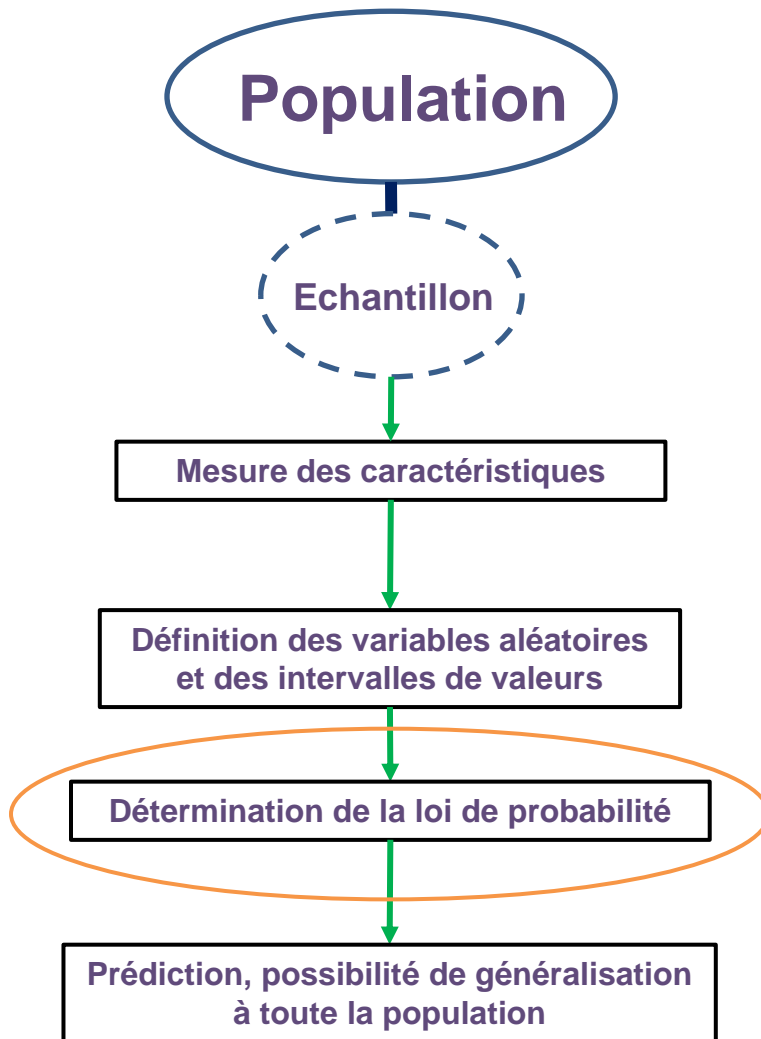




## **B4 – Mathématiques :** Statistique Descriptive

Amine ILMANE

# Le module B4 – Mathématiques



Dans ce chapitre nous allons voir comment déterminer la loi de probabilité ou à défaut des caractéristiques tel que **la moyenne** ou **l'écart-type** ...

Les caractéristiques se regroupent en **deux catégories** :





- Celles qui mesurent **la tendance centrale**
- Celles qui mesurent **la dispersion**

# 206neutrinos

No mean task here...

