IN MEDIA ...

La Statistica

Sai ched'è la statistica? È 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.
Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
secondo le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due.

(Trilussa, Sonetti Romaneschi)

Il grande poeta Trilussa non poteva spiegare meglio in dialetto romanesco il concetto di **media**.

La media riesce a racchiudere in sè un grande senso di ingiustizia o giustizia sociale, dipende dal punto da cui si guarda.

E' un concetto che conosciamo fin da piccolissimi; quando infatti mia figlia aveva pochi anni, 2 o 3 non ricordo, voleva sempre dividere le cose con gli altri, ad esempio le caramelle contenute in un sacchetto, e diceva a me e alla sorella: "Facciamo un po' per ciascuno, metà a te, metà a te e metà a me" ... mumble, mumble, ... aveva qualche problema con le frazioni, ma l'idea di dare ad ognuna di noi tre la stessa quantità di caramelle denotava un certo desiderio di equità, lo stesso che probabilmente avrebbe voluto Trilussa quando ha raccontato la questione del pollo, ma che non trovava nel mondo reale del suo tempo.

Il calcolo della media (**aritmetica**! vi sono vari tipi di medie, ma qui ci riferiremo sempre a quella aritmetica) è noto a tutti: si considera l'ammontare dei valori

osservati (ad esempio, nella poesia di Trilussa, i polli mangiati in un anno da tutta la *"gente"*) e lo si divide per la numerosità della popolazione osservata, trovando così un valore che riassuma in sè tutte le caratteristiche della distribuzione di quel bene sulla popolazione.

Ovviamente ha senso calcolare la media quando su di una popolazione si osservano **caratteri quantitativi**: se l'osservazione riguardasse il colore degli occhi, come si potrebbe calcolare il colore medio? Mentre è sensato parlare di altezza media delle persone, di reddito medio delle famiglie, di voto medio degli alunni di una classe in una verifica.

Ma più che addentrarci sul calcolo della media, risulta interessante porre attenzione sulle sue proprietà:

1. la media è sempre compresa tra i valori minimo e massimo osservati; si dice anche che è compresa nel **campo di variazione** della distribuzione
2. la somma delle differenze di tutti i valori osservati dalla media è nulla; significa che gli **scarti** negativi compensano tutti quelli positivi

Se ogni unità statistica della popolazione possedesse esattamente il valor medio calcolato, cioè il carattere osservato fosse **equidistribuito**, l'ammontare complessivo del fenomeno osservato non cambiarebbe; questa proprietà deriva direttamente dalla definizione di media data da Chisini, un grande matematico italiano morto nel 1967 che ha molto studiato la statistica, ed è strettamente legato al tema della giustizia sociale.

La media è quindi un indice di **posizione** e si può anche considerare il baricentro della distribuzione che rappresenta; sarà molto interessante osservare dove si colloca tale valore all'interno dei campi di variazione di distribuzioni diverse.

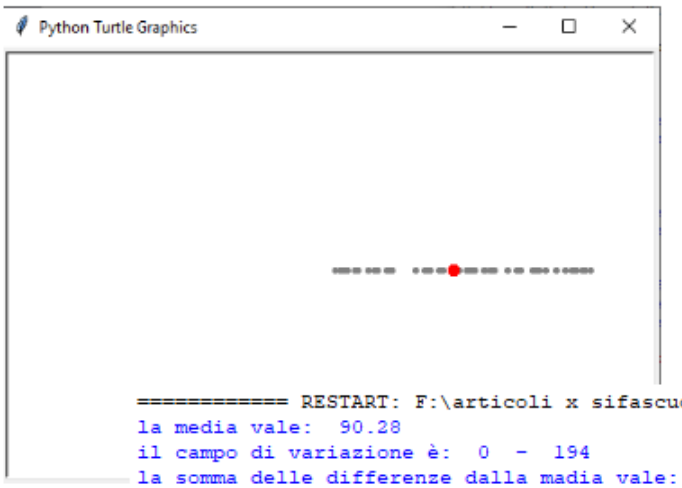
Poichè per noi è impossibile rilevare i dati a cui si riferisce Trilussa sul numero di polli mangiati in un anno dalla *"gente"* della sua epoca, possiamo analizzare distribuzioni generate in modo casuale per osservare le proprietà della media descritte; in particolare sarà evidente che la media è il baricentro della distribuzione e che la somma degli scarti dalla media è sempre nulla.

