UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

PROGETTO DI STATISTICA E ANALISI DEI DATI

Analisi statistica applicata al consumo di Alcol in Italia

Rilevazione Istat 2019

Docente:

Prof.ssa. Amelia G. Nobile

Studente:

Ferrara Carmine

Matr.05225/00990

ANNO ACCADEMICO 2020/2021

Sommario

Introduzione

Considerando differenti banche dati disponibili in rete, sicuramente è di grande importanza l'indagine statistica condotta annualmente dall'ISTAT in merito al consumo di alcol in Italia. Annualmente infatti, l'Istituto Nazionale di Statistica rende disponibili al pubblico tavole di dati molto dettagliate, nelle quali sono riportate informazioni molto dettagliate in materia, prendendo in considerazione la popolazione di 11 anni e più.

Nella banca dati Excel 2019 fornita dall'ISTAT, sono riportati numerosi tabulati in riferimento a vari indici tra cui il consumo di alcol per: fasce d'età, tipologie di bevande, distinzioni per sesso ecc. Ai fini di quest'analisi è stato scelto di utilizzare un Dataset che riporta il consumo di alcol secondo una scala di assiduità che va dal consumo moderato all'eccesso spericolato (valore di Binge Drinking) per migliaia di abitanti per ogni regione o provincia autonoma.

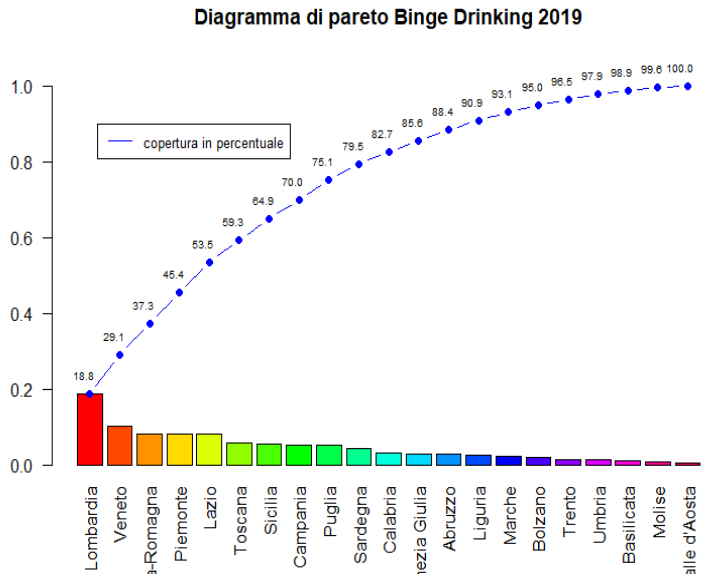
Dataset Utilizzato ai fini dell'indagine

Regioni	Consumo moderato	Comportamento abitudinario	Eccedenza abituale	Eccedenza abutuale a pasto	Binge drinking
Piemonte	2.055	684	463	222	311
Valle d'Aosta	54	28	16	6	16
Liguria	715	258	182	95	96
Lombardia	4.746	1.405	857	377	719
Bolzano	225	*101*	*42*	*12*	*73*
Trento	244	*89*	*44*	*12*	*56*
Veneto	2.390	747	441	179	394
Friuli-Venezia Giulia	578	200	122	36	110
Emilia-Romagna	2.091	715	471	237	316
Toscana	1.769	587	407	218	223
Umbria	418	126	85	42	53
Marche	708	207	140	81	83
Lazio	2.847	685	448	226	310
Abruzzo	596	191	106	39	110
Molise	138	47	33	13	25
Campania	2.565	528	407	211	197
Puglia	1.869	507	372	222	197
Basilicata	264	82	56	27	41
Calabria	903	234	152	73	122
Sicilia	2.330	469	290	174	211
Sardegna	683	267	132	54	167

Analisi Univariata – Colonna Binge – Drinking dataset Excel

Passo 1 – Analisi Univariata – Diagramma di Pareto

Regioni	Binge drinking
Piemonte	311
Valle d'Aosta	16
Liguria	96
Lombardia	719
Bolzano	73
Trento	56
Veneto	394
Friuli-Venezia Giulia	110
Emilia-Romagna	316
Toscana	223
Umbria	53
Marche	83
Lazio	310
Abruzzo	110
Molise	25
Campania	197
Puglia	197
Basilicata	41
Calabria	122
Sicilia	211
Sardegna	167



Dall'analisi del diagramma di Pareto realizzato con il dataset riportato (variabile Binge_Drinking – per valori in termini relativi (ogni valore è stato diviso per la somma totale dei valori in tabella)) si evince che come il tasso alcolemico di pericolosità massima

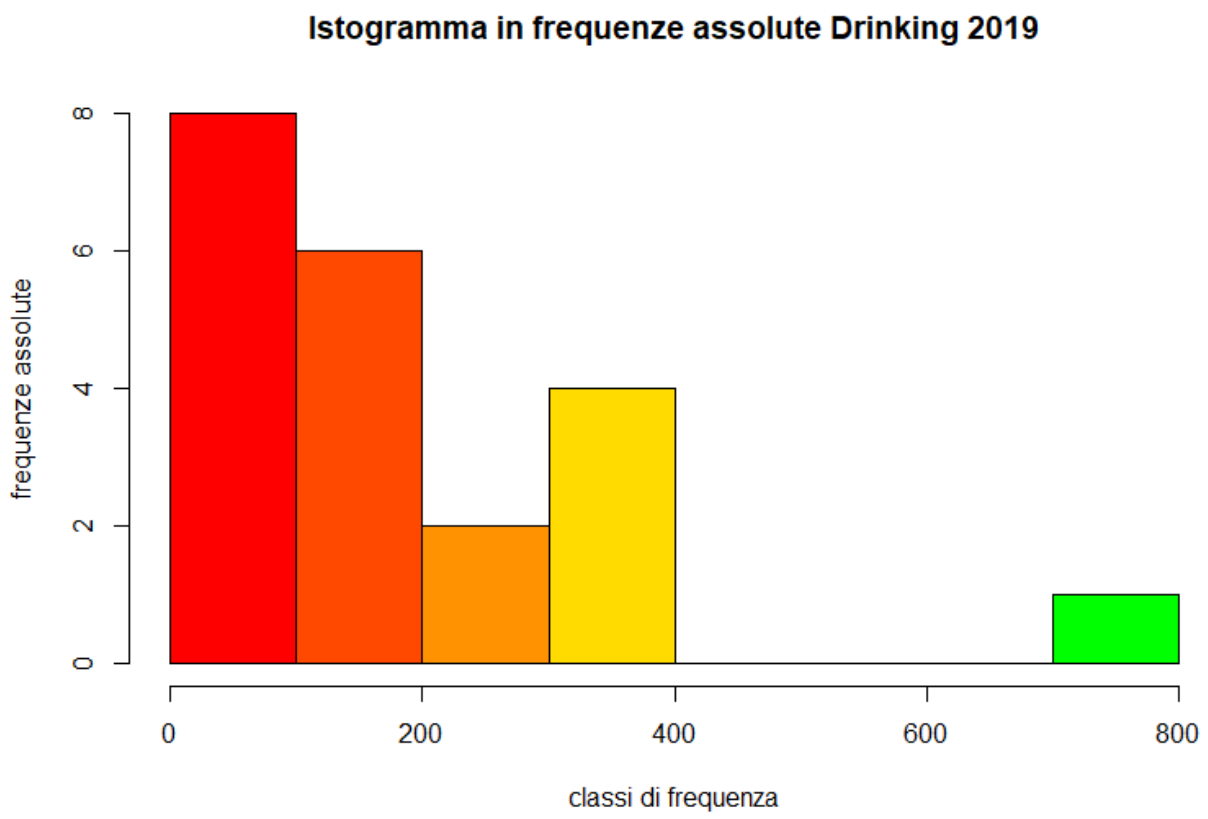
è incentrato in meno della metà delle regioni italiane o province autonome osservate, in particolare

la maggior affluenza di dati in termini relativi è visibile in 5 regioni di maggior rilievo

- Lombardia Picco Massimo
- Veneto Secondo punto
- Emilia Romagna / Piemonte e Lazio dati molto simili

Passo 2 – Analisi univariata – Instogramma – Box Plot ad intaglio

Dal dataset precedente è stato realizzato poi un istogramma in frequenze assolute, (ogni classe considera un intervallo di 100 unità).




```
List of 6
 $ breaks  : num [1:9] 0 100 200 300 400 500 600 700 800
 $ counts  : int [1:8] 8 6 2 4 0 0 0 1
 $ density : num [1:8] 0.00381 0.002857 0.000952 0.001905 0 ...
 $ mids    : num [1:8] 50 150 250 350 450 550 650 750
 $ xname   : chr "array_to_analyze"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```

Parametri estratti dall'istogramma (classi, Counts – frequenze assolute, densità e valore medio di ogni classe)

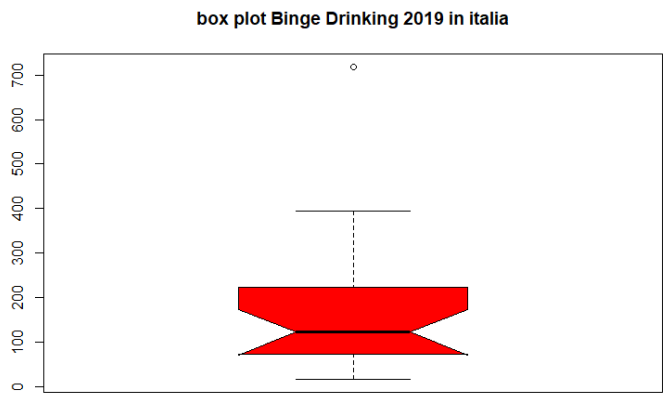
```
> freq_rel <- h$density *  100
> freq_rel
[1] 0.38095238 0.28571429 0.09523810 0.19047619 0.00000000 0.00000000 0.00000000 0.04761905
```

Frequenze relative delle 8 classi calcolate in base alla densità di ogni classe nell'istogramma.

Prime deduzioni:

- Centralità dei dati presente in una buona classe quantitativa delle classi selezionate, (molte regioni sono collocate in un contesto meno pericoloso, permettendo maggior controllo tale indice su base nazionale).
- Presenza di un singolo valore troppo elevato (Lombardia), il quale potrebbe risultare anomalo…

Per approfondire l'aspetto sugli indici di centralità del campione e sulla presenza effettiva di valori anomali nel campione effettuato, è interessante considerare quantitativamente gli indicatori di moda, media, mediana campionaria e soprattutto lo studio di un eventuale box plot in termini di frequenze riportate.



Dal box plot realizzato (considerando tutti i dati del campione), si possono già confermare già alcune deduzioni fatte dall'istogramma e dal diagramma di Pareto, per questa variabile considerata infatti la metà dei dati del campione si concentra su valori molto bassi, ed inoltre effettivamente il valore dato dalla singola regione Lombardia è anomalo rispetto al resto dei dati del campione, venendo riportato esplicitamente al di fuori dei baffi di copertura del diagramma.

Ciò è riscontrabile anche matematicamente considerando dapprima una prima analisi dei quartili riportati dal diagramma di pareto.

```
quantile(dati$`Binge drinking`);
0%  25%  50%  75% 100%
16   73  122  223  719
```

Considerando i dati del campione ordinati in ordine crescente e i quantili calcolati dal diagramma, abbiamo che

Il baffo inferiore sarà posizionato nella prima posizione maggiore rispetto al calcolo:

$Q_1 - 1.5 \cdot (Q_3 - Q_1) = 73 - 1.5\ (223 - 73) = -152$

Quindi esattamente 0 (primo valore superiore al risultato del calcolo) //Non esistono dati anomali, nell'estremo inferiore del campione in analisi.

Mentre per il baffo superiore:

Q3 + (1.5 * (Q3 – Q1)) = 223 + 1.5(223 - 73) = 448

Posizione = 394 – Veneto

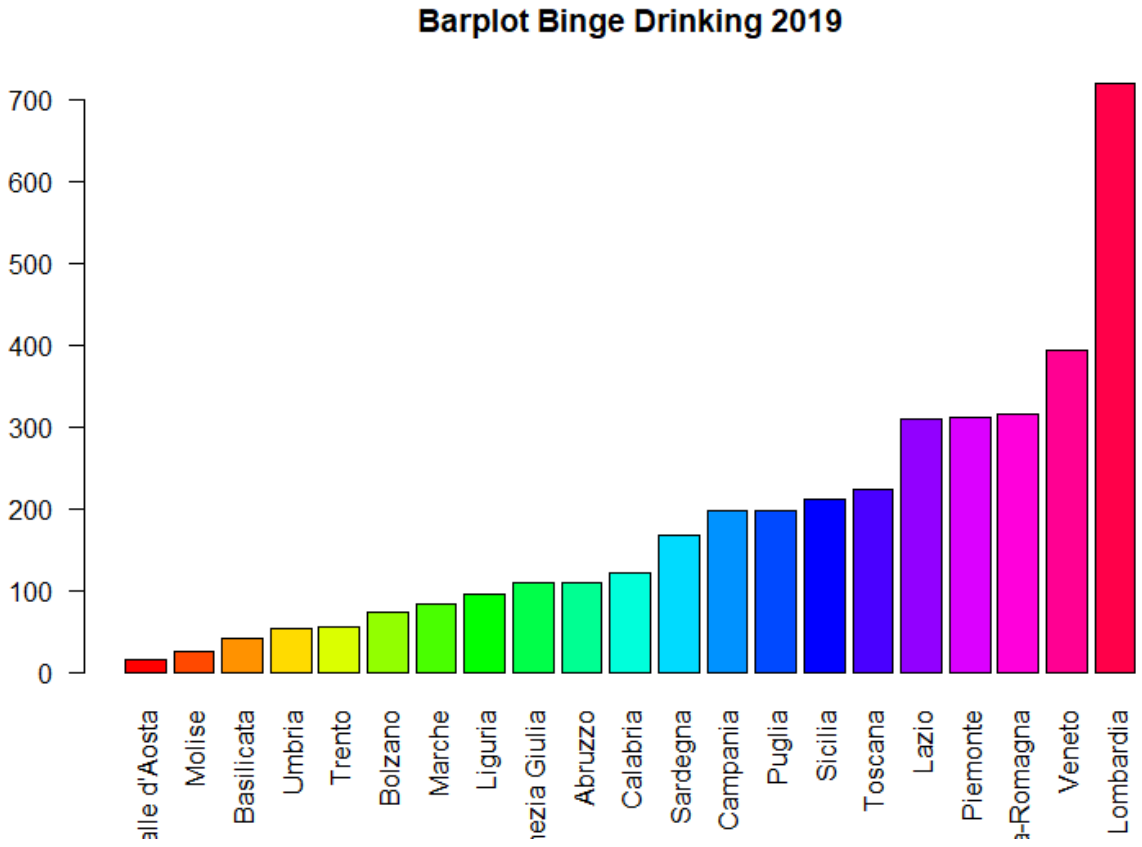
(primo valore inferiore al risultato del calcolo)

Matematicamente si dimostra che per il box plot il valore della Lombardia effettivamente è anomalo, e potrebbe provocare dispersione dei dati nel campione in analisi.

