**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: Studente:**

***Prof.ssa. Amelia G. Nobile******Ferrara Carmine***

***Matr.05225/00990***

**ANNO ACCADEMICO 2020/2021**

Sommario

[Introduzione 3](#_Toc53732395)

[Dataset Utilizzato ai fini dell’indagine 3](#_Toc53732396)

[Analisi Univariata – Colonna Binge – Drinking dataset Excel 4](#_Toc53732397)

[Passo 1 – Analisi Univariata – Diagramma di Pareto 4](#_Toc53732398)

[Passo 2 – Analisi univariata – Instogramma – Box Plot ad intaglio 5](#_Toc53732399)

[Passo 3 – Indici di centralità rispetto al campione 7](#_Toc53732400)

[Passo 4 – Indici di dispersione rispetto alla media campionaria 10](#_Toc53732401)

# 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.

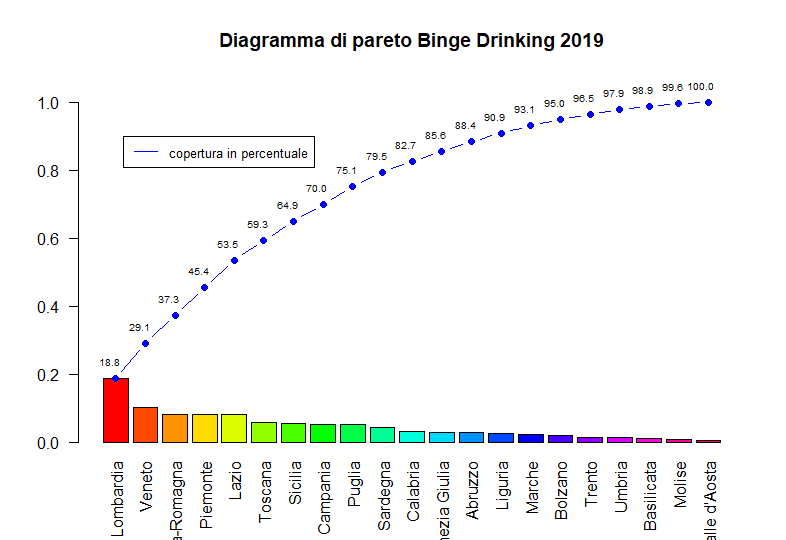
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regioni** | **Consumo moderato** | **Comportamento abitudinario** | **Eccedenza abituale** | **Eccedenza abutuale a pasto** | **Binge drinking** |
| **Piemonte** | **2.055** | **684** | **463** | **222** | **311** |
| **Valle d'Aosta** | **54** | **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** |

## Dataset Utilizzato ai fini dell’indagine

# 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

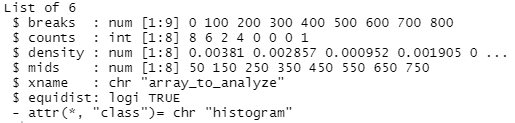
- Veneto Secondo punto

- 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).





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

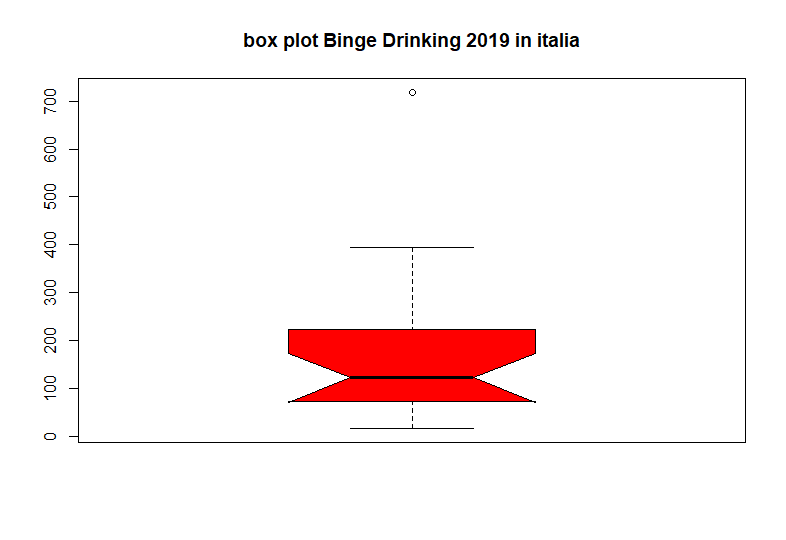


Frequenze relative delle 8 classi calcolate in base alla densità di ogni classe nell’istogramma.

