*25%*	*50%*	*75%*	*100%*
Valori	0.0527	0.5901	1.7137	5.4744	21.1844

- **Confronto media e mediana** : Come possiamo notare il valore della media risulta essere discostato positivamente, il che ci porta a ipotizzare un'asimmetria positiva;
- **Varianza, deviazione standard e coefficiente di variazione**: Dalla deviazione standard vediamo che i valori si discostano circa del 5% dalla media campionaria il che indica una grande variazione dei dati, informazione che è facile estrapolare anche dagli altri 2 valori;
- **Skewness** : Il valore della skewness risulta essere > 0 , il che va a confermare la nostra ipotesi di *asimmetria positiva* ;
- **Curtosi** : dal momento che il valore della curtosi è positiva, possiamo dire che la distribuzione di frequenze sia *leptocurtica*, ossia è più piccata di una normale.

3.3.3 BoxPlot

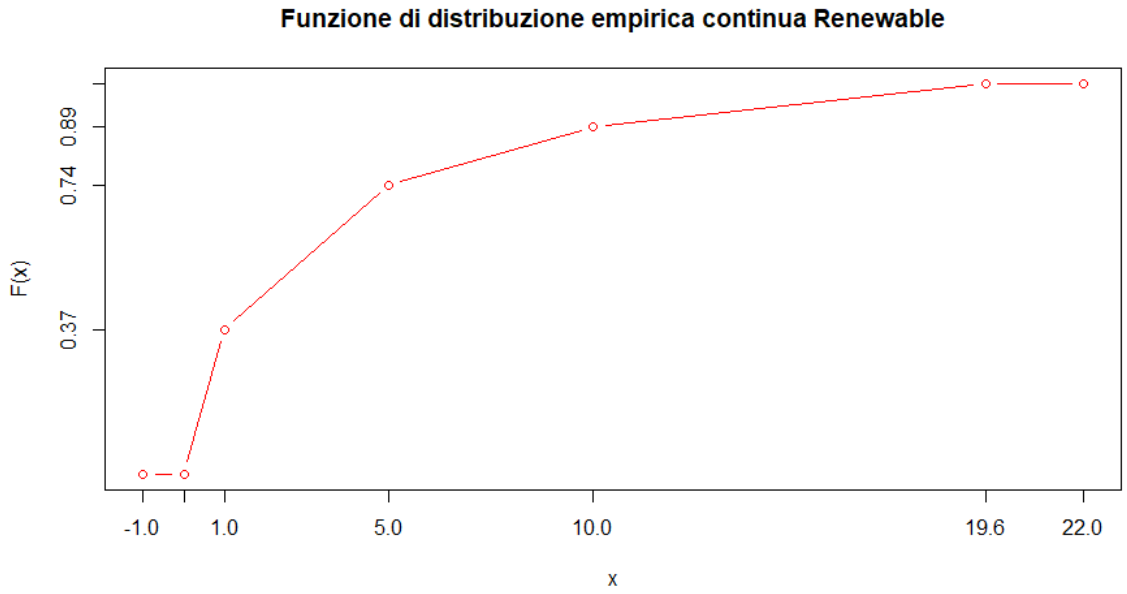


Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.05	0.61	1.71	3.70	3.91	21.18

Tra tutti i paesi solamente 4 sono outlier (valori anomali) poiché hanno una produzione "fuori dalla media" e questi sono , in ordine dal basso verso l'alto , **Italia** , **Spagna** , **Francia** e **Germania** a conferma di quello che avevamo già analizzato nella distribuzione delle frequenze.

3.4 Renewables and Biofuels

3.4.1 Funzione Distribuzione Empirica



[**0.02,1**)	[**1,5**)	[**5,10**)	[**10,19.56**]
0.37	0.74	0.89	1

Come possiamo notare nel seguente grafico i valori sono maggiormente distribuiti nelle prime tre classi, ovvero quasi il 90% dei valori è compreso tra 0 e 10, come già evidenziato dalle precedenti analisi.
Nelle classi successive, il grafico evidenzia una crescita minore dovuta alla ridotta quantità di dati in quell'intervallo.

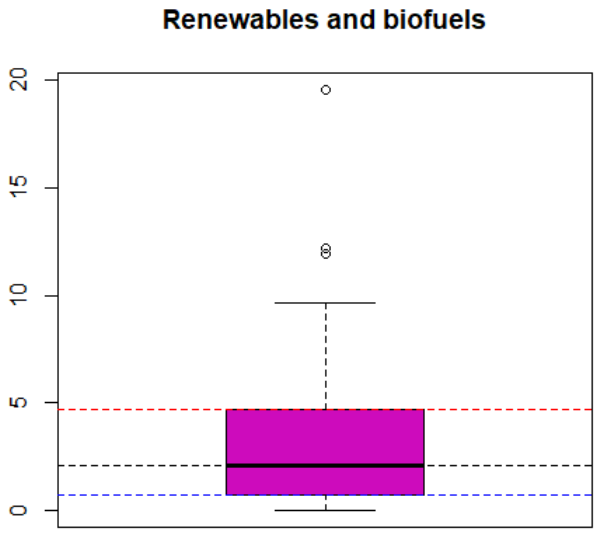
3.4.2 Indici Di Sintesi

Indice di Sintesi	**Valore**
Media Campionaria	3.7037
Mediana Campionaria	2.0537
Classi Modali	[0,1) e [1,5)
Varianza Campionaria	22.3010
Deviazione Standard Campionaria	4.7223
Coefficiente di Variazione	1.2750
Skewness	1.8874
Curtosi Campionaria	3.0933

Quartili	*0%*	*25%*	*50%*	*75%*	*100%*
Valori	0.0238	0.7361	2.0537	5.0080	19.5605

- **Confronto media e mediana** : Come possiamo notare il valore della media risulta essere discostato positivamente, il che ci porta a ipotizzare un'asimmetria positiva;
- **Varianza, deviazione standard e coefficiente di variazione**: Dalla deviazione standard vediamo che i valori si discostano circa del 4% dalla media campionaria il che indica una grande variazione dei dati, informazione che è facile estrapolare anche dagli altri 2 valori;
- **Skewness** : Il valore della skewness risulta essere > 0 , il che va a confermare la nostra ipotesi di *asimmetria positiva* ;
- **Curtosi** : dal momento che il valore della curtosi è positiva, possiamo dire che la distribuzione di frequenze sia *leptocurtica*, ossia è più piccata di una normale.

3.4.3 BoxPlot

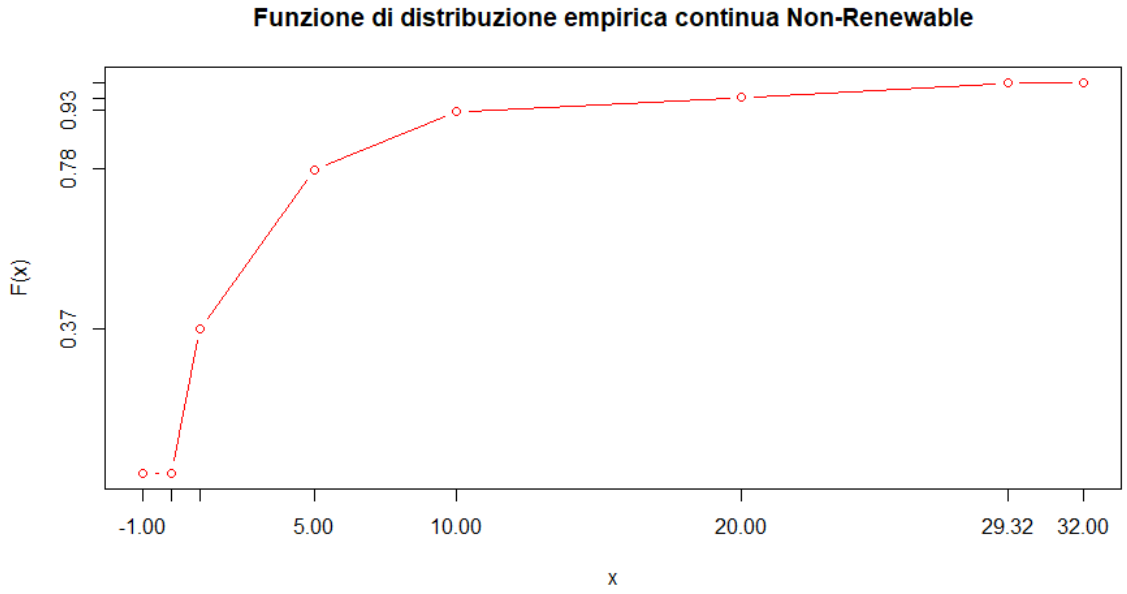


Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
0.02	0.74	2.02	3.70	4.69	19.56

Tra tutti i paesi solamente 3 sono outlier (valori anomali) poiché hanno una produzione "fuori dalla media" e questi sono , in ordine dal basso verso l'alto, **Francia**, **Italia** e **Germania** a conferma di quello che avevamo già analizzato nella distribuzione delle frequenze.

3.5 Non-Renewable Waste

3.5.1 Funzione Distribuzione Empirica



[0,1)	**[1,5)**	**[5,10)**	**[10,20)**	**[20,29.32]**
0.37	0.78	0.93	0.96	1

Come possiamo notare nel seguente grafico i valori sono maggiormente distribuiti nelle prime tre classi, ovvero oltre il 90% dei valori è compreso tra 0 e 10, come evidenziato già dalle precedenti analisi.
Nelle classi successive, il grafico evidenzia una crescita minore dovuta alla ridotta quantità di dati in quell'intervallo.

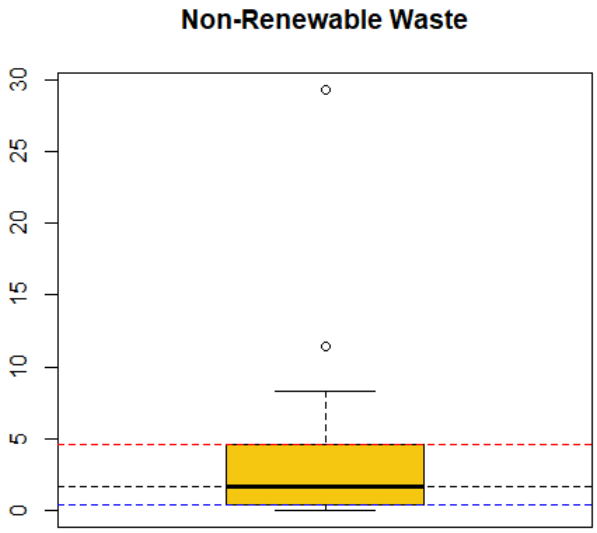
3.5.2 Indici Di Sintesi

Indice di Sintesi	**Valore**
Media Campionaria	3.7037
Mediana Campionaria	1.6440
Classe Modale	[1,5)
Varianza Campionaria	35.1698
Deviazione Standard Campionaria	5.9304
Coefficiente di Variazione	1.6012
Skewness	3.1972
Curtosi Campionaria	11.1127

- **Confronto media e mediana** : Come possiamo notare il valore della media risulta essere discostato positivamente, il che ci porta a ipotizzare un'asimmetria positiva;
- **Varianza, deviazione standard e coefficiente di variazione**: Dalla deviazione standard vediamo che i valori si discostano circa del 6% dalla media campionaria il che indica una grande variazione dei dati, informazione che è facile estrapolare anche dagli altri 2 valori;
- **Skewness** : Il valore della skewness risulta essere > 0 , il che va a confermare la nostra ipotesi di *asimmetria positiva*;
- **Curtosi** : dal momento che il valore della curtosi è positiva, possiamo dire che la distribuzione di frequenze sia *leptocurtica*, ossia è più piccata di una normale.

Quartili	*0%*	*25%*	*50%*	*75%*	*100%*
Valori	0.0000	0.3683	1.6440	4.6779	29.3285

3.5.3 BoxPlot

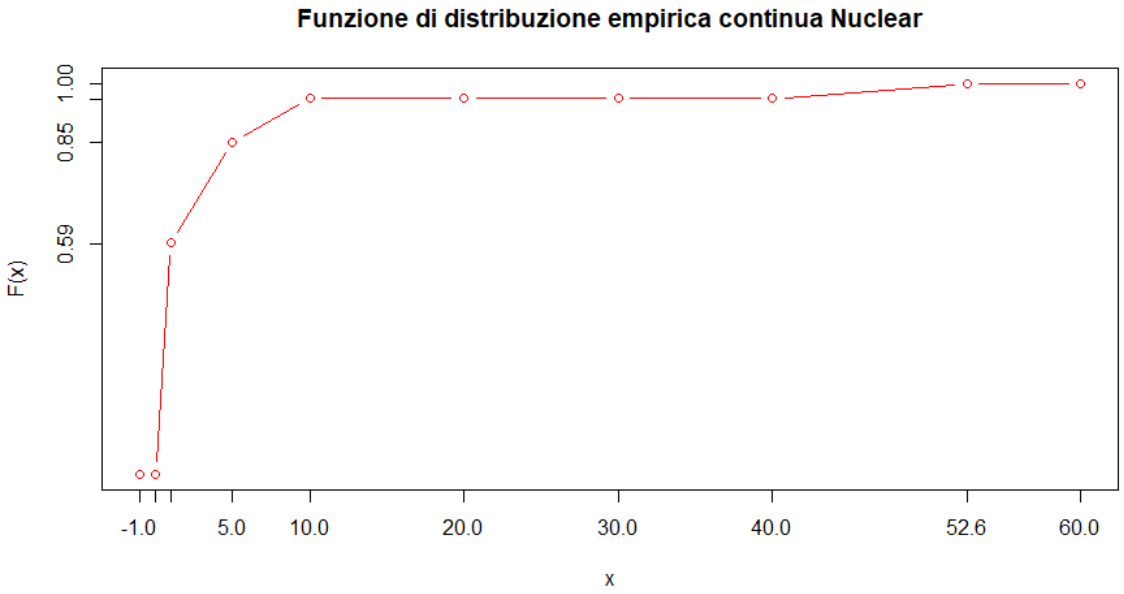


Min	1st Qu.	Median	Mean	3rd Qu.	Max
0.00	0.38	1.64	3.70	4.60	29.32

Tra tutti i paesi solamente 2 sono outlier (valori anomali) poiché hanno una produzione "fuori dalla media" e questi sono , in ordine dal basso verso l'alto , **Francia** e **Germania** a conferma di quello che avevamo già analizzato nella distribuzione delle frequenze.

3.6 Nuclear Heat

3.6.1 Funzione Distribuzione Empirica



[0,1)	**[1,5)**	**[5,10)**	**[10,20)**	**[20,30)**	**[30,40)**	**[40,52.63]**
0.59	0.85	0.96	0.96	0.96	0.96	1

Come possiamo notare nel seguente grafico i valori sono maggiormente distribuiti nelle prime due classi, ovvero oltre il 90% dei valori è compreso tra 0 e 10, come evidenziato già dalle precedenti analisi.
Nelle classi successive, quelle che vanno da 10 a 40, il grafico evidenzia una crescita nulla dovuta all'assenza di dati in quell'intervallo, per poi vedere alla fine una crescita poco pronunciata con l'ultima classe.

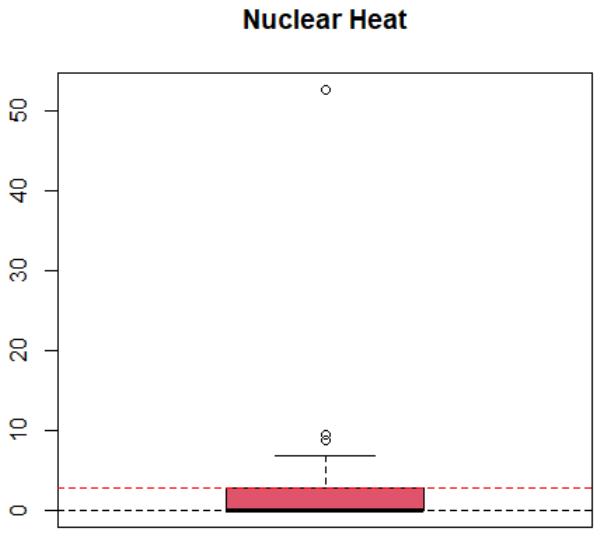
3.6.2 Indici Di Sintesi

Indice di Sintesi	**Valore**
Media Campionaria	3.7037
Mediana Campionaria	0
Classe Modale	[0,1)
Varianza Campionaria	103.2299
Deviazione Standard Campionaria	10.1602
Coefficiente di Variazione	2.7432
Skewness	4.3628
Curtosi Campionaria	18.5108

Quartili	*0%*	*25%*	*50%*	*75%*	*100%*
Valori	0.0000	0.0000	0.0000	3.1668	52.6393

- **Confronto media e mediana** : Come possiamo notare il valore della media risulta essere discostato positivamente, il che ci porta a ipotizzare un'asimmetria positiva;
- **Varianza, deviazione standard e coefficiente di variazione**: Dalla deviazione standard vediamo che i valori si discostano circa del 10% dalla media campionaria il che indica una grande variazione dei dati, informazione che è facile estrapolare anche dagli altri 2 valori;
- **Skewness** : Il valore della skewness risulta essere > 0 , il che va a confermare la nostra ipotesi di *asimmetria positiva* ;
- **Curtosi** : dal momento che il valore della curtosi è positiva, possiamo dire che la distribuzione di frequenze sia *leptocurtica*, ossia è più piccata di una normale.

3.6.3 BoxPlot



Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
0.00	0.00	0.00	3.70	2.82	52.63

Tra tutti i paesi solamente 3 sono outlier (valori anomali) poiché hanno una produzione "fuori dalla media" e questi sono, in ordine dal basso verso l'alto,**Spagna**,**Germania** e la **Francia** a conferma di quello che avevamo già analizzato nella distribuzione delle frequenze.

4 Analisi Descrittiva Bivariata

4.1 Introduzione

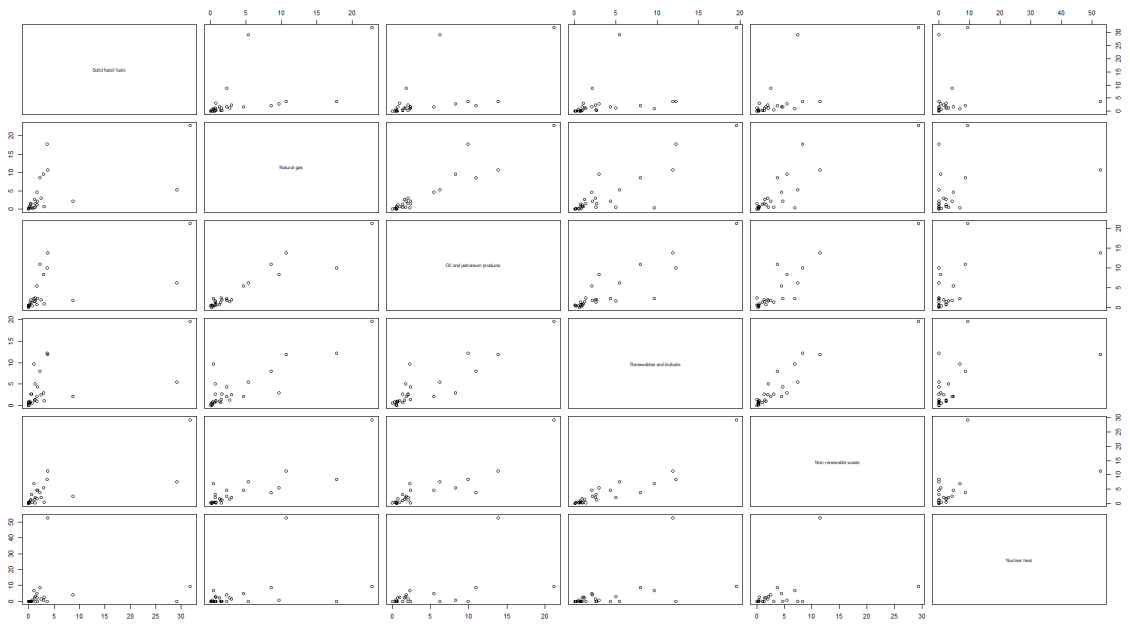
Prima di intraprendere un analisi descrittiva bivariata, la prima operazione che dobbiamo effettuare è quella di identificare le variabili con cui vogliamo effettuare la seguente analisi.
Per far ciò, analizzeremo inizialmente un grafico ottenuto tramite la funzione *pairs()* che mette in relazione tutte le n colonne della matrice tra di loro, e successivamente creeremo e analizzeremo le **tabelle di correlazione** e **covarianza**.

4.2 Grafico Pairs

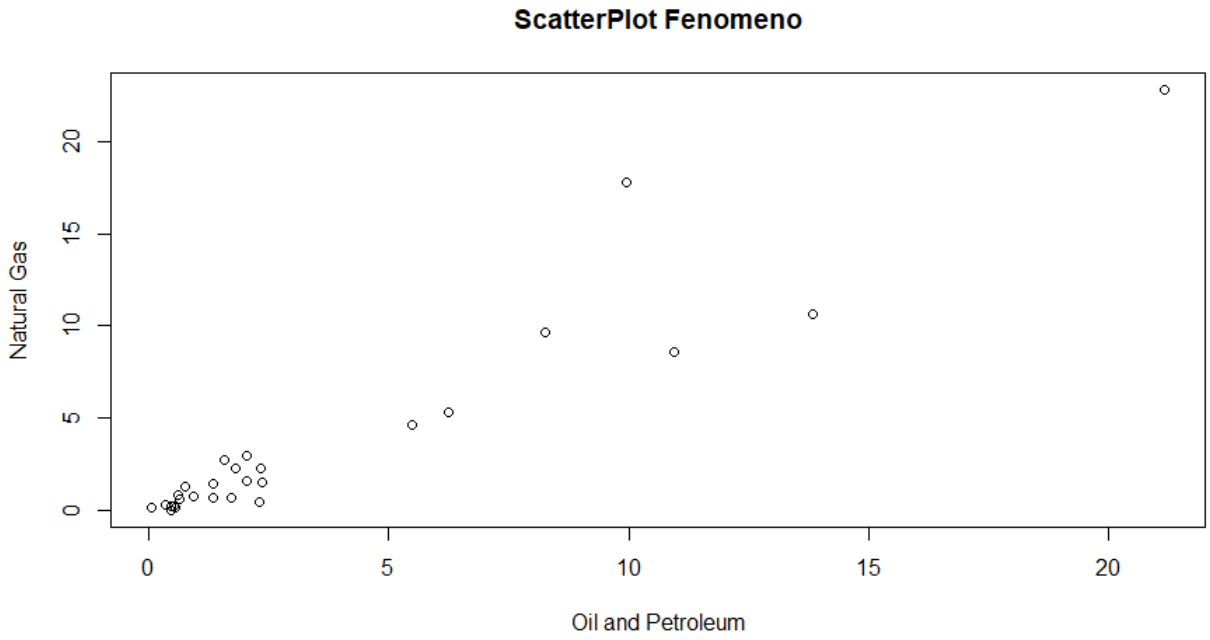
Come introdotto in precedenza eseguiamo quindi il comando

```
pairs(DataframePaesi[,3:8])
```

ed otteniamo il seguente grafico :



L'operazione di analisi che è stata effettuata a partire dal grafico precedente è stata quella di individuare una relazione che risultasse **lineare**, ovvero facilmente rappresentabile tramite una retta di equazione $y = mx + b$.
Tra quelli osservati il più adatto sembra essere quello formato da *Oil and Petroleum Products* e *NaturalGas* con la prima variabile posta come **Indipendente** e la seconda come **Dipendente**.
Il grafico scelto è quindi il seguente:



Per verificare che l'ipotesi sulla relazione lineare sia corretta dobbiamo analizzare le **tabelle di correlazione** e **covarianza**.

