Statistiques descriptives

Atelier théorique #4

Pourquoi faire des statistiques descriptives?

1. Connaître les composantes du jeu de données.
2. Explorer les caractéristiques de chaque variables.
3. Développer une intuition par rapport au jeu de données.
    - Par intuition, on veut dire qu'en

Types de données

Tableau des types de données utilisés par pandas et numpy

Pandas dtype	Python type	NumPy type	Usage
object	str or mixed	string_, unicode_, mixed types	Text or mixed numeric and non-numeric values
int64	int	int_, int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float_, float16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	datetime	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two datetimes
category	NA	NA	Finite list of text values

Variable cible (“*target*”)

En apprentissage automatique, chaque jeu de données est composé de variables pour “***l’entraînement***” et d’une variable “***cible***” ou “***target***” dont on veut ultimement parvenir à prédire la valeur.

Pour développer un algorithme qui parvient à prédire adéquatement la valeur cible, il est est important de savoir si le jeu de données est balancé.

Exemple :

Prenons le cas d’une variable cible binaire (2 catégories possibles, 1 ou 0), on peut regarder le nombre d’observations par catégories pour évaluer le débalancement.

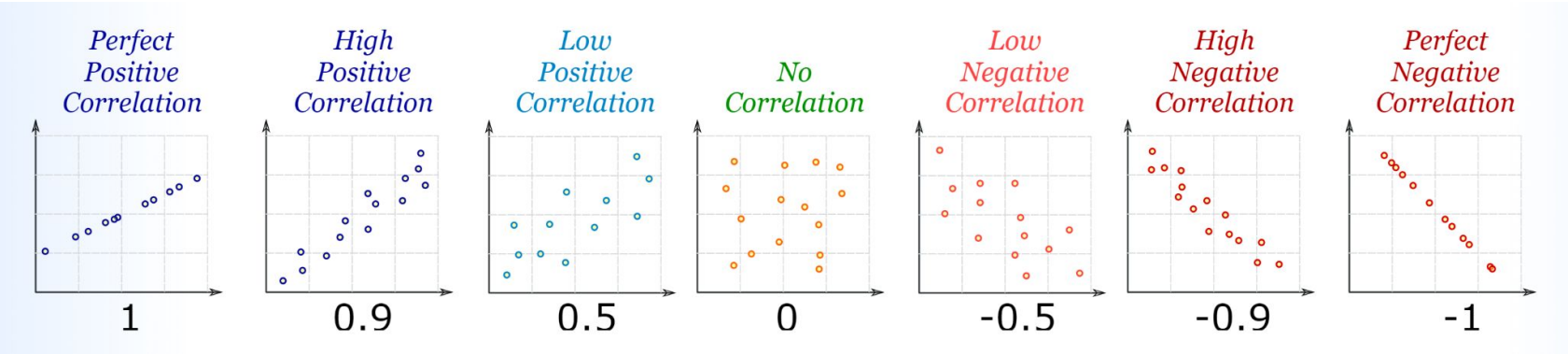
0	200
1	200

Balancé

0	50
1	350

Débalancé

Corrélations



La valeur obtenue donne la force de la corrélation et la direction (positive ou négative).

- 1 = corrélation **positive** parfaite
- 0 = aucune corrélation
- -1 = corrélation **négative** parfaite

Corrélations

Définition : Une corrélation correspond à la relation entre 2 variables qui sont liées entre elles.

Dans ce cours, on utilisera uniquement la **corrélation de Pearson** qui assume une distribution normale des données.

La performance de certains modèles d'apprentissage automatique (tel que les modèles linéaires et logistiques) peut être affecté par de fortes corrélations entre les variables. Il est donc important de vérifier s'il y a des corrélation fortes avant de sélectionner un modèle.