```
> IQR <- quantile(dati$`Binge drinking`, 0.75) - quantile(dati$`Binge drinking`, 0.25);
> M1 <- quantile(dati$`Binge drinking`, 0.5) - 1.57 * IQR/sqrt(length((dati$`Binge drinking`)))
> M2 <- quantile(dati$`Binge drinking`, 0.5) + 1.57 * IQR/sqrt(length((dati$`Binge drinking`)))
> c(M1, M2)
      50%       50%
 70.60969 173.39031
```

Dal box plot analizzato si può evincere anche che la mediana campionaria è posta intorno al valore 125.5, considerando anche l'intaglio riportato (Indice di fiducia posto con alfa = 1.5) si ha che appunto questo valore può variare da 70.61 a 173.39 come intervallo di confidenza.

Passo 3 – Indici di centralità rispetto al campione



Media e moda campionaria

Tenendo in considerazione tutti i dati del campione, abbiamo che la media campionaria assume valore:

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

```
> #mean
>
> M <-  mean(dati$`Binge drinking`)
> M
[1] 182.381
```

Notando tale valore, quindi una stima di centralità che tiene conto di tutti i dati del campione (anche il più estremo anomalo), abbiamo che la media risulta più grande della mediana campionaria calcolata fino ad adesso.

Osservazione: dal dato di media campionaria riportato abbiamo che il valore per la regione Lombardia ha uno scarto molto elevato dalla media campionaria

Scarto Lombardia =

```
> dati$`Binge drinking`[4] - M
[1] 536.619
```

Confermando ancora una volta il distacco di questo dato rispetto alla centralità del campione.

Considerando quindi questo forte scarto, si può presumere che il campione abbia un forte sbilanciamento verso destra, se si considerano i dati ordinati in modo crescente (come nel bar plot riportato).

Per provare a confermare ciò è di rilevanza anche il calcolo della mediana campionaria (tenendo quindi conto solo dei valori centrali del campione in analisi).

#dati = 21 (dispari) – per il calcolo della mediana campionaria -> $Median = C\left[\frac{n+1}{2}\right] = 122$

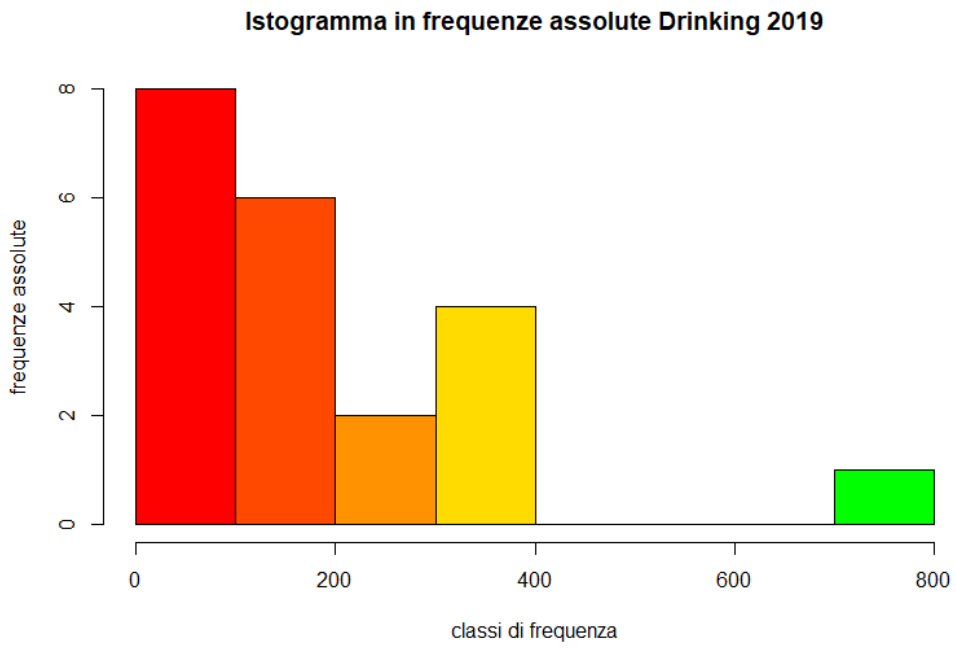
Regioni	Binge drinking
Valle d'Aosta/Vallée d'Aoste	16
Molise	25
Basilicata	41
Umbria	53
Trento	56
Bolzano/Bozen	73
Marche	83
Liguria	96
Abruzzo	110
Friuli-Venezia Giulia	110
Calabria	122
Sardegna	167
Campania	197
Puglia	197
Sicilia	211
Toscana	223
Lazio	310
Piemonte	311
Emilia-Romagna	316
Veneto	394
Lombardia	719

```
> median(dati$`Binge drinking`)
[1] 122
```

Come prima considerato, essendo la media > della mediana campionaria, si deduce che il campione ordinato è fortemente sbilanciato verso destra.

Considerando nuovamente la divisone in classi dell'istogramma precedentemente realizzato

```
List of 6
 $ breaks  : num [1:9] 0 100 200 300 400 500 600 700 800
 $ counts  : int [1:8] 8 6 2 4 0 0 0 1
 $ density : num [1:8] 0.00381 0.002857 0.000952 0.001905 0 ...
 $ mids    : num [1:8] 50 150 250 350 450 550 650 750
 $ xname   : chr "array_to_analyze"
 $ equidist: logi TRUE
 - attr(*, "class")= chr "histogram"
```



E le frequenze relative precedentemente calcolate in funzione di tale suddivisione in classi

```
> freq_rel <- h$density *  100
> freq_rel
[1] 0.38095238 0.28571429 0.09523810 0.19047619 0.00000000 0.00000000 0.00000000 0.04761905
```

Regioni	Binge drinking
Valle d'Aosta/Vallée d'Aoste	16
Molise	25
Basilicata	41
Umbria	53
Trento	56
Bolzano/Bozen	73
Marche	83
Liguria	96
Abruzzo	110
Friuli-Venezia Giulia	110
Calabria	122
Sardegna	167
Campania	197
Puglia	197
Sicilia	211
Toscana	223
Lazio	310
Piemonte	311
Emilia-Romagna	316
Veneto	394
Lombardia	719

Possiamo anche dedurre come per il campione in analisi, il dato di densità riportato è particolarmente interessante in termini della tematica osservata, infatti la prima classe d'intervallo [0:100), risulta avere una frequenza sia assoluta che relativa maggiore rispetto agli altri dati del campione (più un terzo del totale ragionando in termini nazionali, 36%).

Tale intervallo secondo le stime riportate, indica che la classe modale del campione in analisi è proprio la classe [0:100), quindi quella con maggior concentrazione di dati rispetto all'intero campione.

Considerando tale indice di stima, con il resto dell'analisi fin ora condotta (valore anomalo riportato dal box-plot per la regione Lombardia), si può facilmente supporre come il consumo di alcol in

maniera eccessivamente pericolosa nel 2019, sia stato uniformemente stabile per un terzo della nazione su valori bassi (o per più del 60% se si considera anche la seconda classe d'intervallo), ma comunque sono presenti regioni di rilevanza critica, da risultare addirittura anomale rispetto alla centralità dei dati riportati.

Considerando nuovamente il campione diviso in classi, allo stesso modo dell'istogramma,

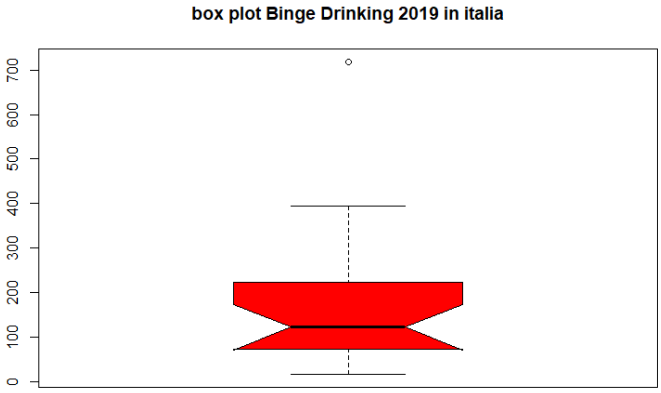
```
> freq_class

(0,100] (100,200] (200,300] (300,400] (400,500] (500,600] (600,700] (700,800]
      8         6         2         4         0         0         0         1
```

E considerandone la distribuzione di frequenza relativa cumulata.

```
> cumsum(freq_class / length(dati$`Binge drinking`))
  (0,100] (100,200] (200,300] (300,400] (400,500] (500,600] (600,700] (700,800]
0.3809524 0.6666667 0.7619048 0.9523810 0.9523810 0.9523810 0.9523810 1.0000000
```

Possiamo osservare come la modalità data classe (100, 200] sia la mediana per frequenze, della nostra analisi.

Passo 4 – Indici di dispersione rispetto alla media campionaria



Se si considera il boxplot precedentemente realizzato, si nota che effettivamente ha senso per il campione analizzato, considerare anche gl'indici di dispersione. In quanto si nota facilmente che alcuni dati del campione si distaccano di molto dalla mediana, ma anche dal precedente valore di media campionaria precedentemente calcolato.

```
> summary(dati$`Binge drinking`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   16.0    73.0   122.0   182.4   223.0   719.0
```

Osservando attentamente il summary del campione analizzato, tale sospetto acquista ancora più fondamento, in quanto il 75% dei dati del campione si attesta al di sotto del valore 223 (terzo quartile).

Ma effettivamente è davvero importante considerare la dispersione dei dati rispetto alla media campionaria?

Essendo che la variabile Binge_Drinking considerata, indica la pericolosità massima per il consumo di alcol in Italia, sarebbe davvero utile capire effettivamente quali regioni necessitano di maggior attenzione.

Ma se la media campionaria è a 182.4, significa che il dato nazionale si mantiene abbastanza stabile, perché procedere con ulteriori analisi?

Osservando il grafico, abbiamo che la media campionaria, non permette di identificare con esattezza quali regioni godono di maggior criticità. Quindi effettivamente è senz'altro doveroso capire in quali regioni è necessario intervenire al fine di ridurre ulteriormente la pericolosità del consumo eccessivo di alcol.

Come è possibile fare ciò?

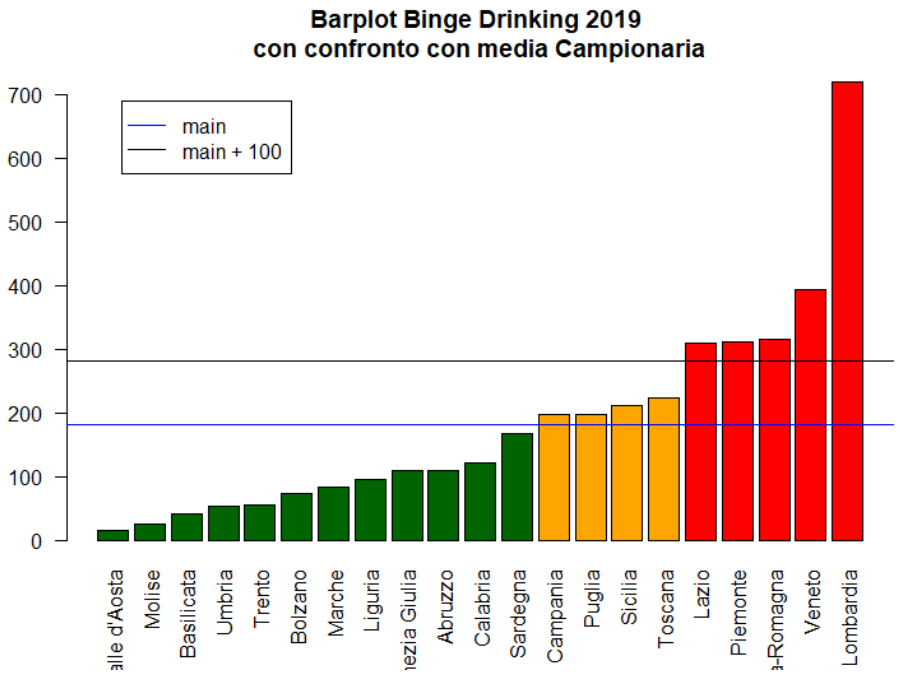
Considerando la media campionaria (media nazionale), è possibile delimitare in che modo i dati si discostino da essa.

Regioni	Binge drinking	Scarto dalla media
Valle d'Aosta/Vallée d'Aoste	16	-166
Molise	25	-157
Basilicata	41	-141
Umbria	53	-129
Trento	56	-126
Bolzano/Bozen	73	-109
Marche	83	-99
Liguria	96	-86
Abruzzo	110	-72
Friuli-Venezia Giulia	110	-72
Calabria	122	-60
Trentino-Alto Adige	129	-53
Sardegna	167	-15
Campania	197	15
Puglia	197	15
Sicilia	211	29
Toscana	223	41
Lazio	310	**128**
Piemonte	311	**129**
Emilia-Romagna	316	**134**
Veneto	394	**212**
Lombardia	719	**537**

Dati che si discostano per più di 100.000 persone dalla media nazionale

Se si considera la media nazionale come valore di guardia (quindi in Italia, la soglia eccessiva per consumo smisurato di alcol in Italia è di 182.400 persone l'anno in media). Dal calcolo degli scarti dalla media campionaria è facile evincere che le 5 regioni (rilevanti dal diagramma di Pareto iniziale) segnalate in rosso possono rappresentare fattore di rischio molto più elevato rispetto al resto della nazione.

Infatti, considerando nuovamente i dati del campione in un barplot ordinato e posizionando anche i valori di media campionaria e di media campionaria + 100 unità.



Osserviamo che i dati riportati in rosso, sicuramente innalzano di molto la soglia di rischio che nel 2019 è stata rilevata in Italia.

Quindi sicuramente, una normalizzazione degli stessi "in vicinanza al valore medio" potrebbe diminuire la dispersione di dati.

Ma in che modo?

Capire effettivamente quali dati debbano variare, al fine di rendere meno elevato il valore di dispersione della media campionaria, è un qualcosa che non si può fare solo con i dati a disposizione in questo momento. A tale scopo, prima di formulare ipotesi si procede a calcolare la varianza e la deviazione standard rispetto al campione.

Varianza campionaria

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad (n = 2, 3, \ldots),$$

```
> var(dati$`Binge drinking`)
[1] 26756.85
```

Deviazione Standard campionaria

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (n = 2, 3, \ldots).$$

```
> sd(dati$`Binge drinking`)
[1] 163.5752
```

In termini teorici questi dati (in particolare la deviazione standard, confrontabile con la media, dato lo stesso ordine di grandezza) indicano effettivamente il grado di dispersione dei dati rispetto alla media.

Effettivamente nell'analisi riportata, la deviazione standard, raggiunge un valore di 163.57, indice effettivo di una forte dispersione di dati all'interno del campione. Considerando anche il coefficiente di variazione rispetto alla media:

$$CV = \frac{s}{|\overline{x}|}.$$

```
> cv(dati$`Binge drinking`)
[1] 0.8968875
```

Abbiamo un valore minore di 1, che in termini statistici, indica che la media campionaria, è un buon indice di valutazione per il campione analizzato, permettendoci quindi di dare ulteriore supporto alle cose dedotte fino a questo punto.

Effettivamente, si è dimostrato utilizzare la media come valore di riferimento, potrebbe diminuire il tasso di pericolosità che si cela dietro al campione in esame, e soprattutto è possibile affermare che i dati più distanti dalla media (meglio evidenziati nel diagramma a barre), devono essere "tenuti sotto controllo", al fine di migliorare l'andamento complessivo della media nazionale.

Essendo che gli scarti dalla media contribuiscono in primo luogo al calcolo degli indicatori di dispersione, questi valori diminuiscono se i valori critici prima individuati, assumessero valori più simili alla media attuale?

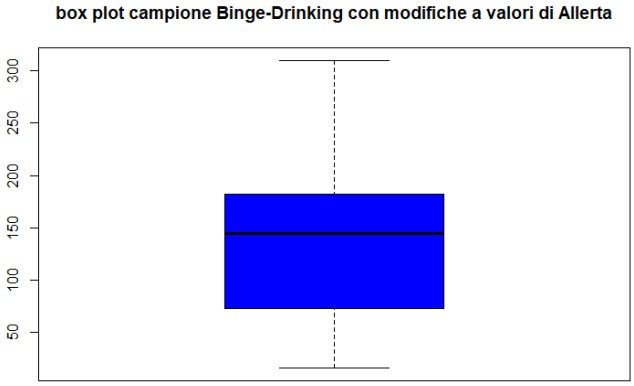
Supponendo di sostituire i valori critici delle 5 regioni con il valore medio (valore effettivo – scarto dalla media), il campione assumerebbe una forma del tipo:

Regioni	Binge drinking	Scarto dalla media	Nuovo Valore Ideale	Vettore Ideale
Valle d'Aosta/Vallée d'Aoste	16	-166		16
Molise	25	-157		25
Basilicata	41	-141		41
Umbria	53	-129		53
Trento	56	-126		56
Bolzano/Bozen	73	-109		73
Marche	83	-99		83
Liguria	96	-86		96
Abruzzo	110	-72		110
Friuli-Venezia Giulia	110	-72		110
Calabria	122	-60		122
Trentino-Alto Adige	129	-53		129
Sardegna	167	-15		167
Campania	197	15		197
Puglia	197	15		197
Sicilia	211	29		211
Toscana	223	41		223
Lazio	310	**128**	182	182
Piemonte	311	**129**	182	182
Emilia-Romagna	316	**134**	182	182
Veneto	394	**212**	182	182
Lombardia	719	**537**	182	182

Se ora si ricalcolano nuovamente i valori di media campionaria, varianza e deviazione standard su quello che è stato definito Vettore Ideale.

