

UE 012 – Bases méthodologiques de l'analyse statistique de données

Kjerstin Torre

kjerstin.torre@umontpellier.fr

Semestre 1 SA1MRAPA

[illegible]

Intro

Statistiques

Recueil, traitement, interprétation, et présentation d'un ensemble de données

Consigne : Indiquez dans quelle mesure chacun des énoncés suivants correspond actuellement à l'une des raisons pour lesquelles vous poursuivez des études supérieures.

Ne correspond pas du tout	Correspond très peu	Correspond un peu	Correspond moyennement	Correspond assez	Correspond fortement	Correspond très fortement
1	2	3	4	5	6	7

e du plaisir et de la
endre de nouvelles



Recueil, traitement, interprétation, et présentation d'un ensemble de données

Consigne : Indiquez dans quelle mesure chacun des énoncés suivants correspond actuellement à l'une des raisons pour lesquelles vous poursuivez des études supérieures.

Pourquoi poursuivez-vous des études supérieures ?

Ne correspond pas du tout	Correspond très peu	Correspond un peu	Correspond moyennement	Correspond assez	Correspond fortement	Correspond très fortement
1	2	3	4	5	6	7
1. Parce que juste avec un baccalauréat, je ne pourrais pas me trouver un emploi assez payant.						
☐	☐	☐	☐	☐	☐	☐
2. Pour le plaisir et de la recherche de nouvelles connaissances.						
☐	☐	☐	☐	☐	☐	☐

Consigne : Indiquez dans quelle mesure chacun des énoncés suivants correspond actuellement à l'une des raisons pour lesquelles vous poursuivez des études supérieures.

Pourquoi poursuivez-vous des études supérieures ?

Ne correspond pas du tout	Correspond très peu	Correspond un peu	Correspond moyennement	Correspond assez	Correspond fortement	Correspond très fortement
1	2	3	4	5	6	7
1. Parce que juste avec un baccalauréat, je ne pourrais pas me trouver un emploi assez payant.						
☐	☐	☐	☐	☐	☐	☐
2. Pour le plaisir et de la recherche de nouvelles connaissances.						
☐	☐	☐	☐	☐	☐	☐

Le traitement statistique ne permet pas de rattraper d'éventuelles boulettes

Conditionne les types d'analyses faisables

La notion de « choix » d'un test statistique plutôt qu'un autre n'est que très relative



Recueil, traîtement, interprétation, et **présentation** d'un ensemble de données

TOP 3 DES SNEAKERS LES PLUS VENDUES

EN FRANCE, AU 1ER SEMESTRE 2023



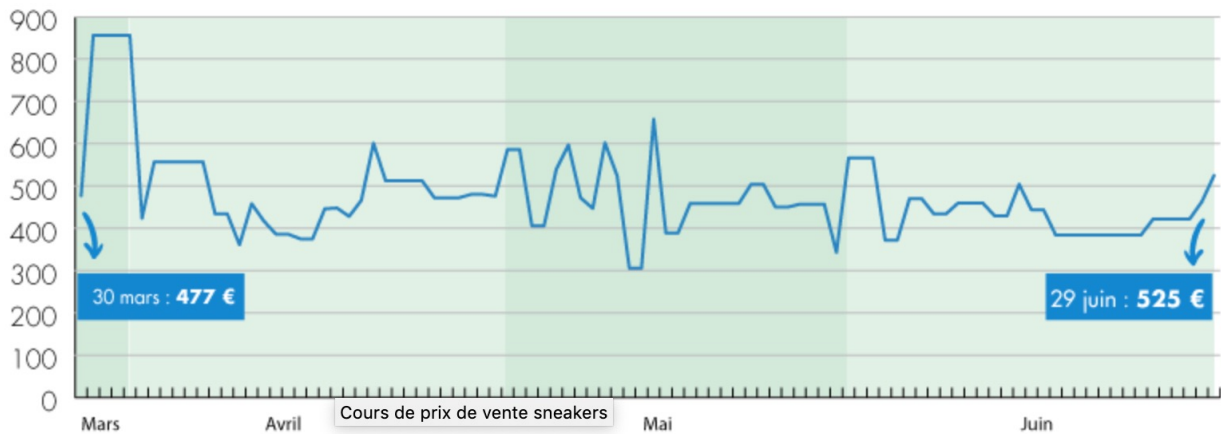
Source : lafinancepourtous.com d'après StockX

Tout n'est pas permis

Mais précision et clarté sont de mise

EXEMPLE DE COURS DE PRIX DE VENTE

AIR JORDAN 1 MID SE FEARLESS MAISON CHÂTEAU ROUGE, EN EUROS



Moyenne de la série : **476 €**
Écart-type : **106 €**
Prix « retail » (prix de vente sur le marché initial) : **118 €**

Recueil, traitement, interprétation, et présentation d'un ensemble de données



“Les faits sont têtus. Il est plus facile de s'arranger avec les statistiques.” Mark Twain



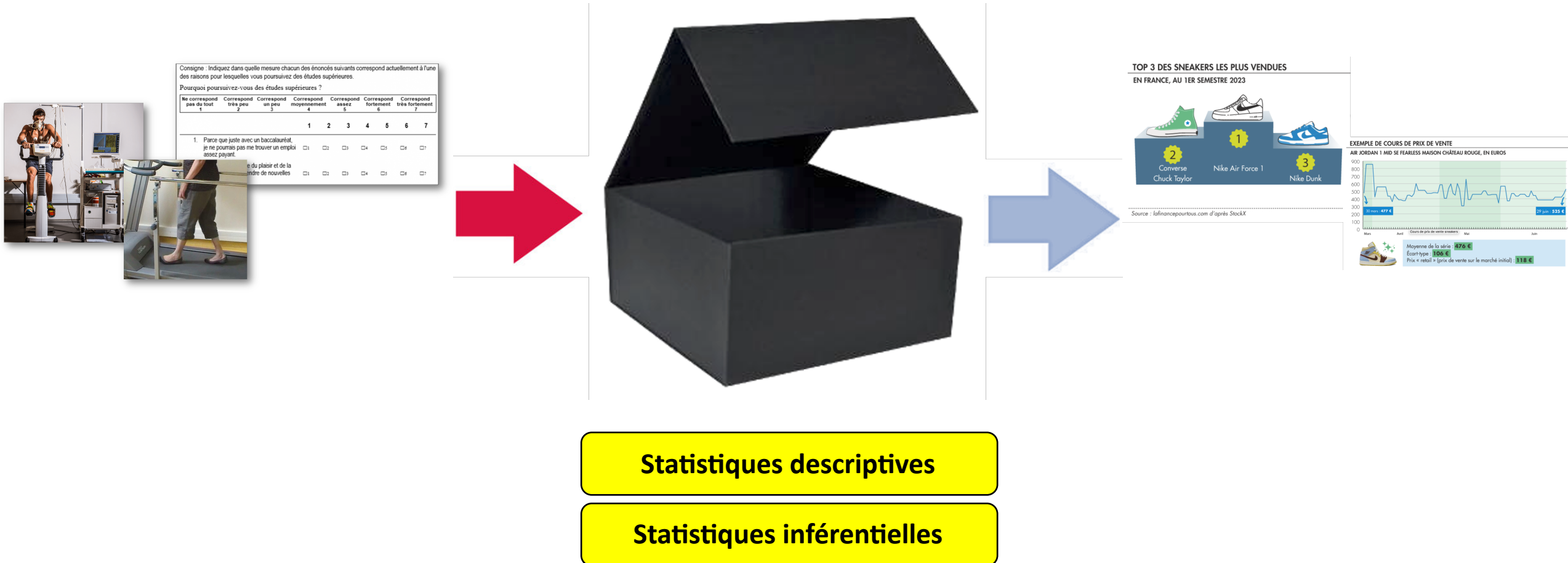
Comme vous le voyez, 100% de moi-même est d'accord avec mon projet !