## Desperately seeking sterile

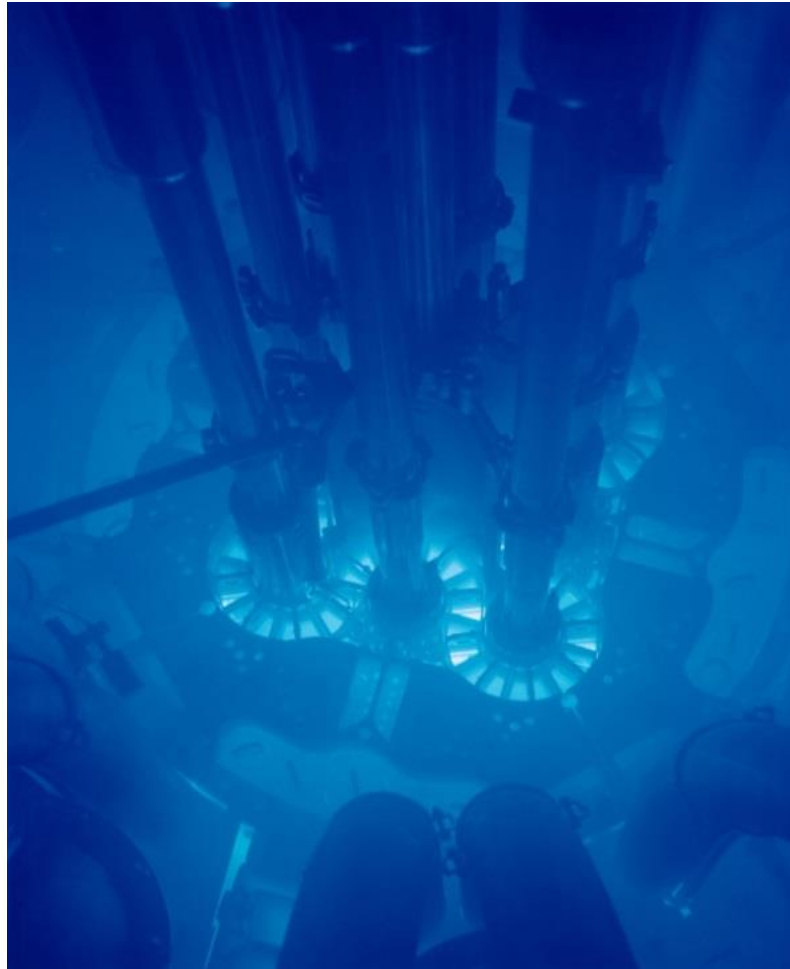
The three known types of neutrino might be "balanced out" by a bashful fourth type

ELECTRON NEUTRINO	MUON NEUTRINO	TAU NEUTRINO	STERILE NEUTRINO
			
MASS	< 1 electronvolt		> 1 electronvolt
FORCES THEY RESPOND TO	Weak force Gravity		Gravity
DIRECTION OF SPIN	All three "left handed"		"Right handed"

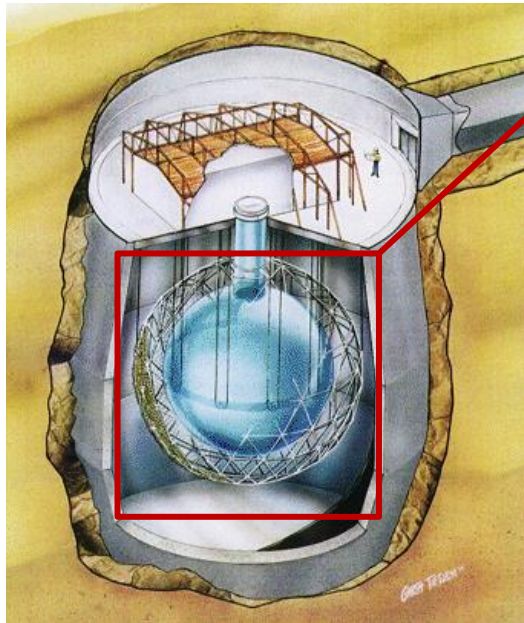


# Mesure de la vitesse des neutrinos

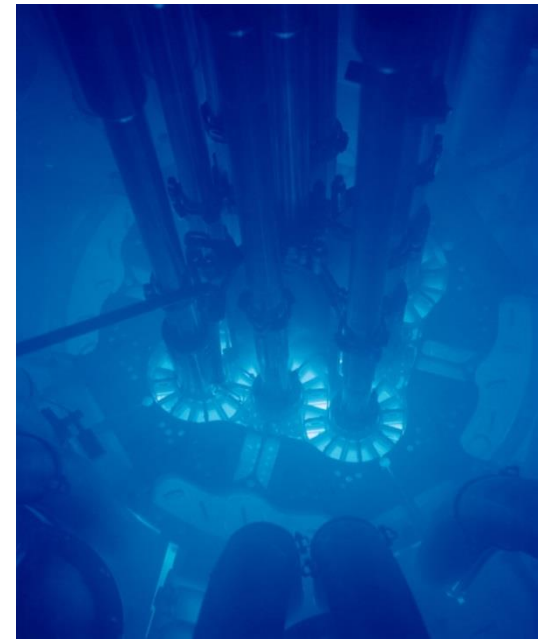
**Effet Cerenkov** dans un réacteur nucléaire



