ÉCOLE NORMALE SUPÉRIEURE DU BURUNDI

DÉPARTEMENT DES LANGUES ET SCIENCES HUMAINES

SECTION : HISTOIRE

STATISTIQUE DESCRIPTIVE

CODE : HIST2201

UE1 : COURS D'APPUI

VOLUME : 45H (3 ECTS)

COURS MAGISTRAL (CM) : 30H

TRAVAUX DIRIGÉS (TD) : 5H

TRAVAUX PRATIQUES (TP) : 10H

SYLLABUS DE L'ÉLÉMENT CONSTITUTIF DE L'UNITÉ D'ENSEIGNEMENT (ECUE) DESTINÉ AUX ÉTUDIANTS DE BACCALAURÉAT 2 EN HISTOIRE

Titulaire : Prof. Emmanuel BARANKANIRA

Docteur en Biostatistique de l'Université de Montpellier (France, 2016)

Master en Mathématiques-Biostatistique de SupAgro (France, 2012)

DEA en Statistique, parcours Épidémiologie et Biostatistique de l'UCL (Belgique, 2008)

Licencié en Pédagogie Appliquée, Agrégé de l'Enseignement en Maths (UB, 2004)

Bujumbura, 10 octobre 2024

Descriptif du cours

Processus	Paramètres	Description
Élaboration	Titre de l'ECUE	**Statistique descriptive**
	Objectif général	Développer des aptitudes pour réaliser une analyse exploratoire des données issues d'enquêtes ou des données expérimentales.
	Objectifs spécifiques	À la fin de l'ECUE, l'étudiant devrait être capable de : - collecter les données, les dépouiller et synthétiser l'information qu'elles contiennent par des tableaux et des graphiques ; - résumer les données par des indices descriptifs ; - développer un esprit de synthèse sur différents aspects du raisonnement statistique.
	Prérequis	Initiation à l'informatique
	Organisation de l'ECUE	VHP=45h, CM=30h, TD=5h,TP=10h, TPE=30h, TGE=75h (3 crédits)
	Bref contenu du cours	Ce cours vise à initier l'étudiant aux techniques d'échantillonnage et à la façon de synthétiser l'information que contiennent les données par des indices descriptifs et des graphiques. Des concepts clés (population, individu, échantillon, caractère, modalités) sont abordés. Des tableaux des statistiques descriptives et des tableaux complets des fréquences sont construits. En statistique descriptive univariée, les paramètres de tendance centrale (moyennes, médiane, mode, quartiles, percentiles, déciles, quintiles, etc), les paramètres de dispersion (étendue, variance, coefficient de variation, écart interquartile, etc), les paramètres de forme (skewness, kurtosis, coefficient de Yule et Kendall, etc) et les paramètres de concentration (médiale, indice de Gini) sont calculés et les graphiques (histogramme, boîte

		à moustaches, diagramme en barres, camembert, etc) construits. En statistique descriptive bivariée, l'analyse de régression (covariance, coefficient de corrélation, coefficient de détermination) est abordée. Les paramètres du modèle sont estimés et le nuage de points ajusté par la méthode des Moindres Carrés Ordinaires sans oublier les prévisions. L'application sur ordinateur est faite avec le logiciel R. Le cours est complété par le calcul des indices élémentaires et des indices synthétiques (indice de Paasche, indice de Laspeyres et indice de Fisher).
Méthodologie et supports pédagogiques	Méthodologie	Méthode expositive et participative
	Supports	Syllabus de l'ECUE Logiciel R
Modes d'évaluation	Évaluation formative	Travaux pratiques sur ordinateur et travaux dirigés sous forme d'exercices : 40 %
	Évaluation sommative	Examen final écrit : 60 %

Table des matières

Liste des tableaux

Liste des figures

Introduction

Dans cette brève introduction, il s'agit de jeter des lumières sur les concepts de statistique et de probabilités. La statistique est une branche des mathématiques appliquées concernant la planification, le résumé et l'interprétation d'observations. Il s'agit d'un terme qui n'a pas de définition universelle mais il existe des approches de définitions de ce concept oh combien intéressant et qui varient d'un auteur à l'autre. Il est primordial de distinguer, avant d'entrer au fond de l'Élément Constitutif de l'Unité d'Enseignement (ECUE), les notions de statistiques (au pluriel) et de statistique (au singulier). Une statistique est un mot employé également pour désigner un paramètre utilisé dans la statistique. Les statistiques (au pluriel) sont donc toutes les mesures calculées sur base d'un échantillon comme la moyenne, l'écart-type et correspondent en général à l'idée de dénombrement alors que la statistique est une science qui a pour objet de rassembler, d'ordonner, de traiter et d'analyser les données sans oublier l'interprétation des résultats et leur présentation afin de rendre les données compréhensibles par tous. C'est à la fois une science, une méthode et un ensemble de techniques.

Étymologiquement parlant, la statistique vient du mot latin « *status* » qui signifie « *État* » ou « *Gouvernement* ». La statistique est donc née lorsque les États ont senti le besoin de collecter des informations concernant les affaires de l'État à savoir les besoins de l'armée et les impôts payés par les citoyens. Il s'agissait de bien connaître la population pour administrer sa répartition sur les territoires, collecter les impôts et gérer les aspects militaires. Autrement dit, le terme « statistique » dont le nom est dérivé de « *state* » en référence à tout ce qui est étatique, a été introduit en Allemagne au 17ème siècle. Cependant, la pratique de la statistique date d'il y a longtemps puisqu'elle fut utile aux grands empires en Mésopotamie, dans l'Égypte ancienne, ainsi que chez les romains et les empires indiens et chinois. Plus exactement, la statistique est née vers les années 1743 au moment où les sociétés ont compris qu'il est nécessaire de faire le dénombrement de la population (recensement) et des biens possédés par la population (le bétail, la surface des terrains à cultiver, etc) afin d'en tenir compte lors de la planification. Dans le domaine de la santé, ce n'est qu'au 18ème siècle que la table de mortalité (en démographie) se voit apparaître. Plus tard, la statistique s'est étendue à d'autres domaines de la vie du pays tels que l'économie, la médecine, l'agriculture, la météorologie. Actuellement, au niveau sociétal, l'état de santé des populations est décrit et les facteurs de risque recherchés (en santé), les précipitations sont modélisées (en environnement), la production agricole est étudiée sous divers angles (en agriculture), le risque de rembourser le crédit est modélisé (en économie et en assurance), la répartition spatiale de la teneur en or est étudiée (en géostatistique), les discours

politiques sont statistiquement étudiés avec des logiciels de traitement des données langagières (en politique), les recettes des sites touristiques sont modélisées (en histoire), etc.

Il convient également de distinguer la statistique des probabilités. La statistique part de l'expérience ou de l'enquête (il faut des données par calculer des statistiques) alors que les probabilités permettent de prédire ce qui va arriver après l'expérience. En effet, il est possible par exemple de prédire la probabilité de réussir en Statistique descriptive avant la passer l'examen. Ces prédictions peuvent être faites notamment en fonction des conditions de travail dans lesquelles se trouvent les étudiants ou de la façon qu'ils jugent les méthodes d'enseignement utilisées. Il est aussi possible de se demander ce qui a précédé l'autre entre la statistique et les probabilités. Eh bien, au milieu du 17ème siècle, la théorie des probabilités s'intéressait aux jeux de hasard. Plus tard, les probabilités sont allées au-delà des jeux de hasard avec le développement des mathématiques et des probabilités (Blaise Pascal de la France, Pierre Fermat de la France, Jacob Bernoulli du Pays-Bas, Abraham de Moivre d'Angleterre, Pierre Simon de Laplace de la France, Carl Gauss d'Allemagne, Simon Poisson de la France, Kolmogorov de la Russie, Thomas Bayes d'Angleterre, Augustin Louis Cauchy de la France, etc) permettant de dégager des règles pour le traitement des données et donnant ainsi naissance à la statistique. Par ailleurs, le recensement, les chiffres du chômage, mais également les données concernant l'état de santé d'une population, sont des exemples de statistiques.

La statistique correspond à une notion différente : il s'agit d'un ensemble de méthodes, de raisonnements, de règles de décision permettant de présenter, d'analyser des données et d'extrapoler à une population les résultats obtenus dans le cadre d'une étude ponctuelle en prenant en compte la variabilité des phénomènes observés.

L'utilisation des statistiques est très ancienne, et remonte probablement aux premières sociétés structurées. En effet, très vite, les sociétés ont eu besoin de dénombrer les individus (recensements), mais également les biens (par exemple le nombre de têtes de bétail, la surface des terrains, etc.). Les Etats sont les premiers à avoir utilisé les statistiques, en particulier dans le but de collecter les impôts. Dans le domaine de la santé, l'utilisation des statistiques est assez tardive, puisqu'il faut attendre le 18ème siècle pour voir apparaître les premières tables de mortalité.

Le développement de la méthode statistique est quant à lui relativement récent, et ne débute véritablement que dans le courant du 18ème siècle. Le développement des mathématiques et des probabilités (Pascal, Fermat, Laplace, Bayes, et bien d'autres) permet de dégager des règles pour

le traitement des données, donnant ainsi naissance à la statistique. Parallèlement, les domaines ayant recours aux méthodes statistiques se multiplient : tout d'abord dans le domaine de l'agronomie, puis de la biologie, pour gagner peu à peu l'économie, l'industrie, et le domaine de la santé. De nombreux chercheurs ont permis le développement de la méthode statistique, dont les plus connus sont probablement K Pearson et RA Fisher. Au 20ème siècle, le principal déterminant du développement de la statistique a été sans conteste l'apparition et la diffusion de l'outil informatique. L'informatique a permis le développement de nouveaux outils statistiques qu'il était peu envisageable d'utiliser « à la main », mais a également entraîné une certaine démocratisation de l'utilisation des statistiques.

La statistique est formée de trois grandes classes : la statistique descriptive, la statistique analytique ou mathématique et la statistique inférentielle. La statistique descriptive est une méthode utilisée pour construire des tables, des graphiques et des résumés numériques des données. La statistique descriptive comme son nom l'indique, se propose de décrire les données, de les classer et de les présenter sous des formes claires et compréhensibles. Elle est à la base par exemple de toute organisation du système d'information d'une entreprise : statistiques de la production ou des ventes, statistiques financières, statistiques des ressources humaines,… Elle est aussi une importante composante en sciences humaines de ce qu'on appelle les méthodes quantitatives.La statistique analytique ou mathématique est une partie de la statistique qui est née à l'aube du 20ème siècle lorsque, Karl Pearson, Professeur au London University College, inventa une méthode de test qui permet de savoir si une distribution s'ajuste bien à une distribution théorique donnée (souvent, en pratique, c'est la loi normale). La statistique inférentielle est une partie de la statistique qui permet de tirer une conclusion (inférence) objective à propos d'une population en se servant de ce qui a été calculé sur base de l'échantillon. Cet ECUE est exclusivement consacré à la statistique descriptive.

L'objectif de cet ECUE est également de développer chez l'étudiant futur historien un esprit critique vis-à-vis des résultats des analyses statistiques. Les outils qui seront développés permettent d'examiner la tendance centrale, la dispersion, la distribution des données, les valeurs extrêmes ou aberrantes, la représentation graphique, l'analyse de régression et les indices élémentaires et les indices synthétiques. On pourrait penser par exemple à la représentativité des données, à la source des données qui pourrait affecter leur qualité.

Chapitre 1 : Types de variables et techniques d'échantillonnage

Il importe de commencer par définir le lexique qui sera utilisé tout le long de cet ECUE.

1.1. Vocabulaire de base

1.1.1. Population

Au début de tout travail statistique, il faut cerner avec précision sur quoi va porter l'étude. L'ensemble de tous les éléments ou objets sur lesquels porte l'étude s'appelle **population** [1]. Une population peut être un ensemble d'êtres vivants (humains, oiseaux, poissons, bactéries,…) ou un ensemble de choses (maisons, voitures, rivières,…) ou un ensemble de faits (pannes, accidents, divorces,…). Il s'agit par exemples de l'ensemble des arbres fruitiers se trouvant à l'École Normale Supérieure (ENS), de l'ensemble des ordinateurs de la salle informatique de l'ENS, de l'ensemble des réserves naturelles du Burundi, de l'ensemble des poissons du Lac Tanganyika, de l'ensemble des livres de statistique descriptive, de l'ensemble des minerais se trouvant au Burundi, de l'ensemble des entreprises installées au Burundi et de l'ensemble des étudiants de l'ENS.

Les exemples ci-après concernent le domaine de l'histoire :

- Ensemble des rois qui ont régné sur le Burundi ;
- Ensemble des nécropoles des reines-mères ;
- Ensemble des personnes tuées en 1972 au Burundi ;
- Ensemble d'askalis qui ont participé à la première guerre mondiale ;
- Ensemble des pays colonisés ;
- Ensemble des arbres sacrés du Burundi ;
- Ensemble des lieux touristiques du Burundi.

Les éléments d'une population possèdent en commun le caractère d'être tous membres d'une population mais ils varient selon d'autres critères. Le nombre d'éléments de la population s'appelle **taille de la population** et sera noté par **N**.

1.1.2. Individu

Chaque élément d'une population s'appelle **individu** ou **unité statistique**. Une population peut être finie (population d'un pays) ou presque infinie (population d'insectes). Généralement, les populations sont considérées comme finies mêmes si elles sont très grandes. Le nombre d'unités statistiques dans une population s'appelle **taille de la population** et se note par **N**.

Les exemples ci-après concernent le domaine de l'histoire :

- Un roi qui a régné sur le Burundi ;
- Une nécropole d'une reine-mère ;
- Une personne tuée en 1972 au Burundi ;
- Un askali qui a participé à la première guerre mondiale ;
- Un pays colonisé ;
- Un arbre sacré du Burundi ;
- Un lieu touristique du Burundi.

1.1.3. Échantillon

Quand une étude porte sur toute la population, alors il s'agit d'un **recensement** (enquête globale ou totale). Mais pour des raisons techniques ou économiques, il n'est généralement pas possible de collecter des données sur tous les éléments d'une population. Alors, une partie de la population appelée **échantillon** est extraite de cette population et ainsi l'étude est restreinte à cet échantillon. Un échantillon est un sous-ensemble d'observations représentatif de cette population.

Exemples :

- Un échantillon de 6 rois qui ont régné au Burundi ;
- Un échantillon de 4 pays colonisés par l'Allemagne avant la deuxième guerre mondiale ;
- Un échantillon de 20 pays indépendants en Afrique ;
- Un échantillon de 4 rebelles sous le roi Mwezi Gisabo ;
- Un échantillon de 50 personnes d'origine ivoirienne utilisées dans les restaurants en Europe.

Il existe des méthodes spécifiques permettant de s'assurer que l'échantillon est représentatif de la population, c'est-à-dire une réplique en miniature de ce qui se passe dans la population. Autrement dit, toutes les caractéristiques présentes dans la population doivent aussi l'être dans l'échantillon. Pour l'instant, à supposer qu'il s'agisse d'un échantillon sur lequel porte l'étude (sans savoir comment il a été extrait). Le nombre d'éléments dans l'échantillon s'appelle **taille de l'échantillon** et sera noté par **n**. Le fait que l'échantillon soit représentatif de la population garantit la validité externe d'une recherche, et le fait que l'échantillon soit représentatif permet de minimiser le biais. Il est à noter que la validité interne d'une recherche est montrée par la fiabilité des résultats issus des analyses statistiques.

1.1.4. Caractère

Appelé également « variable », un **caractère** est toute caractéristique observée ou mesurée sur chacun des individus de la population ou de l'échantillon [2]. Un caractère est symbolisé par une lettre de l'alphabet comme par exemple X, Y, Z et U.

1.1.5. Modalité

Les différentes valeurs que prend une variable s'appellent **modalités**. Afin que le classement d'une unité statistique soit toujours possible sans ambiguïté, les différentes modalités doivent être à la fois incompatibles (un individu ne peut avoir plusieurs modalités à la fois) et exhaustives (tous les cas doivent être prévus).

1.2. Calcul de taille d'échantillon

Il existe plusieurs façons permettant de calculer la taille de l'échantillon. Ces façons dépendent du type d'étude, du fait que la taille de la population est connue ou inconnue et de l'objectif visé par l'étude. Le calcul de taille d'échantillon se fait aussi en tenant compte des aspects éthiques, logistiques et financiers. Dans le cas où la taille de la population est inconnue, la taille de l'échantillon peut être calculée selon Cochran (1977) comme suit :

$$n = \frac{t^2 \times p \times (1-p)}{d^2} \tag{1.1}$$

où ***n*** est la taille d'échantillon minimale pour l'obtention de résultats significatifs pour un événement et un niveau de risque fixé, ***t*** le quantile de la loi de Student pour des grands échnatillons avec un

coefficient de risque de 5 % (la valeur type du niveau de confiance de 95 % sera 1,96), ***p*** la proportion estimée de la population qui présente la caractéristique considérée et ***d*** la marge d'erreur (généralement fixée à 5 %). Le calcul conduit à n=384.

Dans le cas où la taille de cette population est connue, la formule pouvant être utilisée est celle proposée par Krejcie et Morgan (1970) :

$$n = \frac{\chi^2 Np(1-p)}{d^2(N-1)+\chi^2 p(1-p)} \tag{1.2}$$

χ^2 étant le quantile de la loi du khi-deux à un degré de liberté (3,841), N la taille de la population, n la taille de l'échantillon, p la proportion estimée de la population qui présente la caractéristique considérée et d la marge d'erreur acceptable (5 %). Le calcul conduit à n=278.

La théorie d'échantillonnage (des sondages) est un ensemble d'outils statistiques permettant l'étude d'une population statistique à partir de l'examen d'un échantillon tiré de celle-ci. Le taux de sondage (ou d'échantillonnage) dans le cas d'une population finie est le rapport :

$$k = \frac{taille\ de\ la\ poputaion}{taille\ de\ l'échantillon} \tag{1.3}$$

Quand la population statistique est observée complètement, c'est-à-dire que l'échantillon est la population statistique toute entière, nous parlons d'échantillonnage exhaustif ou recensement. Le taux de sondage est alors 100 %. Pour des raisons financières ou techniques, il est impossible de faire un recensement. L'utilisation de sondages est alors incontournable. Nous distinguons deux grandes catégories de plans d'échantillonnage.

D'une part, les plans probabilistes, dits aussi plans stochastiques qui se caractérisent par le fait que les individus statistiques, qui font partie de l'échantillon, sont sélectionnés par tirages probabilistes. Chaque individu de la population statistique a une probabilité connue d'être inclus dans l'échantillon. Les plans probabilistes classiques sont les suivants: plan aléatoire simple, plan aléatoire systématique, plan aléatoire stratifié, plan aléatoire en groupes. L'échantillonnage probabiliste permet l'utilisation des méthodes d'estimation, et des méthodes d'inférence et d'analyse statistiques, qui toutes, sont basées sur la théorie des probabilités. Il permet aussi de connaitre et donc de contrôler les biais.

D'autre part, il y a des plans non probabilistes, dits aussi plans empiriques ou plans par choix raisonné. L'échantillon est construit par des procédés comportant une part arbitraire et ne permettant pas l'évaluation de la précision d'estimation. Les plans non probabilistes sont utilisés dans les études qualitatives où il n'est pas envisagé une extrapolation à la population statistique.

1.3. Techniques d'échantillonnage probabilistes

Les techniques d'échantillonnage probabilistes ou stochastiques sont des techniques d'échantillonnage basées sur des probabilités. Ces techniques sont notamment l'échantillonnage aléatoire simple, l'échantillonnage aléatoire stratifié, l'échantillonnage systématique ou périodique, l'échantillonnage en grappes et l'échantillonnage à plusieurs degrés.

1.3.1. Échantillonnage aléatoire simple

L'échantillonnage aléatoire simple est le modèle d'échantillonnage en apparence le plus simple que nous puissions imaginer. Il consiste à considérer que, dans une population de taille N, tous les échantillons de n unités sont possibles avec la même probabilité. Cependant, l'échantillonnage aléatoire simple fournit un cadre de référence indispensable pour deux raisons: C'est par rapport à ses propriétés que nous jugeons les autres modèles d'échantillonnage; il sert en quelque sorte d'étalon. Mais aussi, il constitue en général « la brique » élémentaire des plans usuels, par exemple, l'échantillonnage stratifié et l'échantillonnage à deux degrés sont des assemblages de sondages simples.

Un échantillonnage est aléatoire si tous les individus de la population ont la même chance de faire partie de l'échantillon. Il est simple si les prélèvements des individus sont réalisés indépendamment les uns des autres.

Les caractéristiques d'un échantillonnage aléatoire simple sont :

- Le caractère aléatoire minimise le risque de non représentativité de l'échantillon ;
- Nous pouvons anticiper sur le degré de précision de l'échantillon obtenu, et ainsi éviter une enquête inutile ;
- La méthode d'échantillonnage aléatoire simple permet la comparaison d'études similaires dans le temps.

Il consiste aussi à tirer, dans la population de taille N, un échantillon de taille n sans remise de telle sorte que chaque individu ait la même probabilité d'inclusion, sans regroupement préalable, ni manipulation sur la base de sondage. L'allocation sera soit à probabilité égale, soit à probabilité proportionnelle à la taille. Une façon simple de mettre en œuvre ce tirage est d'utiliser **le tirage systématique**.

1.3.2. Échantillonnage systématique

L'échantillonnage systématique est une technique qui consiste à prélever des unités d'échantillonnage situées à intervalles égaux; il consiste à tirer sur la base de sondage un individu sur k avec k donné par la relation (1.3). Le choix du premier individu détermine la composition de tout échantillon. Si nous connaissons l'effectif total de la population N et que nous souhaitons prélever un échantillon d'effectif n, l'intervalle entre deux unités successives à sélectionner est donné par : $k = \frac{N}{n}$.

Connaissant k, nous choisissons le plus souvent pour débuter, un nombre aléatoire i compris entre 1 et k, le rang des unités sélectionnées est alors :

$$i, i+k, i+2k, i+3k, \ldots$$

L'échantillonnage systématique réduit le temps consacré à la localisation des unités sélectionnées. Si les éléments de la population se présentent dans un ordre aléatoire (pas de tendance), l'échantillonnage systématique est équivalent à l'échantillonnage aléatoire simple. Par contre, si les éléments de la population présentent une tendance, l'échantillonnage systématique est plus précis que l'échantillonnage aléatoire.

Remarquons que l'échantillonnage systématique est utilisé lorsque nous ne connaissons pas la taille de la population. Dans cette situation, c'est le **pas** qui est établi à l'avance et par conséquent, la taille de l'échantillon sera aléatoire. Dans le cas où les variables étudiées présentent une dispersion forte dans la population, la précision des résultats d'un sondage aléatoire simple peut être améliorée par l'utilisation du **sondage aléatoire stratifié**.

1.3.3. Échantillonnage stratifié

L'échantillonnage stratifié est une technique qui consiste à subdiviser une population hétérogène, d'effectif N, en p sous-populations ou « strates » plus homogènes d'effectifs N_i N_i de telle sorte que $N = N_1 + N_2 + ... + N_p$. Les unités semblables sont regroupées [3]. Les strates doivent être mutuellement exclusives et exhaustives.

L'idée de base de la stratification est d'obtenir une estimation précise d'une moyenne de strate quelconque à partir d'un petit échantillon prélevé dans cette strate ainsi qu'une estimation précise pour l'ensemble de la population en combinant ces estimations.

Dans l'échantillonnage stratifié, la variance de l'estimateur ne comprend que la variation à l'intérieur des strates. Le degré de précision augmente avec le nombre de strates de la population car, plus elles sont nombreuses, plus les unités qu'elles contiennent sont nombreuses. Pour un échantillon de taille n, nous avons :

$$n = \sum_{h=1}^{H} n_h \tag{1.4}$$

Nous choisissons n_h individus dans la strate h de façon que nous ayons le même taux de sondage dans chaque strate que dans la population pour toutes les strates h :

$$\frac{n_h}{N_h} = \frac{n}{N} \tag{1.5}$$

Pour réaliser un sondage aléatoire stratifié, il est nécessaire de disposer de la base de sondage. D'autre part, même si nous disposons de cette base de sondage, le coût de l'enquête peut être prohibitif en raison des coûts de déplacement lorsque la population est dispersée géographiquement et que l'enquête est réalisée par enquêteur à domicile. Nous pouvons alors avoir recours à deux autres méthodes de sondage aléatoire : le **sondage en grappes** et **le sondage à plusieurs degrés**.

1.3.4. Échantillonnage à plusieurs degrés

Les sondages à plusieurs degrés utilisent une succession de regroupements des unités statistiques pour tirer l'échantillon. Le sondage à deux degrés met en œuvre un double échantillonnage : sur les unités primaires et secondaires.

Les sondages à deux degrés possèdent les propriétés d'invariance et d'indépendance :

- L'invariance signifie que les sondages du deuxième degré ne dépendent pas de ce qui s'est passé au premier degré;
- L'indépendance signifie que les tirages du deuxième degré sont indépendants les uns des autres (comme en stratification).

Les méthodes de sondages aléatoires qui ont été décrites précédemment supposent le tirage aléatoire de l'échantillon à partir d'une base de sondage, c'est-à-dire d'une liste exhaustive des individus composant la population étudiée. Lorsque de telles bases sont inexistantes ou indisponibles, lorsqu'il est trop coûteux de réaliser un sondage aléatoire, nous avons recours aux méthodes dites non aléatoires, ou encore méthodes empiriques ou à choix raisonné.

1.4. Techniques d'échantillonnage non probabilistes

Les techniques d'échantillonnage non probabilistes ou déterministes sont des techniques d'échantillonnage qui ne sont pas basées sur des probabilités. L'échantillonnage non probabiliste est un moyen de sélectionner des unités d'une population à l'aide d'une méthode subjective, c'est-à-dire non aléatoire. Il pose un problème : il n'est pas évident qu'il est possible de généraliser et d'appliquer les résultats de l'échantillon à toute la population. La raison de cette constatation est que la sélection d'unités statistiques dans une population pour un échantillon non probabiliste peut donner des biais d'importance. Puisque nous choisissons arbitrairement des unités, il n'existe aucune façon d'estimer la probabilité pour une unité quelconque d'être incluse dans l'échantillon. Également, comme la méthode en question ne fournit aucunement l'assurance que chaque unité aura une chance d'être incluse dans l'échantillon, nous ne pouvons ni estimer la variabilité de l'échantillonnage, ni identifier le biais possible.

