Rapport du TP de Statistique

Mehdi Bahi 22014246 (G.I)

Professor: Phd. Aniss Moumen

École Nationale des Sciences Appliquées Kénitra
Année Académique 2024-2025

Contents

Chapter 1

Introduction à l'Analyse de Données

1.1 Questionnaire du chapitre

1. **Le logiciel développé répond-il aux besoins du client ?**
2. **Comment peut-on définir l'analyse des données dans le domaine de l'informatique ?**

1.2 Réponse à la première question

Analyse de la question :

- **Client** : L'utilisateur de l'application, identifié de manière unique dans le système.
- **Besoins** : Les exigences du client, incluant leurs demandes, problèmes ou aspirations.
- **Logiciel** : Un ensemble d'interfaces, de programmes et de fonctionnalités implémentées à travers un code source.

Conclusion :

- Les réponses possibles à la question sont "Oui" ou "Non".
- Lorsqu'on teste ces propositions statistiquement (par exemple, déterminer la proportion de "Oui" ou "Non"), elles forment une hypothèse. Cela relève des statistiques inférentielles, qui consistent à calculer des probabilités et à les comparer à un seuil de signification.
- Les calculs comme l'écart-type relèvent des statistiques descriptives.

1.3 Analyse des données

L'analyse des données consiste à examiner des observations pour soutenir la prise de décisions tout en tenant compte des risques associés à ces décisions. Son objectif principal est d'extraire des informations utiles à partir des données.

Étapes clés de l'analyse des données :

1. **Définir la problématique et les données** :
   - Formuler des questions de recherche et lister les réponses possibles.
   - Définir les concepts et les caractériser en créant un *Modèle Conceptuel de Données (MCD)*. Cela implique d'identifier les entités et leurs attributs.
2. **Collecte des données** :
   - Recueillir des données en consultant les clients et structurer les observations de manière systématique.
   - Capturer et organiser les données pour une analyse ultérieure.
3. **Pré-traitement** :
   - Évaluer la qualité des données collectées afin de garantir leur validité et leur fiabilité.
   - Traiter les données manquantes ou aberrantes à l'aide de méthodes de pré-traitement appropriées.
4. **Traitement des données** :
   - Se concentrer sur les données validées pour effectuer des statistiques descriptives en calculant les indicateurs de distribution aléatoire.
   - Étudier différents types de variables (statistiques univariées, bivariées, multivariées).
   - Réaliser des tests d'hypothèses pour tirer des conclusions.

Objectifs de l'analyse des données :

- Extraire des informations significatives à partir de données brutes.
- Évaluer l'utilité des informations extraites en examinant les indicateurs de distribution.
- Favoriser la réflexion critique et faciliter la prise de décisions.

Informations supplémentaires :

- Les statistiques descriptives incluent souvent le calcul des corrélations, qui sont essentielles pour identifier les relations entre les variables.

1.4 Index des termes

- **Exigence** : Une variable qualitative avec des réponses possibles telles que “Oui” ou “Non”.
- **Interface** : La frontière de communication entre deux entités.
- **Programme** : Un ensemble d’instructions et d’opérations exécutées par un ordinateur.
- **Fonctionnalité** : Une action ou capacité spécifique qui apporte une utilité supplémentaire.
- **Risque** : La probabilité d’un événement et le seuil de signification utilisé pour l’évaluer.
- **MCD (Modèle Conceptuel de Données)** : Un modèle qui définit les entités et leurs attributs, servant de base à la conception de bases de données.

Chapter 2

Formulation de la problématique

2.1 Problématique

Ensemble de difficultés ou de situations problématiques concernant le sujet ou l'individu analysé. Exemple : une mauvaise organisation du stock. Le problème correspond à une situation défavorable ou critique dont souffre le sujet ou l'individu étudié. (L'analyse des données a pour but d'aider le décideur, c'est-à-dire celui qui subit cette problématique. Il est celui qui sollicite notre intervention pour résoudre cette difficulté.) Le problème doit être suffisamment marquant et significatif pour justifier le recours à une aide à la décision (nous, en tant qu'analystes de données). En somme, un problème est une constatation négative affectant directement le sujet.

2.2 Formulation de la problématique

Il s'agit d'exprimer ou de formuler la problématique, souvent sous forme d'une question de recherche. Exemple : Pourquoi le stock est-il mal organisé ? Avant de poser cette question, il faut d'abord se demander :

- Qu'est-ce que le stock ?
- Qu'entend-on par "mauvaise organisation" ?

Ensuite, il s'agit de formuler des questions préliminaires et de proposer des réponses potentielles. Si ces réponses sont validées, elles deviennent des hypothèses. La formulation de la problématique est un document structuré qui clarifie la problématique. Exemple de structure :

- **Paragraphe 1** : Décrire la constatation de départ.
- **Paragraphe 2** : Formuler les questions de recherche (QR), les sous-questions (SQ), et proposer des réponses. Les QR et SQ servent à enrichir la problématique grâce à une recherche documentaire.
- **Paragraphe 3** : Présenter les résultats de la recherche documentaire.

