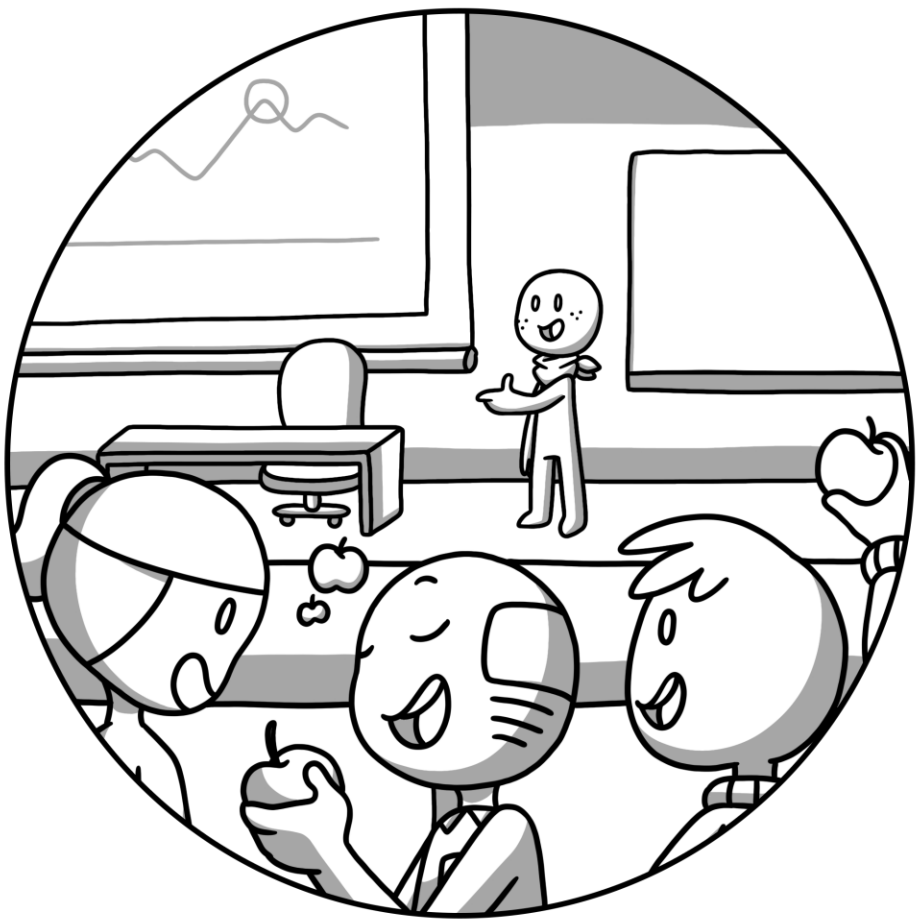
Indici di Tendenza Centrale



COMIXPLAIN

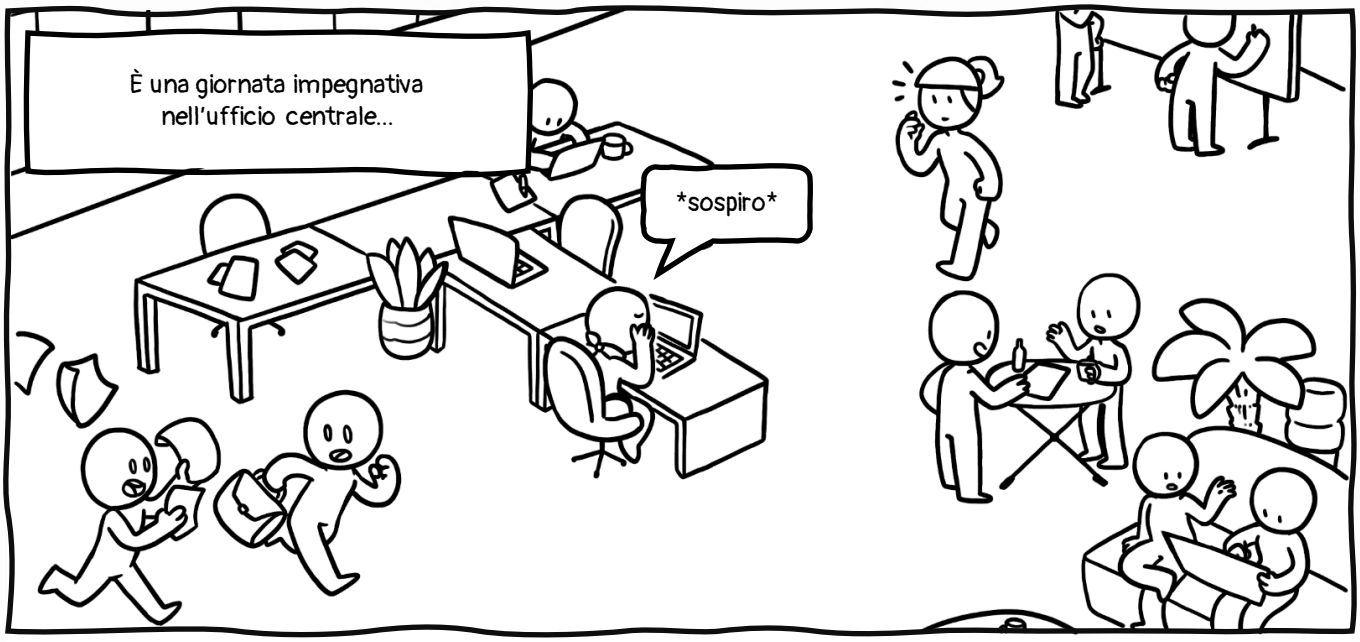
Questo fumetto è stato creato nell'ambito del progetto di ricerca Comixplain, finanziato dall'Innovation Call 2022 dell'Università di Scienze Applicate di St. Pölten, in Austria.

Squadra:
Victor-Adriel De-Jesus-Oliveira
Hsiang-Yun Wu
Christina Stoiber
Magdalena Boucher
Alena Ertl

Contatto:
victor.oliveira@fhstp.ac.at

Illustrazioni:
Magdalena Boucher & Alena Ertl

https://fhstp.github.io/comixplain


