

Les Mesures de Tendance Centrale



COMIXPLAIN

Cette bande dessinée a été créée dans le cadre du projet de recherche Comixplain, financé par l'Innovation Call 2022 de l'Université des Sciences Appliquées de St. Pölten, en Autriche.

Équipe:

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

Contact:

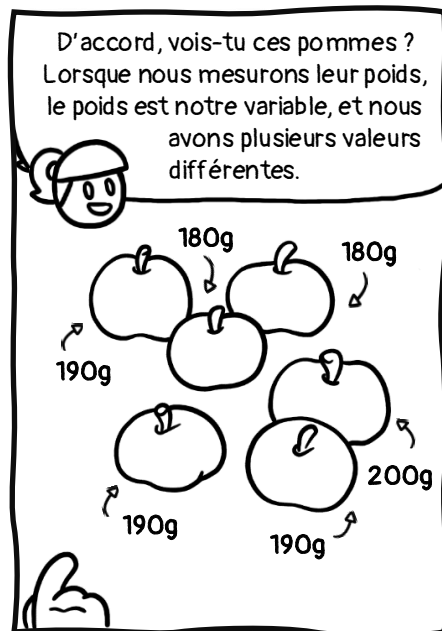
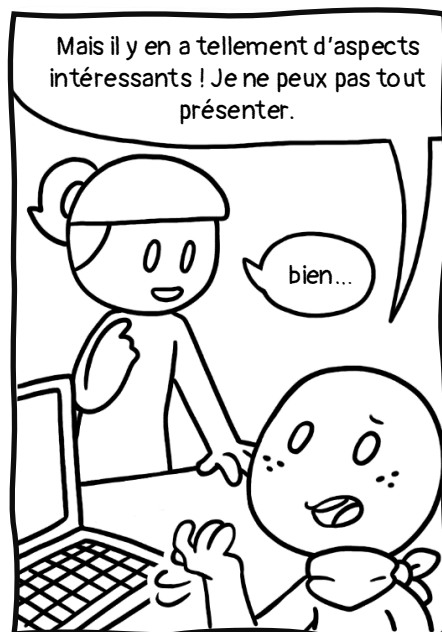
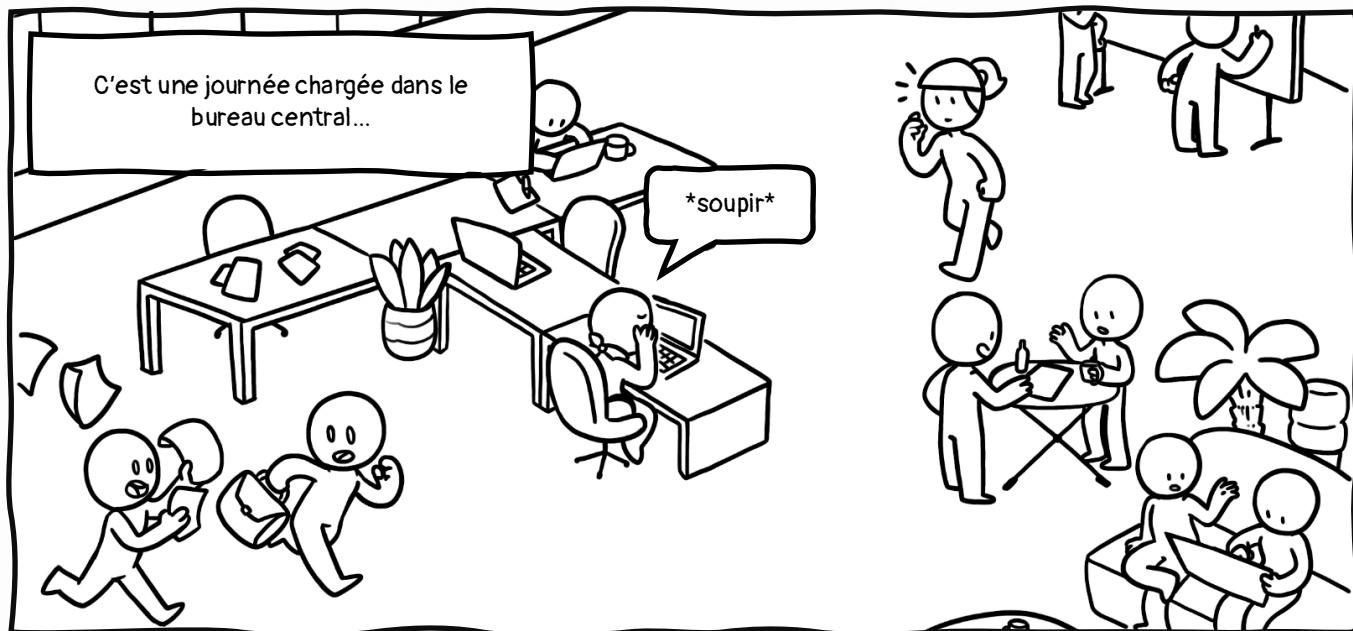
victor.oliveira@fhstp.ac.at