2.3 Analyse de la problématique

2.3.1 Définition et méthodologie

Analyser la formulation de la problématique pour en extraire les concepts, entités, et relations possibles. Étapes :

1. Formulation de la problématique (flèche implication).
2. Analyse de la problématique :
   - Lecture et identification des concepts/entités.
   - Association (catégorisation, classification textuelle).
   - Identification des caractéristiques et attributs (variables de mesure).
3. Création du dictionnaire de données et, ensuite, du MCD (Modèle Conceptuel des Données).

2.3.2 Recherche documentaire

La recherche documentaire enrichit la problématique en définissant les concepts selon la littérature. Exemple :

Concept	Définition	Source

2.3.3 Dictionnaire de données

Recense les variables de mesure associées au sujet. Exemple :

Nom de la variable	Type	Description

2.4 Contextualisation

La contextualisation consiste à adapter et redéfinir les concepts en fonction du cadre de l'étude. Exemple :

- **Thème** : L'employabilité des étudiants.
- **Problème** : Le manque de compétences des étudiants constaté à travers les études sur les difficultés d'intégration au marché du travail.
- **Contexte** : Étudiants de l'ENSA, Université IBN Tofail.
- **Question de recherche principale** : Pourquoi les étudiants manquent-ils de compétences pour intégrer le marché du travail ?

- **Autres questions** :
  - Est-ce lié à un problème de formation ?
  - À un manque d'outils techniques ?
  - À une mauvaise organisation des étudiants ?
  - À l'absence d'expérience professionnelle ?

Après l'analyse documentaire et la formulation de la problématique, on obtient un dictionnaire de données non contextualisé. Exemple :

Concept	Définition (littérature)	Source	Définition (étude)	Source

2.4.1 Contextualisation via interview

Pour contextualiser les concepts, des interviews doivent être menées avec les individus concernés.

- **Définition** : L'interview est un échange structuré entre un enquêteur et un participant sur un sujet donné. Elle permet de collecter des données qualitatives.
- **Préparation** :
  - Préparer un guide d'entretien avec des questions ouvertes.
  - Organiser le cadre de l'interview (lieu, ambiance, etc.).

2.4.2 Qualité de l'interview

- La qualité dépend de la richesse du contenu et de la durée.
- Durée normale :
  - Individuelle : > 20 min.
  - En groupe : $\geq$ 1h30.
  - Pour un mini-projet : ≥ 10 min.

2.4.3 Caractéristiques des participants

- Genre, âge, filière, etc.
- Méthode : Utiliser une approche en entonnoir (questions ouvertes $\rightarrow$ approfondissement).
- Taille de l'échantillon : Nombre impair (ex. 3, 5).
- Taille minimale : dépend du niveau de similarité des réponses ($\geq 70\%$).

2.4.4 Niveau de similarité

$$\text{Niveau de similarité} = \frac{\text{Somme des occurrences de mots similaires}}{\text{Total des occurrences des mots}}$$

2.4.5 Exemple de guide d'entretien

- **Thème** : L'entrepreneuriat étudiant.
- **Page 1** :
  - Introduction : Présentation de l'objectif de l'entretien.
  - Première question : "Présentez-vous."
- **Page 2** : Questions thématiques (démarche en entonnoir) :
  - **Phase introductive** : "Pour vous, qu'est-ce que l'entrepreneuriat ?"
  - **Phase de développement** :
    - "Pourquoi êtes-vous intéressé par l'entrepreneuriat ?"
    - "Quelles sont vos motivations principales ?"
  - **Phase d'approfondissement** :
    - "Comment l'entrepreneuriat contribue-t-il au développement personnel ?"
    - "En quoi permet-il d'augmenter vos revenus ?"
    - "Quels obstacles rencontrez-vous face à l'entrepreneuriat ?"
  - **Phase de conclusion** :
    - "Quelles solutions proposeriez-vous pour surmonter ces obstacles ?"
    - "Si un guichet automatique simplifiait les démarches administratives, cela vous intéresserait-il toujours ?"

2.5 Le modèle conceptuel

C'est une représentation graphique des concepts et de leurs associations Le concept à