4.3 Tabelle Covarianza e Correlazione campionarie

4.3.1 Covarianza Campionaria

Siano X ed Y due variabili, e siano $(x_1, x_2, ..., x_n)$ e $(y_1, y_2, ..., y_n)$ campioni con n osservazioni relativi ad esse, con media campionaria rispettivamente pari ad $\overline{x}$ ed $\overline{y}$.
La **Covarianza Campionaria** è una misura del grado di correlazione lineare tra le due variabili, ed è definita come:

$$c_{xy} := \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

In base al valore di c_{xy}, si può definire il tipo di relazione che lega le variabili :

- $c_{xy} < 0 \implies$ X ed Y sono *linearmente correlate negativamente*;
- $c_{xy} = 0 \implies$ X ed Y *non sono linearmente correlate*;
- $c_{xy} > 0 \implies$ X ed Y sono *linearmente correlate positivamente*;

Si noti inoltre che la covarianza campionaria tra X ed X è pari alla *varianza campionaria*.

4.3.2 Tabella delle covarianze

Tramite la funzione *cov()* applicata sulla matrice dei dati è stata ottenuta la seguente tabella.

	Solid Fossil	Natural Gas	Oil	Renewables	Non-Renewable	Nuclear
Solid Fossil	62.98	27.42	25.85	22.61	36.14	8.66
Natural Gas	27.42	32.06	27.05	23.07	29.00	20.71
Oil	25.85	27.05	25.63	21.36	27.05	27.89
Renewables	22.61	23.07	21.36	22.30	25.50	24.07
Non Renewable	36.14	29.00	27.05	25.50	35.16	24.74
Nuclear	8.66	20.71	27.89	24.07	24.74	103.22

Tabella 1: Tabella delle covarianze

Nel caso preso in analisi, ossia *'Oil and petroleum products'* e *'Natural Gas'*, la covarianza campionaria è pari a 27.05. Quindi le due colonne sono linearmente correlate positivamente.

4.3.3 Correlazione Campionaria

Un indice strettamente legato alla covarianza campionaria è il **coefficiente di correlazione campionario**, una misura normalizzata del grado di relazione lineare tra le variabili X ed Y.
Esso è definito come:

$$r_{xy} := \frac{c_{xy}}{s_x s_y}$$

ed assume sempre valori compresi tra -1 e 1.
Il valore del coefficiente di correlazione campionaria indica il tipo e l'intensità della relazione lineare presente tra le variabili X ed Y:

- $r_{xy} = -1 \implies$ X ed Y sono in *correlazione negativa perfetta*;
- $-1 < r_{xy} < 0 \implies$ X ed Y sono *linearmente correlate negativamente*;
- $r_{xy} = 0 \implies$ X ed Y *non sono linearmente correlate*;
- $0 < r_{xy} < 1 \implies$ X ed Y sono *linearmente correlate positivamente*;
- $r_{xy} = 1 \implies$ X ed Y sono in *correlazione lineare positiva perfetta*;

4.3.4 Tabella delle Correlazioni

Tramite l'applicazione della funzione *cor()* sulla matrice dei dati si ottiene la seguente tabella:

	Solid Fossil	Natural Gas	Oil	Renewables	Non-Renewable	Nuclear
Solid Fossil	1.00	0.61	0.64	0.60	0.76	0.10
Natural Gas	0.61	1.00	0.94	0.86	0.86	0.36
Oil	0.64	0.94	1.00	0.89	0.90	0.54
Renewables	0.60	0.86	0.89	1.00	0.91	0.50
Non Renewable	0.76	0.86	0.90	0.91	1.00	0.41
Nuclear	0.10	0.36	0.54	0.50	0.41	1.00

Tabella 2: Tabella Delle Correlazioni

Nel caso preso in analisi il coefficiente di correlazione risulta essere pari al 0.94, ossia molto vicino all'uno e dunque con una buona correlazione lineare positiva. Inoltre risulta essere la combinazione di colonne con la correlazione maggiore, e per tale motivo si è ritenuto opportuno continuare nell'analisi della regressione lineare.

4.4 Regressione Lineare

4.4.1 Regressione Lineare Semplice

Siano X ed Y due variabili, e siano $(x_1, x_2, ..., x_n)$ e $(y_1, y_2, ..., y_n)$ campioni con n osservazioni relativi ad esse.
Un **modello di regressione lineare semplice** è una funzione lineare della variabile X che approssima Y :

$$\hat{Y} = \beta \cdot X + \alpha$$

dove α e β sono detti rispettivamente **intercetta** e **coefficiente angolare** di X, e sono tali da minimizzare la somma dei quadrati degli errori:

$$\alpha, \beta = \arg\min \left(\text{Q:=} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)$$
$$\text{con : } \hat{y}_i = \beta \cdot x_i + \alpha$$

I valori di α e β vengono ottenuti ponendo a 0 le derivate parziali della funzione Q rispetto ad esse, e quindi risolvendo il sistema lineare:

$$\begin{cases} \dfrac{\partial Q}{\partial a} = 0 \\ \\ \dfrac{\partial Q}{\partial a} = 0 \end{cases} \tag{1}$$

$$\beta = \frac{s_x}{s_y} \cdot r_{xy}$$
$$\alpha = \overline{y} - \beta \cdot \overline{x}$$

Si può dimostrare infine che la media dei valori predetti $(\hat{y_1}, \hat{y_2}.., \hat{y_n})$ è pari a quella dei valori osservati $(y_1, y_2, ..., y_n)$.

4.4.2 Coefficiente di Determinazione

Una metrica dell'efficienza di un modello di regressione è il **coefficiente di determinazione**, che possiede 3 diverse definizioni :

- $D_1^2 := 1 - \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$
- $D_2^2 := \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$
- $D_3^2 := \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

Per i modelli di regressione lineare si può calcolare anche come il quadrato del coefficiente di correlazione e viene denotato come *r-square*. Inoltre per le regressioni di questo tipo è vero che :

$$D_1^2 = D_2^2 = D_3^2$$
e che
$$0 \leq D_1^2 = D_2^2 = D_3^2 \leq 1$$

In generale si ha che :

$$0 \leq D_1^2 \leq 1$$

per tutti i tipi di modelli lineari e non, mentre D_2^2 e D_3^2 possono assumere anche valori strettamente maggiori di 1 per alcuni modelli non lineari, che verranno descritti in seguito, su cui è quindi obbligatorio usare D_1^2.

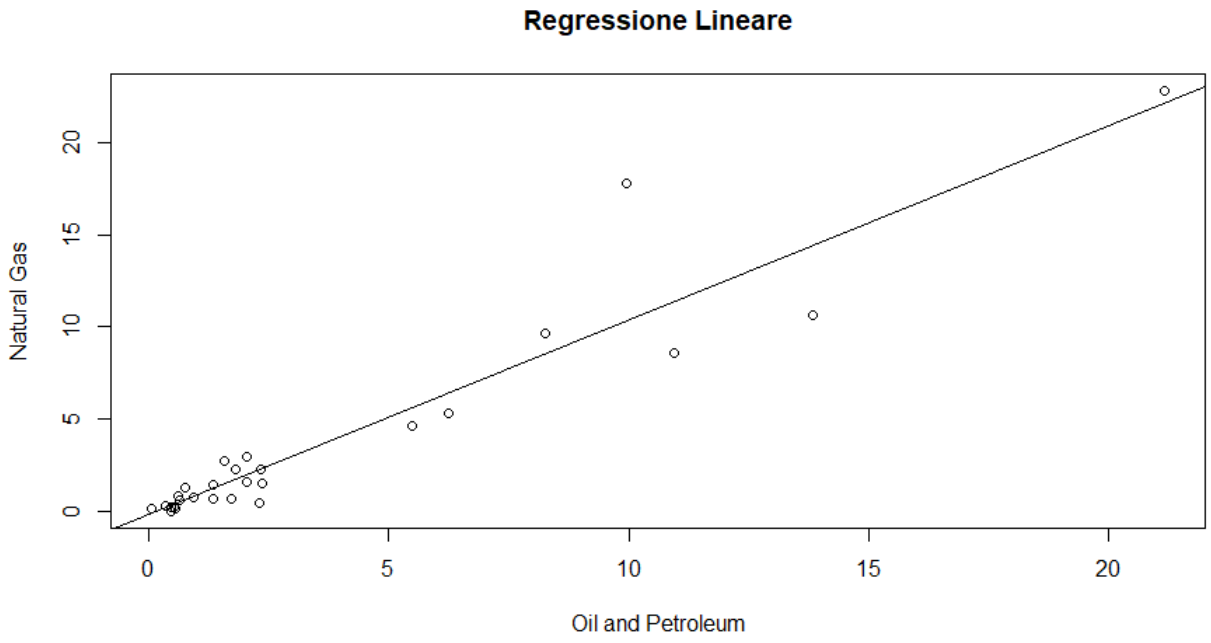
4.4.3 Modello Lineare Natural Gas & Oil and Petroleum

Inizialmente, calcoliamo i coefficienti α e β, per poter disegnare la retta di regressione, tramite la funzione offertaci da R *lm(Y∼X)*.
Successivamente calcoliamo il coefficiente di determinazione della seguente regressione per verificarne l'efficienza.
Per il calcolo del seguente coefficiente possiamo utilizzare sia *summary(modello)$r.square* che calcolare il quadrato della correlazione.

```
> linearmodel <- lm( DataFramePaesi$`Natural gas` ~ DataFramePaesi$`
    Oil and petroleum products`)
> summary(linearmodel)$r.square
[1] 0.890734
> cor(DataFramePaesi$`Natural gas`,DataFramePaesi$`Oil and petroleum
    products`)^2
[1] 0.890734
```

Come vediamo entrambi i metodi restituiscono lo stesso coefficiente di regressione, il quale risulta essere molto alto, il che ci notifica che come avevano ipotizzato inizialmente le due variabili sono efficientemente approssimabili con una retta.
Tracciamo quindi la retta di approssimazione sullo scatterplot visto all'inizio di questo capitolo avvalendoci dei coefficienti calcolati e della funzione *abline()*.

```
> plot(DataFramePaesi$`Oil and petroleum products`,DataFramePaesi$`
    Natural gas`,xlab="Oil and Petroleum" , ylab = "Natural Gas" , main
     = "Regressione Lineare")
> abline(linearmodel)
```



4.4.4 Residui

Siano $(\hat{y_1}, \hat{y_2}.., \hat{y_n})$ i valori predetti dal modello di regressione lineare; definiamo l' **i-esimo residuo** come il valore :

$$E_i := y_i - \hat{y}$$

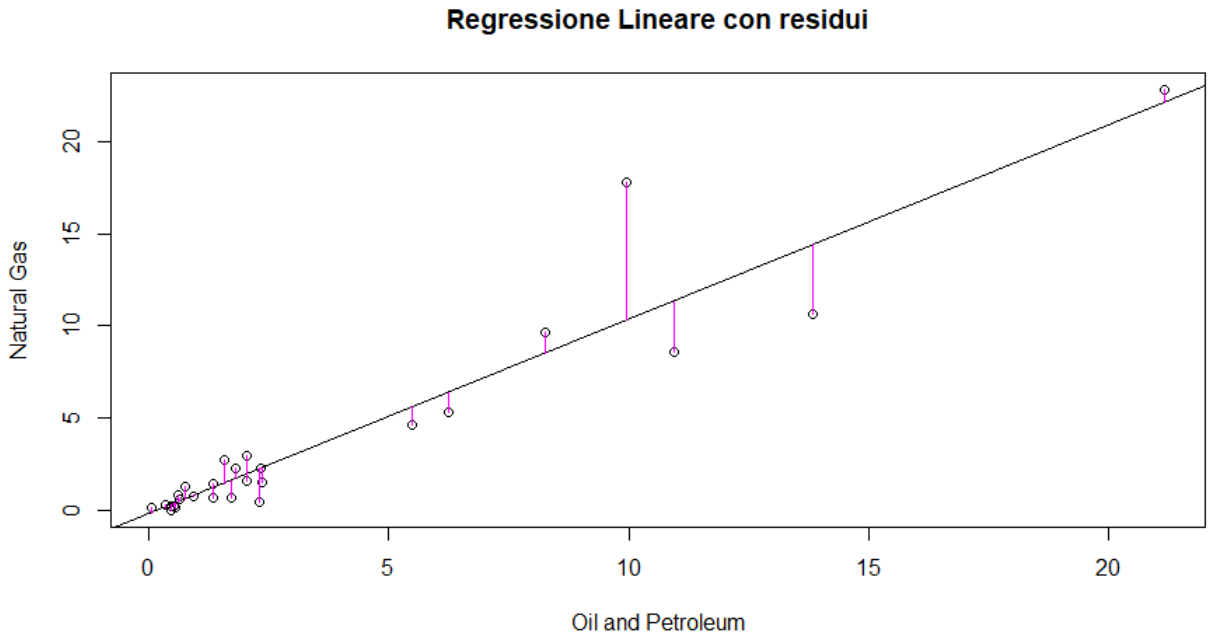
Si può notare poi che la media campionaria dei residui è sempre pari a 0.
Si può definire anche una versione normalizzata dei residui,il cui valore è indipendente dalle unità di misura.
Definiamo quindi **i-esimo residuo standardizzato** il valore :

$$E_i^{(s)} := \frac{E_i - \overline{E}}{s_E} = \frac{E_i}{s_E}$$

4.4.5 Residui della Regressione Lineare

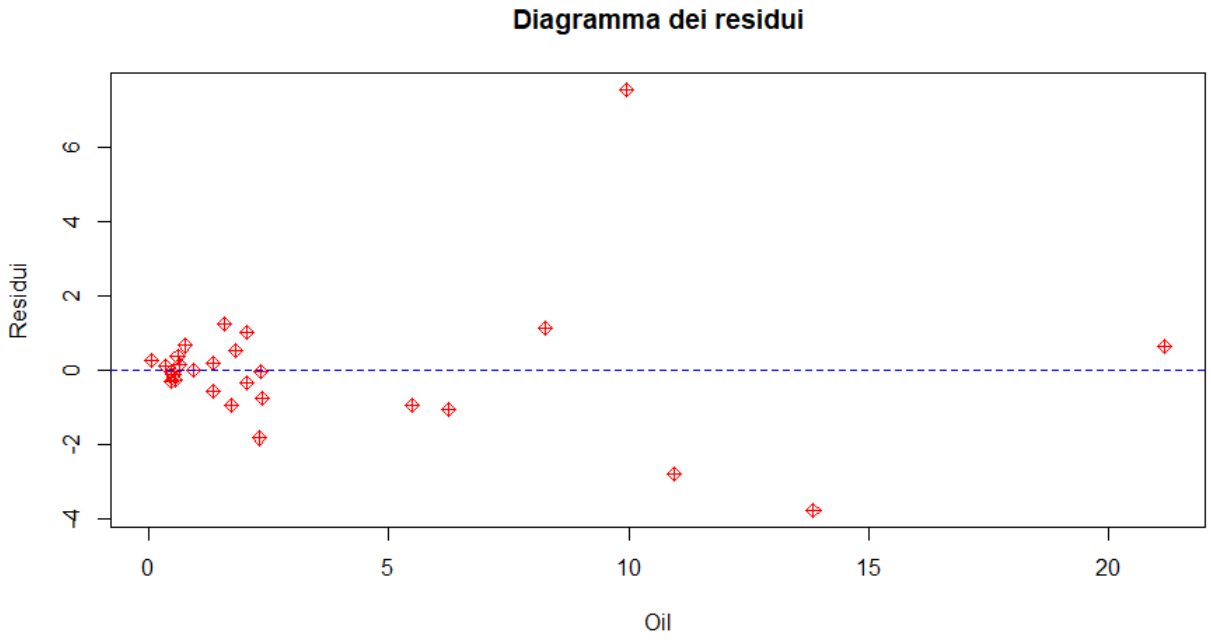
Dalla retta di regressione lineare trovata in precedenza per le variabili *Oil and petroleum product* e *Natural gas* si è dunque proceduto a descriverne i residui.
Si è realizzato un grafico ottenuto aggiungendo, al grafico contenente lo scatterplot e la retta di regressione, dei segmenti verticali che visualizzano i residui, ossia la distanza tra valore osservato e valore stimato.
Tramite il seguente codice è stato realizzato il grafico presente nella figura successiva.

```
> plot(DataFramePaesi$`Oil and petroleum products`,DataFramePaesi$`
    Natural gas`,xlab="Oil and Petroleum" , ylab = "Natural Gas" , main
     = "Regressione Lineare con residui")
> abline(linearmodel)
> stime <- fitted(linearmodel)
> segments(DataFramePaesi$`Oil and petroleum products`,stime,
    DataFramePaesi$`Oil and petroleum products`,DataFramePaesi$`Natural
     gas`, col = "magenta")
```



Per approfondire tale relazione è stato realizzato il seguente diagramma dei residui, il quale è un grafico in cui i valori dei residui sono posti sull'asse delle ordinate e quelli della variabile indipendente sull'asse delle ascisse.
La retta orizzontale è posizionata nello zero e corrisponde alla media campionaria dei residui.
Calcoliamo dunque i residui del modello tramite la funzione *resid* e come evidenziato dal seguente codice otteniamo il grafico dei residui seguente.

```
> residui <- resid(linearmodel)
> plot(DataFramePaesi$`Oil and petroleum products`,residui,main="
   Diagramma dei residui",xlab="Oil",ylab="Residui",pch=9,col="red")
> abline(h=0,col="blue",lty=2)
```

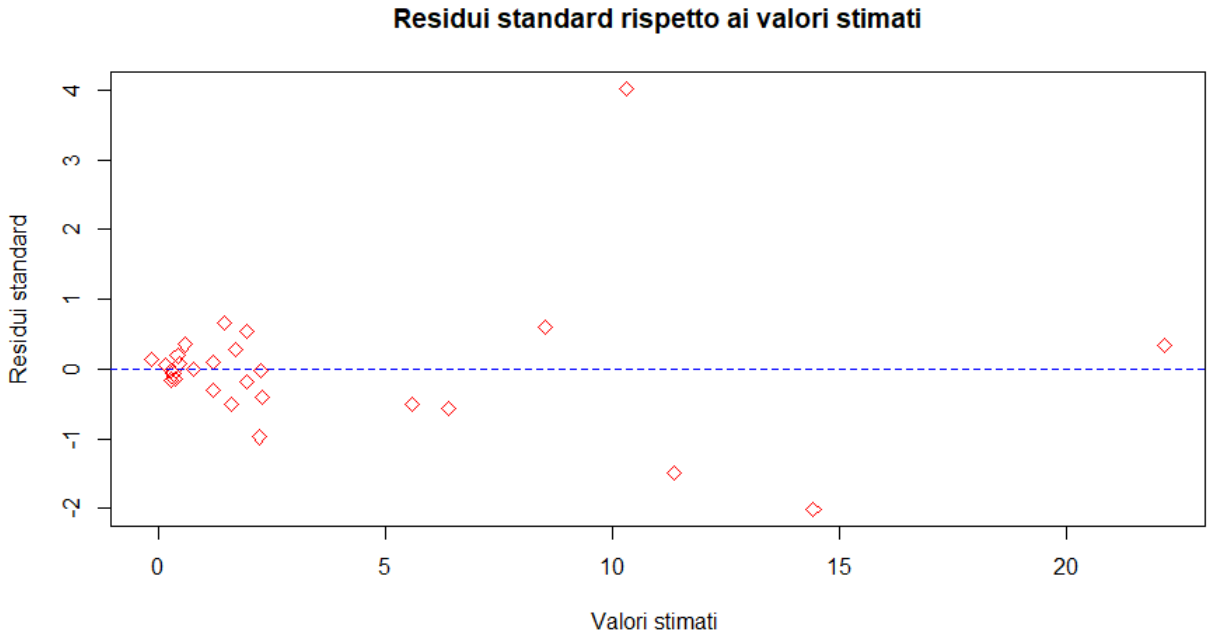


Ad affiancare il grafico dei residui e per favorirne l'interpretazione sono stati calcolati anche diversi indici di posizione relativi ai residui. La media è pari a zero ed è stata dunque omessa.

```
> median(residui)
[1] -0.03399809
> var(residui)
[1] 3.503646
> sd(residui)
[1] 1.871803
```

Per terminare l'analisi mostriamo di seguito il codice utilizzato per creare un grafico dei residui standardizzati, ottenuti mettendo in relazione i valori stimati e i residui standard, ossia i residui divisi la loro deviazione standard.

```
> linearmodel <- lm( DataFramePaesi$`Natural gas` ~ DataFramePaesi$`
    Oil and petroleum products`)
> residui <- resid(linearmodel)
> stime <- fitted(linearmodel)
> residuisd <- residui/sd(residui)
> plot(stime, residuisd, main="Residui standard rispetto ai valori
    stimati", xlab="Valori stimati ",ylab =" Residui standard ",pch =5,
     col ="red ")
> abline (h=0, col ="blue ",lty =2)
```



4.5 Regressione lineare multipla

4.5.1 Definizione

Siano $X_1, X_2, \ldots, X_p$ ed Y variabili, siano $(x_{1,j}, x_{2,j}, \ldots, x_{n,j})$, $j = 1, 2, \ldots, p$ e $(y_1, y_2, \ldots, y_n)$ campioni con n osservazioni relative ad esse.
Un modello di regressione lineare multipla è una funzione lineare delle variabili $X_1, X_2, \ldots, X_n$ che approssima Y:

$$\hat{Y} = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 \cdots + \beta_p \cdot X_p + \alpha$$

con α e β_j detti rispettivamente intercetta e coefficiente angolare di $X_j (\forall j = 1, 2, \ldots, p)$, e sono tali da minimizzare la somma dei quadrati degli errori

$$\alpha, \beta = \arg\min \left(\text{Q:=} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right)$$
$$\text{dove : } \hat{y}_i = \beta_1 \cdot x_{i,1} + \beta_2 \cdot x_{i,2} + \cdots + \beta_p \cdot x_{i,p} + \alpha$$

i valori di α e di $\beta_1, \beta_2, \ldots, \beta_p$ vengono ottenuti ponendo a 0 le derivate parziali della funzione Q rispetto ad esse, e risolvendo quindi il sistema lineare:

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha} = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \\ \dfrac{\partial Q}{\partial \beta_2} = 0 \\ \dots \\ \dfrac{\partial Q}{\partial \beta_p} = 0 \end{cases} \tag{2}$$

I risultati sui residui ottenuti per la regressione lineare semplice sono estendibili anche a quella multipla, mentre il valore del coefficiente di determinazione non può essere calcolato come il quadrato del coefficiente di correlazione, in quanto nella relazione modellata sono coinvolte più di due variabili.

