# Notes de cours - Statistique Descriptive

Licence Mathématiques et informatique appliquées aux sciences humaines et sociales

Université de Bordeaux
UE : Bases en statistiques - Première année Licence MIASHS

**Rédaction** : Brigitte Patouille et Jérôme Poix

**Enseignants impliqués dans l'UE**
Jérémie Bigot, Marie Chavent, Vincent Couallier, Brigitte Patouille

2016-2017

# Table des matières

# 1 Les données

Les **données statistiques** se présentent sous la forme d'**individus** pour lesquels sont mesurés un certain nombre de **caractères** (ou **"variables statistiques**"). L'ensemble des individus constitue un **échantillon** (ou encore une **série statistique**), formant ainsi un sous-ensemble d'un groupe (beaucoup) plus grand appelé **"population"**.
Les caractères statistiques peuvent être de plusieurs natures :

• Les **variables qualitatives** :
Exemples : nationalité d'une personne, profession.
Les valeurs possibles d'une variable qualitative sont appelées les **modalités** ;
Exemples :"Français", "Britannique", "Professeur", "Médecin".

Remarque : Certaines variables qualitatives peuvent parfois comporter un ordre "naturel" ;
Exemple : variable "activité sportive" avec les modalités "peu sportif", "assez sportif", "très sportif".

• Les **variables quantitatives** (qu'on peut mesurer) ;
– Les **variables quantitatives discrètes** : variables prenant leurs valeurs dans un ensemble fini ou dénombrable (en général $\mathbb{N}$) ;
Exemples : âge d'une personne (en années), nombre d'enfants dans une famille.
Pour ces variables, les valeurs possibles sont également appelées **modalités**.
– Les **variables quantitatives continues** : variables qui peuvent prendre **n'importe quelle valeur** dans un intervalle de $\mathbb{R}$, c'est à dire mesurées avec une très grande précision.
Exemples : taille d'un individu (en mm), poids (en g), temps d'une réaction chimique (en micro-secondes).

**En résumé :**
Soit un **caractère statistique** $X$.
Un **échantillon** (ou **série statistique**) pour ce caractère peut se représenter par la donnée d'une suite de nombres :

$$\{x_1, x_2, \ldots, x_n\},$$

où :
— $n$ est le nombre d'observations (ou encore la **taille de l'échantillon**)
— et $\forall i$, $1 \leq i \leq n$, $x_i$ représente la valeur du caractère $X$ pour l'individu $i$.
Les données présentées sous cette forme sont appelées les **données brutes**.
On peut également les représenter sous la forme d'un tableau :

| Individu | X |
|---|---|
| 1 | $x_1$ |
| 2 | $x_2$ |
| $\vdots$ | $\vdots$ |
| n | $x_n$ |

Exemple de tableau de données à 4 individus et 3 variables :

| | Sexe | Nombre de frères et soeurs | Taille(cm) |
|---|---|---|---|
| 1 | M | 1 | 143 |
| 2 | M | 2 | 149 |
| 3 | F | 0 | 144 |
| 4 | F | 2 | 146 |

Pour un tableau de données à $n$ individus et $k$ variables, on peut étudier les données de plusieurs façons :

- **1 variable à la fois** → statistique descriptive à 1 variable, étude d'un caractère statistique ;
- **2 variables à la fois** → statistique descriptive à 2 variables, étude **simultanée** de 2 caractères statistiques ;
- **+ de 2 variables à la fois** → statistique exploratoire, analyse des données (méthodes plus lourdes et plus complexes, calculs intensifs).

# 2 Statistique descriptive à une variable

## 2.1 Distribution statistique associée à un échantillon

• **Variables qualitatives ou quantitatives discrètes.**

**Définition 1 :** Une distribution statistique d'un échantillon de $n$ observations d'une variable qualitative ou quantitative discrète $X$ (ou bien d'un caractère qualitatif ou quantitatif discret) est constituée par la donnée d'un regroupement

$$\{(a_1, n_1), (a_2, n_2), \ldots, (a_p, n_p)\},$$

où :
– les $a_i$, $1 \leq i \leq p$ représentent les modalités de la variable statistique $X$ (classées le plus souvent en ordre croissant, $a_1 < a_2 < \cdots < a_p$), $p$ étant le nombre de modalités ;
– le nombre $n_i$, $1 \leq i \leq p$, appelé **effectif de la modalité** représente le nombre d'individus pour lesquels la variable $X$ a la modalité $a_i$, $1 \leq i \leq p$.
– $\sum_{i=1}^{p} n_i = n$

On peut représenter une distribution statistique discrète par un tableau :

| X(modalités) | $a_1$ | $a_2$ | ... | $a_p$ |
|---|---|---|---|---|
| Effectif | $n_1$ | $n_2$ | ... | $n_p$ |

Exemple :

**Données brutes**

| $i$ | Âge $x_i$ |
|---|---|
| 1 | 20 |
| 2 | 18 |
| 3 | 18 |
| 4 | 19 |
| 5 | 17 |
| 6 | 18 |
| 7 | 22 |
| 8 | 18 |
| 9 | 20 |
| 10 | 18 |

**Distribution statistique**
(**Données regroupées** ou **"triées à plat"**)

| Âge | Effectif |
|---|---|
| | |

• **Variables quantitatives continues : deux approches possibles.**

**Première approche :** On peut considérer un échantillon d'un caractère continu comme celui d'un caractère discret à valeur dans un sous-ensemble fini de $\mathbb{R}$.
→ même définition de la distribution que dans le cas discret.

| X(modalités) | $a_1$ | $a_2$ | ... | $a_p$ |
|---|---|---|---|---|
| Effectif | $n_1$ | $n_2$ | ... | $n_p$ |

On appellera cette distribution la **"distribution empirique"** du caractère $X$.
Mais si l'hypothèse de continuité du caractère est pertinente, alors $p \simeq n$, et (presque) tous les effectifs $n_i$ seront égaux à 1. On se retrouve donc avec les données brutes dans ce cas.
$\rightarrow$ Les "modalités" sont les différentes valeurs prises par le caractère. **Elles dépendent ici des observations de l'échantillon.**

Remarque 1 : Cette définition de la distribution d'un caractère continu correspond seulement au classement dans l'ordre croissant des données et au regroupement éventuel des quelques observations qui pourraient être identiques.

**Seconde approche :** Pour obtenir une représentation synthétique des données, on a besoin d'effectuer un **regroupement en classes**.

**Définition 2 :** Une distribution statistique d'un échantillon de $n$ observations d'une variable quantitative continue (ou d'un caractère quantitatif continu) $X$ **regroupée en classes** est constituée par la donnée d'un regroupement

$$\{(C_1, n_1), (C_2, n_2), \ldots, (C_p, n_p)\},$$

où :

- les $C_i$, $1 \leq i \leq p$ sont des intervalles appelés **classes** et formant une partition de l'ensemble des valeurs possibles pour $X$, c'est à dire

$$C_1 = [a_0, a_1[,\ C_2 = [a_1, a_2[, \ldots,\ C_p = [a_{p-1}, a_p[,$$

ou bien

$$C_1 = ]a_0, a_1],\ C_2 = ]a_1, a_2], \ldots,\ C_p = ]a_{p-1}, a_p],$$

avec $a_0 < a_1 < \cdots < a_p$.

- le nombre $n_i$, $1 \leq i \leq p$, appelé **effectif de la classe** représente le nombre d'individus pour lesquels la valeur de la variable $X$ est dans la classe $C_i$, $1 \leq i \leq p$.

On peut représenter une distribution statistique continue par un tableau :

| X(classes) | $]a_0, a_1]$ | $]a_1, a_2]$ | ... | $]a_{p-1}, a_p]$ |
|---|---|---|---|---|
| Effectif | $n_1$ | $n_2$ | ... | $n_p$ |

Remarque 2 : On a bien sûr pour un échantillon de taille $n$,

$$n_1 + n_2 + \cdots + n_p = n.$$

Remarque 3 : Quand on passe des données brutes à la distribution, il y a perte d'information, surtout dans le cas des variables continues, puisqu'on ne connaît plus les valeurs effectivement prises par la variable.

