Université M'Hamed BOUGARA de Boumerdès

Faculté des Hydrocarbures et de la Chimie

Département d'Économie
et
Commercialisation des Hydrocarbures

Statistique descriptive

Razika TAHI

Année universitaire **2021-2022**

Ce polycopié correspond, mis à part le début du premier chapitre, à la partie 'Statistique' du programme officiel du Tronc Commun en Sciences et Technologies. Il s'agit du quatrième module de mathématique (Probabilité Statistique).

Théoriquement, les mathématiques appartiennent aux sciences exactes, et pourtant d'après Edmond et Jules de Goncourt « La statistique est la première des sciences inexacte » !! Alors exact ou inexacte ? Mark Twain, nous permet d'avoir un premier avis avec sa célèbre citation « Il y a trois sortes de mensonges : les mensonges, les sacrés mensonges et les statistiques ». Puis, il précise que « Les faits sont têtus. Il est plus facile de s'arranger avec les statistiques ». Et même Abe Burrows le confirme en considérant que « La raison des statistiques, c'est de vous donner raison ». Alors, si vous voulez vraiment apprendre à 'mentir scientifiquement', enfin disons « s'arranger avec les statistiques », il suffit de suivre ce cours. Parole d'une statisticienne ☺.

Avant d'aborder le programme officiel du module de statistique, un bref aperçu introductif de l'évolution de la statistique nous permettra de mieux comprendre ses applications dans notre environnement quotidien en général, et dans le monde scientifique en particulier.

SOMMAIRE

0. Introduction

La science commence généralement par la mesure. La statistique comme beaucoup d'autres sciences, s'est développé sous l'influence du besoin d'évaluation quantitatif qu'a ressenti l'Homme. Pour agir, il lui fallait d'abord connaître, et pour cela, il devait savoir compter. Cette nécessité de compter a donc correspondu à une pensée purement pratique et utilitaire. Dans les sociétés humaines organisées les plus ancienne, les données chiffrées portaient essentiellement sur la population, et ses conditions matérielles d'existences. Les premiers recensements de la population, dont les plus anciens remontent à plusieurs milliers d'années, avaient pour objectifs majeur la connaissance du nombre de personnes valides dont disposait l'Etat, pour d'éventuelles guerres, ainsi que le nombre de citoyens en mesure de payer un impôt. Les premières statistiques humaines ont été élaborées surtout pour répondre tout d'abord à des préoccupations démographiques, puis militaires et ensuite fiscales.

Ce n'est que vers le XVIIe siècle que la statistique commence à être considérée comme une science autonome, avec la formation de deux écoles (allemande et anglaise).

L'École allemande, dite 'descriptive', est fondée par Herman Conring[1] dont les travaux seront poursuivi pas Gottfried Achenwal[2] qui proposa la première fois, en 1748, le terme 'Statistik'[3]. Ce n'est qu'un siècle plus tard, en 1835, que l'académie française le reconnaîtra.

L'École anglaise, plus connue sous le nom de l'École de 'l'arithmétique politique', et fondé par John Graunt[4] et William Petty[5]. Cette École met en évidence, au-delà de la description, certaines

[1] **Conring Herman** : érudit allemand né en 1606 et mort en 1681. Professeur de médecine à Helmestedt (1636), il contribua à répandre les idées de Hawey sur la circulation sanguine, et défendit les idées modernes en chimie. Son ardeur novatrice le conduisit à critiquer certaines théories mercantilistes, et à aborder la question des statistiques.

[2] **Gottfried Achenwal** : économiste allemand né en 1719, et mort en 1772.

[3] En allemand, le mot 'Statistik' vient du latin 'Status', qui signifie Etat.

[4] **John Graunt** : marchand d'étoffes du XVIIe siècle à Londres, est né en 1620 et mort en 1674. Il s'intéresse à la connaissance des hauts faits de l'État anglais, non seulement du point de vue démographiques, mais aussi du point de vue sanitaire, commercial et religieux, qu'il décrit dans un bulletin (en 1962), représentant une sorte d'inventaire des qualités de l'État.

[5] **Wililam Petty (Sir)** : chirurgien, homme d'affaire et économiste anglais est né en 1623 et mort en 1687. Il fit un peu de tous les métiers, tel que garçon de cabine dans une entreprise de navigation, puis médecin de bord des troupes anglaises en Irlande. Il fut aussi organisateur du cadastre en Irlande, où il s'agissait d'enregistrer, et de numéroter les terres. Influencé par les travaux de J. Graunt, dont il était ami, il devint un remarquable statisticien. Appartenant au courant mercantiliste, avec un esprit curieux, et remarquable, il est le précurseur des physiocrates par son analyse du circuit économique, mais aussi du courant libéral classique par sa théorie de la valeur, par le rôle qu'il attribue à l'intérêt personnel, et à la démographie. Son ouvrage 'Aritmética politica', écrit en 1680, et parfois considéré comme le premier livre de statistique politique et économique.

permanences statistiques, tel que le rapport du nombre de naissances masculine à celui des naissances féminines (Graunt 1662).

Quelques années plus tard, Edmond Halley[6] présente une table de mortalité qui est à la base des travaux actuariels contemporains, puis John Peter Sismilch[7] publie d'importants travaux sur le taux de masculinité à la naissance, et son évolution jusqu'à l'âge de 20 ans. Il cherche à interpréter les résultats, et à remonter aux causes.

Jusqu'au XVIIIe siècle, l'enregistrement des faits conservait un caractère passif, on accumulait sans méthode. Par la suite, avec le développent de la science, les savants se soucient de plus en plus de l'exactitude dans la connaissance, ce qui les amènent à s'occuper des erreurs dans les calculs, à supputer les chances (probabilités) d'arriver à la connaissance exacte.

C'est l'étude des jeux du hasard (ils étaient très en vogue à la Cour de France au XVII et XVIIIe siècle) qui va permettre de développement d'une méthode d'interprétation, avec la théorie des probabilités.

Au début du XIXe siècle, Pierre Simon de Laplace[8], dans sa théorie analytique des probabilités (1812), met en évidence les avantages que l'on peut tirer de cette théorie dans l'étude des phénomènes naturels, dont les causes sont trop complexes pour qu'on puisse les connaître toutes, et les analyser individuellement. Adolphe Quételet[9] étend le champ d'application de la méthode à l'étude anthropométrique, psychologique, et sociale des êtres humains. Il écrit, en 1835, un ouvrage 'Sur l'Homme et le développement de ses facultés', dont le sous-titre est révélateur 'Essai de physique sociale'. Il écrit ensuite, en 1846, des lettres sur la théorie des probabilités appliquées aux sciences morales et politiques. Pour lui, il est possible, par la statistique appliquée aux actes humains, de constituer une science véritable qu'il appelle la 'physique sociale'. Sur son initiative, se réunit à Bruxelles, en 1853, le premier congrès international de statistique, précurseur de l'actuel Institut International de Statistique, fondé à Londres en 1885. À la suite de ces travaux, et ceux de Sir Francis Galton[10], qui ont porté essentiellement à développer des méthodes de mesures devenues classique par la suite, tel que les tests, l'étalonnage et les méthodes de corrélation, Kark Pearson [11], considéré comme l'un des fondateurs de la science des statistiques, à laquelle il donne une base institutionnelle à

[6] **Edmond Halley** : astronome britannique né en 1656 et mort en 1742.
[7] **John Peter Sismilch** : démographe allemand né en 1707 et mort en 1797.
[8] **Pierre Simon de Laplace** : astronome, mathématicien et physicien français né en 1749 et mort en 1827.
[9] **Adolphe Quételet** : astronome, mathématicien, statisticien et démographe belge né en 1796 et mort en 1874.
[10] **Sir Francis Galton** : homme de science britannique né en 1822 et mort en 1911.
[11] **Karl Pearson** : mathématicien et statisticien britannique né en 1857 et mort en 1936.

l'Universtity College de Londres, fonde la biostatistique, ou biométrie, prolongé maintenant dans le domaine de l'expérimentation thérapeutique. De même la liaison entre l'observation statistique et l'économie aboutit à la création de l'économétrie.

Les travaux de James Maxwel[12] aboutissant à la théorie cinétique des gaz, ont été le point de départ de la mécanique statistique. Ceux de Sir Ronald Aylmer Fisher[13] sur l'expérimentation en agronomie ont été la base d'une théorie générale des plans d'expériences.

Dans le domaine des sciences humaines, les études de Charles Edward Spearman[14] sur le comportement des individus, sont à l'origine du développement des méthodes d'analyses factorielles, prolongement logique de l'étude des corrélations.

Ces dernières décennies, la méthode statistique s'est révélé un auxiliaire indispensable pour la gestion des entreprises : études de marché, contrôle budgétaire ou autres, gestion des stocks, etc. Prolongée par la théorie des jeux et la théorie de la décision, elle a donné naissance aux méthodes de la recherche opérationnelle.

L'apparition de puissants moyens de calculs, a permis de mettre en œuvre de nouvelles méthodes de statistiques descriptives ne recourant pas à des modèles, et ni même à des hypothèses, mais à de grands tableaux de données multidimensionnelles. Ces méthodes sont regroupées sous le nom d'analyse des données.

Ce bref historique met en évidence les domaines aussi variés que nombreux, qui ont contribués au développement de la statistique (médecine, commerce, astronomie, géologie, transport, automatisation, démographie, mathématique, physique, agronomie, économie, biologie, psychologie, gestion des entreprises, informatique, etc.).

1. Objectifs et terminologie de base de la statistique

Les applications de la statistique, et du calcul des probabilités, sont nombreuses. Il serait difficile aujourd'hui de citer une branche des sciences (physique, chimie, biologie, sciences sociales, etc.), une branche technique quelconque, ou même un aspect banal de la vie quotidienne, qui ne soit pas, d'une façon ou d'une autre, concerné par le développement qu'a connu , surtout au XXe siècle, la science du hasard. Si, il y a quelques décades, on pouvait citer des exemples d'application de la statistique probabiliste, dans différents domaines, l'expansion

[12] **James Clerk Maxwell** : physicien écossais né en 1831 et mort en 1879.
[13] **Sir Ronald Aylmer Fisher** : biologiste et statisticien britannique né en 1890 et mort en 1962.
[14] **Carles Edward Spearman** : psychologue britannique né en 1863 et mort en 1945.

de la méthode a été si large, et si rapide, qu'aujourd'hui, on serait plus bref en cherchant les branches où elle ne sert pas.

Dans ce chapitre nous présentons dans un premier paragraphe les principaux objectifs de la statistique. Un second paragraphe, concerne quelques définitions de base en terminologie de la statistique.

1.1. Objectifs de la statistique

Au début de toute démarche scientifique, ou technique, il y a identification de l'objet d'étude, une action de repérage, sinon de mesure qui s'impose. Par sa première intervention, la statistique permet de savoir ce que l'on traite (quoi ?).

Mais, comme les observations sont en générales limitées en nombre, il faut savoir comment on peut généraliser, passer de la partie au tout, de l'échantillon à la population totale. La seconde tâche de la statistique est celle de l'estimation.

Lorsqu'on confronte plusieurs populations, ou des populations, et des échantillons qui en sont tirés (les populations sont ici des ensembles d'Hommes ou d'objets, de qualité, ou de quantités relatives à certains caractères), il faut pouvoir éprouver la validité des comparaisons. Rechercher si les différences, ou les similitudes, sont significatives, c'est faire des tests de signification, la troisième tâche de la statistique.

Juger d'une évolution, d'un phénomène, dégager une tendance cachée par d'innombrables fluctuations à court terme, et par des variations accidentelles, c'est encore un traitement statistique important. Le travail sur série temporelle, le dégagement des évolutions chronologiques, est une quatrième tâche de la statistique.

Enfin, la recherche des liaisons entre deux, ou plusieurs, phénomènes représente la cinquième tâche de la statistique. C'est les cas, par exemple, de la longueur et la température pour le physicien, la température et les rhumes de cerveau pour le médecin, la vitesse des voitures et le nombre d'accidents pour l'assureur, l'industrialisation et le niveau de revenu pour l'économiste. Ces cas impliquent des travaux délicats, dont la technique est toujours à peu près la même, quelques soit le domaine d'application.

Il n'est pas sans intérêt de revenir sur quelques un de ces travaux avec plus de détails.

1.1.1- Identification. Problème de mesure.

L'expérience montre, et différents arguments théoriques confirment, que les courbes de densité de distribution, établies pas les données de l'observation d'une variable, ont des formes particulières, dont certaines se rencontrent si souvent, qu'elles ont reçu la dénomination de loi.

Ainsi, la taille des soldats, les points d'impacts des balles tirés par un fusil tenu par un étau fixe, et bien d'autres phénomènes vont nous montrer des distributions semblables. Très tôt, on a été amené à conclure que de telles distributions de population (soldats, points d'impacts, etc.) obéissent à certaines régularités dont l'expression est possible : tel est le but des lois de statistiques. Il existe de très nombreuses lois statistiques.

Les lois les plus couramment utilisées ont une forme analytique connue. On peut alors valablement remplacer l'ensemble des données par quelques grandeurs caractéristiques. Les meilleures sont celles qui définissent la position et la dispersion des données, et qui sont les paramètres dont la loi dépend. Les paramètres de position caractérisent la valeur centrale des données (moyenne, médiane ou mode), et les paramètres de dispersion précisent les écarts entre elles, et une valeur type comme la moyenne. On verra, dans les cours qui suivent, que pour une distribution normale, la moyenne arithmétique, et une unité de dispersion appelée écart type, sont les paramètres qui permettent d'exprimer la loi, et donnent à eux seuls une information très riche, et très suffisante sur l'ensemble des données.

1.1.2 - Estimation. De l'échantillon à la population.

Tout observateur n'effectue jamais qu'un nombre limité d'observations, et cet ensemble d'observation qu'il effectue, n'est qu'un échantillon, parmi la population infiniment grande de série d'observations, qu'il conviendrait de faire pour atteindre la réalité.

Une opinion répandue, chez les non spécialistes, est que les théories de l'échantillonnage ne s'appliquent que dans les cas de sondages, volontairement limité à une partie de la population (d'Hommes ou d'objets). Il faut généraliser, et dire que l'observateur n'atteint jamais que des sous populations pour lesquelles il a l'ambition d'induire des relations valables au niveau de la population entière. Même si l'on interroge 20 millions d'algériens, on n'atteint qu'une sous population de la vraie population algérienne, qui seule et justifiable des lois statistiques. Ce passage de l'échantillon à la population est le passage de la mesure expérimentale à la mesure vraie, de l'être ordinaire à l'être scientifique, de la relation expérimentale à la loi. C'est une démarche essentielle de l'activité scientifique, qui consiste à calculer une estimation, ou valeur approchée, pour une population globale, dont on ne connaît qu'une partie. On détermine des

intervalles de confiance, dans lesquels les vraies valeurs à estimer sont une forte probabilité de se trouver.

1.1.3 - Tests d'hypothèses.

Les tests d'hypothèses représentent une branche décisive de l'analyse statistique. Leurs techniques progressent très rapidement, et leurs champs d'application est immense. Une certaine hypothèse concerne une, ou plusieurs, populations, une, ou plusieurs, lois de distribution ayant été formulées. L'utilisation d'un test statistique approprié permet de déterminer dans quelle mesure cette hypothèse apparaît infirmée par l'information apportée par un, ou plusieurs, échantillons, c'est-à-dire, dans quelle mesure cette information peut être moyens dont la statistique dispose pour étudier des données échelonnées dans le temps sont également considérée comme incompatible, avec l'exactitude de l'hypothèse formulée.

1.1.4 - Etude des évolutions chroniques.

Les nombreux, et efficaces. Contrairement à une opinion fréquente, ces études ne sont pas exclusivement limitées au domaine économique. On les trouve dans toutes les sciences de l'Homme, et de la nature, par exemple en météorologie, ou bien en médecine où les maladies ont à la fois leurs pointes saisonnières, et de véritables cycles qui s'étendent sur plusieurs années.

1.1.5 - Recherche des liaisons entre phénomènes.

Partant de la liaison fonctionnelle entre deux variables (y étant expliquée par x, et à toute valeur de x correspond une valeur précise de y), on s'aperçoit dans tous les domaines de la recherche, soit que l'observation de ces variables n'a pu être conduite de façon parfaite, c'est-à-dire qu'elle est affectée d'erreurs de mesure, soit que la liaison établie n'est valable que dans les grandes lignes, c'est-à-dire qu'elle est affectée par des facteurs incontrôlables, par des nombreuses causes de variabilité. Le but des méthodes statistiques (corrélation-régression) est de mesurer l'ampleur de ces perturbations, et éventuellement de 'filtrer' les phénomènes étudiées pour que la nature, et la solidité de leurs liaisons apparaissent au grand jour.

Il est certain, par exemple, que la consommation d'essence dans les unités territoriales données (comme une wilaya), est fonction du parc automobile. Mais, la liaison est affectée par de multiples autres variable : revenu de la population, répartition par profession (un représentant de commerce et un travailleur sédentaire), état des routes, structure du commerce de l'essence, mouvement du tourisme, etc. De même, la charge de rupture d'un acier dépend de sa teneur en

carbone. Mais, la liaison peut être plus ou moins troublée par les variations de sa teneur en manganèse, et en silicium. Dans tous les domaines, on trouve des cas où l'expérience reste brute, et où faute de pouvoir physique, éliminer les facteurs secondaires, ou perturbateurs, on est obligé de ruser, et de chercher par le calcul, quelle est leur influence probable.

La conclusion à tirer, pour cette section méthodologique, et qu'on est loin du tableau de Claude Bernard[15] (Introduction à l'étude de la médecine expérimentale), qui est resté longtemps un modèle pour les scientifiques. La succession 'observation-hypothèse-expérimentation' paraît moins simple qu'on ne le pensait au début du siècle passé. Non pas, certes, que l'étude de l'aléatoire entraîne la négation des moyens élaborés par le positivisme[16], mais plutôt, dans la même ligne, un élargissement de la recherche, la découverte de nouveaux outils au nom de l'inquiétude. Seul, un certain déterminisme[17] tranquille est condamné, et dans les grandes lignes d'objectifs de la connaissance, et méthodes, restent les mêmes.

La méthode, dite expérimentale, qui consiste à isoler certains facteurs, et à laisser varier quelques un d'entre eux, en laissant les autres constants, reconnait maintenant ses limites. Son application, dit A. Laurent, « exige la réalisation de conditions susceptibles de caractériser seulement des expériences purement idéales, qu'on peut tout au plus, quand elles constituent autre chose que de simples abstractions dépourvus de caractères opérationnel, comme des cas limites d'expériences réalisable ». La méthode statistique reprend un contact plus réaliste, à la fois plus timide, puisqu'elle reconnaît que l'observateur ne peut prendre les observations que comme elles sont, et comme elles viennent, dans un écheveau embrouillé de paramètres, et de variation de facteurs non contrôlés, et plus ambitieux aussi, puisqu'elle s'attache à réduire l'incertitude des conclusions tenant à l'imprévision des mesures, à l'impureté des observations, à l'information fragmentaire ou à l'enchevêtrement des relations.

