



@SDAConsulting

FORMATION DEVENIR DATA ANALYST





1 Fondements en Statistiques avec les logiciels



Cette première partie de la formation s'étendra sur 12 séances, au cours desquelles nous traiterons les points suivants :





Généralités et vocabulaire de base



Dans cette partie, nous allons définir les concepts liés à la statistique



Statistique

Objet

La statistique descriptive sert à décrire une population [un (gros) ensemble d'unités statistiques élémentaires] à l'aide d'indicateurs numériques ou de techniques graphiques.

Pertinence

C'est dans le souci d'avoir les statistiques qu'est née la Statistique.

- Les statistiques, ce sont des données chiffrées. Elles sont créées par :
- Les services de l'Etat
- Les organismes spécialisés
- Les entreprises publiques
- Les particuliers

Etape

Après avoir créée les statistiques, il faut les :

- Présenter
- Analyser
- Et, interpréter

Statistique

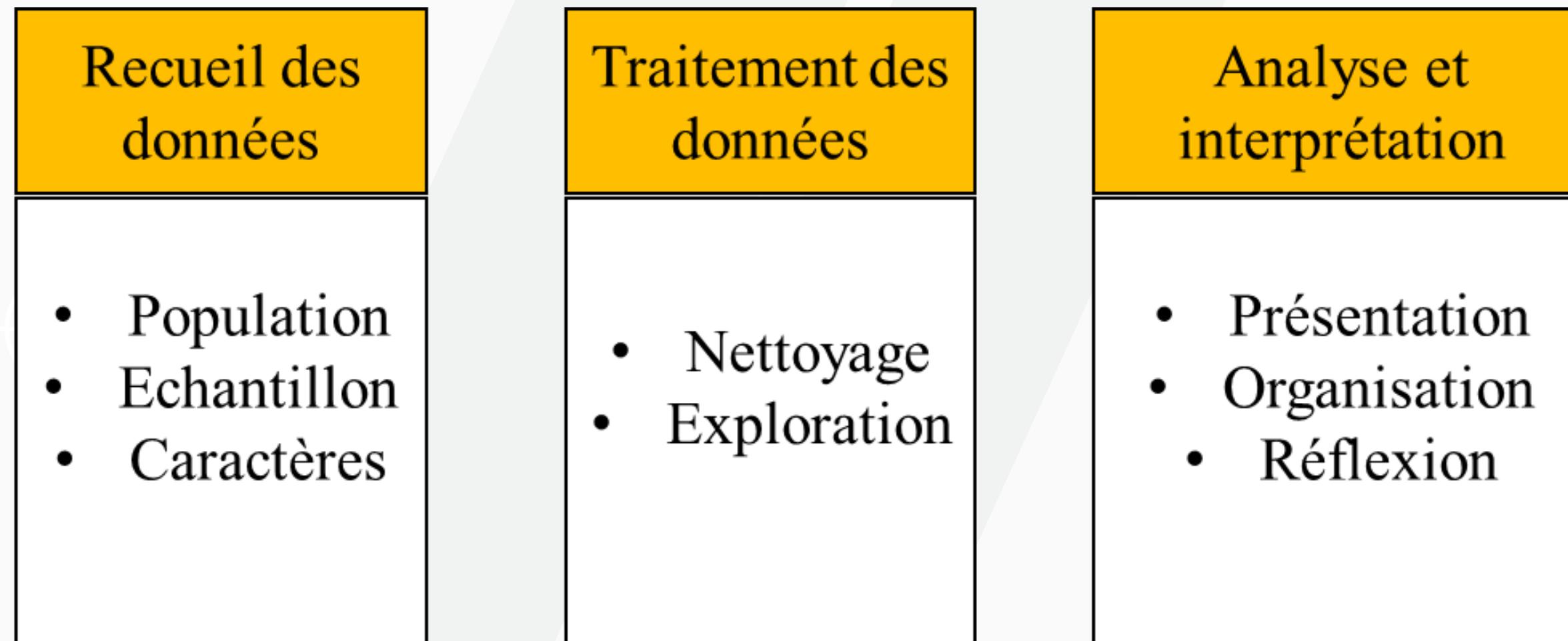
Définition

Ce traitement fait allusion à la Statistique, qui est une discipline étudiant des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles.

« Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation »
(Encyclopédia Universalis)

Statistique

Démarche statistique



Concepts

Variables

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées variables

Population

Le groupe ou l'ensemble d'éléments sur lequel porte l'étude statistique est appelé la population.

Individu

Chaque élément constitue une unité statistique

Concepts

Modalités

En statistique, chaque individu est décrit par un ensemble de variables X . Ces dernières doivent avoir des modalités (ce sont les situations dans lesquelles les individus ou l'individu se retrouve) qui seront à la fois exhaustives et incompatibles.



Modalité incompatible

Càd chaque individu doit appartenir à une et une seule modalité étudiée

Modalité exhaustive

Il n'y a aucune autre possibilité que celle prévue dans le caractère; en un mot ça signifie suffisant.

Concepts

Echantillon

Généralement, la population est trop vaste pour pouvoir être observée exhaustivement. On étudie alors la variable sur un sous ensemble de la population appelée échantillon

Exemple

Supposons qu'on mène une étude sur un caractère X prenant ses valeurs dans Ω , sur une population P. Si l'échantillon est un groupe d'apprenants de SDA

- Un individu est un apprenant.
- Une population, c'est tous les apprenants de SDA
- Les variables peuvent être le sexe, l'âge, le statut matrimonial, la profession, etc.

Classification variables

Variables qualitatives

Ce sont des variables non mesurables ou celles qui s'expriment par l'appartenance à des modalités non mesurables.

Exemple: Sexe

Variables quantitatives

Variable quantitative s'exprimant par des nombres réels, ayant des modalités mesurables par exemple la taille des individus ou les résultats d'un examen.

Exemple: Poids

Classification variables

Variable nominale: On dit que X est une variable nominale s'il n'est pas possible de définir de façon naturelle un ordre sur l'ensemble de ses modalités.

Exemple : Statut matrimonial={Célibataire, marié, Veuf, divorcé}

Variables qualitatives

Variable ordinaire: On dit que X est une variable ordinaire s'il est possible de définir un ordre sur l'ensemble des ses modalités ; Ex: Mention={Très bien, Bien, Assez bien, passable, sans mention}

Classification variables

Variable discrète: c'est lorsque Ω est une suite finie ou infinie d'éléments de N . Ex:
Nombre d'enfant= {1, 2, 3...}

Variables qualitatives

Variable continue: c'est lorsque toutes les valeurs d'un intervalle de R sont acceptables.
Ex: Poids={1,40; 1,50; 2...}

Classification variables

Concept clé en STATISTIQUE

La variabilité, qui signifie que des individus en apparence semblables peuvent prendre des valeurs différentes. Et, le rôle d'un Data analyste est d'étudier cette variabilité pour connaître sa cause et tirer des conclusions.

Autres carrières

Série statistique

C'est un ensemble de données qui représente des observations relatives à une même variable, organisées de manière à pouvoir être analysées.

Ex: 2, 5, 6, 4, 5, 4, 2, 1, 6, 5, 1, 2

Fréquence absolue

c'est un nombre absolu d'individu et symbolisé par ni

Fréquence relative

C'est le rapport entre l'effectif et l'effectif total

Exercices

ex01

Voici une série de variables : nombre d'enfants dans une famille, couleur des yeux, catégorie socio-professionnelle, Province de naissance, Niveau d'études, Revenu, Poids, Sexe, Age, Type de voiture, Taille.

T.D: Spécifier pour chacune de ces variables si elle est qualitative, quantitative, continue, discrète. Proposer des modalités adéquates pour chacune de ces variables.

Exercices

exo2

On dispose des résultats d'une enquête concernant l'âge des apprenants de SDAConsulting.

12, 14, 40, 35, 26, 30, 50, 75, 30, 45, 25, 55, 30, 28, 25, 50, 40, 25, 35, 50.

T.D: Donner la population (Taille de la population), l'unité statistique, le caractère, les modalités du caractère, les fréquences absolues et relatives.