È una giornata impegnativa nell'ufficio centrale...
sospiro

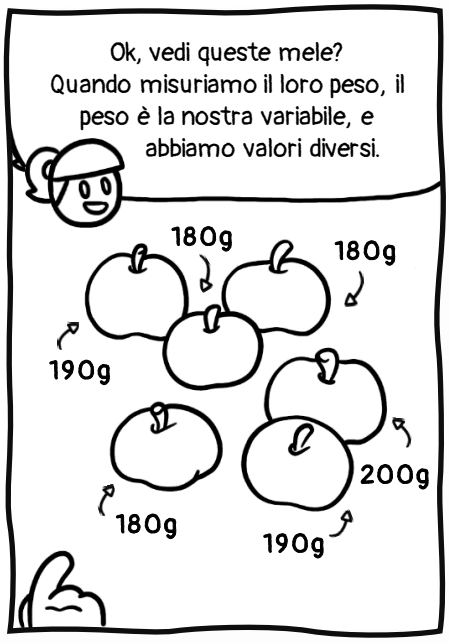

Ehi novellino, sei pronto per la tua presentazione di oggi?


Non proprio... Ho raccolto molti dati, ma ho solo pochi minuti per presentarli tutti!
Perché non mostrate solo gli aspetti più interessanti?

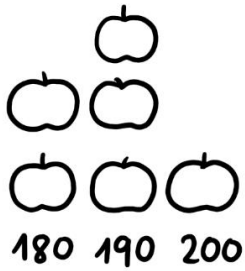

Ma ci sono così tanti aspetti interessanti! Non posso presentare tutto.
beh...


Che ne dici di iniziare con la TENDENZA CENTRALE? È un buon modo per avere una panoramica dei dati raccolti.
Cosa vuoi dire?