Vous pouvez interpréter ça de deux façons différentes.



Apprendre à faire soi-même pour **comprendre**
et ne pas raconter n'importe quoi

Recueil, **traitement**, **interprétation**, et **présentation** d'un ensemble de données



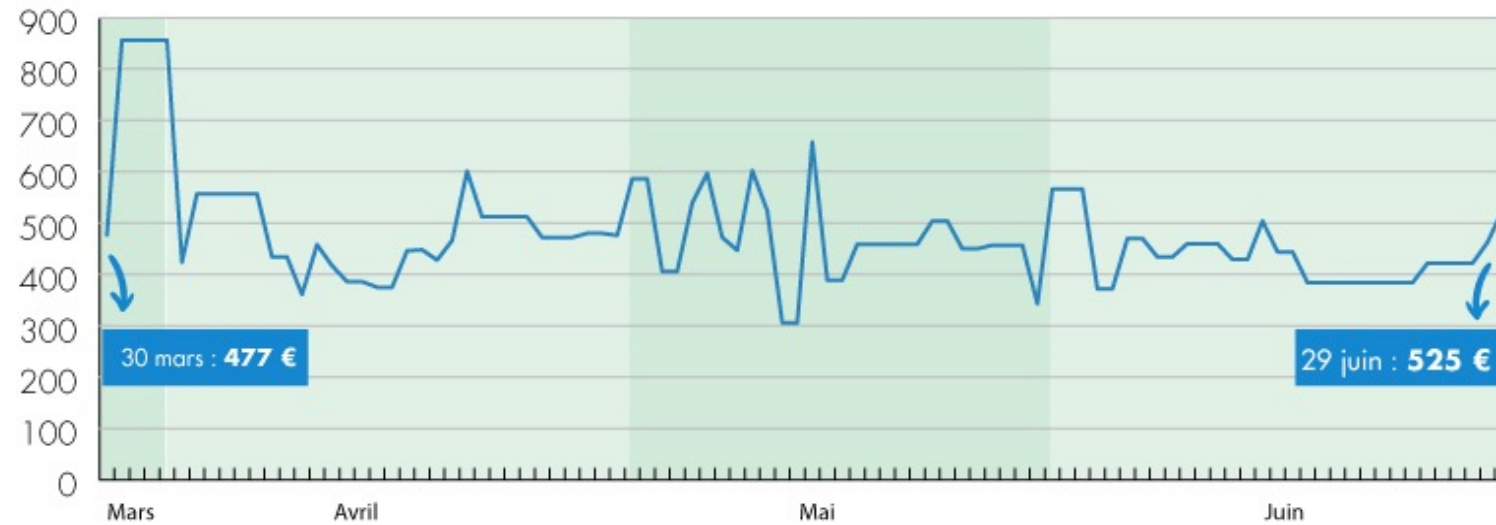
Statistiques descriptives

Réduire/résumer une information pour mieux la comprendre



EXEMPLE DE COURS DE PRIX DE VENTE

AIR JORDAN 1 MID SE FEARLESS MAISON CHÂTEAU ROUGE, EN EUROS



Moyenne de la série : **476 €**

Écart-type : **106 €**

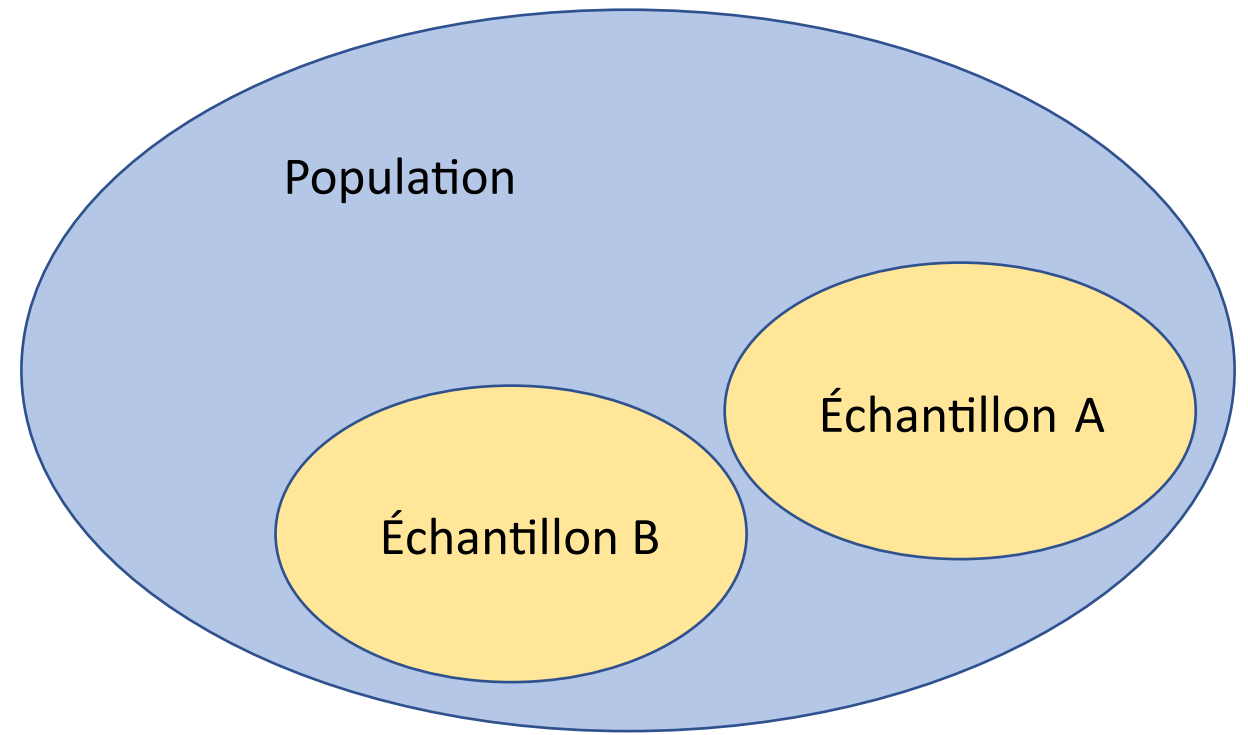
Ces deux notions sont appelées **paramètres**

Une statistique renvoie plus globalement à toute estimation d'un ou plusieurs **paramètres** concernant une population (depuis Fisher 1925).

