**Décrivez et nettoyez votre jeu de données**

Individu : unité d’observation, définis par des variables

Population : ensemble des individus, noté N

Échantillon : sous-ensemble d’individus, noté n

On parle souvent de jeu de données (ou data set). Pas de définition précise mais considérés comme un échantillon.

**Les différents domaines de la statistique**

***Les statistiques descriptives***

On présente, décrit et résume le jeu de données, à l’aide de graphiques et de mesures (moyenne, écart-type…).

En stat descr, chaque graphique (ou chaque mesure) est calculé(e) sur 1 ou 2 variables à la fois, pas plus.

***L’analyse de données***

Prolongement des stat descr., mais on étudie plutôt relations entre 3 variables ou plus. Représenter des graphiques avec 3, 4 ou 100 dimensions n’est plus possible sur du papier à deux dimensions. Il faut donc utiliser des techniques spéciales (voir [cours](https://openclassrooms.com/fr/courses/4379436-explorez-vos-donnees-avec-des-algorithmes-non-supervises) ).

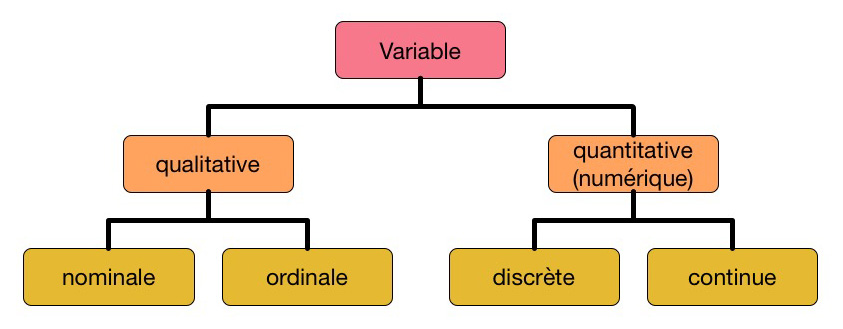
***Les statistiques inférentielles***

Analyse de données d’un sous-ensemble d’une population pour en d.duire caractéristiques globale de la population. On parle d’estimateurs ou de tests statistiques.

***La modélisation statistique***

Observation des caractéristiques d’un échantillon, puis formalisation de ces observations par des règles mathématiques. Cette formalisation s’appelle un modèle probabiliste. Une fois que l’on a décrit un phénomène par un modèle, on peut faire de la prédiction ou de la prévision.

**Variables**



Variable qualitative : les valeurs qu’elle prend sont appelés catégorie ou modalités.

Autres variables : booléennes. Variables qualitatives qui ne prennent que deux modalités (0 ou 1, True ou False, oui ou non..)

Cas des dates : variables qualitatives ordinales

**Distribution empirique**

Modalités : différentes possibilités d’une variable qualitative

Valeurs : différentes possibilités d’une variable quantitative

Effectif : ensemble des valeurs ou (modalités) d’une variable

Fréquence : division d’un effectif par le nombre d’individus de l’échantillon

Distribution empirique d’une variable : ensemble des valeurs (ou modalités) prises par cette variable, ainsi que leur effectif associé et parfois, leur fréquence associé.

