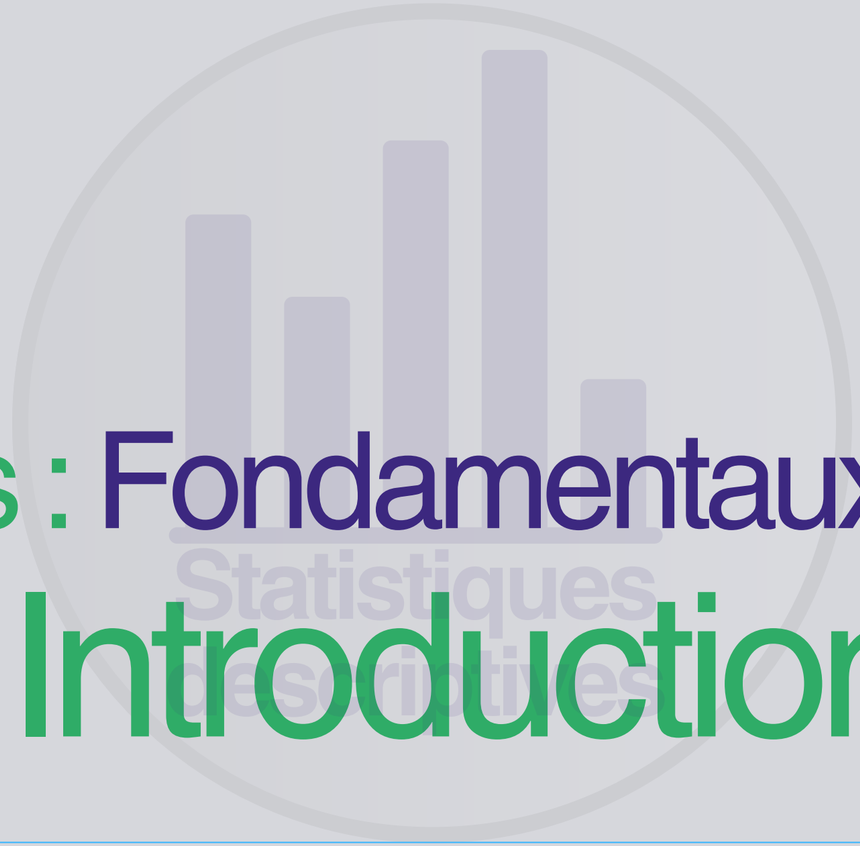
bstorm TRAINING

Statistiques descriptives : Fondamentaux

bstorm TRAINING

Statistiques descriptives : Fondamentaux

Introduction



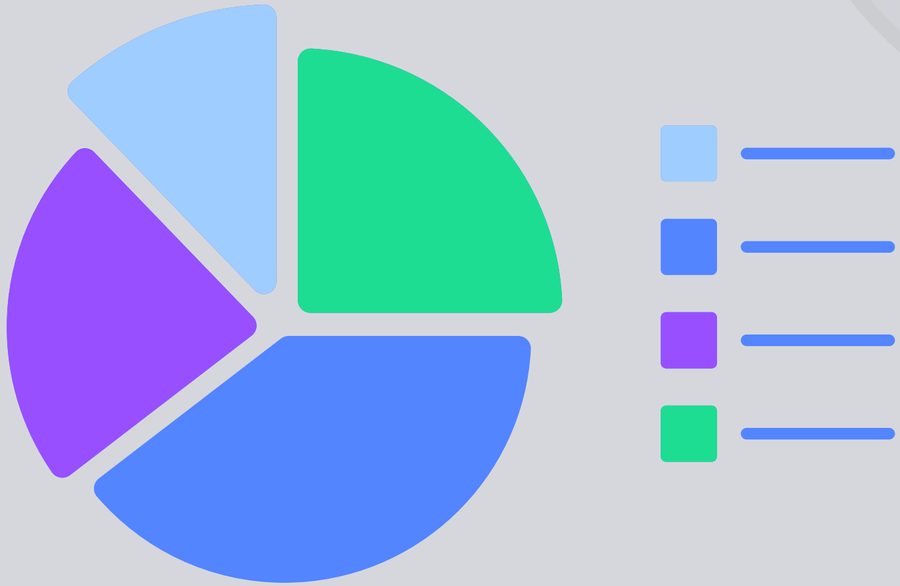
Statistiques descriptives
Introduction

La statistique descriptive:

Méthode permettant de récolter des données, de les ordonner, de les analyser, de les interpréter et de les représenter graphiquement

Les statistiques descriptives :

Quand on parle de statistiques descriptives au pluriel, on parle de données statistiques.



Statistiques descriptives
Introduction

Sélectionner son jeu de données:

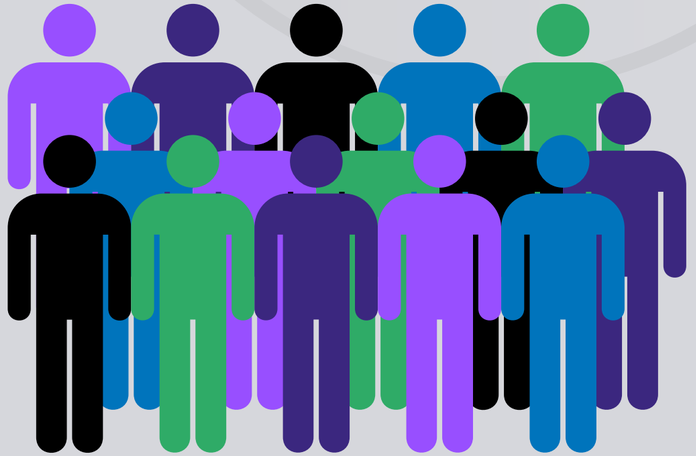
Lorsqu'on veut appliquer la méthode descriptive, la 1 ère étape est de définir sur quoi nous allons baser notre analyse. Pour se faire nous allons définir ce qu'on appelle population ou échantillon.

Population :

En statistique, la population représente l'ensemble des individus ou objets que l'on souhaite étudier. Cela peut être une classe d'élèves, tous les habitants d'un pays, ou toutes les voitures vendues en 2023,...

Individu : unité statistique

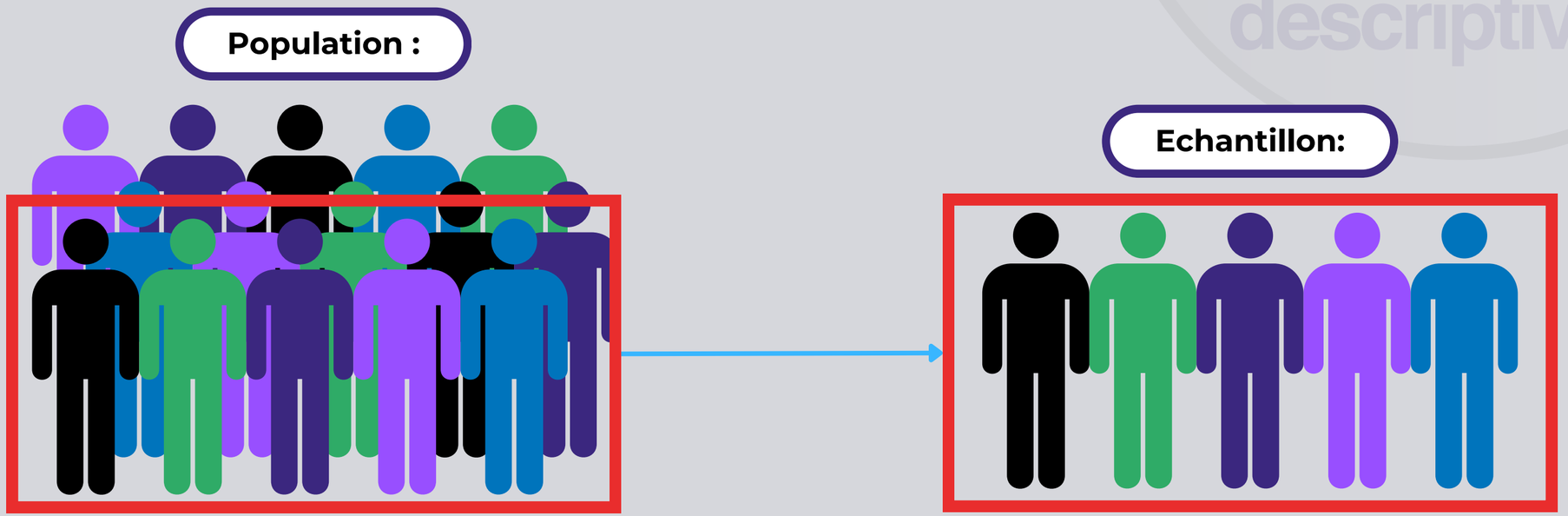
Population :