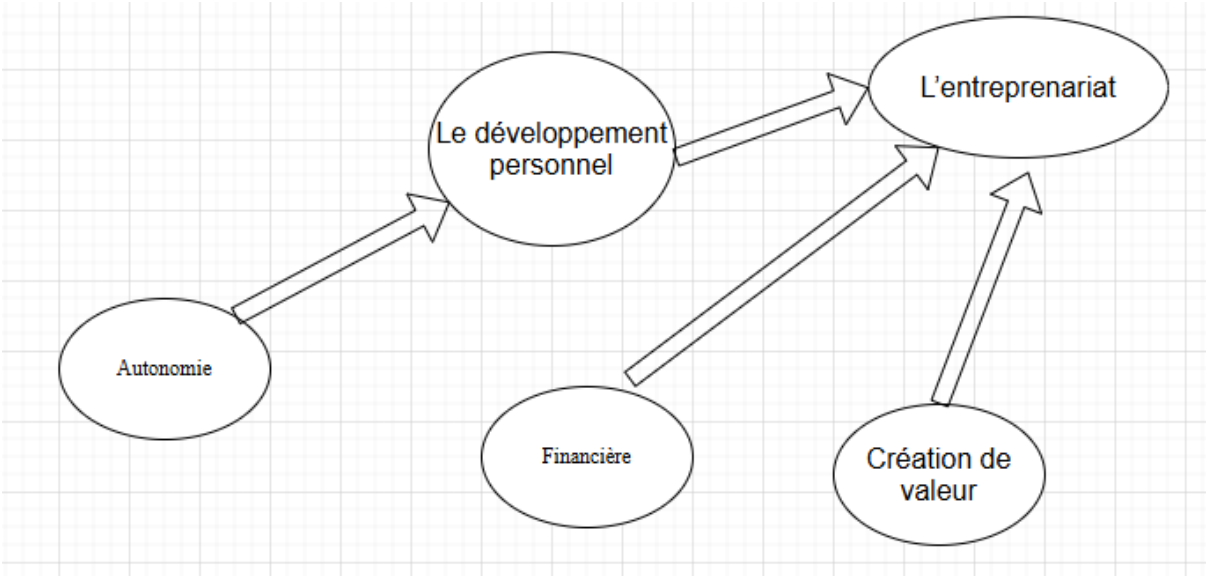


Figure 2.1: Conceptual Model Diagram

Les autres notions en entrée (autonomie, création de valeur, indépendance financière) représentent des antécédents ou des facteurs explicatifs. Au milieu (développement personnel), nous avons un concept intermédiaire.

Un modèle conceptuel contextualisé (MCC) est élaboré à partir de l'analyse des questions issues des interviews. Un modèle théorique est construit sur la base d'une analyse documentaire. Ce modèle précède la construction du MCC.

2.6 Conception de l'instrument de mesure

La grille d'analyse permet d'examiner les réponses des participants sous forme de tableau :

Répondants/Interviewés	**Question 1**	**Question 2**
I1	Texte de I1 (verbatim)	

Un **instrument de mesure** est un outil conçu pour évaluer un phénomène par l'observation. Il regroupe les éléments que l'on cherche à mesurer ou à observer, soit l'ensemble des variables de mesure.

2.6.1 Exemple : Instruments de mesure selon le type d'études

- **Études quantitatives** : questionnaire
- **Études qualitatives** : guide d'entretien

2.6.2 Types de variables de mesure

- **Quantitatives** : chiffres d'affaires, capital, nombre de salariés, nombre d'actionnaires. Ces réponses sont de nature numérique.
- **Qualitatives** : type d'actionnaires, degré d'autonomie, perception de l'entrepreneuriat. Ces réponses prennent la forme de texte.

2.7 Grille d'analyse

Concernant la formulation des réponses, il est nécessaire d'utiliser des phrases complètes.

Interviewé	**Entrepreneuriat**
I1	Pour moi, c'est le fait de prendre des risques.
I2	C'est la mise en œuvre de projets préalablement établis.
I3	C'est l'art de générer des revenus.
I4	Se stabiliser financièrement.
I5	Transformer une idée en une réalité concrète.

Ces éléments de réponse constituent ce que l'on appelle les **i-thèmes de l'entrepreneuriat** (les perceptions des participants sur le concept d'entrepreneuriat).

2.8 Conception des questionnaires

Le questionnaire se compose d'un ensemble de questions fermées, regroupées dans un formulaire structuré en trois parties :

2.8.1 Partie 1 : Préambule

Cette section comprend :

- Une introduction expliquant l'objectif du questionnaire,
- Une note de remerciement,
- Une mention sur la confidentialité des réponses (par exemple : *Toutes vos réponses resteront confidentielles et ne seront utilisées qu'à des fins de recherche dans le cadre de cette étude.*),
- Les coordonnées de l'enquêteur et une signature officielle.

2.8.2 Partie 2 : Corps du questionnaire

Elle regroupe les questions en lien direct avec la thématique étudiée.

2.8.3 Partie 3 : Identification des participants

Elle inclut des informations telles que l'âge, la filière, et l'année d'études.

2.8.4 Exemple : Entrepreneuriat chez les étudiants

Partie 1 : Préambule

Ce questionnaire vise à étudier l'entrepreneuriat chez les étudiants dans le cadre d'un mini-projet. Nous vous remercions pour le temps que vous y consacrerez. Toutes vos réponses resteront confidentielles.

Pour vous, qu'est-ce que l'entrepreneuriat ? Items proposés (cases à cocher) :

- Prise de risque
- Mise en œuvre de projets concrets
- Stabilisation financière

Partie 3 : Identification des participants

Nom	Item 1	Item 2	Item 3	Âge	Genre	Filière
Salma	Non	Oui	Oui	20	F	GI
Joyce	Oui	Oui	Oui	20	F	GI
Idriss	Oui	Non	Oui	20	Autres	GI

Que pensez-vous de l'entrepreneuriat ?

Item	Tout à fait d'accord	D'accord (>70%)	Neutre (50%)	Pas d'accord	Tout à fait pas d'accord
Prise de risque					
Mise en œuvre de projets					
Stabilisation financière					

Chapter 3

Collecte des données

Définitions

Les données représentent l'ensemble des observations recueillies dans le cadre de l'étude d'une problématique donnée, à l'aide d'instruments de mesure adaptés.

