UNIVERSITE DE TOULOUSE LE MIRAIL

# MIS2OP1X

# Statistique pour les Sciences Humaines I

**Polycopié de statistique**

**à l'usage des étudiants inscrits au S.E.D.**

———

**Agnès Lagnoux** **lagnoux@univ-tlse2.fr**

Les étudiants inscrits en première année de Licence dans toute filière ou en deuxième année de Licence d'Histoire peuvent suivre des cours de statistique descriptive, dans le cadre des U.E. optionnelles MIS2OP1X (1er semestre) et uniquement pour les Historiens MIHO15X (2ème semestre) intitulées "Statistique pour les Sciences Humaines I" et "Statistique pour les Sciences Humaines II (parcours Histoire)". L'U.E. optionnelle MIS2OP2X "Statistique pour les Sciences Humaines II" destinée aux étudiants de première année de Licence dans toute filière traitera de Probabilités et Statistique.

L'objectif de ce cours est de donner les outils nécessaires à la compréhension et à l'analyse de documents comportant des données numériques, en liaison avec les Sciences Humaines. Il s'agit d'une initiation à la statistique descriptive qui ne nécessite pas de connaissances spécifiques préalables ; cependant, ce sera l'occasion de revoir avec un peu de recul des notions mathématiques élémentaires qui font partie de la culture générale (calculer des taux de variation, résoudre une équation, utiliser un repère cartésien,...). Ces enseignements sont assurés par des professeurs de mathématiques dépendant du département de Mathématiques et Informatique de l'UFR SES.

Les exercices proposés (situés en fin de polycopié) s'appuient le plus souvent sur des données réelles mais la quantité de données est parfois réduite pour permettre de faire les calculs en un temps raisonnable. Une calculatrice est nécessaire, tous les modèles sont autorisés mais une calculatrice scientifique pour le collège suffit. Une des premières difficultés est de se familiariser avec le vocabulaire spécifique de la statistique. Les mots ayant un sens mathématique précis définis dans ce cours sont en gras. Les premiers exercices fournissent une liste d'exemples permettant d'assimiler ce vocabulaire.

Nous présentons au premier semestre des généralités sur la statistique descriptive concernant une seule varible puis les couples de variables. Nous étudierons aussi l'existence de liaison entre deux variables quelconques. Le deuxième semestre est consacré aux taux de variation, à un type de liaison particulier entre deux variables qui est la corrélation linéaire et enfin à l'étude des séries temporelles. Il est nécessaire d'avoir suivi les cours du premier semestre pour aborder le deuxième. A la fin de chaque semestre, nous illustrerons toutes les notions abordées dans le cours en utilisant le tableur Excel.

N'hésitez pas à me contacter par courrier électronique ou téléphone si vous avez une question

précise concernant le cours ou bien l'organisation de l'U.E.. Il y aura un regroupement en fin de semestre (la date vous sera communiquée plus tard), je vous conseille vivement d'y particper. Enfin, les remarques et suggestions concernant ce nouveau polycopié sont les bienvenues.

Bon courage !

Responsable des U.E. MIS2OP1X, MIHO11X et MIHO15X :
**Agnès Lagnoux**
**U.F.R. S.E.S.**
**Département de Mathématiques et Informatique**
**Bureau 1039, bâtiment 13**
**Tél : 05-61-50-46-11, e-mail : lagnoux@univ-tlse2.fr.**

**Organisation** : Pour les étudiants inscrits en contrôle continu, il y a 12 séances de Cours/TD (2 heures par semaine pendant 12 semaines) au premier semestre. Si vous souhaitez (et pouvez) assister à quelques séances, n'hésitez pas à vous renseigner sur l'horaire du cours : il n'est pas facile de comprendre seul certaines notions.
Vous **devez** renvoyer le devoir qui se trouve à la fin du polycopié avant le 30 novembre 2013 (il sera corrigé et noté à titre indicatif).

**Evaluation** : Une épreuve écrite de statistique aura lieu en janvier. La calculatrice ainsi qu'une feuille manuscrite recto-verso sont autorisées à l'examen (vous pouvez vous inspirer des fiches proposées à la fin du polycopié).

# Table des matières

# Introduction

## Petit historique

Les premières manifestations de la **statistique** remontent à l'Antiquité : par exemple, dans l'ancienne Egypte, le niveau des crues du Nil était enregistré de manière systématique ; on peut citer aussi le recensement ordonné par l'empereur romain CESAR-AUGUSTE.

Le recours à la statistique touche désormais des domaines très divers dans les sciences humaines, physiques,... : démographie, psychologie, pharmacologie, climatologie, météorologie, économie, astronomie,...

Jusqu'à la fin du dix-neuvième siècle, la statistique était essentiellement une technique de comptage ou dénombrement. C'est à partir du siècle dernier qu'elle a connu un essor considérable grâce :

- au développement de techniques statistiques utilisant notamment le calcul des probabilités (théorie mathématique dont le but est d'étudier les lois régissant des phénomènes ou expériences aléatoires, c'est-à-dire dont on ne peut pas prévoir de manière certaine le résultat) ;

- à la collecte importante de données (au travers notamment d'organismes comme l'INSEE ou l'INED en France) ;

- au développement d'ordinateurs permettant le traitement de grands tableaux de données et l'utilisation de logiciels performants.

## Le concept de statistique

Selon la définition du Petit Robert, on désigne par le terme **statistique**, *l'ensemble des techniques d'interprétation mathématique appliquées à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible, à cause de leur grand nombre ou de leur complexité.* Remarquons que cette définition est très différente du sens parfois donné dans le langage courant au mot statistique, c'est-à-dire un ensemble de données numériques comme par exemple "la statistique

de la mortalité".

Une étude statistique comprend quatre étapes :

**1. Le recueil des données** : il s'effectue lors d'une **enquête**. Celle-ci peut-être **exhaustive**, c'est-à-dire porter sur tous les sujets concernés par l'étude et est appelée dans ce cas **recensement**. Dans la plupart des cas, compte tenu du coût trop lourd d'un recensement, on effectue une enquête partielle (portant sur une partie des sujets) appelée **sondage**. Dans ce second cas, pour que les informations obtenues soient intéressantes, il faut que le sondage soit effectué avec certaines règles. La **théorie des sondages** s'intéresse à la manière de choisir, parmi tous les sujets, ceux sur lesquels faire porter l'étude statistique. Dans ce cours, cet aspect ne sera pas abordé.

**2. Le dépouillement des données** : il consiste à rassembler les données et à les organiser par exemple sous forme de **tableaux**, les classer, les **coder** (par exemple, dans le cas du sexe, on attribue 1 aux hommes et 2 aux femmes)...

**3. Traitement des données** : c'est lors de cette étape qu'interviennent les techniques statistiques. Le but de cette étape est de retirer un maximum d'information des données. On distingue deux branches principales en statistique :

**a. La statistique descriptive** vise principalement deux objectifs :

- d'une part, la **représentation graphique** des données en alliant à la fois la lisibilité de la représentation et la fidélité aux données ;

- d'autre part, le résumé des données par des **caractéristiques numériques**.

**b. La statistique inductive ou inférentielle** consiste à faire des extrapolations à partir d'un échantillon. Elle suppose que le phénomène étudié peut être décrit par un **modèle mathématique** (donc théorique) permettant d'approcher les propriétés de ce phénomène. Le choix de ce modèle est bien sûr un problème important puisqu'il doit représenter au mieux la réalité. Les méthodes utilisées en statistique inductive font appel au **calcul des probabilités**.

**4. L'interprétation des résultats** : il s'agit d'une étape délicate nécessitant une bonne connaissance du phénomène étudié et des méthodes statistiques utilisées.

La première partie de ce cours sera consacrée aux généralités et à l'analyse statistique d'une variable et la deuxième partie à l'analyse statistique de deux variables.

# Chapitre 1

# Généralités sur la statistique descriptive

Dans le cadre d'une étude préalablement définie, la statistique descriptive a pour but, à partir d'un recueil de données, de mettre en forme ces données (tableaux, graphiques) et de les résumer à l'aide de valeurs caractéristiques (moyenne, médiane, écart-type,...).

## 1.1 Premiers éléments de vocabulaire

### 1.1.1 Population, Individu, Echantillon

Avant toute enquête statistique, il faut définir avec précision la **population** que l'on souhaite étudier. Il s'agit de l'ensemble concerné par l'étude effectuée. Une population peut être un ensemble de personnes, d'objets, de situations, de pays,... On notera souvent $\Omega$ (oméga majuscule) la population étudiée.

**Exemple 1.1.** *a) Si on fait le recensement des Français, la population est l'ensemble de tous les Français.*
*b) Si on fait une étude géologique portant sur les chaînes montagneuses, la population est l'ensemble des chaînes de montagne...*

Un élément de la population s'appelle un **individu** ou une **unité statistique**, on notera souvent $\omega$ (oméga minuscule) un individu quelconque.

**Exemple 1.2.** *La chaîne des Pyrénées est un individu de la population des chaînes de montagne.*

On affecte en général un numéro à chaque individu : 1, 2,... et on note de manière générique un individu par la lettre $i$. Ce numéro n'a pas de valeur numérique car il s'agit d'un codage. Le nombre d'éléments de la population s'appelle la **taille** de la population.

Lorsque la population est trop vaste pour que l'on puisse réaliser une étude exhaustive (par exemple pour des raisons de "coût" trop important ou par manque de temps), on est amené à effectuer un sondage et à ne considérer qu'une partie de la population. Cette partie est appelée **échantillon**. Le nombre d'individus de l'échantillon est appelé **taille de l'échantillon**, et notée $N$. (C'est dans ce cadre qu'interviendra ultérieurement la statistique inférentielle qui permettra de dire si on peut étendre les résultats obtenus sur l'échantillon étudié à la population entière.)

### 1.1.2 Variable

Un renseignement demandé, une question posée est appelé **variable** en statistique (**champ** en Informatique et **caractère** en Sciences Humaines). On notera souvent $X$ (ou $Y$) la variable étudiée.

Pour qu'une question posée soit considérée comme une variable, il faut que toutes les réponses possibles soient exprimables, et que chaque individu ne puisse donner qu'une seule réponse concernant le renseignement demandé.

**Une variable (renseignement) associe à chaque individu une réponse et une seule.** En statistique les réponses possibles, c'est-à-dire les valeurs prises par la variable, sont appelées **modalités**. Un couple (individu ; modalité associée) est appelé **donnée** ou **observation**. Pour une variable étudiée sur un échantillon de taille $N$, le nombre d'observations est donc $N$. L'ensemble des données pour une variable $X$ s'appelle une **série statistique**.

Le nombre de modalités $k$ d'une variable est en général inférieur à $N$. Les modalités seront souvent notées $x_j$ pour $j = 1, \cdots, k$. Pour une variable $X$ donnée, à chaque individu de la population $\Omega$ est associée une et une seule modalité de la variable $X$. (On dit que $X$ est une application de l'ensemble de départ $\Omega$ dans l'ensemble des modalités).

**Important :** Pour définir une variable, il faut, après avoir indiqué sur quelle population on travaille, préciser l'ensemble des modalités de la variable.

Nous allons présenter cette notion à l'aide d'exemples concrets. Ces exemples, notés $A$ et $B$ seront repris à plusieurs reprises dans ce cours.

**Exemple A.** *Dans la population d'une ville, on a prélevé un échantillon de 20 personnes de moins de 30 ans. La population $\Omega$ est donc composée de ces personnes de moins de 30 ans. Elle*

*a pour taille $N = 20$. Le questionnaire utilisé est le suivant :*

| |
|---|
| Prénom : |
| Age : |
| Diplôme le plus élevé : Supérieur □ ; Baccalauréat □ ; Brevet □ ; Aucun □ |
| Sexe : Masculin □ ; Féminin □ |
| Nombre d'enfants : |
| Taille (en cm) : |
| Spécialité scolaire : L □ ; ES □ ; S □ ; T□ |
| Goût pour la lecture : Faible □ ; Moyen □ ; Fort □ |

*Ce questionnaire fait donc apparaître 7 variables :*

*- l'âge dont les modalités sont les entiers allant de 0 à 30 et qui associe à chaque personne son âge,*

*- le diplôme le plus élevé dont les modalités sont "Supérieur", "Baccalauréat", "Brevet", "Aucun" et qui à chaque individu associe son plus haut diplôme,*

*- le sexe dont les modalités sont "masculin", "féminin" et qui associe à chaque personne son sexe,*

*- le nombre d'enfants dont les modalités sont les entiers à partir de 0 et qui associe à chaque individu son nombre d'enfants,*

*- la taille dont les modalités sont des entiers allant à partir de 0 et qui associe à chaque individu sa taille en cm,*

*- la spécialité scolaire dont les modalités sont "L", "ES", "S" et "T" (pour Technique) et qui associe à chaque personne sa série de préférence,*

*- le goût pour la lecture dont les modalités sont "faible", "moyen" et "fort" et qui associe à chaque individu son goût pour la lecture. .*

**Remarque 1.1.** *a) La variable "Prénom" joue un rôle à part : elle sert ici essentiellement à identifier chaque individu. Ce n'est donc pas une variable que l'on étudie. Il n'en serait pas de même si l'objet de l'enquête était de faire une étude des différents types de prénoms.*
*b) La question "Diplôme" n'est pas une variable car un individu peut avoir obtenu plusieurs diplômes (une personne ayant le baccalauréat a aussi son brevet). Cependant on peut parler de la variable "Diplôme le plus élevé". Les réponses possibles étant données dans le questionnaire, il était nécessaire de proposer la réponse "aucun", sinon certains individus auraient pu être dans l'impossibilité de cocher une réponse et on n'aurait alors pas eu affaire à une variable.*

*Les résultats sont consignés dans un tableau, le tableau des données "brutes", ci-dessous.*

| Individu | Âge | Diplôme | Sexe | Nbre d'enfants | Taille | Spécialité | Goût lecture |
|---|---|---|---|---|---|---|---|
| Elise | 22 | Baccalauréat | f | 1 | 161 | L | Fort |
| Claire | 25 | Diplôme supérieur | f | 1 | 165 | S | Faible |
| Etienne | 30 | Baccalauréat | m | 2 | 172 | L | Fort |
| Thierry | 25 | Aucun | m | 1 | 183 | ES | Fort |
| Bertrand | 30 | Diplôme supérieur | m | 3 | 175 | L | Fort |
| Carine | 22 | Diplôme supérieur | f | 0 | 158 | ES | Moyen |
| Lucien | 24 | Brevet | m | 0 | 174 | S | Faible |
| Paulette | 29 | Baccalauréat | f | 2 | 169 | L | Fort |
| Gaston | 25 | Diplôme supérieur | m | 0 | 181 | ES | Moyen |
| Francis | 23 | Diplôme supérieur | m | 0 | 168 | S | Fort |
| Carole | 22 | Aucun | f | 1 | 157 | L | Fort |
| Cécile | 24 | Brevet | f | 2 | 163 | L | Fort |
| Eric | 25 | Diplôme supérieur | m | 0 | 168 | L | Fort |
| Jules | 24 | Aucun | m | 1 | 178 | ES | Moyen |
| Vincent | 25 | Brevet | m | 1 | 184 | S | Faible |
| Monique | 28 | Baccalauréat | f | 3 | 164 | S | Faible |
| Adam | 25 | Baccalauréat | m | 0 | 173 | L | Fort |
| Nicolas | 27 | Diplôme supérieur | m | 2 | 180 | L | Fort |
| Audrey | 22 | Brevet | f | 1 | 158 | ES | Moyen |
| Victor | 27 | Diplôme supérieur | m | 2 | 174 | L | Fort |

Plus généralement, le tableau comprend $N$ lignes et au moins 2 colonnes : la première colonne est formée par les individus (repérés par un nom ou un numéro) et la deuxième colonne comprend les valeurs de la variable pour chaque individu. Sur la ligne $i$ de ce tableau figure donc l'individu $i$ et la valeur de $X$, $X(i)$, pour cet individu. Il s'agit là du cas le plus simple et le plus souvent, on étudie plusieurs variables sur une même population (âge, sexe, taille,...) : on ajoute dans ce cas autant de colonnes au tableau que de variables supplémentaires. Ce premier tableau fait en général l'objet d'un premier "nettoyage" : correction d'erreurs de saisie, suppression de lignes dans le cas de valeurs **manquantes** à cause de non réponse ou encore d'impossibilité à réaliser une mesure, ... Ce tableau "brut" est la base de l'étude statistique : c'est à partir de celui-ci qu'on génère d'autres tableaux, puis des graphiques, des résumés, ...

**Remarque 1.2.** *Les notions de "modalité" et de "donnée", bien qu'étant très proches ne sont pas identiques. Ainsi pour la variable "Spécialité scolaire", on a 4 modalités : "L", "ES", "S" et "T". Par contre, la liste des données est composée des 20 mots : L, S, L, ES, L, ES, S, L, ES, S, L, L, L, ES, S, S, L, L, ES, L (il y a une donnée par individu). La série statistique de la variable "Spécialité scolaire" est donc composée de 20 couples, cette série correspond aux deux colonnes "Individu"(qui permet d'identifier l'individu) et "Spécialité scolaire" du tableau ci-dessus.*

*Une autre différence entre modalité et donnée réside dans le fait que les données sont effectivement observées alors qu'une modalité peut ne pas l'être. Ainsi pour la variable âge de l'exemple A, 26 est une modalité (c'est en effet un âge possible) pourtant aucun de nos 20 individus n'a 26 ans. 26 n'est donc pas une donnée mais seulement une modalité.*

**Exemple B. L'Union Européenne et les pays candidats**

*Voici un tableau de données de 1996, concernant l'ensemble des pays de l'Union Européenne ainsi que les pays candidats à l'entrée dans l'U.E. en 2002 (Atlaséco 1999).*

| | *Membre de l'UE* | *Population* | *PNB global* | *PNB/habitant* | *F* | *Rang* |
|---|---|---|---|---|---|---|
| *Luxembourg* | *Oui* | *416* | *19* | *45673* | *4* | *69* |
| *Danemark* | *Oui* | *5262* | *170* | *32307* | *2* | *26* |
| *Allemagne* | *Oui* | *81912* | *2342* | *28592* | *9* | *3* |
| *Autriche* | *Oui* | *8059* | *226* | *28043* | *7* | *22* |
| *Suède* | *Oui* | *8843* | *240* | *27140* | *2* | *21* |
| *France* | *Oui* | *58375* | *1535* | *26296* | *6* | *4* |
| *Belgique* | *Oui* | *10159* | *267* | *26282* | *5* | *20* |
| *Pays-Bas* | *Oui* | *15517* | *399* | *25714* | *3* | *13* |
| *Finlande* | *Oui* | *5125* | *120* | *23415* | *2* | *35* |
| *Italie* | *Oui* | *57380* | *1193* | *20791* | *4* | *5* |
| *Royaume-Uni* | *Oui* | *58782* | *1148* | *19530* | *2* | *6* |
| *Irlande* | *Oui* | *3626* | *62* | *17099* | *2* | *46* |
| *Espagne* | *Oui* | *39260* | *574* | *14620* | *3* | *9* |
| *Chypre* | *Non** | *740* | *9* | *12162* | *1* | *83* |
| *Grèce* | *Oui* | *10475* | *125* | *11933* | *3* | *33* |
| *Portugal* | *Oui* | *9930* | *103* | *10373* | *2* | *36* |
| *Slovénie* | *Non** | *1991* | *19* | *9543* | *4* | *70* |
| *Malte* | *Non** | *373* | *3* | *8043* | *1* | *131* |
| *République Tchèque* | *Non** | *10315* | *54* | *5235* | *5* | *49* |
| *Hongrie* | *Non** | *10193* | *43* | *4219* | *5* | *51* |
| *Slovaquie* | *Non** | *5343* | *19* | *3556* | *5* | *68* |
| *Pologne* | *Non** | *38618* | *134* | *3470* | *5* | *32* |
| *Turquie* | *Non* | *62697* | *184* | *2935* | *3* | *24* |
| *Estonie* | *Non** | *1466* | *4* | *2729* | *2* | *116* |
| *Lituanie* | *Non** | *3709* | *8* | *2157* | *3* | *91* |
| *Lettonie* | *Non** | *2490* | *5* | *2008* | *3* | *107* |
| *Roumanie* | *Non* | *22608* | *35* | *1548* | *3* | *56* |
| *Bulgarie* | *Non* | *8356* | *9* | *1077* | *4* | *84* |

** Pays en passe d'adhérer à l'Union Européenne*

*On souhaite étudier et comparer les pays membres de l'Union Européenne (il y en a 15) et les*

*pays candidats en 2002 à l'entrée dans l'Union (il y en a 13). Pour cela, on recueille pour chaque pays un certain nombre d'indicateurs :*

- *Est-il membre ou pas de l'Union Européenne ?*
- *Quelle est sa population (en milliers d'habitants) ?*
- *Quel est son PNB (Produit National Brut) global (en milliards de dollars) ?*
- *Quel est son PNB par habitants (en dollars) ?*
- *Combien a-t-il de frontières communes avec l'ensemble des 28 pays (on compte le pays lui-même) ?*
- *Quel est son rang mondial pour le PNB global ?*

*La colonne "F" indique le nombre de frontières communes avec les 28 pays (le tunnel sous la Manche ne compte pas comme frontière entre la France et le Royaume-Uni, Gibraltar n'induit pas une frontière entre l'Espagne et le Royaume-Uni,... Par exemple, le Luxembourg touche la France, la Belgique et l'Allemagne donc F=4).*
*La colonne "Rang" donne le rang mondial pour le PNB global.*
*Dans notre étude, les variables sont : Membre (oui ou non), Population, PNB global, PNB/habitant, F (frontières), Rang (mondial pour le PNB global).*
*Les modalités de la variable "Membre" sont "oui", "non" et "non*".*
*Les modalités (ou valeurs) de la variable "F" sont les nombres entiers compris entre 1 et 28.*

**Remarque 1.3.** *Dans cet exemple la population étudiée, au sens statistique, est l'ensemble des 28 pays, sa taille est 28 (un individu est un pays). Il ne faut pas la confondre avec la variable "population " qui à chaque pays associe son nombre d'habitants.*

## 1.2 Classement des différents types de variables

En considérant l'ensemble des modalités, on distingue différents types de variables.

### 1.2.1 Le type quantitatif

Une variable est dite **quantitative** lorsque ses modalités sont des nombres qui résultent d'une mesure, d'un comptage. On parlera alors plus souvent de valeurs de la variable plutôt que de modalités.

- Lorsqu'il s'agit d'un comptage, on parlera de type **quantitatif discret**. C'est le cas en particulier lorsque l'ensemble des modalités est un ensemble fini ou bien une partie de l'ensemble des entiers naturels.

**Exemples** :

- le nombre d'enfants par femme ;

- le nombre de personnes par foyer ;

- le nombre de lettres dans une ligne ;

- le nombre de parts fiscales pour les impôts. Attention, pour cet exemple, les modalités ne sont pas des nombres entiers ;

- le nombre de mots mémorisés par des enfants pendant deux minutes parmi une liste de 50 mots ;

- le nombre de médailles remportées aux JO ;

- le nombre de villes de plus de 100 000 habitants ;

- le nombre de fautes dans une dictée...

• Lorsqu'il s'agit de la mesure d'une grandeur physique, on parlera de type **quantitatif continu**.

**Exemples** :

- la taille en cm ;

- le poids en kg ;

- l'âge en années ;

- les précipitations en mm...

**Remarque 1.4.** *1. Bien entendu, la notion de variable quantitative continue est théorique en ce sens que toute mesure quantitative est soumise à une imprécision plus ou moins importante : on donne par exemple l'âge d'un individu en années en précisant parfois le nombre de mois mais rarement de manière plus " fine " ; cependant le temps est un phénomène continu.*

*2.* ***Attention*** *la variable "début du numéro d'INSEE" qui associe à chaque individu d'une population de personnes 1 s'il s'agit d'un homme et 2 s'il s'agit d'une femme n'est pas quantitative : 1 et 2 ne mesurent rien du tout : il s'agit simplement d'un codage.*

L'ensemble des modalités d'une variable de type quantitatif forme ce que l'on appelle une **échelle d'intervalle** (expression employée particulièrement en Psychologie).

**Exemple A.**

*La variable "nombre d'enfants" est de type quantitatif discret (ou discrète).*

*Les variables "âge", "taille" sont de type quantitatif continu (ou continues).*

**Exemple B.**

*La variable "F" est de type quantitatif discret (ou discrète).*

*La variable "population" est de type quantitatif discret tandis que les variables "PNB global" et " PNB/habitant" sont de type quantitatif continu (ou continues).*

### 1.2.2 Le type qualitatif ordinal

Une variable est dite **ordinale** lorsque qu'elle n'est pas quantitative, mais que **ses modalités sont naturellement ordonnées.**

L'ensemble des modalités d'une variable de type ordinal forme ce que l'on appelle une **échelle ordinale** (expression employée particulièrement en Psychologie).

**Exemple A.** *Les variables "Diplôme le plus élevé" et "Goût pour la lecture" sont ordinales : les modalités ne sont pas "réellement" des nombres mais sont naturellement ordonnées.*

**Exemple B.** *La variable "Rang" est ordinale : ses modalités ne sont pas "réellement" des nombres, en effet, elles ne représentent qu'un classement, l'écart entre le troisième et le quatrième n'est pas forcément le même qu'entre le quatrième et le cinquième.*

*On pourrait construire d'autres variables ordinales en considérant par exemple "l'avis (majoritaire) de la population sur une Europe à 28" avec les modalités : très favorable, plutôt favorable, plutôt défavorable et très défavorable.*

**Exemple 1.3.** *a) Considérons sur une population de personnes, la variable "Taille", qui associe à chaque personne, l'un des trois adjectifs "petit", "moyen", "grand", suivant une convention établie à l'avance.*

*On a bien ici, un ordre naturel entre les trois modalités. La variable taille ainsi définie est donc une variable ordinale.*

*b) En prenant maintenant pour population l'ensemble des départements Français, on construit la variable "Numéro" qui associe à chaque département son numéro. Les modalités sont ici des nombres, mais il s'agit là d'un simple codage lié à l'ordre alphabétique des noms donnés aux départements.*

*(Il aurait suffit d'appeler autrement le département du Tarn pour que son numéro change.)*

*c) De même, la variable "début du numéro d'INSEE" n'est pas ordinale. En effet, il n'y a pas ici*

*d'ordre naturel entre les modalités "1" et "2" qui ne représentent que "homme" et "femme".*

*d) Prenons enfin comme population un ensemble de couleurs. Une personne classe les couleurs de la plus claire à la plus sombre. On peut alors considérer la variable "Nuance" qui associe à chaque couleur son numéro d'ordre dans le classement réalisé précédemment.*
*Les modalités sont ici naturellement ordonnées ; la variable Nuance est donc ordinale.*
*Il est à noter qu'ici la variable dépend de la personne ayant réalisé le classement. En effet, pour des nuances très proches deux personnes peuvent avoir des visions différentes. il s'agit donc d'un ordre naturel, mais pour une personne donnée.*

### 1.2.3 Le type qualitatif nominal

Une variable est dite **nominale** lorsque qu'elle n'est ni quantitative, ni ordinale, c'est-à-dire lorsque ses modalités sont des catégories non hiérarchisées. Chaque modalité est simplement désignée par son nom. Ainsi les variables "marque de voiture possédée", "début du numéro d'INSEE" sont nominales.

**Exemple A.** *Les variables " Sexe" et "Spécialité scolaire" sont nominales.*

**Exemple B.** *La variable "Membre" est nominale. Sur cette même population, on aurait pu considérer les variables nominales : "Langue officielle", "Monnaie", "Nature du régime".*

L'ensemble des modalités d'une variable de type nominal forme ce que l'on appelle une **échelle nominale** (expression employée particulièrement en Psychologie). Lorsque les modalités ne sont pas des nombres on dit également "variable qualitative nominale".

**Remarque 1.5.** *Dans le cas d'une variable qualitative, il est fréquent de remplacer les valeurs de la variable par des nombres ; on dit que l'on fait un* **codage**. *Il s'agit uniquement de faciliter le traitement (informatique notamment) de la variable, mais ces nombres n'ont aucune valeur numérique (en particulier cela n'a aucun sens d'envisager des opérations telles que l'addition). Par exemple dans le cas du sexe, l'INSEE code masculin par 1 et féminin par 2.*

**IMPORTANT : il est capital avant toute étude statistique de bien définir la population sur laquelle porte l'enquête et les variables avec leurs modalités et leur type. En effet, les traitements statistiques sont différents selon le type de la variable concernée.**

**Pour déterminer la population et la variable, on se pose la question A QUI (population) ON DEMANDE QUOI (variable) ?**

## 1.3 Effectifs et fréquences

### 1.3.1 Effectifs

On s'intéresse ici à une seule variable. Pour chaque modalité de la variable on compte le nombre d'individus ayant cette modalité. Le résultat obtenu s'appelle l'**effectif de la modalité**.

**Exemple A.** *On considère la variable "Spécialité scolaire" que l'on notera $X$. L'effectif de la modalité "L" est donc 10 (Il suffit de compter à partir du tableau de données).*
*En faisant cela pour chaque modalité de la variable, on peut construire un nouveau tableau appelé* **tableau d'effectifs**, *qui comporte une ligne par modalité :*
*(On ne fait pas figurer la modalité "T" car son effectif est 0.)*

| $x_i$ | $n_i$ |
|---|---|
| $L$ | 10 |
| $ES$ | 5 |
| $S$ | 5 |
| | $N = 20$ |

*L'en-tête de colonne $x_i$ signifie que $x_1$ désignera la modalité de la ligne 1 ($x_1$ =L), $x_2$ désignera la modalité de la ligne 2 ($x_2$ =ES), et ainsi de suite (i est donc le numéro de la ligne). Bien sûr, si on avait utilisé la lettre $Y$ pour nommer la variable, on aurait appelé $y_i$ les modalités.*
*Attention, l'ordre des modalités dans le tableau est totalement arbitraire. Cependant lorsqu'on a une variable ordinale ou quantitative, on respecte l'ordre natrurel des modalités. L'en-tête de colonne $n_i$ signifie que $n_1$ désignera l'effectif de la modalité $x_1$, et ainsi de suite.*
*($n_i$ désigne l'effectif de la modalité $x_i$). Ainsi $n_1 = 10$, $n_2 = 5$, $n_3 = 5$.*

La colonne intitulée $n_i$ est parfois globalement appelée **distribution des effectifs** de la variable $X$.

**Exemple B.** *Pour la variable $F$ dont les modalités sont des nombres entiers, l'effectif de la modalité 2 est le nombre d'individus (de pays) qui ont exactement 1 pays frontalier parmi les 27 autres. Ici, l'effectif est 7.*

**Remarque 1.6.** *L'intérêt du tableau d'effectifs est d'être beaucoup plus lisible en général que le tableau de données. (Ainsi, par exemple, avec une population de 100 personnes, la variable "loisir principal" comporterait toujours 4 modalités. On aurait donc au plus 4 lignes pour le tableau d'effectifs, alors que le tableau de données comporterait lui 100 lignes.)*

*Par contre, avec le tableau d'effectifs, on perd de l'information : ainsi par exemple on ne sait plus en regardant ce tableau quelle est la donnée de Victor pour la variable "Spécialité scolaire".*

*Conclusion : plus on "gagne" en clarté et plus on a tendance à "perdre" de l'information !*

### 1.3.2 Fréquences

L'effectif d'une modalité ne suffit pas à rendre compte de l'importance de cette modalité dans la population. Ainsi par exemple, la modalité "L" a pour effectif $n_1 = 10$ ; mais ici la population est de taille $N = 20$, l'importance de cette modalité est beaucoup plus grande que si la taille de la population était de 100. Il faut donc pour chaque modalité comparer $n_i$ et $N$.

On appellera **fréquence de la modalité** $x_i$, le quotient $\dfrac{n_i}{N}$, que l'on notera $f_i$ : $\boxed{f_i = \dfrac{n_i}{N}}$.

(C'est la première formule de statistique.)

Ainsi la fréquence de la modalité "ES" sera $f_2 = \dfrac{n_2}{N} = \dfrac{5}{20} = 0,25$.

(Le dernier signe "=" ci-dessus est en fait un abus d'écriture qu'on accepte dans ce type de situation.)

La colonne intitulée $f_i$, ajoutée au tableau précédent, est parfois globalement appelée **distribution des fréquences** de la variable $X$.

L'utilisation des pourcentages, rendant plus parlant les résultats de fréquences, on les utilisera donc parfois pour représenter les fréquences, d'où une colonne "pourcentage" que l'on peut rajouter au tableau précédent.

(On a donc par exemple l'égalité : $f_2 = 0,25 = 25\%$.)

| $x_i$ | $n_i$ | $f_i$ | *Pourcentage* |
|---|---|---|---|
| $L$ | 10 | 0,50 | 50,00% |
| $ES$ | 5 | 0,25 | 25,00% |
| $S$ | 5 | 0,25 | 25,00% |
| | $N = 20$ | 1 | 100,00% |

**Comme pour une variable on a une donnée et une seule par individu, la somme de la colonne "$n_i$" vaudra toujours $N$, et la somme de la colonne "$f_i$" vaudra toujours 1.**

**Exemple B.** *En reprenant la variable F, la fréquence de la modalité 2 est l'effectif de 2 divisé par la taille de la population (28). Ici, la fréquence de 2 est donc :* $\frac{7}{28} = 0,25$.

*Cette fréquence peut être exprimée en pourcentage :* $0,25 = \frac{25}{100} = 25\%$.

*En faisant de même pour chaque modalité de F, on peut remplir le tableau d'effectifs et de fréquences de F :*

| *Modalités* | *Effectifs* | *Fréquences* | *Pourcentages* |
|---|---|---|---|
| 1 | 2 | 0,0714 | 7,14% |
| 2 | 7 | 0,2500 | 25,00% |
| 3 | 7 | 0,2500 | 25,00% |
| 4 | 4 | 0,1429 | 14,29% |
| 5 | 5 | 0,1786 | 17,86% |
| 6 | 1 | 0,0357 | 3,57% |
| 7 | 1 | 0,0357 | 3,57% |
| 8 | 0 | 0,0000 | 0,00% |
| 9 | 1 | 0,0357 | 3,57% |
| | 28 | 1,0000 | 100,00% |

*Colonne Effectifs : Distribution des effectifs de X.*

*Colonne Fréquences : Distribution des fréquences de X.*

### 1.3.3 Effectifs et fréquences cumulés

Nous aurons besoin par la suite des effectifs cumulés et des fréquences cumulées dont voici les définitions :

• L'**effectif cumulé** noté $n_i^*$ de la modalité $i$ est la somme des effectifs des modalités qui lui sont inférieures ou égales.

• La **fréquence cumulée** noté $f_i^*$ de la modalité $i$ est la somme des fréquences des modalités qui lui sont inférieures ou égales.

**Exemple B.** *Calculons les effectifs cumulés pour la variable F. Nous obtenons*

| *Modalités* | *Effectifs* | *Effectifs cumulés* |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 7 | 9 |
| 3 | 7 | 16 |
| 4 | 4 | 20 |
| 5 | 5 | 25 |
| 6 | 1 | 26 |
| 7 | 1 | 27 |
| 8 | 0 | 27 |
| 9 | 1 | 28 |

*Notons que le dernier effectif cumulé est la taille de la population et naturellement la dernière fréquence cumulée est* 100%.

## 1.4 Regroupement en classes

Lorsqu'une variable a beaucoup de modalités (c'est souvent le cas avec les variables quantitatives), on est amené à regrouper de façon cohérente des modalités avant de faire un traitement statistique. On dit que l'on fait un **regroupement en classes**. L'**amplitude** d'une classe est la longueur de l'intervalle.

Un regroupement par classes doit vérifier :

- deux classes quelconques sont disjointes
- la réunion des classes recouvre l'ensemble des modalités.
- les classes n'ont pas forcément la même amplitude.

Pour les calculs ultérieurs, on considérera que la classe est représentée par son centre (le centre de la classe $]10\ ;\ 20]$ est $\dfrac{10+20}{2} = 15$).

**Exemple A.** *Si on considère la variable " âge" de l'exemple A., étant donné le nombre important de modalités, on décide de les regrouper en 5 classes : [22 ;24[, [24 ;26[, [26 ;28[, [28 ;30[, [30 ;32[ qui sont ici 5 intervalles d'âges (les deux nombres de chaque intervalle sont appelés "origine" et "extrémité" ou* **bornes** *de la classe : ce sont les valeurs minimales et maximales des classes).*
*On rappelle que lorsque le crochet est tourné vers l'intérieur de l'intervalle, la valeur correspondante est comprise dans l'intervalle, alors que lorsque le crochet est tourné vers l'extérieur la valeur est exclue.*

*Par exemple, [22 ; 24[ représente tous les âges possibles allant de 22 ans à moins de 24 ans. Attention, "moins de 24 ans" ne signifie pas que l'on s'arrête à 23 ans : ainsi la valeur 23,999 fait partie de l'intervalle même si cette valeur n'est pas observée ici.*

*Avec ce groupement en classes, on aura alors le tableau d'effectifs et de fréquences suivant :*

| *classes* | $n_i$ | $f_i$ | *Pourcentages* |
|---|---|---|---|
| $[22;24[$ | 5 | $0,25$ | $25,00\%$ |
| $[24;26[$ | 9 | $0,45$ | $45,00\%$ |
| $[26;28[$ | 2 | $0,10$ | $10,00\%$ |
| $[28;30[$ | 2 | $0,10$ | $10,00\%$ |
| $[30;32[$ | 2 | $0,10$ | $10,00\%$ |
| | $N = 20$ | 1 | $100\%$ |

Tableau 1 : Répartition des individus suivant leur âge

**Remarque 1.7.** *On ne fera désormais qu'une colonne pour fréquences et/ou pourcentages.*

**Exemple B.** *Considérons la variable PNB/h. On peut faire un tableau d'effectifs et de fréquences après regroupement en classes d'amplitude 10 (en milliers de dollars) en prenant comme première classe l'intervalle* $]0\ ;\ 10]$ *(les crochets indiquent que 10 est dans la classe* $]0\ ;\ 10]$ *et non dans la classe*$]10\,;\,20]$*).*

| *PNB/h* | *Effectifs* | *Fréquences* |
|---|---|---|
| $]0;10]$ | 12 | $42,86\%$ |
| $]10;20]$ | 6 | $21,43\%$ |
| $]20;30]$ | 8 | $28,57\%$ |
| $]30;40]$ | 1 | $3,57\%$ |
| $]40;50]$ | 1 | $3,57\%$ |
| | 28 | $100\%$ |

**Remarque 1.8.** *Un groupement en classes permet d'obtenir un tableau d'effectifs plus lisible car comportant moins de lignes ; mais bien sûr, on perd encore une fois de l'information. Il faut bien comprendre que bien que les tableaux ci-dessus présentent des classes, les modalités de la variable sont toujours des nombres et pas des intervalles.*

**Exemple B.** *Considérons de nouveau la variable PNB/h et calculons ses effectifs cumulés.*

| *PNB/h* | *Effectifs* | *Effectifs cumulés* |
|---|---|---|
| ]0; 10] | 12 | 12 |
| ]10; 20] | 6 | 18 |
| ]20; 30] | 8 | 26 |
| ]30; 40] | 1 | 27 |
| ]40; 50] | 1 | 28 |

***L'effectif cumulé d'une classe est en fait l'effectif cumulé de la borne supérieure de l'intervalle représentant la classe.*** *Par exemple, 18 est l'effectif cumulé de* ]10 ; 20] *signifie que parmi les 28 pays, 18 ont un PNB/hab inférieur à 20000$.*

*Pour terminer cette section, on reprend l'exemple A. et on établit, pour chaque variable, le tableau des effectifs et des fréquences, avec regroupement en classes dans le cas de la taille.*

**Exemple A.**

| $x_j$ | $n_j$ | $f_j$ |
|---|---|---|
| *Aucun* | *3* | *0,15* |
| *Brevet* | *4* | *0,2* |
| *Baccalauréat* | *5* | *0,25* |
| *Diplôme supérieur* | *8* | *0,4* |
| | *N=20* | *1* |

*Tableau 2 : Répartition des individus suivant leur diplôme*

| $x_j$ | $n_j$ | $f_j$ |
|---|---|---|
| *1* | *12* | *0,6* |
| *2* | *8* | *0,4* |
| | *N=20* | *1* |

*Tableau 3 : Répartition des individus suivant leur sexe (1 : masculin, 2 : féminin)*

| $x_j$ | $n_j$ | $f_j$ |
|---|---|---|
| *0* | *6* | *0,3* |
| *1* | *7* | *0,35* |
| *2* | *5* | *0,25* |
| *3* | *2* | *0,1* |
| | *N=20* | *1* |

*Tableau 4 : Répartition des individus suivant le nombre d'enfants*

| $[b_j; b_{j+1}[$ | $n_j$ | $f_j$ |
|---|---|---|
| $[150; 160[$ | *3* | *0,15* |
| $[160; 170[$ | *7* | *0,35* |
| $[170; 180[$ | *6* | *0,3* |
| $[180; 190[$ | *4* | *0,2* |
| | *N=20* | *1* |

*Tableau 5 : Répartition des individus suivant leur taille*

| $x_j$ | $n_j$ | $f_j$ |
|---|---|---|
| *Faible* | *4* | *0,20* |
| *Moyen* | *4* | *0,20* |
| *Fort* | *12* | *0,60* |
| | *N=20* | *1* |

*Tableau 6 : Répartition des individus suivant leur goût pour la lecture*

## 1.5 Représentations graphiques

*Une fois le tableau statistique établi, on cherche à rendre celui-ci plus "lisible" ou, en d'autres termes, à représenter les informations qu'il contient sous forme de graphiques. C'est un procédé largement utilisé dans tous les médias. Ces graphiques ont pour seul but de représenter de manière plus attrayante le tableau statistique : à partir de chaque graphique on peut reconstruire le tableau statistique (il n'y a donc pas de perte d'information). Les représentations graphiques sont diverses et dépendent principalement du type de variable étudiée.*

### 1.5.1 Représentations des variables quantitatives discrètes

**Les diagrammes en bâtons**

*Les* **diagrammes en bâtons** *servent à représenter les effectifs ou les fréquences de l'ensemble des modalités d'une variable* **quantitative discrète**.