Les statisticiens hésitent à utiliser les méthodes d'échantillonnage non probabilistes parce qu'il n'existe aucun moyen de mesurer la précision des échantillons qui en découlent. Malgré ces inconvénients, les méthodes d'échantillonnage non probabilistes peuvent être utiles lorsque nous désirons des commentaires descriptifs au sujet des échantillons eux–mêmes. En plus, leur utilisation prend peu de temps tout en étant plus économique et plus pratique.

L'échantillonnage non probabiliste peut être appliqué à des études qui servent :

- d'outils pour donner des idées ;
- d'étape préliminaire à l'élaboration d'une enquête par échantillonnage probabiliste ;
- d'étape de suivi pour aider à comprendre les résultats d'une enquête par échantillonnage probabiliste.

Les méthodes d'échantillonnage non probabiliste les plus couramment utilisées sont : l'échantillonnage de commodité, l'échantillonnage volontaire, l'échantillonnage au jugé, l'échantillonnage par quotas, l'échantillonnage « Boule de neige ».

1.4.1. Échantillonnage de commodité

Parfois appelé échantillonnage à l'aveuglette, accidentel ou de convenance, cet échantillonnage n'est pas normalement représentatif de la population cible, parce que nous ne sélectionnons des unités d'échantillonnage que si nous pouvons y avoir facilement et commodément accès. Celui qui fait l'échantillonnage à l'aveuglette présume que la population est homogène : si les unités de la population sont toutes semblables, n'importe quelle unité peut être choisie pour l'échantillon. Il s'agit d'un échantillon d'individus qui se trouvaient accidentellement à l'endroit et au moment où l'information a été collectée.

Les échantillons accidentels ne peuvent être considérés comme représentatifs d'aucune population. L'avantage évident de la méthode est qu'elle est facile à utiliser, mais la présence de biais annule énormément ce dernier. Même si ses applications utiles sont limitées, la technique peut donner des résultats exacts lorsque la population est homogène.

1.4.2. Échantillonnage volontaire

Cette méthode fait appel à des répondants volontaires. Elle sert souvent à sélectionner des particuliers pour des groupes de discussions approfondis, c'est-à-dire une mise à l'essai qualitative qui exclut la généralisation appliquée à la population complète. Le fait d'échantillonner des participants volontaires plutôt que la population en général peut introduire des biais. Souvent, à l'occasion des sondages d'opinion, seuls les gens qui se soucient assez fortement d'une façon ou d'une autre de la question étudiée ont tendance à y répondre. La majorité silencieuse n'y répond généralement pas, ce qui entraine un important biais sur le plan de la sélection.

1.4.3. Échantillonnage au jugé

La méthode d'échantillonnage au jugé est utilisée lorsque nous prélevons un échantillon en se fondant sur certains jugements au sujet de l'ensemble de la population. L'hypothèse qui sous-tend son utilisation est que l'enquêteur sélectionnera des unités qui seront les caractéristiques de la population. La méthode consiste à sélectionner des individus dont nous pensons avant de les interroger, qu'ils peuvent détenir l'information.

Les statisticiens utilisent souvent cette méthode dans le cadre d'études préparatoires comme des tests préalables de questionnaires et des discussions en groupe. La réduction du coût et du temps qu'exige l'acquisition de l'échantillon est l'un des avantages de l'échantillonnage au jugé. Le risque de ce type d'échantillonnage est de considérer des individus, apparemment représentatifs de la population étudiée.

1.4.4. Échantillonnage par quotas

L'échantillonnage par quotas est l'une des formes les plus courantes d'échantillonnage non probabiliste. Il s'effectue jusqu'à ce qu'un nombre précis d'unités (quotas) pour diverses sous-populations ait été sélectionné. Puisqu'il n'existe aucune règle qui régirait la façon dont il faudrait s'y prendre pour remplir ces quotas, l'échantillonnage par quotas est réellement un moyen de satisfaire aux objectifs en matière de taille d'échantillon pour certaines sous-populations.

L'échantillonnage par quotas ressemble à l'échantillonnage stratifié parce que les unités semblables sont regroupées. La méthode de sélection des unités est cependant différente. Les unités sont sélectionnées aléatoirement par l'échantillonnage probabiliste mais lors l'échantillonnage par quotas, une méthode non aléatoire est appliquée, c'est-à-dire les unités sollicitées qui ne sont pas disposées à participer sont remplacées par d'autres qui le sont, et nous ignorons en fait le biais de non-réponse.

La précision des estimateurs par quotas n'est pas calculable, puisqu'aucune probabilité n'est connue. En tenant compte du résultat numérique de précision, nous pouvons utiliser la formule de la variance d'un sondage stratifié, assimilant à une strate chaque sous-population sur laquelle nous devons respecter un quota.

1.4.5. Échantillonnage « Boule de neige »

Cette méthode est réservée aux populations composées d'individus dont l'identification est difficile ou qui possèdent des caractéristiques rares. La méthode consiste à faire construire l'échantillon par les individus eux-mêmes. Il suffit d'en identifier un petit nombre initial et de leur demander de faire appel à d'autres individus possédant les mêmes caractéristiques.

1.5. Types de variables

Il existe deux types de variables : les **variables qualitatives** et les **variables quantitatives**. Une variable est dite qualitative si elle ne peut être mesurée ou quantifiée, mais peut être classée en catégories comme le sexe, la race, l'espèce, le niveau scolaire,…. Une variable est de type quantitatif si elle peut être mesurée ou quantifiée, comme le poids, la hauteur, le revenu, le nombre d'enfants, le nombre de pannes.

1.5.1. Variables qualitatives

Les variables qualitatives sont des variables qu'il n'est possible ni de mesurer, ni de compter comme :

- Date des indépendances des pays africains
- Le roi du Burundi
- Le président de la République du Burundi
- Nationalité des africains
- Les Etats unis d'Amérique
- Niveau d'étude des historiens
- Grade militaire
- Grade académique
- Religion, état-civil, nationalité, profession

Les variables qualitatives sont constituées de deux sous-classes :

- Les variables qualitatives **nominales** : ce sont celles dont les modalités ne peuvent qu'être constatées, nommées.

  Exemple : Le sexe (masculin, féminin), la nationalité (Burundaise, Rwandaise, Congolaise, Gabonaise, Française, ...), les cours suivis durant une session (Statistique, Base de données, Anglais, Économétrie, ,..), …

- Les variables qualitatives **ordinales** : ce sont des variables qualitatives dont les modalités appellent naturellement un ordre dans leur rangement. **Exemple :** Le niveau d'instruction (aucun, primaire, secondaire, universitaire), le comportement lors d'une réception (incongru, correct, parfait,..), degré de satisfaction d'être étudiant de l'ENS (pas du tout satisfait, pas satisfait, satisfait, très satisfait, extrêmement satisfait), …

Les statistiques faites sur des variables qualitatives sont très limitées. Il s'agit du calcul des fréquences et de la construction des graphiques.

1.5.2. Variables quantitatives

Les variables quantitatives sont des variables qu'il est possible de mesurer ou compter comme :

- Quantité de matières premières exploitées annuellement au Congo Belge
- Quantité de café exporté par le Burundi de 1980 à 2023
- Âge au décès des rois burundais
- Nombre d'enfants par roi du Burundi
- Nombre de communes par province du Burundi
- Recettes quotidiennes des sites touristiques

Les variables quantitatives sont elles aussi subdivisées en deux sous-classes :

- Les variables quantitatives **discrètes :** ce sont celles dont les modalités sont des valeurs isolées. Autrement dit, ce sont des variables qui prennent des valeurs isolées dans l'intervalle des observations. Le support de ces variables est l'ensemble des naturels.

  Exemple : Le nombre d'ordinateurs, le nombre de banques, le nombre de pannes, le nombre d'accidents, le nombre d'enfants,…
- Les variables quantitatives **continues**, ce sont celles dont les modalités forment un continuum. Ce sont celles qui peuvent prendre n'importe quelle valeur dans un intervalle raisonnable des observations.

  Exemple : La taille, le poids, le revenu, la note obtenue à l'examen, …

1.6. Échelles de mesure

1.6.1. Échelle nominale

Pour les variables qualitatives, il existe deux échelles de mesure. **L'échelle nominale** qui s'adresse aux variables qualitatives nominales, elle ne sert qu'à coller une étiquette aux unités statistiques, elle ne les classe pas sur une échelle à une dimension.

Exemple :

- Soit X la variable qui désigne le sexe. Alors, X est une variable qualitative nominale et son échelle est nominale.
- Soit Y la variable qui représente le numéro du dossard d'un joueur de hockey. Même si Y prend des valeurs numériques, ce n'est qu'une variable nominale et son échelle est nominale, car il est possible tout aussi bien de mettre des lettres sur leur dossard ou des dessins.

1.6.2. Échelle ordinale

L'autre **échelle est l'échelle ordinale** et s'adresse aux variables qualitatives ordinales, on l'appelle comme cela car il y a un ordre entre ses modalités.

Exemple :

- Soit X la variable qui désigne le niveau scolaire d'une personne adulte. Alors, ses modalités peuvent être : primaire, secondaire, collégial, universitaire. Il y a un ordre chronologique entre ces modalités.
- Soit Y la variable qui désigne la note finale obtenue dans un cours de statistique. Ses modalités seront : F, E, D, C, B, A ou A+. Il y a un ordre de mérite entre ces modalités.

1.6.3. Échelle d'intervalle

Pour les variables quantitatives, il existe aussi deux types d'échelles, la première échelle est **l'échelle d'intervalle**. Cela s'appelle ainsi car la seule opération possible est la différence. Il y a une échelle d'intervalle par l'absence du zéro absolu (c'est-à-dire que si X=0, cela ne veut pas dire absence de ce qu'on mesure). Par exemple, lorsque la température est de 0 °C, cela ne veut pas dire qu'il n'y a pas de température. De plus, le passage de 5 °C à 10 °C est semblable au passage de 30 °C à 35 °C. C'est l'échelle d'intervalle. Cependant, lorsqu'une température est de 10 °C, cela ne

signifie pas qu'elle est deux fois plus élevée qu'une température de 5 °C ou encore deux fois moins élevée qu'une température de 20 °C.

Exemples :

- Soit T la variable qui désigne la température en degrés Celsius. Le jour où T=0 °C, cela ne veut pas dire absence de température. Si un expérimentateur observe deux journées où la température est respectivement égale à 10 et 30 degrés, cela veut seulement dire qu'il y a un écart de 20 degrés entre ces deux journées. S'il prend deux sots d'eau où la température est respectivement égale à 35 et 45 degrés, s'il les mélange, il ne va pas obtenir une eau chauffée à 80 degrés. Alors, l'échelle de cette variable est une échelle d'intervalle.
- Soit X la variable qui désigne la date de naissance. Si l'année qui nous intéresse est 2010 et si une personne est née en 1950 et une autre en 1980, tout ce qu'il est possible de dire est qu'il y a une différence d'âge de 30 ans entre elles. Il n'est pas possible de dire que l'une est deux fois plus âgée que l'autre, car l'année prochaine ce ne serait plus vrai. Alors l'échelle de cette variable est une échelle d'intervalle.

1.6.4. Échelle de rapports

L'autre échelle est **l'échelle de rapports**. C'est l'échelle la plus maniable, la plus riche. Elle admet un zéro absolu, c'est-à-dire si la variable est nulle, cela signifie l'absence de ce qu'on mesure. On peut faire toutes les opérations algébriques avec une telle échelle.

Exemple :

- Soit X le revenu familial annuel (en FBu). Si X=0, cela veut dire qu'il n'y a pas eu de revenu. Si deux familles ont des revenus respectifs de 30 000 et 120 000 FBu, alors il est possible de dire qu'il y a un écart de 90 000 FBu entre ces deux revenus. Il est aussi possible de dire que la deuxième famille gagne 4 fois plus que la première. Si ces deux revenus sont additionnés, alors le revenu global est de 150 000 FBu. Alors, l'échelle de cette variable est une échelle de rapports.
- Soit Y le nombre d'enfants dans un ménage. Si Y=0 cela veut dire que cette famille n'a pas d'enfant. On peut faire toutes les opérations algébriques avec les modalités de cette variable, donc son échelle est une échelle de rapports.

Exercices :

Donnez la nature des variables suivantes :

- Nombre d'actions vendues chaque jour à la bourse ;
- Rémunérations des enseignants d'un lycée ;
- Indicateur du moral des ménages ;
- Écart de rémunération entre hommes et femmes ;
- Pays de l'Union Européenne ;
- Niveau de formation des salariés ;
- Forme de contrat ;
- Taux de croissance du PIB ;
- Prix à la consommation ;
- Solde commercial ;
- Nombre de personnes par ménage.

Chapitre 2 : Paramètres de tendance centrale, de dispersion et de forme

Dans ce chapitre, nous allons nous intéresser aux paramètres de tendance centrale, aux paramètres de dispersion, aux paramètres de forme et aux paramètres de concentration, ces derniers étant obtenus à partir des paramètres de tendance centrale. Les paramètres de tendance centrale, appelés aussi paramètres de position ou de localisation, donnent une idée de l'ordre de grandeur des observations. Les paramètres de dispersion, quant à eux, quantifient la fluctuation des observations autour d'un paramètre de tendance centrale. Les paramètres de forme vont montrer la forme de la distribution (symétrique, asymétrique, normale, pas normale). Les paramètres de concentration vont permettre de calculer des grandeurs comme la médiale et l'indice de GINI en pondérant les observations par des valeurs centrales des classes. Le traitement des variables quantitatives discrètes est différent de celui des variables quantitatives continues. C'est ainsi que la description de ces variables se fera différemment.

Dans ce chapitre, nous allons nous intéresser à la description d'une seule variable.

2.1. Paramètres de tendance centrale

Les mesures de tendance centrale sont des valeurs de la variable susceptibles de nous donner une idée sur la donnée qui occupe le centre d'une série statistique. Ici, il sera question décrire les données à l'aide des paramètres à savoir le mode, la médiane, la moyenne, les quartiles, les percentiles (pourcentiles), les déciles et les quintiles sans oublier le minimum, le maximum et la fréquence. Les trois plus importantes mesures de tendance centrale sont **le mode, la moyenne et la médiane**.

2.1.1. Mode

Le mode d'une variable X est la valeur de la variable qui a la plus grande fréquence et se note Mo(X). Le mode est une importante mesure de tendance centrale aussi pour les variables qualitatives nominales. C'est l'observation la plus dominante, la plus fréquente.

Remarque : Une distribution peut avoir un seul mode. Dans ce cas, la distribution est unimodale. Elle peut aussi avoir deux modes (distribution bimodalme) ou plusieurs modes (distribution multimodale ou plurimodale).

Exemple 1 : Considérons l'exemple de la cote sur 20 obtenue par 20 étudiants.

15 7 12 10 8 11 14 10 11 11 15 6 9 8 14 16 13 11 12 10

Le mode vaut : $Mo = 11$

Exemple 2 : Considérons l'exemple des boissons non alcoolisées dont les données sont consignées dans le tableau des fréquences ci-après de source fictive :

Tableau 1 : Fréquences des boissons non alcoolisées

Boisson	Fréquences absolues (n_i)	Fréquences relatives (f_i)
CC	19	0,38
CL	8	0,16
PC	13	0,26
P	5	0,10
S	5	0,10
Total	50	1,00

Alors, le mode de cette variable est Mo(X)=Coca-Cola (CC), cela signifie que dans cet échantillon, la boisson la plus fréquemment achetée est Coca-Cola.

Exemple 2 : En reprenant l'exemple des recettes quotidiennes d'un petit magasin, où la variable est quantitative continue avec des données groupées en classes, on avait le tableau des fréquences ci-après.

Tableau 2 : Recettes d'un petit magasin

Classe (C_i)	Fréquences absolues (n_i)	Fréquences relatives (f_i)
[10, 100[	5	0,125
[100, 190[	3	0,075
[190, 280[	11	0,275
[280, 370[	6	0,150
[370, 460[	11	0,275
[460, 550[	3	0,075
[550, 640]	1	0,025
Total	40	1,000

Source : M'HAMMED MOUNTASSIR, 2018

Ici, il est clair qu'il y a deux classes qui ont la plus haute fréquence qui vaut 11. Ce sont des classes modales. Alors, la distribution de données est dite bimodale, et les deux modes sont les milieux des deux classes modales, à savoir Mo(X)=235 et Mo(X)=415 en faisant la moyenne. Cela veut dire que dans cet échantillon les recettes quotodiennes les plus fréquentes sont soit de 235 FBu ou de 415 FBu. En principe, il est possible de faire autrement en faisant des interpolations à l'intérieur des

classes modales pour trouver le mode, quoique dans le cas d'une variable quantitative le mode joue un rôle très marginal.

Le premier mode se calcule comme suit :

$$Mo = L_i + \frac{D_1}{D_1 + D_2} \times (L_s - L_i) = 190 + \frac{8}{8+5}(280 - 190) \simeq 245{,}38 \tag{2.1}$$

où L_i est la borne inférieure de la classe modale, L_s la borne supérieure de la classe modale, D_1 la fréquence de la classe modale diminuée de la fréquence de la classe précédant directement la classe modale et D_2 la fréquence de la classe suivant directement la classe modale diminuée de la fréquence de la classe modale.

Le deuxième mode se calcule comme suit :

$$Mo = L_i + \frac{D_1}{D_1 + D_2} \times (L_s - L_i) = 370 + \frac{5}{5+8}(460 - 370) \simeq 404{,}62 \tag{2.2}$$

Le mode d'une variable est une mesure de tendance centrale facile à déterminer et s'applique à tous les types de variables, mais sa portée comme mesure d'analyse est très limitée. Lorsque les classes ne sont pas de mêmes amplitudes, une fréquence minimale est prise comme référence. Sachant que l'amplitude a_i de chaque classe a été calculée, elle sera divisée par la fréquence a qui a été considérée comme référence. Les fréquences fictives h_i qui vont permette de retrouver la classe modale pour des classes d'amplitudes différentes seront calculées en partant des fréquences n_i des classes : $h_i = \frac{n_i}{\frac{a_i}{a}} = \frac{a\,n_i}{a_i}$

Si les classes ne sont pas de même amplitude, le mode se calcule comme suit :

Classe (C_i)	**Fréquence absolue (n_i)**	**a_i**	**a_i/a**	**$h_i=n_i/(a_i/a)$**
[10, 20[	10 (a)	10	1	10
[20, 30[	40	10	1	40
[30, 50[	220	20	2	**110**
[50, 80[	240	30	3	80
[80, 100[	10	20	2	5
[100, 130[	24	30	3	8

Le mode vaut :

$$Mo = L_i + \frac{D_1}{D_1 + D_2} \times (L_s - L_i) = 30 + \frac{180}{180 + 20} \times 20 = 48$$

2.1.2. Moyenne

La moyenne arithmétique ou tout simplement **la moyenne** est la mesure de tendance centrale la plus connue. Elle ne s'applique qu'aux variables quantitatives. Nous allons décrire la méthode pour calculer la moyenne d'une variable quantitative selon que les données sont en vrac, groupées par valeurs ou groupées par classes. La série statistique simple est la série des valeurs présentées en vrac. Soit X une variable quantitative dont les valeurs observées sur un échantillon forment une série en vrac $x_1, x_2, ..., x_n$ alors la moyenne de cet échantillon est [2] :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.3}$$

Exemple 1 : Un commerçant a l'habitude de noter dans son registre le nombre de clients qui se présentent quotidiennement à son magasin. Considérons que l'échantillon est de taille 10 et que les valeurs enregistrées sont suivantes :

120 105 90 201 196 65 88 163 103 116

Alors, dans cet échantillon, le nombre moyen des clients qui se présentent à ce magasin par jour est donné par la formule suivante :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = 124,7 \simeq 125 \text{ clients par jour}$$

Exemple 2 : Reprenons l'exemple de la cote sur 20 obtenue par 20 étudiants.

15 7 12 10 8 11 14 10 11 11 15 6 9 8 14 16 13 11 12 10

La note moyenne vaut :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{223}{20} = 11,15$$

La série statistique pondérée est la série des données groupées par valeurs. Soit X une variable quantitative discrète dont les données se présentent sous forme d'un tableau où elles sont classées par valeurs, supposons que la taille de l'échantillon est n et qu'il y a *k* valeurs différentes pour cette variable. Alors la moyenne d'un tel échantillon de données est [1] :

$$\overline{x} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i \tag{2.4}$$

Exemple : Reprenons les données de l'exemple ci-haut, où X est le nombre d'accidents de travail par semaine (**Source :** M'HAMMED MOUNTASSIR, 2018) :

Tableau 3 : Fréquences du nombre d'accidents par semaine

x_i	Fréquences absolues
0	4
1	2
2	10
3	7
4	10
5	4
6	3
Total	40

Alors, la moyenne de cet échantillon est égale à :

$$\overline{x} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i = \frac{127}{41} = 3{,}097 \simeq 3{,}00 \text{ accidents par semaine}$$

La série statistique classée représente les données groupées par classes. La formule de calcul est celle utilisée pour une série pondérée, saut que les observations seron les centres des classes. Considérons un tableau où les données provenant d'un échantillon sont groupées par classes. Alors pour calculer la moyenne de cet échantillon, une formule approximative sera utilisée, où chaque classe est assimilée à son centre et la formule utilisée est celle utilisée pour le cas où les données sont groupées par valeurs. En notant par x_i, le milieu de la i^{ème} classe et en supposant que la taille de l'échantillon soit *n* et qu'il y ait *k* classes, alors la moyenne de l'échantillon est donnée par la relation (2.2), où les x_i sont maintenant les valeurs centrales des classes, c'est-à-dire la moyenne des bornes (borne supérieure plus borne inférieure divisé par deux).

Exemple : En reprenant l'exemple ci-haut où X est la recette quotidienne d'un petit magasin, en y ajoutant une colonne à droite contenant le milieu des classes :

Classe (C_i)	Fréquences absolues (n_i)	Centre de classe (x_i)
[100, 190[	3	145
[190, 280[	11	235
[280, 370[	6	325
[370, 460[	11	415
[460, 550[	3	505
[550, 640]	1	595
Total	40	////

Il vient le tableau ci-après :

C_i	x_i	n_i	$n_i x_i$
[10, 100[	55	5	275
[100, 190[	145	3	435
[190, 280[	235	11	2585
[280, 370[	325	6	1950
[370, 460[	415	11	4565
[460, 550[	505	3	1515
[550, 640]	595	1	595
Total	/////////	40	**11920**

Alors, la moyenne de cet échantillon est :

$$\overline{x} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i = \frac{11920}{40} = 298$$

Il existe d'autres types de moyennes à savoir :

- **la moyenne géométrique** :

$$G = M_g = \sqrt[n]{x_1 \times ... \times x_n} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} \text{ (série simple)}$$

$$G_p = \sqrt[n]{x_1^{n_1} \times ... \times x_k^{n_k}} = \left(\prod_{i=1}^{k} x_i^{n_i} \right)^{\frac{1}{n}} \text{ (série pondérée ou classée)}$$

- **la moyenne harmonique** :

$$H = M_h = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{x_i} \right)^{-1} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \text{ (série simple)}$$

$$H_p = \left(\frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i \left(\frac{1}{x_i} \right) \right)^{-1} = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \text{ (série pondérée ou classée)}$$

- **la moyenne quadratique** :

$$Q = M_q = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}} \text{ (série simple)}$$

$$Q_p = \left(\frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i^2 \right)^{\frac{1}{2}} \text{ (série pondérée ou classée)}$$

- **la moyenne cubique** :

$$C = M_c = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^3 \right)^{\frac{1}{3}} \text{ (série simple)}$$

$$C_p = \left(\frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i^3 \right)^{\frac{1}{3}} \text{ (série pondérée ou classée)}$$

Remarquons que ces 5 moyenne sont telles que :

$$M_h < M_g < \bar{x} < M_q < M_c \Leftrightarrow H < G < M < Q < C$$

Exercices d'application

Exercice 1 :

Soit la série statistique suivante :

4 1 12 27

Calculez toutes les moyennes.

Solution

	x_i	$log(x_i)$	$\frac{1}{x_i}$	x^2	x^3
	4	0,602	0,250	16	64
	1	0,000	1,000	1	1
	12	1,079	0,083	144	1728
	27	1,431	0,037	729	19683
Σ	44	3,112	1,370	890	21476

Source : Auteur à partir des données fictives

La moyenne arithmétique de x vaut :

$$M = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{44}{4} = 11$$

La moyenne géométrique de x est donnée par :

$$log(G) = \frac{1}{n}\sum_{i=1}^{n} log(x_i) = \frac{3,112}{4} = 0,778$$
$$\Rightarrow G = 10^{0,778} = 5,997 \simeq 6,00$$

La moyenne harmonique de x vaut :

$$H = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{4}{1,370} = 2,919 \simeq 2,92$$

La moyenne quadratique de x est donnée par :

$$Q = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}} = \left(\frac{890}{4}\right)^{\frac{1}{2}} = (222,5)^{\frac{1}{2}} \simeq 14,92$$

La moyenne cubique de x est donnée par :

$$C = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^3 \right)^{\frac{1}{3}} = \left(\frac{21476}{4} \right)^{\frac{1}{3}} = (5369)^{\frac{1}{3}} \simeq 17,51$$

Résumé :

$$\left.\begin{array}{l} H = 2,92 \\ G = 6,00 \\ M = 11 \\ Q = 14,92 \\ C = 17,51 \end{array}\right\} \Rightarrow H < G < M < Q < C$$

Exercice 2 :

Soit la série statistique suivante :

x_i	1	2	3	4
n_i	27	47	3	3

Calculez toutes les moyennes.

	x_i	n_i	$n_i x_i$	$log(x_i)$	$n_i log(x_i)$	$\frac{n_i}{x_i}$	x_i^2	$n_i x_i^2$ [,6]	x_i^3	$n_i x_i^3$
	1	27	27	0,000	0,000	27,00	1	27	1	27
	2	47	94	0,301	14,147	23,50	4	188	8	376
	3	3	9	0,477	1,431	1,00	9	27	27	81
	4	3	12	0,602	1,806	0,75	16	48	64	192
Σ	///	80	142	////////	17,384	52,25	////	290	////	676

Source : Auteur à partir des données fictives

Solution :

La moyenne arithmétique pondérée de x vaut :

$$M_p = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i = \frac{142}{80} = 1,775 \simeq 1,77$$

La moyenne géométrique de x est donnée par :

$$log\left(G_p\right)=\frac{1}{\sum_{i=1}^{k} n_i}\sum_{i=1}^{n} n_i\, log\left(x_i\right)=\frac{17,384}{80}=0,2173$$

$$\Rightarrow G_p=10^{0,2173}=1,6493\simeq 1,65$$

La moyenne harmonique pondérée de x vaut :

$$H_p=\frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k}\frac{n_i}{x_i}}=\frac{80}{52,25}=1,5311\simeq 1,53$$

La moyenne quadratique de x est donnée par **:**

$$Q_p=\left(\frac{1}{\sum_{i=1}^{k} n_i}\sum_{i=1}^{n} n_i\, x_i^2\right)^{\frac{1}{2}}=\left(\frac{290}{80}\right)^{\frac{1}{2}}=\left(3,625\right)^{\frac{1}{2}}=1,903\simeq 1,90$$

La moyenne cubique pondérée de x est donnée par :

$$C_p=\left(\frac{1}{\sum_{i=1}^{k} n_i}\sum_{i=1}^{k} n_i\, x_i^3\right)^{\frac{1}{3}}=\left(\frac{676}{80}\right)^{\frac{1}{3}}=2,036\simeq 2,04$$

Résumé :

$$\left.\begin{array}{l} H_p=1,53 \\ G_p=1,65 \\ M_p=1,77 \\ Q_p=1,90 \\ C_p=2,04 \end{array}\right\}\Rightarrow H_p<G_p<M_p<Q_p<C_p$$

S'agissant des propriétés d'une moyenne échantillonnale, soit X une variable quantitative dont la moyenne échantillonnale est $\overline{x}$ et soit y une autre variable quantitative transformée linéaire de x,

c'est-à-dire que $Y = a + bX$ où a et b sont des constantes réelles. Alors, la moyenne échantillonnale de y sera égale à $\overline{y} = a + b\overline{x}$. On dit que la moyenne conserve la transformation linéaire entre les variables.