- **Dans les études quantitatives**, les données sont structurées.
- **Dans les études qualitatives**, les données sont non structurées.

Par exemple :

- **Quantitatif** : Les questionnaires se composent de questions fermées et constituent la base des études quantitatives.
- **Qualitatif** : Les interviews, souvent enregistrées sous forme audio, utilisent des questions ouvertes.

Les données sont collectées auprès des individus concernés.

- **Recensement** : Lorsque les données sont collectées auprès de la population entière.
- **Sondage** : Lorsque les données sont recueillies auprès d'un sous-groupe représentatif de la population.

3.1 Population / Échantillon / Types d'échantillons

- **Population** : Ensemble des individus concernés par la problématique étudiée.
- **Individu** : Une personne affectée par la problématique, caractérisée par des traits spécifiques.
- **Échantillon** : Un sous-ensemble de la population mère, sélectionné tout en respectant ses principales caractéristiques.

1. **Échantillon représentatif** : Reflète fidèlement les caractéristiques de la population. Les hypothèses testées sur cet échantillon peuvent être généralisées à la population.
    - Exemple :
        - Population : 60 % hommes, 40 % femmes ; Échantillon : 60 % hommes, 40 % femmes.
        - Population : Moyenne d'âge de 20 ans ; Échantillon : Moyenne d'âge de 20 ans.
2. **Échantillon équilibré (ou exploratoire)** : Maintient un équilibre sur une caractéristique qualitative spécifique, sans être représentatif.
    - Exemple :
        - Population : 60 % hommes, 40 % femmes ; Échantillon : 50 % hommes, 50 % femmes.
    - Utilisé pour neutraliser l'effet de certaines variables catégorielles.
3. **Échantillon d'étude** : Ni représentatif, ni équilibré.

3.2 Plan d'échantillonnage

Le plan d'échantillonnage décrit les étapes pour constituer un échantillon pertinent.

Étape 1 : Définir les caractéristiques des individus Les membres de l'échantillon doivent partager les mêmes caractéristiques que ceux de la population cible.

Étape 2 : Déterminer la méthode de sélection des individus

- **Si la population est connue** (liste d'individus finie) : sélection aléatoire.
    - **Aléatoire simple** : Aucun critère spécifique pour constituer l'échantillon.
    - **Systématique** : Application d'une règle fixe (ex. : un pas ou un algorithme de sélection).
    - **Stratifié** : Constitution de sous-groupes au sein de l'échantillon.
    - **Par grappes** : Sélection basée sur des zones géographiques.
    - **Multiple** : Combinaison de plusieurs techniques de sélection.
- **Si la population est inconnue** (liste d'individus infinie ou non définie) : sélection non aléatoire.
    - **Par convenance** : Sélection des individus accessibles.
    - **Par jugement** : Appel à des experts pour identifier les individus pertinents.
    - **Boule de neige** : Sélection progressive par recommandations successives.
    - **Volontaire** : Constitution de l'échantillon via un appel à participation ou un événement spécifique.

Étape 3 : Déterminer la taille de l'échantillon

- **Taille minimale** : Environ 30 individus (+/- 5).
- **Taille optimale** : Dépend de la taille de la population :
  - **Population connue** :

$$n_{\text{opt}} = \frac{\frac{Z^2 \times P(1-P)}{E^2}}{1 + \frac{Z^2 \times P(1-P)}{NE^2}}$$

  - **Population inconnue** :

$$n_{\text{opt}} = f(\text{IC}, \text{ME}) \quad \text{(basé sur le théorème des grands nombres)}$$

Paramètres :

- N : Taille de la population.
- IC : Intervalle de confiance (généralement ¿= 70 %).
- ME : Marge d'erreur (généralement ¡= 10 %).
- Z : Niveau de confiance.
- P : Probabilité de réponse (souvent estimée à 0,5 dans le pire des cas).
- E : Marge d'erreur.

Chapter 4

Pré-traitement

4.1 Définition

Le but du pré-traitement est de préparer le traitement, en codifiant, convertissant et nettoyant les données.

Début → Saisir ou importer les données
→ Codifier les données
→ Nettoyer les données
→ Traiter les données

4.2 Codifier les données

Donner des noms aux variables (**quantitatives/qualitatives**) :

- **Variables quantitatives** : `NUM`
- **Variables qualitatives** : `CHAR`

4.3 Convertir les données

Transformer les variables de type numériques en variables quantitatives et les variables de type chaîne de caractère en variables qualitatives.

4.4 Nettoyer les données

Traiter les valeurs ($n \geq 30$) manquantes (vide : `NA`) et les valeurs aberrantes (anormales) :

- **Traiter les variables manquantes** :
  - **Action 1** : Recollecter les données.
  - **Action 2** : Remplacer par une valeur calculée par un modèle statistique (Moyenne, Classification, Régression).
  - **Action 3** : Supprimer l'individu (toute la ligne) si l'effectif des valeurs manquantes est inférieur à 5% et si, après suppression, la taille de l'échantillon reste supérieure à 30 ($n' = n - \text{effmanquant}$).
- **Traiter les valeurs aberrantes** :
  - **Identification des valeurs aberrantes** : Si la valeur est $\geq 3\sigma$.