Intro

Un paramètre statistique sera d'autant plus consistante et fiable que l'échantillon sera suffisamment grand

Si un paramètre est obtenu sur un échantillon réduit, sa représentativité devra être discutée au regard de valeurs de références (si celles-ci existent)



NB: échantillonnage par randomisation

les paramètres issus de plusieurs échantillons d'une même population présenteront des variations

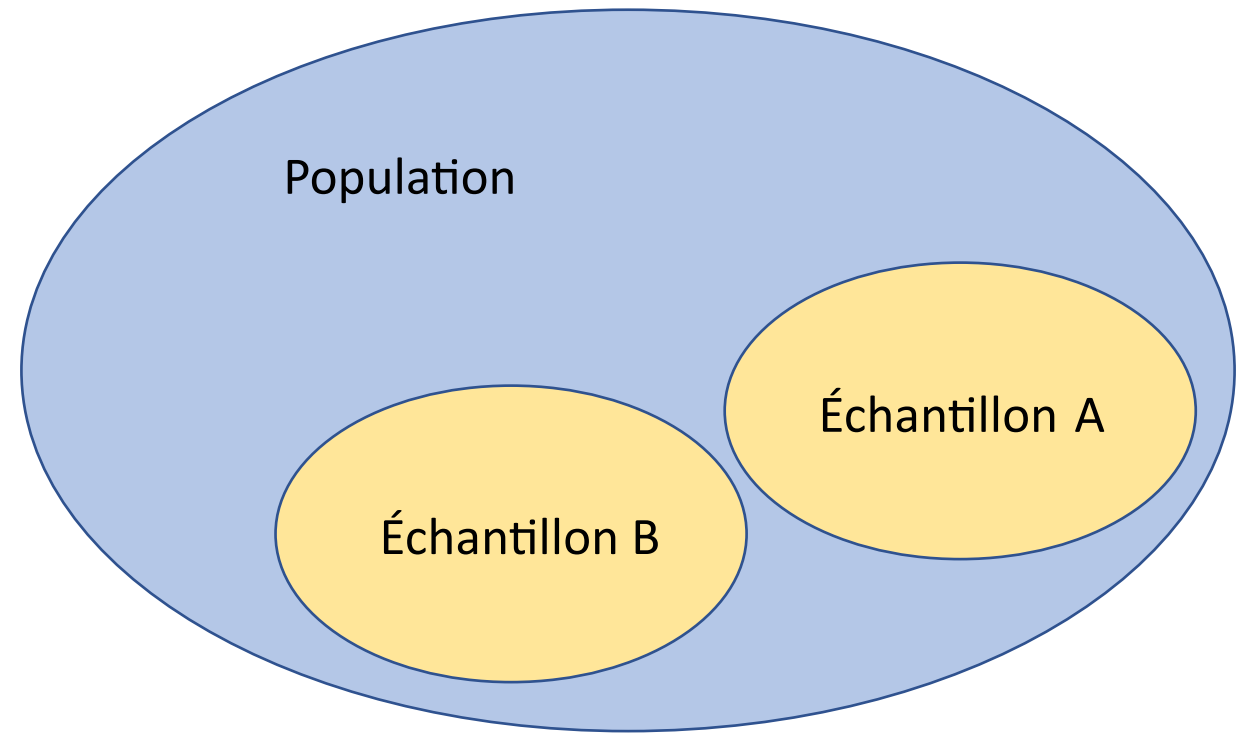
Cela renvoie à la notion **d'intervalle de confiance (IC)** d'un paramètre statistique

L'intervalle de confiance

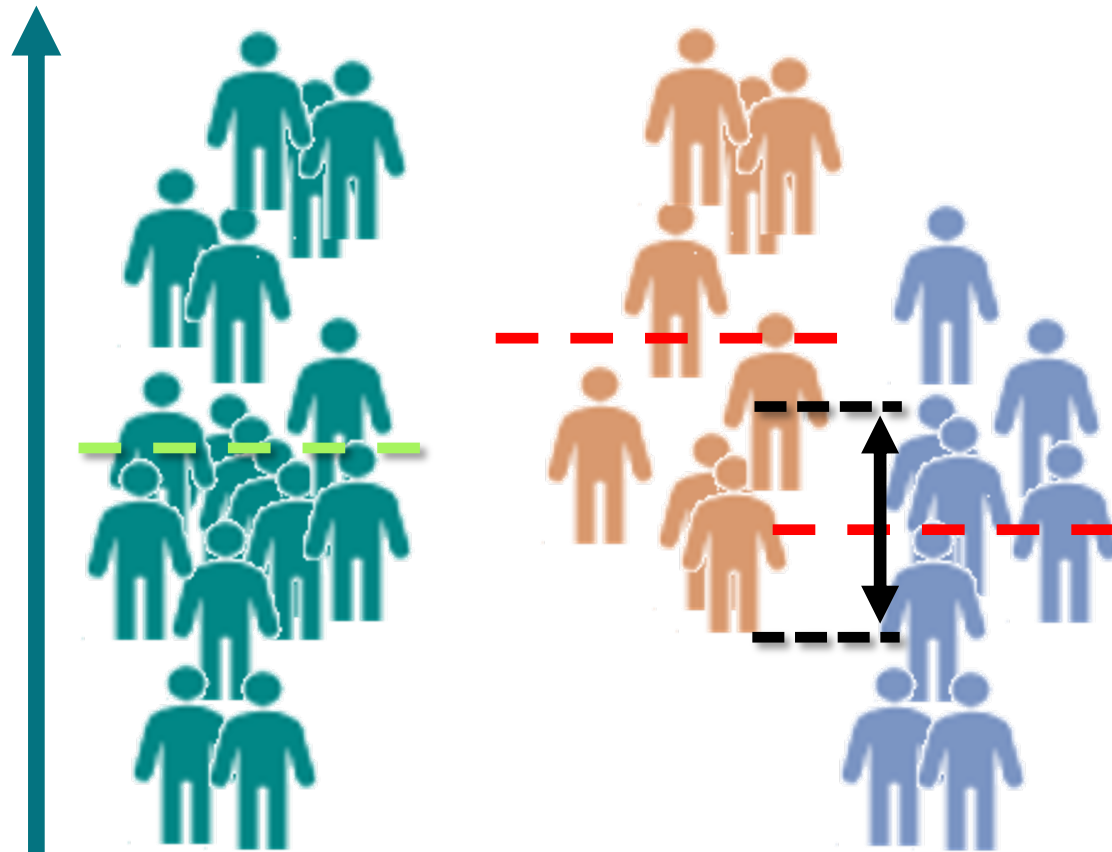
Si plusieurs échantillons sont prélevés dans une population, il est très probable que chaque échantillon ait une valeur moyenne différente

---> Quelle est la valeur moyenne de la population?