Sans la science du hasard, l'induction, ce 'procédé' logique hasardeux qui consiste, dit encore A. Laurent à « tirer des conclusions d'une portée générale à partir d'informations incomplètes et particulières », n'aurait qu'une efficacité limitée. La statistique lui rend toute sa puissance, en mesurant l'incertitude des conclusions inductives, en contrôlant cette incertitude, et même, pourrait-on dire une image assez grossière, en 'domestiquant' l'aléatoire.

[15] **Claude Bernard** : médecin et physiologiste français né en 1813 et mort en 1878. Il est considéré comme le fondateur de la médecine expérimentale.
[16] Le **positivism**e ne cherche pas à connaître la nature intrinsèque des phénomènes, mais seulement les relations entre les phénomènes. Il met l'accent sur les lois scientifiques et conteste la notion de cause.
[17] Le **déterminisme** est la théorie considérant que la succession des évènements et des phénomènes est due au principe de causalité, c'est-à-dire que tout phénomène a une cause.

1.2. Terminologie

Après une définition de la statistique, nous présentons dans ce paragraphe une partie de la terminologie utilisée en statistique, à travers des concepts élémentaires utiles pour la suite de ce cours.

1.2.1. La statistique et les statistiques.

Il y a souvent confusion entre la statistique et les statistiques. Le fait que l'une soit au singulier et l'autre au pluriel n'est pas la seule différence. Nous précisons dans ce qui suit les définitions des deux notions.

- La statistique. Les définitions de la statistique sont nombreuses. Soit la définition donnée par A. Vessereau[18], qui englobe dans les méthodes statistiques « Toute les recherches dans lesquelles le grand nombre, et l'enchevêtrement des facteurs, exigent une technique d'interprétation basée sur la connaissance des lois du hasard ».

Nous pouvons résumer quelques caractéristiques essentielles de la statistique comme suit.

- Méthode générale qui relie des domaines très divers des sciences. Selon l'expression de L. March, c'est en quelque sorte une 'langue commune'.

- Elle porte sur des ensembles et sur leurs relations.

- Elle donne des conclusions probables sur des ensembles imparfaitement connus.

- Elle aboutit à des lois qui sont des propriétés de groupe.

- Elle se distingue des mathématiques, sciences du certain, alors qu'elle est une science de l'incertain (Kendall[19]), et elle cherche à établir les limites de l'incertitude.

- Elle s'appuie sur les lois du hasard, et elle exploite à fond les quelques domaines où le hasard 'organisé' est roi : étude des erreurs et étude des échantillons.

- Les statistiques. Les statistiques représentent le type d'information obtenue en soumettant les valeurs numériques, se rapportant aux affaires humaines ou aux phénomènes naturels, à des opérations mathématiques.

- **Différence entre la statistique et les statistiques** (ou une statistique). Il ne s'agit pas dans ce cas d'une différence grammaticale seulement entre ces deux termes (singulier, pluriel). Il ne

[18] **André Vessereau** : statisticien français né en 1907 et mort en 1990.
[19] **Sir Maurice George Kendall** : statisticien britannique né en 1907 et mort en 1983.

faut pas confondre la statistique, qui est une science dont le domaine d'investigation a pour objet la collecte et l'agencement de données, et une (ou les) statistique(s) qui est un ensemble de données chiffrées sur un sujet précis.

1.2.2. Notions de population, d'échantillon, variables, modalités.

- La population statistique et l'unité statistique. La population statistique est constituée d'un ensemble d'éléments, ou d'individus, appelée unité statistique. Un individu peut être un être humain, un animal ou un objet. Ainsi, lors d'une étude statistique, il faut définir en premier sur qui porte l'analyse.

*Exemple. * Le personnel enseignant d'une université constitue la population statistique, et un enseignant représente l'unité statistique.*

** Le cheptel de moutons constitue la population statistique, et le mouton, l'unité statistique.*

- Echantillon : En statistique, un échantillon est un nombre d'individus, représentatif, choisi à partir d'une population donnée. Ainsi, un échantillon est une partie (un sous-ensemble) de la population. Sa taille est le nombre d'individus qui le composent. Un échantillon sert de support à la statistique descriptive, afin d'étudier et analyser les propriétés de la population.

- Variable statistique (ou caractère statistique). On appelle variable statistique (ou caractère) la chose (ou le phénomène) que l'on étudie, et qui est commune à tous les individus de la population de référence. L'ensemble des résultats s'appelle série statistique. Le caractère est une propriété possédée par les unités statistiques, permettant de les décrire, et de les distinguer les uns des autres. Toute unité statistique peut être étudiée selon un, ou plusieurs, caractères. Après avoir défini en premier la population, il faut préciser ensuite sur quoi porte l'étude statistique, d'où l'intérêt de définir la variable statistique.

*Exemples. * Soit la population d'un pays, dont l'unité est l'habitant, qui peut être décrite selon plusieurs caractères tels que l'âge, le genre, la profession, etc.*

** Pour la population d'une entreprise, on peut considérer comme caractères l'ancienneté, la catégorie professionnelle, l'âge, le salaire, etc.*

** Soit la population de pièces fabriquées dans une usine. On peut considérer comme caractère le diamètre, le poids, la composition de la matière, etc.*

- Les modalités du caractère.

Les modalités du caractère représentent les divers cas, ou situations, susceptibles d'être prises par le caractère. Un caractère peut posséder une, ou plusieurs, modalités.

*Exemple. * Être footballeur : oui/non (deux modalités).*

** Instruction : lettré/illettré.*

Propriété. Les modalités d'un caractère doivent être à la fois exclusives et exhaustives.

- Exclusive : une unité statistique ne peut prendre qu'une seule des modalités du caractère.

Exemple. Une table peut être ancienne ou neuve. La table ne peut être simultanément ancienne et neuve. On dit alors, que les deux modalités s'excluent mutuellement, ou qu'elles sont exclusives.

- Exhaustive : toutes les possibilités doivent être considérées. Chaque individu doit donc posséder une seule des modalités du caractère.

Exemple. Dans le cas de l'exemple précédant, concernant la table, il y a deux possibilités : neuve ou anciennes.

1.2.3. Différentes types de variables statistiques.

Une variable statistique peut être qualitative ou quantitative, discrète ou continue.

- Variable statistique qualitative. Une variable qualitative est une variable où les différentes modalités ne sont pas mesurables. Ses valeurs, ou modalités, s'expriment de façon littérale, ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme, soustraction ou division, n'ont pas de sens.

*Exemple. * Nationalité.*

** Profession.*

** Niveau d'instruction.*

On peut distinguer plusieurs types de variables qualitatives : dichotomique, nominale, ordinale.

- Variable statistique dichotomique. C'est une variable qualitative qui ne peut prendre que deux modalités exclusives entre elles.

*Exemple. * Oui ou non.*

** Lettré ou illettré.*

** Masculin ou féminin.*

- Variable statistique nominale. C'est une variable qualitative dont les modalités ne sont pas ordonnées. Ainsi, ce type de variable peut décrire un nom, une étiquette ou une catégorie sans ordre naturel.

Exemple. La liste des différents modules enseignés lors d'un semestre peuvent être placés dans n'importe quel ordre : mathématique, physique, chimie, statistique, ou bien chimie, physique, statistique, mathématique.

- Variable statistique ordinale. C'est une variable qualitative dont les modalités présentent un ordre naturel de ses valeurs possibles, mais les distances entre les valeurs ne sont pas définies. Ainsi, on peut par exemple, considérer que selon un certain sens la modalité A est moins forte que la B, qui est moins forte que la C, qui est moins forte que la D, qui est moins forte que la E, etc. Généralement, les variables ordinales ont des échelles par catégorie.
Exemple. Dans un questionnaire d'évaluation, il est souvent demandé de choisir entre plusieurs modalités : très bien, bien, moyen, passable, faible. La réponse indique seulement une catégorie, mais il existe un ordre naturel de ces catégories.

- Variable statistique quantitative. Une variable statistique est quantitative quand elle prend des valeurs numériques, et ces valeurs sont des nombres sur lesquels des opérations arithmétiques ont un sens. Une variable statistique quantitative peut être continue ou discrète.

- Variable statistique quantitative continue. Une variable statistique est continue, lorsque les différentes valeurs prises par celle-ci sont en nombre indéterminé à priori, dans un intervalle de valeur, c'est-à-dire que la variable statistique peut prendre n'importe quelle valeur intermédiaire entre deux valeurs données.

*Exemple. * Vitesse d'un objet en déplacement.*

** Salaire mensuel compris entre 1200 et 1600 Unités Monétaire (UM).*

- Variable statistique quantitative discrète. Une variable statistique quantitative est discrète si elle ne peut prendre qu'un nombre fini (ou dénombrable) de valeurs.

*Exemple. * Nombre d'enfants dans un ménage.*

** Nombre de pièces dans un logement.*

Remarque : en pratique, en raison de l'imprécision des mesures, toutes les variables sont considérées comme discrètes, bien que la nature de certaines soit continue.

Les modalités d'un caractère qualitatif peuvent faire l'objet d'une nomenclature, ou énumération. La nomenclature doit être en principe courte (une dizaine de modalités pour une étude statistique simple), mais les exigences de l'étude sont parfois tels que la nomenclature occupe des volumes entiers.

2 - Séries statistiques à une variable

Dans le programme officiel, ce chapitre est composé de huit paragraphes[20]. Mis à part les deux premiers paragraphes que nous avons fusionnés, le plan de ce cours suit le programme officiel. Nous présentons en introduction, l'élaboration d'un tableau statistique à partir du dépouillement de données.

Le premier résultat d'un dépouillement de données est normalement un tableau de nombres. Ces nombres peuvent être de simples résultats de dépouillement par classe. Nous présentons à travers un exemple, l'élaboration d'un tableau statistique à une variable.

Soit le tableau qui suit concernant les effectifs des travailleurs dans une entreprise par tranche d'âge.

[20] **Programme officiel du module Probabilité Statistiques**

II - Séries statistiques à une variable :

1) Effectif. Fréquence. Pourcentage.
2) Effectif cumulé. Fréquence cumulée.
3) Représentations graphiques : diagramme à bande, diagramme circulaire, diagramme en bâton. Polygone des effectifs (et des fréquences). Histogramme. Courbes cumulatives.
4) Caractéristiques de position : mode, moyenne arithmétique, moyenne harmonique, moyenne géométrique, médiane, quantiles.
5) Caractéristiques de dispersion : étendue, variance et écart-type, coefficient de variation, quartiles, étendue interquartile.
6) Caractéristiques de forme.
8) Représentation graphique des résultats à l'aide du box plot.

Titre du tableau : Effectifs des travailleurs dans une entreprise par tranche d'âge.

Âge	Dépouillement	Effectifs	Fréquence
25-29	𝍸	10	0,10
30-34	𝍸𝍸	23	0,23
35-40	𝍸	4	0 ,04
40-44	𝍸𝍸	15	0,15
45-49	𝍸𝍸𝍸	32	0,32
50 et +	𝍸𝍸	16	0,16
Total		100	1,00

Pour noter le dépouillement manuel, on évite les bâtons successifs presqu'impossible à compter sans se tromper. On dessine souvent quatre segments côte à côte, formant un carré, et représentant chacun une observation. Une cinquième observation est signifiée par une diagonale qui barre le carré. Ce tableau est élaboré en lisant les observations dans l'ordre où elles viennent. On évite ainsi les oublis ou les doubles comptes.

Exemple. Soit le tableau qui suit représentant la distribution des ménages selon le nombre d'enfants.

Tableau : Distribution des ménages selon le nombre d'enfants.

Nombre d'enfants/ménage	Fréquence
0	42
1	18
2	33
3	5
4 et plus	2
Total	100

Remarque. Un titre est absolument nécessaire pour tout tableau. La source des chiffres doit être précisée aussi sous le tableau, lorsque cela est possible.

Dans un tableau, on peut faire intervenir deux caractères (X et Y), il est appelé 'tableau de contingence' (cela fait l'objet de la deuxième partie de ce module).

2.1. Effectif, fréquence et pourcentage.

Si l'effectif et la fréquence sont des notions assez facile à comprendre, et surtout utilisés, la notion de pourcentage est plus complexe.

2. 1. 1. Effectif simple, total et cumulé.

Un effectif correspond au nombre d'observations que l'on étudie dans la population. Il répond donc à la question : sur combien d'observations porte l'étude statistique ?

Exemple : le nombre d'étudiants des 11 groupes d'étudiants de la promotion, qui suivent le cours de statistique.

On peut considérer l'effectif simple, et l'effectif total.

- Effectif simple (fréquence absolue). Cet effectif correspond à une seule partie de la population étudiée, que l'on écrit généralement n_i, où i correspond au nombre de toutes les sous-populations étudiées.

Exemple : les effectifs simples, dans l'exemple précédent, sont représentés par les effectifs des 11 groupes d'étudiants qui suivent le cours de statistique : n_1, n_2, n_3, …, n_{11}.

- Effectif total. Cet effectif représente l'effectif total de toutes les sous-populations, c'*est donc la somme de tous les effectifs simples*. Il est symbolisé par la lettre N.

Exemple : l'effectif total est l'ensemble des étudiants de la promotion, soit la somme des 11 groupes : $N = n_1 + n_2 + n_3 + \ldots + n_{11}$.

- Effectif cumulé. Un effectif cumulé est une suite d'addition, de proche en proche, des effectifs simples, classés par ordre, d'une série statistique à caractère quantitatif. L'effectif cumulé peut être croissant, ou décroissant, selon que l'on commence par le premier ou dernier effectif simple.

* **Effectif cumulé croissant** N_i :

$N_1 = n_1 \quad N_2 = n_1 + n_2 \quad N_3 = n_1 + n_2 + n_3 \ldots$ d'où $N_i = n_1 + n_2 + n_3 + \ldots + n_i$

* **Effectif cumulé décroissant** N'_i :

$N'_k = n_k \quad N'_{k-1} = n_k + n_{k-1} \quad N'_{k-2} = n_k + n_{k-1} + n_{k-2}$ ... d'où $N_i' = n_k + n_{k-1} + n_{k-2} + \ldots + n_i$

2. 1. 2. Fréquence.

Lorsqu'on veut résumer une grande quantité de données brutes, il est pratique de les distribuer en classes, ou catégories, et de déterminer le nombre d'individus, ou d'objets, appartenant à chaque classe.

La disposition des données sous forme d'un tableau, où pour chaque classe on a l'effectif correspondant, s'appelle distribution des effectifs ou des fréquences (conférer ci-dessous). Quand on représente des données par une telle distribution, on dit que ces données sont des données groupées.

On distingue la fréquence absolue de la fréquence relative. Dans la mesure où la fréquence absolue correspond en fait à l'effectif, lorsqu'il est écrit fréquence, cela sous-entend la fréquence relative.

- Fréquence (fréquence relative). La fréquence représente le taux de présence d'un effectif, correspondant au nombre d'individus appartenant à chaque classe, par rapport à l'effectif total. Elle est égale à l'effectif d'un caractère divisé par l'effectif total. C'est une valeur comprise entre 0 et 1.

- Fréquence absolue (effectif). Il s'agit de la répartition brute des données. Lorsque les données sont présentées individuellement, chaque donnée a la même fréquence unitaire d'apparition, leur effectif ou fréquence absolue est égal à 1.

- **Fréquences cumulées**. C'est le résultat de l'addition, de proche en proche, des fréquences d'une distribution observée. On distingué la fréquence cumulée croissante de la fréquence cumulée décroissante.

- Fréquence cumulée croissante. La fréquence cumulée croissante d'une modalité est la somme de la fréquence de cette modalité et des fréquences de toutes les modalités qui la précèdent.

- Fréquence cumulée décroissante. La fréquence cumulée décroissante d'une modalité est la somme de la fréquence de cette modalité et des fréquences de toutes les modalités qui la suivent.

- Distribution de fréquence. Une distribution de fréquence consiste à regrouper les données observées dans des classes. Pour cela, il faut tout d'abord déterminer le nombre de classe, puis ensuite la largeur des classes. En pratique, généralement le nombre de classes est d'environ 10 ± 5, c'est-à-dire qu'il varie entre 5 et 15, selon l'étude. La largeur d'une classe peut être calculée par la différence de la valeur maximale et la valeur minimale, divisé par le nombre de classe.

Pour illustre géométriquement une distribution de fréquence, on utilise soit un histogramme de fréquence (conférer paragraphe sur les représentations graphiques) ou soit un polygone de fréquence.

Chaque fréquence relative représente une proportion que l'on peut convertir aisément en pourcentage en la multipliant par 100.

2. 1. 3. Pourcentage.

La pratique des pourcentages est sans doute familière, et il peut paraître curieux de consacrer tout un paragraphe à une notion aussi simple. Cependant, l'expérience a montré que l'usage, souvent erroné, de ces chiffres montre que sa maitrise n'est pas toujours acquise.

Le signe %, ou la locution 'pour cent', sont absolument équivalent à une fraction dont le dénominateur serait 100[21]. Cette évidence est souvent perdue de vue, j'espère que ce n'est pas votre cas, et nombre d'utilisateurs disent 8% comme ils diraient 8 livres, ou 8 crayons, c'est-à-dire comme s'il s'agissait d'un nombre entier. Ainsi, 8 % est égal à 8/100 ou 0,05, et non pas 8. Ces erreurs sont assez fréquentes dans les médias.

Exemple : « Leurs revenus pétroliers ont chutés de 15,5 % à 65,52 dollars le baril, et les… » est une phrase incorrecte. Il faudrait écrire : « Leurs revenus pétroliers ont chutés de 84 à 65,52 dollars, soit une baisse d'environ 15,5 %, et les … ».

On utilise, généralement, les pourcentages soit pour faire apparaître quelle proportion d'un ensemble (ou population) possède un caractère particulier, ou soit pour comparer deux valeurs d'une même grandeur dans deux situations différentes.

Exemple : Soit la répartition des actifs occupés par secteur d'activité.

[21] Ce moyen n'a pas toujours été en usage. Jusqu'au XVII siècle, on préférait énoncer les fractions en les ramenant à avoir pour numérateur l'unité (1). Ainsi, un taux d'intérêt s'énonçait, par exemple, 'au denier vingt', c'est à dire 1/20. Comme 1/20 = 5/100, nous disons actuellement 5%. On emploie encore souvent de nos jours les 1/2, 1/4, ou 1/3.

Tableau 1 : Evolution des actifs occupés par secteur d'activité.

Années / Secteur d'activé	2000	2018
A	504 310	126 960
B	664 040	671 160
C	730 970	1 428 900
∑	1 899 320	2 227 020

L'information brute, exprimée en valeur absolue (ici en nombre de personnes) est difficile à interpréter. On ne peut pas comparer le nombre d'actifs occupés dans chaque secteur en 2000 et 2018, car l'ensemble des actifs occupés a augmenté entre les deux périodes.