4.5.2 Regressione Multipla

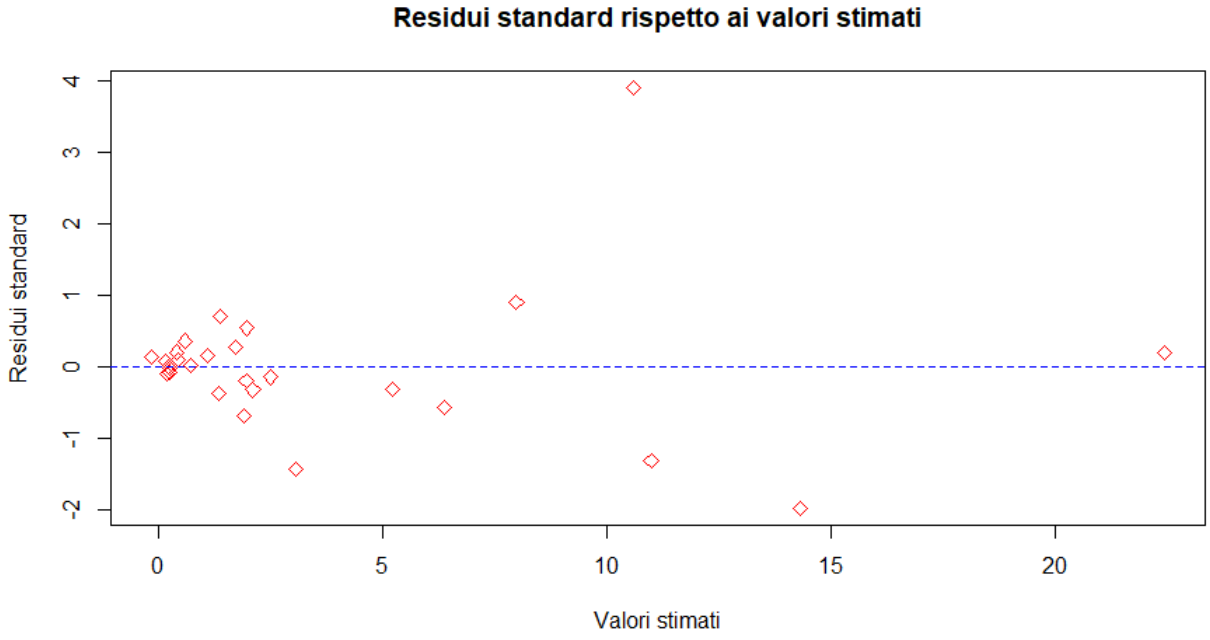
Nel nostro caso si è provato a migliorare l'analisi sulla regressione lineare precedente su *Oil and petroleum products* e *Natural Gas*. In particolare sono state selezionate altre due colonne fortemente correlate alle precedenti, ossia *Renewables and biofuels* e *Non-renewable waste*. Per effettuare l'analisi sono state incluse dunque nell'esecuzione del metodo *lm()* e di seguito sono riportati i risultati e il coefficiente di determinazione, calcolato sia tramite l'attributo r-square che tramite la seconda definizione.

```
> model <- lm( DataFramePaesi$`Natural gas` ~ DataFramePaesi$`Oil and
   petroleum products`+ DataFramePaesi$`Renewables and biofuels`+
   DataFramePaesi$`Non-renewable waste`)
> model

...

Coefficients:
(Intercept)      DataFramePaesi$`Oil and petroleum products`
-0.26368                      0.94667
DataFramePaesi$`Renewables and biofuels`    DataFramePaesi$`Non-
   renewable waste`
   0.10153                                          0.02299
> summary(model)$r.square
[1] 0.8927216
> stimemult <- fitted(model)
> num <- sum((stimemult - mean(DataFramePaesi$'Natural gas'))^2)
> den <- sum((DataFramePaesi$'Natural gas' - mean(DataFramePaesi$'
   Natural gas'))^2)
> d2 <- num/den
> d2
```

```
[1] 0.8927216
```

Il coefficiente di determinazione è leggermente migliorato rispetto al modello precedente, da 0.890734 a 0.8927216.
Inoltre si nota che il coefficiente della colonna *Non-Renewable* risulta essere molto basso, questo va ad indicare che la colonna influisce inferiormente sulla regressione rispetto alle altre colonne.
Per analizzare dunque i residui è stato prodotto il seguente grafico dei residui standardizzati tramite il seguente codice.

```
> residuimult <- resid(model)
> residuimultstandard <- residuimult/sd(residuimult)
> plot(stimemult, residuimultstandard, main="Residui standard rispetto
    ai valori stimati", xlab="Valori stimati", ylab="Residui standard"
   , pch=5, col="red")
> abline(h=0, col="blue", lty=2)
```



4.6 Regressione Non Lineare

In generale dunque, un modello di regressione (non necessariamente lineare) è una funzione che approssima il valore della variabile dipendente Y a partire da quello di quella indipendente X, minimizzando la somma dei quadrati degli errori:

4.6.1 Definizione

$$\hat{Y} = f(X) \text{ t.c. } \min \mathrm{Q} = \sum_{i=1}^{n} (y_i - f(x_i))$$

Alcuni modelli di regressione non lineari possono essere linearizzati, ovvero ridotti a modelli lineari semplici con l'ausilio di apposite trasformazioni.
Regressione quadratica
Un **modello di regressione quadratico** è una funzione :

$$\hat{Y} = \gamma \cdot X^2 + \beta \cdot X + \alpha$$

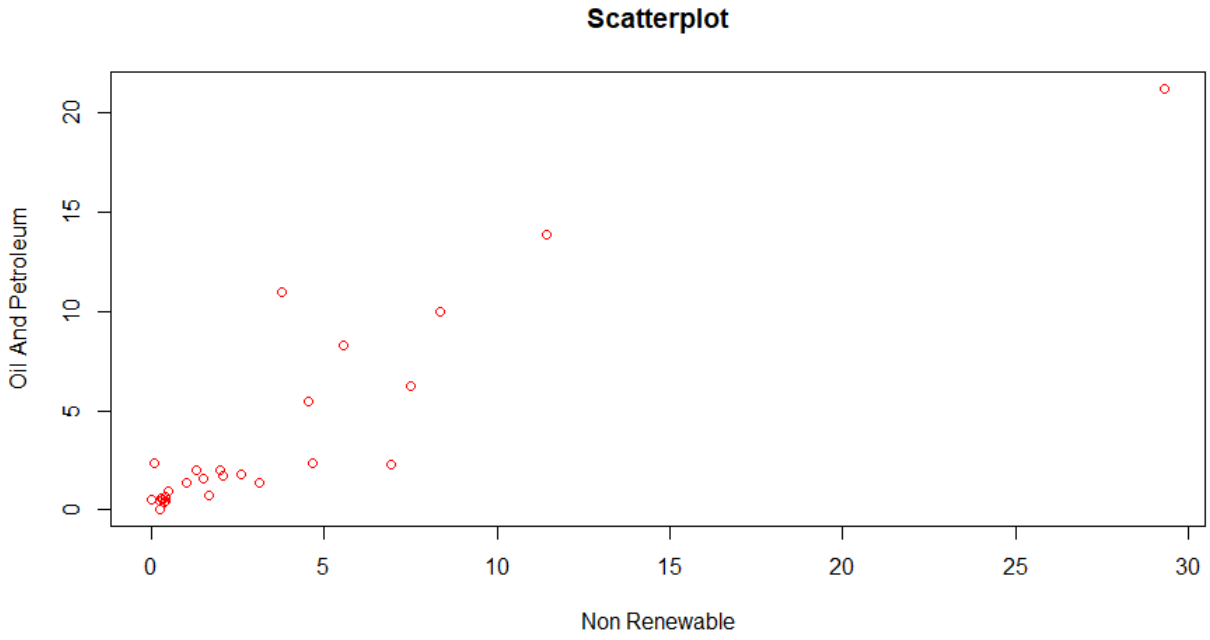
Per linearizzarla, poniamo $X_1 = X$ e $X_2 = X^2$, ottenendo cosi' il modello di regressione lineare multipla:

$$\hat{Y} = \alpha + \beta \cdot X_1 + \gamma \cdot X_2$$

Calcolando ora α, β e γ come i coefficienti di un modello di regressione lineare multipla, otteniamo i coefficienti per il modello quadratico.
Essendo la $\hat{Y}$ inalterata nell'espressione linearizzata, ed essendo quest'ultima un modello di regressione lineare multipla, le proprietà del coefficiente di determinazione vengono ereditate dal modello quadratico.

4.6.2 Regressione Quadratica

Per quanto riguarda i dati a disposizione, dal momento che nei casi precedenti la regressione lineare è risultata sufficiente a descrivere il fenomeno, si è deciso di analizzare altre combinazioni di variabili. In particolare una possibile regressione non lineare, ed in dettaglio quadratica è stata ipotizzata su *Non-renewable waste* come variabile dipendente e *Oil and petroleum products* come variabile indipendente. In seguito è riportato lo scatterplot per i dati selezionati.

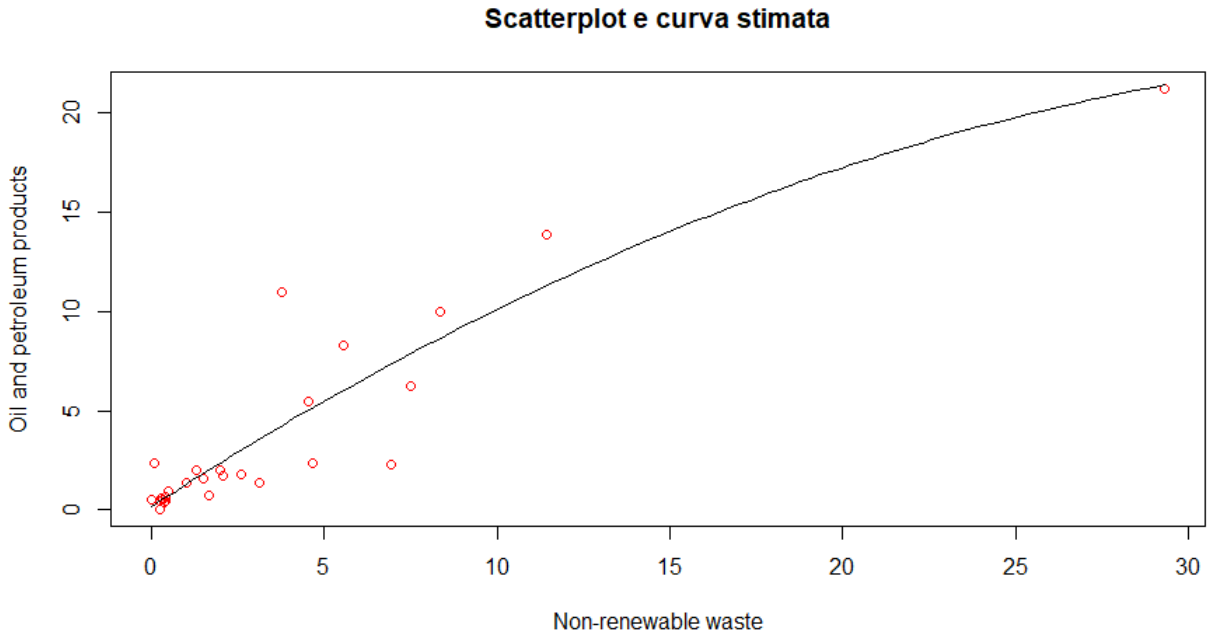


In seguito si è proceduto con la definizione del modello non lineare utilizzando la funzione *lm()* e gli identificatori *I()* per linearizzare il modello della regressione quadratica. Inoltre sono stati determinati i coefficienti ed è stato calcolato il coefficiente di determinazione tramite la prima definizione.

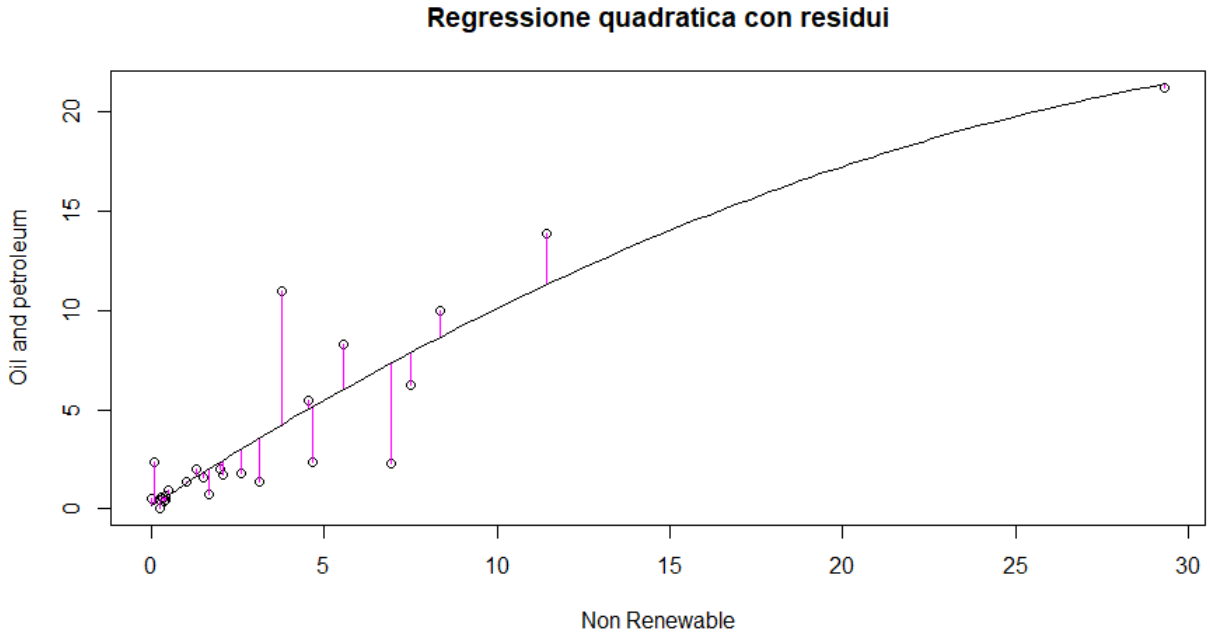
```
> pol2 <-lm(DataFramePaesi$`Oil and petroleum products` ~
   DataFramePaesi$`Non-renewable waste`+ I((DataFramePaesi$`Non-
   renewable waste`)^2))
> alpha <- pol2$coefficients[[1]]
> beta <- pol2$coefficients[[2]]
> gamma <- pol2$coefficients[[3]]
> stime <-alpha + beta*DataFramePaesi$`Non-renewable waste`+gamma*(
   DataFramePaesi$`Non-renewable waste`)^2
> num1 <-sum ((( DataFramePaesi$`Oil and petroleum products` - stime )
   ^2))
> den1 <- sum(((DataFramePaesi$`Oil and petroleum products`-mean(
   DataFramePaesi$`Non-renewable waste`))^2))
> d1 <- 1 - num1/den1
> d1
[1] 0.8376927
```

Il coefficiente di determinazione è alto e la curva stimata potrebbe essere una buona approssimazione. Di seguito è riportato lo scatterplot precedente con l'aggiunta della curva stimata e il relativo codice.

```
> plot(DataFramePaesi$`Non-renewable waste`,DataFramePaesi$`Oil and
   petroleum products`, col ="red ",main=" Scatterplot e curva stimata
    ")
> curve ( alpha +beta *x+ gamma *x^2, add = TRUE)
```



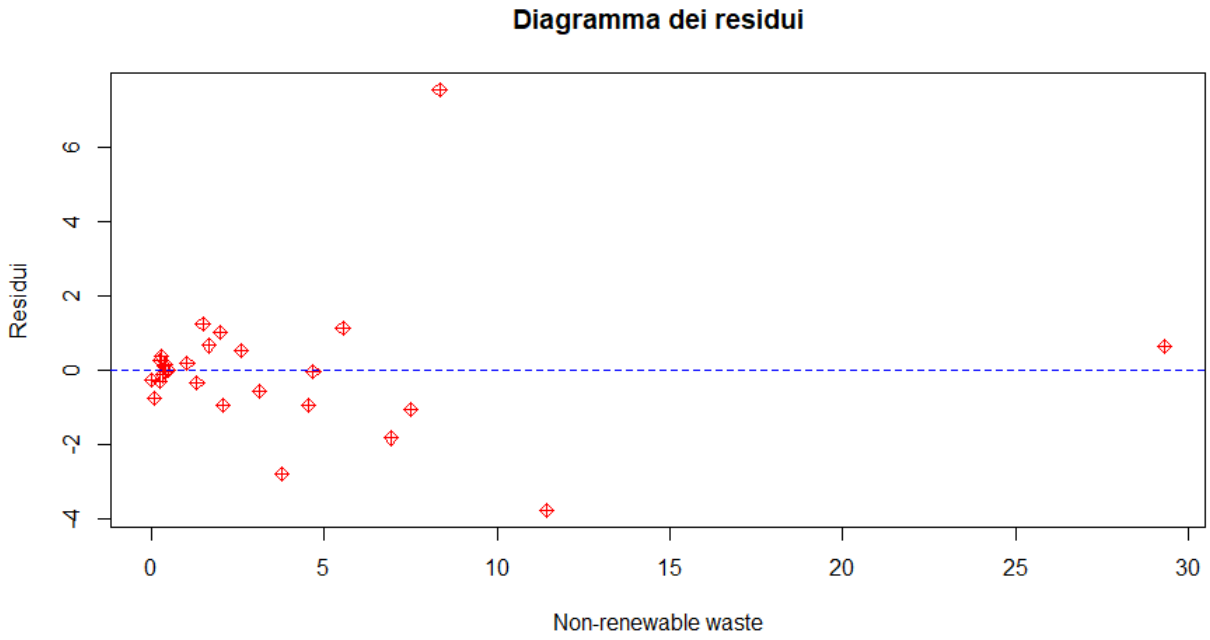
Inoltre ,abbiamo realizzato ,come nel caso precedente , un grafico che traccia i segmenti dei residui.



