

Université de Parakou

~~~~~

**École Nationale de Statistique, de Planification et de  
Démographie (ENSPD)**

---

**ECUE : Statistiques Descriptives**

**Première année de Master**

**Enseignant : Mouftaou AMADOU SANNI, Démographe Ph.D.,  
Professeur Titulaire**

**Assistant : Olaïtan Elihou ADJE, Analyste Statistique,  
Doctorant en Démographie**

**Année académique : 2024-2025**

## TABLE DES MATIERES

|                                                                                                 |    |
|-------------------------------------------------------------------------------------------------|----|
| I – Fondements et concepts en Statistique .....                                                 | 3  |
| 1.1- Fondement de la Statistique .....                                                          | 3  |
| 1.2- Concepts de base en statistique.....                                                       | 4  |
| II – Observation Statistique.....                                                               | 5  |
| 2.1- Intérêt de l'observation statistique .....                                                 | 6  |
| 2.2- Types d'observations .....                                                                 | 6  |
| 2.2.1- Observation instantanée .....                                                            | 6  |
| 2.2.2- Observation rétrospective.....                                                           | 7  |
| 2.2.3- Observation continue .....                                                               | 7  |
| 2.3- Outil d'observation collecte en statistique : le questionnaire.....                        | 8  |
| 2.4- La Base de données .....                                                                   | 8  |
| III – Synthèse des données Statistiques .....                                                   | 8  |
| 3.1- Tableau statistique d'une variable statistique .....                                       | 9  |
| IV – Analyses Statistiques Descriptives – Travaux Personnels de groupes d'Étudiants (TPE) ..... | 10 |
| 4.1- Statistiques descriptives univariés.....                                                   | 10 |
| 4.1.1 – Analyse d'une variable qualitative .....                                                | 10 |
| <i>Les diagrammes en bande .....</i>                                                            | 10 |
| <i>Les diagrammes circulaires et semi-circulaires.....</i>                                      | 11 |
| 4.1.2 – Analyse d'une variable quantitative.....                                                | 11 |
| 4.1.2.1 Représentations graphiques usuelles .....                                               | 11 |
| 4.1.2.2 Caractéristiques de tendance centrale .....                                             | 14 |
| 4.1.2.2 Caractéristiques de dispersion .....                                                    | 16 |
| 4.2- Statistiques descriptives bivariés.....                                                    | 18 |
| 4.2.1 Cas de deux variables qualitatives .....                                                  | 18 |
| 4.2.2 Cas de deux variables quantitatives .....                                                 | 23 |
| 4.2.3 Cas d'une variable qualitative et d'une variable quantitative.....                        | 25 |

## I – Fondements et concepts en Statistique

### 1.1- Fondement de la Statistique

On fait de la statistique parce qu'il y a un enjeu. Faire de la statistique suppose qu'il y a un enjeu c'est-à-dire qu'on a un résultat à obtenir. Le résultat est lié à un contexte de vie. **Quel est ce contexte ?**

C'est un contexte dans lequel il y a des objets, des individus, et d'autres espèces vivantes. C'est donc un contexte dans lequel il y a des vivants et des non vivants. Dire que la statistique a un enjeu suppose que la statistique à un résultat à obtenir à l'issue d'une expérience. Tout résultat attendu découle d'un intérêt et tout intérêt porte sur un sujet. En résumé donc faire de la statistique suppose :

- Un contexte
- Un / des sujets
- Un intérêt sur un sujet
- Un enjeu
- Des défis à relever

En statistique, le contexte correspond à un espace et à un moment précis.

Les sujets sont soit des êtres vivants sur cet espace (hommes, animaux, arbres, ...) ou des non vivants (tables, ordinateurs, voitures, ...).

L'intérêt c'est de dénombrer tout particulièrement un des sujets présents sur cet espace.

L'enjeu de la statistique est la connaissance du nombre total de sujet étudié sur l'espace considéré. Les défis sont les obstacles à surmonter pour l'obtention de l'enjeu de la statistique.

#### Exemple :

**Contexte :** ENSPD, en 2022-2023

**Espace :** ENSPD

**Temps :** Année académique 2022-2023.

**Les sujets :** Etudiants, Enseignants, Personnel administratif, Technicien de surface, matériel de bureau, disciplines enseignées, ...

**L'intérêt :** Connaitre selon le cas le nombre de chacun. Le choix de dénombrer un des sujets du contexte.

**L'enjeu :** La connaissance du nombre total de sujet qui m'intéresse dans cet espace au moment considéré.

**Défis :** Il s'agit des obstacles identifiés qu'il convient de surmonter pour l'atteinte de l'enjeu statistique.

**Remarque :** La statistique est une science, plus ou moins une discipline scientifique, une science par ce qu'elle a un fondement, un enjeu, ou un objet d'étude, et elle à une méthode, c'est-à-dire des défis qui gouvernent, toute la démarche vers la satisfaction de l'objet d'étude depuis son fondement c'est-à-dire le besoin de base.

Cette démarche est appelée la méthode statistique, La statistique est une discipline par ce qu'elle est organisée à l'aide des règles et principes qui favorisent la réalisation de l'enjeu statistique ciblé.

## 1.2- Concepts de base en statistique

**Population statistique :** C'est l'ensemble des sujets ou objets étudié dans le contexte de l'étude. Plus précisément, c'est l'ensemble des sujets ou objets d'intérêts, de l'étude statistique dans un espace donné à un moment précis.

**Individu statistique/ sujet étudié :** C'est chaque sujet d'intérêt statistique dans un contexte donné.

**Exemple :** Supposons que nous nous intéressons, aux notes obtenues en mathématique par les étudiants de l'ENSPD en 2022-2023.

- **Population :** ensemble des notes obtenues en mathématique par les étudiants de l'ENSPD pendant l'année académique 2022-2023.
- **Individu Statistique :** Chaque note obtenue en mathématique par les étudiants de l'ENSPD pendant l'année académique 2022-2023.
- **Enjeu :** la connaissance du nombre total de note obtenue en mathématique par les étudiants de l'ENSPD pendant l'année académique 2022-2023.

Rappelons qu'en tant que science la statistique à un objet et des objectifs :

- **Objet de la statistique :** C'est l'enjeu de la Statistique c'est-à-dire la connaissance du nombre total d'individus statistique dénombré dans un contexte. C'est donc l'effectif, la taille ou encore le volume de la population statistique. De façon simple, l'objet de la statistique est l'effectif, la taille ou le volume de la population statistique étudiée.
- **Objectif fondamental de la statistique :** Dénombrer une population statistique dans un contexte bien précis.