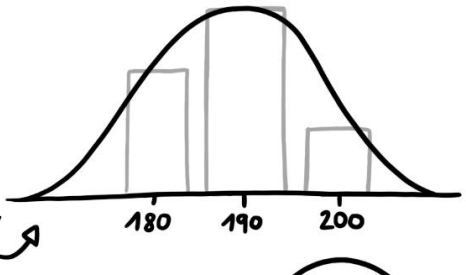

Ciao, accetteresti una mela?
Serviranno da esempio!


Ok, vedi queste mele? Quando misuriamo il loro peso, il peso è la nostra variabile, e abbiamo valori diversi.
180g
180g
190g
180g
200g
190g

Il modo migliore per descrivere una variabile è riportare i valori e la frequenza con cui appaiono. Questa è chiamata la DISTRIBUZIONE della variabile.

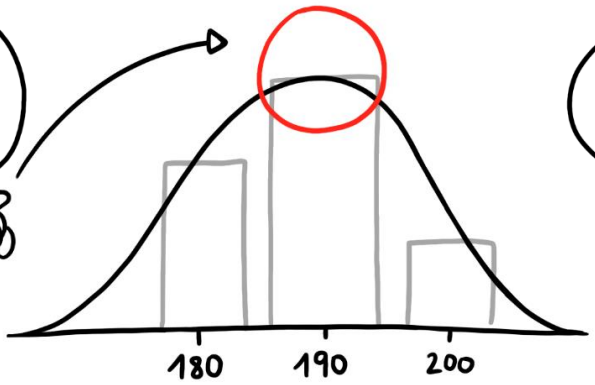


Se visualizziamo il peso delle nostre mele, la distribuzione sarebbe simile a questa, poiché le mele in questo cesto pesano all'incirca lo stesso.



Ha la forma di una campana...

Esatto! Se ci fossero altre mele nel paniere, la nostra migliore ipotesi per il loro peso sarebbe un valore intorno a quel punto.



Ah, capisco. Ma... Come si calcola questo valore?

Questo punto della nostra distribuzione rappresenta bene i nostri dati: lo chiamiamo TENDENZA CENTRALE.

Possiamo indicare la tendenza centrale in diversi modi. La MEDIA è il metodo più comune ed è anche facile da calcolare.

Queste sono le nostre sei mele:

200 180 190 190 190 180

Per calcolare la media, sommiamo tutti i valori di peso...

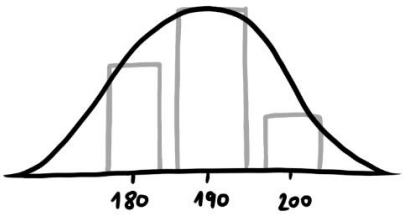
$$200 + 180 + 190 + 190 + 190 + 180$$

... E poi dividiamo per il numero di mele che abbiamo...

$$\frac{(200 + 180 + 190 + 190 + 190 + 180)}{6}$$

Il risultato è un valore medio di

188.3g



Ah, è quasi la parte superiore del nostro grafico a campana

Esatto. Ma la media non è sempre il modo migliore per descrivere i dati.

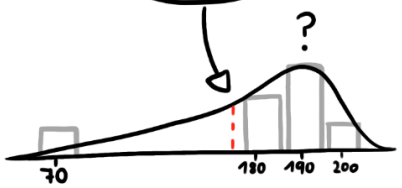
Ragazzi, guardate questa piccola mela!

Pesa, tipo, solo 70 grammi!

Ancora una volta, esempio perfetto!

