Module 2 : Mathematics for data science

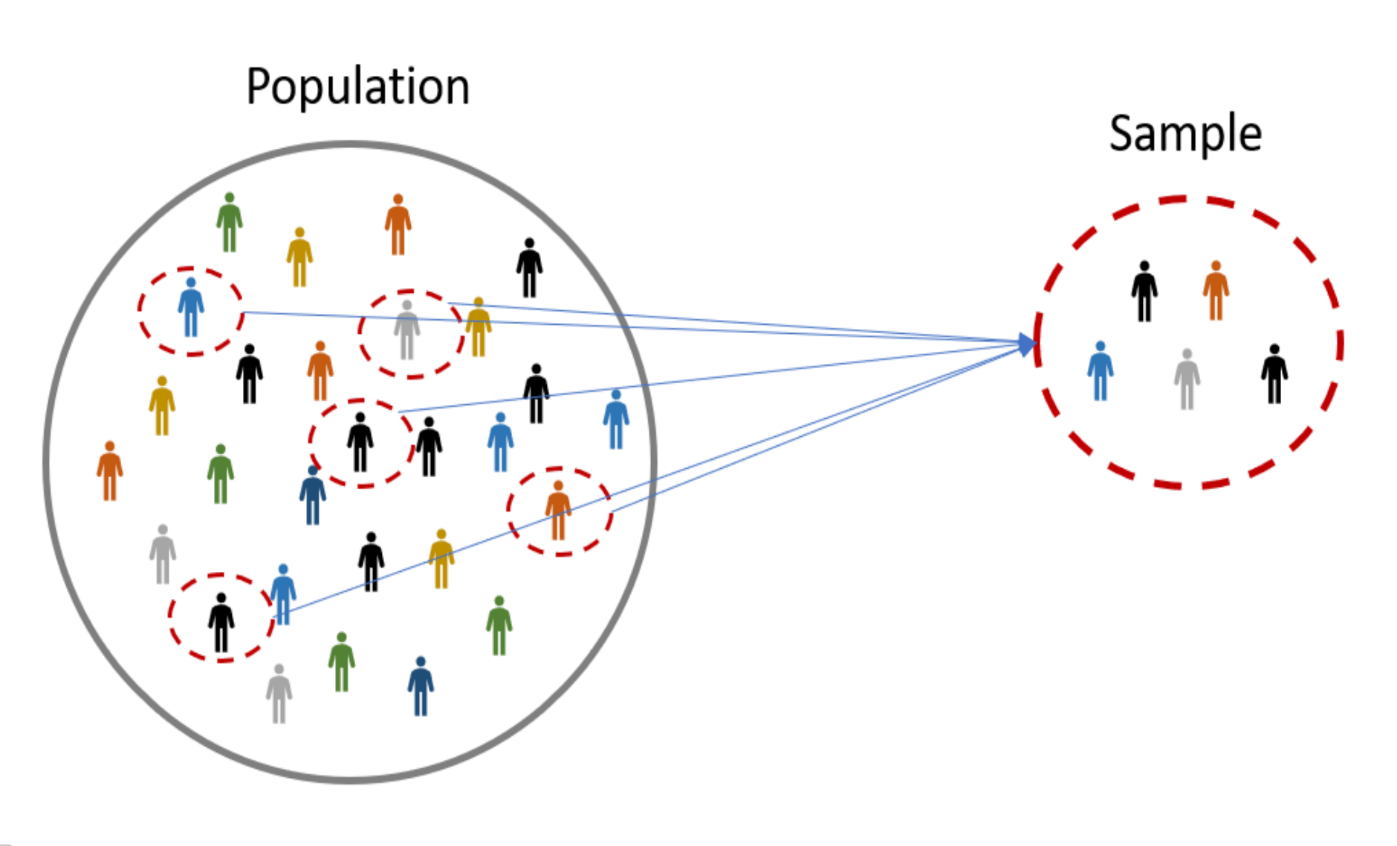
**Qu'est-ce qu'une population ?**

Une population est l'ensemble complet d'où est tiré un échantillon statistique.