Statistiques descriptives
Introduction

Echantillon:

- **Échantillon : Il est souvent impossible ou difficile d'étudier toute la population, c'est pourquoi on utilise un échantillon qui est un sous-ensemble de la population. Cet échantillon doit être représentatif pour que les résultats soient fiables.**

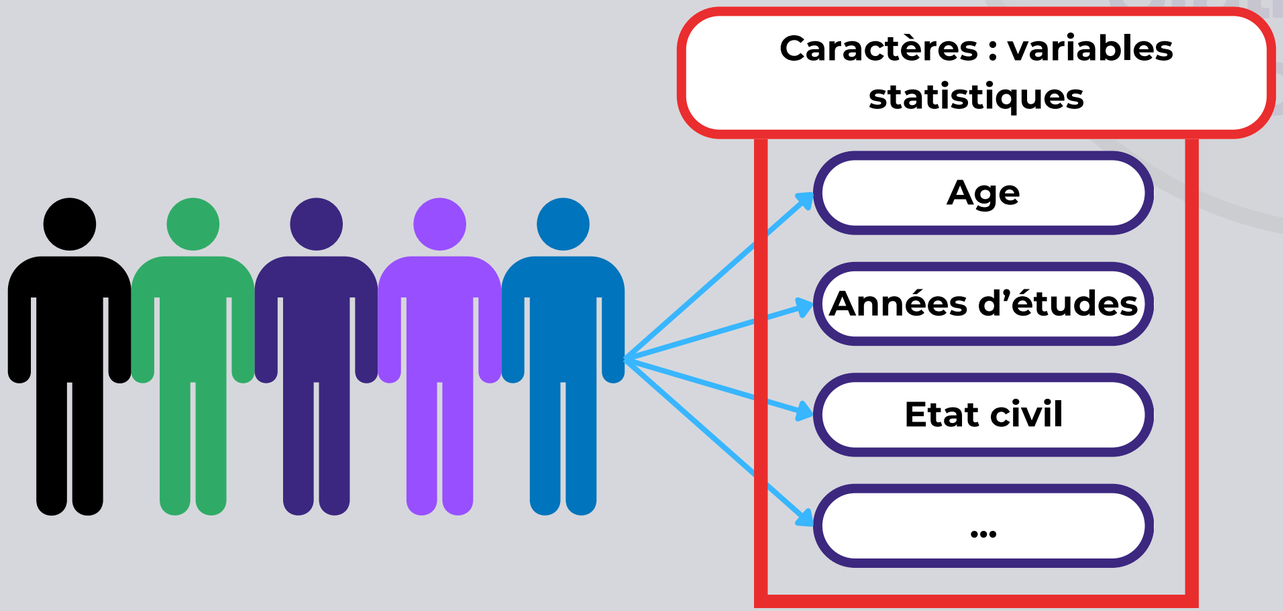


Statistiques descriptives
Introduction

Que faut-il étudier :

Lorsqu'on a définit notre population/échantillon, on doit se poser la question : "Avec les données que possède ma sélection, que peut-on étudier ?"

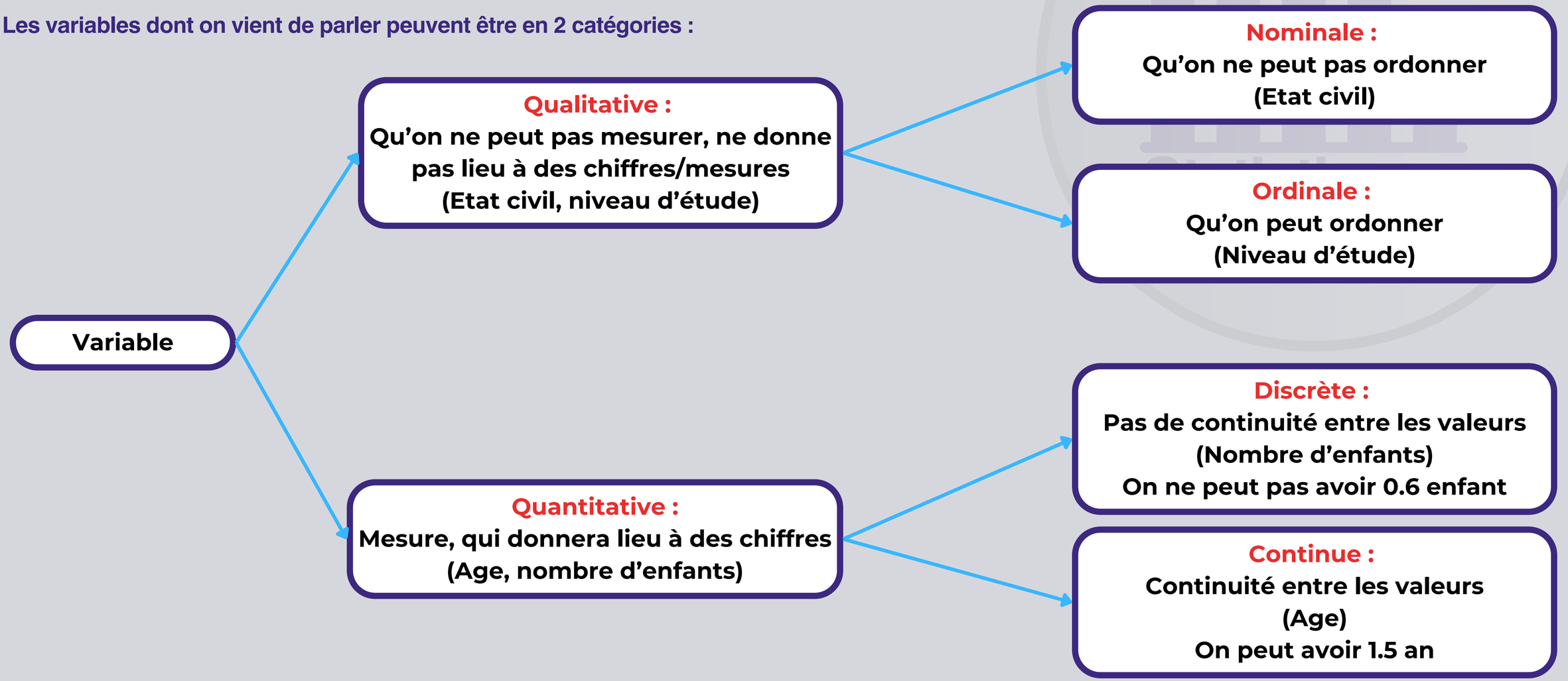
Imaginons un jeu de données basé sur des individus humains, je pourrai par exemple étudier des données comme l'âge, les années d'études,..
Tous ces éléments s'appellent des caractères ou encore des variables statistiques.
On utilise le terme variable, car dans le cas de l'étude de l'âge des individus, chaque individu à un âge différent donc je suis bien dans une variable.