Exercices

exo3

Un individu qui désire d'arrêter de fumer note le nombre de cigarette qu'il allume quotidiennement pendant 1 mois. A la fin de cette période, il obtient la compilation suivante:

- Deux fois, il a réussi à ne pas fumer pendant la journée;
- Une fois, il n'a fumé qu'une cigarette;
- Une fois, il n'en a fumé deux;
- Sept fois, il n'en a fumé trois;
- Douze fois, il n'en a fumé quatre;
- Huit fois, il en a fumé cinq dans la journée.

Exercices

exo3

T.D:

- Les observations de cet individu sont-elles porté sur une N ou un n choisi hasard à l'intérieur de celle-ci?
- Quel est n ou N observé ici ?
- Quel est le caractère étudié?
- De quel type de caractère s'agit-il?
- Présenter le tableau de cette distribution en y indiquant les modalités et leurs effectifs respectifs.

Exercices

exo3

T.D:

- Les observations de cet individu sont-elles porté sur une N ou un n choisi hasard à l'intérieur de celle-ci?
- Quel est n ou N observé ici ?
- Quel est le caractère étudié?
- De quel type de caractère s'agit-il?
- Présenter le tableau de cette distribution en y indiquant les modalités et leurs effectifs respectifs.



B

Traitement des données

Cette section aura deux points:

- Exploration des données
- Nettoyage des données

Exploration des données

Ouverture

Ouvrir le fichier de données à traiter à l'aide d'un logiciel approprié (par exemple SPSS si fichier avec extension .sav, Stata si fichier avec extension .dta, etc.)

Parcours

Parcourir le fichier de données pour confronter la liste des variables aux questions figurant dans le questionnaire.

Exploration des données

Identification

Identifier l'unité d'observation (souvent en lignes) et son identifiant unique. Il est à noter qu'en général, les observations sont en lignes (1 ligne=1 observation) et les variables sont en colonnes.

Identifiant unique

Déterminer l'identifiant unique de chaque individu : c'est la ou les variables permettant d'identifier de façon univoque chaque individu figurant dans le fichier de données. Lorsqu'il y a un jeu de variables servant d'identifiant, on peut décider de créer une seule variable servant d'identifiant unique par exemple en concaténant ces variables ou à l'aide d'une formule de calcul appropriée) ;

Exploration des données

Tri à plats

Sortir des tris à plats (tableaux à simple entrée) ou des statistiques descriptives (min, max, moy, médiane, écart-type, etc.) pour toutes les variables. Ceci permet de se faire une idée de la complétude des données par variable, des problèmes de doublons, des non réponses, des plages de valeurs, de certaines valeurs erronées et/ou atypiques, de la variabilité des réponses à chaque question, les variables à codifier, etc. ;

B Nettoyage

Cette section aura deux points:

- Données manquantes
- Les doublons

Nettoyage

Données manquantes

Les données manquantes se produisent lorsqu'une ou plusieurs valeurs d'une variable ou d'un ensemble de données ne sont pas enregistrées ou ne sont pas disponibles.



Doublons

Les doublons sont des lignes ou des enregistrements dans un jeu de données qui sont identiques ou quasi identiques

Nettoyage

Données manquantes

Supprimer les valeurs manquantes si c'est moins de 5%

Solutions

Remplacer les valeurs manquantes par une valeur constante ou la moyenne

Nettoyage

Les doublons

Solution

Supprimer les doublons



C Analyses univariées

Les analyses univariées sont un type d'analyse statistique qui porte sur une seule variable à la fois.

Plan

Dans ce point, nous allons apprendre:

1

Tableaux à une dimension

2

Graphiques

5

Mesures de formes

3

Mesures de tendance centrale

4

Mesures de dispersion



STATISTICAL
DATA
ANALYSIS
CONSULTING

Tableau

Tableaux à une dimension

Un tableau à une dimension (ou tableau unidimensionnel) est un tableau où les données sont présentées en fonction d'une seule variable.

Eléments

- **Titre:** Le tableau doit avoir un titre explicite indiquant clairement le sujet ou le type de données qu'il présente.
- **Colonnes et lignes :** Les données sont organisées en colonnes (pour les catégories ou les variables) et en lignes (pour les observations ou les catégories).).
- **Étiquettes des colonnes et des lignes :** Chaque colonne et chaque ligne doit être étiquetée avec des noms explicites pour indiquer le contenu ou les catégories.
- **Unité de mesure :** Si les données impliquent des unités (par exemple, des dollars, des kilogrammes, des pourcentages), celles-ci doivent être mentionnées clairement.

