
UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

PROGETTO DI STATISTICA E ANALISI DEI DATI

# Analisi statistica sulle popolazioni delle nazioni del dataset "Metropolitan Areas"

**Mario Cicalese 0522501799**

Anno Accademico 2023-2024

# Indice

# Elenco delle figure

# Elenco delle tabelle

CAPITOLO 1

# Introduzione

Oggi la popolazione mondiale ha superato gli 8 miliardi di abitanti, ma questo **trend di crescita** proseguirà? Oppure, in futuro, si assisterà a un **declino demografico globale**? Da anni si sente parlare di *sovrappopolazione*, cioè, una densità di popolazione troppo elevata per il tipo di territorio in cui ci si trova, portando ad effetti negativi come la distruzione dell'ambiente e l'eccessivo consumo di risorse. **Ma per tutte le nazione è così?** O esistono nazioni che in un trend di crescita mondiale, si trovano ad affrontare la **diminuzione** della propria popolazione?
**L'obiettivo** di tale analisi statistica sarà proprio questo: **individuare** tra un sottoinsieme mondiale di nazioni, chi ha un avuto un **trend di crescita**, e chi di **decrescita**, mostrando i dati anno per anno, considerando: **picchi, tendenze, costanza e anomalie**. Ma quali sono i **motivi** che portano alla decrescita demografica di una nazione in un trend di crescita mondiale? I motivi possono essere vari, ma i più comuni sono: **invecchiamento, diminuzione delle nascite, urbanizzazione e migrazione verso paesi più sviluppati**. Queste cause combinate, possono portare a una **continua decrescita negli della popolazione** di una nazione, creando un **trend negativo**. D'altro canto, una nazione sviluppata e ben equilibrata, caratterizzata da una qualità di vita elevata, attrarrà **nuovi residenti** e favorirà un **aumento delle nascite**, dando luogo a un **trend positivo di crescita demografica**.

## 1.1 Descrizione del dataset

L'analisi statistica verrà effettuata su un dataset messo a disposizione dall'**OECD** (organizzazione per la cooperazione e lo sviluppo economico). Il dataset scelto si presenta con 21 colonne (periodo di tempo che va dal 2000 al 2021) e con un elevato numero di righe (7 pagine). Questo perchè il dataset non fa riferimento solo alle nazioni, ma indica anche la popolazione delle città più importanti di ogni nazione. Tuttavia, lo scopo di quest'analisi statistica è quella di **lavorare sulle nazioni**, quindi le **città non sono state considerate**. Eliminate le città, il dataset contiene 51 righe, che rappresentano 51 nazioni sparse per il mondo. Di queste 51, sono state scelte le nazioni che: sono geograficamente situate in **Europa** (tranne per L'Australia), e che, negli anni, presentavano **cambiamenti/trend** interessanti da analizzare (**21 in totale**). Le nazioni scelte e che verranno analizzate durante l'intera analisi statistica sono: Australia, Austria, Belgio, Bulgaria, Svizzera, Germania, Danimarca, Spagna, Finlandia, Francia, Regno Unito, Grecia, Croazia, Ungheria, Italia, Olanda, Norvegia, Polonia, Portogallo, Romania e Svezia. In Figura 1.1 è possibile osservare il dataset finale utilizzato durante l'intera analisi statistica.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Nazione | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
| 2 | AUS: Australia | 19026200 | 19272100 | 19492800 | 19718400 | 19930400 | 20174500 | 20448600 | 20825100 | 21246500 | 21688800 | 22028700 | 22336900 | 22730400 | 23125200 | 23472800 | 23813100 | 24186300 | 24597200 | 24982700 | 25365700 | 25697300 | 25738100 |
| 3 | AUT: Austria | 8002190 | 8020950 | 8063640 | 8100270 | 8142570 | 8201360 | 8254300 | 8282960 | 8307390 | 8335000 | 8351640 | 8375160 | 8408120 | 8451860 | 8507790 | 8584930 | 8700470 | 8772870 | 8822270 | 8858780 | 8901060 | 8932660 |
| 4 | BEL: Belgium | 10239100 | 10263400 | 10309700 | 10355800 | 10396400 | 10445900 | 10511400 | 10584500 | 10666900 | 10753100 | 10839900 | 11000600 | 11075900 | 11138000 | 11180800 | 11237300 | 11311100 | 11351700 | 11398600 | 11455500 | 11522400 | 11566000 |
| 5 | BGR: Bulgaria | 8190680 | 8149470 | 7868820 | 7805510 | 7745150 | 7688570 | 7623370 | 7572670 | 7518000 | 7467120 | 7421770 | 7369430 | 7327220 | 7284550 | 7245680 | 7202200 | 7153780 | 7101860 | 7050030 | 7000040 | 6951480 | 6916550 |
| 6 | CHE: Switzerland | 7164440 | 7204060 | 7255650 | 7313850 | 7364150 | 7415100 | 7459130 | 7508740 | 7593430 | 7701860 | 7785810 | 7870130 | 7954660 | 8039060 | 8139630 | 8237670 | 8327130 | 8419550 | 8484130 | 8544530 | 8606030 | 8670300 |
| 7 | DEU: Germany | 82163500 | 82259500 | 82440300 | 82536700 | 82531700 | 82500800 | 82438000 | 82314900 | 82217800 | 82002400 | 81802300 | 80222100 | 80327900 | 80523700 | 80767500 | 81197500 | 82175700 | 82521700 | 82792400 | 83019200 | 83166700 | 83155000 |
| 8 | DNK: Denmark | 5330020 | 5349210 | 5368350 | 5383510 | 5397640 | 5411410 | 5427460 | 5447080 | 5475790 | 5511450 | 5534740 | 5560630 | 5580520 | 5602630 | 5627240 | 5659720 | 5707250 | 5748770 | 5781190 | 5806080 | 5822760 | 5840050 |
| 9 | ESP: Spain | 40470200 | 40665500 | 41035300 | 41827800 | 42547500 | 43296300 | 44010000 | 44784700 | 45668900 | 46233300 | 46486600 | 46667200 | 46818200 | 46727900 | 46512200 | 46449600 | 46440100 | 46528000 | 46658400 | 46937100 | 47332600 | 47398700 |
| 10 | FIN: Finland | 5171300 | 5181120 | 5194900 | 5206300 | 5219730 | 5236610 | 5255580 | 5276960 | 5300480 | 5326310 | 5351430 | 5375280 | 5401270 | 5426670 | 5451270 | 5471750 | 5487310 | 5503300 | 5513130 | 5517920 | 5525290 | 5533790 |
| 11 | FRA: France | 60545000 | 60973300 | 61424000 | 61864100 | 62232200 | 62772900 | 63229600 | 63645100 | 64007200 | 64350200 | 64658900 | 64978700 | 65277000 | 65600400 | 66166000 | 66458200 | 66638400 | 66809800 | 67026200 | 67177600 | 67320200 | 67656700 |
| 12 | GBR: United Kingdom | 58785200 | 58999800 | 59239600 | 59501400 | 59793800 | 60182100 | 60620400 | 61073300 | 61571600 | 62042300 | 62510200 | 63022500 | 63495100 | 63905300 | 64351200 | 64853400 | 65379000 | 65844100 | 66273600 | 66647100 | 67081700 | 67330000 |
| 13 | GRC: Greece | 10775600 | 10836000 | 10888300 | 10915800 | 10940400 | 10969900 | 11004700 | 11036000 | 11060900 | 11094700 | 11113300 | 11123400 | 11086400 | 11003600 | 10926800 | 10858000 | 10783700 | 10768200 | 10741200 | 10724600 | 10718600 | 10678600 |
| 14 | HRV: Croatia | 4497740 | 4295410 | 4305490 | 4305380 | 4305730 | 4310860 | 4312490 | 4313530 | 4311970 | 4309800 | 4302850 | 4289860 | 4275980 | 4262140 | 4246810 | 4225320 | 4190670 | 4154210 | 4105490 | 4076250 | 4058170 | 4036360 |
| 15 | HUN: Hungary | 10221600 | 10200300 | 10174900 | 10142400 | 10116700 | 10097500 | 10078600 | 10066200 | 10045400 | 10031000 | 10014300 | 9985720 | 9931930 | 9908800 | 9877370 | 9855570 | 9830490 | 9797560 | 9778370 | 9772760 | 9769530 | 9730770 |
| 16 | ITA: Italy | 56923500 | 56960700 | 56987500 | 57130500 | 57495900 | 57874800 | 58064200 | 58223700 | 58652900 | 59000600 | 59190100 | 59364700 | 59394200 | 59685200 | 60782700 | 60795600 | 60665600 | 60589400 | 60484000 | 59816700 | 59641500 | 59236200 |
| 17 | NLD: Netherlands | 15864000 | 15987100 | 16105300 | 16192600 | 16258000 | 16305500 | 16334200 | 16358000 | 16405400 | 16485800 | 16575000 | 16655800 | 16730300 | 16779600 | 16829300 | 16900700 | 16979100 | 17081500 | 17181100 | 17282200 | 17407600 | 17475400 |
| 18 | NOR: Norway | 4478500 | 4503440 | 4524070 | 4552250 | 4577460 | 4606360 | 4640220 | 4681130 | 4737170 | 4799250 | 4858200 | 4920310 | 4985870 | 5051280 | 5107970 | 5166490 | 5210720 | 5258320 | 5295620 | 5328210 | 5367580 | 5391370 |
| 19 | POL: Poland | 38263300 | 38254000 | 38242200 | 38218500 | 38190600 | 38173800 | 38157100 | 38125500 | 38115600 | 38135900 | 38022900 | 38062700 | 38063800 | 38062500 | 38017900 | 38005600 | 37967200 | 37973000 | 37976700 | 37972800 | 37958100 | 37840000 |
| 20 | PRT: Portugal | 10249000 | 10330800 | 10394700 | 10444600 | 10473100 | 10494700 | 10512000 | 10532600 | 10553300 | 10563000 | 10573500 | 10572700 | 10542400 | 10487300 | 10427300 | 10374800 | 10341300 | 10309600 | 10291000 | 10276600 | 10295900 | 10298300 |
| 21 | ROU: Romania | 22455500 | 22430500 | 21833500 | 21627500 | 21521100 | 21382400 | 21257000 | 21130500 | 20635500 | 20440300 | 20294700 | 20199100 | 20096000 | 20020100 | 19947300 | 19870600 | 19760600 | 19643900 | 19533500 | 19414500 | 19328800 | 19201700 |
| 22 | SWE: Sweden | 8861430 | 8882790 | 8909130 | 8940790 | 8975670 | 9011390 | 9047750 | 9113260 | 9182930 | 9256350 | 9340680 | 9415570 | 9482860 | 9555890 | 9644860 | 9747360 | 9851020 | 9995150 | 10120200 | 10230200 | 10327600 | 10379300 |

**Figura 1.1:** Dataset dell'OECD utilizzato per l'analisi statistica

## 1.2 Definizione del dataframe

Per poter utilizzare la vasta gamma di funzionalità offerte da R, è necessario, come prima operazione, importare il dataset all'interno dell'ambiente di lavoro R. Il

dataset è stato memorizzato all'interno di una struttura dati definita come **Dataframe** in R, poichè era quella che più si adattava alla sua struttura. Infatti, i dataframe, sono strutture bidimensionali composte da colonne che rappresentano variabili o attributi e da righe che rappresentano (istanze/campioni). Esse sono facilmente accessibili e flessibili grazie alle funzioni messe a disposizione da R; infatti, in R le singole colonne e righe del dataframe vengono considerate come singoli vettori. In Figura 1.2 viene mostrato come si presenta il Dataframe in R contenente il dataset da studiare. Da questo momento in poi il dataset è all'interno dell'ambiente di lavoro R, e potrà essere utilizzato per effetuare l'analisi statistica al fine di raggiungere i nostri obiettivi, utilizzando le funzionalità disponibili in R.

| | Nazione | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AUS: Australia | 19026200 | 19272100 | 19492800 | 19718400 | 19930400 | 20174500 | 20448600 | 20825100 | 21246500 | 21688800 | 22028700 | 22336900 | 22730400 | 23125200 | 23472800 | 23813100 | 24186300 | 24597200 | 24982700 | 25365700 | 25697300 | 25738100 |
| 2 | AUT: Austria | 8002190 | 8020950 | 8063640 | 8100270 | 8142570 | 8201360 | 8254300 | 8282980 | 8307990 | 8335000 | 8351640 | 8375160 | 8408120 | 8451860 | 8507790 | 8584930 | 8700470 | 8772870 | 8822270 | 8858780 | 8901060 | 8932660 |
| 3 | BEL: Belgium | 10239100 | 10263400 | 10309700 | 10355800 | 10396400 | 10445900 | 10511400 | 10584500 | 10666900 | 10753100 | 10839900 | 11000600 | 11075900 | 11138000 | 11180800 | 11237300 | 11311100 | 11351700 | 11398600 | 11455500 | 11522400 | 11566000 |
| 4 | BGR: Bulgaria | 8190880 | 8149470 | 7868820 | 7805510 | 7745150 | 7688570 | 7629370 | 7572670 | 7518000 | 7467120 | 7421770 | 7369430 | 7327220 | 7284550 | 7245680 | 7202200 | 7153780 | 7101860 | 7050030 | 7000040 | 6951480 | 6916550 |
| 5 | CHE: Switzerland | 7164440 | 7204060 | 7255650 | 7313850 | 7364150 | 7415100 | 7459130 | 7508740 | 7593490 | 7701860 | 7785810 | 7870130 | 7954660 | 8039060 | 8139630 | 8237670 | 8327130 | 8419550 | 8484130 | 8544530 | 8606030 | 8670300 |
| 6 | DEU: Germany | 82163500 | 82259500 | 82440300 | 82536700 | 82531700 | 82500800 | 82438000 | 82314900 | 82217800 | 82002400 | 81802300 | 80222100 | 80327900 | 80523700 | 80767500 | 81197500 | 82175700 | 82521700 | 82792400 | 83019200 | 83166700 | 83155000 |
| 7 | DNK: Denmark | 5330020 | 5349210 | 5368350 | 5383510 | 5397640 | 5411410 | 5427460 | 5447080 | 5475790 | 5511450 | 5534740 | 5560630 | 5580520 | 5602630 | 5627240 | 5659720 | 5707250 | 5748770 | 5781190 | 5806080 | 5822760 | 5840050 |
| 8 | ESP: Spain | 40470200 | 40665500 | 41035300 | 41827800 | 42547500 | 43296300 | 44010000 | 44784700 | 45668900 | 46239300 | 46486600 | 46667200 | 46818200 | 46727900 | 46512200 | 46449600 | 46440100 | 46528000 | 46658400 | 46937100 | 47332600 | 47398700 |
| 9 | FIN: Finland | 5171300 | 5181120 | 5194900 | 5206300 | 5219730 | 5236610 | 5255580 | 5276960 | 5300480 | 5326310 | 5351430 | 5375280 | 5401270 | 5426670 | 5451270 | 5471750 | 5487310 | 5503300 | 5513130 | 5517920 | 5525290 | 5533790 |
| 10 | FRA: France | 60545000 | 60979300 | 61424000 | 61864100 | 62292200 | 62772900 | 63229600 | 63645100 | 64007200 | 64350200 | 64658900 | 64978700 | 65277000 | 65600400 | 66166000 | 66458200 | 66638400 | 66809800 | 67026200 | 67177600 | 67320200 | 67656700 |
| 11 | GBR: United Kingdom | 58785200 | 58999800 | 59239600 | 59501400 | 59793800 | 60182100 | 60620400 | 61073300 | 61571600 | 62042300 | 62510200 | 63022500 | 63495100 | 63905300 | 64351200 | 64853400 | 65379000 | 65844100 | 66273600 | 66647100 | 67081700 | 67330000 |
| 12 | GRC: Greece | 10775600 | 10836000 | 10888300 | 10915800 | 10940400 | 10969900 | 11004700 | 11036000 | 11060900 | 11094700 | 11119300 | 11123400 | 11086400 | 11003600 | 10926800 | 10858000 | 10783700 | 10768200 | 10741200 | 10724600 | 10718600 | 10678600 |
| 13 | HRV: Croatia | 4497740 | 4295410 | 4305490 | 4305380 | 4305730 | 4310860 | 4312490 | 4313530 | 4311970 | 4309800 | 4302850 | 4289860 | 4275980 | 4262140 | 4246810 | 4225320 | 4190670 | 4154210 | 4105490 | 4076250 | 4058170 | 4036360 |
| 14 | HUN: Hungary | 10221600 | 10200300 | 10174900 | 10142400 | 10116700 | 10097500 | 10076600 | 10066200 | 10045400 | 10031000 | 10014300 | 9985720 | 9931930 | 9908800 | 9877370 | 9855570 | 9830490 | 9797560 | 9778370 | 9772760 | 9769530 | 9730770 |
| 15 | ITA: Italy | 56923500 | 56960700 | 56987500 | 57130500 | 57495900 | 57874800 | 58064200 | 58223700 | 58652900 | 59000600 | 59190100 | 59364700 | 59394200 | 59685200 | 60782700 | 60795600 | 60665600 | 60589400 | 60484000 | 59816700 | 59641500 | 59236200 |
| 16 | NLD: Netherlands | 15864000 | 15987100 | 16105300 | 16192600 | 16258000 | 16305500 | 16334200 | 16358000 | 16405400 | 16485800 | 16575000 | 16655800 | 16730300 | 16779600 | 16829300 | 16900700 | 16979100 | 17081500 | 17181100 | 17282200 | 17407600 | 17475400 |
| 17 | NOR: Norway | 4478500 | 4503440 | 4524070 | 4552250 | 4577460 | 4606360 | 4640220 | 4681130 | 4737170 | 4799250 | 4858200 | 4920310 | 4985870 | 5051280 | 5107970 | 5166490 | 5210720 | 5258320 | 5295620 | 5328210 | 5367580 | 5391370 |
| 18 | POL: Poland | 38263300 | 38254000 | 38242200 | 38218500 | 38190600 | 38173800 | 38157100 | 38125500 | 38115600 | 38135900 | 38022900 | 38062700 | 38063800 | 38062500 | 38017900 | 38005600 | 37967200 | 37973000 | 37976700 | 37972800 | 37958100 | 37840000 |
| 19 | PRT: Portugal | 10249000 | 10330800 | 10394700 | 10444600 | 10473100 | 10494700 | 10512000 | 10532600 | 10553300 | 10563000 | 10573500 | 10572700 | 10542400 | 10487300 | 10427300 | 10374800 | 10341300 | 10309600 | 10291000 | 10276600 | 10295900 | 10298300 |
| 20 | ROU: Romania | 22455500 | 22430500 | 21833500 | 21627500 | 21521100 | 21382400 | 21257000 | 21130500 | 20635500 | 20440300 | 20294700 | 20199100 | 20096000 | 20020100 | 19947300 | 19870600 | 19760600 | 19643900 | 19533500 | 19414500 | 19328800 | 19201700 |
| 21 | SWE: Sweden | 8861430 | 8882790 | 8909130 | 8940790 | 8975670 | 9011390 | 9047750 | 9113260 | 9182930 | 9256350 | 9340680 | 9415570 | 9482860 | 9555890 | 9644860 | 9747360 | 9851020 | 9995150 | 10120200 | 10230200 | 10327600 | 10379300 |

**Figura 1.2:** Dataframe in R contenente il dataset di partenza

CAPITOLO 2

# Rappresentazione grafica dei dati

Nell'ambito della statistica e dell'analisi dei dati, l'uso di rappresentazioni grafiche è un ottimo strumento di presentazione che affianca quella tabellare e favorisce la comprensione del fenomeno statistico in esame. La visualizzazione dei dati tramite l'uso di grafici rende la loro interpretazione più semplice ed intuitiva rispetto alla forma tabellare, permettendo di riconoscere in maniera più veloce pattern, trend ed anomalie presenti nei nostri dati.

## 2.1 Serie Temporali

Le serie temporali vengono utilizzate nell'ambito della rappresentazione grafica dei dati per osservare il loro comportamento in un certo periodo di tempo. Nelle Serie temporali Il tempo rappresenta una variabile significativa per i dati, quest'utlimi infatti, verranno disposti in maniera cronologica. Su uno degli assi è sempre presente la variabile del tempo, che, in base all'analisi da fare, può rappresentare minuti, ore, giorni, settimane, etc... Per definire una serie temporale in R è necessario specificare vari parametri come: i valori della serie, istante iniziale temporale, istante finale (che può essere dedotto automaticamente considerando il numero di valori) e il numero di osservazioni nell'unità di tempo (frequenza). Nel caso del dataset considerato,

si ha che le colonne rappresentano i singoli anni, dal 2000 al 2021, di conseguenza, come istante iniziale temporale si indicherà l'anno 2000, mentre come istante finale si indicherà l'anno 2021, considerando un anno come frequenza (un'unica osservazione per ogni anno). In figura 2.1 è possibile osservare la serie temporale appena descritta.

```
Time Series:
Start = 2000
End = 2021
Frequency = 1
 [1] 19026200 19272100 19492800 19718400 19930400 20174500 20448600 20825100 21246500 21688800
[11] 22028700 22336900 22730400 23125200 23472800 23813100 24186300 24597200 24982700 25365700
[21] 25697300 25738100
```

**Figura 2.1:** Serie Temporale in R

Come valori della serie verranno considerati la popolazione di ogni nazione dall'anno 2000 all'anno 2021, per analizzare, nazione per nazione, se sono presenti cambiamenti anomali, cambi di tendenza, aumenti costanti o diminuzioni graduate.

### 2.1.1 Serie temporale popolazione Australia

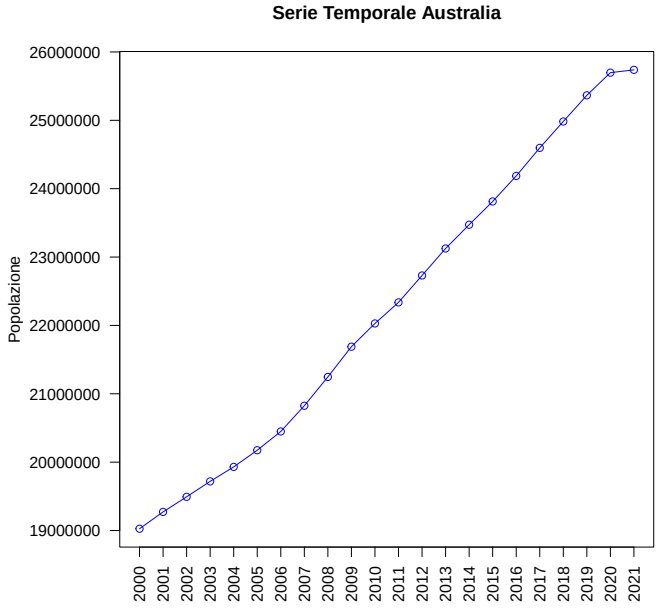


**Figura 2.2:** Serie Temporale Australia

Osservando il grafico in Figura 2.2 si può dedurre che:

- La popolazione **minima** si è registrata nel 2000, mentre la **massima** nel 2021
- La **crescita** della popolazione è stata **costante**, non diminuendo mai rispetto gli anni precedenti. L'unico anno in cui la popolazione ha avuto una **minor crescita** rispetto gli anni precedenti è stato il 2021, dove l'incremento è stato minimo (~40.000)
- la popolazione dal 2000 (19.026.200) al 2021 (25.738.100) è **aumentata** del **~35%**.