**En résumé :**

Echantillon
(données brutes)
$\downarrow$
Distribution
$\swarrow$ $\qquad\qquad$ $\searrow$

Regroupement
en modalités
(qual. ou quant. discret)

Regroupement
en classes
(quant. continu)

**Exemples de distributions statistiques :**

1. Groupe sanguin de 100 personnes (variable ..

| Groupe sanguin | A | O | B | AB |
|---|---|---|---|---|
| Effectif | 40 | 33 | 14 | 13 |

2. Nombre d'enfants dans 80 familles (

| Nb d'enfants | Effectif |
|---|---|
| 0 | 6 |
| 1 | 20 |
| 2 | 27 |
| 3 | 17 |
| 4 | 6 |
| 5 | 3 |
| 6 | 1 |

3. Taille (en cm) de 24 individus (

| Taille | Effectif | Taille | Effectif | Taille | Effectif |
|---|---|---|---|---|---|
| 154 | 1 | 171 | 1 | 182 | 1 |
| 156 | 1 | 174 | 1 | 184 | 1 |
| 157 | 1 | 176 | 2 | 185 | 1 |
| 160 | 1 | 177 | 1 | 186 | 1 |
| 163 | 1 | 179 | 1 | 188 | 1 |
| 167 | 2 | 180 | 2 | 191 | 1 |
| 170 | 1 | 181 | 1 | 194 | 1 |

4. Poids (en kg) de 92 personnes (

| Classe | $]45,55]$ | $]55,65]$ | $]65,75]$ | $]75,85]$ | $]85,95]$ | $]95,105]$ |
|---|---|---|---|---|---|---|
| Effectif | 7 | 20 | 40 | 18 | 6 | 1 |

Avec Regroupement/Découpage de classes :

| Classe | $]45,55]$ | $]55,65]$ | $]65,70]$ | $]70,75]$ | $]75,85]$ | $]85,105]$ |
|---|---|---|---|---|---|---|
| Effectif | 7 | 20 | 19 | 21 | 18 | 7 |

REMARQUE 4 : Dans le cas continu, la distribution statistique dépend du choix des classes, qui est arbitraire.
En pratique, les classes seront choisies de manière "raisonnable" :
• en fonction du nombre d'observations.
• en fonction de l'amplitude de l'échantillon = +grande valeur - (+petite valeur).
• pas trop petites (sinon à la limite $n_i = 0$ ou $n_i = 1$, $\forall i$ : ce n'est plus un regroupement en classes !)
• pas trop grandes (sinon à la limite il ne reste qu'une ou deux classes : plus aucune information sur les données).
• On choisit souvent des classes de même amplitude, mais il est fréquent de regrouper des classes d'effectifs trop faibles, et de couper des classes d'effectifs trop grands.

## 2.2 Fréquence, effectif cumulé et fréquence cumulée

**Définition 3 :** Soit $n = n_1 + \cdots + n_p$ l'effectif total d'un échantillon.
• On appelle **fréquence** de la modalité $a_i$ (ou de la classe $C_i$) le nombre

$$f_i = \frac{n_i}{n}$$

$f_i$ est la proportion des sujets de l'échantillon qui ont une modalité de $X$ égale à $a_i$ (ou qui sont dans la classe $C_i$).

• Si $X$ est quantitative ou qualitative ordinale et que les $a_i$ sont classés par ordre croissant ($a_1 < a_2 < \cdots < a_p$), on appelle **effectif cumulé** de la modalité $a_i$ (ou de la classe $C_i$) le nombre :

$$N_i = n_1 + \cdots + n_i = \sum_{j=1}^{i} n_j$$

$N_i = n_1 + \cdots + n_i$ est le nombre de sujets de l'échantillon qui ont une modalité de $X$ inférieure ou égale à $a_i$.

• On appelle **fréquence cumulée** de la modalité $a_i$ (ou de la classe $C_i$) le nombre

$$F_i = \frac{N_i}{n} = f_1 + \cdots + f_i = \sum_{j=1}^{i} f_j.$$

$F_i$ est la proportion des sujets de l'échantillon qui ont une modalité de $X$ inférieure ou égale à $a_i$.

Ces deux dernières quantités n'ont pas de sens pour les variables purement qualitatives.

Remarque 5 : Le nombre $f_i$ calculé sur un échantillon de taille $n$ est une estimation de la fréquence $p_i$ de la modalité $a_i$ (ou de la classe $C_i$) pour l'ensemble de la population. On appelle $p_i$ la "fréquence théorique" ou encore probabilité de la modalité ou classe en question.
On peut montrer sous certaines conditions que (loi dite des "grands nombres")

$$\lim_{n \to +\infty} f_i = p_i$$

## 2.3 Représentations graphiques

Elles représentent
• les effectifs, les fréquences (tout type de variables),
• les effectifs cumulés, les fréquences cumulées (variables quantitatives ou ordinales),
• les valeurs d'un caractère statistique pour chaque individu.

### 2.3.1 Diagramme circulaire (camembert)

Cette représentation graphique est à réserver aux variables (purement) qualitatives.
On divise le disque en $p$ secteurs représentant les modalités $(a_1, a_2, \ldots, a_p)$ et proportionnels aux effectifs correpondants $(n_1, n_2, \ldots, n_p)$. L'angle (en degrés) du secteur représentant la modalité $a_i$ sera égal à :

$$\alpha_i = \frac{n_i}{n} \times 360$$

**Exemple :** Diagramme circulaire de la variable Groupe sanguin.

### 2.3.2 Diagramme des effectifs et des fréquences

• *Cas discret* : modalités $a_i$, effectifs $n_i$.
Sur l'axe des abscisses, on reporte les modalités $a_i$ ; sur l'axe des ordonnées on reporte les effectifs ou les fréquences. Au-dessus de chaque modalité, on trace un segment vertical dont la hauteur est égale (ou proportionnelle) à la fréquence (à l'effectif) associé.
Ce diagramme est aussi appelé **Diagramme en bâtons**.

**Exemple :** Diagramme en bâtons de la variable Âge.

• *Cas continu* : classes $C_i$, effectifs $n_i$.
Sur l'axe des abscisses, on reporte les classes $C_i$ ; sur l'axe des ordonnées on reporte les effectifs ou les fréquences. Au-dessus de chaque classe, on trace un rectangle dont l'**aire** est égale (ou proportionnelle) à la fréquence (à l'effectif) associé. Ce diagramme est aussi appelé **Histogramme**.

Remarque 6 : La hauteur du rectangle représente la **densité d'effectif** ou la **densité de fréquence** d'une classe.
→ Si $a$ représente l'amplitude de référence, alors la hauteur $h_i$ du rectangle correspondant à la classe $C_i = ]a_{i-1}, a_i]$ vaut

$$h_i = \frac{n_i \times a}{a_i - a_{i-1}}.$$

→ Pour une classe d'amplitude k fois plus grande (resp. k fois plus petite), on divise (resp. on multiplie) l'effectif ou la fréquence par un facteur k pour obtenir la hauteur du rectangle.

**Exemple :** Histogramme de la variable Poids.

### 2.3.3 Diagrammes d'effectifs et de fréquences cumulés

• **Variables quantitatives discrètes.**

**Définition 4 :** On appelle **fonction de répartition empirique ou observée**, ou encore diagramme des fréquences cumulées d'un échantillon $\{x_1, \ldots, x_n\}$ (ou d'une distribution statistique $\{(a_1, n_1), (a_2, n_2), \ldots, (a_p, n_p)\}$) la fonction définie $\forall x \in \mathbb{R}$ par :

$$F(x) = \frac{\text{Nombre d'observations} \leq x}{n},$$

c'est-à-dire :

$$F(x) = \begin{cases} 0, & \forall x < a_1, \\ F_i, & \forall x \in [a_i, a_{i+1}[, \\ 1, & \forall x \geq a_p. \end{cases}$$

REMARQUE 7 : Dans le cas d'un caractère $X$ discret, la longueur des marches de la fonction de répartition est fixée au départ (en général elle vaut 1), mais les sauts sont de hauteurs variables.