En statistique les résultats intermédiaires sont nécessaires pour obtenir convenablement et plus méthodiquement l'enjeu de l'étude statistique. Ce sont des défis qui peuvent être des obstacles à surmonter ou soit des passages stratégiques vers l'enjeu statistique ciblé. À ces défis sont associés des objectifs appelés objectifs intermédiaires en statistiques.

**Remarque :** Dans une population statistique les sujets étudiés sont identiques d'un point de vue donné. C'est la caractéristique commune des individus de cette population qui a permis de définir et de déterminer cette population.

**Exemple** : Ensemble des étudiants de l'ENSPD en 2022-2023

**Caractéristique commune** : est étudiant de l'ENSPD en 2022-2023

Mais dans une population plusieurs caractéristiques peuvent ne pas être identiques

**Exemple** : Dans l'ensemble des étudiants de l'ENSPD entre 2022-2023, les caractéristiques tels que : l'âge, le poids, la taille, le sexe, la religion, sont des caractéristiques non identiques.

### **Caractère statistique / Variable statistique**

Dès l'instant que ces caractéristiques sont variables dans la population statistique on ne les appellera plus caractéristique mais caractère. Techniquement, on va les appeler **variables statistiques**. Stratégiquement, les objectifs courants de la statistique sont de décrire la population statistique selon chaque caractère statistique dans cette population. Cela consiste pour une variable donnée de dénombrer la population par modalité ou par possibilité de la variable considéré dans la population. On distingue deux catégories de variable dans une population statistique : les variables qualitatives et quantitatives.

#### **Les variables qualitatives**

Les variables qualitatives se réfèrent à des attributs c'est-à-dire des caractéristiques individuelles attribuées.

**Exemple** : Chez les êtres vivants : le sexe, l'état matrimonial, la religion, sont des variables non mesurables.

#### **Les variables quantitatives**

Ce sont des variables auxquels sont associées des nombres. Si les nombres associés sont des entiers naturels elles sont dites discrètes. Sinon elle est continue.

**Remarque** : Parfois, les variables quantitatives continues sont difficilement exploitable dans les états bruts. Pour contourner la difficulté de leur exploitation on les regroupe en des classes de valeur.

Du coup, la variable quantitative fonctionne comme une variable qualitative. Mais si l'on considère que dans chaque classe de valeur les valeurs sont réparties de façon linéaire ou continue alors chaque classe de valeurs peut être remplacé par son centre. Dans ce cas la variable devient à nouveau exploitable comme une variable quantitative.

## **II – Observation Statistique**

L'observation statistique constitue la méthode de la statistique. Autrement dit, elle repose sur l'observation. Observer c'est regarder attentivement pendant un moment nécessaire. Du coup, une observation s'effectue en trois temps :

1. On regarde
2. On accorde une attention ou un intérêt à ce qu'on regarde
3. On accorde un temps ou une période suffisante pour satisfaire son intérêt.

#### **Qu'est-ce qu'on observe en statistique ?**

En statistique on observe les individus selon une ou plusieurs variables. On observe les individus statistiques et plus précisément les caractères / variables statistiques.

#### **2.1- Intérêt de l'observation statistique**

Il y en a deux :

- Premièrement : l'individu statistique
- Deuxièmement : les caractères ou variables statistiques associées aux individus statistiques.

L'observation en statistique consiste à dénombrer les sujets selon différentes possibilités des caractères ou variable statistique ciblé dans la population. Dénombrer au cours d'une période suffisante, c'est-à-dire jusqu'à satisfaction totale. La satisfaction totale suppose que soit le résultat attendu est obtenu, soit la manifestation du caractère décompté est achevé.

#### **2.2- Types d'observations**

Il y a différents types d'observation, chaque type dépend du contexte, des atouts et des contraintes. On en distingue couramment trois :

- Observation instantanée
- Observation rétrospective
- Observation continue

Quel que soit le type d'observation, elle est faite à l'aide d'un outil ou d'un formulaire appelé questionnaire. Ces outils sont remplis soit oralement (interview, entretien téléphonique, ...) soit par remplissage (écrit / électronique, ...). Ces outils prennent différentes formes selon la nature et le type d'observation utilisé.

##### **2.2.1- Observation instantanée**

Dans une observation instantanée la période d'observation se réfère au moment où à l'instant actuel où l'on regarde. Elle signifie donc que la période de collecte des données est le moment actuel, généralement l'année en cours. Généralement, une telle observation porte sur les caractéristiques individuelles actuel du moment du sujet étudié. C'est le cas du recensement de la population.

##### **Avantages :**

Les avantages de cette méthode d'observation sont que les données sur les caractéristiques du moment sont appréhendées sans trop de biais.

### Inconvénients :

- Cette méthode ne permet pas de savoir les données relevant du passé des sujets
- Elle ne permet pas de saisir la situation des migrations
- La période d'observation étant très courte elle risque d'être onéreuse c'est-à-dire coûteuse pour réaliser l'exhaustivité de l'observation.

### **2.2.2- Observation rétrospective**

Elle consiste à un moment t à interroger les individus sur des évènements ou des faits qu'ils ont vécus par le passé ou depuis leurs naissances. Ainsi, la période d'observation est fonction de la collecte, des objectifs spécifiques de chaque opération d'observation.

### Avantages :

En un laps de temps, cette méthode permet d'effectuer une observation sur une très longue période. Par ailleurs, elle n'assujettit pas l'observation à des perturbations, consistant à des sorties ou des entrées (phénomène perturbateurs) dans la population étudiée au cours de la période d'observation.

### Inconvénients :

Mais c'est une méthode qui a plusieurs inconvénients.

- L'Effet de sélection découlant de l'hypothèse de la représentativité ou de la non représentativité des personnes enquêtés en tant que résidus des phénomènes de la mortalité et/ou de migration dans leurs générations respectives.
- l'Effet de télescopage, qui se traduit par la possibilité d'omission ou de double compte ou de compte multiple quant aux informations fournis par les personnes enquêtées du fait de l'effet de mémoire.

La méthode rétrospective est généralement utilisée dans les Enquêtes Démographiques et de Santé, par exemple.

### **2.2.3- Observation continue**

Elle consiste à enregistrer les évènements au fur et à mesure qu'ils surviennent dans la population étudiée. On les appelle encore des observatoires de population très souvent la période d'observation est illimitée.

### Avantages :

- Pas de biais de collecte

### Inconvénients :

- L'observation est perturbée par les entrées ou les sorties de la population étudiée
- Elle est fastidieuse : cet inconvénient est d'autant plus important que la période est longue.