L'intervalle de confiance indique la plage dans laquelle se situe la véritable valeur moyenne de la population avec une certaine probabilité



L'intervalle de confiance



Intervalle de confiance à 95% :
À 95% de chances, la valeur moyenne
de la population se situe à l'intérieur
de cette fourchette

Probabilités



Statistiques

En **théorie des probabilités**:

Résultats théoriques d'une expérience régie par l'aléatoire

L'aléatoire est modélisé par des lois qui correspondent le mieux au modèle réel

"Connaissance *a priori*" (étude sans données)

"Construction idéalisée des notions et définitions abstraites"

En **statistiques**:

Étude de données empiriques (issues de situations concrètes, réelles)

Analyse + interprétation + extrapolation

"Connaissance *a posteriori*"

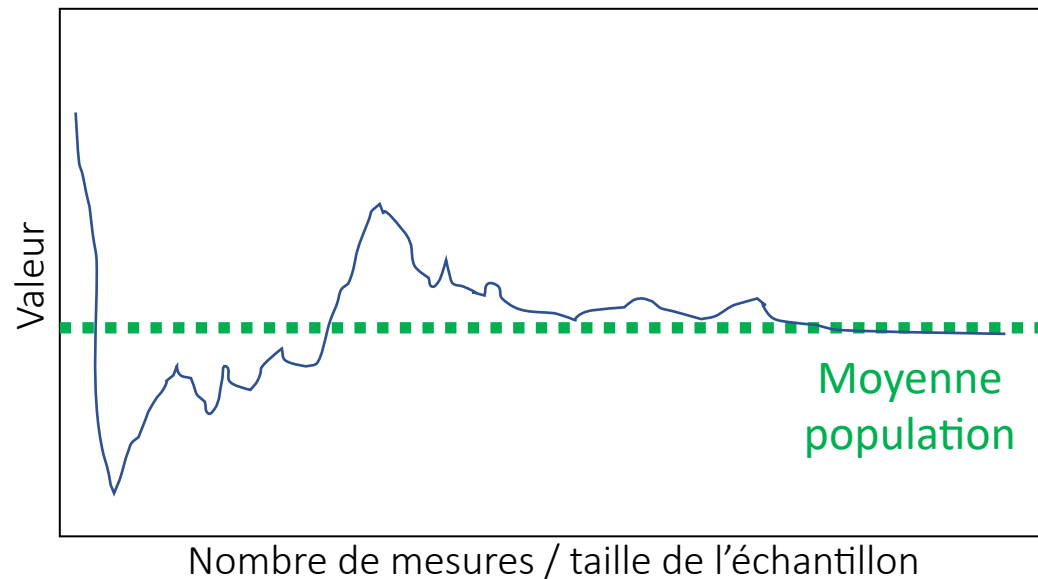
Probabilités



Statistiques

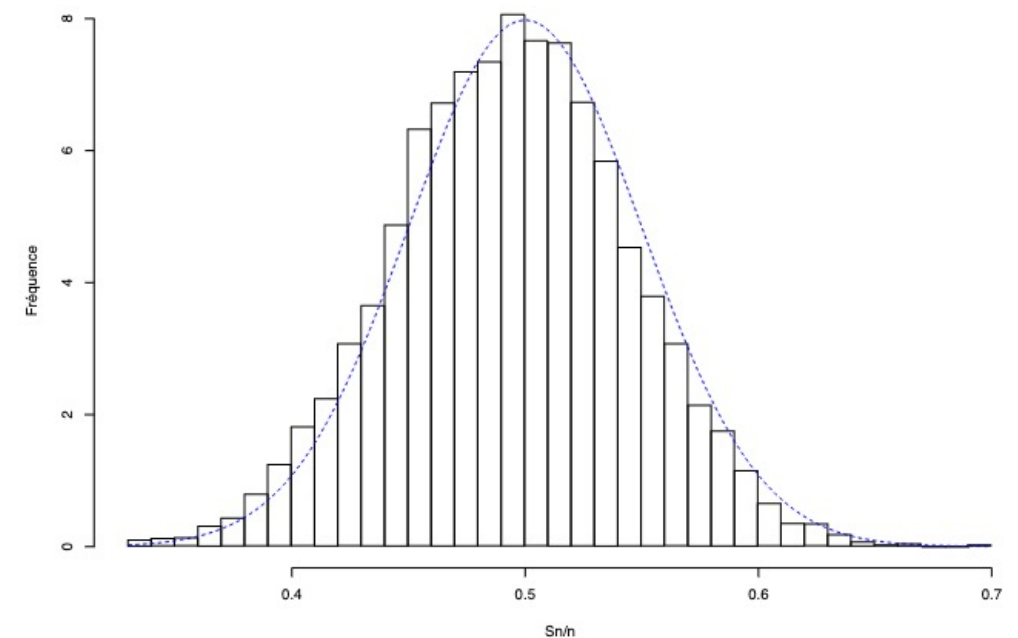
2 théorèmes essentiels en théorie des probabilités

La Loi des Grands Nombres



Quand la taille de l'échantillon augmente, la moyenne de cet échantillon se rapproche de la véritable moyenne de la population

Le Théorème central limite



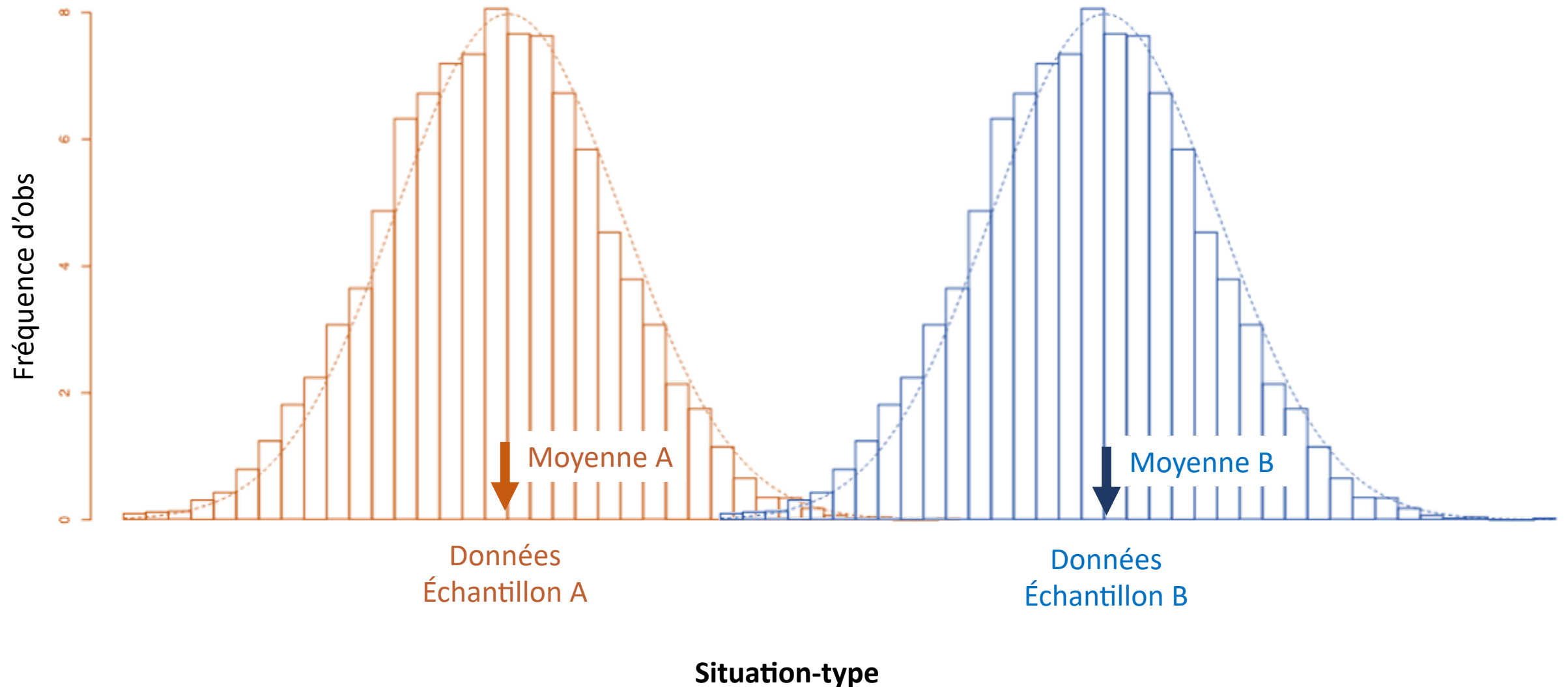
Avec un échantillon de taille suffisante, la distribution d'échantillonnage de la moyenne sera *normalement* distribuée

Intro

~~Est-ce qu'il existe une différence entre les moyennes des deux échantillons?~~

→ Est-ce qu'il existe une différence **significative** entre les moyennes des deux échantillons?