### 2.1.2 Serie temporale popolazione Bulgaria

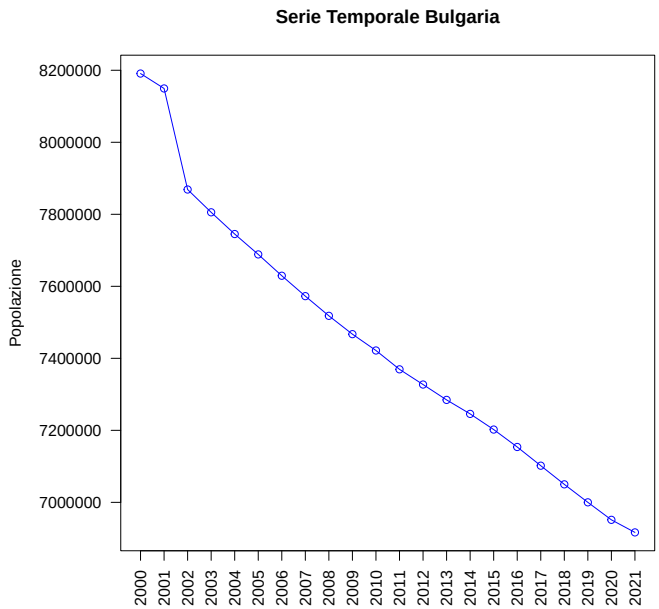


**Figura 2.3:** Serie Temporale Bulgaria

Osservando il grafico in Figura 2.3 si può dedurre che:

- Rispetto la maggior parte delle altre nazioni che vedremo, la Bulgaria ha avuto un **decrescita** costante nel tempo. Il **picco** discendente maggiore è tra il 2001 e 2002, con un decremento di **~280.000** abitanti.
- La popolazione **massima** (8.190.880) si è registrata nel 2000, mentre la **minima** (6.916.550) nel 2021, con una **decrescita** totale del **~16%**.

### 2.1.3 Serie temporale popolazione Germania

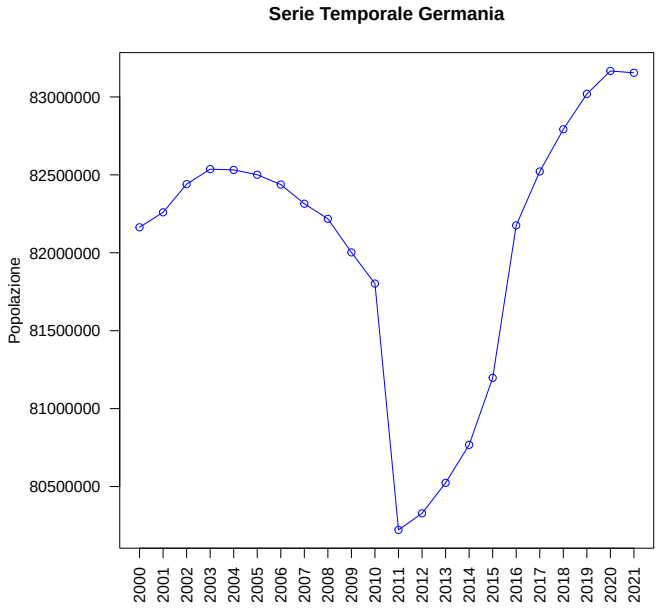


**Figura 2.4:** Serie Temporale Germania

Osservando il grafico in Figura 2.4 si può dedurre che:

- dal 2000 al 2003 la **crescita** è costante. Dal 2004 inizia una costante **discesa** fino all'anno 2011.
- Nel 2011 si presenta una **discesa anomala**, con una perdita di **~1.6 Milioni abitanti**
- Dal 2012 in poi la popolazione è in **costante crescita**. Il **picco maggiore** si è ottenuto nel 2016 con un **incremento** di quasi **1 milione** di abitanti.
- Nel 2021, dopo 10 anni di crescita costante, la popolazione ha una **lieve discesa** rispetto l'anno precedente, con una perdita di ~11.000 abitanti. **In generale**, la popolazione è leggermente **aumentata** rispetto al 2000 (82.163.500), registrando un $+\mathbf{1}\%$.

### 2.1.4 Serie temporale popolazione Spagna

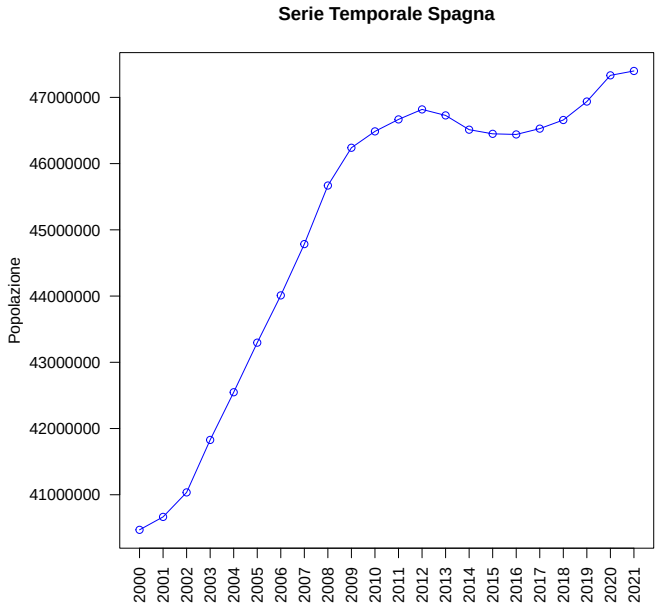


**Figura 2.5:** Serie Temporale Spagna

Osservando il grafico in Figura 2.5 si può dedurre che:

- dal 2000 al 2012 la **crescita** della popolazione è **graduale**.
- dal 2013 al 2016 è visibile una **lieve discesa** della popolazione.
- dal 2017 fino al 2021 la popolazione è in **costante crescita**, con un **incremento totale** del **17%** rispetto al 2000.

### 2.1.5 Serie temporale popolazione Grecia

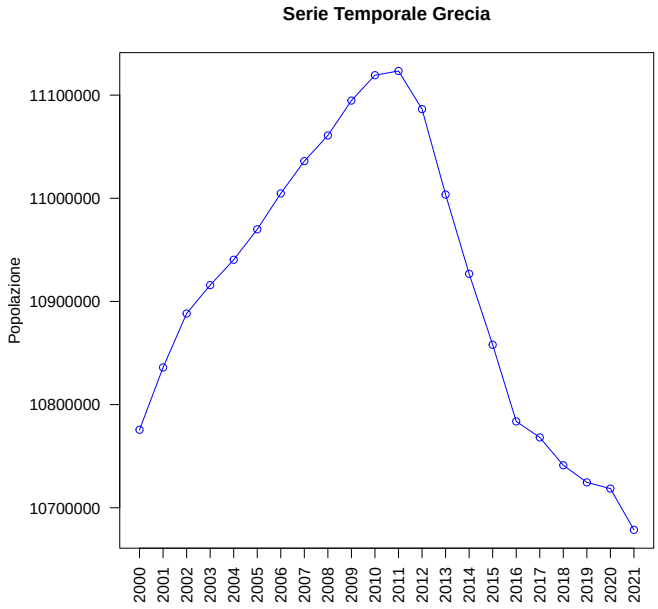


**Figura 2.6:** Serie Temporale Grecia

Osservando il grafico in Figura 2.6 si può dedurre che:

- dal 2000 al 2011 la **crescita** della popolazione è**graduale**.
- dal 2012 al 2021 la popolazione inizia a **decrescere**, fino ad arrivare, nel 2016, allo stesso numero del 2000.
- Nel 2021 (10.678.600) la popolazione raggiunge il **minimo storico** dal 2000 (10.775.600), con una **perdita totale** del **-1%**.

### 2.1.6 Serie temporale popolazione Croazia

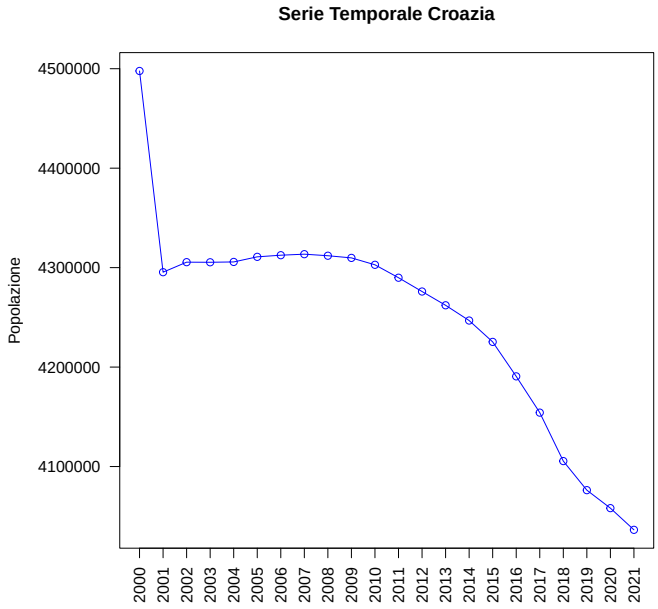


**Figura 2.7:** Serie Temporale Croazia

Osservando il grafico in Figura 2.7 si può dedurre che:

- La popolazioze, considerando l'intero periodo di tempo, è in **forte discesa**, anche se, dal 2001 a 2009 è **rimasta costante**, per poi **riscendere nuovamente**.
- Il **picco discendente maggiore** si è avuto nel 2001, dove la nazione ha perso ~**200.000 abitanti** (~5% della popolazione totale)
- Nel 2021 (4036360) la popolazione raggiunge il **minimo storico** dal 2000 (4.497.740), con una **perdita** del **~10%** degli abitanti.

### 2.1.7 Serie temporale popolazione Ungheria

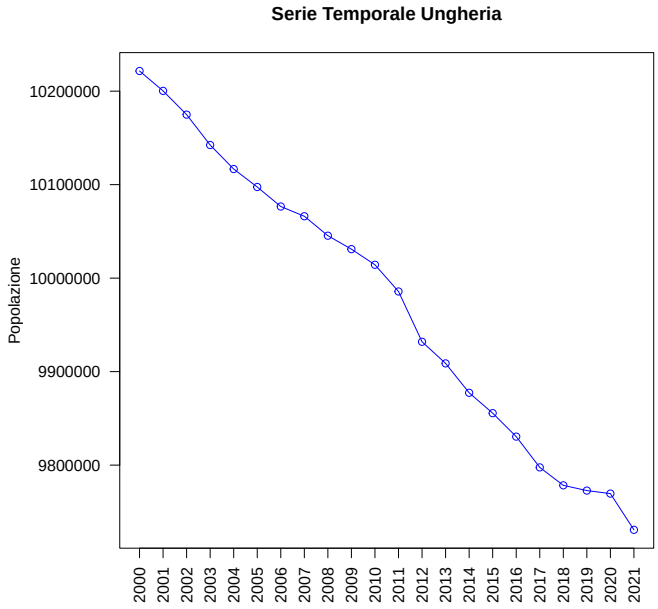


**Figura 2.8:** Serie Temporale Ungheria

Osservando il grafico in Figura 2.8 si può dedurre che:

- La popolazione è in una **costante discesa graduale** dal 2000, non registrando **nessun anno in crescita** (trend negativo).
- La popolazione **massima** si è avuta nel 2000 (10.221.600), mentre la **minima** nel 2021 (9.730.770), con una **perdita totale** del **5%** degli abitanti.

### 2.1.8 Serie temporale popolazione Italia

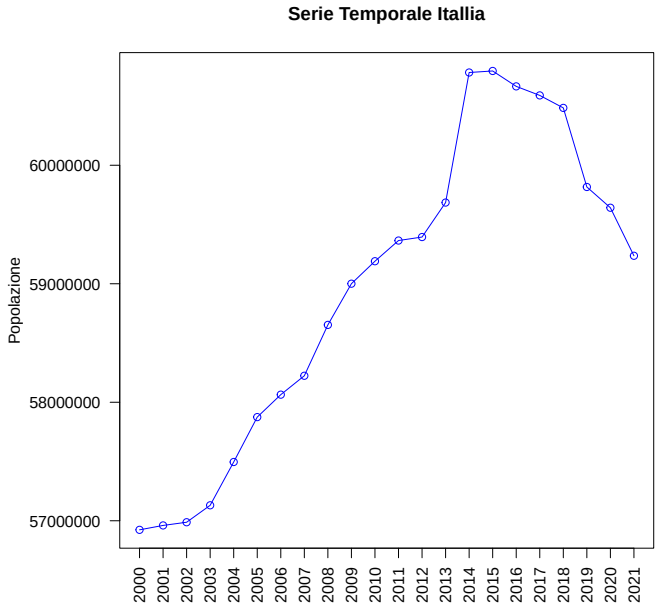


**Figura 2.9:** Serie Temporale Italia

Osservando il grafico in Figura 2.9 si può dedurre che:

- La popolazione dal 2000 al 2015 **(picco massimo)** ha avuto una **forte crescita**, dove il **l'incremento maggiore** si è registrato nel 2014 con un aumento di **~1.1 milioni** di abitanti rispetto l'anno precedente.
- dal 2015 al 2021 la popolazione italiana ha avuto una **discesa notevole**, raggiungendo, nel 2021, lo stesso numero di abitanti che aveva tra il 2010 e il 2011.
- Tuttavia, la popolazione italiana dal 2000 (56.923.500) al 2021 (59.236.200) è **aumentata** del **4%**.

### 2.1.9 Serie temporale popolazione Polonia

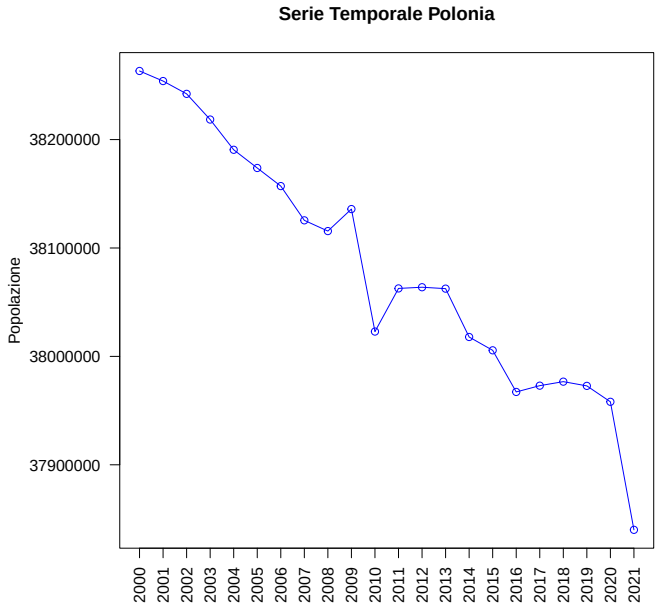


**Figura 2.10:** Serie Temporale Polonia

Osservando il grafico in Figura 2.10 si può dedurre che:

- La popolazione è in una **discesa graduale**, nonostante in vari periodi di tempo, abbia avuto delle crescite.
- I **picchi in negativo** più importanti in negativo si sono avuti nel 2010 e nel 2021. Nel 2021 (37.840.000) la nazione ha raggiunto il **minimo storico** dal 2000 (38.263.300), con una **perdita totale** del **1%** degli abitanti.

### 2.1.10 Serie temporale popolazione Portogallo

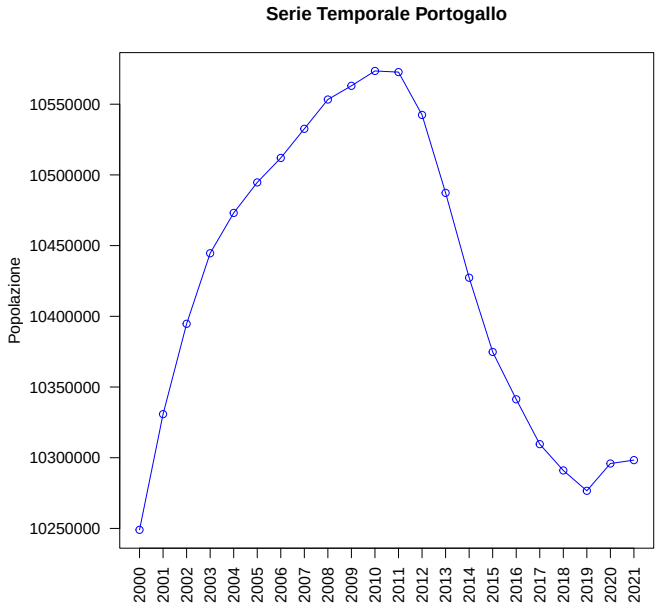


**Figura 2.11:** Serie Temporale Portogallo

Osservando il grafico in Figura 2.11 si può dedurre che:

- Dal 2000 al 2010 ha avuto un **elevata crescita**, con **grandi picchi** visibili dal 2000 al 2003.
- Dal 2011 al 2019 la popolazione ha avuto una **forte discesa**, per poi, **ricrescere** nel 2020 e 2021. Tuttavia la crescita in questi ultimi due anni non è stata significativa, infatti, il numero di abitanti nel 2021 sono gli stessi che la nazione aveva tra il 2000 e il 2001, registrando un **incremento totale** poco significativo del **0,5%**.

### 2.1.11 Serie temporale popolazione Romania

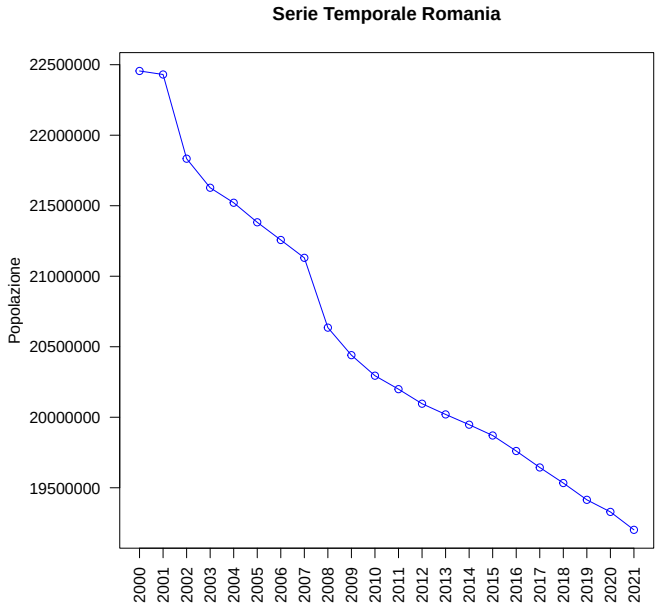


**Figura 2.12:** Serie Temporale Romania

Osservando il grafico in Figura 2.12 si può dedurre che:

- La popolazione è in una **costante discesa graduale** dal 2000, non registrando **nessun anno in crescita (trend negativo)**.
- I **picchi** di **decrescita** maggiori si sono ottenuti nel 2002 e nel 2008, dove in entrambi casi, c'è stata una decrescita del circa **3%** rispetto l'anno precedente.
- La Romania in 21 anni ha perso quasi il **15%** della propria popolazione.

### 2.1.12 Serie temporale nazioni rimanenti

Le restanti nazioni: Austria, Belgio, Svizzera, Danimarca, Finlandia, Francia, UK, Olanda, Norvegia, Svezia, presentano un grafico **molto simile tra loro**, che si avvicina molto a quello in Figura 2.2; cioè un grafico che presenta un **trend di crescita** della popolazione, con **nessuna annata in negativo**, dove il valore minimo di popolazione si è registrato nel 2000 e il massimo nel 2021. In Figura 2.13 è possibile osservare un esempio di **confronto** tra due serie temporale di nazioni presenti nella lista

appena menzionata. In questo caso le nazioni considerate sono Norvegia e UK, dove nonostante la grande differenza di popolazione, è possibile osservare un tipo di **crescita molto simile tra loro**.

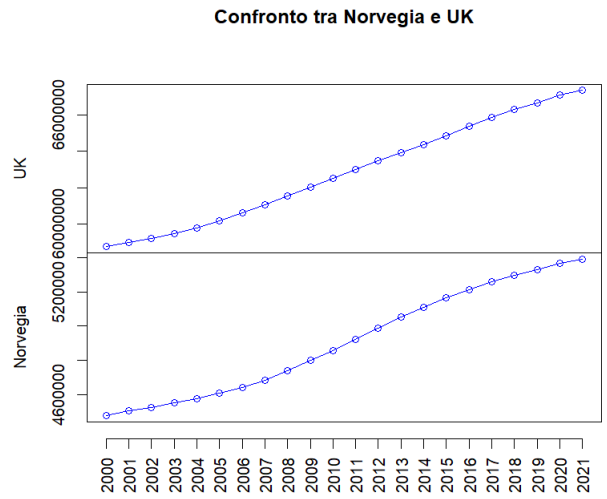


**Figura 2.13:** Confronto Serie Temporale UK e Norvegia

## 2.2 Diagramma a barre (Barplot)

Un diagramma a barre, anche conosciuto come Barplot, è un tipo di grafico che mostra la relazione tra le variabile quantitative e qualitative (numeriche e categoriche). Ogni valore assunto dalla variabile categorica viene rappresentano sul grafico con una barra, la quale altezza (taglia) sarà uguale al valore quantitativo a essa associata. Su un asse vengono posti i valori che può assumere la variabile, mentre, sull'asse opposto vengono posti i valori numerici a essi associati. Viene utilizzato molto spesso per la visualizzazione della distribuzione delle variabile categoriche. In questo caso il barplot verrà utilizzato per visualizzare la popolazione di ogni nazione per ogni anno, andando ad individuare in maniera veloce e rappresentatova le nazioni con popolazione massima, minima e simili tra loro. Verranno utilizzate, a supporto del lettore, delle linee aggiuntive sul grafico che aiuteranno ad individuare correttamente le nazioni con popolazione minima e massima dell'anno a cui fa riferimento il barplot. Sull'asse delle X verranno mostrate le nazioni, mentre sull'asse delle Y la popopolazione.

### 2.2.1 Barplot Anno 2000

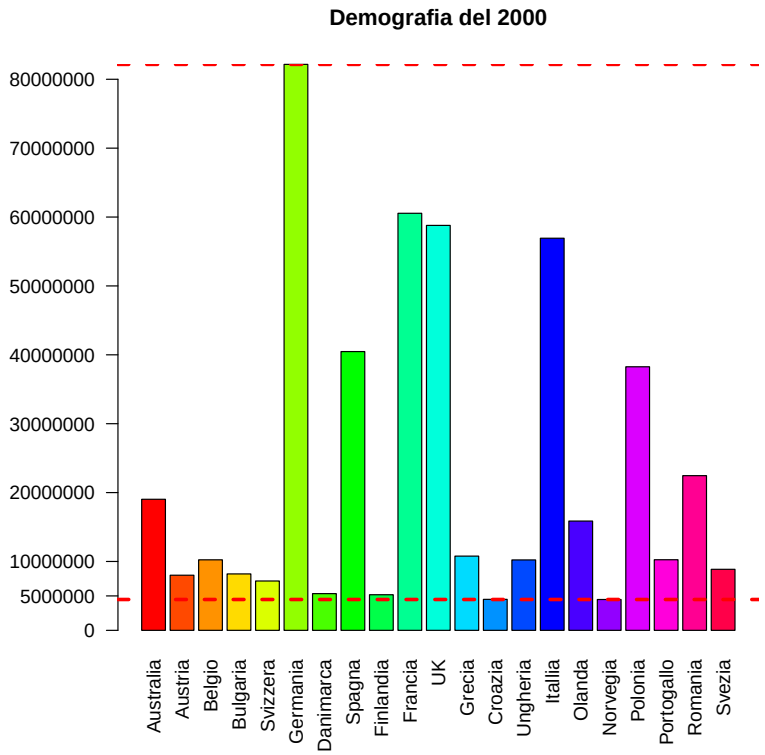


**Figura 2.14:** Barplot Anno 2000

Osservando il grafico in Figura 2.14 si può dedurre che:

- Nonostante Norvegia e Croazia sembrano avere la stessa popolazione, è la **Norvegia** ad avere il **minimo numero di abitanti** rispetto le restanti nazioni, con una differenza di ~2000 persone rispetto la Croazia.
- La nazione con il **massimo numero di abitanti** è la **Germania**.