Se aggiungiamo la mela piccola al nostro set di dati e calcoliamo la media...

$$\frac{(200+180+190+190+190+180+70)}{7 \text{ apples}} = 171.4\,g$$

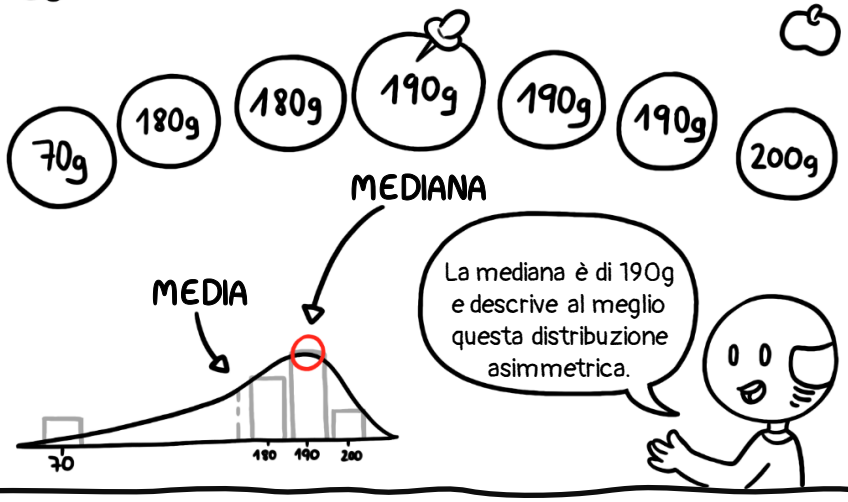


... Il grafico ora assomiglia più a un topo e la media è lontana dall'alto, vicina alla coda.

Vero. Ora la media non è molto rappresentativa...

Ma è ancora possibile descrivere la tendenza centrale con la MEDIANA.

Se ordiniamo tutte le nostre mele in base al peso in ordine crescente, la mediana è il valore posizionato al centro.



La mediana è di 190g e descrive al meglio questa distribuzione asimmetrica.

Vedo che hai trovato una delle mie piccole mele. Sentiti libero di prenderne altre. Sono contento che ti sia piaciuto!

In effetti, possiamo prendere l'intero cestino per un momento?

Se aggiungiamo questo intero cestino di piccole mele al nostro set, la prima piccola mela non è più un'outlier.

*Outliers sono valori anomali che possono essere errori di misurazione o eventi rari.

In questo caso, il nostro grafico ha improvvisamente due picchi:

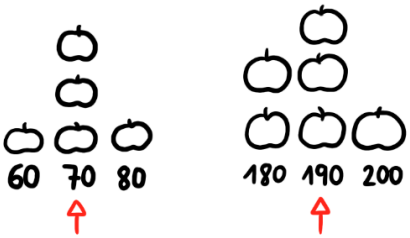
Peso delle mele grandi e piccole

Come topi innamorati!

Ah... Sì.

Ad ogni modo, possiamo usare qualcosa chiamato MODA per descrivere la tendenza centrale della nostra distribuzione se ha più picchi.

La moda definisce i valori che si verificano più frequentemente in un set di dati.



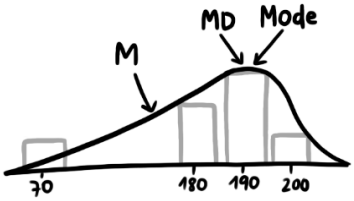
In questo caso, abbiamo più di una moda. Ma a volte ce ne sarà solo una o addirittura nessuna moda.

È possibile applicare la media, la mediana e la moda a diversi campioni di mele.
Ma spesso, alcune misure rappresentano i dati meglio di altre.