*Pour tracer un diagramme en bâtons, on choisit tout d'abord deux axes perpendiculaires et une échelle pour chacun de ces axes. L'axe des abscisses (ou axe horizontal) sert à porter les modalités de la variable et l'axe des ordonnées (axe vertical) est l'axe des effectifs ou des fréquences suivant le cas. Il suffit ensuite de tracer en chaque modalité un trait vertical (bâton) dont la hauteur correspond à la valeur de l'effectif ou de la fréquence.*

**Remarque 1.9.** *a) Lorsqu'on positionne les modalités sur l'axe des abscisses, il faut, bien en-*

*tendu, respecter l'échelle choisie pour cet axe. Par exemple, si les modalités sont 1, 2 et 4, l'espacement entre 1 et 2 est d'une unité alors qu'il est de deux unités entre 2 et 4.*
*b) Les diagrammes en bâtons des effectifs et des fréquences d'une même variable diffèrent simplement par l'échelle des ordonnées : on passe, par exemple, du diagramme en bâtons des effectifs au diagramme en bâtons des fréquences en divisant l'échelle des ordonnées par N (taille de l'échantillon).*

**Exemple A.** *Si on considère la variable nombre d'enfants, le tableau 4 nous donne les deux graphiques suivants :*

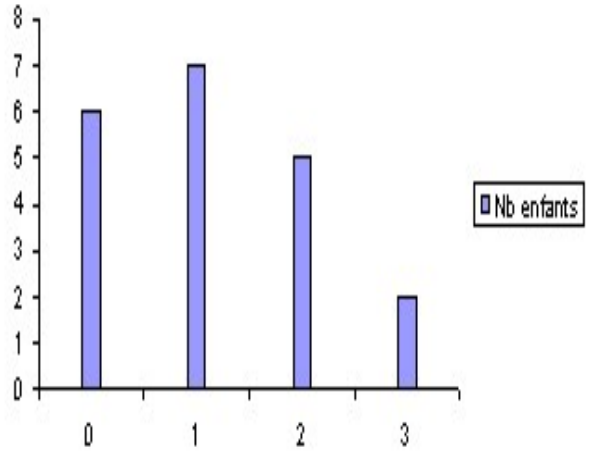



FIGURE 1.1 – *Diagramme en bâtons des effectifs de la variable nb d'enfants de l'exemple A.*

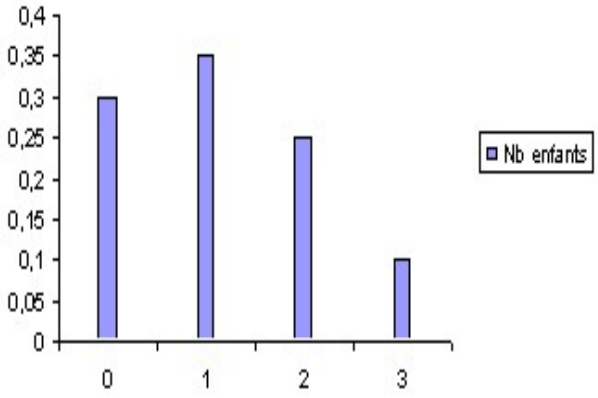



FIGURE 1.2 – *Diagramme en bâtons des fréquences de la variable nb d'enfants de l'exemple A.*

**Exemple B.**

### 1.5.2 Représentations des variables quantitatives continues

**Les histogrammes**

*Les* **histogrammes** *servent à représenter les effectifs ou les fréquences d'une variable* **quanti-**

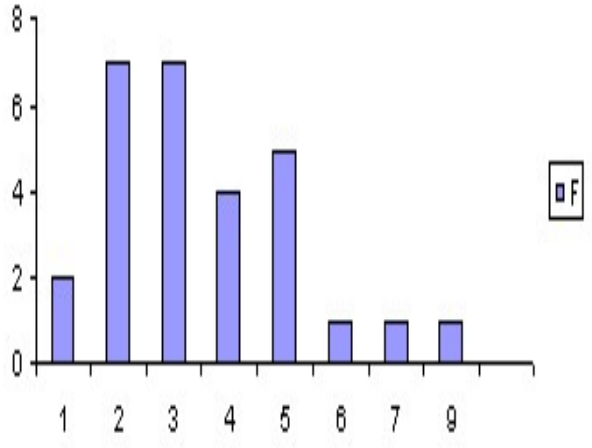



FIGURE 1.3 – *Diagramme en bâtons des effectifs de la variable F de l'exemple B.*

**tative continue**.

*Comme pour les diagrammes précédents, on choisit, pour tracer un histogramme, deux axes perpendiculaires et une échelle pour chacun. Sur l'axe des abscisses (axe horizontal) sont portées les valeurs de la variable, c'est-à-dire les différentes classes de cette variable. Une fois choisie une échelle pour cet axe, les positions des bornes de classes doivent respecter cette échelle. Sur l'axe des ordonnées sont portées les valeurs des effectifs ou des fréquences. En face de chaque classe, on trace un "rectangle" dont la hauteur est égale à la* **densité d'effectifs** *(ou à la* **densité de fréquences***) de cette classe qui vaut* $\frac{n_j}{b_{j+1}-b_j}$ *si* $b_j$ *et* $b_{j+1}$ *sont les bornes de la classe et* $n_j$ *son effectif (ou* $\frac{f_j}{b_{j+1}-b_j}$ *si* $f_j$ *est la fréquence de la classe).*

*Avant de tracer l'histogramme, il convient donc de calculer au préalable les densités d'effectifs et/ou de fréquences. Prenons un exemple simple pour illustrer et expliquer le principe de construction de l'histogramme (des effectifs). Considérons une variable quantitive continue et deux de ses classes, soient* $[5\ ;\ 6[$ *et* $[6\ ;\ 8[$*. La première a une amplitude de 1 (une unité) et la seconde une amplitude de 2 (deux unités). Supposons que, dans notre échantillon, on ait 10 individus dans chaque classe. Le "rectangle" correspondant à la classe* $[5\ ;\ 6[$ *a une hauteur de 10 sur l'axe des ordonnées. On voit très facilement que tracer un rectangle de la même hauteur pour la classe* $[6\ ;\ 8[$ *ne conviendrait pas et conduirait à une interprétation fausse. En effet, si on partageait alors dans le sens de la hauteur ce rectangle en deux parties égales, on obtiendrait deux nouveaux rectangles ayant une hauteur égale à 10 et correspondant aux classes* $[6\ ;\ 7[$ *et* $[7\ ;\ 8[$*. Cela indiquerait, suivant ce graphique, que chacune de ces classes possède 10 individus soit un total de 20 individus pour la classe* $[6\ ;\ 8[$ *! On voit bien que l'on arrive à une absurdité en procédant ainsi.*

*Au contraire, en choisissant de diviser l'effectif par 2 et de tracer un rectangle d'une hauteur de 5 pour la classe* [6 ; 8[*, on fait de manière implicite, l'hypothèse que les individus sont "uniformément" répartis dans la classe* [6 ; 8[ *: suivant ce principe, il y en aurait le même nombre dans chaque "sous-classe"* [6 ; 7[ *et* [7 ; 8[*, soit 5. Le graphe devient ainsi cohérent et interprétable : dans cet exemple simple, on pourra observer qu'il y a une "baisse" d'effectifs entre les classes* [5 ; 6[ *et* [6 ; 8[*. Ces arguments sont bien entendu valables pour l'histogramme des fréquences.*

**Remarque 1.10.** *Le choix de l'amplitude et de la position des classes est un problème important en pratique : deux choix distincts de classes peuvent conduire à des histogrammes d'allures très différentes. Même si ce problème sort du cadre de ce cours, il faut noter que des statisticiens ont proposé des méthodes de choix des classes (position et amplitude)* **automatiques**, *c'est-à-dire des méthodes (dites* **data-driven***) qui ne reposent pas sur un choix " subjectif " mais sur les observations elles-mêmes.*

**Exemple A.** *La variable taille a été divisée en classes d'amplitude constante 10cm. Cf. Figure 1.4.*

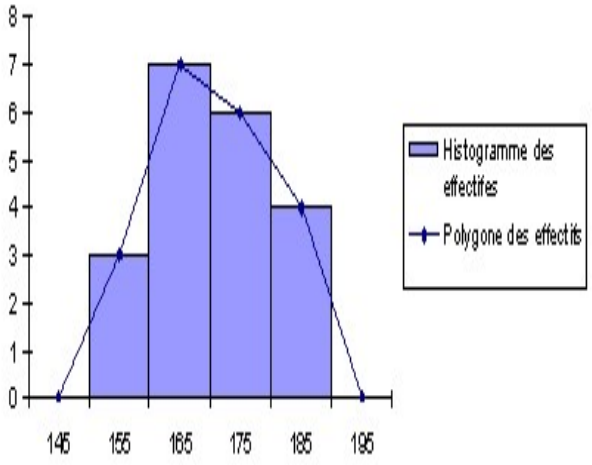



FIGURE 1.4 – *Histogramme et polygône des effectifs de la variable taille.*

*Considérons à nouveau la variable taille mais cette fois avec le regroupement en classes suivant* [157; 160[, [160; 166[, [166; 176[, [176; 184[ *et* [184; 185[.

| *Classe* | $n_i$ | *amplitude* : $a_i$ | *Hauteur du rectangle* : $10\frac{n_i}{a_i}$ |
|---|---|---|---|
| [157; 160[ | *3* | *3* | *10* |
| [160; 166[ | *4* | *6* | *6,67* |
| [166; 176[ | *8* | *10* | *8* |
| [176; 184[ | *4* | *8* | *5* |
| [184; 185[ | *1* | *1* | *10* |
| | $N = 20$ | | |

*Cela donne l'histogramme en Figure 1.5.*

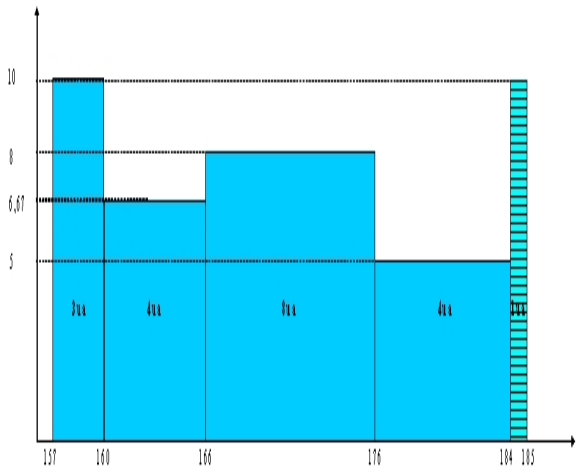



FIGURE 1.5 – *Histogramme des effectifs de la variable taille avec un regroupement par classes différent.*

**Exemple B.** *Pour la variable PNB/hab avec des classes d'amplitude constante de 10, on obtient la Figure 1.6.*

**Le polygône des effectifs**

*On adjoint parfois à l'histogramme le* **polygône des effectifs** *ou des fréquences suivant le cas. Celui-ci est la ligne brisée qui joint le centre des sommets de chaque rectangle à laquelle on ajoute deux segments : l'un joignant le centre du sommet du premier rectangle au point de l'axe des abscisses se situant à une demi-amplitude de la première classe, l'autre joignant le centre du sommet de la dernière classe au point de l'axe des abscisses se situant à une demi-amplitude de la dernière classe.*

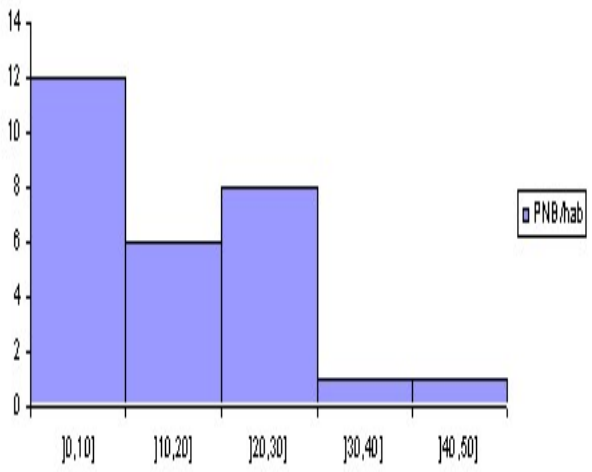



FIGURE 1.6 – *Histogramme des effectifs de la variable PNB/h.*

**Exemple A.** *Cf. Figure 1.4.*

**Les diagrammes cumulatifs**

*Dans le cas des variables quantitatives continues avec regroupement en classes de même amplitude, il est possible de représenter les effectifs ou fréqences cumulés par des histogrammes cumulatifs et le polygône cumulatif. Au lieu de reporter sur l'axe des ordonnées les effectifs ou les fréquences comme pour les histogrammes traditionnels, on reporte les effectifs ou fréquences cumulés. On trace ensuite le polygône cumulatif de la même façon que précédemment.*

**Exemple A.** *On considère la variable taille en classes d'amplitude constante 10cm. La Figure 1.7. représente les effectifs cumulés et le polygône cumulatif.*

***Attention*** *: Dans le cas des variables quantitatives continues avec regroupement en classes d'amplitudes différentes, un tel type de graphique n'est pas possible. Cependant, en se rappelant que l'effectif cumulé d'une classe est en fait l'effectif cumulé de la borne supérieure de l'intervalle, on peut tracer le polygône cumulatif.*

**Exemple A.** *Considérons à nouveau la variable taille mais avec le regroupement en classes suivant* $[157; 160[$, $[160; 166[$, $[166; 176[$, $[176; 184[$ *et* $[184; 185[$. *Cf. Figure 1.8.*

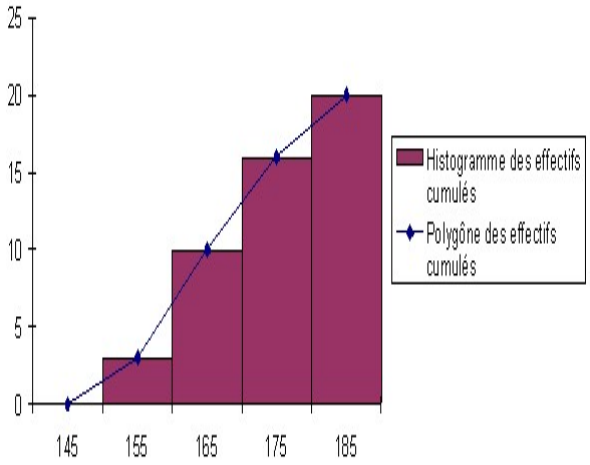



FIGURE 1.7 – *Histogramme et polygône des effectifs cumulés de la variable taille.*

### 1.5.3 Représentations graphiques de variables qualitatives

**Les diagrammes en colonnes**

*Les* **diagrammes en colonnes** *servent à représenter les effectifs ou les fréquences d'une variable* **qualitative**.

*Pour tracer un diagramme en colonnes, dans un repère dont les axes sont orthogonaux, on gradue l'axe vertical en partant de 0 pour le point d'intersection des 2 axes. Chaque modalité de la variable est représentée par un rectangle dont la base est située sur l'axe horizontal et dont la hauteur est égale à l'effectif de la modalité. (La variable étant nominale, il n'y a aucun ordre privilégié dans la disposition des rectangles représentant les modalités ni aucune contrainte dans la largeur de chaque rectangle, cependant l'usage est de donner à chaque rectangle la même largeur). L'axe horizontal n'est pas muni d'une échelle, puisque les modalités n'ont pas ici de valeurs numériques. Les modalités sont régulièrement espacées sur cet axe. L'axe des ordonnées (axe vertical) est l'axe des effectifs ou des fréquences suivant le cas. En face de chaque modalité figure une colonne (un rectangle) dont la hauteur correspond à la valeur de l'effectif ou de la fréquence.*

**Exemple A.** *Considérons la variable diplôme. Cette variable a 4 modalités. Le diagramme en colonnes comprend donc 4 colonnes de hauteurs respectives 3, 4, 5, 8.*

**Remarque 1.11.** *On emploie très souvent abusivement le mot "histogramme" à la place de "diagramme en colonnes", en particulier dans de nombreux logiciels informatiques.*

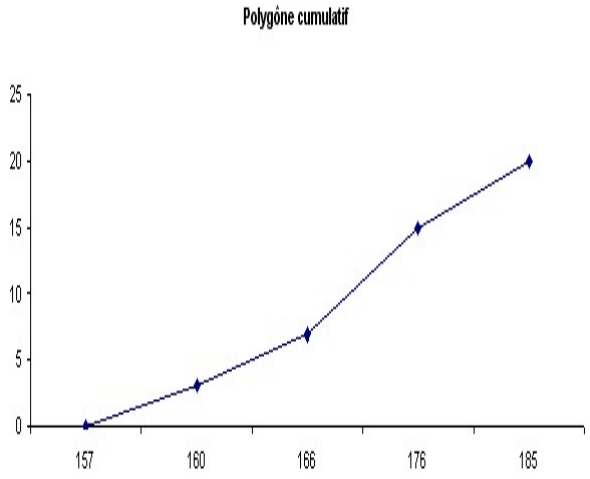



FIGURE 1.8 – *Polygône des effectifs cumulés de la variable taille.*

**Les diagrammes en secteurs et les diagrammes en barre**

*Les* **diagrammes en secteurs** *et les* **diagrammes en barre** *servent (comme les diagrammes en colonnes) à représenter les effectifs ou les fréquences d'une variable* **qualitative**. *Les diagrammes en secteurs sont plus souvent appelés* **camemberts**.

*Les diagrammes en secteurs se présentent sous la forme d'un disque (ou d'un demi-disque) divisé en k secteurs (k étant le nombre de modalités de la variable) : l'angle (ou l'aire ce qui revient au même) de chaque secteur est proportionnel à l'effectif ou à la fréquence de la modalité qu'il représente. Il suffit donc de construire le tableau de proportionnalité consistant à passer de la colonne des fréquences à la colonne des angles (en degrés) en multipliant par 360. (L'angle correspondant à un disque complet ayant pour mesure 360 degrés).*

**Attention :** *lorsque les valeurs $f_i$ sont arrondies, on utilise, pour calculer l'angle correspondant, non pas la valeur arrondie, mais la valeur "exacte" mise en mémoire dans la calculatrice. Pour tout calcul, il faudra procéder ainsi afin de ne pas cumuler les erreurs d'arrondis : on donne comme résultat une valeur arrondie, mais on travaille avec la valeur exacte lorsqu'on la réutilise.*

**Exemple A.** *Reprenons l'exemple de la variable qualitative diplôme et traçons le diagramme en secteurs pour cette variable.*

**Exemple B.** *Reprenons l'exemple de la variable ordinale frontière et traçons le diagramme en secteurs pour cette variable.*

*Les* ***diagrammes en barre*** *sont construits sur le même principe mais sous la forme d'un rec-*

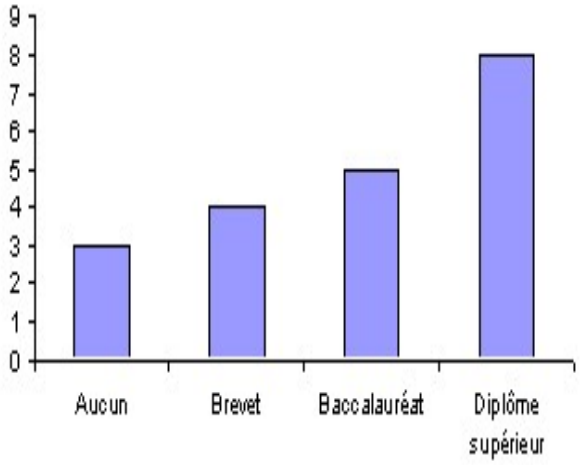



FIGURE 1.9 – *Diagramme en colonnes des effectifs de la variable Diplôme.*

*tangle divisé en k sous-rectangles dont les aires sont proportionnelles aux effectifs ou fréquences des modalités qu'ils représentent.*

## 1.6 Un premier indice de tendance centrale : le mode

*De façon générale, on appelle* **résumé** *(ou* **indice***) d'une variable, un mot, une valeur (pas nécessairement numérique) qui représente globalement la variable.*
*Un exemple très connu de résumé est la moyenne, mais ce résumé n'a évidemment de sens que pour une variable quantitative.*
*On appelle* **mode** *d'une variable qualitative ou quantitative discrète toute modalité ayant le plus grand effectif (et donc également la plus grande fréquence). Le mode fait partie des résumés que l'on appelle "indice de tendance centrale" (ils donnent une idée globale des données de la variable). De plus, dans le cas d'un regroupement en classes, on parlera de* ***classe modale*** *: la classe modale de la variable âge dans l'exemple A. est* $[24; 26[$.

**Remarque 1.12.** *a) La notion de mode est relativement simple : elle indique les valeurs de la variable les plus "présentes" dans l'échantillon.*
*b) Une série statistique peut avoir plusieurs modes. Dans le cas d'une série ayant deux modes, on parle de série statistique* **bimodale**.

**Exemple A.** *Si on reprend la variable nombre d'enfants, le mode est égal à 1 : l'effectif de cette modalité est de 7. Si on prend la variable âge, le mode est dans ce cas 25.*

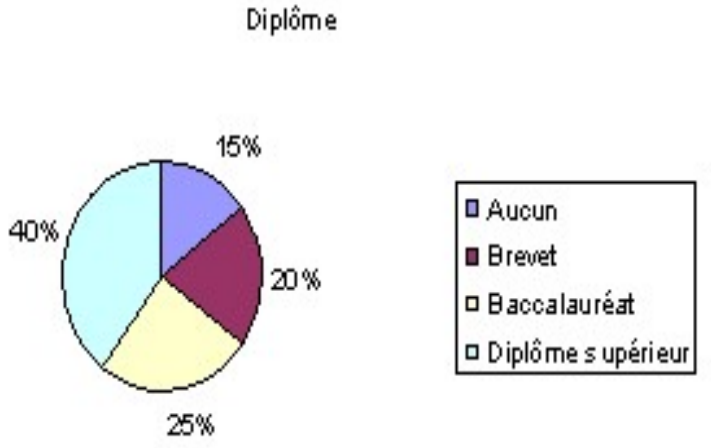



FIGURE 1.10 – *Diagramme en secteurs des fréquences pour la variable Diplôme.*

**Exemple B.**

- *Le mode de la variable "Membre" est OUI, c'est-à-dire tout simplement qu'il y a plus de pays membres de l'U.E. que de pays candidats.*
- *Les modes de la variable F sont 2 et 3 (et non pas 7).*
- *Toute modalité de la variable "rang" est mode : ça ne présente aucun intérêt.*
- *Lorsqu'une variable quantitative continue est regroupée en classes de même amplitude, la classe modale est celle qui a le plus grand effectif : la classe modale de la variable PNB/habitant regroupée en classes d'amplitude 10 (milliers de dollars) est l'intervalle* $]0;10]$.

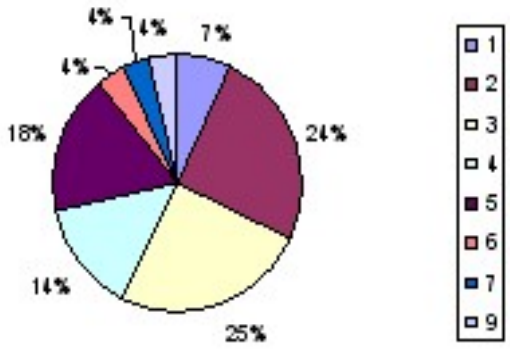



FIGURE 1.11 – *Diagramme en secteurs de la variable F.*

# Chapitre 2

# Médiane et autres quantiles des variables ordinales

**Remarque 2.1.** *1. Rappelons qu'une variable est dite* **ordinale** *lorsque ses modalités peuvent être naturellement ordonnées.*
*2. Les variables quantitatives pouvant être traitées comme ordinales, ce chapitre concerne aussi les variables quantitatives.*

## 2.1 Un indice de tendance centrale : la médiane

### 2.1.1 Principe général

*Étant donnée une variable ordinale sur une population $\Omega$ de taille $N$, on cherche la donnée (que l'on appellera "médiane") située au milieu de la liste des données écrites par ordre croissant.*
*On a ainsi constitué 2 groupes de données de même taille : celui dont les données sont inférieures (ou égales) à la médiane et celui dont les données sont supérieures (ou égales) à la médiane.*
**La médiane sera la donnée située "au milieu" de la liste des données écrites par ordre croissant.**
*C'est la modalité qui partage la population en deux parties égales :*

- *L'une présentant des modalités inférieures à la médiane*
- *L'autre présentant des modalités supérieures.*

*Les méthodes de détermination de la médiane ne sont pas les mêmes selon les types de variables bien que l'objectif reste le même : trouver la modalité Méd telle que la moitié des individus prennent des valeurs supérieures à Méd et l'autre moitié des valeurs inférieures à Méd.*

### 2.1.2 Un exemple introductif

*Premier cas.*

*On considère une population composée de 11 personnes et la variable âge qui associe à chaque individu son âge en nombre d'années révolues.*

*Voici la liste des âges :*

$$8 \quad 9 \quad 15 \quad 20 \quad 21 \quad 12 \quad 15 \quad 16 \quad 22 \quad 25 \quad 13$$

*On cherche maintenant l'âge situé au milieu de la liste des données. Cet âge va permettre de partager les données en deux groupes de même taille, un groupe dont les données sont inférieures à cet âge et un groupe dont les données sont supérieures.*

*Il suffit donc pour déterminer cet âge (appelé âge médian), d'écrire la liste des données par ordre croissant :*

$$8 \quad 9 \quad 12 \quad 13 \quad 15 \quad \boxed{15} \quad 16 \quad 20 \quad 21 \quad 22 \quad 25$$

*Il y a ici 11 données, la donnée située au milieu de cette liste est donc la 6ième donnée : c'est donc 15 ans ( le deuxième 15 de notre liste).*

*On a donc bien 5 données avant l'âge médian et 5 données après l'âge médian. Les données sont donc bien partagées en 2 groupes de même taille (si on ne tient pas compte de la valeur 15 retenue pour médiane).*

*On peut remarquer que dans cet exemple, le nombre de données est impair, ce qui a joué un rôle important dans la détermination de la médiane.*

*Deuxième cas.*

*On considère maintenant une population composée de 12 personnes et toujours la même variable âge. Voici la liste des données (des âges) écrites par ordre croissant :*

$$8 \quad 12 \quad 14 \quad 15 \quad 16 \quad \boxed{17} \quad \boxed{18} \quad 18 \quad 19 \quad 19 \quad 20 \quad 20$$

*On cherche de nouveau l'âge situé au milieu de la liste des données. Il y a ici 12 données.*

*On n'a donc pas une donnée située au milieu, il faut en prendre deux : la 6-ième et la 7-ième.*

*On a donc le choix pour la médiane entre 17 ans et 18 ans.*

*On décide alors de prendre la donnée dont le rang est immédiatement après* $\frac{N}{2}$.

*Ici* $\frac{N}{2} = \frac{12}{2} = 6$, *donc la médiane sera la 7-ième donnée, c'est-à-dire 18 ans.*

*Ce choix pouvant s'étendre sans difficulté au cas où $N$ est impair, on décide de le prendre pour définition.*

*Étant donnée une variable ordinale ou quantitative discrète sur une population $\Omega$ de taille $N$, on appelle* **médiane** *(notée Méd) la donnée dont le rang est situé immédiatement après $\frac{N}{2}$ dans la liste des données écrites par ordre croissant.*

**Remarque 2.2.** *a) On peut voir que cela ne change rien pour le cas impair. En effet, si on reprend le premier cas, $\frac{N}{2} = \frac{11}{2} = 5,5$, donc la médiane est la 6-ième donnée, dans la liste des données écrites par ordre croissant. Donc, Méd = 6-ième donnée = 15.*

*b) Le nombre de données inférieures ou égale à la médiane est bien $\frac{N}{2}$, à un arrondi près, de même que le nombre de données supérieures ou égales à la médiane.*

*c) Lorsque la variable étudiée est quantitative et $N$ pair (comme pour le deuxième cas), certains auteurs prennent pour médiane la moyenne arithmétique des deux données situées au milieu de la liste des données écrites par ordre croissant. Nous n'avons pas retenu ce choix car il ne s'applique pas aux variables ordinales en général et de plus cette définition de la médiane ne pourrait pas se généraliser simplement aux situations que nous allons aborder ensuite.*

*Il n'est évidemment pas possible, lorsque la taille la population augmente, d'écrire la liste explicitement la liste des données. On a donc besoin d'une technique plus pratique pour trouver la médiane.*

### 2.1.3 Utilisation du tableau d'effectifs pour déterminer la médiane

**Cas des variables qualitatives ordinales et des variables quantitatives discrètes**

*Pour déterminer la médiane, on calcule d'abord les effectifs cumulés (ou les fréquences cumulées). Puis, on lit la valeur de la médiane dans le tableau des effectifs cumulés ou on en détermine une approximation graphiquement sur le diagramme des effectifs cumulés (ou des fréquences cumulées) en traçant la droite (horizontale) passant par l'effectif $\frac{N}{2}$ (ou la fréquence 0,5).*

*Afin de mieux comprendre, considérons les cas suivants :*

▷ *Cas a. Une population est composée de $N = 61$ personnes et toujours la même variable*

*âge notée $X$. Les données sont présentées ici avec le tableau d'effectifs.*

| $x_i$ | $n_i$ |
|---|---|
| 11 | 15 |
| 13 | 10 |
| 14 | 10 |
| 17 | 5 |
| 18 | 5 |
| 19 | 8 |
| 23 | 8 |
| | $N = 61$ |

*Ici* $\frac{N}{2} = \frac{61}{2} = 30,5$ *; la médiane est donc la 31-ième donnée, dans la liste des données écrites par ordre croissant (Méd = 31-ième donnée). Afin de déterminer la 31-ième donnée, il faut donc ajouter les effectifs. Cela nous amène à rajouter la colonne des effectifs cumulés (notés* $n_i^*$ *)*

| $x_i$ | $n_i$ | $n_i^*$ |
|---|---|---|
| 11 | 15 | 15 |
| 13 | 10 | 25 |
| 14 | 10 | 35 |
| 17 | 5 | 40 |
| 18 | 5 | 45 |
| 19 | 8 | 53 |
| 23 | 8 | 61 |
| | $N = 61$ | |

$n_2^*$ *est le nombre d'individus ayant une modalité inférieure ou égale à* $x_2$ *(c'est-à-dire à 13 ans) d'où* $n_2^* = 15 + 10 = 25$*. De façon générale,* $n_i^*$ *est le nombre d'individus ayant une modalité inférieure ou égale à* $x_i$*.*

*On peut également lire les effectifs cumulés de la façon suivante :*

| $x_i$ | $n_i$ | $n_i^*$ | *Interprétation* |
|---|---|---|---|
| *11* | *15* | *15* | *de la 1ère à la 15ième la donnée est 11 ans* |
| *13* | *10* | *25* | *de la 16ième à la 25ième la donnée est 13 ans* |
| *14* | *10* | *35* | **de la 26ième à la 35ième la donnée est 14 ans** |
| *17* | *5* | *40* | *de la 36ème à la 40ième la donnée est 17 ans* |
| *18* | *5* | *45* | *de la 41ième à la 45ième la donnée est 18 ans* |
| *19* | *8* | *53* | *de la 46ième à la 53ième la donnée est 19 ans* |
| *23* | *8* | *61* | *de la 54ième à la 61ième la donnée est 23 ans* |
| | *N=61* | | |

*D'après la troisième ligne du tableau ci-dessus, de la 26ième à la 35ième la donnée est 14 ans. Donc, Méd = 31ième donnée = 14 ans.*

▷ *Cas b. Considérons maintenant une population composée de* $N = 70$ *personnes et toujours la même variable âge notée* $X$ *dont le tableau d'effectifs complété par la colonne des effectifs cumulés est :*

| $x_i$ | $n_i$ | $n_i^*$ |
|---|---|---|
| 8 | 8 | 8 |
| 9 | 7 | 15 |
| 11 | 20 | 35 |
| 45 | 12 | 47 |
| 46 | 10 | 50 |
| 47 | 13 | 70 |
| | N=70 | |

*Ici* $\frac{N}{2} = \frac{70}{2} = 35$. *La médiane est donc la 36ième donnée, dans la liste des données écrites par ordre croissant, d'où : Méd = 36ième donnée = 45 ans.*

▷ *Cas c. Parmi les 70 personnes de l'exemple précédent, une des personnes de 47 ans se retire.*

*On obtient alors une nouvelle population de 69 personnes pour laquelle on va de nouveau déterminer l'âge médian. La situation n'ayant presque pas changé par rapport au cas b., on peut s'attendre à trouver à peu près la même médiane.*

*Le tableau d'effectifs complété par la colonne des effectifs cumulés est :*

| $x_i$ | $n_i$ | $n_i^*$ |
|---|---|---|
| 8 | 8 | 8 |
| 9 | 7 | 15 |
| 11 | 20 | 35 |
| 45 | 12 | 47 |
| 46 | 10 | 50 |
| 47 | 12 | 69 |
| | N=69 | |

*Ici* $\frac{N}{2} = \frac{69}{2} = 34,5$ *; la médiane est donc la 35ième donnée, dans la liste des données écrites par ordre croissant, d'où : Méd = 35ième donnée = 11 ans.*

*Contrairement à ce qui était attendu, l'âge médian est donc très différent de l'exemple précédent. Cela vient du fait que la population, dans les cas b. et c. n'est pas du tout homogène du point de vue de l'âge, mais est au contraire composée de deux sous populations, une d'enfants et une d'adultes, toutes deux sensiblement de même taille.*

*Dans ce genre de situation, il n'est en fait pas judicieux de calculer la médiane car elle n'est pas représentative de la situation. Il serait préférable, par exemple, de calculer les médianes de chacune des deux sous populations. Pour le cas b., la médiane du groupe des enfants est 11 et celle du groupe des adultes 46.*

**Remarque 2.3.** *a) De la même façon et pour les variables quantitatives continues, on détermine la* ***classe médiane****. Nous verrons plus loin comment calculer la médiane pour de telles variables.*

*b) La médiane est un indice très peu connu du public, aussi la plupart des gens l'assimilent à la moyenne. Cela peut donner naissance à des cas de manipulation de l'information. Ainsi par exemple, si dans une entreprise le salaire médian est plus élevé que le salaire moyen, un chef d'entreprise pourra-t-il être tenté dans un exposé de parler du salaire médian, sachant que ses auditeurs feront la confusion avec le salaire moyen.*

*c) La médiane peut se calculer pour toute variable ordinale tandis que la moyenne ne se calcule que pour les variables quantitatives (discrètes ou continues).*

*d) Un autre avantage de la médiane est d'être insensible aux valeurs extrêmes. Ainsi, en cas d'erreur lors d'une expérience, sur une valeur très grande (ou très petite), la médiane ne sera pas perturbée. Par exemple, si dans le cas a., on avait noté par erreur 51 ans pour une des personnes de 23 ans, la médiane aurait été inchangée. De la même façon, un individu ayant une donnée exceptionnelle n'influera pas sur la médiane alors qu'il aurait eu une grande influence sur la moyenne.*

*e) Il est inutile de calculer à la fois les effectifs cumulés et les fréquences cumulées ; il faut choisir.*

**Exemple A.** *Considérons la variable nombre d'enfants :*

| $x_i$ | $n_i$ | $n_i^*$ |
|---|---|---|
| 0 | 6 | 6 |
| 1 | 7 | 13 |
| 2 | 5 | 18 |
| 3 | 2 | 20 |
| | N=20 | |

*La taille de l'échantillon est de 20. Les effectifs cumulés 10 et 11 n'apparaissent pas dans le tableau : le premier effectif cumulé supérieur à 11 est 13. La médiane est donc la valeur correspondant à cet effectif cumulé c'est-à-dire* $Méd = 1$.

**Exemple B.** *Considérons les variables F*

| *Modalités de F* | *Effectifs* | *Effectifs cumulés* | *Fréquences* | *Fréquences cumulées* |
|---|---|---|---|---|
| 1 | 2 | 2 | 7,14% | 7,14% |
| 2 | 7 | 9 | 25,00% | 32,14% |
| 3 | 7 | 16 | 25,00% | 57,14% |
| 4 | 4 | 20 | 14,29% | 71,43% |
| 5 | 5 | 25 | 17,86% | 89,29% |
| 6 | 1 | 26 | 3,57% | 92,86% |
| 7 | 1 | 27 | 3,57% | 96,43% |
| 8 | 0 | 27 | 0,00% | 96,43% |
| 9 | 1 | 28 | 3,57% | 100,00% |
| | $N = 28$ | – | 100,00% | |

*Pour la variable F,* $\frac{N}{2} = \frac{28}{2} = 14$ .

*Dans le tableau des effectifs cumulés, 16 est le premier effectif cumulé qui dépasse 14 (*$57,14\%$ *est la première fréquence cumulée qui dépasse* $50\%$*). La médiane est donc la valeur 3. Concrètement, cela signifie que dans cette population de 28 pays la moitié ont 2 pays frontaliers (parmi les 27 autres) ou moins, la moitié en ont 2 ou plus.*

*On peut aussi utiliser le diagramme des effectifs cumulés :*

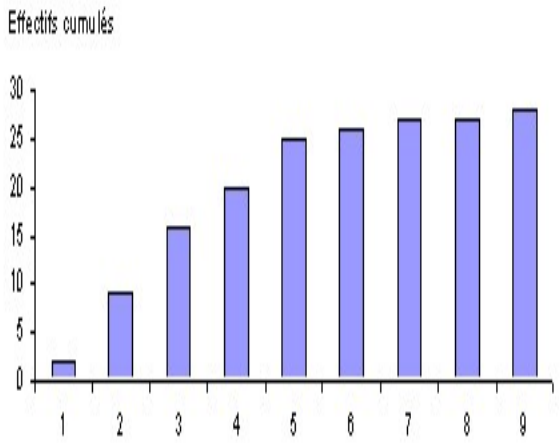



FIGURE 2.1 – *Diagramme des effectifs cumulés de F.*

**Cas d'une variable quantitative $X$ continue regroupée en classes de même amplitude**

*Prenons l'exemple de la variable PNB/h : on peut lire une valeur approximative de la médiane en utilisant le diagramme cumulatif suivant :*

*On lit sur l'axe des abscisses la médiane :* $Méd \approx 13$ .

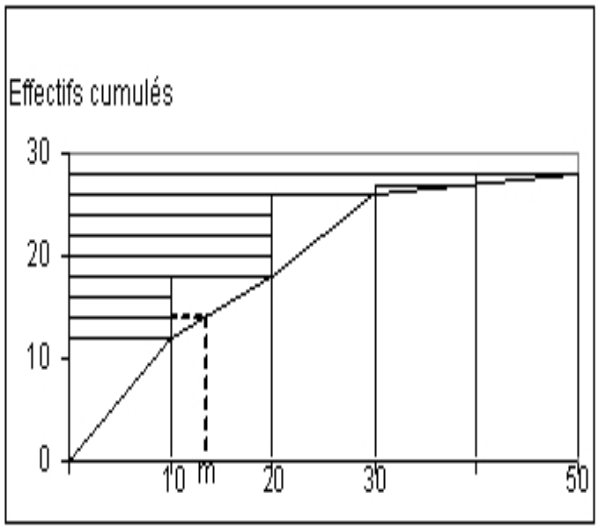



FIGURE 2.2 – *Diagramme des effectifs cumulés pour le PNB/h.*

*Plus généralement, pour calculer la médiane, on repère d'abord la classe dans laquelle elle se trouve : c'est celle dont l'effectif cumulé est immédiatement supérieur à* $\frac{N}{2}$*, notons la* $]x_1\ ;\ x_2]$ *; notons* $N_2$ *l'effectif cumulé de cette classe et* $N_1$ *l'effectif cumulé de la classe qui précède . En faisant l'hypothèse que les valeurs sont uniformément réparties à l'intérieur des classes, on a d'après le théorème de Thalès l'équation suivante :*

$$\frac{Méd - x_1}{x_2 - x_1} = \frac{\frac{N}{2} - N_1}{N_2 - N_1}$$

*qui équivaut à*

$$Méd = x_1 + (x_2 - x_1)\frac{\frac{N}{2} - N_1}{N_2 - N_1}.$$

**Remarque 2.4.** *a) Ce calcul fournit une valeur approchée de la médiane. Pour avoir la valeur exacte, il aurait fallu ne pas faire de regroupement par classes et lire la médiane directement sur les données réorganisées pas ordre croissant. Comme de coutume, en faisant un regroupement par classes, on gagne en lisibilité et en aisance pour les calculs mais on perd de l'information.*

*b) Cette technique de calcul repose sur l'hypothèse que les données sont uniformément réparties à l'intérieur des classes. D'où le choix des classes qui doit satisfaire cette hypothèse.*

*c) Insistons aussi sur le fait que les classes doivent être de même amplitude. En effet, rappelons que l'on peut calculer l'effectif cumulé d'une classe : c'est l'effectif cumulé de la borne supérieure de la classe. Mais on représente graphiquement les densités d'effectifs et non les effectifs ; et les densités d'effectifs cumulés n'ont aucun sens. On ne peut donc pas faire de diagramme cumulatif, appliquer le théorème de Thales et calculer la médiane.*

*d) Enfin si on dispose des fréquences cumulées et non des effectifs cumulés, on a la formule*

*suivante*

$$Méd = x_1 + (x_2 - x_1)\frac{0,5 - F_1}{F_2 - F_1},$$

*où la classe médiane contenant Méd est notée* $[x_1; x_2[$, $F_2$ *en est la fréquence cumulée et* $F_1$ *est la fréquence cumulée de la classe avant.*

**Exemple B.** *Considérons le PNB par habitant*

| *PNB/habitant* | *Effectifs* | *Effectifs cumulés* | *Fréquences* | *Fréquences cumulées* |
|---|---|---|---|---|
| ]0 ; 10] | 12 | 12 | 42,86% | 42,86% |
| ]10 ; 20] | 6 | 18 | 21,43% | 64,29% |
| ]20 ; 30] | 8 | 26 | 28,57% | 92,86% |
| ]30 ; 40] | 1 | 27 | 3,57% | 96,43% |
| ]40 ; 50] | 1 | 28 | 3,57% | 100,00% |
| | $N = 28$ | – | 100,00% | |

$\frac{Méd - 10}{20 - 10} = \frac{14 - 12}{18 - 12}$ *équivaut à :*

$$Méd = 10 + (20 - 10)\frac{14 - 12}{18 - 12} = 10 + 10 \times \frac{2}{6} = 13,333.$$

*Parmi les 28 pays, 14 ont un PNB par habitant inférieur à 13333 dollars et 14 ont un PNB par habitant supérieur à 13333 dollars.*

*Remarquons qu'ici, la population est de taille relativement petite et on pourrait se passer de regroupement en classes. Dans ce cas, la détermination de la médiane se fait comme dans le cas discret et la valeur obtenue est* $Méd = 12162$ *(cf tableau des données). Cette valeur est "plus exacte" que celle fournie par la méthode précédente mais il faut bien comprendre que lorsqu'on fait un regroupement en classes, on perd des informations (les valeurs réellement observées à l'intérieur d'une classe) et on gagne en lisibilité (c'est très important lorsque la population étudiée est de grande taille, ce qui est souvent le cas dans les études statistiques) : pour calculer la médiane, on est alors amené à faire l'hypothèse que la répartition à l'intérieur des classes est uniforme (approximation de la réalité), pour calculer la moyenne et l'écart-type (chapitre suivant) on considèrera les centres des classes.*

## 2.2 Généralisation de la médiane : les quantiles

*On considère ici une variable ordinale* $X$ *sur une population* $\Omega$*. Dans tous les schémas ci-dessous, le trait horizontal représentera la liste des données de* $X$ *écrites par ordre croissant.*

### 2.2.1 La médiane (bref rappel)

*Rappelons, sur un schéma, le principe général concernant la médiane :*

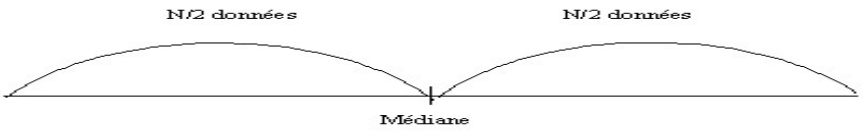



FIGURE 2.3 – *Détermination de la médiane.*

### 2.2.2 Les quartiles

*On peut aussi découper la population en 4 parties avec les* **quartiles** *:*

FIGURE 2.4 – *Détermination des quartiles.*

*Le principe est le même que pour la médiane, mais on va ici partager la liste des données en 4.*
*Par analogie avec la définition donnée pour la médiane,*
*-* $Q_1$ *sera la donnée dont le rang est immédiatement après* $\frac{N}{4}$.
*-* $Q_2$ *sera la donnée dont le rang est immédiatement après* $2 \times \frac{N}{4} = \frac{N}{2}$ *(donc* $Q_2$ *= Méd).*
*-* $Q_3$ *sera la donnée dont le rang est immédiatement après* $3 \times \frac{N}{4}$.

*On peut dire aussi que :*

*Le premier quartile* $Q_1$ *est la valeur telle que* 25% *des individus sont au-dessous.*
*Le deuxième quartile* $Q_2$ *est la médiane.*
*Le troisième quartile* $Q_3$ *est la valeur telle que* 75% *des individus sont au-dessous.*

*Ainsi dans le cadre des variables quantitatives, entre* $Q_1$ *et* $Q_3$*, on aura environ* 50% *des données de la population. La différence* $Q_3 - Q_1$ *est appelée l'***écart interquartile**.

**Remarque 2.5.** *1) De même que pour la médiane et dans le cas des variables quantitatives continues avec un regroupement par classes de même amplitude, on peut déterminer une valeur approchée de* $Q_1$ *et* $Q_2$ *:*

$$Q_1 = x_1 + (x_2 - x_1)\frac{\frac{N}{4} - N_1}{N_2 - N_1},$$

*où la classe de $Q_1$ est notée $[x_1; x_2[$, $N_2$ en est l'effectif cumulé et $N_1$ est l'effectif cumulé de la classe avant. On a aussi*

$$Q_3 = x_1 + (x_2 - x_1)\frac{\frac{3\times N}{4} - N_1}{N_2 - N_1}.$$

*où la classe de $Q_3$ est notée $[x_1; x_2[$, $N_2$ en est l'effectif cumulé et $N_1$ est l'effectif cumulé de la classe avant.*

*2) Si on dispose des fréquences cumulées au lieu des effectifs cumulés, on remplacera $\frac{N}{4}$, $\frac{N}{2}$ et $3\frac{N}{4}$ respectivement par 25%, 50% et 75%.*

**Exemple B.** *Pour F, on a $Q_1 = 2$ et $Q_3 = 5$. L'écart interquartile de la variable F : $5 - 2 = 3$ .*
*On peut calculer les quartiles du PNB par habitant :*

- *$Q_1$ est dans la classe $]0\ ;\ 10]$ $(\frac{N}{4} = 7)$ , on a donc :*

$$\frac{Q_1 - 0}{10 - 0} = \frac{7 - 0}{12 - 0} \text{ d'où } Q_1 = 10 \times \frac{7}{12} = 5,833.$$

*Un quart des pays étudiés ont un PNB par habitant inférieur à 5833 dollars.*

- *$Q_3$ est dans la classe $]20\,;\ 30]$ $(3 \times \frac{N}{4} = 21)$, on a donc :*

$$\frac{Q_3 - 20}{30 - 20} = \frac{21 - 18}{26 - 16} \text{ d'où } Q_3 = 20 + 10 \times \frac{3}{8} = 23,75.$$

*Un quart des pays étudiés ont un PNB par habitant supérieur à 23750 dollars.*
*L'écart interquartile de la variable PNB/h est : $23,75 - 5,833 = 17,917$.*

### 2.2.3 Les boîtes à moustaches

*Pour faire apparaître graphiquement ces paramètres de dispersion, on utilise souvent des* **boîtes à moustaches** *(Box plots en anglais) qui mettent en évidence les 3 quartiles ainsi que les extrémums $x_{\min}$ et $x_{\max}$. Cette représentation graphique est construite sur une échelle verticale (ou horizontale) de la façon suivante :*

| *Sur un segment gradué s'étendant de $x_{\min}$ à $x_{\max}$, tracer un rectangle (la boîte), de largeur arbitraire, qui s'étend du premier au troisième quartile et partager ce rectangle par une ligne tracée au niveau de la médiane.* |
|---|

**Exemple 2.1.** *Considérons un échantillon de 80 enfants sur lequel on a étudié la taille en cm et le poids en kg. Pour ces deux variables on a regroupé les données en classes :*

- *pour la variable taille notée X on a utilisé les classes $]80; 90]$, $]90; 100]$, $]100; 110]$.*


FIGURE 2.5 – *Boîte à moustaches de F et de PNB/hab.*

- *pour la variable poids notée* $Y$ *on a utilisé les classes* $]10;12]$, $]12;14]$, $]14;16]$, $]16;18]$.

*La table de contingence ci-dessous nous donne les résultats de l'enquête.*

| $Y$ / $X$ | $]10;12]$ | $]12;14]$ | $]14;16]$ | $]16;18]$ | *Marge de* $X$ |
|---|---|---|---|---|---|
| $]80;90]$ | *20* | *2* | *1* | *0* | *23* |
| $]90;100]$ | *3* | *31* | *3* | *0* | *37* |
| $]100;110]$ | *0* | *4* | *12* | *4* | *20* |
| *Marge de* $Y$ | *23* | *37* | *16* | *4* | $N = 80$ |

*Nous étudions les trois sous-populations en taille suivantes :* $Pop_1$ $]80;90]$, $Pop_2$ $]90;100]$ *et* $Pop_3$ $]100;110]$. *On montre que*

1. *pour* $P_1$ *:* $x_{min} = 10$, $x_{max} = 18$, *méd=11,15,* $Q_1 = 10,58$ *et* $Q_3 = 11,73$.
2. *pour* $P_2$ *:* $x_{min} = 10$, $x_{max} = 18$, *méd=13,* $Q_1 = 12,4$ *et* $Q_3 = 13,6$.
3. *pour* $P_3$ *:* $x_{min} = 10$, $x_{max} = 18$, *méd=15,* $Q_1 = 14,17$ *et* $Q_3 = 15,83$.

### 2.2.4 Les déciles

*Par analogie avec les définitions précédentes,*
$D_1$ *sera la donnée dont le rang est immédiatement après* $\frac{N}{10}$ *;*
$D_2$ *sera la donnée dont le rang est immédiatement après* $2 \times \frac{N}{10}$
*etc*
$D_5$ *sera la donnée dont le rang est immédiatement après* $5 \times \frac{N}{10} = \frac{N}{2}$ *(donc* $D_5$ *=Méd)*

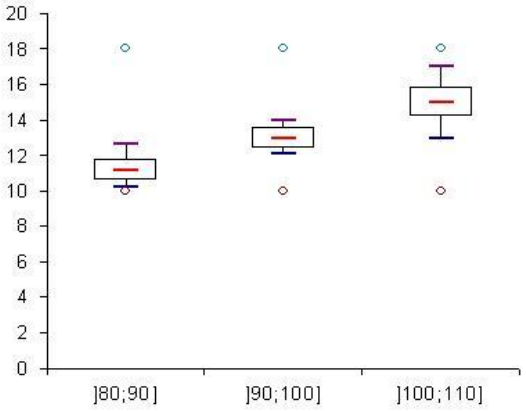



FIGURE 2.6 – Boîte à moustaches pour les variables poids/taille

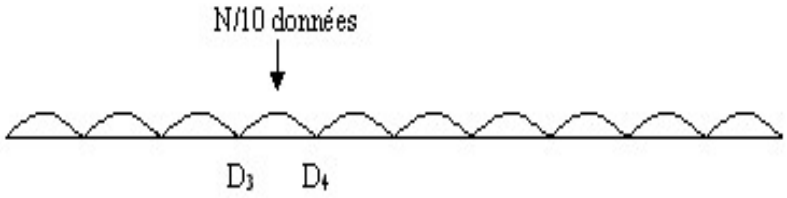



FIGURE 2.7 – *Détermination des déciles.*

*etc*

$D_9$ *sera la donnée dont le rang est immédiatement après* $9 \times \dfrac{N}{10}$.

*Ainsi donc, entre* $D_1$ *et* $D_9$*, on aura environ* 80% *des données de la population. La différence* $D_9 - D_1$ *est appelée l'écart* **interdécile**.

### 2.2.5 Les centiles

*Le principe étant toujours le même, on partage ici la liste des données en 100. On aura donc 99 centiles :* $C_1, C_2, \cdots, C_{99}$.

$C_i$ *étant la donnée dont le rang est immédiatement après* $i \times \dfrac{N}{100}$.

*Les centiles sont relativement peu utilisés en Psychologie, leur usage est plutôt réservé à la géographie.*

**Remarque 2.6.** *Le terme de* **quantile** *est le mot général pour désigner la médiane, les quartiles, les déciles et les centiles.*

# Chapitre 3

# Moyenne et variance des variables quantitatives

*On rappelle qu'une variable est dite* **quantitative** *lorsque ces modalités sont des nombres qui correspondent à la mesure d'une grandeur ou à un comptage.*

## 3.1 Un indice de tendance centrale : la moyenne

*Il s'agit d'un indice de tendance centrale spécifique aux variables quantitatives. La* **moyenne** *d'une variable quantitative $X$ , notée $\overline{X}$, est la somme des valeurs prises par $X$ divisée par la taille de la population (notée $N$).*

### 3.1.1 Un exemple introductif

*Considérons la variable note à un contrôle, sur un échantillon de $N = 12$ enfants, où la liste des données est :*

$$5 \quad 10 \quad 10 \quad 7 \quad 13 \quad 13 \quad 14 \quad 14 \quad 10 \quad 14 \quad 8 \quad 8$$

*Le total des notes des 12 enfants sera donc :*

$$5 + 10 + 10 + 7 + 13 + 13 + 14 + 14 + 10 + 14 + 8 + 8 = 126$$

*On peut également calculer ce total est regroupant les notes apparaissant plusieurs fois.*
*Cela donne :*

$$5 + 7 + (8 + 8) + (10 + 10 + 10) + (13 + 13) + (14 + 14 + 14) = 126$$

*La moyenne des notes sera :*

$$\frac{somme\ des\ notes}{N} = \frac{126}{12} = 10,5.$$

*On peut voir dans le calcul du total ci-dessus que chaque modalité de $X$ est multipliée par son effectif.*

### 3.1.2 Utilisation du tableau d'effectifs

*On peut donc utiliser le tableau d'effectifs de $X$ pour obtenir le total des notes.*

| $x_i$ | $n_i$ | $n_i \times x_i$ |
|---|---|---|
| 5 | 1 | 5 |
| 7 | 1 | 7 |
| 8 | 2 | 16 |
| 10 | 3 | 30 |
| 13 | 2 | 26 |
| 14 | 3 | 42 |
| | $N = 12$ | 126 |

*La somme des données (somme des notes ici) est donc la somme de la colonne $n_i x_i$. Elle s'écrit* $\sum n_i x_i$.