4.5 Langage R

C'est un langage de données dédié à l'analyse et à l'étude statistique. Il est un langage interprété, orienté objet, structuré, fonctionnel, et réfléxif. Il a vu le jour en 1997 par Robert et Ross. La machine virtuelle est **RVM**.

4.5.1 Prétexte de l'instrument de mesure

Il s'agit de vérifier la **fiabilité** et la **validité** du questionnaire, c'est-à-dire vérifier si le questionnaire n'introduit pas d'erreur à travers des questions qui pourraient induire le répondant en erreur. On doit faire la vérification des questions posées au répondant.

- **Fiabilité** : La réponse doit se trouver dans le contexte de la question. La fiabilité signifie que la question doit mesurer ce qui doit être mesuré. Ici, on touche à la formulation de la réponse.
- **Validité** : La question devrait signifier ce qu'elle doit signifier. Ici, on touche au sens. On vérifie la formulation de la question pour qu'elle précise un sens.

Exemple :

- Si on pose une question sur les voitures et on donne dans les réponses l'avion, la **fiabilité** du questionnaire est touchée.
- Si on pose une question sur les moyens de transport et la performance académique, on touche à la **validité**.

4.6 Étude des données avec R

4.6.1 Commentaires en R

En R, les commentaires sont précédés d'un `#`.

4.6.2 Étape 1 : Problématique et définition des données

On souhaite étudier les notes des étudiants du module **probastatistique**.

- **Question de recherche (QR)** : Quelle est la performance des étudiants dans ce module ?
- **Dictionnaire de données** (basé sur le Modèle Conceptuel) :
  - **Étudiant** :
    - **Pays** : qualitative, texte.
    - **Genre** : qualitative, texte.
    - **Filière** : qualitative, texte.
    - **Niveau** : qualitative, texte.
    - **Âge** : quantitative, nombre.
  - **Performance** :
    - **Note** : quantitative, nombre.
- **Hypothèse** : Proposition de réponse possible à la question de recherche.
- **Reformulation de la QR** : Y a-t-il un lien entre les caractéristiques des étudiants et leur performance académique ?

4.6.3 Étape 2 : Collecte des données

- **M1 : Saisie des données** Exemple de saisie :

```
# vecteur contenant une valeur aberrante (40)
Age = c(21, 20, 19, 21, 40)
Genre = c('H', 'F', 'H', 'H', 'F')
Niveau = c('1A', '1A', '1A', '1A', '1A')
Filière = ...
```

- **M2 : Importation d'un fichier** Exemple d'utilisation de bibliothèques :

```
library(readxl)
d2 = read_excel("Chemin_d'accès")
View(d2)
```

4.6.4 Étape 3 : Prétraitement

Codification et conversion

- Voir le fichier R pour la codification.
- Une variable qualitative est à la base un texte qui crée des catégories :
  - **Nominale** : Représente simplement des noms.
  - **Ordinale** : Représente des noms avec un ordre implicite.

  Ces variables sont de type `factor`.

Nettoyage des données

- **Traitement des valeurs aberrantes** :
  - Méthode de l'écart type : Vérifier si la valeur est plus de trois fois l'écart type.
  - Graphiquement via **la boîte à moustache** : En R :

    ```
    Boxplot(d2$age)
    out = Boxplot.stats(d2$age)
    ```

    Les valeurs aberrantes identifiées peuvent être rendues manquantes ou non.
  - Pour les notes, des valeurs comme 7 et 20 peuvent être considérées comme normales.
- **Traitement des valeurs manquantes** :
  - Recollecter les données manquantes.
  - Estimer les valeurs manquantes à l'aide d'un modèle statistique.
  - Supprimer la ligne contenant les valeurs manquantes si :
    - L'effectif des valeurs manquantes est inférieur à 5%.
    - La taille de l'échantillon reste supérieure à 30 après suppression.

4.7 TD1 - Exercice 2 / Partie 2

Problématique à l'issue de ces avis :

- **La problématique :** Après les retours négatifs des clients concernant notre service, et suite à l'analyse de ces avis, nous avons constaté un problème au niveau de la livraison, et les clients expriment leur insatisfaction à cet égard.
- *Pourquoi les clients sont-ils insatisfaits ?*

Les hypothèses possibles pour expliquer la problématique

Parmi les raisons pouvant expliquer l'insatisfaction liée au problème de livraison :

- **La mauvaise logistique** (catégorie 5)
- **La mauvaise communication entre le client et le livreur** (catégorie 5)
- **Le mauvais état du produit à la réception** (catégorie 2)
- **Le manque de suivi par téléphone** (catégorie 4)
- **La mauvaise qualité du produit**
- **Le retard dans le traitement des requêtes** (catégorie 3)
- **Absence de retour malgré relance** (catégorie 1)
- **La livraison en retard** (catégorie 1)

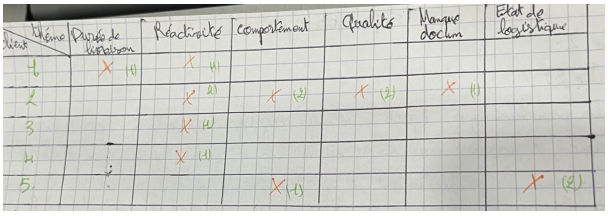
Synthèse du niveau de saturation

Le niveau de saturation est globalement élevé dans plusieurs catégories.