```
> IdealVector
 [1]  16  25  41  53  56  73  83  96 110 110 122 167 197 197 211 223 310 182 182 182 182 182
> summary(IdealVector)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   16.0    75.5   144.5   136.4   182.0   310.0
> var(IdealVector)
[1] 5781.481
> sd(IdealVector)
[1] 76.03605
```

Si ottiene che il valore di media effettivamente varia addirittura diminuendo, ma cosa più importante è la diminuzione dei valori di varianza e deviazione standard, i quali rispetto al caso reale hanno subito una forte diminuzione. Quindi nel complesso per il campione "Ideale", si crea meno dispersione di dati, portando il fattore di pericolo in una situazione migliore rispetto a quella analizzata dai dati ISTAT.

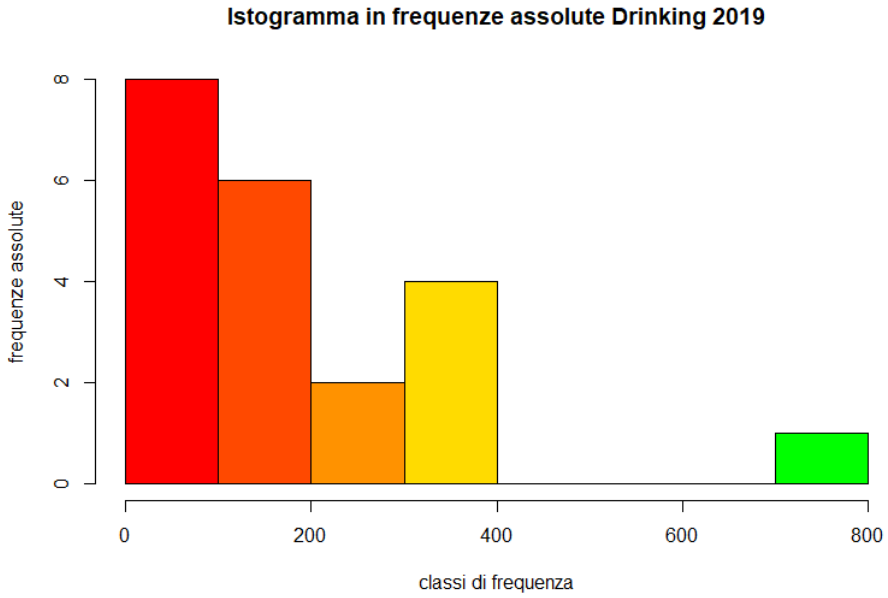


Considerando anche il box plot realizzato con i valori del vettore ideale, si nota ancora di più come la dispersione dei dati sia diminuita rispetto al caso ideale, addirittura non segnalando la presenza di valori anomali, al suo interno.

Passo 5 – Analisi di simmetrie e concentrazione di dati nel campione

Come ultimo passo di studio nell'analisi univariata rispetto al dato Binge Drinking considerato fino a questo momento, è senz'altro molto utile fare anche qualche considerazione aggiuntiva sulla forma che i dati assumono quando vengono disposti all'interno di un grafico, quindi in qualche modo andare ad analizzare le caratteristiche della loro distribuzione ordinata.

Se si considera l'istogramma precedentemente realizzato, si può notare che i dati (per la presenza di valori molto elevati), hanno una forte asimmetria verso destra, fattore che è stato anche confermato in precedenza considerando il confronto tra media e mediana campionaria.



Ma effettivamente di quanto questa asimmetria è forte nel campione in analisi. A tale scopo, è senz'altro doveroso calcolare il valore di skewness campionaria per il vettore, quantizzando effettivamente quanto la dispersione dei dati verso destra è forte (e quindi di quanto i dati estremi, creano dispersione rispetto alla centralità del campione).

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

Dall'analisi della skewness campionaria, quindi dal confronto dei momenti centrali di secondo e di terzo ordine:

```
skw <- function (x){
   n <-length (x)
    m2 <-(n -1) *var (x)/n
    m3 <- (sum ( (x-mean(x))^3) )/n
    m3/(m2 ^1.5)
}

skw(dati$`Binge drinking`)
```

```
> skw(dati$`Binge drinking`)
[1] 1.770653
```

Si ottiene un valore di molto superiore ad uno, che non solo dimostra formalmente che il campione ha un'asimmetria verso destra, ma soprattutto essa è anche molto forte, quindi i dati superiore alla media campionaria, ed in particolar modo i 5 rilementi critici, analizzati nella fase precedente, creano una forte dispersione di dati verso l'alto (Ulteriore prova della loro pericolosità ai fini dell'analisi condotta).

Mettendo poi a confronto i momenti di secondo e quarto ordine del campione in analisi, è possibile anche considerare se ci sono picchi elevati di dati rispetto alla norma.

In particolare avendo come metro di riferimento la curva di distribuzione Normale che in statistica è un ottimo indice di simmetricità, è possibile calcolare quello che viene definito come valore di curtosi campionaria:

$$\gamma_2 = \beta_2 - 3,$$

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

```
> curt <- function (x){
+     n<-length (x)
+     m2 <-(n -1) *var (x)/n
+     m4 <- (sum ( (x- mean(x))^4) )/n
+     m4/(m2 ^2) -3
+  }
> |
```

```
> curt(dati$`Binge drinking`)
[1] 3.487394
```

che si attesta ad un valore molto superiore allo 0 per il campione in analisi, quindi effettivamente è presente una forte piccatezza dei dati nel campione in analisi, il che è senz'altro da riattribuire al valore anomalo Lombardo, che non solo supera la media, ma come già considerato più volte ha uno scarto dalla media campioanaria molto elevato rispetto al resto del campione.

Analisi Bivariata – Confronti tra diverse colonne

Dopo aver condotto una prima analisi statistica su un singolo parametro del dataset di riferimento, è senz'altro interessante procedere con lo studio combinato del dataset, per capire se tra le colonne del dataset fornito dall'ISTAT, ci sono una o più relazioni di dipendenza. Mettere a confronto colonne differenti del dataset in esame, può essere molto utile per capire in ogni regione, come si rapportano tra di loro differenti categorie di consumatori di bevande alcoliche in Italia.

Passo 1 – confronto tra Consumi Moderati e Binge Drinking

Come primo esempio è molto interessante osservare le colonne di consumi moderati e di consumatori pericolosamente eccessivi del dataset iniziale.

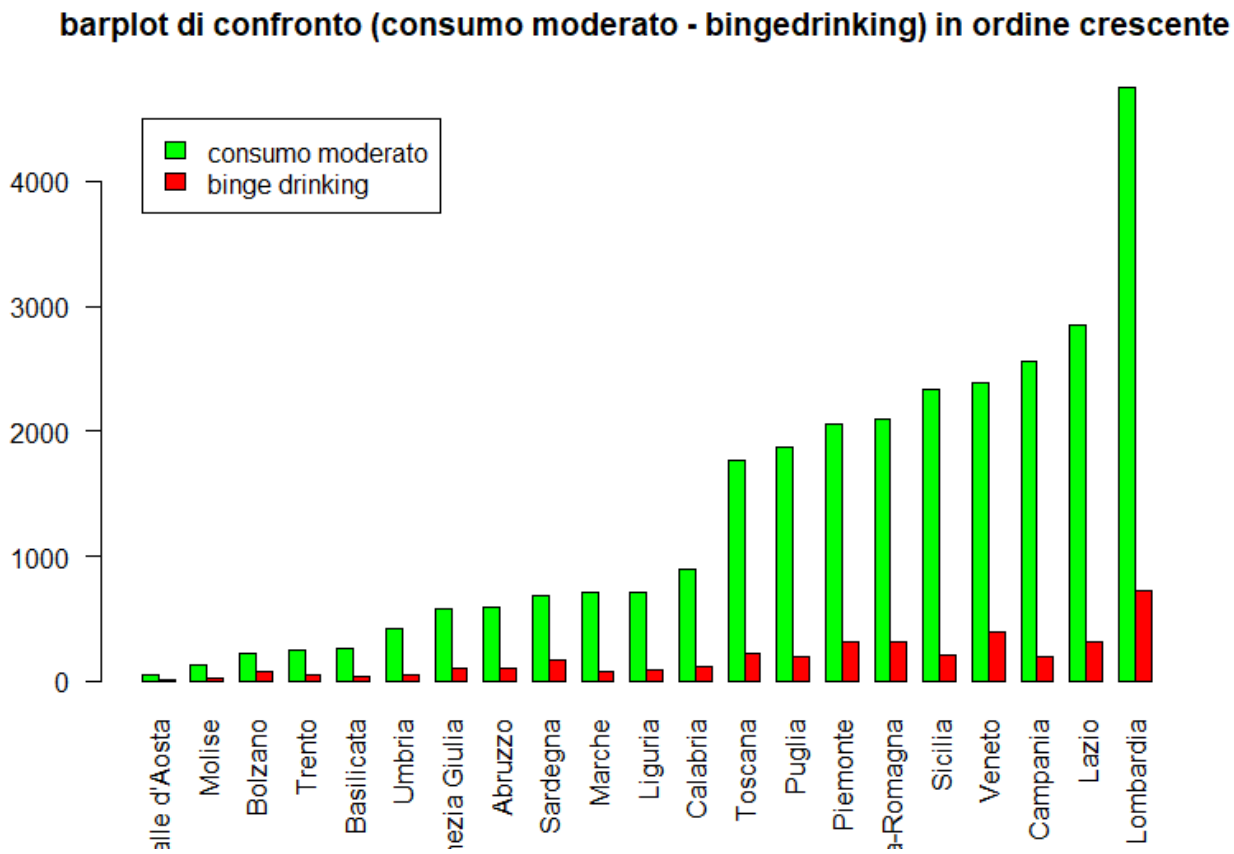
Considerando queste due colonne, risulta facilmente osservabile, come (per la maggioranza dei dati) i dati ordinati, rispetto all'indice di consumo moderato, rispettino pressoché lo stesso ordine anche per la colonna di Binge drinking (a meno di alcuni valori in controtendenza).

Regioni	Consumo moderato	Binge drinking
Valle d'Aosta	54	16
Molise	138	25
Bolzano	225	73
Trento	244	56
Basilicata	264	41
Umbria	418	53
Friuli-Venezia Giulia	578	110
Abruzzo	596	110
Sardegna	683	167
Marche	708	83
Liguria	715	96
Calabria	903	122
Toscana	1.769	223
Puglia	1.869	197
Piemonte	2.055	311
Emilia-Romagna	2.091	316
Sicilia	2.330	211
Veneto	2.390	394
Campania	2.565	197
Lazio	2.847	310
Lombardia	4.746	719

Data questa osservazione, è facile chiedersi se tra di loro effettivamente possa esistere un qualche tipo di correlazione lineare crescente.

- Ma in che misura potrebbe essere veritiera questa deduzione?
- Che informazioni si potrebbero dedurre da un confronto tra le due colonne?

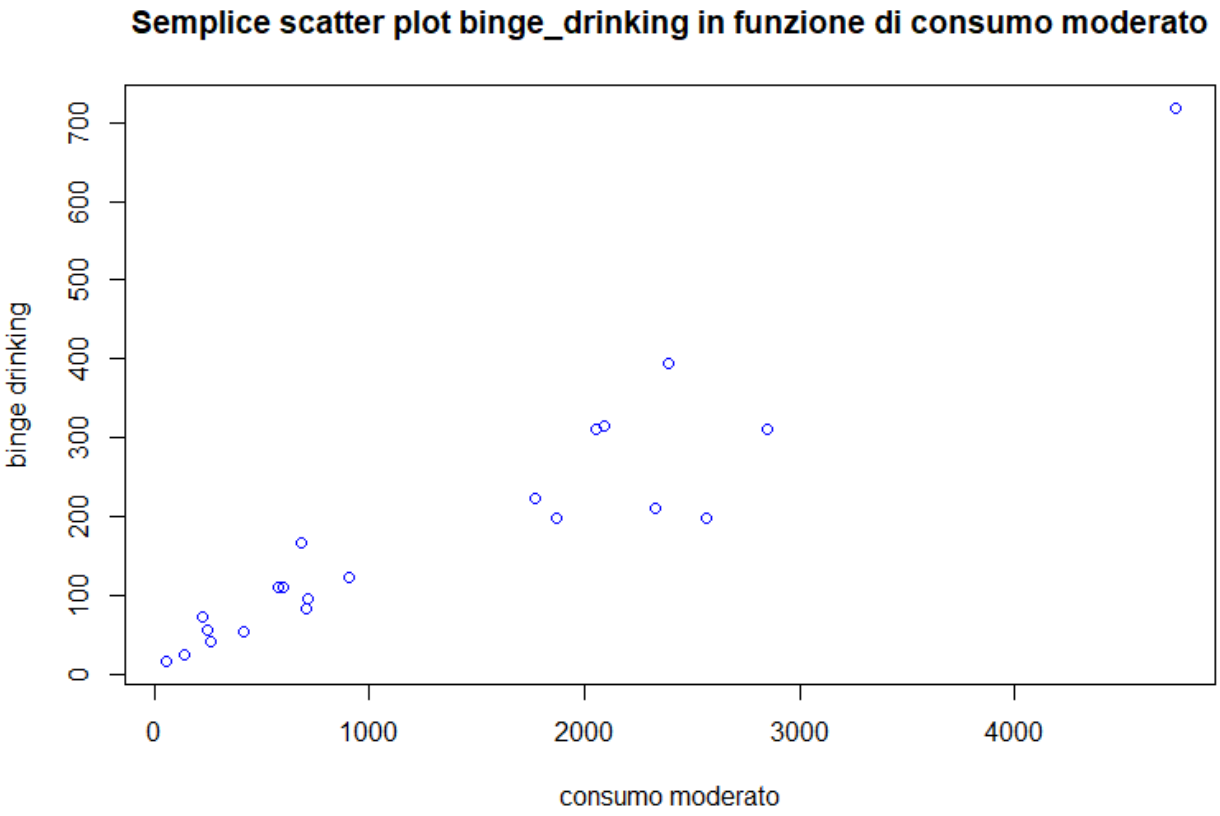
Al fine di verificare quest'osservazione, è senz'altro utile visualizzare i dati in maniera diversa, ad esempio tramite il meccanismo del Barplot che si è rilevato molto utile all'interno dell'analisi univariata.



Da questo diagramma di confronto (anche se in modo molto tenue, data l'enorme differenza di dati tra il consumo moderato e il binge drinking), si può già dedurre come nel 2019 in Italia, in numero di consumatori di bevande alcoliche in modo eccessivamente pericoloso siano progressivamente di più in quelle regioni dove il numero di consumatori moderati è contestualmente più elevato, a meno di alcune piccole differenze dove appunto è il dato è in controtendenza.

Se si osserva il diagramma a barre si nota come alcune regioni di spicco siano particolarmente interessanti, come ad esempio la Lombardia, che sembra avere il tasso di crescita più alto rispetto al resto delle regioni italiane, mentre la Campania invece si pone quasi all'estremo opposto, avendo sì uno dei valori più alti per consumo moderato, ma in proporzione a questo valore, il dato di consumo eccessivo sembra essere quello più basso.

Al fine di avere le idee un pochino più chiare rispetto a questa considerazione, è senz'altro interessante creare uno Scatter Plot dove la variabile indipendente in questo caso è il consumo moderato di alcol, e di conseguenza verificare se e come il consumo eccessivo dipenda da esso.



Già da questo primo e semplice scatterplot, si può notare come il consumo moderato di alcool, sia effettivamente in costante crescita verso l'alto rispetto, quindi è plausibile presumere che tra queste due variabili, ci possa essere un fattore di dipendenza linearmente crescente.

Come verificare effettivamente questo fattore di Correlazione?

Nell'analisi statistica in due variabili, esistono due indici molto importanti per il calcolo della correlazione tra due variabili quantitative, la covarianza e il coefficiente di correlazione, in particolare questi strumenti sono molto utili, per verificare se due variabili siano tra di loro linearmente dipendenti e quanto sia rilevante la loro dipendenza lineare nel caso esista.

Si definisce covarianza tra due variabili quantitative:

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y}) \qquad (n = 2, 3, \ldots).$$

E per il confronto binge drinking in funzione di consumo moderato, essa assume valore:

```
> cov(orderedDF$`consumo moderato`, orderedDF$`binge drinking`)
[1] 185163.9
```

Effettivamente dal calcolo della covarianza, si ottiene un valore positivo, ma arbitrariamente molto alto, al fine di ottenere altre informazioni utili. Teoricamente esso però ci indica che le due variabili sono tra di loro positivamente correlate (valore di covarianza non nullo e maggiore di 0).

Al fine di ottenere qualche informazione in più sul grado di correlazione lineare tra le due variabili è senz'altro più significativo il calcolo del coefficiente di correlazione, che in termini teorici risulta essere, il rapporto tra la covarianza delle due variabili fratto il prodotto tra le due dispersioni standard.

Coefficiente di Correlazione

$$r_{xy} = \frac{C_{xy}}{s_x\, s_y}.$$

```
> cor(orderedDF$`consumo moderato`, orderedDF$`binge drinking`);
[1] 0.9385731
```

Si nota che il valore applicato ai due campioni di dati, non solo è positivo, ma tra l'altro è molto vicino al valore 1, quindi è senz'altro necessario approfondire la conoscenza di questo legame.

Passo 2 – Regressione lineare semplice

Essendo il coefficiente di correlazione tra le variabili in esame compreso tra 0 e 1, si ha che i punti di intersezione tra le due variabili sono disposti attorno ad una retta che rappresenta al meglio la loro correlazione lineare crescente.

- Ma quale retta meglio interpola i dati messi a confronto?
- Che relazione c'è tra gl'indici calcolati e la retta stessa?

In termini teorici, l'indice di correlazione è strettamente legato ad una retta di proporzionalità così formata:

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

dove

- α è l'intercetta;
- β è il coefficiente angolare.

Dove alfa e beta sono due indici, che rappresentano rispettivamente l'ordinata del punto di intersezione della retta di correlazione con l'asse delle ordinate e la retta stessa e il coefficiente angolare della retta.

Essi sono chiamati coefficienti di regressione, e sono calcolati applicando ai campioni il teorema dei minimi quadrati. Ma da questo calcolo, si dimostra che alfa e beta sono calcolabili come:

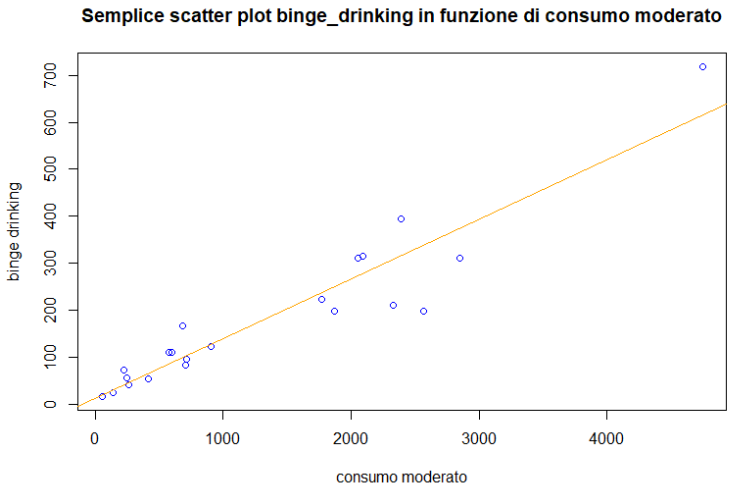
$$\beta = \frac{s_y}{s_x} r_{xy}, \qquad \alpha = \overline{y} - \beta\, \overline{x}.$$

E quindi strettamente correlati al coefficiente di correlazione tra le due variabili.

Avendo calcolato questo valore, possiamo quantificare in R anche i valori di alfa e beta per i campioni in esame.