### 2.2.2 Barplot Anno 2001

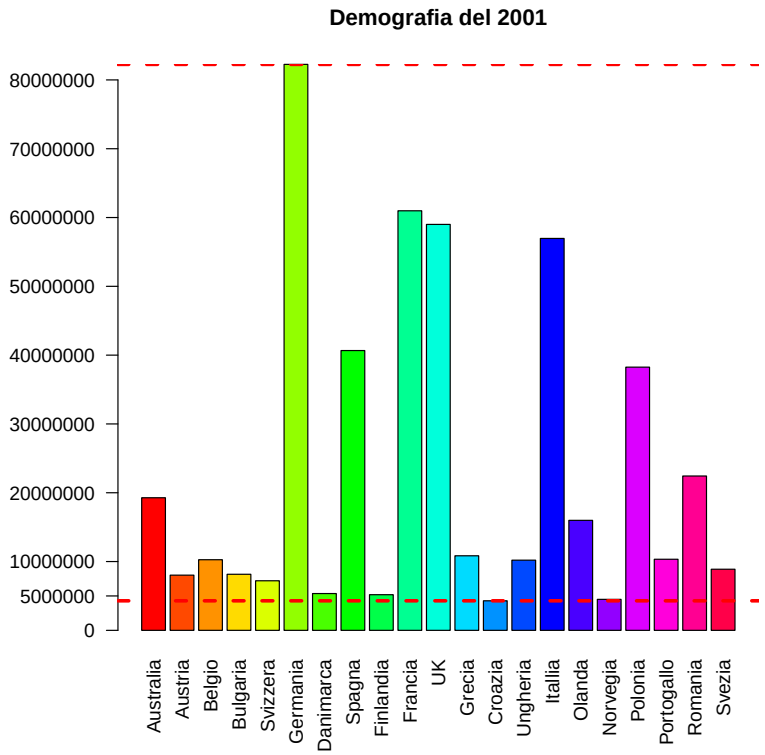


**Figura 2.15:** Barplot Anno 2001

Osservando il grafico in Figura 2.15 si può dedurre che:

- La **Croazia** diventa la nazione con il **minor numero di abitanti** a seguito di una perdita del 5% della propria popolazione rispetto l'anno precedente (osservabile in Figura 2.7), mentre la Norvegia aumenta il numero della propria popolazione.
- La nazione con il massimo numero di abitanti è sempre Germania.

### 2.2.3 Barplot Anno 2002

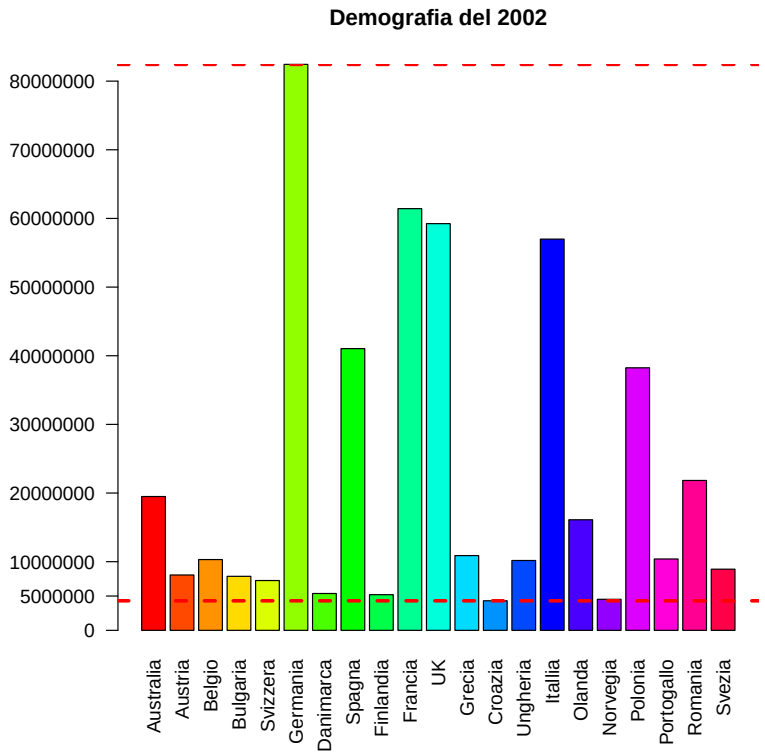


**Figura 2.16:** Barplot Anno 2002

Osservando il grafico in Figura 2.16 si può dedurre che:

- L'Austria **supera** il numero di abitanti della Bulgaria, in seguito ad un **picco di perdita** registrato nel 2002 da quest'ultima (osservabile in Figura 2.3).
- **Croazia** e **Germania** continuano ad essere le nazioni con il **minor** e **maggior** numero di abitanti.

### 2.2.4 Barplot dall'anno 2003 al 2007

Dal 2003 al 2007 la situazione generale non cambia, la "posizione" delle nazioni rispetto le altre è la stessa, con Germania che rappresenta la nazione con il maggior numero di abitanti e la Croazia con il minore.

### 2.2.5 Barplot Anno 2008

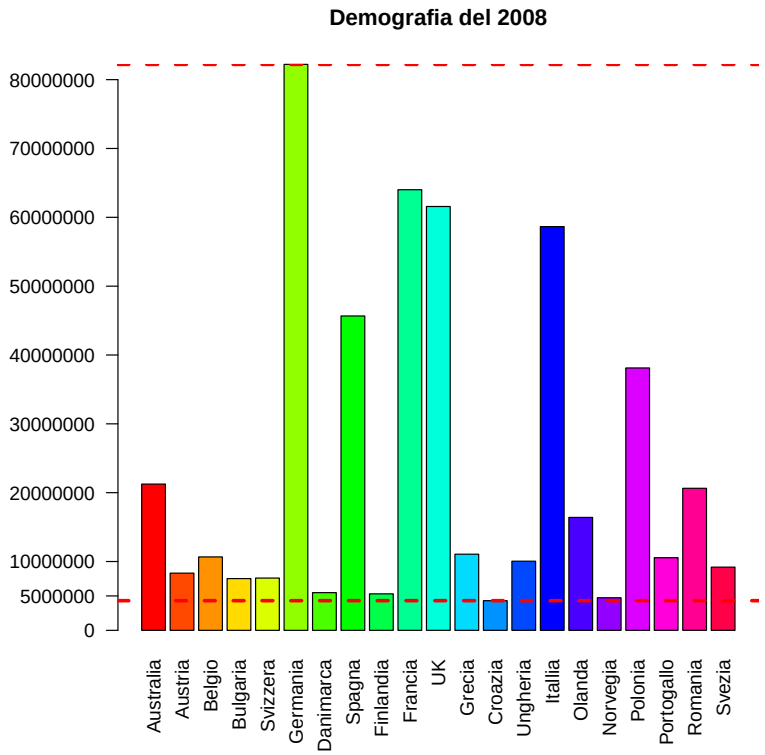


**Figura 2.17:** Barplot Anno 2008

Osservando il grafico in Figura 2.17 si può dedurre che:

- In seguito a un trend negativo costante della Bulgaria e ad un trend positivo costante della Svizzera, quest'ultima supera il numero di abitanti della Bulgaria.

### 2.2.6 Barplot Anno 2009 e 2010

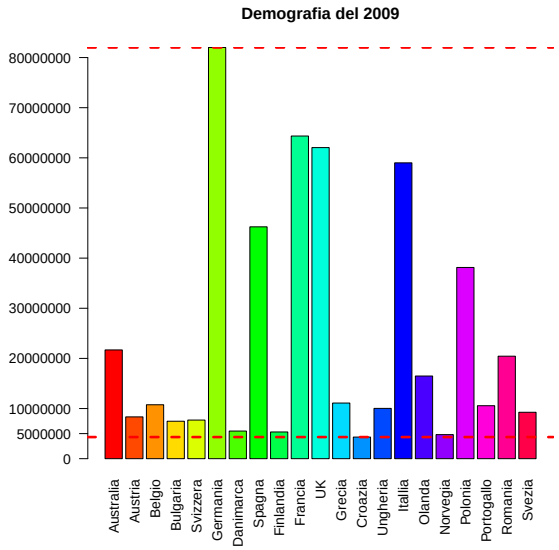


**Figura 2.18:** Barplot anno 2009

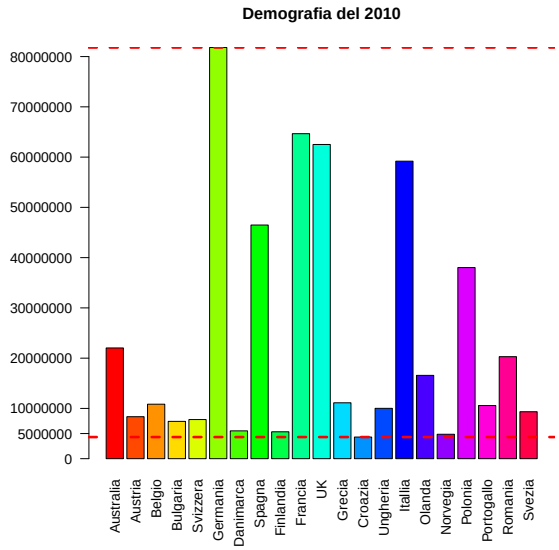


**Figura 2.19:** Barplot anno 2010

Nel 2009 e 2010 la situazione generale non cambia,

### 2.2.7 Barplot Anno 2011

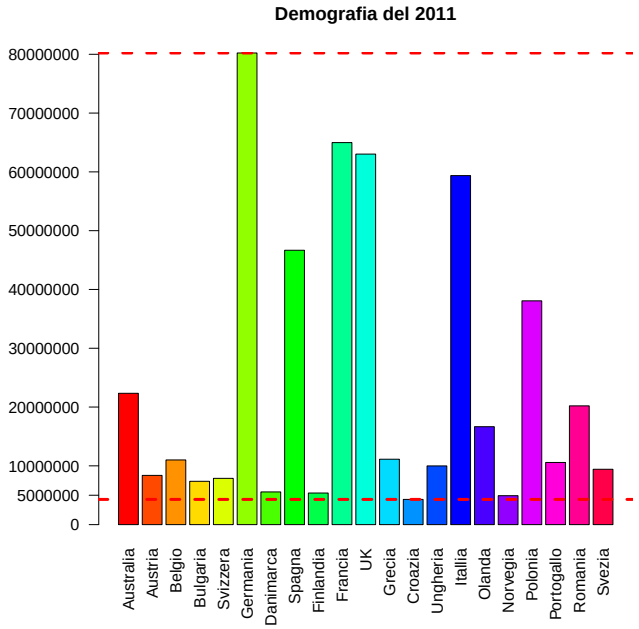


**Figura 2.20:** Barplot Anno 2011

Osservando il grafico in Figura 2.20 si può osservare la **discesa anomala** che la **Germania** ha avuto nel 2011 (osservabile in Figura 2.4). Tale picco in discesa è visibile dal grafico poichè la **distanza** tra la Germania e la **Francia** (Seconda nazione per popolazione) **si accorcia** in maniera significativa, e inoltre, per la prima volta, l'altezza della barra rappresentante la Germania è a livello con il valore 80 Milioni.

### 2.2.8 Barplot dall'anno 2012 al 2019

Dal 2012 al 2019 la situazione generale non cambia.

### 2.2.9 Barplot Anno 2020

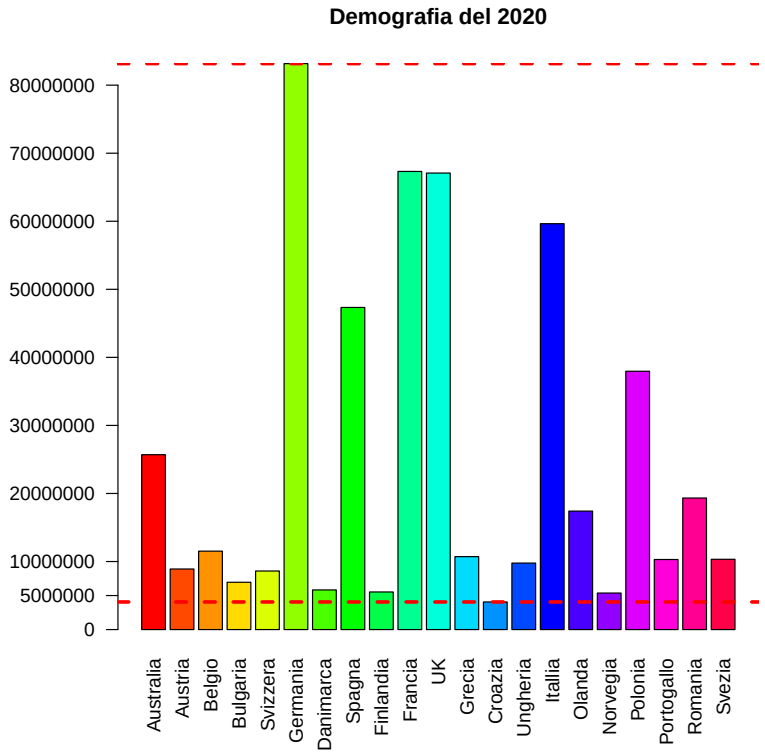


**Figura 2.21:** Barplot Anno 2020

Osservando il grafico in Figura 2.21 si può osservare che per la prima volta le barre della **Francia** e **UK** alla stessa altezza, con una differenza di soli ~250.000 abitanti tra le due (Francia 67320200, UK 67081700), in seguito ad una **crescita più rapida** del Regno Unito (UK) rispetto alla Francia negli anni precedenti.

### 2.2.10 Barplot Anno 2021

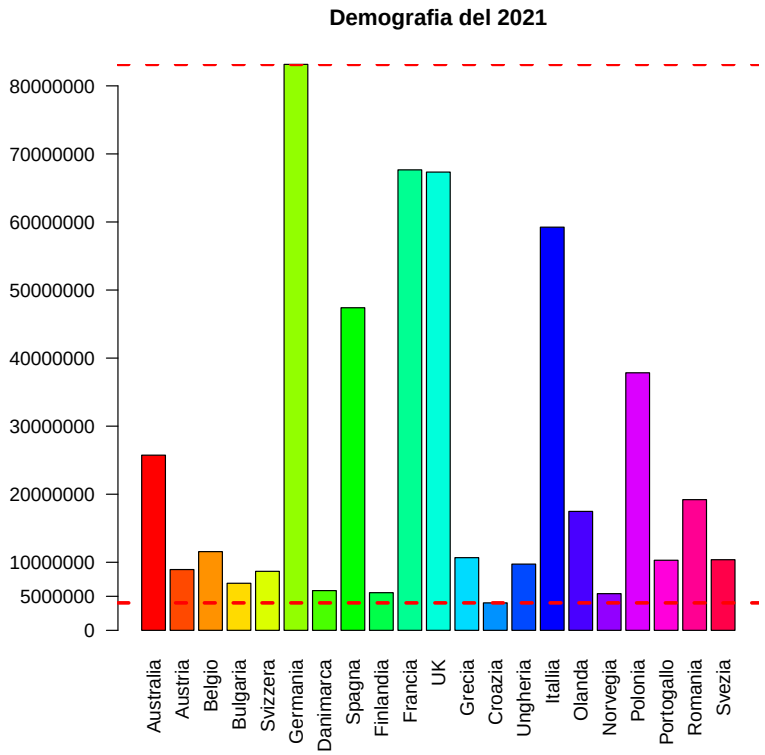


**Figura 2.22:** Barplot Anno 2021

Osservando il grafico in Figura 2.22 si può dedurre la **Francia aumentare il divario sul Regno Unito** rispetto l'anno precedente. Considerando il quadro generale nulla è cambiato. In conclusione: La **Germania** è stata la nazione con il **maggior numero di abitanti dal 2000**, mentre la **Croazia** dal 2001 è stata la nazione con il **minor numero** (nel 2000 è stata la **Norvegia**).

## 2.3 Distribuzione di frequenza

La distribuzione di frequenza consiste nel calcolo della frequenza assoluta e relativa delle singole modalità che una variabile X può assumere. In caso di variabile qualititative le modalità rappresentano delle qualità distinte che la variabile può assumere. Se la variabile X è invece quantitativa, le modalità rappresentano dei numeri reali distinti. La frequenza assoluta e relativa vengono calcolate per descrivere e capire la distribuzione di un insieme di dati. La differenza tra le due frequenze è:

- la **frequenza assoluta** rappresenta il numero di volte che un determinato valore/qualità appare nell'insieme di dati e viene espressa come un conteggio numerico, che indica quante volte quel valore/qualità appare nel campione considerato (insieme di dati).
- la **frequenza relativa** invece rappresenta la proporzione o percentuale di volte che un determinato valore/qualità appare nel campione rispetto al totale degli elementi. Viene espressa come una frazione o una percentuale.

Nel caso del dataset considerato, si hanno variabili **quantitative continue**, in quanto, ogni colonna X rappresenta il numero di abitanti che una nazione Y contava nell'anno X. Di conseguenza, trattandosi di numeri continui e di popolazione, è **improbabile avere più occorrenze dello stesso valore**, quindi, calcolare la frequenza relativa e assoluta non **avrebbe senso** (**le modalità distinte non sono ben definite**). Proprio per questo, l'approccio seguito è stato quello di **raccogliere** le informazioni in **classi**, in modo da avere **modalità ben definite** in cui i singoli valori continui possono essere inseriti. Le **classi** definite sono **5**, e rappresentano degli intervalli di popolazione: **0-5M, 5M-10M, 10M-20M, 20M-60M, 60M-100M**. La fraquenza assoluta e relativa sono state calcolate per ogni anno (colonna) considerando tali intervalli di popolazione.

### 2.3.1 Barplot e grafico a torta

Avendo trasformato la distribuzione dei dati da quantitativa a qualitativa aggiungendoli in delle classi, possiamo utilizzare grafici consigliati per la rappresentazione della distribuzione di frequenze di variabili qualitative, cioè: **Barplot** e **grafico a torta**. Per rappresentare graficamente come vengono distribuite le popolazioni delle varie

nazioni nei vari intervalli ogni anno (frequenza assoluta) verranno utilizzati i barplot. Per rappresentare graficamente la frequenza relativa per ogni anno verrà utilizzato il grafico a torta. Esso è un grafico circolare diviso in settori la cui ampiezza è proporzionata alle frequenze. Solitamente tale grafico viene utilizzato per **dati categorici**, e non numerici, per evitare che, per la grande quantità di numeri, il grafico venga diviso in settori **troppo piccoli e difficili da leggere**. Vicino ad ogni settore circolare, verranno aggiunte le percentuali che rappresentano le frequenze relative al fine di evitare ambiguità nella lettura.

### 2.3.2 Frequenza assoluta e relativa anno 2000

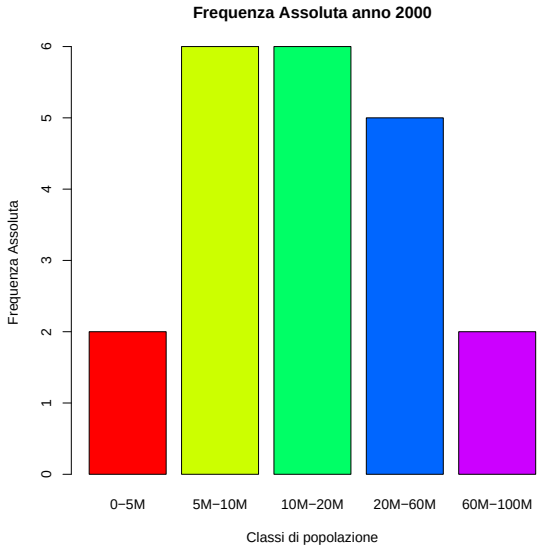


**Figura 2.23:** Frequenza assoluta 2000

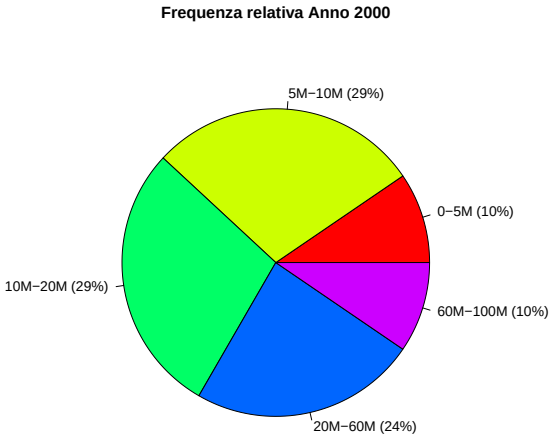


**Figura 2.24:** Frequenza relativa 2000

Osservando il grafico 2.23 e 2.24 si può dedurre che:

- Gli intervalli **più popolati** sono: 5M-10M e 10M-20M, ciascuno comprendente 6 nazioni e rappresentando il ~29% del totale delle nazioni. Entrambi gli intervalli comprendono il **~58%** delle nazioni totali.
- Gli intervalli **meno popolati** sono: 0-5M e 60M-100M, ciascuno comprendente 2 nazioni e rappresentando il **~10%** totale delle nazioni.

### 2.3.3 Frequenza assoluta e relativa dal 2001 al 2004

Dal 2001 e 2004 la distribuzione delle frequenze è rimasta invariata rispetto ai grafici in Figura 2.23 e Figura 2.24.

### 2.3.4 Frequenza assoluta e relativa anno 2005

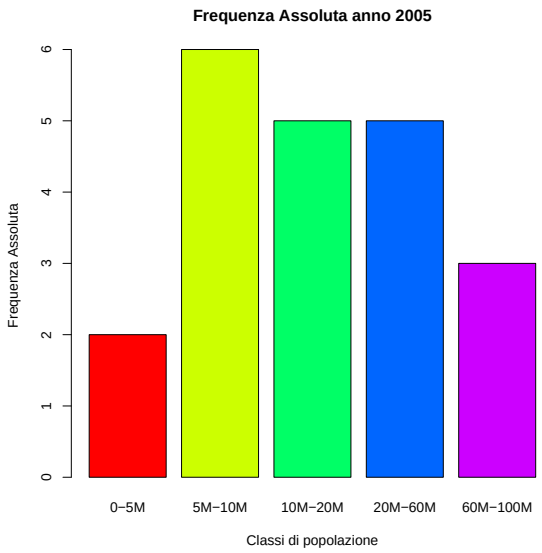


**Figura 2.25:** Frequenza assoluta 2005

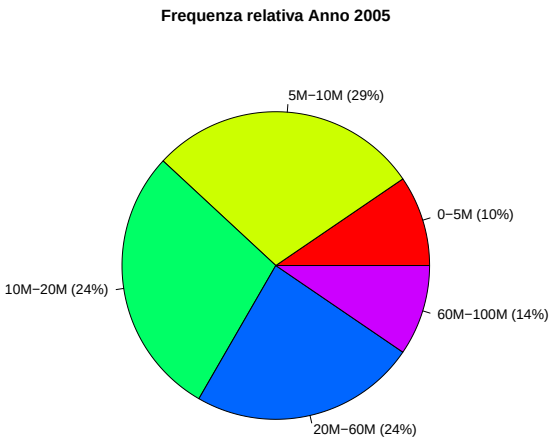


**Figura 2.26:** Frequenza relativa 2005

Osservando il grafico 2.25 e 2.26 si può dedurre che:

- L'intervallo **10M-20M** registra una **diminuzione** di una nazione (5 su 21, ~24%) rispetto gli anni precedenti (6 su 21, ~29%), a causa dell'**aumento** di popolazione dell'Australia, che ha raggiunto i ~20.1 milioni di abitanti, collocandosi nell'intervallo **20M-60M**.