Statistiques descriptives
Introduction

Variables :

Les variables dont on vient de parler peuvent être en 2 catégories :

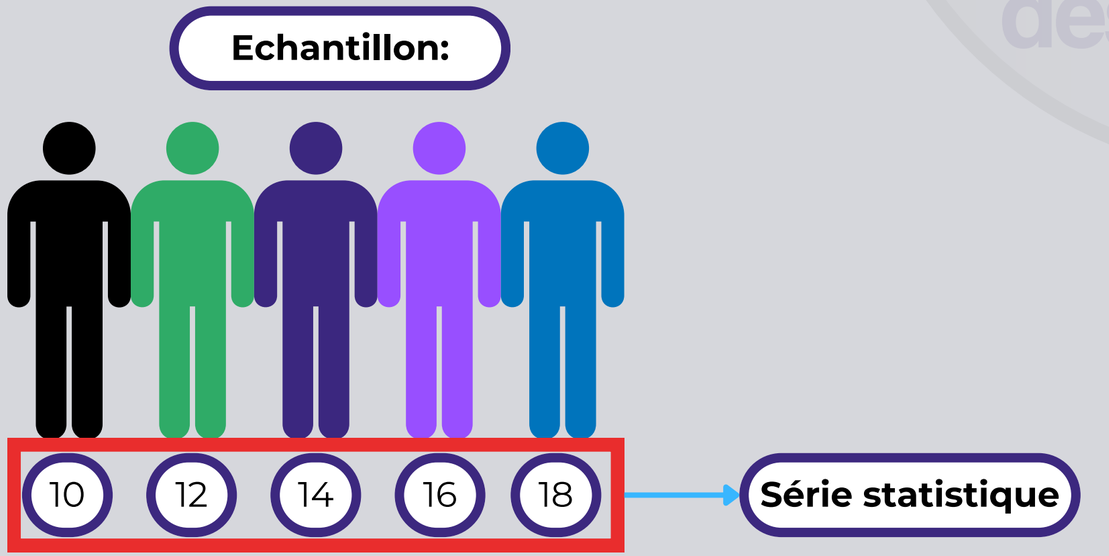


Statistiques descriptives
Introduction

Moyenne :

C'est la somme de toutes les valeurs divisée par le nombre de valeurs. La moyenne donne une idée générale des valeurs dans l'ensemble, mais elle peut être influencée par des valeurs extrêmes (outliers).

- **Exemple : Si vous avez les notes suivantes : 10, 12, 14, 16, 18 la moyenne est (10+12+14+16+18)/5=14**

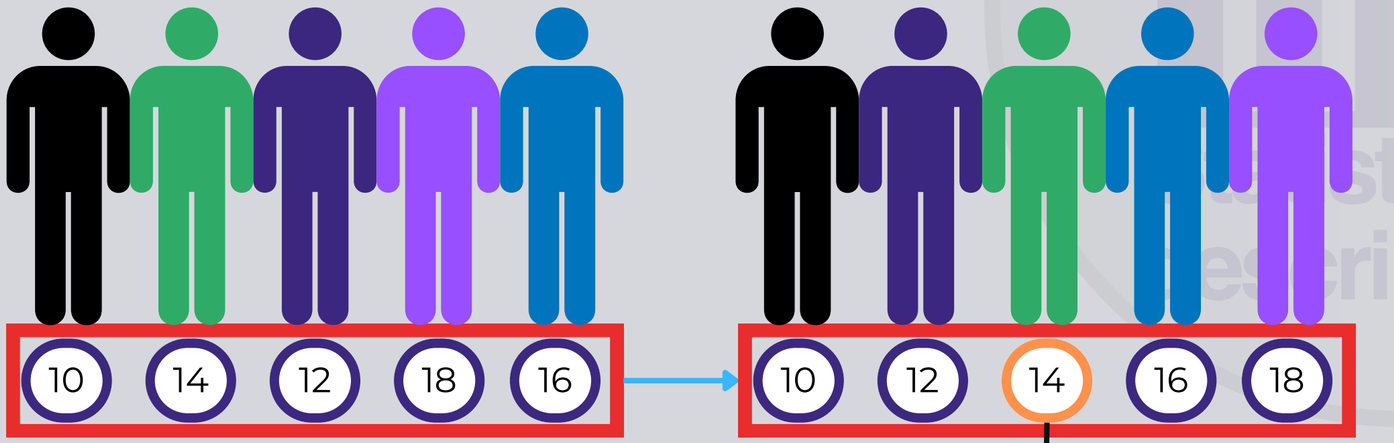


Statistiques descriptives
Introduction

Médiane :

C'est la valeur centrale d'une série de valeurs ordonnées.

Nombres impairs :



1: Ordonner la série

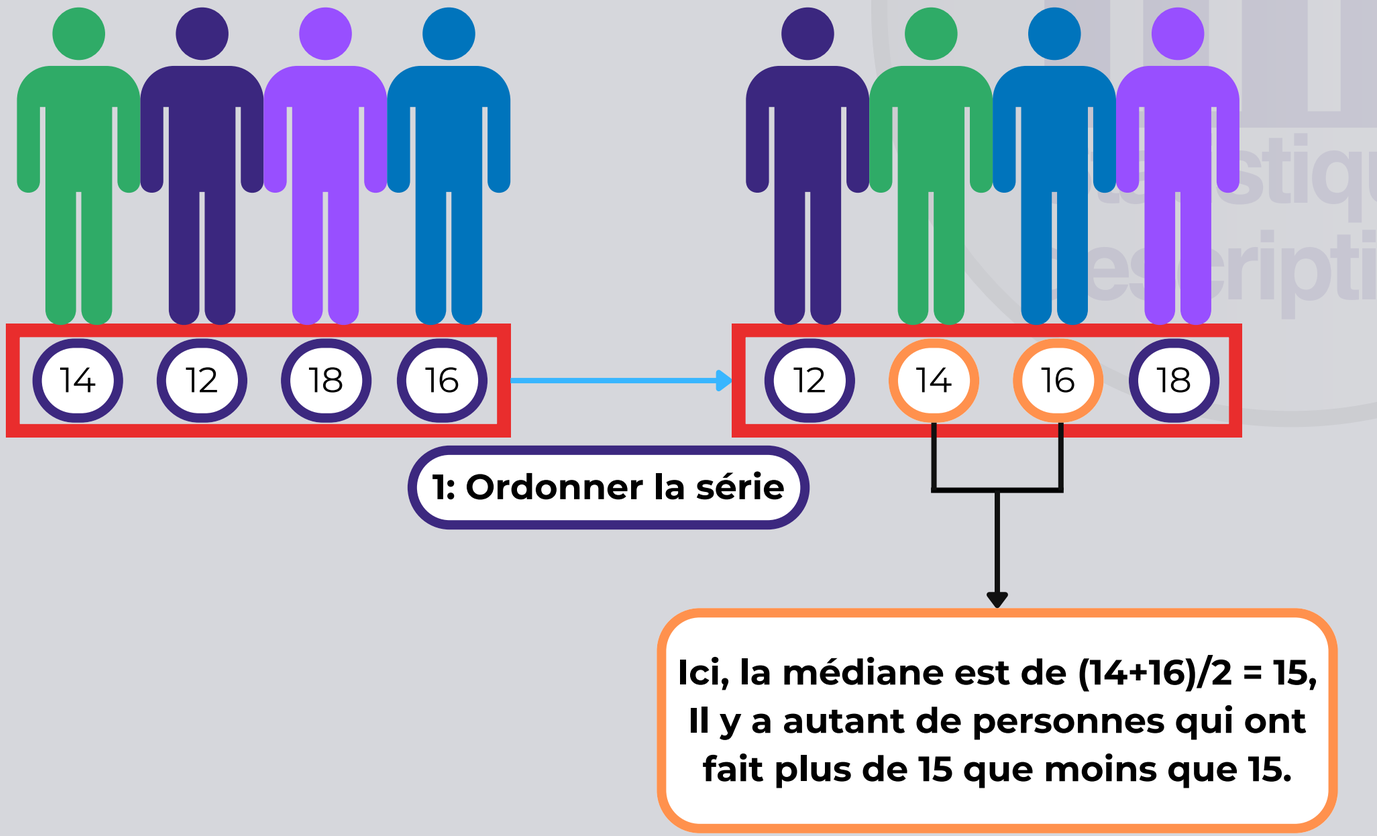
Ici, 14 est la médiane, Il y a autant de personnes qui ont fait plus de 14 que moins que 14.