- Elle est coûteuse

### **2.3- Outil d'observation collecte en statistique : le questionnaire**

Le questionnaire est un document structuré dont la structure est strictement déterminée par les objectifs de la statistique ou de l'observation statistique. Il ne s'agit pas de l'objectif central c'est-à-dire celle de dénombrer la population étudiée. Mais des autres objectifs spécifiques notamment les objectifs relatifs aux caractères statistiques d'intérêts. Ainsi, dans une observation statistique ciblant quatre caractères ou variable statistiques, alors la structure élémentaire du questionnaire comprend quatre éléments. Chacun des éléments est une question du questionnaire. Très souvent une série de caractère statistique être regroupés sous une seule dénomination. Ce groupe de caractère doit former une rubrique ou une composante du questionnaire. Cette composante est meublée par une série de question bien organisé tel que chacune des questions soient associées à chaque caractère statistique.

De ce point un questionnaire à trois parties :

- La première est relative aux identifiants du contexte, du sujet étudié et de l'enquêteur.
- La seconde est relative aux caractéristiques naturelles ou sociaux économiques du sujet étudié
- La troisième partie est structurée en composante ou rubrique relative à la problématique de l'étude statistique.

### **2.4- La Base de données**

La base de données est un dispositif informatique permet d'enregistrer l'ensemble des données collectées, données qui sont stockés soit dans le même dispositif informatique par une programmation informatique. Ces données sont mises en relation entre elle dans un système conçu à cet effet dénommé base des données collectées.

## **III – Synthèse des données Statistiques**

Une fois l'observation statistique terminée, c'est-à-dire une fois les données collectées il est dit précédemment que ces données sont stockées dans un dispositif appelé base de données. Elle est un réservoir dont l'utilité est de stocker les données grâce à l'observation pour une exploitation opérationnelle de ces données. Il faut procéder à une synthèse, synthèse est dictée par les objectifs de l'observation statistique. Rappelons que l'observation statistique cible d'une part les individus statistiques puis d'autres part des caractères statistiques pertinents au sein de la population étudiée. Cette synthèse s'effectue toujours objectif par objectif puisque chaque objectif de l'observation statistique cible une variable alors que la synthèse de données s'effectue variable par variable.

La synthèse pour une variable statistique donnée consiste à déterminer les effectifs des sujets ou des individus appartenant à chacune des possibilités de la variable.

**Exemple :** Si l'on s'intéresse à la variable sexe dans la population des étudiants de master de l'ENSPD en 2022-2023, la synthèse de données se présente comme suit :

Masculin = 13 étudiants

Féminin = 1 étudiante

L'effectif total des étudiants est la somme des effectifs des différentes possibilités de la variable dans la population étudiée. Dans notre exemple, l'effectif des étudiants est  $13 + 1 = 14$  étudiants.

De façon plus pragmatique et aisée, cette synthèse des données statistique se résume à l'aide d'un tableau appelée tableau statistique de la variable étudiée.

### 3.1- Tableau statistique d'une variable statistique

Le tableau statistique d'une variable statistique est une matrice dont les lignes sont définies par les possibilités de la variable étudiée et les colonnes sont définis par les fréquences ou les effectifs ou les nombres d'individu de la population étudiée appartenant à chaque modalité. La première ligne du tableau est définie par le libellé de la variable. Alors que la dernière ligne est décrite par le total. Quant aux colonnes du tableau, la première colonne décrit la variable étudiée dont la première cellule correspond au libellé ou nom de la variable ; la dernière cellule correspond au total ; puis les autres cellules de cette colonne correspondent aux différentes possibilités de la variable. Outre cette première colonne, les autres colonnes sont déterminées par les différents types de fréquence, associées aux différentes possibilités de la variable étudiée. Les fréquences sont de deux types, soit elles sont absolues (effectifs ou nombres), soit elles sont relatives (pourcentages ou proportions).

Un tableau statistique a aussi idéalement trois colonnes :

- La première colonne décrit la variable concernée
- La deuxième décrit les fréquences absolues
- La troisième décrit les fréquences relatives

Un tableau statistique a plusieurs lignes. Le nombre de lignes est égal au nombre total de possibilités + 2. Ces deux lignes supplémentaires sont la première (libellé) et la dernière (total). Le nombre de colonnes d'un tableau statistique est minimalement de trois.

**Remarque :** Tout tableau statistique à un titre qu'il faut rigoureusement préciser au-dessus du tableau. Le libellé du titre d'un tableau statistique se présente comme suit :

Répartition des individus statistiques de la population statistique selon la variable nom de la variable »

**Exemple :** Répartition des étudiants du Master de l'ENSPD au cours de l'année académique 2022-2023 selon l'âge.

Si dans votre tableau statistique vous n'avez qu'un seul type de fréquence alors le type de votre tableau doit préciser le type de fréquence décrit dans ce tableau selon qu'il s'agit des effectifs ou des pourcentages ou des proportions.

**Exemple :**

- Répartition des effectifs des étudiants du Master 1 de l'ENSPD au cours de l'année académique 2022-2023 selon l'Age.
- Répartition en pourcentage des étudiants du Master 1 de l'ENSPD au cours de l'année académique 2022-2023 selon l'Age.

#### IV – Analyses Statistiques Descriptives – Travaux Personnels de groupes d'Étudiants (TPE)

Une fois les statistiques collectés, la base de données mise en place, l'analyse statistique descriptive commence par la production des tableaux des variables ciblé. Les analyses statistiques descriptives portent sur ces tableaux.

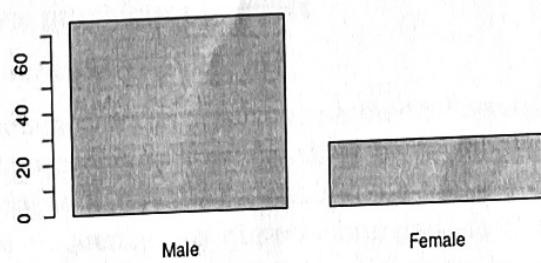
##### 4.1- Statistiques descriptives univariées

###### 4.1.1 – Analyse d'une variable qualitative

Lorsqu'on a une variable qualitative, l'analyse descriptive se réduit quasi exclusivement à un commentaire du tableau statistique, ces commentaires sont idéalement illustrés par des graphiques.

###### *Les diagrammes en bande*

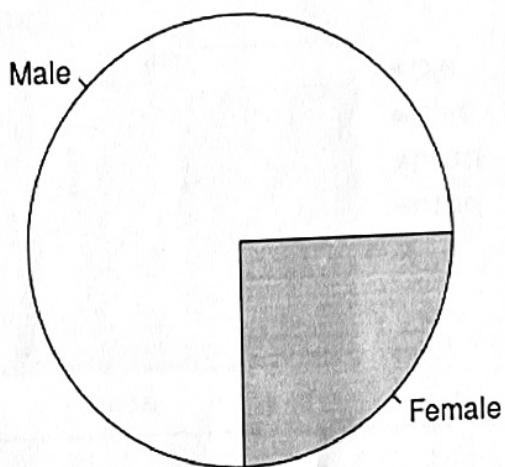
La construction de ce graphique consiste à porter en abscisses les modalités de la variable statistique, de façon arbitraire. Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux fréquences relatives.