- La frequenza assoluta e relativa dell'intervallo 20M-60M **non cambia** nonostante l'ingresso dell'Australia a causa del Regno Unito, che ha raggiunto i ~60.2 milioni di abitanti, collocandosi nell'intervallo 60M-100M, determinandone un aumento di frequenza (3 su 21, ~14% del totale)

- L'intervallo **5M-10M** diventa l'intervallo **più popolato**, con 6 nazioni su 21 che rappresentano il **~29%** delle nazioni totali. L'intervallo **0-5M**, a seguito della crescita dell'intervallo 60M-100M, diventa il **meno popolato** (2 su 21, ~10%).

### 2.3.5 Frequenza assoluta e relativa dal 2006 al 2010

Dal 2001 e 2004 la distribuzione delle frequenze è rimasta invariata rispetto ai grafici in Figura 2.25 e Figura 2.26.

### 2.3.6 Frequenza assoluta e relativa anno 2011

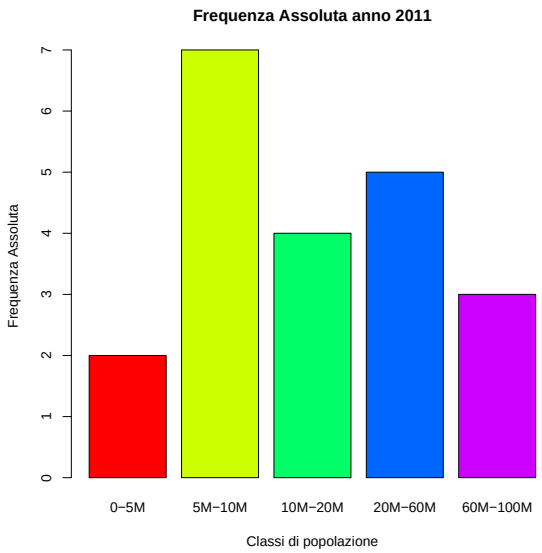


**Figura 2.27:** Frequenza assoluta 2011

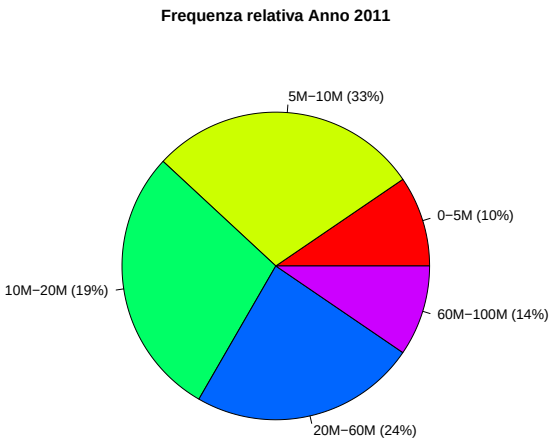


**Figura 2.28:** Frequenza relativa 2011

Osservando il grafico 2.27 e 2.28 si può dedurre che:

- L'intervallo 10M-20M registra nuovamente la **diminuzione** di una nazione (4 su 21, ~19%), a causa della continua decrescita di popolazione dell'Ungheria che raggiunge i ~9,9M di abitanti, collocandosi nell'intervallo 5M-10M, determinandone un **aumento di frequenza** (7 su 21, ~33% del totale)
- L'intervallo **5M-10M** è diventato **notevolmente ampio**, contenendo più di un **terzo delle nazioni totali**.

### 2.3.7 Frequenza assoluta e relativa anno 2012

Nel 2012 la distribuzione delle frequenze è rimasta invariata rispetto ai grafici in Figura 2.27 e Figura 2.28.

### 2.3.8 Frequenza assoluta e relativa anno 2013

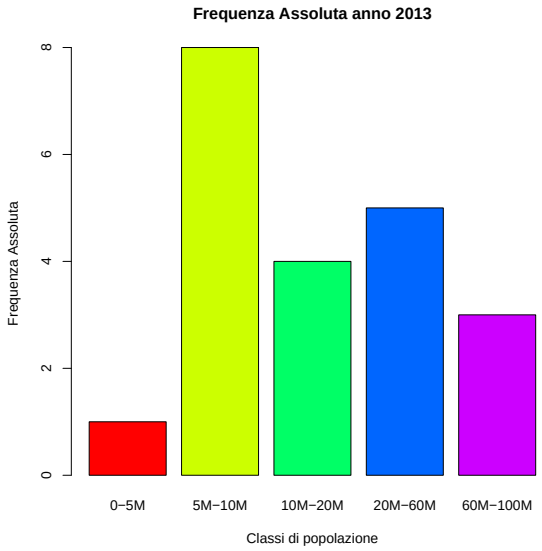


**Figura 2.29:** Frequenza assoluta 2013

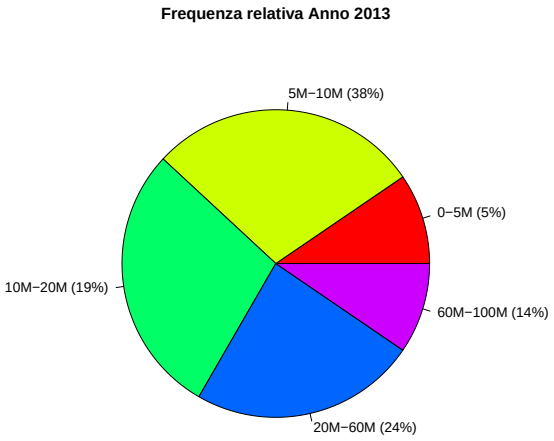


**Figura 2.30:** Frequenza relativa 2013

Osservando il grafico 2.29 e 2.30 si può dedurre che:

- **A causa della crescita graduale e costante della Norvegia**, che raggiunge i ~5.05 Milioni di abitanti, l'intervallo 0-5M registra la **diminuzione** di una nazione, mentre l'intervallo 5-10M registra una **crescita**.
- Di conseguenza, l'intervallo **0-5M** diventa **l'unico** ad avere un **unica nazione**, con una **frequenza relativa** del **~5%.** L'intervallo **5M-10M diventa ancor più ampio**, con una **frequenza relativa** del **~38%.**

### 2.3.9 Frequenza assoluta e relativa anno 2014

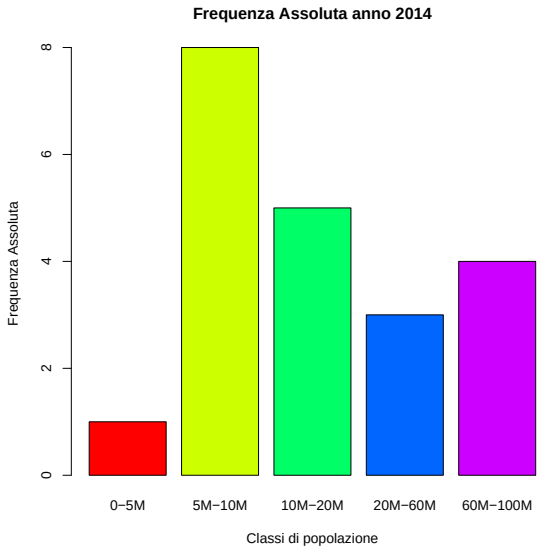


**Figura 2.31:** Frequenza assoluta 2014

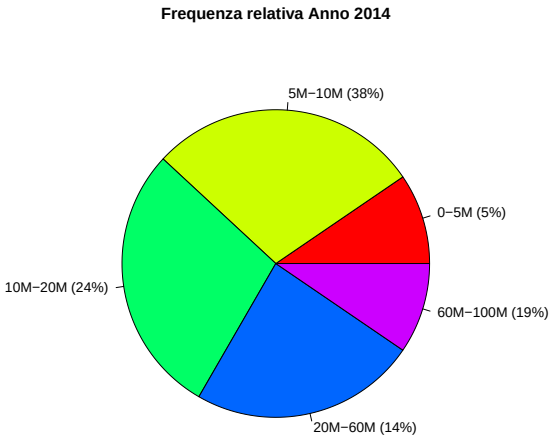


**Figura 2.32:** Frequenza relativa 2014

Osservando il grafico 2.31 e 2.32 si può dedurre che:

- L'intervallo **20M-60M** registra per la prima volta una **doppia diminuzione** a causa del picco di crescita avuto dall'Italia (Fig 2.9), che l'ha portata a raggiungere una popolazione di ~60.7M di abitanti, collocandosi nell'intervallo 60-100M e a causa della decrescita graduale e costante della Romania che ha raggiunto i ~19.94M di abitanti, collocandosi nell'intervallo 10M-20M.
- Di conseguenza: l'intervallo 20M-60M contiene 3 nazioni, ~14% del totale, l'intervallo 60M-100M contiene 4 nazioni, ~19% del totale e l'intervallo 10M-20M contiene 5 nazioni, ~24% del totale.

### 2.3.10 Frequenza assoluta e relativa dal 2015 al 2017

Dal 2014 al 2017 la distribuzione delle frequenze è rimasta invariata rispetto ai grafici in Figura 2.31 e Figura 2.32.

### 2.3.11 Frequenza assoluta e relativa anno 2018

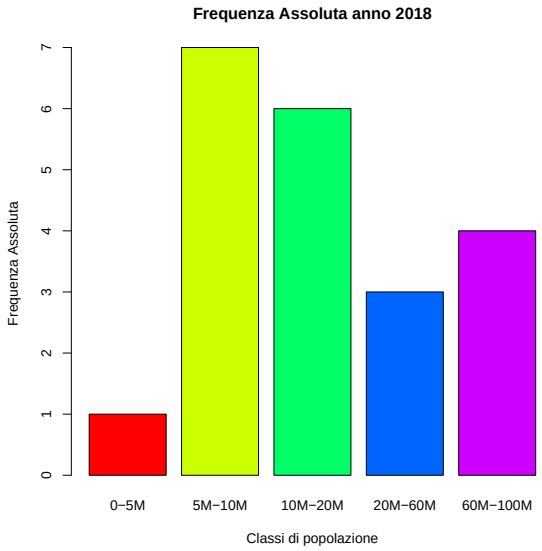


**Figura 2.33:** Frequenza assoluta 2018

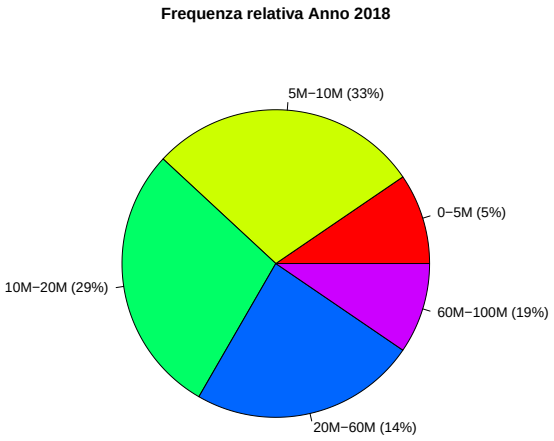


**Figura 2.34:** Frequenza relativa 2018

Osservando il grafico 2.33 e 2.34 si può dedurre che:

- La crescita graduale e costante negli anni della Svezia ha portato la nazione ad avere nel 2018 ~10.1M di abitanti, collocandosi nell'intervallo 10M-20M e facendo registrare una **diminuzione nell'intervallo 5M-10M.**
- Di conseguenza, l'intervallo 5M-10M contiene 7 nazioni su 21, ~33% del totale, mentre, l'intervallo 10M-20M è composto da **6 nazioni**, **~29% del totale**.

### 2.3.12 Frequenza assoluta e relativa anno 2019

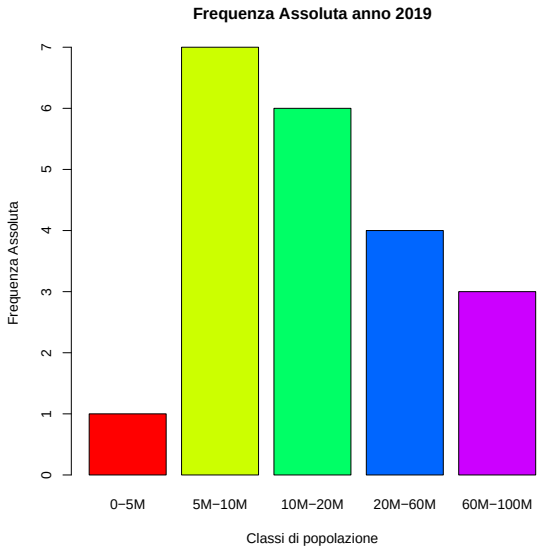


**Figura 2.35:** Frequenza assoluta 2019

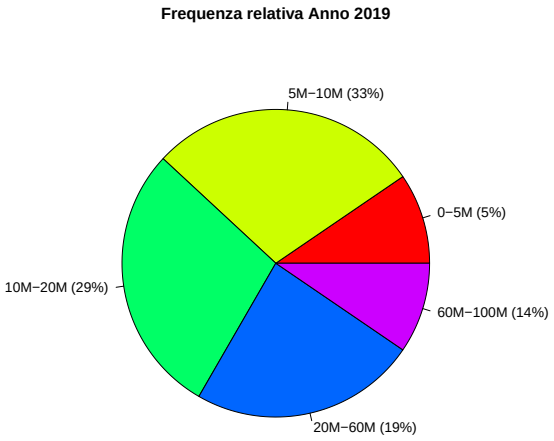


**Figura 2.36:** Frequenza relativa 2019

Osservando il grafico 2.35 e 2.36 si può dedurre che:

- L'intervallo 60M-100M ha registrato una diminuzione a causa dell'Italia che nel 2019 ha avuto un picco di decrescita che ha portato la popolazione a ~59.8M abitanti, collocandola nell'intervallo 20M-60M.
- L'intervallo 60M-100M contiene 3 nazioni, ~14% del totale, mentre, l'intervallo 20M-60M contiene 4 nazioni, ~19% del totale.

### 2.3.13 Frequenza assoluta e relativa del 2020 e 2021

Nel 2020 e 2021 la distribuzione delle frequenze è rimasta invariata rispetto ai grafici in Figura 2.35 e Figura 2.36.

### 2.3.14 Conclusioni

In conclusione, considerando le distribuzioni di frequenze dal 2000 al 2021, si può affermare che la **maggior parte delle nazioni si colloca negli intervalli compresi tra i 5 milioni e i 20 milioni (5M-10M e 10M-20M)**, infatti, la somma di quest'ultimi due

intervalli ha compreso **più del 50% delle nazioni totali**, fino ad arrivare ad **oltre il 60% dal 2014 in poi**. L'intervallo **meno ampio e popolata** invece è stato l'intervallo **0-5M**, che rappresentava una fascia di popolazione **estremamente bassa**, e che quindi, ha contenuto al **massimo 2 nazioni** (fino al 2012) per poi passare a **1 dal 2013 in poi**, rappresentando la **minima parte all'interno delle frequenze**. L'intervallo **60M-100M** rappresentava una fascia di popolazione **estremamente alta**, infatti, non era **comune e ampio** come le fasce intermedie, registrando una frequenza relativa minima di 10% dal 2000 al 2004 e massima di 19% dal 2014 al 2019, per poi riscendere al 14% dal 2019 in poi.
Al termine di tale analisi si può affermare che le nazioni considerate hanno una **tendenza generale verso una dimensione della popolazione medio-piccola.**

## 2.4 Boxplot

Il boxplot, noto anche come "Scatola con baffi", è una rappresentazione grafica di un insieme di dati statistici attraverso la visualizzazione di quartili. La scatola rappresenta i valori che si trovano in corrispondenza tra Q1 e Q3. La linea orrizontale all'interno della scatola rappresenta Q2. L'estremo del baffo inferiore rappresenta il valore più piccolo tra le osservazioni che risulta maggiore o uguale di:

$$Q_1 - 1.5 \cdot (Q_3 - Q_1)$$

L'estremo del baffo superiore rappresenta il valore più grande tra le osservazioni che risulta minore o uguale di:

$$Q_3 + 1.5 \cdot (Q_3 - Q_1)$$

Possono esserci valori al di fuori dell'intervallo superiore ed inferiore, definiti come **anomali** o **outlier**. Essi sono rappresentati con dei punti nel grafico e sono dati che si **discostono** notevolmente dalla maggior parte degli altri. Tale grafico viene utilizzato per mostrare alcune caratteristiche di una distribuzione di frequenza, come: **centralità (mediana), forma, dispersione e presenza di valori anomali**.

### 2.4.1 Boxplot anno 2000

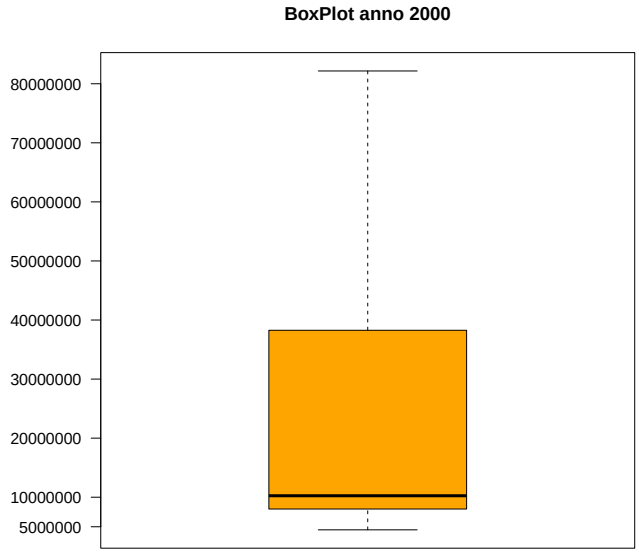


**Figura 2.37:** Boxplot anno 2000

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  4478500  8002190 10249000 23222771 38263300 82163500
```

Osservando il grafico 2.37 possiamo dedurre che:

- il baffo inferiore e superiore sono pari al minimo e al massimo dei valori, **non sono quindi presenti outlier** (ogni valore è compreso tra i due baffi).
- La forma è **molto assimetrica** in quanto la distanza tra $Q_3$ e $Q_2$ è molto **più grande** rispetto quella tra $Q_2$ e $Q_1$ (la **mediana** è molto vicina a $\mathbf{Q_1}$).

### 2.4.2 Boxplot dal 2001 al 2016

Durante il periodo che va dal 2000 al 2016 i boxplot hanno mantenuto una forma e una disposizione simile a Figura 2.37, con una **forte asimmetria** e con **l'assenza di valori anomali.**

### 2.4.3 Boxplot anno 2017

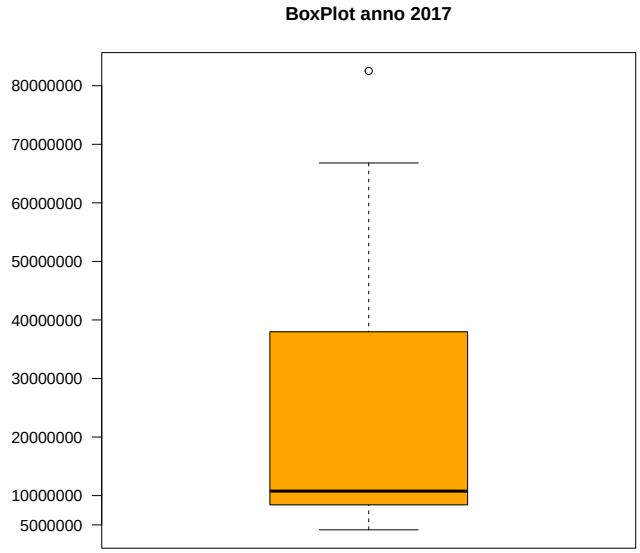


**Figura 2.38:** Boxplot anno 2017

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  4154210  8419550 10768200 24703319 37973000 82521700
```

Osservando il grafico 2.38 possiamo dedurre che:

- Il baffo inferiore è pari al valore minimo, mentre il baffo superiore è pari a 66.809.800, che non corrisponde al valore massimo: **c'è un valore anomalo.**
- Il valore anomalo si trova in corrispondenza del **valore massimo**: **82.521.700**, che rappresenta la popolazione della **Germania** nel 2017. Viene considerato anomalo poichè è maggiore rispetto al valore:

$$Q_3 + 1.5 \cdot (Q_3 - Q_1) = 82.303.175$$

  Il baffo superiore si trova in corrispondenza di 66.809.80 poichè rappresenta il maggior valore tra quelli minori di 82.303.175. La **causa** principale è dovuto **all'enorme picco di crescita** che la popolazione tedesca ha avuto nel 2016 e 2017.

- Il grafico presenta sempre una **forte asimmetria**.

### 2.4.4 Boxplot dal 2018 al 2021

Durante il periodo che va dal 2018 al 2021 i boxplot hanno mantenuto una forma e una disposizione simile a quella in Figura 2.38,

### 2.4.5 Confronto tra Baxplot anno 2000 e 2021

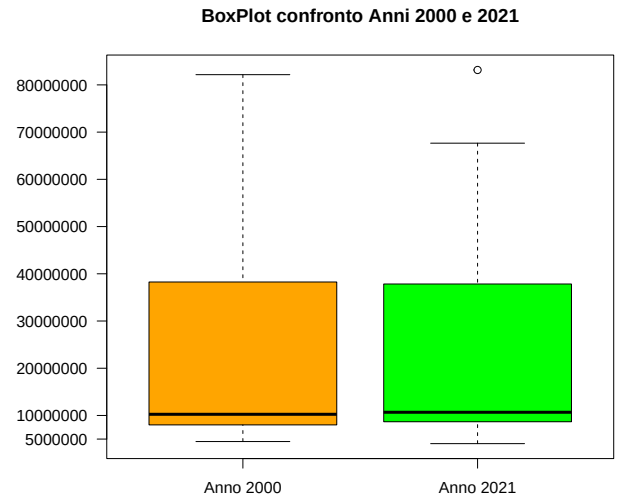


**Figura 2.39:** Boxplot confronto anno 2000 e 2021

Osservando il grafico 2.39 possiamo dedurre che:

- La differenza principale tra i due boxplot sta nel **valore anomalo** presente esclusivamente nell anno 2021, a causa di un picco di crescita della popolazione tedesca avuto dal 2016 in poi.
- A seguito di un **trend in crescita** registrato dalla **maggiore parte delle nazioni**, il valore rappresentanto da $Q_1$ e da $Q_2$ è **maggiore nel 2021**, mentre il valore di $Q_3$ è **diminuito**.
- In entrambi i casi la forma è **notevolemente asimmetrica**.

### 2.4.6 Conclusioni

Ciò che si nota sin da subito in tutti i boxplot è la **mediana estremamente bassa** rispetto all'altezza del boxplot. Infatti, la mediana, è quasi al livello di $Q_1$ nonostante rappresenti un punto centrale rispetto alla distribuzione dei valori. Questo accade

perchè, come già visto nell'analisi fattan nella Sezione 2.3, le **nazioni** hanno una **tendenza generale** verso una dimensione della **popolazione medio-piccola**, con circa il ~50% delle nazioni che ha una popolazione minore o uguale **10M/11M di abitanti**. Quindi è giusto che la mediana sia così vicina a $Q_1$, in quanto la metà dei valori è distribuita in un intervallo così piccolo (0-10M). D'altro canto, le restanti nazioni (il restante 50% dei valori), **hanno popolazioni con valori più dispersivi**, distribuiti in un intervallo molto più ampio rispetto la prima metà dei dati. Ciò rappresenta la causa del perchè il boxplot, dalla mediana in su, raggiunge altezze estremamente alte, creando una forma asimmetrica.