Statistiques descriptives
Introduction

Médiane :

C'est la valeur centrale d'une série de valeurs ordonnées.

Nombres pairs :

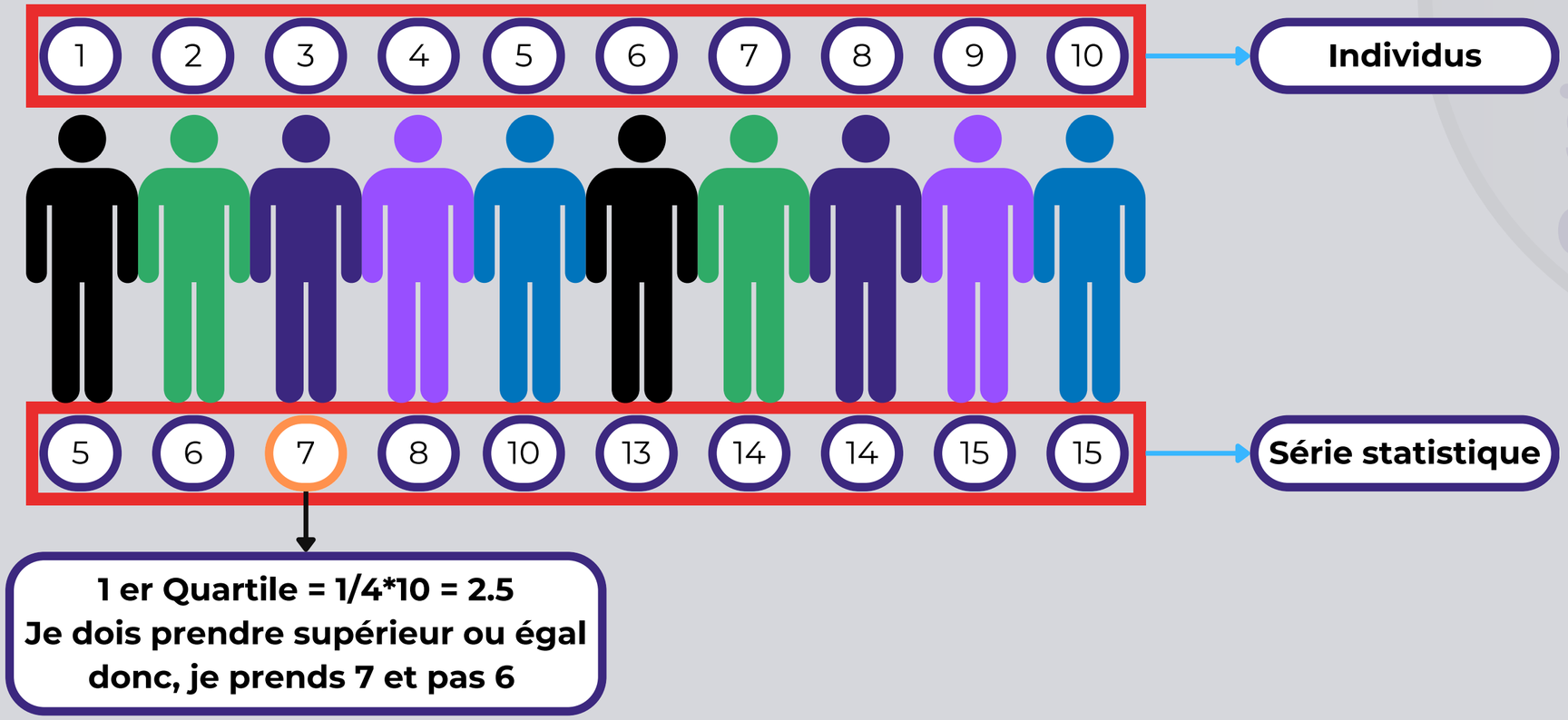


Statistiques descriptives
Introduction

Quartiles :

Un quartile fait référence à 1/4, si la médiane est la valeur du milieu qui sépare une série statistique en 2, alors un quartile va le faire en 4.

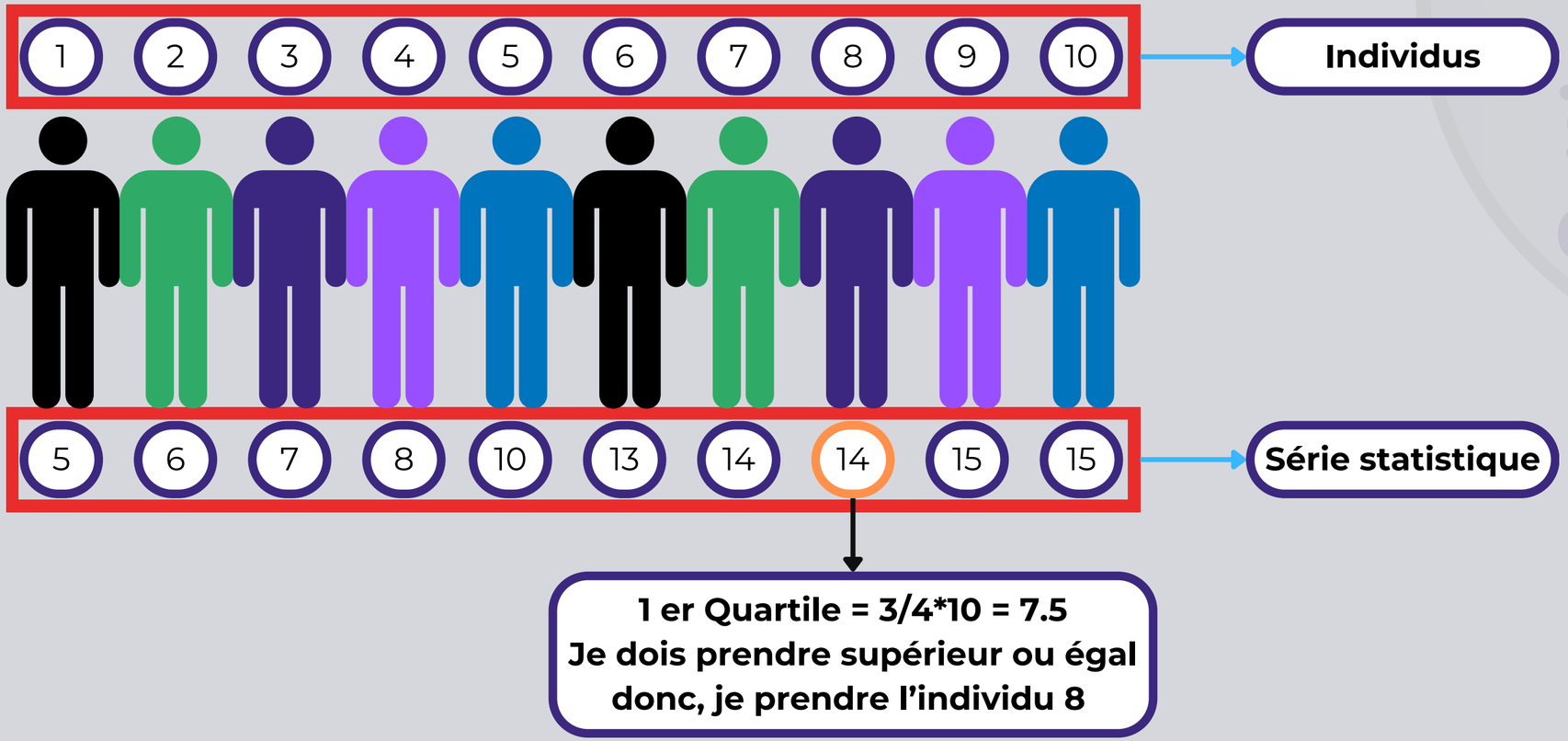
- **Premier quartile (Q1) : Valeur de la série supérieure ou égale à au moins 25% des données de la série ordonnée.**