Infine sono stati analizzati i residui del modello tramite il grafico dei residui mettendo in relazione i residui e la colonna *Non-renewable waste*. Di seguito sono riportati il

codice e il grafico.

```
> residuipol2 <- resid(pol2)
> plot(DataFramePaesi$‘Non-renewable waste‘ ,residui , main ="
   Diagramma dei residui ", xlab="Non-renewable waste",ylab =" Residui
    ",pch =9, col ="red ")
> abline (h=0, col =" blue",lty =2)
```



5 Analisi Dei Cluster

5.1 Definizioni preliminari

5.1.1 Clustering, Metriche e Similarità

Il **clustering** è un'operazione che consiste nell'individuare una partizione di un insieme di n individui I = $I_1, I_2, ..., I_n$, suddividendolo in k sottoinsiemi $G_1, G_2, ..., G_k$ detti appunto *cluster*, in modo tale che gli individui più simili o meno distanti risultino appartenere allo stesso sottoinsieme.
Supponiamo che ogni individuo I_j possa essere rappresentato mediante un vettore di p caratteristiche X_j , costruiamo allora una matrice X:

$$X := ||X_{ij}||_{nxp} \text{ t.c. } x_{ij} := \text{j-esima componente di } X_j \text{ dunque :}$$
$$\text{X} := [X_1, X_2, ..., X_n]$$

Naturalmente, i valori delle varie features di ciascun vettore X_j potrebbero spesso avere unità di misura differenti influenzando, come vedremo, il calcolo delle distanza o delle similarità tra gli individui; per tale ragione, è possibile definire una *matrice dei dati normalizzata*, sottraendo a ciascun valore $x_{i,j}$ della j-esima feature la sua media campionaria x_j e dividendola per la sua deviazione standard s_j :

$$\text{Z:= } ||z_{ij}||_{nxp} \text{ t.c. } z_{ij} := (x_{ij} - \overline{x}_j)/s_j$$

Misure di Distanza:
La **distanza** tra due individui I_r e I_s è definita sulla base di una funzione di distanza o metrica:

$$\text{d: I x I} \longrightarrow \text{R}$$
$$(I_r, I_s) \longrightarrow d(I_r, I_s)$$

1. $\text{d}(I_r, I_s) \geq 0$ (non negatività);
2. $\text{d}(I_r, I_s) = d(I_s, I_r)$ (simmetria);
3. $\text{d}(I_r, I_s) = 0 \longleftrightarrow I_r = I_s$ (Identità);
4. $\text{d}(I_r, I_s) \leq d(I_r, I_t) + d(I_t, I_s), \forall I_t \in I$ (Diseguaglianza Triangolare)

Per convenzione diremo che la distanza tra due individui I_r, I_s è pari alla distanza tra i vettori che li rappresentano :

$$d(I_r, I_s) := d(X_r, X_s)$$

Esistono diverse tipologie di metriche :

- *Distanza Euclidea* : $d_2(X_i, X_y) := (\sum_{i=1}^{p}(x_{ik} - x_{jk})^2)^{1/2}$
- *Distanza Manhattan* : $d_2(X_i, X_y) := \sum_{i=1}^{p} |v_i - w_i|$
- *Metrica del massimo (Chebycev)* : $d_\infty(X_i, X_y) := max\{|v_i - w_i| : i \in \{1, 2, ..., p\}\}$
- *Metrica di Minkowsky* : $d_r(X_i, X_y) := (\sum_{i=1}^{p}(v_i - w_i)^r)^1/r$
- *Distanza di Jaccard* : $d(X_i, X_y) := 1 - \frac{\sum_{i=1}^{p} min(v_i, w_i)}{\sum_{i=1}^{p} max(v_i, w_i)}$

5.1.2 Matrice delle distanze

Sia d una *metrica* e sia X *la matrice degli individui*; definiamo **matrice delle distanze** di x :

$$D_X := ||d_{ij}||_{nxn} \text{ t.c. } d_{ij} := d(X_I, X_J)$$

Passo preliminare ai metodi di clustering, in particolare per i metodi di clustering gerarchico, è la generazione della matrice delle distanze degli elementi in analisi. La metrica selezionata in questo caso è la metrica euclidea.
Questa metrica è fortemente influenzata dall'unità di misura dei dati, ma nel caso preso in esame non si è però effettuata alcuna scalatura dei dati dal momento che nel dataset utilizzato i dati sono stati riportati in percentuale.
La matrice viene calcolata con il metodo *dist* a cui sono passati come parametri la matrice dei dati, la metrica tramite *method='euclidean'* e per risparmiare spazio, dal momento che la matrice è simmetrica e la diagonale contiene tutti 0, i dati ridondanti sono ignorati tramite i parametri *diag* e *upper* uguali a FALSE.

```
d <- dist(X, method='euclidean', diag=FALSE, upper=FALSE)
```

5.1.3 Misure di Similarità

Una **similarità** è una funzione :

$$s : IxI \longrightarrow R$$
$$(I_r, I_t) \longrightarrow s(I_r, I_t)$$

1. $0 \leq d(I_r, I_t) \leq 1$ (Normalizzazione)
2. $s(I_r, I_t) = s(I_t, I_r)$(Simmetria)
3. $d(I_r, I_t) = 1 \iff I_r = I_t$ (Identità)

La mancanza della disuguaglianza triangolare tra le proprietà di una similarità fa sì che una distanza possa essere convertita in una similarità, ma **non viceversa.**

5.1.4 Matrice delle covarianze e misure di non-omogeneità

Matrice Covarianze :
Data la matrice degli individui X; definiamo la **matrice delle covarianze** :

$$W_X := ||w_{r,s}||_{nxn} \text{ t.c. } w_{r,s} := c_{r,s}$$

con $c_{r,s}$ *varianza campionaria* delle features r-esima e s-esima.
Si può inoltre notare come tale matrice abbia le varianze campionarie di ciascuna feature sulla diagonale :

$$w_{i,i} = c_{i,i} = s_i^2$$

Matrice e Misure di non-omogeneità :
Sia G $\subseteq$ I, un sottoinsieme di individui; definiamo allora **matrice di non-omogeneità** la matrice delle covarianze rispetto al sottoinsieme G moltiplicata per (t - 1), dove t = $|G|$:

$$H_G := ||h_{r,s}||_{pxp} \text{ t.c. } h_{r,s} := \sum_{i=1}^{n}(x_{i,r} - \overline{x_r})(x_{i,s} - \overline{x_s}) = (t-1) \cdot c_{r,s}$$

La traccia della matrice di non omogeneità, definisce la misura di non omogeneità del sottoinsieme C:

$$tr(H_G) := \sum_{i=1}^{p} h_{i,i} = (t-1) \cdot \sum_{i=1}^{p} s_i^2$$

Siano $G_1, G_2, ..., G_k$ partizioni di I; definiamo allora **matrice di non omogeneità totale**:

$$T := H_I$$

mentre definiamo **matrice di non omogeneità interna**:

$$\text{S} := H_{G1} + H_{G2} + ... + H_{Gk}$$

infine definiamo **matrice di non omogeneità between (o intra-cluster)**:

$$B := T - S$$

Sulla base di queste tre matrici, definiamo quindi:

- **tr(T)** misura di non omogeneità *totale*
- **tr(S)** misura di non omogeneità *interna*
- **tr(B)** misura di non omogeneità *between*

Un clustering ideale dovrebbe minimizzare tr(S) e massimizzare tr(B).
Vista la linearità dell'operatore tr, risulta che:

$$\begin{aligned} \text{tr(T)} &= \text{tr(S)} + \text{tr(B)} \text{ da cui si ottiene} \\ 1 &= \text{tr(S)/tr(T)} + \text{tr(B)/tr(T)} \end{aligned}$$

una possibile metrica normalizzata per valutare un clustering è il rapporto tr(B)/tr(T), che dovrebbe essere sempre superiore a 0.7.

5.2 Metodo di enumerazione completa

Il metodo più banale per individuare un clustering ottimale sarebbe quello dell'enumerazione completa di tutte le possibili partizioni di I, scegliendo quella più performante; questo metodo però risulta impraticabile, in quanto questo approccio è particolarmente inefficiente dal punto di vista prestazionale.

5.3 Metodi non gerarchici

I metodi non gerarchici sono caratterizzati da un procedimento che mira a ripartire direttamente le n unità in cluster, fornendo come prodotto finale una sola partizione delle n osservazioni. Nel caso dei metodi gerarchici il numero di cluster viene deciso a priori, come ad esempio nel metodo del **k-means**, descritto in seguito.

1. Fissato il numero k di cluster da ottenere, si formano k cluster singleton contenenti altrettanti punti di riferimento $C_1, C_2, ..., C_k$, scelti casualmente tra gli individui;
2. Forma k nuovi cluster $G_1, G_2, ..., G_k$, assegnando ogni individuo I_j il cluster G_i con il centroide C_i più vicino, ovvero t.c. min $d(X_j, C_i)$;
3. Ricalcola i centroidi di ciascun cluster $C_1, C_2, ..., C_k$, tenendo conto dei nuovi elementi aggiunti ad ognuno di essi;
4. Ripeti dal punto 2 finché non ci sono più cambiamenti nei centroidi, finché il loro cambiamento non supera un ϵ prestabilito, o finché non si raggiunge un numero massimo di iterazioni (a seconda di come si imposta l'algoritmo);

Se si esegue l'algoritmo in modo approssimato (scegliendo una soglia ϵ o impostando un numero massimo di iterazioni), le performance del K-Means sono molto influenzate dalla scelta dei punti di riferimento iniziali; per tale ragione, spesso si tende a ripetere tale scelta più volte, scegliendo poi la migliore partizione individuata sulla base della misura di non omogeneità intra-cluster.

5.3.1 Applicazione K-Means

La prima scelta che abbiamo affrontato nell'analisi dei cluster è stata la scelta del numero corretto di gruppi da considerare.
Tenendo conto dei risultati ottenuti nelle analisi precedenti, possiamo ipotizzare che la maggior parte dei paesi contribuiscano in maniera limitata al bilancio complessivo, per questo motivo un cluster sarà molto numeroso.
Dal momento che gli individui sono 27 scegliere un numero di cluster k troppo ridotto non descriverebbe correttamente i nostri dati, così come un numero troppo grande.
Per selezionare un numero di classi che descrivesse la complessità dei nostri dati, si è proceduto inizialmente eseguendo il **k-means** incrementando di volta in volta il numero di cluster e analizzando i risultati ottenuti.
Il metodo è stato applicato con **k**, ossia il numero di cluster da trovare, compreso tra 2 e 9, modificando il parametro **iter.max**, che indica il numero massimo di iterazioni a 50 e **nstart**, ossia il numero di insiemi di centroidi iniziali da generare randomicamente, pari a 10.

```
km <- kmeans(X, k=K, iter.max = 50, nstart = 10)
```

Nella Tabella 3 sono riportati i risultati ottenuti dalle varie esecuzioni del metodo.

k	trB/trT	#Elementi/Cluster	Cluster
2	51.1 %	2, 25	$G_1 = \{DE,FR\}$ $G_2 = \{AT, BE, BG, CY, CZ, DK, EE, EL, ES, FI, HR, HU, IE, IT, LT, LU, LV, MT, NL, PL,PT, RO, SE, SI, SK\}$
3	75.1 %	24, 1, 2	$G_1 = \{AT, BE, BG, CY, CZ, DK, EE, EL, ES, FI, HR, HU, IE, IT, LT, LU, LV, MT, NL, PT, RO, SE, SI, SK\}$ $G_2 = \{FR\}$ $G_3 = \{DE, PL\}$
4	84.9 %	1, 21, 4, 1	$G_1 = \{FR\}$ $G_2 = \{AT, BE, BG, CY, CZ, DK, EE, EL, FI, HR, HU, IE, LT, LU, LV, MT,PT, RO, SE, SI, SK\}$ $G_3 = \{ES,IT,NL,PL\}$ $G_4 = \{DE\}$
5	92.7 %	1, 1, 1, 3, 21	$G_1 = \{DE\}$ $G_2 = \{PL\}$ $G_3 = \{FR\}$ $G_4 = \{ES, IT, NL\}$ $G_5 = \{AT, BE, BG, CY, CZ, DK, EE, EL, FI, HR, HU, IE, LT, LU, LV, MT, PT, RO, SE, SI, SK\}$
6	95.0 %	6, 3, 1, 1, 1, 15	$G_1 = \{AT, BE, CZ, FI, RO, SE\}$ $G_2 = \{ES, IT, NL\}$ $G_3 = \{FR\}$ $G_4 = \{DE\}$ $G_5 = \{PL\}$ $G_6 = \{BG, CY, DK, EE, EL, HR, HU, IE, LT, LU, LV, MT, PT, SI, SK\}$
7	96.4 %	6, 1, 15, 1, 1, 2, 1	$G_1 = \{AT, BE, CZ, FI, RO, SE\}$ $G_2 = \{PL\}$ $G_3 = \{BG, CY, DK, EE, EL, HR, HU, IE, LT, LU, LV, MT, PT, SI, SK\}$ $G_4 = \{FR\}$ $G_5 = \{DE\}$ $G_6 = \{ES,NL\}$ $G_7 = \{IT\}$

k	trB/trT	#Elementi/Cluster	Cluster
8	97.3 %	1, 2, 1, 1, 5, 15, 1, 1	$G_1 = \{FR\}$ $G_2 = \{ES,NL\}$ $G_3 = \{PL\}$ $G_4 = \{SE\}$ $G_5 = \{AT, BE, CZ, FI, RO\}$ $G_6 = \{BG, CY, DK, EE, EL,HR, HU, IE, LT, LU, LV, MT, PT, SI, SK\}$ $G_7 = \{IT\}$ $G_8 = \{DE\}$
9	98.0 %	8, 1, 1, 1, 1, 3, 1, 1, 10	$G_1 = \{AT, BG, DK,FI,HU, PT, RO,SK\}$ $G_2 = \{SE\}$ $G_3 = \{IT\}$ $G_4 = \{DE\}$ $G_5 = \{PL\}$ $G_6 = \{BE, ES, NL\}$ $G_7 = \{FR\}$ $G_8 = \{CZ\}$ $G_9 = \{CY, EE, EL,HR, IE, LT, LU, LV, MT,SI\}$

