

Analyse de l'information chiffrée : Statistiques bivariées

Exomaths.fr

Juin 2025

Contents

1	Introduction aux statistiques bivariées	2
2	Tableaux croisés d'effectifs	2
2.1	Exemple commenté : Genre et choix d'option	3
3	Représentations graphiques de deux caractères	3
3.1	Exemple commenté : Nuage de points	4
3.2	Exemple commenté : Diagramme en barres groupées	5
4	Filtrage et sous-ensembles de données	5
4.1	Exemple commenté : Filtrage de données d'une enquête	6
5	Fréquences conditionnelles et marginales	7
5.1	Exemple commenté : Analyse du choix d'option par genre (suite)	8

1 Introduction aux statistiques bivariées

L'analyse de l'information chiffrée est essentielle pour comprendre les problématiques contemporaines telles que le développement durable, le changement climatique, la démographie ou la santé publique. Elle permet de développer un sens critique face aux débats actuels. En prolongement du programme de seconde, qui a introduit des indicateurs pour l'analyse d'un seul caractère statistique, cette section aborde l'analyse statistique bivariable. Il s'agit d'une première sensibilisation aux bases de données.

Point de méthode

Les données étudiées peuvent provenir d'autres enseignements dispensés au lycée (enseignement scientifique, enseignement moral et civique, enseignement de spécialité). Les outils informatiques actuels permettent de stocker et manipuler des données massives. Des sites institutionnels comme l'INSEE, l'INED ou data.gouv, ainsi que des rapports publics (par exemple ceux du GIEC), sont de bonnes sources de données.

L'analyse bivariable a pour objectif d'étudier les relations éventuelles entre deux caractères. Cela s'illustre par l'utilisation de représentations de données réelles, que l'on peut obtenir en traitant des fichiers de taille raisonnable avec un tableur.

2 Tableaux croisés d'effectifs

Lorsqu'on étudie deux caractères sur une même population, on peut organiser les données dans un tableau croisé d'effectifs. Ce tableau permet de visualiser la répartition des individus selon les modalités de chacun des deux caractères.

Définition

Un **tableau croisé d'effectifs** (ou tableau de contingence) est un tableau qui présente la distribution conjointe de deux variables catégorielles. Les lignes représentent les modalités d'une variable et les colonnes représentent les modalités de l'autre variable. Chaque cellule du tableau contient l'effectif (le nombre d'individus) correspondant à l'intersection des deux modalités.

Point de méthode

Dresser un tableau croisé de deux caractères à partir d'un fichier de données

1. Identifiez les deux caractères que vous souhaitez analyser.
2. Listez toutes les modalités (valeurs possibles) de chaque caractère.
3. Créez un tableau avec les modalités du premier caractère en lignes et celles du second caractère en colonnes.
4. Parcourez le fichier de données et pour chaque individu, incrémentez le compteur de la cellule correspondant à ses deux modalités.
5. Calculez les totaux pour chaque ligne et chaque colonne (effectifs marginaux), ainsi que l'effectif total.

2.1 Exemple commenté : Genre et choix d'option

Considérons un groupe de 50 élèves de première générale et leur choix entre l'option A et l'option B. Nous souhaitons analyser la relation entre le genre des élèves (Fille/Garçon) et leur choix d'option.

Voici les données brutes :

- 15 Filles ont choisi l'Option A
- 10 Filles ont choisi l'Option B
- 12 Garçons ont choisi l'Option A
- 13 Garçons ont choisi l'Option B

Étape 1 : Identifier les caractères et modalités. Caractère 1 : Genre (modalités : Fille, Garçon) Caractère 2 : Option choisie (modalités : Option A, Option B)

Étape 2 : Créer le tableau. Nous allons créer un tableau avec le genre en lignes et l'option en colonnes.

Genre \ Option	Option A	Option B	Total
Fille			
Garçon			
Total			

Étape 3 : Remplir les cellules avec les effectifs.

Genre \ Option	Option A	Option B	Total
Fille	15	10	
Garçon	12	13	
Total			

Étape 4 : Calculer les totaux (effectifs marginaux). Pour les lignes : Total Fille = $15 + 10 = 25$ Total Garçon = $12 + 13 = 25$

Pour les colonnes : Total Option A = $15 + 12 = 27$ Total Option B = $10 + 13 = 23$

Effectif total = $25 + 25 = 50$ (ou $27 + 23 = 50$)

Tableau croisé d'effectifs final :

Genre \ Option	Option A	Option B	Total
Fille	15	10	25
Garçon	12	13	25
Total	27	23	50

3 Représentations graphiques de deux caractères

La visualisation des données est cruciale pour l'analyse bivariable. Un tableur est un outil puissant pour représenter ces données.