Tableau

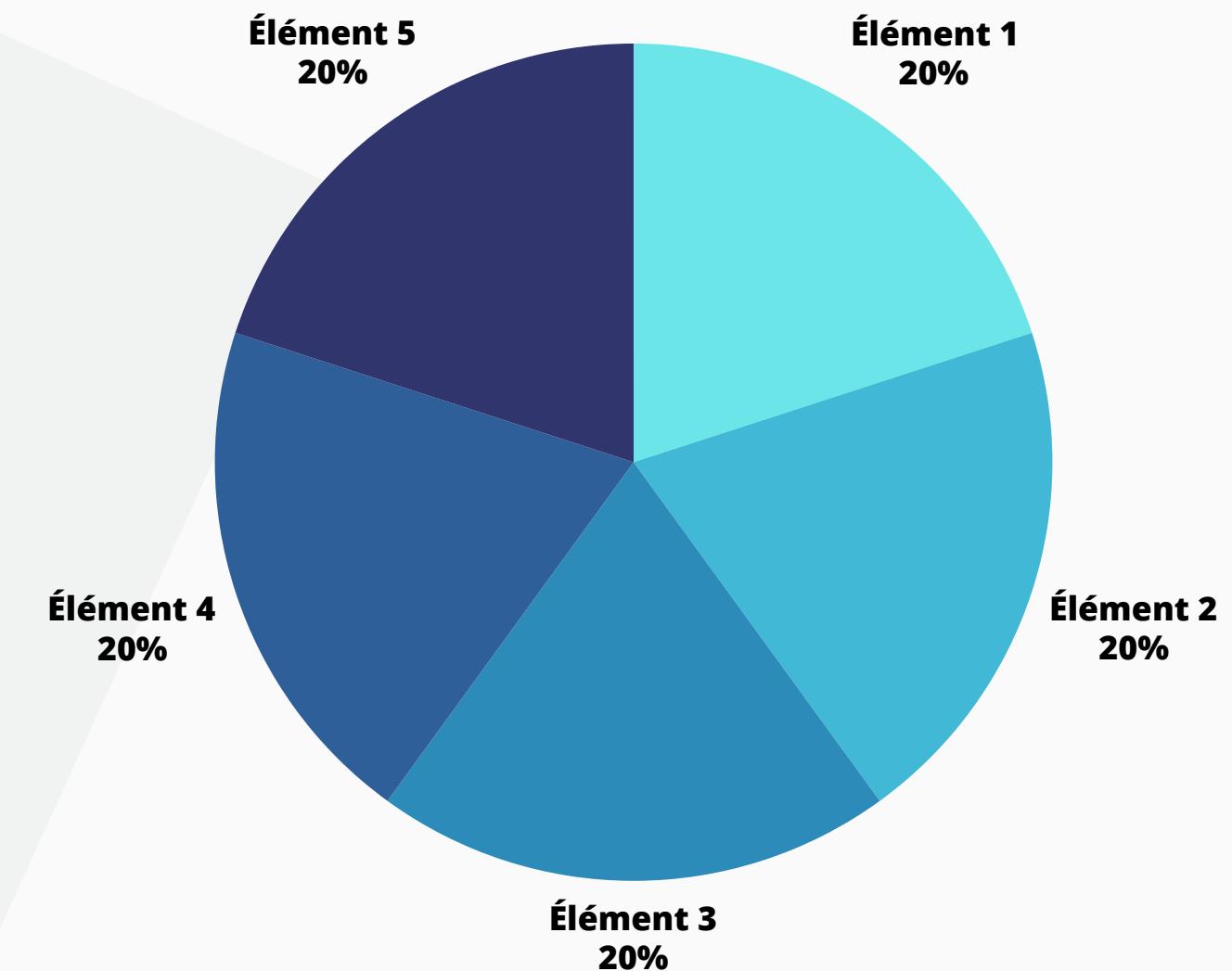
Eléments

- **Source:** c'est l'origine ou le point de départ des informations, des idées, des données ou des faits. Dans le contexte des recherches, des études ou de la rédaction, une source fait référence à tout document, personne, publication, ou autre forme de média d'où proviennent les informations utilisées pour soutenir une argumentation, une analyse ou une présentation.
- **Numéro**