Illustrations:


Magdalena Boucher & Alena Ertl



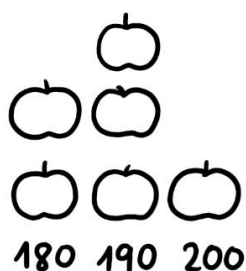
<https://fhstp.github.io/comixplain>



La meilleure façon de décrire une variable est de rapporter les valeurs et la fréquence à laquelle elle apparaît. C'est ce qu'on appelle la **DISTRIBUTION** de la variable.

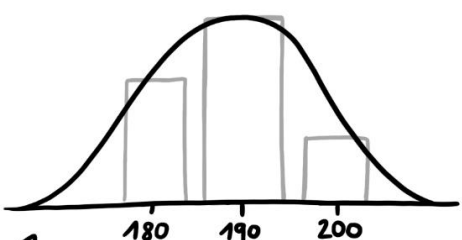


180 190 200





Si nous visualisons le poids de nos pommes, la répartition ressemblerait à ceci, puisque toutes les pommes de ce panier ont environ le même poids.

Poids des pommes


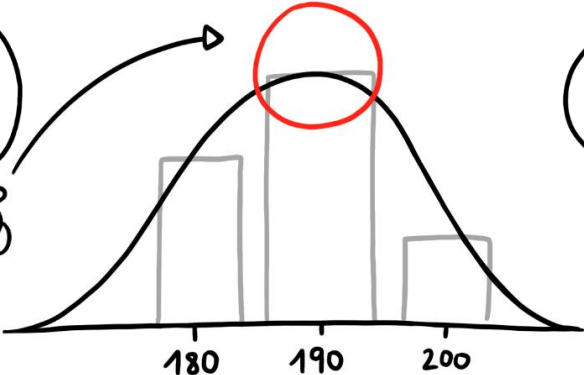


180 190 200

Elle a la forme d'une cloche...





S'il y avait d'autres pommes dans le panier, notre meilleure estimation de leur poids serait une valeur autour de ce point.

180 190 200

Ah, je vois. Mais... Comment puis-je calculer ce montant ?



Ce point de notre distribution représente bien nos données – nous l'appelons **TENDANCE CENTRALE**.

On peut résumer la tendance centrale de plusieurs façons. La **MOYENNE** est le moyen le plus courant, et elle est également facile à calculer.



Voici nos six pommes :

200 180 190 190 190 180

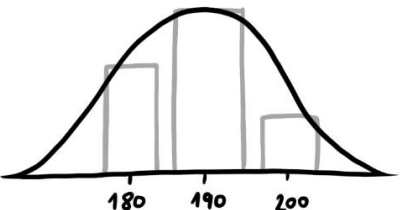
Nous allons faire l'addition de toutes les valeurs de poids...

200 + 180 + 190 + 190 + 190 + 180

... et divisons par le nombre de pommes que nous avons.