Tabella 3: Risultati esecuzione k-means al variare di k

Come trade-off tra il rapporto tra misura di omogeneità tra i cluster e la misura di omogeneità totale, tenendo conto del tipo di raggruppamenti, si è optato per la scelta di k = 6.
Con un numero inferiore si ottiene un raggruppamento in un cluster principale molto corposo che non tiene conto correttamente dei comportamenti all'interno del gruppo di paesi che producono intorno alla media.
Con un numero superiore si vanno a lasciare grandi numeri di singleton o di gruppi molto piccoli con pochi gruppi molto grandi, questo tipo di rappresentazione va a dettagliare troppo le singole differenze tra i paesi facendo perdere informazioni.
Un comportamento che abbiamo notato nella formazione dei cluster è quello di isolare in classi i paesi con comportamenti anomali in almeno una categoria.
Nel caso di k = 6, si riescono a raggruppare dei paesi che hanno outlier simili come **Italia**, **Spagna** e **Olanda**, lasciando isolati **Polonia** (outlier in solid fossil fuel), **Germania** (outlier in rinnovabili e solid fossil fuel) e **Francia** (outlier nel nucleare).
Inoltre, rispetto al caso di k = 5, si va a creare un altro cluster di produttori nella fascia medio-alta, scorporato dal gruppo più grande di paesi che producono poco, o tanto da poche fonti o da tante fonti ma in minore quantità.
Nella Figura 5.3.1 è riportata l'esecuzione del metodo k-means con k=6.

```
> km
K-means clustering with 6 clusters of sizes 6, 1, 1, 3, 15, 1

Cluster means:
  Solid fossil fuels Natural gas Oil and petroleum products Renewables and biofuels Non-renewable waste Nuclear heat
1          2.8254717   2.1812318                 2.6105333              4.2900409          3.7901727    3.4564331
2          3.7763968  10.6670715                13.8563047             11.9200206         11.4325370   52.6393381
3         29.1589469   5.3313572                 6.2507877              5.3970687          7.4849329    0.0000000
4          2.9230369  12.0008288                 9.7190072              7.7034706          5.8812914    3.0692894
5          0.6373617   0.8071432                 0.9258834              0.9514439          0.7579377    0.5300803
6         31.7822903  22.8045455                21.1844360             19.5605955         29.3285540    9.4629901

Clustering vector:
AT BE BG CY CZ DE DK EE EL ES FI FR HR HU IE IT LT LU LV MT NL PL PT RO SE SI SK
 1  1  5  5  1  6  5  5  5  4  1  2  5  5  5  4  5  5  5  5  4  3  5  1  1  5  5

Within cluster sum of squares by cluster:
[1] 155.45686   0.00000   0.00000 157.03667  55.85442   0.00000
 (between_SS / total_SS =  95.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

Figura 13: Oggetto risultante dall'esecuzione del metodo kmean

5.4 Metodi gerarchici

Una famiglia di algoritmi che risolvono in modo efficiente il problema del clustering è quella dei **metodi gerarchici**, che si basano sull'idea di creare n partizioni differenti, costruendo però la partizione dell'iterazione i basandosi induttivamente su quella dell'iterazione precedente i -1, sfruttando una metrica che permette di definire una distanza tra cluster (indicheremo quest'ultima con $d(G_u, G_v) = d_{u,v}$).
Esistono, in generale, due tipologie di metodi gerarchici:

- **Divisivi**, che all'iterazione 1 partono con 1 solo cluster che comprende tutti gli individui $G_1 = I$, per poi ad ogni iterazione i dividere i due sottoinsiemi più distanti tra loro in cluster distinti, fino a formare n sottoinsiemi singleton all'ultima iterazione;

- **Agglomerativi**, che operano come segue:

  1. Alla prima iterazione, crea n cluster singleton $G_1 = I_1, G_2 = I_2, ..., G_n = I_n$;
  2. ad ogni iterazione i, calcola i due cluster G_u, G_v più vicini, sulla base della matrice delle distanze D dell'iterazione precedente, e li unisce in un nuovo cluster G_{uv};
  3. calcola la distanza $d(G_{uv}, G_w)$ per ogni altro cluster G_w, ed aggiorna la matrice delle distanze D inserendole;

Uno dei principali svantaggi dei metodi gerarchici per il clustering è l'impossibilità di riassegnare un individuo ad un cluster differente che, col procedere delle iterazioni, si scopre essere più adatto a quest'ultimo. In questi casi viene utilizzato il metodo non gerarchico del kmeans, come descritto nella sezione precedente.

5.4.1 Single-Linking

Il metodo **single-linking** (o metodo del legame singolo) definisce la distanza tra due cluster G_u e G_v come la minima distanza tra due elementi di tali cluster; supponendo di aggregare due cluster G_u e G_v, per ogni altro cluster G_z:

$$d_{uv,z} = \min d_{u,z}, d_{v,z}$$

Vantaggi

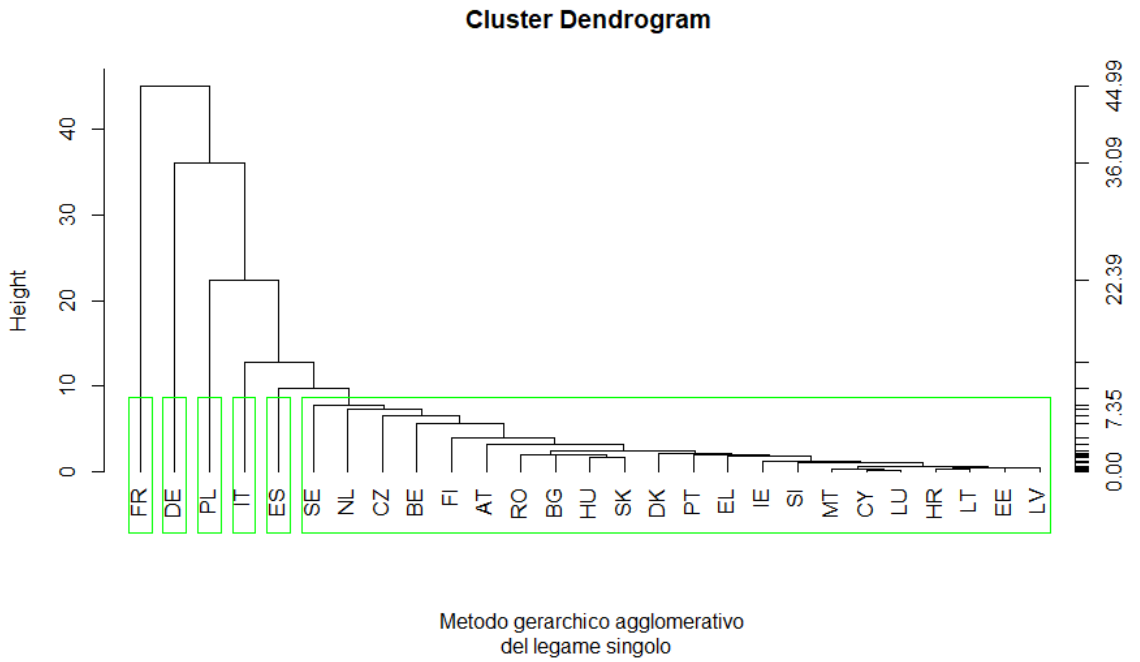
- Individua cluster di ogni forma e dimensione;
- Mette facilmente in evidenza gli outlier isolandoli;
- Favorisce la differenziazione tra gruppi.

Svantaggi

- Considerare solo la distanza tra gli elementi più vicini dei due cluster può portare ad un effetto "catena", per cui vengono agglomerati spesso cluster contenenti individui molto differenti tra loro; questo può avvenire specialmente in presenza di cluster non ben separati, o con outlier che risultano essere vicini tra loro.

5.4.2 Applicazione metodo del legame singolo

Per k = 6, i cluster ottenuti sono di cardinalità 1, 1, 1, 1, 1, 22. I cluster ottenuti sono i seguenti:
$G_1 = \{FR\}$
$G_2 = \{DE\}$
$G_3 = \{PL\}$
$G_4 = \{IT\}$
$G_5 = \{ES\}$
$G_6 = \{SE, NL, CZ, BE, FI, AT, RO, BG, HU, SK, DK, PT, EL, IE, SI, MT, CY, LU, HR, LT, EE, LV\}$
Questa clusterizzazione ha come rapporto tra la misura di omogeneità tra i cluster (*tr(B) = 6808.131*) e la misura di omogeneità totale (*tr(T) = 7315.925*) in percentuale è uguale a (tr(B)/tr(T))*100 = 93.06%.
Il dendogramma risultante dall'esecuzione è nella figura successiva in cui sono evidenziati i sei cluster.



I cluster sono distinti rispetto al metodo del k-means e si vanno a creare cinque cluster da un singolo elemento che non rendono la descrizione delle caratteristiche molto significativa.
In termini di metriche la bontà della clusterizzazione è comunque alta ma ridotta rispetto al metodo non gerarchico (95%).

5.4.3 Metodo Full-linking

Il metodo **full-linking** (o del legame completo) definisce la distanza tra due cluster G_u e G_v come la massima distanza tra due elementi di tali cluster; supponendo di aggregare due cluster G_u e G_v, per ogni altro cluster G_z:

$$d_{uv,z} = \max d_{u,z}, d_{v,z}$$

Vantaggi

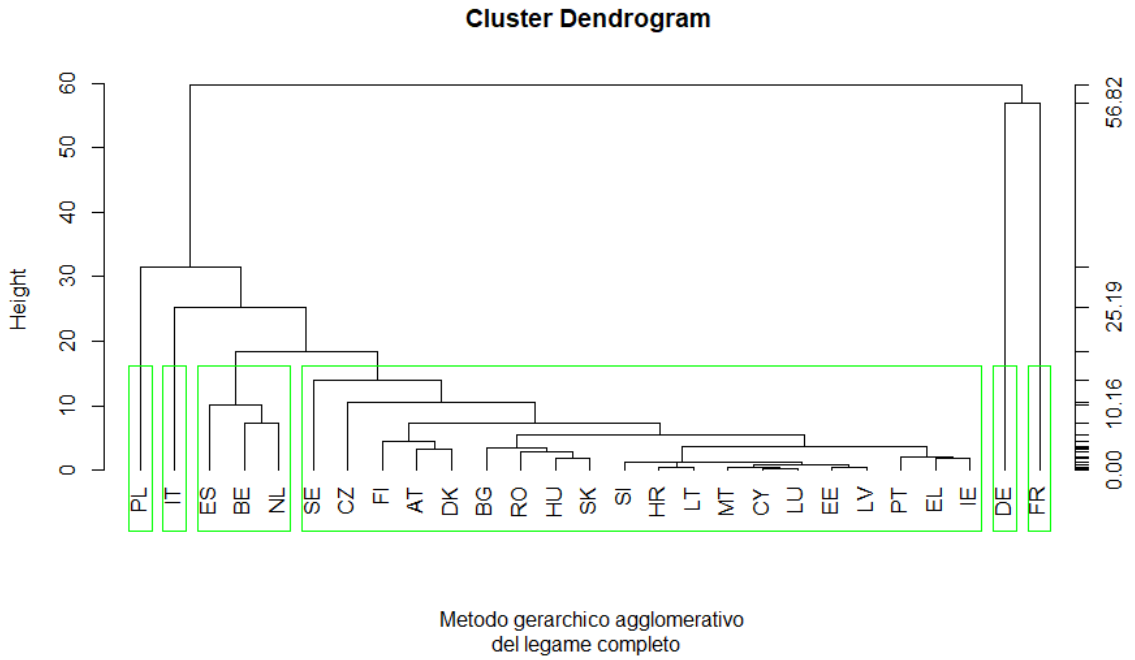
- Favorisce la densità interna dei cluster;
- Considerare la distanza massima tra gli individui consente di individuare cluster anche in presenza di outlier molto vicini tra essi, evitando quindi l'effetto a catena caratteristico del single-linking;

Svantaggi

- Non favorisce la differenziazione tra i cluster;
- Ha migliori performance in presenza di cluster dalla forma ellissoidale;

5.4.4 Applicazione metodo del legame completo

Per k = 6, i cluster ottenuti sono di cardinalità 1, 1, 3, 20, 1, 1.
I cluster ottenuti sono:
$G_1 = \{PL\}$
$G_2 = \{IT\}$
$G_3 = \{ES, BE, NL\}$
$G_4 = \{SE, CZ, FI, AT, DK, BG, RO, HU, SK, SI, HR, LT, MT, CY, LU, EE, LV, PT, EL, IE\}$
$G_5 = \{DE\}$
$G_6 = \{FR\}$
La metrica per valutare la bontà della clusterizzazione ossia tr(B)/tr(T) è pari al 94.40%.
Nella figura successiva è riportato il dendogramma e sono evidenziati i sei cluster ottenuti.



I cluster sono distinti rispetto al metodo non gerarchico ma risulta essere comunque più rappresentativo rispetto al metodo del legame singolo, come evidenziato anche dalle metriche ottenute. Infatti è migliore del legame singolo e ha una percentuale molto simile con il k-means.

5.4.5 Mean-Linking

Il metodo **mean-linking** (o del legame medio) definisce la distanza tra due cluster G_u e G_v come la media aritmetica delle distanze tra tutte le coppie di elementi in tali cluster;

supponendo di aggregare due cluster G_u e G_v, per ogni altro cluster G_z, si dimostra che:

$$d_{uv,z} = \frac{n_u}{(n_u + n_v)} \cdot d_{uz} + \frac{n_v}{(n_u + n_v)} \cdot d_{vz}$$

con n_u ed n_v cardinalità rispettivamente di G_u e G_v.
Vantaggi

- Individua cluster di qualsiasi forma e dimensione;
- Rappresenta un buon compromesso tra la differenziazione tra cluster e densità interna.
- $d_{uv,z}$ viene influenzata dalla dimensione dei cluster G_u e G_v, dunque se ad esempio $n_u >> n_v$ allora $d_{v,z}$ ha poca importanza nell'agglomerazione, dunque eventuali outlier in cluster piccoli hanno poca influenza;

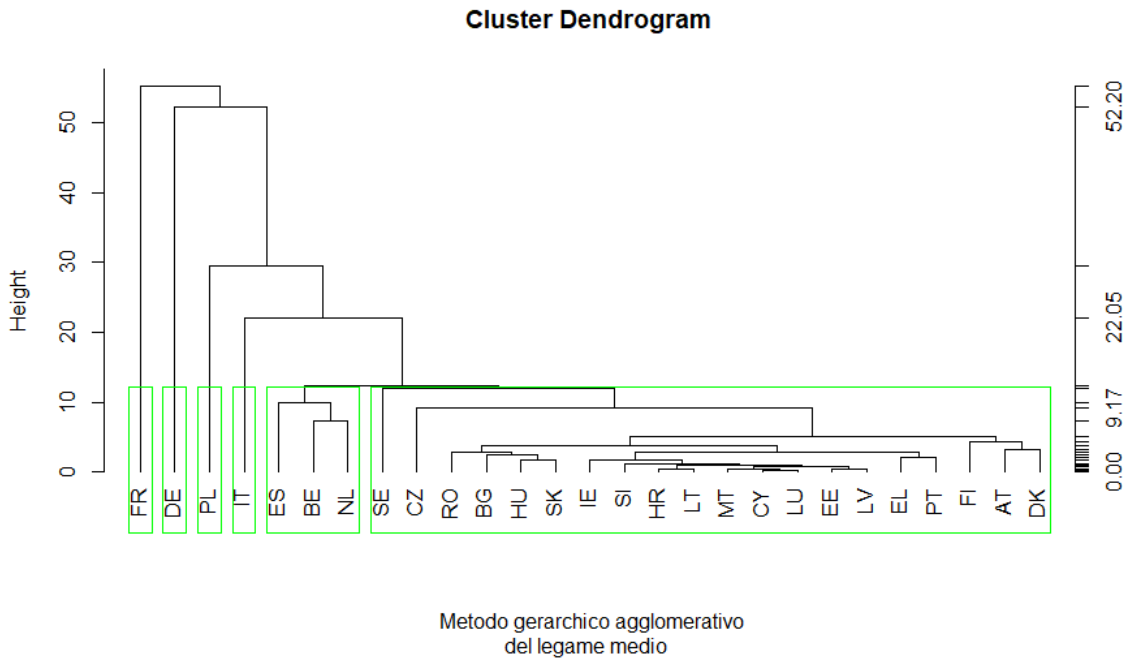
Svantaggi

- Essendo d_{uv} una media aritmetica, può essere influenzata dalla presenza di eventuali outlier in cluster molto numerosi, portando talvolta ad agglomerare elementi molto distanti, anche se non quanto il single-linking;
- I cluster piccoli tendono ad essere inglobati in quelli più ampi per via dell'influenza di n_u ed n_v su $d_{uv,z}$;

5.4.6 Applicazione del metodo del legame medio

Per k = 6, i cluster ottenuti sono di cardinalità 1, 1, 1, 1, 3, 20.
I cluster ottenuti sono:
$G_1 = \{FR\}$
$G_2 = \{DE\}$
$G_3 = \{PL\}$
$G_4 = \{IT\}$
$G_5 = \{ES, BE, NL\}$
$G_6 = \{SE, CZ, RO, BG, HU, SK, IE, SI, HR, LT, MT, CY, LU, EE, LV, EL, PT, FI, AT, DK\}$

Nella figura successiva è descritto il dendogramma e i sei cluster ottenuti.



Esso fornisce gli stessi cluster del metodo del legame completo.
Infatti tr(B)/tr(T) è pari al 94.40% e valgono le stesse considerazioni del metodo precedente.

5.4.7 Metodo del Centroide

Sia $G_u \subseteq I$; Il **centroide** di G_u è definito come la media campionaria dei suoi individui :

$$\overline{x}_u = \sum_{j \in G_u} x_j$$

Il metodo del **centroide** definisce la distanza tra due cluster G_u e G_v come il quadrato della distanza euclidea tra i rispettivi centroidi:

$$d_{u,v} := d_2^2(\overline{x}_u, \overline{x}_v)$$

si dimostra che, agglomerando i cluster G_u e G_v in $G_{u,v}$, il suo centroide viene dato da :

$$\overline{x}_{uv} = \frac{n_u}{(n_u + n_v) \cdot \overline{x}_u + \frac{n_v}{(n_u + n_v)} \cdot \overline{x}_v}$$

mentre la sia distanza da un cluster G_z è :

$$d_{uv,z} = \frac{n_u}{(n_u + n_v)} \cdot d_{uz}^2 + \frac{n_v}{(n_u + n_v)} \cdot d_{vz}^2 + \frac{n_u n_v}{(n_u + n_v)^2} \cdot d_{u,v}^2$$

con n_u ed n_v cardinalità di G_u e G_v

Vantaggi

- Rappresenta, come il mean-linking, un buon compromesso tra la densità interna dei cluster e la differenziazione tra questi ultimi;
- Ancora una volta, $d_{uv,z}$ viene influenzata dalla dimensione dei cluster G_u e G_v, dunque se ad esempio $n_u >> n_v$ allora $d_{v,z}$ ha poca importanza nell'agglomerazione, dunque eventuali outlier in cluster piccoli hanno poca influenza.

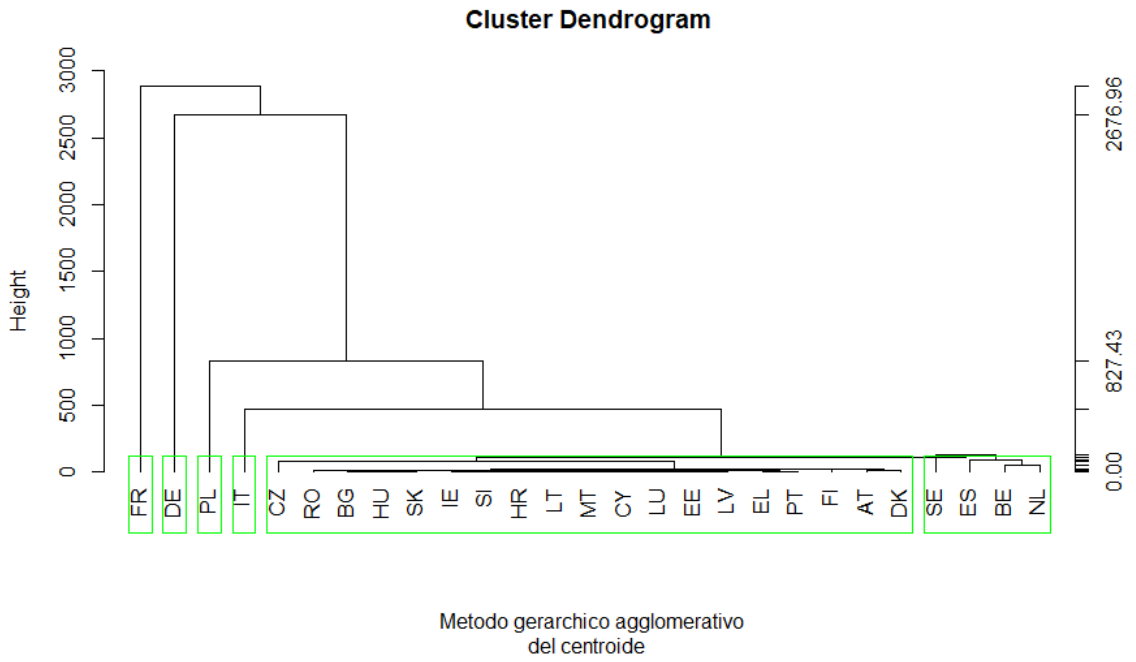
Svantaggi

- Come nel mean-linking, la media aritmetica può essere influenzata dalla presenza di eventuali outlier in cluster numerosi, portando talvolta ad unire elementi molto distanti in un unico cluster;
- Anche in questo caso i cluster poco numerosi tendono ad essere inglobati in quelli più grandi, a causa dell'influenza di n_u ed n_v su $d_{uv,z}$;

5.4.8 Applicazione metodo del centroide

Per k = 6, i cluster ottenuti sono di cardinalità 1, 1, 1, 1, 19, 4.
I cluster ottenuti sono:
$G_1 = \{FR\}$
$G_2 = \{DE\}$
$G_3 = \{PL\}$
$G_4 = \{IT\}$
$G_5 = \{CZ, RO, BG, HU, SK, IE, SI, HR, LT, MT, CY, LU, EE, LV, EL, PT, FI, AT, DK\}$
$G_6 = \{SE, ES, BE, NL\}$

In questo caso tr(B)/tr(T) = 94.88%, è molto simile al metodo non gerarchico.
Nella figura successiva è descritto il dendogramma risultante ed evidenziati i cluster ottenuti.



Questo metodo risulta essere il più efficace dei metodi gerarchici ed evidenzia un risultato simile al metodo del k-means ma permette di ricollocare un elemento del gruppo principale in un cluster minore (Svezia).

5.4.9 Metodo della Mediana

Come il metodo del centroide, quello della **mediana** definisce la distanza tra due cluster come il quadrato della metrica euclidea tra i rispettivi centroidi, tuttavia quando G_u e G_v vengono agglomerati in G_{uv}, calcola il nuovo centroide come media aritmetica tra i due precedenti, non tenendo conto della dimensione dei due cluster:

$$\overline{x}_{uv,z} = \frac{1}{2} \cdot (\overline{x}_u + \overline{x}_v)$$

questo, si dimostra, porta la distanza tra G_{uv} ed un cluster G_z a non essere a sua volta influenzata dalla dimensione dei cluster:

$$d_{uv,z} = \frac{1}{2} \cdot d^2_{u,z} + \frac{1}{2} \cdot d^2_{v,z} - \frac{1}{4} \cdot d^2_{u,v}$$

Vantaggi

- Non essendo la distanza influenzata dalla numerosità dei cluster, i sottoinsiemi poco numerosi non tendono sempre ad essere inglobati da quelli più grandi;

- I risultati sono meno influenzati dagli outlier presenti in cluster numerosi, non utilizzando la media campionaria degli individui.

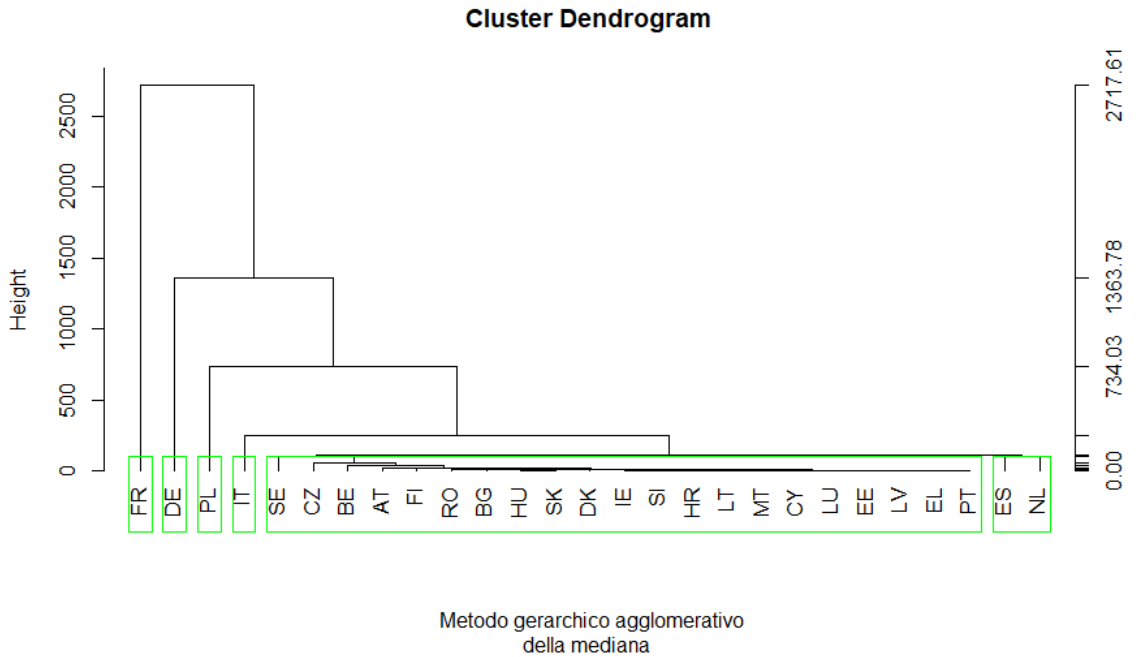
Svantaggi

- Eventuali outlier presenti in cluster piccoli potrebbero influenzare i risultati o portare ad effetti a catena simili a quelli del single-linking.

5.4.10 Applicazione del metodo della mediana

Per k = 6, i cluster ottenuti hanno cardinalità 1, 1, 1, 1, 21, 2.
I cluster ottenuti sono i seguenti:
$G_1 = \{FR\}$
$G_2 = \{DE\}$
$G_3 = \{PL\}$
$G_4 = \{IT\}$
$G_5 = \{SE, CZ, BE, AT, FI, RO, BG, HU, SK, DK, IE, SI, HR, LT, MT, CY, LU, EE, LV, EL, PT\}$
$G_6 = \{ES, NL\}$

In questo caso, il rapporto tr(B)/tr(T) = 94.15% e il comportamento è simile al metodo del legame singolo.
Nella figura successiva è descritto il dendogramma risultante ed evidenziati i cluster ottenuti.



Per poter operare al meglio il confronto tra i vari metodi di clustering utilizzati, sono state riportate le performance nella tabella 4:

Metodo	**Performance**
Legame singolo	93.06%
Legame completo	94.40%
Legame medio	94.40%
Centroide	94.87%
Mediana	94.15%
K-means	**95.0%**

Tabella 4: Confronto tra le metodologie in base a **tr(B)/tr(T)**

5.5 Evoluzioni dell'analisi

Dai risultati dei cluster precedenti si nota come alcuni paesi siano particolarmente diversi per produzione e si è ipotizzato di poterli analizzare a prescindere dagli altri. In particolare, per migliorare i risultati della clusterizzazione, si è pensato di eliminare i due paesi più anomali, ossia la Francia e la Germania.
Inoltre, dal momento che la colonna riguardante la produzione di energia tramite la fonte nucleare è principalmente rappresentata dalla Francia, si è ritenuto opportuno omettere tale colonna dall'analisi dal momento che sarebbe stata principalmente vuota o con valori molto vicini allo zero per tutti gli altri paesi.
Per questo motivo, si è filtrato il dataset iniziale e si è ottenuto un oggetto *dataFrameCluster* in cui sono presenti 25 paesi e 5 fonti energetiche.

5.5.1 K-means

Analogamente all'analisi precedente si è scelto di eseguire il metodo non gerarchico del k-means con 6 centroidi iniziali.

```
1 > km <- kmeans(dataFrameCluster, centers=6 , iter.max = 50 , nstart
    =10)
```

L'oggetto risultante è descritto nella seguente figura.

```
> km
K-means clustering with 6 clusters of sizes 15, 5, 1, 2, 1, 1

Cluster means:
  Solid fossil fuels Natural gas Oil and petroleum products Renewables and biofuels Non-renewable waste
1          0.6373617   0.8071432                  0.9258834              0.9514439           0.7579377
2          3.1833929   2.5396544                  2.6735720              3.2136040           3.1627124
3          3.6307649  17.8174605                  9.9529616             12.2289737           8.3320247
4          2.5691728   9.0925130                  9.6020301              5.4407191           4.6559247
5          1.0358659   0.3891189                  2.2953396              9.6722253           6.9274744
6         29.1589469   5.3313572                  6.2507877              5.3970687           7.4849329

Clustering vector:
AT BE BG CY CZ DK EE EL ES FI HR HU IE IT LT LU LV MT NL PL PT RO SE SI SK
 2  2  1  1  2  1  1  1  4  2  1  1  1  3  1  1  1  1  4  6  1  2  5  1  1

Within cluster sum of squares by cluster:
[1] 42.53345 71.71594  0.00000 18.65433  0.00000  0.00000
 (between_SS / total_SS =  92.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

La clusterizzazione ha dunque prodotto 6 cluster di dimensione 15, 5, 1, 2, 1, 1. I cluster sono descritti nella seguente scomposizione:
$G_1 = \{BG, CY, DK, EE, EL, HR, HU, IE, LT, LU, LV, MT, PT, SI, SK\}$
$G_2 = \{AT, BE, CZ, FI, RO\}$
$G_3 = \{IT\}$
$G_4 = \{ES, NL\}$
$G_5 = \{SE\}$
$G_6 = \{PL\}$
Per quanto riguarda il rapporto tra la misura di non omogeneità tra i cluster e quella totale, ci si attesta intorno al 92.7%.

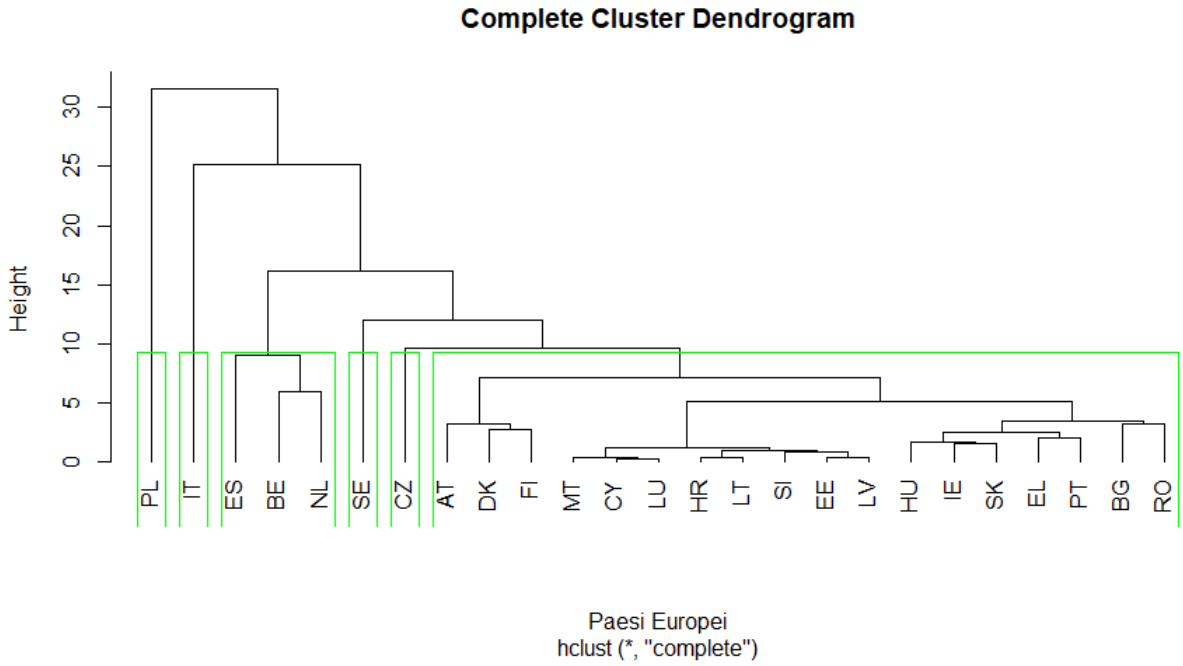
Come per l'analisi precedente non si riduce l'effetto catena dal momento che anche in questo caso, i paesi con la maggiore produzione sono molto diversi tra loro per fonti energetiche e dunque vengono isolati dagli altri paesi tramite le diverse applicazioni dell'analisi dei cluster.