**Exemple :** Fonction de répartition du caractère Nombre d'enfants

• **Variables quantitatives continues.**

**a. Fonction de répartition empirique.**

En général, le terme de fonction de répartition empirique d'un échantillon $\{x_1, \ldots, x_n\}$ désigne toujours la fonction en escalier

$$F(x) = \frac{\text{Nombre d'observations} \leq x}{n}.$$

Dans le cas d'une variable $X$ continue, on ne peut tracer cette fonction que si on connaît le détail des données (données brutes), ce qui permet de considérer $X$ comme un caractère discret à valeurs dans un sous-ensemble fini de $\mathbb{R}$. Cette fonction correspond donc bien à la **fonction de répartition de la "distribution empirique"** du caractère $X$.

**Exemple :**

| Ind. | Taille(cm) |
|---|---|
| 1 | 149,5 |
| 2 | 157,3 |
| 3 | 172 |
| 4 | 168,2 |

Quand $X$ est continu, alors en général toutes les observations de ce caractère ont des valeurs différentes. Dans ce cas les sauts de $F(x)$ seront tous de $1/n$ ; en revanche **la longueur des marches est variable et dépend des observations**.

**Exemple :** Fonction de répartition empirique du caractère continu Taille

**b. Fonction de répartition "Polygone des fréquences (ou des effectifs) cumulé(e)s".**

Soit une distribution d'un caractère continu $X$ :

| X(classes) | $]a_0, a_1]$ | $]a_1, a_2]$ | ... | $]a_{p-1}, a_p]$ |
|---|---|---|---|---|
| Effectif | $n_1$ | $n_2$ | ... | $n_p$ |

Si on définit $F(x)$ comme précédemment, c'est-à-dire :

$$F(x) = \frac{\text{Nombre d'observations} \leq x}{n},$$

on voit qu'on ne connaît a priori les valeurs de $F(x)$ que pour $x = a_0, a_1, \ldots, a_p$. En effet :

$$\begin{cases} F(a_0) = 0, \\ F(a_i) = F_i = \frac{N_i}{n}, \\ F(a_p) = 1, \end{cases}$$

et d'autre part $F(x) = 0,\ x < a_0$ et $F(x) = 1,\ x > a_p$.

Comment définir $F(x)$, $\forall x \in ]a_{i-1}, a_i[,\ 1 \leq i \leq p$ ?
On suppose en fait que les observations sont uniformément réparties à l'intérieur de chaque classe et on peut alors définir $F(x)$ comme une droite sur chaque intervalle $]a_{i-1}, a_i[$. Cela revient donc à relier les points $M_i(a_i, F_i)$ par des segments de droite.

Ainsi, $\forall x \in ]a_{i-1}, a_i[$,

$$\begin{aligned} F(x) &= F(a_{i-1}) + (x - a_{i-1})\tfrac{F(a_i) - F(a_{i-1})}{a_i - a_{i-1}} \\ &= F_{i-1} + (x - a_{i-1})\tfrac{f_i}{a_i - a_{i-1}}. \end{aligned}$$

Le graphique ainsi obtenu est appelé **polygone des fréquences cumulées**. On peut aussi définir la fonction $G(x) = nF(x)$ dont la représentation graphique s'appelle le **polygone des effectifs cumulés**.

**Exemple :** Polygone des fréquences cumulées du caractère Poids

## 2.4 Paramètres caractérisant une variable statistique

### 2.4.1 Paramètres (ou indicateurs) de position (ou de centralité)

• **La moyenne (arithmétique)**

**a. Calcul à partir des données de l'échantillon (données brutes).**

**Définition 5 :** Soit un échantillon d'un caractère statistique $X$ : $\{x_1, \ldots, x_n\}$.
On appelle **moyenne** (arithmétique) de l'échantillon le nombre :

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

**Exemple :** Moyenne de Âge (échantillon) :

**b. Calcul à partir d'une distribution.**

**Définition 6 :**
– **cas discret** : Soit une distribution d'un caractère statistique quantitatif discret $X$ :

$$\{(a_1, n_1), (a_2, n_2), \ldots, (a_p, n_p)\}.$$

On appelle **moyenne** (arithmétique) de la distribution le nombre :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} n_i a_i = \sum_{i=1}^{p} f_i a_i.$$

**Exemple :** Moyenne de Âge (distribution) :

– **cas continu** : Soit une distribution d'un caractère statistique quantitatif continu $X$,

$$\{(C_1, n_1), (C_2, n_2), \ldots, (C_p, n_p)\}.$$

On appelle **moyenne** (arithmétique) de la distribution le nombre

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} n_i c_i = \sum_{i=1}^{p} f_i c_i,$$

où $\forall\ 1 \leq i \leq p$, $c_i$ est le centre de la classe $C_i = ]a_{i-1}, a_i]$, autrement dit : $c_i = \dfrac{a_{i-1} + a_i}{2}$.

**Exemple :** Moyenne du Poids (distribution avec regroupement et découpage) :

Remarque 8 : La moyenne vérifie la **propriété de linéarité**. étant donné deux caractères statistiques $X$ et $Y$ de moyennes respectives $\bar{x}$ et $\bar{y}$, et deux nombres réels $a$ et $b$, alors la moyenne du

caractère $Z = aX + bY$ (c'est à dire de l'échantillon formé des observations $z_i = ax_i + by_i$, $1 \leq i \leq n$) vaut

$$\bar{z} = a\bar{x} + b\bar{y}.$$

REMARQUE 9 : La moyenne calculée à partir de la distribution est appelée **"moyenne pondérée"** (par les effectifs ou les fréquences).

REMARQUE 10 : Dans la cas d'un caractère discret la moyenne calculée sur les données brutes est la même que celle calculée sur la distribution, en revanche ces deux quantités diffèrent dans le cas d'un caractère continu.

| | Données Brutes | Distribution |
|---|---|---|
| $X$ discret | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ | $= \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i a_i$ |
| $X$ continu | $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ | $\neq \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i c_i$ |

- **LA MÉDIANE**

La médiane est le nombre tel que la moitié des valeurs de l'échantillon lui sont inférieures ou égales (et par conséquent la moitié des valeurs de l'échantillon lui sont supérieures). Elle divise l'échantillon en deux parties égales (en nombre de valeurs).

**Définition 7 :**
– **cas discret** : Soit une distribution d'un caractère statistique quantitatif discret $X$ :

$$\{(a_1, n_1), (a_2, n_2), \ldots, (a_p, n_p)\},$$

de fonction de répartition $F$. On appelle **médiane** de la distribution la modalité $\tilde{x} = a_i$ telle que

$$\begin{array}{llll} F(a_{i-1}) & = F_{i-1} & = f_1 + \cdots + f_{i-1} & < 0,5, \\ F(a_i) & = F_i & = f_1 + \cdots + f_i & > 0,5. \end{array}$$

Dans le cas où il existe $a_i$ tel que $F(a_i) = F_i = 0,5$, alors on posera

$$\tilde{x} = \frac{a_i + a_{i+1}}{2}.$$

– **cas continu** : Soit une distribution d'un caractère statistique quantitatif continu $X$,

$$\{(C_1, n_1), (C_2, n_2), \ldots, (C_p, n_p)\},$$

de fonction de répartition $F$. On appelle **médiane** de la distribution le nombre $\tilde{x}$ tel que $F(\tilde{x}) = 0,5$.

**Détermination pratique de la médiane d'une variable quantitative (discrète ou continue) à partir des données brutes (échantillon) :**

On commence par ordonner toutes les valeurs $x_i$ de l'échantillon en les répétant autant de fois qu'elles sont observées. Alors :

- si $n$ est pair ($n = 2k$) : $\tilde{x} = \dfrac{x_{n/2} + x_{n/2+1}}{2} = \dfrac{x_k + x_{k+1}}{2}$ ,
- si $n$ est impair ($n = 2k + 1$) : $\tilde{x} = x_{(n+1)/2} = x_{k+1}$.

**Exemples :**

**Détermination grahique de la médiane d'une variable quantitative (discrète ou continue) :**
– **cas discret** : on utilise la fonction de répartition empirique.

– **cas continu** : on utilise le polygone des fréquences cumulées.