Exemple : Soit X, le nombre d'heures qu'un étudiant travaille à temps partiel par semaine. Supposons qu'à partir d'un échantillon d'étudiants, on a pu trouver qu'en moyenne le nombre d'heures travaillées par ces étudiants est égale à $\overline{x} = 14,5$ heures/semaine. Si le salaire horaire est de 10 (en milliers de FBu) et que les patrons de ces étudiants leur offrent 30 par semaine pour leurs déplacements, quel est le gain net moyen hebdomadaire de ces étudiants ? Posons Y, le gain net hebdomadaire de ces étudiants alors $y = 30 + 10x$, donc le gain moyen hebdomadaire de cet échantillon d'étudiants est égal à $\overline{y} = 30 + 10\overline{x} = 30 + 10 \times 14,5 = 175$.

2.1.3. Médiane

La médiane est la valeur de la variable qui divise l'échantillon en deux groupes d'égal effectif. Il y a 50 % des données qui sont inférieures à la médiane et 50 % des données qui sont supérieures ou égales à la médiane [4]. La médiane se calcule pour des variables qualitatives ordinales et pour des variables quantitatives. La médiane d'une variable X se note par Med(X) ou par $\tilde{x}$. Dans la suite, les façons de calculer une médiane dans les différents cas possibles seront présentées.

Puisque les modalités d'une variable qualitative ordinale sont déjà ordonnées par nature, alors pour déterminer la médiane, il suffit de calculer $l = 50\,\% \times n$, et donc :

$$Me = \begin{cases} \dfrac{x_{(l)} + x_{(l+1)}}{2} \; si\; l\; entier \\ x_{(l)+1} \; si\; l\; \; pas\; entier \end{cases} \tag{2.5}$$

où $x_{(l)}$ signifie l'observation occupant le rang immédiatement supérieur à *l*.

Exemple : Reprenons les données de l'exemple ci-haut, où X est le degré de satisfaction de la clientèle :

Tableau 4 : Fréquences du degré de satisfaction des clients

Degré de satisfaction	Fréquences absolues
1	0
2	0
3	2
4	3
5	12
6	25
7	18
Total	60

Ici, n=60 et $l = 50\,\% \times n = 30$ est un entier, alors la médiane vaut : $Me = \frac{x_{(30)} + x_{(31)}}{2} = \frac{6+6}{2} = 6$

Le degré de satisfaction médian de la clientèle est égal à 6, ce qui veut dire que dans cet échantillon 50 % des clients ont un degré de satisfaction de moins de 6 et l'autre 50 % un degré de satisfaction de 6 ou plus.

Pour les données quantitatives en vrac ou groupées par valeurs**,** il faut d'abord ordonner les données par ordre croissant avant d'appliquer la même procédure comme pour les variables qualitatives ordinales. Ci-après nous donnerons un exemple pour chacun de ces deux cas.

Exemple : Reprenons les données de l'exemple ci-haut où la variable est le nombre de clients qui se présentent quotidiennement au magasin.

Les données se présentaient comme suit (données en vrac) :

120 105 90 201 196 65 88 163 103 116

La série ordonnée est : 65 88 90 103 105 116 120 163 196 201

Ici, n=10 et $l = 50\,\% \times n = 5$ est un entier, alors $Me = \frac{x_{(5)} + x_{(6)}}{2} = \frac{105+116}{2} = 110,5$, ce qui veut dire qu'à partir de cet échantillon, on peut affirmer que dans 50 % des journées, ce magasin reçoit 110 clients ou moins par jour et dans l'autre 50 % des journées, il reçoit 110 clients ou plus.

La médiane peut aussi se calculer comme suit :

$$Me = \begin{cases} \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \; si \; n \; pair \\ x_{\frac{n+1}{2}} \; si \; n \; impair \end{cases} \tag{2.6}$$

Exemple : Reprenons les données de l'exemple ci-haut, où X est le nombre d'accidents de travail par semaine. Ajoutons une donnée supplémentaire au tableau de données où les modalités de la variable sont groupées par valeurs.

Tableau des fréquences du nombre d'accidents par semaine

Nombre d'accidents par semaine	Fréquences absolues
0	4
1	2
2	10
3	7
4	10
5	4
6	4
Total	41

Ici, n=41 et $l = 50\,\% \times n = 20,5$ n'est pas un entier, alors $Me = x_{(20,5)+1} = x_{(21)}$ qui est l'observation occupant la 21ème position, c'est-à-dire 3, c'est-à-dire que dans cet échantillon, dans 50 % des semaines, 3 accidents ou moins par semaine sont observés et l'autre 50 % des semaines où 3 accidents ou plus par semaine le sont.

Voyons maintenant ce qu'il en est des données groupées par classes. Dans le cas où un tableau de fréquences complet (incluant les fréquences cumulées) des données groupées par classes est présenté, il faut d'abord déterminer la classe médiane, i.e. la classe pour laquelle les fréquences cumulées dépassent pour la première fois 50 %. Cette classe aura la forme : [L_i, L_s] et alors la médiane s'obtient par interpolation à l'intérieur de cette classe médiane à l'aide de la formule suivante (méthode des fréquences relatives):

$$Me = L_i + \frac{0,50 - F_i}{f_i} \times \left(L_s - L_i \right) \tag{2.7}$$

où L_i est la borne inférieure de la classe médiane, L_s la borne supérieure de la classe médiane, F_i la fréquence cumulée avant la classe médiane et f_i la fréquence relative de la classe médiane.

Exemple : En reprenant les données où X donne la recette quodienne d'un petit magasin, il est facile de retrouver le tableau des fréquences suivant :

Classe	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
[10, 100[	5	0,125	0,125
[100, 190[	3	0,075	0,200
[190, 280[	11	0,275	0,475
[280, 370[	6	0,150	0,625
[370, 460[	11	0,275	0,900
[460, 550[	3	0,075	0,975
[550, 640]	1	0,025	1,000
Total	40	1,000	

Alors ici, la classe médiane est $[L_i, L_s]$=[280 ; 370[

La médiane vaut alors :

$$Me = L_i + \frac{0,50 - F_i}{f_i} \times (L_s - L_i) = 280 + \frac{0,50 - 0,475}{0,150} \times 90 = 295$$

Cela veut dire qu'en se basant sur cet échantillon de données, 50 % des recettes quotidiennes de ce petit magasin sont inférieures à 295 et les autres 50 % sont supérieures ou égales à 295.

Remarque 1 : Le calcul de la médiane est basé sur l'ordre des observations et non sur leur valeur. Contrairement à la moyenne, la médiane est insensible aux données extrêmes. Dans le cas où les données sont très différentes, la médiane est une meilleure mesure de tendance centrale.

Remarque 2 : Si pour une variable X quantitative les 3 mesures de tendance centrale sont presque égales, alors la variable est dite symétrique et n'importe laquelle de ces mesures peut être utilisée comme mesure de cette tendance centrale. S'il y a un grand écart entre ces mesures, alors c'est la médiane qu'on doit priviligier.

Le calcul de la médiane peut aussi se faire par *interpollation linéaire*. En effet, l'équation cartésinenne d'une doite est [1] :

$$y = ax + b \tag{2.8}$$

En remplaçant la valeur de y par la fréquence relative cumulée de la classe qui précède la classe médiane, on obtient :

$$0,475 = a(280) + b$$

De même, en la remplaçant par la fréquence relative cumulée de la classe médiane, il vient :

$$0,625 = a(370) + b$$

La soustraction de ces deux équations membre à membre et après quelques développements donne :

$$a = \frac{0,150}{90} = 0,00167$$

L'injection de cette valeur dans la première équation donne :

$$b = 0,475 - 280(0,00167) = 0,0074$$

Ainsi, l'équation de la droite s'écrit :

$$y = 0,00167x + 0,0074$$

Le remplacement de y par la valeur 0,50 (50 %) donne :

$$0,50 = 0,00167x + 0,0074$$
$$\Leftrightarrow 0,00167x = 0,50 - 0,0074$$
$$\Leftrightarrow 0,00167x = 0,4926$$

soit

$$\Leftrightarrow 1,67x = 492,6$$
$$\Leftrightarrow x = \frac{492,6}{1,67} = 294,97$$

La valeur trouvée est proche de la précédente.

La médiane peut aussi se calculer à l'aide de la formule (méthode des fréquences absolues) :

$$Me = L_i + \frac{\frac{n}{2} - \left(\sum_i n_i\right)_{Me-1}}{(n_i)_{Me}} \times (L_s - L_i) \qquad (2.9)$$

où L_i est la borne inférieure de la classe médiane, L_s la borne supérieure de la classe médiane, $\left(\sum_{i=1}^{k} n_i\right)_{Me-1}$ la somme des effectifs avant la classe médiane et $(n_i)_{Me}$ l'effectif de la classe médiane.

2.1.4. Quartiles

Les quartiles sont des observations qui partagent la distribution en quatre parties d'effectifs égaux ; il y en a quatre : le premier quartile Q_1, le deuxième quartile Q_2 et le troisième quartile Q_3. La médiane est le deuxième quartile note aussi Méd(X).

Le premier quartile est tel que 25 % des observations lui sont inférieures et 75 % des observations lui sont supérieures ou égales. Le deuxième quartile est tel que 50 % des observations lui sont inférieures et 50 % des observations lui sont supérieures ou égales. Le troisième quartile est tel que 75 % des observations lui sont inférieures et 25 % des observations lui sont supérieures ou égales. Les quartiles peuvent se calculer pour une série statistique simple, une série statistique pondérée et une série une série statistique classée représentant une variable quantitative.

Avant de calculer la médiane d'une série statistique simple, il faut d'abord commencer par ordonner les données par ordre croissant. La deuxième étable consiste à calculer les rangs des quartiles en fonction du nombre *n* d'observations :

$$rang(Q_1) = \frac{1n+1}{4}$$
$$rang(Q_2) = \frac{2n+2}{4}$$
$$rang(Q_3) = \frac{3n+3}{4}$$

La troisième étape consiste à calculer ces quartiles en se basant sur leurs rangs, c'est-à-dire les positions occupées par ces quartiles.

Exemple 1 : Considérons la série statistique suivante de 12 observations sont :

-2 -3 10 12 120 11 4 8 6 13 130 200

La série ordonnée est :

-3 -2 4 6 8 10 11 12 13 120 130 200

Le rang du premier quartile vaut :

$$rang\left(Q_1\right)=\frac{1n+1}{4}=\frac{13}{4}=3,25$$

Le premier quartile vaut donc :

$$Q_1=4+0,25\left(6-4\right)=4,5$$

Le rang du deuxième quartile vaut :

$$rang\left(Q_2\right)=\frac{2n+2}{4}=\frac{25}{4}=6,25$$

Le deuxième quartile vaut donc :

$$Q_2=10+0,25\left(11-10\right)=10,25$$

Le rang du troisième quartile vaut :

$$rang\left(Q_3\right)=\frac{3n+3}{4}=\frac{39}{4}=9,75$$

Le troisième quartile vaut donc :

$$Q_3=13+0,75\left(120-13\right)=93,25$$

Exemple 2 : Considérons la série statistique suivante de 10 observations :

3 10 12 8 6 100 15 6 3 14

La série ordonée est :

3 3 6 6 8 10 12 14 15 100

Le rang du premier quartile vaut :

$$rang\left(Q_1\right)=\frac{1n+1}{4}=\frac{11}{4}=2,25$$

Le premier quartile vaut donc :

$$Q_1=3+0,25\left(6-3\right)=3,75$$

Le rang du deuxième quartile vaut :

$$rang\left(Q_2\right)=\frac{2n+2}{4}=\frac{22}{4}=4,5$$

Le deuxième quartile vaut donc :

$$Q_2=6+0,5\left(8-6\right)=7$$

Le rang du troisième quartile vaut :

$$rang\left(Q_3\right)=\frac{3n+3}{4}=\frac{33}{4}=8,25$$

Le troisième quartile vaut donc :

$$Q_3=14+0,25\left(15-14\right)=14,25$$

Remarque : La procédure décrite pour trouver les quartiles est une convention parmi d'autres. Il n'y a pas d'accord général sur la méthode à utiliser pour déterminer les quartiles. Il est aussi possible de chercher la valeur centrale entre deux valeurs pour trouver le quartile. Si vous utilisez des logiciels, les valeurs trouvées diffèrent d'un logiciel à l'autre, à l'exception de la médiane. Par exemple, en considérants la série statistique suivante :

1 3 6 10 15 21 28 36

alors la calculatrice TI-83 et plus et les logiciels suivants donnent :

Tableau 5 : Quartiles

Logiciel	Q_1	Q_2	Q_3
SPSS	3,75	12,5	26,25
SAS	4,5	12,5	24,5
STATDISK	4,5	12,5	24,5
Excel	5,25	12,5	22,75
R	5,25	12,5	22,75
Splus	5,25	12,5	22,75
Minitab	3,75	12,5	26,25
TI-83 et plus	4,5	12,5	24,5

Source : M'HAMMED MOUNTASSIR, 2018

Heureusement, dans la pratique, les échantillons sont très grands et ces fluctuations ne changent pas grand-chose dans les analyses des données. La même démarche suivie précédemment et qui consiste à calculer les rangs des quartiles est celle qui sera utilisée pour une série statistique pondérée.

Exemple : En reprenant le tableau de l'exemple relatif au nombre d'accidents par semaine,

déterminons les 3 quartiles de la variable X=le nombre d'accidents par semaine.

Tableau des fréquences du nombre d'accidents par semaine

X	Fréquences absolues (n_i)	Fréquences absolues cumulées (N_i)
0	4	0
1	2	0
2	10	2
3	7	5
4	10	17
5	4	42
6	4	60
Total	41	////

Le rang du premier quartile vaut :

$$rang\left(Q_1\right)=\frac{1n+1}{4}=\frac{42}{4}=10,5$$

Le premier quartile vaut :

$$Q_1=2+0,5(2-2)=2$$

Le rang du deuxième quartile vaut :

$$rang\left(Q_2\right)=\frac{2n+2}{4}=\frac{84}{4}=21$$

Le deuxième quartile vaut :

$$Q_2=3$$

Le rang du troisième quartile vaut :

$$rang\left(Q_3\right)=\frac{3n+3}{4}=\frac{126}{4}=31,5$$

Le troisième quartile vaut :

$$Q_3 = 4 + 0,5(4 - 4) = 4$$

Réponse :

$Q_1 = 2$ signifie que dans cet échantillon, durant 25 % des semaines, 2 accidents ont été observés par semaine ou moins.

$Q_2 = 3$ signifie que dans cet échantillon, durant 50 % des semaines, 3 accidents ont été observés par semaine ou moins.

$Q_3 = 4$ signifie que dans cet échantillon, durant 75 % des semaines, 4 accidents ont été observés par semaine ou moins.

La même démarche utilisée pour calculer la médiane quand les données sont groupées par classes est celle qui sera utilisée pour une série statistique classée. D'abord, la classe qui contient le quartile ou autrement dit pour lequel le pourcentage relatif à chaque quartile a été dépassé est repérée et ensuite, une interpolation se fait à l'intérieur de cette classe. La même formule que celle qui a été utilisée pour calculer la médiane où seul le poucentage est à adapter est utilisée.

Exemple : En reprenant les données de l'exemple relatif aux recettes quotidiennes d'un petit dépanneur, calculons et interprétons ces mesures.

Classe	Fréquences absolues (n_i)	Fréquences relatives (f_i)	Fréquences relatives cumulées (F_i)
[10, 100[	5	0,125	0,125
[100, 190[	3	0,075	0,200
[190, 280[	11	0,275	0,475
[280, 370[	6	0,150	0,625
[370, 460[	11	0,275	0,900
[460, 550[	3	0,075	0,975
[550, 640]	1	0,025	1,000
Total	40	1,000	

Réponse :

(a) Pour déterminer le premier quartile, les fréquences relatives cumulées ont dépassé 25 % pour la première fois au niveau de la classe [190, 280[, donc :

$$Q_1 = L_i + \frac{0,25 - F_{Q_1-1}}{f_{r,Q_1}} \times (L_s - L_i) = 190 + \frac{0,25 - 0,20}{0,275} \times (280 - 190) = 206,36$$

Cela signifie que dans cet échantillon de données, 25 % des journées, les recettes quotidiennes de ce petit magasin ont été de 206,36 ou moins.

(b) Pour déterminer le deuxième quartile (médiane), les fréquences relatives cumulées ont dépassé 50 % pour la première fois au niveau de la classe [280, 370[, donc :

$$Q_2 = L_i + \frac{0,50 - F_{Q_2-1}}{f_{r,Q_2}} \times (L_s - L_i) = 280 + \frac{0,50 - 0,475}{0,150} \times (370 - 280) = 295$$

Cela signifie que dans cet échantillon de données, 50 % des journées, les recettes quotidiennes de ce petit magasin ont été de 295 ou moins.

(c) Pour déterminer le troisième quartile, les fréquences relatives cumulées ont dépassé 75% pour la première fois au niveau de la classe [370, 460[, donc :

$$Q_3 = L_i + \frac{0,75 - F_{Q_3-1}}{f_{r,Q_3}} \times (L_s - L_i) = 370 + \frac{0,75 - 0,625}{0,275} \times (460 - 370) = 410,91$$

Cela signifie que dans cet échantillon de données, 75 % des journées, les recettes quotidiennes de ce petit magasin ont été de 410,91 ou moins.

Il est à remarquer que les autres méthodes utilisées plus haut pour calculer la médiane sont aussi d'application (méthode d'interpolation linéaire, méthode des fréquences absolues).

2.1.5. Percentiles, quintiles et déciles

Précédemment, nous avons vu que les quartiles subdivisent la population en 4 parties d'effectifs égaux (contenant chacun le même pourcentage d'observations) ; il y en a 4 à savoir Q_1, Q_2 et Q_3. Quant aux percentiles, ils partagent la distribuion en 100 parties de taille égale, il y en a 99 à savoir P_1, P_2, …, P_{99} et entre deux percentiles (centiles ou pourcentiles) consécutifs, il y a 1 % des observations.Il est à noter que le premier quartile est le percentile 25 %, le deuxième quartile est le percentile 50 % et le troisième quartile le percentile 75 %. Pour les quintiles, ils subdivisent la distribution en 5 parties d'efectifs égaux, il y en a 4 à savoir V_1, V_2, V_3 et V_4 respectivement. Ces quintiles correspondent aux percentiles 20 %, 40 %, 60 % et 80 %. Autrement dit, entre deux

quintiles consécutifs, il y a 20 % d'observations. Les déciles sont des observations qui partagent la population en 10 parties d'effctifs égaux, il y en a 10 à savoir D_1, D_2, …, D_9. Ces déciles correspondent aux percentiles 10 %, 20 %, 30 %, …, 90 % respectivement et entre deux déciles consécutifs, il y a 10 % d'observations.

De même, les quintiles V_1, V_2, V_3 et V_4 correspondent aux déciles D_2, D_4, D_6 et D_8 respectivement. Le calcul de ces différentes mesures de position est identique à ce qui a été fait pour déterminer les quartiles ; il n'y a que le pourcentage de la mesure à adapter à chaque fois.

Exemple : En reprenant les données de l'exemple ci-après, déterminez le deuxième quintile, le septième décile et le quatre vingt quinzième centile de la variable X, les recettes quotidiennes d'un petit dépanneur et interprétez chacune de ces mesures.

Réponse :

(a) Les fréquences cumulées dépassent pour la première fois 40 % au niveau de la classe [190, 280[ainsi le deuxième quintile est égal à :

$$V_2 = 190 + \frac{0,40 - 0,20}{0,275} \times 90 = 255,45$$

Classe	Fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
[10, 100[	5	0,125	0,125
[100, 190[	3	0,075	0,200
[190, 280[	11	0,275	0,475
[280, 370[	6	0,150	0,625
[370, 460[	11	0,275	0,900
[460, 550[	3	0,075	0,975
[550, 640]	1	0,025	1,000
Total	40	1,000	

Ceci signifie que dans cet échantillon de données, 40 % des journées, les recettes quotidiennes de ce petit magasin ont été de 255,45 ou moins.

(b) Les fréquences relatives cumulées dépassent pour la première fois 70 % au niveau de la classe [370, 460[, ainsi le septième décile est égal à :

$$D_7 = 370 + \frac{0,70 - 0,625}{0,275} \times 90 = 394,55$$

Cela signifie que dans cet échantillon de données, 70 % des journées, les recettes quotidiennes de ce petit magasin ont été de 394,55 ou moins.

(c) Les fréquences relatives cumulées dépassent pour la première fois 95 % au niveau de la classe [460 ; 550[, ainsi le quatre vingt quizième centile est égal à :

$$P_{95} = 460 + \frac{0,95 - 0,90}{0,075} \times 90 = 520$$

Cela signifie que dans cet échantillon de données, 95 % des journées, les recettes quotidiennes de ce petit magasin ont été de 520 ou moins.

2.2. Paramètres de dispersion

Rappelons que les données qui nous intéressent sont celles issues d'un échantillon et que le choix de cet échantillon est fait au hasard mais sensé refléter ce qui se passe dans la population. Cela fait que le comportement d'une variable diffère d'un échantillon à l'autre tout en espérant qu'il correspond au profil de cette variable dans la population. Lorsqu'une variable mesurable est manipulée et que les mesures calculées sont des mesures de tendance centrale, alors l'analyse

perd de vue au jiveau de la variabilité des données autour de ces mesures centrales. Cela prouve l'utilité des mesures de dispersion qui, jumulées avec les mesures de tendance centrale, vont nous donner une idée plus exacte sur l'ensemble de ce qui a été observé dans une série échantillonnale. Dans ce paragraphe, il sera décrit quelques unes de ces mesures de dispersion.

2.2.1. Étendue

C'est la mesure de dispersion la plus simple à calculer. Lorsqu'une variable quantitative X est mesurée sur un échantillon de taille n, alors l'étendue est égale à la différence entre la plus grande et la plus petite opbservation :

$$E = x_{max} - x_{min} \tag{2.10}$$

Puisque l'étendue est basée seulement sur les deux observations extrêmes, alors elle est très peu utilisée dans les applications.

Exemple :

Soit la série statistique : -4 -3 -1 1 3 5

L'étendue vaut $E = x_{max} - x_{min} = 5 - (-4) = 9$

Dans le cas d'une série classée, l'étendue totale est la différence entre la borne supérieure

de la classe la plus haute et la borne inférieure de la classe la plus basse :

$$E = L_s(H) - L_i(B) \tag{2.11}$$

Exemple : En reprenant les données classées de l'exemple ci-haut, on a que :

$$E = L_s(H) - L_i(B) = 640 - 10 = 630$$

Il est possible de calculer l'étendue à un certain pourcentage central. On fait la différence entre deux pourcentiles. Par exemple, l'étendue réduite au 60 % central vaut :

$$E = P_{80} - P_{20} \tag{2.12}$$

2.2.2. Variance

La variance d'une variable mesurée sur un échantillon est égale à la moyenne des carrés des écarts qui séparent chaque observation de la moyenne échantillonnale. Son calcul diffère selon la nature des données (en vrac, goupées par valeurs ou classées).

Soit X une variable quantitative mesurée sur un échantillon de taille n, et dont les valeurs sont $x_1, x_2, ..., x_n$, alors la variance de l'échantillon est [1] :

$$s_x^2 = Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.13}$$

La sommation ci-dessus est divisée par (n-1) pour que cette variance échantillonnale soit une bonne estimation de la variance de toute la population (variance corrigée). Pour la variance-population, il suffit de diviser uniquement par *N*, la taille de la population. La variance se prête mal à l'interpétation car vue son calcul, son unité est égale au carré de l'unité de la variable X. Si par exemple X est égal au nombre d'enfants par ménage, alors l'unité de la variance serait $(nombre\ d'enfants)^2$ qui n'a aucune signification.

La variance est surtout utile lorsque les mesures d'une variable sont faites dans plusieurs groupes (analyse de la variance) ou dans le cas où l'analyste veut comparer plusieurs variables mesurées sur le même échantillon ou comme étape de calcul pour calculer d'autres mesures.