Per completare l'analisi si è scelto di utilizzare anche alcuni metodi gerarchici, in particolare il metodo del legame completo e il metodo del centroide, ossia quelli che hanno performato meglio sui dati precedenti.

5.5.2 Legame completo

Il metodo considerato è il metodo gerarchico agglomerativo del **legame completo**.
Per k = 6, i cluster ottenuti sono di cardinalità 1, 1, 3, 1, 1, 18.
I cluster ottenuti sono:
$G_1 = \{PL\}$
$G_2 = \{IT\}$
$G_3 = \{ES, BE, NL\}$
$G_4 = \{SE\}$
$G_5 = \{CZ\}$
$G_6 = \{AT, DK, FI, MT, CY, LU, HR, LT, SI, EE, LV, HU, IE, SK, EL, PT, BG, RO\}$

La metrica per valutare la bontà della clusterizzazione ossia tr(B)/tr(T) è pari al 91.80%.
Nella successiva è riportato il dendogramma e sono evidenziati i sei cluster ottenuti.

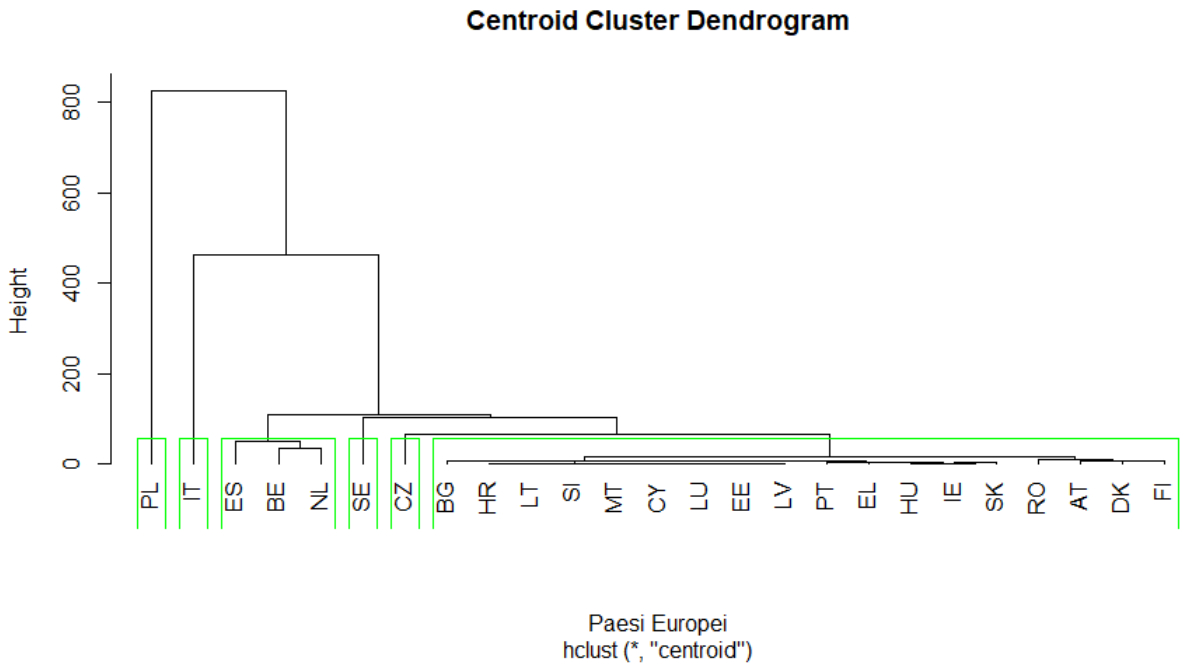


I cluster sono distinti rispetto al metodo non gerarchico e questa particolare suddivisione risulta peggiore anche dal punto di vista della metrica presa in esame, dal momento che aumentano i singleton.

5.5.3 Centroide

Altro metodo considerato è il metodo gerarchico agglomerativo del **centroide**.
Per k = 6, i cluster ottenuti sono di cardinalità 1, 1, 3, 1, 1, 18.
I cluster ottenuti sono:
$G_1 = \{PL\}$
$G_2 = \{IT\}$
$G_3 = \{ES, BE, NL\}$
$G_4 = \{SE\}$
$G_5 = \{CZ\}$
$G_6 = \{BG, HR, LT, SI, MT, CY, LU, EE, LV, PT, EL, HU, IE, SK, RO, AT, DK, FI, \}$
In questo caso tr(B)/tr(T) = 91.80% e si può notare che in questo caso la suddivisione è la stessa ottenuta con il metodo del legame completo.
In particolare però i paesi del cluster più grande sono ordinati in modo differente, quindi con l'aumentare di k potremmo ottenere risultati leggermente diversi.

Nella figura successiva è descritto il dendogramma risultante ed evidenziati i cluster ottenuti.



5.5.4 Osservazioni Finali

In definitiva, nonostante l'eliminazione dei due paesi più anomali rispetto agli altri, ossia Francia e Germania, e l'eliminazione di una colonna che poteva risultare ridondante su molti dei paesi descritti, i risultati sono molto simili se non lievemente peggiori.
Questo fenomeno è molto probabile si sia verificato dal momento che i paesi più produttivi si differenziano molto tra di loro per combinazione delle fonti energetiche e dunque non rendono evidente una separazione in cluster.
Per questo motivo si ritengono soddisfacenti i risultati ottenuti nella precedente analisi.

6 Statistica Inferenziale: Introduzione

La variabile aleatoria selezionata per lo svolgimento di questo elaborato è la variabile aleatoria discreta di Poisson. Il caso di studio preso in analisi che questa variabile permette di descrivere riguarda il numero di goal effettuati per partita durante i mondiali di calcio. Per descrivere questo fenomeno verranno utilizzati come dati di riferimento la somma dei goal effettuati per partita nell'edizione dei mondiali del 2022.
Il campione è strutturato nel seguente modo:

```
> campione <- c(2, 8, 2, 2, 3, 0, 0, 5, 0, 3, 7, 1, 1, 0, 5, 2, 2, 4,
   2, 0, 1, 2, 3, 2, 1, 2, 5, 2, 6, 5, 1, 2, 2, 3, 3, 1, 1, 1, 2, 3,
   0, 3, 3, 6, 2, 3, 5, 1, 4, 3, 4, 3, 2, 5, 0, 7, 2, 4, 1, 3, 3, 2,
   3, 6)
> n <- length(campione)
> n
[1] 64
> freq<-table(campione)
> freq
campione
 0  1  2  3  4  5  6  7  8
 7 10 17 14  4  6  3  2  1
 > mean(campione)
 [1] 2.6875
```

Questo campione verrà utilizzato per effettuare le operazioni di stima dei parametri, la verifica di ipotesi ed infine il test del chi-quadrato per verificare che sia effettivamente descrivibile tramite una popolazione di Poisson.
In questo capitolo verrà utilizzata la media del campione preso in esame come valore di lambda ($\lambda = 2.6875$) per presentare le funzioni fornite da R per descrivere la popolazione di Poisson. Questo calcolo è un processo proprio della stima puntale per la popolazione di Poisson, le cui fondamenta teoriche sono mostrate nel capitolo successivo.

6.1 Descrizione Funzione di distribuzione di Poisson

Questa variabile aleatoria è innanzitutto di tipo discreto, infatti X può assumere un numero finito o al più numerabile di valori $x_1, x_2, \ldots, x_n$con rispettive probabilità $p_X(x_1), p_X(x_2), \ldots, p_X(x_n)$ essendo $p_X(x_1) = P(X = x_1)$.
R permette di determinare per questo tipo di variabili aleatorie e in particolare per la variabile aleatoria discreta di Poisson:

- La **funzione di probabilità** in uno specifico punto o in un insieme di punti (*dpois*);
- La **funzione di distribuzione** in uno specifico punto o in un insieme di punti (*ppois*);
- La **funzione per calcolare i quantili**, tramite la definizione di quantile per una distribuzione di frequenza (*qpois*);

- Una **funzione che simula la variabile** aleatoria mediante la generazione di sequenze di numeri pseudocasuali (*rpois*).

Utilizzeremo queste funzioni per descrivere al meglio la distribuzione di Poisson.
Questo tipo di distribuzione viene spesso utilizzata in contesti che prevedono un conteggio come il numero di arrivi ad un centro di calcolo o in questo caso, il numero di goal totali effettuati in una partita. Ed inoltre viene utilizzata per descrivere eventi rari, come ad esempio il numero di incidenti, inondazioni o terremoti.
Di seguito è riportata la definizione di tale distribuzione:
Una variabile aleatoria X avente funzione di probabilità

$$p_X(x) = P(X = x) = \begin{cases} \dfrac{\lambda^x}{x!} e^{-\lambda} & \text{x = 0,1,.. } (\lambda > 0) \\ \\ 0 & \text{Altrimenti} \end{cases}$$

è detta di **distribuzione di Poisson** di parametro λ.
Nel resto della notazione si utilizzerà $X \sim P(\lambda)$ per indicare il fatto che X sia una variabile aleatoria avente distribuzione di Poisson di parametro λ.
Dalla definizione precedente si ricava:

$$\frac{p_X(x)}{p_X(x-1)} = \frac{\lambda}{x} \text{ (x=1,2,...)}$$

e questo permette di calcolare in modo ricorsivo le probabilità di Poisson, come segue:

$$p_X(0) = e^{-\lambda},$$

$$p_X(x) = \frac{\lambda}{x} p_X(x-1) \text{ (x=1,2,...)}$$

Inoltre una variabile aleatoria di Poisson ha come valore medio $E(X) = \lambda$ e varianza $Var(X) = \lambda$.

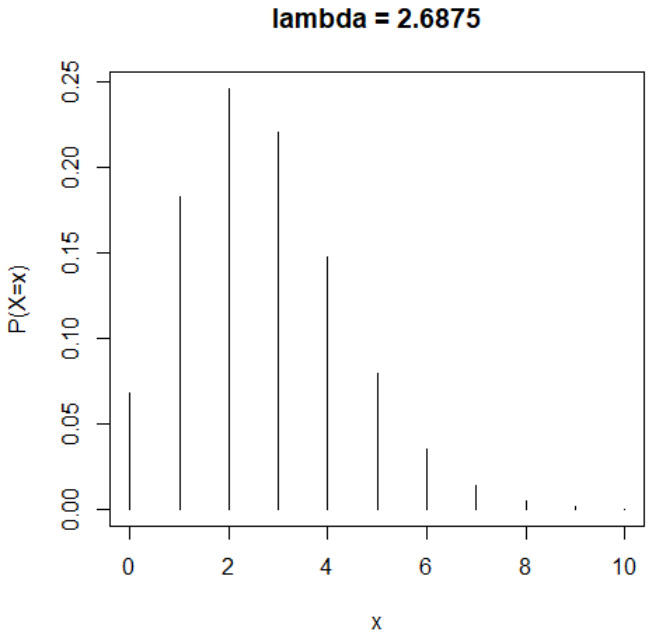
6.2 Calcolo Probabilità Poissoniana

Per calcolare le probabilità di Poisson in R, come accennato in precedenza si utilizza la funzione

dpois(x,lambda)

dove x rappresenta il valore assunto (o l'insieme dei valori assunti) dalla variabile aleatoria considerata e *lambda* il vettore dei valori medi (non negativi).
Nel caso studio preso in esame, supponendo che il numero di goal possa variare tra 0 e 10 e considerando $\lambda = 2.6875$, otteniamo i seguenti risultati che possono essere descritti dal successivo grafico a bastoncini.

```
> x <- 0:10
> lambda <- 2.6875
> distr <- dpois(x,lambda)
> distr
 [1] 0.068050854 0.182886670 0.245753963 0.220154592 0.147916366
   0.079505047 0.035611636 0.013672324
 [9] 0.004593046 0.001371535 0.000368600
```



6.3 Calcolo Funzione Di Distribuizione Poissoniana

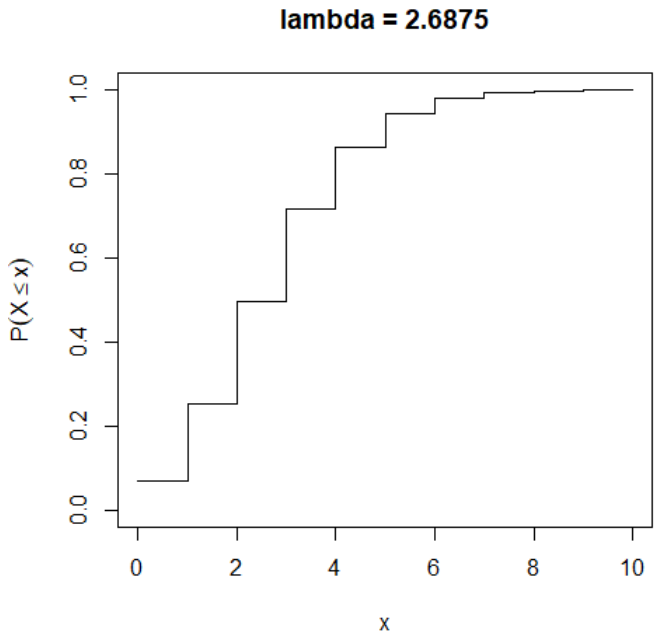
Per calcolare la funzione di distribuzione di Poisson in R si utilizza la funzione

$$ppois(x,\ lambda,\ lower.tail = TRUE)$$

dove x è il valore assunto (o i valori assunti) dalla variabile aleatoria, *lambda* il vettore dei valori medi (non negativi) e se *lower.tail* è TRUE (di default) viene calcolata $P(X <= x)$, altrimenti $P(X > x)$.
Nel nostro esempio, continuando ad utilizzare x e lambda definiti in precedenza, otteniamo i seguenti risultati che sono descritti dal seguente grafico per una funzione di distribuzione discreta.

```
> funz_dist <- ppois(x,lambda)
> funz_dist
[1] 0.06805085 0.25093752 0.49669149 0.71684608 0.86476245 0.94426749
   0.97987913 0.99355145 0.99814450
```

```
[10] 0.99951603 0.99988463
> plot(x,funz_dist, xlab="x" , ylab = expression(P(X<=x)),ylim=c(0,1)
    , type="s" , main = "lambda = 2.6875")
```



6.4 Calcolo Quantili Poissoniani

Per calcolare i quantili (o i percentili) della distribuzione di Poisson si utilizza la funzione

qpois(z, lambda)

dove z è il valore assunto dalle probabilità relative al percentile z*100-esimo e *lambda* il vettore dei valori medi (non negativi). Il risultato è il percentile z*100-esimo, ossia il numero intero k assunto dalla variabile aleatoria di Poisson più piccolo, tale che

$$P(X \leq k) \geq z \text{ (k=0,1,...)}$$

Volendo considerare i quartili del nostro caso studio, otteniamo

```
> z <- c(0,0.25,0.5,0.75,1)
> qpois(z,lambda)
[1]   0   1   3   4 Inf
```

6.5 Simulazione Variabile Poissoniana

E' possibile simulare in R la variabile aleatoria di Poisson generando una sequenza di numeri pseudocasuali tramite la funzione

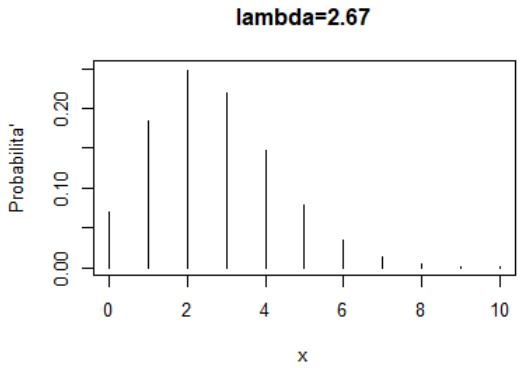
rpois(N, lambda)

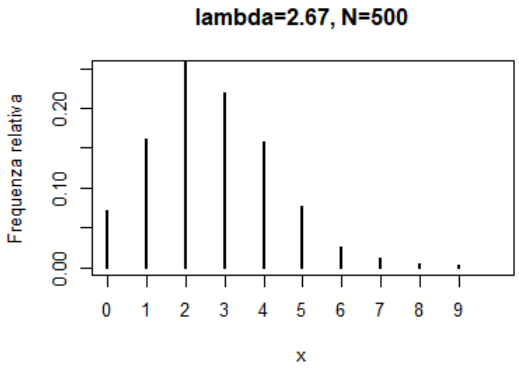
dove N è la lunghezza della sequenza da generare e *lambda* è il vettore dei valori medi (non negativi).
In questo caso, si è definita una sequenza di 64 elementi, ossia il numero di partite eseguite nel contesto di una edizione della competizione dei mondiali di calcio, e si è fornita una descrizione di tale campione tramite le frequenze assolute e relative.

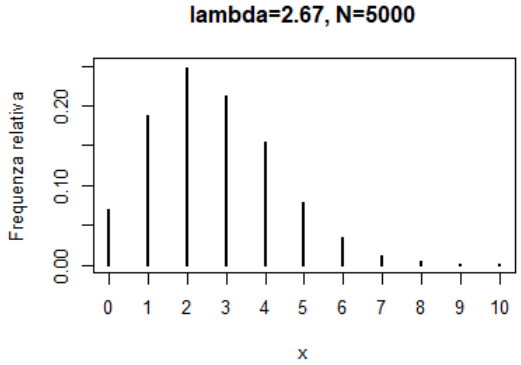
```
> sim <- rpois(64, lambda)
> sim
[1] 3 2 4 3 2 1 0 4 3 4 3 4 1 3 1 2 5 0 2 5 3 1 0 4 5 2 3 1 3 0 1 2 0
[34] 3 2 4 1 2 4 4 0 2 4 3 2 1 1 3 1 6 2 1 2 3 2 5 4 4 1 1 3 0 2 4
> table(sim)
sim
0   1   2   3   4   5   6
7   13  14  13  12  4   1
> table(sim)/length(sim)
sim
0         1         2         3         4         5         6
0.109     0.203     0.218     0.203     0.187     0.062     0.015
```

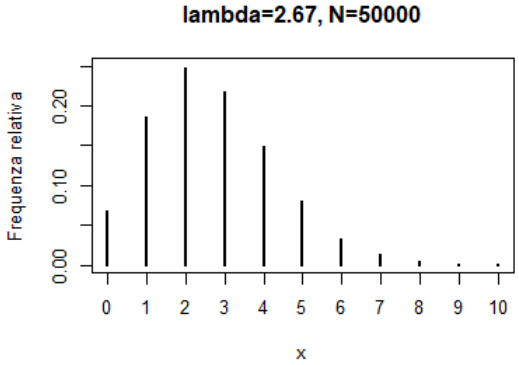
Dal momento che questo campione è abbastanza ridotto, sebbene superiore ai 30 individui, non approssima alla perfezione la funzione di probabilità di Poisson teorica, ma come si può evincere dai grafici all'aumentare della grandezza della sequenza simulata i risultati sono più affidabili.

```
> par(mfrow=c(2,2))
> plot(x, distr, xlab="x", ylab="Probabilita'", type="h", main="lambda
  =2.67", xlim=c(0,10), ylim=c(0,0.25))
> sim1 <- rpois(500, lambda)
> plot(table(sim1)/length(sim1),xlab="x", type="h", ylab="Frequenza
  relativa", xlim=c(0,10), ylim=c(0,0.25), main="lambda=2.67, N=500")
> sim2 <- rpois(5000, lambda)
> plot(table(sim2)/length(sim2),xlab="x", type="h", ylab="Frequenza
  relativa", xlim=c(0,10), ylim=c(0,0.25), main="lambda=2.67, N=5000"
  )
> sim3 <- rpois(50000, lambda)
> plot(table(sim3)/length(sim3), xlab="x", type="h", ylab="Frequenza
  relativa", xlim=c(0,10), ylim=c(0, 0.25), main="lambda=2.67, N
  =50000")
```


lambda=2.67
Probabilita'
0.00
0.10
0.20
0 2 4 6 8 10
x


lambda=2.67, N=500
Frequenza relativa
0.00
0.10
0.20
0 1 2 3 4 5 6 7 8 9
x


lambda=2.67, N=5000
Frequenza relativa
0.00
0.10
0.20
0 1 2 3 4 5 6 7 8 9 10
x


lambda=2.67, N=50000
Frequenza relativa
0.00
0.10
0.20
0 1 2 3 4 5 6 7 8 9 10
x

7 Stima dei parametri

Uno dei problemi della statistica inferenziale, data una variabile aleatoria X osservabile con funzione di distribuzione $F_X(x;\theta_1,\theta_2,\ldots,\theta_k)$ con k parametri non noti, è quello di stimare i parametri tramite una stima puntuale, ossia fornendo come risultato un unico valore reale, o con una stima intervallare, fornendo un intervallo di confidenza in cui collocarli.

7.1 Stima puntuale

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione con funzione di probabilità (nel discreto) oppure densità di probabilità (nel continuo) dove $f(x;\theta_1,\theta_2,\ldots,\theta_k)$ dove $\theta_1,\theta_2,\ldots,\theta_k$ denotano i parametri non noti della popolazione.
Per poter stimare un valore discreto per tali parametri si utilizzano due metodi principali: il *metodo dei momenti* e il *metodo della massima verosimiglianza*.

7.1.1 Metodo dei momenti

Definiamo innanzitutto il concetto di *momento campionario*. Si definisce **momento campionario** r-esimo relativo ai valori osservati $(x_1, x_2, \ldots, x_n)$ del campione casuale il valore

$$M_r(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r \text{ (r = 1,2,..)}$$

Come si può notare il momento campionario r-esimo è la media aritmetica delle potenze r-esime delle n osservazioni effettuate sulla popolazione.
In particolare se r = 1, il momento campionario coincide con il valore della media campionaria $\overline{x}$, ossia $M_1 = (x_1 + x_2 + \cdots + x_n)/n$.
Quindi, nel caso in cui esistano k parametri da stimare, il metodo dei momenti consiste nell'eguagliare i primi k momenti della popolazione in esame con i corrispondenti momenti del campione casuale.
Quindi, se i primi k momenti esistono e sono finiti, tale metodo consiste nel risolvere il sistema di k equazioni

$$E(X^r) = M_r(x_1, x_2, \ldots, x_n) \text{ (r=1,2,..,k)}$$

Le stime dei parametri ottenute con tale metodo, indicate con $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, dipendono dal campione osservato $(x_1, x_2, \ldots, x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli *stimatori*[4] $\hat{\Theta}_1, \hat{\Theta}_2, \ldots \hat{\Theta}_k$ dei parametri non noti $\theta_1, \theta_2, \ldots, \theta_k$ della popolazione, detti *stimatori del metodo dei momenti*.