**1er cas :** $\exists i$ tel que $F(a_i) = 0,5 \Rightarrow \tilde{x} = a_i$.

**2nd cas :** $\exists i$ tel que $F(a_{i-1}) < 0,5,\ F(a_i) > 0,5$ d'où $\tilde{x} \in ]a_{i-1}, a_i[$.
On calcule alors la médiane par interpolation linéaire :
Sur $]a_{i-1}, a_i[$ la fonction de répartition $F$ est une droite d'équation :

$$\begin{aligned} F(x) &= F(a_{i-1}) + (x - a_{i-1})\frac{F(a_i) - F(a_{i-1})}{a_i - a_{i-1}} \\ &= F(a_{i-1}) + (x - a_{i-1})\frac{f_i}{a_i - a_{i-1}}. \end{aligned}$$

Pour trouver la valeur $\tilde{x}$ telle que $F(\tilde{x}) = 0,5$ , il suffit de remplacer $F(x)$ par $0,5$ et $x$ par $\tilde{x}$ dans l'équation ci-dessus :

$$\begin{aligned} 0,5 &= F(a_{i-1}) + (\tilde{x} - a_{i-1})\frac{f_i}{a_i - a_{i-1}} \\ \Leftrightarrow \tilde{x} &= \frac{0,5 - F(a_{i-1})}{f_i}(a_i - a_{i-1}) + a_{i-1}. \end{aligned}$$

On peut aussi écrire en posant $G(x) = nF(x)$,

$$G(x) = G(a_{i-1}) + (x - a_{i-1})\frac{n_i}{a_i - a_{i-1}},$$

d'où finalement comme $F(\tilde{x}) = 0,5 \Leftrightarrow G(\tilde{x}) = n/2$,

$$\tilde{x} = \frac{n/2 - G(a_{i-1})}{n_i}(a_i - a_{i-1}) + a_{i-1}.$$

**Exemples :**

*** Généralisation : les quantiles.**

**Définition 8 :**
– **cas discret** : Soit une distribution d'un caractère statistique quantitatif discret $X$ :

$$\{(a_1, n_1), (a_2, n_2), \ldots, (a_p, n_p)\},$$

de fonction de répartition $F$, et soit $\alpha \in ]0, 1[$. On appelle **quantile** d'ordre $\alpha$ de la distribution la modalité $q_\alpha = a_i$ telle que

$$\begin{array}{lll} F(a_{i-1}) & = F_{i-1} & = f_1 + \cdots + f_{i-1} \quad < \alpha, \\ F(a_i) & = F_i & = f_1 + \cdots + f_i \quad > \alpha. \end{array}$$

Dans le cas où il existe $a_i$ tel que $F(a_i) = F_i = \alpha$, alors on posera :

$$q_\alpha = (1 - \alpha)a_i + \alpha a_{i+1}.$$

– **cas continu** : Soit une distribution d'un caractère statistique quantitatif continu $X$ :

$$\{(C_1, n_1), (C_2, n_2), \ldots, (C_p, n_p)\},$$

de fonction de répartition $F$. On appelle **quantile** d'ordre $\alpha$ de la distribution le nombre $q_\alpha$ tel que $F(q_\alpha) = \alpha$.

**Définition 9 :** On appelle **premier quartile** d'une distribution statistique et on notera $Q_1$ le quantile d'ordre $\alpha = 0,25$ ($Q_1 = q_{0,25}$).
On appelle **troisième quartile** d'une distribution statistique et on notera $Q_3$ le quantile d'ordre $\alpha = 0,75$ ($Q_3 = q_{0,75}$).

Remarque 11 : La médiane n'est autre que le deuxième quartile ($\tilde{x} = Q_2 = q_{0,5}$) et on a bien sûr :

$$Q_1 \leq \tilde{x} \leq Q_3.$$

**Exemples :**

• **Le mode**

**Définition 10 :** On appelle **mode** (cas qualitatif ou quantitatif discret, noté $\hat{x}$) ou **classe modale** (cas quantitatif continu, notée $\hat{C}$) d'une distribution statistique **la modalité ou classe de plus grand effectif**.

**Exemples :**

Remarque 12 : S'il existe un seul mode, on dit que la distribution statistique est **unimodale**. Dans le cas contraire on parle de distribution **multimodale** (bimodale, trimodale,...)
Dans le cas d'un caractère continu, on peut prendre comme mode le centre de la classe modale.

*** Avantages et inconvénients des différents paramètres de position :**

| | Moyenne | Médiane | Mode |
|---|---|---|---|
| + | *toujours calculable<br>*calculs faciles<br>(linéarité) | *toujours définie<br>*peu sensible aux<br>valeurs extrêmes | *permet de détecter<br>plusieurs "pics"<br>dans une distribution |
| - | *sensible aux valeurs<br>extrêmes | *pas de linéarité | *pas toujours<br>bien défini |

### 2.4.2 Paramètres de dispersion

Ils mesurent la tendance des données à s'étaler ou au contraire à se concentrer autour d'une valeur centrale (donnée par un paramètre de position).

• **L'étendue.**

**Définition 11 :** On appelle **étendue** d'une série statistique d'un caractère $X$ la différence entre la plus grande valeur de la série pour ce caractère et la plus petite.

Remarque 13 : Cette quantité est très sensible aux valeurs aberrantes ou extrêmes.

• **L'écart inter-quartiles.**

**Définition 12 :** On appelle **écart inter-quartiles** d'une série statistique d'un caractère $X$ la différence entre le troisième quartile de la série pour ce caractère et le premier quartile, autrement dit le nombre noté $q$ donné par

$$q = Q_3 - Q_1.$$

Remarque 14 : Cette quantité est peu sensible aux valeurs aberrantes ou extrêmes.

• **La variance**

**a. Calcul à partir des données de l'échantillon (données brutes)**

**Définition 13 :** Soit un échantillon d'un caractère statistique $X$ : $\{x_1, \ldots, x_n\}$.
On appelle **variance** de l'échantillon le nombre

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

**b. Calcul à partir d'une distribution**

**Définition 14 :**
– **cas discret** : Soit une distribution d'un caractère statistique quantitatif discret $X$ :

$$\{(a_1, n_1), (a_2, n_2), \ldots, (a_p, n_p)\}.$$

On appelle **variance** de la distribution le nombre :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(a_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(a_i - \bar{x})^2.$$

– **cas continu** : Soit une distribution d'un caractère statistique quantitatif continu $X$ :

$$\{(C_1, n_1), (C_2, n_2), \ldots, (C_p, n_p)\}.$$

On appelle **variance** de la distribution le nombre :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{p} n_i(c_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(c_i - \bar{x})^2,$$

où $\forall 1 \leq i \leq p$, $c_i$ est le centre de la classe $C_i = ]a_{i-1}, a_i]$, et $\bar{x} = \bar{c}$ est la moyenne des centres des classes.

Remarque 15 : La variance est la moyenne des carrés des écarts à la moyenne. Son unité de mesure est le carré de celle du caractère $X$.

• **L'écart-type.**

**Définition 15 :** On appelle **écart-type** le nombre défini comme la racine carrée de la variance, c'est-à-dire :

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ (échantillon)},$$

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{p} n_i(a_i - \bar{x})^2} \text{ (distribution)}.$$

Remarque 16 : L'écart-type s'exprime dans la même unité de mesure que le caractère $X$.

### 2.4.3 Propriétés de la variance

• **Formule de calcul de la variance dite de "Koenig" :**

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \overline{x^2} - \bar{x}^2 \text{ (échantillon)},$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{p} n_i a_i^2 - \left(\frac{1}{n}\sum_{i=1}^{p} n_i a_i\right)^2 \text{ (distribution)}.$$

REMARQUE 17 : Si $s_x^2 = 0$ alors le caractère $X$ ne prend qu'une seule valeur dans l'échantillon considéré (il est constant).

REMARQUE 18 : Soit $X$ un caractère statistique de variance $s_x^2$. On définit pour $a \in \mathbb{R}$ le caractère $Y = aX$ (c'est-à-dire l'échantillon formé des observations $y_i = ax_i$, $1 \leq i \leq n$). Alors

$$s_y^2 = a^2 s_x^2 \ \ (\Leftrightarrow s_y = |a| s_x).$$

**Exemples :** Calculer la variance et l'écart-type

1. pour la distribution du nombre d'enfants
2. pour la distribution des poids (données initiales)

• **Calcul de la moyenne et de la variance pour des échantillons regroupés**

Soient $k$ échantillons d'une même variable quantitative $X$ :

Echantillon 1 : $\{x_1^1, x_2^1, \ldots, x_{n_1}^1\}$ de taille $n_1$, de moyenne $\bar{x}^1$ et de variance $s_1^2$,

Echantillon 2 : $\{x_1^2, x_2^2, \ldots, x_{n_2}^2\}$ de taille $n_2$, de moyenne $\bar{x}^2$ et de variance $s_2^2$,

$\vdots$

Echantillon k : $\{x_1^k, x_2^k, \ldots, x_{n_k}^k\}$ de taille $n_k$, de moyenne $\bar{x}^k$ et de variance $s_k^2$,

et soit $n$ le nombre total de sujets :

$$n = n_1 + n_2 + \cdots + n_k = \sum_{i=1}^{k} n_i.$$

Notons $\bar{x}$ la moyenne de l'échantillon regroupé et $s_x^2$ sa variance.
On a alors les relations suivantes :

**a. Moyenne de l'échantillon regroupé**

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

La moyenne globale est donc égale à la moyenne des moyennes partielles.