$(200 + 180 + 190 + 190 + 190 + 180) / 6$

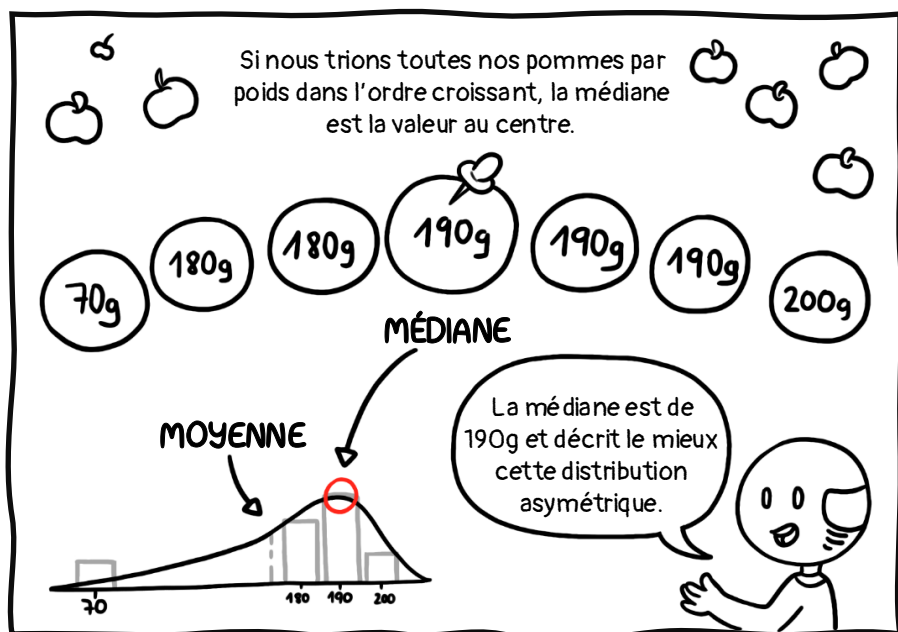
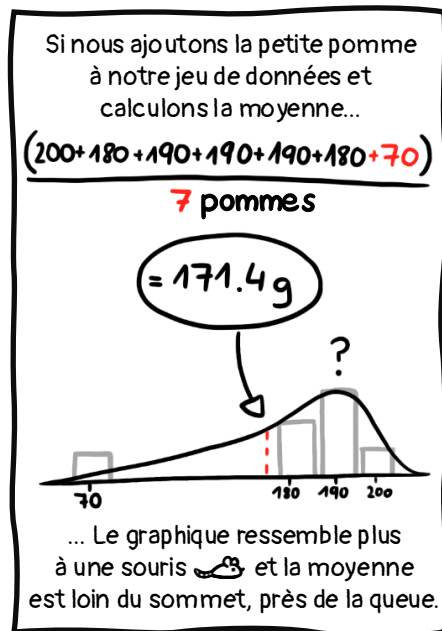
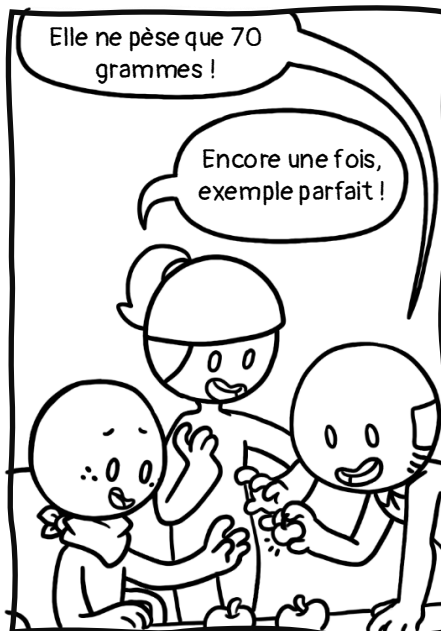
Cela nous donne une valeur moyenne de **188.3g**



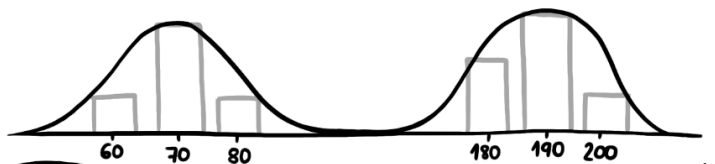
180 190 200

Ah, c'est presque le haut de notre graphique en cloche !





Dans ce cas, notre graphique a soudainement deux sommets :



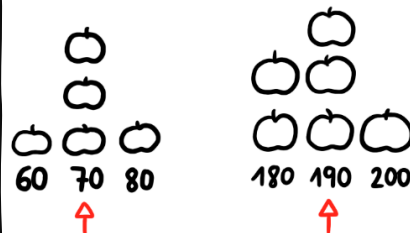
Deux souris amoureuses !

Ah... Oui.

Poids des grosses et petites pommes

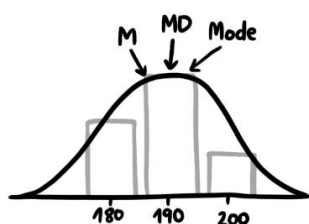
Quoi qu'il en soit, nous pouvons utiliser ce qu'on appelle **MODE** pour décrire la tendance centrale de notre distribution si elle a plusieurs sommets.