Il faut donc transformer les valeurs absolues en valeurs relatives. Seules, ces dernières, exprimées en pourcentage, permettent d'effectuer des comparaisons dans le temps, et dans l'espace.

De la même façon, on ne peut pas comparer les variations du nombre des actifs occupés dans chacun des trois secteurs, car on ne part pas d'un nombre d'actifs occupés identique dans les trois secteurs.

Il faut donc transformer également les variations absolues en variations relatives. Ces dernières, exprimées en pourcentage, permettent de comparer les évolutions du nombre des actifs occupées dans chacun des trois secteurs d'activités entre les deux dates.

On peut donc déjà considérer deux sortes de pourcentages : les pourcentages de répartition (pourcentage instantané), et les pourcentages de variation (taux de croissance).

2. 1. 3. 1. Pourcentage de répartition.

Deux présentations sont possibles pour les pourcentages de répartition. Il faut éviter de mélanger les deux. La confusion, entre ces deux façons de procéder, est source d'un grand nombre d'erreurs. Nous allons préciser ces présentations à travers un exemple.

Exemple : Si le tout est égal à 250, et si la partie est égale à 50.

- Première méthode :

On appelle x la grandeur tel que : $x = \frac{\text{Valeur absolue de la partie}}{\text{Valeur absolue du tout}}$

Donc, pour notre exemple, on a : $x = \frac{50}{250} = 0{,}20 = \frac{20}{100} = 20\ \%$

- Deuxième méthode :

On appelle y la grandeur tel que x = y % avec $y = \frac{\text{Valeur absolue de la partie}}{\text{Valeur absolue du tout}}$ x 100

Donc, pour notre exemple, on a : $y = \frac{50}{250}$ *x 100 = 20.*

On peut facilement ce chiffre en considérant :

$$x = 20\ \%\ \text{et}\ x = y\ \%, \text{d'où}\ y = x/\% = 20\% / \% = 20.$$

Ainsi, pour un tout de 100, la partie serait de 20, soit 20 %.

Remarque : Bien que non rigoureuse d'un point de vue mathématique, on rencontre souvent la formule suivante : $\frac{50}{250}$ x 100 = 20 %

Cette formule a l'inconvénient de mélanger les deux méthodes présentées ci-dessus.

En reprenant le tableau **1**, *la part des actifs occupés dans le secteur A en 2018 est égale à :*

$$\frac{Nombre\ d'actifs\ occupés\ dans\ le\ secteur\ A\ en\ 2018}{Nombre\ total\ d'actifs\ occupés\ en\ 2018} = \frac{126\,960}{2\,227\,020} = 0{,}057 = 5{,}7\ \%.$$

Tableau ***2*** *: Population active occupée dans les trois secteurs d'activité (en %)*

Secteur d'activité / Années	*A*	*B*	*C*	*Σ*
2000	*26,5*	*35,0*	*38,5*	*100*
2018	*5,7*	*30,1*	*64,2*	*100*

On peut faire la comparaison de deux séries de pourcentage de répartition dans l'esspace (comparer la situation économique de deux entreprises situées dans deux villes différentes, à un moment donné), ou dans le temps (cas des années 2000 et 2018 de l'exemple ci-dessus).

La différence entre le pourcentage de répartition d'arrivée (2018 dans l'exemple ci-dessus), et le pourcentage de répartition de départ (2000) s'exprime, non pas en pourcentage, mais en points.

Exemple : Entre 2000 et 2018, la part de la population active occupée dans le secteur A a diminuée de 20,8 points, celle occupée dans le secteur B a diminuée de 4,9 points, et celle de C a augmentée de 25,7 points.

Puisque la somme est toujours égale à 100, la somme des diminutions est égale à la somme des augmentations.

En effet, on a bien 20,8 + 4,9 = 25,7.

Remarque : Il ne faut déduire de la baisse d'un pourcentage de répartition, entre deux dates, que la valeur absolue du sous-ensemble, à partir de laquelle il a été calculé, a forcément diminuée elle aussi (*cas du secteur B où les actifs passent de 664 040 à 671 160*).

2. 1. 3. 2. Pourcentage de variation (taux de croissance ou pourcentage d'évolution).

Les pourcentages de variation, appelés aussi taux de croissance, ou pourcentage d'évolution, permettent de mesurer la vitesse à laquelle varient les grandeurs dont on mesure la croissance entre deux dates.

- Calcul du pourcentage de variation. Il est égal à :

$$\frac{\text{Valeur finale}-\text{Valeur initiale}}{\text{Valeur initiale}} = \frac{Valeur\ finale}{Valeur\ initiale} - \frac{Valeur\ initiale}{Valeur\ initiale} = \frac{V_f}{V_i} - 1 = \Delta$$

Si $\Delta > 0$, il s'agit d'une augmentation.

Si $\Delta < 0$, il s'agit d'une diminution.

Si $\Delta = 0$, il s'agit d'une stagnation.

Le rapport $\frac{Valeur\ finale}{Valeur\ initiale}$ ou $\frac{Vf}{Vi}$ est appelé 'multiplicateur'.

Exemple : Le taux de croissance du nombre d'actifs occupés dans le secteur B, entre 2000 et 2018 est de : $\frac{671\,160}{664\,040} - 1 = 1{,}01 - 1 = 0{,}011$ *ou 1,1 %.*

- Calcul du taux de croissance successifs.

Pour déterminer le taux de croissance, lorsqu'une grandeur a subi deux, ou plusieurs, variations successives, il suffit de multiplier les multiplicateurs, et retrancher 1 au résultat final. Dans le cas de deux variations Δ_1 et Δ_2, on a :

$$\Delta' = (\Delta_1 \times \Delta_2) - 1.$$

Exemple : Si la vente du bien A a augmenter de 10 % au cours de la période 1, puis a été suivie de 15 % à la période 2, son taux de croissance est égal à :

$$\Delta' = (1{,}1 \times 1{,}15) - 1 = 26{,}5\ \% \quad \text{avec} \quad 1{,}1 = 1 + 0{,}10 \quad \text{et} \quad 1{,}15 = 1 + 0{,}15.$$

Evidemment, il serait faux de calculer $\Delta' = 15 - 5 = 10$.

Remarque : Lorsqu'une variable est égale au rapport de deux autres variables ($z = x/y$), son taux de croissance est égal au rapport des multiplicateurs des deux variables, diminué de 1.

Exemple : Si les salaires des ouvriers ont augmentés de 40%, à la suite de revendications syndicales, et le niveau général des prix a augmenté de 30%, le pouvoir d'achat a augmenté de : $\Delta = \left(\frac{1+0{,}4}{1+0{,}3}\right) - 1 = 6{,}92 \approx 7\ \%$.

- Dissymétrie des taux de croissance.

Les taux de croissances ne sont pas symétriques à la hausse, et à la baisse. La division par la valeur correspondante à la situation de départ, introduit dans la comparaison en valeur relative une dissymétrie que ne comporte pas la comparaison en valeur absolue : l'écart entre A et B est le même (au signe près) que celui entre B et A. Mais la variation relative entre A et B est différente de celle existante entre B et A, puisqu'on divise le même écart absolu par une utilisation de départ différente. L'effet d'une hausse est annulée (la grandeur revient à son point de départ) par une baisse de taux plus petite.

Le terme 'taux de croissance' peut induire en erreur puisqu'un taux de croissance peut très bien être négatif. Toutefois, la baisse ne peut aller au-delà de 100 % pour des grandeurs positives par nature (comme la plupart des grandeurs économiques). Une baisse de 100 % ramène la

valeur de la grandeur à zéro. Inversement, les hausses peuvent être illimitées. Cette confusion est très souvent faite, y compris dans les médias.

Exemple : Si une population agricole est passée de 900 à 300, il est faux de dire qu'elle a été réduite de 300 %. Elle a été réduite du tiers seulement, soit 33,33 %.

Si l'on s'intéresse à des pourcentages de plus en plus petits, l'écart entre les deux taux correspondants aux deux sens de comparaison (augmentation et diminution), est lui-même de plus en plus petit.

* Si A > B de moitié (50 %) $\Rightarrow$ B < A du tiers seulement (33,33 %).

* Si A > B du tiers (33,33 %) $\Rightarrow$ B < A de 1/4 seulement (25 %).

* Si A > B de 25 % $\Rightarrow$ B < A de 20 %.

* Si A > B de 10 % $\Rightarrow$ B < A de 9,1 %.

* Si A > B de 3 % $\Rightarrow$ B < A de 2,9 %.

* Si A > B de 1 % $\Rightarrow$ B < A de 0,99 %.

Il est donc acceptable pour les petits pourcentages de faire comme si les comparaisons par pourcentage étaient symétriques. Evidemment la définition du mot 'petit' est fonction de la précision recherchée, et ne serait pas la même en physique qu'en économie.

Il faut également éviter de confondre évolution et niveau. A partir de la seule connaissance des taux de variation, on ne peut pas en déduire le niveau auquel se situent les grandeurs étudiées, les unes par rapport aux autres. Le pourcentage de variation ne nous renseigne que sur la vitesse à laquelle augmentent les grandeurs étudiées. Cette confusion, souvent faite, entre le niveau et l'évolution, provoque parfois des effets surprenants (conférer exercices).

2. 2. Représentation graphique.

Les représentations graphiques résument les données statistiques pour une exploitation rapide, et donnent une vision synthétique et globale des observations. Elles sont nombreuses, et varient selon que la variable soit discrète ou continue, et que le caractère soit quantitatif,

ou qualitatif. Dans ce qui suit, nous nous limitons aux représentations graphiques du programme officiel[22].

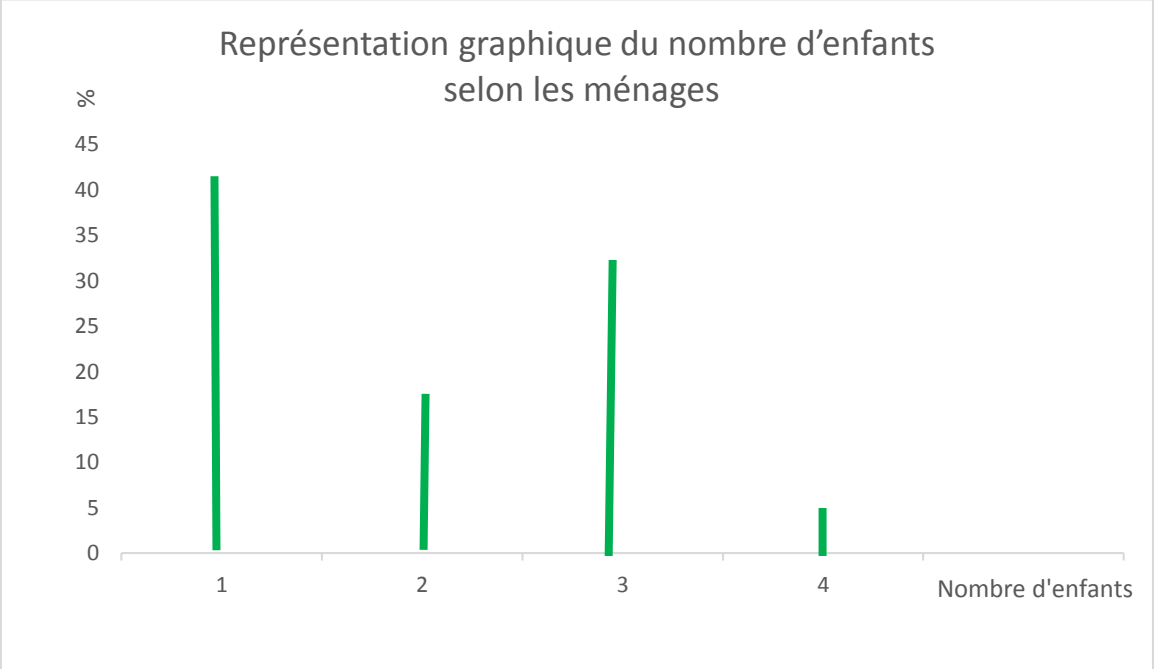
2. 2. 1. Cas de distribution quantitative.

L'abscisse représente la valeur observée, et l'ordonnée l'effectif, ou la fréquence.

2. 2. 1. 1. Diagramme en bâton.

Dans le cas d'une variable discrète, le graphique représentant la répartition est un diagramme en bâton. On porte en abscisse les valeurs du caractère et en ordonné l'effectif du caractère correspondant. L'effectif est représenté par un segment de droite. Ainsi, apparaît la discontinuité entre deux valeurs de la variable.

Pour le cas de la distribution des ménages selon le nombre d'enfants, on aura le diagramme en bâton qui suit.



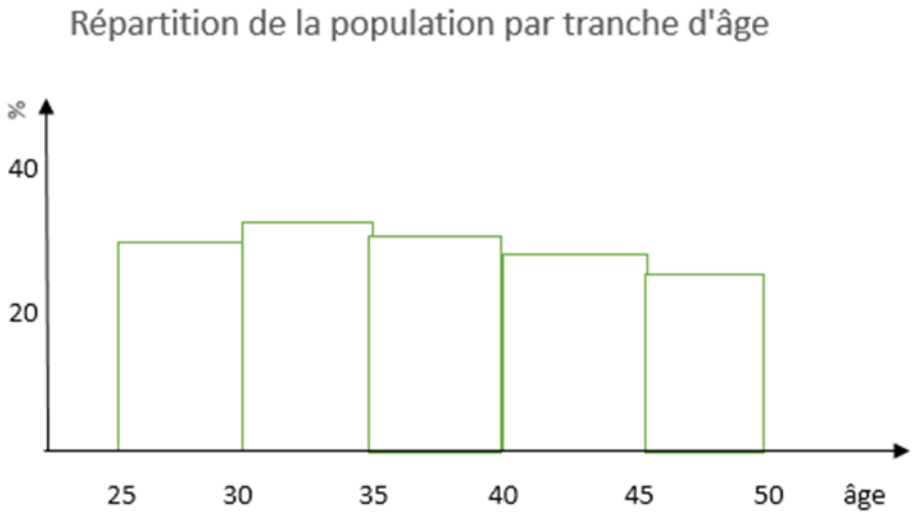
2. 2. 1. 2. Histogramme.

Dans le cas d'une variable continue, où les données sont regroupées en classes, la représentation la plus courante est l'histogramme, dans lequel chaque sous population relative à une modalité est représenté par un rectangle. On porte en abscisse les classes et on réalise des rectangles dont l'aire est proportionnelle à l'effectif de la classe. La représentation est donc effectuée comme si la distribution était uniforme à l'intérieur d'une classe. La surface limitée

[22] 3) Représentations graphiques : diagramme à bande, diagramme circulaire, diagramme en bâton. Polygone des effectifs (et des fréquences). Histogramme. Courbes cumulatives.

par l'histogramme doit être proportionnelle à l'effectif (ou à la fréquence). Il convient de prendre garde à l'amplitude des classes. On se ramène à la plus petite amplitude, appelée amplitude élémentaire, et on divise la hauteur du rectangle par la mesure de l'amplitude de la classe par rapport à cette amplitude élémentaire.

Exemple.



2. 2. 2. Cas de distribution qualitative.

Il n'est plus possible dans ce cas d'utiliser un diagramme cartésien, puisque les données ne sont pas numériques. Différentes méthodes sont possibles, dont nous présentons quelques-unes dans ce qui suit (diagrammes à bandes et diagrammes circulaires).

Dans le tableau qui suit, la variable statistique est qualitative (nature de la voiture). De plus, une comparaison dans le temps est souhaitée, ce qui permet de présenter quelques exemples des multiples possibilités offerte par les représentations graphiques.

Exemple. Production d'une entreprise de fabrication d'automobiles

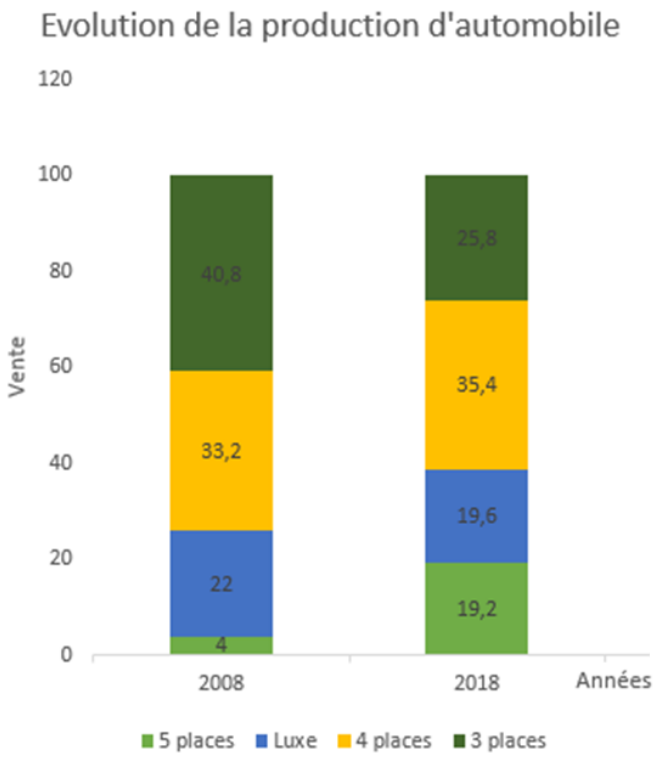
Tableau : Evolution de la production d'automobiles (unité 10^3)

Véhicule :	2008		2018	
	Effectif	Pourcentage	Effectif	Pourcentage
À 2 places	10,2	40,8	25,8	25,8
À 4 places	8,3	33,2	35,4	35,4
De luxe	5,5	22,0	19,6	19,6
À 5 places	1,0	4,0	19,2	19,2
	25		100	

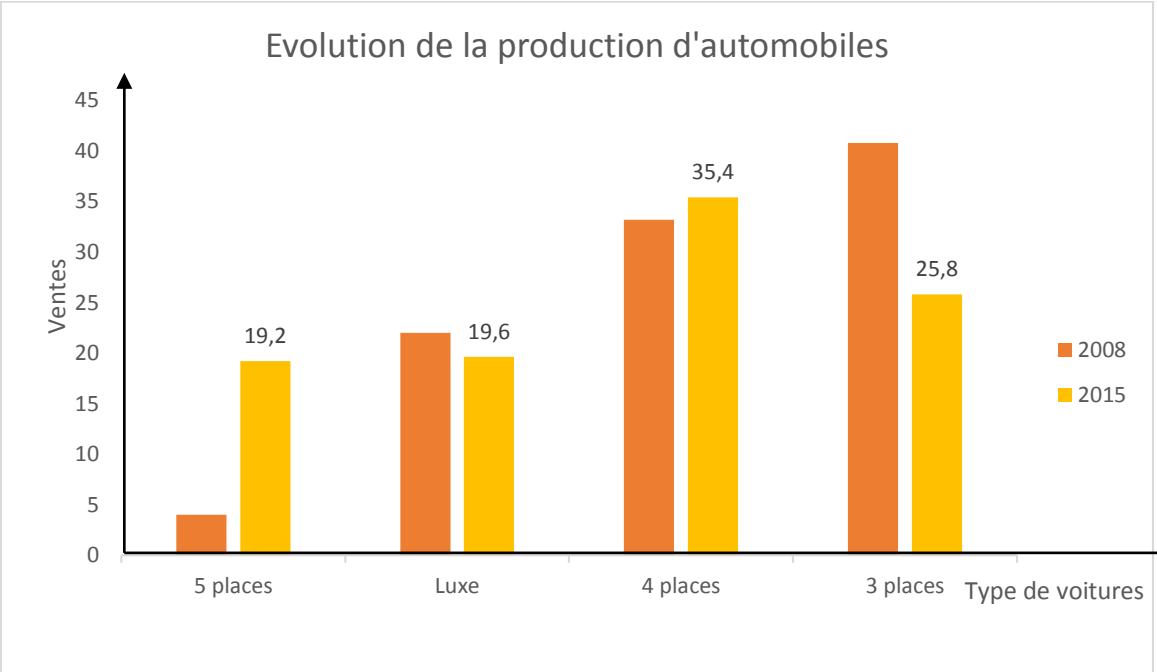
2. 2. 2. 1. Diagramme à bandes.