*Figure 1: Diagramme en bande des pourcentages des ménages selon le sexe d'après les données combinées du RGPH Bénin et du RGPH Togo en 2013.*

### **Les diagrammes circulaires et semi-circulaires**

Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle aux fréquences relatives des différentes modalités de la variable statistique ciblée.



*Figure 2: Diagramme en circulaire des pourcentages des ménages selon le sexe d'après les données combinées du RGPH Bénin et du RGPH Togo en 2013.*

#### **4.1.2 – Analyse d'une variable quantitative**

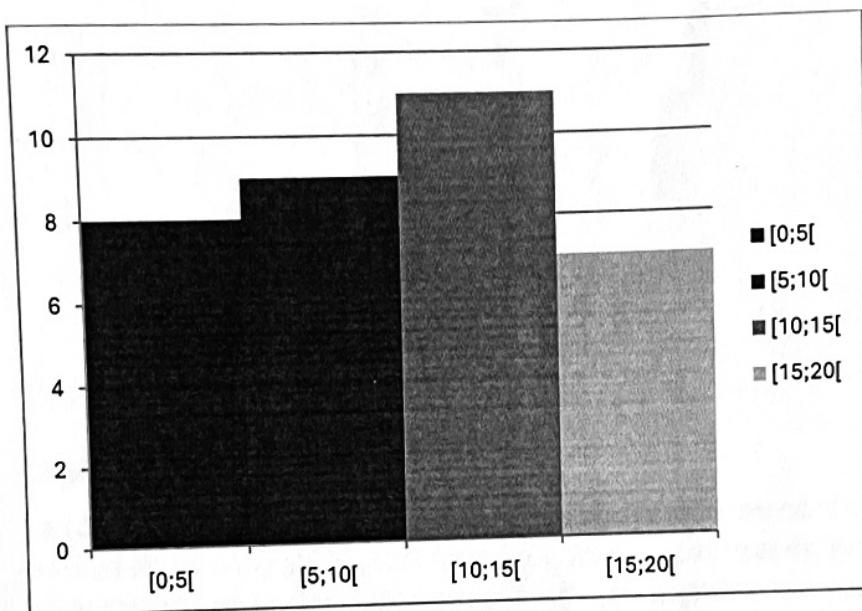
Lorsqu'il s'agit d'une variable quantitative, on procède non seulement à un commentaire du tableau statistique et des représentations graphiques associées, mais également au calcul de quelques indices synthétiques appelés caractéristiques de tendance centrale. Ces caractéristiques sont interprétées à l'aide de certains paramètres de dispersion de la distribution autour des caractéristiques de tendance centrale.

##### **4.1.2.1 Représentations graphiques usuelles**

###### **Histogramme de fréquences**

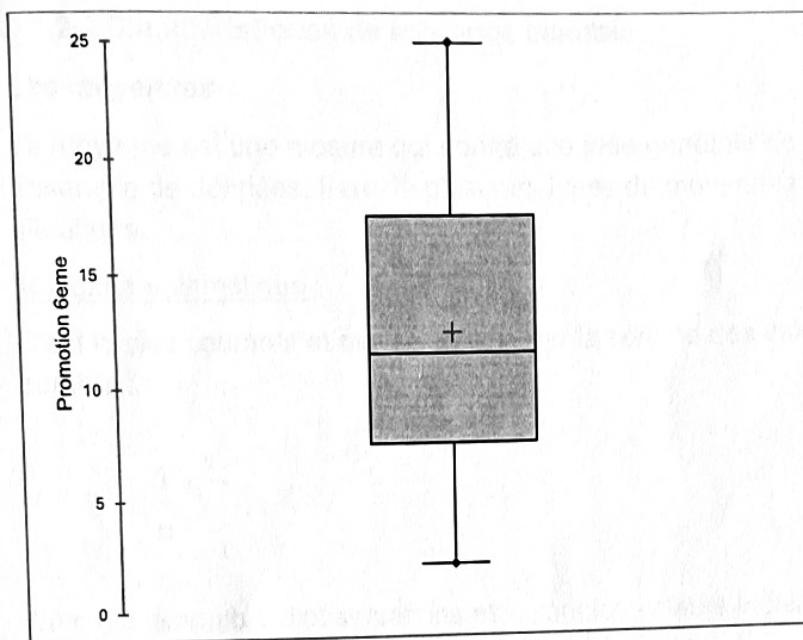
Un histogramme de fréquence est un graphique permettant de visualiser la distribution d'une variable numérique continue. Il est constitué de rectangles adjacents, chacun correspondant à une classe (intervalle de valeurs). La hauteur de chaque rectangle est proportionnelle à la fréquence des observations dans la classe associée. L'axe horizontal (abscisse) représente les plages de valeurs de la variable, divisées en intervalles de largeur fixe ou variable, tandis que l'axe vertical (ordonnée) indique la fréquence absolue ou relative. Les rectangles, dépourvus d'espaces entre eux (sauf

pour des classes vides), mettent en évidence la densité des données. Cet outil est particulièrement utile pour identifier la forme d'une distribution (symétrie, asymétrie, multimodalité) et estimer des paramètres comme le mode. Toutefois, son interprétation dépend du choix de la largeur des classes : des classes trop étroites peuvent fragmenter la visualisation, tandis que des classes trop larges masquent les détails.



### **Boîte à moustaches**

La boîte à moustache, ou diagramme en boîte, résume une distribution de données à l'aide de cinq valeurs clés : le minimum, le premier quartile ( $Q_1$ ), la médiane ( $Q_2$ ), le troisième quartile ( $Q_3$ ) et le maximum. La « boîte » centrale s'étend de  $Q_1$  à  $Q_3$ , englobant 50 % des données, avec une ligne marquant la médiane. Les « moustaches », prolongeant la boîte, atteignent les valeurs minimale et maximale non aberrantes, définies conventionnellement comme les données situées dans l'intervalle  $[Q_1 - 1,5 \times IQR ; Q_3 + 1,5 \times IQR]$ , où  $IQR$  est l'écart interquartile ( $Q_3 - Q_1$ ). Les valeurs en dehors de cette plage sont représentées par des points isolés, identifiés comme aberrantes. Ce graphique synthétique permet de comparer rapidement plusieurs distributions, d'évaluer leur symétrie (via la position de la médiane) et de détecter des valeurs extrêmes. Cependant, il ne révèle pas les particularités comme la multimodalité ou les creux dans la distribution.



