Statistiques Descriptives Bivariées

1 Introduction

La statistique descriptive regroupe un ensemble de méthodes permettant de résumer, organiser et présenter des données sous forme de tableaux, graphiques et indicateurs numériques. Elle est essentielle pour comprendre les tendances générales d'un ensemble de données avant de passer à des analyses plus avancées.

Dans nos analyses, nous utiliserons la base de données relative aux étudiants du Master 1 ENSPD pour l'année 2024-2025. Notre base de données montre que le genre masculin est majoritaire, représentant 80% des individus, tandis que le genre féminin représente 20%. En outre, la majorité des étudiants ont opté pour l'option Planification et Suivi Évaluation (PSE), soit 65%, alors que 35% ont choisi la Statistique Appliquée (SA).

2 Analyse de la liaison entre deux variables qualitatives

Ici, nous allons utiliser la statistique descriptive pour mettre en évidence les relations entre deux variables qualitatives. Nous décrirons également le processus de mise en œuvre de cette analyse, en utilisant les outils statistiques appropriés pour garantir une approche méthodique et efficace.

2.1 Représentations graphiques usuelles (Présenté par AKPAKI Kamila)

Pour les variables qualitatives, nous disposons principalement de deux types de visualisation graphique : le diagramme en bâton simple et le diagramme en bâton empilé.

2.1.1 Diagramme en bâton simple

FIGURE 1 – Répartition conjointe des étudiants du Master 1 de l'ENSPD 2024-2025 selon le genre et la filière

Source : Données des étudiants du Master 1 de l'ENSPD 2024-2025.

2.1.2 Diagramme en bâton empilé

FIGURE 2 – Répartition conjointe des étudiants du Master 1 de l'ENSPD 2024-2025 selon le genre et la filière

Source : Données des étudiants du Master 1 de l'ENSPD 2024-2025.

3 2.1.2 Tableau de distribution statistique (Présenter par DEGBEVI John)

Un tableau de distribution statistique est un outil permettant d'organiser et de présenter un ensemble de données en regroupant les valeurs observées dans des classes ou des catégories, accompagnées de leurs effectifs et fréquences.

3.1 a. Tableau de contingence

Un tableau de contingence est un tableau statistique encore appelé tableau croisé ou tableau à double entrée qui permet d'analyser la relation entre deux variables catégorielles. Il est souvent utilisé en analyse de données pour observer la distribution conjointe de ces variables.

Exemple d'application sur notre base de données

TABLE 1 – Répartition conjointe des étudiants du Master 1 de l'ENSPD 2024-2025 selon le genre et la filière

Genre\Filières	**PSE**	**SA**	**Total Général**
F	7	6	13
M	35	17	52
Total général	42	23	65

Source : Données des étudiants du Master 1 de l'ENSPD 2024-2025

Dans le tableau de contingence, les lignes représentent les modalités de la première variable, ici le sexe, et les colonnes représentent les modalités de la seconde variable, qui est généralement la variable dépendante (ici la filière). Les cellules contiennent le nombre d'occurrences, fréquence ou répétition correspondant à l'intersection de deux modalités.

3.2 b. Distributions marginales et conditionnelles (Expliquer par DOSSI Jean)

En statistique, les distributions marginales et distributions conditionnelles sont des outils pour analyser les relations entre les variables dans un tableau de contingence.

Distribution marginale

La distribution marginale correspond aux totaux des lignes ou des colonnes du tableau de contingence. Elle permet d'observer la répartition d'une seule variable sans tenir compte de l'autre.

Exemple d'application sur notre base de données

Source : Données des étudiants du Master 1 de l'ENSPD 2024-2025

Interprétation : La répartition des étudiants indique que 42 sont inscrits en Planification et Suivi Évaluation (PSE) et 23 en Statistique Appliquée (SA). En outre, parmi ces étudiants, 13 sont des femmes et 52 sont des hommes.

Il faut noter que la distribution marginale ne permet pas de détecter une liaison entre les variables.

TABLE 2 – Distribution conjointe des étudiants du Master 1 de l'ENSPD 2024-2025 selon le Genre et la Filière

Genre\Filières	**PSE**	**SA**	**Total Général**
F	7	6	13
M	35	17	52
Total général	42	23	65

3.3 Distribution conditionnelle

La distribution conditionnelle mesure la répartition d'une variable en fonction d'une autre, sous forme de proportions. On peut calculer la distribution conditionnelle de la filière en fonction du genre, c'est-à-dire la proportion de chaque filière au sein d'un même genre.

La différence principale réside dans le fait que la distribution marginale concerne le total des occurrences d'une variable sans considérer l'autre, tandis que la distribution conditionnelle concerne la proportion d'une modalité dans une catégorie donnée.

En réalisant les tableaux des profils colonnes et des profils lignes, nous pourrons mieux visualiser les distributions conditionnelles et comprendre comment chaque catégorie influence les autres.