****

En statistique, une population est un ensemble complet, c’est-à-dire chaque membre d’un groupe. Elle peut se référer à un groupe entier de personnes, d’objets, d’événements, de visites à l’hôpital ou de mesures. Parfois, il est possible d'enquêter sur chaque membre du groupe. Un exemple classique est le recensement, où il est obligatoire de répondre.

## Qu'est-ce qu'un échantillon ?

Un échantillon est simplement une partie d'une population. Il est utilisé dans les tests statistiques lorsque les tailles de population sont trop grandes pour que le test puisse inclure tous les membres ou observations possibles. Par exemple, si vous souhaitez savoir combien gagne en moyenne un Tunisien, vous ne voudriez pas sonder toute la population (plus de 12 millions de personnes), donc vous choisiriez un petit nombre de personnes dans la population. Par exemple, vous pourriez sélectionner 10 000 personnes.  
  


## Qu'est-ce qu'un paramètre ?

En mathématiques, un paramètre est une valeur qui est utilisée dans une équation. En statistiques, c'est une valeur qui vous renseigne sur une population.

## https://i.imgur.com/HVA7ewC.png

Un paramètre ne change jamais, car tout le monde (ou tout) a été sondé pour déterminer ce paramètre. Par exemple, si vous êtes intéressé par l'âge moyen de tout le monde en Tunisie, et que vous avez demandé à tout le monde et trouvé que l'âge moyen était de 38 ans, c'est un paramètre, car vous avez interrogé tout le monde dans le pays.

## Qu'est-ce qu'une statistique ?

Une statistique est une information que vous obtenez à partir d'un échantillon de la population, contrairement à un paramètre qui se réfère à une mesure basée sur tous les éléments de cette population.

**Qu'est-ce qu'une statistique ?**

Une statistique est une information que vous obtenez à partir d'un échantillon de la population, contrairement à un paramètre qui se réfère à une mesure basée sur tous les éléments de cette population.



**Ces ressources peuvent vous aider**

**population**

[https://en.wikipedia.org/wiki/Statistical\_population](https://en.wikipedia.org/wiki/Statistical_population" \t "_blank)

**sample**

[https://en.wikipedia.org/wiki/Sample\_(statistics)](https://en.wikipedia.org/wiki/Sample_(statistics)" \t "_blank)