```
> #coefficienti di dispersione
> beta <-(sd(df$`binge drinking`)/sd(df$`consumo moderato`))*cor (df$`consumo moderato` ,df$`binge drinking` )
> alpha <-mean (df$`binge drinking` )-beta*mean (df$`consumo moderato` )
>  c(alpha , beta)
[1] 11.513314  0.127296
>
> lm <- lm(df$`binge drinking`~ df$`consumo moderato`)
> lm

Call:
lm(formula = df$`binge drinking` ~ df$`consumo moderato`)

Coefficients:
          (Intercept)  df$`consumo moderato`
              11.5133                 0.1273
```

In R questi valori sono automaticamente calcolati, considerando il modello lineare (lm) tra le due variabili.



Conoscendo ora più in dettaglio come derivare la retta di interpolazione più consona alle due variabili, è possibile inserirla all'interno del grafico precedentemente realizzato.

Da questo nuovo Scatterplot, si osserva facilmente come molti punti (soprattutto iniziali) non si discostino di molto rispetto al modello ideale. E soprattutto è importante considerare come la maggior parte dei punti sia posta sul lato superiore della retta.

Questo studio appena effettuato conferma l'ipotesi iniziale, e si può senz'altro sostenere che secondo l'ISTAT, nel 2019 in Italia, all'aumentare progressivo del consumo moderato di alcol, sussiste anche un progressivo aumento del consumo eccessivo, e senz'altro la progressione lineare tra le due variabili, è un buon indice di approssimazione tra le due variabili.

Passo 3 – Analisi dei residui

Come appena visto, il diagramma fornito, conferma il tipo di relazione supposto tra le due variabili analizzate, però si nota anche come alcuni valori siano ben distanti dalla retta, e queste distanze senz'altro possono essere di buon interesse al fine di dare qualche informazione in più in merito alla comparazione lineare precedentemente affrontata.

Nella teoria dell'analisi bivariata, si definiscono valori ideali (o valori stimati), i cooefficienti di ordinata:

$$\hat{y}_i = \alpha + \beta\, x_i$$

, che appunto rappresentano il valore di ordinata, per ogni valore Xi della variabile indipendente della relazione lineare.

Il differenziale tra il valore effettivo Yi e il valore ideale, viene definito Residuo del punto iesimo delle coppie di campioni.

Ogni residuo Ei, viene calcolato come:

$$E_i = y_i - \widehat{y}_i = y_i - (\alpha + \beta\, x_i) \qquad (i = 1, 2, \ldots, n)$$

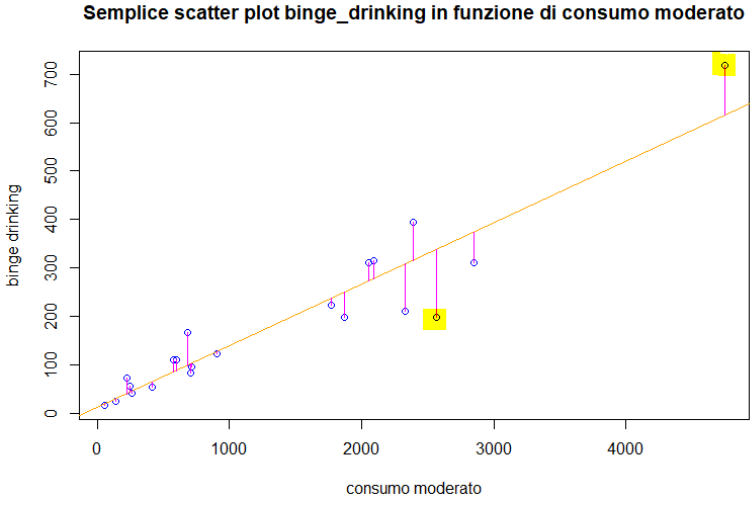
E graficamente rappresentano il segmento di distanza tra il punto reale (xi, yi) e il corrispettivo valore ideale sulla retta.

```
> residui
        1         2         3         4         5         6         7         8         9        10        11
37.893351 -2.387300 -6.529973 103.339743 32.845080 13.426455 78.249182 24.909582 38.310694 -13.699985 -11.723053
       12        13        14        15        16        17        18        19        20        21
-18.638901 -63.925102 22.618254 -4.080166 -141.027623 -52.429588 -4.119465 -4.461626 -97.113056 68.543500
```

```
> data.frame(lm$residuals, row.names = dati$Regioni)
                lm.residuals
Piemonte          37.893351
Valle d'Aosta     -2.387300
Liguria           -6.529973
Lombardia        103.339743
Bolzano           32.845080
Trento            13.426455
Veneto            78.249182
Friuli-Venezia Giulia 24.909582
Emilia-Romagna    38.310694
Toscana          -13.699985
Umbria           -11.723053
Marche           -18.638901
Lazio            -63.925102
Abruzzo           22.618254
Molise            -4.080166
Campania        -141.027623
Puglia           -52.429588
Basilicata        -4.119465
Calabria          -4.461626
Sicilia          -97.113056
Sardegna          68.543500
```

Mettendo a confronto i residui così definiti, con le regioni corrispondenti, si nota come una supposizione iniziale di quest'analisi in due variabili sia effettivamente vera.

In particolare, è utile osservare quanto i valori di Lombardia e Campania si discostino dall'andamento generale, la prima infatti ha un andamento che supera di molto la linearità rispetto agli altri valori (quasi ad assumere un comportamento esponenziale), mentre la seconda, assume un comportamento quasi in controtendenza, pur avendo dei valori effettivi è abbastanza alti rispetto alla media nazionale, quasi a voler simboleggiare un andamento del tutto opposto. In altri termini quasi possibile affermare che in Campania, è quasi stato riscontrato un andamento decrescente per il consumo eccessivo di alcool rispetto al consumo moderato, ma al fine di approfondire questo aspetto, sarebbe utile utilizzare un dataset di tipo regionale.



Considerando anche il diagramma precedente costruito con l'aggiunta dei segmenti residui tra i punti effettivi, è facile notare come per molte regioni il valore il valore di proporzionalità si discosti dalla linea di proporzionalità lineare, in particolare i valori evidenziati (Lombardia e Campania), risultano essere i valori con residui più rilevanti, il primo verso l'alto (quindi ancora una volta forse il più estremo nella pienezza dei dati fin ora analizzati), mentre il secondo (il

valore campano) detiene l'estremo opposto, validando maggiormente quanto affermato al paragrafo precedente.

Dall'analisi dei residui portata avanti, si osserva come alcuni valori effettivi del campione, si discostino anche di parecchio rispetto al valore atteso. In particolare, il caso Lombardo e Campano si distaccano quasi in modo estremo dal valore atteso.

Per meglio identificare questo andamento, è utile ricorrere al diagramma dei Residui Standard in relazione del valore atteso.

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

Il calcolo di ogni residuo standard avviene dividendo l'iesimo Residuo della correlazione con la deviazione standard dei residui (la sottrazione con la media campionaria dei residui è trascurabile dato che per Correlazioni di tipo lineare essa vale sempre 0). Ai fini del diagramma da realizzare essi possono sicuramente essere più significativi dei residui generici precedentemente visti (dato che si porranno in un intervallo di analisi ridotto).

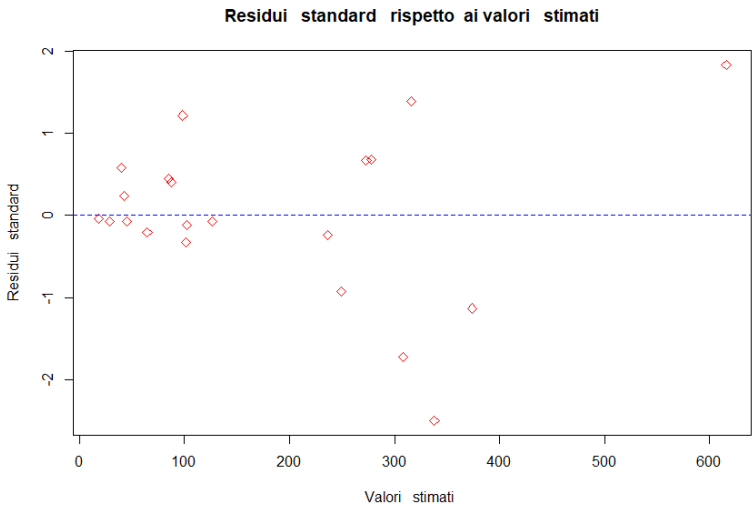
```
> residui <-resid (lm)
> residuistandard <-residui / sd( residui )
> residuistandard
          1           2           3           4           5           6           7           8           9          10
 0.67131304 -0.04229305 -0.11568405  1.83075172  0.58187861  0.23786111  1.38625102  0.44129450  0.67870663 -0.24270693
         11          12          13          14          15          16          17          18          19          20
-0.20768389 -0.33020404 -1.13248773  0.40070167 -0.07228362 -2.49842467 -0.92883489 -0.07297984 -0.07904152 -1.72044066
         21
 1.21430658
```

	dati.Regioni	residuistandard
1	Piemonte	0.67131304
2	Valle d'Aosta	-0.04229305
3	Liguria	-0.11568405
4	Lombardia	1.83075172
5	Bolzano	0.58187861
6	Trento	0.23786111
7	Veneto	1.38625102
8	Friuli-Venezia Giulia	0.44129450
9	Emilia-Romagna	0.67870663
10	Toscana	-0.24270693
11	Umbria	-0.20768389
12	Marche	-0.33020404
13	Lazio	-1.13248773
14	Abruzzo	0.40070167
15	Molise	-0.07228362
16	Campania	-2.49842467
17	Puglia	-0.92883489
18	Basilicata	-0.07297984
19	Calabria	-0.07904152
20	Sicilia	-1.72044066
21	Sardegna	1.21430658

Sapendo che i residui standard, rappresentano in forma standardizzata, la distanza tra il valore reale e il valore atteso, possiamo senz'altro affermare che l'analisi di regressione lineare effettuata sicuramente interpola al meglio i dati analizzati, in quanto se pur con diverse eccezioni, molti di questi valori si approssima allo 0.

Ciò appunto significa che per le variabili considerate la regressione di tipo lineare positiva individuata è quella che interpola meglio la relazione "Binge Drinking" in funzione del dato di consumo moderato.

Come è possibile osservare anche dal grafico realizzato, infatti la maggior parte dei residui standardizzati si concentra tra i valori 1 e -1, ciò significa che nel grafico precedentemente analizzato quindi molti valori riportati rispettano pressappoco lo stesso andamento dei loro corrispettivi valori attesi. A far eccezione ci sono alcuni dati che nuovamente anche in questo caso possono risultare anomali dato che sono posti al di fuori dell'intervallo [-1, 1], in particolare si nota come il valore Lombardo e Campano possono essere quasi considerati in controtendenza, dato che (come visibile dalla tabella soprariportata), essi hanno valori molto sfasati rispetto al resto del campione. In particolare fatto che il dato campano si approssimi al valore -2.5, è un ulteriore conferma del comportamento in controtendenza precedentemente emerso (Ciò motiverebbe ancora di più un analisi più approfondita a livello locale).



Analisi Multivariata – Binge Drinking in funzione di più campi del dataset

Quando in statistica si cerca di mettere in relazione più variabili tra di loro, potrebbe capitare che non basti analizzare una singola variabile indipendente al fine di trovare il modo migliore di rappresentare le dipendenze tra i dati. A tal proposito interviene il coefficiente di Determinazione, definito nel caso di analisi lineari come:

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

la varianza tra i valori attesi fratto la varianza dei valori effettivi (si ricorda che per modelli lineari, la media campionaria dei valori attesi e quella dei valori effettivi coincide).

Se si considera lo studio in due variabili (binge drinking in funzione di consumo moderato), che in R si ottiene tramite il parametro r.square del modello lineare. Si ottiene

```
> summary (lm(df$`binge drinking`~df$`consumo moderato`))$r.square
[1] 0.8809194
```

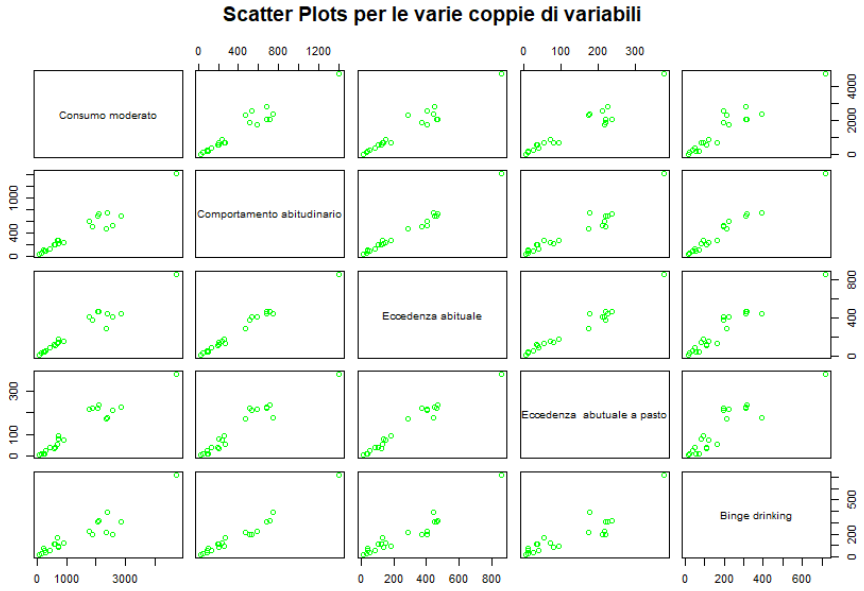
Un valore di 0.88, sapendo che il coefficiente di determinazione è un valore che intervalla tra 0 e 1, è automatico chiedersi, se l'analisi bivariata con il consumo moderato sia esaustiva per rappresentare al meglio le dipendenze del tasso pericoloso Binge Drinking.

- *Per caso il coefficiente di determinazione può aumentare ulteriormente?*
- *Per caso altre colonne del dataset possono influire nella rappresentazione delle dipendenze della colonna Binge Drinking?*

Al fine di determinare questo fattore, ed in particolare determinare quali altre variabili potrebbero essere utili a rappresentare un coefficiente di correlazione più simile ad 1, è utile mettere a confronto lo studio bivariato precedentemente fatto con altre colonne del dataset originale.

Passo 1 – Confronto delle singole dipendenze e modello multivariato

Al fine di verificare se ci sono altre dipendenze tra la Variabile Binge Drinkig e le altre variabili del dataset, si procede con il confrontare tutte le possibili combinazioni di Scatter Plot tra le colonne presenti nel dataset.



Considerando tutte le possibili combinazioni, si nota come il consumo di alcol in maniera eccessivamente pericolosa (posto come variabile indipendente – ultima riga dei diagrammi in figura), presenti (con tutte le altre variabili in esame), una dipendenza simile all'andamento lineare studiato per il caso precedente con Consumo Modetato come variabile indipendente.

Infatti, se si considera Binge Drinking in funzione, ad esempio, della seconda variabile "consumo abituinario" oppure della terza "eccedenza abitudinaria", si nota come anche in questo caso siano presenti delle regressioni lineari molto simili a quella studiata nella precedente analisi.

- *Quindi oltre "consumo moderato", anche "consumo abitudinario" e "eccedenza abitudinaria" sono due variabili che influenzano l'andamento di "Binge Drinking"?*

Al fine di verificare questa proprietà, sarebbe utile andare a mettere in relazione queste tre variabili secondo una relazione polinomiale del tipo:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$

Che in statistica mette in relazione una singola variabile dipendente con più variabili indipendenti.

Dove alfa è detto anche in questo caso valore "intercetta" ossia il valore di Y quando X1 = X2 = . . . = Xp = 0;

e i valori Beta rappresentano in questo caso i Regressori, che anche in questo caso rappresentano l'inclinazione di Y, rispetto alla variabile Xi corrispondente, quando tutte le altre sono considerate costanti.