Client	Avis	Thèmes
Client 1	1. Retard de livraison	Durée de livraison
	2. Pas de réponse du service de livraison	La réactivité
Client 2	1. Refus du livreur de laisser le client tester le produit	
	2. Produit reçu défectueux	Qualité
	3. Le client a dû se débrouiller pour retourner le produit	La réactivité
	4. Manque de flexibilité des livreurs	Comportement de livreur
	5. Absence de protocoles pour tester les produits à la livraison	Manque de documentation
	6. Mauvais contrôles qualité des produits avant expédition	Qualité
Client 3	1. Lecteur de traitement de demandes de clients	La réactivité
Client 4	1. Appels téléphoniques non répondus	La réactivité
Client 5	1. Livreurs perçus comme vulgaires	Comportement de livreur
	2. Camions de livraison en mauvais état	État de la logistique
	3. Emballages souvent ouverts ou abîmés	État de la logistique

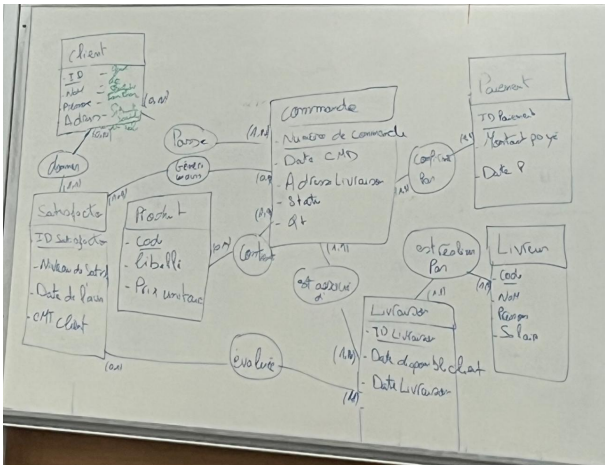
Figure 4.1: Image illustrating the saturation level.

$$\text{niveau de saturation} = \frac{4+3+2+2}{13} = \frac{11}{13} = 84\%$$



Ce qui explique la taille de l'échantillon.

Le modèle conceptuel des données à la lumière des avis des clients



4.7.1 Formulaire

`https://docs.google.com/forms/d/1nC1QsYppFoTSLqRjY963dB7bSnly42SQlkVuyqCWDkE/viewform?edit_requested=true`

- Les questions doivent toujours être ouvertes.
- **Q1 :** Présentez-vous (Afin d'extraire les caractéristiques de l'individu).
- **Q2 :** Que pensez-vous du service de livraison ?
- **Q3 :** Quels sont les problèmes que vous avez rencontrés avec notre service client ?
- **Q4 :** Que suggérez-vous pour améliorer le service de livraison ? (Question d'approfondissement, de relance)

Chapter 5

Traitement

Après un nettoyage des données, nous avons des données valides prêtes pour le traitement selon les étapes ci-dessous.

5.1 Statistiques descriptives univariées (variable par variable)

Objectif : Calculer les indicateurs de la distribution de chaque variable.

Type de variable	**Indicateurs**
Quantitative	- Indicateur de fréquence - Indicateurs de position : min, max, moyenne, médiane, etc. - Indicateurs de dispersion : étendue, écart-type, variance
Qualitative	- Fréquence absolue et relative

5.1.1 Graphiques:

Variable	**Représentation**
Qualitative	Diagramme circulaire, histogramme
Quantitative	Boîte à moustache, diagramme en bâton, nuage de points

En R:

```
summary(d)
plot(d$variable)
hist(d$variable)  # Histogramme
pie(d$variable)   # Diagramme circulaire
```

5.1.2 Statistiques inférentielles

Il s'agit de vérifier des hypothèses à l'issue d'une question de recherche.

Exemple : Tester si la fréquence des femmes par rapport aux hommes (freq(femme)/freq(homme)) a une **valeur significative**.

1. Formuler les hypothèses statistiques à partir de l'hypothèse de recherche.
2. Calculer la valeur significative (p-valeur) du test.
3. Comparer la p-valeur avec le seuil de significativité (α, généralement fixé à 5%).
4. Décider quelle hypothèse retenir en fonction de la p-valeur et α.

Conclusion : Les tests d'hypothèses peuvent porter sur :

- **Comparaison** : Différences entre groupes (exemple : une variable ou plusieurs).
- **Association** : Relations entre deux variables (correspondances).

5.1.3 Formulation des hypothèses :

- **Hypothèse nulle (H_0)** : Formulée de manière négative, représente l'absence d'effet ou de différence.
- **Hypothèse alternative (H_1)** : Formulée de manière positive, représente la présence d'un effet ou d'une différence.

5.2 La règle d'or de la prise de décision par rapport au test d'hypothèse

- H_0 : Il n'y a pas.
- H_1 : Il y a.

5.2.1 Décision:

- **Significatif** : H_0 est rejeté.
- **Non significatif** : H_0 est accepté.

α : Seuil de signification = 5%.