## 2.5 Diagramma di Pareto

Il diagramma di Pareto è uno strumento grafico utilizzato per identificare e visualizzare le principali cause di un problema o di un insieme di dati. E' molto utile in quanto permette di visualizzare, mostrando visivamente la frequenza relativa di ciascuna variabile, il sottoinsieme di variabili che influenzano in modo significativo i risultati finali di un determinato fenomeno. E' un diagramma a barre verticali con le modalità ordinate in ordine decrescente rispetto la loro frequenza relativa. Nel grafico è presente anche una linea cumulativa, che rappresenta la frequenza relativa cumulata.
Nel caso di questa analisi statistica il diagramma di pareto verrà utilizzato considerando le frequenze relative della Sezione 2.3 per visualizzare in **ordine significativo** le **fasce di popolazioni** che contribuiscono alla maggior parte della frequenza cumulativa. Queste sono le fasce che rappresentano un maggior numero di nazioni.

### 2.5.1 Diagramma di Pareto anno 2000

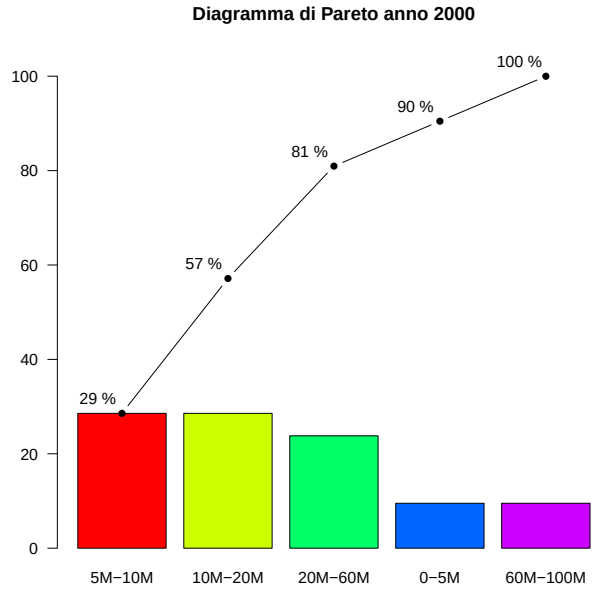


**Figura 2.40:** Diagramma di Pareto anno 2000

Osservando il diagramma di Pareto in Figura 2.40 si può dedurre che:

- Gli intervalli più significativi, contenendo l' **80%** delle nazioni sono: **5M-10M,10M-20M,20-60M**.
- **Non esiste un intervallo più significativo**, infatti, gli intervalli 5M-10M e 10M-20M hanno la medesima frequenza relativa, con un totale che rappresenta più del 50% delle nazioni totali.
- Gli intervalli **meno significativi**, contenendo il 20% delle nazioni, sono: 0-5M e 60M-100M,

### 2.5.2 Diagramma di Pareto dal 2001 al 2004

Durante il periodo che va dal 2001 al 2004 la distribuzione di frequenze è rimasta invariata, di conseguenza, anche i diagrammi di Pareto che fanno riferimento a tale periodo non sono variati, rimanendo uguali a quello presente in Figura 2.40.

### 2.5.3 Diagramma di Pareto anno 2005

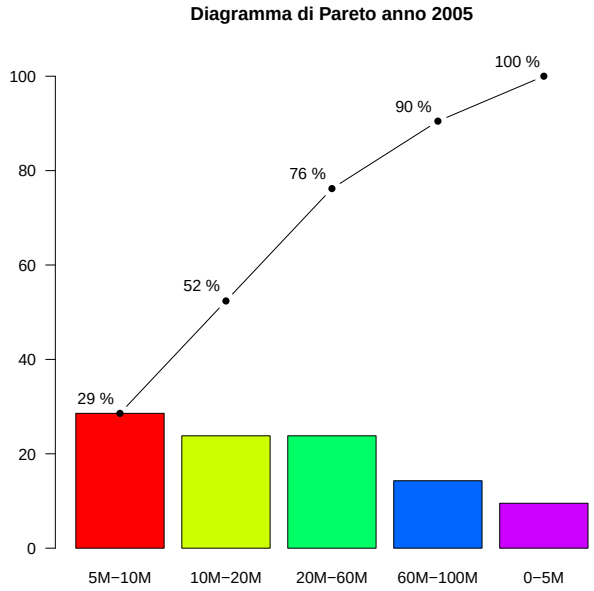


**Figura 2.41:** Diagramma di Pareto anno 2005

Osservando il diagramma di Pareto in Figura 2.41 si può dedurre rispetto gli anni precedenti che:

- I valori sono stati **più dispersivi** verso gli intervalli di minor significato (60M-100M e 0-5M), che nel 2005 incrementano la propria frequenza relativa totale (24%).
- Gli intervalli più significativi non sono variati, tuttavia, **la frequenza relativa cumulata è minore**.
- L'intervallo **più significativo** diventa **5M-10M**, contenente circa **1/3 delle nazioni totali**.

### 2.5.4 Diagramma di Pareto dal 2006 al 2010

Durante il periodo che va dal 2006 al 2010 la distribuzione di frequenze è rimasta invariata, di conseguenza, anche i diagrammi di Pareto che fanno riferimento a tale periodo non sono variati, rimanendo uguali a quello presente in Figura 2.41.

### 2.5.5 Diagramma di Pareto anno 2011

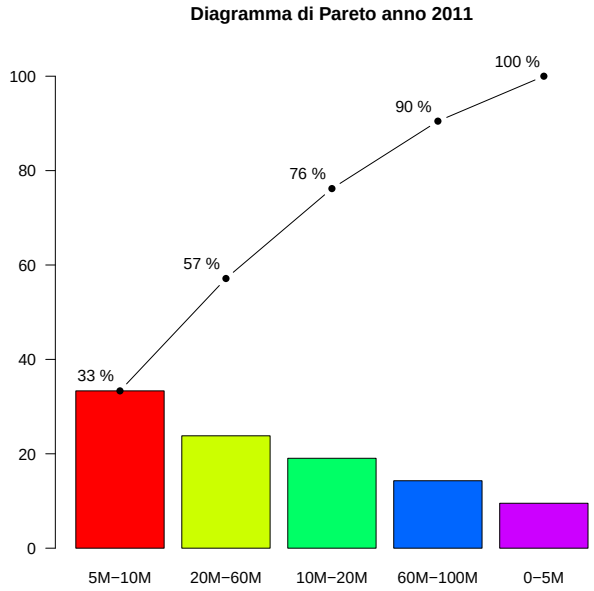


**Figura 2.42:** Diagramma di Pareto anno 2011

Osservando il diagramma di Pareto in Figura 2.42 si può dedurre rispetto gli anni precedenti che:

- Lintervallo **più significativo** 5M-10M ha **incrementato** la propria frequenza relativa, contenendo **più di 1/3 delle nazioni**.
- Gli intervalli più significativi e meno significativi non sono variati.

### 2.5.6 Diagramma di Pareto anno 2012

Nel 2012 la distribuzione di frequenze è rimasta invariata, di conseguenza, anche il diagramma di Pareto che fa riferimento a tale periodo non è variato, rimanendo uguale a quello presente in Figura 2.42.

### 2.5.7 Diagramma di Pareto anno 2013

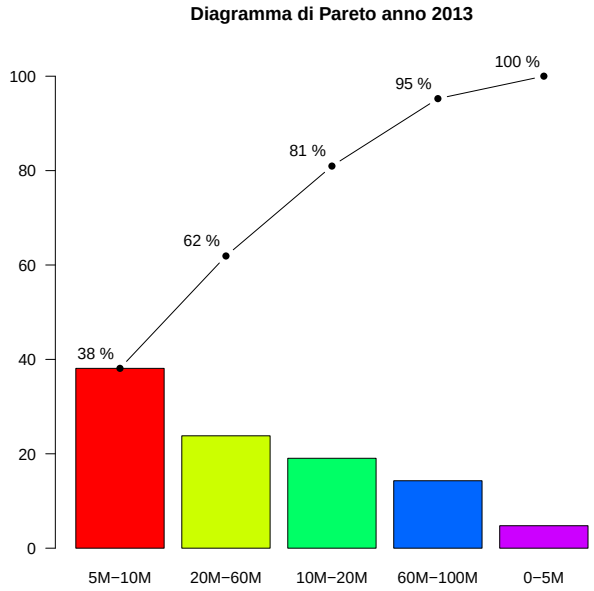


**Figura 2.43:** Diagramma di Pareto anno 2013

Osservando il diagramma di Pareto in Figura 2.43 si può dedurre rispetto gli anni precedenti che:

- I dati si sono **concentrati** nuovamente negli **intervalli più significativi**: 5-10M,20M-60M,10M-20M, che contengono più dell'80% delle nazioni totali.
- L'intervallo più significativo **5M-10M** ha nuovamente **incrementato** la propria frequenza relativa, contenendo quasi il **40% delle nazioni totali**.
- Gli intervalli 5M-10M e 20M-60M rappresentano più del 60% delle nazioni totali.

### 2.5.8 Diagramma di Pareto dal 2014 al 2017

Durante il periodo che va dal 2014 al 2017 la distribuzione di frequenze è rimasta invariata, di conseguenza, anche i diagrammi di Pareto che fanno riferimento a tale periodo non sono variati, rimanendo uguali a quello presente in Figura 2.43.

### 2.5.9 Diagramma di Pareto anno 2018

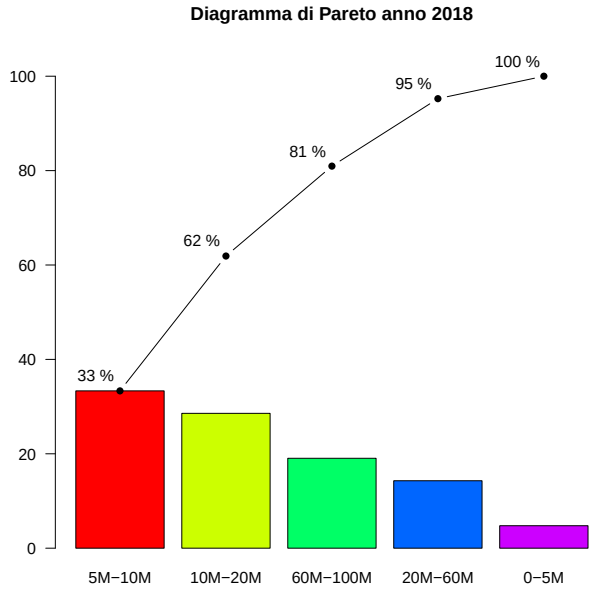


**Figura 2.44:** Diagramma di Pareto anno 2018

Osservando il diagramma di Pareto in Figura 2.44 si può dedurre rispetto gli anni precedenti che:

- I dati si **distribuiscono verso altri intervalli**, infatti, l'intervallo più significativo 5M-10M **decrementa** la propria frequenza relativa, ma rimane comunque l'intervallo maggiore.
- L'intervallo **60M-100M** diventa il terzo per frequenza relativa, **rappresentando uno dei tre intervalli significativi**.
- L'intervallo **20M-60M** a causa di un picco in discesa, rappresenta insieme l'intervallo **0-5M**, **gli intervalli meno significativi**.

### 2.5.10 Diagramma di Pareto anno 2019

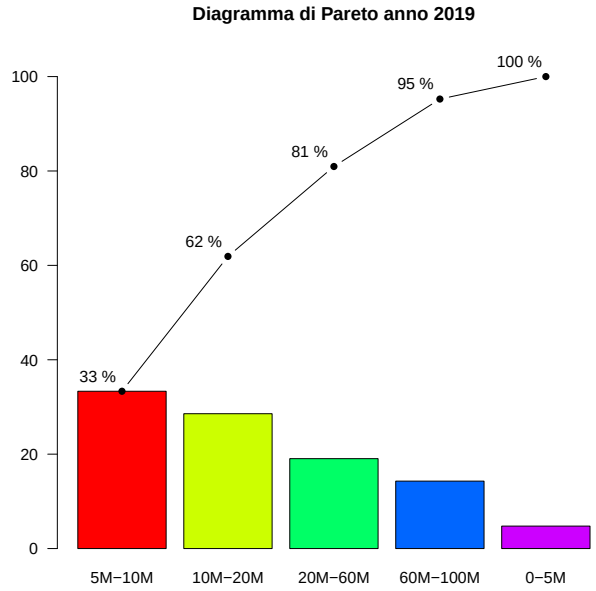


**Figura 2.45:** Diagramma di Pareto anno 2019

Osservando il diagramma di Pareto in Figura 2.45 si può dedurre rispetto gli anni precedenti che:

- A seguito di un incremento della frequenza relativa dell'intervallo **20M-60M**, che rientra nuovamente tra **i tre intervalli più significativi** che rappresentano all'incirca **l'80% delle nazioni totali.**
- L'intervallo **60M-100M**, di seguito a un decremento della propria frequenza relativa, **rappresenta insieme l'intervallo 0-5M, gli intervalli meno significativi.**

### 2.5.11 Diagramma di pareto anno 2020 e 2021

Durante gli anni 2020 e 2021 la distribuzione di frequenze è rimasta invariata, di conseguenza, anche i diagrammi di Pareto che fanno riferimento a tale periodo non sono variati, rimanendo come quello presente in Figura 2.45

### 2.5.12 Conclusioni

Osservando i diagrammi di Pareto dell'intero periodo a cui fa riferimento il dataset (2000-2021), si può dedurre che gli **intervalli più significativi**, cioè coloro

che contenevano all'incirca **l'80% dei dati**, erano **tre**. Tali intervalli si durante il corso degli anni sono cambiati, ma negli anni finali sono tornati nella stessa disposizione rispetto i primi anni. Stesso discorso per gli intervalli **meno significativi (0-5M e 60M-100M)**. L'intervallo **più significativo** durante l'intero periodo è stato **5M-10M**, che ha sempre rappresentato almeno il **30% delle nazioni**, con **picchi** che arrivavano quasi al **40%**. Considerando i risultati di tale analisi, si raggiungono gli **stessi risultati della Sezione 2.3**, cioè che le nazioni considerate hanno una **tendenza generale verso una dimensione della popolazione medio-piccola** in quanto gli intervalli con popolazione (media-bassa) 5M-10M e 10M-20M hanno rappresentato costantemente più del 50% delle nazioni totali.

CAPITOLO 3

# Statistica Descrittiva

## 3.1 Statistica descrittiva univariata

La statistica descrittiva **univariata** descrive la distribuzione di una **singola variabile** e include gli indici di posizione **centrali** (**media**, **mediana**, **moda**) e **non centrali** (**quantili**: quartili, decili, percentali) e gli indici di dispersione (**varianza**, **deviazione**, **standard**, **coefficiente di varianza**) che misurano quanto si disperdono i dati rispetto alla media). La **forma** della distribuzione invece viene descritta attraverso gli indici di **skewness** e **curtosi**.

## 3.2 Indici di sintesi

Gli **indici di sintesi**, detti anche "Statistiche", sono misure statistiche che sintetizzano o riassumono le informazioni contenute in un insieme di dati, fornendo una **visione compatta e rappresentativa** delle **caratteristiche principali** della distribuzione dei dati, senza dover considerare ogni singolo valore. Sono utilizzati solitamente per estrarre informazioni chiavi su un insieme di dati.

### 3.2.1 Media campionaria

Dato un insieme X composto da n valori numerici -> X = $\{x_1,x_2,...,x_n\}$, la media campionaria è la media aritmetica dei valori in X (somma di tutti i valori divisa per il numero totale di osservazioni. Rappresenta il valore medio di una distribuzione). Bisogna porre attenzione quando si analizza la media, poichè essa è **influenzata** da tutti i dati e in particolar modo da valori **particolarmente** grandi o piccoli rispetto il resto dei dati, cioè da **valori anomali**, che possono **variare significatamente** il suo valore. Questo è uno dei motivi per il quale la media potrebbe non essere la misura di tendenza centrale **più robusta** in presenza di **dati anomali**. Di seguito verrà riportato la media campionaria per ogni anno, che rappresenterà il valore medio di popolazione avuta in quell'anno: Analizzando la Tabella 3.1 si osserva

**Tabella 3.1:** Media Campionaria 2000-2021

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| Media | 23.222.771 | 23.286.926 | 23.336.102 | 23.432.570 | 23.534.090 | 23.645.370 | 23.747.148 |

| | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|---|---|---|
| Media | 23.852.212 | 23.965.510 | 24.073.073 | 24.145.882 | 24.160.404 | 24.237.430 | 24.316.270 |

| | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|
| Media | 24.439.544 | 24.522.162 | 24.623.188 | 24.703.319 | 24.775.706 | 24.820.208 | 24.895.281 | 24.905.040 |

un andamento **costantemente crescente** della media campionaria, non **registrando nessun anno in decrescita rispetto quello precedente**. Ciò indica che la **popolazione media delle nazioni** considerate è **aumentata** anno dopo anno nel periodo preso in considerazione, indicando una **tendenza generale di crescita demografica**. Infatti, il valore della media campionaria del 2021 è cresciuta del ~7% rispetto quella del 2000.

### 3.2.2 Mediana Campionaria

Dato un insieme di dati X = $\{x_1,x_2,...,x_n\}$ e siano $x_1<x_2<...<x_n$ ordinati in ordine crescente, si definisce mediana campionaria il valore che bipartisce i dati in due gruppi di uguale numerosità, in maniera tale che lo stesso numero di dati cada sia a sinistra che a destra della mediana stessa. Per calcolare la mediana bisogna:

- Ordinare gli n elementi in **ordine crescente**.
- Se **n è dispari**, allora la mediana corrisponde al valore in posizione $\frac{n+1}{2}$.
- Altrimenti, se **n è pari**, la mediana rappresenta la media aritmetica dei valori in posizione $\frac{n}{2}$ e $\frac{n}{2}+1$.

La mediana campionaria rispetto la media **è più robusta** in quanto **non risente dei valori anomali**, visto che dipende solo da uno o due valori centrali. Analizzando

**Tabella 3.2:** Mediana Campionaria 2000-2021

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| Mediana | 10.249.000 | 10.330.800 | 10.394.700 | 10.444.600 | 10.473.100 | 10.494.700 | 10.512.000 |

| | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|---|---|---|
| Mediana | 10.584.500 | 10.666.900 | 10.753.100 | 10.839.900 | 11.000.600 | 11.075.900 | 11.003.600 |

| | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|
| Mediana | 10.926.800 | 10.858.000 | 10.783.700 | 10.768.200 | 10.741.200 | 10.724.600 | 10.718.600 | 10.678.600 |

la Tabella 3.2 si osserva un **andamento costantemente crescente** della mediana campionaria dal 2000 al 2012. Dal 2013 al 2021 la mediana ha avuto un **andamento in costante decrescita**. Tuttavia, la mediana campionaria del 2021 è **maggiore** rispetto quella del 2000, con un aumento del **~4%.**

Un ulteriore analisi utile da fare è il confronto tra media e mediana campionaria, in quanto può offrire caratteristiche sulla distribuzione dei dati, sulla loro forma e valutare quanto la presenza di valori estremi influenzi le misure di tendenza centrale.

**Tabella 3.3:** Confronto tra media e mediana Campionaria 2000-2021

| | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|---|
| Media | 23.222.771 | 23.286.926 | 23.336.102 | 23.432.570 | 23.534.090 | 23.645.370 | 23.747.148 |
| Mediana | 10.249.000 | 10.330.800 | 10.394.700 | 10.444.600 | 10.473.100 | 10.494.700 | 10.512.000 |

| | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 |
|---|---|---|---|---|---|---|---|
| Media | 23.852.212 | 23.965.510 | 24.073.073 | 24.145.882 | 24.160.404 | 24.237.430 | 24.316.270 |
| Mediana | 10.584.500 | 10.666.900 | 10.753.100 | 10.839.900 | 11.000.600 | 11.075.900 | 11.003.600 |

| | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|
| Media | 24.439.544 | 24.522.162 | 24.623.188 | 24.703.319 | 24.775.706 | 24.820.208 | 24.895.281 | 24.905.040 |
| Mediana | 10.926.800 | 10.858.000 | 10.783.700 | 10.768.200 | 10.741.200 | 10.724.600 | 10.718.600 | 10.678.600 |

Analizzando la Tabella 3.3 si nota immediatamente la **significativa differenza tra media e mediana campionaria**, presente **costantemente** durante **l'intero periodo** preso in considerazione. Ciò può indicare la presenza di una distribuzione dei dati **fortemente asimmetrica**, con una **coda lunga verso i valori superiori** (stessa conclusione fatta nella Sezione 2.4.6). Ecco alcune motivazioni:

- A causa di **valori estrememamenti alti (outlier)**, il valore della media campionaria (che è sensibile ai dati a differenza della mediana) è **stato notevolmente influenzato**. Infatti, considerando come esempio l'anno 2000, **eliminando i 3 valori maggiori**, la **media** varia da **23.222.771** a **15.899.139**.
- La maggior parte delle nazioni ha una popolazione **media-bassa**, tuttavia, c'è un **limitato numero di nazioni** con popolazione **estremamente alta**, che oltre ad **influenzare la media**, creano **un'assimetria dei dati con una coda lunga verso i valori superiori**.

### 3.2.3 Moda campionaria

La moda identifica la modalità che si presenta con la maggiore frequenza (assoluta o relativa) in un campione. Se ci sono più modalità con frequenza massima, ciascuna viene definita come **valore modale**. Quando la moda è unica la distribuzione è detta **uni-modale**, in caso contrario è detta multi-modale o plurimodale. Nel caso di qust'analisi statistica il calcolo della moda **non fornisce alcun informazione utile** in quanto è **improbabile** avere più occorrenze dello stesso valore.

### 3.2.4 Varianza, deviazione standard e coefficiente di variazione

Gli indici di posizione sono importanti, ma le informazioni che ci danno **non considerano la variabilità dei dati**. Infatti, anche se la media e mediana campionaria coincidono, la dispersione dei dati può essere **differente**. Gli indici più significativi per misurare la variabilità dei dati sono: **Varianza, deviazione standard e coefficiente di variazione.** Tuttavia, nel nostro caso, la media e mediana campionaria erano significatavente discostate tra di loro in ogni anno analizzato, di conseguenza **si ha già conoscenza dell'assimetria dei dati e della loro variazione**.

**Tabella 3.4:** Indici di dispersione 2000,2010 e 2021

| | 2000 | 2010 | 2021 |
|---|---|---|---|
| Media | 23.222.771 | 24.145.882 | 24.905.040 |
| Mediana | 10.249.000 | 10.839.900 | 10.678.600 |
| Deviazione Standard | 23.231.451 | 24.130.702 | 24.892.946 |
| Coefficiente di variazione | 1.0003 | 0.9993 | 0.9995 |

Infatti, anche gli indici di dispersione calcolati nella Tabella 3.4 **suggeriscono** che la **distribuzione dei dati** è caratterizzata da una **notevole variabilità rispetto alla media**. Nella tabella non è stata considerata la **varianza** poichè è molto **influenzata dalla scala dei dati**, di conseguenza, è stata considerata la **deviazione standard** che rappresenta un indice più **robusto** e **interpretabile**.

### 3.2.5 Forma della distribuzione di frequenza

Gli indici trattati fino ad ora ci hanno già suggerito un'asimmetria della distribuzione dei dati considerati. Tuttatia, esistono degli indici che permettono di **misurare** precisamente la **simmetria** della funzione di distribuzione e la piccatezza: **skewness** e **curtosi campionaria.**

### 3.2.6 Skewness

La skewness misura l'asimmetria della distribuzione dei dati rispetto alla media.

- Se la skewness è **zero**, la distribuzione è **simmetrica** rispetto alla media.
- Se la skewness è **positiva**, la coda della distribuzione è più **lunga verso destra**.
- Se la skewness è **negativa**, la coda della distribuzione è più **lunga verso sinistra**.

**Tabella 3.5:** Skewness anno 2000,2010 e 2021

| | 2000 | 2010 | 2021 |
|---|---|---|---|
| Skewness | 1.229 | 1.155 | 1.148 |

La skewness in questo caso risulta sempre essere **positiva**, in Tabella 3.5 è possibile vedere degli anni presi in esempio. Ciò indica che la distribuzione dei dati **non è mai simmetrica** durante il periodo che va dal **2000** al **2021**, con una **forte assimetria verso destra** e quindi, con una **coda allungata verso destra** in quanto i valori della skewness è sempre positiva.