Après quelques développements, la variance s'écrit [5] :

$$s_x^2 = Var(x) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \tag{2.14}$$

En effet :

Sachant que :

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

il vient :

$$\begin{aligned} s_x^2 = Var(x) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2\bar{x}\,x_i + \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \end{aligned}$$

Exemple : Soit X une variable quantitative mesurée sur un échantillon de taille n=6 et les valeurs suivantes ont été obtenues : -2 5 10 7 8 8

Alors, la moyenne et la variance valent respectivement :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 6$$

$$s^2 = Var(x) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 18$$

Dans le cas des données groupées par valeurs, soit X une variable quantitative mesurée sur un échantillon de taille n, et dont les k valeurs sont : $x_1, x_2, ..., x_k$ avec des fréquences absolues respectivement égales à $n_1, n_2, ..., n_k$. Alors, la variance de X dans cet échantillon est égale à :

$$s_x^2 = Var(x) = \frac{1}{\left(\sum_{i=1}^{k} n_i\right) - 1} \sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^2$$

$$= \frac{1}{\left(\sum_{i=1}^{k} n_i\right) - 1} \sum_{i=1}^{k} n_i \left(x_i^2 - 2\bar{x}\, x_i + \bar{x}^2\right)$$

$$= \frac{1}{\left(\sum_{i=1}^{k} n_i\right) - 1} \left(\sum_{i=1}^{k} n_i x_i^2 - 2\bar{x}\sum_{i=1}^{k} n_i x_i + \bar{x}^2 \sum_{i=1}^{k} n_i\right)$$

$$= \frac{1}{\left(\sum_{i=1}^{k} n_i\right) - 1} \left(\sum_{i=1}^{k} n_i x_i^2 - 2\left(\sum_{i=1}^{k} n_i\right)\bar{x}^2 + \bar{x}^2 \sum_{i=1}^{k} n_i\right)$$

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

$$\Rightarrow \sum_{i=1}^{k} n_i x_i = \bar{x} \sum_{i=1}^{k} n_i$$

$$s_x^2 = Var(x) = \frac{1}{\left(\sum_{i=1}^{k} n_i\right) - 1} \sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^2 = \frac{1}{\left(\sum_{i=1}^{k} n_i\right) - 1} \left(\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \left(\sum_{i=1}^{k} n_i\right)\right)$$

Exemple : En reprenant le tableau de l'exemple utilisé plus haut, déterminer la variance de la variable X=le nombre d'accidents par semaine.

Tableau des fréquences du nombre d'accidents par semaine

x_i	Fréquences absolues (n_i)
0	4
1	2
2	10
3	7
4	10
5	4
6	3
Total	40

Réponse :

La moyenne de cette variable est :

$$\bar{x} = \frac{1}{\sum_{i=1}^{k} n_i} \sum_{i=1}^{k} n_i x_i = \frac{121}{40} = 3,025$$

x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
0	4	0	0	0
1	2	2	1	2
2	10	20	4	40
3	7	21	9	63
4	10	40	16	160
5	4	20	25	100
6	3	16	36	108
///	40	121	////	473

Donc, sa variance sera égale à :

$$s_x^2 = \frac{1}{\left(\sum_{i=1}^{k} n_i\right)-1}\left(\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \sum_{i=1}^{k} n_i\right) = \frac{473 - 40(3,025)^2}{6} = 2,74$$

Pour des données groupées par classes, soit maintenant X une variable quantitative mesurée sur un échantillon de taille n, et dont les observations sont groupées en k classes avec des fréquences absolues respectivement égales à $n_1, n_2, ..., n_k$ et dont les milieux des classes sont respectivement égaux à $x_1, x_2, ..., x_k$. Alors, la variance échantillonnale de cette variable est :

$$s_x^2 = Var(x) = \frac{1}{\left(\sum_{i=1}^{k} n_i\right)-1}\sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \frac{1}{\left(\sum_{i=1}^{k} n_i\right)-1}\left(\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \sum_{i=1}^{k} n_i\right) \quad (2.15)$$

Exemple : En reprenant les données de l'exemple ci-haut, déterminer la variance de la variable X, les recettes quotidiennes d'un petit dépanneur.

Réponse : La moyenne trouvée pour cette variable est $\bar{x} = 298$.

x_i	Classe	Fréquences absolues n_i	$n_i x_i^2$
55	[10, 100[	5	15125
145	[100, 190[	3	63075
235	[190, 280[	11	607475
325	[280, 370[	6	633750
415	[370, 460[	11	1894475
505	[460, 550[	3	765075
595	[550, 640]	1	354025
	Total	40	4333000

Alors, la variance de cet échantillon est égale à :

$$s_x^2 = Var(x) = \frac{4333000 - 40 \times (298)^2}{39} = 20021,54$$

2.2.3. Déviation standard

La déviation standard (ou écart-type échantillonnal) d'une variable quantitative mesurée sur un échantillon est égale à la racine carrée de sa variance. Son unité de mesure étant la même que celle de la variable, l'écart type se prête alors aisément à l'interprétation et est considéré comme la mesure de dispersion par excellence. La variance n'est donc qu'une étape de calcul pour déterminer l'écart-type, quand les calculs se font à la main. Maintenant que tout est programmé, aucune calculatrice et aucun logiciel ne parle de variance comme telle.

Exemple : L'écart-type échantillonnal pour les 3 précédents exemples où les variances échantillonnales ont été calculées est respectivement égal à :

$S_x = \sqrt{18} = 4,24$ pour les données en vrac

$S_x = \sqrt{274} = 1,655$ pour les données groupées par valeurs

$S_x = \sqrt{20021,54} = 141,497$ pour les données groupées par classes

Interprétation de l'écart-type échantillonnal :

L'écart-type mesure la dispersion entre toutes les valeurs observées. Des valeurs proches donneront un plus petit écart-type, alors que des données très séparées donneront un plus grand écart-type. Lorsque la distribution des données (histogramme ou polygone des fréquences ou autre) a une forme en cloche et que la taille de l'échantillon est supérieure à 100, alors 68 % des données observées sont comprises entre la moyenne plus ou moins un écart-type et 95 % des données observées sont comprises entre la moyenne plus ou moins deux écarts-types. Il est possible d'estimer l'écart- type par la formule suivante :

$$S_x \approx \frac{\text{Étendue de } x}{4} \qquad (2.16)$$

S'agissant des propriétés de l'écart-type échantillonnal, soit X une variable quantitative dont l'écart-type échantillonnal est S_x et soit Y une autre variable quantitative telle que $Y = a + bX$ où a et b sont des constantes réelles. Alors, l'écart-type échantillonnal de Y sera égal à :

$$S_Y = |b| S_X \tag{2.17}$$

Exemple : Reprenons le contexte de l'exemple ci-haut, où X est le nombre d'heures qu'un étudiant travaille à temps partiel par semaine. Supposons qu'à partir d'un échantillon d'étudiants, l'écart-type du nombre d'heures travaillées par ces étudiants soit égal à $S_x = 3,2$ heures/semaine. Si le salaire horaire est de 10 et que les patrons de ces étudiants leur offrent 30 par semaine pour leurs déplacements, quel est l'écart-type du gain net hebdomadaire de ces étudiants ? Posons Y, le gain net hebdomadaire de ces étudiants alors $Y = 30 + 10X$, donc l'écart-type du gain net de cet échantillon d'étudiants sera égal à $S_Y = 10S_X = 32$ par semaine.

2.2.4. Coefficient de variation

L'unité de l'écart-type d'une variable est la même que celles des données et qu'alors il s'interprète mieux que la variance. Mais, si l'objectif est de comparer la dispersion de deux variables ou plus ayant des unités différentes mesurées sur le même échantillon ou sur des échantillons différents, il nous faut une mesure de dispersion sans unité. Cette mesure est le coefficient de variation.

Pour un échantillon de données dont la moyenne est non négative, le coefficient de variation d'une variable X se définit par :

$$CV = \frac{s_x}{\bar{x}} \tag{2.18a}$$

ou

$$CV = \frac{s_x}{\bar{x}} \times 100\,\% \tag{2.18b}$$

Pour un échantillon, si le coefficient de variation de X est inférieur à 15 %, alors la variable est homogène, sinon elle est dite hétérogène. Pour deux échantillons (sur une ou deux variables) ou plus, alors celui qui a le plus petit coefficient de variation est le plus homogène.

Exemple : Prenons un échantillon de taille n=50 d'hommes d'âge adultes sur lesquels le poids et la taille ont été mesurés. Les résultats sont résumés dans le tableau suivant :

Variable	Moyenne	Écart type
X=taille	$\overline{x} = 173,59\ cm$	$s_x = 7,86\ cm$
Y=poids	$\overline{y} = 78,42\ kg$	$s_y = 11,98\ kg$

Le coefficient de variation qui permet de comparer l'homogénéité de ces deux variables est :

$$CV = \frac{7,86}{173,59} \times 100\,\% = 4,53\,\%$$

$$CV = \frac{11,98}{78,42} \times 100\,\% = 15,28\,\%$$

Donc, la taille des hommes adultes est plus homogène que leur poids, ce qui correspond à l'intuition. Par exemple, il est très rare de voir deux hommes adultes dont l'un serait deux fois plus grand que l'autre, alors qu'il est fréquent de voir un homme adulte dont le poids est le double d'un autre.

2.2.5. Écart moyen absolu

L'écart moyen absolu (EMA) est donné par :

$$e_m = \begin{cases} \dfrac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n} \ \textit{pour une série simple} \\ \dfrac{\sum_{i=1}^{k} n_i |x_i - \overline{x}|}{\sum_{i=1}^{k} n_i} \ \textit{pour une série pondérée} \end{cases} \qquad (2.19)$$

Les autres paramètres de dispersion sont :

L'intervalle interquartile : $IIQ = [Q_1,\ Q_3]$

L'écart interquartile : $EIQ = Q_3 - Q_1$

La déviation quartile : $DQ = \dfrac{Q_3 - Q_1}{Q_2}$

L'écart semi-interquartile : $ESIQ = \frac{Q_3 - Q_1}{2}$

Ces mesures peuvent aussi concerner les déciles, les percentiles et les quintiles.

2.3. Paramètres de forme

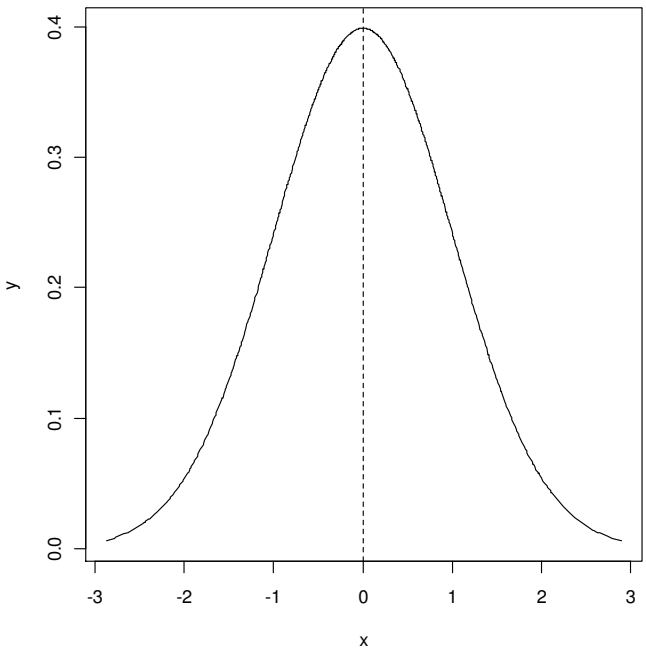
Les paramètres de forme, qui sont des moments d'ordre supérieur à 2, donnent la forme de la distribution. Ils montrent si la distribution est symétrique, asymétrique, normale ou pas normale.

2.5.1. Skewness

Le skewness est un moment centré d'ordre 3. Appelé également coefficient d'asymétrie, il est donné par :

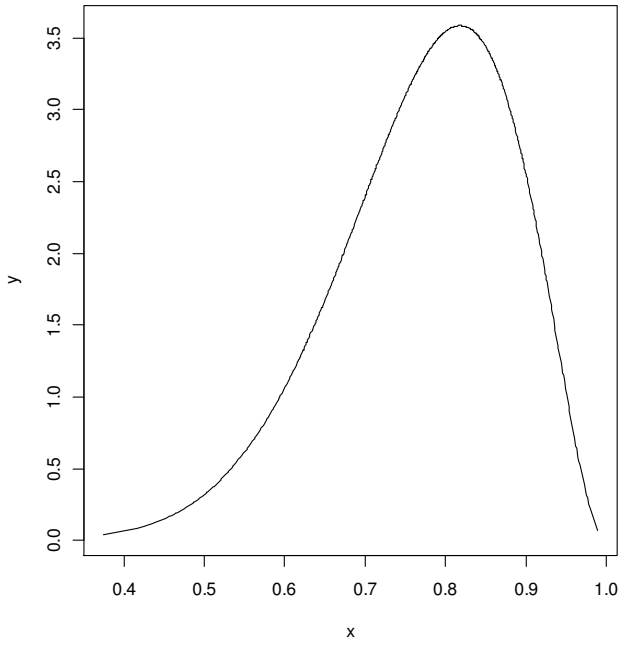
$$\alpha_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}\left(\sqrt{\frac{n}{n-1}}\right)^3 \tag{2.20}$$

- Si $\alpha_3 = 0$, alors la distribution est symétrique (moyenne, mode, médiane confondus)



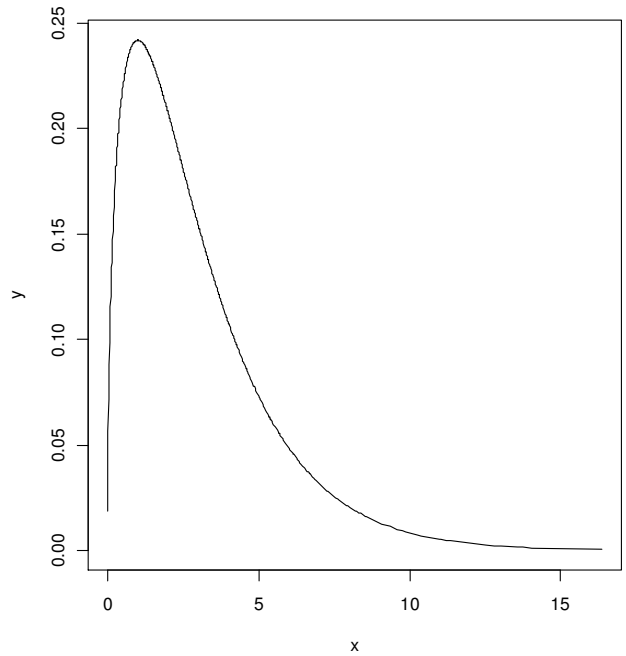
Source : Auteur à partir des données simulées

- Si $\alpha_3 < 0$, alors la distribution est dissymétrique vers les valeurs basses ;



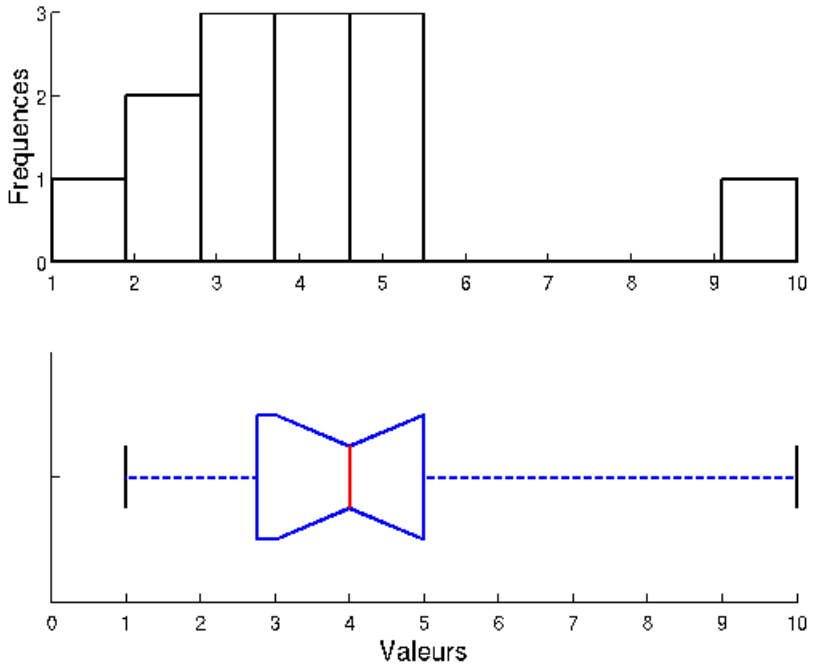
Source : Auteur à partir des données simulées

- Si $\alpha_3 > 0$, alors la distribution est dissymétrique vers les valeurs très élevées.



Source : Auteur à partir des données simulées

Le signe du skewness peut être remarqué en regardant l'histogramme ou la boîte à moustaches :



2.5.2. Coefficient d'asymétrie de Fisher

Les moments centrés d'ordre 3 et 4 s'écrivent respectivement :

$$\mu_3 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^3 \tag{2.21}$$

$$\mu_4 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4 \tag{2.22}$$

Le coefficient d'asymétrie de Fisher est donné par :

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \sqrt{n}\,\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{3}{2}}} \tag{2.23}$$

où μ_3 est le moment centré d'ordre 3, μ_2 le moment centré d'ordre 2, σ l'écart-type et n le nombre d'observations.

- Si $\gamma_1 = 0$, alors la distribution est symétrique ;
- Si $\gamma_1 < 0$, alors la distribution est plus étalée sur la gauche ;
- Si $\gamma_1 > 0$, alors la distribution est plus étalée sur la droite.

2.5.3. Coefficient d'asymétrie de Yule et Kendall

Il est donné par :

$$Y_K = \frac{Q_1 + Q_3 - 2Me}{2Me} \tag{2.24}$$

Il peut aussi s'écrire :

$$Y_K = \frac{Q_1 + Q_3 - 2Me}{2Me} = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Me + Q_1}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Me}{2Me}$$

2.5.4. Coefficients de dissymétrie de Pearson

Les premier et deuxième coefficients de dissymétrie de Pearson sont donnés par :

$$\begin{aligned} CD_1 &= \frac{\bar{x} - Mo}{s} \\ CD_2 &= \frac{3(\bar{x} - Me)}{s} \end{aligned} \tag{2.25}$$

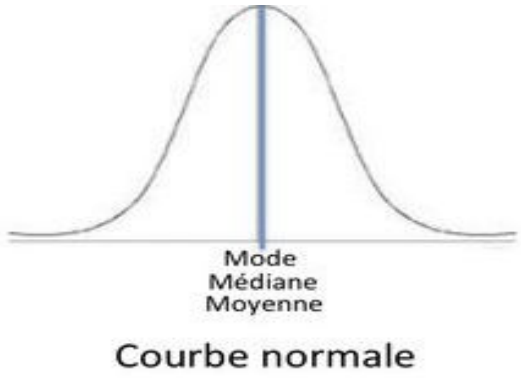
2.5.5. Kurtosis

Le kurtosis (kurtose), appelé aussi aplatissement ou coefficient d'aplatissement, est un paramètre de forme donné par :

$$\alpha_4 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)^2 s^4} - 3 \tag{2.26}$$

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

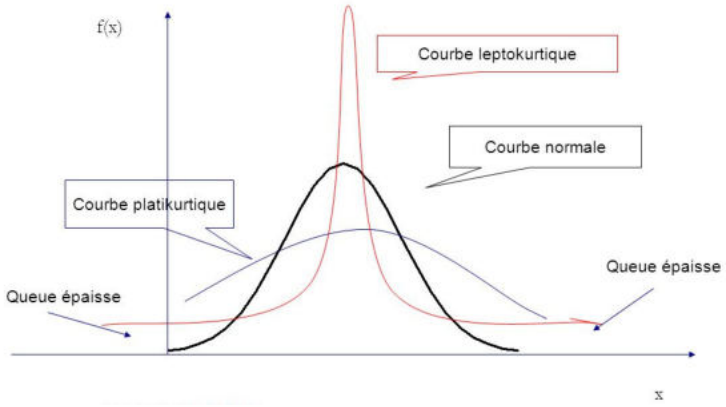
- Si $\alpha_4 = 0$, alors la distribution de la variable suit une loi normale.



Courbe normale

- Si $\alpha_4 < 0$, alors la distribution de la variable (courbe bleue) passe en dessous de la courbe normale, ce qui veut dire que la courbe est plus aplatie que la courbe normale (courbe noire).

-Courbes platikurtique, leptokurtique et normale



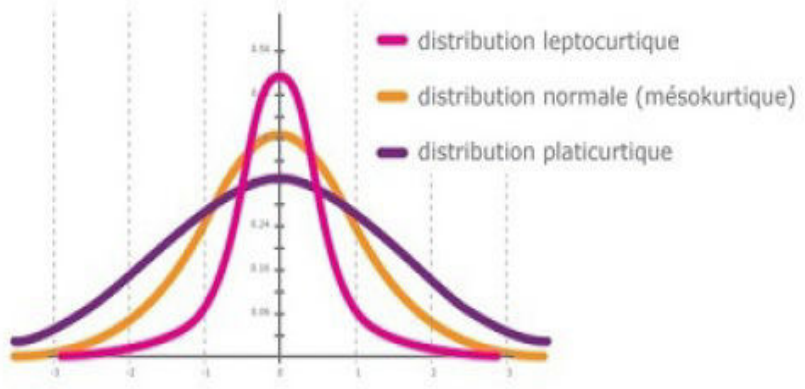
- Si $\alpha_4 > 0$, alors la distribution de la variable (courbe rouge) passe au-dessous de la courbe normale, ce qui veut dire que la courbe est moins aplatie que la courbe normale (courbe noire).

2.5.6. Coefficient d'aplatissement de Fisher

Il est donné par :

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3 \qquad (2.27)$$

- Si $\gamma_2 = 0$, alors la distribution est normale (courbe mesokurtique) ;
- Si $\gamma_2 > 0$, alors la distribution est moins aplatie (courbe leptokurtique) ;
- Si $\gamma_2 < 0$, alors la distribution est plus aplatie (courbe platikurtique).



2.4. Paramètres de concentration

2.4.1. Médiale

La médiale se calcule pour une série classée. Pour la calculer, il suffit de pondérer les effectifs par les valeurs centrales des classes. Par exemple, considérons la série statistique suivante représentant les tranches de salaire des ouvriers.

Tranche de salaire (en milliers de Fbu)	Nombre de salariés
De 0 à moins de 10	4
De 10 à moins de 20	3
De 20 à moins de 30	2
De 30 à moins de 40	1

Le tableau ci-dessous est un tableau montrant les centres de classes, les effectifs qui tombent dans chaque classe, les centres de classes, les effectifs pondérés par les centres de classe et le cumul de cette dernière colonne.

C_i	n_i	x_i	n_ix_i	n_ix_i cum
[0, 10[	4	5	20	20
[10, 20[	3	15	45	65
[20, 30[	2	25	50	115
[30, 40[	1	35	35	150
Somme	////	////	**150**	//////////////

La classe médiale est [20, 30[.

La médiale vaut :

$$Ml = L_i + \frac{\frac{\sum_{i=1}^{n} n_i x_i}{2} - \left(\sum_{i=1}^{n} n_i x_i\right)_{Ml-1}}{\left(\sum_{i=1}^{n} n_i x_i\right)_{Ml}} \times \left(L_s - L_i\right) = 20 + \frac{75-65}{50} \times 10 = 20 + \frac{10}{50} \times 10 = 22$$

2.4.2. Indice de GINI

L'indice de GINI et la courbe de LORENTZ sont deux indicateurs complémentaires qui permettent notamment l'analyse de la répartition d'une masse de salaire, de revenu, de richesse et de ressources dans une population.

Considérons par exemple le tableau ci-après.

Salaire mensuel (C_i)	Nombre d'ouvriers (n_i)	Fréquence relative (f_i)	Fréquence relative
[500, 1500[	50	0,250	0,250
[1500, 2500[	125	0,625	0,875
[2500, 5500[	25	0,125	1,000
Total	200	//////////////////////////////////	//////////////////////////////

De ce tableau, il vient :

Salaire mensuel (C_i)	Nombre d'ouvriers (n_i)	Centre de classe (x_i)	Masse salariale (n_ix_i)	Pourcentage de masse salariale (g_i)	Pourcentage cumulé de masse salariale $F(n_ix_i)$
[500, 1500[	50	1000	50000	0,125	0,125
[1500, 2500[	125	2000	250000	0,625	0,750
[2500, 5500[	25	4000	100000	0,25	1,000
Total	200	//////////////	400000	//////////////////////////	//////////////////////////////

En mettant $F(x_i)$ en abscisses et $F(n_i x_i)$ en ordonnées dans un carré de superficie unité, nous obtenons la courbe appelée « courbe de LORENTZ » (courbe en dessous de la bissectrice). Le double de l'aire de surface du polygone (en rouge) donne l'indice de GINI.

L'aire du triangle A vaut :

$$S_A = \frac{0,25 \times 0,125}{2} = 0,0156$$

L'aire du trapèze B vaut :

$$S_B = \frac{(0,875 - 0,25) \times (0,125 + 0,75)}{2} = 0,2734$$

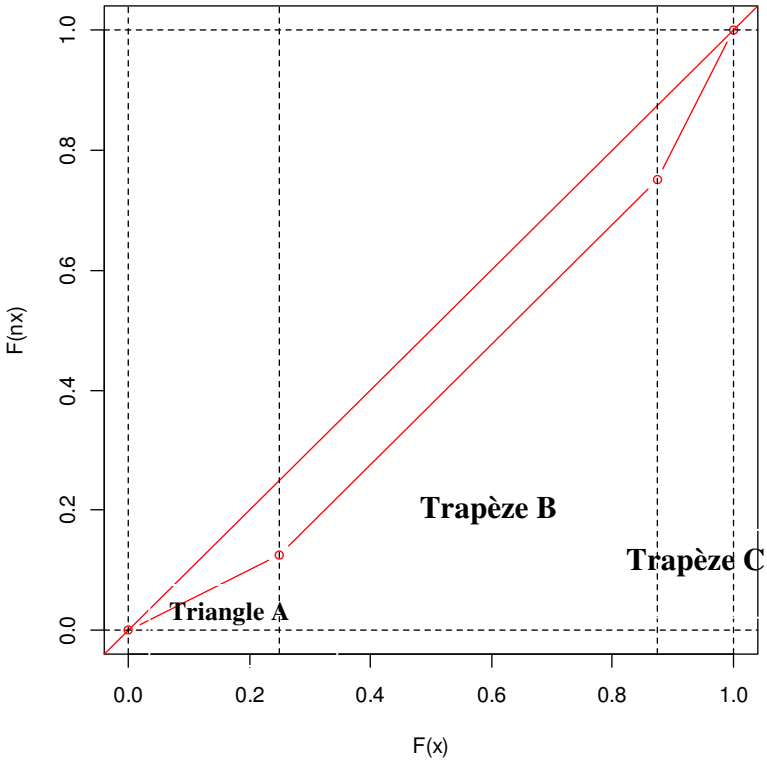
L'aire du trapèze C vaut :

$$S_C = \frac{(1 - 0,875) \times (0,75 + 1)}{2} = 0,1094$$

L'indice de GINI vaut :

$$G = 2 \times \left(\frac{1}{2} - 0,0156 - 0,2734 - 0,1094 \right) = 0,2030 \simeq 0,20$$

L'indice de Gini appartient dans l'intervalle [0,1]. Plus il est proche de 1, alors plus la répartition des salaires est inégalitaire. Comme la valeur de cet indice (0,20) est faible, alors nous concluons que la concentration des salaires est faible. Autrement dit, les salaires sont assez bien répartis sur l'ensemble des salariés.