# Mesure de la vitesse des neutrinos



**Effet Cerenkov**

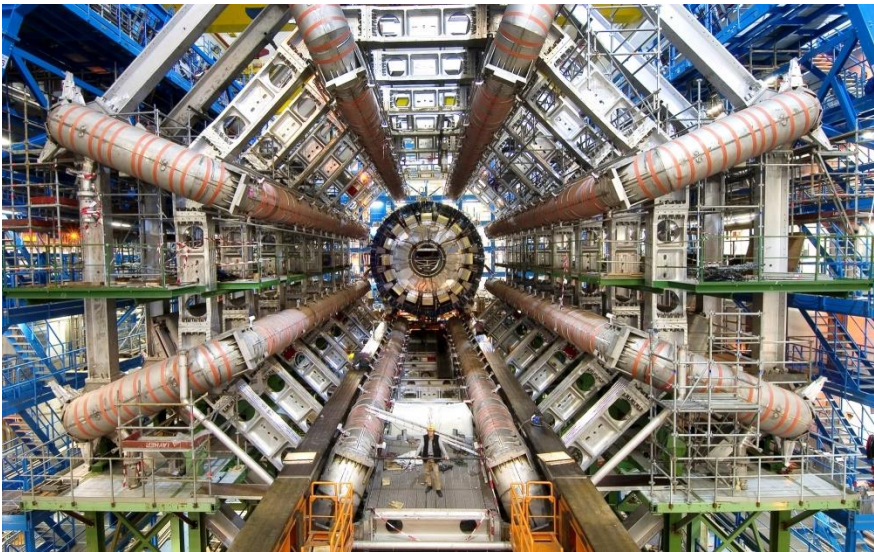


**Réacteur nucléaire**



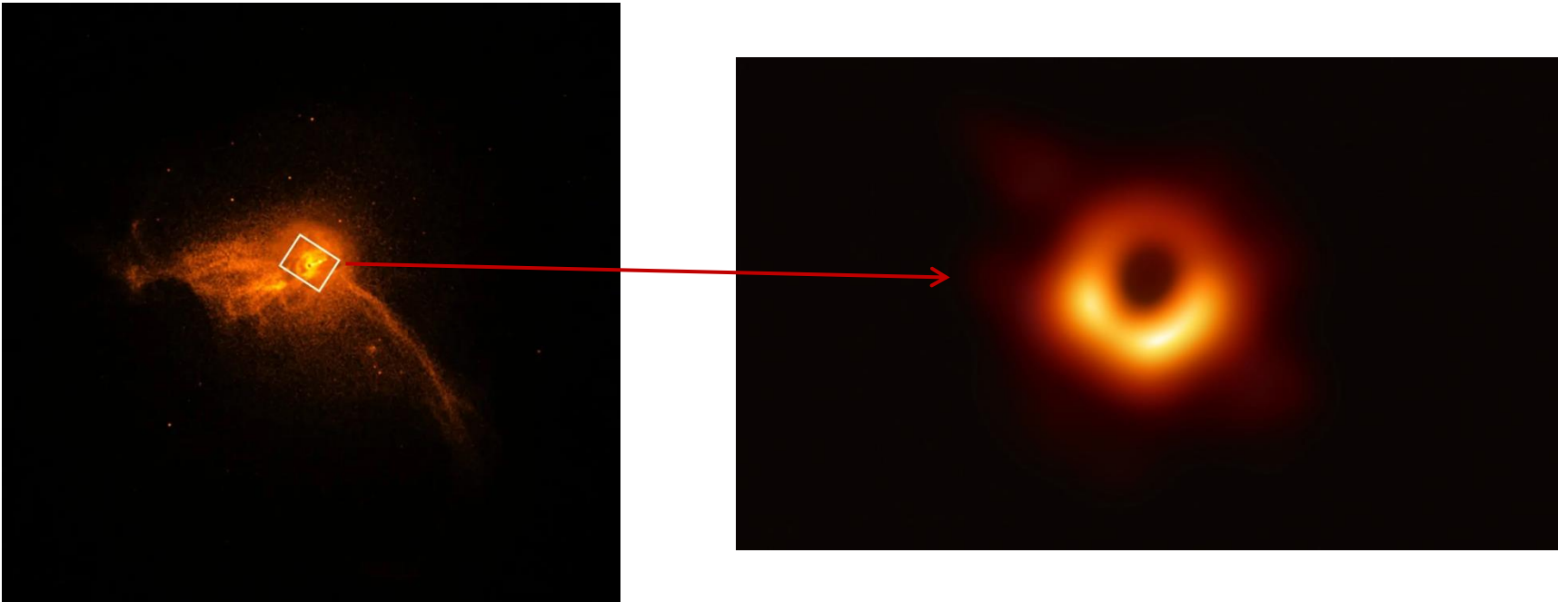
# Acquisition des données

Le problème " **how to manage Big data ?** " se résout en partie en utilisant des techniques de statistiques, comme nous allons le voir dans de ce chapitre.