```
Call:
lm(formula = dati$`Binge drinking` ~ dati$`Consumo moderato` +
    dati$`Comportamento abitudinario` + dati$`Eccedenza abituale`)

Coefficients:
                    (Intercept)           dati$`Consumo moderato`  dati$`Comportamento abitudinario`
                      3.030e-01                        -1.223e-05                          1.118e+00
      dati$`Eccedenza abituale`
                     -1.006e+00
```

Verificando questa tipologia di regressione multipla con R, effettivamente si nota come i dati di eccedenza abitudinaria, e di comportamento abituale, modifichino l'andamento complessivo della variabile Binge Drinking, dato che il loro Regressore corrispondente nel modello lineare, assume valore differente da 0 e in particolare in ogni caso supera 1 o -1.

Al fine di determinare quanto effettivamente questa tipologia di relazione vada bene per rappresentare le dipendenze della variabile Binge Drinking, nella loro completezza, è interessante verificare che valore assume il coefficiente di determinazione, per questo nuovo modello lineare multiplo in esame. Che anche nel caso di correlazione lineare multipla si calcola come:

$$D^2 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

```
> summary(lm2)$r.square
[1] 0.995926
```

Si nota come questo valore, assuma un valore molto simile ad 1, quindi effettivamente, il modello lineare multiplo identificato, rappresenta al meglio le dipendenze che la variabile Binge Drinking possiede, rispetto alle altre variabili del dataset.

Essendo tale modello basato su più variabili dipendenti, non è ovviamente possibile confrontare i dati tramite una retta di regressione. Ma in che modo variano i valori effettivi calcolabili da questa relazione, rispetto ai valori reali di Binge Drinking?. Per fare ciò è utile considerare nuovamente sia i valori attesi rispetto a quelli effettivi della variabile, ed effettuare una nuova analisi dei residui tra i due valori, sia normali che standardizzati.

Passo 2 – Analisi dei residui del modello lineare multivariato

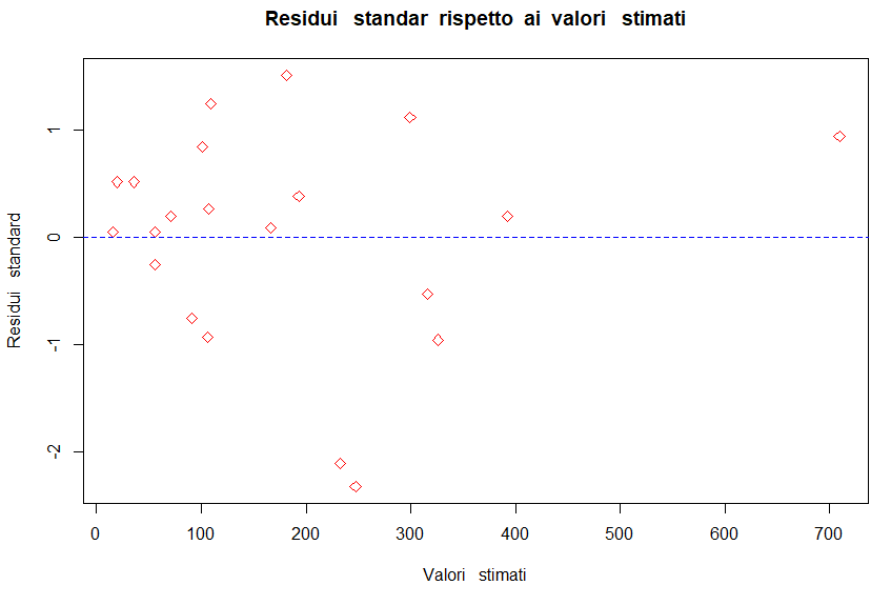
Considerando il dato rilevato dal calcolo del coefficiente di determinazione, ci si aspetta che i valori reali della variabile binge drinking siano se non uguali, molto simili a quelli calcolati dal modello lineare multivariato preso in esame.

	dati.Regioni	dati..Binge.drinking.	fitted.lm2.
1	Piemonte	311	299.38645
2	Valle d'Aosta	16	15.51711
3	Liguria	96	105.71295
4	Lombardia	719	709.24056
5	Bolzano	73	70.98891
6	Trento	56	55.55880
7	Veneto	394	391.95685
8	Friuli-Venezia Giulia	110	101.21339
9	Emilia-Romagna	316	326.00240
10	Toscana	223	247.25632
11	Umbria	53	55.68799
12	Marche	83	90.93314
13	Lazio	310	315.58318
14	Abruzzo	110	107.24371
15	Molise	25	19.66142
16	Campania	197	181.27420
17	Puglia	197	193.00692
18	Basilicata	41	35.66073
19	Calabria	122	109.05091
20	Sicilia	211	232.99298
21	Sardegna	167	166.07106

Mettendo questi due dati a confronto si nota in effetti un'enorme somiglianza tra i due vettori. Ciò è un primo step che dimostra effettivamente la veridicità, di quanto affermato rispetto al modello multivariato in esame.

È quasi possibile affermare che le dipendenze del valore di binge drinking 2019 in Italia, siano Linearmente crescenti al crescere delle tre variabili considerate fino a questo punto: Consumo moderato, consumo abitudinario e eccedenza abitudinaria.

Come già anticipato essendo 3 le variabili indipendenti, non si può procedere alla realizzazione della retta d'interpolazione, dato che le variabili indipendenti in questo caso sono 3. Ma cosa accade se consideriamo i residui standardizzati?



Se si fa riferimento ai residui, si nota come molti più valori ricadano nell'intervallo [-1, 1], però qualche divergenza con quanto affermato fino a questo momento è maggiormente analizzabile. Il che fa concludere, che per quanto questo modello multivariato sia un buon indice di stima per i valori raccolti, non è del tutto possibile approssimare allo zero questa tipologia di errori, però senz'altro esso fornisce uno strumento più completo per la rappresentazione delle dipendenze tra i dati considerati in forma molto simile al contesto reale.

Passo 3 – Considerazione finale analisi multivariata

Dati i risultati dell'analisi multivariata appena effettuata, sarebbe facile concludere dicendo che le dipendenze della variabile Binge Drinking siano pienamente soddisfatte dalle 3 variabili indipendenti considerate.

```
Call:
lm(formula = dati$`Binge drinking` ~ dati$`Consumo moderato` +
    dati$`Comportamento abitudinario` + dati$`Eccedenza abituale`)

Coefficients:
                      (Intercept)           dati$`Consumo moderato`  dati$`Comportamento abitudinario`
                        3.030e-01                        -1.223e-05                          1.118e+00
       dati$`Eccedenza abituale`
                       -1.006e+00
```

Nell'analisi statistica multivariata, però è anche utile considerare il peso che ogni variabile indipendente fornisce nella relazione che si è appena realizzata. Un buon metro di stima per questa cosa sono appunto i regressori (beta) visibili dal modello lineare, considerando quelli del modello lineare utilizzato, si nota come il fattore Beta del valore di consumo moderato, incida molto di meno rispetto ai valori di regressione di eccedenza abituale e di comportamento abitudinario, dato che appunto questo valore è molto più piccolo degli altri due.

Secondo tale osservazione, il legame tra Binge Drinking e Consumo Moderato osservato è molto meno forte rispetto agli altri legami analizzati nel Dataset. Al fine di confermare questa ipotesi, si è provato a realizzare

un ulteriore modello lineare in assenza della variabile indipendente consumo moderato.

```
> lm3 <- lm(dati$`Binge drinking`~dati$`Eccedenza abituale`+dati$`Comportamento abitudinario`)
> lm3

Call:
lm(formula = dati$`Binge drinking` ~ dati$`Eccedenza abituale` +
    dati$`Comportamento abitudinario`)

Coefficients:
                      (Intercept)          dati$`Eccedenza abituale`  dati$`Comportamento abitudinario`
                           0.3032                            -1.0059                              1.1182

> summary(lm3)$r.square
[1] 0.995926
```

Da questo nuovo modello, si ottiene un coefficiente di correlazione pressoché identico rispetto a quello analizzato per il modello lineare precedente. A fronte di questo dato, si evince come nel modello realizzato ed analizzato, sia effettivamente superflua la presenza della variabile indipendente Consumo Moderato, al fine di rappresentare le dipendenze della variabile dipendente.

Concludendo, ciò comunque non invalida il lavoro di analisi fatto per l'analisi bivariata e quella multivariata, che comunque restano un ottimo metro di stima, però a ciò è importante aggiungere, il differente peso che la variabile consumo moderato riporta, rispetto alle altre variabili considerate.

Analisi dei Cluster

Nel corso dei precedenti capitoli del documento, il dataset di riferimento è stato osservato ed analizzato, sempre in funzione delle varie caratteristiche, ed in particolare l'ottica di analisi è stata rivolta al parametro di Binge Drinking o tasso alcolemico eccessivo pericoloso, analizzandolo in prima battuta con indici di centralità e dispersione, al fine di osservare come questo tasso possa essere "normalizzato", intervenendo su altri elementi del campione. Poi con l'analisi bivariata e multivariata, si è voluto osservare come questo valore dipendesse strettamente da altre variabili del dataset, in particolare si è osservato che all'aumentare di parametri come "comportamento abitudinario" ed "eccedenza abituale", anche il tasso pericoloso ne risentisse in maniera forte. Durante queste analisi però, sono saltate all'occhio anche alcune caratteristiche di interesse per le singole regioni, ad esempio l'anomalia del valore Lombardo, che da sempre all'interno delle varie analisi è risultata molto differente rispetto alle altre.

Questo incipit, è senz'altro un buon punto di partenza per porsi un altro quesito di studio:

Nel dataset relativo ai consumi alcolemici in Italia del 2019 esistono regioni tra di loro simili in base alle caratteristiche che ci sono a disposizione?

Dare una risposta a questo quesito senza i dovuti strumenti, non è un'operazione semplice, in quanto, per ogni regione, influiscono ben 5 caratteristiche ben distinte. Per ovviare a questa problematica però in statistica si può far ricorso all'analisi dei cluster, che ha proprio il compito di raggruppare un insieme di elementi con una o più caratteristiche comuni (come ad esempio, differenti variabili di un dataset), secondo quelli che vengono definiti misure di similarità o distanza statistica tra individui.

Passo 1 – Misure di Similarità o di Distanza?

Dato un insieme di caratteristiche, e due individui a cui corrisponde un valore numerico quantitativo per ognuna delle caratteristiche, l'analisi dei cluster fornisce due tipologie di misurazioni che permettono di confrontare i due individui rispetto alle caratteristiche comuni:

- Le misure di similarità, che indicano analiticamente quanto due individui siano tra di loro simili, in particolare questa tipologia di indice varia tra 0 e 1 e indica che la forza del legame di similarità tra due individui (persone, regioni, ecc.) è più forte, quanto più il valore del loro indice di similarità si avvicina al valore 1;

Le misure di similarità sono soggette alle seguenti proprietà:

$$(i)\ s(X_i, X_i) = 1;$$
$$(ii)\ 0 \leq s(X_i, X_j) \leq 1;$$
$$(iii)\ s(X_i, X_j) = s(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j.$$

- Le misure di distanza invece seguono il paradigma inverso (il legame di similarità tra due individui, tanto più un indice di distanza si approssima allo 0), ma a differenza delle misure di similarità questi valori intervallano tra 0 ed infinito, quindi permettono dei confronti molto più grandi.

Le misure di distanza sono soggette alle seguenti proprietà:

$$(i)\ d(X_i, X_j) = 0\ se\ e\ solo\ se\ X_i = X_j,\ con\ X_i\ e\ X_j\ in\ E_p;$$
$$(ii)\ d(X_i, X_j) \geq 0\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$
$$(iii)\ d(X_i, X_j) = d(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$
$$(iv)\ d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)\ per\ ogni\ X_i,\ X_j\ e\ X_k\ in\ E_p.$$

Come si può osservare le misure di distanza, godono di una proprietà in più rispetto a quelle di similarità, in particolare la proprietà degli indici di distanza indicata al punto 4 detta disuguaglianza triangolare, indica il concetto di distanza, anche a distanze combinate tramite delle addizioni (fattore molto importate nei successivi paragrafi, perché sarà molto utile per alcuni metodi di divisione in cluster). A tal proposito quando si parla di analisi dei cluster, si preferisce utilizzare le misure di distanza, al posto delle misure di similarità, in quanto più omogenee nei calcoli.

È importante anche ricordare che qualora si combinino tra di loro misure di distanza per fattori moltiplicativi, i valori risultanti non sono più considerabili come delle metriche di distanza.

Passo 2 – Metrica euclidea e Standardizzazione del Dataset

Avendo determinato l'utilizzo delle misure di distanza per l'analisi che si sta per cominciare, è buona norma considerare che in statistica esistono differenti metriche per il calcolo delle distanze, e che ognuna di esse, permette di ottenere delle misure di stima leggermente differenti tra di loro, dato l'approccio differente di calcolo.

In particolare, per l'analisi del dataset, si considererà quella che in statistica viene definita come metrica euclidea, che calcola l'indice di distanza tra due individui nel seguente modo:

$$d_2(X_i, X_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{1/2}$$

Considerando due individui, la metrica euclidea, va a sommare le differenze quadre tra i due valori per ogni singola caratteristica del Dataset considerato (corrispondenti ai due individui), elevando il tutto al quadrato, generalizzando quello che è il calcolo dell'ipotenusa con il teorema del triangolo rettangolo.

Altre metriche, come quella del valore assoluto (o di Manhattan), quella del valore massimo di Chebychev o la più generica di Minkowski, utilizzano dei procedimenti algebrici molto simili. Ma ne esistono anche di altri tipi che si basano sul peso che le singole variabili danno nel calcolo, come quella di Camberra (tutte quest'altre metriche, non sono però applicabili in tutti gli algoritimi di Clustering con la stessa versatilità della metrica euclidea).

La metrica euclidea, così come le altre metriche simili, sono molto legate alla tipologia di dati analizzati, ed in particolare, quando i dati sono legati da un'unità di misura, oppure sono comunque molto grandi, si potrebbero avere delle misurazioni di distanza inconsistenti (nel primo caso), oppure difficili da gestire (nel secondo). A tal proposito si consiglia sempre prima di procedere con l'analisi dei cluster, ad effettuare per ogni valore un'operazione di standardizzazione, sottraendo ad esso la media campionaria e dividendo per la deviazione standard del campione, tale operazione prende il nome di Scalatura di un valore rispetto alla media campionaria e alla deviazione standard.

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{s_j} \qquad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p)$$

Considerando il dataset di partenza (se bene ogni dato sia riportato in migliaia di abitanti per regione), è comunque molto utile effettuare l'operazione di scaling del dataset, dato che alcuni valori sono molto grandi.

In R è possibile effettuare quest'operazione per un intero dataset tramite la funzione scale.

Dataset prima dell'operazione di scaling

```
> dataset
                     Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
Piemonte                         2055                       684                463                       222            311
Valle d'Aosta                      54                        28                 16                         6             16
Liguria                           715                       258                182                        95             96
Lombardia                        4746                      1405                857                       377            719
Bolzano                           225                       101                 42                        12             73
Trento                            244                        89                 44                        12             56
Veneto                           2390                       747                441                       179            394
Friuli-Venezia Giulia             578                       200                122                        36            110
Emilia-Romagna                   2091                       715                471                       237            316
Toscana                          1769                       587                407                       218            223
Umbria                            418                       126                 85                        42             53
Marche                            708                       207                140                        81             83
Lazio                            2847                       685                448                       226            310
Abruzzo                           596                       191                106                        39            110
Molise                            138                        47                 33                        13             25
Campania                         2565                       528                407                       211            197
Puglia                           1869                       507                372                       222            197
Basilicata                        264                        82                 56                        27             41
Calabria                          903                       234                152                        73            122
Sicilia                          2330                       469                290                       174            211
Sardegna                          683                       267                132                        54            167
```

Dataset dopo l'operazione di scaling