Source : Auteur à partir des données des tableaux ci-dessus

Exercices d'application - 1

Exercice 1 :

Comparer la moyenne, la médiane et le mode à l'aide d'un tableau comparatif.

Exercice 2 :

Une étude sur le budget consacré aux vaca
suivants :

Budget X	Fréquence cumulée	Fréquences
[800, 1000[	0,08	?
[1000, 1400[	0,18	?
[1400, 1600[	0,34	?
[1600, β [	0,64	?
[β , 2400[	0,73	?
[2400, α [	1	?

Le travail demandé :

— Certaines données sont manquantes. Calculer la borne manquante α sachant que l'étendue de la série est égale à 3200.
— Calculer les fréquences dans le tableau.
— Calculer la borne manquante β dans les deux cas suivants :

1. Le budget moyen est égal à 1995.

Exercice 3 :

Le tableau statistique suivant représente les montants, en FBu, des ventes réalisées au mois de Janvier par des employés d'une société :

Modalité C_i	Effectifs n_i	Fréquences f_i	Effectifs cumulés **?**	Fréquences cumulées
[0, 1000[	20	**?**	**?**	**?**
[1000, 2000[	**?**	0,35	**?**	**?**
[2000, **?**[	**?**	**?**	**?**	**?**
[**?**, **?** [	15	**?**	100	**?**
Total	**?**	**?**	**?**	**?**

1. Complétez le tableau statistique précédent.
2. Déterminer la variable statistique étudiée et son type.
3. Déterminer le pourcentage des ventes qui ont réalisé un montant entre 0 et 2000 FBu.
4. Représenter graphiquement les effectifs et les fréquences cumulées.
5. Calculer la moyenne arithmétique et donner sa signification.
6. Trouver le mode et la médiane graphiquement, puis par le calcul, et donner leurs significations.
7. Calculer la variance et l'écart-type de cette série statistique.

Exercice 4 :

Pour déterminer le type de logement (F2, F3, ...) à construire, on étudie 20 familles selon leur nombre d'enfants. Durant l'expérience, on note les résultats suivants :

1, 3, 5, 5, 3, 2, 4, 4, 7, 0, 2, 4, 3, 7, 0, 5, 4, 2, 3, 2

– Déterminer la population, l'unité (individu), la variable statistique et les modalités.

– Déterminer le tableau statistique avec x_i, n_i, f_i et F_i.

Modalité	0	1	2	3	4	5	6	7	$\sum$
n_i									
N_i									
f_i									
F_i									

Exercice 5 :

Soit la série statistique suivante :

Âge (années) (X)	2	8	6	8	10	4	x_7	2	6	10

où x_7 est une donnée manquante.

a) Calculez la valeur de x_7 pour que la moyenne soit égale à 7.
b) En partant de la valeur x_7 trouvée en a), calculez la déviation standard.
c) En supposant que la valeur de x_7 vaut 25, que vaudra le mode ?
d) En utilisant la valeur trouvée en a), que vaudra l'étendue totale de la série ?

Exercice 6 :

Les 35 élèves d'une classe ont composé et le tableau ci-dessous donne la répartition des diverses notes.

Note	2	4	5	6	9	11	12	14	15	16	18
Effectif	1	3	2	2	6	4	4	5	3	3	2

Calculez la note moyenne, la note dominante, la déviation standard et l'étendue.

Exercice 7 : Choix multiple (Justifiez votre choix par des calculs)

Dans le but de décider si oui ou non il va inscrire son école à une compétition interscolaire d'athlétisme dans laquelle il y a une course relais, un professeur propose de chronométrer 21 élèves de l'école sur un parcours de mille mètres. Les résultats obtenus sont les suivants (en minutes et dixièmes de minutes) :

7,4	7,2	7,0	6,8	6,6	6,4	7,6	7,8	8,0	8,4
14,2	13,0	11,6	10,2	10,0	9,6	9,4	9,2	9,0	8,8
15,0									

Le jour de ce test, cinq élèves qui avaient été sélectionnés par le professeur étaient absents. Ces cinq élèves ont réalisé le test le lendemain et le professeur a enregistré un temps de parcours moyen de 10,0 minutes avec ces cinq élèves.

a) Pour l'ensemble des 26 élèves, le temps moyen de parcours était de :

A: 9,85 min B: 9,20 min C: 9,35 min D: 10,00 min E: 9,60 min

b) Trois quarts des 21 élèves ont réalisé ce parcours en moins de :

A: 10,1 min B: 10,0 min C: 9,6 min D: 13,0 min E: 11,6 min

c) La déviation standard des temps de parcours des 21 élèves ayant participé au test est de :

A: 5,99 min B: 2,39 min C: 0,53 min D: 2,45 min E: 9,20 min

d) Sur ce parcours, la médiane des temps relevés pour ces 21 élèves est :

A: 8,8 min B: 9,0 min C: 8,4 min D: 8,0 min E: 10,0 min

De nombreux exercices d'entrainement sur le calcul des statistiques de tendance centrale, de dispersion et de forme peuvent être trouvés dans le livre de **Jean-Louis Monino**, Professeur à l'Université de Montpellier (France) [6].

Chapitre 3 : Représentations graphiques

Dans ce chapitre, nous allons détailler la manière de résumer l'information contenue dans une série de données non pas par des tableaux comme il a été fait pour le chapitre précédent, mais plutôt par des graphiques. *Un bon graphique vaut mieux qu'une montagne de chiffres*, se dit en statistique. De même, *un bon graphique vaut mieux qu'un long discours*.

3.1. Variables discrètes

3.1.1. Diagramme en bâtons

Soit X une variable quantitative discrète dont le nombre de modalités n'est pas trop grand. Alors, il est possible de dresser un tableau des fréquences comme celui utilisé pour les variables qualitatives auquel une colonne supplémentaire relative aux fréquences relatives et une colonne relative aux fréquences relatives cumulées peuvent être ajoutées. En ce qui concerne la représentation graphique, un seul graphique s'associe avec les variables quantitatives discrètes : **le diagramme à bâtons.**

Le diagramme en bâtons, appelé également diagramme en bâtonnets, se construit en portant en abscisse les valeurs de la variables discrète et en ordonnée la fréquence de chaque observation. Il suffit alors de tracer des bâtons dont la hauteur est proportionnelle à la fréquence de chaque modalité. Remarquons que les bâtons ne doivent pas avoir d'épaisseur, car la variable prend exactement les valeurs 0, 1, 2,…. Il est possible, cependant, d' ajouter les effectifs ou les fréquences relatives sur les bâtons.

Exemple 1 : Un inspecteur en contrôle de qualité a extrait de sa base de données, un échantillon de 40 semaines où il a noté X, le nombre d'accidents de travail enregistrés par semaine. Il a obtenu les résultats suivants :

2 0 4 2 2 1 3 2 0 5 4 3 2 4 5 6 6 4 2 0
3 4 4 2 6 2 4 3 0 4 3 4 3 3 5 5 4 2 2 1

Il est possible de dresser le tableau des fréquences ci-après.

Tableau 6 : Fréquences du nombre d'accidents par semaine

Nombre d'accidents par semaine	Fréquences absolues (n_i)	Fréquences relatives (f_i)	Fréquences relatives cumulées (F_i)
0	4	0,100	0,100
1	2	0,050	0,150
2	10	0,250	0,400
3	7	0,175	0,575
4	10	0,250	0,825
5	4	0,100	0,925
6	3	0,075	1,000
Total	40	1,000	///////

Pour cette série pondérée, les modes (nombres d'accidents dominants) sont :

Mo=2

Mo=4

La série est donc bimodale. Ces modes peuvent aussi être visualisés sur un graphique appelé diagramme en bâtons. Pour le faire avec le logiciel R, il suffit de créer une séquence portant les valeurs de cette variable qui représente le nombre d'accidents par semaine et de répéter chaque valeur de cette séquence autant de fois que cela est précisé dans la colonne des fréquences absolues ou effectifs. Cette nouvelle variable sera tabulée pour avoir une série statistique pondérée dont la distribution sera facile à représenter, soit par un diagramme en bâtons, soit par un diagramme en points ou un polygone des fréquences. La figure 1 montre le diagramme en bâtons du nombre d'accidents par semaine.

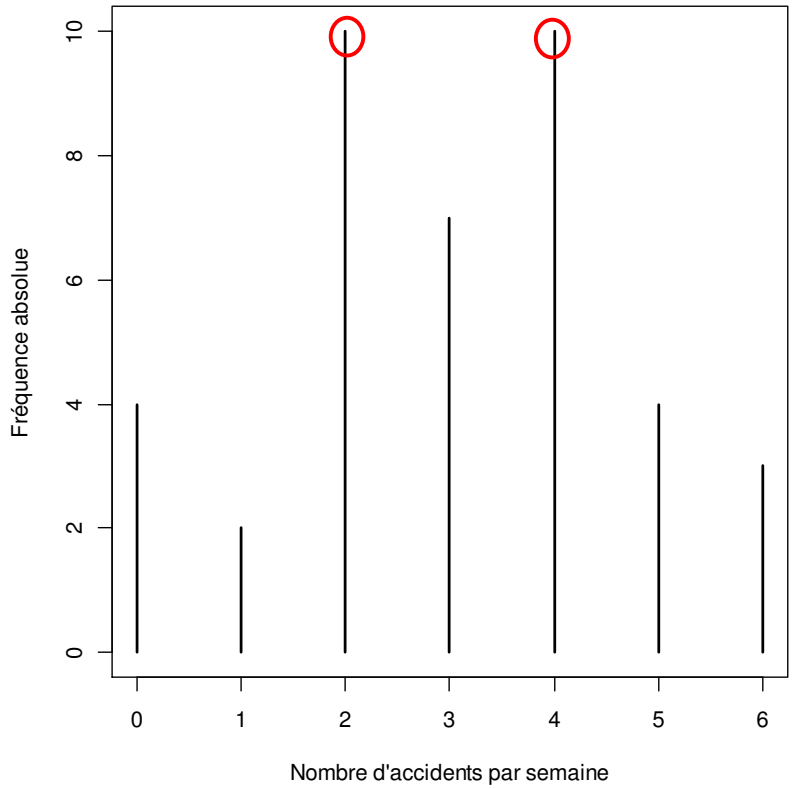


Figure 1 : Diagramme en bâtons du nombre d'accidents par semaine

Exemple 2 : Considérons l'exemple ci-après.

Nombre de pièces	1	2	3	4	5	6	7	8		Total
Effectifs	10	15	15	30	10	12	5	3		100

Tableau 7 : Fréquences du nombre de pièces d'un logement

	Nombre de pièces	Effectifs	Effectifs cumulés
	1	10	10
	2	15	25
	3	15	40
	4	30	70
	5	10	80
	6	12	92
	7	5	97
	8	3	100
Σ	////	100	/////

La fréquence maximale vaut 30, ce qui montre que le mode vaut 4.

La figure 2 montre le diagramme en bâtons du nombre de pièces d'un logement.

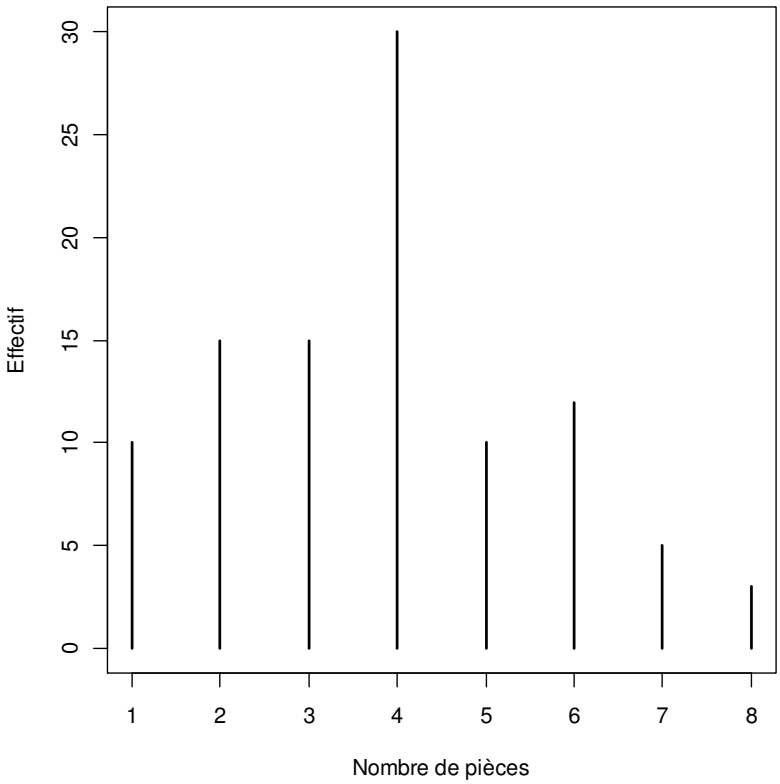


Figure 2 : Diagramme en bâtons du nombre de pièces d'un logement

Sur ce diagramme, le pic est observé au point d'abcisse4, ce qui montre que le mode vaut 4.

3.1.2. Diagramme en points

Le diagramme en points est semblable au disgramme de dispersion. La seule différence est que le diagramme de dispersion croise deux variables x (en abscisse) et y (en ordonnée), alors que pour le diagramme en point, ce qui est mis en abscisse est la valeur de la variable à représenter sous forme de points.

Exemple : Un inspecteur en contrôle de qualité a extrait de sa base de données, un échantillon de 40 semaines où il a noté X, le nombre d'accidents de travail enregistrés par semaine. Il a obtenu les résultats suivants :

2 0 4 2 2 1 3 2 0 5 4 3 2 4 5 6 6 4 2 0
3 4 4 2 6 2 4 3 0 4 3 4 3 3 5 5 4 2 2 1

Partons du tableau des fréquences ci-après :

Tableau des fréquences du nombre d'accidents par semaine

Le nombre d'accidents par semaine	Fréquences absolues (n_i)	Fréquences relatives (f_i)	Fréquences relatives cumulées (F_i)
0	4	0,100	0,100
1	2	0,050	0,150
2	10	0,250	0,400
3	7	0,175	0,575
4	10	0,250	0,825
5	4	0,100	0,925
6	3	0,075	1,000
Total	40	1,000	////////

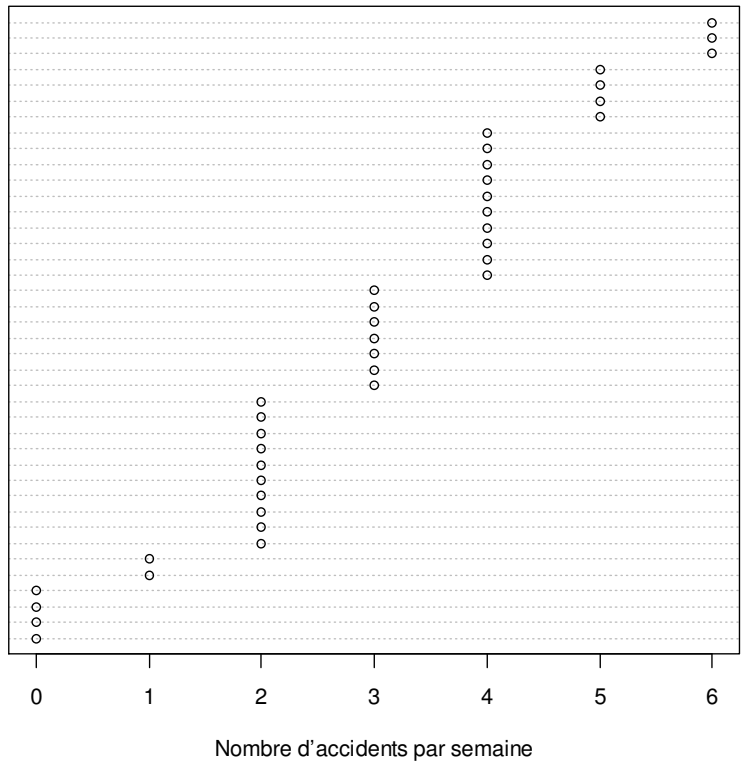


Figure 3 : Diagramme en points du nombre d'accidents par semaine

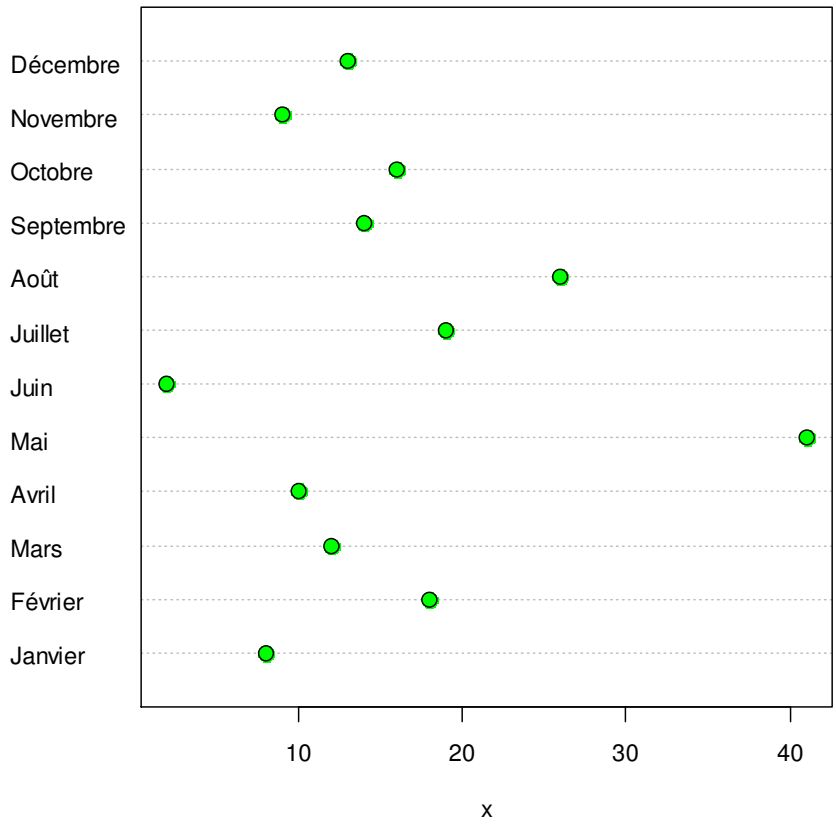


Figure 4 : Diagramme en points du nombre d'accidents par mois

3.1.3. Polygone des fréquences

Le polygone des fréquences consiste à joindre le milieux des sommets des rectangles d'un histogramme par une ligne en zig-zag et cette ligne se ferme en ajoutant aux deux extrémités deux classes fictives de même amplitude que les autres. Comme cela, la surface délimitée par l'histogramme est identique à celle délimitée par le polygone des fréquences. Le polygone de fréquences est très utile quand l'analyse veut comparer le comportement de la même variable mesurée sur plusieurs groupes (il est possible de penser à comparer les recettes des deux sites touristiques par exemple) ou la même variable mesurée sur le même échantillon à différents instants (il est possible de comparer les recettes d'un site touristique selon des années).

Un exemple d'illustration est le polygone des fréquences du nombre d'accidents par semaine (figure 5).

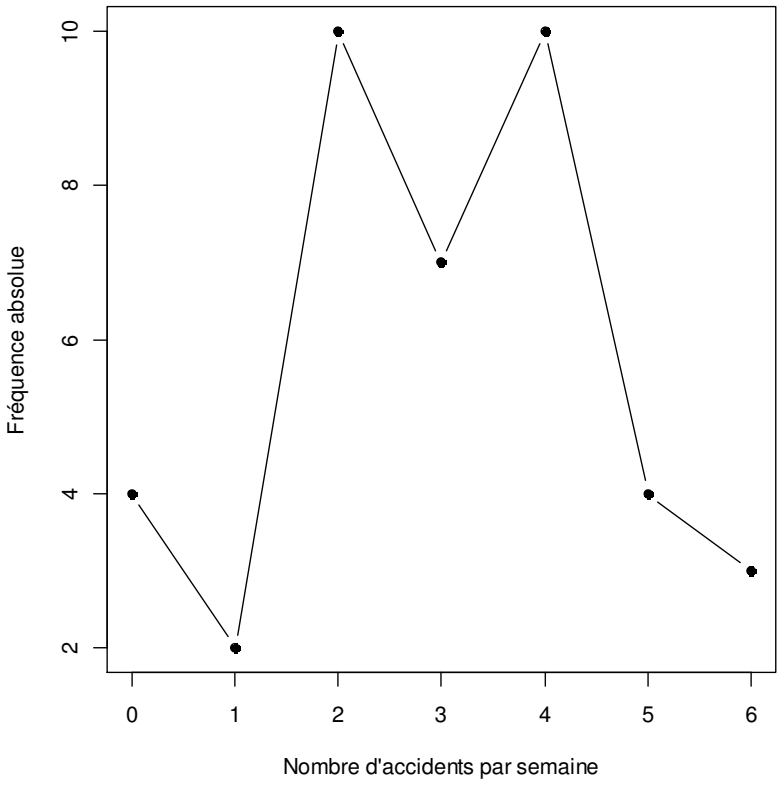


Figure 5 : Polygone des fréquences du nombre d'accidents par semaine

3.1.4. Fonction de répartition

La fonction de répartition d'une variable x est une fonction cumulative. Pour la construite, les effectifs ou les fréquences relatives sont cumulées de façon croissante. En abscisse figure les valeurs de la variable à représenter et en ordonnée, les fréquences cumulées N_i.

Cette fonction est telle que :

$$F(x)=\begin{cases} 0\ si\ x < x_1 \\ N_1\ si\ x_1 \le x < x_2 \\ N_2\ si\ x_2 \le x < x_3 \\ \vdots \\ N_i\ si\ x_{i-1} \le x < x_i \\ \vdots \\ n\ si\ x \ge x_n \end{cases} \qquad (3.1)$$

La figure 6 montre la fonction de répartition du nombre de pièces d'un logement. La valeur modale vaut 4.

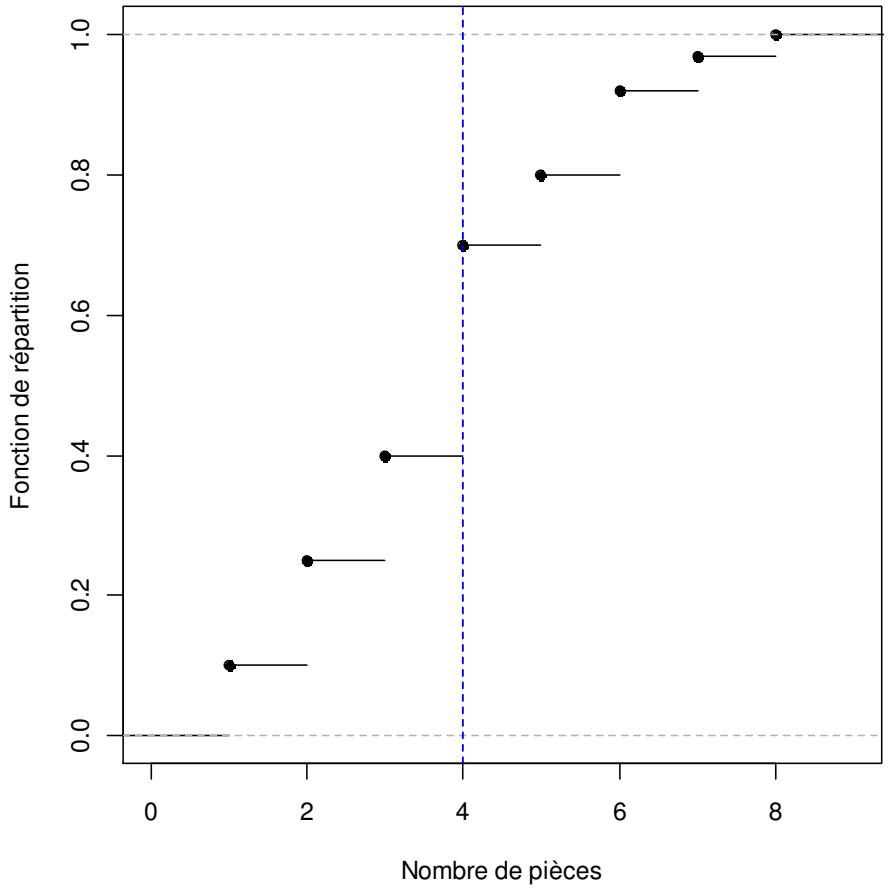


Figure 6 : Fonction de répartition du nombre de pièces d'un logement

3.2. Variables continues

Une variable quantitative continue est, dans la plupart des cas, représentée soit par un histogramme, soit par une boîte à moustaches. Un histogramme montre la symétrie de la distribution, alors qu'une boîte à moustaches permet d'examiner la symétrie et de détecter les observations aberrantes (outliers), c'est-à-dire des observations qui ne sont pas du même ordre de grandeur que les autres [7]. Ce sont des observations qui sont soit anormalement petites, soit anormalement grandes.

3.2.1. Histogramme

L'histogramme est un graphique qui s'adapte à la représentation d'une variable quantitative continue avec un grand nombre de modalités. Il s'agit d'une suite de rectangles juxtaposés les uns

aux autres et dressés au-dessus de chacune des classes, dont la largeur est égale à l'amplitude de la classe (prise comme unité de mesure) et dont la surface reflète la fréquence absolue (histogramme de fréquences absolues) ou relative (histogramme de fréquences relatives) de la classe qu'il représente [1]. Autrement dit, les classes de la variable à représenter sont en abscisse et les rectangles dont les hauteurs sont proportionnelles à la fréquence de chaque classe sont tracés. Il est aussi possible de construire un histogramme de densité de fréquences, ces dernières étant calculées en divisant chaque fréquence relative par l'amplitude de la classe.

En outre, il est possible de l'utiliser dans de rares cas pour un échantillon de données provenant d'une variable quantitative discrète. L'idée principale est de grouper ces données en classes de valeurs.

Avant de le faire, deux questions majeurs se posent alors :

- Combien de classes faut-il former ?
- Quelles seront les largeurs de chacune des classes ?

La réponse à la première question dépend de la taille *n* de l'échantillon. Le nombre de classes *K* à former est donné par exemple par la formule de **Sturges** suivante :

$$K = 1 + \frac{10}{3} log(n) \tag{3.2}$$

Ainsi, par exemple, si n=150, il faut former $K = 1 + \frac{10}{3} log(150) = 8,2536 \simeq 9$ (l'arrondissement se fait à l'entier immédiatement supérieur). Une fois que le nombre de classes à former est trouvé, il reste à former des classes de même amplitude (largeur) et cette amplitude sera égale à différence entre l'observation la plus grande (maximum) et l'observation la plus petite (minimum) divisé par le nombre de classes (K) selon la formule :

$$A = \frac{x_{max} - x_{min}}{K} \tag{3.3}$$

Cette amplitude sera arrondie selon les données pour avoir des bornes de classes faciles à manipuler. Il est possible de coller un polygone des fréquences absolues ou un polygone des fréquences relatives sur l'histogramme afin d'examiner la symétrie de la distribution.