**b. Variance de l'échantillon regroupé**

$$s_x^2 = \underbrace{\frac{1}{n} \sum_{i=1}^{k} n_i s_i^2}_{(1)} + \underbrace{\frac{1}{n} \sum_{i=1}^{k} n_i (\bar{x}^i - \bar{x})^2}_{(2)}$$

La variance globale est donc égale à la moyenne des variances partielles (1) + la variance des moyennes partielles (2).

**Exemple :**
On mesure la taille d'enfants de maternelle.
L'échantillon disponible est composé d'une part de 35 filles de taille moyenne $\bar{x}_F = 1.21\text{m}$ et de variance $s_F^2 = 0.25\text{m}^2$, et d'autre part de 41 garçons de taille moyenne $\bar{x}_G = 1.24\text{m}$ et de variance $s_G^2 = 0.3\text{m}^2$.
Calculer la taille moyenne de ce groupe d'enfants ainsi que la variance associée.

# 3 Statistique descriptive à deux variables

Dans le cas de l'étude **simultanée** de deux caractères statistiques $X$ et $Y$, chaque observation est caractérisée simultanément par sa valeur pour $X$ et sa valeur pour $Y$, c'est-à-dire par un couple de réels :

$$(x_k, y_k),\ 1 \leq k \leq n,$$

où $n$ est la taille de l'échantillon et $k$ le numéro de l'observation.
Cette étude comporte non seulement l'étude du caractère $X$ et celle du caractère $Y$ (2 études à une variable) mais également l'étude de la dépendance, des relations possibles entre les 2 variables.

**Exemple :** $X$ = Taille et $Y$ = Poids. Existe-t-il une relation, et si oui de quelle nature, entre ces deux variables ?

## 3.1 Distribution statistique d'un couple de variables

**Définition 1 :** Soient $X$ et $Y$ deux variables statistiques.
Un échantillon de ces deux variables est donné par l'ensemble des couples $(x_k, y_k)$ pour $1 \leq k \leq n$. Soit $\{a_1, \ldots, a_p\}$ l'ensemble des modalités dans lequel les $x_k$ prennent leur valeur et soit $\{b_1, \ldots, b_q\}$ l'ensemble des modalités dans lequel les $y_k$ prennent leur valeur.
On appelle **distribution statistique du couple** $(X, Y)$ le regroupement $\{(a_i, b_j), n_{ij}\}$ où :
- $a_1, \ldots, a_p$ sont les modalités du caractère $X$
- $b_1, \ldots, b_q$ sont les modalités du caractère $Y$
- $n_{ij}$ est **l'effectif croisé** de $(a_i, b_j)$, c'est-à-dire le nombre d'observations ayant la modalité $a_i$ pour $X$ et $b_j$ pour $Y$.

Une telle distribution se représente par un **tableau croisé** ou **tableau de contingence** :

| X\Y | $b_1$ | $\cdots$ | $b_j$ | $\cdots$ | $b_q$ | Total |
|---|---|---|---|---|---|---|
| $a_1$ | $n_{11}$ | $\cdots$ | $n_{1j}$ | $\cdots$ | $n_{1q}$ | $n_{1\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_i$ | $n_{i1}$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $n_{iq}$ | $n_{i\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_p$ | $n_{p1}$ | $\cdots$ | $n_{pj}$ | $\cdots$ | $n_{pq}$ | $n_{p\bullet}$ |
| Total | $n_{\bullet 1}$ | $\cdots$ | $n_{\bullet j}$ | $\cdots$ | $n_{\bullet q}$ | $n$ |

**Définition 2 :** Les nombres $n_{i\bullet}$ (resp. $n_{\bullet j}$) sont appelés les **effectifs marginaux** des modalités $a_i$ (resp. $b_j$), autrement dit les effectifs de la distribution de la variable $X$ (resp. $Y$) et sont définis par :

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \qquad \text{et} \qquad n_{\bullet j} = \sum_{i=1}^{p} n_{ij}.$$

On peut également définir :
– la **fréquence du couple** $(a_i, b_j)$ : $f_{ij} = \frac{n_{ij}}{n}$,
– la **fréquence marginale** de $a_i$ : $f_{i\bullet} = \frac{n_{i\bullet}}{n} = \sum_{j=1}^{q} f_{ij}$,
– la **fréquence marginale** de $b_j$ : $f_{\bullet j} = \frac{n_{\bullet j}}{n} = \sum_{i=1}^{p} f_{ij}$.

**Exemples :**

**1.** Soient $X$ = Statut tabagique de la mère et $Y$ = Catégorie de poids du bébé à la naissance :

| X\Y | Poids faible | Poids normal | Total |
|---|---|---|---|
| Fumeuse | 35 | 39 | 74 |
| Non fumeuse | 30 | 85 | 115 |
| Total | 65 | 124 | 189 |

**2.**

| Taille\Poids | $]50,60]$ | $]60,70]$ | $]70,80]$ | $]80,90]$ | $]90,100]$ | Total |
|---|---|---|---|---|---|---|
| $]150,160]$ | 2 | 1 | 0 | 0 | 0 | 3 |
| $]160,170]$ | 7 | 2 | 3 | 0 | 0 | 12 |
| $]170,180]$ | 1 | 4 | 9 | 2 | 1 | 17 |
| $]180,190]$ | 0 | 0 | 1 | 4 | 3 | 8 |
| Total | 10 | 7 | 13 | 6 | 4 | 40 |

## 3.2 Distributions conditionnelles

**Définition 3 :** On appelle **distribution conditionnelle de $Y$ sachant** $X = a_i$ la distribution :

$$\{(b_1, n_{i1}), \ldots, (b_q, n_{iq})\}.$$

→ Il s'agit donc de la distribution à une variable donnée par la **i-ème ligne** du tableau de contingence.
On peut de même définir la **distribution conditionnelle** de $X$ **sachant** $Y = b_j$ :

$$\{(a_1, n_{1j}), \ldots, (a_p, n_{pj})\}.$$

→ C'est la **j-ème** colonne du tableau.

**Exemple :** Donner la distribution conditionnelle de la variable "Catégorie de poids du bébé" sachant que "Statut tabagique" = Fumeuse.

**• Fréquence conditionnelle**

**Définition 4 :** On appelle **fréquence conditionnelle** de $b_j$ sachant $x = a_i$ le nombre :

$$f_{j|x=a_i} = \frac{n_{ij}}{n_{i\bullet}}.$$

On appelle **fréquence conditionnelle** de $a_i$ sachant $y = b_j$ le nombre :

$$f_{i|y=b_j} = \frac{n_{ij}}{n_{\bullet j}}.$$

### 3.2.1 $Y$ quantitative et $X$ qualitative (ou quantitative)

On suppose dans ce paragraphe que $Y$ est une variable quantitative prenant les valeurs $b_1, b_2, \ldots, b_q$ et que $X$ est une variable qualitative à $p$ modalités $a_1, a_2, \ldots, a_p$ (ou bien quantitative prenant les valeurs $a_1, a_2, \ldots, a_p$).