Si può procedere con la costruzione di una applicazione Python che:

- generi un certo numero di valori casuali inserendoli in una lista
- determini il minimo e il massimo di tali valori, provveda a sommarli uno ad uno e a rappresentarli su di una linea come piccoli punti grigi
- calcoli la loro media e la rappresenti con un punto rosso tra i dati che essa riassume
- calcoli la somma degli scarti di ciascun valore dalla media per verificare che tale somma sia nulla.

Ecco il codice Python e una prova di elaborazione ottenuta nella IDLE (Integrated Development and Learning Environment) in cui vengono generati numeri casuali compresi tra 0 e 200, per cui ci si aspetta che il valor medio risulti attorno al valore 100; la numerosità dei dati generati è stata impostata a 50:

```
import random
import turtle

# inizializzo le variabili
dato = []
somma = 0
media = 0
minimo = 400
massimo = 0
n = 50
# costruisco una turtle
t = turtle.Turtle()
t.hideturtle()
t.penup()


for i in range(n):
    # genero n dati random
    dato.append(random.randint(0,200))
    # trovo il valore minimo, il massimo
    if dato[i] > massimo: massimo = dato[i]
    if dato[i] < minimo: minimo = dato[i]
    # li rappresento sulla retta dei numeri reali con un puntino grigio
    t.goto(dato[i] ,0)
    t.dot(5,'gray')
    # sommo i dati uno ad uno
    somma += dato[i]
# calcolo la media
media = somma / n
# rappresento la media col colore rosso
t.goto(media ,0)
t.dot(10,'red')

print("la media vale: ", media)
print("il campo di variazione è: ", minimo," - ",massimo)

# calcolo la somma degli scarti dei dati dalla media
diff = []
som_diff = 0
for i in range(n):
    som_diff += media - dato[i]

print("la somma delle differenze dalla madia vale: ", som_diff)
```

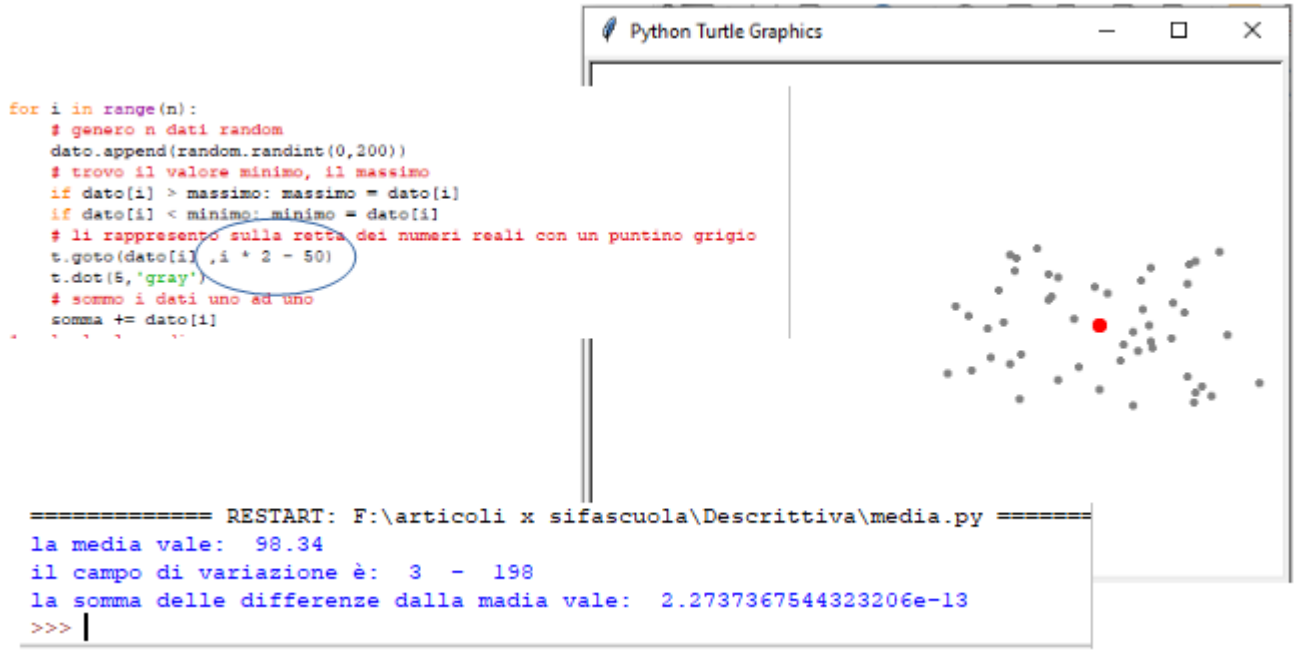



```
============ RESTART: F:\articoli x sifascuola\Descrittiva\media.py ============
la media vale:  90.28
il campo di variazione è:  0  -  194
la somma delle differenze dalla madia vale:  5.684341886080802e-14
>>>
```