3.3.1 Profils-lignes et Profils colonnes

Le terme profils-lignes fait référence à l'analyse des données le long des lignes d'un tableau. Le terme profils-colonnes fait référence à l'analyse des données le long des colonnes d'un tableau.

Dans un tableau de contingence, on distingue généralement deux types de profils :

Profil ligne Il correspond à la distribution des modalités d'une ligne donnée par rapport au total de cette ligne. Chaque effectif de la ligne est divisé par le total de la ligne. Cela permet d'analyser la répartition des observations d'une catégorie de la variable en ligne parmi les différentes modalités de la variable en colonne. La somme des proportions d'une ligne est toujours égale à 1.

Formule :

$$P_{ij} = \frac{n_{ij}}{n_{i.}}$$

Où : n_{ij} est l'effectif de la case (i, j) et $n_{i.}$ est le total de la ligne i.

Exemple d'application sur notre base de données

TABLE 3 – Profil ligne des étudiants du Master 1 de l'ENSPD 2024-2025 selon le Genre et la Filière

Genre\Filières	**PSE**	**SA**	**Total général**
F	54%	46%	100%
M	67%	33%	100%

Source : Données des étudiants du Master 1 de l'ENSPD 2024-2025

Interprétation : 54% des femmes sont inscrites en Planification et Suivi Évaluation (PSE), alors que 67% des hommes ont choisi cette filière. Inversement, 33% des hommes ont opté pour la Statistique Appliquée (SA), contre 46% des femmes.

Profil colonne Il correspond à la distribution des modalités d'une colonne donnée par rapport au total de cette colonne. Chaque effectif de la colonne est divisé par le total de la colonne. Cela permet d'étudier comment se répartissent les observations d'une catégorie de la variable en colonne parmi les différentes modalités de la variable en ligne. La somme des proportions d'une colonne est toujours égale à 1.

Formule :

$$P_{ij} = \frac{n_{ij}}{n_{.j}}$$

Où : n_{ij} est l'effectif de la case (i, j) et $n_{.j}$ est le total de la colonne j.

Exemple d'application sur notre base de données

TABLE 4 – Profil colonne des étudiants du Master 1 de l'ENSPD 2024-2025 selon le Genre et la Filière

Genre\Filières	PSE	SA
F	17%	26%
M	83%	74%
Total général	100%	100%

Source : Données des étudiants du Master 1 de l'ENSPD 2024-2025

Interprétation : Dans la filière Planification et Suivi Évaluation (PSE), 83% des individus sont des hommes, tandis que 17% sont des femmes. En Statistique Appliquée (SA), 74% des individus sont des hommes, et 26% sont des femmes.

Ces profils sont particulièrement utiles dans l'analyse des correspondances et l'interprétation des relations entre les modalités des variables qualitatives.

4 2.1.3. Mesure de liaison (Présenter par MAMA Moukadas)

4.1 Caractérisation de la situation d'indépendance

La caractérisation de l'indépendance entre deux variables qualitatives repose sur l'analyse de leur distribution conjointe et la vérification de l'absence de relation entre elles.

Deux variables qualitatives A et B sont indépendantes si la connaissance de la modalité d'une variable n'apporte aucune information sur la modalité de l'autre variable. Mathématiquement, cela se traduit par :

$$P(A = a_i \text{ et } B = b_j) = P(A = a_i) \times P(B = b_j)$$

4.2 Test statistique

4.2.1 b.1 Test d'independance de Khi-deux (X2)

Le Khi-deux (X2) est une statistique qui mesure l'écart entre les effectifs observes dans un tableau de contingence et les effectifs théoriques attendus sous l'hypothèse d'indépendance entre les deux variables qualitatives.

Conditions d'application

- Les données doivent être organisées dans un tableau de contingence.
- Les effectifs théoriques e_{ij} doivent être suffisamment grands, au moins 80% des effectifs théoriques e_{ij} doivent être ≥ 5.
- Les effectifs observés doivent être supérieurs ou égaux à 5.

Si ces conditions ne sont pas respectées, le test de Fisher est préférable.

Étapes du test d'indépendance de Khi-2 Le test d'indépendance de Khi-2 se fait en six (07) étapes définies ci-dessous :

1. **Définir les hypothèses :**
   - Hypothèse nulle H_0 : Il n'y a pas d'association entre les deux variables (elles sont indépendantes).
   - Hypothèse alternative H_1 : Il y a une association entre les deux variables (elles ne sont pas indépendantes).
2. **Construire le tableau de contingence :** Il s'agit d'un tableau croisant les modalités des deux variables. Chaque case contient les fréquences observées.
3. **Calculer les fréquences théoriques :** Les fréquences théoriques sont calculées selon la formule ci-dessous :
$$e_{ij} = \frac{n_{(i.)} \times n_{(.j)}}{n}$$
Où :
   - $n_{(i.)}$ est le total de la ligne i,
   - $n_{(.j)}$ est le total de la colonne j,
   - n est l'effectif total.

   Les valeurs théoriques calculées désignent les effectifs attendus en considérant que les deux variables sont indépendantes.