```
> scaling_dataset <- scale(dataset)
> scaling_dataset
                     Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
Piemonte                   0.5909418                  0.8777305          0.9821777                 0.9543746     0.78629917
Valle d'Aosta             -1.0681727                 -1.0703306         -1.0864115                -1.1012014    -1.01715265
Liguria                   -0.5201094                 -0.3873214         -0.3182106                -0.2542280    -0.52808097
Lombardia                  2.8221649                  3.0188159          2.8054978                 2.4294406     3.28056473
Bolzano                   -0.9263893                 -0.8535494         -0.9660908                -1.0441021    -0.66868908
Trento                    -0.9106356                 -0.8891847         -0.9568354                -1.0441021    -0.77261681
Veneto                     0.8687047                  1.0648156          0.8803679                 0.5451627     1.29371104
Friuli-Venezia Giulia     -0.6337019                 -0.5595585         -0.5958735                -0.8157047    -0.44249342
Emilia-Romagna             0.6207910                  0.9697882          1.0191994                 1.0971229     0.81686615
Toscana                    0.3538070                  0.5896788          0.7230256                 0.9163083     0.24832032
Umbria                    -0.7663648                 -0.7793093         -0.7670990                -0.7586054    -0.79095699
Marche                    -0.5259134                 -0.5387712         -0.5125747                -0.3874598    -0.60755512
Lazio                      1.2476229                  0.8807001          0.9127619                 0.9924408     0.78018577
Abruzzo                   -0.6187774                 -0.5862849         -0.6699170                -0.7871551    -0.44249342
Molise                    -0.9985247                 -1.0139081         -1.0077403                -1.0345855    -0.96213208
Campania                   1.0138046                  0.4144720          0.7230256                 0.8496924     0.08937203
Puglia                     0.4367213                  0.3521103          0.5610555                 0.9543746     0.08937203
Basilicata                -0.8940527                 -0.9099719         -0.9013028                -0.9013537    -0.86431775
Calabria                  -0.3642305                 -0.4585919         -0.4570421                -0.4635922    -0.36913267
Sicilia                    0.8189561                  0.2392653          0.1815828                 0.4975799     0.17495957
Sardegna                  -0.5466419                 -0.3605949         -0.5495964                -0.6444068    -0.09402985
attr(,"scaled:center")
          Consumo moderato  Comportamento abitudinario          Eccedenza abituale Eccedenza  abutuale a pasto
                 1342.2857                    388.4286                    250.7619                  121.7143
            Binge drinking
                  182.3810
attr(,"scaled:scale")
          Consumo moderato  Comportamento abitudinario          Eccedenza abituale Eccedenza  abutuale a pasto
                 1206.0650                    336.7451                    216.0893                  105.0800
            Binge drinking
                  163.5752
```

Con dati di tipo scalato, sicuramente risulterà più semplice osservare i parametri di distanza tra le varie regioni, avendo a disposizione dei valori standardizzati (si nota come la funzione scale di R riporti anche per ogni categoria la media campionaria e la deviazione standard, al fine di poter tornare ai dati iniziali).

Nel caso si volesse considerare, una matrice di distanze tra i singoli individui di un dataset, R mette a disposizione la funzione DIST, nella quale va fornito un dataset di riferimento, e la modalità (euclidea, massimo, etc.). Tale funzione, mette a disposizione una matrice quadrata (triangolare inferiore dato che per ogni distanza vale la proprietà commutativa e se si considera due volte lo stesso individuo si ha distanza 0), di ordine N (dove N è la numerosità del campione. Se si volesse fare questo ragionamento per il dataset in analisi, si otterrebbe una matrice delle distanze di *ordine 21* (considerando le 21 righe del dataset), troppo elevata per provare ad effettuare una divisione in cluster ottimale senza ulteriori strumenti.

A tale scopo, l'analisi dei cluster e il linguaggio di statistica R mettono a disposizione differenti metodologie per effettuare secondo una logica algoritmica l'analisi dei cluster.

La prima metodologia, forse anche quella più imminente, comprende quella che viene definita la cosiddetta classe di metodi ad enumerazione completa, che si pone l'obbiettivo di trovare la divisione in cluster migliore (*secondo quelle che vengono definite in questo ambito misure di non omogeneità statistica*), tra tutte le possibili divisioni in cluster tra gli individui del dataset. Il numero di possibili partizioni di un dataset in cluster è legato al concetto di Numeri di Stirling del secondo tipo (se si considera un numero fissato di cluster) ed ai numeri di Bell (se si considera un numero variabili di cluster da 1 ad N, dove N è l'ampiezza del campione).

Numero di modi per dividere n individui in m cluster

$$R(n,m) = \sum_{k=0}^{m} \binom{m}{k} (-1)^k (m-k)^n$$

Numero di modi per dividere n individui in m cluster (con m che varia da 1 ad N)

$$B_n = \sum_{m=1}^{n} S(n,m).$$

Come si può facilmente osservare, i numeri di Stirling, sono legati da un calcolo esponenziale in N, quindi il numero di cluster da analizzare, cresce esponenzialmente rispetto al numero di individui (ciò rende i metodi non gerarchici non computabili in tempo polinomiale).

A puro scopo informativo, se si volesse considerare un'analisi dei cluster, fatta tramite metodi di enumerazione completa:

```
> stirling2 <-function (n,m){
+   s<-0
+   if ((m >=1)&(m <=n)){
+     for (k in seq (0,m)){
+       s<-s+( choose (m,k)*(-1)^k*(m-k)^n/ factorial (m))}
+     return (c(s))
+   }}
>
> sumstirling2 <- function (n){
+   s <-0
+   for (k in seq (1,n))
+     s <-s+ stirling2 (n,k)
+   return (c(s))
+ }
> sumstirling2(21)
[1] 4.748698e+14
```

Si avrebbe da considerare un numero di cluster di una grandezza spropositata, quindi anche per tale dataset, è necessario ricorrere ad altre metodologie di divisione in cluster, al fine di trovare una soluzione migliore

rispetto alle metriche di *non omogeneità statistica*, quali ad esempio i metodi gerarchici e i metodi non gerarchici.

Passo 3 – Divisione in cluster tramite metodologie gerarchiche

Al fine di dividere in cluster il dataset relativo al consumo alcolico in Italia del 2019, è possibile procedere considerando quelli che in statistica vengono considerati metodi gerarchici, questa tipologia di metodi, non si occupa fin da subito di trovare una divisione in cluster migliore per un insieme di individui correlati da un insieme di caratteristiche, ma permettono di costruire un diagramma *chiamato dendrogramma*, che fornisce una divisione in cluster ad albero (dove ogni nodo corrisponde ad un cluster), ad ogni livello (altezza) partendo dalla base dell'albero dove per ogni cluster ci sarà un singolo individuo, fino ad arrivare alla radice ad avere un singolo cluster per tutti gli individui.

Gli algoritmi di clustering di tipo gerarchico si dividono in due macrocategorie:

- Gli algoritmi di tipo agglomerativo: che partono dalle foglie del dendrogramma per arrivare alla radice;
- Gli algoritmi di tipo divisivo, che lavorano in logica opposta.

Ai fini dell'analisi sul dataset si utilizzeranno algoritmi di tipo agglomerativo che avranno alla base, una metrica di distanza (in particolare la distanza euclidea), e una metodologia di accorpamento al fine di ottenere una nuova metrica per un'agglomerazione di due singoli cluster (importante ricordare che alcune metriche perderanno la fattezza di distanza, data appunto la perdita della proprietà di disuguaglianza triangolare).

La metodologia di tipo agglomerativo considerata è sintetizzabile in modo seguente:

- Si consideri la matrice delle distanze (secondo la metrica scelta) di ordine N x N, e si accorpino nello stesso cluster i due individui, i due Cluster o l'individuo e il cluster con distanza minima;
- Si elimini dalla matrice delle distanze le righe e le colonne relative alla coppia selezionata;
- Si inserisca una nuova riga e una nuova colonna relative al nuovo cluster formato all'interno della matrice delle distanze (secondo il metodo di agglomerazione scelto);
- Si iteri il procedimento fino ad ottenere una matrice quadrata di ordine 2 dove il singolo valore di distanza consistente corrisponde agli ultimi due cluster da unificare;

Come si nota, ad ogni iterazione del procedimento, si ottiene un nuovo cluster, che può essere inserito all'interno del dendrogramma ad un'altezza superiore (in particolare la nuova altezza corrisponderà alla distanza tra i due cluster sottostanti unificati);

È importante considerare che a seconda del metodo agglomerativo scelto (fissando una singola distanza – nel caso di analisi si opta per la distanza euclidea), si ottiene un algoritmo gerarchico differente, e potrebbe capitare che i singoli metodi possano portare ad una selezione ad una divisione in cluster di tipo gerarchico in maniera differente.

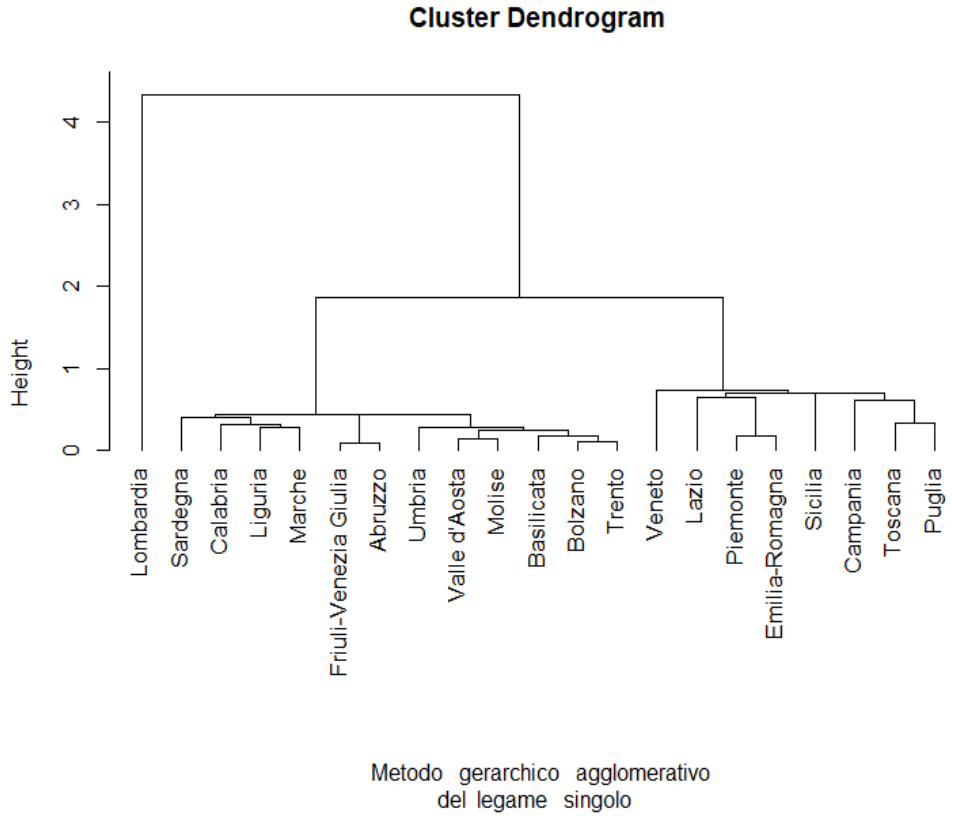
Il metodo del legame singolo

Il primo metodo che si intende usare ai fini dell'analisi è il metodo di accorpamento e ricalcolo delle distanze del valore singolo, che per il nuovo cluster formato, crea una nuova distanza rispetto agli altri individui, andando a selezionare la minima distanza tra Dik e Djk dove i e j sono i valori accorpati in un unico cluster e h è arbitrario tra i valori restanti (ovviamente per determinare la nuova riga e la nuova colonna della matrice delle distanze, il metodo del legame singolo va applicato a tutti i valori esterni al cluster appena individuato).

$$d_{(ij),k} = \min(d_{ik}, d_{jk}).$$

In R è possibile generare un dendrogramma, attraverso la funzione Plot, applicata al risultato della funzione HClust che fornisce in output, dettagli utili sull'esecuzione dell'algoritmo gerarchico, come le coppie unite ad ogni iterazione (parametro merge), oppure la distanza che simboleggia l'altezza dell'agglomerazione;

```
> #metodo del legame singolo
> d <- dist(scaling_dataset);
> d <- dist(scaling_dataset);
> hlssingle <-hclust (d,method = "single");
> str(hlssingle);
List of 7
 $ merge      : int [1:20, 1:2] -8 -5 -2 -1 -18 3 -11 -3 -19 -10 ...
 $ height     : num [1:20] 0.0851 0.1114 0.1473 0.179 0.1805 ...
 $ order      : int [1:21] 4 21 19 3 12 8 14 11 2 15 ...
 $ labels     : chr [1:21] "Piemonte" "Valle d'Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "single"
 $ call       : language hclust(d = d, method = "single")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```



Metodo gerarchico agglomerativo del legame singolo

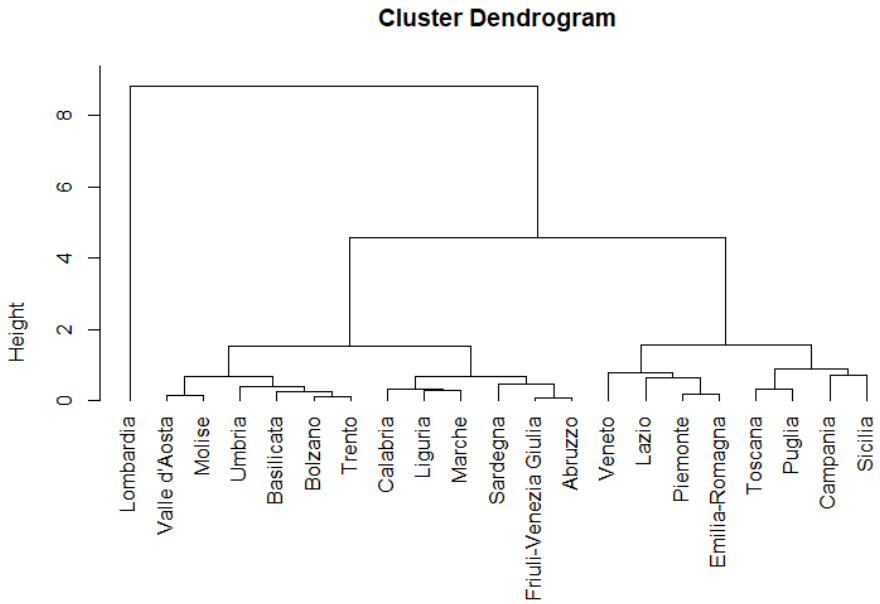
Come visibile dal dendrogramma, si potrebbe suppore una possibile suddivisione del dataset in 3 Cluster, uno per la singola regione Lombardia (che anche in questo caso assume un comportamento anomalo rispetto alle atre regioni), e gli altri due secondo l'ordine riportato dal diagramma. Però come già anticipato a seconda del metodo utilizzato gli algoritmi di tipo agglomerativo potrebbero generare dendrogrammi differenti, e soprattutto l'algoritmo del valore singolo potrebbe produrre una soluzione molto distante da quella ottimale, dato che esso considera le distanze minime tra i singoli cluster (regioni molto differenti potrebbero risultare posizionate nello stesso cluster).

A tal proposito è utile realizzare altri dendrogrammi con R secondo altri metodi (in particolare altri 4), per poi provarne ad analizzare le differenze.

Il metodo del legame completo

Tale metodo, effettua un'analisi opposta rispetto a quella proposta dal legame singolo, ed in particolare predispone l'algoritmo a considerare le distanze massima tra gli elementi del cluster formato e quelli esterni ad esso.

```
> hlsComplete <-hclust (d,method = "complete");
> str(hlsComplete);
List of 7
 $ merge      : int [1:20, 1:2] -8 -5 -2 -1 -18 -3 -10 -19 -11 -21 ...
 $ height     : num [1:20] 0.0851 0.1114 0.1473 0.179 0.259 ...
 $ order      : int [1:21] 4 2 15 11 18 5 6 19 3 12 ...
 $ labels     : chr [1:21] "Piemonte" "Valle d'Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "complete"
 $ call       : language hclust(d = d, method = "complete")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
>
> plot(hlsComplete ,hang =-1, xlab="Metodo gerarchico agglomerativo del legame completo");
```



Dal metodo del legame singolo, si nota già una piccola differenza nella formazione del dendrogramma, in particolare nelle altezze (cioè le distanze) a cui sono avvenute creazioni dei cluster al livello superiore, in particolare si nota che anche alcune regioni sono state agglomerate tramite combinazioni di cluster differenti, ma spicca ancora il valore singolo della Lombardia agglomerato solo alla fine.