Dal risultato dell'elaborazione si individua che il minimo valore generato è 0 e il massimo è 194, la media dei valori vale 90.28, quindi distante dal valore atteso, ma ricordiamo che i dati generati sono solo 50!!!! Si potrebbero ripetere più prove aumentando la numerosità dei dati generati.

La somma degli scarti è un valore piccolissimo, prossimo a zero e sarebbe molto interessante approfondire anche l'aspetto relativo alla rappresentazione numerica interna alla macchina per comprendere pienamente il perchè tale valore non sia esattamente 0.

Una piccola modifica al codice consente di rappresentare i punti corrispondenti ai valori generati non più su una stessa linea, ma a livelli diversi, così da osservare in modo ancora più evidente che la media rappresenta il baricentro della distribuzione.



Generare in modo casuale i dati da eleborare è stata una scelta arbitraria, dettata solo dalla comodità di avere a disposizione dei valori in modo veloce, e ovviamente i dati così generati non possono essere considerati rappresentativi della quantità dipollo consumata. Tale scelta ci può consentire però alcune riflessioni di natura molto diversa dal punto di vista statistico, e che meritano opportuni approfondimenti in altra sede:

- lo studio della **Distribuzione Uniforme**, un modello statistico che rappresenta valori che abbiano tutti la stessa probabilità di presentarsi (molto adatto per le uscite dei numeri del lotto, meno per il consumo di pollo) e di cui è noto a priori il valor medio

- lo studio delle **indagini campionarie:** laddove non sia possibile intervistare tutta la popolazione ci si "accontenta" di estrarre con tecniche opportune un campione rappresentativo e le informazioni ricavate dal campione si estendendo a tutta la popolazione mediante l'**Inferenza Statistica.**

Quindi, se disponessimo dei dati effettivi derivanti da un'indagine statistica su di una certa popolazione, indagine che ad esempio abbia davvero indagato sulla quantità di pollo consumata in un anno, potremmo elaborarli ottenendo campo di variazione e media aritmetica per quella popolazione così come abbiamo fatto per i dati generati in modo casuale.

Se questi dati fossero contenuti in un file, ad esempio di tipo .csv (Comma Separated Value), tipologia spesso utilizzata per memorizzare i dati grezzi di indagini statistiche (dataset), potremmo effettuare poche modifiche al codice proposto, al fine di importare i dati dal file anzichè generarli in modo casuale.

Di seguito il codice Python modificato per la lettura dei dati da file .csv, file che per essere letto deve trovarsi nella stessa cartella in cui si trova l'applicazione Python.

```
import turtle
# inizializzo le variabili
somma = 0
media = 0
minimo = 200.0
massimo = 0.0
n = 50
# costruisco una turtle
t = turtle.Turtle()
t.hideturtle()
t.penup()
# Apro File Csv che contiene i dati dell'indagine
data = open("data.csv")

for i in range(50):
    # leggo il dato dal file e lo trasformo in numero reale
    dato = float(data.readline())
    # trovo il valore minimo, il massimo
    if dato > massimo: massimo = dato
    if dato < minimo: minimo = dato
    # lo rappresento sulla retta dei numeri reali con un puntino grigio
    t.goto(dato ,i * 2 - 50)
    t.dot(5,'gray')
    # sommo il dato ai precedenti
    somma += dato
# chiudo il file
data.close()
# calcolo la media
media = somma / n
# rappresento la media col colore rosso
t.goto(media ,0)
t.dot(10,'red')
print("la media vale: ", media)
print("il campo di variazione è: ", minimo," - ",massimo)
# calcolo la somma degli scarti dei dati dalla media
# apro nuovamente il file
data = open("data.csv")
diff = []
som_diff = 0
for i in range(n):
    dato = float(data.readline())
    som_diff += media - dato
print("la somma delle differenze dalla madia vale: ", som_diff)
data.close()
```