*Le symbole $\Sigma$ (on lit sigma) est le S grec.*

$\sum n_i x_i$ *signifie donc "la somme de la colonne $n_i x_i$". La moyenne $\overline{X}$ de $X$ s'écrit donc :*

$$\overline{X} = \frac{\textit{somme des données}}{N} = \frac{\sum n_i x_i}{N} = \frac{126}{12} = 10,5.$$

### 3.1.3 Définition et propriété

*Nous avons donc*

$$\overline{X} = \frac{\textit{somme des données}}{N} = \frac{\sum n_i x_i}{N} = \frac{126}{12} = 10,5.$$

*Se rappelant le fait que diviser par un nombre c'est multiplier par son inverse on préfère écrire $\overline{X}$ sous la forme*

$$\boxed{\overline{X} = \frac{1}{N} \sum n_i x_i}$$

*(cette écriture prend moins de place en hauteur !) Une façon de définir par une phrase la moyenne de $X$ est de dire que c'est le nombre tel que, si tous les individus avaient eu cette valeur pour modalité, on aurait retrouvé le même total. (Pour notre exemple cela revient à dire que 10,5 est la note telle que, si tous les enfants avaient eu cette note, on aurait retrouvé le total de 126.)*

*Plus précisément, si chaque valeur $x_i$ de $X$ a pour effectif $n_i$, la moyenne est :*

$$\overline{X} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{n_1 x_1 + n_2 x_2 + \cdots + n_k x_k}{N}$$

*où $k$ est le nombre de valeurs prises par $X$.*

**Exemple B.** *La moyenne de la variable $F$ est* $3,57$.

**Remarque 3.1.** *a)* $\sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k = N$ *(taille de la population)*

*b)* $\overline{X} = \frac{\sum_{i=1}^{k} n_i x_i}{N}$ *équivaut à* $N\overline{X} = \sum_{i=1}^{k} n_i x_i$, *c'est-à-dire à*

$$\sum_{i=1}^{k} n_i \overline{X} = \sum_{i=1}^{k} n_i x_i,$$

*soit encore à :*

$$\sum_{i=1}^{k} n_i (x_i - \overline{X}) = 0.$$

*La somme des écarts algébriques des valeurs à la moyenne est nulle, c'est une propriété caractéristique de la moyenne.*

**Cas d'un groupement en classes**

*Si on a fait un regroupement en classes, on prend pour le calcul (à la place des $x_i$) les centres $c_i$ des classes :*

$$\overline{X} = \frac{\sum_{i=1}^{k} n_i c_i}{N} = \frac{n_1 c_1 + n_2 c_2 + \cdots + n_k c_k}{N}.$$

*L'amplitude d'une classe étant la différence (extrémité de la classe - origine de la classe), pour obtenir le centre de la classe il suffit d'ajouter à l'origine de la classe la moitié de son amplitude. Ainsi par exemple, la classe* $[4\ ;\ 9[$ *a pour amplitude* $9 - 4 = 5$, *donc le centre de la classe est* $4 + \frac{5}{2} = 6,5$.

**Exemple A.** *Calculons la taille moyenne des 20 individus :*

| $[b_i\ ;\ b_{i+1}[$ | $c_i$ | $n_i$ | $n_i \times c_i$ |
|---|---|---|---|
| $[150\ ;\ 160[$ | *155* | *3* | *465* |
| $[160\ ;\ 170[$ | *165* | *7* | *1155* |
| $[170\ ;\ 180[$ | *175* | *6* | *1050* |
| $[180\ ;\ 190[$ | *185* | *4* | *740* |
| *Totaux* | - | *20* | *3410* |

*La taille moyenne est donc :*

$$\overline{X} = \frac{3410}{20} = 170,5cm$$

*Ici les classes ayant même amplitude, il suffit d'ajouter l'amplitude à un centre pour trouver le centre de la classe suivante.*

**Exemple B.** *Calculons la moyenne de la variable PNB/habitant (regroupée en classes), notée $X$, du tableau de données.*
*La moyenne de la variable $X$ est 15357 avec le regroupement en classes*

$$\frac{12 \times 5 + 6 \times 15 + 8 \times 25 + 1 \times 35 + 1 \times 45}{28} = \frac{430}{28} = 15,357 \ ;$$

*sans regroupement, on trouve 14875. Notons bien que ce dernier résultat n'est pas le PNB par habitant de l'ensemble des 28 pays, en effet, il faudrait tenir compte de la population représentée dans chaque classe. Pour calculer le PNB par habitant de l'ensemble des 28 pays, il suffit d'additionner les PNB globaux et diviser par la somme des populations.*

*On a arrondi le résultat à deux chiffres après la virgule. (Il faut toujours arrondir au plus proche, c'est-à-dire en tenant compte de la décimale suivante.)*

**Remarque 3.2.** *a) Remplacer une classe par son centre revient à faire comme si tous les individus dont la donnée est dans cette classe avaient la même modalité : son centre. Ainsi comme pour le calcul de la médiane, on fait une hypothèse sur la répartition des données à l'intérieur des classes. Par exemple dans l'exemple A, cela revient à faire comme si les 3 personnes dont la taille est dans* $[150 \ ; \ 160[$ *mesuraient 155cm.*
*On peut avoir l'impression de "fausser" légèrement les résultats, par contre on gagne en simplicité (moins de classes que de modalités) et dans l'optique d'une généralisation d'un échantillon à une population, cela n'a pas d'importance.*
*b) Lorsqu'il y a peu de données, plutôt que d'utiliser la méthode par tableau, il est plus simple d'utiliser directement la formule :*

$$\overline{X} = \frac{somme \ des \ données}{N}.$$

**Remarque sur la moyenne et la médiane :**

*La moyenne et la médiane sont deux caractéristiques de tendance centrale de la série statistique, c'est-à-dire qu'elles résument chacune la "position centrale" de la série. La moyenne est très*

*simple et très rapide à calculer. Le calcul de la médiane est moins aisé. Cependant la médiane est moins sensible que la moyenne à des observations "exceptionnelles" appelées observations* **aberrantes**.

*Prenons par exemple une série de notes obtenues par un étudiant :*

14 15 16

*Sa moyenne est alors de 15 et la médiane de ses notes est également de 15. Cet étudiant passe une quatrième épreuve : il n'était pas dans son élément et a obtenu 0 à cette épreuve. Sa moyenne est désormais de 11,25 tandis que la médiane de ses notes est toujours 15. Sur cet exemple très simple, on voit que la valeur de la médiane a été peu affectée par cette nouvelle note (qui peut être considérée comme exceptionnelle sur l'ensemble des notes de l'étudiant) tandis que la moyenne a été considérablement diminuée. Ne pas en déduire toutefois que les étudiants auraient toujours intérêt à avoir recours à la médiane dans leur évaluation : on peut construire un exemple symétrique d'un étudiant ayant obtenu 4 mauvaises notes et une excellente. Dans ce cas, le calcul de la moyenne lui sera "favorable". On dit que la médiane est plus* **robuste** *que la moyenne en ce sens qu'elle "résiste" mieux aux observations aberrantes. Notons enfin que dans le cas d'une distribution dissymétrique, la médiane est un paramètre de position plus pertinent que la moyenne.*

**Proposition 3.1.** *Soit $X$ et $Y$ deux variables définies sur une même population $\Omega$ et $a$ un nombre fixé (positif ou négatif). On a les trois propriétés suivantes :*

- $\overline{X+Y} = \overline{X} + \overline{Y}$ *(la moyenne de la somme est la somme des moyennes)*
- $\overline{aX} = a\overline{X}$ *(si on multiplie les valeurs de $X$ par $a$, la moyenne est multipliée par $a$).*
- $\overline{X+a} = \overline{X} + a$ *(si on ajoute $a$ à chaque valeur de $X$, on ajoute $a$ à la moyenne).*

*Nous verrons que ces propriétés permettent dans certains cas de faire des* ***changements de variables*** *et de simplifier les calculs.*

**Exemple A.** *Pour le calcul de la taille moyenne, introduisons la variable* $Y = \dfrac{X - 150}{5}$. *Les classes de $Y$ sont* $[0;2[$, $[2;4[$, $[4;6[$ *et* $[6;8[$.

| $[b_i\ ;\ b_{i+1}[$ | $c_i$ | $n_i$ | $n_i \times c_i$ |
|---|---|---|---|
| [0; 2[ | 1 | 3 | 3 |
| [2 ; 4[ | 3 | 7 | 21 |
| [4 ; 6[ | 5 | 6 | 30 |
| [6 ; 8[ | 7 | 4 | 28 |
| *Totaux* | - | 20 | 102 |

*La moyenne de $Y$ est donc :*

$$\overline{Y} = \frac{82}{20} = 4,1cm$$

*Ensuite $X = 5 * Y + 150$ et donc d'après les propriétés précédentes $\overline{X} = 5 * \overline{Y} + 150 = 170,5$.*

**Une autre expression de la moyenne :**

*À partir de la définition de la moyenne et en notant $f_j, j = 1, \ldots, k$ les fréquences, on obtient facilement une autre expression pour la moyenne :*

$$\boxed{\overline{X} = \sum_{j=1}^{k} \frac{1}{N} n_j x_j = \sum_{j=1}^{k} f_j x_j}$$

*pour une variable quantitative discrète ou continue sans regroupement par classes et :*

$$\boxed{\overline{X} = \sum_{j=1}^{k} \frac{1}{N} n_j c_j = \sum_{j=1}^{k} f_j c_j}$$

*pour une variable quantitative continue avec regroupement par classes.*

## 3.2 Un indice de dispersion : l'étendue

*Pour compléter l'information fournie par la médiane et la moyenne que l'on appelle caractéristiques de tendance centrale, on a besoin, lorsqu'on étudie une variable quantitative, de mesurer la dispersion de la série statistique. On va voir trois façons d'aborder la question :*

- *une se rapportant aux données extrêmes ;*
- *une autre se rapportant aux quartiles ;*
- *la dernière permettant de mesurer la dispersion autour de la moyenne.*

*On veut fabriquer un nombre (un paramètre) qui rende compte de l'éloignement (la dispersion) entre les différentes données d'une variable quantitative.*

*L'idée la plus simple consiste à mesurer l'écart (absolu) entre la plus grande donnée et la plus petite. L'***étendue** *est simplement la différence entre la valeur maximum observée (plus grande donnée $x_{\max}$) et la valeur minimum observée (plus petite donnée $x_{\min}$).*

$$\boxed{\text{Étendue} = x_{\max} - x_{\min}}$$

*Ainsi dans l'exemple introductif précédent on aura : étendue = 14-5 = 9.*

**Exemple A.** *L'étendue de la variable âge est* $30-22=8$ *et celle de la variable nombre d'enfants* $3-0=3$.

*Dans le cas d'un groupement en classes, on procède de même avec les centres de classes.*

**Exemple B.** *L'étendue de la variable F est* $9-1=8$. *L'étendue de la variable PNB/h (regroupée en classes) est* $50-0=50$.

*L'étendue est très simple à calculer mais reste cependant très élémentaire et limitée : elle ne tient compte que des deux données extrêmes et ne reflète pas en particulier la répartition des observations entre les deux valeurs extrêmes. Elle peut ainsi donner une vision totalement fausse de la variable étudiée.*
*Considérons par exemple un échantillon composé de 40 enfants de 10 ans accompagnés d'un adulte de 49 ans tenant dans ses bras un bébé de 1 an.*
*Pour ce groupe de 42 personnes l'étendue de l'âge sera* $49-1=48$, *alors que presque tous les membres du groupe ont 10 ans. Cet exemple montre que l'on a besoin d'un autre paramètre de dispersion plus significatif.*

## 3.3 Un indice de dispersion : l'écart et l'intervalle interquartiles

L'**écart interquartile** *(vu au chapitre précédent), est la différence entre le troisième quartile et le premier quartile.*

$$\boxed{\text{Écart interquartile} = Q_3 - Q_1}$$

*Rappelons que l'écart interquartile peut se calculer aussi pour les variables ordinales.*

L'**intervalle interquartile** *est l'intervalle délimité par le premier quartile et le troisième quartile. Dans cet intervalle on trouve 50% de la population.*

## 3.4 Un indice de dispersion : la variance

### 3.4.1 Exemples introductifs

***Exemple 1*** *Voici les notes obtenues par deux élèves lors de contrôles :*

*Pierre : 2 ; 3 ; 14 ; 13. La moyenne est :* $\dfrac{2+3+14+13}{4} = 8.$

*Paul : 7 ; 7 ; 8 ; 10. La moyenne est :* $\dfrac{7+7+8+10}{4} = 8.$

*Pierre et Paul ont la même moyenne mais les notes de Pierre s'écartent beaucoup de cette moyenne tandis que celles de Paul sont regroupées autour de 8.*

*On a donc besoin d'un indice permettant de mesurer l'hétérogénéïté des données et la dispersion de $X$ autour de sa moyenne. Il faudra donc calculer $\overline{X}$ avant de calculer la variance.*

***Exemple 2*** *Considérons maintenant la variable âge, notée $X$, sur deux groupes d'enfants $A$ et $B$. Chaque groupe est composé de $N = 9$ enfants, les deux listes de données étant :*

*Groupe A :* 4 9 9 10 10 10 11 11 16 .

*Groupe B :* 4 4 5 10 10 10 15 16 16 .

*De rapides calculs nous permettent d'obtenir le mode, la médiane, la moyenne et l'étendue pour chaque groupe :*

*Groupe A : mode = 10* *Méd = 10* $\overline{X} = 10$ *étendue = 12*

*Groupe B : mode = 10* *Méd = 10* $\overline{X} = 10$ *étendue = 12.*

*Ces deux groupes sont identiques au regard de ces résultats ; pourtant on voit bien qu'ils sont de nature fort différente. Ainsi par exemple, un moniteur de colonie de vacances préférera nettement travailler avec le groupe A qui est plus homogène, alors que le groupe B comporte lui des enfants d'âges bien plus dispersés.*

*L'objectif est donc de construire un nombre qui rende compte de cette différence de constitution entre les deux groupes.*

*Reprenons le groupe B.*

*Pour faire apparaître la nature dispersée des données, on peut regarder l'écart (dit algébrique) entre chaque donnée et la moyenne. Cet écart est obtenu en soustrayant à chaque donnée sa moyenne.*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *donnée :* | 4 | 4 | 5 | 10 | 10 | 10 | 15 | 16 | 16 |
| *donnée − moyenne :* | -6 | -6 | -5 | 0 | 0 | 0 | 5 | 6 | 6 |

*On voit donc bien des écarts importants. Comme on souhaite avoir un nombre "résumant" la situation on va simplement calculer la moyenne de ces écarts algébriques ; on trouve 0 :*

$$\frac{-6-6-5+0+0+0+5+6+6}{9} = \frac{0}{9}$$

*(On trouverait le même résultat avec le groupe A.)*

*L'objectif visé n'est donc pas atteint. Cela vient du fait qu'on a des écarts négatifs et des écarts positifs, ils se "neutralisent" donnant un total de 0.*
*Pour éviter cela on pourrait tout simplement enlever les signes (en mathématique on dit prendre les valeurs absolues), mais pour des raisons liées aux calculs on préfère élever chaque écart au carré, ce qui permettra également de n'avoir que des nombres positifs.*
*Cela donne toujours pour le groupe B :*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *donnée :* | *4* | *4* | *5* | *10* | *10* | *10* | *15* | *16* | *16* |
| *donnée − moyenne :* | *-6* | *-6* | *-5* | *0* | *0* | *0* | *5* | *6* | *6* |
| *(donnée − moyenne)$^2$ :* | *36* | *36* | *25* | *0* | *0* | *0* | *25* | *36* | *36* |

*Il reste alors à faire la moyenne des valeurs de la dernière ligne, c'est-à-dire la moyenne des carrés des écarts entre les données et leur moyenne :*

$$\frac{36+36+25+0+0+0+25+36+36}{9} = \frac{194}{9} = 21,56.$$

*Cette quantité est appelée la* **variance de** $X$ *(sur le groupe B).*
*Le même travail avec le groupe A donne :*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *donnée :* | *4* | *9* | *9* | *10* | *10* | *10* | *11* | *11* | *16* |
| *donnée - moyenne :* | *-6* | *-1* | *-1* | *0* | *0* | *0* | *1* | *1* | *6* |
| *(donnée - moyenne)$^2$ :* | *36* | *1* | *1* | *0* | *0* | *0* | *1* | *1* | *36* |

*Variance de $X$ sur le groupe A :*

$$\frac{36+1+1+0+0+0+1+1+36}{9} = \frac{76}{9} = 8,44.$$

***Interprétation :*** *La variance représente donc globalement l'ensemble des carrés des écarts entre les données et leur moyenne. L'objectif est donc rempli.*

### 3.4.2 La variance : définition et formule simplifiée

*On définit donc la variance d'une variable quantitative $X$ comme la moyenne des carrés des écarts entre les données de $X$ et leur moyenne.*

$$\boxed{\text{variance de } X = \frac{1}{N}\sum n_i(x_i - \overline{X})^2}$$

*En développant le carré $(x_i - \overline{X})^2$ et en regroupant différemment les termes, on peut montrer que :*

$$\boxed{\textit{variance de } X = \left(\frac{1}{N}\sum n_i(x_i^2)\right) - \overline{X}^2}$$

*ou encore*

$$\boxed{\textit{variance de } X = \frac{\textit{somme des carrés des données}}{N} - \left(\frac{\textit{somme des données}}{N}\right)^2}$$

**Remarque 3.3.** *a) Pour calculer la variance de $X$ on a besoin de $\overline{X}$, or un simple arrondi sur la valeur de $\overline{X}$ peut entraîner dans certains cas une erreur importante sur le résultat de la variance. Aussi, il peut arriver, bien qu'une variance soit toujours positive, d'obtenir à cause d'un arrondi sur la moyenne, un résultat négatif. Pour éviter cela, il faut utiliser la valeur exacte de la moyenne (en pratique on met le résultat de la moyenne en mémoire dans la machine à calculer).*

*b) L'équivalence entre les deux formules n'est pas évidente mais la démonstration ne demande que des connaissances mathématiques élémentaires et un étudiant à l'aise en calcul littéral peut la faire en exercice.*

*c) La première formule permet de bien comprendre ce que mesure la variance : c'est la moyenne des carrés des écarts à la moyenne, en particulier, c'est un nombre positif.*

*d) La deuxième formule est souvent plus pratique pour faire les calculs. En effet, cette expression ne demande que deux calculs par modalité $x_i$ (élévation au carré puis multiplication par $n_i$). Cependant, deux erreurs apparaissent fréquemment : il faut bien comprendre que dans cette formule, seules les valeurs $x_i$ (pas les $n_i$) sont élevées au carré et ne pas oublier de soustraire $\overline{X}^2$.*

### 3.4.3 Utilisation du tableau d'effectifs

*Bien sûr avec un grand nombre de données, la méthode décrite ci-dessus deviendrait très lourde même avec la formule simplifiée. Aussi va-t-on en partant du tableau d'effectifs rajouter une colonne intitulée "$n_i(x_i^2)$" dont la somme est égale au carré des données et permettant de faire plus facilement le calcul de la variance.*

*Reprenons l'exemple précédent avec le groupe B. On a le tableau suivant :*

| $x_i$ | $n_i$ | $n_i(x_i)^2$ |
|---|---|---|
| 4 | 2 | 32 |
| 5 | 1 | 25 |
| 10 | 3 | 300 |
| 15 | 1 | 225 |
| 16 | 2 | 512 |
| | N=9 | 1094 |

*La variance pour le groupe B est donc*

$$\frac{1094}{9} - 10^2 \approx 21,56.$$

*Reprenons l'exemple précédent avec le groupe A. On a le tableau suivant :*

| $x_i$ | $n_i$ | $n_i(x_i)^2$ |
|---|---|---|
| *4* | *1* | *16* |
| *9* | *2* | *162* |
| *10* | *3* | *300* |
| *11* | *2* | *242* |
| *16* | *1* | *256* |
| | *N=9* | *976* |

*La variance pour le groupe A est donc*

$$\frac{976}{9} - 10^2 \approx 8,44.$$

### 3.4.4 Exemples de calcul

**Exemple A.** *Calculons la variance pour la variable nombre d'enfants. Comme le calcul de la variance nécessite de connaître la moyenne, on va reprendre les calculs comme si la moyenne n'était pas déjà connue.*

*Pour calculer la variance, il faudra rajouter la colonne $n_i(x_i^2)$. On peut remarquer que pour obtenir cette colonne il suffit de faire le produit des colonnes "$x_i$" et "$n_i x_i$". La somme de cette colonne correspond au total des carrés des données.*

| $x_j$ | $n_j$ | $n_j \times x_j$ | $n_j \times x_j^2$ |
|---|---|---|---|
| *0* | *6* | *0* | *0* |
| *1* | *7* | *7* | *7* |
| *2* | *5* | *10* | *20* |
| *3* | *2* | *6* | *18* |
| *Totaux* | *20* | *23* | *45* |

*La variance est donc :*

$$\mathrm{Var}(X) = \frac{45}{20} - \left(\frac{23}{20}\right)^2 = 0,93.$$

**Remarque 3.4.** *Lorsqu'il y a peu de données, plutôt que d'utiliser la méthode par tableau, il est plus simple d'utiliser directement la formule :*

$$\textit{variance de } X = \frac{\textit{somme des carrés des données}}{N} - \overline{X}^2.$$

*Par exemple,*

*Variance de la variable "note de Pierre" :*

$$\frac{(2-8)^2+(3-8)^2+(14-8)^2+(13-8)^2}{4} = \frac{6^2+5^2+6^2+5^2}{4} = 30,5.$$

*Variance de la variable "note de Paul" :*

$$\frac{2\times(7-8)^2+(8-8)^2+(10-8)^2}{4} = \frac{2\times 1^2+0+2^2}{4} = 1,5.$$

*Exercice : Faire les calculs avec la deuxième formule pour vérifier sur ces exemples simples que l'on trouve bien les mêmes résultats.*

**Exemple B.** *Calculer la variance des variables F et PNB/h, notée X. (Pour la variable PNB/h, il faut prendre les centres $c_i$ des classes à la place des $x_i$ comme on l'a fait pour la moyenne). Voici deux tableaux de calculs que l'on peut utiliser pour présenter les calculs de la moyenne et de la variance :*

| $x_i$ | $n_i$ | $n_i \times x_i$ | $ni(x_i^2)$ |
|---|---|---|---|
| 1 | 2 | 2 | 2 |
| 2 | 7 | 14 | 28 |
| 3 | 7 | 21 | 63 |
| 4 | 4 | 16 | 64 |
| 5 | 5 | 25 | 125 |
| 6 | 1 | 6 | 36 |
| 7 | 1 | 7 | 49 |
| 8 | 0 | 0 | 0 |
| 9 | 1 | 9 | 81 |
| | N=28 | 100 | 448 |

| *PNB/habitant* | *Centres $c_i$* | *Effectifs $n_i$* | $n_ic_i$ | $n_i(c_i^2)$ |
|---|---|---|---|---|
| ]0 ; 10] | 5 | 12 | 60 | 300 |
| ]10 ; 20] | 15 | 6 | 90 | 1350 |
| ]20 ; 30] | 25 | 8 | 200 | 5000 |
| ]30 ; 40] | 35 | 1 | 35 | 1225 |
| ]40 ; 50] | 45 | 1 | 45 | 2025 |
| | | N=28 | 430 | 9900 |

*La variance de la variable F est : 3,24. Celle de la variable PNB/habitant est : 117,73 .*

## 3.5 Un indice de dispersion : l'écart-type

*La variance représentant globalement les carrés des écarts entre les données et leur moyenne, il serait souhaitable de construire un nombre représentant globalement simplement les écarts entre les données et leur moyenne. Il suffit pour cela de prendre la racine carrée de la variance que l'on appellera l'écart-type.*

*L'***écart-type** *d'une variable quantitative $X$, noté $\sigma_X$ est la racine carrée de la variance :*

$$\boxed{\sigma(X) = \sqrt{\mathrm{Var}(X)}.}$$

*On le note indifféremment $\sigma_X$ ou $\sigma(X)$.*

**Exemple 3.1.**
*Écart-type de la variable "note de Pierre" :* $\sqrt{30,5} = 5,52$.
*Écart-type de la variable "note de Paul" :* $\sqrt{1,5} = 1,22$.

*Exercice : Calculer l'écart-type des variables $F$ et $PNB/habitant$.*

**Exemple A.** *L'écart-type de la variable nombre d'enfants vaut :* $\sqrt{0,9275} = 0,96$.

**Remarque 3.5.** *a) On a donc aussi : variance de $X = \sigma_X^2$. Aussi la variance de $X$ sera-t-elle notée parfois $\sigma_X^2$.*
*b) $\sigma_X$ représente globalement l'écart entre les données de $X$ et leur moyenne. Contrairement à la variance (qui n'a pas réellement d'unité), l'écart-type s'exprime dans l'unité de la variable $X$. Il s'interprète donc plus simplement.*
*c) Certains auteurs souhaitent différencier les notations de moyenne, de variance et d'écart-type pour une variable quantitative $X$ suivant qu'on travaille sur une population $\Omega$ ou seulement sur un échantillon de cette population. Ils adoptent alors les notations suivantes :*

| | *Moyenne de $X$* | *Écart-type de $X$* | *Variance de $X$* |
|---|---|---|---|
| *Population* | $\mu$ | $\sigma$ *ou* $\sigma_X$ | $\sigma^2$ *ou* $\sigma_X^2$ |
| *Échantillon* | $\overline{X}$ | $s$ *ou* $s_X$ | $s^2$ *ou* $s_X^2$ |

*d) Pour certaines variables $X$ telles que le QI, la taille, le poids... sur des populations de taille suffisamment grande qui sont assez "régulières" dans un sens à préciser, on trouve environ 68% des observations entre $\overline{X} - \sigma_X$ et $\overline{X} + \sigma_X$ ; on trouve également environ 95% des observations entre $\overline{X} - 2\sigma_X$ et $\overline{X} + 2\sigma_X$.*

## 3.6 Changement de variable

*Dans certains cas, il peut s'avérer utile de transformer les données (issues d'une variable quantitative discrète ou continue) afin de simplifier les calculs de certaines caractéristiques numériques ou encore de ramener les observations de plusieurs séries à une même échelle et ainsi pouvoir les comparer. La transformation utilisée est une* **transformation affine** *que nous décrivons ci-dessous en donnant les expressions de certaines caractéristiques numériques (moyenne, variance et écart-type) de la variable transformée.*

### 3.6.1 Transformation affine des données : cas général

*Soient $a$ et $b$ deux nombres réels quelconques. À partir de la variable (quantitative) $X$ on définit sur la population $\Omega$ une nouvelle variable quantitative $Z = aX + b$. On dit alors que l'on a réalisé une* **transformation affine** *des données. La variable $Z$ est définie de la manière suivante :*

*- dans le cas discret : si $x_1, \ldots, x_k$ sont les modalités de la variable $X$ avec les effectifs $n_1, \ldots, n_k$, la variable $Z$ est une variable discrète ayant pour modalités $z_1 = ax_1 + b, \ldots, z_k = ax_k + b$ avec pour effectifs les mêmes effectifs que ceux de la variable $X$, c'est-à-dire $n_1, \ldots, n_k$. Les fréquences, effectifs cumulés et fréquences cumulés sont donc les mêmes que ceux de la variable $X$ : seules les modalités de la variable $X$ sont transformées ;*

*- dans le cas continu : si $[b_1 \ ; \ b_2[, \ldots, [b_k \ ; \ b_{k+1}[$ sont les classes de la variable $X$ avec les effectifs $n_1, \ldots, n_k$, la variable $Z$ est une variable quantitative continue ayant pour classes $[b'_1 \ ; \ b'_2[, \ldots, [b'_k \ ; \ b'_{k+1}[$ où pour $j = 1, \ldots, k+1$ on a $b'_j = ab_j + b$. Les effectifs, effectifs cumulés, fréquences et fréquences cumulées associés aux classes de la variable $Z$ sont les mêmes que ceux associés aux classes corespondantes de la variable $X$.*

**Exemple 3.2.** *Soit par exemple une variable $X$ dont les observations sont regroupées dans le tableau suivant :*

| $x_j$ | $n_j$ |
|---|---|
| 5889500 | 4 |
| 5889600 | 2 |
| 5889900 | 1 |
| 5890100 | 1 |
| 5890200 | 2 |

*Calculons la variable $Z$ définie par* $Z = \dfrac{X}{100} - \dfrac{5890000}{100}$
*(on a donc* $a = \dfrac{1}{100}$, $b = \dfrac{5890000}{100}$ *et* $X = 100Z + 5890000$*) :*

| $z_j$ | $n_j$ |
|---|---|
| -5 | 4 |
| -4 | 2 |
| -1 | 1 |
| 1 | 1 |
| 2 | 2 |

***Expression de la moyenne de la variable transformée :***

*On a les relations suivantes liant les moyennes des variables $X$ et $Z$ :*

$$\overline{Z} = a\overline{X} + b,$$

$$\overline{X} = \frac{\overline{Z} - b}{a}.$$

*La seconde relation se déduit directement de la première. Montrons cette relation dans le cas d'une variable quantitative discrète :*

$$\begin{aligned}
\overline{Z} &= \frac{1}{N}\sum_{j=1}^{k} n_j z_j = \frac{1}{N}\sum_{j=1}^{k} n_j(ax_j + b) \\
&= \frac{1}{N}\sum_{j=1}^{k} n_j a x_j + \frac{1}{N}\sum_{j=1}^{k} n_j b = a\frac{1}{N}\sum_{j=1}^{k} n_j x_j + b\frac{1}{N}\sum_{j=1}^{k} n_j \\
&= a\overline{X} + b,
\end{aligned}$$

*puisque* $\sum_{j=1}^{k} n_j = N$*. La démonstration est analogue dans le cas d'une variable quantitative continue.*

**Exemple 3.3.** *La moyenne de la variable $Z$ se calcule facilement :*

$$\begin{aligned}
\overline{Z} &= \frac{1}{10}(-5 \times 4 - 4 \times 2 - 1 \times 1 + 1 \times 1 + 2 \times 2) \\
&= \frac{-24}{10} = -2,4.
\end{aligned}$$

*On en déduit la moyenne de $X$ :*

$$\overline{X} \;=\; -2,4 \times 100 + 5890000 \;=\; 5889760.$$

***Expression de la variance de la variable transformée :***

*On a la propriété suivante que nous admettrons sans démonstration :*

$$\sigma_Z^2 \;=\; a^2\sigma_X^2.$$

**Conséquences** *: Si on particularise la propriété précédente au cas où $a = 1$, on obtient $\sigma_{X+b}^2 = \sigma_X^2$, c'est-à-dire que la variance d'une variable n'est pas affectée par une translation des données. Si on pose $b = 0$, on a $\sigma_{aX}^2 = a^2\sigma_X^2$ : la variance est alors multipliée par un facteur $a^2$.*

*Comme pour la moyenne, cette propriété peut être utile pour simplifier les calculs de la variance d'une variable dont les valeurs sont par exemple très "grandes".*

***Expression de l'écart-type de la variable transformée :*** *On a la propriété suivante qui se déduit de la propriété de la variance :*

$$\sigma_Z \;=\; |a|\sigma_X.$$

*On a en effet :*

$$\sigma_Z \;=\; \sqrt{\sigma_Z^2} \;=\; \sqrt{a^2\sigma_X^2} \;=\; |a|\sqrt{\sigma_X^2} \;=\; |a|\sigma_X.$$

***Vocabulaire*** *: On appelle variable* **centrée** *la variable statistique $Y = X - \overline{X}$. On montre facilement que cette variable a pour moyenne 0.*

*On appelle variable* **centrée et réduite**, *la variable $Z = \dfrac{X - \overline{X}}{\sigma_X}$. Cette variable a pour moyenne 0 et pour variance (et pour écart-type) 1.*

### 3.6.2 Changement de variable afin de simplifier des calculs

*Lorsque les calculs de moyenne de variance et d'écart-type pour une variable $X$ sont très lourds, on peut utiliser une autre variable $Z$ définie à partir de $X$ en posant $z_i = \dfrac{x_i - b}{a}$. On parle de changement de variable affine : à chaque $x_i$ on enlève $b$ puis on divise le résultat obtenu par $a$. $a$ et $b$ doivent être judicieusement choisis afin que les calculs de moyenne de variance et d'écart-type pour $Z$ soient beaucoup plus simples que pour $X$. En particulier on prendra toujours $a > 0$.*

**Exemple 3.4.** *Si toutes les modalités sont multiples de 100 on prendra $b = 0$ et $a = 100$, ce qui revient à poser $z_i = \dfrac{x_i}{100}$. Cela revient à changer d'unité : c'est-à-dire à travailler avec des centaines. Les $x_i$ étant divisés par 100, la moyenne et l'écart-type sont automatiquement divisés eux aussi par 100. Par contre, la variance sera divisée par $100^2$.*

### 3.6.3 Changement de variable dans un but de comparaison

*Il s'agit ici de relativiser des valeurs en tenant compte de leur environnement afin de pouvoir les comparer, alors que leur comparaison directe n'aurait aucun sens.*

*Posons-nous par exemple la question suivante :*
*Dans une population de fourmis dont le poids moyen est 10 milligrammes avec un écart-type de 5 milligrammes, on observe une fourmi de 12,5 milligrammes. Dans une population d'éléphants dont le poids moyen est 3,5 tonnes avec un écart-type de 0,5 tonnes, on observe un éléphant de 3,7 tonnes. Qui de la fourmi et de l'éléphant est le plus lourd ?*
*Bien sûr, prise au premier degré cette question est absurde. Évidemment, un éléphant est plus lourd qu'une fourmi ! Il faut donc comprendre la question dans le sens suivant : qui est le plus lourd, chacun relativement à sa population.*
*Déjà on peut remarquer que la fourmi comme l'éléphant a un poids supérieur au poids moyen de sa population. Ensuite, si on prend en compte l'écart-type, plus l'écart entre le poids étudié et la moyenne est grand par rapport à l'écart-type et plus ce poids se "détache" des poids de l'ensemble de la population. Aussi, notant $X$ la variable poids pour les fourmis, on construit la variable*

$$Z = \frac{X - \overline{X}}{\sigma_X} = \frac{X - 10}{5}.$$

*$Z$ sera appelée la variable poids réduit.*
*Le poids réduit de la fourmi de 12,5 milligrammes sera donc*

$$\frac{12,5 - 10}{5} = 0,5.$$

*De la même façon, le poids réduit de l'éléphant de 3,7 tonnes est*

$$\frac{3,7 - 3,5}{0,5} = 0,4.$$

*Il suffit maintenant de comparer les poids réduits pour dire qui est le plus lourd dans le sens précisé ci-dessus : ici la fourmi de 12,5 milligrammes est donc plus lourde que l'éléphant de 3,7 tonnes.*

**Exemple 3.5.** *Un groupe d'étudiants passe deux examens notés de 0 à 20. La moyenne des notes obtenues par ce groupe au premier examen est 13 et l'écart-type est 1,6. Au second examen la moyenne est 10,5 et l'écart-type est 2,4. Un étudiant obtient 11 au premier examen et 9 au second : pour quel examen se situe-t-il le mieux par rapport au groupe des étudiants ?*

*On compare pour cela les notes centrées et réduites de cet étudiant qui sont* $\frac{11-13}{1,6} = -1,25$ *pour le premier examen et* $\frac{9-10,5}{2,4} = -0,625$. *C'est donc au second examen que l'étudiant est le meilleur relativement au groupe.*

*On pourra utiliser cette notion de variable réduite chaque fois qu'il faudra relativiser une valeur par rapport à son contexte. Cela est souvent utilisé en sciences humaines. Ainsi par exemple, un salaire de 2000 euros pour une personne doit être relativisé par rapport au pays dans lequel vit cette personne. En effet en France c'est un salaire permettant de vivre assez difficilement, alors que dans des pays où on vit aisément avec 1000 euros, cela représente un salaire élevé.*

**Remarque 3.6.** *En reprenant les explications contenues dans le paragraphe précédent sur le changement de variable, on peut voir que la variable réduite $Z$ a pour moyenne 0 et pour écart-type 1.*

## 3.7 Conclusion

*Les différents indicateurs de tendance centrale et de dispersion que nous avons définis (mode, médiane, moyenne - étendue, écart interquartile, variance, écart-type) sont complémentaires et permettent de décrire et résumer une série statistique qui peut être de taille importante, ils fournissent des informations claires et concises à partir de documents comportant des tableaux de chiffres parfois difficiles à exploiter sous leur forme brute. Cependant, l'interprétation de ces résultats ne peut se faire que par comparaison, on ne peut pas dire dans l'absolu qu'une moyenne est grande ou petite, qu'un écart-type traduit une homogénéité des données ou pas : il faut tenir compte du contexte et regarder l'évolution au cours du temps ou bien comparer des résultats obtenus pour différentes populations...*

*Reprenons l'exemple des pays de l'U.E. et des pays candidats en distinguant les pays en passe d'adhérer à l'U.E. des trois autres (Turquie, Bulgarie et Roumanie). Peut-on considérer que le PNB par habitant est un critère qui distingue ces trois groupes ?*

*On partage l'ensemble $\Omega$ des 28 pays en 3 groupes :*

- $\Omega_1$ *: Les pays membres de l'UE.*
- $\Omega_2$ *: Les pays en passe d'adhérer à l'UE.*
- $\Omega_3$ *: Turquie, Bulgarie et Roumanie.*

*On s'intéresse à la variable "PNB/habitant", notée $X$ :*

*Sur $\Omega_1$, la moyenne est $\mu_1 = 2385$ et la variance est $\sigma_1^2 = 73\,110\,246$.*

*Sur $\Omega_2$ , la moyenne est $\mu_2 = 5312$ et la variance est $\sigma_2^2 = 10\,750\,677$.*

*Sur* $\Omega_3$*, la moyenne est* $\mu_3 = 1853$ *et la variance est* $\sigma_3^2 = 621\,733$.

*Une simple comparaison des moyennes permet de penser qu'un tel écart n'est pas fortuit mais que le PNB par habitant est bien un des critères qui distingue les trois groupes. On peut également calculer la variance de* $X$ *sur* $\Omega$ *(en reprenant le résultat donné précédemment, il suffit de le multiplier par* $1000^2$ *puisque la variable était exprimée en milliers de dollars) et constater que* $117\,730\,000 > 73\,110\,246$*, c'est-à-dire que l'hétérogénéité de l'U.E. pour le PNB/h serait bien plus grande avec ces 28 pays.*

*On peut aussi mesurer le lien entre* $X$ *et la variable nominale* $Y$ *dont les modalités sont "membre de l'U.E.", "en passe d'adhérer à l'U.E.", "autre candidat" en calculant le rapport entre la variance intergroupe (variance des moyennes de chaque groupe* $\Omega_i$ *affectées des effectifs des groupes) et la variance totale. Ce rapport, toujours compris entre 0 et 1, indique un lien fort lorsqu'il est proche de 1 et faible lorsqu'il est proche de 0. On trouve une variance intergroupe égale à* $99\,094\,043$ *soit un rapport de 0,84 ; on peut donc considérer que le lien entre* $X$ *et* $Y$ *est fort.*

*Evidemment, ces conclusions ne sont pas très surprenantes et les moqueurs pourront dire que la statistique sert à démontrer qu'il y a plus de morts en temps de guerre qu'en temps de paix.*

*Néanmoins, une analyse statistique pertinente à partir de données multiples permet souvent de mettre en évidence une idée, d'étayer un raisonnement de façon convaincante. Contrairement au lieu commun "on peut faire dire aux chiffres ce qu'on veut", les chiffres sont têtus : placés entre les mains d'un chercheur ou d'un étudiant initié à la statistique, ils sont une source de données objectives dont il faut savoir extraire prudemment l'information.*

# Chapitre 4

# Distributions conjointes, marginales et conditionnelles

*"Conjointes" signifie "mises ensemble". Le but est donc d'étudier le comportement simultané de deux variables sur une même population. L'objectif visé à terme sera de voir s'il existe un lien entre les deux variables, mais ceci fera l'objet des deux chapitres suivants.*

*Soient deux variables $X$ et $Y$ définies sur $\Omega$. Le couple $(X, Y)$ est une variable définie sur $\Omega$ dont les modalités sont les couples $(x, y)$ où $x$ est une modalité de $X$ et $y$ est une modalité de $Y$.*

## 4.1 Distribution conjointe

### 4.1.1 Effectifs conjoints

*En reprenant l'exemple A, considérons sur la population des jeunes de moins de 30 ans de taille $N = 20$, la variable "goût pour la lecture" notée $Y$ et la variable "âge" notée $X$ avec le regroupement en classes déjà utilisé.*

| $X$ \ $Y$ | *faible* | *moyen* | *fort* |
|---|---|---|---|
| *[22 ;24[* | *0* | ***2*** | *3* |
| *[24 ;26[* | *3* | *2* | *4* |
| *[26 ;28[* | *0* | *0* | *2* |
| *[28 ;30[* | *1* | *0* | *1* |
| *[30 ;32[* | *0* | *0* | *2* |

*Pour ce tableau, comme pour les tableaux d'effectifs précédents, $y_j$ désigne les différentes modalités de Y ($y_1$=faible, $y_2$=moyen, $y_3$ =fort) et $x_i$ désigne les classes de modalités de X ($x_1$=[22 ;24[, $x_2$=[24 ;26[, $x_3$=[26 ;28[, $x_4$=[28 ;30[, $x_5$=[30 ;32[).*

*Dans la case située à l'intersection de la ligne 1 correspondant à $x_1$ et de la colonne 2 correspondant à $y_2$, on écrit le nombre d'individus ayant simultanément les modalités $x_1$ et $y_2$. Cette valeur est appelée* ***effectif conjoint*** *des modalités $x_1$ et $y_2$ ; on le note $n_{12}$. Ici $n_{12} = 2$ (en gras dans le tableau ci-dessus).*

***Attention, il faut lire " n un deux " et pas " n douze ".***

*En fait, $n_{12}$ est l'effectif conjoint situé à l'intersection de 1ère ligne et de la 2ième colonne. C'est comme pour la bataille navale !*

*Plus généralement, dans la case située à l'intersection de la ligne $x_i$ et de la colonne $y_j$ , on écrit le nombre d'individus ayant simultanément les modalités $x_i$ et $y_j$. Cette valeur est appelée* ***effectif conjoint des modalités*** *$x_i$* ***et*** *$y_j$ ; on le note $n_{ij}$.*

*Par exemple, $n_{51} = 0$ est le nombre d'individus dont l'âge est dans [30 ;32[ et dont le goût pour la lecture est faible.*

*Ce tableau est appelé* ***tableau des effectifs conjoints*** *ou* ***table de contingence*** *ou encore* ***tableau croisé****, on dit également parfois* ***distribution conjointe des effectifs****.*

### 4.1.2 Effectifs marginaux

| X \ Y | *faible* | *moyen* | *fort* | *Marge de X* |
|---|---|---|---|---|
| *[22 ;24[* | *0* | *2* | *3* | ***5*** |
| *[24 ;26[* | *3* | *2* | *4* | *9* |
| *[26 ;28[* | *0* | *0* | *2* | *2* |
| *[28 ;30[* | *1* | *0* | *1* | *2* |
| *[30 ;32[* | *0* | *0* | *2* | *2* |
| *Marge de Y* | *4* | *4* | *12* | *N=20* |

*On complète la table de contingence en indiquant la somme de chaque ligne et de chaque colonne. La somme de la i ième ligne est notée $L_i$ (c'est le nombre d'individus ayant la modalité $x_i$) ; la somme de la j ème colonne est notée $C_j$ (c'est le nombre d'individus ayant la modalité $y_j$).*

*Par exemple : $L_1 = 5$ et $C_2 = 4$.*

*La colonne intitulée "Marge de X" est alors la table d'effectifs de X, d'où son nom.*

*La ligne intitulée "Marge de $Y$" est alors la table d'effectifs de $Y$, d'où son nom.*
*Ces deux marges sont parfois appelées* ***distributions marginales des effectifs****.*

**Exemple B.** *Notons $X$ la variable PNB/h (regroupée en classes) et $Y$ la variable "Membre de l'U.E." (avec les modalités OUI et NON). À partir du tableau de données on obtient le tableau des effectifs conjoints :*

| | | *PNB par habitant en milliers de dollars* | | | | | |
|---|---|---|---|---|---|---|---|
| | | *]0 ;10]* | *]10 ;20]* | *]20 ;30]* | *]30 ;40]* | *]40 ;50]* | *Marge de Y* |
| *Membre* | *OUI* | *0* | *5* | *8* | *1* | *1* | *15* |
| *de l'U.E.* | *NON* | *12* | *1* | *0* | *0* | *0* | *13* |
| | *Marge de X* | *12* | *6* | *8* | *1* | *1* | *N=28* |

*On retrouve dans les marges du tableau croisé les distributions des effectifs de $X$ et de $Y$ en additionnant les effectifs de chaque colonne (pour $X$) et les effectifs de chaque ligne (pour $Y$).*

**Exemple B.** *Représentons la table de contingence du couple de (Population, Membre).*

| | | *Population* | | | | | |
|---|---|---|---|---|---|---|---|
| | | $< 1$ | *[1 ;10[* | *[10 ;30[* | *[30 ;50[* | $50 \leq$ | *Marge de Y* |
| *Membre* | *OUI* | *1* | *6* | *3* | *1* | *4* | *15* |
| *de l'U.E.* | *NON* | *2* | *6* | *3* | *1* | *1* | *13* |
| | *Marge de X* | *3* | *12* | *6* | *2* | *5* | *N=28* |

### 4.1.3 Distributions conjointes et marginales de fréquences

*En divisant chacun des effectifs du tableau précédent, on obtient le* ***tableau des fréquences conjointes*** *(appelé aussi parfois* ***distribution conjointe des fréquences****) complété par les deux marges appelées parfois* ***distributions marginales des fréquences****.*

**Exemple A.**

| *Y* / *X* | *faible* | *moyen* | *fort* | *Marge de X* |
|---|---|---|---|---|
| *[22 ;24[* | *0* | ***0,1*** | *0,15* | *0,25* |
| *[24 ;26[* | *0,15* | *0,1* | *0,2* | *0,45* |
| *[26 ;28[* | *0* | *0* | *0,1* | *0,1* |
| *[28 ;30[* | *0,05* | *0* | *0,05* | *0,1* |
| *[30 ;32[* | *0* | *0* | *0,1* | *0,1* |
| *Marge de Y* | *0,2* | *0,2* | *0,6* | *1* |

*Les **fréquences conjointes** sont notée $f_{ij}$ (même principe que pour les $n_{ij}$). Ainsi*

$$f_{ij} = \frac{n_{ij}}{N}$$

*$f_{ij}$ est donc la proportion d'individus dans la population ayant simultanément la modalité $x_i$ et la modalité $y_j$.*

**Exemple A.** *Par exemple, $f_{12} = \frac{n_{12}}{N} = \frac{2}{20} = 0,10$ (en gras dans le tableau précédent).*

**Exemple B.** *On obtient les fréquences conjointes et les **fréquences marginales** en divisant chaque nombre du tableau par 28 (taille de la population).*

## 4.2 Distributions conditionnelles

**Exemple A.** *On considère uniquement les individus dont le goût pour la lecture est faible (on restreint la population aux individus aimant peu lire) et on détermine les fréquences en pourcentage de la variable âge sur cette partie de la population. Les individus dont le goût pour la lecture est faible représenteront donc 100%.*
*On a donc obtenu les fréquences de la variable âge **si on ne considère que les individus dont le goût pour la lecture est faible**; pour cette raison, on appelle ces fréquences, les **fréquences de $X$ conditionnellement à $Y$=faible**.*

| $X$ \ $Y$ | *faible* |
|---|---|
| *[22 ;24[* | *0,00%* |
| *[24 ;26[* | *75,00%* |
| *[26 ;28[* | *0,00%* |
| *[28 ;30[* | *25,00%* |
| *[30 ;32[* | *0,00%* |
| *Total* | *100,00%* |

*De la même façon on peut déterminer les fréquences de $X$ conditionnellement à $Y$=moyen. Ce sont les fréquences de la variable âge **si on ne considère que les individus dont le goût pour la lecture est moyen**.*
*On peut enfin déterminer les fréquences de $X$ conditionnellement à $Y$=fort. Ce sont les fréquences de la variable âge **si on ne considère que les individus dont le goût pour la lecture est fort**.*

*Le tableau ci-dessous regroupe les fréquences conditionnelles par rapport aux 3 modalités de $Y$. On l'appelle* ***tableau de fréquences de $X$ conditionnellement à $Y$.***

| $X$ \ $Y$ | *faible* | *moyen* | *fort* |
|---|---|---|---|
| *[22 ;24[* | *0,00%* | *50,00%* | *25,00%* |
| *[24 ;26[* | *75,00%* | *50,00%* | *33,33%* |
| *[26 ;28[* | *0,00%* | *0,00%* | *16,67%* |
| *[28 ;30[* | *25,00%* | *0,00%* | *8,33%* |
| *[30 ;32[* | *0,00%* | *0,00%* | *16,67%* |
| *Total* | *100,00%* | *100,00%* | *100,00%* |

*La fréquence de $x_i$ conditionnellement à $y_j$ sera la proportion d'individus ayant la modalité $x_i$ parmi ceux ayant la modalité $y_j$.*

***Attention, ajouter les valeurs d'une même ligne n'aurait aucun sens : en effet, chaque pourcentage d'une même ligne est un pourcentage par rapport à une partie différente de la population.***

*De la même façon, on peut déterminer les fréquences de $Y$ conditionnellement à $X$.*

**Remarque 4.1.** *Pour éviter les confusions et pour la résolution des exercices, on regarde la variable par laquelle on conditionne, on détermine ses modalités et cela nous donne ainsi le nombre de sous-populations à considérer.*