Graphique

Un graphique est une représentation visuelle des données qui permet de mieux comprendre et interpréter des informations numériques ou statistiques. Il sert à illustrer des tendances, des relations ou des comparaisons de manière plus intuitive et accessible que les tableaux de données brutes. En mettant en évidence les schémas et les variations, les graphiques facilitent l'analyse et la prise de décision.



Graphique

VARIABLES	GRAPHIQUES CORRESPONDANTS
Quantitative discrète	Diagramme en bâton
Quantitative continue	Histogramme
Qualitative nominale et Ordinale	Diagramme en secteur et en barre



Mesures de tendance centrale

Moyenne arithmétique est la somme de toutes les valeurs d'une variable, divisée par le nombre total d'observations.

Médiane c'es la valeur médiane est la valeur qui sépare les données en deux moitiés égales

Le mode est la valeur qui apparaît le plus souvent dans un ensemble de données.

Mesures de dispersion

Les mesures de dispersion indiquent dans quelle mesure les données sont étalées ou concentrées autour de la moyenne ou d'une autre valeur centrale.

Étendue c'est la différence entre la valeur maximale et la valeur minimale dans un ensemble de données.

Variance : c'est la moyenne des carrés des écarts entre chaque valeur et la moyenne. La variance mesure à quel point les données sont dispersées autour de la moyenne. Plus la variance est grande, plus la dispersion est importante.

Mesures de dispersion

L'écart-type mesure la dispersion des données par rapport à la moyenne. Plus l'écart-type est élevé, plus les données sont dispersées.



Interprétation : Un écart-type faible indique que les valeurs sont proches de la moyenne, tandis qu'un écart-type élevé indique une forte dispersion.

Mesures de dispersion

Écart interquartile (IQR): L'écart interquartile mesure la dispersion au sein du milieu des données, c'est-à-dire la différence entre le premier et le troisième quartile (Q3 et Q1).

Interprétation: Il mesure la dispersion des 50 % centraux des données, moins sensible aux valeurs extrêmes.