Le mode définit la ou les valeurs qui apparaissent le plus fréquemment dans un ensemble de données.



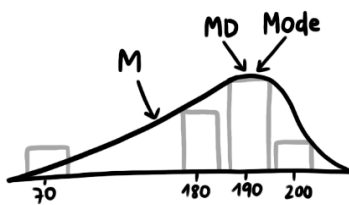
Dans ce cas, nous avons plusieurs modes. Mais parfois, il n'y aura qu'un mode, voire aucune.

Vous pouvez appliquer la moyenne, la médiane et le mode à différents échantillons de pommes. Mais souvent, certaines mesures représentent mieux les données que d'autres.



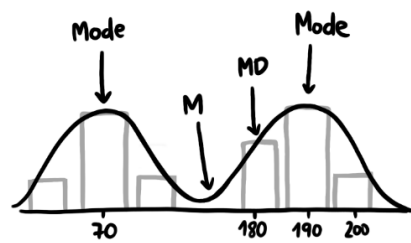
180, 180, 190, 190, 190, 200

$M = 188.3$ ↪ Meilleurs
 $MD = 190$ ↪ paramètres
 $Mode = 190$ ↪



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

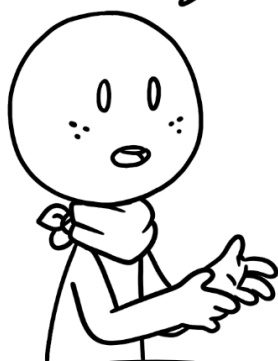
$M = 171.8$ ↪ Meilleurs
 $MD = 190$ ↪ paramètres
 $Mode = 191$ ↪



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

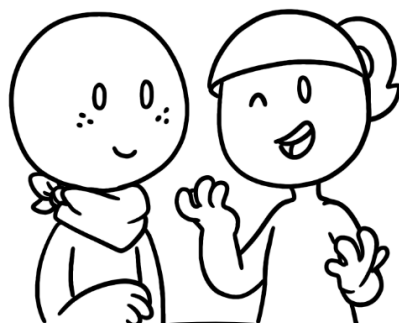
$M = 134.5$ ↪ Meilleur
 $MD = 180$ ↪ paramètre
 $Mode = 70 \text{ \& } 190$ ↪

J'ai beaucoup appris ! Il ne me reste plus qu'à l'appliquer aux données que je dois présenter. Elles proviennent d'une application qui suit les mesures de la fréquence cardiaque.



| 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 |

C'est faisable. Regarde tes données et suis les exemples des pommes! Tu pourras utiliser la page suivante.





Avant de tourner la page, essayez de calculer la moyenne, la médiane et le mode pour chaque variable, et vérifiez quel paramètre est le mieux adapté pour décrire la tendance centrale. Vous pouvez prendre des notes sur cette page !

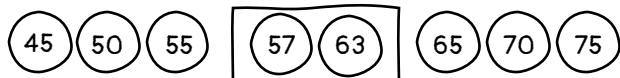
N'hésitez pas à vérifier vos calculs.
Vous pouvez prendre plus de notes sur cette page !

FRÉQUENCE CARDIAQUE

Calcul de la MOYENNE :

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

Calcul de la MÉDIANE :



S'il y a deux valeurs centrales, la moyenne des deux valeurs est la médiane:
 $(57+63)/2 = 60 \text{ bpm}$

Calcul du MODE :

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

Chaque valeur n'existe qu'une seule fois – le mode n'existe pas !

Si la distribution des valeurs est symétrique, sans distorsion, la moyenne est généralement égale à la médiane.



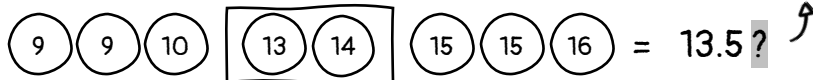
TEMPS D'ÉCRAN PLUS FRÉQUENT

Calcul de la MOYENNE :

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

Le temps d'écran n'est pas une valeur quantitative – donc calculer la moyenne et la médiane n'a aucun sens !