**Exemple A.**

| $X$ \ $Y$ | *faible* | *moyen* | *fort* | *Total* |
|---|---|---|---|---|
| *[22 ;24[* | *0,00%* | *40,00%* | *60,00%* | *100,00%* |
| *[24 ;26[* | *33,33%* | *22,22%* | *44,45%* | *100,00%* |
| *[26 ;28[* | *0,00%* | *0,00%* | *100,00%* | *100,00%* |
| *[28 ;30[* | *50,00%* | *0,00%* | *50,00%* | *100,00%* |
| *[30 ;32[* | *0,00%* | *0,00%* | *100,00%* | *100,00%* |

*Ainsi par exemple, la deuxième ligne représente les fréquences de la variable " goût pour la lecture " conditionnellement à $X = [24; 26[$ (c'est-à-dire si on ne considère que les individus dont l'âge est dans $[24; 26[$).*

**Exemple B.** *Reprenons la table de contingence du couple de (Population, Membre) et représentons la distribution de la variable Population conditionnellement à la variable Membre.*

| | | Population | | | | | |
|---|---|---|---|---|---|---|---|
| | | $< 1$ | $[1\,;10[$ | $[10\,;30[$ | $[30\,;50[$ | $50 \leq$ | *Total* |
| *Membre de l'U.E.* | *OUI* | *6,67%* | *40%* | *20%* | *6,67%* | *26,67%* | *100%* |
| | *NON* | *15,38%* | *46,15%* | *23,08%* | *7,69%* | *7,69%* | *100%* |

*Représentons maintenant la distribution de la variable Membre conditionnellement à la variable Population.*

| | | Population | | | | |
|---|---|---|---|---|---|---|
| | | $< 1$ | $[1\,;10[$ | $[10\,;30[$ | $[30\,;50[$ | $50 \leq$ |
| *Membre de l'U.E.* | *OUI* | *33,33%* | *50%* | *50%* | *50%* | *80%* |
| | *NON* | *66,67%* | *50%* | *50%* | *50%* | *20%* |
| | *Total* | *100%* | *100%* | *100%* | *100%* | *100%* |

## 4.3 Représentations graphiques

### 4.3.1 Histogramme de distribution conjointe

*On utilisera le même principe que pour une seule variable mais chaque couple $(x_i, y_j)$ de modalités est représenté par un rectangle.*

**Exemple A.** *Pour les variables "goût pour la lecture" et "âge", on obtient la Figure 4.1.*

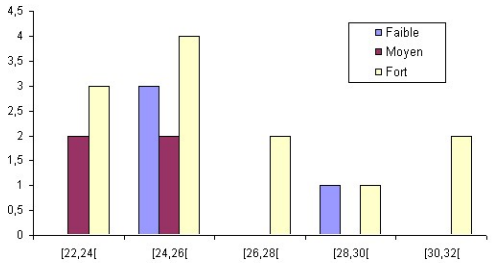



FIGURE 4.1 – Distribution conjointe des effectifs des variables âge et goût de l'exemple A

*Si on "range" différemment les rectangles, on obtient la Figure 4.2. Tout dépend de ce que l'on veut mettre en lumière.*

**Exemple B.** *Pour les variables PNB/hab et Membre, on obtient les Figures 4.3 et 4.4.*

### 4.3.2 Histogramme des distributions conditionnelles

*Pour construire un graphique représentant une variable $X$ conditionnellement à une variable $Y$, on trace pour chaque modalité de $Y$ un rectangle de hauteur 100. Ensuite, chaque rectangle est*

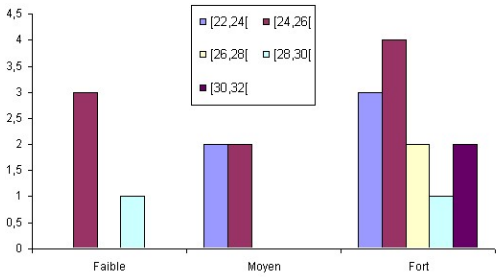



FIGURE 4.2 – Distribution conjointe des effectifs des variables âge et goût de l'exemple A

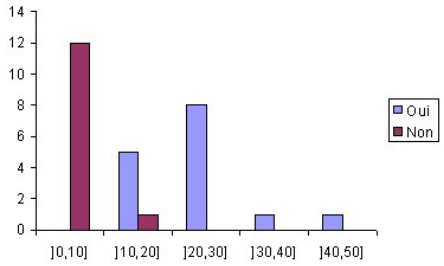



FIGURE 4.3 – Distribution conjointe des effectifs des variables PNB/hab et Membre de l'exemple B

*partagé proportionnellement suivant les modalités de $X$.*

**Exemple A.** *On obtient alors la Figure 4.5. On a donc ici une barre de hauteur 100 par modalité de la variable goût pour la lecture. On remarque que selon les classes d'âge, la répartition du goût pour la lecture est très différente. On a donc mis en évidence l'existence d'un lien entre le goût pour la lecture et l'âge. Puis si on conditionne par l'âge, on obtient la Figure 4.6. On a donc ici une barre de hauteur 100 par classe de modalités de la variable âge.*

**Exemple B.** *De même, un diagramme en bandes représentant pour chaque classe de la variable*

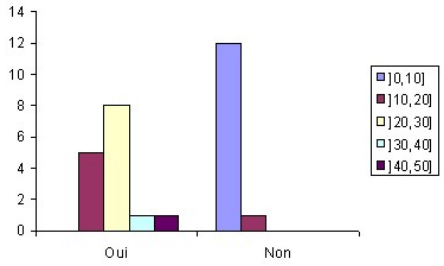



FIGURE 4.4 – Distribution conjointe des effectifs des variables PNB/hab et Membre de l'exemple B

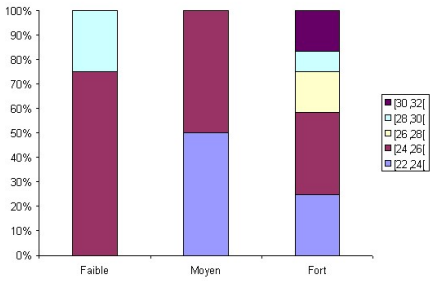



FIGURE 4.5 – Représentation graphique de l'âge conditionnellement au goût pour la lecture

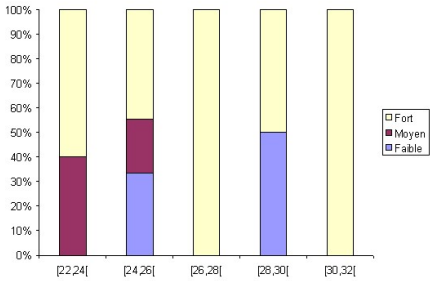



FIGURE 4.6 – Représentation graphique du goût pour la lecture conditionnellement à l'âge

*PNB/hab la proportion de pays membres de l'U.E. permet de mettre en évidence le lien entre X et Y (ici de façon évidente) comme le montre la Figure 4.7.*

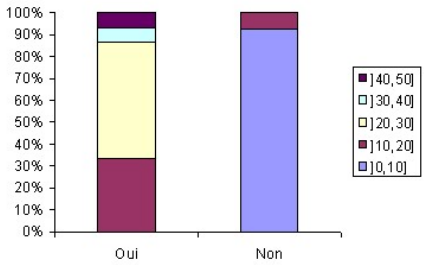



FIGURE 4.7 – Représentation graphique du PNB/hab conditionnellement à Membre

*On a aussi la Figure 4.8 en choisissant de conditionner par la variable PNB par habitant. Un diagramme en bandes représentant pour chaque classe de la variable Population la proportion de pays membres de l'U.E. met en évidence le faible lien entre X et Y (cf. Figure 4.9). Concrètement, on ne peut pas dire que les pays membres sont en général plus peuplés que les pays candidats ni l'inverse. On a aussi la Figure 4.10.*

**Remarque 4.2.** – *On vient donc de voir que les distributions conditionnelles permettent de mettre en évidence graphiquement l'existence d'un lien entre deux variables. Nous allons*

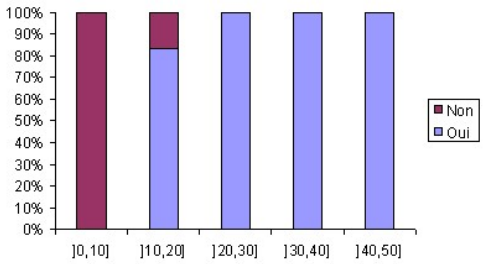



FIGURE 4.8 – Représentation graphique de Membre conditionnellement à PNB/hab

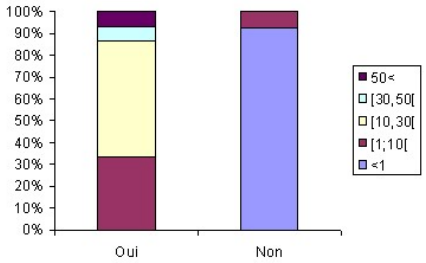



FIGURE 4.9 – Représentation graphique du Population conditionnellement à Membre

*construire dans le prochain chapitre un nombre permettant de déterminer analytiquement l'existence d'un lien entre deux variables et le cas échéant nous quantifierons l'importance de ce lien.*

- *Pour mettre en évidence l'existence d'un lien entre les variables $X$ et $Y$, il n'est pas nécessaire de représenter à la fois la distribution de $X$ conditionnellement à $Y$ et la distribution de $Y$ conditionnellement à $X$. Une seule suffit et les conclusions seront les mêmes.*
- *Une autre façon de mettre en évidence l'existence d'un lien entre les variables $X$ et $Y$ est de faire des boîtes à moustaches pour chacune des sous-populations correspondant aux différentes modalités de $X$ par exemple. Mais ceci nécessite d'avoir de grands échantillons*

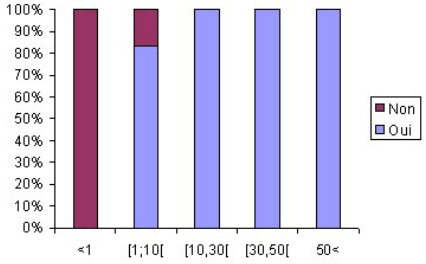



FIGURE 4.10 – Représentation graphique de Membre conditionnellement à Population

*(et de grandes sous-populations) et que $Y$ soit une variable quantitative (pour pouvoir déterminer la médiane et les quartiles et les représenter).*

# Chapitre 5

# Indices de liaison entre deux variables quelconques

**Remarque 5.1.** *Les variables ordinales ou quantitatives pouvant être traitées comme nominales, ce paragraphe concerne aussi les variables ordinales ou quantitatives.*

*On travaille ici sur l'exemple A : avec la variable X qui associe à chaque individu son intérêt pour la lecture, avec les modalités : "Fort", "Moyen", "Faible" et Y celle qui associe à chaque individu sa spécialité scolaire, sachant que seules les modalités " L ", " ES " et " S " ont été observées. On peut remarquer que la variable Y est nominale et la variable X ordinale.*
*Voici le tableau des effectifs conjoints observés.*

***Notation*** *: Comme les effectifs conjoints sont qualifiés d'observés, on les notera à partir de maintenant $O_{ij}$ au lieu de $n_{ij}$.*

| $X$ \ $Y$ | $L$ | $ES$ | $S$ | Marge de $X$ |
|---|---|---|---|---|
| Fort | 10 | 1 | 1 | 12 |
| Moyen | 0 | 4 | 0 | 4 |
| Faible | 0 | 0 | 4 | 4 |
| Marge de $Y$ | 10 | 5 | 5 | $N$=20 |

*En examinant ce tableau, on peut constater l'existence d'un lien entre les deux variables. En effet, les individus ayant pour spécialité L lisent tous beaucoup. De même, la spécialité ES est plutôt associée à "Moyen" pour la lecture et la spécialité S à "Faible" pour la lecture même si cela est moins net que pour la série L.*

*Afin de se conforter dans cette opinion, nous pourrions représenter la distribution de $Y$ conditionnellement à $X$ ou de façon équivalente la distribution de $X$ conditionnellement à $Y$. Par exemple, représentons la distribution de $Y$ conditionnellement à $X$. Remarquons que $X$ a trois modalités "Fort", "Moyen" et "Faible" : on va donc considérer ces trois sous-populations. On obtient*

| $X$ \ $Y$ | $L$ | $ES$ | $S$ | |
|---|---|---|---|---|
| *Fort* | *86,67%* | *8,67%* | *8,67%* | *100%* |
| *Moyen* | *0* | *100%* | *0* | *100%* |
| *Faible* | *0* | *0* | *100%* | *100%* |

*Traçons maintenant l'histogramme associé.*

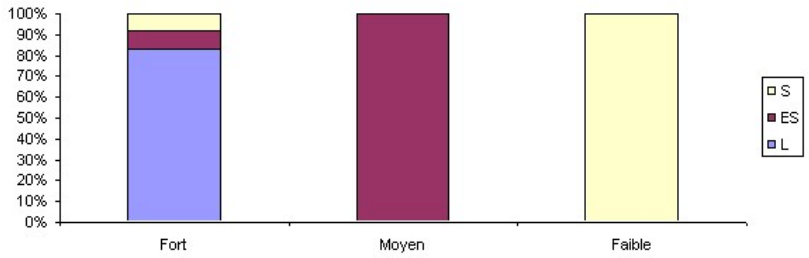



FIGURE 5.1 – Histogramme de la spécialité scolaire conditionnellement au goût pour la lecture

*De la même façon, nous aurions pu étudier la distribution de $X$ conditionnellement à $Y$ :*

| $X$ \ $Y$ | $L$ | $ES$ | $S$ |
|---|---|---|---|
| *Fort* | *100%* | *20%* | *20%* |
| *Moyen* | *0* | *80%* | *0* |
| *Faible* | *0* | *0* | *80%* |
| | *100%* | *100%* | *100%* |

*L'analyse de n'importe quel de ces histogrammes nous conduit à la même conclusion : il semble exister un lien fort entre les variables goût pour la lecture et spécialité scolaire.*

*Nous voudrions quantifier ces deux remarques. Le but de ce chapitre est de construire un outil permettant de détecter l'existence d'un lien entre deux variables $X$ et $Y$ à partir de la table de contingence et d'en mesurer l'importance le cas échéant.*

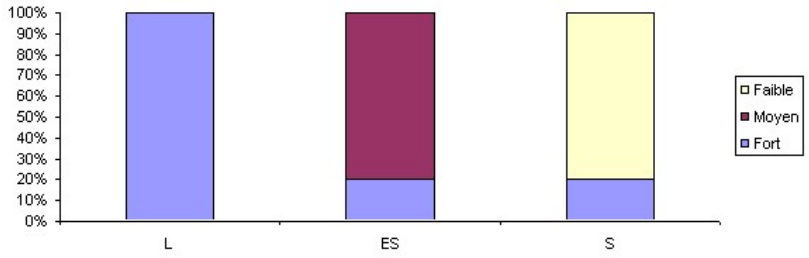



FIGURE 5.2 – Histogramme du goût pour la lecture conditionnellement à la spécialité scolaire

## 5.1 Effectifs théoriques

*On va maintenant construire un tableau d'effectifs de mêmes dimensions que celui des effectifs conjoints ci-dessus et avec les mêmes marges. La comparaison de ce nouveau tableau avec celui des effectifs conjoints nous permettra de dire s'il y a un lien entre les variables "intérêt pour la lecture" et "spécialité scolaire". Si ce n'est pas le cas, on dira alors que les deux variables sont* ***indépendantes****.*

***Attention, on s'impose de ne pas modifier les marges du tableau précédent****.*

*Il s'agit donc de compléter le tableau suivant :*

| $Y$ / $X$ | *Série L* | *Série ES* | *Série S* | *Marge de X* |
|---|---|---|---|---|
| *Fort* | $T_{11} =$ | $T_{12} =$ | $T_{13} =$ | $L_1 = 12$ |
| *Moyen* | $T_{21} =$ | $T_{22} =$ | $T_{23} =$ | $L_2 = 4$ |
| *Faible* | $T_{31} =$ | $T_{32} =$ | $T_{33} =$ | $L_3 = 4$ |
| *Marge de Y* | $C_1 = 10$ | $C_2 = 5$ | $C_3 = 5$ | $N = 20$ |

*Les effectifs conjoints que l'on cherche ne sont pas observés ; ce sont des* ***effectifs théoriques*** *: ceux que l'on aurait obtenu s'il y avait eu indépendance entre $Y$ et $X$ sur $\Omega$. Aussi, on les note $T_{ij}$.*

*Maintenant, dire qu'il n'y a pas de lien entre la spécialité scolaire et l'intérêt pour la lecture, signifie par exemple que l'on va trouver autant d'individus aimant beaucoup la lecture (en proportion) dans chaque spécialité scolaire. Or, la proportion d'individus aimant beaucoup la lecture dans $\Omega$ est*

$$\frac{L_1}{N} = \frac{12}{20} = 0,6,$$

*il faudra donc que, pour chaque spécialité scolaire, la proportion d'individus aimant beaucoup la lecture soit 0,6.*

• *Mais la proportion d'individus aimant beaucoup la lecture pour la spécialité scolaire L est* $\frac{T_{11}}{C_1} = \frac{T_{11}}{10}$. *On aura donc* $\frac{T_{11}}{10} = 0,6$ *soit* $T_{11} = 6$.
*Remarquons que l'on a exploité la relation :*

$$\frac{T_{11}}{C_1} = \frac{T_{11}}{10} = \frac{L_1}{N} = 0,6$$

• *De même, la proportion d'individus aimant beaucoup la lecture pour la spécialité scolaire ES est* $\frac{T_{12}}{C_2} = \frac{T_{12}}{5}$. *On aura donc* $\frac{T_{12}}{5} = 0,6$ *soit* $T_{12} = 3$.
*Remarquons que l'on a exploité la relation :*

$$\frac{T_{12}}{C_2} = \frac{T_{12}}{5} = \frac{L_1}{N} = 0,6$$

• *De même, la proportion d'individus aimant beaucoup la lecture pour la spécialité scolaire S est* $\frac{T_{13}}{C_3} = \frac{T_{13}}{5}$. *On aura donc* $\frac{T_{13}}{5} = 0,6$ *soit* $T_{13} = 3$.
*Remarquons que l'on a exploité la relation :*

$$\frac{T_{13}}{C_3} = \frac{T_{13}}{5} = \frac{L_1}{N} = 0,6$$

*En faisant le même raisonnement pour les individus aimant moyennement puis ceux aimant peu la lecture, on détermine alors les autres effectifs théoriques. Cela donne le tableau des effectifs conjoints théoriques suivant :*

| Y / X | L | ES | S | Marge de X |
|---|---|---|---|---|
| Fort | 6 | 3 | 3 | 12 |
| Moyen | 2 | 1 | 1 | 4 |
| Faible | 2 | 1 | 1 | 4 |
| Marge de Y | 10 | 5 | 5 | N=20 |

*En regardant les trois relations utilisées ci-dessus, on voit que l'on a d'une façon générale :*

$$\frac{T_{ij}}{C_j} = \frac{L_i}{N} \quad \text{ce qui donne} \quad T_{ij} = \frac{L_i C_j}{N}$$

*Donc, pour obtenir un effectif théorique, en regardant la case où il se trouve, il suffit de multiplier la somme de sa ligne par la somme de sa colonne et diviser le résultat par N.*
*Ainsi par exemple :* $T_{23} = \frac{L_2 C_3}{N} = \frac{4 \times 5}{20} = 1$.

**Remarque 5.2.** *(i) Notez bien que les marges doivent être conservées.*

*(ii)* ***Attention, bien que les*** $T_{ij}$ ***représentent des effectifs, il ne faut surtout pas les***

***arrondir en nombres entiers. En général, le résultat calculé ne tombe pas juste ; aussi il faut garder au moins 2 chiffres après la virgule afin d'éviter de cumuler ces erreurs d'arrondis par la suite.***

## 5.2 Le chi-deux d'indépendance noté $\chi^2$

*Idée : Puisque le second tableau contient les effectifs que l'on devrait avoir si les variables $X$ et $Y$ étaient indépendantes, on voudrait le comparer à la table de contingence pour déterminer l'existence d'une liaison entre $X$ et $Y$. Or l'absence de liaison (l'indépendance) est caractérisée par les $T_{ij}$ (les effectifs théoriques). Donc plus les $O_{ij}$ (les effectifs observés) sont éloignés des $T_{ij}$ et plus la liaison est grande.*

*Pour cela, on va fabriquer un nombre qui permette de comparer ces deux tableaux afin de déterminer s'il y a une liaison entre deux variables nominales (et éventuellement en mesurer l'importance).*

***Construction du $\chi^2$ d'indépendance***

*Etape 1 : Pour chaque $O_{ij}$ on mesure l'écart avec le $T_{ij}$ correspondant en calculant la différence : $O_{ij} - T_{ij}$.*

*Etape 2 : On pourrait ensuite faire la somme de toutes ces différences pour avoir un résultat général. Seulement, on aura alors des différences positives et des différences négatives qui pourront éventuellement se neutraliser donnant un résultat global de 0. Pour éviter cela, il faut ne pas tenir compte des signes de ces différences (de la même façon on dit que la distance de Paris à Toulouse est 700 km et la distance de Toulouse à Paris est 700 km et pas -700 km). Comme généralement en mathéatiques (et de même que pour le calcul de la variance), plutôt que de supprimer les signes négatifs des différences, on préfère élever au carré chaque différence. Ceci permet également de n'avoir que des résultats positifs.*

*Etape 3 : Tenir compte uniquement de $(O_{ij} - T_{ij})^2$ ne suffit pas. Il faut relativiser le résultat par rapport aux valeurs en jeu : ainsi un écart au carré de 25 est grand comparativement à un effectif de 10, alors que 25 est petit par rapport à un effectif de 100. Aussi pour chaque $(O_{ij} - T_{ij})^2$, on fait le rapport avec le $T_{ij}$ correspondant qui présente la valeur de référence. Cela donne pour*

*chaque case du tableau le nombre* $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$ *que l'on appelle une* ***contribution****.*

*Dernière étape : Il reste alors à faire la somme des contributions de chacune des cases. Le résultat est appelé* $\chi^2$*. Cela s'écrit :*

$$\chi^2 = \Sigma\frac{(O_{ij}-T_{ij})^2}{T_{ij}}.$$

**Remarque 5.3.** *1)* $\chi^2$ *s'écrit chi-deux et se lit " ki-deux ".*
*2)* $\chi^2 = \Sigma\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$ *se lit : "ajouter pour chaque case du tableau le nombre* $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$*.*
*3) Le* $\chi^2$ *remplit bien l'objectif fixé : il ne peut être nul que si chaque* $O_{ij}$ *est égal au* $T_{ij}$ *correspondant. Plus les* $O_{ij}$ *sont éloignés des* $T_{ij}$ *et plus le* $\chi^2$ *est grand. Le* $\chi^2$ *représente donc bien globalement l'écart entre les effectifs théoriques et les effectifs observés, c'est-à-dire qu'il donne une mesure de l'écart entre la situation observée et l'indépendance entre* $X$ *et* $Y$ *sur* $\Omega$*.*
*4)* ***Attention, rien ne permet à ce stade de dire si la valeur obtenue en calculant*** $\chi^2$ ***est petite ou grande !*** *On ne dispose en effet pas d'éléments de référence (cet aspect sera abordé au paragraphe suivant). Ainsi par exemple un écart d'un gramme est faible si on pèse un éléphant, alors qu'il énorme si on pèse un insecte ! Aussi le seul commentaire possible à ce stade concerne le fait que le* $\chi^2$ *est nul (les deux variables sont dont indépendantes) ou bien au contraire n'est pas nul (les deux variables sont donc dépendantes ou liées).*

***Méthode de calcul par tableau sur l'exemple A***

*Chaque ligne du tableau ci-dessous correspond à une case du tableau d'effectifs conjoints précédent.*

| $O_{ij}$ | $T_{ij}$ | $(O_{ij}-T_{ij})^2$ | $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$ |
|---|---|---|---|
| 10 | 6 | 16 | 2,67 |
| 0 | 2 | 4 | 2 |
| 0 | 2 | 4 | 2 |
| 1 | 3 | 4 | 1,33 |
| 4 | 1 | 9 | 9 |
| 0 | 1 | 1 | 1 |
| 1 | 3 | 4 | 1,33 |
| 0 | 1 | 1 | 1 |
| 4 | 1 | 9 | 9 |
| $N = 20$ | $N = 20$ | - | $\chi^2 = 29{,}33$ |

*La somme de la dernière colonne nous donne donc la valeur du* $\chi^2 = 29{,}33$*.*

**Remarque 5.4.** *Afin d'éviter de cumuler les erreurs d'arrondis, il est préférable de calculer la contribution d'une case dès que l'on a son effectif théorique, en mettant cet effectif théorique en mémoire et en conservant ainsi toutes les décimales du calcul. (Pour cela, il faut identifier sur votre machine à calculer les touches de mise en mémoire et de rappel de mémoire.)*

## 5.3 Le coefficient phi noté $\varphi$

*Lorsque le $\chi^2$ des variables $X$ et $Y$ calculé sur toute la population $\Omega$ n'est pas nul, on sait qu'il existe un lien entre les deux variables mais pour connaître l'importance de ce lien, on a besoin de savoir quelle est la valeur maximale que peut prendre le $\chi^2$. On peut montrer que l'on a toujours :*

$$\underset{ind\acute{e}pendance}{0} \leq \chi^2 \leq \underset{d\acute{e}pendance\ totale}{N\left[\min(p,q)-1\right]}$$

*où*

- *$p$ est le nombre de lignes de la table de contingence de $X$ et $Y$ (ou nb de modalités de $X$) ;*
- *$q$ est le nombre de colonnes de la table de contingence de $X$ et $Y$ (ou nb de modalités de $Y$) ;*
- *$\min(p,q)$ est simplement le plus petit des deux nombres $p$ et $q$ ;*
- *$N$ est toujours la taille de la population.*

*Pour l'exemple précédent, $p$=3 et $q$=3 donc $\min(p,q)=3$. Donc*

$$\underset{ind\acute{e}pendance}{0} \leq \chi^2 \leq \underset{d\acute{e}pendance\ totale}{40}$$

*On peut maintenant mesurer l'importance du lien entre $X$ et $Y$ sur $\Omega$ en comparant $\chi^2$ et $N\left[\min(p,q)-1\right]$ en faisant le rapport des deux nombres. Pour cela, on construit le coefficient $\varphi$ où*

$$\boxed{\varphi = \sqrt{\frac{\chi^2}{N\left[\min(p,q)-1\right]}}.}$$

**Remarque 5.5.** *1) $\varphi$ s'écrit "phi" et se lit "fi".*
*2) On peut expliquer la racine carrée en remarquant qu'elle permet de "compenser" le fait que dans $\chi^2$ interviennent les carrés des écarts entre les $O_{ij}$ et $T_{ij}$.*

***Propriétés** : L'avantage de ce nouveau coefficient est qu'il est toujours compris entre 0 et 1.*

$$\underset{ind\acute{e}pendance}{0} \leq \varphi \leq \underset{d\acute{e}pendance\ totale}{1}$$

***Convention*** *: On considère en général que le lien est*

- *faible lorsque $\varphi$ est inférieur à 0,3 ;*
- *moyen lorsque $\varphi$ est compris entre 0,3 et 0,5 ;*
- *fort lorsque $\varphi$ est supérieur à 0,5.*

*Toujours pour notre exemple, on a*

$$\varphi = \sqrt{\frac{\chi^2}{N\left[\min(p,q)-1\right]}} = \sqrt{\frac{29,33}{40}} = 0,86.$$

*0,86 étant proche de 1, on peut conclure à l'existence d'un lien fort entre $X$ et $Y$ sur $\Omega$ c'est-à-dire entre l'intérêt pour la lecture et la spécialité scolaire pour les 20 individus de l'enquête.*

**Remarque 5.6.** *a) Avoir une dépendance ou un lien total signifierait que connaissant la série de n'importe lequel des 20 individus on peut en déduire sans erreur son intérêt pour la lecture, ou bien réciproquement, que connaissant l'intérêt pour la lecture n'importe lequel des 20 individus on peut sans erreur en déduire sa spécialité scolaire.*
*b) On rencontre aussi les expressions "variable indépendante" et "variable indépendante" en méthodologie. Ces expressions n'ont rien à voir avec les notions de dépendance et d'indépendance abordées dans ce chapitre. Il est donc important de bien différencier les deux types d'emploi des termes "indépendante" et "dépendante". Se reporter au cours de méthodologie pour bien comprendre le sens des deux expressions évoquées dans cette remarque et qui n'ont donc rien à voir avec leur utilisation en statistique dans ce chapitre.*
*c) Il existe un coefficient appelé coefficient de corrélation linéaire qui permet de déterminer l'existence d'un lien particulier : un lien linéaire entre deux variables $X$ et $Y$ quantitatives. Le cas échéant, il est ensuite possible de déterminer une droite appelée droite de régression linéaire qui sera une bonne approximation de nos données et qui permettra de faire des prédictions. Ces notions et techniques ne font pas l'objet de ce cours.*

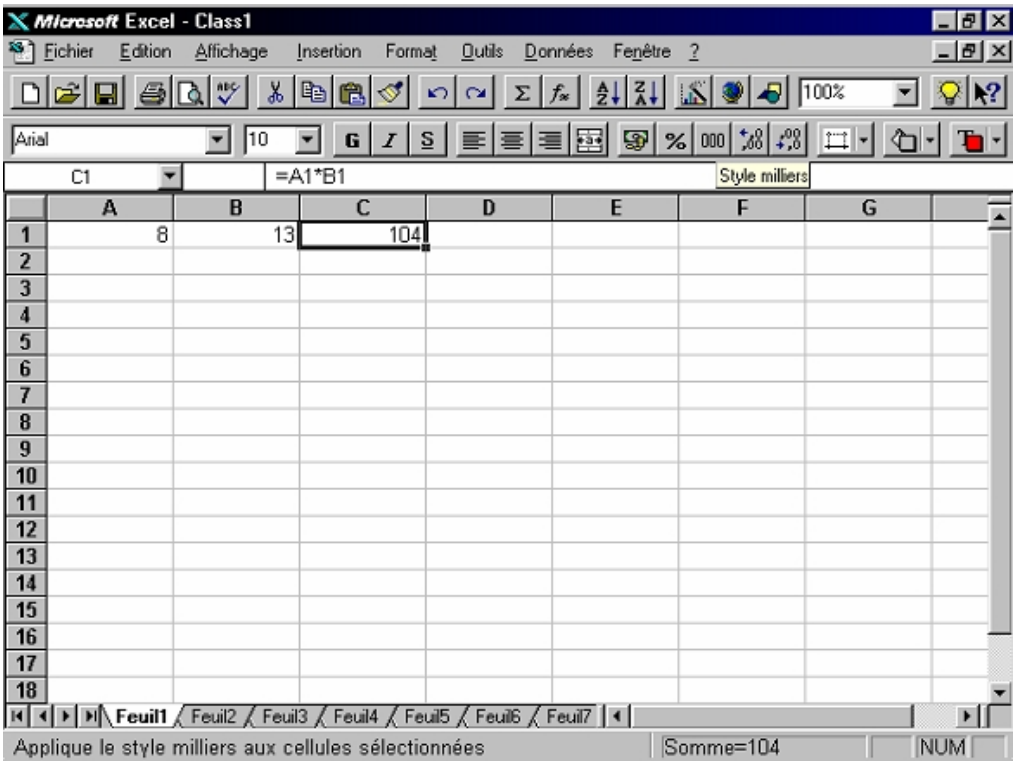
# Chapitre 6

# Utilisation d'un tableur

*EXCEL est un logiciel qui permet d'effectuer des calculs, particulièrement des calculs répétitifs et d'avoir instantanément la mise à jour des résultats lors de changements des données.*

## 6.1 Découverte du logiciel

*Premiers pas avec le logiciel*



*Pour lancer EXCEL, cliquer sur* ***démarrer*** *puis* ***tous les programmes*** *puis* ***Microsoft EXCEL****.*

***Les cellules***

*Le quadrillage situé sous les barres d'outils s'appelle une* **feuille de calcul***. Les cases sont appelées des* ***cellules****. Chaque cellule est repérée par ses coordonnées : colonne, ligne (et non pas ligne, colonne comme cela est fait en mathématique en général).*

**Exemple** *:* ***D5*** *est le nom de la cellule située à l'intersection de la colonne D et de la ligne 5.*

***Plage de cellules***

*Chaque plage rectangulaire de cellules est repérée par la cellule située à l'angle supérieur gauche et par la cellule située à l'angle inférieur droit.*
*Par exemple,* ***B3 :E8*** *désigne une plage rectangulaire de cellules à 4 colonnes et 6 lignes. (Ne pas oublier le symbole : entre* ***B3*** *et* ***E8*** *.)*

***Cliquer sur une cellule***

*Cela signifie amener le curseur de la souris (une croix assez grosse) sur la cellule, puis appuyer sur le bouton gauche de la souris.*

***Copier une plage de cellules***

*On va ici copier la plage de cellules* ***A1 :B7*** *à l'emplacement de la plage* ***E1 :F7****.*

- *Sélectionner la plage* ***A1 :B7****.*
- *Dans le menu* ***Edition*** *cliquer sur* ***copier****. La plage* ***A1 :B7*** *est maintenant mémorisée.*
- *Cliquer sur la cellule* ***E1****. (C'est la première cellule de la plage où l'on désire faire la copie)*
- *Dans le menu* ***Édition*** *cliquer sur* ***coller****.*

*Dans chaque cellule on peut rentrer du texte, un nombre ou encore une formule.*

***Rentrer un nombre ou du texte dans une cellule***

*Pour rentrer du texte ou un nombre dans une cellule, il suffit de*

- *Cliquer sur la cellule*
- *Taper le texte ou le nombre et enfin de valider, soit en tapant sur la touche* ***Entrée*** *(ou* ***Enter****, ou* ***Return****) soit en se déplaçant avec les flèches soit encore en cliquant sur une*

*autre cellule avec la souris.*

**Exemple** *: Rentrer le nombre 8 dans la cellule* ***A1*** *et le nombre 13 dans la cellule* **B1**.

***Rentrer une formule dans une cellule***

*Il suffit de suivre les étapes suivantes :*

- *Cliquer sur la cellule dans laquelle on veut rentrer la formule.*
- *Taper le signe = qui indique au programme qu'il va s'agir d'une formule.*
- *Taper ensuite la formule et valider en tapant sur la touche* ***Entrée****.*

**Exemple** *: Pour obtenir dans la cellule* ***C1*** *le produit des nombres rentrés dans les cellules* ***A1*** *et* ***B1*** *: cliquer sur la cellule* ***C1*** *puis taper* ***=A1*B1*** *et enfin taper sur la touche* ***Entrée****.*

**Remarque 6.1.** *Il apparaît alors dans la cellule* ***C1*** *le nombre 104, tandis que dans la ligne située juste sous les barres d'outils on peut lire la formule qui vient d'être tapée.*

*Intérêt de l'utilisation d'une formule*

*Changer les valeurs des cellules* ***A1*** *et* ***B1****. On remarque alors que le résultat dans la cellule* ***C1*** *est automatiquement mis à jour.*

*Autre méthode pour écrire la formule*

*Au lieu de taper* ***=A1*B1****, il est recommandé de taper =, puis cliquer sur la cellule* ***A1****, taper* ***** *et cliquer ensuite sur la cellule* ***B1*** *et enfin taper sur touche* ***Entrée****.*

*L'avantage de cette méthode est d'éviter d'avoir à écrire soi-même les noms des cellules.*

***Utilisation de la fonction de remplissage***

*Taper des nombres dans les cellules suivantes :*

***A2 A3 A4 A5 A6 A7***

***B2 B3 B4 B5 B6 B7***

*Comme pour la ligne 1, on veut obtenir le produit des 2 cases de chaque ligne dans la colonne C.*

**Exemple** *: Cliquer sur la cellule* ***C1*** *puis sans relâcher le bouton de la souris la déplacer pour sélectionner les cellules de* ***C1*** *jusqu'à* ***C7*** *(la cellule* ***C1*** *reste en blanc ce qui est normal.). Dans le menu* ***Édition*** *cliquer sur* ***Remplissage*** *et choisir* ***en bas*** *en cliquant dessus. Le programme a alors recopié la formule de la cellule* ***C1*** *mais en l'adaptant à chaque nouvelle ligne. (Lorsqu'on recopie la formule sur les lignes en dessous, le programme modifie les noms de cellules de la formule en effectuant le même décalage sur les numéros de lignes.) On a maintenant les formules suivantes :*

| | | | |
|---|---|---|---|
| ***dans C1 :*** | ***A1*B1*** | ***dans C5 :*** | ***A5*B5*** |
| ***dans C2 :*** | ***A2*B2*** | ***dans C6 :*** | ***A6*B6*** |
| ***dans C3 :*** | ***A3*B3*** | ***dans C7 :*** | ***A7*B7*** |
| ***dans C4 :*** | ***A4*B4*** | | |

*Autre méthode : Cliquer sur la cellule **C1**, relâcher le bouton de la souris, mettre le curseur de la souris sur le coin inférieur droit de la cellule **C1** (le curseur prend alors la forme d'une petite croix très fine), appuyer alors sur le bouton gauche de la souris et sans relâcher le bouton déplacer la souris pour sélectionner les cellules de **C1** jusqu'à **C7**, relâcher enfin la souris.*

**Remarque 6.2.** *Selon le même principe, on peut utiliser la fonction **Remplissage à droite** pour recopier une formule qui s'adaptera à chaque nouvelle colonne.*

***Sommer une plage de cellules***

*L'icône $\Sigma$ de la barre d'outils signifie : **somme automatique** et permet d'écrire automatiquement la formule donnant la somme des nombres d'une plage rectangulaire de cellules.*

***Exemple** : Pour obtenir dans la cellule **C8** la somme des 14 nombres de la plage **A1 :B7**, cliquer sur la cellule **C8**, puis cliquer sur l'icône $\Sigma$ de la barre d'outils, taper alors **A1 :B7** (ou bien sélectionner cette plage avec la souris), puis taper sur la touche **Entrée**.*

***Nous allons utiliser et illustrer les outils statistiques proposés par Excel sur l'exemple B du cours. Ouvrir le fichier DocUE.xls***

## 6.2 Trier

*Lorsqu'on dispose d'un tableau de données avec plusieurs variables, on peut souhaiter trier les individus en fonction de chaque variable.*

***Exemple** : Pour l'instant, les pays de l'exemple sont classés par PNB/habitant décroissant. Trions-les du plus peuplé au moins peuplé.*

- *Sélectionner la plage **A1 :G29** (c'est-à-dire tout le tableau).*
- *Dans le menu **Données**, cliquer sur **Trier**.*
- *Trier par **Population**, **Décroissant**. Cliquer sur **OK**.*

*On peut utiliser le TRI pour dresser facilement un tableau d'effectifs ou bien réaliser un regroupement en classes. A partir du tableau d'effectifs on pourra ensuite calculer la moyenne, la variance et l'écart-type de la variable. Nous verrons plus loin que EXCEL permet aussi d'avoir directement un tableau d'effectifs à partir d'un tableau de données.*

*Exercice 1 : Nous allons d'abord étudier la variable* ***F*** *qui est de type quantitatif discret; voici un tableau d'effectifs et de fréquences, avec des colonnes de calculs que nous allons utiliser pour déterminer la moyenne et la variance de* ***F*** *.*

| $F(x_i)$ | *Effectifs* $n_i$ | *Fréquences en %* | $n_i x_i$ | $n_i(x_i)^2$ |
|---|---|---|---|---|
| *1* | | | | |
| *2* | | | | |
| *3* | | | | |
| *4* | | | | |
| *5* | | | | |
| *6* | | | | |
| *7* | | | | |
| *8* | | | | |
| *9* | | | | |
| *Somme* | | | | |

*- Après avoir trié les pays selon la variable* ***F****, compléter les effectifs dans le tableau ci-dessus puis recopier ce tableau sur la feuille de calcul sous le tableau de données (cellule* ***A40****).*

*- Calculer la taille de la population (utiliser* $\Sigma$*).*

*- Compléter la colonne "Fréquences" en calculant la fréquence de la première modalité puis en utilisant la fonction de remplissage (pour que la cellule* ***B49*** *contenant la taille de la population soit conservée comme diviseur lors du remplissage, taper* ***B$49*** *au lieu de* ***B49*** *: le $ fixe la ligne 49).*

*- Sur le même principe, saisir les formules* $n_i x_i$ *et* $n_i(x_i)^2$ *pour la première ligne puis utiliser le remplissage vers le bas pour compléter le tableau. Utiliser le remplissage, vers la droite pour obtenir la somme de chaque colonne puis terminer les calculs de moyenne et de variance dans les cellules* ***G44*** *et* ***G46*** *(on écrira " moyenne = " dans la cellule* ***F44*** *et " variance= " dans la cellule* ***F46****).*

## 6.3 Utilisation des fonctions de calcul

*On peut calculer la moyenne, la variance, l'écart-type mais aussi la médiane en utilisant directement les fonctions d'EXCEL.*

***Exemple** : Retrouvons les résultats de l'exercice 1.*

- *Sous la série statistique correspondant à la variable **F**, taper =.*
- *Parmi les fonctions proposées (à gauche de la barre de formules) choisir **MOYENNE** (puis **VAR.P**, puis **ECARTYPEP**, puis **MEDIANE**, on rentrera les résultats dans les 4 cases sous la série statistique (on peut faire de même pour chacune des variables quantitatives du tableau de données) .*
- *Sélectionner avec la souris la série statistique correspondant à la variable **F**. Cliquer sur **OK**.*

*Exercice 2 : Nous allons voir comment des calculs de moyennes et variances permettent de mettre en évidence le lien entre les variables "Membre de l'UE" et "PNB/habitant".*
*On partage l'ensemble $W$ des 28 pays en 3 groupes :*

- $\Omega_1$ *: les pays membres de l'UE.*
- $\Omega_2$ *: les pays en passe d'adhérer à l'UE.*
- $\Omega_3$ *: Turquie, Bulgarie et Roumanie.*

*On s'intéresse à la variable "PNB/habitant", notée $X$.*

*- Calculer la moyenne de $X$, notée $m$ et la variance de $X$, notée $v$, sur toute la population $O$ (utiliser les fonctions d'EXCEL). Cette moyenne est-elle le PNB/habitant de l'ensemble des 28 pays ?*

*- Sur chaque groupe $\Omega_i$, calculer la moyenne de $X$, notée $m_i$ et la variance de $X$, notée $v_i$ (utiliser le tri et les fonctions d'EXCEL). Compléter les phrases suivantes :*

$$\text{Sur } \Omega_1, \text{ la moyenne est } m_1 = \quad \text{et la variance est } v_1 =$$
$$\text{Sur } \Omega_2, \text{ la moyenne est } m_2 = \quad \text{et la variance est } v_2 =$$
$$\text{Sur } \Omega_2, \text{ la moyenne est } m_3 = \quad \text{et la variance est } v_3 =$$

*Une simple comparaison des moyennes permet de penser qu'un tel écart n'est pas fortuit mais que le PNB par habitant est bien un des critères qui distingue les trois groupes. On peut également constater que la variance de $X$ sur $\Omega$ est bien plus grande que la variance de $X$ sur $\Omega_1$ c'est-à-dire que l'hétérogénéité de l'U.E. pour le PNB/h serait bien plus grande avec ces 28 pays.*
*On peut aussi mesurer le lien entre $X$ et la variable nominale $Y$ dont les modalités sont "membre*

*de l'U.E.", "en passe d'adhérer à l'U.E.", "autre candidat" en calculant le rapport entre la variance intergroupe (variance des moyennes de chaque groupe $\Omega_i$ affectées des effectifs des groupes) et la variance totale. Ce rapport, toujours compris entre 0 et 1, indique un lien fort lorsqu'il est proche de 1 et faible lorsqu'il est proche de 0. On trouve une variance intergroupe égale à* $\left(var\ inter = \dfrac{15m_1^2 + 10m_2^2 + 3m_3^2}{28} - m^2\right)$ *soit un rapport de 0,84 ; on peut donc considérer que le lien entre $X$ et $Y$ est fort.*

## 6.4 Couple de variables

*Une autre façon de mettre en évidence un lien entre 2 variables est d'étudier le couple de variables à partir d'une table de contingence (tableau de distribution des effectifs conjoints). EXCEL permet de dresser des tables de contingence à partir d'un tableau de données mais lorsqu'on traite des variables quantitatives continues, il est nécessaire de commencer par faire un regroupement en classes.*

***Exemple** : Les variables "PNB/habitant" et "Population" sont quantitatives continues ; nous allons faire un regroupement en classes avec les intervalles suivants :*

| | | | | | |
|---|---|---|---|---|---|
| *PNB/habitant (en milliers de dollars) :* | *]0 ;10]* | *]10 ;20]* | *]20 ;30]* | *]30 ;40]* | *]40 ;50]* |
| *Population (en millions d'habitants) :* | *]0 ;1]* | *]1 ;10]* | *]10 ;30]* | *]30 ;50]* | *]50 ;100]* |

*Pour effectuer ce regroupement en classes :*

- *Rajouter 2 colonnes au tableau intitulées "PNB/hab (classes)" et "Population (classes)".*
- ***Trier** par "PNB/habitant" puis utiliser la fonction **remplissage** (menu **Edition**) pour compléter rapidement la nouvelle colonne "PNB/hab (classes)".*
- ***Trier** par "Population" puis utiliser la fonction **remplissage** (menu **Edition**) pour compléter rapidement la nouvelle colonne "Population (classes)".*

*Nous allons maintenant regarder des tableaux croisés (tables de contingence) pour les couples de variables (Membre de l'UE , PNB/hab (classes)) et (Membre de l'UE , Population (Classes)). On peut se douter que le lien est fort dans le premier cas et faible dans le deuxième, nous allons voir des graphiques qui mettent cela en évidence.*

***Exemple** :*

- *Sélectionner tout le tableau de données (plage **A1 :I29**).*
- *Dans le menu **Données**, choisir **Rapport de tableau croisé dynamique**.*

• *Cliquer sur* ***Suivant*** *(la nouvelle fenêtre précise alors que vous avez bien sélectionné la plage* ***A1 :I29****).*

• *Cliquer de nouveau sur* ***Suivant****, la dernière fenêtre, étape 3/3, propose de placer le tableau croisé sur une nouvelle feuille, cliquer sur* ***Terminer*** *pour confirmer ce choix.*

• *Apparaît alors un tableau constitué de 3 parties :* ***Colonne****,* ***Ligne*** *et* ***Données*** *: avec la souris, placer le mot "****pays****" dans la partie* ***Données****, la variable "****Membre de l'UE****" dans la partie****Colonne*** *(on obtient alors le tableau d'effectifs de la variable "Membre de l'UE"), la variable "* ***PNB/hab (classes)*** *" dans la partie* ***Ligne*** *(on obtient alors la table de contingence de ce couple de variable, avec les distributions marginales des effectifs).*

• *Cliquer sur l'icône* ***Assistant Graphique*** *du* ***Tableau croisé dynamique****.*

• *Cliquer sur le bouton droit de la souris sur le graphique obtenu pour sélectionner le* ***Type de graphique*** *souhaité et dans* ***Histogramme****, choisir* ***Histogramme empilé 100****%. On obtient pour chaque classe la proportion de pays appartenant à chacune des trois catégories. Ici, il n'y a que des pays candidats dans la première classe alors qu'il n'y a plus de pays candidats à partir de la troisième classe, le lien entre les 2 variables est évident.*

• *Revenir sur la feuille où se trouve le* ***Tableau croisé dynamique*** *et remplacer à l'aide de la souris la variable "PNB/hab (classes)" par "Population (classes)". Cela modifie automatiquement le graphique qui cette fois ne permet pas de parler de lien entre les deux variables (les 5 rectangles "se ressemblent").*

## 6.5 Représentations graphiques

### 6.5.1 Etude d'une variable qualitative

*Nous avons vu dans le cours qu'il est possible de représenter graphiquement*

- *une variable qualitative nominale par un* ***diagramme en colonnes*** *ou* ***en bâtons*** *ou un* ***diagramme en secteurs****.*
- *une variable qualitative ordinale par une* ***boîte à moustaches****.*

*Les variables quantitatives discrètes pouvant être considérées comme des variables qualitatives nomibales ou ordinales, nous illustrerons cette section en étudiant la variable F.*

#### A) Le diagramme en colonnes ou en bâtons

***Exemple*** *: Nous allons représenter la variable "F".*

• *Sélectionner le tableau d'effectifs de la variable F.*

• *Cliquer sur l'icône* ***Assistant Graphique*** → ***Histogramme****. En appuyant sur l'onglet* ***Maintenir appuyé pour visualiser****, on peut voir le graphique.*

• *Cliquer sur* ***Suivant*** *(la nouvelle fenêtre précise alors la plage de données sélectionnée). Cliquer sur l'onglet* ***Série****.*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***Variable F****.*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de F en sélectionnant la bonne plage de données dans le fichier Excel.*

• *Cliquer sur* ***Suivant****. Vous pouvez alors modifier les options de légende, quadrillage, titre... On peut entre autres faire figurer les valeurs.*