Deux types de diagrammes à bandes sont possibles.

Première représentation.

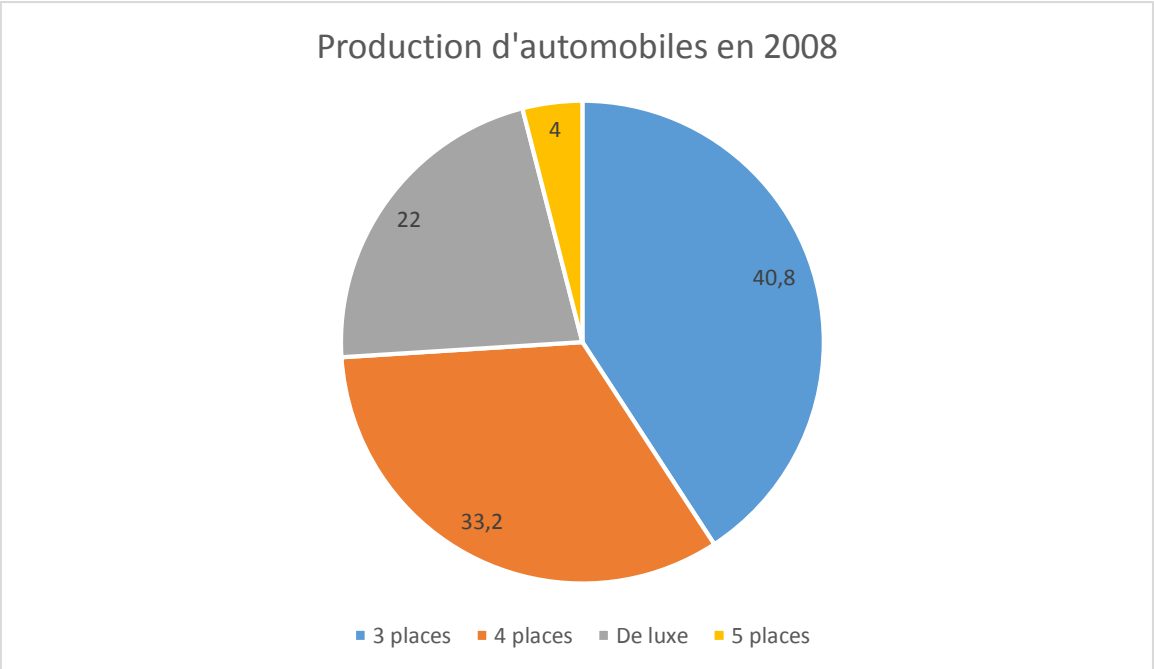


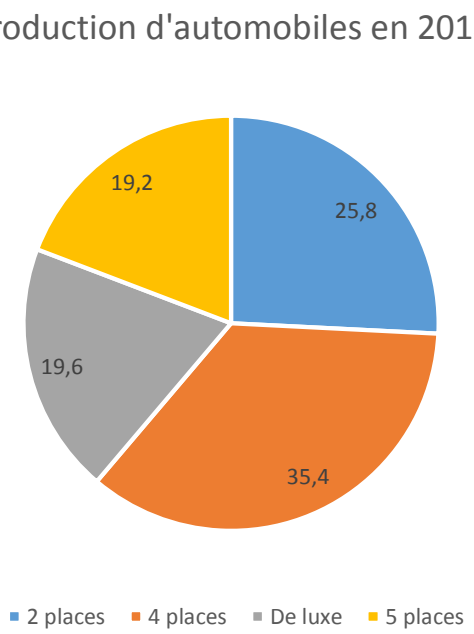
Deuxième représentation.



2. 2. 2. 2. Diagrammes circulaires

Il est également possible de réaliser des *diagrammes circulaires* (appelé aussi à secteur). Les effectifs, ou les fréquences, des diverses modalités, sont représenté par des secteurs d'un cercle (cas ci-dessous), ou d'un demi-cercle. Le cercle est découpé en secteurs dont l'aire est proportionnelle à l'effectif de la valeur considérée.





Les représentations au moyen de tableaux et de graphiques sont essentielles dans la mesure où elles constituent une mise en ordre nécessaire et une possibilité de se faire une idée globale sur le problème étudié. Cependant, elles ne suffisent pas si l'on désire approfondir l'analyse. Pour cela, on leur attribue des valeurs caractéristiques, appelées paramètres, qui ont pour but de résumer dans une certaine mesure les informations disponibles.

2. 3. Caractère de position

Trois catégories de paramètre sont habituellement considérées : les paramètres de position (objet de ce paragraphe), les paramètres de dispersion (objet du paragraphe 4), et les paramètres de forme (objet du paragraphe 5).

Nous abordons dans ce paragraphe, le cas de séries statistiques simples.

Les paramètres de position (appelés aussi paramètres de localisation) sont destinés à définir des valeurs centrales, ou caractéristiques, de la série étudiée.

Lorsqu'on est en présence d'un groupe individualisé de données, et qu'on veut connaître une valeur qui puisse en quelque sorte le caractériser, on prendra une mesure de localisation. Cela signifie que ce que l'on recherche est, au fond, un nombre exprimant une valeur moyenne de la suite de ces données. En terme statistique, ce genre de valeurs moyenne porte le nom de 'mesure de tendance centrale' relative au groupe donné. Nous présentons dans ce qui suit quatre mesures de tendance centrale : les moyennes, le mode, la médiane et les quantiles.

2. 3. 1. Les moyennes.

Si l'on désire représenter un ensemble de données par un seul nombre, comme les notes durant un semestre universitaire, la première mesure à laquelle on pense est la moyenne des notes. La moyenne est la mesure de tendance centrale la plus connue, et la plus utilisée pour résumer les nombreuses données d'une série statistique. Il existe plusieurs types de moyenne, dont chacune à des applications bien distinctes. Nous présentons dans ce qui suit la moyenne arithmétique, la moyenne harmonique et la moyenne géométrique (dans les trois cas nous présentons la valeur pour des valeurs simples, puis pour des valeurs pondérées).

2. 3. 1. 1. Moyenne arithmétique.

Nous présentons tout d'abords la moyenne arithmétique simple, ensuite la moyenne arithmétique pour des données groupés, puis la moyenne arithmétique pondérée.

- Moyenne arithmétique simple

Notion très connue, la moyenne arithmétique est définie comme le quotient de la somme des valeurs des données par le nombre de ces données. Ainsi, la moyenne arithmétique relative à une population entière sera représentée par le symbole μ (mu), tandis que la moyenne arithmétique relative à un échantillon sera représentée par $\bar{X}$ (X barre).

$$\mu = \frac{\sum X}{N} \qquad\qquad \bar{X} = \sum \frac{X}{n}$$

Exemple : Durant un certain mois de l'été passé, les huit vendeurs d'une entreprise de climatisation ont effectué les ventes suivantes d'appareils : 8, 11, 5, 14, 8, 11, 16 et 11. Si on considère qu'il s'agit d'une population entière de valeurs, on peut dire que le nombre moyen d'unités vendues est égal à :

$\mu = \frac{\sum X}{N} = \frac{84}{8} =$ *10,5 unités, soit environ 11 appareils par vendeur.*

- Moyenne arithmétique pour des données groupées

Lorsque des données ont été regroupées en une distribution d'effectifs, et qu'à chaque valeur X de la variable, on peut attribuer l'effectif partiel correspondant f, ou encore lorsque des données ont été distribuées en classe, et que chaque classe peut globalement être représentée par la valeur X_c du centre de cette classe, étant entendu que f représente l'effectif observé de la classe, on voit que les formules précédentes pourraient s'écrire comme suit.

Soit $\mu = \sum \frac{(f.X)}{N}$ ou $\mu = \sum \frac{(f.Xc)}{\sum f}$

Soit $\bar{X} = \sum \frac{(fX)}{n}$ ou $\bar{X} = \sum \frac{(fXc)}{\sum f}$

Tandis que, dans tous les cas, le dénominateur représente l'effectif total soit de la population, soit de l'échantillon (pour $\bar{X}$), on voit que le numérateur représente une sommation effectuée sur des produits. Chaque produit est celui d'un effectif par la valeur de la variable qui lui est associée du fait du regroupement préalable des données.

Exemple : Considérons les internautes de 15 à 25 ans (lycéens et étudiants), et leur temps d'utilisation d'internet par semaine. Après une enquête sur cette population, on obtient les résultats regroupés dans le tableau qui suit.

1. En moyenne, pendant combien d'heures par semaine, un internaute de 15 à 25 ans est-il connecté à l'Internet ?
2. Quel pourcentage d'internautes de 15 à 25 ans sont connectés à Internet durant 10 heures ou plus par semaine ?
Tableau : Temps d'utilisation d'internet par les lycéens et étudiants de 15 à 25 ans.

Nombre d'heures par semaine	*Pourcentage d'internautes*
Moins de 5	*23,4*
[5 - 10[	*27,1*
[10 – 20[	*25,4*
[20 – 30[	*12,1*
[30 et plus.	*12,0*

1. Pour calculer la moyenne, il faut tout d'abord définir les centres de classe, d'où le tableau qui suit. On attribue à la dernière classe ouverte, une amplitude de 10.

Nombre d'heures par semaine	*Centre de classe*	*Pourcentage d'internautes*
Moins de 5	*2,5*	*23,4*
[5 - 10[	*7,5*	*27,1*
[10 – 20[	*15*	*25,4*
[20 – 30[	*25*	*12,1*
[30 et plus.	*35*	*12,0*

A partir de ces données, nous pouvons calculer la moyenne arithmétique ;

(2,5 x 23,4 +7,5 x 27,1 + 15 x 25,4 + 25 x 12,1 + 35 x 12,0) / 100 =

= 1365,25 /100 = 13,6525 ≈13,5 h.

Ainsi donc, en moyenne, un internaute de 15 à 25 ans est connecté à Internet durant environ 13 h 30 durant une semaine.

2. Le pourcentage d'internautes de 15 à 25 ans connectés à Internet durant 10 heures ou plus par semaine est égal à :
25,4 + 12,1 + 12,0 = 49,5 %
Ainsi, près de la moitié des internautes de 15 à 25 ans sont connectés à Internet plus de 10 heures par semaine.

- Moyenne arithmétique pondérée.

Une moyenne arithmétique pondérée est très semblable à une moyenne arithmétique simple, mais ici, à chaque valeur de la variable est attribué un poids au lieu d'un effectif. Ce poids W est un nombre multiplicateur qui sert à quantifier l'importance accordée à la valeur ainsi désignée de la variable. On peut alors donner comme formule d'une moyenne pondérée, tant pour une population que pour un échantillon :

$$\mu_w \text{ ou } \bar{X}_m = \frac{\sum(WX)}{\sum W}$$

Naturellement, si toutes les valeurs possibles de la variable étaient affectées d'un facteur 'poids' égal à 1, on se retrouverait avec la moyenne arithmétique.

*Exemple : Dans u*ne usine fabricant plusieurs produits, les marges bénéficiaires relatives à chacun des types de produits fabriqués étaient, durant l'année écoulée, respectivement les suivantes :

Tableau : Marges bénéficiaires selon le type de produit.

Type de produit	A	B	C	D
Marge bénéficiaire (en %)	4,2	5,5	7,4	10,1

A moins que l'on ait réalisé un chiffre d'affaire à peu près égal pour chacun des types de produits fabriqué, le calcul de la moyenne arithmétique n'a aucune signification. Pour pouvoir calculer une moyenne pondérée, nous avons besoin d'informations supplémentaires concernant

le facteur ‘poids’ à chacune des marges. Dans ce cas, il s’agit du chiffre d’affaire pour les différents types de produit, donnés dans le tableau qui suit.

Tableau : Marges bénéficiaires et chiffres d’affaires des différents types de produits.

Types de produits	Marges bénéficiaires (X)	Chiffres d’affaires (W)	W.X
A	4,2 %	30 000	1 260
B	5,5 %	20 000	1 100
C	7,4 %	5 000	370
D	10,1 %	5 000	303
		$\sum$ = 58 000	$\sum$ = 3033

Unité : 1000 Unités Monétaires

La moyenne pondérée est égale à :

$$\mu_w = \frac{\sum(WX)}{\sum W} = \frac{3033}{58\,000} = 5{,}2\ \%$$

2. 3. 1. 2. Moyenne harmonique.

Bien que moins utilisée que la moyenne arithmétique, la moyenne harmonique a beaucoup d’applications, dont principalement :

- aux grandeurs inversement proportionnelles ;
- lorsque la somme des inverses à un sens important ;
- lorsque les x_i varient en sens inverse tels que le prix et la quantité
- lorsque aucune valeur de x_i ne doit être nulle ;
- dans des études relatives à des questions de vitesse ou d’allure de travail tels que les kilomètres par heures, kilomètres par litre ou le coût par kilomètre.

- Moyenne harmonique simple

La moyenne harmonique simple, généralement représentée par la lettre H, est calculée par la formule :

$$H = \frac{n}{1/X1 \ + \ 1/X2 \ + \ \ldots \ + \ 1/Xk}$$

- Moyenne harmonique pondérée

La moyenne harmonique pondérée, représentée aussi par la lettre H, est calculée par la formule :

$$H = \frac{n}{n1/X1 \ + \ n2/X2 \ + \ \ldots \ + \ nk/Xk}$$

2. 3. 1. 3. Moyenne géométrique.

Les principales applications de la moyenne géométrique sont :

- pour le calcul de moyenne de coefficient multiplicateur,
- les x_i doivent être strictement positives (les logarithmes n'ont pas de valeurs négatives),
- lorsqu'une population s'accroît suivant une progression géométrique,
- dans les moyennes de pourcentages tels que le taux d'intérêt, le taux de croissance, le taux d'accroissement ou l'indice des prix,
- lorsque les valeurs de x_i sont liées de façon multiplicative les unes par rapport aux autres.

- Moyenne géométrique simple

La moyenne géométrique simple, représentée par la lettre G, est calculée par la formule :

$$G^n = X_1 \, . \, X_2 \, . \, \ldots \, . \, X_k \qquad G = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$$

. Moyenne géométrique pondérée

La moyenne géométrique pondérée, représentée par la lettre G, est calculée par la formule :

$$G = \left[\prod_{i=1}^{h} x_i^{n_i}\right]^{\frac{1}{n}}$$

$G^n = 1/n \, . \, (n_1 \log x_1 + n_2 \log x_2 + \ldots + n_k \log x_k)$

Exemple : Supposons que les taux d'intérêt pour quatre années consécutives soient respectivement de 5, 10, 15, et 10 %. Que va-t-on obtenir après 4 ans si je place 100 U.M ?

- Après un an on a, 100 × 1,05 = 105 U.M.

- Après deux ans on a, 100 × 1,05 × 1,1 = 115,5 U.M.

- Après trois ans on a, 100 × 1,05 × 1,1 × 1,15 = 132,825 U.M.

- Après quatre ans on a, 100 × 1.05 × 1,1 × 1,15 × 1,1 = 146,1075 U.M.

Si on calcule la moyenne arithmétique des taux on obtient :

$$\bar{X} = \frac{1{,}05+1{,}10+1{,}15+1{,}10}{4} = 1{,}10$$

Si on calcule la moyenne géométrique des taux, on obtient :

$$G = (1.05 \times 1{,}10 \times 1{,}15 \times 1{,}10)^{1/4} = 1{,}099431377.$$

Le bon taux moyen est bien G et non pas $\bar{X}$, car si on applique quatre fois le taux moyen G aux 100 U.M, on obtient :

$$100 \times G^4 = 100 \times 1{,}099431377^4 = 146{,}1075 \text{ U.M.}$$

2. 3. 2. Le mode.

Le mode d'une série statistique est la valeur du caractère ayant le plus grand effectif. Le mode n'est pas toujours unique. Quand il existe plusieurs modes, la distribution statistique est dite multimodale. Lorsque la série à deux modes, elle est dite bimodale. Le calcul du mode dépend du type de données (variable discrète ou continue).

Pour une série statistique à caractère quantitatif discret, le mode est la valeur du caractère qui correspondant au plus grand effectif.

Si le caractère est une variable quantitative continue, on regroupe ses valeurs en classes. La classe qui a le plus grand effectif (effectif ramené à l'unité d'amplitude) est appelée classe modale**.** Si l'effectif n'est pas ramené à l'unité d'amplitude, il peut arriver que la classe modale ne soit pas celle où l'effectif apparait le plus élevé sur le tableau. Effectivement, cette dernière classe peut avoir une amplitude plus grande qu'une autre dont l'effectif par unité d'amplitude est plus élevé.

Pour calculer le mode d'une distribution de données groupées, on peut utiliser le milieu de l'amplitude de la classe modale. Le mode d'une distribution aux données groupées est calculé en tenant compte de l'effectif de la classe modale et des classes précédente et suivante. La formule qui suit donne une valeur théorique plus précise du mode.

$$\text{Mode} = L_{\text{Mode}} + \left(\frac{D_1}{D_1+D_2}\right) . \text{Am}$$

Où :

- L_{Mode} représente la limite inférieure de la classe modale,

- D_1 représente la différence entre l'effectif de la classe modale et l'effectif de la classe précédent,

- D_2 représente la différence entre l'effectif de la classe modale et l'effectif de la classe suivante,

- Am et l'amplitude de la classe modale.

L'utilisation du mode comme mesure de tendance centrale pour des variables quantitatives continues est assez rare. Pour des variables qualitatives, le mode est plus utile, parce que la moyenne et la médiane n'ont pas de sens.

2. 3. 3. La médiane.