### **Courbe cumulative**

La courbe cumulative, ou ogive, illustre l'évolution des fréquences cumulées (absolues ou relatives) d'une variable numérique. Sur l'axe des abscisses figurent les valeurs de la variable (ou les bornes supérieures des classes pour des données groupées), tandis que l'axe des ordonnées affiche les fréquences cumulées, allant de 0 à 100 % ou au total des observations. La courbe, tracée sous forme de marches d'escalier (pour des classes discrètes) ou d'une ligne continue, est toujours croissante. Elle permet de déterminer des percentiles (par exemple, la médiane correspond au point où la courbe atteint 50 %) ou d'estimer la proportion de données inférieures à un seuil donné. Bien qu'utile pour analyser la répartition globale, elle ne fournit pas d'informations détaillées sur la densité locale des observations.

### **Courbe de densité**

La courbe de densité est une représentation lissée d'un histogramme, estimant la fonction de densité de probabilité d'une variable continue. L'axe horizontal présente les valeurs de la variable, et l'axe vertical indique la densité (l'aire totale sous la courbe étant égale à 1). Générée par des méthodes de lissage, cette courbe est idéale pour comparer une distribution empirique à une distribution théorique (par exemple, une loi normale) ou pour visualiser des structures complexes comme des modes multiples. Cependant, son apparence dépend fortement du paramètre de lissage (bandwidth) : un lissage trop faible crée un tracé irrégulier, tandis qu'un lissage excessif atténue les particularités de la distribution.

#### 4.1.2.2 Caractéristiques de tendance centrale

##### Les moyennes

La moyenne est une mesure qui donne une idée générale de la valeur centrale d'un ensemble de données. Il existe plusieurs types de moyennes adaptées à différentes situations.

##### Moyenne arithmétique

C'est la plus courante et elle se calcule par la somme des valeurs divisée par leur nombre :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Où n est le nombre d'observations et  $x_i$  sont les valeurs individuelles.

Exemple : Supposons qu'on dispose des données sur les salaires annuels de 160 employés d'une entreprise X au cours de l'année 2024-2025. Ces données sont récapitulées dans le tableau statistique suivant :

Titre du tableau : Répartition des salariés de l'entreprise X au cours de l'année 2024-2025 selon leurs salaires

| Salaires | Effectifs des salariés | Pourcentage des salariés % | Fréquence cumulés |
|----------|------------------------|----------------------------|-------------------|
| 52000    | 58                     | 36                         | 58                |
| 60000    | 20                     | 13                         | 78                |
| 70000    | 32                     | 20                         | 110               |
| 120000   | 50                     | 31                         | 160               |
| Total    | 160                    | 100                        |                   |

Calculons la moyenne arithmétique de la distribution des salariés

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Avec n le nombre de valeur possible de la variable et  $x_i$  la valeur pour le salariés i et  $i \in \{1, \dots, n\}$

$$\text{donc } \bar{x} = \frac{1}{4} (52000 + 60000 + 70000 + 12000) = 75500$$

Alors le salaire moyen des salariés de cet entreprise est de 75500f

### Moyenne pondérée

Lorsque certaines valeurs ont plus d'importance que d'autres, on leur attribue des poids :

$$\bar{X}_p = \frac{\sum w_i X_i}{\sum w_i}$$

Exemple : Calculons la moyenne pondérée de la distribution des salariés

$$\bar{X}_p = \frac{\sum w_i X_i}{\sum w_i}$$

avec  $w_i$  l'effectif des salariés  $i$  et  $X_i$  la valeur pour les salariés  $i$  et  $i \in \{1, \dots, N\}$  et  $N$  le nombre total de salariés

$$\text{Donc } \bar{X} = \frac{1}{160} (3016000 + 1200000 + 2240000 + 6000000) = 77850$$

Alors le salaire moyen dans cet entreprise est de 77850f

### Moyenne géométrique

Utilisée pour des taux de croissance (économie, finance) :

$$G = \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}}$$

### Moyenne harmonique

Appropriée pour des rapports (vitesse moyenne, rendement) :

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

Où  $n$  est le nombre d'observations et  $x_i$  sont les valeurs individuelles.

### La médiane

La médiane est la valeur qui partage une série de données ordonnées en deux parties égales : 50 % des valeurs sont inférieures et 50 % sont supérieures.

#### Calcul :

- Si  $n$  est impair, la médiane est la valeur centrale.
- Si  $n$  est pair, la médiane est la moyenne des deux valeurs centrales.

**Définition avec la moyenne :** Robustesse (La médiane n'est pas influencée par les valeurs extrêmes contrairement à la moyenne.)

Exemple : Calculons la médiane de la distribution des salariés . Soit  $M$  cette médiane

A partir du tableau ci-dessus, 50% des salariés a une fréquence cumulée égale à 80.

Par interpolation linéaire on a :

$$\frac{70000 - 60000}{110 - 78} = \frac{M - 60000}{80 - 78}$$

Donc  $M = \frac{10*2+60000*32}{32} = 60625$ . Alors 50% des salariés de cet entreprise gagnent au moins 60625f

#### 4.1.2.2 Caractéristiques de dispersion

L'analyse des caractéristiques de dispersion est une méthode statistique utilisée pour évaluer la variabilité ou l'étalement des données autour de la tendance centrale. Contrairement aux mesures de tendance centrale (moyenne, médiane, mode, fractile), qui résument la valeur centrale des données, les mesures de dispersion quantifient à quel point les données sont dispersées ou concentrées. Voici une analyse détaillée des principales mesures de dispersion :

##### *L'Étendue (Range)*

L'étendue est la différence entre la valeur maximale et la valeur minimale d'un ensemble de données

**Etendue = Valeur maximale - Valeur minimale**

**Exemple :** Calculons l'étendue de cette distribution des salariés

$$E=120000 - 52000 = 68000$$

Alors l'écart entre le salaire maximal et le salaire minimal dans cette entreprise est de 68000f

##### *La Variance*

La variance mesure la moyenne des carrés des écarts par rapport à la moyenne. Elle quantifie à quel point les données s'écartent de la moyenne.

➤ Pour la population

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Où  $x_i$  est chaque valeur,  $\mu$  est la moyenne, et  $N$  est le nombre de valeurs.

Pour l'échantillon

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

**Exemple :** Calculons la variance de cette distribution

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

avec  $N$  l'effectif total des salariés et  $x_i$  la valeur possible du salarié  $i$ .