• *Cliquer sur* ***Suivant****. Vous pouvez alors choisir de faire apparaître le graphique sur la même feuille de calcul ou dans une autre fenêtre. Cliquer sur* ***Terminer****.*

### B) Le diagramme en secteurs

***Exemple*** *: Nous allons représenter la variable "F" cette fois par un diagramme en secteurs.*

• *Sélectionner le tableau d'effectifs de la variable F.*

• *Cliquer sur l'icône* ***Assistant Graphique*** → ***Secteurs*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***Variable F****.*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de F en sélectionnant la bonne plage de données dans le fichier Excel. Cliquer sur* ***Terminer****.*

### C) La boîte à moustaches (Box-Plot en anglais)

| | *F* |
|---|---|
| *Minimum* | *1* |
| $D_1$ | *2* |
| $Q_1$ | *2* |
| *Médiane* | *3* |
| $Q_3$ | *5* |
| $D_9$ | *6* |
| *Maximum* | *9* |

• *Sélectionner les deux colonnes du tableau ci-dessus sans les deux premières et dernière lignes. Cliquer sur l'icône* ***Assistant Graphique****.*

1. *sélectionner un graphique en courbes avec marques puis cliquer sur* ***Suivant****.*

2. *sélectionner l'option* ***Lignes*** *→ cliquer sur l'onglet* ***Série*** *→ cliquer dans la partie* ***Étiquettes des abscisses (X)*** *puis sélectionner la cellule contenant F → cliquer sur* ***Suivant****.*

3. *sélectionner* ***Légende****; ne pas cocher* ***Afficher la légende****, cliquer sur* ***Terminer*** *pour créer le graphique.*

• *Activer la deuxième série de données "1er Quartile", ouvrir la boîte de dialogue* ***Format de série de données*** *:*

1. *Cliquer sur l'onglet* ***Motifs****, sélectionner Aucun pour l'option* ***Trait****; Aucune pour l'option* ***Marque****.*

2. *Cliquer sur l'onglet* ***Options****, sélectionner Lignes haut/bas et Barres hausse/baisse.*

3. *Cliquer sur l'onglet* ***Ordre des séries*** *: la série "1 er Quartile" est sélectionnée. Cliquer sur Déplacer vers le haut, sélectionner la série "3e Quartile", cliquer sur déplacer vers le bas. L'ordre final des séries doit être : 1er Quartile, 1er Décile, Médiane, 9e Décile, 3e Quartile.*

• *Activer la série de données "1er Décile", ouvrir la boîte de dialogue* ***Format de série de données****, cliquer sur l'onglet* ***Motifs****, sélectionner Aucun pour l'option* ***Trait****; Barre horizontale pour l'option* ***Marque****, taille 10 pts . On peut modifier les couleurs à cet endroit.*

• *recommencer la même manipulation avec les séries "Médiane" et "9e Décile", ouvrir la boîte de dialogue Format de série de données, cliquer sur l'onglet Motifs. Sélectionnez Aucun pour l'option Trait; Barre horizontale pour l'option Marque, taille 10 pts.*

• *Activer la série de données "3e Quartile", ouvrir la boîte de dialogue* ***Format de série de données****, cliquer sur l'onglet* ***Motifs****. Sélectionnez Aucun pour l'option* ***Trait****; Aucune pour l'option* ***Marque****.*

• *Activer le graphique Dans la barre de menu Graphique → cliquer sur* ***Ajouter des données****. Quand la boîte de dialogue* ***Ajouter des données*** *apparaît sélectionner la cellule contenant le minimum → cliquer sur OK.*

1. *Activer la série de données "Minimum". Ouvrir la boîte de dialogue* ***Type de graphique*** *et choisir* ***Nuages de points*** *→ cliquer sur OK.*

2. *Activer la série de données "Minimum", ouvrir la boîte de dialogue* ***Format de série de données*** *→ cliquer sur l'onglet* ***Motifs****. Sélectionnez Aucun pour l'option* ***Trait*** *et un cercle pour l'option* ***Marque****.*

• *Recommencer avec la série de données "Maximum".*

### 6.5.2 Etude d'une variable quantitative

*Nous avons vu dans le cours qu'il est possible de représenter graphiquement*

- *une variable quantitative discrète par un* ***diagramme en bâtons****.*
- *une variable quantitative continue par un* ***histogramme****. Dans le cas où les classes ont même amplitude, on peut de plus tracer le polygone des effectifs. Dans le cas contraire, on ne représente pas les effectifs mais les densités d'effectifs.*

**A) L'histogramme avec des classes de même amplitude**

***Exemple*** *: Nous allons représenter la variable "PNB/hab" avec le regroupement en classes suivant* $]0;10]$*,* $]10;20]$*,* $]20;30]$*,* $]30;40]$ *et* $]40;50]$*. On dispose du tableau suivant*

| *Classes* | *Effectifs* |
|---|---|
| $]0;10]$ | *12* |
| $]10;20]$ | *6* |
| $]20;30]$ | *8* |
| $]30;40]$ | *1* |
| $]40;50]$ | *1* |

• *Sélectionner le tableau d'effectifs de la variable PNB/hab.*

• *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Histogramme*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***PNB/hab****.*

*Dans le cadre* ***Valeurs****, sélectionner les différents effectifs de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (deuxième colonne du tableau ci-dessus).*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (première colonne du tableau). Cliquer sur* ***Terminer****.*

• *Cliquer sur une des colonnes de l'histogramme. Cliquer sur l'onglet* ***Options****, régler la largeur de l'intervalle sur 0. Ainsi les différentes barres de l'histogramme sont collées.*

**B) Le polygone des effectifs** *On va maintenant ajouter sur le même graphique le polygone des effectifs. On modifie le tableau des données en ajoutant des lignes fictives :*

| *Classes* | *Effectifs* |
|---|---|
| | *0* |
| ]0; 10] | *12* |
| ]10; 20] | *6* |
| ]20; 30] | *8* |
| ]30; 40] | *1* |
| ]40; 50] | *1* |
| | *0* |

• *Sélectionner le tableau d'effectifs de la variable PNB/hab (deuxième colonne du tableau).*

• *Cliquer sur l'icône* ***Assistant Graphique*** *puis sur l'onglet* ***Types personnalisés*** *et choisir* ***Courbes-Histogramme*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***Histogramme des effectifs****.*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (première colonne du tableau ci-dessus).*

• *Cliquer sur* ***Ajouter****.*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***Polygone des effectifs****.*

*Dans les cadres* ***Valeurs*** *et* ***Etiquettes des abscisses****, sélectionner les mêmes plages de données que pour l'histogramme (première et deuxième colonnes du tableau ci-dessus)* → ***Terminer****.*

• *Cliquer sur une des colonnes de l'histogramme. Cliquer sur l'onglet* ***Options****, régler la largeur de l'intervalle sur 0. Ainsi les différentes barres de l'histogramme sont collées.*

### C) L'histogramme avec des classes d'amplitudes différentes

***Exemple*** *: Nous allons représenter la variable "PNB/hab" avec un regroupement en classes différents* ]0; 5], ]5; 10], ]10; 20], ]20; 30] *et* ]30; 50]. *On dispose donc du tableau suivant*

| *Classes* | *Effectifs* |
|---|---|
| ]0; 5] | *9* |
| ]5; 10] | *3* |
| ]10; 20] | *6* |
| ]20; 30] | *8* |
| ]30; 50] | *2* |

*On transforme ce tableau en un tableau un peu artificiel avec les densités d'effectifs*

| *Début des classes* | *Densités d'effectifs* | *Début des classes* | *Densités d'effectifs* |
|---|---|---|---|
| *0* | *0* | *20* | *0,6* |
| *0* | *1,8* | *20* | *0* |
| *5* | *1,8* | *20* | *0,8* |
| *5* | *0* | *30* | *0,8* |
| *5* | *0,6* | *30* | *0* |
| *10* | *0,6* | *30* | *0,1* |
| *10* | *0* | *50* | *0,1* |
| *10* | *0,6* | *50* | *0* |

• *Sélectionner les deux colonnes du tableau ci-dessus.*

• *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Nuage de points reliés par une courbe sans marquage des données*** *puis sur* ***Suivant****.*

• *Cliquer sur l'onglet* ***Série****.*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***PNB/hab****.*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab en sélectionnant la bonne plage de données dans le fichier Excel (première colonne du tableau ci-dessus)* → ***Terminer****.*

• *Cliquer sur une des colonnes de l'histogramme. Cliquer sur l'onglet* ***Options****, régler la largeur de l'intervalle sur 0. Ainsi les différentes barres de l'histogramme sont collées.*

### 6.5.3 Etude de deux variables

*Dans le cas de deux variables, il est possible de représenter la distribution conjointe et les distributions conditionnelles.*

#### A) Distribution conjointe

***Exemple*** *: Nous allons représenter les variables "Membre" et "PNB/hab" avec le regroupement en classes suivant* $]0;10]$, $]10;20]$, $]20;30]$, $]30;40]$ *et* $]40;50]$. *On dispose du tableau suivant*

| | | *PNB par habitant en milliers de dollars* | | | | |
|---|---|---|---|---|---|---|
| | | *]0 ;10]* | *]10 ;20]* | *]20 ;30]* | *]30 ;40]* | *]40 ;50]* |
| *Membre de l'U.E.* | *OUI* | *0* | *5* | *8* | *1* | *1* |
| | *NON* | *12* | *1* | *0* | *0* | *0* |

• *Sélectionner le tableau d'effectifs.*

• *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Histogramme*** *puis sur* ***Suivant****.*

- *Dans l'onglet* ***Plage de données****, choisir* ***Série en lignes*** *plutôt que* ***colonnes****.*
- *Cliquer sur l'onglet* ***Série****.*

*Pour la série 1 :*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***Oui****.*

*Dans le cadre* ***Valeurs****, sélectionner les différents effectifs de PNB/hab pour les membres (première ligne de données du tableau ci-dessus).*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab.*

*Pour la série 2 :*

*Dans le cadre* ***Nom****, vous pouvez rentrer* ***Non****.*

*Dans le cadre* ***Valeurs****, sélectionner les différents effectifs de PNB/hab pour les non membres (deuxième ligne de données du tableau ci-dessus).*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de PNB/hab.*

- *Cliquer sur* ***Terminer****.*

**Remarque 6.3.** *On peut aussi inverser les séries.*

### B) Distributions conditionnelles

***Exemple*** *: Nous allons représenter la distribution de la variable "PNB/hab" conditionnellement à "Membre". On travaille sur le même tableau.*

- *Sélectionner le tableau d'effectifs.*
- *Cliquer sur l'icône* ***Assistant Graphique*** *et choisir* ***Histogramme*** *puis* ***Histogramme empilé 100%*** *puis sur* ***Suivant****.*
- *Dans l'onglet* ***Plage de données****, choisir* ***Série en lignes*** *plutôt que* ***colonnes****.*
- *Cliquer sur l'onglet* ***Série****. Pour chacune des séries, rentrer dans nom la classe correspondante du PNB/hab.*

*Dans le cadre* ***Etiquettes des abscisses****, vous pouvez rentrer les différentes modalités de Membre : oui et non.*

- *Cliquer sur* ***Terminer****.*

# Chapitre 7

# Fiches récapitulatives

*Voici quatre fiches récapitulatives synthétisant les différentes notions vues dans ce cours. Elles vous aideront dans vos révisions et vous pourrez vous en inspirer pour réaliser la feuille recto-verso manuscrite A4 autorisée à l'examen.*

***Fiche résumé 1 : Généralités sur la statistique descriptive***

*Nous étudions un ensemble d'**individus** $\omega$ qui constitue une **population** $\Omega$ et dont le nombre total $N$ est appelé la **taille de la population**. Une partie de $\Omega$ est appelée un **échantillon**. L'étude de ces individus se fait par divers renseignements appelés **variables** que l'on regroupe dans un tableau de données.*

*Types de variables*

***Qualitatif nominal** : les valeurs étudiées (**modalités**) sont des catégories non hiérarchisées, désignées par leur nom ;*

***Qualitatif ordinal** : il existe un ordre naturel entre les modalités, même si ce ne sont pas des nb. ;*

***Quantitatif discret** : les valeurs étudiées sont des nombres qui résultent d'un comptage.*

***Quantitatif continu** : les valeurs étudiées sont des nombres qui résultent de la mesure d'une grandeur physique.*

*Quelques quantités*

***Effectif d'une modalité** $n_i$ : nombre d'individus ayant la modalité.*

***Effectif cumulé d'une modalité** $N_i$ pour les variables qualitatives ordinales et quantitatives : nombre d'individus ayant la modalité ou une modalité inférieure.*

***Densité d'effectif d'une modalité** $\frac{n_i}{a_i}$ pour une variable quantitative continue avec un regroupement en classes : nombre d'individus ayant la modalité divisé par l'amplitude $a_i$ de la classe de l'individu.*

***Fréquence d'une modalité** $f_i$ : effectif de la modalité divisé par la taille.*

*Les différents types de graphiques*

***Diagramme en bâtons et en barres** : pour les variables qualitatitives et les variables quantitatives discrètes.*

***Diagramme en secteurs** : pour toutes les variables.*

***Histogramme** : pour les variables quantitatives continues.*

*Si les classes ont même amplitude, on peut aussi tracer le **polygone des effectifs**. Si les classes n'ont pas la même amplitude, on ne représente pas les effectifs mais les densités d'effectifs.*

***Boîte à moustache** : pour les variables quantitatives continues. On fait figurer sur ce graphique, la moyenne, la médiane, les premiers et troisième quartiles et enfin l'étendue.*

***Fiche résumé 2 : Les indices***

*Les indices de tendance centrale*

***Le mode*** *: c'est la modalité ayant le plus grand effectif (et donc également la plus grande fréquence).*

***La médiane*** *: Pour déterminer la médiane, on calcule d'abord les effectifs cumulés. L'effectif cumulé d'une modalité est la somme des effectifs des modalités qui lui sont inférieures ou égales.*

• *Si $X$ est une variable ordinale ou quantitative discrète, étudiée sur une population de taille $N$, la médiane est la modalité dont l'effectif cumulé est immédiatement supérieur à $\frac{N}{2}$.*

• *Si $X$ est une variable quantitative continue regroupée en classes, on repère d'abord la classe dans laquelle elle se trouve : c'est celle dont l'effectif cumulé est immédiatement supérieur à $\frac{N}{2}$, notons la $[x_1; x_2]$, notons $N_2$ l'effectif cumulé de cette classe appelée* ***classe médiane*** *et $N_1$ l'effectif cumulé de la classe qui précède, la médiane est :*

$$\boxed{m = x_1 + (x_2 - x_1)\frac{\frac{N}{2} - N_1}{N_2 - N_1}}$$

*On définit de même le* ***1er*** *et le* ***3ème quartiles*** *en comparant les effectifs cumulés à $\frac{N}{4}$ et $\frac{3 * N}{4}$.*

*La* ***moyenne*** *: La moyenne d'une variable quantitative $X$, notée $\overline{X}$, est la somme des valeurs prises par $X$ divisée par la taille de la population $N$. Plus précisément, si chaque valeur $x_i$ de $X$ a pour effectif $n_i$, la moyenne est*

$$\boxed{\overline{X} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{N}}$$

*où $k$ est le nombre de valeurs prises par $X$.*

*Les indices de dispersion*

*L'****étendue*** *: Différence entre la valeur maximum observée et la valeur minimum observée.*

*L'****écart interquartile*** *: Différence entre le troisième quartile et le premier quartile.*

*La* ***variance*** *: Soit $X$ une variable quantitative définie sur une population $\Omega$ de taille $N$ dont les valeurs sont $x_1$, $x_2$, …, $x_k$ d'effectifs respectifs $n_1$, $n_2$, …, $n_k$ :*

$$\boxed{\mathrm{Var}(X) = \frac{\sum_{i=1}^{k} n_i (x_i)^2}{N} - \overline{X}^2 = \frac{n_1 (x_1)^2 + n_2 (x_2)^2 + \ldots + n_k (x_k)^2}{N} - \overline{X}^2}$$

*L'****écart-type*** *: L'écart-type d'une variable quantitative $X$, noté $\sigma(X)$, est la racine carrée de la variance.*

***Fiche résumé 3 : Les couples de variables***

***Distribution conjointe*** *: on note*

*- $n_{ij}$ l'effectif associé à la modalité du couple de variables $(X, Y)$ ;*

*- $p$ est le nb de lignes de la table de contingence de $X$ et $Y$ (ou nb de modalités de $X$) ;*

*- $q$ est le nb de colonnes de la table de contingence de $X$ et $Y$ (ou nb de modalités de $Y$).*

*La distribution conjointe est donnée par les $n_{ij}$ ou de façon équivalente par les fréquences conjointes $f_{ij} = \frac{n_{ij}}{N}$.*

***Effectifs marginaux*** *: On complète la* ***table de contingence*** *contenant les $n_{ij}$ en indiquant la somme de chaque ligne et de chaque colonne. La somme de la $i$ ième ligne est notée $L_i$ (c'est le nombre d'individus ayant la modalité $x_i$) ; la somme de la $j$ ème colonne est notée $C_j$ (c'est le nombre d'individus ayant la modalité $y_j$).*

*La colonne intitulée* ***Marge de*** *$X$ est alors la table d'effectifs de $X$, d'où son nom.*

*La ligne intitulée* ***Marge de*** *$Y$ est alors la table d'effectifs de $Y$, d'où son nom.*

*Ces deux marges sont parfois appelées* ***distributions marginales des effectifs****.*

***Distributions conditionnelles*** *:*

*La fréquence de $x_i$ conditionnellement à $y_j$ sera la proportion d'individus ayant la modalité $x_i$ parmi ceux ayant la modalité $y_j$.*

***Attention, ajouter les valeurs d'une même ligne n'aurait aucun sens : en effet, chaque pourcentage d'une même ligne est un pourcentage par rapport à une partie différente de la population.***

*De la même façon, on peut déterminer les fréquences de $Y$ conditionnellement à $X$.*

*Représentations graphiques :*

- *Distribution conjointe : On trace $q$ rectangles par modalité $x_i$ de $X$ représentant chacun l'effectif du couple $(x_i, y_j)$. On peut aussi tracer $p$ rectangles par modalité $y_j$ de $Y$ représentant chacun l'effectif du couple $(x_i, y_j)$.*

*Chaque couple de modalités est donc bien représenté par un seul rectangle.*

- *Distribution de $X$ conditionnellement à $Y$ : on trace $q$ rectangles de hauteur 100 (autant que de modalités de $Y$) et on divise chaque rectangle selon la répartition de $X$ pour la modalité concernée.*

***Fiche résumé 4 : Indice de liaison entre deux variables quelconques***

*On note désormais l'effectif $O_{ij}$ associé à la modalité $(x_i, y_j)$ du couple de variables $(X, Y)$ plutôt que $n_{ij}$ afin de rappeler que ce sont des effectifs observés et marquer la différence avec les effectifs $T_{ij}$ théoriques définis ci-dessous.*

*Afin de déterminer l'existence d'une liaison entre les variables $X$ et $Y$, on commence par calculer les* ***effectifs théoriques*** *; ce sont les effectifs que l'on doit obtenir à partir des marges observées dans le cas où les variables $X$ et $Y$ seraient indépendantes. Ils sont donnés par*

$$\boxed{T_{ij} = \frac{L_i \times C_j}{N}.}$$

*Ensuite on calcule le coefficient $\chi^2$ donné par*

$$\boxed{\chi^2 = \sum_{i,j} \frac{(O_{ij} - T_{ij})^2}{T_{ij}}.}$$

*Naturellement,*

- *si $\chi^2$ est proche de 0, les variables sont indépendantes.*
- *si $\chi^2$ est significativement différent de 0, les variables sont liées.*

*Propriété* : $\underset{\text{indépendance}}{0} \leq \chi^2 \leq \underset{\text{dépendance totale}}{N\,[\min(p,q) - 1]}.$

- *$p$ est le nb de lignes de la table de contingence de $X$ et $Y$ (ou nb de modalités de $X$) ;*
- *$q$ est le nb de colonnes de la table de contingence de $X$ et $Y$ (ou nb de modalités de $Y$) ;*
- $\min(p,q)$ *est simplement le plus petit des deux nombres $p$ et $q$ ;*

***Attention rien ne permet à ce stade de dire si le lien est important ou non.***

*Dans le cas de variables dépendantes, on calcule le coefficient $\varphi$ pour déterminer l'importance du lien*

$$\boxed{\varphi = \sqrt{\frac{\chi^2}{N\,[\min(p,q) - 1]}}.}$$

*Propriété : L'avantage de ce nouveau coefficient est qu'il est toujours compris entre 0 et 1.*

$$\underset{\text{indépendance}}{0} \leq \varphi \leq \underset{\text{dépendance totale}}{1}$$

*Convention : On considère en général que le lien est*

- *faible lorsque $\varphi$ est inférieur à 0,3 ;*
- *moyen lorsque $\varphi$ est compris entre 0,3 et 0,5 ;*
- *fort lorsque $\varphi$ est supérieur à 0,5.*

# Chapitre 8

# Devoir à rendre

***Le tribut de la ligue de Délos*** *(Source : Patrice Brun, Impérialisme et démocratie à Athènes, Ed. Armand Colin, 2005, p. 35-36.)*

*Le tableau ci-dessous donne le montant des sommes exigées chaque année des communautés membres de la ligue de Délos (sorte d'OTAN en Grèce au Vème siècle* ***AVANT*** *J.C.).*
*La première colonne concerne le montant annuel (stable) de la décénie 450-440. La deuxième colonne concerne le montant de l'année 425, qui a subi une révision exceptionnelle. L'unité de monnaie est le millier de drachme (dr.).*

| *Lieux* | *450-440* | *425* | *Lieux* | *450-440* | *425* |
|---|---|---|---|---|---|
| *ANAPHE* | *1* | *1* | *MYRINA* | *9* | *24* |
| *ANDROS* | *12* | *90* | *NAXOS* | *16* | *90* |
| *ATHENA DIADES* | *4* | *6* | *PAROS* | *39* | *78* |
| *CARYSTOS* | *18* | *30* | *RHENE* | *0,3* | *1* |
| *ERETRIE* | *18* | *90* | *SERIPHOS* | *6* | *12* |
| *GRYNCHE* | *1* | *2* | *SIPHNOS* | *18* | *54* |
| *IMBROS* | *6* | *6* | *STYRA* | *6* | *12* |
| *IOS* | *3* | *6* | *SYME* | *1,8* | *3* |
| *KEOS* | *24* | *60* | *SYROS* | *1* | *6* |
| *KYTHNOS* | *18* | *36* | *TENOS* | *12* | *60* |
| *LEMNOS* | *24,3* | *24* | *THERA* | *18* | *30* |
| *MYCONOS* | *6* | *12* | | | |

***Partie A : Questions générales***

1. *Quelle est la population considérée ? Quelle est sa taille ?*
2. *Quelles sont les variables $X$ et $Y$ étudiées ? Précisez leur type ?*

***Partie B : Etude du montant annuel (stable) de la décénie 450-440 et différents regroupements par classes***

***Dans cette partie, nous nous intéressons au montant annuel (stable) de la décénie 450-440 selon ou non différents regroupements par classes.***

1. ***Dans cette question, nous regroupons les données selon les classes suivantes :*** $[0;5[$, $[5;10[$, $[10;20[$ ***et*** $[20;40[$.
   *Représenter graphiquement cette variable. Vous choisirez le type de graphique le mieux adapté.*
2. ***Dans cette question, nous ne faisons pas de regroupement par classes.***
   (a) *Déterminer la médiane. Que représente ce nombre ?*
   (b) *Déterminer les premier et troisième quartiles.*
3. ***Dans cette question, nous regroupons les données selon les classes suivantes :*** $[0;10[$, $[10;20[$, $[20;30[$ ***et*** $[30;40[$.
   (a) *Calculer la médiane. Que représente ce nombre ?*
   (b) *Calculer les premier et troisième quartiles.*
   (c) *Conclure sur les résultats de cette question et celle de la question précédente.*

***Partie C : Corrélation entre les variables***

***Dans cette partie, nous nous intéressons à l'existence éventuelle d'une corrélation entre le montant annuel (stable) de la décenie 450-440 et celui de l'année 425.***
***Nous ne considérons pas de regroupements par classes et afin de ne pas faire de trop lourds calculs nous prendrons seulement quelques données et nous travaillerons sur le tableau de données suivant :***

| *Lieux* | *450-440* | *425* | *Lieux* | *450-440* | *425* |
|---|---|---|---|---|---|
| *ANAPHE* | *1* | *1* | *IOS* | *3* | *6* |
| *ATHENA DIADES* | *4* | *6* | *KYTHNOS* | *18* | *36* |
| *CARYSTOS* | *18* | *30* | *MYCONOS* | *6* | *12* |
| *GRYNCHE* | *1* | *2* | *MYRINA* | *9* | *24* |
| *IMBROS* | *6* | *6* | *SYROS* | *1* | *6* |

***De la même façon que dans la partie B, nous faisons un regroupement par classes :*** $[0;5[$, $[5;10[$, $[10;20[$ ***et*** $[20;40[$***. Nous obtenons donc le tableau des effectifs observés*** $O_{ij}$ ***suivant***

| $Y$ \ $X$ | $[0;5[$ | $[5;10[$ | $[10;20[$ | $[20;40[$ |
|---|---|---|---|---|
| $[0;5[$ | *2* | *3* | *0* | *0* |
| $[5;10[$ | *0* | *1* | *1* | *1* |
| $[10;20[$ | *0* | *0* | *0* | *2* |

*Déterminer s'il y a un lien entre les variables X et Y. Si oui, en donner l'importance.*

# Chapitre 9

# Enoncé des exercices

## 9.1 Exercices du chapitre 1

**Exercice 1.** *Étant donnée une variable étudiée sur une population, à chaque individu est associé :*

- □ *un nombre de modalités qui dépend de la variable*
- □ *au moins une modalité de la variable*
- □ *une et une seule modalité de la variable.*

---

**Exercice 2.** *Sur la population des membres d'une famille de cinq personnes (le père, la mère, et trois enfants), la relation qui associe à chaque personne son père est-elle une variable ? Pourquoi ?*

---

**Exercice 3.** *"Pour les prochaines élections législatives, il y a en Haute-Garonne 8 sièges à pourvoir et un total de 78 candidats" (LA DÉPÊCHE 1 Mars 93).*

*1. Si les individus sont les électeurs, quelle est la variable, quelles sont ses modalités, quelle est la taille de la série statistique ? (La série statistique est l'ensemble des couples (individus ; donnée de l'individu pour la variable étudiée).)*

*2. Même question si les individus sont les candidats.*

---

**Exercice 4.** *Vous avez dû remarquer que dans les magasins de bricolage "Cassetout", à la caisse, on demande aux clients leur code postal.*
*Pour cette situation, indiquer ce que sont les individus, quelle est la variable en précisant son type.*

---

**Exercice 5.** *Le mode d'une variable est :*

- □ *la modalité ayant le plus petit effectif*
- □ *la modalité ayant le plus grand effectif*

□ *le plus grand des effectifs.*

---

**Exercice 6.** *Dans les exemples suivants, donner la population, un individu de la population, la variable étudiée et le type de cette variable.*

*1. On souhaite connaître les catégories socioprofessionnelles de la population active en France en 1930.*

*2. On veut étudier la répartition des entreprises françaises selon leur taille (petite, moyenne ou grande) en 1973.*

*3. On mesure le P.N.B. dans chaque pays en 1991.*

*4. On s'intéresse aux billets en circulation dans le monde et à leur valeur à la Bourse (convertie en Euros) au 1/1/2000.*

*5. On s'intéresse à l'âge des étudiants inscrits en licence d'histoire.*

*6. On compte le nombre de personnes par logement dans la ville de Toulouse en 1999.*

---

**Exercice 7.**

1. *Voici un tableau d'effectifs donnant la production mondiale d'or en 1937.*

| *Continents* | *Europe* | *Asie* | *Afrique* | *Amérique* | *Océanie* |
|---|---|---|---|---|---|
| *Production en tonnes* | *176* | *87* | *431* | *350* | *56* |

*Préciser la population, sa taille, un individu de la population, la variable et son type.*

*Faire le tableau de fréquence puis un diagramme circulaire.*

2. *Voici un tableau de données donnant la production mondiale d'or en 1937.*

| *Continents* | *Europe* | *Asie* | *Afrique* | *Amérique* | *Océanie* |
|---|---|---|---|---|---|
| *Production en tonnes* | *176* | *87* | *431* | *350* | *56* |

*Préciser la population, sa taille, un individu de la population, la variable et son type.*

---

**Exercice 8.** *Le tableau suivant indique le nombre de chômeurs (exprimé en milliers) au sens du BIT (Bureau International du Travail), selon le sexe et l'âge, en France (Mars 1989) :*

| | *Hommes* | *Femmes* |
|---|---|---|
| *Moins de 25 ans* | *249* | *342,6* |
| *De 25 à moins de 50 ans* | *565,1* | *827,9* |
| *50 ans et plus* | *168,5* | *155,2* |

1. *Définir la population étudiée. Quelles sont les variables et leur type ?*
2. *Calculer les fréquences marginales.*
3. *a) Quel est le pourcentage de femmes parmi les chômeurs ?*

   *b) Quel est le pourcentage de jeunes de moins de 25 ans parmi les chômeurs ?*

   *c) Quel est le pourcentage de femmes parmi les chômeurs de moins de 25 ans ?*

4. *Représenter graphiquement ces données.*

---

**Exercice 9.** *Population des pays de l'Union Européenne en 1995 (en millions d'habitants).*

| | | | | | |
|---|---|---|---|---|---|
| *Allemagne* | *81,59* | *Luxembourg* | *0,41* | *Portugal* | *9,82* |
| *France* | *57,98* | *Royaume-Uni* | *56,26* | *Grèce* | *10,45* |
| *Italie* | *57,19* | *Danemark* | *5,12* | *Suède* | *8,78* |
| *Pays-Bas* | *15,5* | *Irlande* | *3,55* | *Autriche* | *7,97* |
| *Belgique* | *10,11* | *Espagne* | *39,62* | *Finlande* | *5,11* |

1. *Déterminer la variable étudiée, ses modalités et son type.*
2. *Quelle est la population et sa taille ?*
3. *Représenter ce tableau à l'aide d'un diagramme en bandes.*

---

**Exercice 10.** *Des élèves ont participé à une dictée. Voici le nombre de fautes relevées pour chacun :*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *8* | *3* | *5* | *1* | *4* | *2* | *8* | *7* | *9* | *4* |
| *3* | *1* | *0* | *2* | *8* | *2* | *1* | *8* | *4* | *0* |
| *2* | *8* | *5* | *1* | *0* | *5* | *7* | *4* | *3* | *9* |
| *6* | *8* | *4* | *9* | *2* | *8* | *8* | *7* | *0* | *3* |
| *6* | *8* | *4* | *9* | *2* | *8* | *7* | *7* | *0* | *3* |

1. *Expliquer la méthode de dépouillement de ces valeurs.*
2. *Déterminer la variable et son type. Quelle est la taille de l'échantillon ?*
3. *Déterminer le mode.*
4. *Déterminier l'étendue.*

---

**Exercice 11.** *Le tableau suivant donne la répartition des veufs français de 20 à 30 ans qui se sont à nouveau mariés en 1967 (`source I.N.S.E.E.`).*

| *Âge* | *20* | *21* | *22* | *23* | *24* | *25* | *26* | *27* | *28* | *29* | *30* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Effectifs* | *4* | *7* | *11* | *13* | *23* | *32* | *37* | *45* | *70* | *56* | *82* |

1. *Représenter ces données.*
2. *Quel est le mode de la variable étudiée ?*
3. *Calculer les effectifs cumulés puis faire le diagramme cumulatif.*

---

**Exercice 12.** *Une étude sur l'âge des femmes mariées d'un quartier d'une grande ville a fourni les informations suivantes :*

| *Âge* | *[18 ;28[* | *[28 ;38[* | *[38 ;48[* | *[48 ;58[* | *[58 ;68[* | *[68 ;78[* | *[78 ;88[* | *[88 ;98[* |
|---|---|---|---|---|---|---|---|---|
| *Effectifs* | *4* | *15* | *16* | *9* | *8* | *4* | *3* | *1* |

1. *Décrire cette étude statistique : population, échantillon (taille), variable (type).*
2. *Calculer les fréquences, puis les fréquences cumulées.*
3. *Construire un histogramme de la variable âge.*
4. *Construire le polygone cumulatif.*

---

**Exercice 13.** *À partir du graphique ci-dessous décrire l'étude statistique : taille de l'échantillon, variable statistique et reconstruire le tableau statistique (modalités, effectifs et effectifs cumulés).*

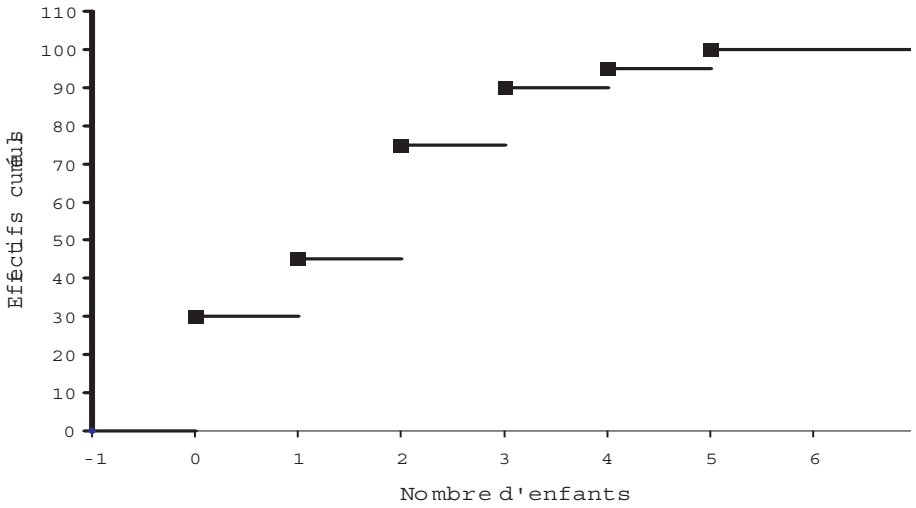



FIGURE 9.1 – *Diagramme cumulatif.*

---

## 9.2 Exercices des chapitres 2 et 3

**Exercice 14.** *Pour un groupe de 25 garçons, le n° d'INSEE médian est 1780574120547. Quel est ici le sens de ce nombre ?*

---

**Exercice 15.** *Concernant une variable X sur un échantillon, on a relevé les pourcentages cumulés suivants :*

| *Modalité de X* | *Pourcentage cumulé* |
|---|---|
| $x_1$ | 12% |
| $x_2$ | 32% |
| $x_3$ | 32% |
| $x_4$ | 54% |
| $x_5$ | 80% |
| $x_6$ | 90% |
| $x_7$ | 96% |
| $x_8$ | 100% |

1. *La variable X est elle nominale ?*
2. *Déterminer l'effectif de chaque modalité si la taille de l'échantillon est de 250.*
3. *Déterminer le mode.*
4. *Déterminer la médiane.*

---

**Exercice 16.** *Le QUID de 1991 donne le tableau suivant qui concerne les distinctions dans l'ordre de la légion d'honneur.*

| | *Effectifs au 1-1-90* |
|---|---|
| *Chevaliers* | *176751* |
| *Officiers* | *43152* |
| *Commandeurs* | *5220* |
| *Grands Officiers* | *459* |
| *Grand Croix* | *64* |

1. *Indiquer avec précision la population et la variable étudiée ainsi que son type.*
2. *Indiquer la médiane et les quartiles des distinctions.*

---

**Exercice 17.** *Mr Colec possède 50 albums BD rares de la série "épuisée". Le tableau suivant donne l'année de parution de ces albums :*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1970 | 1965 | 1975 | 1962 | 1981 | 1973 | 1963 | 1974 | 1985 | 1979 |
| 1980 | 1977 | 1986 | 1971 | 1965 | 1982 | 1972 | 1984 | 1962 | 1986 |
| 1974 | 1982 | 1960 | 1979 | 1961 | 1968 | 1977 | 1965 | 1967 | 1976 |
| 1965 | 1971 | 1987 | 1977 | 1978 | 1964 | 1983 | 1975 | 1986 | 1973 |
| 1983 | 1976 | 1970 | 1961 | 1969 | 1979 | 1962 | 1976 | 1968 | 1978 |

1. *Quel est le type de variable ?*
2. *Dépouiller ce tableau pour obtenir les effectifs et les effectifs cumulés.*
3. *Déterminer le mode.*
4. *Déterminer la médiane, les quartiles et les déciles.*

---

**Exercice 18.** *On reprend l'exercice du chapitre précédent sur les veufs français de 20 à 30 ans qui se sont à nouveau mariés en 1967 (*`source I.N.S.E.E.`*).*

1. *Déterminer la médiane et les quartiles. Représenter ces résultats par une boîte à moustaches.*
2. *Calculer la moyenne, la variance et l'écart-type.*

---

**Exercice 19.** *On reprend l'exercice du chapitre précédent sur l'âge des femmes mariées d'un quartier d'une grande ville.*

1. *Déterminer la médiane et les quartiles.*
2. *Calculer la moyenne, la variance et l'écart-type.*

---

**Exercice 20.** *Dans un quartier, la répartition par âge et par sexe des personnes de moins de 50 ans est la suivante :*

| *Âge* | *[0 ;2[* | *[2 ;6[* | *[6 ;10[* | *[10 ;20[* | *[20 ;30[* | *[30 ;40[* | *[40 ;50[* |
|---|---|---|---|---|---|---|---|
| *Hommes* | *11* | *40* | *49* | *80* | *100* | *70* | *60* |
| *Femmes* | *10* | *39* | *48* | *75* | *95* | *70* | *65* |

1. *Quelle est la population ? Quelle est la taille de l'échantillon ?*
2. *Quelles sont les variables étudiées ? Quels sont leur type ?*
3. *Construire un histogramme de la variable "Âge" pour l'ensemble de l'échantillon.*
4. *Construire un histogramme de cette variable pour le sous-échantillon des hommes puis pour celui des femmes. Comparer ces deux histogrammes.*
5. *Faire une représentation graphique de la variable "sexe".*
6. *Déterminer la médiane et les quartiles $Q_1$ et $Q_3$.*

---

**Exercice 21.** ***Mobilité dans la ville de Versailles entre 1830 et 1880.*** *`Extrait d'un article de Claire Lévy-Vroélan, du centre de recherche sur l'Habitat.` , la médiane, les quartiles Si la mobilité extra-communale est un facteur classique des études historiques et*

*économiques, en particulier pour le XIXième siècle, le facteur de la mobilité interne dans une même région ou commune est lui moins utilisé. Pourtant, il participe à la compréhension du fonctionnement et de l'évolution des ségrégations urbaines, comme la constitution d'espaces spécifiques -populaires ou bourgeois. Dans ce type d'études, la ville de Versailles offre une exceptionnelle occasion, compte tenu de la qualité de ses registres tenus annuellement et comportant de nombreuses informations permettant d'effectuer le recensement de chaque déménagement dans la même ville sur toute la période.*

*Le tableau d'effectifs suivant donne le temps de séjour moyen dans une même habitation pour 100 ménages :*

| *Temps de séjour moyen (en année)* | *Nombre de ménages* |
|---|---|
| $[0;3[$ | *45* |
| $[3;5[$ | *18* |
| $[5;10[$ | *16* |
| $[10;20[$ | *11* |
| $[20;50[$ | *10* |

1. *Préciser quelle est la population $\Omega$ étudiée et donner sa taille.*
2. *Quelle est la variable étudiée ? Préciser son type.*
3. *Représenter la variable temps de séjour.*
4. *Calculer les effectifs cumulés associés à cette variable.*
5. *Calculer la médiane de cette variable.*
6. *Déterminer les 1er et 3ième quartiles (en précisant bien quelle est la lecture graphique). Que signifient concrètement ces valeurs ?*

---

**Exercice 22.** *Dans un groupe de 20 personnes, la variable âge a pour moyenne 22 ans et pour écart-type 0 an. Qu'en déduisez-vous concernant les âges de ces personnes ?*

---

**Exercice 23.** *Dans un groupe de 20 personnes, la variable âge a pour moyenne 22 ans et pour variance 0 an. Qu'en déduisez-vous concernant les âges de ces personnes ?*

---

**Exercice 24.** *Que représente l'écart-type pour une variable quantitative ?*

---

**Exercice 25.** *Dans un groupe de 25 personnes, le numéro d'INSEE moyen est 1720974120045. Quel est ici le sens de ce nombre ?*

---

**Exercice 26.** *Le tableau ci-dessous indique la répartition des couples suivant le nombre d'enfants.*

| *Nombre d'enfants* | *0* | *1* | *2* | *3* | *4* | *5* |
|---|---|---|---|---|---|---|
| *Effectifs* | *31* | *16* | *27* | *15* | *7* | *4* |

*Calculer :*

1. *le mode de cette série statistique ;*
2. *la moyenne et la médiane ;*
3. *l'étendue et l'intervalle interquartile ;*
4. *la variance puis l'écart-type.*

---

**Exercice 27.** *On reprend l'exercice du chapitre précédent des élèves ayant participé à une dictée. Voici le nombre de fautes relevées sur chaque dictée :*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *8* | *3* | *5* | *1* | *4* | *2* | *8* | *7* | *9* | *4* |
| *3* | *1* | *0* | *2* | *8* | *2* | *1* | *8* | *4* | *0* |
| *2* | *8* | *5* | *1* | *0* | *5* | *7* | *4* | *3* | *9* |
| *6* | *8* | *4* | *9* | *2* | *8* | *8* | *7* | *0* | *3* |
| *6* | *8* | *4* | *9* | *2* | *8* | *7* | *7* | *0* | *3* |

*1. Déterminer la médiane et les quartiles de cette série statistique ainsi que l'intervalle interquartile.*

*2. Déterminer la moyenne et l'écart-type.*

*3. À partir de ces deux derniers indices que peut-on dire de ce groupe d'élèves ? Est-il homogène ?*

---

**Exercice 28.** *On a mesuré en centièmes de seconde (noté cs) le temps (variable $X$ nécessaire pour identifier un moy simple par un enfant scolarisé en cours élémentaire de 2ème année (CE2). L'expérience a porté sur 100 sujets et on a obtenu les résultats suivants :*

| *Classes* | $[10;30[$ | $[30;40[$ | $[40;45[$ | $[45;50[$ | $[50;60[$ | $[60;70[$ | $[70;90[$ |
|---|---|---|---|---|---|---|---|
| $n_i$ | *12* | *13* | *10* | *15* | *23* | *17* | *10* |

*1. Donner le type de la variable $X$ puis préciser la population ainsi que la taille de l'échantillon.*

*2. Compléter le tableau suivant :*

| *Classes* | $x_i$ *(centre des classes)* | $n_i$ | $n_i x_i$ | $n_i(x_i)^2$ | $a_i$ *(amplitude des classes)* | $\frac{n_i}{a_i}$ | $N_i$ *(eff. cumulés)* |
|---|---|---|---|---|---|---|---|
| $[10;30[$ | | | | | | | |
| $[30;40[$ | | | | | | | |
| $[40;45[$ | | | | | | | |
| $[45;50[$ | | | | | | | |
| $[50;60[$ | | | | | | | |
| $[60;70[$ | | | | | | | |
| $[70;90[$ | | | | | | | |
| *TOTAL* | - | | | | | | |

*3. Représenter l'histogramme.*

*4. Calculer la moyenne, la variance et l'écart-type de $X$.*

*5. Déterminer la médiane ; que représente cette valeur ?*

*6. On a réalisé la même expérience sur un groupe d'adultes pour lequel le temps moyen d'identification est de 25,3 cs et l'écart-type est de 14,7 cs. Un adulte ayant un temps de 37,4 cs est-il plus rapide qu'un enfant ayant un temps de 61.3 cs, ceci relativement bien sûr à leur population d'origine ?*

---

**Exercice 29.** *Une enquête sur le nombre annuel $X$ d'infractions à la législation du travail porte sur des entreprises (E) numérotées de 1 à 48.*

| *E* | *X* | *E* | *X* | *E* | *X* | *E* | *X* |
|---|---|---|---|---|---|---|---|
| *1* | *13* | *13* | *17* | *25* | *17* | *37* | *15* |
| *2* | *22* | *14* | *1* | *26* | *16* | *38* | *24* |
| *3* | *11* | *15* | *15* | *27* | *22* | *39* | *28* |
| *4* | *23* | *16* | *12* | *28* | *4* | *40* | *5* |
| *5* | *15* | *17* | *12* | *29* | *24* | *41* | *22* |
| *6* | *15* | *18* | *5* | *30* | *7* | *42* | *6* |
| *7* | *0* | *19* | *23* | *31* | *27* | *43* | *21* |
| *8* | *16* | *20* | *16* | *32* | *8* | *44* | *26* |
| *9* | *14* | *21* | *23* | *33* | *16* | *45* | *14* |
| *10* | *5* | *22* | *25* | *34* | *23* | *46* | *16* |
| *11* | *24* | *23* | *7* | *35* | *13* | *47* | *3* |
| *12* | *6* | *24* | *13* | *36* | *2* | *48* | *23* |

*1. Quels sont les individus ? Quelle est la variable ? Préciser ses modalités et son type. Quelle est la taille de l'échantillon ?*

*2. Déterminer médiane et moyenne puis indiquer quel est de ces deux indices celui qui fournit l'information la plus accessible à un profane.*

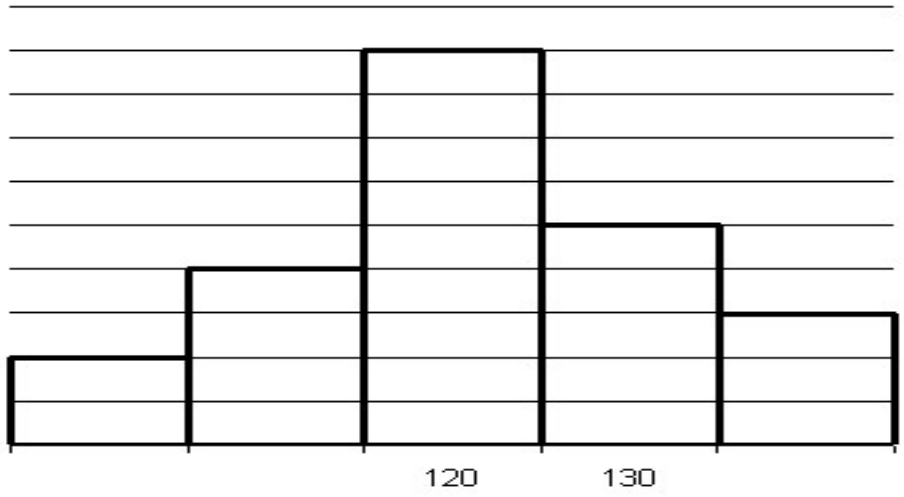
**Exercice 30.** *Sur les 150 copies à un examen :*

- *50 ont été corrigées par l'enseignant A donnant une moyenne de 9 ;*
- *65 l'ont été par l'enseignant B donnant une moyenne de 11 ;*
- *35 par l'enseignant C donnant une moyenne de 13.*

*Donner la moyenne sur l'ensemble des 150 copies.*

---

**Exercice 31.** *Le graphique ci-dessous représente l'histogramme des effectifs da la variable Q.I. (notée $X$) sur un échantillon de personnes. Les modalités de $X$ sont regroupées en classes de même amplitude 10. La classe de centre 120 a pour effectif 63.*




1. *Déterminer la fréquence de chaque classe.*
2. *Calculer la moyenne et l'écart-type de $X$ sur cet échantillon.*

---

**Exercice 32.** *Un client a acheté un stylo à 1,90 euros un livre à 9 euros et un répertoire à 3,20 euros dans une librairie de Toulouse.*

*On a constaté, après enquête dans l'ensemble des librairies de Toulouse, que le prix moyen du stylo est de 1,50 euros avec un écart-type de 0,40 euro ; le prix moyen du livre est de 8,5 euros avec un écart-type de 1 euro et le prix moyen du répertoire est de 3,50 euros avec un écart-type de 0,60 euros.*

*Déterminer l'article le moins cher, puis le plus cher acheté par le client. (Il faudra bien sûr donner un sens à cette question qui prise au 1er degré serait triviale).*

**Exercice 33.** *Le tableau de données ci-dessous indique la proportion de femmes mariées sur l'ensemble des femmes au moment du déclin de la fécondité pour chaque pays européen (c'est un des paramètres explicatifs dans la théorie de la transition démographique).*

*Nous noterons $X$ la variable "pourcentage de femmes mariées parmi les femmes ".*

| *Pays* | *X (en %)* | *Pays* | *X (en %)* |
|---|---|---|---|
| *France* | *51* | *Pays-Bas* | *45* |
| *Belgique* | *44* | *Danemark* | *47* |
| *Suisse* | *44* | *Norvège* | *42* |
| *Allemagne* | *50* | *Autriche* | *51* |
| *Hongrie* | *70* | *Finlande* | *46* |
| *Angleterre* | *48* | *Italie* | *54* |
| *Suède* | *42* | *Espagne* | *51* |
| *Ecosse* | *42* | *Irlande* | *35* |

*Source : Histoire des populations de l'Europe.*

*1. Quelle est la population étudiée et la taille de cette population ?*

*2. Construire le tableau d'effectifs de la variable $X$ puis calculer la moyenne, la variance et l'écart-type de $X$.*

*3. Calculer la médiane de $X$ et expliquer ce que représente ce nombre.*

*4. Calculer le premier et le troisième quartile.*

---

**Exercice 34.** *Les résultats d'un concours sont donnés dans le tableau suivant :*

| *Notes* | $[0;5[$ | $[5;8[$ | $[8;10[$ | $[10;12[$ | $[12;16[$ | $[16;20]$ |
|---|---|---|---|---|---|---|
| *Effectifs* | *17* | *16* | *26* | *30* | *24* | *11* |

*(notes sur 20).*

*1. Quelle est la population étudiée ? Quelle est la variable (ou caractère) et le type de cette variable ?*

*2. Faire le tableau de distribution des fréquences (en %).*

*3. On sait que 25% des candidats sont admis. Déterminer à l'aide d'un graphique approprié la note à partir de laquelle un candidat est admis. Peut-on calculer ce nombre ?*

*4. Calculer la note moyenne $\bar{m}$ et l'écart- type $\sigma$.*

**Exercice 35.** ***Répartition des chômeurs en avril 1975.***

| *Tranches d'âge* | *Nombre de chômeurs* |
|---|---|
| *moins de 18 ans* | *41700* |
| *18-24 ans* | *277200* |
| *25-39 ans* | *246600* |
| *40-49 ans* | *134100* |
| *50-59 ans* | *88600* |
| *60 ans et plus* | *38900* |

*Source : Enquête emploi INSEE*

*1. Calculer l'âge moyen des chômeurs et l'écart-type. On prendra la classe* $[16;18[$ *pour les moins de 18 ans et* $[60;65[$ *pour les plus de 60 ans.*

*2. Représenter les données par un histogramme puis estimer graphiquement le nombre de chômeurs entre 35 et 45 ans.*

*3. Calculer les trois quartiles :* $Q_1$, $Q_2$ *(la médiane) et* $Q_3$.