Statistiques descriptives
Introduction

Quartiles :

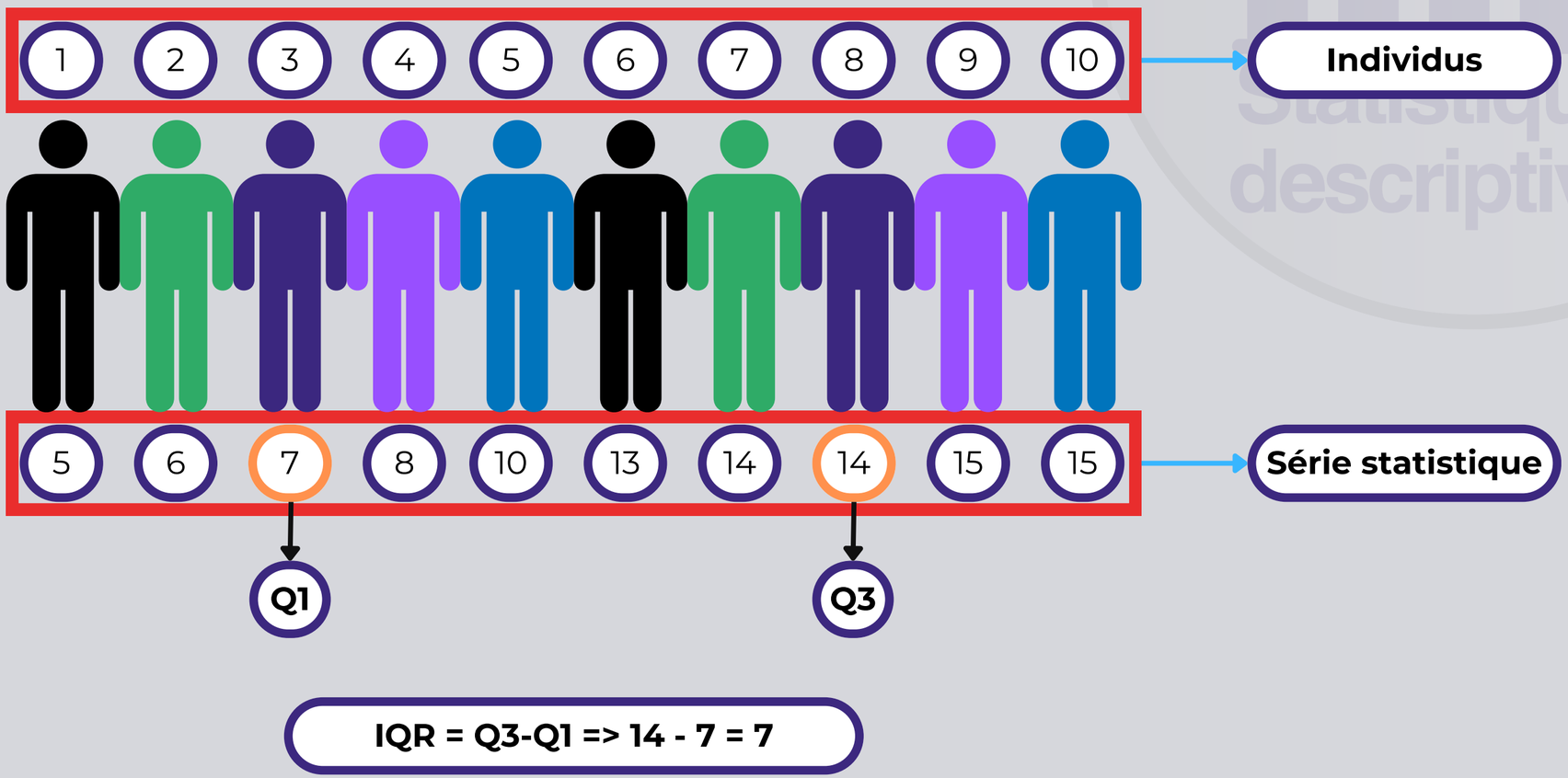
3ème quartile (Q3) : Valeur de la série supérieure ou égale à au moins 75% des données de la série ordonnée.



Statistiques descriptives
Introduction

Caractéristiques de dispersion autour de la médiane - L'écart interquartile :

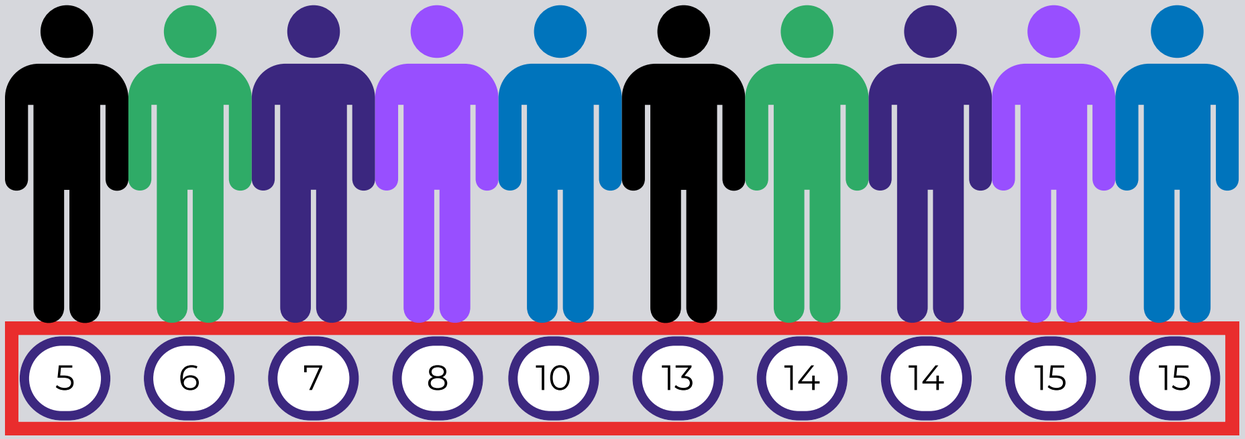
C'est la différence entre le 1er et le 3ème quartile. Il est un indicateur de la dispersion de la valeur autour de la médiane, c'est-à-dire, de combien les 50 % de valeurs autour de la médiane s'écartent de celle-ci.