$$\text{Donc } \sigma^2 = \frac{1}{4}((52000 - 75500)^2 + (60000 - 75500)^2 + (70000 - 75500)^2 + (120000 - 75500)^2) = 700750000$$

### **L'Écart-Type (Standard Déviation)**

L'écart-type est la racine carrée de la variance. Il exprime la dispersion dans la même unité que les données.

➤ Pour la population

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

➤ Pour l'échantillon

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

**Exemple :** Calculons l'écart-type de cette distribution.

$$s = \sqrt{700750000} = 26471.68$$

### **L'Écart Interquartile (IQR)**

L'écart interquartile est la différence entre le troisième quartile (Q3) et le premier quartile (Q1). Il mesure la dispersion des 50 % centraux des données.

$$IQR = Q3 - Q1$$

### **Le Coefficient de Variation (CV)**

Le coefficient de variation est une mesure relative de dispersion, exprimée en pourcentage. Il compare l'écart-type à la moyenne.

➤ Pour la population

$$CV = \left( \frac{\sigma}{\mu} \right) \times 100$$

➤ Pour l'échantillon

$$CV = \left( \frac{s}{\bar{x}} \right) \times 100$$

**Exemple :** Calculons le coefficient de variation de cette distribution.

$$CV = \left( \frac{26471.6}{75500} \right) * 100 = 35$$

### **Intervalle de confiance**

L'intervalle de confiance pour un paramètre statistique est un intervalle qui, avec une certaine probabilité, contient la valeur réelle de ce paramètre

Exemple : Le salaire moyen des salariés de cette entreprise est compris dans l'intervalle ]49028,101972[

## 4.2- Statistiques descriptives bivariées

L'analyse descriptive des variables prennent certaines dimensions avancées tel que l'analyse simultanée de deux ou trois variables statistiques. Cette approche d'analyse consiste en un croisement des tableaux de manière à ressortir les comportements d'une ou de deux variables en présence d'une autre.

### 4.2.1 Cas de deux variables qualitatives

#### 4.2.1.1. 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 : Tableau de répartition 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            |

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 correspondant à l'intersection de deux modalités.

#### 4.2.1.2. Distributions marginales et conditionnelles

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

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.

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.

#### 4.2.1.3. Profils-lignes et Profils colonnes

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 :**

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

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

*Exemple : 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%          |

**Interprétation :** 54 % des femmes sont inscrites en PSE, contre 67 % des hommes et 33 % des hommes choisissent la filière 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 :**

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

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

Exemple : 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% |

Interprétation : 83% des individus en PSE sont des hommes, contre 17% qui sont des femmes. 74% des individus en SA sont des hommes, contre 26% qui 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

##### *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 :

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

##### *Test d'indépendance de Khi-deux ( $\chi^2$ )*

Le Khi-deux ( $\chi^2$ ) est une statistique qui mesure l'écart entre les effectifs observés dans un tableau de contingence et les effectifs théoriques attendus sous l'hypothèse d'indépendance entre les deux variables qualitatives. Il s'applique sous les conditions suivantes :

- 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  $\geq 5$ .
- Les effectifs observer doivent être supérieur ou égale à 5.

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

Le test d'indépendance de khi-2 se fait en six à (07) étapes définir ci-dessous.

Étape 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).

#### Etape 2. Construire le tableau de contingence

Il s'agit d'un tableau croisant les modalités des deux variables. Chaque case contient les fréquences observées.

#### Etape 3. Calculer les fréquences théoriques

Les fréquences théoriques sont calculées selon la formule ci-dessous :

$$e_{ij} = \left( \frac{n_i \times n_j}{n} \right)$$

Où :  $n_i$  est le total de la ligne  $i$ ,  $n_j$  est le total de la colonne  $j$ , et  $n$  est l'effectif total.  
Les valeurs théoriques calculées désignent les effectifs attendus en considérant que les deux variables sont indépendantes.

#### Etape 4. Calculer la statistique du khi-deux

La statistique du  $\chi^2$  est calculée selon la formule ci-dessous :

$$\chi^2 = \sum_{i=1}^n \sum_{j=i}^n \left( \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \right)$$

Où:  $n_{ij}$  est l'effectif observé pour la combinaison  $(A_i, B_j)$  et  $e_{ij}$  l'effectif théorique sous l'hypothèse d'indépendance.

#### Etape 5. Déterminer les degrés de liberté (dl)

Le degré de liberté est calculé selon la formule ci-dessous :

$$dl = (r - 1)(c - 1)$$

Où :  $r$  est le nombre de lignes et  $c$  est le nombre de colonnes.

#### Etape 6. Comparer avec la valeur critique ou calculer la p-valeur

On utilise la table de la loi du  $\chi^2$  pour trouver la valeur critique correspondant à un seuil  $\alpha$  fixé.

Après avoir déterminer le niveau de signification ( $\alpha$ ) et déterminer le degré de liberté (dl), pour trouver la valeur critique il faut :

- Cherchez la ligne correspondant aux degrés de liberté (dl).
- Cherchez la colonne correspondant au niveau de signification ( $\alpha$ ).
- L'intersection donne la valeur critique du  $\chi^2$ .

Plus les degrés de liberté augmentent, plus la valeur critique augmente. Si le niveau de signification ou le degré de liberté ne sont pas directement dans la

table, on peut faire une interpolation ou utiliser un logiciel statistique (comme Excel, R ou Python).

#### Étape 7. Règle de Décision :

On compare le  $\chi^2$  calculer à la valeur critique  $I_u$ .

- Si  $\chi^2$  calculer est supérieur à la valeur critique  $I_u$ , on rejeter l'hypothèse nulle ( $H_0$ ). Il y a une différence significative. On conclut que les variables sont dépendantes.
- Si  $\chi^2$  calculer est inférieur ou égale à la valeur critique  $I_u$ , on ne peut pas rejeter l'hypothèse nulle ( $H_0$ ). On n'a donc aucune preuve statistique pour conclure à une dépendance entre les variables, on n'accepte donc l'hypothèse privilégiée ( $H_0$ ).

#### **Coefficient de Cramer (V de Cramer)**

Le coefficient de Cramer est une mesure de liaison qui quantifie l'intensité de la relation entre deux variables qualitatives. Il est basé sur la statistique du  $\chi^2$ , mais il est normalisé pour être compris entre 0 et 1. Il est calculé à travers la formule suivante :

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}$$

Où  $\chi^2$  est la statistique de Khi-deux,  $n$  l'effectif total,  $r$  le nombre de ligne dans le tableau de contingence,  $c$  le nombre de colonne dans le tableau de contingence et  $\min(r-1, c-1)$  est le minimum entre  $(r-1)$  et  $(c-1)$ .