Point de méthode

Utiliser un tableur pour représenter des données sous forme de tableau ou de diagramme

1. Saisissez vos données brutes ou votre tableau croisé d'effectifs dans le tableur.
2. Sélectionnez les données que vous souhaitez représenter.
3. Utilisez la fonction "Insertion de graphique" de votre tableur.
4. Choisissez le type de diagramme adapté à votre analyse :
 - **Nuage de points** pour deux caractères quantitatifs (ex: température et niveau des océans).
 - **Diagrammes en barres groupées ou empilées** pour un caractère quantitatif et un qualitatif, ou pour deux caractères qualitatifs (ex: genre et option, revenus par catégorie d'âge).
 - **Diagrammes circulaires** (ou camemberts) pour représenter des proportions au sein d'une modalité (moins courant pour l'analyse bivariable directe, mais peut être utilisé pour les fréquences conditionnelles).
5. Personnalisez le graphique (titres, légendes, axes) pour une meilleure lisibilité.

3.1 Exemple commenté : Nuage de points

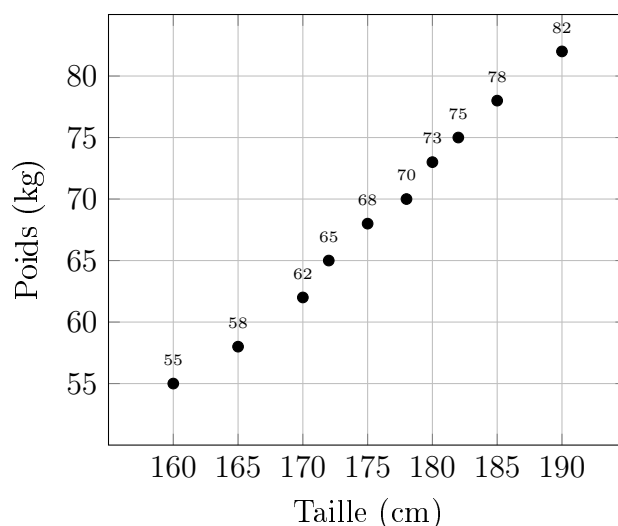
Un nuage de points est particulièrement utile pour visualiser la relation entre deux variables numériques.

Situation : On étudie la taille (en cm) et le poids (en kg) de 10 personnes.

Taille (cm)	160	165	170	172	175	178	180	182	185	190
Poids (kg)	55	58	62	65	68	70	73	75	78	82

Représentation graphique (Nuage de points) : Pour tracer le nuage de points, chaque couple (taille, poids) correspond à un point dans un repère. La taille est généralement placée sur l'axe des abscisses et le poids sur l'axe des ordonnées.

Nuage de points: Poids en fonction de la Taille



Point de méthode

Ce nuage de points suggère une relation linéaire positive entre la taille et le poids : plus la taille augmente, plus le poids a tendance à augmenter. L'analyse bivariée peut ensuite chercher à quantifier cette relation (par exemple, par le calcul d'un coefficient de corrélation ou l'ajustement d'une droite de régression).

3.2 Exemple commenté : Diagramme en barres groupées

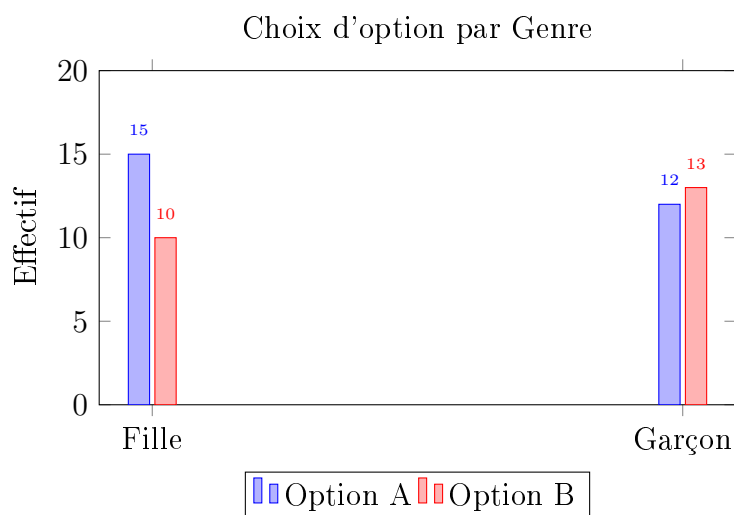
Un diagramme en barres groupées est utile pour comparer la distribution d'un caractère qualitatif selon les modalités d'un autre caractère qualitatif.

Reprenons l'exemple du choix d'option par genre.

Tableau croisé d'effectifs :

Genre \ Option	Option A	Option B	Total
Fille	15	10	25
Garçon	12	13	25
Total	27	23	50

Représentation graphique (Diagramme en barres groupées) : Pour chaque catégorie de genre, nous aurons deux barres (une pour l'Option A, une pour l'Option B).