180, 180, 190, 190, 190, 200

M = 188.3
MD = 190
Mode = 190

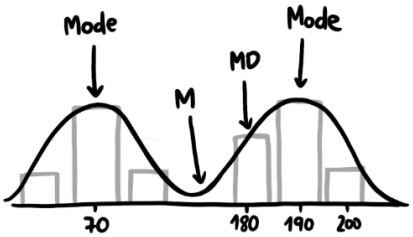
Parametri migliori



70, 180 ,181, 190, 191, 191, 200

M = 171.8
MD = 190
Mode = 191

Parametri migliori

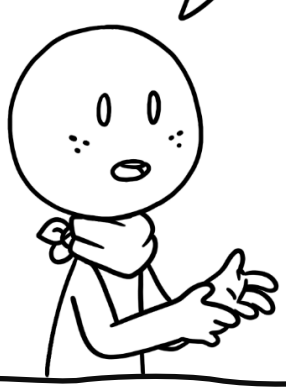


60, 70, 70, 70, 80, 180, 180, 190, 190, 190, 200

M = 134.5
MD = 180
Mode = 70 & 190

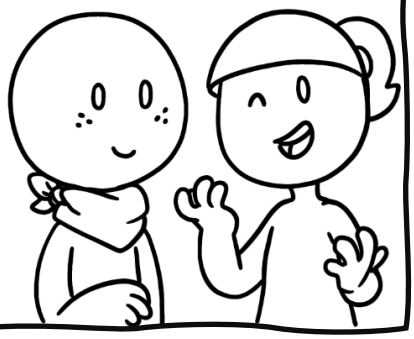
Miglior parametro

Ok grazie... Ho imparato molto! Ora vorrei applicarlo ai dati che devo presentare. Provengono da un'app che tiene traccia delle misurazioni della frequenza cardiaca.



User ID	Heart Rate (bpm)	Time of Use	User Rating
1	45	13:00	1
2	50	9:00	5
3	55	10:00	3
4	57	9:00	4
5	63	14:00	5
6	70	15:00	5
7	65	16:00	4
8	75	15:00	2

È fattibile.
Guarda i tuoi dati e segui l'esempio delle mele!
È possibile utilizzare la pagina successiva per i calcoli.



Prima di voltare pagina, prova a calcolare la media, la mediana e la moda per ogni variabile e controlla quale parametro è più adatto a descrivere l'andamento centrale.
Puoi prendere appunti su questa pagina!

Sentiti libero di controllare i tuoi calcoli.
Puoi prendere altri appunti in questa pagina!

FREQUENZA CARDIACA

Calcolo della MEDIA:

$$\frac{45+50+55+57+63+70+65+75}{8 \text{ utenti}} = \frac{480}{8} = 60 \text{ bpm}$$

Calcolo della MEDIANA:

45 50 55 | 57 63 | 65 70 75

Se ci sono due valori nel mezzo, la media dei due valori è la mediana:
(57+63)/2 = 60bpm)

Calcolo del MODA:

45, 50, 55, 57, 63, 70, 65, 75

Ogni valore esiste una volta sola: la moda non esiste!

Se la distribuzione dei valori è simmetrica, senza distorsioni, la Media è solitamente uguale alla Mediana.

FREQUENZA D'USO

Calcolo della MEDIA:

$$\frac{9+9+10+13+14+15+15+16}{8 \text{ utenti}} = \frac{101}{8} = 12.6?$$

Il tempo trascorso davanti allo schermo non è un valore quantitativo, quindi calcolare la media e la mediana non ha alcun senso!

Calcolo della MEDIANA:

9 9 10 | 13 14 | 15 15 16 = 13.5?

Calcolo del MODA:

9:00	10:00	13:00	14:00	15:00	16:00
2x	1x	1x	1x	2x	1x

= 2 mode: 9:00 & 15:00

La Moda non è adatta solo per le distribuzioni multimodali, ma anche quando si lavora con dati ordinali e categorici.

VALUTAZIONE A STELLE

Calcolo della MEDIA:

$$\frac{1+2+3+4+4+5+5+5}{8 \text{ utenti}} = \frac{29}{8} = 3.6 \text{ Stelle}$$

Calcolo della MEDIANA:

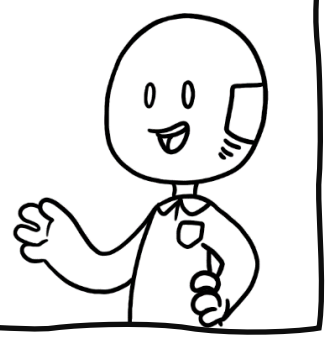
1 2 3 | 4 4 | 5 5 5 = 4 Stelle

Calcolo del MODA:

1	2	3	4	5
1x	1x	1x	2x	3x

= 5 Stelle

Per i set di dati con una distribuzione asimmetrica, la Mediana è un modo migliore per descrivere la tendenza centrale.