**• Moyenne conditionnelle**

**Définition 5 :** On appelle **moyenne conditionnelle** de $Y$ sachant $X = a_i$ la moyenne de la distribution conditionnelle de $Y$ sachant $X = a_i$, autrement dit la quantité :

$$\bar{y}_{|x=a_i} = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} b_j,$$

**• Variance conditionnelle**

**Définition 6 :** On appelle **variance conditionnelle** de $Y$ sachant $X = a_i$ la variance de la distribution conditionnelle de $Y$ sachant $X = a_i$, autrement dit la quantité :

$$s^2_{y|x=a_i} = \frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} (b_j - \bar{y}_{|x=a_i})^2,$$

**• Ecart-type conditionnel**

**Définition 7 :** Il est défini comme la racine carrée de la variance conditionnelle :

$$s_{y|x=a_i} = \sqrt{\frac{1}{n_{i\bullet}} \sum_{j=1}^{q} n_{ij} (b_j - \bar{y}_{|x=a_i})^2},$$

**Dans la pratique**, on trie les valeurs de $Y$ selon les $p$ modalités de $X$ et on obtient ainsi $p$ échantillons (conditionnels) :

$$(b^1_1, \ldots, b^1_{n_{1.}}), (b^2_1, \ldots, b^2_{n_{2.}}), \ldots, (b^p_1, \ldots, b^p_{n_{p.}})$$

Puis on calcule la moyenne et la variance de chaque échantillon (conditionnel) : ce sont les $p$ moyennes et les $p$ variances conditionnelles.

**Exemples :**

**1.** On dispose d'un échantillon de 120 enfants de maternelle composé de 55 filles et 65 garçons. Soient $Y$ la variable Poids (mesurée en kg) et $X$ la variable Sexe.
Supposons que la moyenne du poids des filles est de 19.3 kg (avec une variance de 1 kg$^2$) et que la moyenne du poids des garçons est de 20,2 kg (avec une variance de 1.4 kg$^2$).

**2.** Tableau des données brutes correspondant à la répartition en classes donnée en page 19 :

| | Sexe | Taille | Poids |
|---|---|---|---|
| 1 | 1 | 165 | 70.0 |
| 2 | 1 | 180 | 72.5 |
| 3 | 1 | 184 | 80.0 |
| 4 | 1 | 183 | 95.0 |
| 5 | 1 | 173 | 77.5 |
| 6 | 1 | 183 | 82.5 |
| 7 | 1 | 180 | 75.0 |
| 8 | 1 | 185 | 95.0 |
| 9 | 1 | 180 | 97.5 |
| 10 | 1 | 178 | 69.0 |
| 11 | 1 | 178 | 85.0 |
| 12 | 1 | 175 | 65.0 |
| 13 | 1 | 175 | 65.0 |
| 14 | 1 | 188 | 92.5 |
| 15 | 1 | 185 | 80.0 |
| 16 | 1 | 180 | 77.5 |
| 17 | 1 | 175 | 76.5 |
| 18 | 1 | 168 | 72.5 |
| 19 | 1 | 177 | 85.0 |
| 20 | 1 | 180 | 72.5 |
| 21 | 1 | 173 | 75.0 |
| 22 | 1 | 183 | 82.0 |
| 23 | 1 | 185 | 70.0 |
| 24 | 1 | 165 | 71.0 |
| 25 | 2 | 159 | 62.0 |
| 26 | 2 | 165 | 60.0 |
| 27 | 2 | 170 | 60.0 |
| 28 | 2 | 170 | 64.0 |
| 29 | 2 | 158 | 52.5 |
| 30 | 2 | 175 | 62.5 |
| 31 | 2 | 170 | 58.0 |
| 32 | 2 | 173 | 70.0 |
| 33 | 2 | 173 | 65.0 |
| 34 | 2 | 157 | 56.0 |
| 35 | 2 | 170 | 62.5 |
| 36 | 2 | 164 | 57.0 |
| 37 | 2 | 170 | 59.0 |
| 38 | 2 | 168 | 54.0 |
| 39 | 2 | 164 | 55.0 |
| 40 | 2 | 174 | 60.0 |

**Sexe** : 1 = “Homme”, 2 = “Femme”
**Taille** : taille de l’individu en cm
**Poids** : poids de l’individu en kg

On peut en déduire 4 échantillons conditionnels du Poids en fonction de (la classe) de la Taille :

| | ]150,160] $c_1 = 155$ | ]160,170] $c_2 = 165$ | ]170,180] $c_3 = 175$ | ]180,190] $c_4 = 185$ |
|---|---|---|---|---|
| 1 | 62.0 | 70.0 | 72.5 | 80.0 |
| 2 | 52.5 | 72.5 | 77.5 | 95.0 |
| 3 | 56.0 | 71.0 | 75.0 | 82.5 |
| 4 | | 60.0 | 97.5 | 95.0 |
| 5 | | 60.0 | 69.0 | 92.5 |
| 6 | | 64.0 | 85.0 | 80.0 |
| 7 | | 58.0 | 65.0 | 82.0 |
| 8 | | 62.5 | 65.0 | 70.0 |
| 9 | | 57.0 | 77.5 | |
| 10 | | 59.0 | 76.5 | |
| 11 | | 54.0 | 85.0 | |
| 12 | | 55.0 | 72.5 | |
| 13 | | | 75.0 | |
| 14 | | | 62.5 | |
| 15 | | | 70.0 | |
| 16 | | | 65.0 | |
| 17 | | | 60.0 | |
| Total | 3 | 12 | 17 | 8 |

Et on peut alors caluler :

**Moyennes, Variances et Ecarts-types conditionnels de Poids sachant Taille** :

| | | Moyenne | Variance | Ecart-type |
|---|---|---|---|---|
| Taille | $]150, 160]$ | | | |
| | $]160, 170]$ | 61.92 | 35.87 | 5.99 |
| | $]170, 180]$ | 73.56 | 84.73 | 9.20 |
| | $]180, 190]$ | 84.63 | 68.17 | 8.26 |

**Propriété 1 :** La moyenne des moyennes conditionnelles (pondérées par les effectifs $n_{i\bullet}$) est égale à la moyenne globale :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{p} n_{i\bullet} \bar{y}_{|x=a_i}$$

**Exemples :**

**1.** Calculer le poids moyen des 120 enfants de maternelle.

**2.** Vérifier l'égalité sur le Poids des 40 individus (on donne $\sum_{i=1}^{40} y_i = 2841$).

**Définition 8 :** On appelle **moyenne des variances conditionnelles** de $Y$ sachant $X$ le nombre :

$$\overline{s^2_{y|x}} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet} s^2_{y|x=a_i}$$

**Définition 9 :** On appelle **variance des moyennes conditionnelles** de $Y$ sachant $X$ le nombre :

$$s^2_{\bar{y}|x} = \frac{1}{n}\sum_{i=1}^{p} n_{i\bullet}(\bar{y}_{|x=a_i} - \bar{y})^2$$

**Propriété 2 :**
Variance totale = Variance des moyennes conditionnelles + Moyenne des variances conditionnelles :

$$s^2_y = s^2_{\bar{y}|x} + \overline{s^2_{y|x}}$$

**Exemples :**

1. Calculer la variance du poids des 120 enfants de maternelle.
2. Vérifier l'égalité sur le Poids des 40 individus (on donne $\sum_{i=1}^{40} y_i^2 = 207433$).

**Définition 10 :** On appelle **rapport de corrélation** de $Y$ sachant $X$ la quantité :

$$\eta^2_{y|x} = \frac{s^2_{\bar{y}|x}}{s^2_y}$$

Remarque 1 : Pour l'étude des distributions conditionnelles et du rapport de corrélation :
- la variable $Y$ (appelée variable dépendante) est une variable quantitative,
- la variable $X$ (appelée variable explicative) peut être qualitative, quantitative discrète ou bien quantitative continue regroupée en classes.

Remarque 2 :

$$0 \leq \eta^2_{y|x} \leq 1$$

**Interprétation :**
Le rapport $\eta^2_{y|x}$ représente la **part de variance** de $Y$ **expliquée** par $X$ :

- $\eta^2_{y|x} = 1 \Leftrightarrow s^2_{\bar{y}|x} = s^2_y \Leftrightarrow \overline{s^2_{y|x}} = 0 \Leftrightarrow \forall i,\ s^2_{y|x=a_i} = 0$

Si $\eta^2_{y|x} = 1$, alors toutes les variances conditionnelles sont nulles ; c'est-à-dire que conditionnellement à $x = a_i$, $(\forall\, 1 \leq i \leq p,)$ le caractère $Y$ est constant. Si on connaît $x = a_i$ alors on connaît la valeur de $Y$.
Dans ce cas, $X$ **"explique totalement"** $Y$.