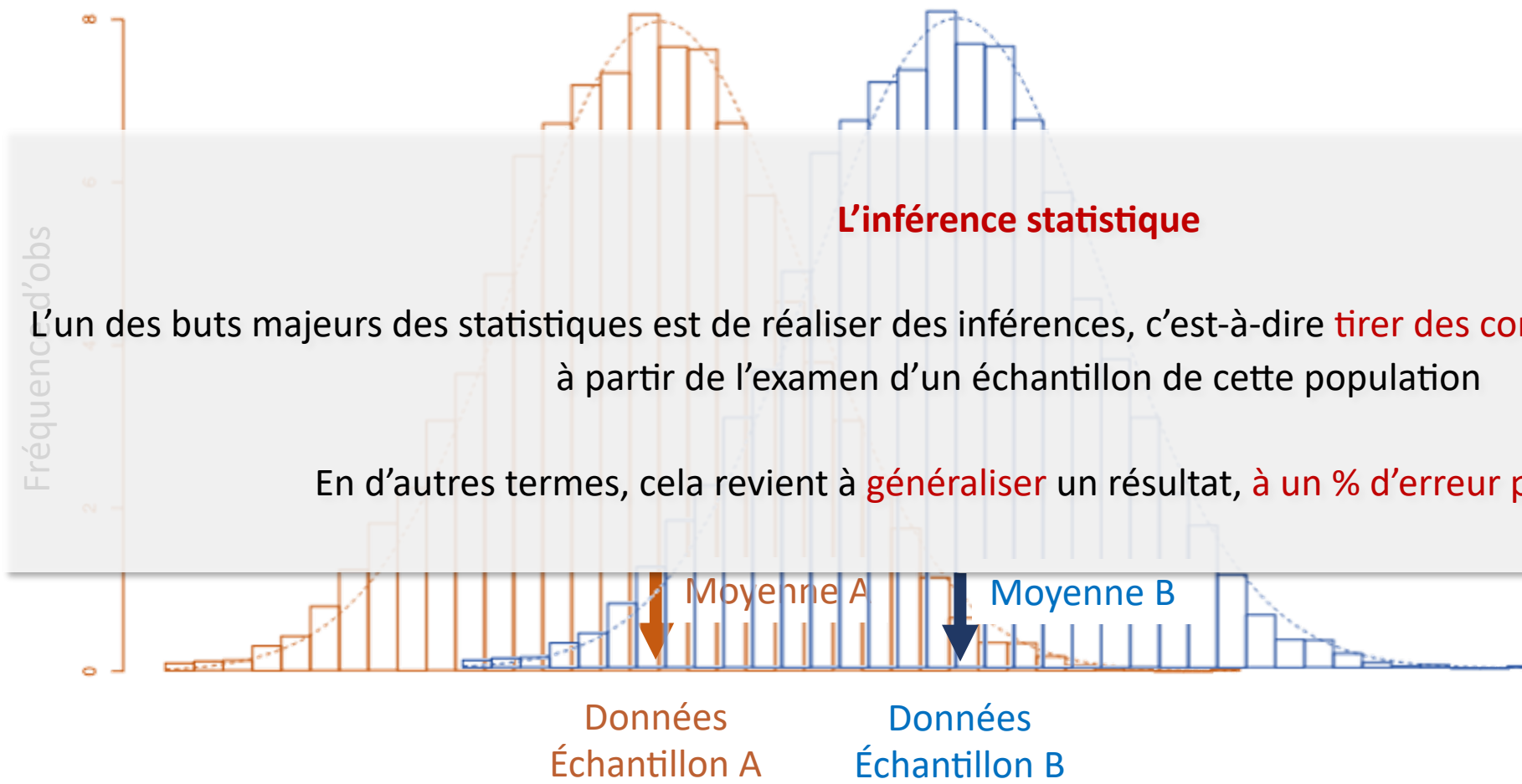
→ À quoi est due cette différence? (*que signifie-t-elle ?*)



Intro — ~~Est-ce qu'il existe une différence entre les moyennes des deux échantillons?~~

→ Est-ce qu'il existe une différence **significative** entre les moyennes des deux échantillons?

→ À quoi est due cette différence? (« que signifie-t-elle »?)



L'inférence statistique

L'un des buts majeurs des statistiques est de réaliser des inférences, c'est-à-dire **tirer des conclusions** sur des populations, à partir de l'examen d'un échantillon de cette population

En d'autres termes, cela revient à **généraliser** un résultat, à un % d'erreur possible près

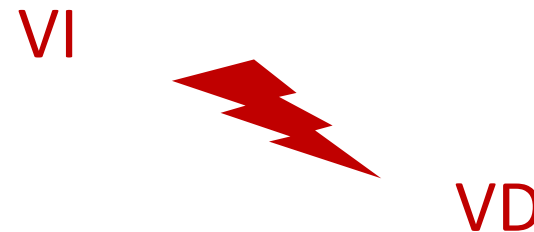
Situation-type

Statistiques inférentielles

Objectif

description $\neq$ explication
corrélation $\neq$ relation de cause à effet

En général, une expérimentation vise à tester l'effet d'un ou plusieurs **facteur(s) expérimental(aux) (Variable Indépendante, VI)** sur une/plusieurs variables d'intérêt (**Variables Dépendantes, VD**)



Variables recueillies - Échelles de mesure



Variables recueillies - Échelles de mesure

Données recueillies → Tableau de contingence

VARIABLES				
	1	2	3	
INDIVIDU 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	.
INDIVIDU 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	.
INDIVIDU 3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	.
INDIVIDU 4	$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

D'une manière générale: $X_{i,j}$ = valeur numérique de la $j^{\text{ème}}$ variable du $i^{\text{ème}}$ sujet

Variables recueillies - Échelles de mesure

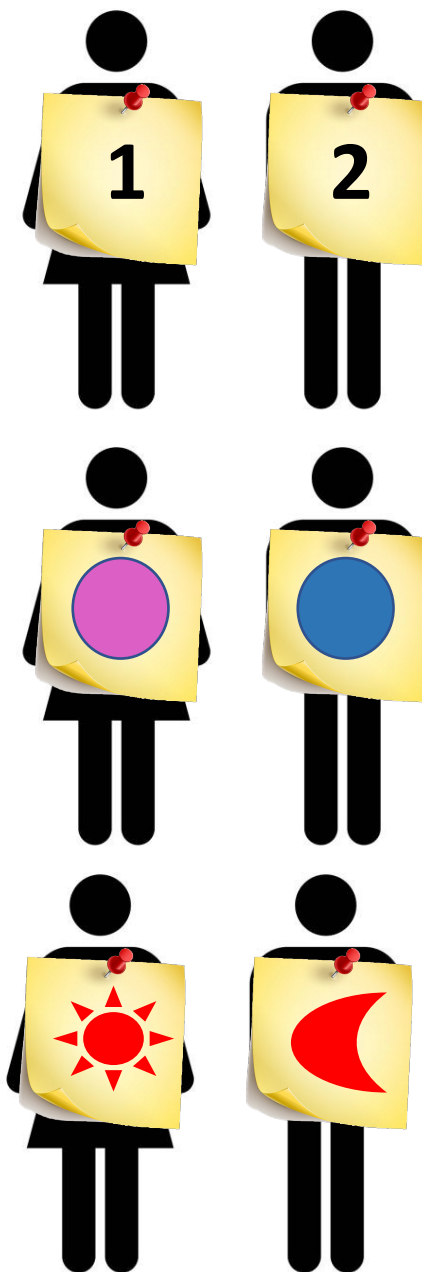
Variables **nominales**

Mot clé: *l'étiquette*

Opérations autorisées:

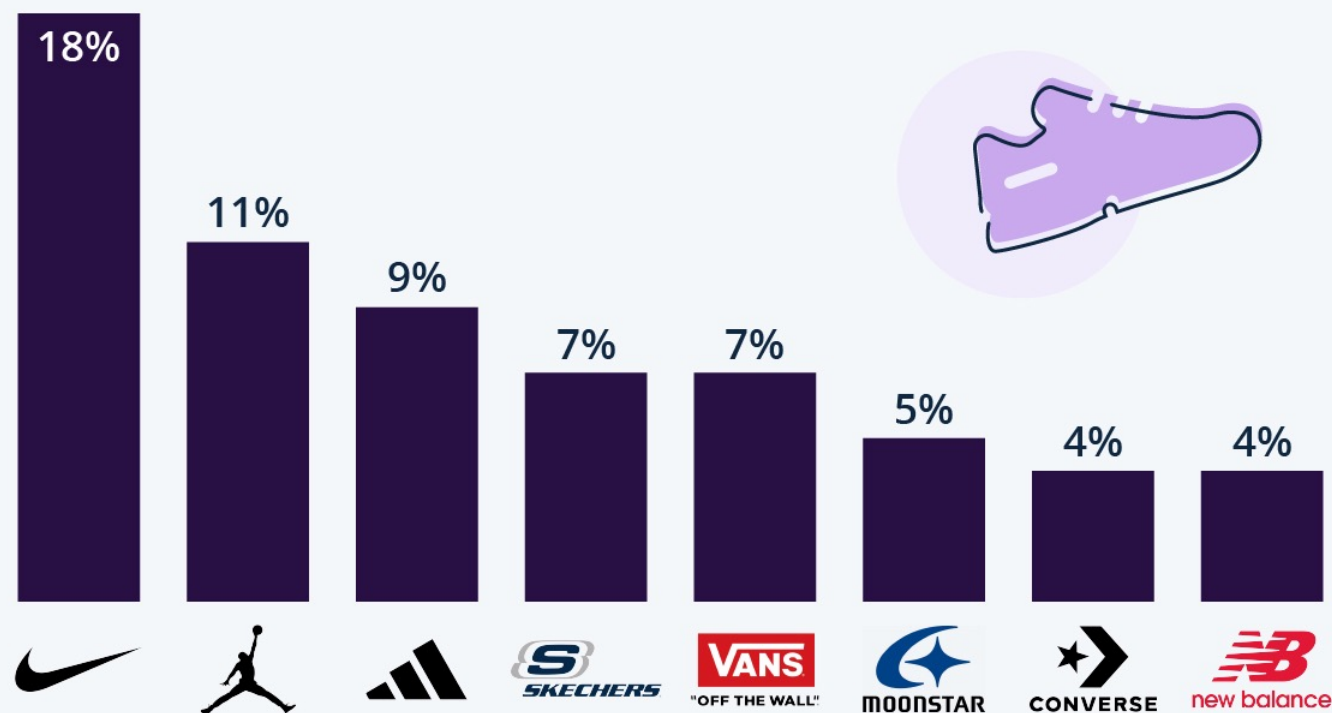
=

≠



The World's Most Popular Sneaker Brands

Estimated global market share of the world's most popular sneaker brands in 2022*



* excl. functional athletic footwear, e.g. dedicated running, football or tennis shoes

Source: Statista Market Insights

Variables recueillies - Échelles de mesure

Variables **ordinales**

Mot clé: *le rang*

Opérations autorisées:

=

≠

>

<

Classement

Les meilleures universités de France

Classement des universités françaises selon le score obtenu dans l'étude Times Higher Education 2022 (100 = le meilleur)

1		Université Paris Sciences & Lettres	74,1
2		Sorbonne Université	63,6
3		Institut Polytechnique de Paris	62,9
4		Université Paris-Saclay	60,4
5		Université de Paris	57,2
6		École des Ponts ParisTech	50,3
7		Université Aix-Marseille	48,0
7		Université de Bordeaux	48,0
7		ENS de Lyon	48,0
7		Université de Montpellier	48,0
11		Université Grenoble Alpes	46,0
12		IMT Atlantique	46,0

Étude couvrant 1 600 universités dans 99 pays et territoires et basée sur 13 indicateurs regroupés en 4 domaines : enseignement, recherche, transfert de connaissances et ouverture internationale.

Source : Times Higher Education "World University Ranking 2022"

Variables recueillies - Échelles de mesure

Variables **d'intervalle**

Mot clé: *Zéro absolu: NON*

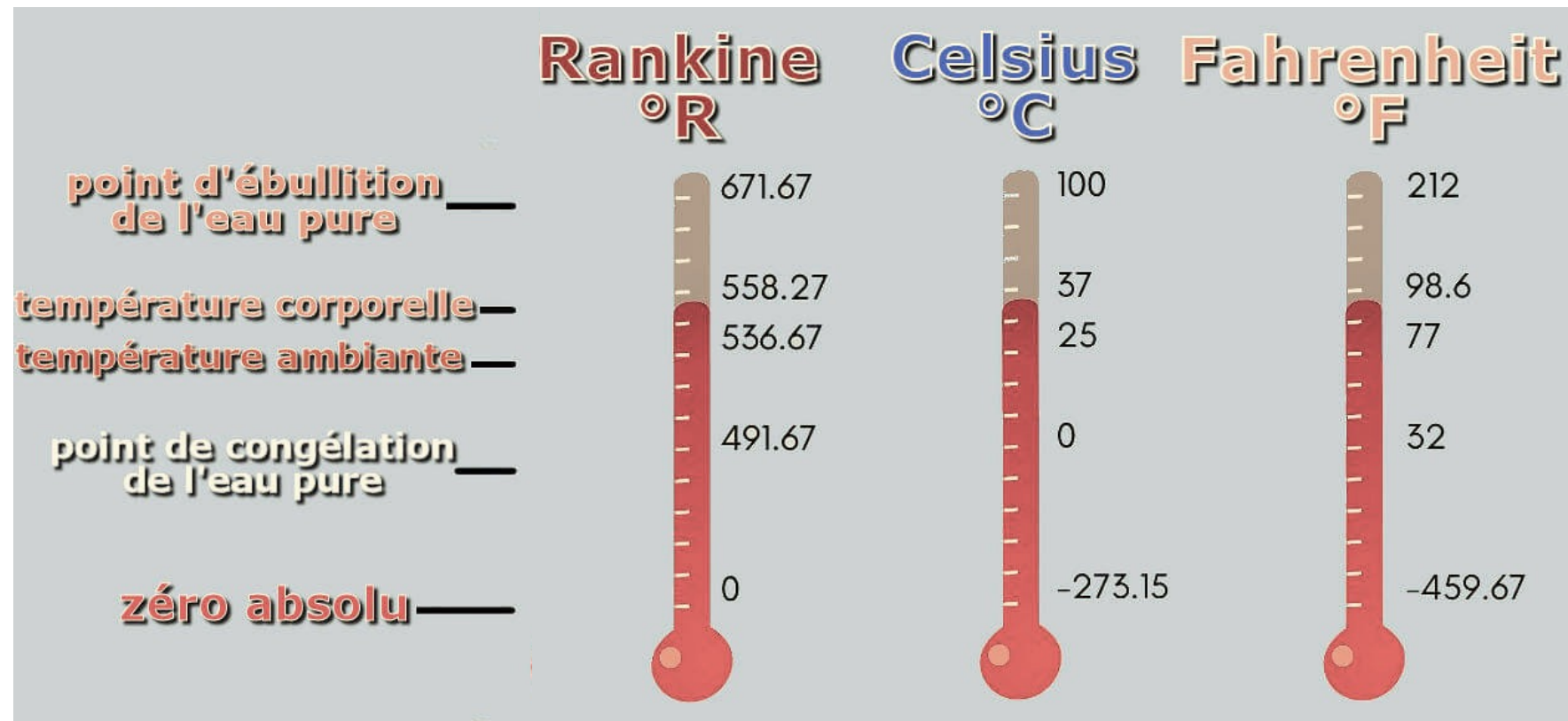
Opérations autorisées:

$=$ $+$

$\neq$ $-$

$>$

$<$



Variables recueillies - Échelles de mesure

Variables **proportionnelles**

Mot clé: *Zéro absolu: OUI*

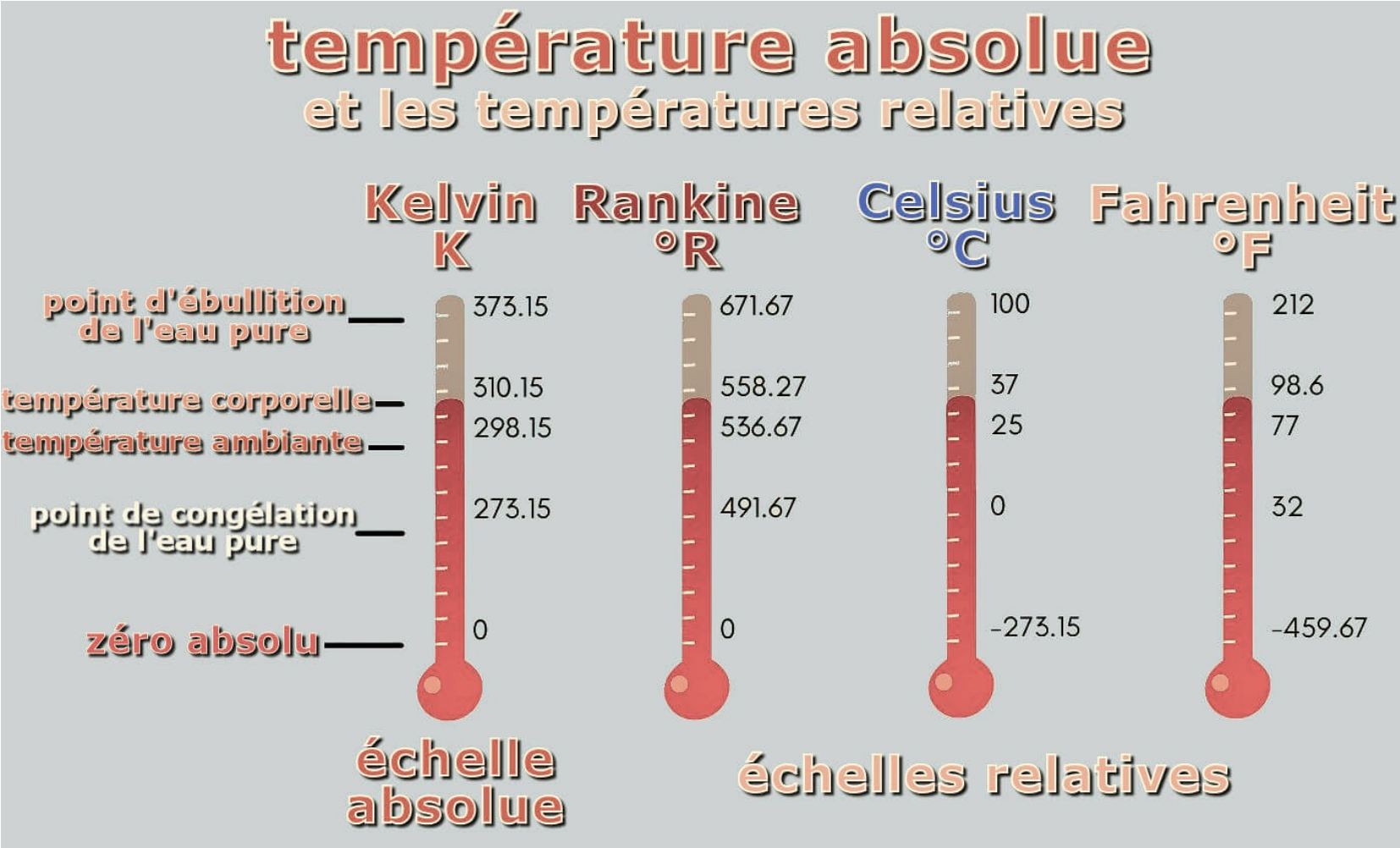
Opérations autorisées:

= +

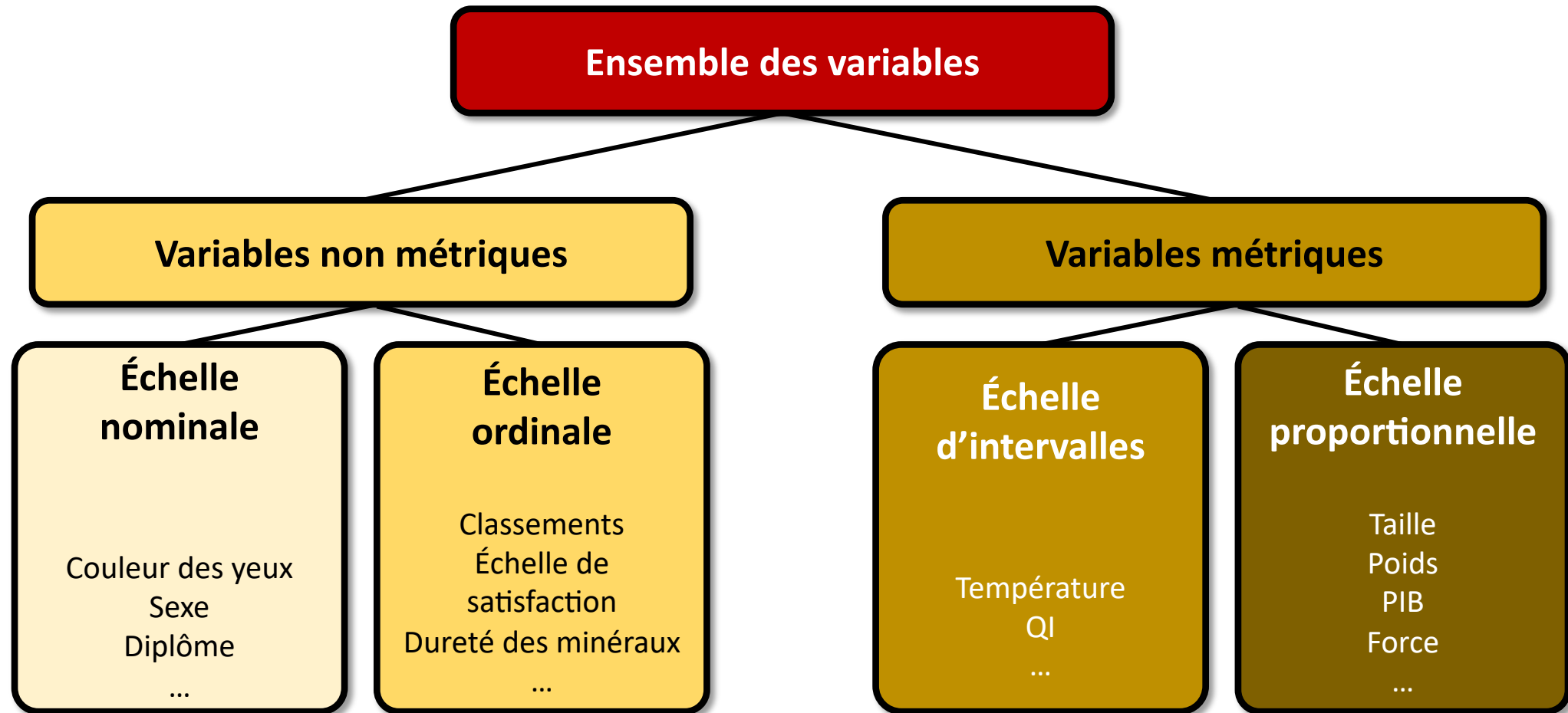
≠ -

> x

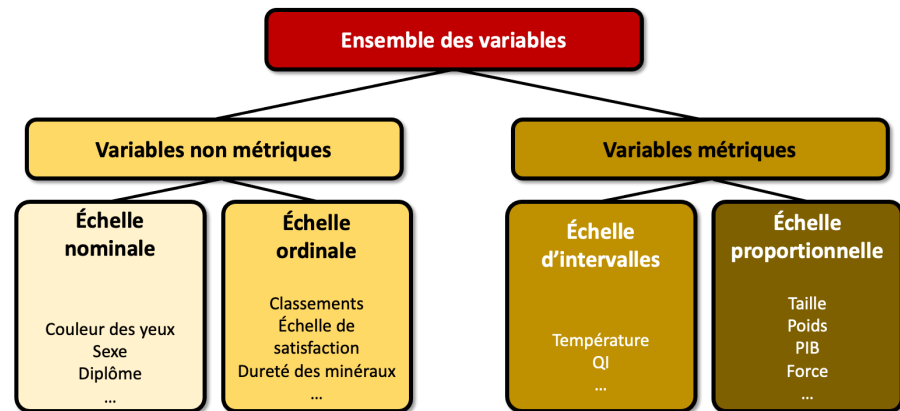
< ÷



Variables recueillies - Échelles de mesure

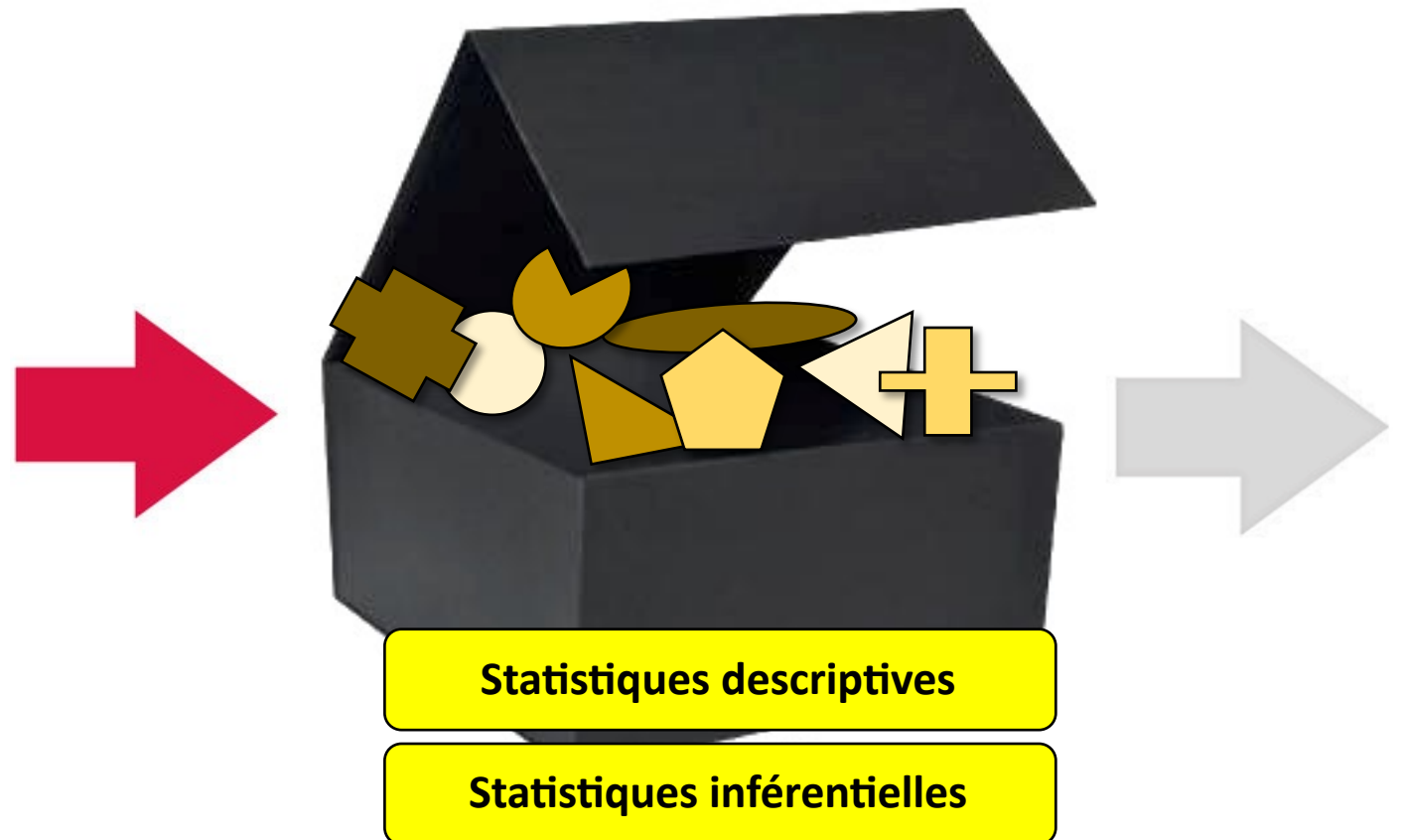


Variables recueillies - Échelles de mesure



On peut faire
peu de choses

On peut faire un
tas de choses



FATTO A MANO

Intro

Variables recueillies – échelles de mesure

Statistiques descriptives

Représentation graphique des résultats

Normalité

Statistiques inférentielles