5.2.2 Formule:

$P(H_0) = 1 - P(H_1)$ $H_0 = H_1$

5.2.3 Exemple : Genre

Catégorie	Effectif	Pourcentage (%)
F1	12	56%
F2	13	44%

- **H** : 95%
- α : 5%

Code en R:

```
# Importer le fichier
library(readxl)
d1 <- read_excel("C:/Users/hp/Desktop/statistique/Questionnaire vierge (réponses).xls
View(d1)

# Pre-traitement partie identification
d1$age[13] = 20
d1$age[20] = 19
d1$age[23] = 20
d1$age[59] = 25

if(!is.numeric(d1$age)){
  d1$age = as.numeric(d1$age)
}

if(!is.character(d1$genre)){
  d1$genre = as.character(d1$genre)
}

# Nettoyage des données
boxplot(d1$age)
boxplot.stats(d1$age)
out = boxplot.stats(d1$age)$out
for(i in 1:length(d1$age)){
  for(j in 1:length(out)){
    if(d1$age[i] == out[j] && !is.na(d1$age[i]))
      d1$age[i] <- NA
  }
}

# Traitement des valeurs manquantes
for(i in 1:length(d1$age)){
  if(is.na(d1$age[i])){
    c <- c + 1
  }
}
```

```
probNA = (c / length(d1$age)) + 100
if (probNA >= 5){
  print("estimer")
  for(i in 1:length(d1$age)){
    if(is.na(d1$age[i])){
      d1$age[i] = mean(d1$age, na.rm=TRUE)
    }
  }
}

# Statistique univariée
summary(d1)
plot(d1$age)
hist(d1$genre)
pie(d1$status_sociale)
hist(d1$status_familial)
```

1. Rappel

Test d'hypothèse	
Association (X, Y)	
H_0:	Il n'y a pas d'association significative entre X et Y.
H_1:	Il y a une association significative entre X et Y.
Comparaison (X, Y) $(\alpha = 5\%)$	
H_0:	Il n'y a pas une différence significative entre X et Y.
H_1:	Il y a une différence significative entre X et Y.

Table 5.1: Rappel des hypothèses

5.3 Test d'Hypothèse de Fréquence

5.3.1 Matrice Théorique

	R	N
H	$\frac{1}{4}$	$\frac{1}{4}$
F	$\frac{1}{4}$	$\frac{1}{4}$

5.3.2 Matrice Observée

	R	N
H	f_1	f_2
F	f_3	f_4

5.3.3 Formule du Khi-Deux

$$\chi^2 = \sum \frac{(f_0 - f_i)^2}{f_T}$$

Lorsque $f_i \sim f_0$, $\chi^2 \to 0$.

Remarques sur le Khi-deux :

- **Définition :** Le Khi-deux est un test qui mesure l'écart entre les valeurs théoriques et observées.
- **Utilisation :** Il est utilisé pour les variables qualitatives.
- **Conclusion :** Quand les valeurs observées $\sim$ théoriques, cela signifie une indépendance des variables.

5.3.4 Exemple de Répartition

	1	2	3	4	5
item 1	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$
item 2	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

5.4 En R : Le Test d'Hypothèse de Khi-Deux

Voici un exemple de code R pour effectuer le test d'hypothèse de Khi-deux :

```
# Création de la table de contingence
table_data <- table(data$itemX)

# Test du Khi-deux
result <- chisq.test(table_data)

# Vérification de la p-valeur
if (result$p.value > 0.05) {
  cat("H0 est rejetée. Pas d'association significative.")
} else {
  cat("H0 est acceptée. Association significative.")
}
```

Interprétation des résultats :

- **p-valeur $> 5\%$:** Hypothèse H_0 est rejetée. Cela signifie qu'il n'y a pas de différence significative.
- **p-valeur $\leq 5\%$:** Hypothèse H_0 est acceptée. Cela signifie qu'il y a une différence significative.

- Le test de fréquence de Khi-deux est utilisé pour les variables qualitatives.
- Lorsque la proportion significative est $> 5\%$, on utilise le test exact de Fischer.
- Un exemple de cas où le Khi-deux s'applique : "H : mauvaise logistique lors de livraison."

5.5 Exercice 6 - TD 1

Mehdi Bahi

1 - Formulation de la problématique

L'inégalité des salaires entre les hommes et les femmes ayant des niveaux de scolarité et des fonctions égaux.

Variable	**Quali. / Quanti.**	**Type**	**Échelle**	**Codage**
Sexe	Quali.	Nominal	Nominale	Homme = 1 / Femme = 2
Âge	Quanti.	Numérique	Continu	
Scolarité (années d'étude)	Quanti.	Numérique	Continu	
Fonction	Quali.	Nominal	Ordinale	Responsable = 1 / Ing. = 2 / Tech. = 3
Revenu	Quanti.	Numérique	Continu	
Q1 (Oui/Non)	Quali.	Nominal	Nominale	Oui = 1 / Non = 0

Table 5.2: Description des variables

2 - Question de recherche

Est-ce qu'il existe une différence significative entre les salaires des hommes et des femmes ayant des niveaux de scolarité et des fonctions égaux ?