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.

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

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:

Q1 -1.5·(Q3-Q1) = 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.

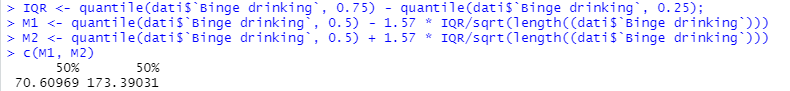
Mentre per il baffo superiore:

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

Posizione = 394 – Veneto

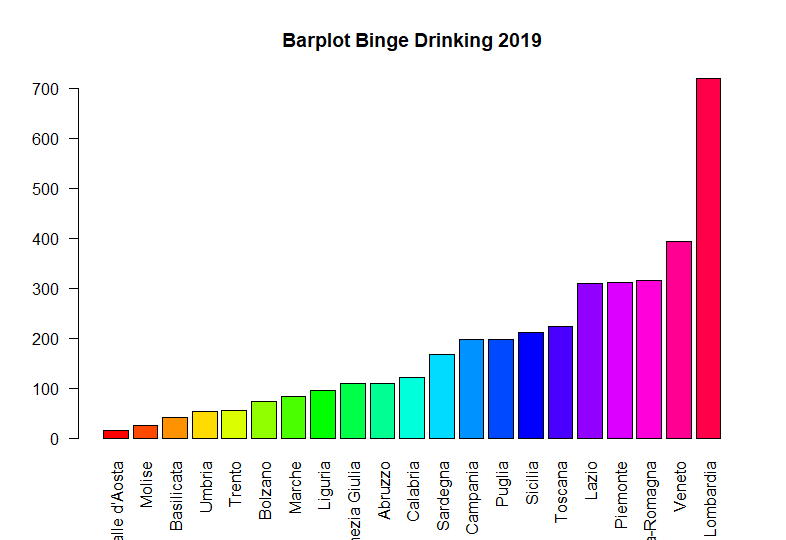
(primo valore inferiore al risultato del calcolo)

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.



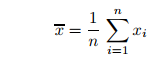
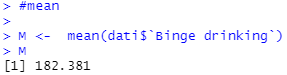
Dal box plot analizzato si può evincere anche che la mediana campionaria è posta intorno al valore 125.5, considerando anche l’intaglio riportato (Indice di fiducia posto con alfa = 1.5) si ha che appunto questo valore può variare da 70.61 a 173.39 come intervallo di confidenza.

## 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:

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.

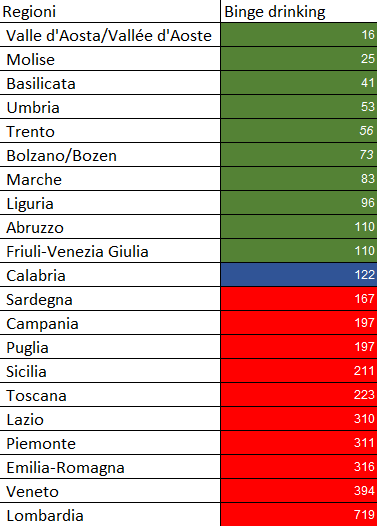
Osservazione: dal dato di media campionaria riportato abbiamo che il valore per la regione Lombardia ha uno scarto molto elevato dalla media campionaria

Scarto Lombardia =

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).

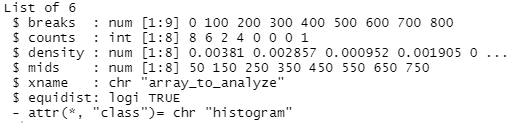
Per provare a confermare ciò è di rilevanza anche il calcolo della mediana campionaria (tenendo quindi conto solo dei valori centrali del campione in analisi).

 #dati = 21 (dispari) – per il calcolo della mediana campionaria ->



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





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



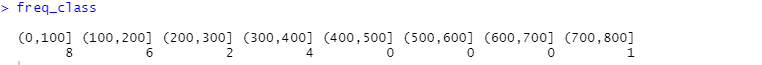
Possiamo anche dedurre come per il campione in analisi, il dato di densità riportato è particolarmente interessante in termini della tematica osservata, infatti la prima classe d’intervallo [0:100), risulta avere una frequenza sia assoluta che relativa maggiore rispetto agli altri dati del campione (più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



E considerandone la distribuzione di frequenza relativa cumulata.