---

**Exercice 36.** *Le graphique suivant est le diagramme cumulatif du taux de boisement (exprimé en pourcentage et réparti en 5 classes distinctes) des différentes communes du département de la Corrèze :*

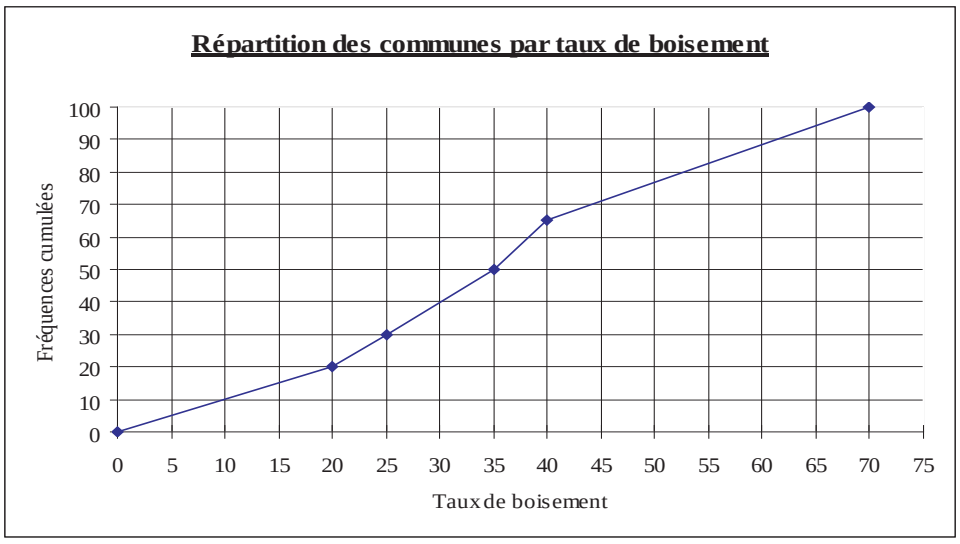



FIGURE 9.2 – *Diagramme cumulatif.*

***Remarque*** *: Taux de boisement = (Superficie de forêts de la commune) / (Superficie totale de la commune).*

*1. Quelle est la population étudiée ? Quelle est la variable ? Préciser le type de la variable.*

*2. Lire sur le graphique le pourcentage de communes dont le taux de boisement est supérieur à 50%. (Les traits permettant la lecture doivent apparaître).*

*3. A l'aide du graphique, compléter le tableau des fréquences cumulées suivant :*

| *Taux de boisement* | [0; 20[ | | | | [40; 70[ |
|---|---|---|---|---|---|
| *Fréquence cumulée* | | | | | |
| *Fréquence* | | | | | |

*4. Déterminer graphiquement la médiane. Que représente ce nombre ?*

*5. Sachant que le nombre total de communes, dans le département de la Corrèze est 286, retrouver le tableau des effectifs :*

| *Taux de boisement* | [0; 20[ | | | | [40; 70[ |
|---|---|---|---|---|---|
| *Effectif* | | | | | |

*6. Déterminer la moyenne de cette série statistique. Cette moyenne est-elle le taux de boisement du département de la Corrèze ? Justifier.*

*7. Déterminer l'écart-type de cette série statistique.*

---

**Exercice 37.** *Le document de la page suivante permet de voir l'évolution de la pyramide des âges de la population française depuis la fin de la deuxième guerre mondiale.*

*1. Déterminer la population et les variables étudiées.*

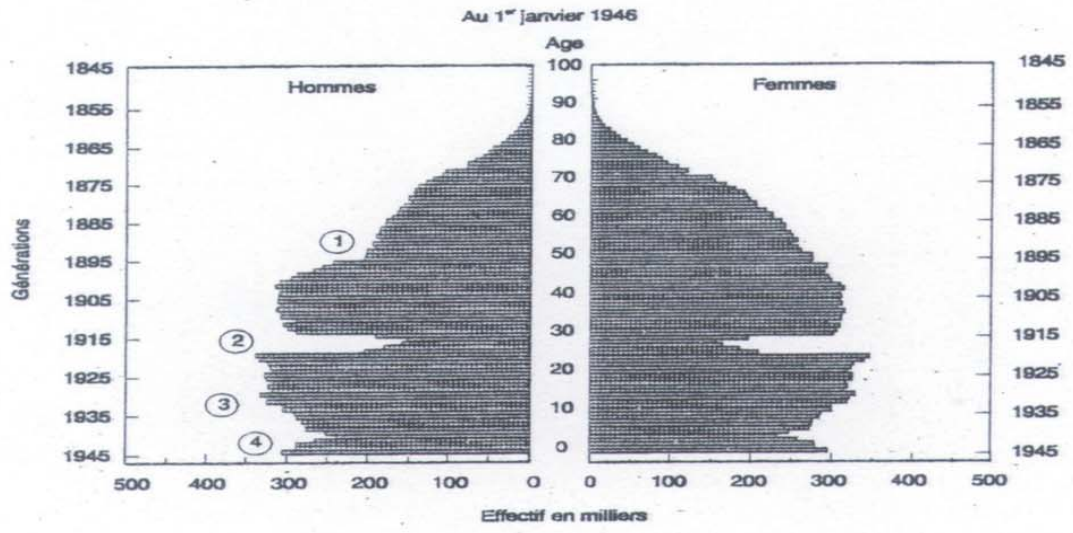
*2. Quel est le type de tableau représenté par la Pyramide des âges ?*

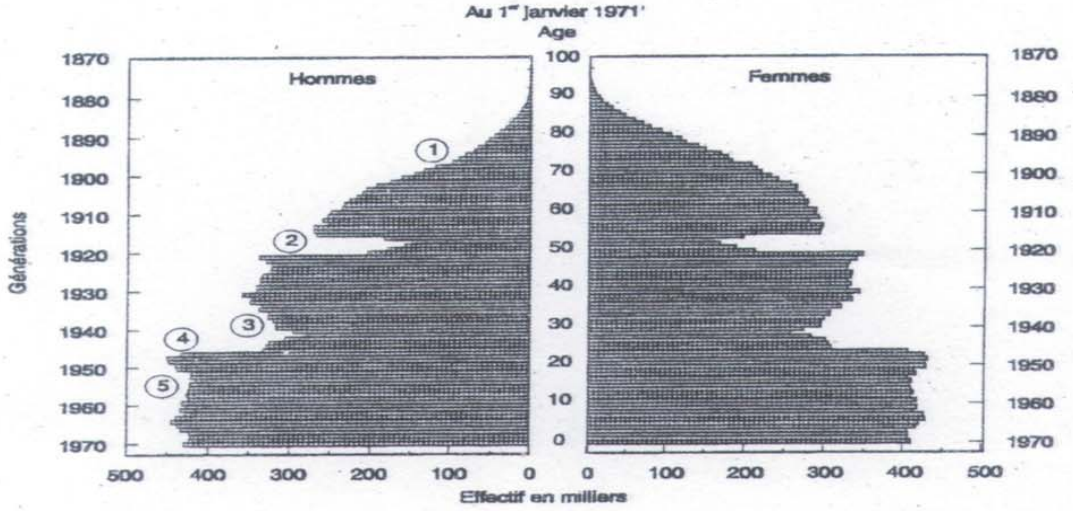
*3. Quelle est l'amplitude des classes pour la variable quantitative ?*

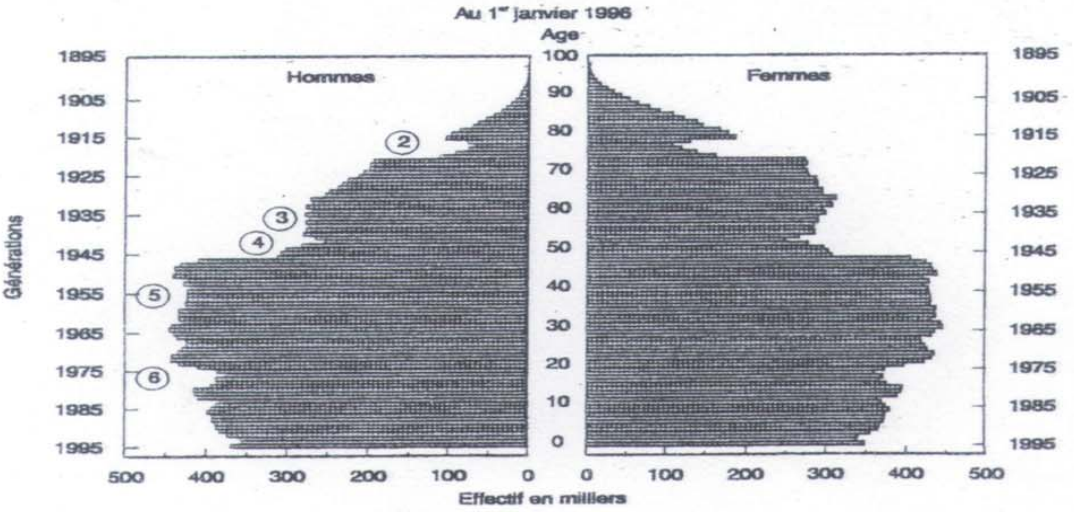
---

**Exercice 38.** ***Exploitation d'un tableau complexe***

*On s'intéresse au taux d'activité des femmes selon l'âge et le nombre d'enfants en 1982. Le taux d'activité est le pourcentage d'actifs par rapport à la population totale. Les effectifs sont donnés en milliers d'actives.*




Lecture :
① Pertes militaires de la guerre 1914-1918 (classes creuses)
② Déficit des naissances dû à la guerre 1914 - 1918
③ Passage des classes creuses à l'âge de la fécondité
④ Déficit des naissances dû à la guerre 1939-1945
⑤ "Baby-boom"
⑥ Baisse de la fécondité dans les années soixante-dix.

*Source : Statistiques de l'état civil, Insee*

| *Âge* | *15/19* | *20/24* | *25/29* | *30/34* | *35/39* | *40/44* | *45/49* | *50/54* | *55 et +* | *TOTAL* |
|---|---|---|---|---|---|---|---|---|---|---|
| *0 enfant* | *28,8* | *399,9* | *314,5* | *143,5* | *121,6* | *235,7* | *434,6* | *531,6* | *640* | *2850,2* |
| *Taux (%)* | *69,0* | *84,3* | *86,6* | *82,9* | *77,9* | *70,0* | *61,0* | *52,2* | *17,9* | *41,6* |
| *1 enfant* | *5,7* | *192,8* | *448,6* | *379,6* | *293,6* | *259,9* | *183,2* | *84,1* | *23,6* | *1871* |
| *Taux (%)* | *35,5* | *64,0* | *76,7* | *79,4* | *74,7* | *62,4* | *51,0* | *41,3* | *30,9* | *66,1* |
| *2 enfants* | *0,4* | *34,9* | *245,5* | *474,1* | *396,4* | *151,2* | *54,5* | *16,4* | *3,7* | *1350* |
| *Taux (%)* | *16,8* | *35,6* | *54,3* | *64,8* | *64,1* | *52,9* | *40,9* | *32,8* | *28,9* | *57,7* |
| *3 enfants* | *0,1* | *3* | *36,7* | *126,7* | *139,2* | *50,7* | *15,6* | *3,6* | *0,8* | *376,3* |
| *Taux (%)* | *14,3* | *11,9* | *20,8* | *31,9* | *37,3* | *32,3* | *24,2* | *19,0* | *31,1* | *30,9* |
| *TOTAL* | *34,9* | *630,5* | *1045,3* | *1123,9* | *923,8* | *697,4* | *687,9* | *635,7* | *668,1* | *6447,5* |
| *Taux (%)* | *57,5* | *70,2* | *66,3* | *63,1* | *61,6* | *58,3* | *54,2* | *49,2* | *18,3* | *48,7* |

*1. Définir clairement la population étudiée, les variables et leur type.*

*2. Que signifie le nombre 396,4 ? Que signifie le nombre 77,9 ? Que signifie le nombre 34,9 (ligne TOTAL) ?*

*3. Dans la colonne TOTAL, que représente les nombres 2850,2 et 41,6 ? Retrouver ces nombres à partir des données du tableau.*

*4. Calculer la moyenne d'âge des femmes actives sans enfant. On prendra la classe* $[55; 60[$ *pour les 55 ans et plus.*

*5. Faire l'histogramme de la variable "Age" (considérer les effectifs marginaux), calculer la moyenne et la médiane.*

*6. Faire le tableau de fréquence de la variable "Nombre d'enfants " puis représenter les résultats par un diagramme circulaire.*

*7. Faire un diagramme circulaire représentant la variable " Nombre d'enfants " définie non sur la population des femmes actives mais sur la population totale des femmes françaises.*

---

**Exercice 39.** ***L'explosion Urbaine***

*Population des villes (en millions) de plus de 10 millions d'habitants dans le monde en 2000 :*

| | | | |
|---|---|---|---|
| *Los Angeles* | *13,2* | *Bombay* | *16,1* |
| *Mexico* | *18,1* | *Calcutta* | *13,1* |
| *New York* | *16,7* | *Dacca* | *12,5* |
| *Buenos Aires* | *12,1* | *Beijong* | *10,8* |
| *Sao Paulo* | *18* | *Shangaï* | *12,9* |
| *Rio de Janeiro* | *10,7* | *Osaka* | *11* |
| *Paris* | *10* | *Tokyo* | *26,4* |
| *Karashi* | *10* | *Manille* | *10* |
| *Delhi* | *12,4* | *Jakarta* | *11* |

*Source : National Geographic - Novembre 2002.*

*1. Quelle est la population étudiée et sa taille ? Quelle est la variable étudiée et son type ?*

*2. Regrouper les données en classes (*$[10;12[$, $[12;14[$, $[14;17[$, $[17;27[$ *) et déterminer les centres des classes. Utiliser le tableau ci-dessous.*

| *Population (en millions d'habitants)* | *Effectifs* $n_i$ | *Centres* $c_i$ | $n_i c_i$ | $n_i(c_i)^2$ | *Effectifs cumulés* $n_i^*$ |
|---|---|---|---|---|---|
| $[10;12[$ | | | | | |
| $[12;14[$ | | | | | |
| $[14;17[$ | | | | | |
| $[17;27[$ | | | | | |

*3. Calculer la moyenne et l'écart-type (après regroupement en classes).*

*4. Calculer les effectifs cumulés et faire le diagramme cumulatif.*

*5. Calculer la médiane et expliquer ce que représente ce nombre. Déterminer graphiquement le premier et le troisième quartile (mettre en évidence la lecture graphique).*

---

**Exercice 40.** *Dans un livre en hommage à Ernest Labrousse, François Bérabida a écrit un texte intitulé : " Peuple ou classe ouvrière ? Un quartier de l'East End au XIXème siècle ", qui débute ainsi et qui contient un tableau à propos duquel vous sont posées quelques questions. " A l'extrémité Est du comté de Londres, le district de Poplar, encadré entre la Tamise et la Lea, entrecoupé de docks, de canaux et de voies ferrées, aligne sur 9 km*$^2$ *de terres basses et plates, en majorité alluviales et par endroit marécageuses, ses petites maisons pauvres et grises, ternes et monotones, qu'habite une population à peu près entièrement ouvrière. Son développement, qui date principalement du milieu du XIXème siècle, s'inscrit dans la grande poussée de l'East End. "*

| *Localités* / *Nombre de salariés* | *BOW* | *BROMLEY* | *POPLAR* |
|---|---|---|---|
| $[100;200[$ | *6* | *3* | *15* |
| $[200;500[$ | *4* | *4* | *8* |
| $[500;1000[$ | *1* | *3* | *2* |
| $[1000;5000[$ | *2* | *1* | *1* |

*1. Définir la population étudiée.*

*2. Quelles sont les variables et leur type ?*

*3. Calculer les effectifs marginaux et la taille de la population.*

*4. Quel est le pourcentage d'usines et entreprises dans la localité de Bow, parmi celles du district de Poplar ?*

*5. Quel est le pourcentage d'usines et entreprises de plus de 1000 salariés, parmi celles du district de Poplar ?*

*6. Quel est le pourcentage d'usines et entreprises de 100 à 200 salariés de la localité de Poplar, parmi celles de 100 à 200 salaries du district de Poplar ?*

*7. On considère à présent les effectifs marginaux de la variable " nombre de salariés ", notée $X$. Calculer la médiane, la moyenne et l'écart-type de $X$.*

---

## 9.3 Exercices du chapitre 4

**Exercice 41.** *Le tableau ci-dessous est extrait d'une enquête réalisée sur 2000 élèves de classe de troisième en 1963. Les valeurs ont été arrondies afin de simplifier leur traitement.*

| | *Effectifs* | | | *% par colonne* | | | *% par ligne* | | |
|---|---|---|---|---|---|---|---|---|---|
| *Choix professionnel* | *G* | *F* | *Ens.* | *G* | *F* | *Ens.* | *G* | *F* | *Ens.* |
| *Médecin* | *30* | *30* | *60* | *3,0* | *3,0* | *3,0* | *50,0* | *50,0* | *100* |
| *Ingénieur* | *130* | *5* | *135* | *13,1* | *0,5* | *6,8* | *96,3* | *3,7* | *100* |
| *Cadre supérieur* | *20* | *5* | *25* | *2,0* | *0,5* | *1,3* | *80,0* | *20,0* | *100* |
| *Instituteur* | *60* | *140* | *200* | *6,1* | *13,9* | *10,0* | *30,0* | *70,0* | *100* |
| *Technicien* | *200* | *35* | *235* | *20,2* | *3,5* | *11,8* | *85,1* | *14,9* | *100* |
| *Cadre moyen* | *60* | *130* | *190* | *6,1* | *12,9* | *9,5* | *31,6* | *68,4* | *100* |
| *Employé* | *50* | *155* | *205* | *5,1* | *15,3* | *10,3* | *24,4* | *75,6* | *100* |
| *Ouvrier qualifié* | *40* | *10* | *50* | *4,0* | *1,0* | *2,5* | *80,0* | *20,0* | *100* |
| *Autre choix* | *400* | *500* | *900* | *40,4* | *49,5* | *45,0* | *44,4* | *55,6* | *100* |
| *Total* | *990* | *1010* | *2000* | *100* | *100* | *100* | *49,5* | *50,5* | *100* |

*1. Indiquer la population, les variables étudiées dans ce tableau en précisant le type de chaque variable.*

*2. Que représente chacun des nombres suivants :*

*a) dans la 3ième colonne de nombres : 200*

*b) dans la 4ième colonne de nombres : 6,1*

*c) dans la 5ième colonne de nombres : 15,3*

*d) dans la 6ième colonne de nombres : 11,8*

*e) dans la 8ième colonne de nombres : 75,6*

*f) dans la ligne "total" : 1010 puis 50,5*

*g) dans la ligne "Ingénieur" : la somme des deux nombres 13,1 et 0,5*

*h) dans la 4ième colonne de nombres : la somme des deux nombres 3,0 et 13,1.*

*3. Sans calcul, par simple lecture du tableau ci-dessus donner les informations suivantes :*

*a) le nombre d'enfants voulant être instituteurs*

*b) le pourcentage d'enfants voulant être ouvrier qualifié*

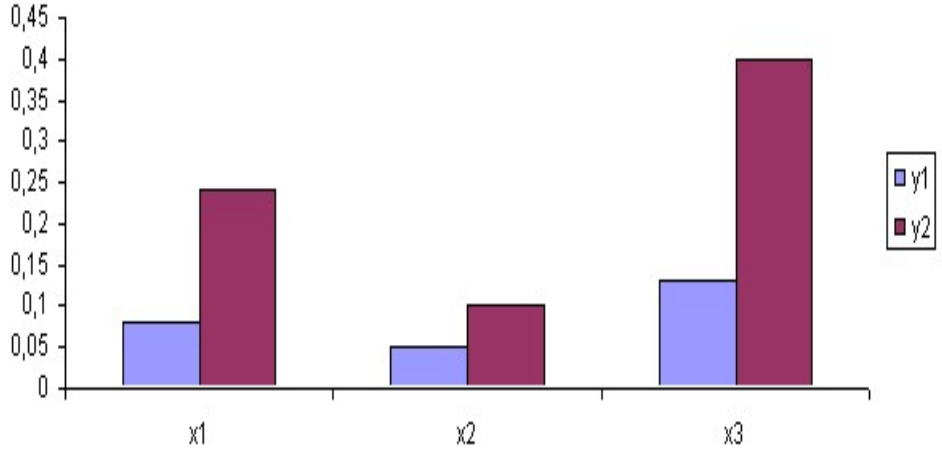
*c) le pourcentage de garçons parmi les enfants voulant être cadre moyen*

*d) le tableau de fréquences du sexe conditionnellement au choix professionnel*

*e) le tableau de fréquences du choix professionnel conditionnellement au sexe.*

---

**Exercice 42.** *Une variable $X$ (3 modalités : $x_1$, $x_2$, $x_3$) et une variable $Y$ (2 modalités : $y_1$ et $y_2$) ont été étudiées sur un échantillon.*




*La figure ci-dessus représente*

- ☐ *la distribution de $X$ conditionnellement à $Y$*
- ☐ *la distribution de $Y$ conditionnellement à $X$*
- ☐ *la distribution conjointe de $X$ et $Y$.*

---

**Exercice 43.** *Sur une population de 15 personnes, on considère les variables "Sexe" et "Âge". Voici le tableau d'effectifs conjoints :*

| *Âge* / *Sexe* | *moins de 20 ans* | *20 ans et plus* |
|---|---|---|
| *masculin* | *3* | *2* |
| *féminin* | *7* | *3* |

*Donner le tableau de distribution de la variable Âge conditionnellement à la variable Sexe. Représenter ce résultat graphiquement.*

---

**Exercice 44.** *Sur une population de 15 personnes, on considère les variables "Sexe" et "Âge". Voici le tableau d'effectifs conjoints :*

| *Âge* / *Sexe* | *moins de 20 ans* | *20 ans et plus* |
|---|---|---|
| *masculin* | *3* | *1* |
| *féminin* | *7* | *4* |

*Donner le tableau de distribution de la variable Sexe conditionnellement à la variable Âge. Représenter ce résultat graphiquement.*

---

**Exercice 45.** *La variable Sexe (2 modalités : homme et femme) et la variable État Civil (2 modalités : marié et célibataire) ont été étudiées sur un échantillon. Après examen des résultats un statisticien déclare que : "sur l'échantillon étudié, la fréquence de la modalité homme conditionnellement à la modalité célibataire est égale à 25%". Cela signifie-t-il que :*

□ *25% de l'échantillon est constitué d'hommes célibataires ?*

□ *parmi les célibataires de l'échantillon 25% sont des hommes ?*

□ *parmi les hommes de l'échantillon 25% sont des célibataires ?*

---

**Exercice 46.** *Soit le tableau d'effectifs ci-dessous.*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ | *Total* |
|---|---|---|---|---|
| $x_1$ | 3 | 5 | 6 | 14 |
| $x_2$ | 1 | 2 | 3 | 6 |
| *Total* | 4 | 7 | 9 | 20 |

*La fréquence de $x_2$ conditionnellement à $y_3$ est égale à :*

3/20 □ 1/3 □ 1/2 □.

---

**Exercice 47.** *Le tableau ci-dessous est :*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ | *Total* |
|---|---|---|---|---|
| $x_1$ | 10% | 25% | 25% | 60% |
| $x_2$ | 10% | 15% | 15% | 40% |
| *Total* | 20% | 40% | 40% | 100% |

□ *le tableau de fréquences conjointes de $X$ et $Y$*

□ *le tableau de distribution de $X$ conditionnellement à $Y$*

□ *le tableau de distribution de $Y$ conditionnellement à $X$.*

---

**Exercice 48.** *Compléter le tableau ci-dessous :*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ | *Total* |
|---|---|---|---|---|
| $x_1$ | *70* | *80* | | *200* |
| $x_2$ | | *40* | *25* | |
| $x_3$ | | | | *100* |
| *Total* | | *160* | *100* | *400* |

---

**Exercice 49.** *Dans la population Toulousaine, on considère un échantillon E composé des habitants d'une rue de Toulouse (80 personnes de sexe masculin et 100 personnes de sexe féminin) dont les numéros vont de 1 à 60. On s'intéresse aux variables suivantes : la variable T qui associe à chaque habitant le n° de son logement, la variable "Sexe", la variable "pointure des chaussures", la variable "parité" obtenue à partir de T en regroupant ses modalités en deux classes : pair et impair. On connaît également les pourcentages suivants sur l'échantillon E :*

$p_1 = 45\%$ *: le pourcentage d'hommes parmi les personnes habitant un n° pair ;*

$p_2 = 50\%$ *: le pourcentage d'hommes parmi les personnes habitant un n° impair ;*

$p_3$ *: le pourcentage de femmes parmi les personnes habitant un n° impair.*

*Choisissez la ou les réponses appropriées à chaque question :*

*1. T est une variable :*

☐ *nominale* ☐ *ordinale* ☐ *quantitative*

*2. La pointure est une variable*

☐ *nominale* ☐ *ordinale* ☐ *quantitative*

*3. Le calcul de* $p_1 + p_2$

☐ *donne le pourcentage d'homme dans l'échantillon*

☐ *n'a aucun sens*

*4. le calcul de* $p_2 + p_3$

☐ *donne le pourcentage de numéros impairs dans l'échantillon*

☐ *donne le pourcentage de personnes habitant un numéro impair dans l'échantillon*

☐ *n'a aucun sens*

☐ *représente la totalité de l'échantillon*

☐ *représente l'ensemble des personnes habitant un numéro impair dans l'échantillon*

*5.* $p_1$ *est une des valeurs*

☐ *de la distribution du sexe conditionnellement à la parité dans l'échantillon*

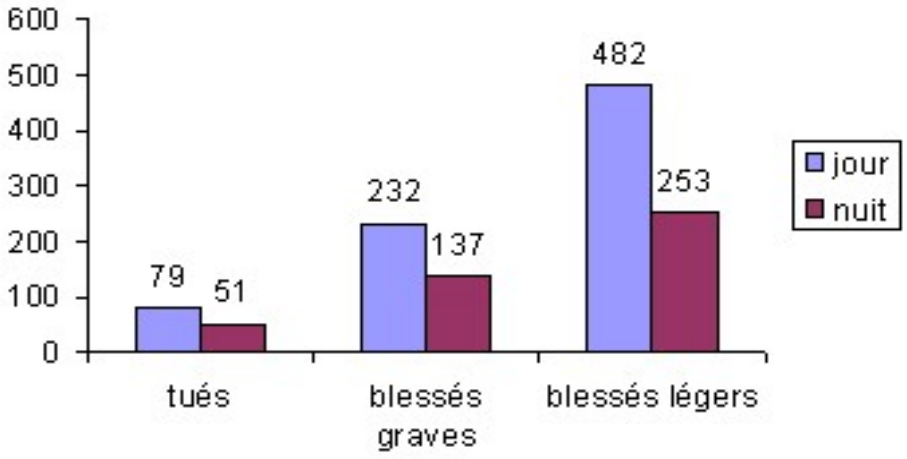
☐ *de la distribution de la parité conditionnellement au sexe dans l'échantillon*

☐ *de la distribution marginale de la parité en pourcentage dans l'échantillon*

□ *de la distribution marginale du sexe en pourcentage dans l'échantillon.*

---

**Exercice 50.** *Dans un article paru dans le journal "La dépêche" du 10 novembre 1999 et intitulé "Des chiffres qui font mal" on trouve le tableau ci-dessous qui concerne l'ensemble des personnes accidentées de la route en 1998 pour la Haute-Garonne.*




*1. Indiquer la population, les variables ainsi que leur type.*

*2. Construire la table de contingence.*

*3. Construire les deux distributions conditionnelles.*

*4. Le journaliste a donné à ce graphique la légende suivante : "Gravité plus importante la nuit". Indiquer les effectifs ou pourcentages écrits précédemment qui justifient une telle légende.*

---

**Exercice 51.** *Le tableau suivant résume les résultats à trois concours A, B, C en tenant compte du sexe :*

| | | Concours / Sexe | A | B | C |
|---|---|---|---|---|---|
| Admission | oui | G | 10 | 10 | 10 |
| | | F | 0 | 35 | 25 |
| | non | G | 43 | 0 | 2 |
| | | F | 4 | 5 | 11 |

*1. Quel est le nombre de modalités du triplet de variables (Sexe, Concours, Admission) ?*

*2. Donner les distributions conjointes et marginales des couples de variables : (Sexe, Concours), (Sexe, Admission), (Admission, Concours)*

*3. On considère le groupe des garçons. Donner la distribution de la variable Admission conditionnellement à la variable Concours.*

*4. On considère le groupe des filles. Donner la distribution de la variable Admission conditionnellement à la variable Concours.*

*5. Pour chacun des trois concours comparer le pourcentage de réussite des filles à celui des garçons.*

*6. Comparer pour l'ensemble des trois concours le pourcentage de réussite des filles à celui des garçons. Que conclure ?*

---

**Exercice 52.** *On reprend les données de l'exercice 8 concernant les chômeurs français en mars 1989.*

*1. Représenter graphiquement la distribution conjointe.*

*2. Donner le tableau de la distribution de l'âge conditionnellement au sexe.*

*Représenter graphiquement ce résultat.*

*3. Donner le tableau de la distribution du sexe conditionnellement à l'âge.*

*Représenter graphiquement ce résultat.*

---

## 9.4 Exercices du chapitre 5

**Exercice 53.** *À partir des données de l'exercice 41. du chapitre précédent, on regroupe les modalités de la variable choix professionnel en 3 classes (les 3 premières, les 3 suivantes, et les 3 dernières).*

1. *Donner la table de contingence des variables sexe et choix professionnel.*
2. *Calculer $\chi^2$ et $\varphi$ puis conclure.*

---

**Exercice 54.** *Soit 2 variables $X$ et $Y$ étudiées sur un échantillon de taille 300 et dont un tableau de distributions conditionnelles est donné ci-dessous :*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ |
|---|---|---|---|
| $x_1$ | 2/3 | 2/3 | 2/3 |
| $x_2$ | 1/3 | 1/3 | 1/3 |
| Total | 1 | 1 | 1 |

1. *Que peut-on conclure pour $X$ et $Y$ ?*
2. *Quelle est la valeur du $\chi^2$ ?*

---

**Exercice 55.** *Soit le tableau d'effectifs observés suivant :*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ | Total |
|---|---|---|---|---|
| $x_1$ | 2 | 4 | 6 | 12 |
| $x_2$ | 3 | 6 | 9 | 18 |
| Total | 5 | 10 | 15 | 30 |

1. *Compléter le tableau des effectifs théoriques d'indépendance ci-dessous :*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ | Total |
|---|---|---|---|---|
| $x_1$ | | 4 | | |
| $x_2$ | | | 9 | |
| Total | | | | |

2. *Sans effectuer de calcul, donner en la justifiant, la valeur de $\chi^2$.*

---

**Exercice 56.** *Le tableau suivant donne la répartition de 2000 personnes selon les deux variables "Age" et "Sport pratiqué".*

| | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* |
|---|---|---|---|---|---|
| *Moins de 20 ans* | *50* | *140* | *20* | *140* | *150* |
| *De 20 à moins de 30 ans* | *80* | *150* | *50* | *170* | *250* |
| *De 30 à moins de 40 ans* | *80* | *50* | *70* | *100* | *200* |
| *40 ans et plus* | *30* | *20* | *60* | *90* | *100* |

*1. Parmi les personnes de moins de 30 ans sur lesquelles l'enquête a été réalisée, existe-t-il un lien entre l'âge et le sport pratiqué ? (si oui, décrire l'importance de ce lien).*

*2. Même question concernant les personnes de plus de 30 ans.*

---

**Exercice 57.** *Deux traitements A et B ont été proposés contre l'acné .*

*Sur 710 personnes ayant de l'acné et soumis au traitement A on a observé 497 guérisons.*

*Sur 1070 personnes ayant de l'acné et soumis au traitement B on a observé 856 guérisons.*

*Peut-on dire que les deux traitements sont de même efficacité pour les 1780 personnes testées ? (en cas de différence d'efficacité en décrire l'importance.)*

---

**Exercice 58.** *En Lidurie, le premier tour des élections au poste de Grand Mamamouchi vient de se dérouler et une étude de Sciences Politiques porte sur l'électorat de Johnny, candidat arrivé en tête. Cette étude concerne les 36 bureaux de vote du pays, répartis sociologiquement en 3 catégories*

*A : ville B : banlieue C : campagne*

*La première colonne est le numéro du bureau de vote, la deuxième la catégorie de ce bureau, la troisième est le pourcentage de voix obtenues par Johnny dans ce bureau et la dernière colonne concerne l'opinion du député de la circonscription : est-il favorable à Johnny ?*

*On a obtenu les résultats suivants :*

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $A$ | 27 | *oui* | 10 | $C$ | 28 | *non* | 19 | $A$ | 27 | *oui* | 28 | $A$ | 25 | *non* |
| 2 | $A$ | 29 | *oui* | 11 | $C$ | 32 | *oui* | 20 | $A$ | 29 | *oui* | 29 | $A$ | 31 | *non* |
| 3 | $A$ | 28 | *oui* | 12 | $C$ | 28 | *non* | 21 | $A$ | 28 | *non* | 30 | $A$ | 29 | *oui* |
| 4 | $A$ | 26 | *non* | 13 | $A$ | 26 | *oui* | 22 | $A$ | 29 | *non* | 31 | $A$ | 30 | *non* |
| 5 | $B$ | 30 | *non* | 14 | $B$ | 26 | *oui* | 23 | $B$ | 28 | *non* | 32 | $B$ | 26 | *non* |
| 6 | $B$ | 29 | *non* | 15 | $B$ | 29 | *oui* | 24 | $B$ | 27 | *non* | 33 | $B$ | 29 | *oui* |
| 7 | $B$ | 31 | *non* | 16 | $B$ | 27 | *oui* | 25 | $B$ | 27 | *oui* | 34 | $B$ | 28 | *oui* |
| 8 | $B$ | 25 | *oui* | 17 | $B$ | 30 | *non* | 26 | $B$ | 28 | *oui* | 35 | $B$ | 27 | *non* |
| 9 | $C$ | 28 | *oui* | 18 | $C$ | 28 | *non* | 27 | $C$ | 30 | *oui* | 36 | $C$ | 32 | *non* |

*1. Identifier les individus et les variables (en indiquant leur type).*

*2. Les collaborateurs de Johnny considèrent qu'un bureau de vote a fourni un bon résultat lorsque le pourcentage en faveur de Johnny est* $> 29$.

*a) Existe-t-il un lien important entre la nature du résultat (bon ou mauvais) et la catégorie sociologique à laquelle appartient le bureau de vote ?*

*b) Existe-t-il un lien important entre la nature du résultat et l'attitude du député de la circonscription pour les 36 bureaux de vote étudiés ?*

---

**Exercice 59.** *Lors d'un référendum organisé par le maire d'un petit village, on a constaté un lien total* ($\varphi = 1$) *entre la réponse à la question posée (les réponses possibles étant : "OUI ", "NON ", "SANS OPINION ") et le sexe.*

*Le nombre de suffrages exprimés étant de 500, imaginer plusieurs tables de contingence correspondant à cette situation, et vérifier par le calcul, pour l'une d'elles qu'on a effectivement* $\varphi = 1$.

---

# Chapitre 10

# Corrigé des exercices

## 10.1 Correction des exercices du Chapitre 1

---

**Exercice 1.** *Étant donnée une variable étudiée sur une population, à chaque individu est associé une et une seule modalité de la variable.*

---

**Exercice 2.** *Oui, chaque membre de la famille a un père unique : il y a 3 pères différents (celui du père, celui de la mère et celui des 3 enfants, en supposant que les 3 enfants ont le même père).*

---

**Exercice 3.** *1. Si la population est l'ensemble des $N$ électeurs de la Haute-Garonne, on peut définir la variable qualitative nominale qui à chaque individu associe un des 78 candidats. La série statistique est de taille $N$ et est la suite des $N$ couples suivants : $(E_1, nom_1)$, $(E_2, nom_2)$,..., $(E_N, nom_N)$, où $nom_k$ est la donnée associée à l'individu $E_k$ c'est-à-dire le candidat choisi par l'électeur $E_k$.*

*2. Si la population est l'ensemble des 78 candidats $C_1$,$C_2$,...,$C_{78}$. On considère la variable qualitative nominale qui, à chaque candidat $C_k$, associe la modalité "élu" ou "non élu". La série statistique est la suite des 78 couples de la forme $(C_k, résultat)$ . Elle est donc de taille 78. Les modalités de la variable sont "élu" ou "non élu". L'effectif de la modalité "élu" est donc 8 puisqu'il y a 8 sièges à pourvoir et l'effectif de la modalité "non élu" est par conséquent 70.*

---

**Exercice 4.** *Les individus sont les clients des magasins de bricolage "Cassetout" (qui passent à la caisse). Variable : code Postal. Type : variable qualitative nominale (éventuellement ordinale).*

---

**Exercice 5.** *Le mode d'une variable est la modalité ayant le plus grand effectif.*

---

**Exercice 6.** *1. $\Omega$ : la population active en France en 1930 ; $\omega$ : un actif français en 1930 ; $X$ : "catégorie socio-professionnelle" de type qualitatif nominal.*

*2. $\Omega$ : les entreprises françaises en 1973 ; $\omega$ : une entreprise française en 1973 ; $X$ : "la taille" avec les trois modalités petite, moyenne et grande, est ordinale.*

*3. $\Omega$ : les pays en 1991 ; $\omega$ : un pays en 1991 ; $X$ : "le P.N.B." , variable quantitative continue.*

*4. $\Omega$ : les billets en circulation dans le monde ; $\omega$ : un billet en circulation ; $X$ : "la valeur en euro" , variable quantitative continue.*

*5. $\Omega$ : les étudiants inscrits en licence d'histoire ; $\omega$ : un étudiant inscrit en licence d'histoire ; $X$ : "l'âge" , variable quantitative qui peut être considérée comme discrète si l'on compte en nombre d'années révolues ou continue si l'on considère qu'il s'agit d'une mesure du temps pouvant prendre toutes les valeurs dans un intervalle.*

*6. $\Omega$ : les logements de la ville de Toulouse ; $\omega$ : un logement de la ville de Toulouse ; $X$ : "le nombre d'habitants", variable quantitative discrète.*

---

**Exercice 7.**

1. *Puisque le tableau donné est un tableau d'effectifs, nous avons donc compté combien d'individus avaient donné la réponse Europe, combien la réponse Afrique... Nous avons donc demandé à chaque tonne d'or où elle avait été produite. La population est donc constituée des tonnes d'or et la variable (de type qualitatif nominal) est le continent où elle a été produite.*

| *Continents* | *Europe* | *Asie* | *Afrique* | *Amérique* | *Océanie* | *Total* |
|---|---|---|---|---|---|---|
| *Production en tonnes* | *176* | *87* | *431* | *350* | *56* | *1100* |
| *Fréquences* | *0,160* | *0,079* | *0,392* | *0,318* | *0,051* | *1* |

2. *Nous disposons maintenant d'un tableau de données. C'est plus facile et cela signifie que nous avons sur une ligne les individus et sur une autre leur réponse. Dans ce cas, cela signifie que nous avons demandé à chaque continent quelle est leur production. Donc la population est constituée des 5 continents. Un individu est donc un continent. La variable est la production d'or en tonnes d'or, variable quantitative continue.*

---

**Exercice 8.** *1. La population est l'ensemble des chômeurs en France, au sens du BIT, en mars 1989.*

*Les variables sont « le sexe », nominale et « l'âge », quantitative continue.*

*2. Il faut calculer les effectifs marginaux puis les diviser par la taille de la population.*

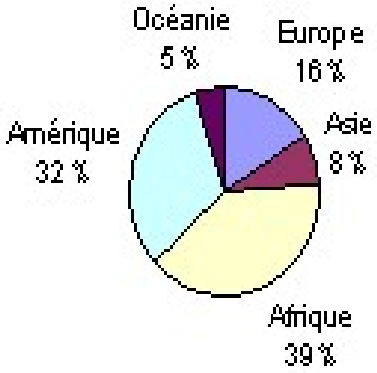



FIGURE 10.1 – *Diagramme circulaire de la production mondiale d'or.*

| | *Hommes* | *Femmes* | *Marge âge* | *Fréquences marginales* |
|---|---|---|---|---|
| *Moins de 25 ans* | *249* | *342,6* | *591,6* | *0,2563* |
| *De 25 à moins de 50 ans* | *565,1* | *827,9* | *1393* | *0,6035* |
| *50 ans et plus* | *168,5* | *155,2* | *323,7* | *0,1402* |
| *Marge sexe* | *982,6* | *1325,7* | *2308,3* | *1* |
| *Fréq. marginales* | *0,4257* | *0,5743* | *1* | |

*3. a) 57,43%*

*b) 25,63%*

*c)* $\dfrac{342,6}{591,6} = 57,91\%$

*4. On a représenté le diagramme en secteurs de la variable Age pour les hommes, l'histogramme de la variable Age pour les femmmes et le diagramme en secteurs de la variable Sexe pour la population des moins de 25 ans. Nous verrons dans le chapitre 4 comment représenter toutes les données de ce tableau de façon conjointe.*

**Exercice 9.** *1. Sur l'ensemble de la population de l'U.E., on étudie la variable « nationalité » dont les modalités sont les 15 pays de l'U.E., c'est une variable nominale.*

*2. La population est donc celle de l'U.E. et sa taille est 371,46 millions.*

*3. 4. Le mode, c'est-à-dire le pays le plus peuplé, est la modalité Allemagne.*

**Exercice 10.** *1. Pour dépouiller, on place les modalités à gauche dans un tableau et on met un petit bâton à côté de chaque modalité rencontrée, en regroupant les bâtons par 5 pour compter plus facilement.*

*2. La variable est le nombre de fautes de type quantitatif discret car c'est un comptage. On travaille sur un échantillon de 50 notes.*

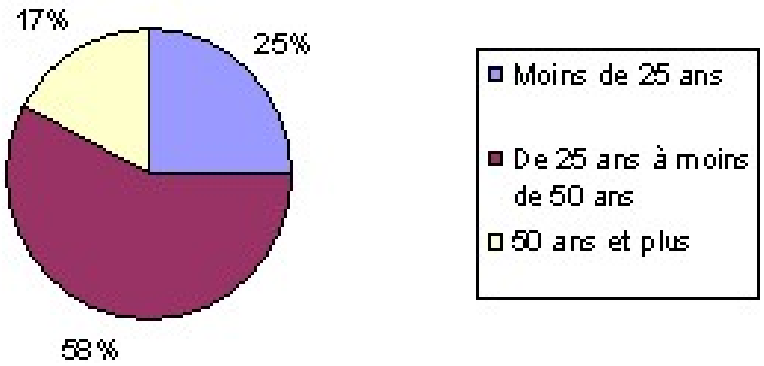



FIGURE 10.2 – *Diagramme circulaire de l'Age pour les Hommes.*

*3. On peut alors donner le tableau des effectifs et des effectifs cumulés :*

| *nombre de fautes* | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 5 | 4 | 6 | 5 | 6 | 3 | 2 | 5 | 10 | 4 |
| $n_i^*$ | 5 | 9 | 15 | 20 | 26 | 29 | 31 | 36 | 46 | 50 |

*Le mode (modalité ayant le plus gros effectif sur l'échantillon) est ici égal à 8.*

*4. L'étendue est 9-0=9.*

---

**Exercice 11.** *1. On considère ici que la variable « âge » est quantitative continue. Mais on considère qu'un individu entre 20 ans (compris) et 21 ans (non compris) a 20 ans (on n'a donc pas de classes mais des valeurs discrètes).*

*2. Le mode est 30 ans.*

*3. On calcule les effectifs cumulés :*

| *Âge* | *20* | *21* | *22* | *23* | *24* | *25* | *26* | *27* | *28* | *29* | *30* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Effectifs* | *4* | *7* | *11* | *13* | *23* | *32* | *37* | *45* | *70* | *56* | *82* |
| *Effectifs cumulés* | *4* | *11* | *22* | *35* | *58* | *90* | *127* | *172* | *242* | *298* | *380* |

*et on trace le diagramme cumulatif.*

---

**Exercice 12.** *1. La population est l'ensemble des femmes mariées d'un quartier ; la taille de l'échantillon est 60 (somme des effectifs). La variable étudiée est l'âge : il s'agit d'une variable quantitative continue.*

*2.*

| *Âge* | [18; 28[ | [28; 38[ | [38; 48[ | [48; 58[ | [58; 68[ | [68; 78[ | [78; 88[ | [88; 98[ |
|---|---|---|---|---|---|---|---|---|
| $f_j$ | *0,067* | *0,25* | *0,27* | *0,15* | *0,13* | *0,067* | *0,05* | *0,017* |
| $f_j^*$ | *0,067* | *0,317* | *0,587* | *0,737* | *0,867* | *0,934* | *0,984* | *1* |

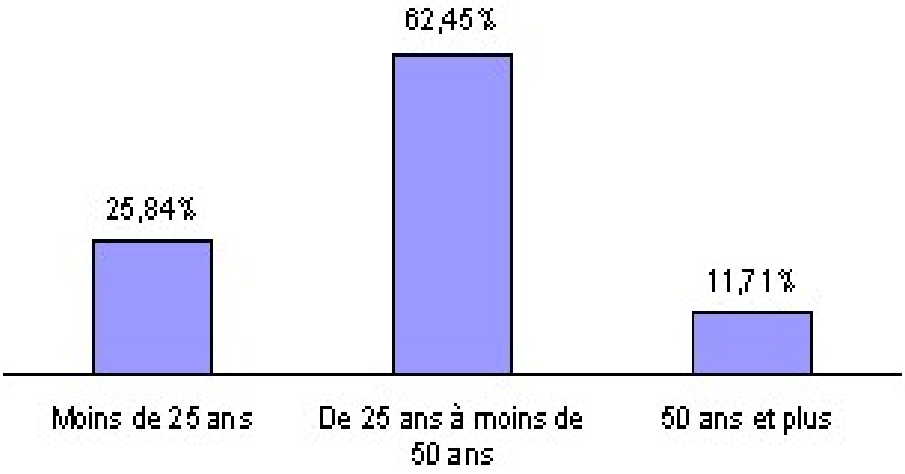



FIGURE 10.3 – *Histogramme de l'Age pour les Femmes.*

*3. Les classes sont d'amplitude constante* $a_i = 10$. *Pour tracer l'histogramme on divise les effectifs par l'amplitude :*

| *Âge* | [18; 28[ | [28; 38[ | [38; 48[ | [48; 58[ | [58; 68[ | [68; 78[ | [78; 88[ | [88; 98[ |
|---|---|---|---|---|---|---|---|---|
| $n_i/a_i$ | *0,4* | *1,5* | *1,6* | *0,9* | *0,8* | *0,4* | *0,3* | *0,1* |
| $f_j$ | *0,067* | *0,25* | *0,27* | *0,15* | *0,13* | *0,067* | *0,05* | *0,017* |
| $f_j^*$ | *0,067* | *0,317* | *0,587* | *0,737* | *0,867* | *0,934* | *0,984* | *1* |