Calcul de la MÉDIANE :



Calcul du MODE :

9:00, 10:00, 13:00, 14:00, 15:00, 16:00
 2x 1x 1x 1x 2x 1x = 2 modes: 9:00 & 15:00

Le mode convient non seulement aux distributions multimodales, mais aussi au travail avec des données ordinales et catégorielles.



ÉTOILES

Calcul de la MOYENNE :

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

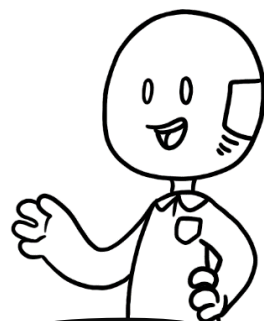
Calcul de la MÉDIANE :



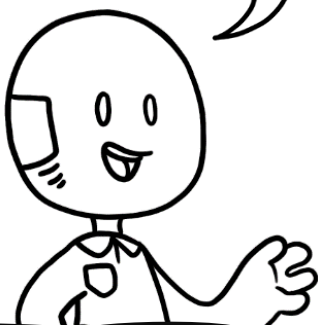
Calcul du MODE :

1 2 3 4 5
 1x 1x 1x 2x 3x = 5 étoiles

Dans les jeux de données dont la distribution est asymétrique, la médiane est une meilleure façon de décrire la tendance centrale.



Les langages de programmation, tels que **R**, vous aident à calculer la tendance centrale des attributs dans les grands ensembles de données. A l'aide de librairies R, telles que **tidyverse**, vous aident à visualiser la distribution des données.



| | 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 |
| + 227 autres articles | | | |

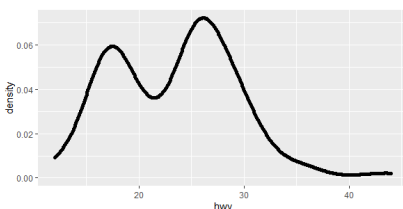
Dans tidyverse, vous avez accès aux données d'économie de carburant (**mpg**). Il comprend 11 attributs, tels que le modèle de voiture, l'année de fabrication et les kilomètres routiers par gallon (route).



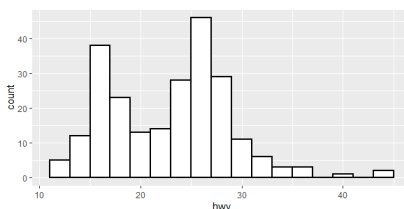
Vous pouvez utiliser **ggplot**, qui est inclus dans le tidyverse, pour visualiser la distribution des données de mile routier par gallon (route) à l'aide d'un histogramme, d'une courbe de densité ou les deux.

`install.packages("tidyverse")` # Installez-le uniquement la première fois que vous utilisez la bibliothèque

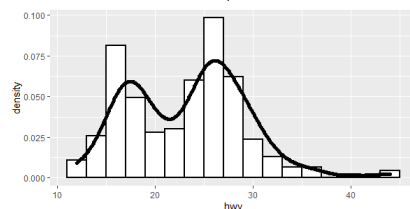
```
library(tidyverse)
plot <- ggplot(mpg, aes(x=hwy))
plot +
  geom_density()
```



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



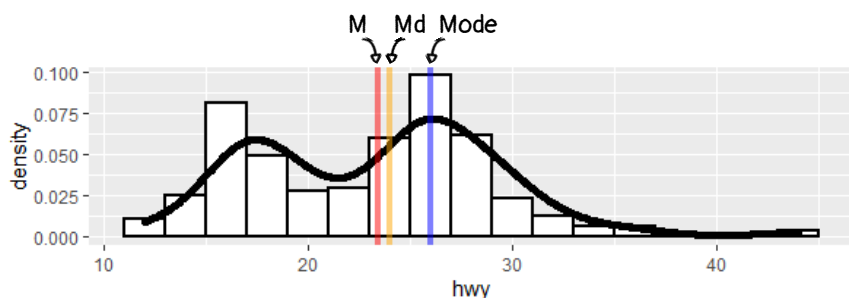
```
plot +
  geom_histogram(aes(y=..density..),
    colour="black",
    fill="white") +
  geom_density()
```



`mean(mpg$hwy)` = 23.4

`median(mpg$hwy)` = 24

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



R inclut des fonctions natives pour le calcul de la moyenne et de la médiane. Pour le mode, vous pouvez créer votre propre fonction ou utiliser les Most Likely Values (**mlv**) de la bibliothèque **modeest**.



Sources:

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

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