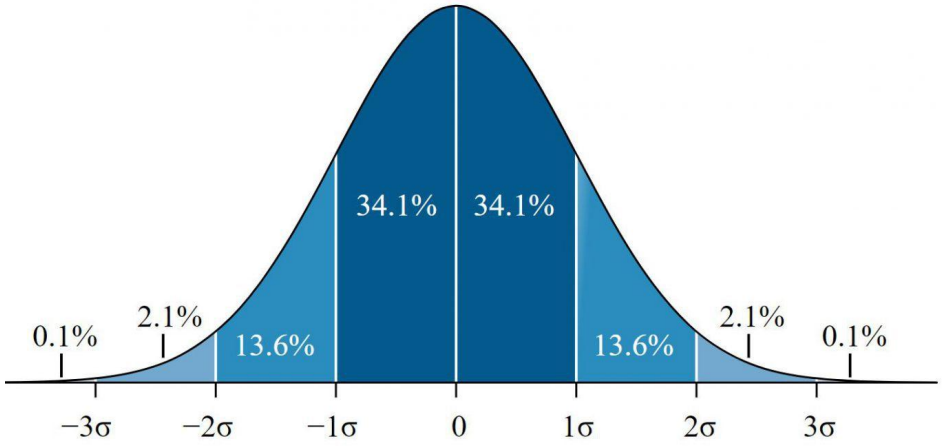
Distribution normale

Propriétés d'une courbe avec une distribution normale "parfaite" (*ne représente pas la réalité*) :

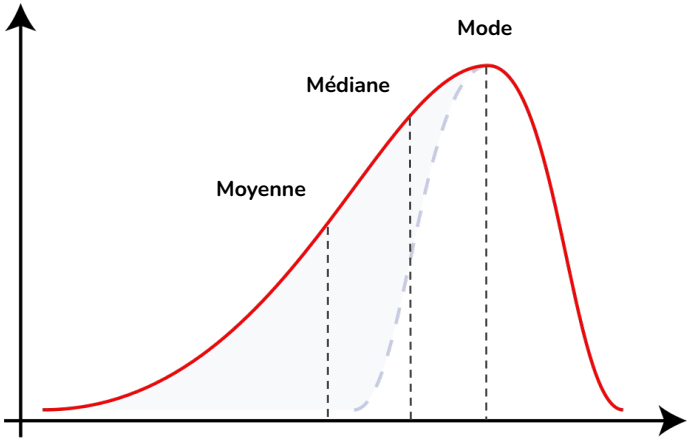
- La courbe de distribution des données est en forme de cloche
- La moyenne et la médiane sont égales
- La courbe est centrée sur la moyenne

Distribution normale des données :

- 68.2% des valeurs sont situés dans un intervalle de +/- 1 écart-type
- 95.4% des valeurs sont situés dans un intervalle de +/- 2 écart-types
- 99.6% des valeurs sont situés dans un intervalle de +/- 3 écart-types

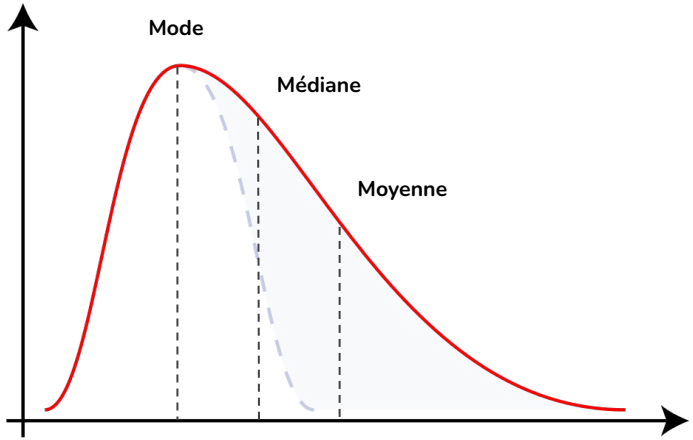


Distribution asymétrique (“*skewed*”)



Asymétrie négative

Moyenne < Médiane < Mode



Asymétrie positive

Moyenne > Médiane > Mode

Distribution asymétrique (*"skewed"*)