Point de méthode

Ce diagramme permet de voir rapidement que, parmi les filles, l'option A est plus choisie que l'option B, tandis que chez les garçons, les choix sont plus équilibrés, avec une légère préférence pour l'option B.

4 Filtrage et sous-ensembles de données

L'analyse de données massives implique souvent de travailler sur des sous-ensembles d'individus répondant à des critères spécifiques. Les tableurs offrent des fonctionnalités de filtrage très puissantes pour cela.

Point de méthode

Détermination dans un fichier de données d'un sous-ensemble d'individus répondant à un sous-caractère (filtre, utilisation des ET, OU, NON)

1. Ouvrez votre fichier de données dans un tableur (Excel, LibreOffice Calc, Google Sheets, etc.).
2. Sélectionnez la plage de données (y compris les en-têtes de colonnes).
3. Activez la fonction "Filtrer" (souvent accessible via le menu "Données" ou une icône en forme d'entonnoir). Des flèches apparaîtront à côté de chaque en-tête de colonne.
4. Cliquez sur la flèche de la colonne correspondant au critère souhaité.
5. Cochez les modalités que vous voulez inclure (filtre simple).
6. Pour des critères plus complexes, utilisez les options de "Filtres textuels" ou "Filtres numériques" (selon le type de données) pour spécifier des conditions (égal à, contient, supérieur à, etc.).
7. Pour combiner plusieurs conditions :
 - **ET** : Appliquez des filtres successifs sur différentes colonnes. Les individus doivent satisfaire toutes les conditions.
 - **OU** : Utilisez la fonction de filtre avancé ou des formules logiques (selon le tableur). Par exemple, dans Excel, une colonne supplémentaire avec une formule 'OU(condition1;condition2)' peut être filtrée.
 - **NON** : Utilisez l'option "ne contient pas", "n'est pas égal à", ou combinez avec un "NON" logique si vous utilisez des formules.
8. Les lignes ne correspondant pas aux critères seront masquées, affichant ainsi le sous-ensemble d'individus.

4.1 Exemple commenté : Filtrage de données d'une enquête

Imaginons une base de données d'une enquête avec les colonnes suivantes : "Âge", "Genre", "Catégorie Socio-Professionnelle (CSP)", "Ville".

Scénario 1 : Filtrer les femmes de moins de 30 ans.

- Appliquer un filtre sur la colonne "Genre" et sélectionner "Femme".
- Appliquer un filtre sur la colonne "Âge" et utiliser "Filtres numériques > Inférieur ou égal à" et entrer "29".

Le tableur affichera uniquement les lignes correspondant aux femmes de 29 ans ou moins. Ceci est un exemple d'utilisation de l'opérateur logique **ET**.

Scénario 2 : Filtrer les personnes résidant à Paris ou Lyon.

- Appliquer un filtre sur la colonne "Ville" et cocher "Paris" et "Lyon".

Le tableur affichera les lignes des personnes vivant à Paris ou à Lyon. Ceci est un exemple d'utilisation de l'opérateur logique **OU**.

Scénario 3 : Filtrer les personnes qui ne sont pas de la CSP "Cadre".

- Appliquer un filtre sur la colonne "CSP" et décocher "Cadre".

Le tableur affichera toutes les personnes dont la CSP n'est pas "Cadre". Ceci est un exemple d'utilisation de l'opérateur logique **NON**.

Point de méthode

La maîtrise des fonctions de filtrage est fondamentale pour l'exploration et l'analyse exploratoire de grandes bases de données, permettant de se concentrer sur des segments spécifiques de la population étudiée.

5 Fréquences conditionnelles et marginales

L'analyse d'un tableau croisé va au-delà des simples effectifs. Le calcul des fréquences (relatives) permet de comparer des proportions, ce qui est souvent plus pertinent que les effectifs bruts, surtout lorsque les populations sont de tailles différentes. L'analyse statistique bivariable permet d'introduire naturellement la notion de fréquence conditionnelle.

Définition

À partir d'un tableau croisé d'effectifs :

- Une **fréquence marginale** est la fréquence d'une modalité d'un caractère, sans tenir compte de l'autre caractère. Elle est calculée en divisant l'effectif marginal de cette modalité par l'effectif total de la population.
- Une **fréquence conditionnelle** est la fréquence d'une modalité d'un caractère *sachant* (ou *étant donné*) une modalité de l'autre caractère. Elle est calculée en divisant l'effectif de la cellule concernée par l'effectif marginal de la modalité "conditionnante".