Exemple : Considérons les données relatives aux recettes quotidiennes (en dollars) d'un petit magasin. Un échantillon de taille n=40 jours a été sélectionné au hasard et a donné les résultats suivants :

16,00 58,50 68,20 78,00 79,45 142,20 145,3 186,70 209,05 216,75

219,70 247,75 249,10 256,00 257,15 262,35 268,60 269,60 270,15 284,45

319,00 332,00 343,29 350,75 354,90 372,60 383,20 389,20 404,55 420,20

428,50 432,40 444,60 446,80 456,10 458,10 493,95 511,95 521,05 621,35

Le nombre de classes à former est $K = 1 + \frac{10}{3} log(40) = 6,34 \simeq 7$ classes d'amplitude chacune égale à $A = \frac{621,35 - 16}{7} = 86,48 \simeq 90$ .Cette amplitude est arrondie à 90, ce qui donne le tableau des fréquences suivant, où les classes sont des intervalles fermés à gauche et ouverts à droite sauf le dernier qui est un intervalle fermé des deux côtés.

Les figures 7 et 8 montrent l'histogramme des fréquences absolues et l'histogramme des fréquences relatives (avec courbe de normalité) respectivement. La classe modale est [200, 300]. De plus, il y a une queue prolongée vers les valeurs élevées.

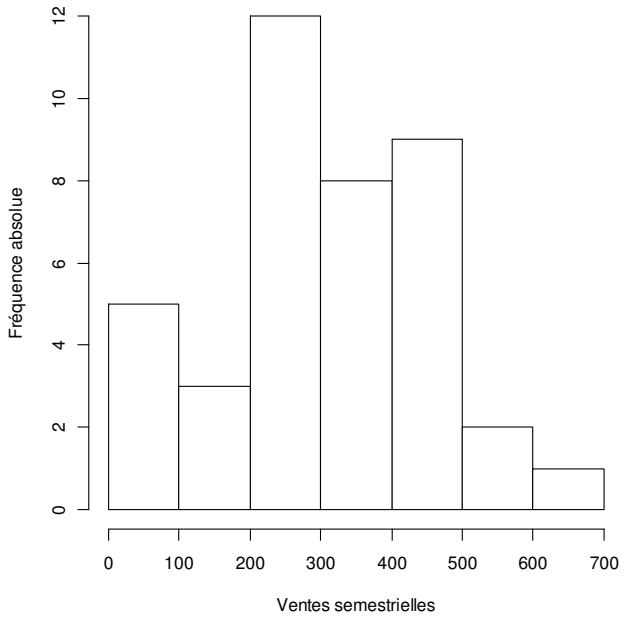


Figure 7 : Histogramme des fréquences absolues
Source : Pr BARANKANIRA Emmanuel

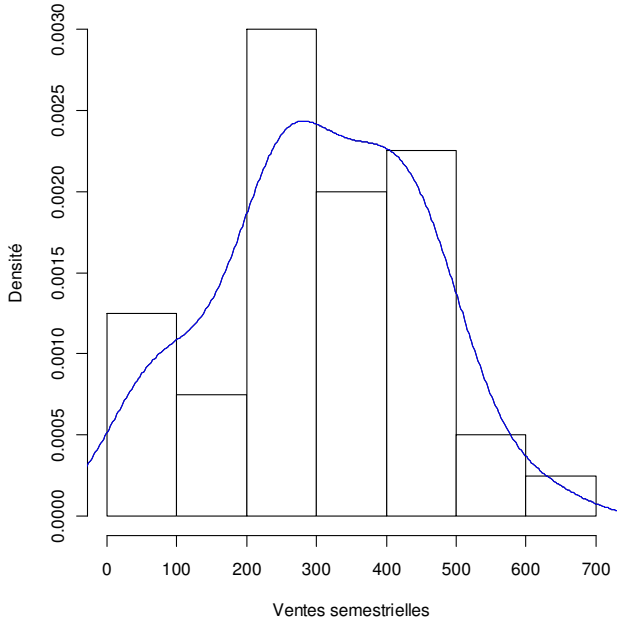


Figure 8 : Histogramme des fréquences relatives
Source : Pr BARANKANIRA Emmanuel

Il existe d'autres règles de découpage des variables continues en variables discrètes (en classes) telles que :

Brooks-Carruthers : $K = 5 \times log_{10}(n)$ (3.4)

Huntsberger : $K = 1 + 3{,}332 \times log_{10}(n)$ (3.5)

Sturges : $K = log_2(n+1)$ (3.6)

Scott : $K = \frac{x_{max} - x_{min}}{3{,}5 \times \sigma \times n^{-\frac{1}{3}}}$ où σ est l'écart-type (voir plus loin) (3.7)

Freedman-Diaconis : $K = \frac{x_{max} - x_{min}}{2 \times EIQ \times n^{-\frac{1}{3}}}$ où EIQ est l'écart interquarile (3.8)

3.2.2. Polygone des fréquences cumulées

La courbe des fréquences cumulées, appelée également **Ogive**, consiste à tracer le graphique des fréquences cumulées, en mettant les limites des classes sur l'axe horizontal et les fréquences cumulées sur l'axe vertical, ces dernières se cumulant à la fin de chacune des classes. Ce graphique

aura l'allure d'une courbe croissante variant entre entre 0 et l'effetcif totale si les fréquences absolues sont utilisées et entre 0 et 1 si les fréquences relatives sont utilisées (figures 9 et 10 respectivement).

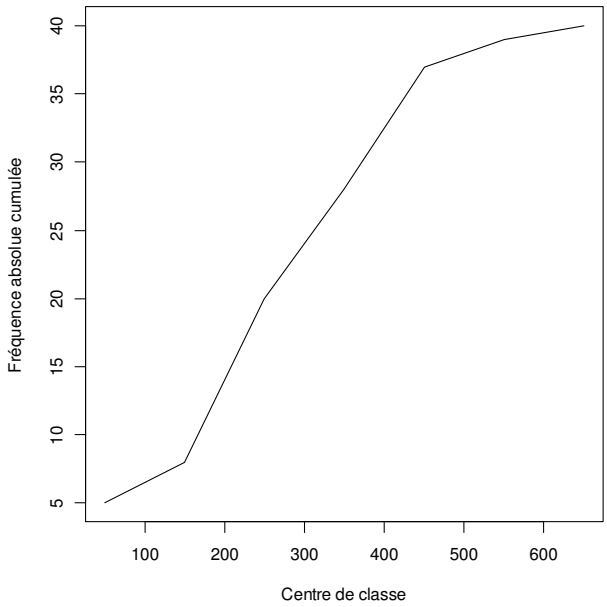


Figure 9 : Polygone des fréquences absolues cumulées des recettes quotidiennes
Source : Pr BARANKANIRA Emmanuel

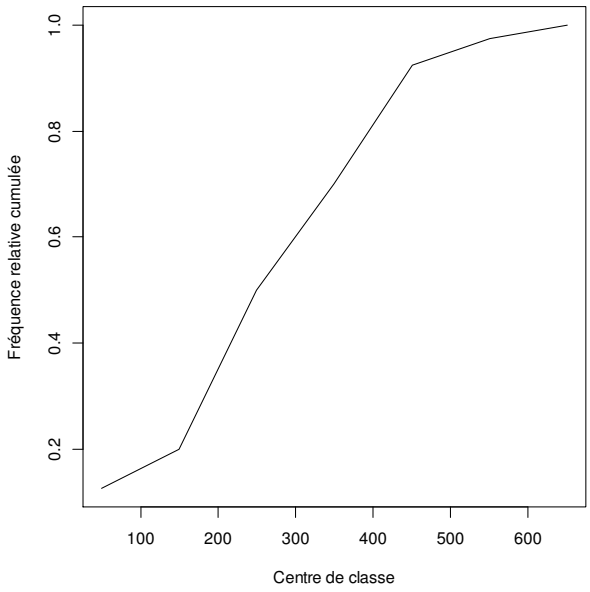


Figure 10 : Polygone des fréquences relatives cumulées des recettes quotidiennes
Source : Pr BARANKANIRA Emmanuel

Remarque : Lorsque les classes ne sont pas de même amplitude, il faut se rappeler que la surface du rectangle d'un histogramme est égale à la fréquence relative de la classe associée à ce rectangle. Alors, si la largeur de cette classe par exemple est le double de la l'amplitude de base, la hauteur du rectangle doit être divisée par deux.

3.2.3. Boîte à moustaches

Une bôîte à moustaches est un graphique qui est construit en utilisant les quartiles. Les quartiles, en plus de leur utilisation comme mesures de position, s'utilisent pour détecter des données aberrantes dans toute série de données. Cette détection se fait à l'aide d'un graphique, appelé graphique en boîte (box-plot) ou hamac ou diagramme à moustaches selon les auteurs. Son principe consiste à calculer les quartiles de la série et deux limites acceptables, soient une limite inférieure $BI = Q_1 - 1{,}5(Q_3 - Q_1)$ et une limite supérieure $BS = Q_3 + 1{,}5(Q_3 - Q_1)$. Toute observation qui ne se trouve pas entre ces deux limites est jugée aberrante (outlier *en* anglais) et doit être exclue de la série ou imputée avant toute analyse statistique proprement dite des données (une interprétation de la présence des données aberrantes éventuelles est faite en fin d'analyse). Il est à noter qu'une donnée aberrante peut avoir un effet catastophique sur la moyenne (paramètre qui est le moins stable par rapport au mode et à la médiane), sur l'écart type et même sur l'allure générale de la distribution des données.

Exemple : Reprenons l'exemple de la série statistique suivante de 12 observations :

-2 -3 10 12 120 11 4 8 6 13 130 200

La figure 11 montre la représentation de la variable sous forme d'une boîte à moustaches.

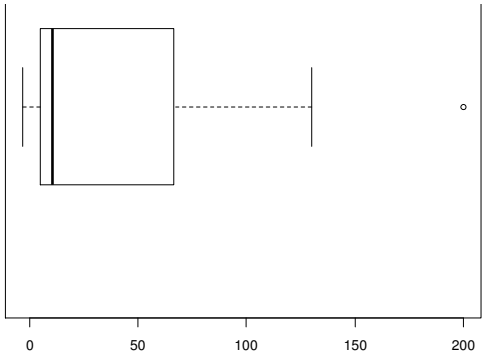


Figure 11 : Boîte à moustaches
Source : Pr BARANKANIRA Emmanuel

La valeur 200 est une observation aberrante qui est anormalement grande. Elle s'écarte sensiblement des autres observations.

Exercice : Soit la série des données déjà ordonnée suivante (n=21 observations) :

8 12 20 27 30 32 35 36 40 40 40 40 41 42 45 47 50 52 61 89* 101*

Déterminez s'il y a des données aberrantes dans cette série à l'aide d'un graphique en boîte (box-plot).

Réponse : Les différentes mesures de cette variable sont obtenues à l'aide du logiciel R :

Nombre d'observations : n=21

Nombre de valeurs manquantes : 0

Moyenne : 42,29

Erreur standard de la moyenne : 4,72

Déviation standard : 21,63

Minimum : 8

Premier quartile : Q_1=32

Médiane : Q_2=40

Troisième quartile : Q_3=47

Maximum : 101

Cela signifie que :

$$BI = Q_1 - 1{,}5\left(Q_3 - Q_1\right) = 32\text{-}1{,}5*(47\text{-}32) = 9{,}5$$

$$BS = Q_3 + 1{,}5\left(Q_3 - Q_1\right) == 47+1{,}5*(47\text{-}32) = 69{,}5$$

La série ordonée est :

8* 12 20 27 30 32 35 36 40 40 40 40 41 42 45 47 50 52 61 89* 101*

Donc, il y a 3 données aberrantes dans cette série, ce sont 8, 89 et 101 (qui sont signalées par le symbole *), ce qui est illustré dans le diagramme en boîte ci-dessous.

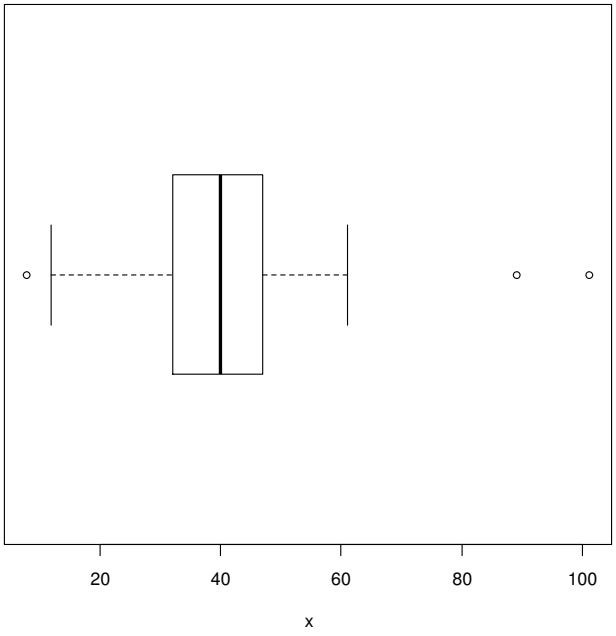


Figure 12 : Détection des valeurs aberrantes
Source : Pr BARANKANIRA Emmanuel

3.2.4. Digramme en tiges et feuilles

Un diagramme en tiges et feuilles ou diagramme en branches et feuilles (**stem-and-leaf** en anglais) est une représentation graphique d'une variable continue ou discrète. Les tiges sont constituées par la partie des dizaines et les feuilles par la partie des unités. Ce graphique permet d'examiner la symétrie d'une distribution.

```
0 | 22244
1 | 001234566
2 | 2344
3 | 00223345
4 | 22346
5 | 3345
6 | 2
7 | 35
8 | 1
9 | 4
```

La figure 13 montre le diagramme en tiges et feuilles des recettes quotidiennes, une variable continue. La série est ordonnée. Pour cet exemple, les tiges sont les chiffres des centaines et les feuilles les chiffres des dizaines. La distribution est asymétrique à droite.

```
 [1]  16.00  58.50  68.20  78.00  79.45 142.20 145.30 186.70 209.05 216.75
[11] 219.70 247.75 249.10 256.00 257.15 262.35 268.60 269.60 270.15 284.45
[21] 319.00 332.00 343.29 350.75 354.90 372.60 383.20 389.20 404.55 420.20
[31] 428.50 432.40 444.60 446.80 456.10 458.10 493.95 511.95 521.05 621.35

0 | 26788
1 | 459
2 | 122556667778
3 | 23455789
4 | 023345669
5 | 12
6 | 2
```

Figure 13 : Diagramme en tiges et feuilles des recettes quotidiennes
Source : Pr BARANKANIRA Emmanuel

3.3. Variables nominales

En ce qui concerne la représentation graphique d'une variable qualitative, il est possible de donner trois graphiques qui résument la même information contenue dans le tableau des fréquences. Le premier type de graphique est le diagramme en bâtons. Le deuxième est le diagramme en barres avec des barres horizontales ou verticales, appelé aussi barchart. Pour construire ce dernier graphique, les modalités de la variable sont mises sur un axe et les fréquences absolues ou les fréquences relatives sur l'autre axe. Le troisième type de graphique est le diagramme sectoriel, appelé aussi camembert ou pie chart. Pour ce dernier graphique, le cercle est découpé en secteurs et l'aire de chaque secteur est proportionnelle à la fréquence de chaque modalité.

Considérons ici deux exemples où des variables qualitatives observées sur un échantillon sont données.

Exemple 1 :

Un échantillon de 50 achats de boissons non alcoolisées achetées dans une grande surface a été pris, en notant par :

CC=Coca-Cola; S=Sprite; CL=Coke-Light; P=Perrier; PC=Pepsi-Cola

Les résultats obtenus sont les suivants :

CC S PC CL CC CC PC CL CC CL CC CC CC CL PC CC
CC P P S CC CL PC CL PC CC PC PC CC PC CC CC PC
P PC PC S CC CC CC S P CL P PC CC PC S CC CL

Alors, ici la variable est X=Boisson non alcoolisée. C'est une variable qualitative nominale. Pour présenter ces données sous forme de tableau, un tableau est dressé. Dans la première colonne, les cinq modalités de la variable sont énumérées. Dans la seconde colonne, la f**réquence absolue** ou l'effectif de chacune des modalités (c'est-à-dire le nombre de fois que cette modalité se répète dans l'échantillon) est donné(e) et dans la troisième colonne, **la fréquence relative** de chacune des modalités est donnée, la **fréquence relative** d'une modalité étant égale à sa fréquence absolue divisée par la taille de l'échantillon. Ce tableau s'appelle **tableau de fréquences** de la variable et montre la répartition des ventes des boissons non alcoolisées selon la marque.

Tableau des fréquences des boissons non alcoolisées

Boisson	Fréquences absolues	Fréquences relatives
CC	19	0,38
CL	8	0,16
PC	13	0,26
P	5	0,10
S	5	0,10
Total	50	1,00

Source : données fictives

Remarque : Pour une présentation complète des tableaux et graphiques, il faut mettre le titre en haut pour le tableau et en bas pour la figure sans oublier la source des données en bas.

Les figures 14, 15 et 16 montrent le diagramme en bâtons, le diagramme en barres et le camembert du type de boisson. Le mode lu vaut **CC**.

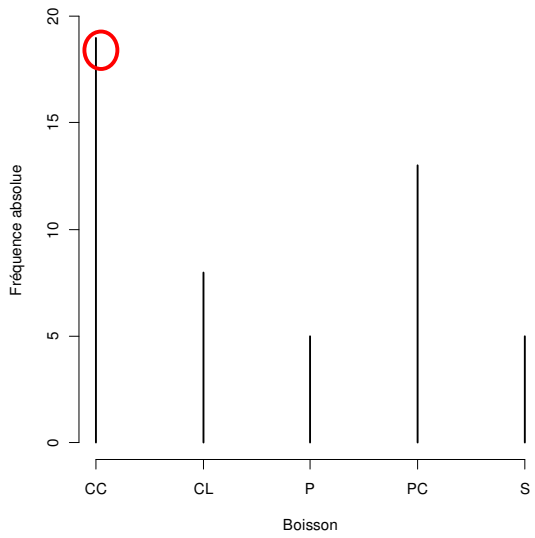


Figure 14 : Diagramme en bâtons du type de boisson
Source : Pr BARANKANIRA Emmanuel

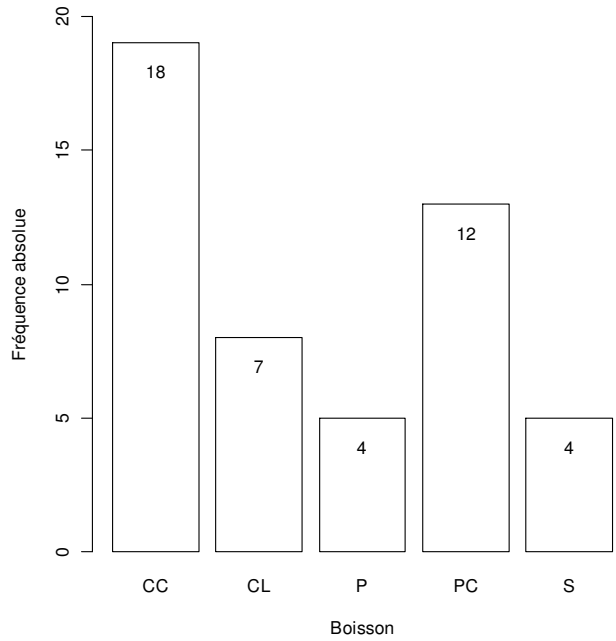


Figure 15 : Diagramme en barres du type de boisson
Source : Pr BARANKANIRA Emmanuel

Remarque : Les largeurs des barres doivent être les mêmes pour une belle esthétique du graphique, ainsi que la distance entre les bandes. Il est possible aussi d'ajouter les fréquences absolues au

dessus des bandes. Ce graphique porte le nom de diagramme en barres ou diagramme en tuyaux d'orgue.

Le troisième graphique est **le diagramme à secteurs** (ou circulaire) ou camembert qui est une sorte de tarte où chaque modalité occupe une partie qui reflète sa fréquence relative.

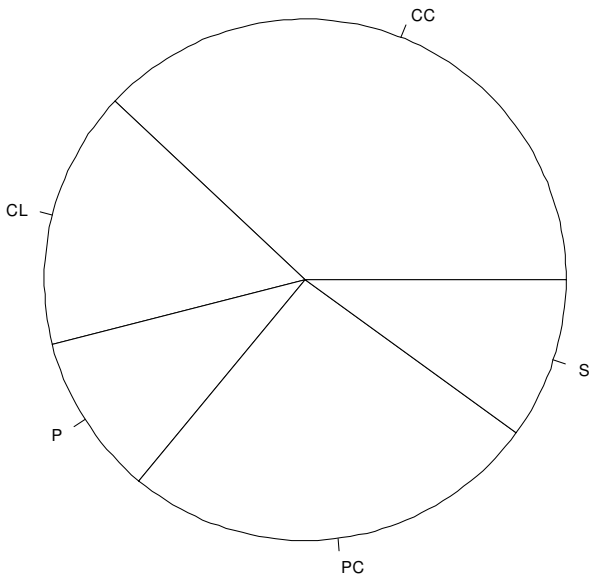


Figure 16 : Camembert du type de boisson
Source : Pr BARANKANIRA Emmanuel

3.4. Variables ordinales

Une variable ordinale est une variable qualitative dont les modalités sont naturellement ordonnées. La représentation graphique de cette variable peut se fait soit à l'aide d'un diagramme en barres, soit à l'aide d'un diagramme sectoriel.

Exemple 2 : Lors d'une enquête de satisfaction de la clientèle, une compagnie de courtage a demandé à un échantillon de 60 clients d'indiquer leur degré de satisfaction vis-à-vis de leur conseiller financier, sur une échelle de 1 à 7, le 1 correspondant à « pas du tout satisfait » et le 7 correspondant à « extrêmement satisfait ». Les résultats sont les suivants :

5 7 6 6 7 5 5 7 3 6 7 7 6 6 6 5 5 6 7 7

6 6 4 4 7 6 7 6 7 6 5 7 5 7 6 4 7 5 7 6

6 5 3 7 7 6 6 6 6 5 5 6 6 7 7 5 6 6 6 6

Ici la variable, « degré de satisfaction » est une variable qualitative ordinale. Il est possible de résumer l'information contenue dans ces données sous forme d'un tableau de fréquences ce qui donne :

Tableau 8 : Fréquences du degré de satisfaction des clients

Degré de satisfaction	Fréquences absolues (n_i)	Fréquences relatives (f_i)
1	0	0,0000
2	0	0,0000
3	2	0,0333
4	3	0,0500
5	12	0,2000
6	25	0,4167
7	18	0,3000
Total	60	1,0000

Source : Données fictives

En ce qui concerne la représentation graphique, les mêmes graphiques utilisés pour une variable qualitative nominale font l'affaire.

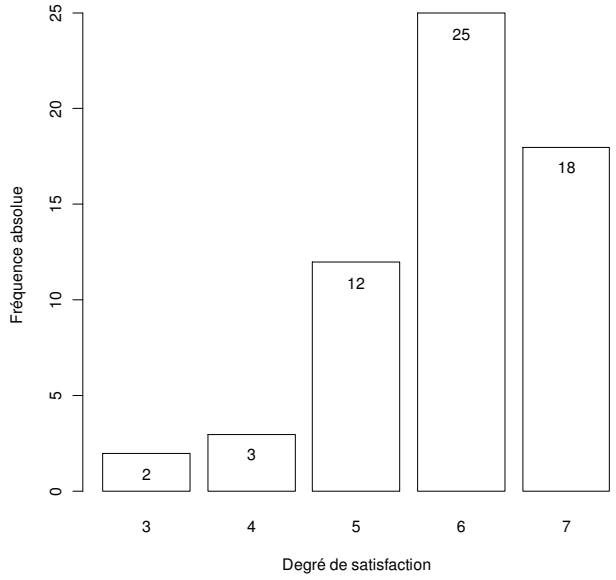


Figure 17 : Diagramme en barres du degré de satisfaction
Source : Pr BARANKANIRA Emmanuel

La figure ci-après donne la répartition du degré de satisfaction des clients sous forme d'un diagramme circulaire ou camembert.

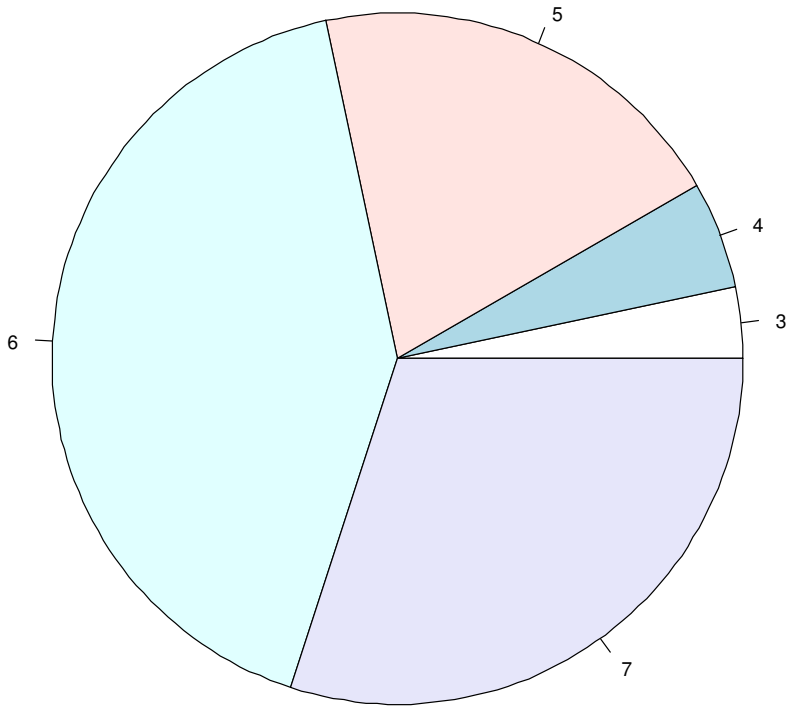
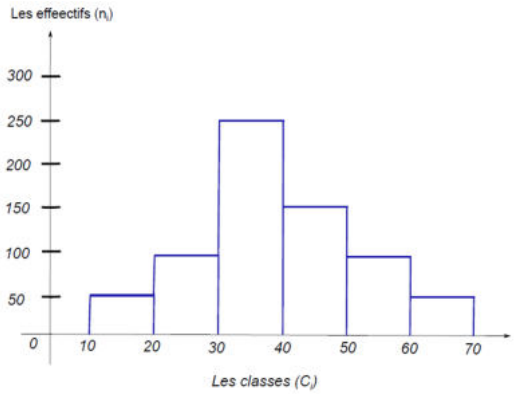


Figure 18 : Camembert du degré de satisfaction
Source : Pr BARANKANIRA Emmanuel

Exercices d'application - 2

Exercice 1 :

Dans une gare routière, on évalue le temps d'attente des voyageurs en minutes. Voici l'histogramme des fréquences absolues de cette variable.