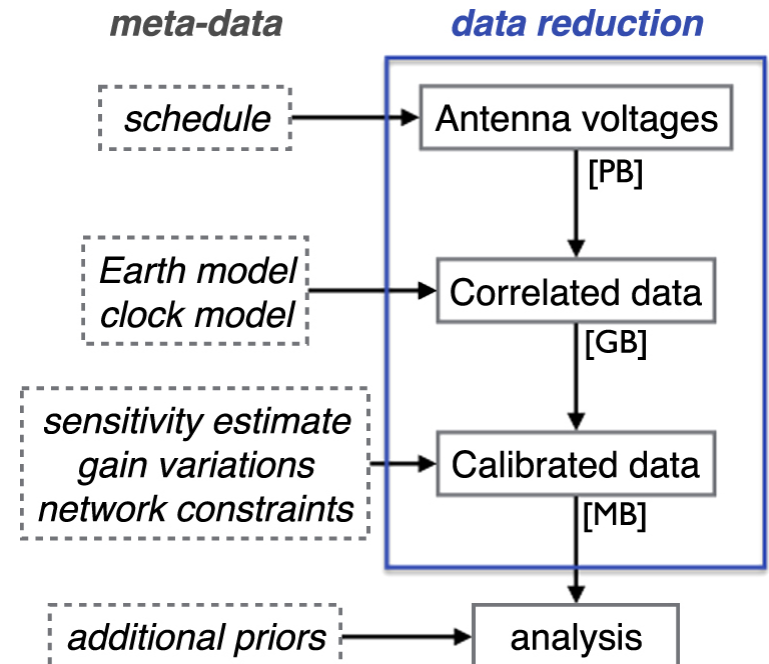
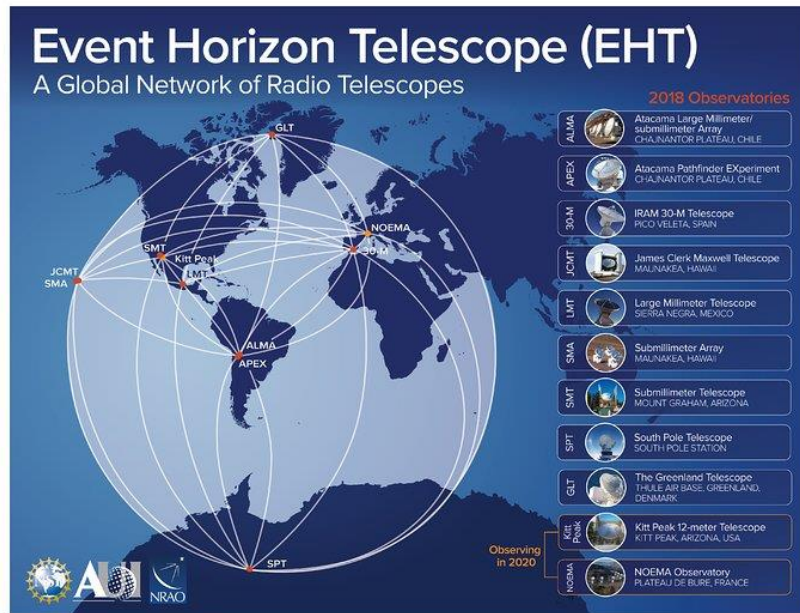
# Acquisition des données

Le problème " **how to manage Big data ?** " se résout en partie en utilisant des techniques de statistiques, comme nous allons le voir dans de ce chapitre.