Mesures de dispersion

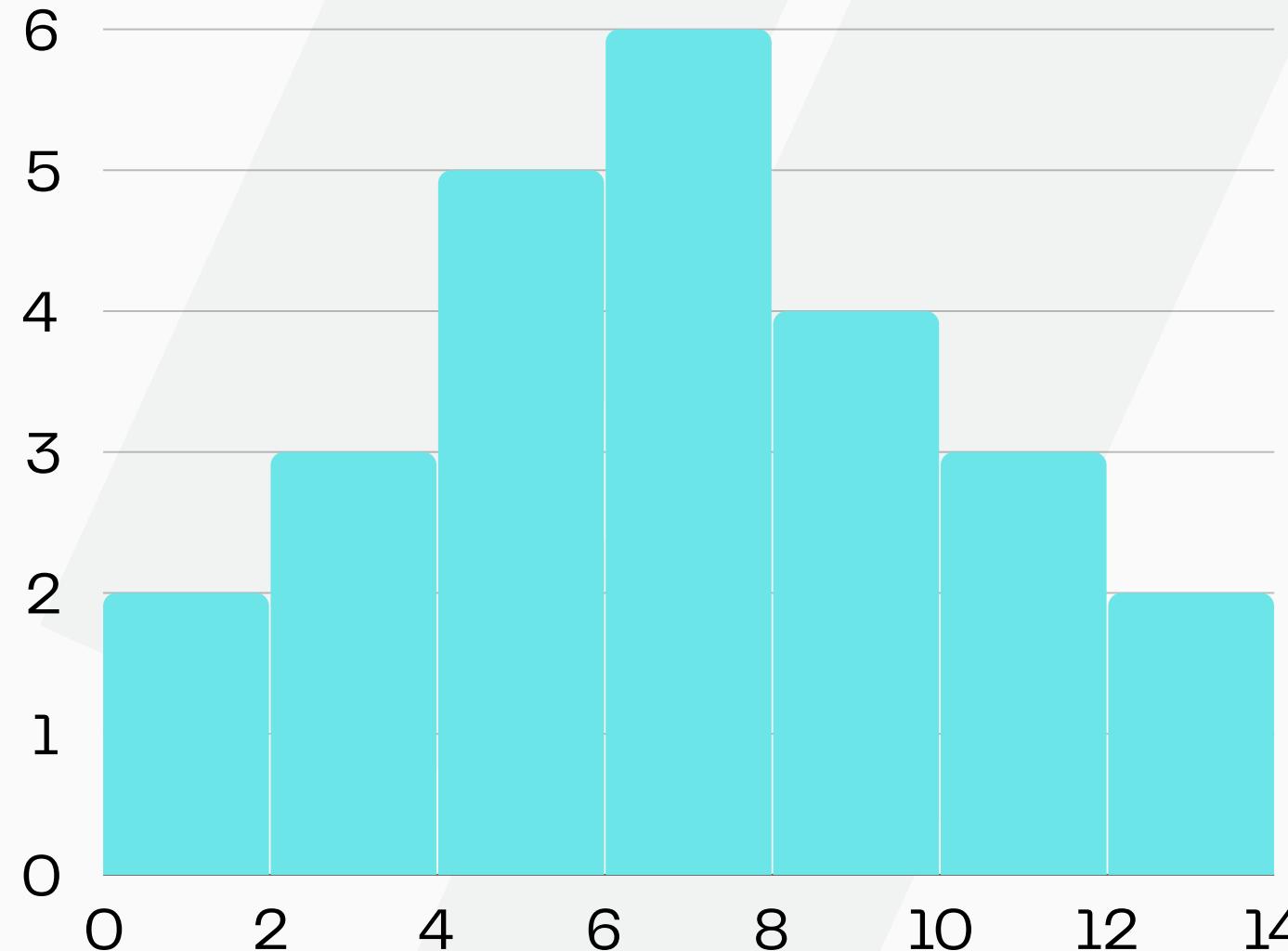
Coefficient de variation (Cv): C'est le rapport de l'écart-type à la moyenne. Il est souvent exprimé en pourcentage.



Interprétation : Il permet de comparer la dispersion relative entre différents ensembles de données.

Mesures de forme

Les mesures de forme caractérisent la forme d'une distribution, c'est-à-dire si elle est symétrique ou asymétrique, et si elle est aplatie ou pointue.



Mesures de forme

Coefficient d'asymétrie (skewness): L'asymétrie mesure le degré et la direction de l'asymétrie d'une distribution.

Interprétation :

- Si le coefficient est proche de 0, la distribution est symétrique.
- Si il est positif, la distribution est asymétrique à droite (longue queue vers la droite).
- Si il est négatif, la distribution est asymétrique à gauche (longue queue vers la gauche).

Mesures de forme

Coefficient d'aplatissement (kurtosis): Le kurtosis mesure la "pointitude" d'une distribution, c'est-à-dire si les valeurs sont concentrées autour de la moyenne ou dispersées dans les queues.

Interprétation :

- Si le kurtosis est égal à 3 (ou 0 après soustraction), la distribution est mesokurtique (comme la distribution normale).
- Un kurtosis supérieur à 3 indique une distribution leptokurtique (concentrée autour de la moyenne, avec des queues épaisses).
- Un kurtosis inférieur à 3 indique une distribution platykurtique (plus aplatie, avec des queues fines).

Projet 1: Analyse des Tendances de Vente du Magasin Maman Wa Bolingo

1. Contexte

Maman Christine, propriétaire de la boutique Maman Wa Bolingo, spécialisée dans la vente en ligne des outils informatiques, cherche à optimiser la performance de ses ventes et à mieux comprendre l'engagement de sa clientèle. Pour ce faire, elle vous sollicite en tant que Data Analyst afin d'analyser sa base de données intitulée « Données transactionnelles ».



2. Objectifs

- Identifier les tendances des ventes
- Comprendre les comportements des clients
- Formuler des recommandations d'amélioration





STATISTICAL
DATA
ANALYSIS
CONSULTING

@SDAConsulting

Merci