Statistiques descriptives
Introduction

Caractéristiques de dispersion autour de la moyenne - Variance :

- **Variance : Somme des écarts entre les valeurs et la moyenne portés au carré, elle permet l'analyse de la dispersion des valeurs autour de la moyenne.**



Moyenne = 10.7

Variance population =
(**(**5-10.7**)²** + **(**6-10.7**)²** + … + **(**15-10.7**)²**) / 10 = 14

💡**Pourquoi la moyenne est-elle portée au carré ?**
Parce qu'on s'intéresse à l'écart qu'il soit positif (15-10,7) ou négatif (5-10,7), le fait de porter les éléments au carré, nous assure que nos différences seront toujours positifs.

Statistiques descriptives
Introduction

Caractéristiques de dispersion autour de la moyenne - Variance par Biais d'estimation :

Quand on travaille avec un échantillon, sa moyenne est utilisée à la place de celle de la population, et cela introduit un biais dans l'estimation de la variance.

Pour vulgariser, la moyenne de l'échantillon est calculée uniquement à partir des données de l'échantillon, et donc, elle est une estimation de la moyenne réelle de la population.
De ce fait, on sous-estime la vraie variance de la population, car elle tend à être trop proche des données de l'échantillon lui-même et pour corriger ce biais, on divise par n–1 (10-1) au lieu de n (10). Cette correction s'appelle le degré de liberté. En diminuant le dénominateur, on compense l'écart créé par l'utilisation de la moyenne de l'échantillon.

Variance échantillon =
(**(**5-10.7**)²** + **(**6-10.7**)²** + … + **(**15-10.7**)²**) / (10-1)

Statistiques descriptives
Introduction

Caractéristiques de dispersion autour de la moyenne - Ecart-type :

Le résultat de la variance étant sensé représenter la dispersion des données autour de la moyenne, se retrouve un peu exagéré de part son portage au carré. De ce fait, nous pouvons effectuer la racine de carré de la variance afin de récupérer ce qu'on appelle l'écart-type.

Ecart-type = Racine carré de la Variance = 3.7

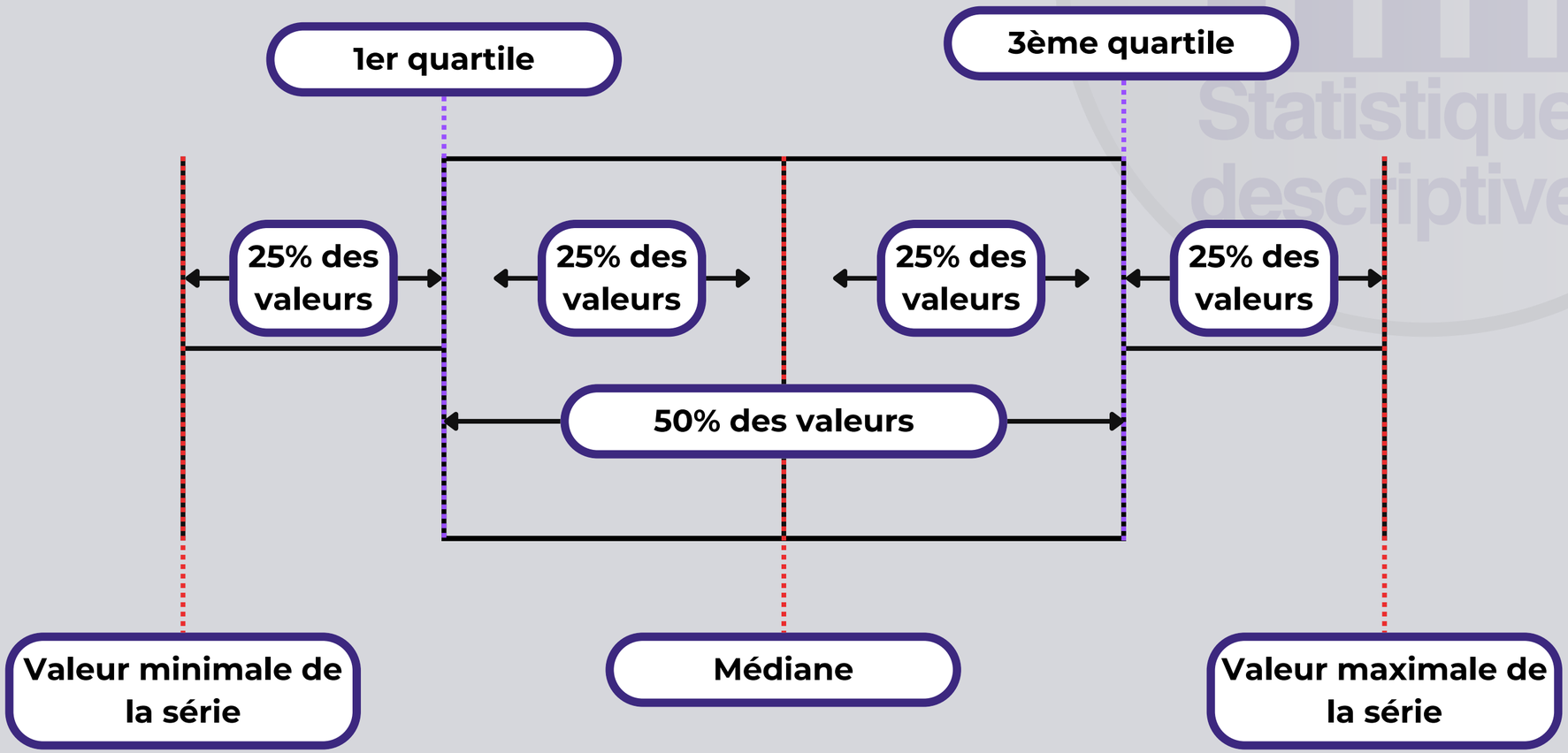
Cela signifie que les points se dispersent autour de la moyenne aux alentours de 3.7

Notez que cet écart-type est lui aussi, une sorte de moyenne, puisqu'on fait la moyenne des écarts entre la moyenne et chaque valeur. C'est-à-dire, en moyenne, les valeurs s'écartent de 3.7 de la moyenne.

Statistiques descriptives
Introduction

La boite à moustaches : sans valeur aberrante.

Montre la distribution des données à travers des quartiles. Il est utile pour visualiser la médiane, l'étendue, et les éventuels outliers (valeurs extrêmes).