### 3.2.7 Curtosi

La curtosi misura la densità dei dati intorno alla media confrontando la distribuzione che stiamo considerando con una distribuzione normale. In pratica la curtosi misura quanto le **code di una distribuzione differiscono da quelle di una distribuzione normale**. Una distribuzione normale è caratterizzata da $\beta_2 = 3$ e indice di curtosi $\gamma_2 = 0$. Se risulta:

- $\beta_2 < 3$, $\gamma_2 < 0$: la distribuzione di frequenza si definisce **platicurtica**, ossia la distribuzione di frequenza è **più piatta** di una normale.
- $\beta_2 > 3$, $\gamma_2 > 0$: la distribuzione di frequenza si definisce **leptocurtica**, ossia la distribuzione di frequenza è **più piccata** di una normale.
- $\beta_2 = 3$, $\gamma_2 = 0$: la distribuzione di frequenza si definisce **normocurtica**, ossia segue la **curva di una normale**.

**Tabella 3.6:** Curtosi anno 2000,2008 e 2021

| | 2000 | 2008 | 2021 |
|---|---|---|---|
| Curtosi | 0.2117964 | -0.04357013 | -0.1516102 |

Considerando il dataset preso in analisi, dal **2000 al 2007** il valore della curtosi è **positivo**, indicando che tali distribuzioni di dati hanno **code più appuntite** rispetto a una distribuzione normale. In altre parole, ci sono valori che si estendono **più lontano dalla media** rispetto a quanto ci si aspetterebbe in una distribuzione normale. Com'è possibile osservare anche in Tabella 3.6, **dal 2008** in poi la **curtosi diventa negativa**, indicando che la distribuzione diventa via via più **"piatta"** rispetto a una distribuzione normale, con **code meno pesanti**. La **diminuzione costante** dal **2008 al 2021** della curtosi indica che la distribuzione sta diventando **meno estrema** e **si sta avvicinando a una distribuzione più normale**.

## 3.3 Statistica Descrittiva bivariata

In questa sezione si tratterà la **statistica bivariata**, ossia il ramo della statistica che si occupa dei metodi grafici e statistici atti a descrivere le relazioni che intercorrono tra due variabili quantitative. L'analisi verrà effettuata considerando gli anni estremi del dataset considerato: **2000 e 2021**. Verrà utilizzato il **coefficiente di correlazione campionario**, piuttosto che la covarianza campionaria, per calcolare le possibile relazioni tra le due distribuzioni di dati in quanto il coefficiente di correlazione è una **misura normalizzata**.

### 3.3.1 Coefficiente di correlazione campionario

Il coefficiente di correlazione campionario è una metrica quantitativa che misura quanto è **forte il legame di natura lineare tra le variabili considerate**. Il coefficiente ci indica se e come i punti sono posizionati attorno ad una retta interpolante, o se c'è una retta che allinea tutti i punti. I valori che può assumere sono limitati nell'intervallo **[-1,1]**.

- Se il coefficiente è **> 0**, indica una **correlazione positiva**, cioè quando una variabile aumenta, l'altra aumenta proporzionalmente.
- Se il coefficiente è **< 0**, indica una **correlazione negativa**, cioè quando una variabile aumenta, l'altra diminuisce proporzionalmente.
- Se il coefficiente è **= 0**, indica **l'assenza** di una correlazione lineare.

```
## coefficiente di correlazione campionario tra il 2000 e 2021
## [1] 0.9942846
```

Il valore ottenuto dal coefficiente di correlazione campionario tra l'anno 2000 e 2021 è **0.9942846**, molto vicino ad 1, che indica una **correlazione positiva perfetta**. In questo caso quindi, è presente una **forte relazione positiva lineare tra le due variabili**, ciò significa che quando una variabile aumenta, l'altra aumenta proporzionalmente (segue una tendenza simile). Questa relazione può essere ben rappresentata da una retta di regressione lineare che approssima l'allineamento lineare quasi perfetto dei dati.

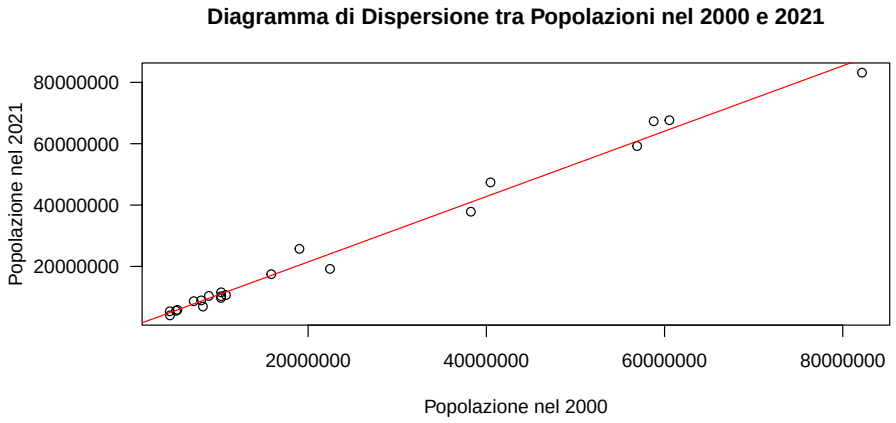


**Figura 3.1:** Diagramma di dispersione tra popolazioni nel 2000 e 2021

In Figura 3.1 è posibile osservare lo scatterplot applicato sulle due distribuzioni con la retta di regressione visibile in rosso. Avendo una correlazione quasi perfetta, **la maggior parte dei punti si trova sulla retta di regressione**. Con una correlazione perfetta (coefficiente = 1) **tutti** i punti giacerebbero sulla retta di regressione. In qusto caso, la retta di regressione è **ascendente** poichè il valore del coefficiente è **positivo**.

### 3.3.2 Regressione lineare semplice

Nella regressione lineare le relazioni sono modellate usando funzioni di predizione lineare i cui parametri del modello sono stimati dai dati. Questi modelli sono detti **modelli lineari**.
Il modello di regressione lineare semplice è esprimibile attraverso l'equazione di una retta che riesce ad interpolare la nuvola di punti dello scatterplot meglio di tutte e altre possibili rette. Il modello lineare viene di solito utilizzato per spiegare, descrivere, o anche prevedere un andamento futuro sulla base della relazione che si instaura tra una variabile Y, chiamata variabile dipendente, e una variabile indipendente X.
Data l'equazione:

$$Y = \alpha + \beta X$$

Dove: $\alpha$ è **l'intercetta**, e $\beta$ è il **coefficiente angolare**, che indica la pendenza della retta. Una volta individuati i valori di $\alpha$ e $\beta$ basterà **sostituire** nell'equazione la variabile X per conoscere il **valore predetto di Y**.

```
## Valori di alpha e beta
## [1] 112875.4152142      0.9279204
```

Considerando la variabile X come la distribuzione di dati del 2000, e la variabile Y, come la distribuzione di dati 2021, si ha che $\alpha = 112875.4152142$ e $\beta = 0.9279204$, con equazione:

$$Y = 112875.4152142 + 0.9279204 \cdot X$$

In **Figura 3.4** è possibile osservare la **retta di regressione ascendente** che riesce ad **interpolare** la nuvola di punti meglio di un'altra qualsiasi retta.

### 3.3.3 Residui

Dopo aver trovato la retta di regressione, è possibile osservare qual è il **discostamento tra i valori osservati e i valori stimati**, questo perchè i valori stimati dalla retta di regressione **non sono sempre coerenti con quelli reali**, creando dei **discostamenti**. I **residui** mostrano di quanto si discostano i valori osservati dai valori stimati con la retta di regressione.

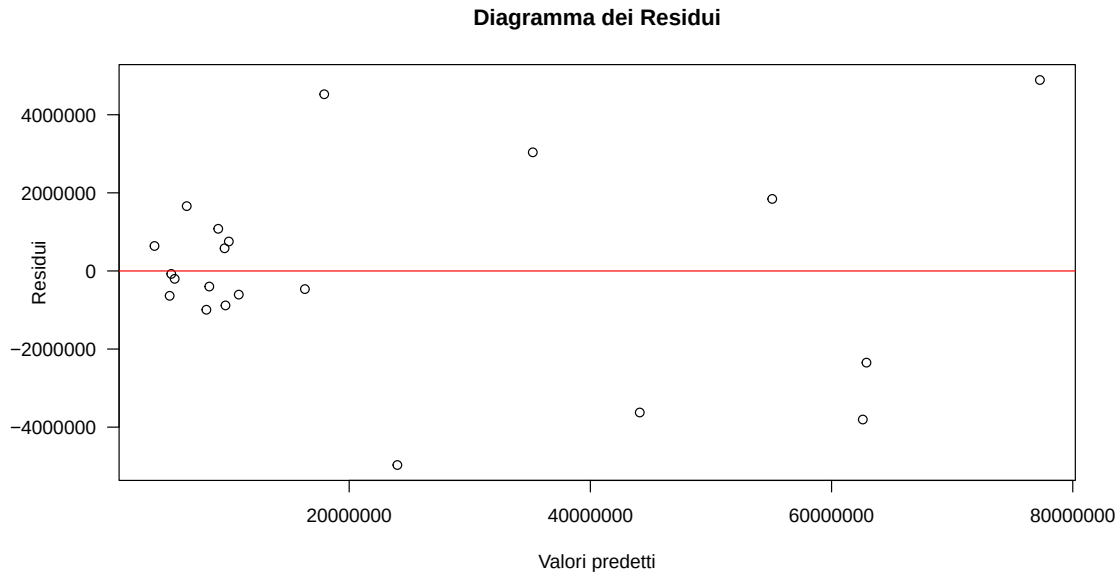


**Figura 3.2:** Diagramma dei residui

In Figura 3.2 è possibile osservare i residui dei valori predetti dalla retta di regressione. La **retta in rosso** indica un residuo uguale a **0** ed indica un **valore**

**perfettamente predetto**. Più i residui si **avvicinano a questa retta**, più il valore predetto **si avvicina a quello reale**. Il **residuo minore ottenuto**, cioè, ll valore predetto che più si è avvicinato a quello reale è quello della **Finlandia**, dove la popolazione reale nel 2021 era di **5.533.790**, mentre il valore predetto dal modello è stato di **5.247.792**, con una differenza del **~5%**. Invece, il **residuo maggiore**, e quindi, il valore predetto che più si è discostato da quello reale è quello degli **UK**, dove la popolazione reale nel 2021 era di **25.738.100**, mentre il valore predetto dal modello è stato di **23.995.784**, con una differenza del **~7%**

CAPITOLO 4

# Analisi dei cluster

L'analisi dei cluster è una metodologia che permette di raggruppare in sottoinsieme (cluster) singoli entità appartenenti ad un insieme più ampio. I raggruppamenti vengono effettuati considerando la **somiglianza**, in modo che elementi di uno stesso cluster siano il **più simile tra loro**, mentre, elementi di cluster differenti **siano diversi il più possibile tra di loro**.
Per determinare quanto due elementi siano simili tra di loro si possono utilizzare delle metriche come i **coefficienti di similatirà** o le **misure di distanza**. I primi hanno la caratteristica di assumere valori tra **0 e 1**, mentre le distanza possono assumete qualunque valore **maggiore o uguale a 0**. Esistono due tipologie di metodi di clustering: **gerarchici e non gerarchici**.

## 4.1 Clustering gerarchico

Il clustering gerarchico è un metodo di clustering che non comporta nè la **scelta a priori del numero di cluster** nè la **scelta di parametri per la determinazione automatica del loro numero**. Fornisce una visione completa dell'insieme in termini di distanza o similarità, ma ha lo **svantaggio** di **non poter riallocare** gli individui che sono stati già classificati ad un livello precedente dell'analisi. L'obiettivo finale dei

metodi gerarchici è quello di ottenere una sequenza di partizioni che possono essere rappresentate graficamente mediante una struttura ad albero: **Dendrogramma**, dove sulle ordinate sono riportati i livelli di distanza, mentre sulle ascissse sono riportati i singoli individui. Il dendrogramma fornisce un quadro completo della struttura dell'insieme in termini delle misure di distanza tra gli individui.

### 4.1.1 Scelta delle metriche e dendogramma

Il dataset preso in considerazione, come già specificato in precedenza, presenta una nazione con una popolazione estremamente alta: la Germania. Tale nazione rappresenta un outlier all'interno della distribuzione dei dati, ed, insieme ad altre nazioni con valori alti di popolazione, rendono la distribuzione dei dati fortemente asimmetrica. Di conseguenza, è stata utilizzata la **metrica di Canberra** per ottenere le **misure di distanza**, in quanto essa risulta essere **meno sensibile all'asimmetria della distribuzioni di dati** e alla presenza di eventuali **valori anomali**. Inoltre, come metodo gerarchico agglomerativo è stato utilizzato il **metodo del legame completo**, poichè rappresenta un metodo standard.

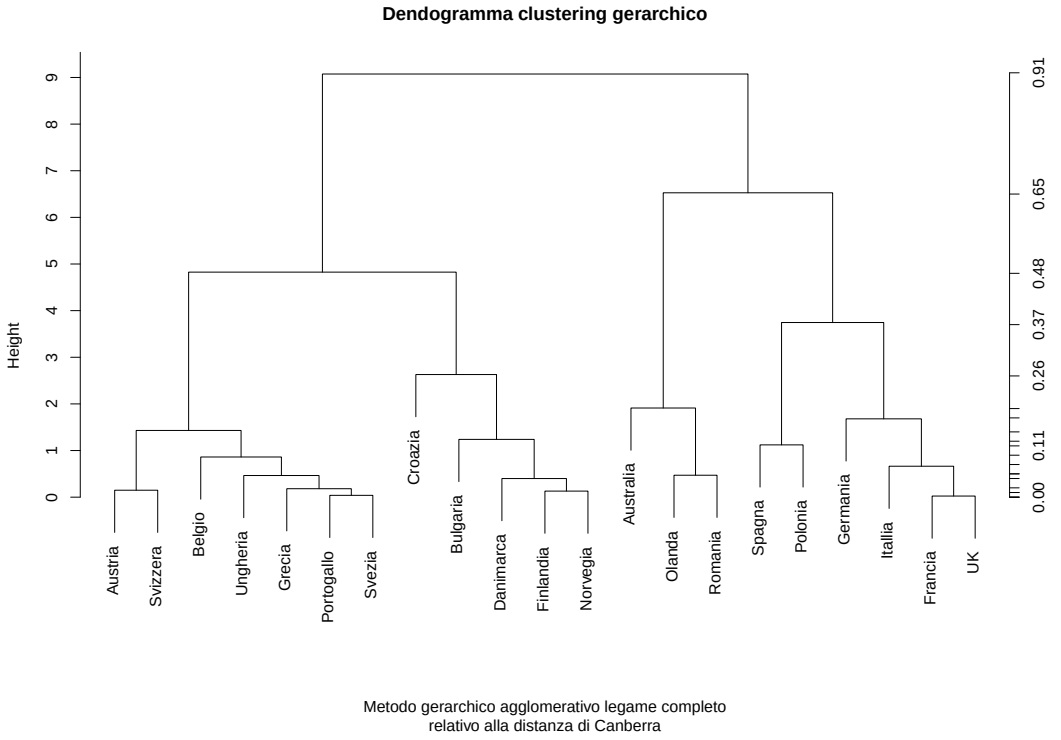


**Figura 4.1:** Dendogramma cluster gerarchico

### 4.1.2 Screeplot

Adesso, dopo aver definito il Dendogramma in Figura 4.1, bisogna determinare il numero **ottimale di cluster**. Per farlo, viene utilizzata una **metodologia euristica** che consiste nel creare un grafico chiamato **screeplot**, dove sull'asse delle ordinate sono posti i numeri di gruppi ottenibili con il clustering gerarchico, mentre sull'asse delle ascisse vengono poste le distanze a cui avvengono le successive aggregazioni tra i gruppi. Se nel passaggio da $k$ gruppi a $k-1$ gruppi si registra un forte incremento della distanza di aggregazione è **consigliabile** tagliare il dendrogramma in $k$ gruppi.

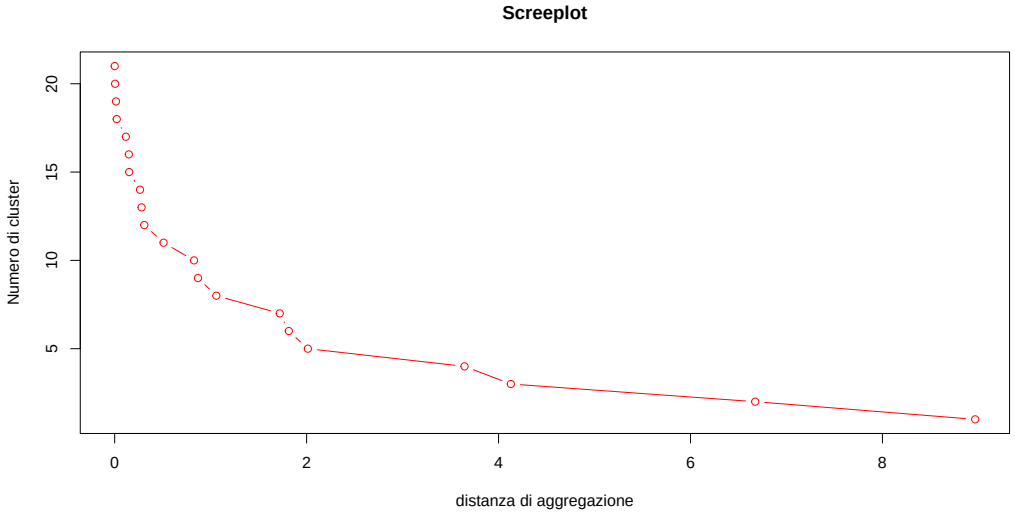


**Figura 4.2:** Screeplot cluster gerarchico

Osservando lo screeplot visibile In Figura 4.2, si deduce che il numero di cluster consigliato è **tre**, in quanto si registra il **massimo incremento** della distanza di aggregazione.

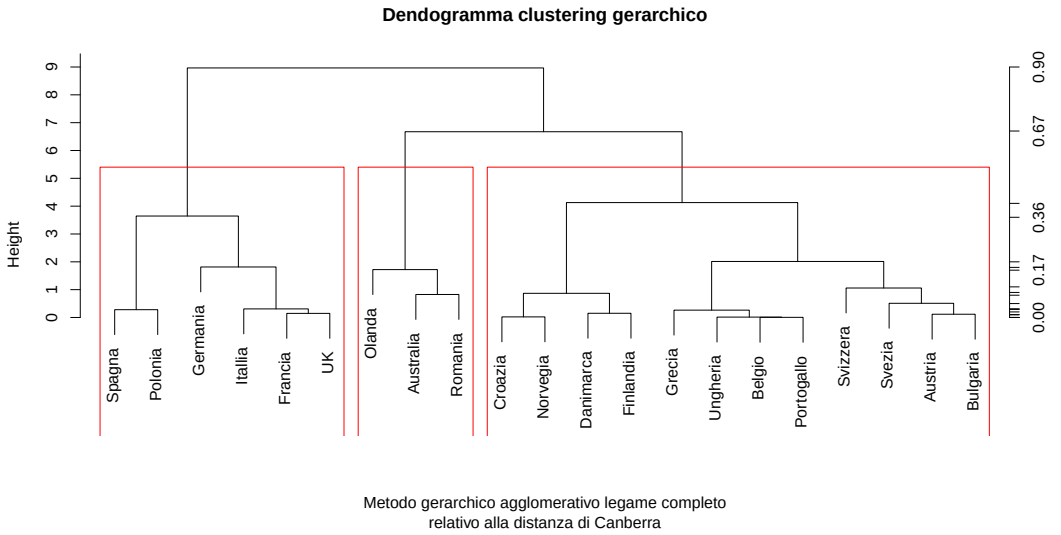


**Figura 4.3:** Dendogramma con i tre cluster

In Figura 4.3 è possibile osservare il dendogramma con i **3 cluster** definiti e racchiusi tra i rettangoli rossi. Le nazioni che si trovano all'interno dello stesso rettangolo rosso fanno parte dello stesso cluster.

### 4.1.3 Misure di non omogeneità

Una volta scelto il numero di cluster da considerare bisogna determinare se questo valore rappresenta un **numero ottimale di cluster**. L'obiettivo è quello di trovare un numero di cluster che vada a **minimizzare** la misura di non omogeneità statica all'interno dei cluster (**within**) e **massimizzare** la misura di non omogeneità statica tra i gruppi (**between**). Di seguito verranno elencati i risultati ottenuti considerando **tre** cluster:

- La misura di non omogeneità statistica totale è: **12393175407239696.**
- La misura di non omogeneità statistica all'interno dei tre gruppi (**within**) è: **1405062235050558.**
- La misura di non omogeneità tra i cluster (**between**) è: **10988113172189138.**

```
## Misura di non omogeneità statistica tra i tre cluster
## [1] 0.8866261
```

Il rapporto tra between e la misura di non omogeneità totale ha restituito il valore **0.8866261**, un valore **ottimo**, il che indica una **buona separazione** tra i cluster e supporta in positivo il numero di cluster definito dallo screeplot.

## 4.2 Clustering non gerarchico

I metodi di clustering non gerarchico hanno l'obiettivo di ottenere un'unica partizione degli $n$ individui di partenza in cluster. A differenza dei metodi gerarchici, i metodi non gerarchici prevedono la **scelta a priori del numero di cluster** e dei **parametri per la determinazione automatica del loro numero**. Tuttavia, presentano il vantaggio di **poter riallocare** gli individui che sono stati già classificati in un livello precedente dell'analisi.

### 4.2.1 Metodo del K-means

Il metodo del **k-means** prevede inizialmente un **assegnazione casuale** dei $k$ centroidi, assegnando i restanti individui al cluster del **centroide più vicino**. Una volta formati i $k$ cluster con i $k$ centroidi casuali, vengono **ricalcoli** i centroidi per ogni gruppo. Vengono poi considerate le distanze di ogni indiviuo; se la distanza minima non è ottenuta in corrispondenza del centroide del cluster di appartenenza, l'individuo **viene spostato** presso il cluster che ha il centroide più vicino, modificando i gruppi. Una volta ripartizionati i cluster, vengono ricalcolati nuovamente i $k$ centroidi dei nuovi gruppi per poi andare a spostare gli individui in caso di un centroide più vicino. Questo procedimento continua finchè non si raggiunge una **configurazione stabile**, ossia, **non avvengono spostamenti** di individui all'interno dei gruppi (**e quindi i centroidi non cambiano durante un'intera iterazione**).

In quest'analisi statistica verrà scelto $\mathbf{k = 3}$, in quanto è il **valore suggerito** dallo screeplot e che ha portato a dei risultati ottimi nei metodi gerarchici. Inoltre, utilizzando lo stesso numero di cluster, si potrà effettuare un confronto tra le due metodologie, verificando chi ha creato dei cluster migliori considerando le misure di non omogeneità. Inoltre, il metodo del k-means offre tre scelte iniziali:

1. Scelta casuale dei punti di riferimento.
2. Ripetizione della procedura di scelta casuale dei punti di riferimento.
3. Scelta dei centroidi come punto di riferimento.

L'analisi procederà considerando la **terza scelta**.