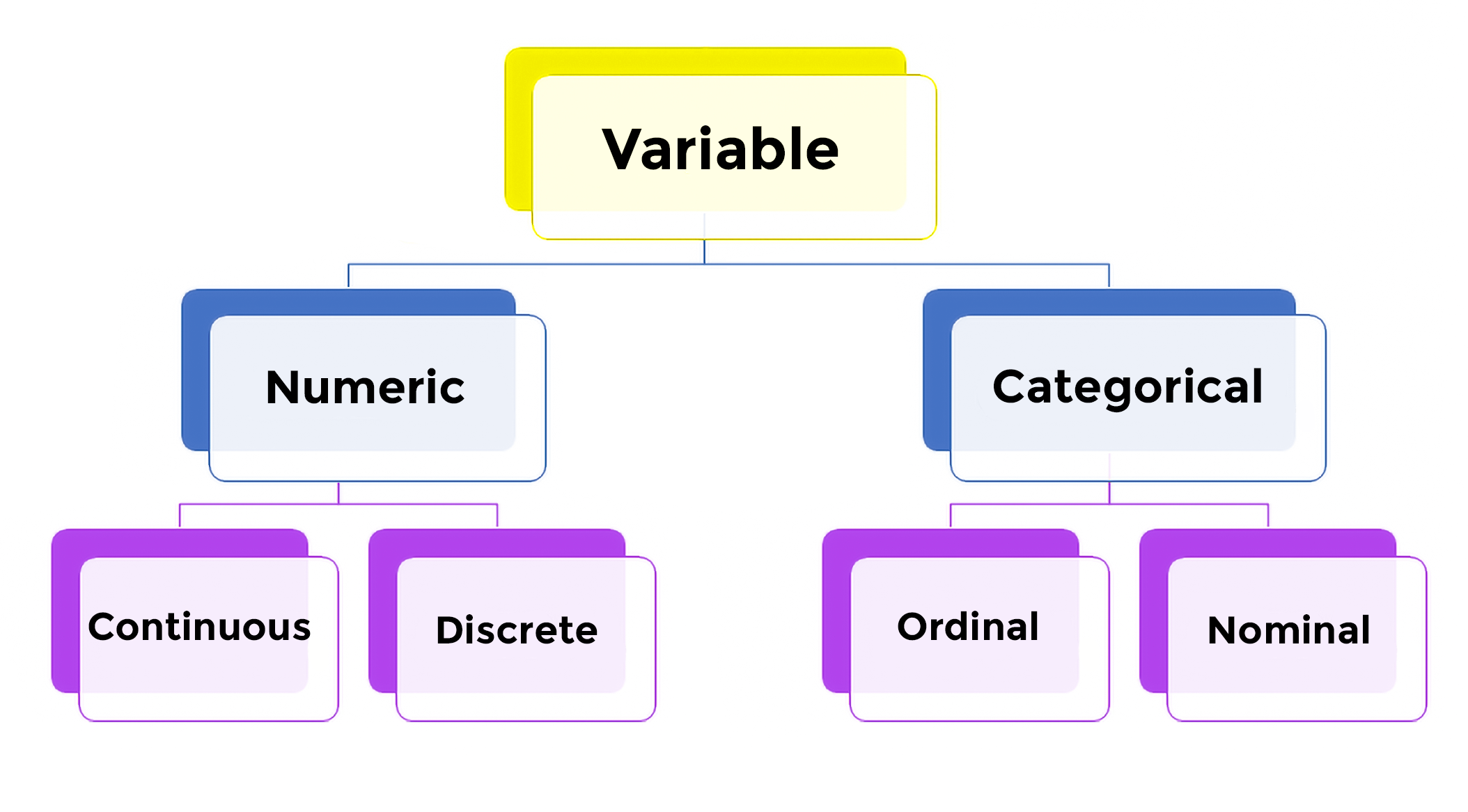
**population\_vs\_sample**

[https://www.scribbr.com/methodology/population-vs-sample/](https://www.scribbr.com/methodology/population-vs-sample/" \t "_blank)

**The Basic Terminologies of Statistics**

[https://statanalytica.com/blog/terminologies-of-statistics/](https://statanalytica.com/blog/terminologies-of-statistics/" \t "_blank)

**Importance des types de données :**

Il est important de comprendre le type de données avec lequel vous travaillez, car cela peut influencer les types d'analyses et de visualisations que vous pouvez effectuer. Il existe deux principaux types de données : numériques et catégoriques.  
  


**Données numériques :**

Les données numériques sont des valeurs qui peuvent être mesurées ou quantifiées, comme la taille, le poids ou l'âge. Il existe deux principaux types de données numériques : continues et discrètes. Les données continues peuvent prendre n'importe quelle valeur dans une plage donnée, comme le poids ou la taille. Les données discrètes ne peuvent prendre que des valeurs spécifiques, comme le nombre d'enfants dans un foyer.

**Données catégorielles :**

Les données catégoriques, également appelées données qualitatives, consistent en des données non numériques qui sont catégorisées en fonction de propriétés, attributs, étiquettes et autres identifiants. Des exemples de données catégoriques incluent la race, le sexe, le groupe d'âge et le niveau d'éducation. Les données catégoriques peuvent être divisées en deux types : les données nominales, qui permettent d'assigner des catégories sans ordre particulier, et les données ordinales, qui possèdent des catégories naturellement ordonnées mais dont les distances entre elles sont inconnues. Des exemples de données ordinales incluent les évaluations de satisfaction des clients et la gravité d'un bug logiciel, tandis que des exemples de données nominales incluent le genre et la couleur des cheveux.