1. Déterminer la variable statistique X et son type et sa population.

2. Déterminer le nombre de voyageurs.

3. Depuis le graphe, déterminer le tableau statistique.

4. Tracer la fonction cumulative.

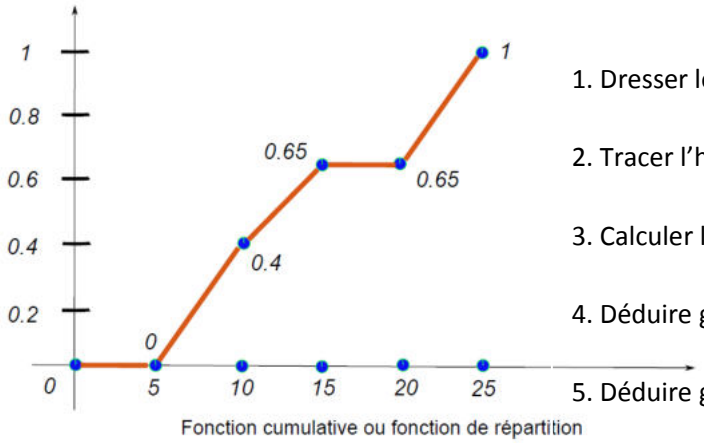
5. Déterminer le mode graphiquement et dire ce que représente cette valeur par rapport à notre étude.

6. Calculer la médiane à partir du graphe de la fonction cumulative.

7. Calculer la moyenne et l'écart-type.

Exercice 2 :

Le traitement de l'information sur un caractère x a permit de dresser sa fonction cumulative (fonction de répartition dans la figure ci-dessous).



Fonction cumulative ou fonction de répartition

1. Dresser le tableau statistique du caractère x.

2. Tracer l'histogramme du caractère x.

3. Calculer la moyenne et l'écart-type.

4. Déduire graphiquement la médiane.

5. Déduire graphiquement le mode.

Exercice 3 :

Le tableau suivant donne la répartition selon le groupe sanguin de 40 individus pris au hasard dans une population.

Groupes sanguins	*A*	*B*	*AB*	*O*
L'effectif	*20*	*10*	n_3	*5*

1. Déterminer la variable statistique et son type.

2. Déterminer l'effectif des personnes ayant un groupe sanguin AB.

3. Donner toutes les représentations graphiques possibles de cette distribution.

Exercice 4 :

Au poste de péage, on compte le nombre de voitures se présentant sur une période de 5 min. Sur 100 observations de 5 min, on obtient les résultats suivants :

Nombre de voitures	1	2	3	4	5	6	7	8	9	10	11	12
Nombre d'observations	2	8	14	20	19	15	9	6	2	3	1	1

1. Construire la table des fréquences et le diagramme en bâtons en fréquences de la série du nombre de voitures.

2. Calculer la moyenne et l'écart-type de cette série.

3. Déterminer la médiane.

Exercice 5 :

On observe le nombre d'arrivées des clients à un bureau de poste pendant un intervalle de temps donné (disant 10 minutes). En répétant 100 fois cette observation, on obtient les résultats suivants :

Nombre d'arrivés	1	2	3	4	5	6	Total
Nombre d'observations	15	25	26	20	7	7	100

– Représenter graphiquement ces résultats (pour les effectifs et pour les fréquences cumulées).

– Calculer la valeur de la moyenne arithmétique, de la médiane, de la variance et de l'écart-type des résultats.

Chapitre 4 : Régression linéaire simple

4.1. Introduction

Le mot « régression » vient de Sir Francis Galton. En 1885, il s'intéresse à expliquer la taille des enfants en fonction de celle des parents. Lorsque le père est plus grand que la moyenne, son fils a tendance d'être plus petit que lui. En plus, lorsque le père est plus petit que la moyenne, son fils a tendance d'être plus grand que lui. Autrement dit, la taille du père influence ou a un effet sur celle de sa génération. Nous disons qu'il y a une relation ou un lien linéaire entre la taille du père et celle de l'enfant. Autrement dit, régresser une variable (ici la taille du père) sur une autre variable (ici la taille de l'enfant) revient à construire un « modèle linéaire ». Comme il y a une seule variable explicative, alors le modèle est dit linéaire simple. Le but de la régression linéaire simple est d'étudier l'association entre deux variables quantitatives [5]. De plus, la régression permet la prédiction de l'une par l'autre.

4.2. Nuage de points

Pour illustrer le modèle linéaire simple, nous partons d'une série statistique double représentant l'âge et la taille des nourrissons. Soient x la variable qui représente l'âge des nourrissons et y leur taille. La figure 19 montre le nuage de points (diagramme de dispersion ou graphe X-Y) qui croise ces deux variables. Elle permet de visualiser ces données. Ce nuage de points est allongé. Autrement dit, il présente une direction envisagée. Il est donc possible de l'ajuster par une droite par la méthode des moindres carrés ordinaires.

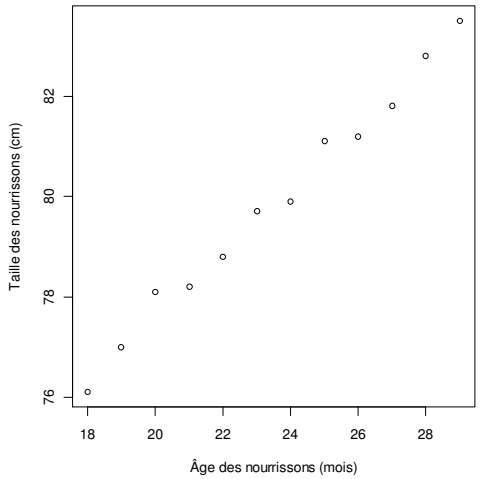


Figure 19 : Diagramme de dispersion
Source : Pr BARANKANIRA Emmanuel

La variable x (à mettre en abscisses) porte le nom de variable indépendante, de variable explicative, de prédicteur, de facteur, de régresseur ou de variable exogène en économie. Quant à la variable y (à mettre en ordonnées), elle porte le nom de variable dépendante, de variable expliquée, de variable à expliquer, de variable à prédire, de variable réponse ou de variable endogène en économie. Lorsqu'il y a plusieurs variables explicatives, alors il s'agit d'un modèle de régression linéaire multiple [8]. Nous n'allons pas nous intéresser à cet aspect dans cet ECUE.

4.3. Spécification du modèle

Notons y_i le revenu mensuel d'un ménage i et x_i ses dépenses. Nous pouvons alors écrire d'une façon matricielle le modèle linéaire simple comme suit :

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4.1}$$

Ce modèle linéaire simple est dépendant. Il a deux paramètres inconnus: β_0 est appelé constante ou intercept et β_1 appelé la pente. Le vecteur aléatoire ε est appelé erreur du modèle [9].

Pour chaque observation, ce modèle s'écrit :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{4.2}$$

L'erreur du modèle s'écrit :

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \tag{4.3}$$

En général, nous supposons qu'il y ait n observations connues des variables y et x, et que les hypothèses suivantes (il y en a autour de 11) soient vérifiées :

- Pour $i = 1,....n$, $E(\varepsilon_i) = 0$ (les erreurs sont centrées) ;
- Pour $i = 1,....n$, $Var(\varepsilon_i) = \sigma^2$ (la variance des erreurs est constante) ;
- Pour $i = 1,....n$, les variables ε_i sont indépendantes de loi de gaussienne ;
- Pour $i, j = 1,....n$, $Cov(\varepsilon_i, \varepsilon_j) = 0 (i \neq j)$ (les erreurs ne sont pas auto corrélées).

Le modèle linéaire simple peut aussi s'écrire :

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon} \tag{4.4}$$

ou encore

$$Y = X\beta + \varepsilon \tag{4.5}$$

avec

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \tag{4.6}$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \tag{4.7}$$

En général, x et y sont des variables observées n fois. L'échantillon est $(x_i, y_i), \ldots\ldots (x_{n,}, y_n)$, c'est-à-dire que (x, y) est observé n fois sur des éléments, $(z_1, \ldots z_n)$ de Ω référentiel de la population sur laquelle sont étudiées les variables x et y. La variable x est supposée non aléatoire. La constante n représente, par exemple, le nombre d'individus sur lesquels les couples (x, y) ont été observés, ou bien le nombre de fois où les valeurs de (x, y) ont été relevées. La figure 20 ci-après montre la représentation schématique du modèle linéaire.

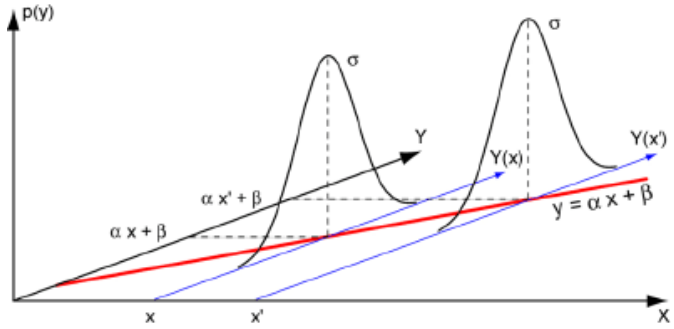


Figure 20 : Représentation schématique du modèle linéaire
Source : Pr BARANKANIRA Emmanuel

4.4. Estimation des paramètres du modèle

Deux situations sont envisageables :

- soit aucune hypothèse probabiliste n'est faite sur ε_i : estimation à l'aide d'une droite des moindres carrés ;
- soit une loi a priori est spécifiée pour ε_i : estimation à l'aide du maximum de vraisemblance.

Pour l'estimation à l'aide d'une droite des moindres carrés, nous cherchons de façon purement descriptive les valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ définissant la droite :

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon \tag{4.8}$$

telle que $\sum_{i=1}^{n} \varepsilon_i^2$ soit minimale.

La figure 21 suivante illustre le principe des moindres carrés ordinaires (MCO).

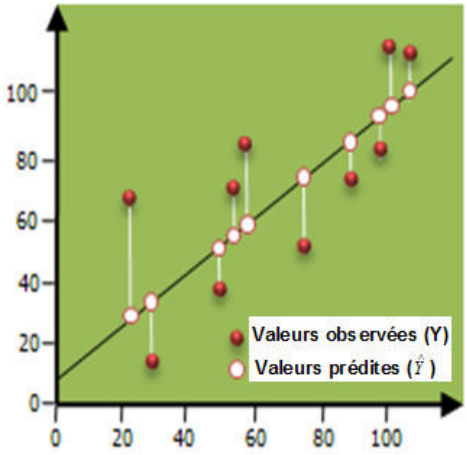


Figure 21 : Illustration du principe des moindres carrés ordinaires
Source : Pr BARANKANIRA Emmanuel

La méthode MCO consiste à minimiser la somme des carrés des erreurs du modèle linéaire, cette somme étant appelée fonction de perte [2,10] :

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_1^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{4.9}$$

La dérivation de la fonction de perte $S(\beta_0, \beta_1)$ par rapport à β_0 et β_1 donne respectivement :

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \dfrac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} (x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \end{cases} \tag{4.10}$$

Sachant que :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.11}$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.12}$$

la résolution du système :

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = 0 \\ \dfrac{\partial S}{\partial \beta_1} = 0 \end{cases} \tag{4.13}$$

conduit aux équations suivantes dites équations normales :

$$\begin{cases} \beta_0 + \beta_1 \overline{x} = \overline{y} \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_i x_i y_i \end{cases} \tag{4.14}$$

Matriciellement, cela s'écrit :

$$\begin{pmatrix} \overline{y} \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix} = \begin{pmatrix} 1 & \overline{x} \\ \sum_i x_i & \sum_{i=}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \tag{4.15}$$

L'estimateur MCO de l'intercept β_0 est :

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \tag{4.16}$$

L'injection de cette relation dans la deuxième équation du système ci-haut donne comme estimateur de la pente β_1, après quelques transformations :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \tag{4.17}$$

Sachant que la covariance entre $x = (x_1 ....., x_n)$ et $y = (y_1, ..., y_n)$ et la variance de $x = (x_1 ....., x_n)$ s'écrivent respectivement :

$$S_{xy} = Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right) \tag{4.18}$$

$$S_x^2 = Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}\right) \tag{4.19}$$

L'estimateur MCO de la pente peut s'écrire :

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{Cov(x, y)}{Va(x)} \tag{4.20}$$

où

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

et

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

désignent la variance de $(x_1, ..., x_n)$ et de $(y_1, ..., y_n)$ respectivement. Une fois $\hat{\beta}_1$ calculé, nous estimons $\hat{\beta}_0$ par la relation ci-haut. La droite ajustée $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ passe par le point moyen $(\bar{x}, \bar{y})$ et est appelée « droite des moindres carrés » de y en x. La valeur $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ est l'estimation de y_i par MCO et $\varepsilon_i = y_i - \hat{y}_i$ est le résidu. Ce résidu est considéré comme une estimation de la perturbation aléatoire ε_i :

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + u_i \tag{4.21}$$

Comme $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, alors y_i est devenu aléatoire et il vient :

$$E(y_i) = \beta_0 + \beta_1 x_i \tag{4.22}$$

$$Var(y_i) = Var(\varepsilon_i) = \sigma^2 \tag{4.23}$$

La variance σ^2 peut être estimée par la variance empirique S_y^2 de $(y_1,.....,y_n)$:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{4.24}$$

Quant à l'estimation à l'aide du maximum de vraisemblance (MMV), cette dernière consiste à rechercher l'estimation du paramètre inconnu qui rend le plus probable ou le plus vraisemblable l'échantillon observé. Puisqu'il s'agit de trouver le maximum, cette méthode fait appel à la notion de dérivée en mathématique, tout au moins pour les cas où la loi de probabilité est une fonction dérivable.

Les estimateurs obtenus par la méthode du maximum de vraisemblance (MMV) ont de bonnes propriétés statistiques. Certains estimateurs habituels obtenus par la méthode des moments sont les mêmes que ceux obtenus par la méthode du maximum de vraisemblance. La connaissance de cette méthode n'est pas indispensable pour traiter l'intervalle de confiance, ce qui dépasse le cadre de la statistique descriptive. La vraisemblance $L(x_i,....,x_n;\theta)$ représente la probabilité d'observer le n-uplet $(x_i,...,x_n)$ pour une valeur fixée de θ dans la situation inverse ici où $(x_i,...,x_n)$ est observé sans connaître la valeur de θ. Nous allons attribuer à θ une valeur qui paraît la plus vraisemblable compte tenu de l'observation à notre disposition, c'est-à-dire celle qui va attribuer la valeur avec la plus forte probabilité.

Par définition, l'estimateur du maximum de vraisemblance est toute fonction $\hat{\theta}_n$ de $(x_i,...,x_n)$ qui vérifie :

$$L(x_1,.... x_n;\hat{\theta}_n) = \max_{\theta\in\Theta} L(x_1,..... x_n;\theta) \tag{4.25}$$

La recherche de l'estimateur du maximum de vraisemblance peut se faire directement par la recherche du maximum de L ou dans le cas particulier où la fonction L, est deux fois dérivable par rapport à θ.

Le maximum est la solution de l'équation :

$$\frac{\partial L}{\partial \theta} = 0 \tag{4.26}$$

et vérifie aussi :

$$\frac{\partial^2 L}{\partial \theta^2} < 0 \tag{4.27}$$

Cependant, comme la vraisemblance se calcule à partir d'un produit, il est préférable de remplacer ce dernier problème par le problème équivalent pour la log-vraisemblance, puisque la fonction logarithme népérien (ln) est strictement croissante. En plus, si $\hat{\theta}_n$ est un estimateur du maximum de vraisemblance du paramètre θ, alors $g(\hat{\theta}_n)$est un estimateur de vraisemblance du maximum du paramètre $g(\theta)$ pour toute fonction g.

4.5. Coefficient de corrélation

Avant tout calcul, il convient de construire un nuage de points. Une fois que le nuage de points est construit et qu'il a une forme allongé, il alors est possible de l'ajuster par une droite de régression. Sinon, il suffit d'appliquer une transformation mathématique aux variables en présence jusqu'à ce que le nuage de points soit bon. Cette étape est suivie par le calcul des statistiques descriptives (nombre d'observations, minimum, moyenne, déviation standard, maximum) pour chaque variable. Il y a deux étapes à suivre lorsque l'analyse veut construire un modèle de régression linéaire simple. La première étape est de faire une analyse de corrélation dans le but d'étudier l'association entre les variables x et y. Le coefficient de corrélation linéaire qui mesure la force de la relation entre ces deux variables quantitatives se calcule somme suit :

$$r = r(x,y) = \frac{S_{xy}}{S_x S_y} = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.28}$$

Sachant que la pente de la droite de régression est :

$$\beta_1 = \frac{Cov(x, y)}{S_x^2} \tag{4.29}$$

Le coefficient de corrélation linéaire est lié à la pente de la droite par la relation :

$$\beta_1 = r\frac{S_y}{S_x} \tag{4.30}$$

- r n'a pas d'unité
- le signe de r nous donne une information sur le sens de la droite (directe / ou inverse)
- r varie entre 1 et -1
  - Si r est compris entre 0 et 0,1 => association nulle
  - Si r est compris entre 0,1 et 0,3 => association faible
  - Si r est compris entre 0,3 et 0,5 => association moyenne
  - Si r est compris entre 0,5 et 0,75 => association bonne
  - Si r est compris entre 0,75 et 1 => association excellente

La deuxième étape est alors d'expliquer une variable par une autre afin de pouvoir faire des prévisions.

4.6. Coefficient de détermination

Le coefficient de détermination (R^2) est un indicateur qui permet de juger la qualité d'une régression linéaire simple ou multiple. D'une valeur comprise entre 0 et 1, il mesure l'adéquation entre le modèle et les données observées. Certes, le R^2 a ses imperfections, mais son utilité n'a d'égale que sa simplicité.

Ce coefficient peut se définir de deux façons. D'une part, il s'agit du carré du coefficient de corrélation dans le cadre d'une régression linéaire simple. D'autre part, une deuxième façon de le définir est beaucoup plus riche en implications car elle s'applique aussi bien à une régression simple qu'à une régression linéaire. Nous savons que la valeur y_i d'une observation peut être décomposée en deux parties : une part expliquée par le modèle et une part résiduelle. La dispersion de l'ensemble des observations se décompose donc en variance expliquée par la régression et en variance résiduelle inexpliquée. Le R^2 se définit alors comme la part de variance expliquée par rapport à la variance totale. C'est également le complément du rapport entre la somme des carrés des résidus et

la somme des carrés totale, la somme des carrés totale étant la somme des carrés des distances entre les points du nuage et une droite horizontale qui passerait par son centre de gravité :

$$R^2 = 1 - \frac{SCE_R}{SCE_T} \tag{4.31}$$

où la somme des carrés des écarts totaux (SCE_T), la somme des carrés des écarts factoriels (SCE_F) et la somme des carrés des écarts résiduels (SCE_R) sont respectivement données par :

$$SCE_T = \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2 \tag{4.32}$$

$$SCE_F = \sum_{i=1}^{n} \left(\hat{y}_i - \overline{y} \right)^2 \tag{4.33}$$

$$SCE_R = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{4.34}$$

Le carré du coefficient de corrélation noté r² se définit comme suit :

$$r^2 = \frac{a^2 \sum_{i=1}^{n} \left(x_1 - \overline{x} \right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2} = \frac{\left[\sum_{i=1}^{n} \left(x_1 - \overline{x} \right) \left(y_i - \overline{y} \right) \right]^2}{\sum_{i=1}^{n} \left(x_1 - \overline{x} \right)^2 \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2} = \frac{S^2(x,y)}{S^2(x) S^2(y)} \tag{4.35}$$

et vaut le coefficient de détermination.

Le coefficient de corrélation a le signe de la covariance [2] :

$$r = \frac{\sum_{i=1}^{n} (x_1 - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_1 - \overline{x}) \sum_{i=1}^{n} (y_i - \overline{y})} = \frac{S(x,y)}{S(x) S(y)} \tag{4.36}$$

De plus :

$$var(y) = var(\hat{y}) + var(\hat{e}) \tag{4.37}$$

Donc [1] :

$$var\left(\hat{y}\right) = var\left(y\right) + var\left(\hat{e}\right) \tag{4.38}$$

$$r^2\left(x,y\right) = \frac{var\left(\hat{y}\right)}{var\left(y\right)} \tag{4.39}$$

Il vient, par simple remplacement :

$$r^2\left(x,y\right) = \frac{var\left(\hat{y}\right)}{var\left(y\right)} - \frac{var\left(\hat{e}\right)}{var\left(y\right)} \tag{4.40}$$

Nous obtenons :

$$r^2\left(x,y\right) = 1 - \frac{var\left(\hat{e}\right)}{var\left(y\right)} \tag{4.41}$$

Nous déduisons de cette décomposition que le coefficient R^2 défini comme le carré du coefficient de corrélation entre x et y est une mesure de la qualité de l'ajustement et est égal au rapport de la variance effectivement expliquée sur la variance à expliquer.

Donc, nous obtenons :

$$R^2 = r^2\left(x,y\right) = \frac{var\left(\hat{y}\right)}{var\left(y\right)} \quad avec\ 0 < R^2 < 1 \tag{4.42}$$

ce qui implique :

$$R^2 = 1 - \frac{var\left(\hat{e}\right)}{var\left(y\right)} \tag{4.41}$$

4.7. Formule fondamentale

Considérons le développement télescopique suivant :

$$y_i - \overline{y} = y_i - \hat{y}_i + \hat{y}_i - \overline{y} \tag{4.42}$$

Il suffit de montrer aisément que :

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SCE_T}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCE_R}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SCE_F} \tag{4.43}$$

Cela signifie que la somme des carrés des écarts totaux est égale à la somme de la somme des carrés des écarts factoriels et de la somme des carrés des écarts dus aux résidus (formule à démontrer).

Le rapport $\frac{SCE_F}{SCE_T}$ exprime le pourcentage variation de y qui est expliqué par la régression sur x et :

$$R^2=\frac{SCE_F}{SCE_T}=(r)^2=coefficient\ de\ détermination \tag{4.44}$$

Ce coefficient n'a pas d'unité. Il donne la force de la relation (y, x) et il varie entre 0 et 1 (proportion ou pourcentage). Il montre le pourcentage de la variabilité de la variable dépendante expliqué par les variations de la variable indépendantes.

La formule fondamentale permet de construire le tableau de l'analyse de la variance mais son interprétation dépasse le cadre de la statistique descriptive.

Le tableau 9 ci-après est un tableau de l'analyse de la variance à compléter.

Tableau 9 : Tableau de l'analyse de la variance (1)

Source de variation	Somme des carrés	Degré de liberté	Carré moyen	F^{obs}	F^{tab}	P-value
Factorielle	SCE_F	1	$CMF=\frac{SCE_F}{1}$	$\frac{CMF}{CMR}$	$F(1,n-2)$	$P(F^{obs}\geq F^{tab})$
Résiduelle	SCE_R	$n-2$	$CMR=\frac{SCE_R}{n-2}$	///////	//////////	/////////
Totale	SCE_T	$n-1$	$CMT=\frac{SCE_T}{n-1}$	///////	//////////	/////////

Source : Pr BARANKANIRA Emmanuel

Application :

Le tableau 10 montre les résultats qui vont permettre de calculer la moyenne, la variance et la déviation standard pour chaque variable.

Tableau 10 : Tableau statistique

n°	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	18	76,1	324	5791,21	1369,8
2	19	77	361	5929	1463
3	20	78,1	400	6099,61	1562
4	21	78,2	441	6115,24	1642,2
5	22	78,8	484	6209,44	1733,6
6	23	79,7	529	6352,09	1833,1
7	24	79,9	576	6384,01	1917,6
8	25	81,1	625	6577,21	2027,5
9	26	81,2	676	6593,44	2111,2
10	27	81,8	729	6691,24	2208,6
11	28	82,8	784	6855,84	2318,4
12	29	83,5	841	6972,25	2421,5
Σ	282	958,2	6770	76570,58	22608,5

Les moyennes sont :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{282}{12} = 23,5$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{958,2}{12} = 79,85$$

Les variances sont :

$$S_x^2 = Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

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

$$= \frac{6770 - 12(23,5)^2}{11}$$

$$= 13$$

$$S_y^2 = Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^2 - n\,\overline{y}^2\right)$$

$$= \frac{76570,58 - 12(79,85)^2}{11}$$

$$= 5,30$$

Les déviations standards (écarts-types) sont :

$$S_x = \sqrt{S_x^2} = \sqrt{13} = 3,61$$
$$S_y = \sqrt{S_y^2} = \sqrt{5,30} = 2,30$$

La covariance entre x et y vaut :

$$S_{xy} = Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right)$$
$$= \frac{22608,5 - 12(23,5)(79,85)}{11}$$
$$= 8,25$$

Le coefficient de corrélation linéaire entre x et y vaut :

$$r = cor(x,y) = \frac{S_{xy}}{S_x S_y} = \frac{8,25}{3,61(2,30)} = 0,99$$

Comme $r > 0$, alors aux plus grandes valeurs de l'âge des nourrissons correspondent les plus grandes valeurs de leur taille. De plus, comme $0,75 \leq r \leq 1$, alors l'association entre ces deux variables est excellente.

L'équation de la droite de régression de y en x s'écrit :

$$D_{y/x} \equiv y = ax + b \qquad (4.45)$$

avec

$$a = \frac{S_{xy}}{S_x^2} = \frac{Cov(x,y)}{Var(x)} = \frac{8,25}{13} = 0,634965 \simeq 0,635$$
$$b = \bar{y} - a\,\bar{x} = 79,85 - 0,635(23,5) = 64,9275 \simeq 64,93$$

L'équation de la droite de régression est :

$$y = 0,635\,x + 64,93$$

Le coefficient de détermination vaut :

$$R^2 = (r)^2 = (0,9944)^2 = 0,9888 \simeq 98,9\,\%$$

On parvient à expliquer 98,9 % de la variabilité de la taille des nourrissons à l'aide des variations de leur âge.