**Risultati ottenuti**

Il metodo del k-means ha restituito i seguenti cluster:

| Australia | Austria | Belgio | Bulgaria | Svizzera | Germania | Danimarca | Spagna | Finlandia | Francia |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 3 | 3 | 3 | 2 | 3 | 2 | 3 | 2 |
| UK | Grecia | Croazia | Ungheria | Itallia | Olanda | Norvegia | Polonia | Portogallo | Romania |
| 2 | 3 | 3 | 3 | 2 | 1 | 3 | 1 | 3 | 1 |
| Svezia | | | | | | | | | |
| 3 | | | | | | | | | |

Producendo i seguenti risultati:

- La misura di non omogeneità totale è: **12393175407239694**.
- La misura di non omogeneità interna ai cluster (**within**) è: **1010003468051892**.
- La misura di non omogeneità tra i cluster (**between**) è: **11383171939187802**.

Quindi:

```
## Misura di non omogeneità statistica tra i tre cluster
## [1] 0.9185033
```

Il rapporto tra between e la misura di non omogeneità totale ha restituito il valore **0.9185033**, un valore **ottimo**, che indica una separazione **quasi perfetta** dei cluster e che continua a supportare il valore $k = 3$ consigliato dallo screeplot in precedenza.

**Conclusioni**

Come precedentemente specificato i cluster dovrebbero essere individuati in modo da minimizzare la misura di non omogeneità statistica all'interno dei cluster (within) e massimizzare la misura di non omogeneità statistica tra i gruppi (between). Sono state applicate entrambe le metodologie di clustering: gerarchico e non gerarchico e proprio **quest'ultima metodologia è risultata essere la migliore**. Infatti confrontando i risultati tra la metodologia gerarchia e non gerarchica (**in grassetto**) si ha che:

- Il valore di within è minore: **1010003468051892** < 1405062235050558
- Il valore di between è maggiore: **11383171939187802** > 10988113172189138

- La misura di non omogeneità statistica tra i tre cluster è maggiore: **0.9185033** > 0.8866261

Considerando i risultati appena descritti, si ha che il metodo non gerarchico **minimizza** la misura di non omogeneità statistica all'interno dei cluster (**within**) e **massimizza** la misura di non omogeneità statistica tra i gruppi (**between**) rispetto quello gerarchico. In Figura 4.4 è possibile osservare come il metodo non gerarchico del K-means ha generato i $k = 3$ cluster che hanno portato ai risultati appena descritti.

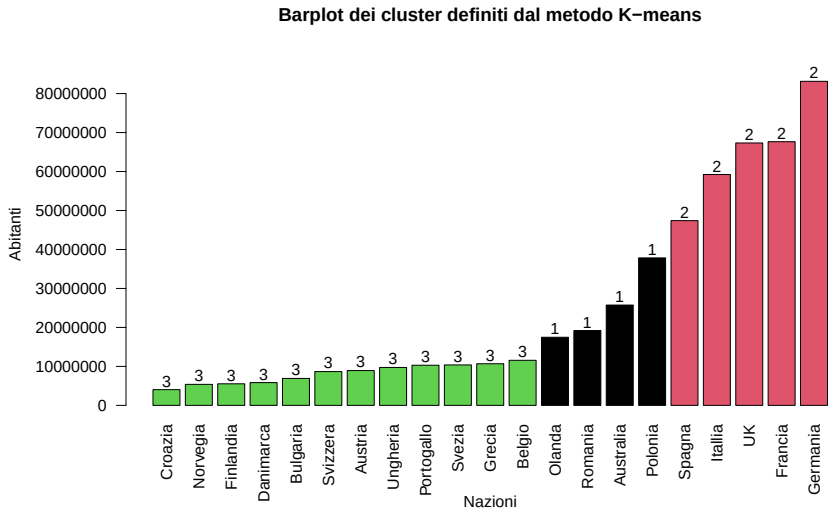


**Figura 4.4:** Barplot dei cluster deifniti dal metodo k-means

In Figura 4.4 viene mostrato anche l'intervallo di popolazione che ricopre ogni cluster:

- Il **cluster 3** è composto dalle nazioni che hanno una popolazione compresa tra i **~4 Milioni** e **~11.5 Milioni.**
- Il **cluster 1** è composto dalle nazioni che hanno una popolazione compresa tra i **~17.5 Milioni** e **~38 Milioni**.
- Il **cluster 2** è composto dalle nazioni che hanno una popolazione compresa tra i **~47 Milioni** e **~83 Milioni.**

Osservando questi dati è importante considerare come la clusterizzazione tende a **raggruppare** i paesi con un numero di **popolazione simile senza commettere errori**.

CAPITOLO 5

# Inferenza statistica

**L'inferenza statistica** ha lo scopo di derivare le caratterisitche di una popolazione tramite un campione estratto da essa. Cioè, si studia una popolazione descritta da una variabile aleatoria X la cui funzione di distribuzione ha una forma nota ma contiene un **parametro non noto**. Per ottenere informazioni su questo parametro non noto della popolazione, si può fare uso dell'inferenza statistica considerando un campione estratto dalla popolazione e effettuando su tale campione delle opportune misure.

## 5.1 Variabili aleatorie

Una **variabile aleatoria X** è una **funzione** definita sullo spazio campionario $\Omega$ che **associa** ogni evento $E \subset \Omega$ un **unico valore reale**. Una variabile aleatoria può essere classificata come:

- **Discreta**: può assumere un insieme discreto (finito o numerabile) di numeri reali.
- **Continuo**: può assumere tutti i valori compresi in un intervallo reale.

### 5.1.1 Scelta della variabile aleatoria

La scelta della variabile aleatorie e della distribuzione da considerare dev'essere fatta in base agli obiettivi che si intendono raggiungere. Nel caso di quest'analisi statistica, **L'obiettivo** sarà quello di calcolare con che **probabilità** delle specifiche nazioni **raggiungeranno** uno specifico **intervallo di abitanti** nel **2022**, che rappresenta una **variabile non nota** all'interno del nostro dataset, in quanto, si ricorda, che il dataset copre il periodo che va dal 2000 al 2021. Per calcolare l'intervallo di valori stimato verrà considerato l'intero campione (popolazione dal 2000 al 2021) della nazione di riferimento.
La variabile aleatoria facendo riferimento alla popolazione di una nazione, può assumere un qualsiasi valore $> \mathbf{0}$, cioè, non esiste un insieme finito di valori che può assumere. Di conseguenza, si farà riferimento a una variabile aleatoria di **tipo continua**.

### 5.1.2 Distribuzione normale

I dati di ogni campione ricoprono un periodo di 22 anni (2000-2021), di conseguenza, in un periodo così ampio è probabile che siano presenti delle tendenze (positivi e/o negativi) che si sono seguiti negli anni e/o che sono tutt'ora presenti, dei picchi, fluttuazioni etc... Quindi, Per ottenere una stima che **tenga conto di tali dinamiche** e approssimi un possibile intervallo di valori **reali** della popolazione, è stata scelta la **Distribuzione normale**.
La distribuzione normale, anche conosciuta come distribuzione gaussiana è completamente descritta da due parametri: la **media ($\mu$)** e la **deviazione standard ($\sigma$)**.
La funzione di densità di probabilità di una variabile casuale distribuita normalmente è data da:

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

dove:

- $x$ è il **valore** (o i valori assunti) dalla **variabile aleatoria**.
- $\mu$ è la **media della distribuzione normale**, che indica il **centro della distribuzione.**

- $\sigma$ è la **deviazione standard**, che rappresenta la **dispersione dei dati intorno alla media.**

### 5.1.3 Probabilità di avere un intervallo di popolazione nel 2022

Come già specificato in precedenza, verrà calcolato con che **probabilità delle specifiche nazioni raggiungeranno uno specifico intervallo di abitanti nel 2022**. Le nazioni considerate saranno un sottoinsieme di quelle analizzate singolarmente nella Sezione 2.1 (e che quindi presentano tendenze e/o forti picchi). La probabilità verrà calcolata considerando gli **incrementi e decrementi** registrati annualmente dal **2000 al 2021** dalla nazione di riferimento. Tramite l'uso di questi dati, utilizzando la **distribuzione normale**, verrà calcolata con che probabilità può avvenire un determinato **incremento o decremento**, e quindi, basandosi su questo, con che probabilità la nazione avrà un determinato intervallo di popolazione nel 2022.
Per ogni nazione, oltre che definire la probabilità, verranno generati anche **X valori casuali** in linea con la media e la deviazione standard della distribuzione dei dati che **fa riferimento agli incrementi e decrementi annuali della nazione** (in modo da **simulare** un possibile **valore reale** di incremento/decremento). Questa generazione verrà poi visualizzata graficamente tramite un **density plot** per verificare se la **probabilità è coerente con i valori generati in linea alla distribuzione dei dati.**

#### Australia

Osservando la serie temporale relativa all'Australia (Figura 2.2) sappiamo che tale nazione **non ha mai registrato un decremento rispetto l'anno precedente**, infatti, la crescita della sua popolazione è stata costante durante l'intero periodo. La popolazione nel 2021 è stata di 25.738.100 abitanti, rappresentando la popolazione massima avuta in 22 anni. Calcoliamo con che probabilità l'Australia arrivi a **26 milioni di abitanti**. Quindi, calcoliamo qual'è la probabilità di avere un **incremento** di **almeno** $26.000.000 - 25.738.100 =$ **261.900 abitanti.**

```
incrementiDecrementi <- c(245900,220700,225600,212000,244100,
274100,376500,421400,442300,339900,308200,393500,394800,
```

```
347600,340300,373200,410900,385500, 383000,331600,40800)

1 - pnorm(261900, mean = mean(incrementiDecrementi),
sd = sd(incrementiDecrementi))

## [1] 0.7272106
```

La probabilità di avere un incremento di almeno 261.900 abitanti è del **~73%**. Una probabilità così alta è sicuramente derivata dal fatto che non si sono mai registrati decrementi e che la **media degli incrementi è di 319.614 abitanti.**
Per verificare la fattibilità di questo risultato, verranno generati $\mathbf{X = 1000}$ valori casuali in linea con la media e la deviazione standard degli incrementi dell'Australia.

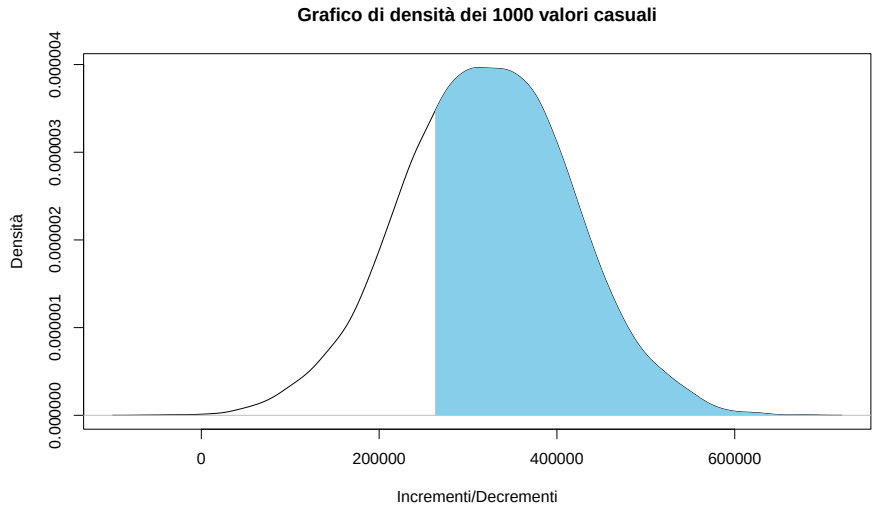


**Figura 5.1:** Density Plot Austrlia dei 1000 valori causali

In Figura 5.1 è possibile osservare in blu l'area che fa riferimento a valori di **incremento maggiori o uguali a 261.900**. **L'area è in linea con la probabilità restituita dalla distribuzione normale**, inoltre, all'interno dei valori causali c'è **qualche valore negativo** in quanto nel 2021 l'Australia ha avuto un incremento minimo di **40.000 abitanti**, quindi è presente nella distribuzione casuale qualche valore che è **vicino allo 0**.

**Bulgaria**

Osservando la serie temporale relativa alla Bulgaria (Figura 2.3) sappiamo che tale nazione **non ha mai registrato un incremento rispetto l'anno precedente**, avendo una significativa decrescita costante nel tempo. Tuttavia, escludendo il 2002 in cui ha avuto il maggior picco discendete, i **decrementi** registrati **sono stati lievi**, con una **media di -60.682 abitanti**. Di conseguenza, la possibilità di registrare il primo incremento non è così infattibile.
Calcoliamo quindi, con che **probabilità** la popolazione bulgara **rimanga invariata o aumenti nel 2022** rispetto l'anno precedenteà (2021). Consideriamo la probabilità di avere un **incremento maggiore o uguale a 0.**

```
incrementiDecrementi <- c(-41410,-280650,-63310,-60360,-56580,
-59200,-56700,-54670,-50880,-45350,-52340,-42210,-42670,
-38870,-43480,-48420,-51920,-51830,-49990,-48560,-34930)

1 - pnorm(0, mean = mean(incrementiDecrementi),
sd = sd(incrementiDecrementi))

## [1] 0.1167901
```

La probabilità di avere un incremento della popolazione bulgara nel 2022 è del **~12%.** Rappresenta un valore basso, tuttavia, è una **buona percentuale** considerando che la nazione **ha solo registrato decrementi**. Una percentuale così "alta" è sicuramente dovuto al fatto che, come detto in precedenza, la media dei decrementi è bassa considerando una popolazione milionaria.

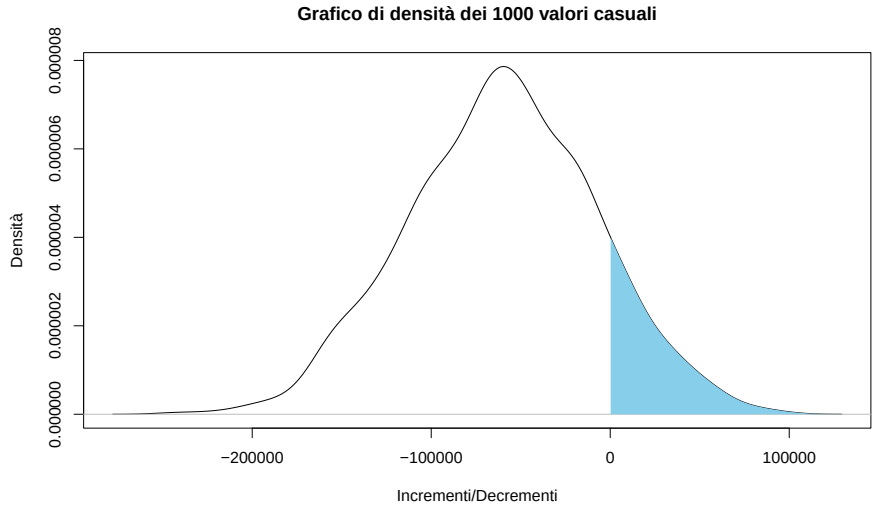


**Figura 5.2:** Density Plot Bulgaria dei 1000 valori causali

In Figura 5.2 è possibile osservare in blu l'area che fa riferimento a **valori di incremento maggiori o uguali a 0**. L'area in blu rappresenta solo una piccola parte rispetto l'intero grafico, di conseguenza, **è in linea con la probabilità restituita dalla distribuzione normale**. Tuttavia, ci sono valori che arrivano a simulare un incremento anche di **100.000 abitanti**.

### Germania

Osservando la serie temporale relativa alla Germania (Figura 2.3) sappiamo che tale nazione ha avuto sia trend positivi che negativi, con un picco anomalo nel 2011, che ha registrato una **perdita di ~1.6 Milioni di abitanti.**
Calcoliamo con che **probabilità** si potrebbe verificare nuovamente un **decremento di questa portata (1 Milione in su).**

```
incrementiDecrementi <- c(96000,180800,96400,-5000,-30900,
-62800,-123100,-97100,-215400,-200100,-1580200,105800,195800,
243800,430000,978200,346000,270700,226800,147500,-11700)

pnorm(-1000000, mean = mean(incrementiDecrementi),
sd = sd(incrementiDecrementi))

## [1] 0.01063215
```

La probabilità di avere un decremento uguale o maggiore di 1 milione di abitanti è uguale **all'~1%**, indicando che tale evento è **praticamente infattibile**, nonostante ci siano già stati decrementi di **200.000 e 1.600.000 abitanti.**

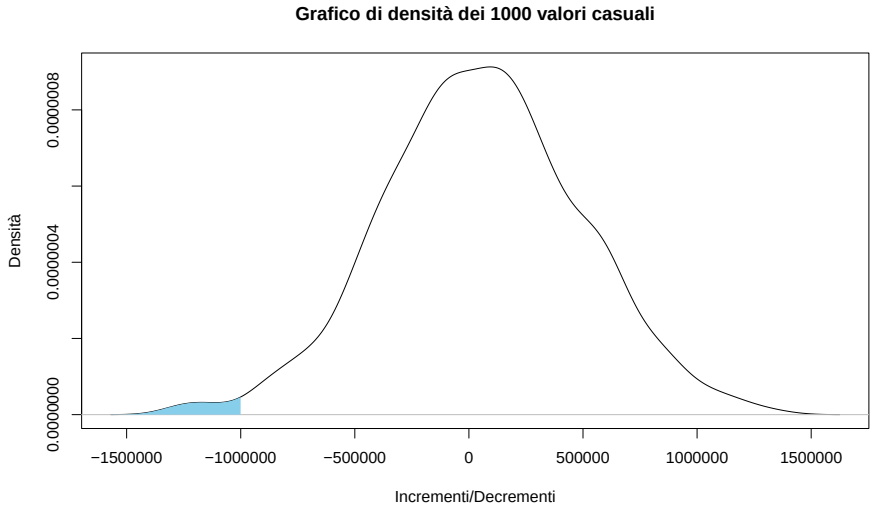


**Figura 5.3:** Density Plot Germania dei 1000 valori causali

In Figura 5.2 è possibile osservare in blu l'area che fa riferimento a **valori di decremento uguale o maggiore di 1 milione. L'area è davvero minima**, infatti, solo **10 valori su 1000 (1%) vanno oltre al meno un milione.** Quindi i valori generati **sono in linea con la probabilità** calcolata dalla distribuzione normale. Inoltre, il **decremento maggiore generato è stato di 1.28 Milioni, lontano dai 1.6 milioni del 2011**. Ciò riconferma che il decremento avuto è stato **è molto anomalo.**

### Grecia

Osservando la serie temporale relativa alla Grecia (Figura 2.6) sappiamo che tale nazione ha avuto un buon trend, inizialmente positivo, per poi dal 2011, diventare negativo, che ha portato alla Grecia a registrare nel 2021 il minimo storico di abitanti dal 2000. Tuttavia, **la differenza tra la popolazione del 2021, rispetto al 2000, è di "solo" 97.000 abitanti.**

Calcoliamo con che probabilità la popolazione greca ritorni allo **stesso numero del 2000**. Calcoliamo quindi, la probabilità di avere un **incremento di almeno 97.000 abitanti.**

```
incrementiDecrementi <- c(60400,52300,27500,24600,29500,
34800,31300,24900,33800,24600,4100,-37000,-82800,-76800,
-68800,-74300,-15500,-27000,-16600,-6000,-40000)

1 - pnorm(97000, mean = mean(incrementiDecrementi),
sd = sd(incrementiDecrementi))

## [1] 0.0115351
```

Nonostante l'incremento (97.000 abitanti) **non fosse così elevato**, considerando che la Grecia ha una popolazione di **oltre 10 milioni di abitanti**, la probabilità restituita **è pari all'1%.** Percentuale molto bassa considerando che la Grecia **per oltre 10 anni ha avuto incrementi** della propria popolazione, quindi perchè è un evento così raro? Visualizzando più attentamente i valori degli incrementi e decrementi, si nota che nonostante gli anni di incrementi, **il valore dei decrementi è proporzionalmente molto più significativo**. Infatti, la **media degli incrementi è di -4700 abitanti**, quindi, basandoci su quesi dati, **è più probabile avere un decremento**, piuttosto che un incremento nel 2022. Ed è questa la ragione per il quale un incremento di 97.000 abitanti ha una **probabilità così bassa.**

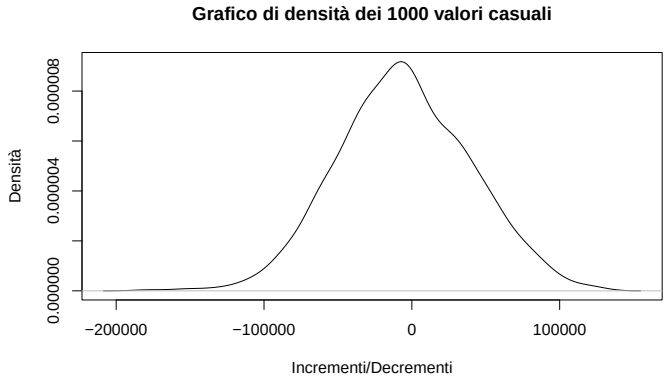


**Figura 5.4:** Density Plot Grecia dei 1000 valori causali

Infatti, anche la generazione dei 1000 valori casuali ha portato lo stesso risultato, dove solo **10 valori su 1000 (1%) hanno un valore** $>=$ **97.0000**, rendendo l'area in blu all'interno della Figura 5.4 praticamente **invisibile.**

### Croazia

Osservando la serie temporale relativa alla Croazia (Figura 2.7) sappiamo che tale nazione ha avuto un trend molto negativo, con un picco significativo nel 2001, che ha portato ad una perdita di ~200.000 abitanti, pari al 5% della popolazione totale. Tuttavia, dal 2001 al 2009 la popolazione è rimasta costante, **con qualche minimo incremento, di massimo 10.000 abitanti.**
Calcoliamo con che probabilità la popolazione croata **aumenti di almeno 10.000 abitanti.**

```
incrementiDecrementi <- c(-202330,10080,-110,350,5130,
1630,1040,-1560,-2170,-6950,-12990,-13880,-13840,
-15330,-21490,-34650,-36460,-48720,-29240,-18080,-21810)

1 - pnorm(10000, mean = mean(incrementiDecrementi),
sd = sd(incrementiDecrementi))

## [1] 0.2339613
```

La probabilità è del **23%**, un valore comunque alto considerando che la **media degli incrementi è di -22.000 abitanti** e che il massimo valore di incremento è stato proprio di 10.000 abitanti (**registrato un unico anno).**

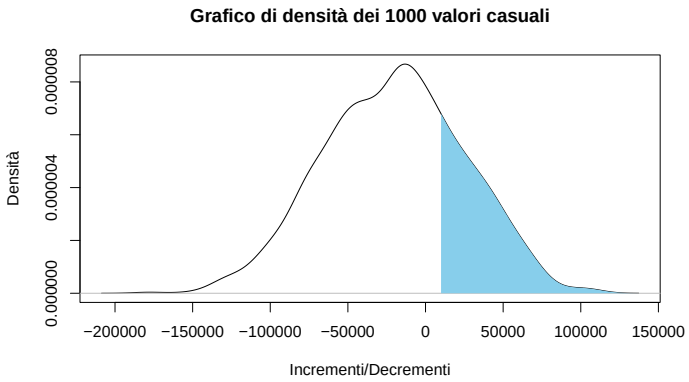


**Figura 5.5:** Density Plot Croazia dei 1000 valori causali

La generazione casuale dei 1000 valori ha portato a una probabilità ancora più alta, con il **25%** dei valori che avevano **almeno valore 10.000** e con **picchi fino a**

**100.000.**

**Italia**

Osservando la serie temporale relativa all'Italia (Figura 2.9) sappiamo che tale nazione fino al 2015 ha avuto una forte crescita, con un picco di 1.1 milioni di abitanti nel 2014. Dal 2015 in poi la popolazione ha avuto una notevole discesa raggiungendo nel 2021 la stessa popolazione (59.236.200) che aveva tra il 2010 e il 2011. Considerando che attualmente è in un trend negativo, calcoliamo con che probabilità la popolazione italiana nel 2022 sia uguale a quella del 2009 (59.000.600). Di conseguenza, definiamo la probabilità di **avere almeno un decremento di** $59.236.200 - 59.000.600 = 235.600$ **abitanti.**

```
incrementiDecrementi <- c(37200,26800,143000,365400,378900,
189400,159500,429200,347700,189500,174600,29500,291000,
1097500,12900,-130000,-76200,-105400,-667300,-175200,-405300)

pnorm(-235600, mean = mean(incrementiDecrementi),
sd = sd(incrementiDecrementi))

## [1] 0.161233
```