[4]Uno stimatore $\hat{\Theta} = \mathrm{t}(X_1, X_2, \ldots, X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, \ldots, X_n$ i cui valori possono essere usati per stimare un parametro non noto θ della popolazione. I valori $\hat{\theta}$ assunti da tale stimatore sono detti stime del parametro non noto θ.

Metodo dei momenti su popolazione di Poisson
Si desidera determinare con il metodo dei momenti lo stimatore del valore medio λ di una popolazione di Poisson descritta da una variabile aleatoria $X \sim P(\lambda)$ con funzione di probabilità:

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ x= 0,1,...} (\lambda > 0)$$

Occorre quindi stimare il parametro λ.
Poiché $E(X) = \lambda$, dalla formula precedente otteniamo:

$$\hat{\lambda} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \overline{x}$$

Il metodo dei momenti fornisce quindi come stimatore del parametro $E(X) = \lambda$ la media campionaria.
Per effettuare dunque la stima puntuale del valore medio per la variabile aleatoria in esame consideriamo il campione descritto in precedenza:

```
> stimalambda <- mean(campione)
> stimalambda
[1] 2.6875
```

7.1.2 Stima a massima verosimiglianza

Per definire tale metodo è necessario descrivere la *funzione di verosimiglianza*.
Sia $X_1, X_2, \ldots, X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La **funzione di verosimiglianza**

$$L(\theta_1, \theta_2, \ldots, \theta_k) = L(\theta_1, \theta_2, \ldots, \theta_k; x_1, x_2, \ldots, x_n)$$

del campione osservato $(x_1, x_2, \ldots, x_n)$ è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale $X_1, X_2, \ldots, X_n$, ossia

$$L(\theta_1, \theta_2, \ldots, \theta_k) = L(\theta_1, \theta_2, \ldots, \theta_k; x_1, x_2, \ldots, x_n) = \\ f(x_1; \theta_1, \theta_2, \ldots, \theta_k) f(x_2; \theta_1, \theta_2, \ldots, \theta_k) \ldots f(x_n; \theta_1, \theta_2, \ldots, \theta_k)$$

L'obiettivo del metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\theta_1, \theta_2, \ldots, \theta_k$.
Si cerca di determinare da quale funzione di probabilità congiunta è più verosimile che provenga il campione osservato.
I valori di $\theta_1, \theta_2, \ldots, \theta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $b\theta_1, b\theta_2, \ldots, b\theta_k$; essi costituiscono le stime di massima verosimiglianza dei parametri non noti $\theta_1, \theta_2, \ldots, \theta_k$ della popolazione.

Tali stime dipendono dal campione osservato $(x_1, x_2, \ldots, x_n)$ e quindi al variare dei possibili campioni osservati si ottengono gli stimatori di massima verosimiglianza $\hat{\Theta}_1, \hat{\Theta}_2, \ldots \hat{\Theta}_n$ dei parametri non noti $\theta_1, \theta_2, \ldots, \theta_k$ della popolazione, detti stimatori di massima verosimiglianza.

Stima a massima verosimiglianza su popolazione di Poisson
Si desidera determinare lo stimatore di massima verosimiglianza del valore medio di una popolazione di Poisson descritta da una variabile aleatoria $X \sim P(\lambda)$ con funzione di probabilità

$$P(X = x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ (x=0,1,..)}$$

Essendo $E(X) = \lambda$, si ha

$$L(\lambda) = \frac{\lambda^{x_1+x_2+\cdots+x_n}}{x_1!x_2!\ldots x_n!} e^{-n\lambda}$$

dove le x_i sono numeri interi non negativi. Si nota che

$$logL(\lambda) = log\lambda \sum_{i=1}^{n} x_i - n\lambda - log[x_1!x_2!\ldots x_n!](\lambda > 0)$$

da cui segue

$$\frac{dlogL(\lambda)}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = \frac{n}{\lambda}(\frac{1}{n} \sum_{i=1}^{n} x_i - \lambda)$$

La stima di massima verosimiglianza del parametro λ è

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$$

Quindi, per una popolazione di Poisson lo stimatore di massima verosimiglianza e dei momenti di $E(X) = \lambda$ è la media campionaria $\overline{X}$
Questo stimatore gode della proprietà della correttezza, della varianza minima e della consistenza.
Uno stimatore si dice ***corretto*** quando il valore medio dello stimatore è uguale al parametro da stimare, ossia $E(\hat{\Theta}) = \theta$.
E uno stimatore si dice ***consistente*** se è asintoticamente corretto (ossia è corretto al crescere del campione) e la sua varianza tende a zero al crescere del campione.

Per effettuare dunque la stima puntuale del valore medio per la variabile aleatoria in esame consideriamo il campione descritto in precedenza:

```
> stimalambda <- mean(campione)
> stimalambda
[1] 2.6875
```

7.2 Stima intervallare

Si utilizza questo tipo di stima per determinare due statistiche come limite superiore $\overline{C}_n$ e limite inferiore $\underline{C}_n$ in cui è compreso il parametro da determinare con un dato grado di fiducia (determinato dal decisore), ossia

$$P(\overline{C}_n < \theta < \underline{C}_n) = 1 - \alpha$$

Si dice che $(\overline{C}_n, \underline{C}_n)$ è un intervallo di confidenza di grado $1 - \alpha$ per θ.
Per determinare gli intervalli di confidenza si utilizza il cosiddetto metodo *pivotale*.
Tale metodo consiste nel determinare una variabile aleatoria di pivot $\gamma(X_1, X_2, \ldots, X_n; \theta)$ che dipende dal campione casuale, dipende dal parametro non noto e la sua funzione di distribuzione non contiene il parametro da stimare.
Essa non è una statistica, dal momento che non è osservabile dal momento che dipende dal parametro non noto.
Per ogni coefficiente α fissato (con $0 < \alpha < 1$) siano α_1 e $\alpha_2 (\alpha_1 < \alpha_2)$ due valori dipendenti soltanto dal coefficiente fissato α tali che per ogni $\theta \in \Theta$ si abbia

$$P(\alpha_1 < \gamma(X_1, X_2, \ldots, X_n; \theta) < \alpha_2) = 1 - \alpha$$

Se per ogni possibile campione osservato $x = (x_1, x_2, \ldots, x_n)$ e per ogni $\theta \in \Theta$, si riesce a dimostrare che

$$\alpha_1 < \gamma(x; \theta) < \alpha_2 \iff g_1(x) < \theta < g_2(x)$$

con $g_1(x)$ e $g_2(x)$ dipendenti soltanto dal campione osservato, allora è equivalente a richiedere che

$$P(g_1(X_1, X_2, \ldots, X_n) < \theta < g_2(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$

Denotando con $\underline{C}_n = g_1(X_1, X_2, \ldots, X_n)$ e $\overline{C}_n = g2(X_1, X_2, \ldots, X_n)$ segue

$$P(\underline{C}_n < \theta < \overline{C}_n) = 1 - \alpha$$

Per quanto riguarda la popolazione di Poisson, così come le altre popolazioni diverse dalla popolazione normale, bisogna considerare una dimensione del campione elevata ($n >= 30$) per utilizzare il teorema centrale di convergenza per determinare l'intervallo di confidenza.

Teorema Centrale Di Convergenza
Sia $X_1, X_2, \ldots$ una successione di variabili aleatorie, definite nello stesso spazio di probabilità, indipendenti e identicamente distribuite con valore medio μ finito e varianza σ^2 finita e positiva. Posto per ogni intero n positivo $Y_n = X_1 + X_2 + \cdots + X_n$, per ogni $x \in R$ risulta :

$$\lim_{n \to +\infty} P\left(\frac{Y_n - n\,\mu}{\sigma\,\sqrt{n}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy = \Phi(x),$$

ossia la successione delle variabili aleatorie standardizzate

$$\frac{Y_n - E(Y_n)}{\sqrt{Var(Y_n}} = \frac{Y_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z$$

converge in distribuzione alla variabile aleatoria normale standard.
E' dimostrabile che sottraendo a Y_n la sua media $n\mu$ e dividendo la differenza per la deviazione standard di Y_n, ossia per $\sigma\sqrt{n}$ si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione è per n sufficientemente grande approssimativamente normale standard.
Quindi per n grande la distribuzione della media campionaria $\overline{X}_n$ è approssimativamente normale con valore medio μ e varianza σ^2/n, ossia :

$$\overline{X}_n \simeq \mu + \frac{\sigma}{\sqrt{n}} Z$$

Con una variabile aleatoria X che descrive la popolazione con $E(X) = \mu$ e $Var(X) = \sigma^2$ e con $(X_1, X_2, \ldots, X_n)$ il campione casuale, il teorema centrale di convergenza afferma che la variabile aleatoria

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard.
Il comportamento di una variabile aleatoria normale standard è descritto dalla seguente figura:

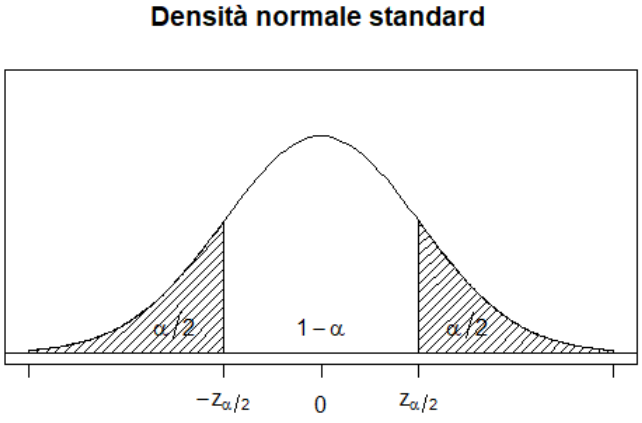


Figura 14: Densità normale standard e grado di fiducia $1 - \alpha$

Dal momento che il valore medio $E(X) = \mu$ e $Var(X) = \sigma^2$ dipendono dal parametro θ non noto, si nota la variabile aleatoria Z_n può essere interpretata come una

variabile aleatoria di pivot poiché dipende dal campione casuale, dipende dal parametro non noto attraverso il valore medio e la varianza e per grandi campioni la sua funzione di distribuzione è approssimativamente normale standard e quindi non contiene nessun parametro da stimare (dal momento che i due parametri per la normale μ la media e σ la deviazione standard sono pari rispettivamente a 0 e 1).
Per questo motivo è possibile applicare il metodo pivotale in forma approssimata ossia

$$P(-z_{\alpha/2} < \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) \simeq 1 - \alpha$$

Intervallo di confidenza per il parametro λ di una popolazione di Poisson
Consideriamo una popolazione di Poisson descritta da una variabile aleatoria $X \sim P(\lambda)$ con funzione di probabilità

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ x= 0,1,...} (\lambda > 0)$$

Il valore medio di una variabile aleatoria di Poisson è $E(X) = \lambda$ e la varianza è $Var(X) = \lambda$ ed entrambi dipendono dal parametro non noto λ.
Ricaviamo che

$$E(\overline{X}_n) = \lambda$$
$$Var(\overline{X}_n) = \frac{\lambda}{n}$$

Applicando il teorema centrale di convergenza si ha che la variabile aleatoria

$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} = \sqrt{n}\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}}$$

converge in distribuzione ad una variabile aleatoria normale standard.
Per campioni sufficientemente numerosi, l'intervallo di confidenza di grado $1 - \alpha$ per il parametro λ può essere determinato richiedendo che

$$P(-z_{\alpha/2} < \frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} < z_{\alpha/2}) \simeq 1 - \alpha$$

La diseguaglianza

$$-z_{\alpha/2} < \frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} < z_{\alpha/2}$$

è equivalente a

$$\left[\sqrt{\frac{n}{\lambda}}(\overline{x}_n - \lambda)\right]^2 < z^2_{\alpha/2}$$

che conduce alla disuguaglianza di secondo grado in λ

$$n\lambda^2 - \lambda(2n\overline{x}_n + z^2_{\alpha/2}) + n\overline{x}^2_n < 0$$

Essendo il coefficiente di λ^2 positivo, le soluzioni della disuguaglianza precedente sono interne all'intervallo delle radici della relativa equazione di secondo grado, ossia $\underline{c}_n < \lambda < \overline{c}_n$.
Se si denota con

$$\begin{gathered} a_2 = n \\ a_1 = -(2n\overline{x}_n + z^2_{\alpha/2}) \\ a_0 = n\overline{x}^2_n \end{gathered}$$

le radici dell'equazione $a_2\lambda^2 + a_1\lambda + a_0 = 0$ possono essere calcolate utilizzando

$$polyroot(c(a_0, a_1, a_2))$$

Nel caso studio in analisi, vogliamo determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per il parametro λ. Sappiamo che n $= 64$, $\alpha = 0.01$ e $\overline{x}_{64} = 2.4$, determiniamo l'intervallo tramite i seguenti comandi R:

```
> alpha <- 1-0.99
> qnorm(1-alpha/2,mean=0,sd=1)
[1] 2.575829
> zalpha <- qnorm(1-alpha/2,mean=0 , sd=1)
> n <- length(campione)
> meancamp <- mean(campione)
> a2 <- n
> a1 <- -(2*n*meancamp+zalpha^2)
> a0 <- n*meancamp^2
> polyroot(c(a0,a1,a2))
[1] 2.208957+0i 3.269713-0i
```

Il valore di λ è compreso nell'intervallo stimato.
Inoltre è possibile notare dall'esempio successivo che. al diminuire del grado di confidenza, l'intervallo diminuisce di dimensione, dal momento che si vuole determinare con minore certezza l'appartenenza del parametro all'intervallo determinato.
Per evidenziare ciò si è determinato l'intervallo di confidenza con $1-\alpha = 0.95$, restando con gli stessi parametri del caso precedente.

```
> alpha <- 1-0.95
> qnorm(1-alpha/2, mean=0, sd=1)
[1] 1.959964
> zalpha <- qnorm(1-alpha/2, mean = 0, sd = 1)
> n <- length(campione)
> meancamp <- mean(campione)
> a2 <- n
> a1 <- -(2*n*meancamp+zalpha^2)
> a0 <- n*meancamp^2
> polyroot(c(a0, a1, a2))
[1] 2.314756-0i 3.120267+0i
```

Nel caso in cui volessimo considerare due edizioni del mondiale di calcio, e quindi ottenere una simulazione di lunghezza 128, l'intervallo di confidenza, nel caso in cui il grado di confidenza sia fissato come $1-\alpha = 0.99$, aumenta rispetto all'intervallo ottenuto con lo stesso grado di confidenza ma su un campione di grandezza minore.
Questo si può evincere dal seguente codice eseguito su un campione ottenuto concatenando il campione due volte.

```
> alpha <- 1-0.99
> qnorm(1-alpha/2, mean = 0, sd = 1)
[1] 2.575829
> zalpha <- qnorm(1-alpha/2, mean = 0, sd = 1)
> campionedoppio <- c(campione, campione)
> n <- length(campionedoppio)
> n
[1] 128
> meancamp <- mean(campionedoppio)
> meancamp
[1] 2.6875
> a2 <- n
> a1 <- -(2*n*meancamp+zalpha^2)
> a0 <- n*meancamp^2
> polyroot(c(a0, a1, a2))
[1] 2.339280-0i 3.087555+0i
```

Confronto tra due Popolazioni di Poisson
Supponiamo di voler confrontare due edizioni dei mondiali di calcio e verificare quale delle due abbia un parametro λ più grande dell'altra. Consideriamo una prima popolazione di Poisson descritta da una variabile $X \sim P(\lambda_1)$ con funzione di probabilità

$$p_X(x) = \frac{\lambda_1^x}{x!} e^{-\lambda_1} \text{ x= 0,1,...} (\lambda_1 > 0)$$

e una seconda popolazione di Poisson descritta da una variabile $Y \sim P(\lambda_2)$ con funzione di probabilità

$$p_Y(x) = \frac{\lambda_2^x}{x!} e^{-\lambda_2} \text{ x= 0,1,...} (\lambda_2 > 0)$$

e siano $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ due campioni casuali indipendenti di ampiezza n_1 e n_2 estratti dalle due popolazioni di Poisson.
Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $\lambda_1 - \lambda_2$ tra i parametri delle due popolazioni per grandi valori di n_1 e n_2.
Denotiamo con $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ rispettivamente le medie campionarie delle due popolazioni. Dal teorema centrale di convergenza segue che la variabile aleatoria

$$\frac{\overline{X}_{n1} - \overline{Y}_{n2} - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1/n_1 + \lambda_2/n_2}} \xrightarrow{d} Z$$

converge in distribuzione ad una variabile aleatoria normale standard. Poiché

$$E(\overline{X}_{n1}) = \lambda_1$$

$$\lim_{n_1 \to +\infty} Var(\overline{X}_{n1}) = 0$$

$$E(\overline{Y}_{n2}) = \lambda_2$$

$$\lim_{n_2 \to +\infty} Var(\overline{Y}_{n2}) = 0$$

ossia le medie campionarie $\overline{X}_{n_1}$ e $\overline{Y}_{n_2}$ sono stimatori corretti e consistenti di λ_1 e λ_2, per campioni sufficientemente numerosi l'intervallo di confidenza di grado $1 - \alpha$ per la differenza $p_1 - p_2$ può essere determinato supponendo che

$$P(-z_{\alpha/2} < \frac{\overline{X}_{n1} - \overline{Y}_{n2} - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1/n_1 + \lambda_2/n_2}} < z_{\alpha/2}) \simeq 1 - \alpha$$

possiamo dire che una stima approssimata per la differenza di $\lambda_1 - \lambda_2$ è

$$\overline{x}_{n1} - \overline{y}_{n2} - z_{\alpha/2}\sqrt{\frac{\overline{x}_{n1}}{n_1} + \frac{\overline{y}_{n2}}{n_2}} < \lambda_1 - \lambda_2 < \overline{x}_{n1} - \overline{y}_{n2} + z_{\alpha/2}\sqrt{\frac{\overline{x}_{n1}}{n_1} + \frac{\overline{y}_{n2}}{n_2}}$$

dove $\overline{x}_{n_1}$ e $\overline{y}_{n_2}$ denotano rispettivamente le medie campionarie delle due osservazioni. Nel nostro caso studio dunque consideriamo due campioni: *campione*, ossia il campione utilizzato fin ora riguardante le partite del 2022 e *campione2018*, ossia il numero di goal per partita dell'edizione dei mondiali di calcio del 2018. Entrambi sono di lunghezza 64 ma con λ_1 uguale a 2.6875 e λ_2 uguale a 2.578125, ne calcoliamo la media campionaria e determiniamo l'intervallo di confidenza per $\lambda_1 - \lambda_2$ di grado $1 - \alpha = 0.99$.
Visualizziamo le due edizioni:

```
> campione
 [1] 2 8 2 2 3 0 0 5 0 3 7 1 1 0 5 2 2 4 2 0 1 2
[23] 3 2 1 2 5 2 6 5 1 2 2 3 3 1 1 1 2 3 0 3 3 6
[45] 2 3 5 1 4 3 4 3 2 5 0 7 2 4 1 3 3 2 3 6
> campione2018 <- c(1, 1, 6, 3, 2, 1, 2, 1, 1, 2, 1, 3, 3, 3, 3, 4, 1,
    1, 1, 2, 1, 3, 2, 2, 3, 7, 3, 3, 7, 4, 3, 3, 3, 4, 2, 2, 0, 3, 3,
   2, 3, 2, 4, 1, 1, 3, 1, 7, 3, 2, 2, 2, 5, 1, 2, 1, 2, 3, 2, 4, 1,
   3, 2, 6)
```

Mostriamo le frequenze assolute delle due edizioni e la media campionaria:

```
> table(campione)
campione
 0  1  2  3  4  5  6  7  8
 7 10 17 14  4  6  3  2  1
> mean(campione)
```

```
[1] 2.6875
> table(campione2018)
campione2018
 0  1  2  3  4  5  6  7
 1 16 17 19  5  1  2  3
> mean(campione2018)
[1] 2.578125
```

Determiniamo ora l'intervallo di confidenza per $\lambda_1 - \lambda_2$ di grado $1 - \alpha = 0.99$:

```
> alpha <- 1-0.99
> qnorm(1-alpha/2, mean=0, sd=1)
[1] 2.575829
> zalpha <- qnorm(1-alpha/2, mean=0, sd=1)
> n1 <- length(campione)
> n2 <- length(campione2018)
> m1 <- mean(campione)
> m2 <- mean(campione2018)
> rad <- sqrt(m1/n1+m2/n2)
> m1 - m2 - zalpha * rad
[1] -0.6294678
> m1 - m2 + zalpha * rad
[1] 0.8482178
> m1 - m2
[1] 0.109375
```

L'intervallo ottenuto è dunque $(-0.63, 0.85)$ e la differenza $\lambda_1 - \lambda_2$, uguale a 0.11, è compresa in tale intervallo.
Si può notare dell'intervallo ottenuto che il valore dello 0 è incluso, questo quindi implica che non è possibile trarre nessuna conclusione sul quale λ sia maggiore dell'altro.

8 Verifica delle Ipotesi

La verifica delle ipotesi viene definita a partire da una variabile aleatoria X caratterizzata da una funzione di probabilità $f(x;\theta)$ su cui si effettua un test su un'ipotesi statistica. Un ipotesi statistica è un'affermazione su un parametro non noto θ che è soggetta a verifica utilizzando un campione casuale $X_1, X_2, \ldots, X_n$ estratto dalla popolazione a cui l'ipotesi si riferisce.
Le ipotesi statistiche possono essere di due tipi:

- **Semplici**: quando l'ipotesi specifica completamente la funzione di probabilità;
- **Composita**: altrimenti.