Hypothèse de recherche

Il existe une différence significative entre le revenu moyen des hommes et des femmes ayant des niveaux de scolarité et des fonctions équivalents.

Si c'est accepté : → Le revenu des hommes est supérieur à celui des femmes à scolarité et fonctions égales.

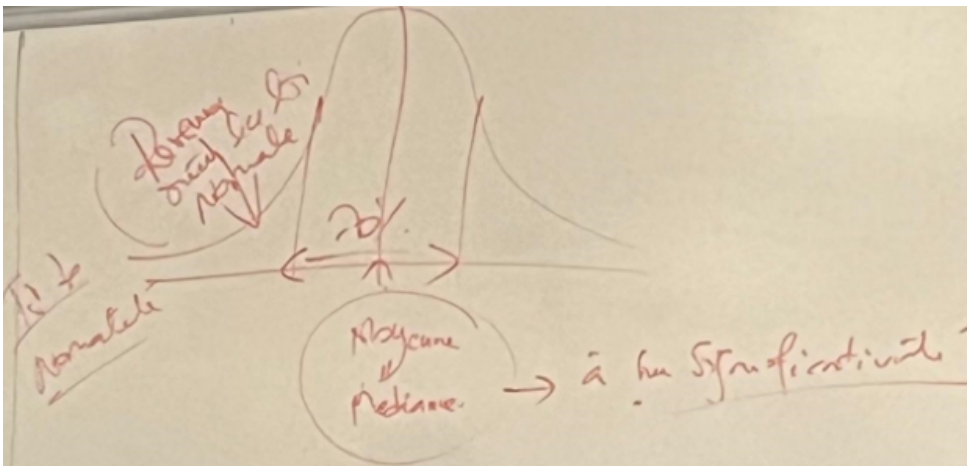


Figure 5.1: Illustration relative à l'hypothèse

5.6 Test de normalité

Définition : C'est un test de comparaison entre la loi théorique normale et la distribution de la variable.

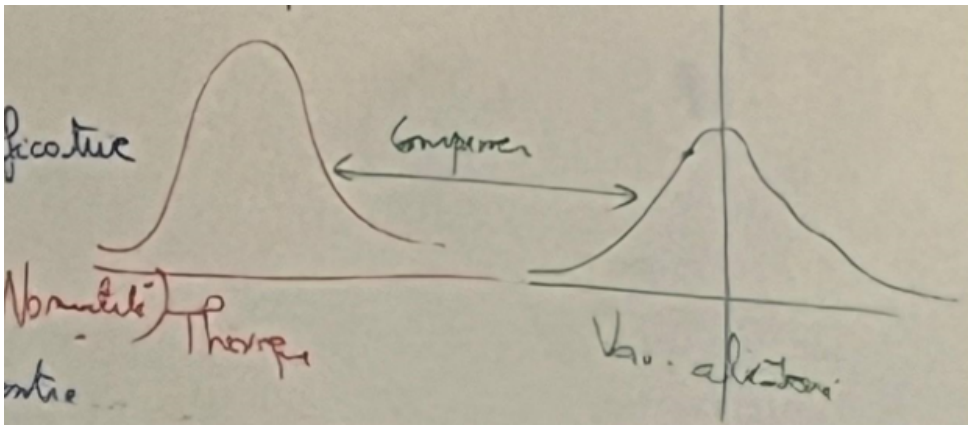


Figure 5.2: Exemple de distribution normale

Hypothèses du test :

- **H_0 :** Il n'y a pas de différence significative entre la loi normale et la distribution observée (normalité).
- **H_1 :** Il y a une différence significative entre la loi normale et la distribution observée.

Remarque : Le test de normalité concerne en priorité les **variables quantitatives**. Pour une variable qualitative, il est possible d'effectuer ce test à condition qu'il y ait **plus de 5 modalités**.

5.6.1 Récapitulatif de la méthodologie

1. Vérifier l'égalité du nombre d'hommes et de femmes.
2. Vérifier les égalités des fréquences des hommes et des femmes par fonction.
3. Vérifier la normalité de la variable *scolarité*.
4. Vérifier la moyenne des années de scolarité entre les hommes et les femmes (elles doivent être égales).

5.7 Tests d'Hypothèse

Rappel

	Quali	**Quanti**
Quali	Correspondance	Comparaison
Quanti	Comparaison	Corrélation

Type de Test	**Description**
Test Paramétrique	Utilisé lorsque la variable suit la loi normale au sens strict.
Test Non Paramétrique	Utilisé lorsque la variable ne suit pas la loi normale ou aucune autre distribution précisée.

Table 5.3: Description des Types de Tests

5.7.1 Exemples d'Items

Test	**Items**
Le Khi-deux	Test du Khi-2
Comparaison de Moyenne / Test de Student (*t.test*)	Test du rang
Comparaison de Variance/Écart-type	Test de signe

Table 5.4: Tests et Items

5.7.2 Test de Comparaison de Moyenne (*t.test*)

Comparer la moyenne entre deux groupes M_1 et M_2 par rapport à la variance (la distance), est appliquée lorsque nous avons deux groupes uniquement.

$$t = \frac{(M_1 - M_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Si $t \to 0$, H_0 est acceptée.
- Sinon, H_1 est rejetée.