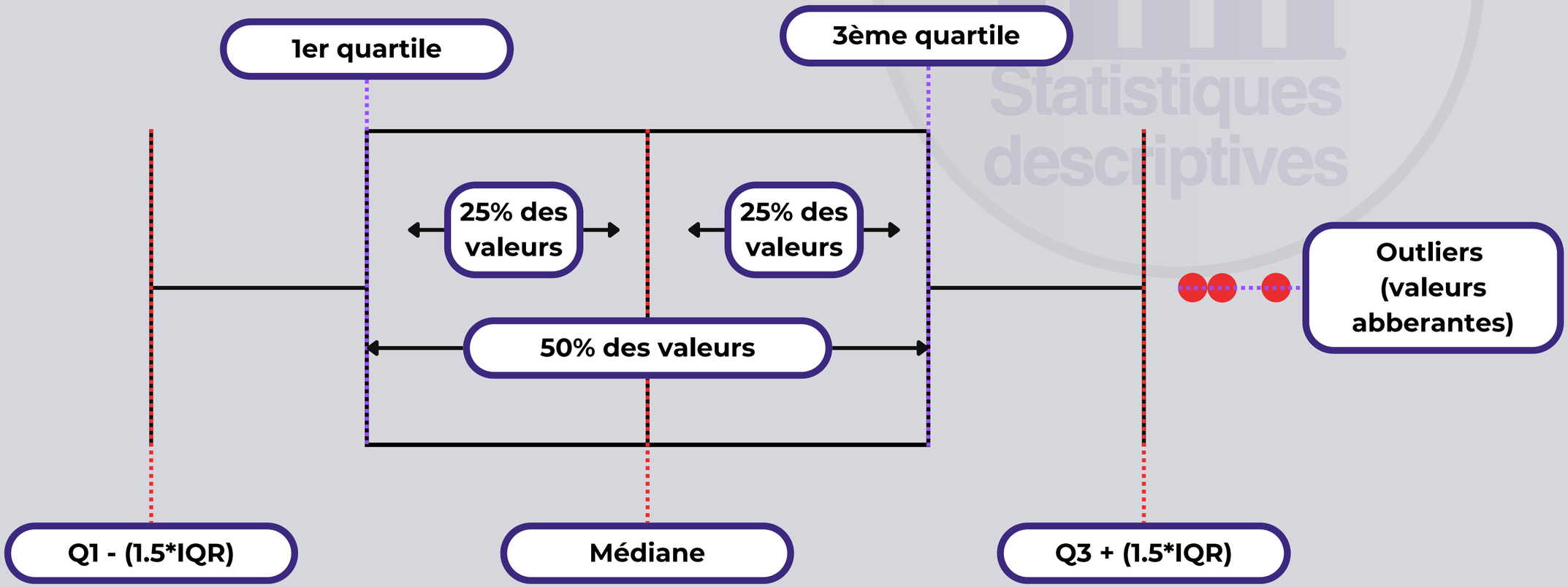


Statistiques descriptives
Introduction

La boite à moustaches : Outliers

Il existe une règle lorsqu'on développe un box plot qui dit :

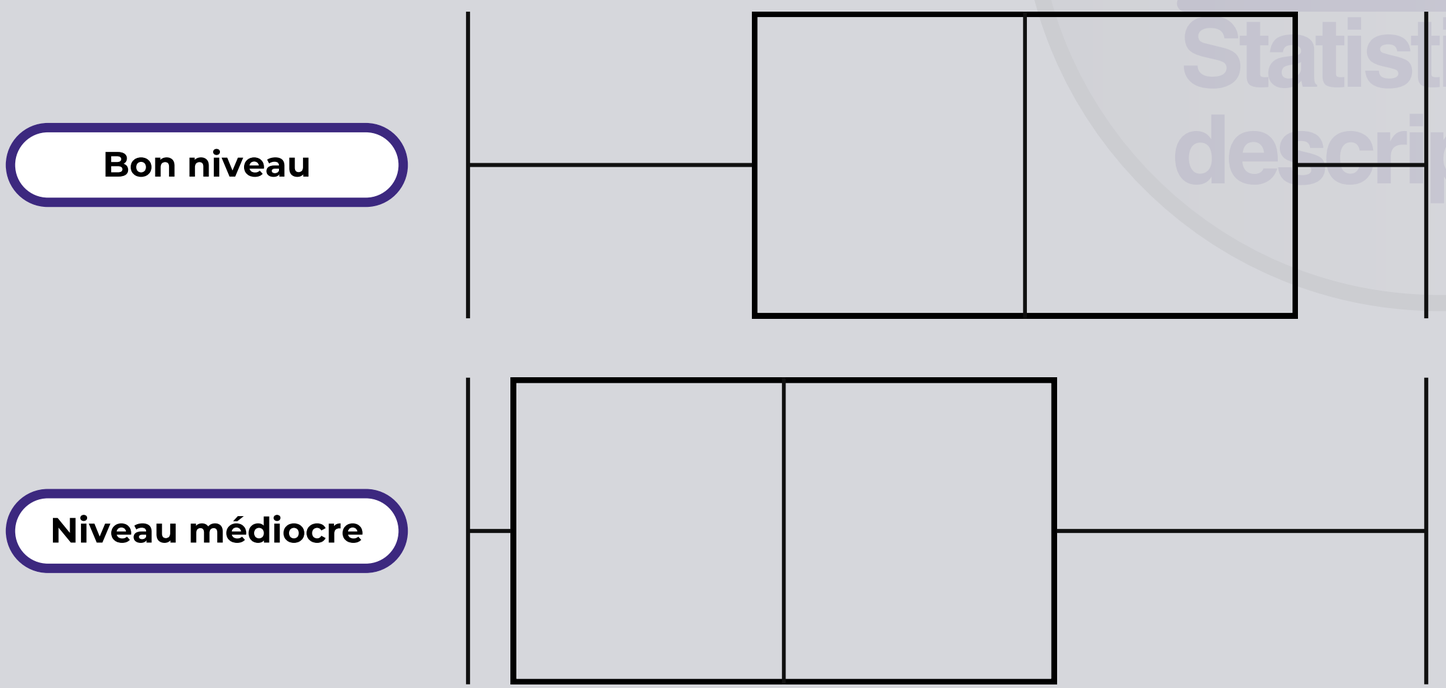
- **Si " abs(x - médiane) > IQR*1.5 " où x est une variable de notre série statistique, alors les limites des boites à moustaches sont définies par les formules ci-dessous.**