Après avoir ordonné les valeurs d'une série statistique par ordre croissant (ou décroissant), c'est-à-dire de la plus petite à la plus grande valeur (ou le contraire), on peut définir la médiane. La médiane représente la valeur de l'élément central qui partage la série statistique en deux sous-séries de même nombre d'éléments. La première sous-série est composée des valeurs inférieures à la médiane, alors que la seconde sous-série est composée des valeurs supérieures à la médiane. Si le nombre d'observations qui composent la série est pair, la médiane est la moyenne des deux observations centrales. Dans le cas où nombre d'observations qui composent la série est impair, la médiane est la valeur centrale de la série ordonnée. Si les données sont regroupées, on parle alors de classe médiane. Dans ce cas la médiane ne peut être estimée graphiquement à partir de la lecture, sur le polygone de la courbe des effectifs cumulés, de l'abscisse du point ayant pour ordonnée N/2 (N étant le nombre total d'observations).

La médiane ne se calcule que pour des données quantitatives et son mode de calcul dépend du type de données. Si la médiane est sensible aux fluctuations d'échantillonnage, elle l'est moins aux valeurs aberrantes. Toutefois elle se prête difficilement à des calculs algébriques. Ainsi, la médiane donne une information plus significative que la moyenne dans les séries où les valeurs extrêmes sont importantes. C'est le cas des salaires en économie, les résultats d'un examen où les notes peuvent varier de 0 à 20.

2. 3. 4. Les quantiles.

Les quantiles sont des mesures de position qui ne tentent pas obligatoirement de déterminer le centre d'une distribution d'observations, mais de décrire une position particulière. Les quantiles

sont des valeurs qui partagent une série de données statistique en un certain nombre de parties égales. La notion de quantile est une extension du concept de la médiane qui divise une distribution d'observations en deux parties. Les quantiles les plus fréquemment utilisés sont les quartiles, les quintiles, les déciles et les centiles.

- Les *quartiles* (Q_1, Q_2 et Q_3) divisent un ensemble d'observations en quatre parties égales, comprenant chacune 25% des données.
- Les *quintiles* (V_1, V_2, V_3 et V_4) divisent un ensemble d'observations en cinq parties égales, comprenant chacune 20% des données.
- Les *déciles* (D_1, D_2, …, D_8 et D_9) divisent un ensemble d'observations en dix parties égales, comprenant chacune 10% des données.
- Les *centiles* (C_1, C_2, …, C_{98} et C_{99}) divisent un ensemble d'observations en cent parties égales, comprenant chacune 1% des données.

2. 4. Caractéristiques de dispersion

Pour présenter les caractéristiques de dispersion, considérons tout d'abord les deux séries de données suivantes :

95	97	100	103	105
50	75	100	125	150

On remarque que pour ces deux séries les moyennes et les médianes sont identiques (100). Cependant, comme on peut le constater elles diffèrent profondément dans leur dispersion. Ainsi, l'exploitation, seule, des caractéristiques de tendance centrale ne suffit pas à décrire convenablement les observations. Nous devrons alors nous intéresser à la manière avec laquelle sont réparties les observations autour de la tendance centrale. C'est ce que l'on désigne par la dispersion. Les caractéristiques de dispersion (appelées aussi paramètres de dispersion ou mesures de dispersion) ont pour objectif de caractériser la répartition des observations les unes par rapport aux autres ou autour d'une valeur centrale. Un paramètre de dispersion permet de décrire un ensemble de données concernant une variable particulière, en fournissant une indication sur la variabilité des valeurs au sein de l'ensemble des données. Ainsi, les caractéristiques de dispersion complète la description fournie par les paramètres de tendance centrale d'une distribution.

Si nous observons différentes distribution, nous pouvons constater que, pour certaines, toutes les données sont groupées à une distance plus ou moins faible de la valeur centrale, alors que pour d'autres, la répartition des données est nettement plus grande.

Les caractéristiques de dispersion peuvent être classées en deux groupes. Le premier, par les mesures définies par la distance entre deux valeurs représentatives de la distribution (l'étendue ou intervalle de variation, et l'intervalle interquartile). Le second groupe, par les mesures calculées en fonction des déviations de chaque donnée par rapport à une valeur centrale (l'écart géométrique, l'écart médian, l'écart moyen, la variance, l'écart type et le coefficient de variation). Parmi les mesures de dispersion, l'écart type est la plus utilisée. Nous présentons dans ce qui suit le calcul de ces différentes caractéristiques de dispersion selon les deux groupes.

2. 4. 1. L'étendue.

- Dispersion définie par la distance entre deux valeurs représentatives de la distribution.

2. 4. 1. 1. L'étendue simple.

L'étendue est la caractéristique de dispersion la plus simple. Elle est égale à la différence entre la plus grande valeur et la plus petite valeur.

$$\text{Etendue} = \text{Valeur la plus grande} - \text{Valeur la plus petite}$$

Bien que l'étendue soit le paramètre de dispersion le plus simple à calculer, elle est rarement utilisée seule car elle est basée uniquement sur deux observations, et donc très influencée par les valeurs extrêmes. A cela s'ajoute qu'elle ne nous renseigne pas sur les autres valeurs, d'où son inconvénient. En fait, l'étendue sert très peu comme mesure de dispersion. On l'utilise surtout en contrôle de la qualité, où les échantillons sont souvent de petite taille (quatre ou cinq éléments). Dans ce cas, l'étendue donne une idée assez précise de la dispersion des résultats de l'échantillon.

2. 4. 1. 2. Etendue interquartile (ou intervalle interquartile).

L'intervalle interquartile est une caractéristique de dispersion qui est égale à la différence entre le premier et le troisième quartile. Ainsi, l'étendue interquartile correspond à l'intervalle comprenant 50% des observations les plus au centre de la distribution.

Soient Q_1 et Q_3 les premier et troisième quartiles d'une distribution. L'intervalle interquartile s'écrit donc :

$$\text{Etendue interquartile} = Q_3 - Q_1$$

L'intervalle interquartile est une caractéristique de variabilité qui ne dépend pas du nombre d'observation, et contrairement à l'étendue, il n'est pas dépendant des valeurs extrêmes.

Exemple. Soit la série statistique qui suit composée de 10 observations

0 1 1 2 2 2 3 3 4 5

Le premier quartile Q_1 est égal à 1, et le troisième quartile Q_3 est égal à 3.

L'intervalle interquartile est donc égal à : $Q_3 - Q_1 = 3 - 1 = 2$

- Dispersion calculée en fonction des déviations de chaque donnée par rapport à une valeur centrale.

2. 4. 1. 3. Ecart géométrique.

L'écart géométrique d'un ensemble d'observations quantitatives est une caractéristique de dispersion qui correspond à l'écart des observations autour de la moyenne géométrique.

Soient $x_1, x_2, \ldots, x_n$ représentant un ensemble de n observations relatives à une variable quantitative X. L'écart géométrique, notée $E_{géom}$ se calcule comme suit :

$$\text{Log } E_{géom} = \frac{1}{n} \sum_{i=1}^{n} (\log xi - \log G)$$

où $G = \sqrt[n]{x1 \,.\, x2 \,.\, \ldots \,.\, xn}$ est la moyenne géométrique des observations.

2. 4. 1. 4. Ecart médian.

L'écart médian d'un ensemble d'observations quantitatives est une caractéristique de dispersion qui correspond à la moyenne des valeurs absolues des écarts de chaque observation par rapport à la médiane.

Soient $x_1, x_2, \ldots, x_n$ représentant un ensemble de n observations relatives à une variable quantitative X. L'écart médian, notée $E_{méd}$ se calcule comme suit :

$$E_{méd} = (\sum_{i=1}^{n} |xi - Md|) / n$$

où Md est la médiane des observations.

2. 4. 1. 5. Ecart moyen.

L'écart moyen d'un ensemble d'observations quantitatives est une caractéristique de dispersion qui correspond à la moyenne des valeurs absolues des écarts de chaque observation par rapport à la moyenne arithmétique.

Soient x_1, x_2, …, x_n représentant un ensemble de n observations relatives à une variable quantitative X. L'écart moyen, notée EM se calcule comme suit :

$$EM = (\sum_{i=1}^{n} |Xi - \bar{X}|) / n$$

Où $\bar{X}$ est la moyenne arithmétique des observations.

Exemple. Considérons cinq étudiants qui ont passé successivement deux examens auxquels ils ont obtenu les notes qui suivent.

Examen 1 : 3,5 4,0 4,5 3,5 4,5. D'où une moyenne $\bar{X}$ = 20/5 = 4.

Examen 2 : 2,5 5,5 3,5 4,5 4,0. D'où une moyenne $\bar{X}$ = 20/5 = 4.

La moyenne arithmétique $\bar{X}$ des notes est identique pour les deux examens. Cependant la dispersion des notes n'est pas identique.

Pour l'examen 1, l'écart moyen est égal à :

$$EM = \frac{|3,5-4|}{5} + \frac{|4-4|}{5} + \frac{|4,5-4|}{5} + \frac{|3,5-4|}{5} + \frac{|4,5-4|}{5} = 2/5 = 0,4.$$

Pour l'examen 2, l'écart moyen est égal à :

$$EM = \frac{|2,5-4|}{5} + \frac{|5,5-4|}{5} + \frac{|3,5-4|}{5} + \frac{|4,5-4|}{5} + \frac{|4-4|}{5} = 4/5 = 0,8.$$

Les notes du deuxième examen sont donc plus dispersées autour de la moyenne arithmétique que les notes du premier examen.

2. 4. 2. Variance.

Si les données sont issues d'une population, la La variance est un paramètre de dispersion qui utilise toutes les observations. La variance est basée sur la différence entre la valeur de chaque observation (x_i) et la moyenne ($\bar{X}$ pour un échantillon, μ pour la population). La différence entre chaque observation x_i et la moyenne est appelée écart par rapport à la moyenne. Pour un échantillon, un écart par rapport à la moyenne s'écrit (x_i - $\bar{x}$). Pour une population, il s'écrit (x_i - μ). Pour calculer la variance, les écarts par rapport à la moyenne sont élevés au carré.

moyenne des écarts au carré est appelée variance de la population. La variance de la population est notée par le symbole grec σ^2. Dans le cadre d'une population comprenant N observations, de moyenne μ, la variance est définie par l'expression :

$$\sigma^2 = \frac{\Sigma\,(x_i - \mu)2}{N}$$

Dans la plupart des études statistiques, les données à analyser sont issues d'un échantillon. Le calcul de la variance d'un échantillon nous permet, généralement ensuite, d'estimer la variance de la population σ^2. La variance de l'échantillon[23], notée S^2 est définie comme suit :

$$S^2 = \frac{\Sigma\,(x_i - \bar{x})2}{n-1}$$

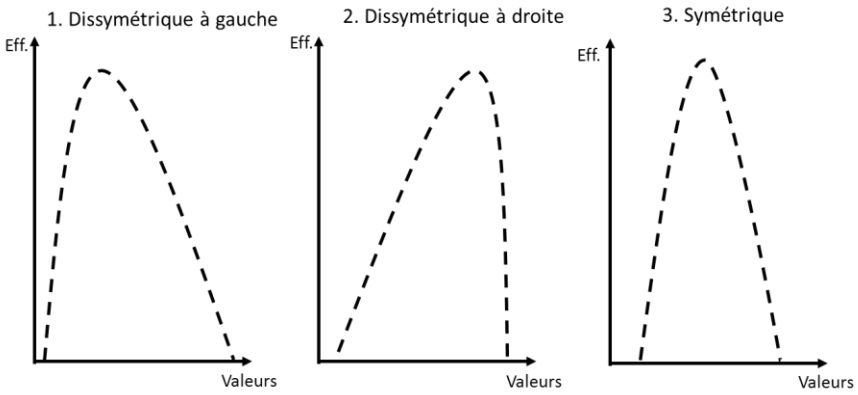
La variance d'échantillon S^2 est l'estimateur de la variance de la population σ^2.

2. 5. Caractéristiques de forme

Les caractéristiques définies dans les paragraphes précédents dépendent des unités de mesures de la variable étudiée, ce qui ne permet pas de comparer la dispersion de séries statistiques hétérogènes. Afin de pallier à cette limite, on utilise des caractéristiques de formes qui nous permettent de comparer la symétrie (ou asymétrie) de plusieurs distributions statistiques et dont on mesure l'intensité par des coefficients qui sont indépendants des unités de mesure.
On décompose les caractéristiques (ou paramètres) de forme en deux catégories : les paramètres de symétrie et les paramètres d'aplatissement.

2. 5. 1. Caractéristiques de symétrie.

On distingué généralement trois types de distribution, selon qu'elles sont dissymétrique à gauche (schéma 1 ci-dessous), dissymétrique à droite (schéma 2 ci-dessous) ou symétriques (schéma 3 ci-dessous)



[23] Bien qu'une explication détaillée ne soit pas l'objet de ce paragraphe, on peut souligner que si la somme des écarts par rapport à la moyenne au carré est divisée par n – 1, et non pas par n, la variance de l'échantillon fournira un estimateur sans biais de la variance de la population. D'où le (n – 1) du dénominateur dans la variance de l'échantillon.

Dans une distribution symétrique, la moyenne arithmétique, la médiane et le mode se confondent, et les fractiles d'ordre σ et $(1 - \sigma)$ sont équidistants de la médiane, ce qui est le cas pour les deux quartiles Q_1 et Q_3.

Une première mesure de dissymétrie provient de la différence entre $(Q_3 - Q_2)$ et $(Q_2 - Q_1)$. Pour obtenir un coefficient de dissymétrie qui soit indépendant de l'unité de mesure, on calcule :

$$C = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

C varie entre – 1 et + 1, et il est égal à zéro pour une distribution symétrique.

Le coefficient C ne dépend pas de toutes les valeurs observées, et pour palier à cet inconvénient, on peut utiliser d'autres coefficients.

Un paramètre pour caractériser l'asymétrie d'une distribution $\{(X_j, n_j) ; j = 1, 2, \ldots, J\}$ est le moment d'ordre 3 défini par la formule suivante :

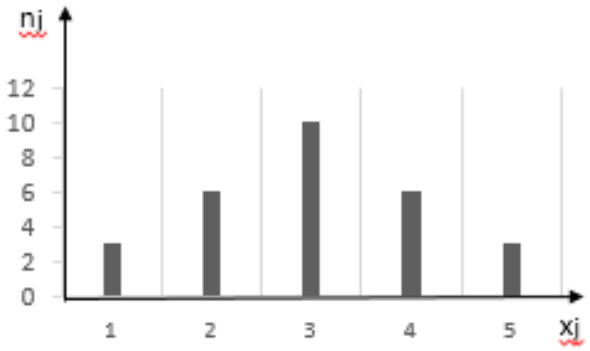
$$m_3 = \frac{1}{n} \sum_{j=1}^{J} n_j \ (X - \vec{x})^3$$

On obtient trois cas pour une distribution :

- Dissymétrie à gauche si $m_3 > 0$,
- Symétrie si $m_3 = 0$,
- Dissymétrie à droite si $m_3 < 0$.

La seconde proposition (symétrie) est évidente. E, effet, quand la distribution est symétrique, à chaque différence ($X - \vec{X}$) correspond une autre différence de même valeur absolue mais de signe opposée, associées toutes deux à un même effectif. Comme l'élévation à la puissance 3 conserve le signe des différences, m_3 est nul.

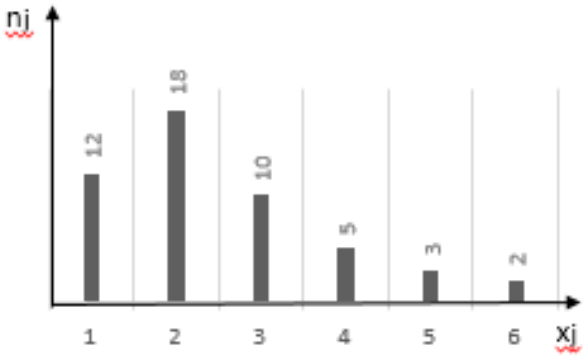
Exemples. 1. Soient les données du graphe et le calcul du moment centré d'ordre 3..



M m3 = 1/28 [$3(1-3)^3 + 6(2-3)^3 + 10(3-3)^3 + 6(4-3)^3 + 3(5-3)^3$] =

= 1/28 [$-24 - 6 + 0 + 6 + 24$] = 0.

2. Considérons dans ce deuxième exemple une distribution dissymétrique à gauche.



Les calculs sont donnés dans le tableau qui suit.

X_j	n_j	$n_j.X_j$	$(X_j - \bar{X})$	$n_j (X_j - \bar{X})$	$(X_j - \bar{X})^3$	$n_j. (X_j - \bar{X})^3$
1	12	12	- 1,5	- 18	- 3,375	- 40,500
2	18	36	- 0,5	- 9	- 0,125	- 2,250
3	10	30	0,5	5	0,125	1,250
4	5	20	1,5	7,5	3,375	16,875
5	3	15	2,5	7,5	15,625	46,875
6	2	12	3,5	7	42,875	85,750
		125		0		108,000

Ainsi, on constate que la moyenne est égale à :

$$\bar{X} = \frac{125}{50} = 2{,}5$$

En considérant les colonnes des écarts ($X_j - \bar{X}$) et des produits n_j ($X_j - \bar{X}$), on constate qu'il y a des écarts positifs plus importants (en valeur absolue) que les écarts négatifs (0,5 ; 1,5 ; 2,5 ; 3,5) contre (- 1,5 ; 0,5). Mais, comme la somme (pondérée par les effectifs) des écarts positifs est compensée par la somme des écarts négatifs, cela signifie que ces derniers sont plus nombreux. Or, en prenant la troisième puissance des écarts, on accroit l'importance des valeurs les plus grandes (en valeur absolue), c'est-à-dire des écarts positifs. Ceci explique pourquoi m_3 est positif. Dans cet exemple, on a :

$$m_3 = \frac{108}{50} = 2{,}16$$

En prenant le cas d'une asymétrie à droite, on obtient, par le même raisonnement, $m_3 < 0$.

Dans cet exemple, m3 est indépendant d'un changement d'origine puisqu'il est basé sur des écarts. Cependant, il dépend des unités choisies, comme in peut le constater. C'est pour cette raison Fisher a introduit le coefficient suivant, et qui porte son nom :

$$F = m_3 / s^3$$

La division de m_3 par le cube de l'écart type (qui est positif) implique que F a le même signe que m_3. De plus, F est nul lorsque le moment m3 est nul (la vérité de la Palice évidemment). Mais cette division permet à F d'être à la fois indépendant d'un changement d'origine et d'unité (m3 a été divisé par une autre expression du troisième degré, donc le rapport est sans dimension).

Il existe d'autres coefficients d'asymétrie, et, notamment les suivant, dont les propriétés résultent de constations empiriques.

- Coefficient de dissymétrie de Pearson qui est égal à :

$$C_P = \frac{\bar{X} - M_0}{S}$$

Où M_0 représente le mode et S l'écart type.

- Coefficient de dissymétrie de Yule et Kendall qui est égal à :

$$Y_K = \frac{X_{1/4} + X_{3/4} - 2X_{1/2}}{X_{3/4} - X_{1/4}}$$

On peut constater que ce coefficient possède des propriétés semblables à F, et qu'il est compris entre – 1 et + 1.