Si  $V=0$  : il y a indépendance parfaite entre les variables. Si non si  $V=1$  : il y a une liaison parfaite entre les variables (dépendance totale). Plus  $V$  est proche de 1, plus la liaison est forte et plus elle est proche de 0 plus l'intensité de la liaison diminuée.  
Intervalles couramment utilisés pour interpréter la force de l'association :

- o  $V < 0,1$  : Association très faible
- o  $0,1 \leq V < 0,30$  : Association faible
- o  $0,3 \leq V < 0,50$  : Association modérée
- o  $V \geq 0,5$  : Association forte

#### Avantages

Le coefficient de Cramer est normalisé, ce qui permet de comparer des liaisons entre des tableaux de contingence de tailles différentes. Il est utile pour interpréter l'intensité de la liaison une fois que le test du Khi-deux a rejeté l'hypothèse d'indépendance.

#### Limites

Il ne donne pas d'information sur la nature de la liaison (par exemple, négative ou positive).

Remarque : Le test du Chi-deux et le coefficient de Cramer sont deux indicateurs sont complémentaires : le Khi-deux permet de détecter une dépendance, et le coefficient de Cramer permet de quantifier son intensité.

#### 4.2.2 Cas de deux variables quantitatives

##### 4.2.2.1. Notion de liaison fonctionnelle

Une liaison fonctionnelle existe entre deux variables lorsque les valeurs de l'une peuvent être déterminées par l'autre à travers une fonction sous la forme :  $Y = f(X)$ . En général, nous examinons la relation linéaire entre deux variables quantitatives

##### 4.2.2.2. Indicateurs numériques de corrélation linéaire

###### La covariance

La covariance entre deux variables  $X$  et  $Y$  mesure la manière dont ces deux variables varient simultanément. Plus précisément, elle indique si, lorsque  $X$  augmente,  $Y$  a tendance à augmenter (ou diminuer) en même temps, et dans quelle proportion. Sa formule est :

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Où :

- $\text{Cov}(X, Y)$  est la covariance entre les variables  $X$  et  $Y$ ,
- $x_i$  et  $y_i$  représentent les valeurs des variables  $X$  et  $Y$  pour l'observation  $i$ ,
- $\bar{x}$  et  $\bar{y}$  sont les moyennes respectives des variables  $X$  et  $Y$ ,

###### Le coefficient de corrélation de Pearson

Elle mesure la force et la direction de la relation linéaire entre deux variables. Sa formule est :

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Où :

- $\text{Cov}(X, Y)$  est la covariance entre les variables  $X$  et  $Y$ ,
- $\sigma_X$  et  $\sigma_Y$  sont les écarts-types des variables  $X$  et  $Y$ , respectivement.

Le coefficient de corrélation de Pearson est une valeur qui varie entre  $-1$  et  $1$ .

###### Interprétation

$r = 1$  : Corrélation parfaitement positive

Il existe une relation linéaire parfaite entre les deux variables. Lorsque X augmente, Y augmente dans la même proportion, sans aucune variation aléatoire. Les points de données sont alignés sur une droite ascendante.

#### ***r= -1 : Corrélation parfaitement négative***

Il existe une relation linéaire parfaite mais inverse entre les deux variables. Lorsque X augmente, Y diminue dans la même proportion, sans aucune variation aléatoire. Les points de données sont alignés sur une droite descendante.

#### ***r=0 : Aucune corrélation linéaire***

Il n'y a aucune relation linéaire entre les deux variables. Cela ne signifie pas qu'il n'y a pas de relation entre X et Y, mais seulement qu'il n'existe pas de relation linéaire. Il peut y avoir une relation non linéaire (par exemple, quadratique ou exponentielle) entre les deux variables.

#### ***0 < r < 1 : Corrélation positive modérée à forte***

Si r est entre 0 et 1, cela signifie que les variables sont positivement corrélées, c'est-à-dire que lorsque X augmente, Y a tendance à augmenter également. Plus r est proche de 1, plus la relation est forte et plus la relation est linéaire.

#### ***-1 < r < 0 : Corrélation négative modérée à forte***

Si r est entre -1 et 0, cela signifie que les variables sont négativement corrélées, c'est-à-dire que lorsque X augmente, Y a tendance à diminuer. Plus r est proche de -1, plus la relation est forte et inversement linéaire.

#### **4.2.2.3 Indicateurs numériques de corrélation de rang :**

##### ***Corrélation de Spearman***

Il mesure la relation monotone entre deux variables. Sa formule est :  $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$

Où :

$d_i$  est la différence entre les rangs des deux variables pour chaque observation i,  
n est le nombre total d'observations.

Une valeur de  $\rho = 1$  indique une corrélation monotone parfaite entre les rangs des années f et les rangs des revenus. Cela confirme la forte relation positive observée avec le coefficient de Pearson.

##### ***Corrélation de Kendall***

Il évalue l'association entre les rangs des valeurs des variables. Sa formule est :  $\tau = \frac{(C-D)}{\frac{1}{2}n(n-1)}$

où :

C'est le nombre de paires concordantes,

D est le nombre de paires discordantes,

n est le nombre total d'observations.

Une paire  $(x_i, y_i)$  et  $(x_j, y_j)$  est :

✓ Concordante si  $(x_i - x_j)(y_i - y_j) > 0$

✓ Discordante si  $(x_i - x_j)(y_i - y_j) < 0$

### Courbe de régression

La régression linéaire permet de quantifier la relation entre une variable dépendante et une variable indépendante. Elle permet de trouver l'équation de la droite qui "correspond le mieux" aux données.

L'équation de type  $Y = B_0 + B_1 X + \varepsilon$ . On démontre que :

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

### 4.2.3 Cas d'une variable qualitative et d'une variable quantitative

#### 4.2.3.1. La variance inter-groupe

La variance inter-groupe (ou variance entre groupes) est un concept statistique qui mesure la variation des moyennes des groupes par rapport à la moyenne globale. Elle est utilisée dans le cadre de l'analyse de variance (ANOVA), une méthode statistique permettant de comparer les moyennes de plusieurs groupes pour déterminer si des différences significatives existent entre deux.

**Définition :** La variance inter-groupe quantifie la dispersion des moyennes des différents groupes par rapport à la moyenne globale de l'échantillon. En d'autres termes, elle mesure la part de la variation totale qui est due aux différences entre les groupes.