Il metodo del legame medio

Questo terzo metodo pone invece il calcolo della nuova distanza, effettuando una media tra tutte le possibili distanze tra i valori interni nel cluster appena creato, e quelli posti all'esterno. In particolare, le nuove distanze vengono calcolate come:

$$d_{(i,j),k} = \frac{1}{2}\left(d_{i,k} + d_{j,k}\right) \qquad (k = 1, 2, \ldots, n; k \neq i, j)$$

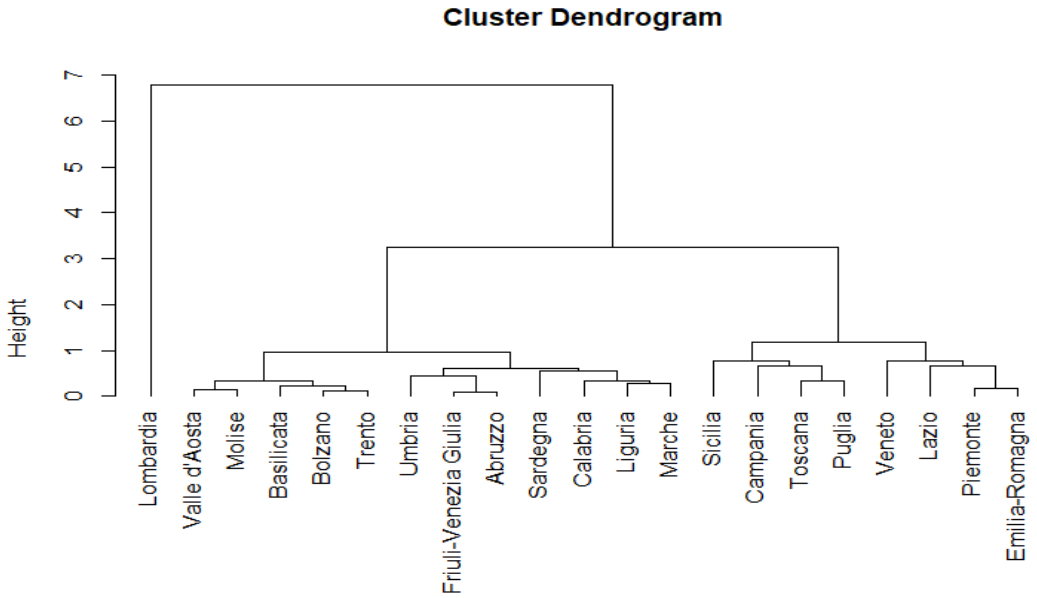
Nel caso ci siano solo due individui nel cluster di nuova formazione;

$$\begin{aligned} d_{(uv),z} &= \frac{1}{(N_u + N_v)\, N_z} \sum_{\{i:I_i \in G_{uv}\}} \sum_{\{j:I_j \in G_z\}} d_{ij} \\ &= \frac{1}{(N_u + N_v)\, N_z} \sum_{\{i:I_i \in G_u\}} \sum_{\{j:I_j \in G_z\}} d_{ij} + \frac{1}{(N_u + N_v)\, N_z} \sum_{\{i:I_i \in G_v\}} \sum_{\{j:I_j \in G_z\}} d_{ij} \\ &= \frac{N_u}{N_u + N_v} d_{uz} + \frac{N_v}{N_u + N_v} d_{vz}, \end{aligned}$$

Nel caso in cui si abbiano più individui da considerare nei cluster che vengono uniti (nella formula N_u Corrisponde alla numerosità di individui nel primo cluster, N_y, la numerosità nel secondo cluster, e N_z, corrisponde alla numerosità degli elementi esclusi dal cluster);

È molto importante considerare come per il metodo del legame medio, il numero di individui di ogni cluster incida molto sul calcolo delle nuove distanze (infatti cluster più grandi, avranno un peso maggiore rispetto ad altri.

```
> #Il metodo del legame medio
> hlsAverage <-hclust (d,method = "average");
> str(hlsAverage);
List of 7
 $ merge      : int [1:20, 1:2] -8 -5 -2 -1 -18 -3 -19 -10 3 -11 ...
 $ height     : num [1:20] 0.0851 0.1114 0.1473 0.179 0.2197 ...
 $ order      : int [1:21] 4 2 15 18 5 6 11 8 14 21 ...
 $ labels     : chr [1:21] "Piemonte" "Valle d'Aosta" "Liguria" "Lombardia" ...
 $ method     : chr "average"
 $ call       : language hclust(d = d, method = "average")
 $ dist.method: chr "euclidean"
 - attr(*, "class")= chr "hclust"
```



Da questo dendrogramma generato, si nota come in effetti i dati siano tra di loro molto più omogenei e appunto salta subito all'occhio come ogni agglomerazione in cluster con più individui, avvenga in modo già più distribuito rispetto al caso del legame minimo, ma nonostante ciò il valore lombardo standardizzato risulta molto distante anche secondo questa metrica dalle altre regioni.

Il metodo del centroide

Secondo questo nuovo metodo, la metrica è definita come la "distanza" tra i centroidi dei due gruppi, ossia la media campionaria tra tutti gli elementi dei due Cluster in considerazione; Nel caso si ragioni in termini di due individui ne cluster di nuova formazione la nuova misura sarà calcolata come:

$$d^2_{(ij),k} = \sum_{r=1}^{p} (\overline{x}_{(i,j),r} - \overline{x}_{k,r})^2 = \frac{1}{2}(d^2_{ik} + d^2_{jk}) - \frac{1}{4}d^2_{ij}, \qquad (k \neq i, j)$$

dove

$$\overline{x}_{(i,j),r} = \frac{1}{2}(x_{i,r} + x_{j,r}) \qquad \overline{x}_{k,r} = x_{k,r} \qquad (r = 1, 2, \ldots, p).$$

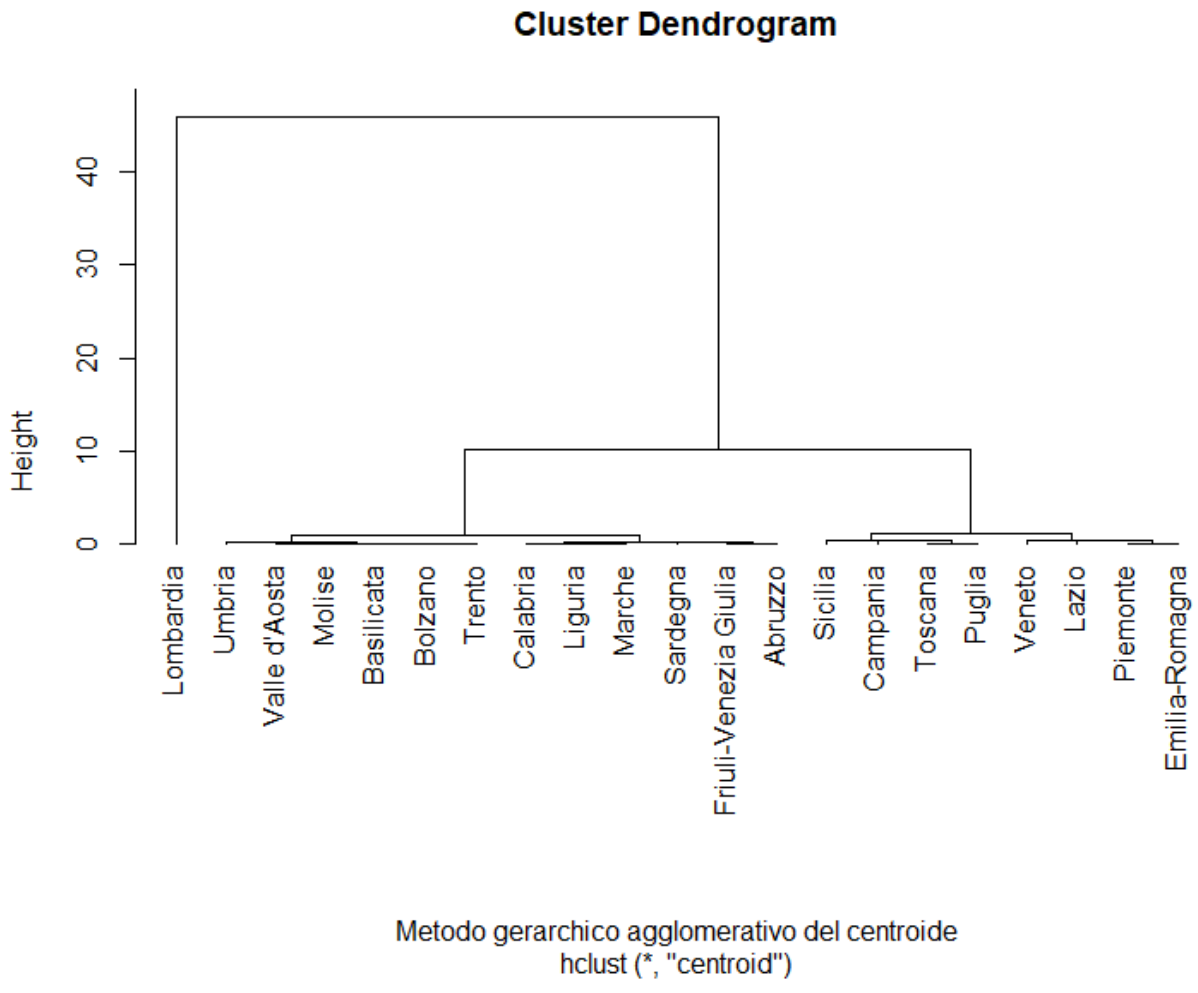
Nel caso in cui invece si voglia generalizzare al caso in cui i due cluster uniti abbiano più elementi ognuno, allora questa formula sarà espressa come:

$$d^2_{(uv),z} = \sum_{k=1}^{p} (\overline{x}_{(u,v),k} - \overline{x}_{(z),k})^2 = \frac{N_u}{N_u + N_v} d^2_{uz} + \frac{N_v}{N_u + N_v} d^2_{vz} - \frac{N_u N_v}{(N_u + N_v)^2} d^2_{u,v},$$

dove

$$\begin{aligned}
\overline{x}_{(u,v),r} &= \frac{1}{N_u + N_v} \sum_{\{i:I_i \in G_{uv}\}} x_{i,r} \\
&= \frac{1}{N_u + N_v} \sum_{\{i:I_i \in G_u\}} x_{i,r} + \frac{1}{N_u + N_v} \sum_{\{i:I_i \in G_v\}} x_{i,r} \\
&= \frac{N_u}{N_u + N_v} \overline{x}_{(u),r} + \frac{N_v}{N_u + N_v} \overline{x}_{(v),r} \\
& \qquad\qquad (r = 1, 2, \ldots, p) \\
\overline{x}_{(z),r} &= \frac{1}{N_z} \sum_{k:I_k \in G_z} x_{kr}
\end{aligned}$$

Dove anche in questo caso il numero di elementi nei due cluster accumunati corrisponde a N_u ed N_v. È importante osservare come anche per il metodo del centroide, la nuova metrica calcolata sia pesata in base alla numerosità dei singoli cluster agglomerati, ed in particolare, per questo e per il successivo metodo della mediana, la metrica di calcolo, perde l'appellativo di distanza, dato che in questo nuovo caso, si considera la distanza quadratica (perdita della proprietà di disuguaglianza triangolare).



Come si osserva, questo nuovo dendrogramma, ha prodotto un risultato molto simile rispetto a quello generato con il metodo del legame medio, dato il ragionamento di selezione molto affine, con l'unica differenza riportata dai valori delle altezze molto più grandi data l'elevazione al quadrato delle distanze (è interessante considerare che se i dati non fossero stati standardizzati, a questo punto le altezze sarebbero state molto più elevate).

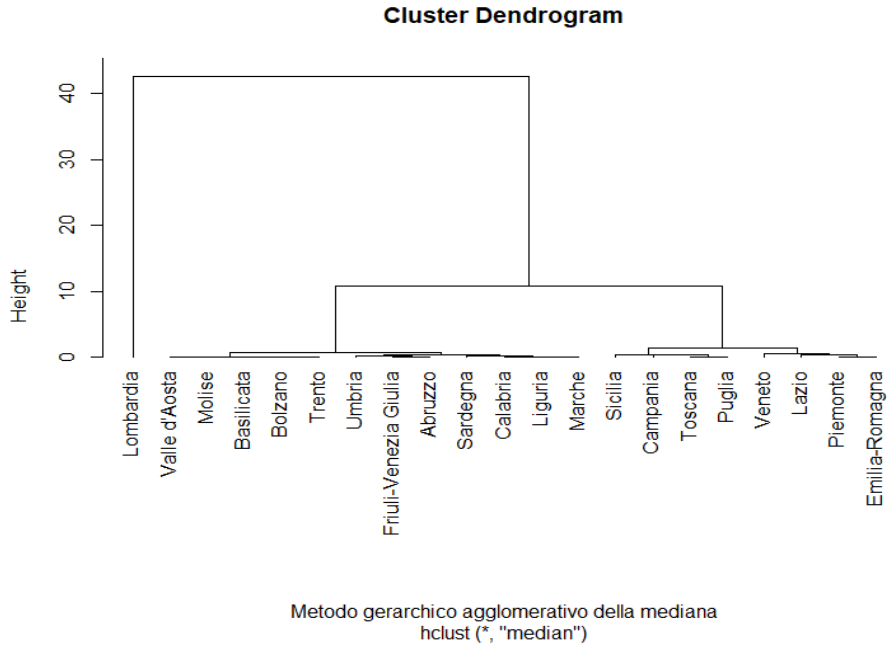
Il metodo della mediana

Come già anticipato, questo metodo come il metodo del centroide, si basa sulle distanze elevate al quadrato, ma rispetto al metodo precedente, non è per nulla condizionato dalla numerosità dei Cluster, infatti per il calcolo delle distanze verrà utilizzata la singola formula:

$$d^2_{(uv),z} = \sum_{k=1}^{p} (\overline{x}_{(u,v),k} - \overline{x}_{(z),k})^2 = \frac{1}{2} d^2_{u,z} + \frac{1}{2} d^2_{v,z} - \frac{1}{4} d^2_{u,v},$$

dove

$$\overline{x}_{(uv),r} = \frac{1}{2} \left(\overline{x}_{(u),r} + \overline{x}_{(v),r} \right) \qquad (r = 1, 2, \ldots, p).$$



Per il dataset in esame, anche il metodo della mediana sembra non influire più di tanto, rispetto alla divisione precedente riportata dal metodo del centroide, a meno dell'ordine di selezione degli elementi.

Passo 4 – Confronti tra i risultati dei vari metodi gerarchici

Come già anticipato, quando viene effettuata un'analisi dei Cluster tramite metodologie, gerarchiche non è opportuno pensare di generare un dendrogramma con un singolo metodo di agglomerazione e accontentarsi di selezionare l'altezza corrispondente al numero di cluster desiderato. Infatti, in alcuni casi potrebbe capitare che alcuni metodi (come quello del legame singolo o della mediana) possano portare a quello che viene definito effetto agglomerativo a Catena, dove elementi molto distanti tra loro risulteranno poi far parte dello stesso Cluster.

Per il Dataset relativo al consumo alcolico in Italia del 2019, per quanto riportato al passo precedente, questa problematica non sembra essersi verificata, in quanto i diagrammi risultanti dai 5 metodi sembrano produrre pressoché la stessa divisione in cluster in particolare quando il numero di Cluster è uguale a 3;

Ma è effettivamente così? Oppure i vari metodi presentano leggere differenze nella produzione dei vari dendrogrammi?

Per dare risposta a questa domanda, viene in aiuto la funzione cutree di R, che dato in input un dendrogramma risultante dalla funzione HClust (applicata ad un dataset con qualsiasi metodo e distanza) e un numero di cluster da ottenere o un'altezza di riferimento, restituisce un vettore dove ogni elemento corrisponde ad un individuo del dataset e l'elemento indica il Cluster di appartenenza, dove appunto il numero dei Cluster corrisponde a quello selezionato oppure a quello derivante dall'altezza scelta nel dendrogramma.

Cosa accade applicando la funzione cutree ai vari dendrogrammi generati con i 5 metodi fissando un numero di cluster pari a 3?

Metodo del legame singolo

```
> #cutree metodo del legame singolo con un numero di cluster pari a 3
> cut3Single = cutree(hlsSingle, k = 3);
> cut3Single
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                Trento 
                    1                     2                     2                     3                     2                     2 
   Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria               Marche 
                    1                     2                     1                     1                     2                     2 
               Lazio              Abruzzo               Molise              Campania               Puglia           Basilicata 
                    1                     2                     2                     1                     1                     2 
           Calabria               Sicilia             Sardegna 
                    2                     1                     2 
```

Metodo del legame completo

```
> #cutree metodo del legame completo con un numero di cluster pari a 3
> cut3Complete = cutree(hlsComplete, k = 3);
> cut3Complete
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                Trento 
                    1                     2                     2                     3                     2                     2 
   Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria               Marche 
                    1                     2                     1                     1                     2                     2 
               Lazio              Abruzzo               Molise              Campania               Puglia           Basilicata 
                    1                     2                     2                     1                     1                     2 
           Calabria               Sicilia             Sardegna 
                    2                     1                     2 
```

Metodo del legame medio

```
> #cutree metodo del legame medio con un numero di cluster pari a 3
> cut3Average = cutree(hlsAverage, k = 3);
> cut3Average
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                Trento 
                    1                     2                     2                     3                     2                     2 
   Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria               Marche 
                    1                     2                     1                     1                     2                     2 
               Lazio              Abruzzo               Molise              Campania               Puglia           Basilicata 
                    1                     2                     2                     1                     1                     2 
           Calabria               Sicilia             Sardegna 
                    2                     1                     2 
```

Metodo del centroide

```
> #cutree metodo del centroide con un numero di cluster pari a 3
> cut3Centroid = cutree(hlsCentroid, k = 3);
> cut3Centroid
             Piemonte         Valle d'Aosta              Liguria             Lombardia              Bolzano                Trento
                    1                     2                    2                     3                    2                     2
              Veneto Friuli-Venezia Giulia       Emilia-Romagna               Toscana               Umbria               Marche
                    1                     2                    1                     1                    2                     2
               Lazio               Abruzzo               Molise              Campania               Puglia           Basilicata
                    1                     2                    2                     1                    1                     2
            Calabria               Sicilia            Sardegna
                    2                     1                    2
> |
```

Metodo della mediana

```
> #cutree metodo del median con un numero di cluster pari a 3
> cut3Median = cutree(hlsMedian, k = 3);
> cut3Median
             Piemonte         Valle d'Aosta              Liguria             Lombardia              Bolzano                Trento
                    1                     2                    2                     3                    2                     2
              Veneto Friuli-Venezia Giulia       Emilia-Romagna               Toscana               Umbria               Marche
                    1                     2                    1                     1                    2                     2
               Lazio               Abruzzo               Molise              Campania               Puglia           Basilicata
                    1                     2                    2                     1                    1                     2
            Calabria               Sicilia            Sardegna
                    2                     1                    2
```

Fortunatamente l'analisi dei cluster condotta tramite metodologia gerarchica ha prodotto dei risultati molto affidabili, dato che con ognuno dei 5 metodi utilizzati, sono stati generati dei dendrogrammi molto simili tra loro ed in particolare, considerando la divisione in 3 Cluster, si ottiene esattamente lo stesso partizionamento di regioni.