La possibilità di avere un decremento nel 2022 di almeno 235.600 abitanti, e quindi, arrivare alla stessa popolazione del 2009 è del **16%**. Una percentuale molto alta considerando che la **media degli incrementi** è di **100.000 abitanti**, e che la probabilità di avere un incremento (**62%**) è maggiore rispetto a quella di un decremento. Tuttavia, bisogna considerare che l'Italia si trova in un **trend negativo**, in cui è arrivata a perdere anche **600.000 abitanti in un anno**, quindi, la probabilità è **influenzata sicuramente da questi forti decrementi.**

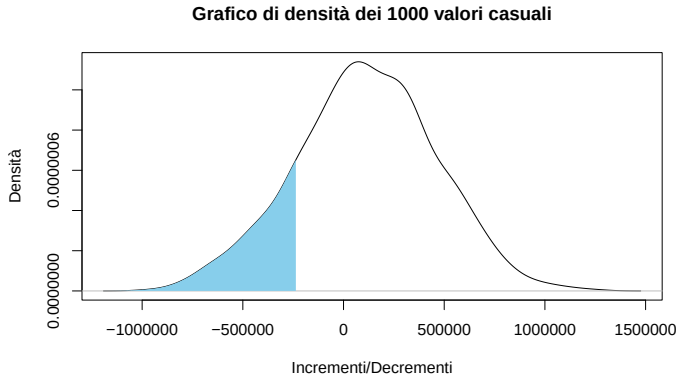


**Figura 5.6:** Density Plot Italia dei 1000 valori causali

La generazione dei 1000 valori causali è **perfettamente in linea** con la probabilità restituita dalla distribuzione normale, infatti, il **16% dei valori è minore o uguale a -235600**. In Figura 5.6 è possibile osservare l'area blu che rappresenta questi valori. Dal grafico è possibile osservare anche valori che hanno simulato un **decremento fino a 1 milione di abitanti.**

### Portogallo

Osservando la serie temporale relativa al Portogallo (Figura 2.12) sappiamo che tale nazione ha avuto una forte crescita dal 2000 al 2010, con anche forti picchi. Dal 2011 al 2019 la popolazione ha avuto una **forte discesa**, arrivando quasi quasi allo stesso numero di abitanti del 2000. Nel 2020 e 2021 la popolazione è tornata a ricrescere, ma si trova comunque allo stesso numero di abitanti che aveva tra il 2000 e il 2001. Calcoliamo con che probabilità si verifichi un **picco discendente** di abitanti che porti la popolazione portoghese allo **stesso livello del 2000.** Calcoliamo quindi, con che probabilità si verifichi un picco decrescente di almeno $10298300 - 10249000 =$ **49.300 abitanti.**

```
incrementiDecrementi <- c(81800,63900,49900,28500,21600,
17300,20600,20700,9700,10500,-800,-30300,-55100,
-60000,-52500,-33500,-31700,-18600,-14400,19300,2400)

pnorm(-49300, mean = mean(incrementiDecrementi),
```

```
sd = sd(incrementiDecrementi))

## [1] 0.08806734
```

La probabilità di avere un picco negativo di **almeno 49.300 abitanti** è del **9%**. Trattandosi di un picco, ci si aspettava di ricevere una probabilità bassa, tuttavia, **non è** un evento così **improbabile.**

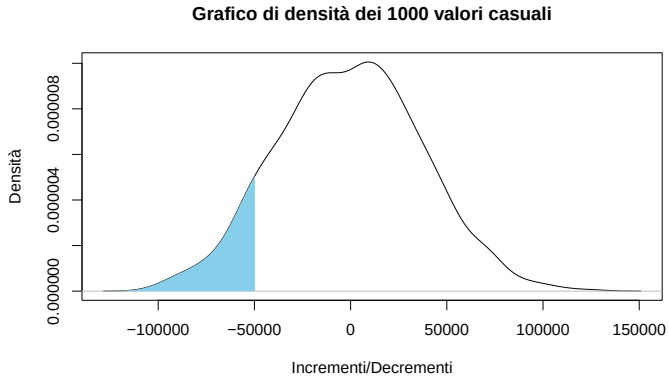


**Figura 5.7:** Density Plot Portogallo dei 1000 valori causali

Infatti, anche la simulazione con 1000 valori casuali ha portato ad avere il **10% dei valori minori o uguali di -49300**, suggerendo che tale picco non è un evento così improbabile, ma anzi, può realmente accadere in quanto ci sono stati valori che hanno toccato anche i **-100.000 abitanti**, visibili nel grafico in Figura 5.7.

## 5.2 Stima puntuale

Quando parliamo di **stime puntuali** quello che vogliamo fare è **ottenere informazioni su un parametro non noto** della popolazione effettuando su un campione estratto da quest'ultima delle opportune misure. Introduciamo quindi gli stimatori. Quando parliamo di uno stimatore si intende una funzione che associa ad ogni possibile campione un valore del parametro che si vuole stimare. Abbiamo dunque una variabile casuale funzione del campione che assume valore tra i possibili valori del parametro che si vuole stimare.

La stima può essere:

- **Puntuale**: si risolve in un **valore** assunto a rappresentare un parametro della popolazione.
- **Intervallare**: si risolve nel fissare **due valori** tra cui si presume sia **compreso** un parametro della popolazione.

Nel contesto di questa analisi statistica, effettuare una stima puntuale, cioè cercare di **determinare un singolo valore**, **risulterebbe impraticabile** a causa della vastità dei dati espressi in milioni.

### 5.2.1 Stima intervallare

Alla stima puntuale di un parametro non noto di una popolazione (costituita da un singolo valore reale) spesso si preferisce sostituire un intervallo di valori, detto **intervallo di confidenza** o **intervallo di fiducia**.
si cerca di determinare in base ai dati del campione un **limite inferiore** ed uno **superiore** entro i quali è compreso il **parametro non noto** con un certo **coefficiente di confidenza** (detto anche grado di fiducia). Utilizzeremo tale stima per stimare **l'intervallo dei possibili incrementi/decrementi di una nazione nel 2022**, considerando il suo **campione** di incrementi/decrementi dal **2000 al 2021.**

### 5.2.2 Intervallo di confidenza per $\mu$ con varianza nota

Verrà stimato l'intervallo dei possibili incrementi/decrementi di una nazione nel 2022. Verrà considerato un intervallo di confidenza del **99%**, che indica, sulla base del campione osservato, che **c'è una probabilità del 99% che l'intervallo calcolato contenga il vero valore del parametro di incremento/decremento**. Tale stima verrà applicata a un sottoinsieme di nazioni della Sottosezione 5.1.3 riferita alle probabilità. Verranno inseriti degli snippet di codice per mostrare i calcoli effettuati in R che hanno portato all'intervallo stimato.

**Australia**

```
incrementiDecrementi <- c(19026200,19272100,19492800,19718400,
19930400,20174500,20448600,20825100,21246500,21688800,
22028700,22336900,22730400,23125200,23472800,23813100,
24186300,24597200,24982700,25365700,25697300,25738100)

campnorm <- diff(incrementiDecrementi)
alpha <- 1-0.99
n <- length(campnorm)

mean(campnorm) - qnorm(alpha/2,mean=0,sd=1)*sd(campnorm)/sqrt(n)

## [1] 373288.7

mean(campnorm) + qnorm(alpha/2,mean=0,sd=1)*sd(campnorm)/sqrt(n)

## [1] 265939.9
```

La stima dell'intervallo di confidenza di grado 1 - $\lambda$ = **0.99** per l'incremento/decremento dell'Australia è di **(265.939,9 ; 373.288,7).**

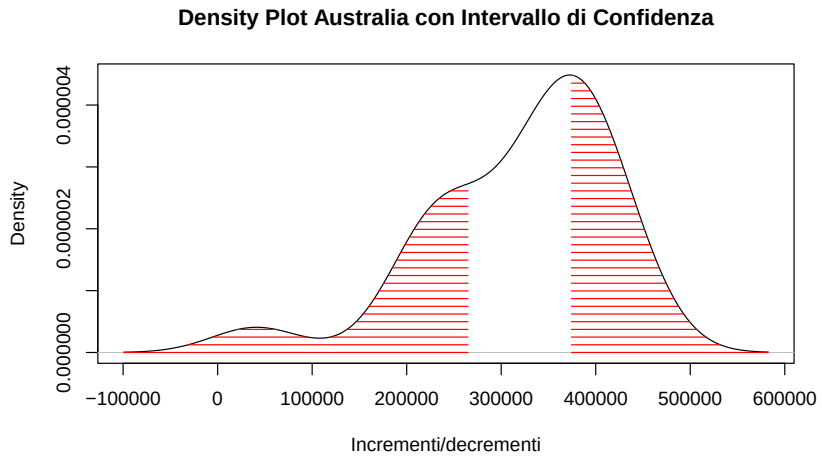


**Figura 5.8:** Density Plot Australia intervallo di confidenza

In Figura 5.8 l'area in bianco rappresenta l'intervallo di confidenza calcolato.

### Bulgaria

```
incrementiDecrementi <- c(8190880,8149470,7868820,7805510,7745150,
7688570,7629370,7572670,7518000,7467120,7421770,7369430,
7327220,7284550,7245680,7202200,7153780,7101860,7050030,
7000040,6951480,6916550)

campnorm <- diff(incrementiDecrementi)
alpha <- 1-0.99
n <- length(campnorm)

mean(campnorm) - qnorm(alpha/2,mean=0,sd=1)*sd(campnorm)/sqrt(n)

## [1] -32047.85

mean(campnorm) + qnorm(alpha/2,mean=0,sd=1)*sd(campnorm)/sqrt(n)

## [1] -89316.91
```

La stima dell'intervallo di confidenza di grado 1 - $\lambda$ = 0.99 per l'incremento/decremento della Bulgaria è di **(-89,316.91 ; -32.047,85).**

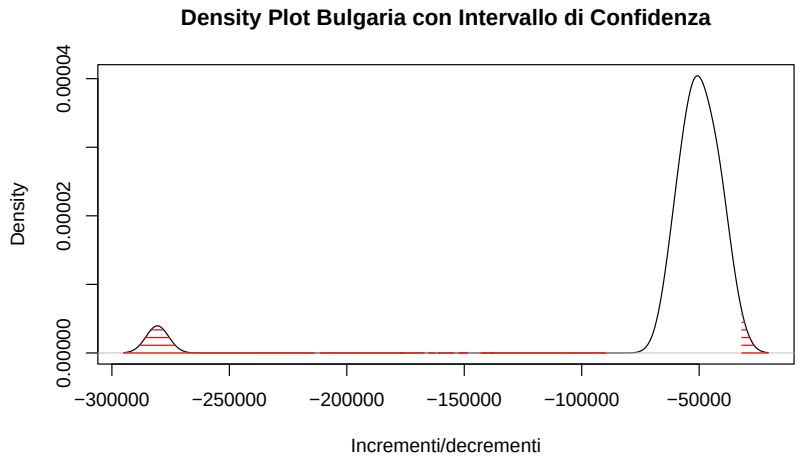


**Figura 5.9:** Density Plot Bulgaria intervallo di confidenza

In Figura 5.9 l'area in bianco rappresenta l'intervallo di confidenza calcolato.

**Italia**

```
incrementiDecrementi <- c(56923500,56960700,56987500,57130500,5749590
57874800,58064200,58223700,58652900,59000600,59190100,
59364700,59394200,59685200,60782700,60795600,60665600,
60589400,60484000,59816700,59641500,59236200)

campnorm <- diff(incrementiDecrementi)
alpha <- 1-0.99
n <- length(campnorm)

mean(campnorm) - qnorm(alpha/2,mean=0,sd=1)*sd(campnorm)/sqrt(n)

## [1] 306541.2

mean(campnorm) + qnorm(alpha/2,mean=0,sd=1)*sd(campnorm)/sqrt(n)

## [1] -86284.07
```

La stima dell'intervallo di confidenza di grado 1 - = 0.99 per l'incremen- to/decremento dell'Italia è di **(-86.284,07 ; 306.541,2).**

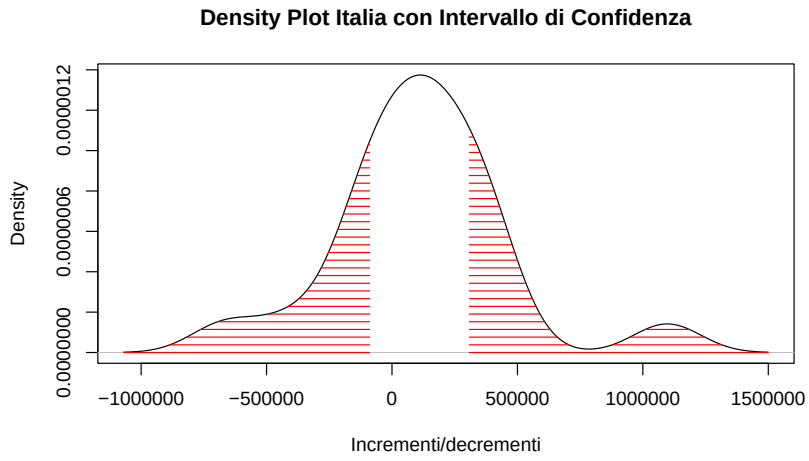


**Figura 5.10:** Density Plot Italia intervallo di confidenza

In Figura 5.10 l'area in bianco rappresenta l'intervallo di confidenza calcolato.

### 5.2.3 Confronto tra popolazioni

Può essere interessante **stimare la differenza tra le medie di due distinte popolazioni** considerando due campioni casuali indipendenti, per verificare chi **avrà un incremento più forte.**

Per fare il confronto si costruisce un **intervallo di confidenza per la differenza tra le due medie** con un certo grado di fiducia 1 - $\alpha$. Confrotiamo considerando i campioni (incrementi/decrementi) indipendenti dell'Italia e della Germania con un grado di fiducia 1 - $\alpha = \mathbf{0.99}$:

```
incrementiDecrementiI <- c(56923500,56960700,56987500,57130500,
57495900,57874800,58064200,58223700,58652900,59000600,
59190100,59364700,59394200,59685200,60782700,60795600,
60665600,60589400,60484000,59816700,59641500,59236200)
incrementiDecrementiG <- c(82163500,82259500,82440300,82536700,
82531700,82500800,82438000,82314900,82217800,82002400,
81802300,80222100,80327900,80523700,80767500,81197500,
82175700,82521700,82792400,83019200,83166700,83155000)

campItalia <- diff(incrementiDecrementiI)
campGermania <- diff(incrementiDecrementiG)

alpha <- 1-0.99
nI <- length(campItalia)
nG <- length(campGermania)
mediaI<- mean(campItalia)
mediaG <- mean(campGermania)
sdI <- sd(campItalia)
sdG <- sd(campGermania)

mediaI-mediaG-qnorm(1-alpha/2,mean=0,sd=1) * sqrt(sdI*2/nI+sdG*2/nG)

## [1] 62201.47
```

```
mediaI-mediaG+qnorm(1-alpha/2,mean=0,sd=1) * sqrt(sdI*2/nI+sdG*2/nG)

## [1] 63627.1
```

Si ha che la stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ per $\mu_1 - \mu_2$ degli incrementi di abitanti delle due nazioni è **(62201.47,63627.1).**
**Poichè il limite inferiore e il limite superiori sono positivi**, si deduce che **L'Italia avrà un incremento superiore rispetto l'incremento della Germania** con un grado di fiducia del **99%**.

## 5.3 Verifica dell'ipotesi

**Dopo la stima dei parametri** il passo successivo è la **verifica delle ipotesi**. La verifica delle ipotesi interviene ogni volta che si ha il bisogno di predirre qualcosa. In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono: una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x;v)$, un'ipotesi su di un parametro non noto della popolazione ed un campione casuale estratto dalla popolazione.
Un'ipotesi statistica è un'affermazione o una congettura su un parametro non noto $\vartheta$. L'ipotesi soggetta a verifica è denotata con $\mathbf{H_0}$ ed è chiamata **ipotesi nulla**. Il test di ipotesi è il procedimento o regola con cui si decide, sulla base dei dati del campione, se **accettare** o **rifiutare** $\mathbf{H_0}$. La costruzione del test richiede la formulazione, in contrapposizione all'ipotesi nulla, di una proposizione alternativa, ed è di solito indicata con $\mathbf{H_1}$. Nel caso si verifichi che **l'ipotesi nulla sia falsa, l'ipotesi alternativa sarà vera e viceversa.**

### 5.3.1 Italia

Considerando il **campione degli incrementi/decrementi dal 2000 al 2021 (ampiezza = 21) della nazione italiana**, con una **media annua di incrementi di 110.128,6 abitanti** e con **deviazione standard di 349.431,5 abitanti, formuliamo l'ipotesi che L'Italia**, nel 2022, **avrà un incremento di almeno 300.000 abitanti.** Si desidera co-

struire il test di misura $a = 0.05$ per verificare l'ipotesi nulla $\mathbf{H_0}$:$\mu \geq 300.000$ in alternativa all'ipotesi $\mathbf{H_1}$:$\mu < 300.000$. Occorre considerare un test **unilaterale destro**. Utilizzando R si ha:

```
alpha <- 0.05
mu0 <- 300000
sigma <- 349431.5
n <- 21

qnorm(alpha,mean=0,sd=1)

## [1] -1.644854

meancamp <- 110128.6
z0 <- (meancamp - mu0) / (sigma / sqrt(n))
z0

## [1] -2.490045

pvalue <- pnorm(z0,mean=0,sd=1)
pvalue

## [1] 0.006386351
```

Si ha che il valore $z_o s = -2.490045$ **cade nella regione di rifiuto** poichè **è minore** del valore $z_a = -1.644854$, occorre quindi**rifiutare l'ipotesi nulla** con un livello di significatività del 5% ($\alpha = 0.05$). Poichè $\rho < \alpha$, l'ipotesi $\mathbf{H_0}$:$\mu \geq 300.000$ deve essere **rifiutata**.

### 5.3.2 Australia

Considerando il **campione degli incrementi/decrementi dal 2000 al 2021 (ampiezza = 21) della nazione australiana**, con una **media annua di incrementi di 319.614,3 abitanti** e con **deviazione standard di 95.490,46 abitanti**, **formuliamo l'ipotesi** che L'Australia, nel 2022, **avrà un incremento di almeno 350.000 abitanti.** Si desidera

costruire il test di misura $a = 0.05$ per verificare l'ipotesi nulla $\mathbf{H_0}$:$\mu \geq 350.000$ in alternativa all'ipotesi $\mathbf{H_1}$:$\mu < 350.000$. Occorre considerare nuovamente un test **unilaterale destro**. Utilizzando R si ha:

```
alpha <- 0.05
mu0 <- 350000
sigma <- 95490.46
n <- 21

qnorm(alpha,mean=0,sd=1)

## [1] -1.644854

meancamp <- 319614.3
z0 <- (meancamp - mu0) / (sigma / sqrt(n))
z0

## [1] -1.458206

pvalue <- pnorm(z0,mean=0,sd=1)
pvalue

## [1] 0.07239187
```

Si ha che il valore $z_o s = -1.458206$ **cade nella regione di accettazione** poichè **è maggiore** del valore $z_a = -1.644854$, occorre quindi **accettare l'ipotesi nulla** con un livello di significatività del 5% ($\alpha = 0.05$). Poichè $\rho > \alpha$, l'ipotesi $\mathbf{H_0}$:$\mu \geq 350.000$ deve essere **accettata**.

## 5.4 Criterio del chi-quadrato

Il criterio del chi-quadrato permette di verificare se un dato campione osservato può essere stato estratto da una popolazione descritta da una variabile aleatoria $X$ con una funzione di distribuzione $FX(x)$. **Si è interessati** alla distribuzione media $\mu$ relativi agli **incrementi di abitanti del Regno Unito**. A questo scopo si osserva

un **campione di ampiezza 21** che fanno riferimento agli **incrementi registrati dalla nazione dal 2000 al 2021.** La **media campionaria degli incrementi** è $\bar{x} = 406.895,2$ mentre la **deviazione standard** è $s = 97.396,97$:

```
uk <- c(214600,239800,261800,292400,388300,438300,
452900,498300,470700,467900,512300,472600,410200,
445900,502200,525600,465100,429500,373500,434600,248300)

n <- length(uk)
n

## [1] 21

m <- mean(uk)
m

## [1] 406895.2

s <- sd(uk)
s

## [1] 97396.97
```

Vogliamo verificare se la popolazione da cui proviene il campione **può essere descritta** da una variabile aleatoria $X$ di densità normale. Applichiamo il test chi-quadrato di misura $\alpha = 0.05$.

**Supponiamo di suddividere** l'insieme dei valori che $X$ può assumere in $r$=5 sottoinsiemi $I_1, I_2, ..., I_5$ in modo che risulti che la probabilità che $X$ assuma un valore appartenente a $I_i (i = 1, 2, ..., 5)$ sia uguale a $p_i = 0.2$

Utilizzando i quantili della normale possiamo determinare i sottoinsiemi $I_1, I_2, ..., I_5$:

```
a <- numeric(4)
for(i in 1:4)
a[i]<- qnorm(0.2*i, mean=m, sd=s)
a
```

```
## [1] 324923.9 382220.0 431570.5 488866.6
```

Si ha che gli intervalli $I_i (i = 1, 2, ..., 5)$ sono:

$I_1 = (-\infty, 324923.9)$, $I_2 = (324923.9, 382220.0)$, $I_3 = (382220.0, 431570.5)$,

$I_4 = (431570.5, 488866.6)$, $I_5 = [488866.6, +\infty)$

**Occorre ora determinare il numero di elementi del campione che cadono negli intervalli $\mathbf{I_i (i = 1, 2, ..., 5)}$:**

```
r <- 5
nint <- numeric(r)
nint[1] <- length(which(uk<a[1]))
nint[2] <- length(which((uk >=a [1]) & (uk < a[2])))
nint[3] <- length(which((uk >= a[2]) & (uk < a[3])))
nint[4] <- length(which((uk >= a[3]) & (uk < a[4])))
nint[5] <- length(which(uk >= a[4]))
nint

## [1] 5 1 3 8 4
```

Le frequenze degli intervalli sono: $\mathbf{n_1 = 5, n_2 = 1, n_3 = 3, n_4 = 8, n_5 = 4}$.

Calcoliamo $x^2$

```
chi2 <- sum(((nint - n *0.2)/sqrt(n*0.2))^2)
chi2

## [1] 6.380952
```

Si ha $x^2 = 6.380952$.

La distribuzione normale ha due parametri non noti $(\mu, \sigma^2)$ e quindi **k = 2**. Pertanto, la funzione di distribuzione della statistica $Q$ è approssimabile con la funzione di distribuzione chi–quadrato con $r - k - 1 = 2$ gradi di libertà. Ora occorre quindi calcolare $X^2_{1-\alpha/2, r-k-1}$ e $X^2_{\alpha/2, r-k-1}$ con $\alpha = 0.05$.

```
r <- 5
k <- 2
alpha <- 0.05
qchisq(alpha/2,df=r-k-1)

## [1] 0.05063562

qchisq(1-alpha/2,df=r-k-1)

## [1] 7.377759
```

Essendo che $0.05063562 < X^2 < 7.377759$, l'ipotesi $\mathbf{H_0}$ di popolazione normale **può essere accettata.**