- $\eta^2_{y|x} = 0 \Leftrightarrow s^2_{\bar{y}|x} = 0 \Rightarrow \bar{y}_{|x=a_1} = \cdots = \bar{y}_{|x=a_p} = \bar{y}$

Si $\eta^2_{y|x} = 0$, alors toutes les moyennes conditionnelles sont égales ; c'est-à-dire que, en moyenne, $Y$ **ne dépend pas** de $X$, autrement dit que $X$ **"n'explique pas"** $Y$.

**En conséquence :**

Plus $\eta^2_{y|x}$ est proche de 1, plus $Y$ dépend de $X$, autrement dit plus on peut dire que $Y$ est une fonction de $X$.

**En pratique :**
- si $\eta^2_{y|x} \leq 0,2$ alors on admet que $Y$ **ne dépend pas** de $X$,
- si $0,2 < \eta^2_{y|x} < 0,9$ alors on pourra dire que $Y$ **dépend partiellement** de $X$,
- si $\eta^2_{y|x} \geq 0,9$ alors on admet que $Y$ **dépend fonctionnellement** de $X$.

**Exemples :**

**1.** Calculer le rapport de corrélation du Poids des 120 enfants de maternelle sachant le Sexe.
**2.** Calculer le rapport de corrélation du Poids des 40 individus sachant la Taille.

• **Notion d'indépendance de deux caractères statistiques**

**Définition 11 :** Soit $(X, Y)$ un couple de caractères statistiques de distribution $\{(a_i, b_j), n_{ij}\}$ et de distributions marginales

$$\{(a_1, n_{1\bullet}), \ldots, (a_p, n_{p\bullet})\}$$

et

$$\{(b_1, n_{\bullet 1}), \ldots, (b_q, n_{\bullet q})\}.$$

On dit que $X$ et $Y$ sont (statistiquement) **indépendants** si $\forall i, j$,

$$f_{ij} = f_{i\bullet} f_{\bullet j} \Leftrightarrow n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

REMARQUE 3 : Si $X$ et $Y$ sont indépendants, alors :

$$\eta^2_{y|x} = \eta^2_{x|y} = 0.$$

REMARQUE 4 : La notion d'indépendance de 2 caractères statistiques dépend de la distribution statistique considérée. Si on procède à d'autres observations ou si on change les classes du regroupement, il se peut que la définition de l'indépendance ne soit plus vérifiée alors qu'elle l'était (ou inversement). Néanmoins on devrait toujours avoir :

$$n_{ij} \simeq \frac{n_{i\bullet} n_{\bullet j}}{n}$$

REMARQUE 5 : Pour l'étude de l'indépendance statistique $X$ et $Y$ peuvent être qualitatifs, quantitatifs discrets ou continus regroupés en classes.

### 3.2.2 $X$ et $Y$ qualitatives

Soit $(X, Y)$ un couple de variables qualitatives dont la distribution du couple est donnée par le tableau de contingence suivant :

| X\Y | $b_1$ | $\cdots$ | $b_j$ | $\cdots$ | $b_q$ | Total |
|---|---|---|---|---|---|---|
| $a_1$ | $n_{11}$ | $\cdots$ | $n_{1j}$ | $\cdots$ | $n_{1q}$ | $n_{1\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_i$ | $n_{i1}$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $n_{iq}$ | $n_{i\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_p$ | $n_{p1}$ | $\cdots$ | $n_{pj}$ | $\cdots$ | $n_{pq}$ | $n_{p\bullet}$ |
| Total | $n_{\bullet 1}$ | $\cdots$ | $n_{\bullet j}$ | $\cdots$ | $n_{\bullet q}$ | $n$ |

• **Mesure de la relation entre deux variables qualitatives**

**Définition 12 :** On appelle **effectif théorique** de la modalité $(a_i, b_j)$ d'un tableau croisé le nombre :

$$\hat{n}_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

REMARQUE 6 : Par opposition, on appelle parfois $n_{ij}$ **l'effectif observé**.

On a vu que si les deux caractères sont statistiquement indépendants alors

$$n_{ij} \simeq \hat{n}_{ij}$$

Pour mesurer la dépendance entre $X$ et $Y$, on va introduire un indice qui mesure une certaine "distance" entre les effectifs observés $n_{ij}$ et les effectifs théoriques $\hat{n}_{ij}$.

**Définition 13 :** On appelle **distance du "khi-deux"** ($\chi^2$) ou encore **indice "khi-deux"** le coefficient défini par :

$$\chi^2 = \text{dist}_{\chi^2}(X,Y) = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{(n_{ij}-\frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}$$

Pour le calcul de cette distance du "khi-deux", on construit donc le tableau de contingence théorique :

**Tableau de contingence théorique**

| X\Y | $b_1$ | $\cdots$ | $b_j$ | $\cdots$ | $b_q$ | Total |
|---|---|---|---|---|---|---|
| $a_1$ | $\frac{n_{1\bullet}n_{\bullet 1}}{n}$ | $\cdots$ | $\frac{n_{1\bullet}n_{\bullet j}}{n}$ | $\cdots$ | $\frac{n_{1\bullet}n_{\bullet q}}{n}$ | $n_{1\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_i$ | $\frac{n_{i\bullet}n_{\bullet 1}}{n}$ | $\cdots$ | $\frac{n_{i\bullet}n_{\bullet j}}{n}$ | $\cdots$ | $\frac{n_{i\bullet}n_{\bullet q}}{n}$ | $n_{i\bullet}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $a_p$ | $\frac{n_{p\bullet}n_{\bullet 1}}{n}$ | $\cdots$ | $\frac{n_{p\bullet}n_{\bullet j}}{n}$ | $\cdots$ | $\frac{n_{p\bullet}n_{\bullet q}}{n}$ | $n_{p\bullet}$ |
| Total | $n_{\bullet 1}$ | $\cdots$ | $n_{\bullet j}$ | $\cdots$ | $n_{\bullet q}$ | $n$ |

Remarque 7 : Ce coefficient est proche de 0 si les deux caractères sont indépendants et il s'éloigne de 0 s'il existe un lien entre les deux. Plus cette distance est élevée, moins les variables sont indépendantes.
La distance du $\chi^2$ dépend aussi du nombre de modalités $p$ et $q$, ainsi que de la taille $n$ de l'échantillon observé.
Par exemple :
- Si $p = q = 2$, on rejette l'hypothèse d'indépendance et on dit donc que les variables sont liées lorsque $\text{dist}_{\chi^2}(X,Y) \geq 3,84$.
- Si $p = 2$ et $q = 3$, on rejette l'hypothèse d'indépendance et on dit donc que les variables sont liées lorsque $\text{dist}_{\chi^2}(X,Y) \geq 5,99$.

**Exemple :** Les variables Statut tabagique de la mère et Poids du bébé sont-elles indépendantes ?

### 3.2.3 $X$ et $Y$ quantitatives

Soit $(X, Y)$ un couple de variables quantitatives pour lequel on dispose d'un $n$-échantillon :

$$\{(x_i, y_i),\ 1 \leq i \leq n\}$$

**• Représentation graphique : le nuage de points ou diagramme de dispersion**

Ce graphe se construit à partir des données brutes de préférence. Dans un repère orthogonal, on reporte les points $(x_i, y_i)$, $1 \leq i \leq n$ et on étiquette éventuellement les points par l'effectif si plusieurs points sont superposés (observations identiques).

**Exemple :** Construire le nuage de points pour les données de Taille et Poids des 40 individus (cf. p. 21).

**• Mesure du lien linéaire entre 2 variables quantitatives**

**Objectif.** Étant donné un $n$-échantillon pour un couple de caractères statistiques $(X, Y)$ ou encore la distribution de ce couple (donnée par un tableau croisé), on cherche à déterminer s'il existe une relation linéaire entre les variables $X$ et $Y$, c'est-à-dire du type :

$$Y = aX + b$$

ou bien

$$X = \alpha Y + \beta,$$

avec $a$, $b$, $\alpha$, $\beta \in \mathbb{R}$.