Statistiques descriptives
Introduction

La boite à moustaches : Analyse visuelle

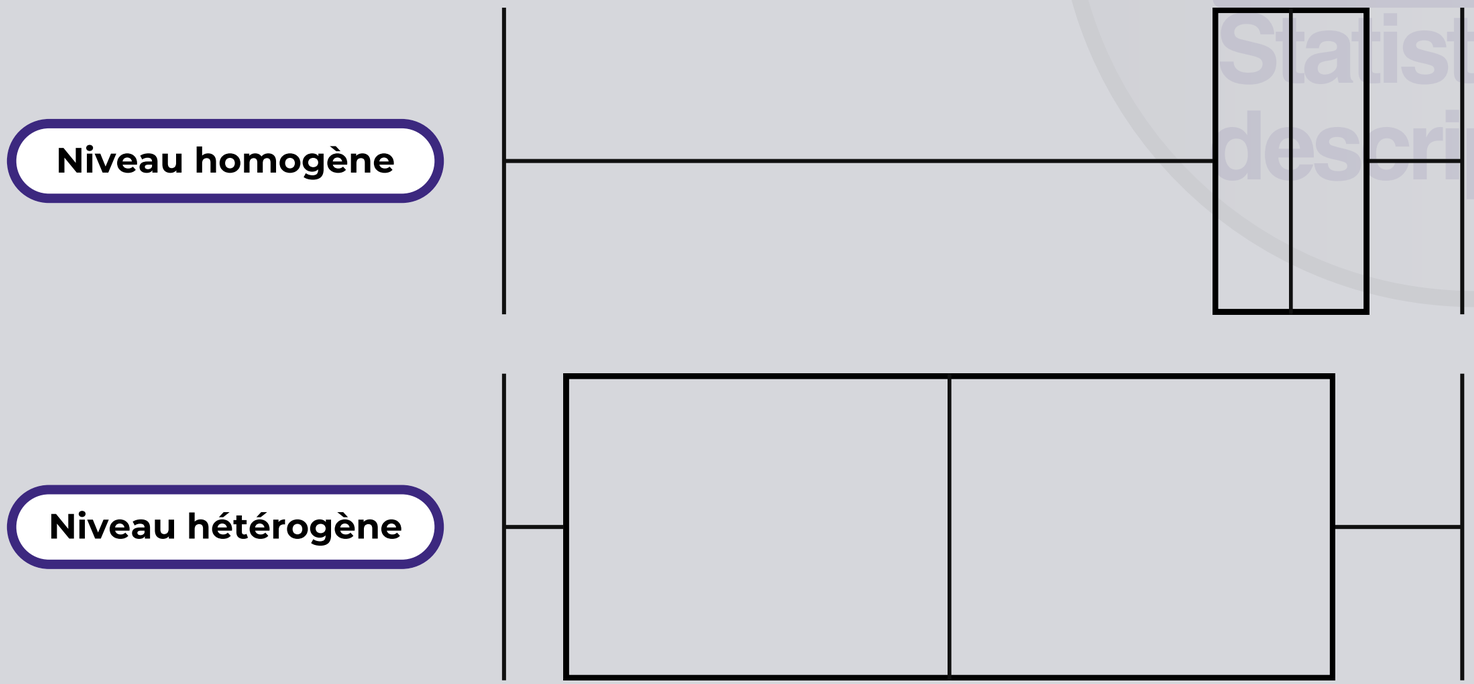
Plus la boîte se rapproche de la limite maximale et plus la classe sera bonne dans sa globalité, plus elle sera proche de la limite minimum et plus le niveau sera médiocre.



Statistiques descriptives
Introduction

La boite à moustaches : Analyse visuelle

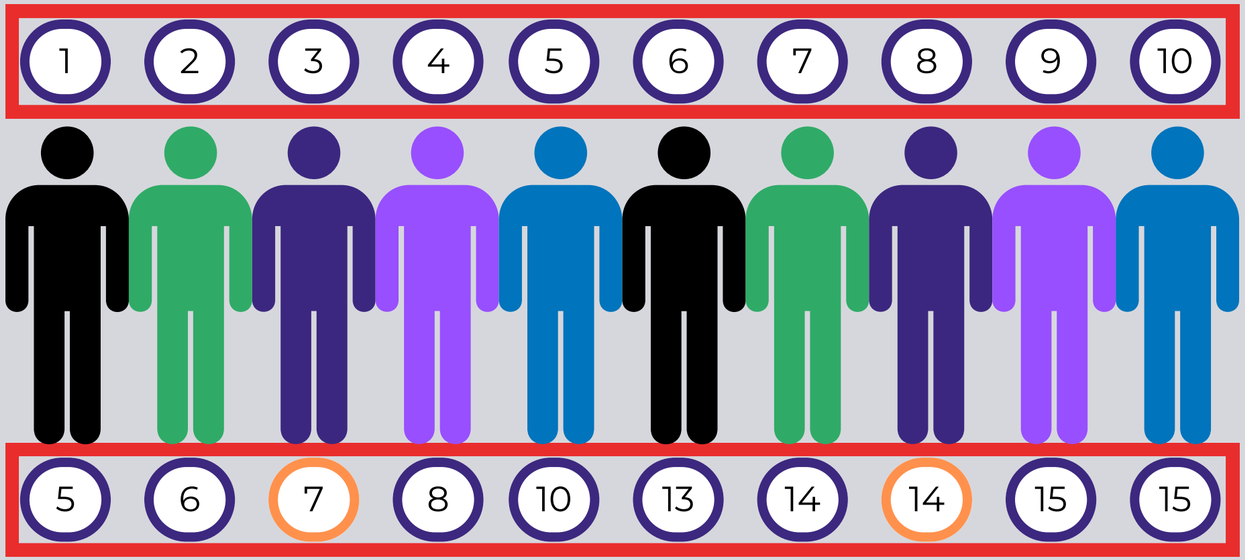
Plus la boîte est petite et plus la classe à un niveau homogène, les élèves se rapprochent globalement d'un même niveau. Au contraire, plus la boite est large et plus le niveau est hétérogène.



Statistiques descriptives
Introduction

La Fréquence :

Si un pourcentage est lié au %, une fréquence est liée à l'unité, elle se calcule en divisant le nombre d'occurrences d'une variable par le nombre total d'occurrences.



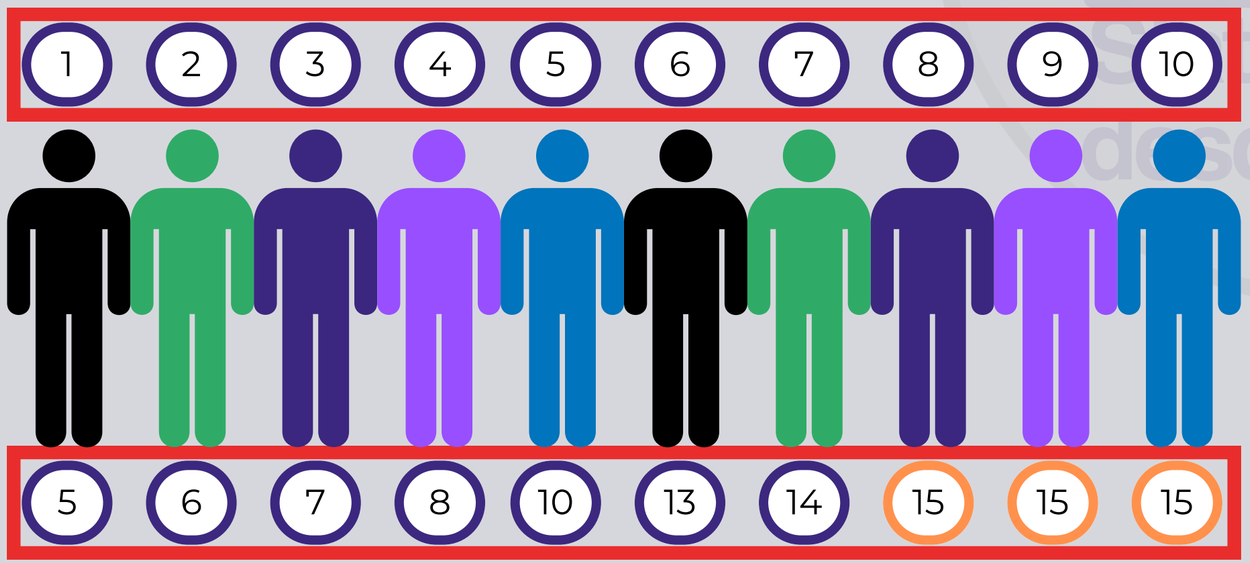
Notes	occurences	fréquence
5	1	1/10 = 0.1
6	1	1/10 = 0.1
7	1	1/10 = 0.1
8	1	1/10 = 0.1
10	1	1/10 = 0.1
13	1	1/10 = 0.1
14	2	2/10 = 0.2
15	2	2/10 = 0.2
Total :	10	10/10 = 1

💡 Dans le cas d'une variable quantitative continue, on définit des intervalles pour classer les valeurs afin de pouvoir établir la fréquence.

Statistiques descriptives
Introduction

Le Mode :

C’est la valeur dominante dans une série statistiques, c’est-à-dire la valeur qui se répète le plus.



Mode = 15