**Ces ressources peuvent vous aider**

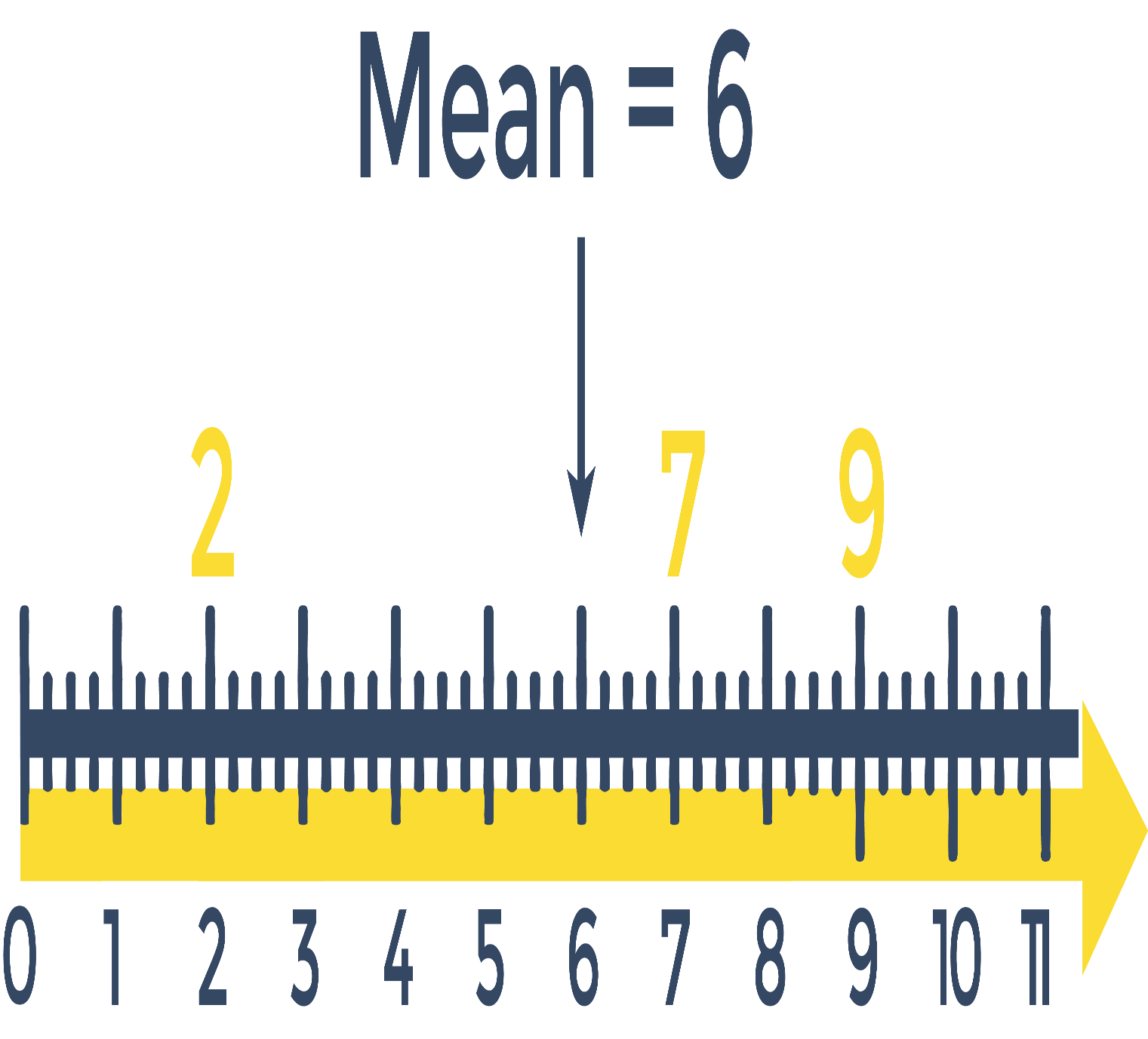
**types de données.**

[https://statisticsbyjim.com/basics/data-types/](https://statisticsbyjim.com/basics/data-types/" \t "_blank)

## ****Qu'est-ce que la moyenne ?****

La moyenne arithmétique est la moyenne d'un ensemble de nombres. Il est facile de la calculer :

* additionnez tous les nombres : 2 + 7 + 9 = 18
* divisez par le nombre de nombres (il y a 3 nombres) : 18 / 3 = 6