La somme des carrés des écarts totaux vaut :

$$SCE_T = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n} y_i^2 - n\,\overline{y}^2 = 6770 - 12\left(79,85\right)^2 = 58,31 \tag{4.3}$$

La figure 22 montre le nuage de points ajusté.

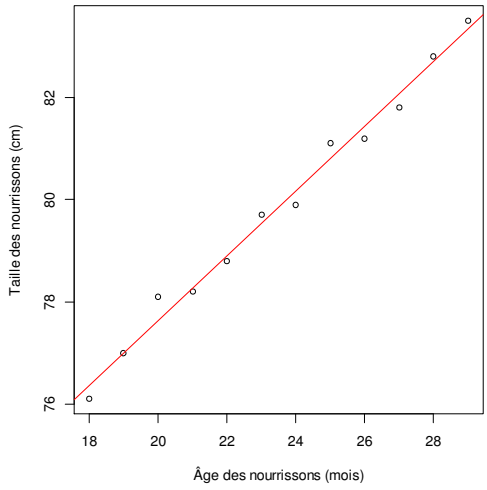


Figure 22 : Nuage de points ajusté
Source : Pr BARANKANIRA Emmanuel

La somme des carrés des écarts factoriels (expliqués) vaut :

$$\begin{aligned}
SCE_F &= \sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2 \\
&= \sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i - \overline{y}\right)^2 \\
&= \sum_{i=1}^{n}\left(\overline{y} - \beta_1\overline{x} + \beta_1 x_i - \overline{y}\right)^2 \\
&= \sum_{i=1}^{n}\left(\beta_1 x_i - \beta_1\overline{x}\right)^2 \\
&= \beta_1^2\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \\
&= \left(n-1\right)\beta_1^2 S_x^2 \\
&= 11\left(0,634965\right)^2\left(13\right) \\
&= 57,65483 \\
&\simeq 57,65
\end{aligned}$$

La somme des carrés des écarts résiduels vaut :

$$
\begin{aligned}
SCE_R &= \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2 \\
&= \sum_{i=1}^{n}\left(y_i-\overline{y}+\beta_1\overline{x}-\beta_1 x_i\right)^2 \\
&= \sum_{i=1}^{n}\left[\left(y_i-\overline{y}\right)+\beta_1\left(x_i-\overline{x}\right)\right]^2 \\
&= \sum_{i=1}^{n}\left[\left(y_i-\overline{y}\right)^2+\beta_1^2\left(x_i-\overline{x}\right)^2-2\beta_1\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)\right] \\
&= \sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2+\beta_1^2\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2-2\beta_1\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right) \\
&= \left(n-1\right)S_y^2+\beta_1^2\left(n-1\right)S_x^2-2\beta_1\left(n-1\right)S_{xy} \\
&= \left(n-1\right)\left(S_y^2+\beta_1^2 S_x^2-2\beta_1 S_{xy}\right) \\
&= \left(n-1\right)\left(S_y^2+\frac{S_{xy}^2}{S_x^4}S_x^2-2\frac{S_{xy}}{S_x^2}S_{xy}\right) \\
&= \left(n-1\right)\left(S_y^2-\frac{S_{xy}^2}{S_x^2}\right) \\
&= \left(n-1\right)S_y^2-\left(n-1\right)\left(\frac{S_{xy}^2}{S_x^2}\right) \\
&= SCE_T-SCE_F
\end{aligned}
$$

soit

$$SCE_R = 58,31-57,65 = 0,65$$

Le coefficient de détermination vaut :

$$R^2 = \frac{SCE_F}{SCE_T} = 1-\frac{SCE_R}{SCE_T} = 1-\frac{0,65}{58,31} = 0,9888 \simeq 98,9\%$$

Le tableau 11 ci-après est un tableau de l'analyse de la variance complétée.

Tableau 11 : Tableau de l'analyse de la variance (2)

Source de variation	Somme des carrés	Degré de liberté	Carré moyen	F^{obs}	F^{tab}	P-value
Factorielle	57,65	1	57,65	879,99	4,96	<0,05
Résiduelle	0,65	10	0,065	///////	//////////	/////////
Totale	58,31	11	5,30	////////	/////////	/////////

La conclusion est que l'âge a un pouvoir explicatif.

Lorsque l'âge du nourrisson vaut 30 mois, alors sa taille prédite vaut :

$$\hat{y}_{30} = 30a + b = 30(0,635) + 64,93 = 83,97727 \simeq 83,98$$

Exercices d'application - 3

Exercice 1 :

Le tableau suivant donne les résultats obtenus à partir de 10 essais de laboratoire concernant la charge de rupture d'un acier en fonction de sa teneur en carbone.

Teneur en carbone x_i	70	60	68	64	66	64	62	70	74	62
Charge de rupture y_i (en kg)	87	71	79	74	79	80	75	86	95	70

a) Représentez graphiquement le nuage de points de coordonnées (x_i, y_i).
b) Déterminez la valeur du coefficient de corrélation linéaire de la série statistique de variables x et y. Interprétez le résultat.
c) Déterminez une équation de la forme y = ax + b de la droite D de régression de y en x par la méthode des moindres carrés. Tracez la droite D sur le graphique du point a).
d) Un acier a une teneur en carbone de 77. Donner une estimation de sa charge de rupture.
e) Calculez le coefficient de détermination. Interprétez le résultat.

Exercice 2 :

Une usine produit des pièces d'une machine. Pour chaque pièce (individu), on dispose du coût de sa production (DA) et du temps nécessaire pour sa réalisation (en heures). Le tableau ci-après (série statistique) donne cette répartition :

Individu	1	2	3	4	5
Temps (x) mesuré en heures	2	3	52	2	4
Coût (y) mesuré en FBu	10	16	23	12	18

1) Tracez le nuage de points.
2) Calculez le coefficient de corrélation, la pente et l'ordonnée à l'origine de la droite de régression.
3) Une nouvelle pièce est réalisée en 6 heures. Estimez le coût de production de cette pièce.

Exercice 3 :

Démontrer que, dans un modèle de régression, la somme des carrés totale est égale à la somme de la somme des carrés expliquée et de la somme des carrés résiduelle : $SCE_T = SCE_F + SCE_R$

Exercice 4 :

À l'aide d'un schéma, établissez l'équation de la droite de régression de y en x et celle de x en y par la méthode des moindres carrés ordinaires.

Exercice 5 :

Soit x et y deux variables statistiques mesurées sur un même individu. Par exemple, pour l'individu n°3, x = 2 et y = 8.

Individu	*1*	*2*	*3*	*4*	*5*
X	*3*	*4*	*2*	*5*	*3*
Y	*12*	*14*	*8*	*19*	*11*

1. Calculer la moyenne de la variable statistique x.
2. Calculer la moyenne de la variable statistique y .
3. Calculer l'écart-type de la variable statistique x.

4. Calculer l'écart-type de la variable statistique y.
5. Calculer la covariance des variables statistiques x et y.
6. En supposant qu'il existe une corrélation linéaire entre x et y (**de deux manières**), **déterminer cette droite de corrélation**.
7. Calculer le coefficient de corrélation. Conclusion ?
8. **Calculer le coefficient de détermination (de deux manières) et interpréter**

Exercice 6 :

Le tableau ci-dessous représente l'évolution de la consommation des ménages en fonction du revenu disponible brut sur une période donnée.

Revenu x_i	56	42	72	36	63	47	55	49	38	42	68	60
Consommation y_i	136	132	136	130	138	132	136	130	142	134	136	140

a) Calculez la moyenne de x et de y
b) Calculez l'écart-type (déviation standard) de x et y
c) Construisez le diagramme de dispersion
d) Estimez les paramètres du modèle linéaire simple (3 manières différentes) et déterminez l'équation de la droite de régression de y sur x
e) Calculez le coefficient de corrélation linéaire et interpréter
f) Calculez le coefficient de détermination (2 manières différentes)
g) Calculez le coefficient de détermination ajusté et non ajusté. Interprétez le résultat.
h) Estimez la consommation pour un ménage ayant reçu une somme égale à 50 ;
h) Tracez la droite de régression sur le nuage de points
i) Construisez le tableau de l'analyse de la variance et interpréter

Exercice 7 :

On cherche à déterminer si, dans la ville de Pompaluile, il existe une relation entre le nombre de véhicules qui passent devant une station d'essence et le nombre de litres d'essence vendus (moyennes par jour, sur un an). On note x la variable représentant le nombre de véhicules (en centaines) et y la variable qui désigne le nombre de litres(en milliers). Les résultats sont donnés par ce tableau :

x_i (en centaines)	3	4	5	7	2	3	2
y_i (en milliers)	100	112	150	210	60	85	77

a) Quelle est la nature des variables x et y ? Calculer la moyenne, la déviation standard, la médiane, le mode et l'étendue pour chaque variable.
b) Construisez le nuage de points correspondant à cette série statistique double.
c) Un ajustement linéaire est-il envisageable ? Justifiez votre réponse.
d) Calculez le coefficient de corrélation linéaire entre le nombre de véhicules et le nombre de litres d'essence vendus. Interprétez.
e) Calculez le coefficient de détermination (**deux manières**) et interprétez.
f) Déterminez les coefficients de la droite de régression par la méthode des moindres carrés et tracez cette droite sur le graphique de la sous-question c).

Exercice 8 :

Le gérant d'une salle de remise en forme vous demande de réaliser une étude permettant de prévoir la rentabilité de son centre en 2009, en suivant les étapes suivantes : En tenant compte de la quantité d'abonnements annuels réalisés entre 2002 et 2007, vous devrez prévoir le nombre d'abonnements annuels que le gérant peut espérer réaliser en 2009.

Le tableau ci-dessous regroupe les nombres d'abonnements annuels réalisés entre 2002 et 2007.

Année	2002	2003	2004	2005	2006	2007
Rang de l'année (x)	1	2	3	5	6	7
Nombre d'abonnements annuels réalisés (y)	306	314	328	339	332	340

a) Calculez le coefficient de corrélation linéaire entre x et y ;
b) Représentez cette série statistique par un nuage de points ;
c) Trouver l'équation de la droite d'ajustement du nuage de points ;
d) Faites des prévisions du nombre d'années d'abonnements pour l'année 2009 ?
e) Placer le point *G (point moyen)* et tracer la droite d'ajustement dans le repère.

Chapitre 5 : Indices élémentaires et indices synthétiques

Dans la plupart des cas, les données sont indexées par le temps et les statistiques à calculer permettent de synthétiser l'évolution au cours du temps et/ou de l'espace d'une variable d'intérêt. Pour cela, des indices élémentaires et des indices composites ou synthétiques sont calculés. La moyenne arithmétique et la moyenne géométrique sont des exemples d'indices synthétiques.

Il est possible de se poser les questions suivantes :

- Le prix d'un produit (fleurs, viande, fruits, céréales, fibres textiles, carburants, loisirs) donné a-t-il augmenté ou diminué entre deux dates données ?
- Comment peut-on appliquer la propriété de circularité ?
- Comment peut-on appliquer la propriété de réversibilité ?

5.1. Indices élémentaires

Si V_t est la valeur d'une grandeur étudiée à l'instant t et V_0 la même valeur à l'instant de référence, alors un indice simple qui permet de comparer ces deux grandeurs dans la base 100 vaut :

$$I_{t/0} = \frac{V_t}{V_0} \times 100 \tag{5.1}$$

Pour se rendre compte de la diminution ou de l'augmentation de ces valeurs entre ces deux dates, il suffit de calculer la différence :

$$\Delta = I_{t/0} - 100 \tag{5.2}$$

SI cette différence est négative, alors il y a une diminution de Δ % et si cette différence est positive, alors il y a eu une augmentation de Δ %.

Par exemple, si un produit donné coûte 2500 FBu en 2020 et 3000 FBu en 2022, alors :

$$\begin{aligned} I^P_{2022/2020} &= \frac{P_{2022}}{P_{2020}} \times 100 = \frac{3000}{2500} \times 100 = 120 \\ \Delta &= I^P_{2022/2020} - 100 = 120 - 100 = 20 > 0 \end{aligned} \tag{5.3}$$

Le prix du produit a donc augmenté de 20 % de 2020 à 2022.

5.1.1. Propriété de circularité

En utilisant la propriété de circularité ou de réversibilité, si le prix de produit est de 2700 FBu en 2024, alors :

$$I^{P}_{2024/2020} = I^{P}_{2024/2020} \times I^{P}_{2022/2020} \times \frac{1}{100} \tag{5.4}$$

En effet :

$$I^{P}_{2024/2020} \times I^{P}_{2022/2020} \times \frac{1}{100} = \frac{P_{2024}}{P_{2022}} \times 100 \times \frac{P_{2022}}{P_{2020}} \times 100 \times \frac{1}{100} = \frac{P_{2024}}{P_{2020}} \times 100 = I^{P}_{2024/2020}$$

D'où :

$$\begin{aligned} I^{P}_{2024/2020} &= I^{P}_{2024/2020} \times I^{P}_{2022/2020} \times \frac{1}{100} \\ &= \frac{P_{2024}}{P_{2022}} \times 100 \times \frac{P_{2022}}{P_{2020}} \times 100 \times \frac{1}{100} \\ &= \frac{2700}{3000} \times 100 \times \frac{3000}{2500} \times 100 \times \frac{1}{100} \\ &= \frac{2700}{2500} \times 100 \\ &= \frac{2700}{2500} \times 100 \\ &= 108 \end{aligned}$$

Le prix du produit a augmenté de 8 % de 2020 à 2024.

De 2020 à 2022, le prix du produit a augmenté de 20 %. L'indice qui permet de comparer l'année 2022 et l'année 2024 vaut $I^{P}_{2024/2022} = \frac{P_{2024}}{P_{2022}} \times 100 = \frac{2700}{3000} \times 100 = 90$. De 2022 à 2024, le prix a diminué de 10 %. Un calcul simple mais qui est faux ferait croire que le prix a augmenté de 20 %-10 %=10 %, alors que la vraie réponse est de 8 %. Les pourcentages ne s'additionnent pas.

5.1.2. Propriété de réversibilité

Connaissant V_t la valeur d'un produit à l'instant t et V_0 la même valeur à l'instant de référence, alors l'indice qui compare 0 à t en base 100 vaut :

$$I_{0/t} = \frac{100^2}{I_{t/0}} = \frac{100^2}{\frac{V_t}{V_0}\times 100} = \frac{V_0}{V_t}\times 100 \tag{5.5}$$

Avec l'exemple précédent, si un produit donné coûte 2500 FBu en 2020 et 3000 FBu en 2022, alors :

$$\begin{aligned} I^P_{2020/2022} &= \frac{P_{2020}}{P_{2022}}\times 100 = \frac{2500}{3000}\times 100 = 83{,}33 \\ \Delta &= I^P_{2020/2022} - 100 = 83{,}33 - 100 = -16{,}67 < 0 \end{aligned} \tag{5.6}$$

Le prix du produit a donc diminué de 16,67 % de 2022 à 2020 et non de 20 %.

5.2. Indices synthétiques

Certaines grandeurs sont des grandeurs complexes (composites) qui proviennent ou qui sont calculées à partir d'autres grandeurs [5]. C'est la cas par exemple d'une assiette de nourriture (haricot, banane, riz, huile, sel, piment, tomate, carotte, oignon, etc) et d'un prix d'un sac de riz (riz, sac, frais de déplacement, location de la boutique, etc).

La valeur d'un produit i à l'instant t vaut :

$$V_i(t) = P_i(t)\,Q_i(t) \tag{5.7}$$

où *P* représente le prix et *Q* la quantité.

La valeur globale de l'ensemble des produits vaut :

$$V(t) = \sum_i V_i(t) = \sum_i P_i(t)\,Q_i(t) \tag{5.8}$$

L'indice de la valeur globale vaut donc :

$$I^V_{t/0} = \frac{V(t)}{V(0)}\times 100 = \frac{\sum_i P_i(t)\,Q_i(t)}{\sum_i P_i(0)\,Q_i(0)}\times 100 \tag{5.9}$$

Par exemple, étudions le panier de trois produits de l'association « Transport de l'Agglomération de Montpellier (TAM) » vendus à une sous-population (non précisée ici) de 2010 à 2012.

	2010		2012	
Produit	Prix	Quantités	Prix	Quantités
Ticket simple	1,5	100	2	100
Abonnement hebdomadaire	35	40	50	50
Abonnement mensuel	130	35	175	40

Quelle est l'évolution des recettes de la TAM pour ce panier et la sous-population étudiée ? Pour répondre à cette question, utilise les indices de Laspeyres, de Paasche et de Fisher.

L'indice qui compare les années 2010 et 2012 est :

$$\begin{aligned} I^{V}_{2012/2010} &= \frac{V(2012)}{V(2010)} \times 100 = \frac{\sum_i P_i(2012) Q_i(2012)}{\sum_i P_i(2010) Q_i(2010)} \times 100 \\ &= \frac{2 \times 100 + 50 \times 50 + 175 \times 45}{1,5 \times 100 + 35 \times 40 + 130 \times 55} \\ &\simeq 159,02 \end{aligned} \tag{5.10}$$

De 2010 à 2012, il y a eu une augmentation de 59,02 % pour les recettes. La question qu'il est alors possible de se poser est de savoir si cette augmentation a été due à l'augmentation des prix ou à l'augmentation des quantités. Pour cela, nous allons calculer l'indice de Laspeyres, l'indice de Paasche et l'indice de Fisher.

L'indice de Laspeyres des prix vaut :

$$L^{P}_{t/0} = \frac{\sum_i P_i(t) Q_i(0)}{\sum_i P_i(0) Q_i(0)} \times 100 \tag{5.11}$$

L'indice de Laspeyres des quantités vaut :

$$L^{Q}_{t/0} = \frac{\sum_i P_i(0) Q_i(t)}{\sum_i P_i(0) Q_i(0)} \times 100 \tag{5.12}$$

L'indice de Paasche des prix vaut :

$$P^{P}_{t/0} = \frac{\sum_i P_i(t) Q_i(t)}{\sum_i P_i(0) Q_i(t)} \times 100 \tag{5.13}$$

L'indice de Paasche des quantités vaut :

$$P_{t/0}^{Q} = \frac{\sum_i P_i(t) Q_i(t)}{\sum_i P_i(t) Q_i(0)} \times 100 \tag{5.14}$$

L'indice de Fisher des prix vaut :

$$F_{t/0}^{P} = \sqrt{L_{t/0}^{P} P_{t/0}^{P}} \tag{5.15}$$

L'indice de Fisher des quantités vaut :

$$F_{t/0}^{Q} = \sqrt{L_{t/0}^{Q} P_{t/0}^{Q}} \tag{5.16}$$

5.2.1. Indice de Laspeyres

L'indice de Laspeyres des prix vaut :

$$\begin{aligned} L_{2012/2010}^{P} &= \frac{\sum_i P_i(2012) Q_i(2010)}{\sum_i P_i(2010) Q_i(2010)} \times 100 \\ &= \frac{2 \times 100 + 50 \times 40 + 175 \times 35}{1,5 \times 100 + 35 \times 40 + 130 \times 35} \\ &\simeq 136,48 \end{aligned}$$

L'indice de Laspeyres des quantités vaut :

$$\begin{aligned} L_{2012/2010}^{Q} &= \frac{\sum_i P_i(2010) Q_i(2012)}{\sum_i P_i(2010) Q_i(2010)} \times 100 \\ &= \frac{1,5 \times 100 + 35 \times 50 + 130 \times 40}{1,5 \times 100 + 35 \times 40 + 130 \times 35} \\ &\simeq 116,39 \end{aligned}$$

À quantités fixées en 2010, les prix ont augmenté de 36,48 % de 2010 à 2012. À prix fixés en 2010, les quantités ont augmenté de 16,38 % de 2010 à 2012. Selon Laspeyres, l'augmentation des recettes a été principalement due à l'augmentation des prix.

5.2.2. Indice de Paasche

L'indice de Paasche des prix vaut :

$$\begin{aligned} P_{2012/2010}^{P} &= \frac{\sum_i P_i(2012)Q_i(2012)}{\sum_i P_i(2010)Q_i(2012)} \times 100 \\ &= \frac{2\times 100 + 50\times 50 + 175\times 40}{1{,}5\times 100 + 35\times 50 + 130\times 40} \\ &\simeq 136{,}62 \end{aligned}$$

L'indice de Paasche des quantités vaut :

$$\begin{aligned} P_{2012/2010}^{Q} &= \frac{\sum_i P_i(2012)Q_i(2012)}{\sum_i P_i(2012)Q_i(2010)} \times 100 \\ &= \frac{2\times 100 + 50\times 50 + 175\times 40}{2\times 100 + 50\times 40 + 175\times 35} \\ &\simeq 116{,}52 \end{aligned}$$

À quantités fixées en 2010, les prix ont augmenté de 36,62 % de 2010 à 2012. À prix fixés en 2010, les quantités ont augmenté de 16,52 % de 2010 à 2012. Selon Paasche, l'augmentation des recettes a été principalement due à l'augmentation des prix.

5.2.3. Indice de Fisher

L'indice de Fisher des prix vaut :

$$F_{t/0}^{P} = F_{2012/2010}^{P} = \sqrt{L_{2012/2010}^{P} P_{2012/2010}^{P}} = \sqrt{136{,}48\times 136{,}62} = 136{,}55$$

L'indice de Fisher des quantités vaut :

$$F_{t/0}^{Q} = F_{2012/2010}^{Q} = \sqrt{L_{2012/2010}^{Q} P_{2012/2010}^{Q}} = \sqrt{116{,}39\times 116{,}52} = 116{,}45$$

À quantités fixées en 2010, les prix ont augmenté de 36,55 % de 2010 à 2012. À prix fixés en 2010, les quantités ont augmenté de 16,45 % de 2010 à 2012. Selon Fisher, l'augmentation des recettes a été principalement due à l'augmentation des prix.

En guise de conclusion, l'augmentation de 59,02 % des recettes globales a été plus due à l'augmentation des prix (36,5 %) qu'à l'augmentation des quantités (16,4 %).

5.3. Relation entre les indices synthétiques

Sachant que les recettes sont données par le produit des prix et des quantités, alors :

$$V = P \times Q \tag{5.17}$$

D'où l'indice pour un seul produit vaut :

$$I^V_{t/0} = I^P_{t/0} \times I^Q_{t/0} \times \frac{1}{100} \tag{5.18}$$

Pour un ensemble de produits et en se basant sur les indices synthétiques ci-hauts, cette formule devient :

$$\begin{aligned} I^V_{t/0} &= L^P_{t/0} \times P^Q_{t/0} \times \frac{1}{100} \\ &= P^P_{t/0} \times L^Q_{t/0} \times \frac{1}{100} \\ &= F^P_{t/0} \times F^Q_{t/0} \times \frac{1}{100} \end{aligned} \tag{5.19}$$

En effet :

$$L^P_{t/0} \times P^Q_{t/0} \times \frac{1}{100} = L^P_{2012/2010} \times P^Q_{2012/2010} \times \frac{1}{100} = 136,48 \times 116,52 \times \frac{1}{100} = 159,0265 \simeq 159,03$$

$$P^P_{t/0} \times L^Q_{t/0} \times \frac{1}{100} = P^P_{2012/2010} \times L^Q_{2012/2010} \times \frac{1}{100} = 136,62 \times 116,39 \times \frac{1}{100} = 159,012 \simeq 159,01$$

$$F^P_{t/0} \times F^Q_{t/0} \times \frac{1}{100} = F^P_{012/2010} \times F^Q_{012/2010} \times \frac{1}{100} = 136,55 \times 116,45 \times \frac{1}{100} = 159,0125 \simeq 159,01$$

En faisant un retour à la propriété de réversibilité et en notant $\bullet = P \, ou \, Q$, alors :

$$\begin{aligned} L^\bullet_{0/t} &= \frac{100^2}{P^\bullet_{t/0}} \\ F^\bullet_{0/t} &= \frac{100^2}{F^\bullet_{t/0}} \end{aligned} \tag{5.20}$$

Ainsi :

$$L^P_{2010/2012} = \frac{100^2}{P^P_{2012/2010}} = \frac{100^2}{136,62} = 73,19573 \simeq 73,20$$

$$F^Q_{2010/2012} = \frac{100^2}{F^Q_{2012/2010}} = \frac{100^2}{116,45} = 85,87377 \simeq 85,87$$

Exercice d'application - 4

Il vous faut produire deux indices composites (Laspeyres et Paasche, base 1997 = 100) du prix des fleurs en vente sur le marché de la capitale en 2000. Vous disposez pour ce faire des trois doubles séries suivantes :

Tableau des ventes (milliers d'unités) et des prix moyens observés durant l'année (€) 1997 - 2000						
	Roses		Orchidées		Oeillets	
Années	Prix	Ventes	Prix	Ventes	Prix	Ventes
1997	1,23	3256	2,45	597	0,56	5698
1998	1,25	4567	2,79	612	0,63	5893
1999	1,19	3972	3,06	624	0,48	5364
2000	1,32	3587	2,98	658	0,67	6971

Références bibliographiques

1. Leboucher L, Voisin M-J. Introduction à la statistique descriptive: cours et exercices avec tableur. 3e éd. Toulouse: Cépaduès-éditions; 2015. 208 p.

2. Rousson V. Statistique appliquée aux sciences de la vie. Paris Berlin Heidelberg [etc.]: Springer; 2013. 327 p. (Statistique et probabilités appliquées).

3. Franklin S, Walker C. Méthodes et pratiques d'enquête. Ottawa: Statistique Canada; 2010. 434 p.

4. Krickeberg K. Petit cours de statistique. Berlin Heidelberg New York [etc.]: Springer; 1996. 147 p.

5. Py B. Statistique descriptive: nouvelle méthode pour bien comprendre et réussir. 5e éd. Paris: Économica; 2007. 353 p.

6. Monino J-L. Statistique descriptive: QCM et exercices corrigés, 4 sujets d'examen corrigés, avec rappels de cours. 5e éd. Malakoff: Dunod; 2017. 288 p. (TD).

7. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. Hoboken, NJ: Wiley-Interscience; 2005. 360 p. (Wiley series in probability and mathematical statistics. Applied probability and statistics).

8. Lejeune M. Statistique: La théorie et ses applications. Deuxiéme èdition avec exercices corrigè. Paris: Springer-Verlag Paris; 2010. 434 p. (Statistique et probabilitès appliquèes).

9. Dodge Y, Rousson V. Analyse de régression appliquée. 2e éd. Paris: Dunod; 2004. 288 p. (Éco sup).

10. Cornillon P-A, Matzner-Løber E. Régression: théorie et applications. Paris: Springer; 2007. 302 p. (Collection Statistique et probabilités appliquées).