Point de méthode

Calculer des fréquences conditionnelles et des fréquences marginales à partir d'un tableau croisé d'effectifs

1. **Fréquences marginales** : Pour une modalité M_1 du caractère 1 (en ligne) : $F(M_1) = \frac{\text{Effectif total de la ligne } M_1}{\text{Effectif total général}}$ Pour une modalité M_2 du caractère 2 (en colonne) : $F(M_2) = \frac{\text{Effectif total de la colonne } M_2}{\text{Effectif total général}}$
2. **Fréquences conditionnelles** : Fréquence de M_1 sachant M_2 : $F(M_1|M_2) = \frac{\text{Effectif de la cellule } (M_1, M_2)}{\text{Effectif total de la colonne } M_2}$ Fréquence de M_2 sachant M_1 : $F(M_2|M_1) = \frac{\text{Effectif de la cellule } (M_1, M_2)}{\text{Effectif total de la ligne } M_1}$

Point de méthode

Interpréter un tableau croisé en utilisant des fréquences conditionnelles

1. Calculez les fréquences conditionnelles pertinentes pour la question posée.
2. Comparez ces fréquences. Si les fréquences conditionnelles sont significativement différentes des fréquences marginales, ou différentes entre elles selon la condition, cela suggère une relation entre les deux caractères.
3. Exprimez vos conclusions en langage clair, en évitant les conclusions hâtives de causalité si l'étude n'est que descriptive.

5.1 Exemple commenté : Analyse du choix d'option par genre (suite)

Reprenons notre tableau croisé d'effectifs :

Genre \ Option	Option A	Option B	Total
Fille	15	10	25
Garçon	12	13	25
Total	27	23	50

Effectif total = 50.

Calcul des fréquences marginales :

- Fréquence des Filles : $F(\text{Fille}) = \frac{25}{50} = 0,5 = 50\%$
- Fréquence des Garçons : $F(\text{Garçon}) = \frac{25}{50} = 0,5 = 50\%$
- Fréquence de l'Option A : $F(\text{Option A}) = \frac{27}{50} = 0,54 = 54\%$
- Fréquence de l'Option B : $F(\text{Option B}) = \frac{23}{50} = 0,46 = 46\%$

Calcul et interprétation des fréquences conditionnelles :

1. Fréquence du choix d'option sachant le genre (profils des genres) :

• Parmi les Filles (Total Filles = 25) :

- Fréquence de l'Option A sachant Fille : $F(\text{Option A} \mid \text{Fille}) = \frac{15}{25} = 0,6 = 60\%$
- Fréquence de l'Option B sachant Fille : $F(\text{Option B} \mid \text{Fille}) = \frac{10}{25} = 0,4 = 40\%$

• Parmi les Garçons (Total Garçons = 25) :

- Fréquence de l'Option A sachant Garçon : $F(\text{Option A} \mid \text{Garçon}) = \frac{12}{25} = 0,48 = 48\%$
- Fréquence de l'Option B sachant Garçon : $F(\text{Option B} \mid \text{Garçon}) = \frac{13}{25} = 0,52 = 52\%$

Interprétation (Profils des genres) :

- Il y a une tendance plus marquée pour les filles à choisir l'Option A (60% des filles ont choisi l'Option A, contre 48% des garçons).
- Inversement, les garçons sont légèrement plus nombreux à choisir l'Option B (52% des garçons ont choisi l'Option B, contre 40% des filles).

Ceci indique que le choix de l'option semble dépendre du genre de l'élève.

2. Fréquence du genre sachant le choix d'option (profils des options) :

- **Parmi les élèves ayant choisi l'Option A (Total Option A = 27) :**

- Fréquence de Fille sachant Option A : $F(\text{Fille} \mid \text{Option A}) = \frac{15}{27} \approx 0,556 = 55,6\%$
- Fréquence de Garçon sachant Option A : $F(\text{Garçon} \mid \text{Option A}) = \frac{12}{27} \approx 0,444 = 44,4\%$

- **Parmi les élèves ayant choisi l'Option B (Total Option B = 23) :**

- Fréquence de Fille sachant Option B : $F(\text{Fille} \mid \text{Option B}) = \frac{10}{23} \approx 0,435 = 43,5\%$
- Fréquence de Garçon sachant Option B : $F(\text{Garçon} \mid \text{Option B}) = \frac{13}{23} \approx 0,565 = 56,5\%$

Interprétation (Profils des options) :

- Parmi les élèves ayant choisi l'Option A, il y a un peu plus de filles que de garçons (55,6% contre 44,4%).
- Parmi les élèves ayant choisi l'Option B, il y a plus de garçons que de filles (56,5% contre 43,5%).

Ces deux types de fréquences conditionnelles (profils en ligne et profils en colonne) sont complémentaires et apportent un éclairage différent sur la relation entre les deux caractères.