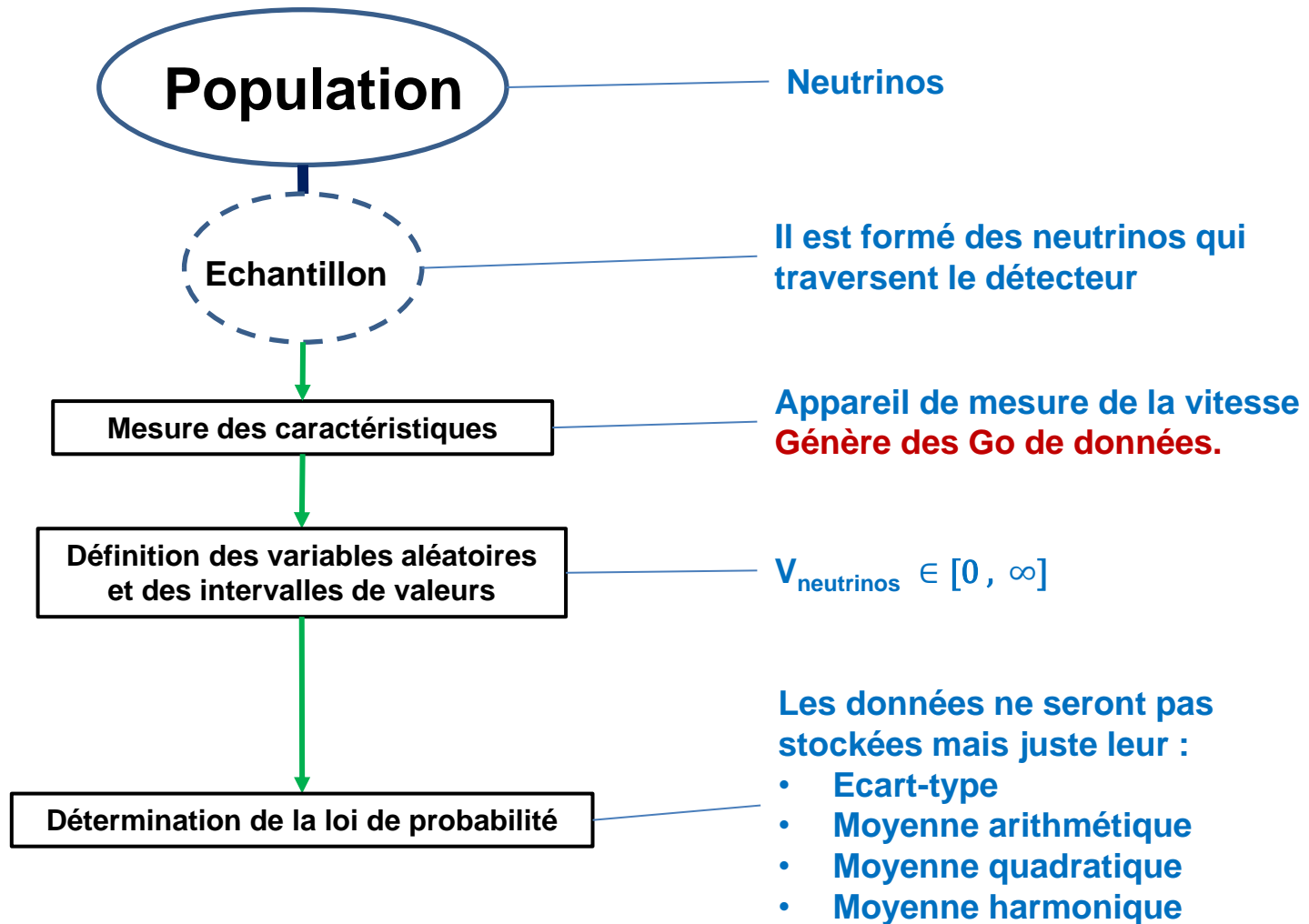
# Acquisition des données

Le problème " **how to manage Big data ?** " se résout en partie en utilisant des techniques de statistiques, comme nous allons le voir dans de ce chapitre.