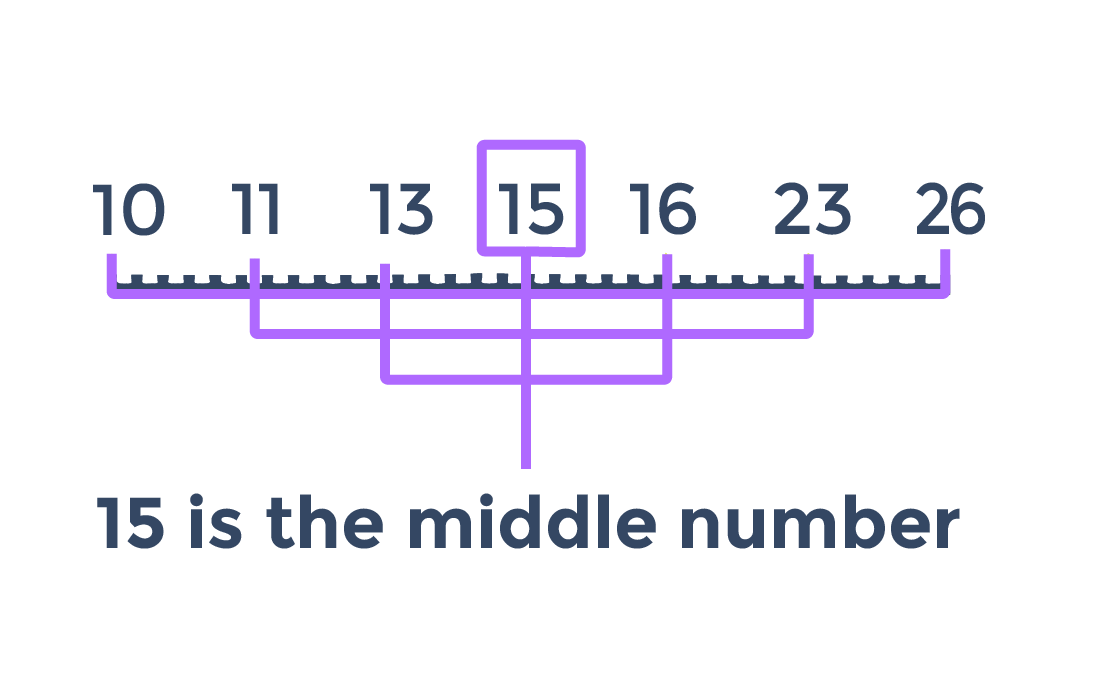


## ****Qu'est-ce que la mode ?****

La mode est la valeur qui apparaît le plus fréquemment dans un ensemble de données.  
Dans l'exemple ci-dessous, la mode de la série {6, 3, 9, 6, 6, 5, 9, 3} est 6, car elle se produit le plus souvent.  
  
  
  
Lorsque deux nombres apparaissent avec la même fréquence, nous utilisons simplement le plus élevé. Par exemple, dans cette série {3, 3, 3, 9, 16, 16, 16, 27, 37, 48}, le nombre 3 et le nombre 16 apparaissent tous deux trois fois. Ici, nous choisissons 16 comme mode.

## Qu'est-ce que la Médiane ?

La médiane est le "milieu" d'une liste de nombres triés.  
Mais avec un nombre pair de valeurs, les choses sont légèrement différentes. Dans ce cas, nous trouvons la paire de nombres du milieu, puis nous trouvons la valeur qui se trouve à mi-chemin entre eux (la moyenne entre les deux valeurs du milieu).



Prenons cette série comme exemple : 3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29. Lorsque nous mettons ces nombres en ordre, nous avons : 3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56.

Il y a maintenant quatorze nombres, donc nous n'avons pas un seul nombre du milieu, mais une paire de nombres du milieu : 3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56.

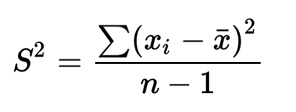
La meilleure façon est de trouver la valeur à mi-chemin entre eux (moyenne) : (21 + 23)/2 = 44/2 = 22. Donc, la médiane est 22.

**Ces ressources peuvent vous aider**

**Mean, Median, and Mode in Statistics**

[https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b#:~:text=Median%20is%20the%20middle%20value,mean%20of%20those%20two%20numbers).](https://medium.com/@nhan.tran/mean-median-an-mode-in-statistics-3359d3774b0b" \l ":~:text=Median%20is%20the%20middle%20value,mean%20of%20those%20two%20numbers)." \t "_blank)

**Variance**

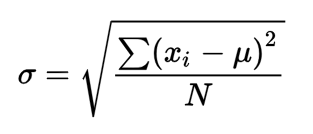
La variance est une mesure de la dispersion entre les nombres dans un ensemble de données. Autrement dit, elle évalue à quelle distance chaque nombre de l'ensemble se trouve par rapport à la moyenne, et donc par rapport à tous les autres nombres de l'ensemble. Elle est calculée en prenant la moyenne des écarts au carré par rapport à la moyenne.  
  