Divisione in 3 Cluster analisi Gerarchica	
	Regioni
1	Lombardia
2	Valle d'Aosta
	Molise
	Basilicata
	Bolzano
	Trento
	Umbria
	Friuli-Venezia Giulia
	Abruzzo
	Sardegna
	Calabria
	Liguria
	Marche
3	Sicilia
	Campania
	Piemonte
	Veneto
	Emilia-Romagna
	Toscana
	Lazio
	Puglia

Da come si evince, anche da questa prima divisione in cluster, la Lombardia è riportata nuovamente come dato anomalo rispetto al resto d'Italia, e dato che 5 metodologie di clustering hanno riportato essa come singola regione non accumunabile alle altre, si può affermare con un certo margine di sicurezza, che tale regione, in termini di Consumo Alcolico è effettivamente da tenere sotto controllo, ad avvalorare ancora di più la tesi, si rammenta che le singole colonne del dataset, sono state standardizzate rispetto la media

campionaria e la deviazione standard di ogni colonna, quindi viene anche meno il fattore di grandezza che evidentemente si poteva attribuire all'elevata popolazione della regione.

Passo 5 – Affidabilità della divisione in Cluster rilevata

Ad affermare la validità della divisione in cluster rilevata, senz'altro è la moltitudine di metodi gerarchici che hanno portato allo stesso risultato, ma per essere sicuri che tale divisione sia effettivamente ottimale rispetto ad altre eventuali divisioni possibili, essa va validata attraverso l'utilizzo di opportuni indici, ed in particolare trattandosi di una divisione in Cluster, allora è opportuno verificare lo stato del clustering attraverso le misure di non omogeneità statistica.

Tali indici sono calcolabili a partire dal calcolo delle varianze delle singole colonne del dataset che in R possono essere riassunte in forma tabulata tramite la funzione COV (utilizzata precedentemente), applicata all'intero Dataset.

Per il dataset scalato, la matrice delle varianze e covarianze è la seguente:

```
> covMatrix <- cov(scaling_dataset);
> covMatrix
                          Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
Consumo moderato                1.0000000                 0.9687964          0.9718822                 0.9599366      0.9385731
Comportamento abitudinario      0.9687964                 1.0000000          0.9919715                 0.9583284      0.9837097
Eccedenza abituale              0.9718822                 0.9919715          1.0000000                 0.9829876      0.9545601
Eccedenza  abutuale a pasto     0.9599366                 0.9583284          0.9829876                 1.0000000      0.8941021
Binge drinking                  0.9385731                 0.9837097          0.9545601                 0.8941021      1.0000000
```

Al fine di calcolare gli indici di interesse, è necessario moltiplicare i singoli individui della matrice per N - 1 dove N è il numero di individui del dataset, in questo caso le 21 righe del dataset.

```
> #calcolo matrice di non omogeneità
> NOMatrix <- 20 * covMatrix;
> #calcolo matrice di non omogeneità
> NOMatrix <- 20 * covMatrix;
> NOMatrix
                          Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
Consumo moderato                 20.00000                  19.37593           19.43764                  19.19873       18.77146
Comportamento abitudinario       19.37593                  20.00000           19.83943                  19.16657       19.67419
Eccedenza abituale               19.43764                  19.83943           20.00000                  19.65975       19.09120
Eccedenza  abutuale a pasto      19.19873                  19.16657           19.65975                  20.00000       17.88204
Binge drinking                   18.77146                  19.67419           19.09120                  17.88204       20.00000
```

Da questo primo conteggio ne risulta quella che in statistica viene definita come matrice di non omogeneità.

Da questa matrice, la diagonale principale (le varianze delle singole categorie * (numero di elementi -1), servirà a calcolare il primo indice di stima necessario, cioè la misura di non omogeneità totale, definita come:

$$\mathrm{tr}H_I = \sum_{r=1}^{p} h_{rr} = (n-1)\sum_{r=1}^{p} s_r^2.$$

dove P è il numero di categorie del dataset.

Tramite R per il dataset standardizzato in analisi, questo valore risulta essere:

```
> #calcolo della misura di non omogeneità totale del dataset standardizzato
> trHI <-(21 - 1) * sum(apply(scaling_dataset ,2, var ))
> trHI
[1] 100
```

Questo stesso calcolo adesso va ripetuto anche per i 3 cluster che l'analisi gerarchica ha generato, ma per fare ciò con R, è senz'altro più conveniente ricorrere all'utilizzo della funzione Aggregate, che permette di calcolare una serie di indici (quali media campionaria o varianza) data un dataset e il relativo partizionamento ottenuto con la funzione ***cutree***.

```
> agvar <- aggregate (scaling_dataset, tagliolist , var)[, -1]
> agvar
  Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
1       0.09041116                0.09878987         0.07572573                0.04638051     0.19819159
2       0.05096740                0.06139238         0.06250091                0.07924971     0.07267015
3               NA                        NA                 NA                        NA             NA
```

*per il terzo cluster, avendo ovviamente un singolo elemento, allora tutte le varianze saranno eguagliate a 0.

Al fine di calcolare la misura di non omogeneità statistica per i due cluster (il terzo non è utile alla stima dato che ha un singolo elemento e per definizione, la misura di omogeneità sarà pari a 0) è utile determinare quanti elementi sono presenti all'interno di ogni Cluster.

In R tale procedura può essere fatta tramite la funzione table, sul taglio fatto con la funzione cutree:

```
> num <-table (cut3Centroid);
> num
cut3Centroid
 1  2  3
 8 12  1
```

Da queste due tabelle si possono evincere le misure di non omogeneità statistica per i cluster 1 e 2:

```
> #misura non omogeneità primo cluster
> trH1 <-(num [[1]] -1) * sum ( agvar [1, ])
> trH1 # visualizza la misura di non omogeneità del primo cluster
[1] 3.566492
>
> #misura non omogeneità secondo cluster
> trH2 <-(num [[2]] -1) * sum ( agvar [2, ])
> trH2 # visualizza la misura di non omogeneità del secondo cluster
[1] 3.594586
```

La somma di non omogeneità statistica tra i cluster, prende il nome di misura di non omogeneità interna ai cluster, e se essa è molto piccola rispetto a quella totale, allora ciò è già un primo segnale che la divisione in Cluster scelta è ottimale, a questa misura, si aggiunge poi quella di non omogeneità tra i cluster che è ottenuta sottraendo alla misura totale quella interna, che per indicare una buona divisione in cluster deve essere grande.

```
> #misura di non omogeneità interna ai cluster
> trHS <- trH1 + trH2
> trHS
[1] 7.161078
>
> #misura di non omogeneità tra i cluster
> trHB <- trHI - trHS
> trHB
[1] 92.83892
```

La relazione tra le misure di non omogeneità interna, tra i cluster e totale, può essere anche vista in termini relativi nel seguente modo:

$$1 = \frac{\text{tr } S}{\text{tr } T} + \frac{\text{tr } B}{\text{tr } T}.$$

```
> #Misure Relative
> trHS / trHI
[1] 0.07161078
> #Misure Relative
> trHB / trHI
[1] 0.9283892
```

Dato che per il dataset ISTAT relativo al consumo alcolico in italia 2019, la divisione in cluster individuata produce degl'indici di non omogeneità interna e tra i cluster ottimali (in particolare il rapporto tra la misura tra i cluster e quella totale riporta un valore di 0,93), si può senz'altro affermare che la divisione in cluster è ottimale per il dataset, ma è effettivamente la migliore?

Affermare che la partizione trovata sia la migliore, non è ancora possibile, dato che essa è derivata da un'analisi di tipo agglomerativo secondo dei criteri specifici(anche se nel caso del dataset in analisi, addirittuta 5 metodi anno portato allo stesso risultato). A questa analisi, va aggiunta la ricerca di una nuova partizione, che consente di riallocare anche individui già assegnati ad un cluster in Cluster differenti, in modo tale da poter individuare le misure di non omogeneità migliori, e capire se 3 Cluster effettivamente sono ottimali per il dataset considerato, o addirittura verificare che la soluzione ottenuta con l'analisi gerarchica sia la migliore.

Tale procedura può essere fatta con quelli che nell'analisi dei Cluster vengono definiti Metodi non Gerarchici.

Passo 6 – Ottimizzazione della soluzione tramite metodi non gerarchici

Come già anticipato, la soluzione proposta tramite l'utilizzo dei metodi gerarchici non è automaticamente definibile la migliori, in quanto se da un dendrogramma si determina una partizione, non è detto che essa possa dare delle misure di non omogeneità ottimali, in quanto (come già ribadito più volte) le partizioni create tramite metodologie gerarchiche non permettono la riallocazione di individui, una volta che gli stessi sono entrati a far parte di un cluster.

Secondo questa seconda logica lavorano invece le metodologie di clustering non gerarchiche, ed in particolare l'algoritmo del k-means. Questo algoritmo si basa su un numero fissato di K cluster e su una partizione iniziale provvisoria da cui partire. Ad ogni iterazione dell'algoritmo, vengono ricalcolati i centrodi delle K partizioni, e gli individui vengono riassegnati alla partizione contenente il centroide a distanza minore dall'individuo stesso. L'algoritmo non appena effettua un'iterazione senza differenze rispetto a quella precedente, restituisce in output la divisione in cluster ottenuta (sicuramente dipendente da quella iniziale immessa, ma plausibilmente anche diversa da essa).

In R, il metodo non gerarchico del k-means è implementato dalla funzione *kmeans,* che permette di definire il clustering di partenza in tre metodologie differenti:

- Generazione automatica dei k cluster a cui sarà associato un centroide per ogni cluster;

- Generazione automatica di p possibili divisioni in k cluster e utilizzo di quella migliore;
- Utilizzo di centroidi derivanti da un partizionamento precedentemente definito tramite altre tecniche di clustering (come ad esempio il metodo gerarchico del centroide).

In particolare, kmeans fornisce in automatico le misure di non omogeneità statistica della partizione finale (senza dover ulteriormente calcolare come in precedenza questi indici).

Fissando un numero di cluster ideale pari a 3, è possibile applicare la funzione kmeans al dataset standardizzato relativo al consumo alcolico in Italia del 2019, tramite le 3 metodologie precedentemente esposte ed osservare se ci sono risultati differenti rispetto alla precedente analisi gerarchica.

Metodo 1 – Generazione automatica di una singola partizione di partenza

```
> #k-means standard sul dataset di consumo alcolico 2019 standardizzato con 3 cluster fissati
> kmeans(scaling_dataset, 3)
K-means clustering with 3 clusters of sizes 12, 8, 1

Cluster means:
  Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
1       -0.7311262                -0.7006147         -0.7323912                -0.7697081     -0.6299709
2        0.7439187                 0.6735701          0.7478995                 0.8508820      0.5348858
3        2.8221649                 3.0188159          2.8054978                 2.4294406      3.2805647

Clustering vector:
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                Trento
                    2                     1                     1                     3                     1                     1
               Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria               Marche
                    2                     1                     2                     2                     1                     1
                Lazio              Abruzzo               Molise              Campania               Puglia           Basilicata
                    2                     1                     1                     2                     2                     1
             Calabria               Sicilia             Sardegna
                    1                     2                     1

Within cluster sum of squares by cluster:
[1] 3.594586 3.566492 0.000000
 (between_SS / total_SS =  92.8 %)
```

Metodo 2 – Generazione automatica di 10 cluster partition da cui estrarre selezionare la partizione iniziale

```
> #k-means con 10 ripetizioni per la scelta della partizione iniziale
> #sul dataset di consumo alcolico 2019 standardizzato con 3 cluster fissati
> k2 <- kmeans(scaling_dataset, 3, nstart = 10)
> k2
K-means clustering with 3 clusters of sizes 12, 1, 8

Cluster means:
  Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
1       -0.7311262                -0.7006147         -0.7323912                -0.7697081     -0.6299709
2        2.8221649                 3.0188159          2.8054978                 2.4294406      3.2805647
3        0.7439187                 0.6735701          0.7478995                 0.8508820      0.5348858

Clustering vector:
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                Trento
                    3                     1                     1                     2                     1                     1
               Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria               Marche
                    3                     1                     3                     3                     1                     1
                Lazio              Abruzzo               Molise              Campania               Puglia           Basilicata
                    3                     1                     1                     3                     3                     1
             Calabria               Sicilia             Sardegna
                    1                     3                     1

Within cluster sum of squares by cluster:
[1] 3.594586 0.000000 3.566492
 (between_SS / total_SS =  92.8 %)
```

Metodo 3 – Utilizzo del precedente clustering ottenuto con il metodo del centroide

```
> #k-means standard sul dataset di consumo alcolico 2019 standardizzato con 3 cluster fissati
> #tramite il precedente clustering ottenuto con il metodo del centroide
> tagliolist <-list(cut3Centroid)
> tagliolist
[[1]]
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                 Trento
                    1                     2                     2                     3                     2                      2
               Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria                 Marche
                    1                     2                     1                     1                     2                      2
                Lazio               Abruzzo               Molise              Campania               Puglia             Basilicata
                    1                     2                     2                     1                     1                      2
             Calabria               Sicilia             Sardegna
                    2                     1                     2

> centroidiIniziali <- aggregate(scaling_dataset, tagliolist, mean)[,-1]
> centroidiIniziali
  Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
1        0.7439187                 0.6735701          0.7478995                0.8508820      0.5348858
2       -0.7311262                -0.7006147         -0.7323912               -0.7697081     -0.6299709
3        2.8221649                 3.0188159          2.8054978                2.4294406      3.2805647
> km3 <- kmeans (scaling_dataset , centers = centroidiIniziali , iter.max = 10)
> km3
K-means clustering with 3 clusters of sizes 8, 12, 1

Cluster means:
  Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
1        0.7439187                 0.6735701          0.7478995                0.8508820      0.5348858
2       -0.7311262                -0.7006147         -0.7323912               -0.7697081     -0.6299709
3        2.8221649                 3.0188159          2.8054978                2.4294406      3.2805647

Clustering vector:
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                 Trento
                    1                     2                     2                     3                     2                      2
               Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria                 Marche
                    1                     2                     1                     1                     2                      2
                Lazio               Abruzzo               Molise              Campania               Puglia             Basilicata
                    1                     2                     2                     1                     1                      2
             Calabria               Sicilia             Sardegna
                    2                     1                     2

within cluster sum of squares by cluster:
[1] 3.566492 3.594586 0.000000
 (between_SS / total_SS =  92.8 %)
```

Come è possibile notare, tutte e 3 le applicazioni del metodo k-means, riportano lo stesso risultato finale, che tra l'altro resta del tutto identico alla precedente soluzione ottenuta tramite metodi gerarchici. Dopo aver ottenuto tale risultato, si può senz'altro affermare che la soluzione precedentemente ricavata tramite analisi gerarchica è davvero ottimale, e con essa possono anche essere maggiormente validate tutte le deduzioni precedentemente effettuate.

Ma effettivamente la soluzione con 3 cluster è la migliore?

Effettuare una divisione in cluster di una particolare gruppo di individui in modo ottimale, significa senz'altro cercare di creare dei sottogruppi di individui il più simili possibili tra loro, ma è anche vero che aumentare il numero di cluster porterebbe nel caso delle regioni d'Italia per il dataset di riferimento ad un partizionamento poco significativo, in quanto per ogni dendrogramma generato con metodologie gerarchiche, le altezze in cui si ottenevano 3 cluster, si sono rivelate abbastanza distanti rispetto al resto delle possibili divisioni.

Sarebbe possibile pensare di ridurre a due il numero di cluster possibili, ma senz'altro è facile presumere che anche la funzione kmeans possa riportare un partizionamento peggiore di quello rilevato, in quanto l'anomalia del valore Lombardo, non porterebbe altro che all'aumento inevitabile della misura di non omogeneità interna.

```
> #k-means in due cluster
> km4 <- kmeans(scalind_dataset, 2)
> km4
K-means clustering with 2 clusters of sizes 12, 9

Cluster means:
  Consumo moderato Comportamento abitudinario Eccedenza abituale Eccedenza  abutuale a pasto Binge drinking
1       -0.7311262                -0.7006147         -0.7323912                -0.7697081     -0.6299709
2        0.9748349                 0.9341530          0.9765216                 1.0262774      0.8399612

Clustering vector:
             Piemonte         Valle d'Aosta               Liguria             Lombardia               Bolzano                Trento
                    2                     1                     1                     2                     1                     1
               Veneto Friuli-Venezia Giulia        Emilia-Romagna               Toscana               Umbria               Marche
                    2                     1                     2                     2                     1                     1
                Lazio               Abruzzo               Molise              Campania               Puglia           Basilicata
                    2                     1                     1                     2                     2                     1
             Calabria               Sicilia             Sardegna
                    1                     2                     1

Within cluster sum of squares by cluster:
[1]  3.594586 24.974133
 (between_SS / total_SS =  71.4 %)
```

Come infatti si presumeva l'aggiunta della regione Lombardia ad uno dei due cluster, riposta ad un abbassamento della misura di non omogeneità tra i cluster, rispetto al 93% fin ora avuto con il precedente partizionamento in 3 cluster, che a fronte di questa deduzione, può senz'altro considerarsi il miglior compromesso tra tutte le tecniche gerarchiche e non gerarchiche utilizzate.