**Exercice 13.** *La variable étudiée est la variable quantitative discrète "nombre d'enfants". La taille de l'échantillon est* $N = 100$. *La tableau statistique est le suivant :*

| *Nombre d'enfants* | *0* | *1* | *2* | *3* | *4* | *5* |
|---|---|---|---|---|---|---|
| *Effectifs* | *30* | *15* | *30* | *15* | *5* | *5* |
| *Eff. cum.* | *30* | *45* | *75* | *90* | *95* | *100* |

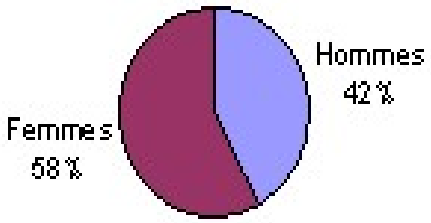



FIGURE 10.4 – *Diagramme circulaire du Sexe pour les moins de 25 ans.*

## 10.2 Correction des exercices des Chapitres 2 et 3

**Exercice 14.** *12 personnes ont un n° INSEE inférieur et 12 personnes ont leur n° plus grand (la moitié des individus sont des garçons nés jusqu'au mois de mai 78).*

**Exercice 15.** *1. La variable est ici ordinale car on considère des effectifs cumulés, ce qui ne peut se faire que s'il existe un ordre entre les différentes modalités.*

*2. À partir des pourcentages cumulés $p_i^*$, comme on connaît la taille de l'échantillon $N = 250$, on en déduit les effectifs cumulés $n_i^* = N \times p_i^*$, puis les effectifs de chaque modalité $n_i = n_i^* - n_{i-1}^*$.*

| *Modalités de X* | *Pourcentages cumulés* | *Effectifs cumulés* $n_i^*$ | *Effectifs* $n_i$ |
|---|---|---|---|
| $x_1$ | 12% | 30 | 30 |
| $x_2$ | 32% | 80 | 50 |
| $x_3$ | 32% | 80 | 0 |
| $x_4$ | 54% | 135 | 55 |
| $x_5$ | 80% | 200 | 65 |
| $x_6$ | 90% | 225 | 25 |
| $x_7$ | 96% | 240 | 15 |
| $x_8$ | 100% | 250 | 10 |

*3. Le mode (modalité ayant le plus gros effectif sur l'échantillon) est ici égal à $x_5$.*

*4. La médiane est la modalité venant immédiatement après $\dfrac{250}{2} = 125$, c'est-à-dire la 126ième donnée : c'est donc $x_4$ (car du 81-ième au 135-ième, c'est la modalité $x_4$).*

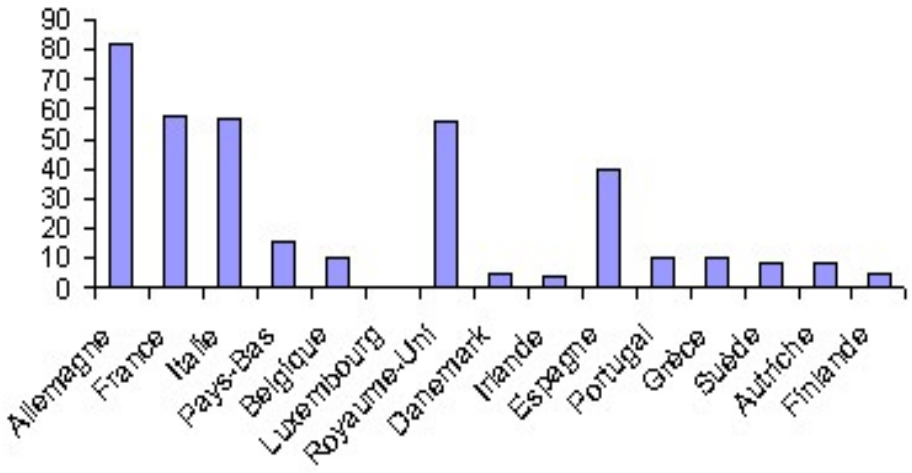



FIGURE 10.5 – *Diagramme en bandes pour la population des pays de l'UE.*

---

**Exercice 16.** *1. La population est ici l'ensemble des 225646 personnes ayant reçu une médaille au 1er janvier 1990. La variable associe à chaque personne le type de médaille obtenue. C'est une variable ordinale.*

*2. On reprend le tableau de l'énoncé en y rajoutant les effectifs cumulés :*

| *modalités* | $n_i$ | $n_i^*$ |
|---|---|---|
| *Chevaliers* | 176751 | 176751 |
| *Officiers* | 43152 | 219903 |
| *Commandeurs* | 5220 | 225123 |
| *Grands Officiers* | 459 | 225582 |
| *Grand Croix* | 64 | 225646 |

*La médiane est la modalité venant immédiatement après* $\frac{50}{2}$*, c'est-à-dire celle du 26-ième : c'est donc 4 (car du 21-ième au 26-ième, il y a exactement 4 fautes).*

*La médiane correspond à la modalité de l'individu qui est juste après le rang* $\frac{225646}{2} = 112823$*, c'est donc la 112824ème donnée Méd = Chevalier.*

*Pour le premier quartile* $\frac{225646}{4} = 56411,5$ *: on regarde donc la 56412ième donnée, soit* $Q_1 =$ *Chevalier.*

*Pour le troisième quartile,* $\frac{3 \times 225646}{4} = 169234,5$*, donc on regarde la 169235ième donnée, soit* $Q_3 =$ *Chevalier.*

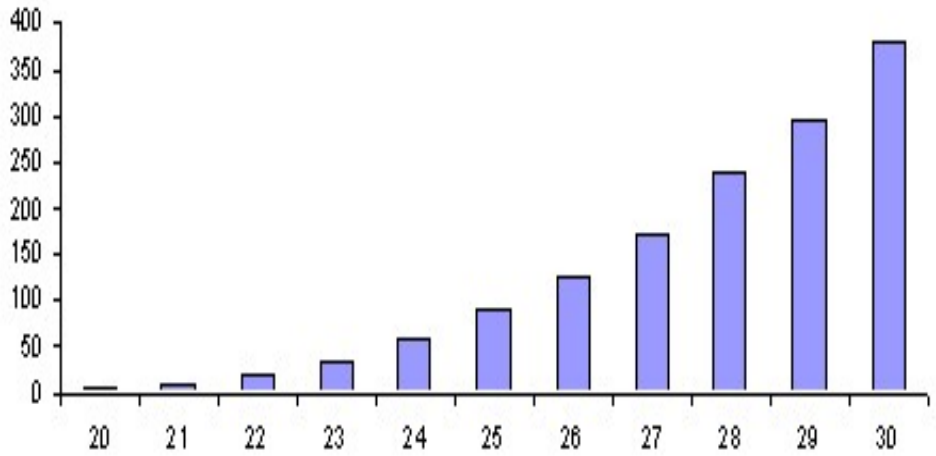



FIGURE 10.6 – *Diagramme cumulatif des veufs.*

---

**Exercice 17.** *1. La variable est l'année de parution de type quantitatif continu.*
*2. Après dépouillement, on obtient le tableau suivant :*

| *Années* | $n_i$ | $n_i^*$ | *Années* | $n_i$ | $n_i^*$ | *Années* | $n_i$ | $n_i^*$ |
|---|---|---|---|---|---|---|---|---|
| 1960 | 1 | 1 | 1970 | 2 | 18 | 1979 | 3 | 38 |
| 1961 | 2 | 3 | 1971 | 2 | 20 | 1980 | 1 | 39 |
| 1962 | 3 | 6 | 1972 | 1 | 21 | 1981 | 1 | 40 |
| 1963 | 1 | 7 | 1973 | 2 | 23 | 1982 | 2 | 42 |
| 1964 | 1 | 8 | 1974 | 2 | 25 | 1983 | 2 | 44 |
| 1965 | 4 | 12 | 1975 | 2 | 27 | 1984 | 1 | 45 |
| 1967 | 1 | 13 | 1976 | 3 | 30 | 1985 | 1 | 46 |
| 1968 | 2 | 15 | 1977 | 3 | 33 | 1986 | 3 | 49 |
| 1969 | 1 | 16 | 1978 | 2 | 35 | 1987 | 1 | 50 |

*3. Le mode est donc l'année 1965 (correspondant au plus gros effectif).*
*4.* $\frac{50}{4} = 12,5$, $\frac{50}{2} = 25$ *et* $3 * \frac{50}{4} = 37,5$ *donc pour les quartiles, on regarde les années classées 13ième, 26ième et 38ième. On a donc* $Q_1 = 1967$, $Q_2 = Méd = D_5 = 1975$ *et* $Q_3 = 1979$.

$\frac{50}{10} = 5$, $2 * \frac{50}{10} = 10$, $3 * \frac{50}{10} = 15$, *... donc pour les déciles, on regarde les années classées en position 6, 11, 16, 21, 26,...*

*Ainsi :* $D_1 = 1962$, $D_2 = 1965$, $D_3 = 1969$, $D_4 = 1972$, $D_6 = 1977$, $D_7 = 1979$, $D_8 = 1982$, $D_9 = 1985$.

---

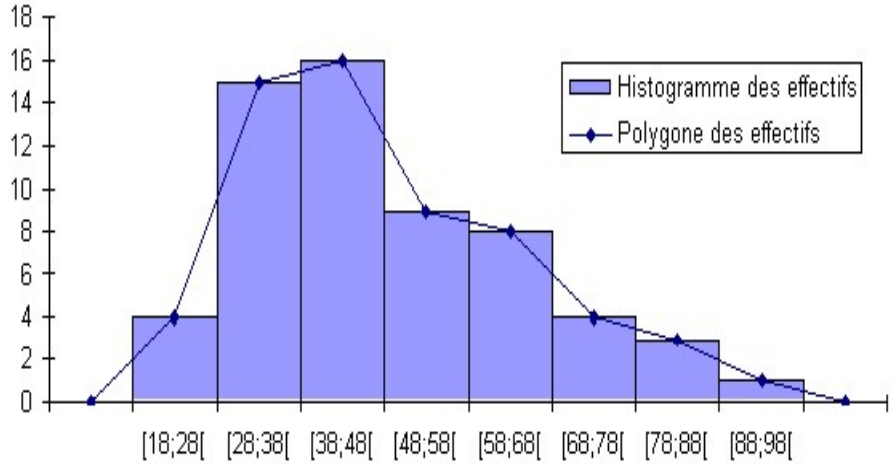



FIGURE 10.7 – *Histogramme de la variable âge pour les femmes mariées.*

**Exercice 18.** *1. On reprend les effectifs cumulés déjà calculés*

| *Âge* $(x_i)$ | *20* | *21* | *22* | *23* | *24* | *25* | *26* | *27* | *28* | *29* | *30* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Effectifs* $(n_i)$ | *4* | *7* | *11* | *13* | *23* | *32* | *37* | *45* | *70* | *56* | *82* | *380* |
| *Eff. cum*$(n_i^*)$ | *4* | *11* | *22* | *35* | *58* | *90* | *127* | *172* | *242* | *298* | *380* | |
| $n_i x_i$ | *80* | *147* | *242* | *299* | *552* | *800* | *962* | *1215* | *1960* | *1624* | *2460* | *10341* |
| $n_i(x_i)^2$ | *1600* | *3087* | *5324* | *6877* | *13248* | *20000* | *25012* | *32805* | *54880* | *47096* | *73800* | *283729* |

*puis* $N/2$*=380/2=190,* $N/4$*=380/4=95 et* $3*N/4$*=3*380/4=285.*

*La médiane est la valeur dont l'effectif cumulé est immédiatement supérieur à 190 : m=28.*

*Le 1er quartile est la valeur dont l'effectif cumulé est immédiatement supérieur à 95 : Q1=26.*

*Le 3ème quartile est la valeur dont l'effectif cumulé est immédiatement supérieur à 285 : Q3=29.*

*2. On complète le tableau précédent en ajoutant la ligne des* $n_i x_i$ *et celle des* $n_i(x_i)^2$*.*

$$Moyenne = \frac{10341}{380} \approx 27,21$$

$$Variance = \frac{283729}{380} - \left(\frac{10341}{380}\right)^2 \approx 6,10$$

$$Ecart\ type = \sqrt{Variance} \approx 2,47.$$

---

**Exercice 19.** *1. On calcule les effectifs cumulés*

| *Âge* | *[18 ;28[* | *[28 ;38[* | *[38 ;48[* | *[48 ;58[* | *[58 ;68[* | *[68 ;78[* | *[78 ;88[* | *[88 ;98[* | |
|---|---|---|---|---|---|---|---|---|---|
| *Effectifs* $(n_i)$ | *4* | *15* | *16* | *9* | *8* | *4* | *3* | *1* | *60* |
| *Eff. cum*$(n_i^*)$ | *4* | *19* | *35* | *44* | *52* | *56* | *59* | *60* | |
| $c_i$ | *23* | *33* | *43* | *53* | *63* | *73* | *83* | *93* | |
| $n_i c_i$ | *92* | *495* | *688* | *477* | *504* | *292* | *249* | *93* | *2890* |
| $n_i c_i^2$ | *2116* | *16335* | *29584* | *25281* | *31752* | *21316* | *20667* | *8649* | *155700* |

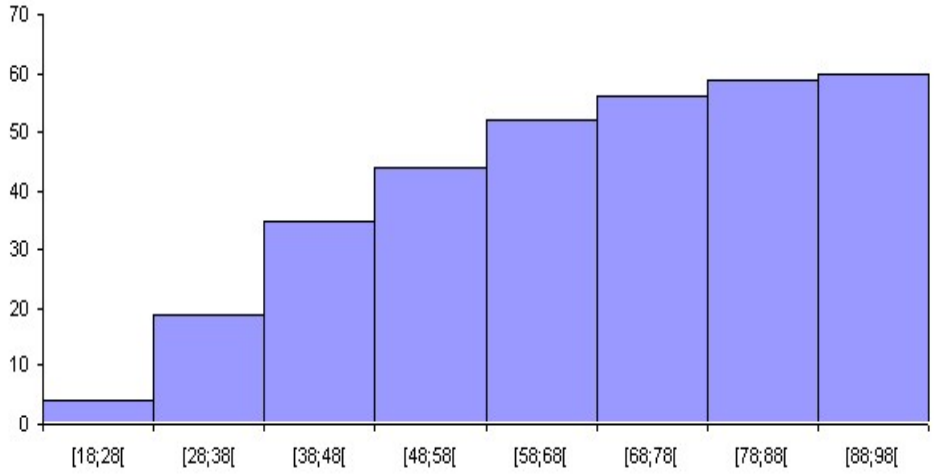



FIGURE 10.8 – *Polygone cumulatif.*

*puis* $N/2$*=60/2=30,* $N/4$*=15 et* $3 * N/4$*=45.*

*La médiane est la valeur dont l'effectif cumulé est immédiatement supérieur à 30. Donc* $m \in [38; 48[$ *et le théorème de Thalès nous donne l'approxiamtion de* $m$ *suivante*

$$m = 38 + (48 - 38)\frac{30 - 19}{35 - 19} \approx 44,875.$$

*Le 1er quartile est la valeur dont l'effectif cumulé est immédiatement supérieur à 15. Donc* $Q_1 \in [28; 38[$ *et le théorème de Thalès nous donne l'approxiamtion de* $Q_1$ *suivante*

$$Q_1 = 28 + (38 - 28)\frac{15 - 4}{19 - 4} \approx 35,33.$$

*Le 3ème quartile est la valeur dont l'effectif cumulé est immédiatement supérieur à 45. Donc* $Q_3 \in [58; 68[$ *et le théorème de Thalès nous donne l'approxiamtion de* $Q_3$ *suivante*

$$Q_3 = 58 + (68 - 58)\frac{45 - 44}{52 - 44} \approx 59,25.$$

*2. On complète le tableau précédent en ajoutant les lignes des* $c_i$*,* $n_i c_i$ *et* $n_i (c_i)^2$*.*

*Moyenne*$=\frac{2890}{60} \approx 48,17$

*Variance*$=\frac{155700}{60} - \left(\frac{2890}{60}\right)^2 \approx 274,97$

*Ecart type*$=\sqrt{Variance} \approx 16,58.$

---

**Exercice 20.** *1. La population est l'ensemble des personnes de moins de 50 ans habitant le quartier. La taille de l'échantillon est 812.*

*2. Les deux variables étudiées sont l'âge et le sexe. La première est quantitative continue et la seconde est qualitative nominale.*

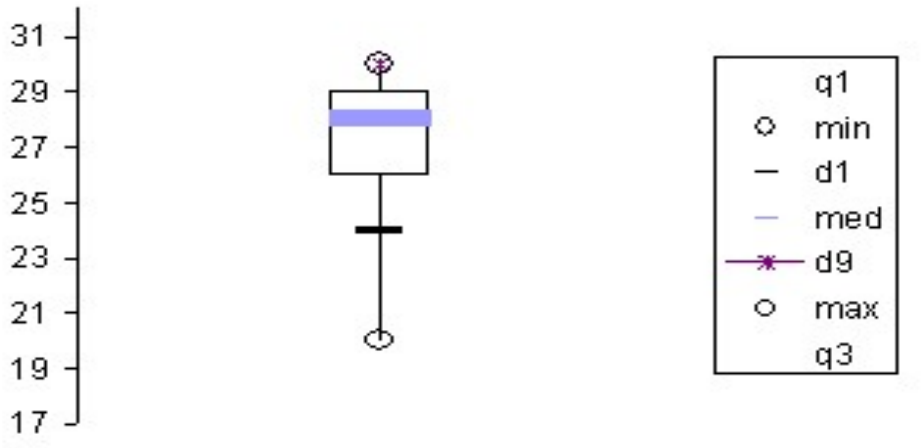



FIGURE 10.9 – *Boîte à moustaches des veufs remariés.*

*3. Complétons tout d'abord le tableau statistique en calculant les effectifs totaux (hommes+femmes), l'amplitude des classes puis les effectifs divisés par l'amplitude afin de pouvoir tracer l'histogramme.*

| *Âge* | [0; 2[ | [2; 6[ | [6; 10[ | [10; 20[ | [20; 30[ | [30; 40[ | [40; 50[ |
|---|---|---|---|---|---|---|---|
| *Amplitude* $a_i$ | *2* | *4* | *4* | *10* | *10* | *10* | *10* |
| *Effectifs (H+F)* $n_i^{HF}$ | *21* | *79* | *97* | *155* | *195* | *140* | *125* |
| $n_i^{HF}/a_i$ | *10,5* | *19,75* | *24,25* | *15,5* | *19,5* | *14* | *12,5* |
| $n_i^{HF}*$ | *21* | *100* | *197* | *352* | *547* | *687* | *812* |

*4. Pour les hommes, on obtient le tableau :*

| *Âge* | [0; 2[ | [2; 6[ | [6; 10[ | [10; 20[ | [20; 30[ | [30; 40[ | [40; 50[ |
|---|---|---|---|---|---|---|---|
| *Amplitude* $a_i$ | *2* | *4* | *4* | *10* | *10* | *10* | *10* |
| *Effectifs (H)* $n_i^H$ | *11* | *40* | *49* | *80* | *100* | *70* | *60* |
| $n_i^H/a_i$ | *5,5* | *10* | *12,25* | *8* | *10* | *7* | *6* |

*Pour les femmes, on obtient le tableau :*

| *Âge* | [0; 2[ | [2; 6[ | [6; 10[ | [10; 20[ | [20; 30[ | [30; 40[ | [40; 50[ |
|---|---|---|---|---|---|---|---|
| *Amplitude* $a_i$ | *2* | *4* | *4* | *10* | *10* | *10* | *10* |
| *Effectifs (F)* $n_i^F$ | *10* | *39* | *48* | *75* | *95* | *70* | *65* |
| $n_i^F/a_i$ | *5* | *9,75* | *12* | *7,5* | *9,5* | *7* | *6,5* |

*5.*

*6. On détermine* $N/2$*=812/2=406,* $N/4$*=812/4=203 et* $3 * N/4$*=3*812/4=609.*

*L'effectif cumulé immédiatement supérieur à 406 est 547 et correspond à la classe* $[20, 30[$*. Donc*

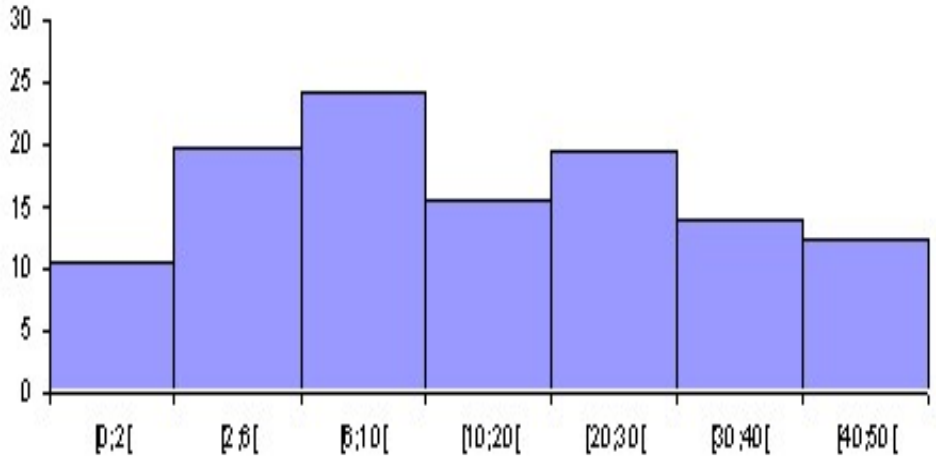



FIGURE 10.10 – *Histogramme des effectifs (hommes et femmes).*

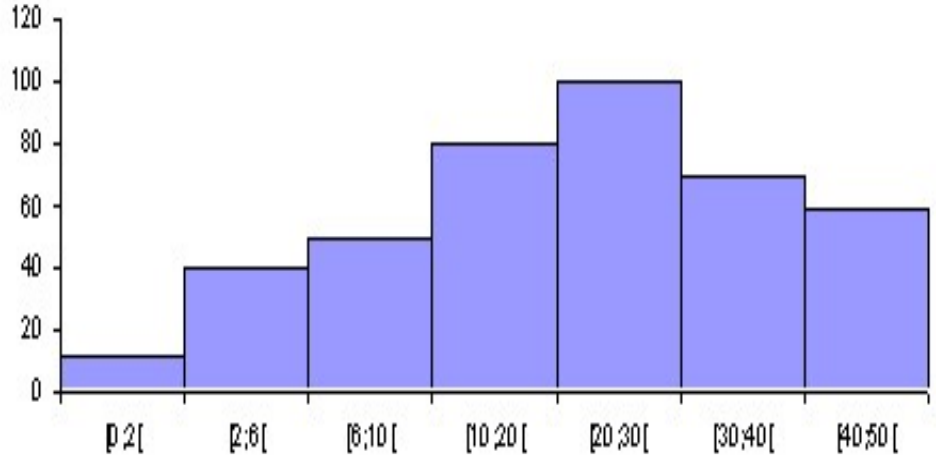



FIGURE 10.11 – *Histogramme des effectifs (hommes).*

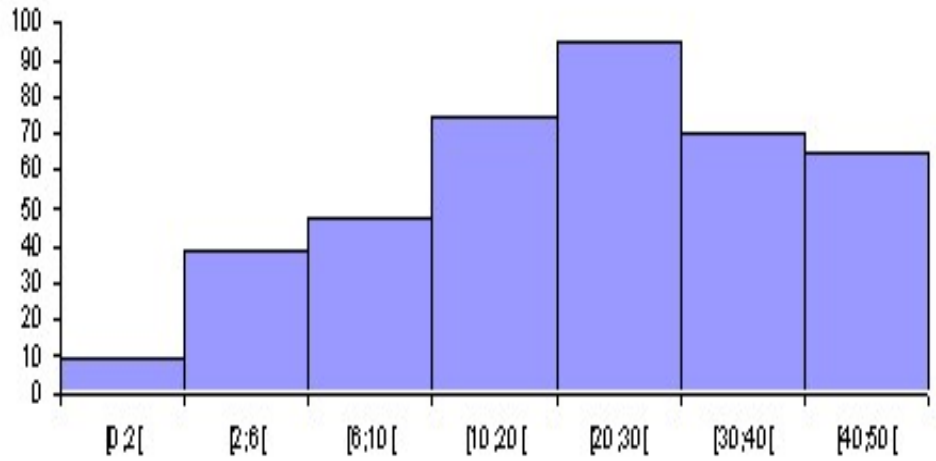



FIGURE 10.12 – *Histogramme des effectifs (femmes).*

$m \in [20, 30[$ *et plus précisépment*

$$m = 20 + (30 - 20) * \frac{406 - 352}{547 - 352} = 22,77$$

*L'effectif cumulé immédiatement supérieur à 203 est 352 et correspond à la classe* $[10, 20[$. *Donc* $Q_1 \in [10, 20[$ *et plus précisépment*

$$Q_1 = 10 + (20 - 10) * \frac{203 - 197}{352 - 197} = 10,39$$

*L'effectif cumulé immédiatement supérieur à 609 est 687 et correspond à la classe* $[30, 40[$. *Donc* $Q_3 \in [30, 40[$ *et plus précisépment*

$$Q_3 = 30 + (40 - 30) * \frac{609 - 547}{687 - 547} = 34,43$$

---

**Exercice 21.** *1. La population* $\Omega$ *est l'ensemble des ménages étudiés à Versailles. La taille est N=100.*

*2. La variable étudiée est le temps de séjour moyen. Son type est quantitatif continue.*

*3. On calcule les densités d'effectifs puisque les classes n'ont pas la même amplitude.*

| *Tps de séjour moyen* | [0; 3[ | [3; 5[ | [5; 10[ | [10; 20[ | [20; 50[ |
|---|---|---|---|---|---|
| *Effectifs* | *45* | *18* | *16* | *11* | *10* |
| *Amplitude* | *3* | *2* | *5* | *10* | *30* |
| *Densités d'effectifs* | *15* | *9* | *3,2* | *1,1* | *0,3* |
| *Eff. cum.* | *45* | *63* | *79* | *90* | *100* |

*4. Cf. tableau ci-dessus.*

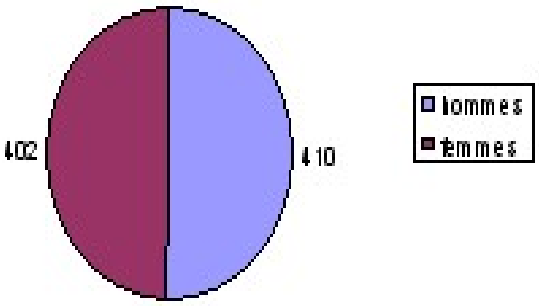



FIGURE 10.13 – *Diagramme circulaire de la variable sexe.*

*5.* $N/2 = 100/2 = 50$ *et l'effectif cumulé immédiatement supérieur à 50 est 63 et correspond à la classe* $[3;5[$*. La médiane est donc dans la classe* $[3;5[$ *et plus précisément*

$$m = 3 + (5-3) * (50-45)/(63-45) = 3,56 \quad \textit{années.}$$

*6. Graphiquement,* $Q_1 = 1,7$ *ans et* $Q_3 = 8,8$ *ans. On le lit sur le graphique du polygône cumulatif (cf. cours). Rappelons qu'ici nous ne pouvons tracer l'histogramme cumulatif puisque les classes n'ont pas la même amplitude.*

*Par le calcul, on obtient*

$$Q_1 = 0 + (3-0)\frac{25-0}{45-0} \approx 1,67 \quad \textit{ans}$$

*et*

$$Q_3 = 5 + (10-5)\frac{75-63}{79-63} \approx 8,75 \quad \textit{ans.}$$

*Ce qui signifie que 25% des couples étudiés restent en moyenne moins de 2 ans ; 75% moins de 8 ans et trois trimestres.*

---

**Exercice 22.** *L'écart-type étant nul, l'écart entre les données et la moyenne est nul. Les données sont donc toutes égales à 22. Toutes les personnes du groupe ont donc 22 ans.*

---

**Exercice 23.** *La variance étant nulle l'écart-type l'est aussi. On a donc la même conclusion que pour l'exercice précédent.*

---

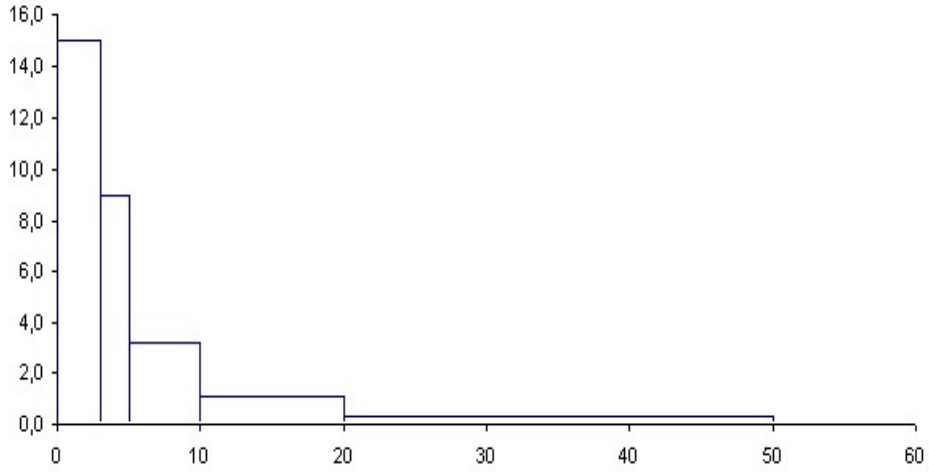



FIGURE 10.14 – *Histogramme de la variable temps de séjour.*

**Exercice 24.** *L'écart-type représente globalement l'écart entre les données de la variable et la moyenne.*

---

**Exercice 25.** *Le numéro INSEE n'est qu'un code, la variable correspondant est donc nominale et la notion de moyenne n'a donc aucun sens pour cette variable.*

---

**Exercice 26.** *1. Le mode de la série est 0.*

*2. On commence par remplir le tableau suivant pour faciliter les calculs*

| $x_j$ | *0* | *1* | *2* | *3* | *4* | *5* | *Totaux* |
|---|---|---|---|---|---|---|---|
| $n_j$ | *31* | *16* | *27* | *15* | *7* | *4* | *100* |
| $n_j^*$ | *31* | *47* | *74* | *89* | *96* | *100* | — |
| $n_j x_j$ | *0* | *16* | *54* | *45* | *28* | *20* | *163* |
| $n_j (x_j)^2$ | *0* | *16* | *108* | *135* | *112* | *100* | *471* |

*Calculons la moyenne :*

$$\overline{X} = \frac{163}{100} = 1,63.$$

*La médiane vaut 2, puisque l'effectif cumulé de 1 est 47 ($\leq 50$) et celui de 2 est 74 ($\geq 50$).*

*3. L'étendue de la série statistique est égal à $5 - 0 = 5$.*

*Le premier quartile $Q_1$ vaut 0 car l'effectif cumulé immédiatement supérieur à $\frac{50}{4} = 25$ est 31 et correspond à la modalité 0.*

*Le troisième quartile $Q_3$ vaut 3 car l'effectif cumulé immédiatement supérieur à $3 * \frac{50}{4} = 75$ est 89 et correspond à la modalité 3.*

*L'intervalle interquartile est donc $3 - 0 = 3$.*

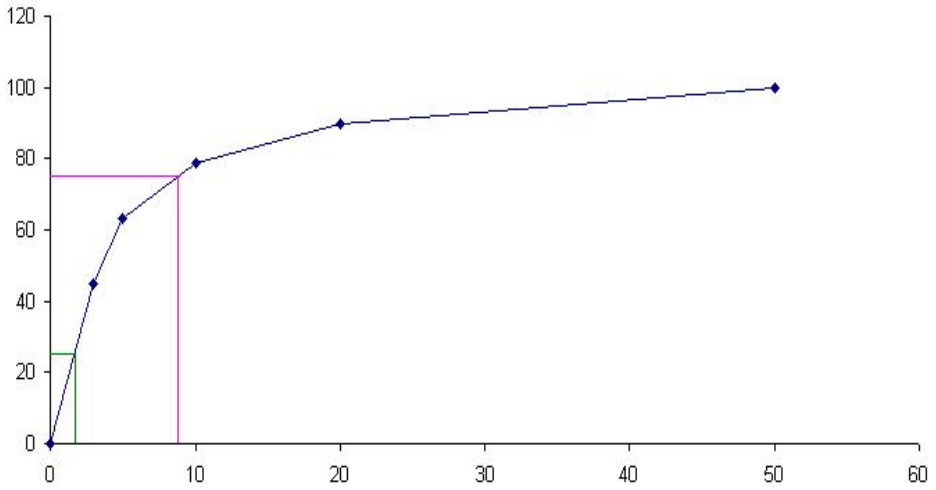



FIGURE 10.15 – *Polygône cumulatif de la variable temps de séjour.*

*4. La variance vaut :*

$$\sigma_X^2 \;=\; \frac{471}{100} - (1,63)^2 \;=\; 2,053.$$

*L'écart-type est donc :*

$$\sigma_X \;=\; \sqrt{2,053} \;\approx\; 1,43.$$

---

**Exercice 27.** *1. On peut alors donner le tableau des effectifs et des effectifs cumulés :*

| *nombre de fautes* | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 5 | 4 | 6 | 5 | 6 | 3 | 2 | 5 | 10 | 4 |
| $n_i^*$ | 5 | 9 | 15 | 20 | 26 | 29 | 31 | 36 | 46 | 50 |

*La médiane est la modalité venant immédiatement après* $\frac{50}{2}$*, c'est-à-dire la 26ième donnée : c'est donc 4 (car du 21ième au 26ième indicidu (après avoir classé les données par ordre croissant), il y a exactement 4 fautes).*

*Pour les quartiles, on a* $\frac{50}{4} = 12.5$ *et* $\frac{3 \times 50}{4} = 37.5$ *donc le premier quartile est le nombre de fautes du 13ième individu (après avoir classé les données par ordre croissant), soit* $Q_1 = 2$*, on a évidemment le deuxième quartile égal à la médiane, soit* $Q_2 = Méd = 4$ *et le troisième quartile est égal au nombre de fautes du 38ième individu (après avoir classé les données par ordre croissant), soit* $Q_3 = 8$*.*

*2. On a le tableau suivant :*

| $x_j$ | *0* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *Totaux* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_j$ | *5* | *4* | *6* | *5* | *6* | *3* | *2* | *5* | *10* | *4* | *50* |
| $x_j \times n_j$ | *0* | *4* | *12* | *15* | *24* | *15* | *12* | *35* | *80* | *36* | *233* |
| $x_j^2 \times n_j$ | *0* | *4* | *24* | *45* | *96* | *75* | *72* | *245* | *640* | *324* | *1525* |

*La moyenne vaut :*

$$\overline{X} \;=\; \frac{233}{50} \;=\; 4,66,$$

*et la variance vaut :*

$$\mathrm{Var}(X) = \sigma_X^2 \;=\; \frac{1525}{50} - \left(\frac{233}{50}\right)^2 \;\approx\; 8,78.$$

*On en déduit l'écart-type :*

$$\sigma_X \;=\; \sqrt{8,78} \;\approx\; 2,96.$$

*3. Le nombre de fautes moyen est d'environ 4,66 et l'écart-type est proche de 3. Compte tenu de l'étendue de la série (9-0=9), on peut en déduire que ce groupe d'élèves n'est pas homogène puisque la série statistique est très dispersée autour de la moyenne (l'écart-type représente environ le tiers de l'étendue du nombre de fautes).*

---

**Exercice 28.** *1. X est une variable quantitative continue car on mesure une grandeur physique. La population de taille 100 est constituée des enfants de CE2.*

*2. Compléter le tableau suivant :*

| *Classes* | $x_i$ *(centre des classes)* | $n_i$ | $n_i x_i$ | $n_i (x_i)^2$ | $a_i$ *(amplitude des classes)* | $\frac{n_i}{a_i}$ | $N_i$ *(eff. cumulés)* |
|---|---|---|---|---|---|---|---|
| [10; 30[ | *20* | *12* | *240* | *4800* | *20* | *0,6* | *12* |
| [30; 40[ | *35* | *13* | *455* | *15925* | *10* | *1,3* | *25* |
| [40; 45[ | *42,5* | *10* | *425* | *18062,5* | *5* | *2* | *35* |
| [45; 50[ | *47,5* | *15* | *712,5* | *33843,75* | *5* | *3* | *50* |
| [50; 60[ | *55* | *23* | *1265* | *69575* | *10* | *2,3* | *73* |
| [60; 70[ | *65* | *17* | *1105* | *71825* | *10* | *1,7* | *90* |
| [70; 90[ | *80* | *10* | *800* | *64000* | *20* | *0,5* | *100* |
| *TOTAL* | - | *100* | *5002,5* | *278031,25* | | | |

*3. On calcule les densités d'effectifs puisque les classes n'ont pas la même amplitude (cf. tableau ci-dessus).*

*4. La moyenne est* $\overline{X} = \dfrac{5002,5}{100} = 50,025$ *cs.*
*La variance est* $\mathrm{Var}(X) = \dfrac{278031,25}{100} - (50,025)^2 = 2780,3125 - 2502,501 \approx 277,81.$
*L'écart-type est* $\sigma(X) = \sqrt{277,81} \approx 16,67$ *cs.*

*5. Puisque* $100/2 = 50$*, la médiane correspond à la 51ème donnée, ce qui correspond à l'effectif cimulé 73 (effectif cumulé immédiatement supérieur à 50). D'où la classe médiane est* $[50; 60[$*. On peut prendre le centre de la classe pour la médiane soit 55 ou utiliser le théorème de Pythagore*

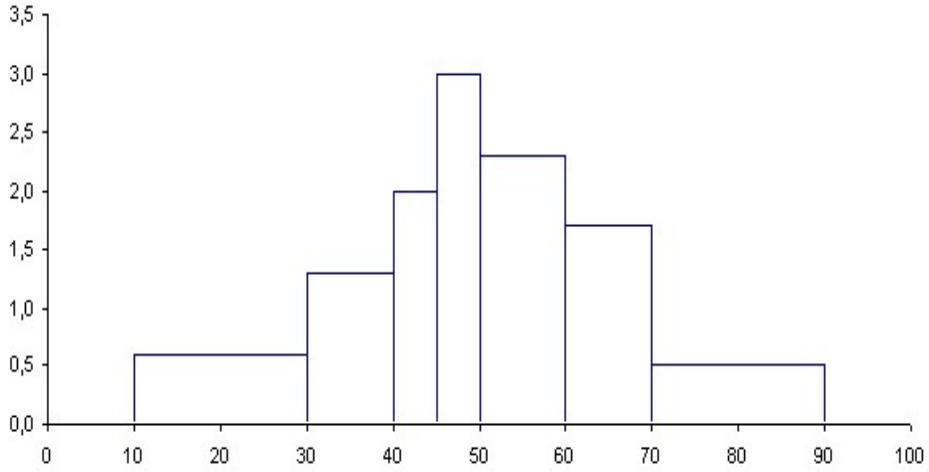



FIGURE 10.16 – *Histogramme de la variable Temps d'identification d'un mot simple.*

*pour donner une valeur plus précise :*

$$m = x_1 + (x_2 - x_1)\frac{N/2 - N_1}{N_2 - N_1} = 50 + (60 - 50)\frac{50 - 50}{73 - 50} = 50.$$

*La médiane partage l'échantillon en deux sous-échantillons de même taille :* 50% *des valeurs sont plus petites que* $m$ *et* 50% *des valeurs sont plus grandes que* $m$.

*6. On calcule les valeurs réduites et on les compare*

$$\frac{37,4 - 25,3}{14,7} = 0,82 > 0,67 = \frac{61,2 - 50,025}{16,67}.$$

*Un adulte ayant un temps de 37,4 cs est donc moins rapide qu'un enfant ayant un temps de 61,2 cs relativement à leur population d'origine.*

---

**Exercice 29.** *1. Les individus sont ici des entreprises et la variable étudiée associe à chaque entreprise le nombre annuel d'infractions à la législation du travail commises. Il s'agit d'une variable quantitative discrète (car on réalise un compatge) que l'on notera* $X$. *L'échantillon est de taille* $n = 48$.

*2. On fait un tableau permettant de déterminer à la fois la médiane et la moyenne.*

| $x_i$ | $n_i$ | $n_i^*$ | $n_i x_i$ | $x_i$ | $n_i$ | $n_i^*$ | $n_i x_i$ |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 14 | 2 | 21 | 28 |
| 1 | 1 | 2 | 1 | 15 | 4 | 25 | 60 |
| 2 | 1 | 3 | 2 | 16 | 5 | 30 | 80 |
| 3 | 1 | 4 | 3 | 17 | 2 | 32 | 34 |
| 4 | 1 | 5 | 4 | 21 | 1 | 33 | 21 |
| 5 | 3 | 8 | 15 | 22 | 3 | 36 | 66 |
| 6 | 2 | 10 | 12 | 23 | 5 | 41 | 115 |
| 7 | 2 | 12 | 14 | 24 | 3 | 44 | 72 |
| 8 | 1 | 13 | 8 | 25 | 1 | 45 | 25 |
| 11 | 1 | 14 | 11 | 26 | 1 | 46 | 26 |
| 12 | 2 | 16 | 24 | 27 | 1 | 47 | 27 |
| 13 | 3 | 19 | 39 | 28 | 1 | 48 | 28 |
| | | | | | 48 | – | 715 |

*On a alors* $\overline{X} = \frac{1}{N}\sum n_i x_i = \frac{715}{48} \approx 14,89$ *et comme* $\frac{N}{2} = \frac{48}{2} = 24$*, la médiane est donc la 25ième donnée, c'est-à-dire 15.*

*Moyenne et médiane sont ici très proches.*

*Si on se contente de citer la valeur trouvée, on donnera plutôt la moyenne (arrondie à 15) car on est plus habitué à entendre parler de moyenne que de médiane.*

*Par contre, si on explicite le résultat, on donnera plutôt la médiane dont le sens est beaucoup plus parlant, en disant : "dans la moitié des entreprises, on a observé moins de 15 infractions à la législation du travail" ; on peut également dire : "dans la moitié des entreprises, on a observé plus de 15 infractions à la législation du travail" (tout dépend de l'impression que l'on veut donner...)*

---

**Exercice 30.** *Le total des notes de l'enseignant A est donc :* $50 \times 9 = 450$

*De même, le total des notes de l'enseignant B est :* $65 \times 11 = 715$

*Le total des notes de l'enseignant C est :* $35 \times 13 = 455$.

*Le total des notes pour l'ensemble des 150 copies est donc :* $450 + 715 + 455 = 1620$.

*La moyenne sur l'ensemble des 150 copies est alors :* $\frac{1620}{150} = 10,8$.

*NB. La moyenne sur l'ensemble des 150 copies n'est pas égale à la moyenne des moyennes sur chaque groupe (c'est-à-dire 11). Ceci serait vrai si les effectifs de chaque groupe étaient identiques.*

---

**Exercice 31.** *1. Les classes ayant toutes même amplitude, il suffit de comparer les hauteurs de chaque rectangle. La classe centrée en 120 ayant pour effectif 63 et 9 "étages", on en déduit que l'effectif d'un "étage" vaut 7. Les effectifs de chacune des classes sont donc* $2 \times 7$, $4 \times 7$, $9 \times 7$,

$5 \times 7$, $3 \times 7$, *soit un effectif total de* $N = 23 \times 7$, *ce qui permet d'obtenir les fréquences en faisant* $f_i = \dfrac{n_i}{N}$ *(voir tableau).*

*2. On a donc le tableau suivant (tableau classique auquel on ajoute la colonne des fréquences pour répondre à la question 1. :*

| *classe* | $x_i$ | $n_i$ | $f_i$ | $n_i x_i$ | $n_i(x_i^2)$ |
|---|---|---|---|---|---|
| $95-105$ | 100 | 14 | 0.09 | 1400 | 140000 |
| $105-115$ | 110 | 28 | 0.17 | 3080 | 338800 |
| $115-125$ | 120 | 63 | 0.39 | 7560 | 907200 |
| $125-135$ | 130 | 35 | 0.22 | 4550 | 591500 |
| $135-145$ | 140 | 21 | 0.13 | 2940 | 411600 |
| | | $N = 161$ | 1 | 19530 | 2389100 |

*On a alors* $\overline{X} = \dfrac{1}{N}\sum n_i x_i = \dfrac{19530}{161}$ *soit* $\overline{X} \approx 121,30$.

$\mathrm{Var}(X) = \dfrac{1}{N}\sum n_i(x_i^2) - \overline{X}^2 = \dfrac{2389100}{161} - \left(\dfrac{19530}{161}\right)^2 \approx 124,39$ *et* $\sigma_X = \sqrt{\mathrm{Var}(X)}$, *soit* $\sigma_X \approx 11,15$.

---

**Exercice 32.** *Pour chaque article, il faut calculer son prix réduit : on pourra alors comparer les différents prix réduits entre eux. Si on note* $X$ *la variable "Prix d'un stylo" sur l'ensemble des librairies de Toulouse et* $x_i$ *le prix constaté, alors la variable réduite d'un stylo est* $Z$ *dont une modalité est* $z_i$ *définie par* $z_i = \dfrac{x_i - \overline{X}}{\sigma_X}$.
*Attention : il y a une variable réduite par article, donc, sur cet exemple, 3 variables réduites. On procède donc de même pour le livre et le répertoire.*

| *article* | *prix* | *prix moyen* | *écart-type du prix* | *prix réduit* |
|---|---|---|---|---|
| *stylo* | 1,9 | 1,5 | 0,4 | 1 |
| *livre* | 9 | 8,5 | 1 | 0,5 |
| *répertoire* | 3,2 | 3,5 | 0,6 | −0,5 |

*L'article le moins cher (dans sa catégorie) est celui dont le prix réduit est le plus petit. Il s'agit ici du répertoire. De même, l'article le plus cher est le stylo.*

---

**Exercice 33.** *1. La population est l'ensemble des 16 pays d'Europe.*

*2.*

| $X$ | *Effectifs* | $n_i x_i$ | $n_i (x_i)^2$ | *Effectifs cumulés* |
|---|---|---|---|---|
| 35 | 1 | 35 | 1225 | 1 |
| 42 | 3 | 126 | 5292 | 4 |
| 44 | 2 | 88 | 3872 | 6 |
| 45 | 1 | 45 | 2025 | 7 |
| 46 | 1 | 46 | 2116 | 8 |
| 47 | 1 | 47 | 2209 | 9 |
| 48 | 1 | 48 | 2304 | 10 |
| 50 | 1 | 50 | 2500 | 11 |
| 51 | 3 | 153 | 7803 | 14 |
| 54 | 1 | 54 | 2916 | 15 |
| 70 | 1 | 70 | 4900 | 16 |
| | 16 | 762 | 37162 | |

*La moyenne est donc approximativement 47,63, la variance 54,48 et l'écart-type 7,38.*

*3.* $N/2 = 8$ *donc la médiane est donc la 9ème valeur et* $m = 47$. *Cela signifie que 8 pays ont moins de 47% de femmes mariées et 8 plus de 47% de femmes mariées.*

*4. On calcule* $N/4 = 4$ *donc le premier quartile est la 5ème valeur :* $Q_1 = 44$.

*On calcule* $N/4 = 12$ *donc le troisième quartile est la 13ème valeur :* $Q_3 = 51$.