Ces coefficients doivent cependant être interprétés avec prudence dans la mesure où les constatations empiriques qui les ont engendrées ne concernent pas nécessairement toutes les distributions. Par exemple, une distribution discrète presque symétrique peut fournir des coefficients C_p et Y_K de signe contraire. Ils ne peuvent donc être considérés que comme des outils d'appréciation, simples à obtenir, mais pouvant parfois être contradictoires

A préciser que la détermination d'un paramètre d'asymétrie n'est pas toujours nécessaire. Une analyse graphique peut parfois être suffisante pour donner la forme des données

2. 5. 2. Caractéristiques d'aplatissement.

En plus des caractéristiques de symétrie, on peut ajouter les paramètres d'aplatissement qui caractérisent l'aplatissement d'une distribution.

On calcule ces paramètres à travers le moment d'ordre 4 qui est égal à :

$$m_4 = 1/n \sum_{j=1}^{j} n_j \, (X_j - \bar{X})^4$$

On obtient à partir de ce moment deux coefficient d'aplatissement ;

- Coefficient d'aplatissement de Pearson :

$$P_a = \frac{m_4}{S^4}$$

- Coefficient d'aplatissement de Fisher :

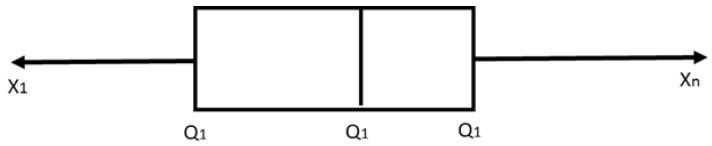
$$F_a = \frac{m_4}{S^4} - 3$$

Le coefficient de Fisher est légèrement plus simple à manipuler. Plus la série est effilée, plus ces coefficients sont grands et plus la série est aplatie, plus ces coefficient sont petits. Leur utilisation est cependant parfois délicate.

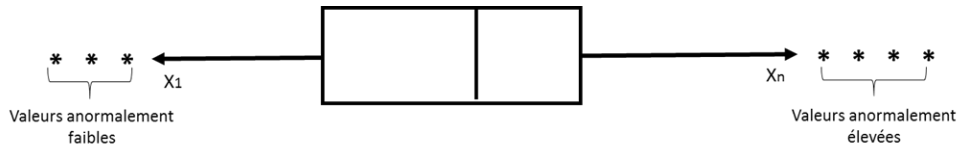
2. 6 - Représentation des résultats à l'aide du box plot.

La représentation graphique à l'aide du box plot, appelée aussi boîte (ou diagramme) à moustaches (ou à pattes) a été créée en 1977 par J.W. Turkey, et constitue un résumé à la fois visuel et numérique d'une série statistique.

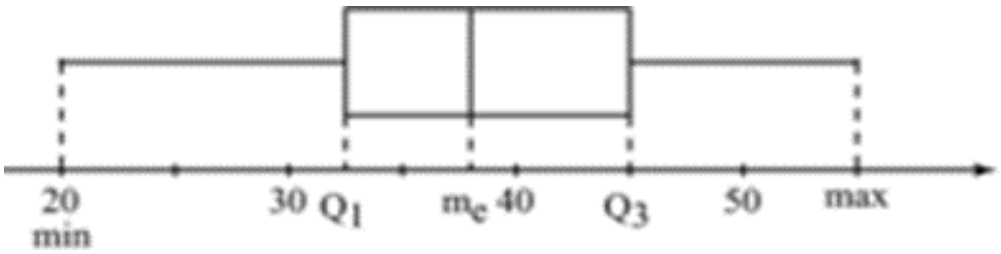
Cette représentation graphique est un diagramme qui fait apparaître la médiane, les quartiles et les valeurs extrêmes du caractère. La boîte est délimitée par les premier et troisième quartiles et partagée par la médiane, sa "longueur" est donc égale à l'intervalle interquartile. Les moustaches sont les segments reliant les quartiles aux valeurs extrêmes de la série. Souvent lorsqu'elles sont connues, on indique les valeurs extrêmes de la série. Ainsi, il s'agit d'un diagramme permettant de positionner les quartiles Q_1 , Q_2 , Q_3 , au moyen de rectangles de largeur arbitraire, prolongés par des "moustaches" de part et d'autre, de longueur au plus égale à une fois et demie $Q_3 - Q_1$.



Si la plus petite ou la plus grande valeur observée se trouvent à l'intérieur, on raccourcit les moustaches correspondantes ; si elles se trouvent à l'extérieur, on positionne à part les valeurs 'aberrantes' qui dépassent des moustaches.



Exemple. Soit une variable statistique *X* dont le maximum est 55, le minimum 20, la médiane 38, le premier quartile 32,5 et le troisième quartile 45, on construit alors le diagramme en boîte suivant :



3 - Séries statistiques à deux variables (bivariée)

Dans le chapitre précédant, nous avons présenté des distributions statistiques qui étudiaient une population selon un seul caractère. Cependant, dans la pratique, il est souvent utile de considérer dans une même population plusieurs caractères. Ces caractères peuvent correspondre à deux ou plusieurs aspects d'une même unité statistique, c'est le cas par exemple de la taille, le volume et la densité d'un objet d'étude ou bien le salaire, la qualification et l'âge d'un ensemble de salariés. Les caractères peuvent aussi concerner des phénomènes distincts, mais plus ou moins liés, tels la production industrielle et les importations ou la consommation et les revenus. Dans certains cas, l'un des caractères peut n'avoir qu'une signification de repère, tel le temps par exemple

Dans ce qui suit nous nous limitons à présenter les séries statistiques à deux variables seulement.

3. 1. Tableaux de données à double entrée. Nuage de points.

Nous présentons dans ce paragraphe des observations sous forme de tableau statistique. Lorsque les données sont regroupées dans un graphe, il s'agit alors d'une représentation de nuage de points.

3. 1. 1. Tableau de contingence.

Lors de l'étude de deux caractères (X et Y dans notre cas), les données sont regroupées dans un tableau à double entrée, appelé tableau de contingence.

Notation des tableaux carrés (tableau de contingence ou tableau à double entrée).

Y \ X	X_1	X_2	…	X_j	…	X_m	Total
Y_1	A_{11}	A_{12}	…	A_{ij}	…	A_{1m}	$A_{1.}$
Y_2	A_{21}	A_{22}	…	A_{2j}	…	A_{2m}	$A_{2.}$
.	.	.	…	.		.	
.	.	.	…	.		.	
.	.	.	…	.		.	
Y_i	A_{i1}	A_{i2}	…	A_{ij}	…	A_{im}	$A_{i.}$
.	.	.	…	.		.	.
.	.	.	…	.		.	.
.	.	.	…	.		.	
Y_n	A_{n1}	A_{n2}	…	A_{nj}	…	A_{nm}	$A_{n.}$
Total	$A_{.1}$	$A_{.2}$	…	$A_{.j}$	…	$A_{.m}$	

A_{ij} représentent le nombre d'individus, ou unités statistiques qui présentent à la fois la modalité X_i et la modalité Y_j.

La dernière ligne de la dernière colonne du tableau représente les distributions marginales (dans la marge), c'est-à-dire la distribution de X, sans tenir compte du caractère Y, ou celle de Y, sans tenir compte de X.

Il est courant qu'un tableau sortant d'un ordinateur ait des dizaines, voire des centaines, de lignes ou colonnes, représentant un nombre imposant de feuillets. Ces tableaux sont utilisés pour l'analyse statistique détaillée. Cependant, il convient de veiller à ne publier que des tableaux faciles à lire. Leur clarté, est en générale plus grande, si les résultats sont suffisamment groupés. Un dizaine de lignes et de colonnes constituent alors un nombre à ne pas dépasser, autant que possible. Des tableaux de cinq ou six lignes, et deux ou trois colonnes sont d'ailleurs souvent plus parlant. Lorsqu'on a un grand nombre de données, on choisit des classes pas trop nombreuses, pour que le tableau soit clair, mais suffisamment pour qu'il n'y ait pas de pertes d'informations. Il importe que les classes recouvrent tous les résultats, et aient une intersection vide, d'où les formules du type « de … à moins de … ». La différence entre les deux extrémités est appelée amplitude de la classe.

L'effectif d'une classe statistique est le nombre d'éléments de la population observé dans cette classe. La fréquence de la même classe est le rapport de cet effectif total de la population.**3. 1. 2. Nuage de points.**

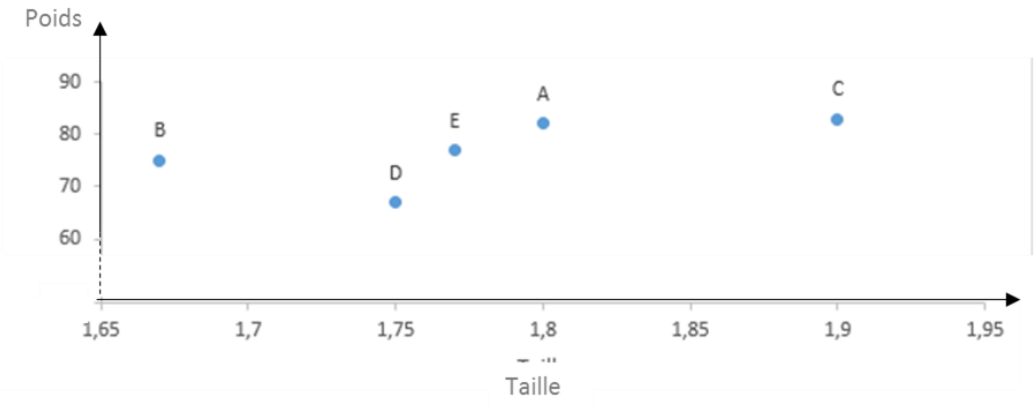
Nous présentons la représentation graphique sous forme de nuage de points à travers l'exemple qui suit et qui concerne deux variables quantitatives.

Exemple. Soient la taille et le poids de cinq étudiants. Unités : mètres pour X et kg pour Y.

Etudiants	A	B	C	D	E
Tailles (x_i)	1,80	1,67	1,90	1,75	1,77
Poids (y_i)	82	75	83	67	77

Lorsque les deux variables sont quantitatives, comme dans cet exemple, chaque individu (étudiant) i est symbolisé par un point défini dans un système d'axes orthogonaux par les coordonnées x_i et y_i.

La représentation graphique de la série statistique considérée ci-dessus est la suivante.



L'ensemble des éléments représentés constituent un nuage de points. L'impression que l'on peut retirer de la vision d'un tel graphique dépend surtout des unités choisies le long de chaque axe.

Cet exemple se caractérise par le faible nombre de valeurs recueillies. Dans le cas où la taille n de l'échantillon est plus élevée, chaque couple d'observations peut apparaitre à plusieurs reprises. Cette situation se présente surtout quand la variable est discrète (ou continue mais donnant lieu à des arrondis importants), et que le nombre de valeurs distinctes de chaque variables est faible. On peut alors effectuer une opération où l'on associe à chaque couple d'observations un effectif représentant le nombre de fois qu'il est apparu. Si les valeurs distinctes de x et y sont notées $X_1, X_2, \ldots X_j$ d'une part, et $Y_1, Y_2, \ldots Y_k$ d'autre part, un couple d'observations peut être représenté par (X_j, Y_k) avec $j \in J$ et $k \in K$, où nous avons posé :

$$J = \{ 1, 2, \ldots, J\} \quad , K = \{ 1, 2, \ldots, K].$$

Nous désignons par n_{jk} l'effectif correspondant à ce couple. La série statistique bivariée permet alors de définir une distribution observée à deux dimensions par l'ensemble des triplets :

$$\{ (X_j, Y_k, n_{jk}) \ , \ j \in J \quad , \quad k \in K \}.$$

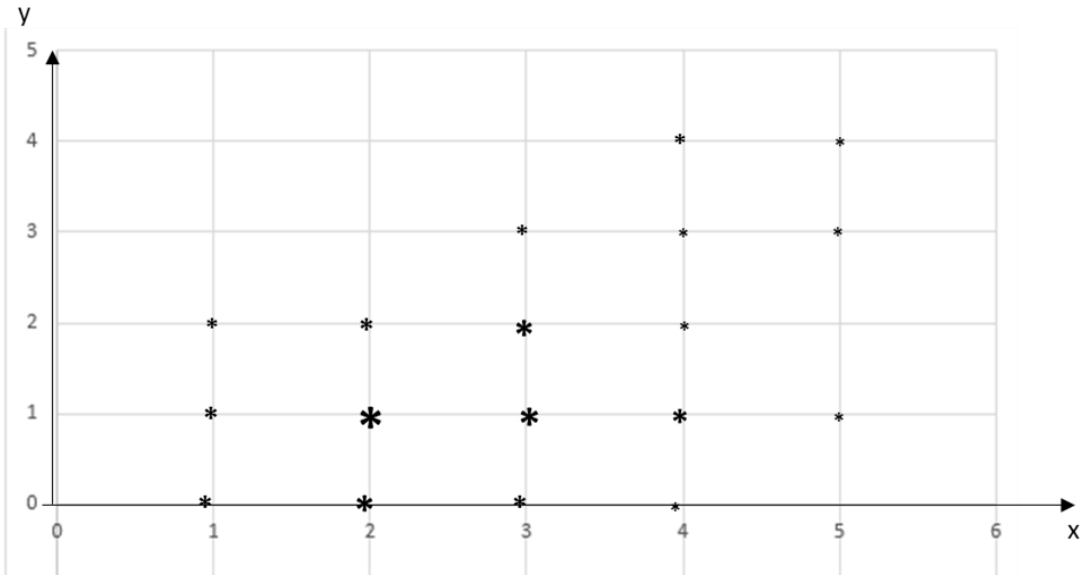
Cette distribution donne lieu à un tableau de contingence. La représentation graphique devient plus difficile à construire. On peut en effet associer un 'point' de coordonnées (X_j, Y_k) au couple défini par la j-ème valeur de x et la k-ième valeur de y, en le dotant d'une surface égale à n_{jk}.

Exemple. Soit un échantillon de 80 adultes, ayant au moins un enfant, et qui a été examiné sous deux aspects. D'une part le nombre d'enfant (variable x) et d'autre part, le nombre de frères et sœurs (variable y) qu'ils ont eus.

Le tableau de contingence résultant de l'enquête est le suivant.

X_j \ Y_k	**0**	**1**	**2**	**3**	**4**
1	4	4	2	0	0
2	9	16	4	0	0
3	4	12	9	2	0
4	1	6	1	1	2
5	0	1	0	1	1

On peut lui associer le graphique qui suit.



Cette construction est encore valable quand les variables sont ordinales. Par ailleurs, on peut aussi recourir à des groupements en classes pour l'une ou l'autre variable si cela s'avère nécessaire. Dans ce cas, plusieurs représentations graphiques sont possibles. On peut en effet construire :

- Un nuage de pooints où ces derniers sont définis à partir des centres de classe . Cependant, on ne fait pas apparaitre explicitement lle caractère continu des classes.

- Un diagramme à trois dimensions, appelé stéréogramme, et composé de parallèlipipèdes rectangles dont la base définie par les couples de classes et le volume par l'effecti (ou la fréquence) correspondant. Cette représentation ne vaut que si les deux variables donnent lieu à un regroupement en classes. En outre, elle se heurte à la difficulté d'interpréter un diagramme à trois dimensions.

- Un découpage du plan à partir des classes, en affectant aux rectangles ainso constitués une couleur, ou un ombrage, d'autant plus sombre que l'effectif (la fréquence) est grand. On peut aussii symboliser les observations par des points. L'effectif n_{jk} est alors défini par le nombre de points par rectangle. Cette façon de découper un plan en régions et de représenter des densités (fréquences) par des ombrages d'intensités différentes est fort répandu, particulièrement dans le domaine de la cartographie.

Il est évident que ce type de réprésentation donne une impression générale. Seule l'analyse du tableau des données permet une étude plus approfondis.

Il faut aussi préciser que si les variables sont nominales, la construction d'un graphique assicié à un tableau de contingence devient plus critiquable, en raison du caractère arbirtraire de la disposition des valeurs les unes par rapport aux autres.

3. 2. Distributions marginales et conditionnelles. Covariance.

L'étude d'une série à deux caractères (x_i, y_i) avec i variant de 1 à n, comporte en particulier l'analyse des séries marginales univariées obtenue en ne considérant qu'une variable à la fois dans le tableau individuel à deux variables.

- Série marginale en x : { xi ; i = 1, 2, …, n }.

- Série marginale en y : { yi ; i = 1, 2, …, n }.

Si l'on dispose d'une distribution observée à deux caractère sous forme d'un tableau de contingence { (X_j, Y_k, n_{jk}) ; j =1, 2, …, J et k = 1, 2, …, K }, on peut, par une démarche anligue définir des distributions marginales et des distributions conditionnelles.

3. 2. 1. Distribution marginale.

Nous distinguons la distribution marginale en x de la distribution marginale en y.

3. 2. 1. 1. Distribution marginale en x.

La distribution marginale en x est définie par l'ensemble des couples { (X_j, n_j) , j ∈ J } où l'on associe X_j les effectifs marginaux :

$$nj = \sum_k n_{jk} \quad \text{avec } j \in J.$$

3. 2. 1. 2. Distribution marginale en y.

La distribution marginale en y, par analogie est définie par { (Y_k, n_k) , k ∈ K } où l'on associe Y_k les effectifs marginaux :

$$n_k = \sum_j n_{jk} \quad \text{avec } k \in K.$$

Exemple. En reprenant les données de l'exemple précédant, on obtient le tableau suivant.

X_j \ Y_k	0	1	2	3	4	$n_{j.}$
1	4	4	2	0	0	10
2	9	16	4	0	0	29
3	4	12	9	2	0	27
4	1	6	1	1	2	11
5	0	1	0	1	1	3
$n_{.k}$	18	39	16	4	3	n = 80

Il est évident que les effectifs marginaux permettent de définir, comme dans toute distribution à une dimension, des fréquences marginales.

3. 2. 2. Distribution conditionnelle.

Le même tableau de contingence du paragraphe précédent permet de définir des distributions conditionnelles qui consistent à fixer à priori la valeur d'une variable et à examiner les variations de l'autre. De manière plus précise, nous avons des distributions conditionnelles de y en x, et des distributions conditionnelles de x en y.

3. 2. 2. 1. Distributions conditionnelles de y en x.

Pour les distributions conditionnelles de y en x, on se fixe une valeur de x, par exemple x = X_j. Si on prend tous les couples observés (X_j, Y_k), k ∈ K, ils définissent une distribution observée univariée appelée distribution conditionnelle de y étant donné que x = Xj :

$$\{ (Y_k , n_{jk}) ; \quad j \text{ fixé} \quad , k \in K \}.$$

On remarque que cette distribution comporte $n_{j.}$observations. Les fréquences conditionnelles auront donc pour valeurs :

$$f_{(j)k} = \frac{n_{jk}}{n_{j.}} \qquad j \text{ fixé}, \quad k \in K$$

Ces fréquences définissent les profils-lignes du tableau de contingence.

3. 2. 2. 2. Distributions conditionnelles de x en y.