**Définition 14 :** On appelle **covariance** de $X$ et $Y$ le paramètre :

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}(a_i - \bar{x})(b_j - \bar{y})$$

ou pour le cas de l'échantillon :

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

**Propriétés de la covariance :**

**1.** Formule de Koenig pour la covariance :

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

ou encore

$$Cov(X,Y) = \overline{xy} - \bar{x}\bar{y}$$

Autrement dit : la covariance est égale à la moyenne des produits moins le produit des moyennes.

**2.** $Cov(X,X) = Var(X) = s_x^2$

**3.** $Cov(X,Y) = Cov(Y,X)$

**4.** Soit $a,\ b,\ c,\ d \in \mathbb{R}$ et $X,\ Y,\ Z$, 3 caractères statistiques. On a :

$$Cov(aX + bY, Z) = a\ Cov(X,Z) + b\ Cov(Y,Z)$$

$$Cov(aX + b, cY + d) = ac\ Cov(X,Y)$$

**5.** Si $X$ et $Y$ sont indépendants, alors $Cov(X,Y) = 0$.
Ou, par contraposée : Si $Cov(X,Y) \neq 0$, alors $X$ et $Y$ ne sont pas indépendants.
Par contre, la réciproque n'est pas vraie en général :
Si $Cov(X,Y) = 0$, cela n'implique pas que $X$ et $Y$ sont indépendantes.

**Problème de la covariance :**

C'est un paramètre très sensible aux unités et aux changements d'échelles.
Par exemple, si $X$ = taille en $cm$ et $Y$ = poids en $kg$, et si $X'$ = taille en $mm$ et $Y'$ = poids en $g$, on a :

$$X' = 10X \text{ et } Y' = 1000Y,$$

et comme $Cov(aX, bY) = ab\ Cov(X,Y)$, on obtient

$$Cov(X',Y') = 10000\ Cov(X,Y),$$

mais on ne peut pas dire pour autant que $X'$ et $Y'$ sont 10000 fois plus liés que $X$ et $Y$.

Ceci nous amène à définir une nouvelle quantité qui n'est pas affectée par les problèmes d'échelle :

**Définition 15 :** On appelle **coefficient de corrélation linéaire (de Pearson)** de deux caractères statistiques $X$ et $Y$ la quantité :

$$\rho(X,Y) = Cov\left(\frac{X}{s_x}, \frac{Y}{s_y}\right) = \frac{Cov(X,Y)}{s_x\ s_y}$$

**Propriétés du coefficient de corrélation :**

**1.** On a : $$-1 \leq \rho(X,Y) \leq 1$$

**2.** $\rho(X,Y)$ est invariant par changement d'origine et d'échelle :

$$\rho(aX + b, cY + d) = \rho(X,Y),\ \forall a > 0,\ c > 0,\ b,\ d \in \mathbb{R}.$$

En effet :

$$\rho(aX+b,cY+d)=\frac{Cov(aX+b,cY+d)}{s_{ax+b}\ s_{cy+d}}=\frac{ac\ Cov(X,Y)}{|a|s_x\ |c|s_y}=\frac{Cov(X,Y)}{s_x\ s_y}=\rho(X,Y)$$

**3.** $|\rho(X,Y)|=1 \iff \exists\ (a,b)\in\mathbb{R}^2$ t.q. $\forall i\ 1\leq i\leq n : y_i=ax_i+b$
$\iff \exists\ (a,b)\in\mathbb{R}^2$ t.q. $Y=aX+b$.

**4.** Si $\rho(X,Y)>0$, alors $Y$ et $X$ varient linéairement dans le même sens.

Si $\rho(X,Y)<0$, alors $Y$ et $X$ varient linéairement en sens inverse.

Si $\rho(X,Y)\simeq 0$, il n'y a pas de tendance linéaire. Il peut cependant exister un lien d'une autre nature entre $Y$ et $X$ : $Y=aX^2+b$, $Y=\ln X,\cdots$

**5.** Si $X$ et $Y$ sont indépendants, alors $\rho(X,Y)=0$.
Ou, par contraposée : Si $\rho(X,Y)\neq 0$, alors $X$ et $Y$ ne sont pas indépendants.
Par contre, la réciproque n'est pas vraie en général :
Si $\rho(X,Y)=0$, cela n'implique pas que $X$ et $Y$ sont indépendantes.

**Coefficient de corrélation linéaire et rapport de corrélation :**
On a les relations suivantes :

$$\rho^2(X,Y)\leq \eta^2_{y|x}$$
$$\rho^2(X,Y)\leq \eta^2_{x|y}$$

**Interprétation pratique du coefficient de corrélation linéaire :**
- Si $\rho^2(X,Y)\leq 0,2$ on dit que $Y$ ne dépend pas linéairement de $X$.
- Si $0,2<\rho^2(X,Y)<0,9$ on dit qu'il existe une dépendance linéaire partielle entre $Y$ et $X$.
- Si $\rho^2(X,Y)\geq 0,9$ alors $Y$ dépend linéairement de $X$ (la liaison linéaire est très forte).

**Exemple :** Calculer le coefficient de corrélation de la Taille et du Poids des 40 individus.

**Interprétation simultanée du coefficient de corrélation linéaire et du rapport de corrélation :**

Pour comparer $\eta^2_{y|x}$ et $\rho^2(X,Y)$, il faut supposer qu'on travaille sur des **données regroupées en classes** et donc que $\eta^2_{y|x}$ et $\rho^2(X,Y)$ sont calculés à partir des **mêmes données**.

- Si $\rho^2(X,Y) \leq 0,2$ et $0,2 < \eta^2_{y|x} < 0,9$, on dit que $Y$ dépend partiellement de $X$ mais que cette dépendance n'est pas linéaire.
- Si $0,2 < \rho^2(X,Y) < 0,9$ et $0,2 < \eta^2_{y|x} < 0,9$, on dit que $Y$ dépend partiellement de $X$ et que cette dépendance est linéaire.
- Si $\rho^2(X,Y) < 0,9$ et $\eta^2_{y|x} \geq 0,9$ , $Y$ dépend fonctionnellement de $X$ mais cette fonction n'est pas une droite.
- Si $\rho^2(X,Y) \geq 0,9$ alors $Y$ dépend linéairement (donc fonctionnellement) de $X$.

**• Régression linéaire**

Lorsque le coefficient de corrélation linéaire $\rho^2(X,Y)$ est suffisamment éloigné de 0, on admet qu'il existe une relation linéaire du type $y = f(x) = ax + b$ (même partielle) entre $Y$ et $X$.
On s'intéresse alors à la famille des droites :

$$\mathcal{F} = \{f(x) = ax + b,\ (a,b) \in \mathbb{R}^2\}$$

et on cherche les valeurs $\hat{a}$ et $\hat{b}$ (de $a$ et $b$) telles que :

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - \hat{a}x_i - \hat{b})^2$$

soit le plus petit possible.

On dit alors que $f$ minimise la **somme des carrés des résidus** et la méthode utilisée porte le nom de **méthode des moindres carrés**.

On peut démontrer que le minimum de cette fonction est atteint pour le couple $(\hat{a}, \hat{b})$ tel que :

$$\hat{a} = \frac{Cov(X,Y)}{s_x^2} \quad \text{et} \quad \hat{b} = \bar{y} - \frac{Cov(X,Y)}{s_x^2}\bar{x}.$$

Et l'équation de la **droite de régression** de $Y$ en $X$ est donc :

$$y = \hat{a}x + \hat{b} = \frac{Cov(X,Y)}{s_x^2}x + \bar{y} - \frac{Cov(X,Y)}{s_x^2}\bar{x}$$

Remarque 8 :

**1.** Notons que le fait que $\bar{y} = \hat{a}\bar{x} + \hat{b}$ indique que la droite de régression passe par le **point moyen** $(\bar{x}, \bar{y})$.

**2.** $\rho(X,Y) = 0 \Leftrightarrow Cov(X,Y) = 0 \Leftrightarrow \hat{a} = 0$ et donc la droite de régression de $Y$ en $X$ est dans ce cas horizontale.

**Interprétation des coefficients :**

- Pente $\hat{a}$ : quand $X$ augmente d'une unité, $Y$ augmente en moyenne de $\hat{a}$ unités.
- Ordonnée à l'origine $\hat{b}$ : $\hat{b}$ est la valeur de $Y$ lorsque $X = 0$.

**Exemple :** Déterminer et tracer la droite de régression du Poids en fonction de la Taille des 40 individus.