**Exercice 34.** *1. La population est l'ensemble des candidats au concours.*

*La variable est la note obtenue, c'est une variable quantitative (discrète ou continue).*

*2.*

| *classes* | $n_i$ | $f_i$(%) | $f_i^*$ *(cumulée)* | $c_i$ | $n_i.c_i$ | $n_i.(c_i)^2$ |
|---|---|---|---|---|---|---|
| [0; 5[ | *17* | *13,71* | *13,71* | *2,5* | *42,5* | *106,25* |
| [5; 8[ | *16* | *12,90* | *26,61* | *6,5* | *104* | *676* |
| [8; 10[ | *26* | *20,97* | *47,58* | *9* | *234* | *2106* |
| [10; 12[ | *30* | *24,19* | *71,77* | *11* | *330* | *3630* |
| [12; 16[ | *24* | *19,35* | *91,12* | *14* | *336* | *4704* |
| [16; 20] | *11* | *8,87* | *99,99* | *18* | *198* | *3564* |
| | *124* | *100,00* | - | - | *1244,5* | *14786,25* |

*3. Ici les classes n'étant pas de même amplitude, on ne peut pas représenter les fréquences cumulées correctement ni appliquer le théorème de Thalès. Par contre, on peut considérer que les fréquences cumulées des classes correspondent aux fréquences cumulées de la fin des classes (par exmple, 26,61 est la fréquence cumulée de la donnée 8), représenter les points (fin de classe, fréquence cumulée), les relier et lire une approximation graphique de la note minimale d'admission*

*soit environ 13 (il s'agit du troisième quartile $Q_3$). 4. Voir tableau. On en déduit les moyenne et écart-type suivants :*

$$\overline{X} = \frac{1244,5}{124} \approx 10,04 \; ; \quad \sigma = \sqrt{\frac{14786,25}{124} - \overline{X}^2} \approx \sqrt{18,44} \approx 4,29$$

---

**Exercice 35.**

| *Classes* | *Effectifs/100* | *Centres* | $n_i ci$ | $n_i(c_i^2)$ | *Eff. Cumulés/100* | *Densités d'effectifs* |
|---|---|---|---|---|---|---|
| [16; 18[ | *417* | *17* | *7089* | *120513* | *417* | *208,5* |
| [18; 25[ | *2772* | *21,5* | *59598* | *1281357* | *3189* | *396* |
| [25; 40[ | *2466* | *32,5* | *80145* | *2604713* | *5655* | *164,4* |
| [40; 50[ | *1341* | *45* | *60345* | *2715525* | *6996* | *134,1* |
| [50; 60[ | *886* | *55* | *48730* | *2680150* | *7882* | *88,6* |
| [60; 65] | *389* | *62,5* | *24312,5* | *1519531* | *8271* | *77,8* |
| | *8271* | - | *280219,5* | *10921789* | | |

*1. L'âge moyen est* $\frac{24312,5}{8271} \approx 33,88$ *ans et l'écart-type est 13,14 ans.*

*2. Pour réaliser l'histogramme, il faut calculer les densités d'effectifs pusique les classes n'ont pas la même amplitude. On lit sur le diagramme qu'il y a environ 150000 chômeurs entre 35 et 45 ans.*

**Remarque :** *On peut calculer la valeur exacte en faisant : 246600/3+134100/2=149250.*

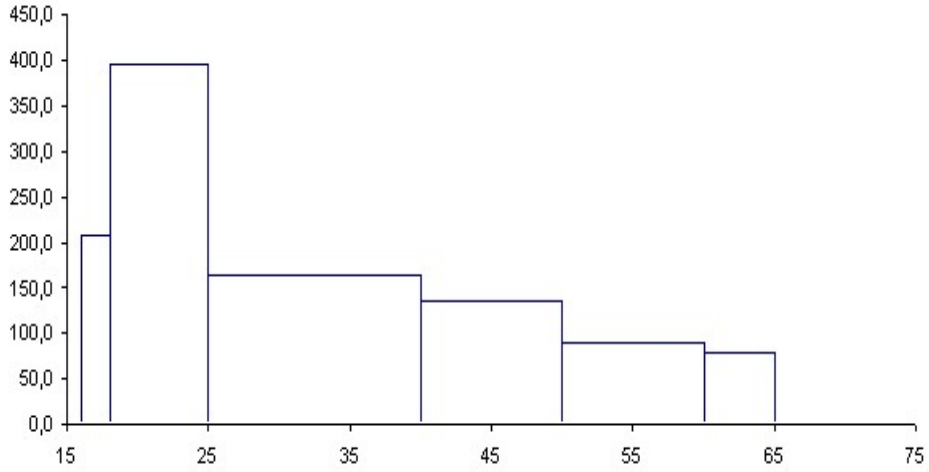



FIGURE 10.17 – *Histogramme de la variable âge.*

*3. On calcule :* $N/4 = 8271/4 = 2067,75$, $N/2 = 4135,5$ *et* $3 * N/4 = 6203,25$. *On en déduit que* $Q_1 \in [18, 25[$, $Q_2 = m \in [25, 40[et Q_3 \in [40, 50[$ *et*

$$Q_1 = 18 + (25 - 18)\frac{2067,75 - 417}{3189 - 417} \approx 22,17$$

$$Q_2 = 25 + (25-18)\frac{4135,5-3189}{5655-3189} \approx 30,76$$

$$Q_3 = 18 + (25-18)\frac{6203,25-5655}{6996-5655} \approx 44,09$$

---

**Exercice 36.** *1. La population est l'ensemble des communes de la Corrèze, la variable est le taux de boisement, de type quantitatif continu.*

*2. Environ 23%.*

*3.*

| *Taux de boisement* | [0; 20[ | [20; 25[ | [25; 35[ | [35; 40[ | [40; 70[ |
|---|---|---|---|---|---|
| *Fréquence cumulée* | *20%* | *30%* | *50%* | *65%* | *100%* |
| *Fréquence* | *20%* | *10%* | *20%* | *15%* | *35%* |

*4. La médiane est= 35%.*

*5.*

| *Taux de boisement* | [0; 20[ | [20; 25[ | [25; 35[ | [35; 40[ | [40; 70[ |
|---|---|---|---|---|---|
| *Effectif* | *57* | *29* | *57* | *43* | *100* |

*6. et 7.*

| *Centre* | *Fréquence* | $f_i c_i$ | $f_i(c_i^2)$ |
|---|---|---|---|
| *10* | *20* | *200* | *2000* |
| *22,5* | *10* | *225* | *5062,5* |
| *30* | *20* | *600* | *18000* |
| *37,5* | *15* | *562,5* | *21093,75* |
| *55* | *35* | *1925* | *105875* |

*D'où l'on déduit :*

- *Moyenne≈35,13*
- *Variance≈286,55*
- *Ecart-type ≈16,93.*

---

**Exercice 37.** *1. $\Omega$ : la population française au 1er janvier 1994.*

*Variable $X$ : l'âge, variable quantitative continue regroupée en classes d'amplitudes 1 an.*

*Variable $Y$ : le sexe, variable nominale.*

*2. La pyramide des âges représente la distribution conjointe du couple $(X, Y)$.*

*3. Vu au 1.*

---

**Exercice 38.** *1. Population : les femmes en France en 1982.*

*Variables :*

*- "Activité" avec les modalités oui et non, on donne ici directement les effectifs de femmes actives avec les taux d'activité (pourcentage de femmes actives parmi les femmes). Variable qualitative nominale.*

*- "Nombre d'enfants", quantitative discrète.*

*- "Âge", quantitative continue.*

*Bien que le tableau permette de retrouver la distribution conjointe pour ces trois variables, on ne s'intéresse ici qu'aux femmes actives (à la distribution des variables "Âge" et "Nombre d'enfants" conditionnellement à la modalité oui de la variable "Activité") et on peut donc aussi considérer que la population étudiée est l'ensemble des femmes actives, le taux d'activité étant une information annexe.*

*2. 396,4 est le nombre (en milliers) de femmes actives avec 2 enfants âgées entre 35 et 39 ans.*

*77,9 est le pourcentage de femmes actives parmi les femmes sans enfant entre 35 et 39 ans.*

*34,9 est le nombre (en milliers) de femmes actives entre 15 et 19 ans.*

*3. 2850,2 est le nombre (en milliers) de femmes actives sans enfant.*

*41,6% est le taux d'activité des femmes sans enfant.*

*Le nombre 2850,2 est simplement la somme de la ligne 1.*

*Pour retrouver 41,6%, il faut d'abord calculer pour chaque classe d'âge le nombre de femmes sans enfant, par exemple, le nombre de femmes sans enfant qui ont entre 35 et 39 ans est :* $\frac{121,6 \times 100}{77,9} = 156,1 milliers.$ *Puis on calcule le nombre total de femmes sans enfant en faisant la somme :*

$$\frac{28,8 \times 100}{69} + \ldots + \frac{640 \times 100}{17,9} = 6851,5.$$

*Le taux d'activité (en %) est donc :* $\frac{2850,2}{6851,5} \times 100 = 41,6.$

*4. Prenons* $[55;60]$ *pour la dernière classe. La moyenne calculée en prenant les centres des classes est :*

$$\frac{28,8 \times 17,5 + 399,9 \times 22,5 + \ldots + 57,5 \times 640}{2850,2} = 43,1 ans.$$

*5. L'âge moyen des femmes actives est 38,9 ans.*

| *Effectifs cumulés* | *34,9* | *665,4* | *1710,7* | *2834,6* | *3758,4* | *4455,8* | *5143,7* | *5779,4* | *6447,5* |
|---|---|---|---|---|---|---|---|---|---|

*On a* $\frac{6447,5}{2} = 3223,75$. *La médiane m est dans la classe* $[35;40[$ *:*

$$m = 35 + (40 - 35)\frac{3223,75 - 2834,6}{3758,4 - 2834,6} \approx 37,1.$$

*L'âge médian des femmes actives est 37,1 ans.*

*6.*

| | *0 enfant* | *1 enfant* | *2 enfants* | *3 enfants et +* |
|---|---|---|---|---|
| *Fréquences* | *44,20%* | *29%* | *20,90%* | *5,90%* |

*7.*

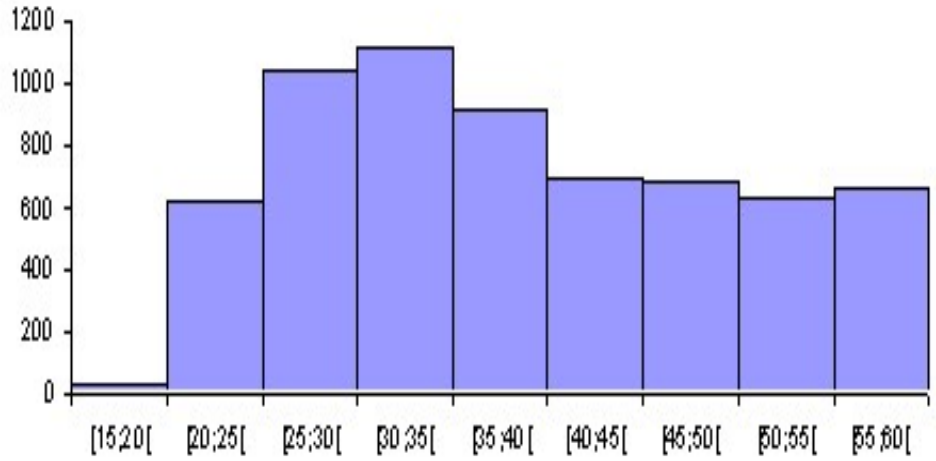



FIGURE 10.18 – *Histogramme de la variable âge (exercice tableau complexe).*

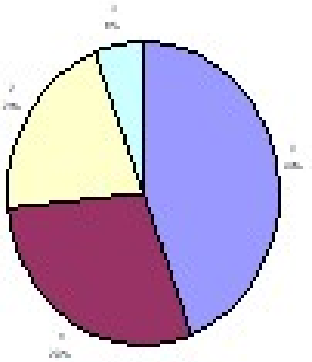


FIGURE 10.19 – *Diagramme circulaire du nombre d'enfants (exercice tableau complexe).*

| | *0 enfant* | *1 enfant* | *2 enfants* | *3 enfants et +* | *Total* |
|---|---|---|---|---|---|
| *Effectifs* | *6851,4* | *2830,5* | *2339,6* | *1217,7* | *13239,2* |
| *Fréquences* | *0,52* | *0,21* | *0,18* | *0,09* | *1,00* |

**Exercice 39.** *1. La population $\Omega$ est l'ensemble des villes du monde de plus de 10 millions d'habitants en 2000. Sa taille est 18.*

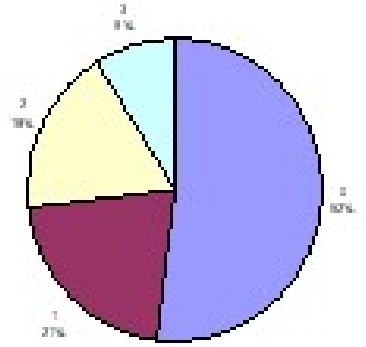


FIGURE 10.20 – *Diagramme circulaire du nombre d'enfants sur toute la population (exercice tableau complexe).*

*La variable est le nombre d'habitants (en millions), variable quantitative discrète (comptage).*
*2.*

| *Population (en millions d'habitants)* | *Effectifs* $n_i$ | *Centres* $c_i$ | $n_i c_i$ | $n_i(c_i)^2$ | *Effectifs cumulés* $n_i^*$ |
|---|---|---|---|---|---|
| [10; 12[ | *7* | *11* | *77* | *847* | *7* |
| [12; 14[ | *6* | *13* | *78* | *1014* | *13* |
| [14; 17[ | *2* | *15.5* | *31* | *480.5* | *15* |
| [17; 27[ | *3* | *22* | *66* | *1452* | *18* |
| | *18* | - | *252* | *3793,5* | |

*3. Moyenne* $= \dfrac{252}{18} = 14$.
*Ecart type* $= \dfrac{3793,5}{18} - \left(\dfrac{252}{18}\right)^2 \approx 3,84$.
*4. cf. tableau.*
*5. D'après le tableau des effectifs cumulés (voir ci-dessus), la médiane* $m$ *est entre 12 et 14 et on a :*

$$m = 12 + (14 - 12) * (8 - 7)/(13 - 7) = 12,67 \quad millions$$

*On lit approximativement* $Q_1 = 11,3$ *et* $Q_3 = 14,5$.

---

**Exercice 40.** *1. La population* $\Omega$ *est l'ensemble des usines et entreprises du district de Poplar employant plus de 100 salariés, en 1898.*
*2.* $X$ *: Nombre de salariés, variable quantitative discrète.*
$Y$ *: Localités, variable qualitative nominale.*
*3.*

| *Localités* / *Nombre de salariés* | *BOW* | *BROMLEY* | *POPLAR* | $n_i$ | $c_i$ | $n_i c_i$ | $n_i (c_i)^2$ | $n_i^*$ |
|---|---|---|---|---|---|---|---|---|
| [100; 200[ | *6* | *3* | *15* | *24* | *150* | *3600* | *540000* | *24* |
| [200; 500[ | *4* | *4* | *8* | *16* | *350* | *5600* | *1960000* | *40* |
| [500; 1000[ | *1* | *3* | *2* | *6* | *750* | *4500* | *3375000* | *46* |
| [1000; 5000[ | *2* | *1* | *1* | *4* | *3000* | *12000* | *36000000* | *50* |
| | *13* | *11* | *26* | *50* | - | *25700* | *41875000* | |

*La taille de la population est donc 50.*

*4. 13/50=26%.*

*5. 4/50=8%.*

*6. 15/24=62,5%.*

*7. D'après le tableau des effectifs cumulés (voir ci-dessus) et puisque $N/2 = 9$, la médiane $m$ est entre 200 et 500 et on a :*

$$m = 200 + (500 - 200)\frac{25 - 24}{40 - 24} \approx 218,75$$

*Moyenne* $= \dfrac{25700}{50}$ *=514.*

*Ecart type* $= \sqrt{\dfrac{41875000}{50} - \left(\dfrac{25700}{50}\right)^2}$ *≈757,17.*

---

## 10.3 Correction des exercices du Chapitre 4

**Exercice 41.** *1. Population : ensemble des élèves de troisième en 1963. Échantillon de taille 2000. Deux variables :*

→ *sexe : qualitative nominale, 2 modalités.*

→ *choix professionnel : qualitative nominale, 9 modalités.*

*2. a) dans la 3ième colonne de nombres, 200 représente le nombre d'enfants de l'échantillon ayant choisi "instituteur".*

*b) dans la 4ième colonne de nombres, 6,1 représente le pourcentage de garçons ayant choisi "instituteur" ou bien le pourcentage de garçon ayant choisi "cadre moyen".*

*c) dans la 5ième colonne de nombres, 15,3 représente le pourcentages de filles ayant choisi "employé".*

*d) dans la 6ième colonne de nombres, 11,8 est le pourcentage d'enfants ayant choisi "technicien".*

*e) dans la 8ième colonne de nombres, 75,6 est le pourcentage de filles parmi les élèves ayant choisi "employé".*

*f) dans la ligne "total", 1010 est le nombre total de filles de l'échantillon et 50,5 représente le pourcentage de filles dans l'échantillon.*

*g) dans la 4ième colonne de nombres : la somme des deux nombres 13,1 et 0,5 n'a pas de sens : 13,1 est un pourcentage de garçons, alors que 0,5 est un pourcentage de filles (les deux pourcentages portent sur des ensembles d'individus disjoints (et même disjoints).*

*h)* $3,0 + 13,1 = 16,1$ *: c'est le pourcentage des garçons ayant choisi d'être "médecin" ou "ingénieur".*

*3. a) Il y a 200 enfants voulant être instituteurs.*

*b)* $2,5\%$ *des enfants veulent être ouvriers qualifiés.*

*c) Parmi les enfants voulant être cadre moyen,* $31,6\%$ *sont des garçons.*

*d) Le tableau de fréquences du sexe conditionnellement au choix professionnel est constitué des 3 dernières colonnes.*

*e) Le tableau de fréquences du choix professionnel conditionnellement au sexe est constitué des 3 colonnes du milieu.*

---

**Exercice 42.** *Il s'agit de la distribution conjointe de* $X$ *et* $Y$*.*

---

**Exercice 43.** *Pour calculer les distributions conditionnelles de l'âge conditionnellement au sexe, il nous faut les totaux par sexe. On a alors :*

| Âge / Sexe | *moins de 20 ans* | *20 ans et plus* | *TOTAL* |
|---|---|---|---|
| *masculin* | 3 | 2 | 5 |
| *féminin* | 7 | 3 | 10 |

*On a alors le tableau de distribution de la variable Âge conditionnellement à la variable Sexe (profil ligne) :*

| Âge / Sexe | *moins de 20 ans* | *20 ans et plus* | *TOTAL* |
|---|---|---|---|
| *masculin* | 60% | 40% | 100% |
| *féminin* | 70% | 30% | 100% |

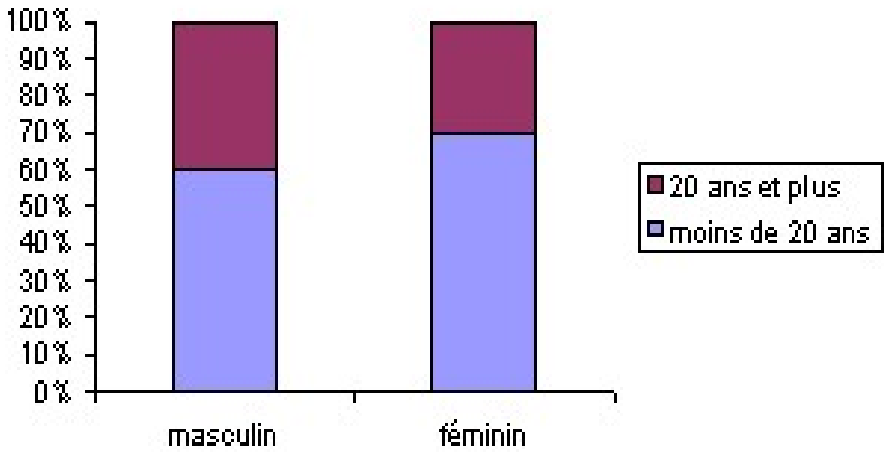



FIGURE 10.21 – *Distribution de l'Âge conditionnellement au Sexe.*

---

**Exercice 44.** *Pour calculer les distributions conditionnelles du sexe conditionnellement à l'âge, il nous faut les totaux des âges. On a alors :*

| Âge / Sexe | *moins de 20 ans* | *20 ans et plus* |
|---|---|---|
| *masculin* | 3 | 1 |
| *féminin* | 7 | 4 |
| *TOTAL* | 10 | 15 |

*On a alors le tableau de distribution de la variable Sexe conditionnellement à la variable Âge (profil colonne) :*

| Sexe \ Âge | *moins de 20 ans* | *20 ans et plus* |
|---|---|---|
| *masculin* | 30% | 20% |
| *féminin* | 70% | 80% |
| *TOTAL* | 100% | 100% |

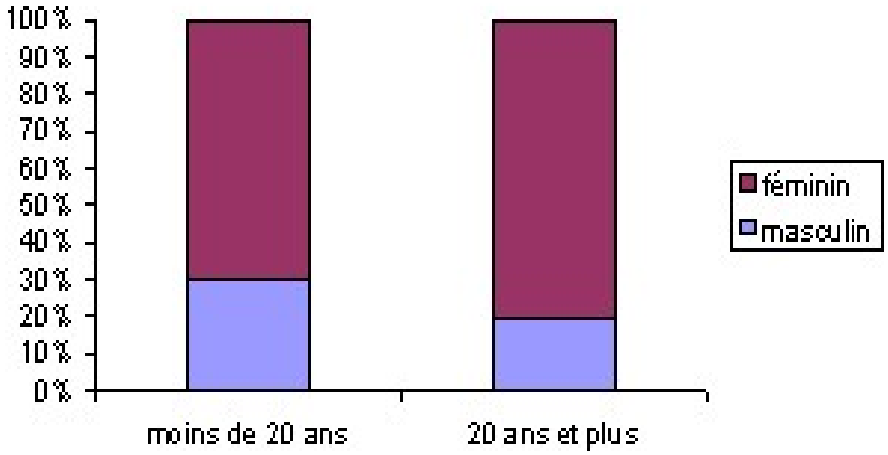



FIGURE 10.22 – *Distribution du Sexe conditionnellement à l'Âge.*

**Exercice 45.** *Parmi les célibataires de l'échantillon* 25% *sont des hommes.*

**Exercice 46.** *La fréquence de* $x_2$ *conditionnellement à* $y_3$ *est égale à* $1/3$.

**Exercice 47.** *Il s'agit du tableau de fréquences conjointes de* $X$ *et* $Y$.

**Exercice 48.** *On a le tableau ci-dessous :*

| $X$ \ $Y$ | $y_1$ | $y_2$ | $y_3$ | *Total* |
|---|---|---|---|---|
| $x_1$ | 70 | 80 | *50* | 200 |
| $x2$ | *35* | 40 | 25 | *100* |
| $x3$ | *35* | *40* | *25* | 100 |
| *Total* | *140* | 160 | 100 | 400 |

**Exercice 49.** *1. T est une variable nominale.*

*2. La pointure est une variable ordinale.*

*3. Le calcul de $p_1 + p_2$ n'a aucun sens.*

*4. le calcul de $p_2 + p_3$ représente l'ensemble des personnes habitant un numéro impair dans l'échantillon.*

*5. $p_1$ est une des valeurs de la distribution du sexe conditionnellement à la parité dans l'échantillon.*

---

**Exercice 50.** *1. Pour cet exercice, la population est constituée des personnes blessées dans un accident en 1998 dans la Haute-Garonne, l'accident ayant été communiqué à "la Dépêche".*

*Le tableau permet de distinguer 2 variables. On peut prendre, par exemple :*

*→ X "moment où l'accident a eu lieu" : variable nominale dont les 2 modalités sont "jour" et "nuit"*

*→ Y "état de l'accidenté" : variable ordinale dont les 3 modalités sont "blessé léger", "blessé grave" et "tué".*

*2. L'histogramme fourni par l'énoncé permet de construire la table de contingence (ou tableau des effectifs croisés) :*

| X \ Y | *bl. léger* | *bl. grave* | *tué* | *total* |
|---|---|---|---|---|
| *jour* | 482 | 232 | 79 | 793 |
| *nuit* | 253 | 137 | 51 | 441 |
| *total* | 735 | 369 | 130 | 1234 |

*3. À partir de la table de contingence, on peut déterminer les deux types de distributions conditionnelles :*

*→ Distribution de X conditionnelle à Y : on détermine, pour une gravité donné, comment sont répartis les moments des accidents (dernière colonne facultative) :*

| X \ Y | *bl. léger* | *bl. grave* | *tué* |
|---|---|---|---|
| *jour* | 66% | 63% | 61% |
| *nuit* | 34% | 37% | 39% |
| *total* | 100% | 100% | 100% |

*→ Distribution de Y conditionnelle à X : on détermine, pour un moment donné, comment sont répartis les états des accidentés (dernière ligne facultative) :*

| $X$ \ $Y$ | *bl. léger* | *bl. grave* | *tué* | *total* |
|---|---|---|---|---|
| *jour* | 61% | 29% | 10% | 100% |
| *nuit* | 57% | 31% | 12% | 100% |

*4. Le journaliste a écrit "Gravité plus importante la nuit" car en regardant la distribution de $X$ conditionnelle à $Y$, plus l'état est grave, plus le pourcentage d'accidents nocturnes correspondant est important. De même, en regardant la distribution de $Y$ conditionnelle à $X$, le pourcentage de tués, par exemple, est plus important la nuit (12%) que le jour ( 10%).*

---

**Exercice 51.** *1. On a 2 modalités pour le sexe, 3 pour le concours et 2 pour le résultat, ce qui fait en tout $2 \times 3 \times 2 = 12$ modalités pour le triplet de variables.*

*2. On donne d'abord la distribution conjointe :*

| *Sexe* \ *Concours* | $A$ | $B$ | $C$ | $TOTAL$ |
|---|---|---|---|---|
| $G$ | 53 | 10 | 12 | 75 |
| $F$ | 4 | 40 | 36 | 80 |
| $TOTAL$ | 57 | 50 | 48 | 155 |

*puis, les distributions marginales (dernière colonne pour l'admission, dernière ligne pour le sexe, dans le tableau qui suit, et dernière ligne pour le concours dans le tableau suivant) :*

| *Admission* \ *Sexe* | $G$ | $F$ | $TOTAL$ |
|---|---|---|---|
| *oui* | 30 | 60 | 90 |
| *non* | 45 | 20 | 65 |
| $TOTAL$ | 75 | 80 | 155 |

| *Admission* \ *Concours* | $A$ | $B$ | $C$ | $TOTAL$ |
|---|---|---|---|---|
| *oui* | 10 | 45 | 35 | 90 |
| *non* | 47 | 5 | 13 | 65 |
| $TOTAL$ | 57 | 50 | 48 | 155 |

*3. Pour le groupe des garçons :*

| *Admission* \ *Concours* | $A$ | $B$ | $C$ |
|---|---|---|---|
| *oui* | 10 | 10 | 10 |
| *non* | 43 | 0 | 2 |
| $TOTAL$ | 57 | 10 | 12 |

*Donc la distribution de la variable Admission conditionnellement à la variable Concours s'obtient par le tableau suivant (profil colonne) :*

| *Admission* \ *Concours* | *A* | *B* | *C* |
|---|---|---|---|
| *oui* | 18,9% | 100% | 83,3% |
| *non* | 81,1% | 0% | 16,7% |
| *TOTAL* | 100% | 100% | 100% |

*4. De même, pour le groupe des filles :*

| *Admission* \ *Concours* | *A* | *B* | *C* |
|---|---|---|---|
| *oui* | 0 | 35 | 25 |
| *non* | 4 | 5 | 11 |
| *TOTAL* | 4 | 5 | 11 |

*Donc la distribution de la variable Admission conditionnellement à la variable Concours s'obtient par le tableau suivant (profil colonne) :*

| *Admission* \ *Concours* | *A* | *B* | *C* |
|---|---|---|---|
| *oui* | 0% | 87,5% | 69,44% |
| *non* | 100% | 12,5% | 30,56% |
| *TOTAL* | 100% | 100% | 100% |

*5. Pour les trois concours, les garçons réussissent mieux.*

*Globalement, le pourcentage de réussite des filles est* 75% *(*60/80*) et celui des garçons est* 40% *(*30/75*).*

*6. Les résultats ne sont pas contradictoires car le concours A est beaucoup plus difficile que le B et beaucoup de garçons passent le concours A, alors que beaucoup de filles passent le concours B.*

---

**Exercice 52.** *1.*

*2. Pour calculer les distributions conditionnelles de l'âge conditionnellement au sexe, il nous faut les totaux par sexe. On a alors :*

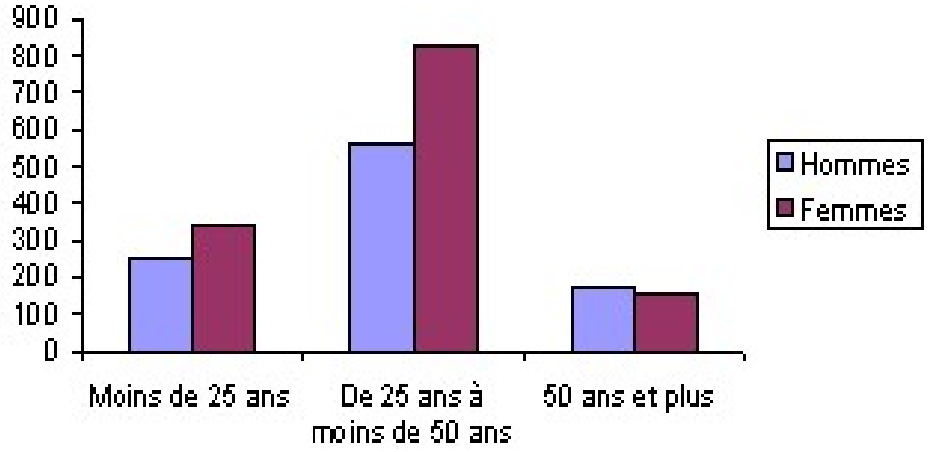



FIGURE 10.23 – *Distribution conjointe du Sexe et de l'Âge.*

| | *Hommes* | *Femmes* | *Marge âge* |
|---|---|---|---|
| *Moins de 25 ans* | *249* | *342,6* | *591,6* |
| *De 25 à moins de 50 ans* | *565,1* | *827,9* | *1393* |
| *50 ans et plus* | *168,5* | *155,2* | *323,7* |
| *Marge sexe* | *982,6* | *1325,7* | *2308,3* |

*On a alors le tableau de distribution de la variable Âge conditionnellement à la variable Sexe (profil ligne) :*

| | *Hommes* | *Femmes* |
|---|---|---|
| *Moins de 25 ans* | *25,34%* | *25,84%* |
| *De 25 à moins de 50 ans* | *57,51%* | *62,45%* |
| *50 ans et plus* | *17,15%* | *11,71%* |
| *Marge sexe* | *100%* | *100%* |

*3. Pour calculer les distributions conditionnelles du sexe conditionnellement à l'âge, il nous faut les totaux des âges. Cf. tableau précédent. On a alors :*

| | *Hommes* | *Femmes* | *Marge âge* |
|---|---|---|---|
| *Moins de 25 ans* | *42,09%* | *57,91%* | *100%* |
| *De 25 à moins de 50 ans* | *40,57%* | *59,43%* | *100%* |
| *50 ans et plus* | *52,05%* | *47,95%* | *100%* |

*On a alors le tableau de distribution de la variable Sexe conditionnellement à la variable Âge*

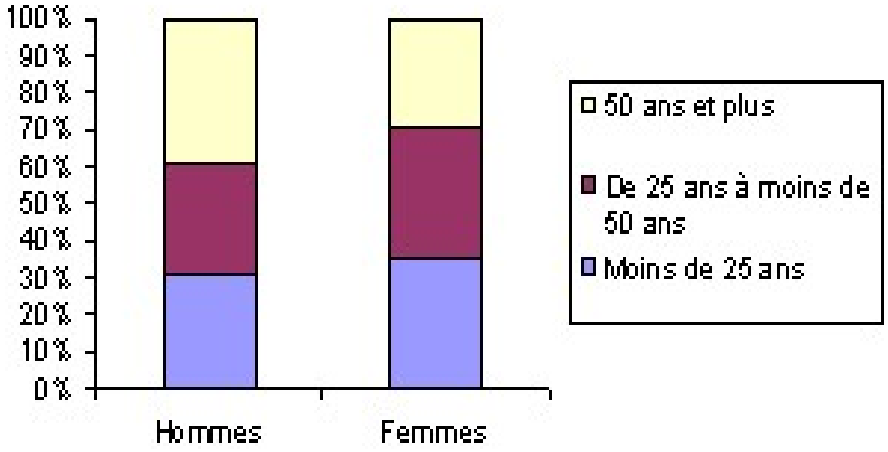



FIGURE 10.24 – *Distribution de l'Âge conditionnellement au Sexe.*

*(profil colonne) :*

| *Âge* / *Sexe* | *moins de 20 ans* | *20 ans et plus* |
|---|---|---|
| *masculin* | 30% | 20% |
| *féminin* | 70% | 80% |
| *TOTAL* | 100% | 100% |

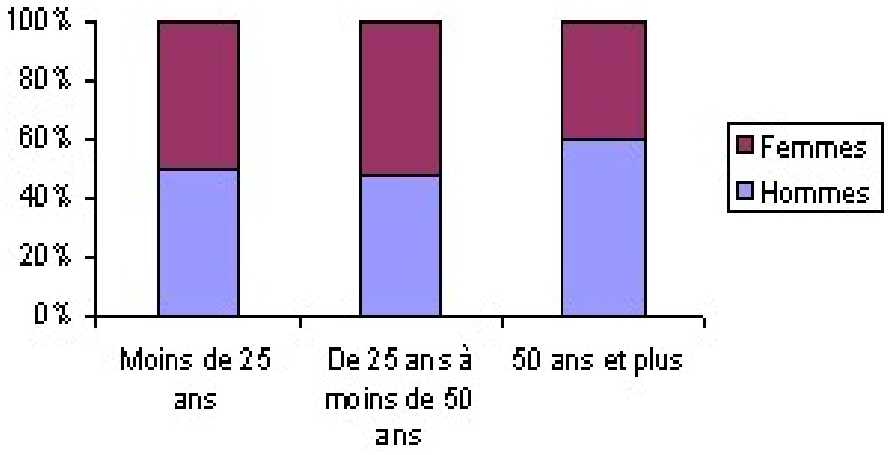



FIGURE 10.25 – *Distribution du Sexe conditionnellement à l'Âge.*

## 10.4 Correction des exercices du Chapitre 5

**Exercice 53.** *On est en présence de 2 variables nominales; X choix professionnel dont les modalités sont regroupées en 3 classes : $x_1$ (médecin, ingénieur, cadre supérieur), $x_2$ (instituteur, technicien, cadre moyen), $x_3$ (employé, ouvrier qualifié, autre choix); Y sexe à deux modalités G et F. À partir de l'exercice 6 du chapitre 1, on construit le tableau des effectifs observés, puis on calcule les effectifs théoriques et le coefficient $\chi^2$ :*

| *Tableau des $O_{ij}$* | *G* | *F* | *Total* |
|---|---|---|---|
| $x_1$ | 180 | 40 | 220 |
| $x_2$ | 320 | 305 | 625 |
| $x_3$ | 490 | 665 | 1155 |
| *Total* | 990 | 1010 | 2000 |

| *Tableau des $T_{ij}$* | *G* | *F* | *Total* |
|---|---|---|---|
| $x_1$ | $108,90$ | $111,10$ | 220 |
| $x_2$ | $309,38$ | $315,63$ | 625 |
| $x_3$ | $571,73$ | $583,28$ | 1155 |
| *Total* | 990 | 1010 | 2000 |

| *Tableau des $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$* | *G* | *F* |
|---|---|---|
| $x_1$ | $46,42$ | $45,50$ |
| $x_2$ | $0,36$ | $0,36$ |
| $x_3$ | $11,68$ | $11,45$ |

*En faisant la somme, on obtient $\chi^2 = 115,78$ et $\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{115.,78}{2000}} \approx 0,24$ : lien plutôt faible.*

---

**Exercice 54.** *1. Les trois colonnes ont les mêmes valeurs : les fréquences de $x_1$ conditionnellement à chacune des modalités de Y sont identiques et ne dépendent donc pas de la modalité de Y. Il en est de même pour $x_2$. Les variables X et Y sont donc indépendantes.*

*2. La valeur du $\chi^2$ est donc 0.*

---

**Exercice 55.** *1.*

| *Y* / *X* | $y_1$ | $y_2$ | $y_3$ | *Total* |
|---|---|---|---|---|
| $x_1$ | $\frac{12\times 5}{30} = 2$ | $\frac{12\times 10}{30} = 4$ | $\frac{12\times 15}{30} = 6$ | *12* |
| $x_2$ | $\frac{18\times 5}{30} = 3$ | $\frac{18\times 10}{30} = 6$ | $\frac{18\times 15}{30} = 9$ | *18* |
| *Total* | *5* | *10* | *15* | *30* |

*2. On constate que les effectifs théoriques sont identiques aux effectifs observés : on conclut que le coefficient $\chi^2$ est nul (car tous les $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$ sont nuls).*

---

**Exercice 56.** *1. Pour les moins de 30 ans :*

| *Tableau des $O_{ij}$* | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* | *Total* |
|---|---|---|---|---|---|---|
| *Moins de 20 ans* | 50 | 140 | 20 | 140 | 150 | 500 |
| *De 20 à moins de 30 ans* | 80 | 150 | 50 | 170 | 250 | 700 |
| *Total* | 130 | 290 | 70 | 310 | 400 | 1200 |

| *Tableau des $T_{ij}$* | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* | *Total* |
|---|---|---|---|---|---|---|
| *Moins de 20 ans* | 54 | 121 | 29 | 129 | 167 | 500 |
| *De 20 à moins de 30 ans* | 76 | 169 | 41 | 181 | 233 | 700 |
| *Total* | 130 | 290 | 70 | 310 | 400 | 1200 |

| *Tableau des $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$* | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* |
|---|---|---|---|---|---|
| *Moins de 20 ans* | 0,32 | 3,04 | 2,88 | 0,91 | 1,67 |
| *De 20 à moins de 30 ans* | 0,23 | 2,17 | 2,06 | 0,65 | 1,19 |

*En faisant la somme, on obtient* $\chi^2 = 15,11$ *et* $\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{15,11}{1200}} \approx 0,11$ *: lien faible.*
*Pour les données de ce tableau, on peut admettre qu'il n'y a presque pas de lien entre l'âge et le sport pratiqué pour les moins de 30 ans.*

*2. Pour les plus de 30 ans :*

| *Tableau des $O_{ij}$* | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* | *Total* |
|---|---|---|---|---|---|---|
| *De 30 à moins de 40 ans* | 80 | 50 | 70 | 100 | 200 | 500 |
| *40 ans et plus* | 30 | 20 | 60 | 90 | 100 | 300 |
| *Total* | 110 | 70 | 130 | 190 | 300 | 800 |

| *Tableau des $T_{ij}$* | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* | *Total* |
|---|---|---|---|---|---|---|
| *De 30 à moins de 40 ans* | 69 | 44 | 81 | 119 | 187 | 500 |
| *40 ans et plus* | 41 | 26 | 49 | 71 | 113 | 300 |
| *Total* | 110 | 70 | 130 | 190 | 300 | 800 |

| *Tableau des $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$* | *Équitation* | *Football* | *Golf* | *Natation* | *Tennis* |
|---|---|---|---|---|---|
| *De 30 à moins de 40 ans* | 1,84 | 0,89 | 1,56 | 2,96 | 0,83 |
| *40 ans et plus* | 3,07 | 1,49 | 2,60 | 4,93 | 1,39 |

*En faisant la somme, on obtient* $\chi^2 = 21,56$ *et* $\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{21,56}{800}} \approx$*0,16 : lien faible (un peu plus fort que dans a) toutefois).*

*Pour les données de ce tableau, on peut admettre qu'il n'y a presque pas de lien entre l'âge et le sport pratiqué pour les plus de 30 ans.*

---

**Exercice 57.** *La population est ici l'ensemble des personnes ayant de l'acné et ayant suivi l'un des deux traitements $A$ ou $B$.*
*On considère ici deux variables :*

- *la variable $X$ "traitement" dont les 2 modalités sont $A$ et $B$ ;*
- *la variable $Y$ "guérison" dont les 2 modalités sont "oui" et "non".*

*On désire connaître l'importance du lien entre le type de traitement et la guérison au moyen du calcul du coefficient $\varphi$ obtenu à partir d'un échantillon.*

*Pour cela, on commence par établir le tableau des effectifs observés pour chacune des modalités du couple de variables $(X, Y)$ :*

| $X$ \ $Y$ | *oui* | *non* | *total* |
|---|---|---|---|
| $A$ | $O_{11}$ | $O_{12}$ | $L_1$ |
| $B$ | $O_{21}$ | $O_{22}$ | $L_2$ |
| *total* | $C_1$ | $C_2$ | $N$ |

*On traduit alors les informations fournies par l'énoncé :*

- *le nombre de personnes de l'échantillon soumises au traitement $A$ est $L_1 = 710$ parmi lesquelles on a $O_{11} = 497$ guérisons ;*

- *le nombre de personnes de l'échantillon soumises au traitement $B$ est $L_2 = 1070$ parmi lesquelles on a $O_{21} = 856$ guérisons.*

*À partir de là, on peut remplir entièrement le tableau des effectifs observés :*
*en particulier, la taille de l'échantillon est $N = 710 + 1070 = 1780$,*
*$O_{12} = 710 - 497 = 213$ et $O_{22} = 1070 - 856 = 214$.*

*Ceci donne le tableau des effectifs observés :*

| $X$ \ $Y$ | *oui* | *non* | *total* |
|---|---|---|---|
| $A$ | 497 | 213 | 710 |
| $B$ | 856 | 214 | 1070 |
| *total* | 1353 | 427 | 1780 |

*On a alors (application de la formule du chi-deux) :*

$$\chi^2 = \sum \frac{(O_{ij} - T_{ij})^2}{T_{ij}}.$$

*On calcule les effectifs théoriques* $T_{ij} = \dfrac{L_i \times C_j}{N}$ *:*
*par exemple* $T_{11} = \dfrac{L_1 \times C_1}{N} = \dfrac{710 \times 1353}{1780} \approx 540.$

*On peut alors,*

*- soit calculer les 3 autres effectifs théoriques par la même méthode :*

- $T_{12} = \dfrac{L_1 \times C_2}{N} = \dfrac{710 \times 427}{1780} \approx 170$ *;*
- $T_{21} = \dfrac{L_2 \times C_1}{N} = \dfrac{1070 \times 1353}{1780} \approx 813$ *;*
- $T_{22} = \dfrac{L_2 \times C_2}{N} = \dfrac{1070 \times 427}{1780} \approx 257$ *;*

*- soit compléter le tableau des effectifs théoriques sachant que les marges sont les mêmes que pour les effectifs observés : alors* $T_{12} = L_1 - T_{11}$, $T_{21} = C_1 - T_{11}$, *puis* $T_{22} = L_2 - T_{21}$ *(ou bien* $C_2 - T_{12}$*).*

*On déduit du tableau des* $T_{ij}$*, le tableau des* $\dfrac{(O_{ij} - T_{ij})^2}{T_{ij}}$ *:*

- $\dfrac{(O_{11} - T_{11})^2}{T_{11}} \approx \dfrac{(497 - 540)^2}{540} \approx 3,4$ *;* $\dfrac{(O_{12} - T_{12})^2}{T_{12}} \approx \dfrac{(213 - 170)^2}{170} \approx 10,8$
- $\dfrac{(O_{21} - T_{21})^2}{T_{21}} \approx \dfrac{(856 - 813)^2}{813} \approx 2,2$ *;* $\dfrac{(O_{22} - T_{22})^2}{T_{22}} \approx \dfrac{(214 - 257)^2}{257} \approx 7,2.$

*Plus précisément, avec 2 décimales :*

*tableau des* $T_{ij}$ *:*

| | |
|---|---|
| 539,68 | 170,32 |
| 813,32 | 256,68 |

*; tableau des* $\dfrac{(O_{ij} - T_{ij})^2}{T_{ij}}$ *:*

| | |
|---|---|
| 3,38 | 10,69 |
| 2,24 | 7,10 |

*puis*

$$\begin{aligned}\chi^2 &= \frac{(O_{11}-T_{11})^2}{T_{11}} + \frac{(O_{12}-T_{12})^2}{T_{12}} + \frac{(O_{21}-T_{21})^2}{T_{21}} + \frac{(O_{22}-T_{22})^2}{T_{22}} \\ &\approx 3,38 + 10,69 + 2,24 + 7,10 \approx 23,41\end{aligned}$$

*et* $\varphi = \sqrt{\dfrac{\chi^2}{N(\min(p,q)-1)}}$ *avec ici* $p = q = 2$ *(*$X$ *et* $Y$ *ont 2 modalités chacune, donc* $\min(p,q) = 2$ *et* $\min(p,q) - 1 = 1$*, et* $N = 1780$ *donc* $\varphi \approx \sqrt{\dfrac{23,41}{1780}} \approx 0,115$.

*On obtient donc un coefficient* $\varphi$ *plutôt faible et on en conclut ainsi que le choix du traitement n'a pas grande importance en ce qui concerne la guérison.*

---

**Exercice 58.** *1. En considérant la population dont les individus sont les bureaux de vote, on peut définir :*

- *la variable nominale qui a chaque bureau associe sa catégorie ; 3 modalités : A, B, C.*
- *la variable quantitative qui à chaque bureau associe le nombre de voix obtenues par Johnny, d'où l'on déduira le pourcentage P ;*
- *la variable nominale qui, à chaque bureau, associe l'attitude du député, favorable ou non (on considère un bureau par circonscription).*

*2. a) On commence par étudier l'existence d'un lien éventuel entre la nature du résultat (bon ou mauvais) et la catégorie sociologique à laquelle appartient le bureau de vote. Pour cela, il faut établir le tableau des effectifs observés, puis calculer les effectifs théoriques et le* $\chi^2$ *:*

| *Tableau des* $O_{ij}$ | $A$ | $B$ | $C$ | *Total* |
|---|---|---|---|---|
| *Mauvais* | 11 | 13 | 4 | 28 |
| *Bon* | 2 | 3 | 3 | 8 |
| *Total* | 13 | 16 | 7 | 36 |

| *Tableau des* $T_{ij}$ | $A$ | $B$ | $C$ | *Total* |
|---|---|---|---|---|
| *Mauvais* | 10 | 12,5 | 5,5 | 28 |
| *Bon* | 3 | 3,5 | 1,5 | 8 |
| *Total* | 13 | 16 | 7 | 36 |

| *Tableau des* $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$ | $A$ | $B$ | $C$ |
|---|---|---|---|
| *Mauvais* | 0,08 | 0,02 | 0,38 |
| *Bon* | 0,27 | 0,09 | 1,34 |

*En faisant la somme, on obtient* $\chi^2 = 2,19$ *et* $\varphi = \sqrt{\dfrac{\chi^2}{N}} = \sqrt{\dfrac{2,19}{36}} \approx$*0,25 : lien plutôt faible.*

*b) On étudie maintenant l'existence éventuelle d'un lien entre la nature du résultat et l'attitude du député de la circonscription.*

| *Tableau des $O_{ij}$* | *oui* | *non* | *Total* |
|---|---|---|---|
| *Mauvais* | 16 | 12 | 28 |
| *Bon* | 2 | 6 | 8 |
| *Total* | 18 | 18 | 36 |

| *Tableau des $T_{ij}$* | *oui* | *non* | *Total* |
|---|---|---|---|
| *Mauvais* | 14 | 14 | 28 |
| *Bon* | 4 | 4 | 8 |
| *Total* | 18 | 18 | 36 |

| *Tableau des $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$* | *oui* | *non* |
|---|---|---|
| *Mauvais* | 2/7 | 2/7 |
| *Bon* | 1 | 1 |

*En faisant la somme, on obtient* $\chi^2 = 2 + \frac{4}{7} = \frac{18}{7} \approx 2{,}57$ *et* $\varphi = \sqrt{\frac{\chi^2}{N}} = \sqrt{\frac{2,57}{36}} \approx 0{,}27$ *: lien plutôt faible.*

---

**Exercice 59.** *Sur la population des 500 personnes ayant exprimé leurs suffrages, appelons $Y$ la variable réponse avec pour modalité "oui", "non", "sans opinion", et $X$ la variable sexe avec pour modalités "G", "F". Pour avoir un lien total entre les deux variables, il faut pour chaque individu, sa donnée suivant l'une des variables détermine automatiquement sa donnée suivant l'autre variable. Or ici, $X$ a 3 modalités et $Y$ seulement 2. On va donc s'arranger pour que chaque modalité de $Y$ corresponde à une seule modalité de $X$. Voici, partant de ce principe, 2 tables de contingence pour lesquelles $\varphi = 1$.*

| *Tableau des $O_{ij}$* | *oui* | *non* | *S.O.* | *Total* |
|---|---|---|---|---|
| *G* | 200 | 0 | 50 | 250 |
| *F* | 0 | 250 | 0 | 250 |
| *Total* | 200 | 250 | 50 | 500 |

| *Tableau des $O_{ij}$* | *oui* | *non* | *S.O.* | *Total* |
|---|---|---|---|---|
| *G* | 150 | 0 | 0 | 150 |
| *F* | 0 | 300 | 50 | 350 |
| *Total* | 150 | 300 | 50 | 500 |

*On calcule alors le coefficient $\varphi$ pour le premier cas par exemple :*

| *Tableau des $T_{ij}$* | *oui* | *non* | *S.O.* | *Total* |
|---|---|---|---|---|
| *G* | 100 | 125 | 25 | 250 |
| *F* | 100 | 125 | 25 | 250 |
| *Total* | 200 | 250 | 50 | 500 |

| *Tableau des $\frac{(O_{ij}-T_{ij})^2}{T_{ij}}$* | *oui* | *non* | *S.O.* |
|---|---|---|---|
| *G* | 100 | 125 | 25 |
| *F* | 100 | 125 | 25 |

*On a bien* $\chi^2 = 500$ *et* $\varphi = 1$.

---