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.

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?*

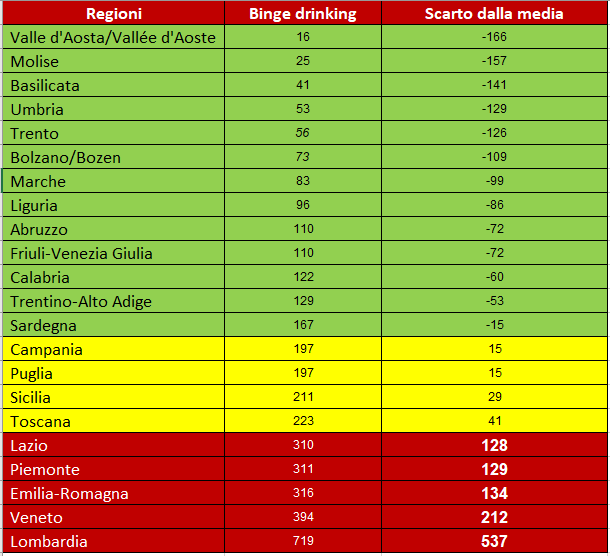
Essendo che la variabile Binge\_Drinking considerata, indica la pericolosità massima per il consumo di alcol in Italia, sarebbe davvero utile capire effettivamente quali regioni necessitano di maggior attenzione.

*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?*

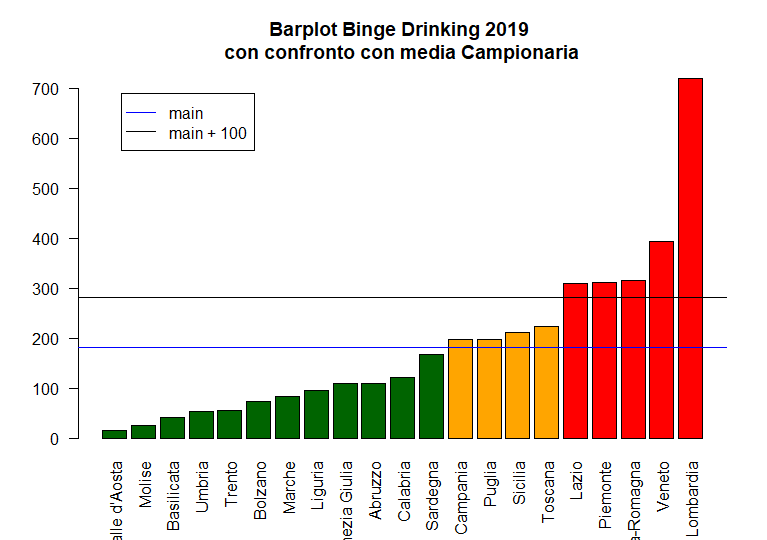
Considerando la media campionaria (media nazionale), è possibile delimitare in che modo i dati si discostino da essa.



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.

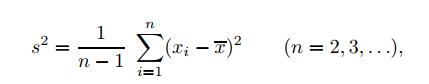


Osserviamo che i dati riportati in rosso, sicuramente innalzano di molto la soglia di rischio che nel 2019 è stata rilevata in Italia.

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

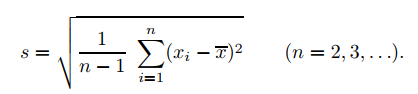
*Ma in che modo?*

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



Deviazione Standard campionaria



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:

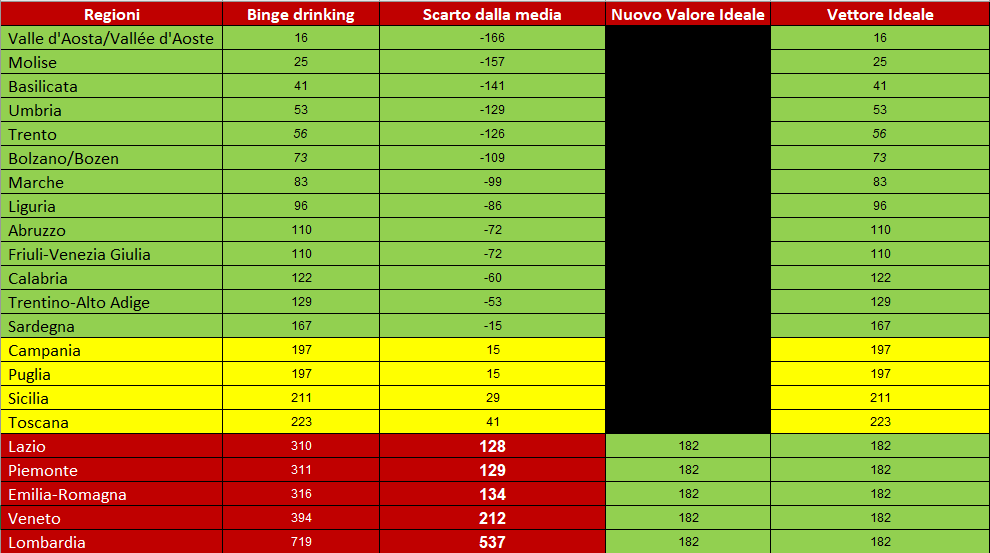


Abbiamo un valore minore di 1, che in termini statistici, indica che la media campionaria, è un buon indice di valutazione per il campione analizzato, permettendoci quindi di dare ulteriore supporto alle cose dedotte fino a questo punto.

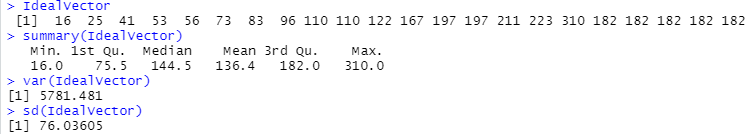
Effettivamente, si è dimostrato utilizzare la media come valore di riferimento, potrebbe diminuire il tasso di pericolosità che si cela dietro al campione in esame, e soprattutto è possibile affermare che i dati più distanti dalla media (meglio evidenziati nel diagramma a barre), devono essere “tenuti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

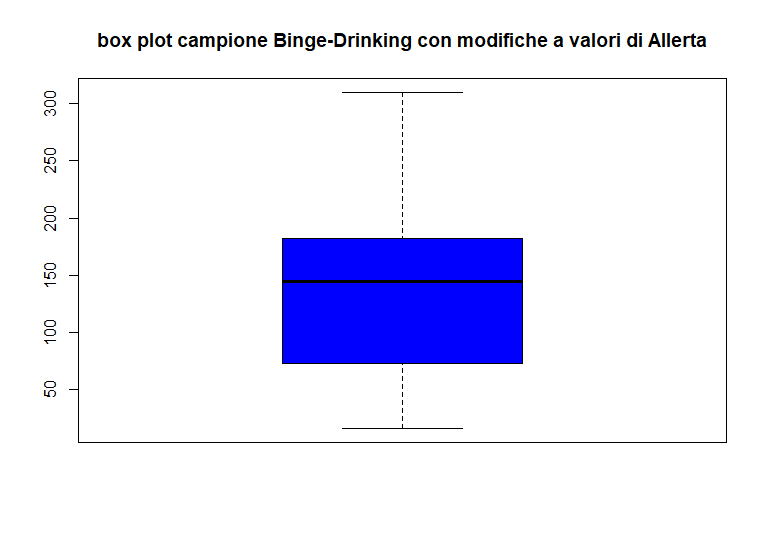
*Essendo che gli scarti dalla media contribuiscono in primo luogo al calcolo degli indicatori di dispersione, questi valori diminuiscono se i valori critici prima individuati, assumessero valori più simili alla media attuale?*

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:



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



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.

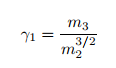
# Passo 5 – Analisi di simmetrie e concentrazione di dati nel campione

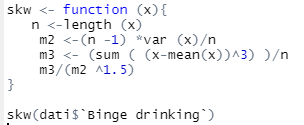
Come ultimo passo di studio nell’analisi univariata rispetto al dato Binge Drinking considerato fino a questo momento, è senz’altro molto utile fare anche qualche considerazione aggiuntiva sulla forma che i dati assumono quando vengono disposti all’interno di un grafico, quindi in qualche modo andare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

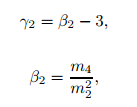
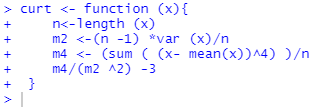
Dall’analisi della skewness campionaria, quindi dal confronto dei momenti centrali di secondo e di terzo ordine:



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:



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.