Toujours par analogie, dans les distributions conditionnelles de x en y, on se fixe une valeur de y, par exemple $y = Y_k$.. La distribution conditionnelle de x étant donné que $y = Y_k$ est alors donné par :

$$\{ (X_j , n_{jk}) ; \quad k \text{ fixé} \ , j \in J \}.$$

De façon analogue, les fréquences conditionnelles sont définies par :

$$f_{j(k)} = \frac{n_{jk}}{n_{.k}} \qquad k \text{ fixé}, \qquad j \in J.$$

Nous obtenons ainsi les profils-colonnes du tableau de contingence.

3. 2. 3. Covariance.

Nous avons présenté dans les paragraphes précédents les distributions à deux variables. Tout d'abord sous forme de tableau à double entrée (tableau de contingence), puis ensuite les distributions marginales et conditionnelles. L'étude se poursuit généralement par la recherche de relation entre les deux variables.

Ces deux variables peuvent être totalement indépendantes l'une de l'autre, comme par exemple la taille et la réussite à un examen d'un candidat. Ces deux variables peuvent être, ou paraitre, liées par une relation fonctionnelle comme c'est le cas pour certaines grandeurs en physique, les deux variables sont totalement dépendantes. Entre les deux situations (dépendance et indépendance), il peut exister une proximité, ou une dépendance plus ou moins marquée entre ces deux variables. On peut utiliser un test appelé le chi-deux (qui ne fait pas l'objet d'un cours dans ce module) pour mesurer cette proximité. Lorsque la dépendance est plutôt de type 'linéaire, la covariance (objet de ce paragraphe) et la corrélation (objet du paragraphe qui suit) sont des caractéristiques algébriques qui permettent de mesure cette dépendance linéaire.

La covariance est une méthode mathématique d'évaluation du sens de variation de deux variables et permettant de qualifier leur indépendance. Les deux variables concernent la moyenne de leurs produits centrés sur leurs espérance mathématique, ou encore la différence entre la moyenne de leurs produits et le produit de leurs moyennes.

Si deux variables sont indépendantes, leur covariance sera nulle, mais l'inverse n'est pas forcément vrai.

Ainsi, la covariance est la moyenne de la somme du produit des écarts des valeurs des deux variables par rapport à leur moyenne arithmétique (μ). Le terme « covariation » désigne cette dernière somme. On peut définir la covariance comme la moyenne de la covariation. Elle se calcule par la formule :

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

Si la covariance permet d'apprécier si deux variables X et Y ont tendance à être liées linéairement, l'inconvénient est que son interprétation est difficile car sa valeur dépend du choix des unités dur X et Y. Pour cette raison, il est préférable d'utiliser le coefficient de corrélation linéaire, et qui fait l'objet du paragraphe qui suit.

3. 3. Coefficient de corrélation linéaire. Droite de régression.

Il y a corrélation, ou dépendance, entre deux variables quantitative X et Y si elles ont généralement tendance à varier toutes deux dans le même sens ou en sens contraire. La forme (linéaire ou non linéaire), le sens (positif ou négatif) et l'intensité (parfaite, imparfaite ou nulle) caractérisent une corrélation entre les variables X et Y.

3. 3. 1. Coefficient de corrélation linéaire.

Le diagramme de dispersion permet une analyse qualitative de la tendance à une relation linéaire entre les variables X et Y. Le coefficient de corrélation linéaire permet de mesurer quantitativement la force de la corrélation (ou de la dépendance) linéaire entre ces deux variables. On note ce coefficient par la lettre r, et on le calcule à partir de la formule suivante :

$$r = \frac{\sum xy - n.\bar{x}\bar{y}}{(n-1)SxSy}$$

Où

- $\sum xy$ représente la somme des produits de chaque valeur de la variable X par la valeur correspondante de la variable Y

- n correspond au nombre de couples (x, y)

- $\bar{x}$ représente la moyenne des valeurs de la variable X

- $\bar{y}$ représente la moyenne des valeurs de la variable Y

- Ss est l'écart type corrigé de la variable X

- Sy est l'écart type corrigé de la variable Y.

Le coefficient de corrélation linéaire est un nombre, sans unité, compris entre – 1 et + 1.

La corrélation linéaire est parfaite et positive pour r = 1, et négative pour r = - 1, puis nulle pour r = 0

Dans le cas d'une corrélation positive, plus la valeur de r est près de 1, plus la corrélation entre X et Y est forte. Il en est de même pour une corrélation négative : plus la valeur de r est près de – 1, plus la corrélation entre X et Y est forte.

Cependant, un coefficient de corrélation linéaire égal à zéro n'implique pas nécessairement que les variables X et Y sont indépendantes. La corrélation linéaire nulle indique seulement qu'il n'y a pas de dépendance linéaire entre les deux variables. Ces derniers pourraient néanmoins entretenir une relation de dépendance non linéaire. Seul le nuage de points peut nous garantir qu'il n'y a aucune autre forme de dépendance entre les variables X et Y.

Lors de l'interprétation de résultats, il faut être prudent. Il n'y a pas nécessairement une relation de cause à effet entre deux variables dépendantes. Une corrélation entre X et Y peut résulter de différents types de liaisons : X peut être la cause de Y, Y peut être la cause de X, les deux variables peuvent être causées par un facteur Z ou par un mélange de ces rapports.

3. 3. 2. Droite de régression

La régression est une méthode statistique par laquelle on essaie de prévoir la valeur d'une caractéristique en étudiant sa relation avec une ou plusieurs autres caractéristiques. Cette relation s'exprime au moyen d'une équation de régression. Le modèle de régression est utilisé pour décrire la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Ce modèle présente des formes et des degrés de complexité très divers.
Nous nous limiterons dans ce qui suit à présenter la droite de régression.
Lorsque le diagramme de dispersion indique qu'il existe une corrélation linéaire entre deux variables, on exprime mathématiquement cette relation par l'équation d'une droite. On appelle droite de régression la droite qui représente le mieux le nuage de points. On considère que la droite qui s'ajuste le mieux aux points est celle pour laquelle la valeur D, égale à la somme des carrés des écarts entre chaque points du diagramme de dispersion et la droite minimale. $D = d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2$

A l'aide de cette méthode, que l'on appelle méthode des moindres carrés, on en arrive à trouver la pente de la droite, notée b, et son ordonnée à l'origine, notée a, ce qui nous donne l'équation suivante : $y = a + bx$

On calcule les valeurs de a et b selon les formules qui suivent.

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

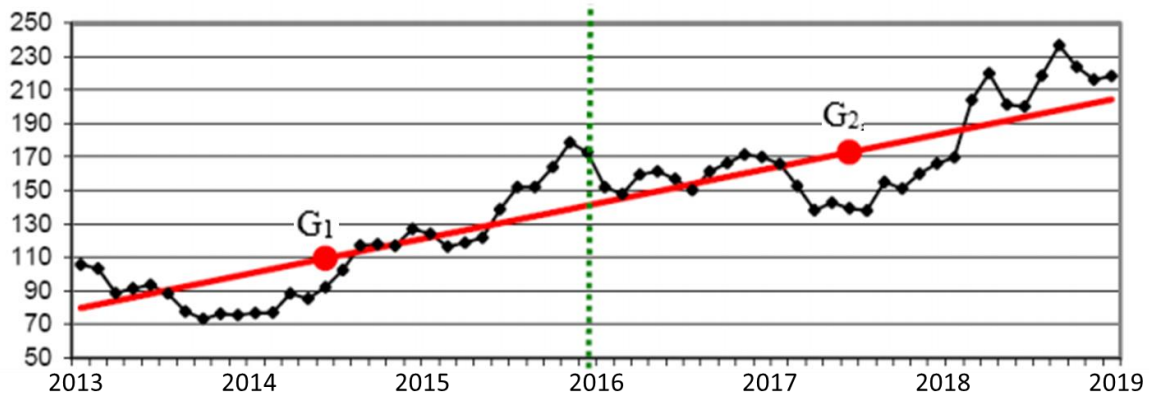
et $b = \frac{\sum xy - n.\bar{x}.\bar{y}}{(n-1)S_x^2}$

On estime la régression pour estimer la valeur y associée à une valeur x donnée. Il suffit de remplacer x dans l'équation y = a + bx pour obtenir la valeur correspondante. On peut aussi faire l'inverse, soit trouver la valeur x correspondant à une valeur y donnée.

3. 3. 3. Droite de Mayer

La droite de Mayer, est une autre représentation graphique de la relation entre deux variables X et Y. Comme toute méthode graphique, cette droite ne donne pas de résultats très précis. Cependant elle est moins subjective par rapport à la droite de régression.

La droite de Mayer consiste à partager un nuage de points en deux sous-ensembles de même effectif (éventuellement à une unité près), classés préalablement par abscisses croissantes. Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et la moyenne de y_i. On obtient ainsi deux points G_1 et G_2, appelés points moyen, de coordonnées $(\overline{x_1}, \overline{y_1})$ et $(\overline{x_2}, \overline{y_2})$. Ensuite on trace la droite passant par le deux points moyens G_1 et G_2. La droite d'ajustement est alors la droite de Mayer G_1 G_2.



La droite passant par les deux points G_1 et G_2 représente la droite de Mayer

Exercices

Exercice 1

En observant le chiffre des ventes réalisé sur un échantillon de 15 clients chez un marchand, on a recueilli les montants suivants : 0,10 – 0,10 – 0,25 – 0,25 – 0,25 – 0,35 – 0,40 – 0,53 – 0,90 – 1,25 – 1,35 – 2,45 – 2,71 – 3,09 et 4,10.

1. Déterminez :

- La moyenne,
- La médiane,
- Le mode.

2. Comment pourrait-on décrire les données du point de vue de la dissymétrie ?

3. Si on vous demandez de fournir la grandeur la plus susceptible de caractériser le montant typique d'achat chez ce commerçant, sur laquelle des mesures de tendance centrale porteriez-vous votre choix ? Pourquoi ?

Exercice 2

Dans deux lycées A et B, on pose la question : « Etes-vous sportif ? ». Les réponses sont données dans les tableaux qui suivent.

Tableau 1

Lycée A	Filles	Garçons
Effectifs	32	72
Oui	11	30
Pourcentages	34	41

Tableau 2

D'après ces résultats, les garçons sont-ils plus sportifs que les filles ? Justifiez votre réponse.

Lycée B	Filles	Garçons
Effectifs	115	64
Oui	72	45
Pourcentages	62	70

Exercice 3

Le prélèvement d'un échantillon de 20 ouvriers à la production, au sein d'une petite entreprise, a permis d'établir la liste suivante de salaire hebdomadaire :

140 – 240 – 140 – 230 – 140 – 140 – 225 – 155 – 165 – 155 – 140 – 205 – 200 – 190 – 140 – 140 – 165 – 140 – 180 – 180.

1. Déterminez :

- La moyenne,
- La médiane,
- Le mode.

2. Comment pourrait-on décrire les données du point de vue de la dissymétrie ?

2. En supposant successivement que vous occupiez les postes (a) ou (b), quelle est la grandeur que vous seriez incité à proclamer comme typique des salaires de l'entreprise ? Justifiez votre réponse.

(a) : vice-président de la compagnie.
(b) : président de l'unité syndicale chargée de la négociation.

Exercice 4

Soit la distribution d'effectifs relative à des taux mensuels de loyers pour 200 appartements. En supposant qu'il s'agisse d'appartements situés dans la même zone géographique, déterminez ce qu'on pourrait appeler le loyer typique du district en question, c'est-à-dire la moyenne, la médiane et le mode. Précisez lequel est le plus représentatif. Quel commentaire peut-on faire sur la forme de la distribution ?

Loyer mensuel	Nombre d'appartements (f)
150 – 179	3
180 – 209	8
210 – 239	10
240 – 269	13
270 – 299	33
300 – 329	40
330 – 359	35
360 – 389	30
390 – 419	16
420 - 449	12

Exercice 5

Lors d'un séminaire sur la lutte contre la pollution, un industriel affirme que : « Notre usine pollue **60%** de moins que la moyenne nationale ». Pour argumenter son chiffre, il fournit les données suivantes : « Mon usine dégage seulement **3** tonnes par année de monoxyde de carbone et **1,2** tonnes par an de monoxyde d'azote, alors que la moyenne nationale des quantités polluantes dégagées dans l'atmosphère de monoxyde de carbone est de **7,5** tonnes par an et celle des quantités polluantes dégagées de monoxyde d'azote est de **3** tonnes par an ».

1 – Cet industriel est-il dans le vrai ? Ces chiffres lui donnent-ils raison ?

Doutant un peu de l'affirmation de cet industriel, on a pu obtenir les informations complémentaires suivantes sur la production.

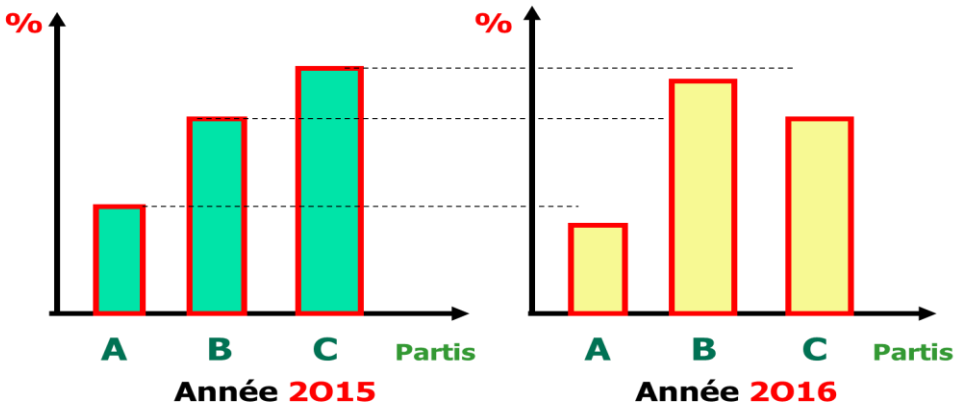
Soient les moyennes nationales des quantités polluantes dégagées de monoxyde de carbone et d'azote, pour des usines de même taille que celles de l 'industriel. Pour les quantités polluantes dégagées de monoxyde de carbone, la moyenne est de **2** tonnes par an, et pour les quantités polluantes dégagées de monoxyde d'azote, la moyenne est de **0,8** tonne par an.

2 - Ces données complémentaires modifient-elles votre réponse précédente ? Justifiez votre réponse.

Exercice 6

Soient trois partis politiques **A**, **B** et **C**. Un politicien du parti **B**, en concurrence avec les deux autres partis **A** et **C** présente aux téléspectateurs, à une heure de très grande écoute, les deux histogrammes du graphe qui suit correspondants à deux périodes consécutives (l'année **2015** et l'année **2016**) pour montrer l'évolution de son parti en pourcentage de votes.

Graphe 1 : Histogrammes d'un politicien



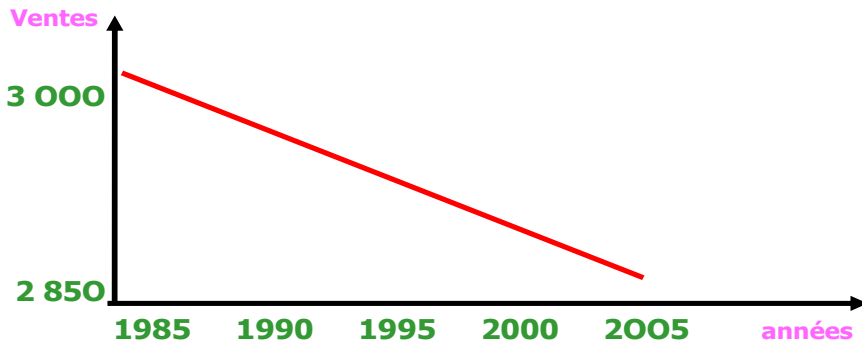
En présentant ces deux histogrammes, le politicien représentant le parti **B** tient le discours suivant : « Comparez l'évolution de notre parti par rapport aux autres, les progressions sont si évidentes que je ne commenterai même pas ces graphes, alors votez pour le parti qui représente l'avenir ».

Commentez la phrase de ce politicien.

Exercice 7

Les représentants des travailleurs de l'**U**nion **G**énérale des **T**ravailleurs **A**lgériens d'une usine fabriquant des portes coulissantes et basculantes, sont venus demander au directeur général une augmentation de salaire. Ce directeur leur affirme que cela n'est pas possible car ces deux dernières décennies les ventes de portes ont chutées. Pour confirmer ses affirmations, il présente aux syndicalistes le graphe qui suit représentant l'évolution des ventes de portes des deux dernières décennies.

Graphe : Evolution des ventes durant deux décennies



D'après ce graphe l'explication de ce directeur vous semble-t-elle justifiée ?

Exercice 8

Soit le tableau **1** correspondant au nombre de chômeurs par niveau d'études suivis.

Tableau 1 : Nombre de chômeurs par niveau d'études.

Niveau d'études	Nb de chômeurs	%
Primaire terminé ou non	55 957	69,91
Moyen	8 552	10,69
Secondaire	5 706	7,13
Ecole normale	1 338	1,67
Supérieur non universitaire	1 978	2,48
Universitaire ou assimilé	1 219	1,52
Artistique	425	0,53
Niveau inconnu	4 861	6,07
Total	80 036	100,00

Source : D.I.N.E.R*

1 - D'après les données de ce tableau[24], quel est le niveau d'étude qui offre le plus de perspectives dans le monde du travail ?

Soit le tableau **2** qui nous renseigne cette fois sur la population active toujours en fonction du niveau d'études.

Tableau 2 : Population active par niveau d'étude.

Niveau d'études	Population active	%
Primaire terminé ou non	2 056 312	56,60
Moyen	626 819	17,25
Secondaire	397 339	10,94
Ecole normale	96 815	2,67
Supérieur non universitaire	156 678	4,31
Universitaire ou assimilé	136 343	3,75
Artistique	17 946	0,49
Niveau inconnu	144 521	3,99
Total	3 632 773	100,00

Source : D.I.N.E.R*

2 - A partir des données supplémentaires du tableau **2**, peut-on savoir avec une plus grande précision quelles sont les niveaux d'études qui offrent le plus de perspectives dans le monde professionnel ?

*** Source : D.I.N.E.R (Données Imaginaires Nécessaires pour ces Exercices et leurs Résolutions).**

[24] *** Source : D.I.N.E.R (Données Imaginaires Nécessaires pour ces Exercices et leurs Résolutions).**

Exercice 9

Mohammed a subi un test d'aptitude professionnel, et a obtenu un résultat se situant au 82° rang centile pour son aptitude aux mathématiques.

1. Quel est le pourcentage des résultats qui ont été inférieurs à son résultat ?

2. Quel est le pourcentage des résultats qui lui ont été supérieurs ?

Exercice 10

Lors d'un concours pour le recrutement d'enseignants, la note minimale de réussite a été fixée au 70°rang centile ou plus. Les résultats ont d'abord été calculés selon un barème de 0 à 100 points.

Est-ce qu'un résultat original de 82 est un résultat suffisant pour être recruté ?