# Projet 206 : neutrinos



# Projet 206 : neutrinos

```
Terminal
~/B-MAT-400> ./206neutrinos -h
USAGE
  ./206neutrinos n a h sd

DESCRIPTION
  n      number of values
  a      arithmetic mean
  h      harmonic mean
  sd     standard deviation
```

Le nombre de données déjà traitées

Caractéristiques extraites de ces données

```
Terminal
~/B-MAT-400> ./206neutrinos 12000 297514 297912 4363
Enter next value: 299042
  Number of values:    12001
  Standard deviation:  4362.84
  Arithmetic mean:     297514.13
  Root mean square:    297546.11
  Harmonic mean:       297912.09

Enter next value: 302420
  Number of values:    12002
  Standard deviation:  4362.89
  Arithmetic mean:     297514.54
  Root mean square:    297546.52
  Harmonic mean:       297912.46

Enter next value: END
```

En se basant sur les données précédentes déterminées les nouvelles valeurs des caractéristiques.

# Chapitre 6 : Statistique descriptive

- Définitions
- Mesure de tendance centrale
- Mesure de dispersion



# Mesure de tendance Centrale

**Au lieu de stocker toutes les données nous les remplaçant par leur moyenne et nous stockerons que la moyenne.**

Pour bien comprendre l'idée, imaginons que ces données sont vos dépenses journalières :

*Au bout de 5 jours, vous aurez dépensé **24 euros**, ce qui revient à dépenser **4,8 euros/jour***

indice	Données	Tendance centrale
1	5	4,8
2	8	4,8
3	1	4,8
4	6	4,8
5	4	4,8
....	...	...

- Ce dernier chiffre est plus simple à retenir, et permet de rapidement d'estimer un total rien qu'on connaissant le nombre de jour !! De plus, les données ne seront pas stockées, juste la valeur moyenne.
- Mais il existe différents des moyennes : **le choix d'une moyenne dépend de la façon dont une caractéristique s'exprime en fonction de la données.**

# Moyenne arithmétique

$$\bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Lorsque chaque donnée  $x_i$  admet un effectif  $e_i$ , on parle de moyenne arithmétique **pondérée**

$$\bar{x} = \frac{e_1 \cdot x_1 + e_2 \cdot x_2 + \dots + e_n \cdot x_n}{e_1 + e_2 + \dots + e_n} = \frac{1}{N} \sum_{i=1}^n e_i \cdot x_i \quad N = e_1 + e_2 + \dots + e_n$$

Trouver la relation de récurrence entre  $\bar{x}_{n+1}$  **et**  $\bar{x}_n$

# Moyenne arithmétique

## Formules de récurrence

$$\bar{x}_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

on pose  $S = x_1 + x_2 + \cdots + x_n$  ce qui donne  $\bar{x}_n = \frac{S}{n} \Rightarrow S = \bar{x}_n \times n$


$$\bar{x}_{n+1} = \frac{x_1 + x_2 + \cdots + x_n + x_{n+1}}{n+1} = \frac{S + x_{n+1}}{n+1}$$

$$\bar{x}_{n+1} = \frac{\bar{x}_n \times n + x_{n+1}}{n+1}$$



# Moyenne quadratique (root mean square)

Elle est liée à l'écart-type  
et la moyenne arithmétique



$$Q = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \dots + x_n^2)}$$

**Exemple**

8, 10, 9, 12, 13

$$Q = \sqrt{\frac{1}{5}(8^2 + 10^2 + 9^2 + 12^2 + 13^2)} = \sqrt{\frac{558}{5}} = \sqrt{111,6} \simeq 10,564$$

# Moyenne quadratique (root mean square)

- Nous pouvons utiliser notre donnée  $\mathbf{x}_i$  pour calculer une caractéristique  $\mathbf{F}$  qui dépend d'elle.
- Dans le cas où la caractéristique est  $\mathbf{F} = \mathbf{x}_i^2$ , il serait faux d'utiliser la moyenne arithmétique pour calculer une valeur moyenne de  $\mathbf{F}$ , **il faut utiliser la moyenne quadratique**

$$var = Q^2 - A^2$$

## Exemple

$$E_{cinétique} = \frac{1}{2} m (v_1^2 + v_2^2 + \dots + v_n^2) = \frac{1}{2} m n V^2$$

$$(v_1^2 + v_2^2 + \dots + v_n^2) = n V^2 \quad \rightarrow \quad V^2 = \frac{v_1^2 + v_2^2 + \dots + v_n^2}{n}$$