Où :

* x\_i = la valeur d'une observation
* x̅= la valeur moyenne de toutes les observations
* n = le nombre d'observations

**Écart type**

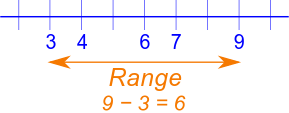
L'écart-type mesure la dispersion d'un ensemble de données par rapport à sa moyenne. Si les points de données sont plus éloignés de la moyenne, il y a une plus grande déviation au sein de l'ensemble de données. Ainsi, plus les données sont écartées, plus l'écart-type est élevé. Il est calculé comme la racine carrée de la variance.



Où :

* N = la taille de la population
* x\_i​ = chaque valeur de la population
* μ = la moyenne de la population

**Étendue**

L'étendue est la différence entre les valeurs les plus basses et les plus élevées. Exemple : Dans {4, 6, 9, 3, 7}, la valeur la plus basse est 3 et la valeur la plus élevée est 9.  
  
  
  


**Ces ressources peuvent vous aider**

[https://www.scribbr.com/statistics/variability/](https://www.scribbr.com/statistics/variability/" \t "_blank)

## ****Comprendre les Types de Données et les Mesures de Tendance Centrale****

### **Instructions de l'exercice :**

**Ensemble de données :**{45, 52, 60, 70, 80, 90, 100, 110}

Examinez l'ensemble de données fourni et identifiez s'il s'agit de données numériques ou catégoriques. S'il s'agit de données numériques, calculez la moyenne, le mode et la médiane. S'il s'agit de données catégoriques, indiquez qu'il n'est pas possible de calculer la moyenne, le mode et la médiane. Comparez vos résultats avec les réponses fournies et expliquez les éventuelles divergences.

**Remarque :** Il s'agit d'un exemple simple ; dans un scénario réel, l'ensemble de données serait plus vaste et plus complexe.

### **Instructions de l'exercice expliquées :**

Dans votre exercice, vous avez un ensemble de données : {45, 52, 60, 70, 80, 90, 100, 110}.

1. **Identifier le Type de Données :**
   * Tout d'abord, déterminez si l'ensemble de données est numérique ou catégorique. Dans ce cas, l'ensemble de données est numérique car il se compose de nombres pouvant représenter des mesures ou des quantités.
2. **Calculer les Mesures de Tendance Centrale :**
   * Puisqu'il s'agit de données numériques, vous pouvez calculer la moyenne, la médiane et le mode.
     + **Moyenne :** Ajoutez tous les nombres ensemble et divisez par le nombre de valeurs.
     + **Médiane :** Arrangez les nombres dans l'ordre et trouvez le nombre du milieu.
     + **Mode :** Recherchez le nombre le plus fréquemment apparaissant dans l'ensemble.
3. **Comparer les Résultats :**
   * Après avoir calculé ces mesures, comparez-les avec les réponses fournies pour vérifier leur exactitude.
   * S'il y a des divergences, essayez de déterminer s'il y a eu une erreur dans les calculs ou dans la compréhension des concepts.

Calculons la moyenne, la médiane et le mode pour l'ensemble de données donné .

Sur la base des calculs pour l'ensemble de données {45, 52, 60, 70, 80, 90, 100, 110} :

* **Moyenne (Moyenne):** La moyenne est de 75,875. Elle est calculée en additionnant tous les nombres (45 + 52 + 60 + 70 + 80 + 90 + 100 + 110 = 607), puis en divisant par le nombre de valeurs de l'ensemble (607 ÷ 8).
* **Médiane (valeur moyenne):** La médiane est 75,0. Pour la trouver, il faut d'abord classer les nombres dans l'ordre, puis trouver le nombre du milieu. Dans ce cas, comme il y a huit nombres, la médiane est la moyenne des quatrième et cinquième nombres de la liste triée (70 et 80), soit (70 + 80) ÷ 2 = 75.
* **Mode (valeur la plus fréquente):** Dans cet ensemble de données, il n'y a pas de mode car toutes les valeurs sont uniques et aucun nombre ne se répète.