Un ipotesi statistica di cui si vuole effettuare il test è detta **ipotesi nulla** ed è denotata con H_0. I test richiedono inoltre la costruzione di un'ipotesi alternativa H_1, ossia di un'ipotesi opposta a quella indicata con H_0. Equivalente è dire che H_1 si trova in un sottoinsieme Θ_1 dello spazio dei parametri Θ disgiunto rispetto al sottoinsieme dello spazio dei parametri Θ_0 dell'ipotesi H_0.
Un test è denominato con ψ e consiste nel suddividere l'insieme dei possibili campioni, ossia l'insieme delle n-uple $(x_1, x_2, \ldots, x_n)$ assumibili dal vettore aleatorio $X_1, X_2, \ldots, X_n$ in due sottoinsiemi: una regione di accettazione A dell'ipotesi nulla ed una regione di rifiuto R dell'ipotesi nulla.
Nel caso in cui l'ipotesi nulla risulti **falsa**, l'ipotesi alternativa sarà **vera** e viceversa.
Due possibili errori potrebbero essere la presenza di falsi negativi (si rifiuta un'ipotesi vera, errore di tipo 1) oppure di falsi positivi (si accetta un'ipotesi falsa, errore di tipo 2).
Per descrivere questo tipo di errori si definisce il concetto di misura della regione critica. Sia ψ un test per verificare l'ipotesi nulla $H_0 : \theta \in \Theta_0$ in alternativa all'ipotesi $H_1 : \theta \in \Theta_1$. Si definisce misura della regione critica del test ψ (o livello di significatività del test ψ) la seguente probabilità

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

Questa misura fornisce la probabilità massima di commettere un errore di tipo 1 al variare di $\theta \in \Theta_0$, ossia la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera, come mostrato dalla tabella successiva.

	Rifiutare H_0	Accettare H_0
H_0 vera	Errore del tipo 1 Probabilità α	Decisione esatta Probabilità 1 - α
H_0 falsa	Decisione esatta Probabilità 1 - β	Errore del 2 tipo Probabilità β

Tabella 5: Errori di tipo 1 e 2

In generale per campioni casuali di fissata ampiezza, se si diminuisce la probabilità di commettere un errore di tipo 1 aumenta la probabilità di commettere un errore di tipo 2 e viceversa.
Nella costruzione del test si va quindi a fissare la probabilità di commettere un errore di tipo 1 e si cerca un test ψ che minimizzi la probabilità di commettere un errore di tipo 2.
Valori tipici per questa probabilità α sono 0.05, 0.1, 0.001 ed il test viene rispettivamente detto statisticamente *significativo*, statisticamente *molto significativo* e statisticamente *estremamente significativo* nei rispettivi casi.
Infatti, quando diminuisce il valore di α la credibilità di un eventuale rifiuto di H_0 aumenta.
I test statistici sono di due tipi:

- **Test Bilaterali**: $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$
- **Test Unilaterali**: i quali si distinguono a loro volta in :
  - **Test Unilaterale Sinistro**: $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$
  - **Test Unilaterale Destro**: $H_0 : \theta \geq \theta_0$, $H_1 : \theta < \theta_0$

Le conclusioni dei test statistici unilaterali e bilaterali dipendono dal livello di significatività α, scelto a priori dal decisore.
Altro metodo è il calcolo del livello di significatività osservato, noto come **p-value**.
Il criterio del p-value indica che:

- se $p > \alpha$, l'ipotesi H_0 non può essere rifiutata;
- se $p <= \alpha$, l'ipotesi H_0 deve essere rifiutata.

8.1 Verifica delle ipotesi per popolazione di Poisson

Quando l'ampiezza del campione è grande, per una popolazione descritta da una variabile aleatoria X caratterizzata da valore medio μ e varianza σ^2, entrambi finiti, si può utilizzare il teorema centrale di convergenza ricordando che la variabile aleatoria

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \underset{\rightarrow}{d} Z$$

converge in distribuzione ad una variabile normale standard.
Consideriamo una popolazione di Poisson descritta dalla variabile aleatoria $X \sim P(\lambda)$.
Siamo interessati a costruire dei test per il valore medio $E(X) = \lambda$.
Essendo $\mu_0 = \lambda$ e $\sigma_0^2 = \lambda$, nei seguenti test occorre considerare

$$z_{os} = \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} = \frac{\overline{x}_n - \lambda_0}{\sqrt{\frac{\lambda_0}{n}}} = \sqrt{n}\frac{\overline{x}_n - \lambda_0}{\sqrt{\lambda_0}}$$

8.1.1 Test bilaterale

Il test bilaterale può essere così formulato:

$$H_0 : \lambda = \lambda_0$$
$$H_1 : \lambda \neq \lambda_0$$

A questo punto per $n \to \infty$ e fissato il livello di significatività del test α, si dimostra che β è minimizzato se:

$$-z_{\alpha/2} < \sqrt{n}(\lambda_0 \overline{X}_n - 1) < z_{\alpha/2} \implies H_0 \text{ accettata}$$
$$\sqrt{n}(\lambda_0 \overline{X}_n - 1) < -z_{\alpha/2} \vee \sqrt{n}(\lambda_0 \overline{X}_n - 1) > z_{\alpha/2} \implies H_0 \text{ rifiutata}$$

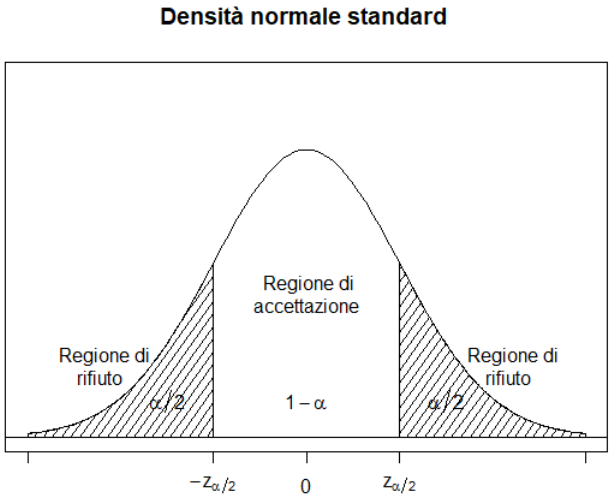


Figura 15: Densità normale standard e regioni di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale

Il *p-value* sarà invece dato da:

$$p = P(\sqrt{n}(\lambda_0 \overline{X}_n - 1) \leq z_{obs}) + P(\sqrt{n}(\lambda_0 \overline{X}_n - 1) \geq z_{obs})) \implies$$
$$p = 2(1 - P(\sqrt{n}(\lambda_0 \overline{X}_n - 1)) \leq |z_{obs}|) \approx 2(1 - \Phi(|z_{obs}|))$$

Nel nostro esempio, dal momento che nella stima intervallare con grado di confidenza $1 - \alpha = 0.99$ per il parametro λ abbiamo ottenuto i valori $(2.20, 3.26)$, costruiamo tre test. Un test bilaterale con λ_0 uguale al valore medio dell'intervallo:

$$H_0 : \lambda = 2.73$$
$$H_1 : \lambda \neq 2.73$$

Fissiamo $\alpha = 0.05$
Con R otteniamo:

```
> lambda0 <- 2.73
> alpha <- 0.05
> qnorm(1-alpha/2, mean=0, sd=1)
[1] 1.959964
> n <- length(campione)
> n
[1] 64
> meancamp <- mean(campione)
> meancamp
[1] 2.6875
> zos <- (meancamp - lambda0)/sqrt(lambda0/n)
> zos
[1] -0.2057774
> pvalue <- 2 * (1 - pnorm(abs(zos),mean = 0, sd = 1))
> pvalue
[1] 0.8369648
```

In questo caso H_0 viene accettata poiché il valore ottenuto è compreso nell'intervallo $(-z_{\alpha/2}, z_{\alpha/2})$.
Inoltre anche per quanto riguarda il test del p-value H_0 viene accettata poiché il risultato ottenuto è maggiore di 0.05.

8.1.2 Test unilaterale sinistro

Il test unilaterale sinistro ha ipotesi

$$H_0 : \lambda \leq \lambda_0$$
$$H_1 : \lambda > \lambda_0$$

per $n \to \infty$ e fissato α, si dimostra che β è minimizzato se:

$$\sqrt{n}(\lambda_0 \overline{X}_n - 1) \leq -z_\alpha \implies H_0 \text{ accettata}$$
$$\sqrt{n}(\lambda_0 \overline{X}_n - 1) > -z_\alpha \implies H_0 \text{ rifiutata}$$

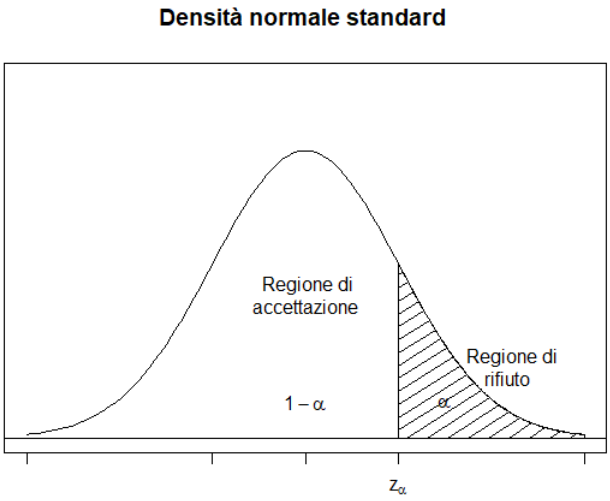


Figura 16: Densità normale standard e regione di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale sinistro

Il *p-value* sarà invece dato da:

$$p = P(\sqrt{n}(\lambda_0 \overline{X}_n - 1) \geq z_{obs}) \implies$$
$$p = 1 - P(\sqrt{n}(\lambda_0 \overline{X}_n - 1) \leq |z_{obs}|) \approx 1 - \Phi(z_{obs})$$

Un test unilaterale sinistro con λ_0 uguale a 3.26 e α fissato a 0.05:

$$H_0 : \lambda \leq 3.26$$
$$H_1 : \lambda > 3.26$$

```
> lambda0 <- 3.26
> alpha <- 0.05
> qnorm(1-alpha, mean=0, sd=1)
[1] 1.644854
> n
[1] 64
> meancamp
[1] 2.6875
> zos <- (meancamp-lambda0)/sqrt(lambda0/n)
> zos
[1] -2.536627
> pvalue <- 1- pnorm(zos, mean = 0, sd = 1)
> pvalue
[1] 0.9944037
```

In questo caso H_0 è verificata poiché il risultato ottenuto nel test, $z_{obs} \leq z_\alpha$. Per quanto riguarda il p-value, anche in questo caso H_0 è accettata poiché $0.99 \geq 0.05$.

8.1.3 Test unilaterale destro

Il test unilaterale destro ha per ipotesi:

$$H_0 : \lambda \geq \lambda_0$$
$$H_1 : \lambda < \lambda_0$$

Per $n \to \infty$ e fissato α, si dimostra che β è minimizzato se:

$$\sqrt{n}(\lambda_0 \overline{X}_n - 1) \geq -z_\alpha \implies H_0 \text{ accettata}$$
$$\sqrt{n}(\lambda_0 \overline{X}_n - 1) < -z_\alpha \implies H_0 \text{ rifiutata}$$

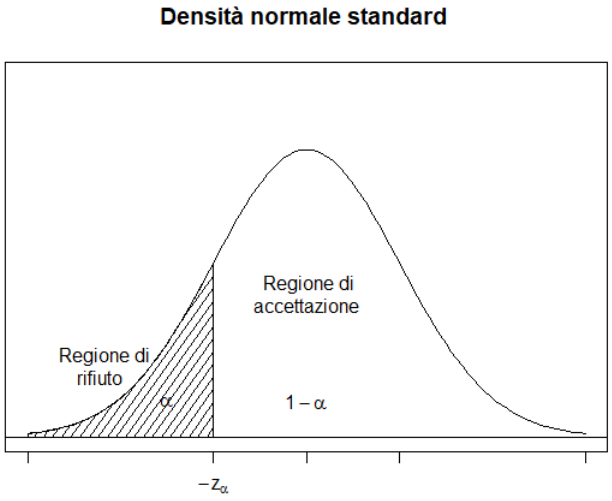


Figura 17: Densità normale standard e regione di accettazione e di rifiuto dell'ipotesi nulla per il test unilaterale destro

Il p-value, per n grande, sarà invece dato da:

$$p = P(\sqrt{n}(\lambda_0 \overline{X}_n - 1) \geq z_{obs}) \approx \Phi(z_{obs})$$

Un test unilaterale destro con λ_0 uguale a 2.20 e α fissato a 0.05:

$$H_0 : \lambda \geq 2.20$$
$$H_1 : \lambda < 2.20$$

```
> lambda0 <- 2.20
> alpha <- 0.05
> qnorm(1-alpha, mean=0, sd=1)
[1] 1.644854
```

```
> n
[1] 64
> meancamp
[1] 2.6875
> zos <- (meancamp-lambda0)/sqrt(lambda0/n)
> zos
[1] 2.629379
> pvalue <- pnorm(zos, mean=0, sd=1)
> pvalue
[1] 0.995723
```

In questo caso H_0 può essere associato alla regione di accettazione dal momento che il valore ottenuto $z_{os} \geq z_{\alpha}$. Inoltre viene accettato anche tramite il metodo del p-value dal momento che $0.99 \geq 0.05$.

9 Test del Chi-Quadrato

Altro problema da poter risolvere con la statistica inferenziale è la verifica del fatto che il campione osservato possa essere stato estratto da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione $F_X(x)$. A questo scopo si utilizza il criterio di verifica delle ipotesi del chi-quadrato e un test secondo tale principio è strutturato come segue:

- H_0: X ha una funzione di distribuzione $F_X(x)$ (avendo stimato k parametri non noti in base al campione);
- H_1: X non ha una funzione di distribuzione $F_X(x)$.

dove α è la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera. Occorre determinare un test ψ con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. Il test di verifica considerato è bilaterale. Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi $I_1, I_2, \ldots, I_r$ (classi o categorie) in modo che risulti essere uguale a p_i la probabilità che la variabile aleatoria assuma un valore appartenente a I_i ossia $p_i = P(X \in I_i)(i = 1, 2, \ldots, r)$. Si estrae poi un campione $x_1, x_2, \ldots, x_n$ di ampiezza n e si osservano le frequenze assolute $n_1, n_2, \ldots, n_r$ con cui gli n elementi si distribuiscono nei rispettivi insiemi $I_1, I_2, \ldots, I_r$. Quindi n_i rappresenta il numero degli elementi del campione che cadono nell'intervallo $I_i(i = 1, 2, \ldots, r)$.
Quindi

$$p_i \geq 0 \ (i = 1, 2, ..., r), \sum_{i=1}^{r} p_i = 1$$

$$n_i \geq 0 \ (i = 1, 2, ..., r), \sum_{i=1}^{r} n_i = n$$

Si noti che la probabilità che esattamente n_1 elementi appartengano ad I_1, n_2 elementi appartengano ad I_2, ... , n_r elementi appartengano ad I_r è

$$p(n_1, n_2, ..., n_r) = \frac{n!}{n_1! n_2! ... n_r!} p_1^{n_1} p_2^{n_2} p_r^{n_r}$$

ossia una funzione di probabilità multinomiale.
Ne segue che il numero medio di elementi nell'intervallo I_i è np_i.
Si calcola poi la quantità

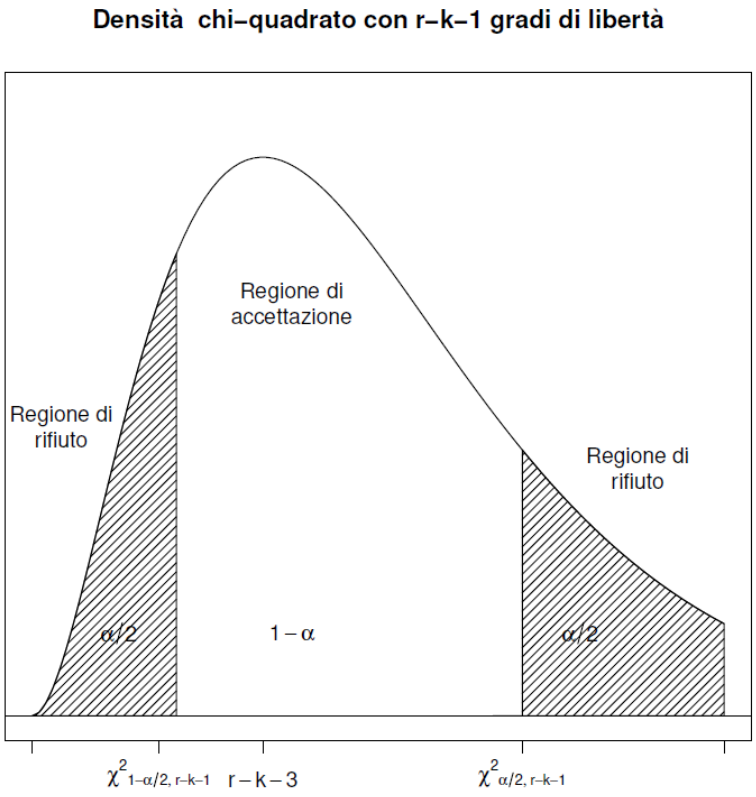
$$\chi^2 = \sum_{i=1}^{r} (\frac{n_i - np_i}{\sqrt{np_i}})^2$$

Il criterio chi-quadrato si basa sulla statistica

$$Q = \sum_{i=1}^{r} (\frac{N_i - np_i}{\sqrt{np_i}})^2$$

dove N_i è la variabile aleatoria che descrive il numero di elementi del campione casuale di variabili aleatorie osservabili, indipendenti ed identicamente distribuite che cadono nell'intervallo I_i con $(i = 1, 2, ...r)$.
Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti, si può dimostrare che per n sufficientemente grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con $r-k-1$ gradi di libertà, descritta nella figura successiva. Si sottrae 1 da r a causa della prima delle condizioni precedenti secondo la quale se conosciamo $r-1$ delle probabilità p_i la rimanente probabilità può essere univocamente determinata e si sottrae k poiché si suppone che siano k i parametri indipendenti non noti sostituiti da stime. Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta

$$min(np_1, np_2, ..., np_r) \geq 5$$



Si giunge quindi alla definizione del test chi-quadrato bilaterale:
Per un campione sufficientemente numeroso di ampiezza n, il test chi-quadrato bilaterale di misura α è il seguente:

- Si accetti l'ipotesi H_0 se $\chi^2_{1-\alpha/2,r-k-1} < \chi^2 < \chi^2_{\alpha/2,r-k-1}$
- Si rifiuti l'ipotesi H_0 se $\chi^2 < \chi^2_{1-\alpha/2,r-k-1}$ oppure $\chi^2 > \chi^2_{\alpha/2,r-k-1}$

con $\chi^2_{1-\alpha/2,r-k-1}$ e $\chi^2_{\alpha/2,r-k-1}$ sono soluzioni delle equazioni :

$$P(Q < \chi^2_{1-\alpha/2,r-k-1}) = \frac{\alpha}{2}$$
$$P(Q < \chi^2_{\alpha/2,r-k-1}) = 1 - \frac{\alpha}{2}$$

Esempio di test del chi-quadrato su popolazione di Poisson
Nel caso preso in esame, supponiamo di avere il numero di goal per le 64 partite dell'edizione del 2022 dei mondiali di calcio. I risultati sono i seguenti :

```
> campione <- c
   (2,8,2,2,3,0,0,5,0,3,7,1,1,0,5,2,2,4,2,0,1,2,3,2,1,2,5,2,6,
    5,1,2,2,3,3,1,1,1,2,3,0,3,3,6,2,3,5,1,4,3,4,3,2,5,0,7,2,4,
    1,3,3,2,3,6)
> n <- length(campione)
> n
[1] 64
> freq<-table(campione)
> freq
campione
 0  1  2  3  4  5  6  7  8
 7 10 17 14  4  6  3  2  1
```

Si desidera verificare se il numero di goal sia descrivibile con una variabile aleatoria X di Poisson di parametro λ, ossia

$$p_X(x) = \frac{\lambda^x}{x!}e^{-\lambda} \text{ (x = 0,1,...)}$$

con $\lambda > 0$.
Dal campione stimiamo il valore di λ con la media campionaria:

```
> stimalambda <- mean(campione)
> stimalambda
[1] 2.6875
```

Supponiamo di considerare 4 categorie corrispondenti agli intervalli $I_1 = 0, 1$, $I_2 = 2, 3$, $I_3 = 4$, $I_4 = 5, 6, 7, 8$.
Le probabilità associate agli intervalli $p_1 = p_X(0) + p_X(1)$, $p_2 = p_X(2) + p_X(3)$, $p_3 = p_X(4)$ e $p_4 = 1 - p_1 - p_2 - p_3$ possono essere così calcolate:

```
> p <- numeric(4)
> p[1] <- dpois(0, stimalambda) + dpois(1, stimalambda)
> p[2] <- dpois(2, stimalambda) + dpois(3, stimalambda)
> p[3] <- dpois(4, stimalambda)
> p[4] <- 1- p[1] - p[2] - p[3]
> p
[1] 0.2509375 0.4659086 0.1479164 0.1352376
> sum(p)
[1] 1
```

Si nota che $p_1 + p_2 + p_3 + p_4 = 1$. Essendo

```
> min (n*p[1], n*p[2],n*p[3],n*p[4])
[1] 8.655203
```

maggiore di 5, si garantisce il fatto che le classi contengono in media 5 elementi. Il numero di elementi del campione appartenente ai quattro intervalli è

```
> r <- 4
> nint <- numeric(r)
> nint[1]<-length(which(campione >= 0 & campione <=1))
> nint[2]<-length(which(campione >= 2 & campione <=3))
> nint[3]<-length(which(campione == 4))
> nint[4]<-length(which(campione > 4))
> nint
[1] 17 31  4 12
> sum(nint)
[1] 64
```

Calcoliamo ora χ^2 definito in precedenza

```
> chi2 <- sum(((nint-n*p)/sqrt(n*p))^2)
> chi2
[1] 4.551247
```

ossia $\chi^2 = 4.55$.
In questo caso le categorie sono r = 4 e poniamo k = 1 dal momento che la distribuzione di probabilità ha un parametro non noto.
Dunque, r-k-1 = 2 e scegliendo $\alpha = 0.01$ occorre calcolare $\chi^2_{1-\alpha/2,2}$ e $\chi^2_{\alpha/2,2}$:

```
> k <- 1
> alpha <- 0.01
> qchisq(alpha/2, df=r-k-1)
[1] 0.01002508
> qchisq(1-alpha/2,df=r-k-1)
[1] 10.59663
```

Essendo $0.010 < \chi^2 < 10.597$, l'ipotesi H_0 di popolazione di Poisson può essere accettata.