$$V_{quadratique} = \sqrt{\frac{v_1^2 + v_2^2 + \dots + v_n^2}{n}}$$

# Moyenne harmonique

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}}$$

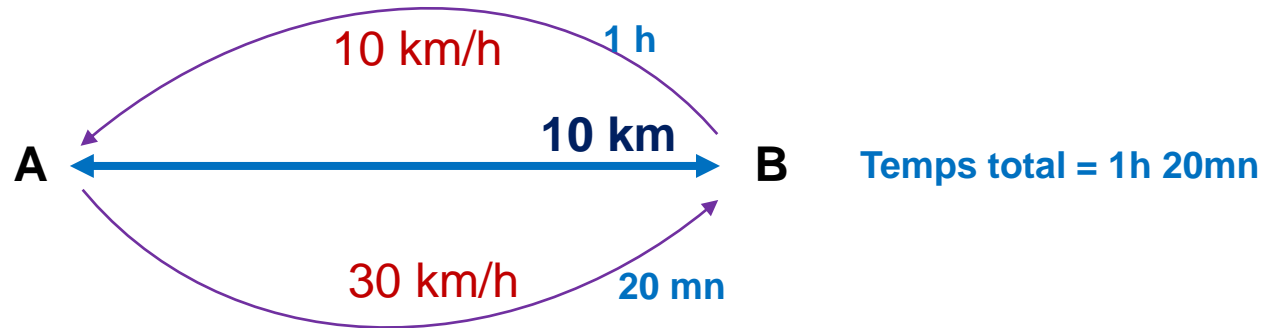
**Exemple**

8, 10, 9, 12, 13

$$H = \frac{5}{\frac{1}{8} + \frac{1}{10} + \frac{1}{9} + \frac{1}{12} + \frac{1}{13}} \simeq \frac{5}{0,496} \simeq 10,073$$

# Moyenne harmonique

Lorsqu'il s'agit de données dont les unités sont composées (comme la vitesse : m/s) la vitesse moyenne n'est pas celle que l'on croit



$$\bar{v} = \frac{v_1 + v_2}{2} = 20 \text{ km/h}$$

20 km en 1 h

$$\bar{v}_{\text{harmonique}} = \frac{2}{\frac{1}{10} + \frac{1}{30}} = 15 \text{ km/h}$$

20 km en en 1h 20mn

# Moyenne géométrique

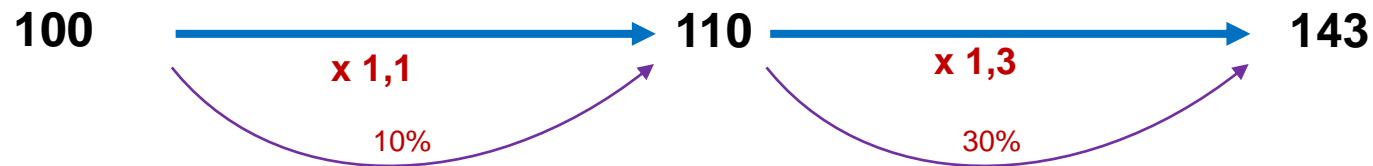
$$G = \sqrt[n]{x_1 \cdot x_2 \dots x_n} = (x_1 \cdot x_2 \dots x_n)^{\frac{1}{n}}$$

$$\log(G) = \frac{1}{n} \cdot (\log(x_1) + \log(x_2) + \dots + \log(x_n)) = \frac{1}{n} \cdot \sum_{i=1}^n \log(x_i)$$

**Exemple**    8, 10, 9, 12, 13     $G = \sqrt[5]{8 \cdot 10 \cdot 9 \cdot 12 \cdot 13} = \sqrt[5]{112320} \simeq 10,235$