**Calcul :** La variance inter-groupe est calculée en prenant la somme des carrés des différences entre les moyennes des groupes et la moyenne globale, multipliée par la taille de l'échantillon de chaque groupe

$$SSB = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})^2$$

Avec,

k : nombre de groupes

n<sub>i</sub> : taille de l'échantillon du groupe i

$\bar{X}_i$  : moyenne du groupe i

$\bar{X}$  : moyenne globale

Ensuite, la **variance inter-groupe** (MSB) est obtenue en divisant cette somme par les degrés de liberté associés (en général  $k-1$ ) :

$$MSB = \frac{SSB}{k - 1}$$

Avec :

- MSB : La moyenne des carrés entre les groupes ;
- k le degré de liberté

Cette formule est utilisée pour évaluer si les différences observées entre les groupes sont dues au hasard ou si elles reflètent des différences réelles dans les populations étudiées.

**Interprétation :**

- Une **variance inter-groupe élevée** indique que les moyennes des groupes sont assez différentes les unes des autres par rapport à la moyenne globale.
- Une **variance inter-groupe faible** suggère que les groupes sont relativement similaires entre eux.

La variance inter-groupe est comparée à la **variance intra-groupe** (variation à l'intérieur des groupes) dans le cadre de l'ANOVA (Analyse de la variance) pour déterminer si les différences observées sont statistiquement significatives.

#### 4.2.3.2. La variance inter-groupe

La **variance intra-groupe** (ou variance à l'intérieur des groupes) est un concept statistique qui mesure la variation des données à l'intérieur de chaque groupe, c'est-à-dire la dispersion des valeurs des individus au sein du même groupe par rapport à leur propre moyenne.

**Définition :**

La variance intra-groupe quantifie la variation des observations à l'intérieur des groupes. Elle permet de savoir si les éléments à l'intérieur d'un même groupe sont proches de leur moyenne (faible variance intra-groupe) ou bien si elles sont dispersées (variance intra-groupe élevée).

**Calcul :**

La variance intra-groupe est calculée en prenant la somme des carrés des différences entre chaque valeur individuelle et la moyenne de son groupe.

$$SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} n_i (X_{ij} - \bar{X}_i)^2$$

Où :

k : nombre de groupes

$n_i$  : taille de l'échantillon du groupe i

$X_{ij}$ : valeur individuelle dans le groupe ij

$\bar{X}_i$  : moyenne du groupe i

Ensuite, la **variance intra-groupe (MSW)** est obtenue en divisant cette somme par les degrés de liberté associés (en général  $N-k$ , où N est le nombre total d'observations et k le nombre de groupes) :

$$MSW = \frac{SSW}{N - k}$$

Interprétation :

- Une **variance intra-groupe faible** signifie que les individus au sein d'un même groupe sont similaires entre eux, avec des valeurs proches de la moyenne du groupe.
- Une **variance intra-groupe élevée** indique une plus grande dispersion à l'intérieur des groupes, avec des observations plus éloignées de la moyenne du groupe.

#### 4.2.3.3 Rapport de corrélation ( $\eta^2$ )

Le rapport de corrélation est une mesure de l'effet de la variable qualitative sur la variable quantitative. Il permet de quantifier la proportion de la variance totale de la variable quantitative qui peut être expliquée par les différences entre les groupes définis par la variable qualitative.

**Calcul :** Le rapport de corrélation est calculé en divisant la somme des carrés entre les groupes (SSB) par la somme des carrés totale (SST). La formule est la suivante :

$$\eta^2 = \frac{SSB}{SST}$$

où :

- SSB est la somme des carrés entre les groupes, qui mesure la variabilité des moyennes des groupes par rapport à la moyenne globale.
- SST est la somme des carrés totale, qui mesure la variabilité totale des valeurs de la variable quantitative.

Interprétation :

Un  $\eta^2$  élevé indique que la variable qualitative explique une grande partie de la variance de la variable quantitative. En d'autres termes, plus le  $\eta^2$  est élevé, plus les différences entre les groupes définis par la variable qualitative ont un impact significatif sur la variable quantitative.

- $\eta^2$  varie entre 0 et 1.
- Plus  $\eta^2$  est proche de 1, plus la variable explicative (facteur) explique une grande part de la variance de la variable dépendante.
- Si  $\eta^2$  est proche de 0, cela signifie que les différences entre les groupes sont faibles et que la variable explicative a peu d'effet.

#### **4.2.3.4. Mesure de liaison : le test de Fisher**

Le test de Fisher, également connu sous le nom de test de l'analyse de la variance (ANOVA) de Fisher, est un outil statistique utilisé pour déterminer si les différences observées entre les groupes sont statistiquement significatives. Ce test compare la variance inter-groupe à la variance intra-groupe à l'aide du rapport de Fisher ( $F$ ).

Le test de Fisher vise à vérifier l'hypothèse nulle selon laquelle les moyennes des différents groupes sont égales. Autrement dit, il cherche à déterminer si les différences entre les groupes sont dues au hasard ou à des facteurs réels.

#### **Rapport F :**

Le rapport de Fisher ( $F$ ) est calculé en divisant la moyenne des carrés entre les groupes (MSB) par la moyenne des carrés intra-groupes (MSW). La formule est la suivante :

$$F = \frac{MSB}{MSW}$$

où :

- MSB est la moyenne des carrés entre les groupes.
- MSW est la moyenne des carrés intra-groupes.

#### **Interprétation du rapport F :**

- Un rapport  $F$  élevé indique que la variance inter-groupe est grande par rapport à la variance intra-groupe, ce qui suggère que les différences entre les groupes sont significatives.
- Si le rapport  $F$  est proche de 1, cela suggère que les différences entre les groupes ne sont pas significatives et sont probablement dues au hasard.

**Table de Fisher :** Pour déterminer si le rapport  $F$  est statistiquement significatif, on compare la valeur calculée du rapport  $F$  à une valeur critique obtenue à partir de la table de Fisher (ou table  $F$ ). La valeur critique dépend du niveau de signification ( $\alpha$ ) choisi et des degrés de liberté.

#### **Hypothèses :**

- **Hypothèse nulle ( $H_0$ )** : Les moyennes des différents groupes sont égales.
- **Hypothèse alternative ( $H_1$ )** : Au moins une des moyennes des groupes est différente.

#### **Étapes du test de Fisher :**

1. Calculer la somme des carrés totale (SST), la somme des carrés entre les groupes (SSB) et la somme des carrés intra-groupes (SSW).
2. Calculer les moyennes des carrés entre les groupes (MSB) et intra-groupes (MSW).

3. Calculer le rapport de Fisher ( $F$ ) en divisant MSB par MSW.
4. Comparer la valeur de  $F$  obtenue à la valeur critique de la table de Fisher pour déterminer si les différences sont significatives.