I linguaggi di programmazione, ad esempio R, consentono di calcolare la tendenza centrale degli attributi in set di dati di grandi dimensioni. Con le librerie R come tidyverse, è possibile visualizzare la distribuzione dei dati.

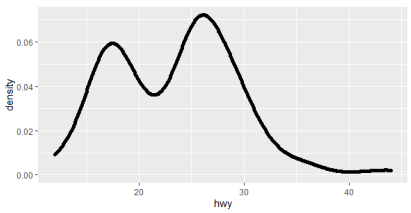
	model	year	hwy
1	jetta	1999	44
2	corolla	2008	37
3	civic	2008	36
4	civic	2008	36
5	corolla	1999	35
6	altima	2008	32
7	sonata	2008	31
+ altri 227 articoli			

Nel tidyverse, hai accesso a set di dati come mpg con dati sul risparmio di carburante. Include 11 attributi, come il modello dell'auto, l'anno di produzione e le miglia stradali per gallone (hwy).

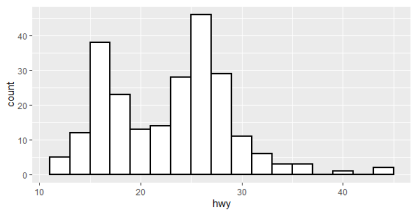
È possibile utilizzare ggplot, incluso nel tidyverse, per visualizzare la distribuzione dei dati relativi al miglio stradale per gallone (hwy) utilizzando un istogramma, una curva di densità o entrambi.

```
install.packages("tidyverse")  # Installalo solo la prima volta che usi la libreria
```

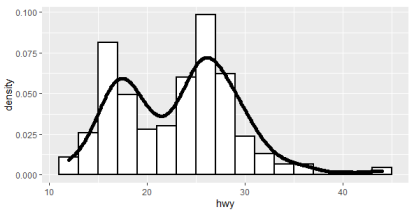
```
library(tidyverse)
plot <- ggplot(mpg, aes(x=hwy))
plot +
 geom_density()
```




```
plot +
 geom_histogram(
  colour="black",
  fill="white" )
```



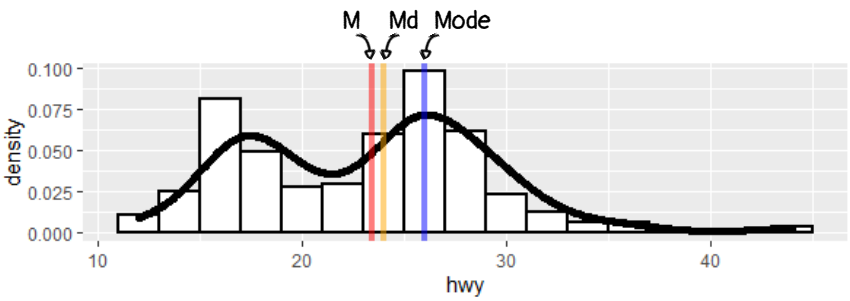

```
plot +
 geom_histogram(aes(y=..density..),
  colour="black",
  fill="white" ) +
 geom_density()
```



mean(mpg$hwy) = 23.4

median(mpg$hwy) = 24

library(**modeest**)
mlv(mpg$hwy) = 26



R include funzioni native per il calcolo della media e della mediana. Per la moda, è possibile creare la propria funzione o la Most Likely Values (mlv) dalla libreria modeest.

Fonti:

Downey, A. (2014). Think stats: exploratory data analysis. O'Reilly Media, Inc.

Field, A. (2022). An adventure in statistics: The reality enigma. Sage.