# Moyenne géométrique

Lorsqu'il s'agit de données qui sont des **coefficient multiplicatif** par exemple les taux de variation d'un indice boursier (en pourcentages)



$$\bar{p} = \frac{1,1+1,3}{2} = 1,2$$

$$100 \times 1,2 = 120$$

$$120 \times 1,2 = 144$$

$$\bar{p}_{\text{géométrique}} = \sqrt[2]{1,1 \cdot 1,3} = 1,19583$$

$$100 \times 1,19583 = 119,583$$

$$119,583 \times 1,19583 = 143,001$$



# Benchmark des moyennes

$$H \leq G \leq \bar{x} \leq Q$$

$$10,073 \leq 10,235 \leq 10,4 \leq 10,564$$

## Exemple

8, 10, 9, 12, 13

$$\bar{x} = 10,4$$

$$Q \simeq 10,564$$

$$G \simeq 10,235$$

$$H \simeq 10,073$$

# Médiane

$$x_1 \leq x_2 \leq \dots \leq x_n$$

Si  $n$  est impair,  $n=2p+1$

$$med = x_{p+1}$$

**Exemple**    8, 9, 10, 12, 13             $p = 2$      $5 = 2 \times 2 + 1$      $med = x_3 = 10$

---

Si  $n$  est pair,  $n=2p$

$$med = \frac{x_p + x_{p+1}}{2}$$

**Exemple**    2, 5, 8, 10, 14, 15             $p = 3$      $6 = 2 \times 3$      $med = \frac{x_3 + x_4}{2} = \frac{8 + 10}{2} = 9$

# Mesure de dispersion

## Variance et écart-type

Trouver le lien avec la moyenne quadratique

$$\text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Pour calculer la moyenne quadratique on peut utiliser le fait que  $\text{Var}(x) = E(x^2) - E(x)^2$

$$\sigma = \sqrt{\text{Var}(x)}$$

Exemple 8, 9, 10, 12, 13

$$\text{Var}(x) = \frac{8^2 + 10^2 + 9^2 + 12^2 + 13^2}{5} - 10,4^2 = \frac{558}{5} - 108,16 = 3,44$$

$$\sigma = \sqrt{3,44} \simeq 1,855$$

## Écart moyen

$$\overline{E} = \frac{1}{n} \cdot \sum_{i=1}^n |x_i - \bar{x}|$$

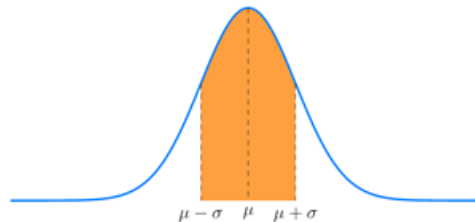
Exemple 8, 9, 10, 12, 13

$$\overline{E} = \frac{1}{5} \cdot (|8-10,4| + |10-10,4| + |9-10,4| + |12-10,4| + |13-10,4|) = \frac{8,4}{5} = 1,68$$

## Concentration autour de la moyenne

Dans l'intervalle  $[\mu - \sigma, \mu + \sigma]$  centré autour de la moyenne  $\mu$ , il y a 68% de la masse de la distribution  $\mathcal{N}(\mu, \sigma^2)$

$$\mathbb{P}[\mu - \sigma \leq X \leq \mu + \sigma] \simeq 0.68.$$



# Variance et écart-type

Trouver le lien avec la  
moyenne quadratique

$$Var(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Pour calculer la moyenne quadratique on peut utiliser le fait que  $Var(x) = E(x^2) - E(x)^2$

$$\sigma = \sqrt{Var(x)}$$

$$var = Q^2 - A^2$$

Exemple

8, 9, 10, 12, 13

$$Var(x) = \frac{8^2 + 10^2 + 9^2 + 12^2 + 13^2}{5} - 10,4^2 = \frac{558}{5} - 108,16 = 3,44$$

$$\sigma = \sqrt{3,44} \simeq 1,855$$

# Écart moyen

$$\overline{E} = \frac{1}{n} \cdot \sum_{i=1}^n |x_i - \overline{x}|$$

**Exemple**     8, 9, 10, 12, 13

$$\overline{E} = \frac{1}{5} \cdot (|8 - 10,4| + |10 - 10,4| + |9 - 10,4| + |12 - 10,4| + |13 - 10,4|) = \frac{8,4}{5} = 1,68$$

# Écart médian

$$E_m = \frac{1}{n} \cdot \sum_{i=1}^n |x_i - med|$$

**Exemple**      8, 9, 10, 12, 13       $8 \leq 9 \leq 10 \leq 12 \leq 13$

$$E_m = \frac{1}{5} \cdot (|8 - 10| + |10 - 10| + |9 - 10| + |12 - 10| + |13 - 10|) = \frac{8}{5} = 1,6$$