Cet exemple illustre comment déterminer et calculer des mesures statistiques de base pour un ensemble de données numériques. Il est important de d'abord identifier le type de données avec lequel vous travaillez (numérique ou catégorique) pour déterminer les mesures statistiques appropriées à utiliser. Dans les scénarios réels, les ensembles de données sont souvent plus vastes et plus complexes, ce qui peut affecter la facilité et la méthode de calcul, mais pas les principes sous-jacents.

**Ces ressources peuvent vous aider**

**types de données.**

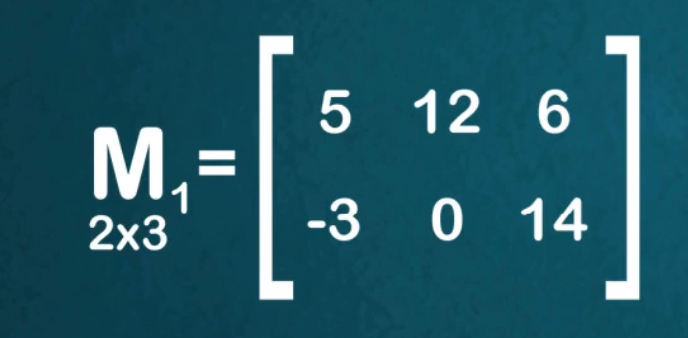
[https://statisticsbyjim.com/basics/data-types/](https://statisticsbyjim.com/basics/data-types/" \t "_blank)

## Qu'est-ce qu'une matrice ?

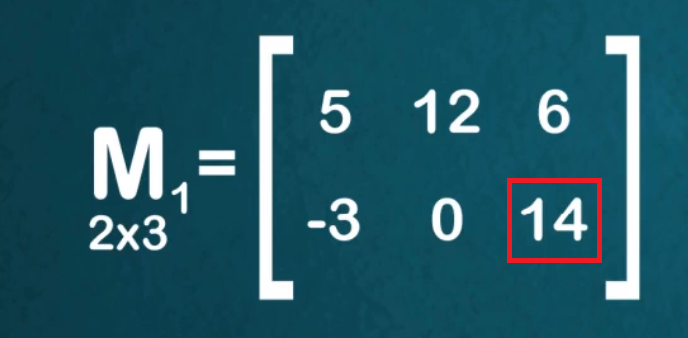
Bienvenue à la première leçon d'algèbre linéaire ! Notre voyage commence avec les matrices. Savez-vous ce qu'est une matrice ? Une matrice est un ensemble de nombres disposés en lignes et en colonnes, comme celle-ci.

## Matrices

Elle est utilisée en mathématiques pour effectuer des opérations telles que l'addition, la soustraction, la multiplication et la division. En algèbre linéaire, les matrices sont généralement désignées par une lettre majuscule, comme M1, et leur taille est décrite par le nombre de lignes et de colonnes, comme 2x3.



Les éléments d'une matrice peuvent également être référencés par leur position, comme l'élément de la deuxième ligne et de la troisième colonne de la matrice A serait M1(23), qui est 14.



Ainsi, c'est le bloc de construction fondamental de l'algèbre linéaire. Il peut avoir n'importe quelle taille et contenir différents éléments tels que des nombres, des expressions ou des symboles, et il est utilisé pour effectuer des opérations mathématiques.

**Ces ressources peuvent vous aider**

**Introduction aux matrices**

[https://www.slideshare.net/shakehandwithlife/introduction-of-matrices](https://www.slideshare.net/shakehandwithlife/introduction-of-matrices" \t "_blank)

**Pourquoi avons-nous besoin des matrices ?**

[https://news.mit.edu/2013/explained-matrices-1206](https://news.mit.edu/2013/explained-matrices-1206" \t "_blank)

**Pourquoi avons-nous besoin des matrices ?**

[https://news.mit.edu/2013/explained-matrices-1206](https://news.mit.edu/2013/explained-matrices-1206" \t "_blank)