Exercice 11

Beaucoup de voyageurs réservent des places, puis ne se présentent pas au vol prévu. La compagnie aérienne a choisi un échantillon de 40, et a enregistré le nombre de personnes qui ne se présentent pas, et qui bénéficient des tarifs réduits. Les résultats ont été les suivants :

3 – 5 – 2 – 5 – 1 – 4 – 0 – 7 – 8 – 10 – 12 – 7 – 5 – 0 – 2 – 7 – 5 – 8 – 6 – 12 – 6 – 10 – 18 – 16 – 21 – 9 – 10 – 3 – 9 – 7 – 9 – 10 – 15 – 8 – 4 – 6 – 5 – 7 – 9 – 9.

1. Faites une représentation graphique de ces données sous forme de box plot.

2. Déterminez la dispersion interquartile.

Pour éviter des pertes, la compagnie aérienne a décidé d'exiger que les tarifs réduits soient payés à l'avance, et qu'une prime soit imposée pour les annulations de dernière minute, ou pour ceux qui ne se présentent pas. Lorsque cette politique était en vigueur, un autre échantillon de 40 vols a été choisi, et le nombre de ceux qui ne se présentent pas et qui bénéficient des tarif réduits a été enregistré. Les résultats obtenus sont les suivants :

1 – 7 – 6 – 2 – 3 – 2 – 1 – 0 – 3 – 9 – 3 – 1 – 5 – 0 – 7 – 2 – 6 – 3 – 3 – 5 – 2 – 9 – 11 – 3 – 6 – 2 – 0 – 4 – 7 – 8 – 12 – 0 – 3 – 7 – 6 – 2 – 1 – 1 – 2 – 4.

3. Faites une représentation graphique de ces nouvelles données sous forme de box plot.

4. Déterminez la dispersion interquartile.

5. Comparez ce graphe avec celui obtenu en question (1). Discutez de l'emplacement des médianes, de l'emplacement de la partie centrale de la banque des données, et de la distribution entre les valeurs Q_1 et Q_3 avec leurs valeurs extrêmes.

Exercice 12

« Nous entrons dans une période de récession », affirment les représentants de l'**U**nion **G**énérale des **T**ravailleurs **A**lgériens, « la moyenne nationale des salaires a baissé ce mois-ci, pour la première fois depuis des années, alors qu'on nous avait promis une augmentation des salaires avec la relance économique ».

« Absurde », déclarent les experts du premier ministère, « les rapports des wilayas montrent que la moyenne des salaires de chaque wilaya, sans exception, a continuée d'augmenter ».

Ces déclarations sont apparemment contradictoires. Alors, qui faut-il croire, les représentants de l'**U.G.T.A** ou ceux du premier ministère ?

Justifiez votre réponse en vous aidant d'un exemple numérique.

Exercice 13

Soit la comparaison des salaires dans deux entreprises : l'une de transport et l'autre de marketing. Les ingénieurs de l'entreprise de transport sont deux fois mieux payés que ceux de l'entreprise de marketing, il en est de même pour les ouvriers. Et pourtant, affirme un syndicaliste de l'entreprise du transport, le salaire moyen où il travaille représente moins de la moitié du salaire moyen de l'entreprise de marketing.

L'affirmation de ce syndicaliste est-elle vraie ? Justifiez votre réponse avec un exemple.

Exercice 14

Soit le tableau qui suit où les prix du baril de pétrole sont indiqués selon la durée (en mois) et les quantités vendues (en million de barils).

Tableau : Prix du baril de pétrole en fonction de la durée et des quantités vendues.

Durée (en mois)	**Prix** (en $)	**Quantité vendue** (10^6)
2	9	5
3	13	7
1	15	4
4	20	15
2	14	10

Source : D.I.N.E.R

Calculez le prix moyen en **19XX** du baril de pétrole d'après ces données.

Exercice 15

Soit une réunion de statisticiens et de gestionnaires qui a pour ordre du jour les prévisions de production pour l'année à venir. D'après certaines estimations, la production de la prochaine année allait être supérieure d'environ **2%** à l'année courante qui allait s'achever.

Conclusion évidente : la production augmentera lentement, mais continuera d'augmenter.
Après cette conclusion et en fin de réunion, un statisticien déclare alors que si la production de l'année prochaine dépassait de **2%** seulement celle de l'année en cours, cela signifiait qu'elle allait en fait diminuer.
Est-ce possible que ce statisticien ait raison malgré la conclusion évidente ?
Justifiez votre réponse.

Exercice 16

Supposons qu'on veuille calculer, pour en suivre l'évolution, un indicateur de durée de chômage. Une idée est d'interroger, directement ou « sur dossier », les personnes en chômage à une date donnée et leur demander depuis combien de temps elles le sont. On obtient ainsi une durée moyenne qu'il est raisonnable d'appeler « ancienneté moyenne de chômage ». Mais, cet indicateur présente un défaut évident, celui de considérer des durées de chômage en cours, non achevées : la durée totale de chômage considérée sera évidemment plus longue que l'ancienneté que l'on prend en compte. Celle-ci semble donc sous-estimer la durée moyenne recherchée.
Pour éviter cet inconvénient, interrogeons l'ensemble des chômeurs qui pendant une période donnée (par exemple une année) ont retrouvé un emploi, ou sont partis à la retraite, et

demandons-leur combien de temps a duré leur période de chômage. Cette fois-ci, il s'agit bien de durées achevées totales, dont on peut faire la moyenne.
Cette moyenne est-elle supérieure ou inférieure à celle calculée précédemment ?

Exercice 17

Supposons qu'une catégorie de travailleurs se plaigne de gagner **50%** de moins que telle autre et obtienne, de haute lutte, d'être augmentée de **50%**.
Cette catégorie de travailleurs est-elle satisfaite ? Justifiez votre réponse.

Exercice 18

Soient les phrases suivantes (où il est question de moyenne et pourcentage) qui ont été lues dans un quotidien, une affiche publicitaire, entendues à la radio ou à la télévision.
1 - « Les coûts de production d'une entreprise ont augmentés de **15%** et ses prix de vente de **10%** seulement, ses bénéfices sont donc amputés ».
2 - « XXX (une marque de lessive) lave **15%** plus propre ! ».
3 - « Vos beaux billets de **1 000 U.M** valent **10%** de plus pendant la journée de promotion commerciale. **10%** de réduction, c'est toujours bon à prendre ! ».
4 - « La population active agricole a été réduite de **300%**. Elle est passée de **6** à **2** millions d'individus ».
5 - « Il y a eu dévalorisation d'une monnaie de **200%**, elle est passée de **400** à **200** ».
6 - « Rabais sur **50%** du stock ! ».
7 - « Les confitures Miam-Miam contiennent **20%** de calories en moins que la moyenne des confitures vendues sur le marché. C'est bon pour votre ligne ! ».
Commentez ces différentes annonces.

Exercice 19

Un concours national est organisé par une association scientifique pour élire le meilleur groupe d'étudiants de l'année, en fonction de leurs connaissances et raisonnements scientifiques. Une des conditions pour participer à ce concours, est que la moyenne annuelle des notes du groupe soit supérieure ou égale à **13,20** sur **20**.
Sous la pression de son supérieur hiérarchique, qui lui-même subit la pression d'un inconnu, un chef de département d'une certaine faculté doit remettre, en urgence au Doyen un dossier pédagogique sur un groupe de cinq étudiants devant participé à ce concours.

Dans ce dossier, doivent être précisées les moyennes des modules suivis durant l'année universitaire par ces étudiants que l'on représentera par les chiffres romains **I**, **II**, **III**, **IV** et **V**. Les documents dont dispose ce chef de département sont les relevés de notes (**RN**) de **11** modules enseignés et représentés par les chiffres arabes **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**, **10** et **11**. Tous les modules sont affectés d'un coefficient identique qui est égal à **1**. Les relevés de notes disponibles sont donnés dans la synthèse qui suit.

Synthèse 10 : Relevés de notes des modules enseignés.

RN_1. **I** : -	**RN_2** . **I** : 16	**RN_3** . **I** : 16	**RN_4** . **I** : -	**RN_5** . **I** : 15,5
II : 08	**II** : 16	**II** : 15	**II** : 15	**II** : 15
III : 10,5	**III** : 15,5	**III** : 14	**III** : 15,5	**III** : 11
IV : 15	**IV** : 15,5	**IV** : 15	**IV** :14	**IV** : 10
V : -	**V** : 12	**V** : 12	**V** : -	**V** : 14,5
RN_6 . **I** : 15	**RN_8** . **I** : 14	**RN_9** . **I** : 09	**RN_{10}** . **I** : 13	**RN_{11}**. **I** : 14
II : 16,5	**II** : 16	**II** : 11	**II** : 10	**II** : 15
III : 16	**III** : 15	**III** : 11	**III** : 11	**III** : 14
IV : 13	**IV** : 16	**IV** : 07,5	**IV** : 07	**IV** : 11
V : 15,5	**V** : 15	**V** : -	**V** : 12	**V** : 13,5

Dans le dossier remis par le chef de département à son supérieur hiérarchique, les moyennes y figurant sont données dans le tableau qui suit.

Tableau : Moyennes des étudiants.

Etudiants	I	II	III	IV	V
Moyennes	13,73	13,59	13,41	12,55	12,91

D'après ces moyennes, pensez-vous que le Doyen de cette faculté pourra inscrire ce groupe d'étudiants au concours national pour représenter l'institution qu'il gère, sachant que la moyenne du groupe d'étudiants doit être supérieure ou égale à **13,20** sur **20** ?

Exercice 20

Ce cas s'est produit au **U**nited **S**tates of **A**merica, à l'époque de l'administration Carter. Au nom de l'égalité, l'administration surveillait soigneusement les proportions d'hommes et de femmes, de blancs et de noirs, dans les entreprises, les écoles et les administrations.

La sous-commission pour les droits de l'administration s'est un jour penchée sur une université de la côte ouest, et on a découvert que la proportion d'étudiants reçus aux examens était de **75%**, alors que la proportion des étudiantes reçues n'était que de **56%**. L'enquête chez les doyens a permis d'obtenir les résultats donnés dans le tableau qui suit.

Tableau : Répartition des étudiants selon la faculté.

Facultés / Etudiants(es)	Lettres	Sciences
Etudiants présents	100	500
Etudiants reçus	50	400
Etudiantes présentes	400	100
Etudiantes reçues	200	80

Dans la faculté de **L**ettres : sur **100** étudiants présents, **50** ont été reçus, et sur **400** étudiantes présentes, **200** ont été reçues.

Dans la faculté des **S**ciences : sur **500** étudiants présents, **400** ont été reçus, et sur **100** étudiantes présentes, **80** ont été reçues.

Ainsi en **L**ettres, il y a **50%** de reçus, tant chez les étudiants que chez les étudiantes ; en **S**ciences de même, dans une proportion de **80%**. Nos deux doyens sont donc irréprochables.

Pourquoi y a-t-il eu alors plainte et enquête ?

Exercice 21

Dans un quotidien, on peut lire : 3 390 personnes ont voté, c'est-à-dire 75% des inscrits sur les listes électorales.

Calculer le nombre d'inscrits.

Exercice 22

La TVA se calcule sur le prix hors taxe.

1. Soit la facture : Total HT ..... 70 UM.

TVA ..... 13,72 UM

Quel est le taux de TVA appliqué ?

2. Soit une autre facture

Prix TTC des réparations ……..548,6 UM

dont TVA…….. 28,6 UM

Quel est le taux de TVA appliqué ?

Solutions des exercices

Solution de l'exercice 1

1. - Moyenne : $\overline{X} = \frac{\sum X}{n} = 18{,}08 = 1{,}21$.

- Médiane : Me = 0,53 (valeur du milieu).

- Mode : Valeur la plus fréquente = 0,25.

2. La moyenne étant plus grande que le mode, tandis que la médiane se situe nettement entre les deux, on peut en déduire que la distribution en question est dissymétrique, étalée vers la droite (côté positif).

3. S'agissant d'une distribution fortement dissymétrique telle que celle-ci, la moyenne ne convient pas pour la caractériser. Le choix entre la médiane et le mode est plus difficile. Toutefois, on peut dire qu'il n'y a pas vraiment une forte concentration autour du mode, de sorte que la médiane 0,53 est la grandeur qui représente le mieux la distribution. N'oublions pas qu'elle signifie que 50% des achats dépassent la valeur 0,53, alors que 50 % lui sont inférieurs.

Solution de l'exercice 2

En lisant la dernière ligne du tableau des énoncés, on serait tenté de penser que les garçons sont plus sportifs que les filles. On lit **41%** et **70%** dans les deux lycées pour les garçons, alors que pour les filles ces pourcentages sont respectivement de **34** et **62** seulement.

Mais, les exercices précédents nous ont appris à nous méfier de l'évident. Si l'on regroupe les deux lycées, on obtient le tableau qui suit.

Tableau : Résultats du sondage selon le genre.

Filles	Garçons
147 83	136 75
56,46%	**55,14%**

Ce tableau nous donne un autre résultat. On constate que les filles sont plus sportives que les garçons (**56,46%** pour les filles et **55,14%** pour les garçons).

Alors selon l'objectif que l'on se fixe, on présentera un tableau ou l'autre (plus exactement deux tableaux ou un seul tableau). Dans cette situation, les différentes structures entre les deux populations observées annulent toute validité à la comparaison des moyennes globales. Cette apparente contradiction est appelée par les statisticiens l'« effet de structure ». C'est un phénomène que nous rencontrons assez souvent dans la pratique.

Solution de l'exercice 3

1. – Comme il s'agit d'une moyenne arithmétique simple, il suffit de faire la somme de tous les salaires, puis ensuite diviser cette somme par le nombre total d'ouvriers.

$$\frac{140.8 + 155.2 + 165.2 + 180.2 + 190 + 200 + 205 + 225 + 230 + 240}{20} = 170{,}5$$

Moyenne : $\overline{X}$ = 170,50.

- Médiane. Dans la mesure où les données brutes ne sont pas données en ordre, il faut commencer par classer la série par ordre croissant.

140 – 140 – 140 – 140 – 140 – 140 – 140 – 140 – 140 – 155 – **155 – 165** – 165 – 180 – 180 - 190 – 200 – 205 – 225 – 230 – 240.

La taille de l'échantillon correspond à un nombre pair, donc la médiane et la moyenne des deux valeurs centrales. Me = $\frac{155+165}{2}$ = 160.

Médiane : Me = 160.

- Mode = 140.

2. La moyenne étant plus grande que le mode, tandis que la médiane se situe nettement entre les deux, on peut en déduire que la distribution en question est dissymétrique, étalée vers la droite (côté positif).

3. Les points de vue concernant les salaires sont totalement opposé selon que l'on soit un vice-président de l'entreprise, qui cherchera à minimiser ses coûts (en partie à travers les salaires), ou que l'on soit un représentant syndical, qui essaiera d'obtenir une augmentation maximal des salaires.

– Vice-président. Vous seriez enclin à considérer la moyenne comme la grandeur caractérisant le mieux la situation de l'entreprise (170). Vous pourriez faire remarquer que l'avantage d'une

moyenne réside dans le fait qu'elle tient compte de chacune des données individuelles. Vous pourriez (à tort) être tenté de dire que c'est la seule grandeur réellement représentative.

- Président de l'unité syndicale chargée de la négociation. Vous avez le choix entre le mode et la médiane. Le mode peut ici être considéré comme très représentatif du fait qu'il représente le niveau de salaire gagné par l'effectif partiel qui est de loin le plus nombreux (8). La médiane représente la valeur qui départage la moitié des ouvriers les mieux payés et la moitié des ouvriers les moins bien payés.

Solution de l'exercice 4

Dans cet exercice, les données sont groupées. Il faudra donc tout d'abord considérer les centres des classes (X_i), ensuite faire la somme des (f).X, puis calculer les effectifs cumulés. Nous présentons les calculs dans le tableau qui suit.

X_i	Nombre d'appartements (f)	(f).X_i	Cumul
165	3	495	-
195	8	1 560	2 055
225	10	2 250	4 305
255	13	3 315	7 620
285	33	9 405	17 025
315	40	12 600	29 625
345	35	12 075	41 700
375	30	11 250	52 950
405	16	6 480	59 430
435	12	5 220	64 650

- Moyenne : $\mu = \frac{\sum fXi}{N} = \frac{64\,650}{200} = 323{,}25$

- Médiane = 324,25.

- Mode = 317.

Les trois valeurs ne sont pas très éloignées les unes des autres, mais ne sont pas égales, ce qui signifie que la distribution n'est pas symétrique. Comme la moyenne est inférieure à la médiane et le mode, on peut dire que l'étalement de la distribution est du côté négatif, donc à gauche

Solution de l'exercice 5

1 - Les chiffres présentés par cet industriel sont insuffisants pour pouvoir donner un avis objectif sur cette pollution. Il faut toujours vérifiez si les données dont vous disposez sont suffisantes pour une analyse complète et objective.

La pollution varie selon la taille de l'usine et la nature du produit fabriqué, évidemment il y a d'autres facteurs que nous ne considérons pas ici. Cet industriel utilise la moyenne nationale, sans aucune précision sur la taille de son usine. Est-ce que la taille de son usine est la même que celle qu'il prend comme référence pour la moyenne nationale ?

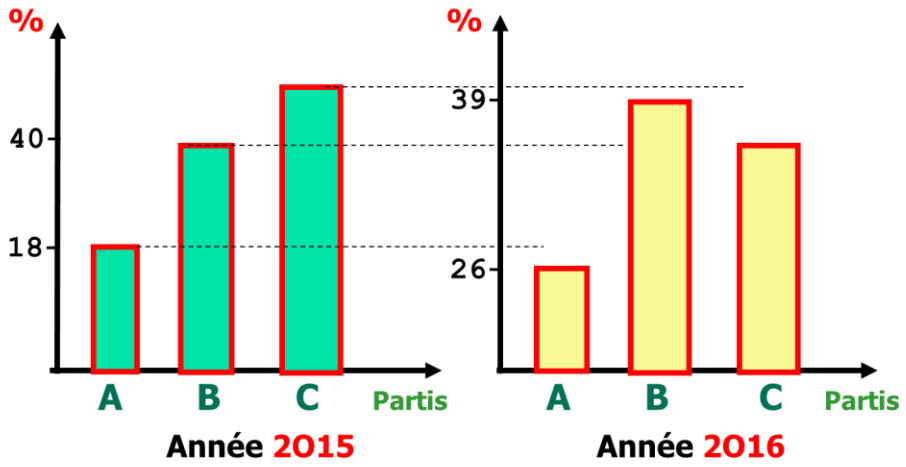
2 - D'après ces données complémentaires concernant les moyennes nationales pour les usines de même taille, on est en mesure de savoir si cet industriel a raison ou pas. Les moyennes nationales pour les usines de même taille sont respectivement de **2 T**/an et **0,8 T**/an pour le monoxyde de carbone et le monoxyde d'azote. L'industriel par contre a comparé **3** à **7,5** et **1,2** à **3** alors qu'il devait normalement comparer **3** à **2** et **1,2** à **0,8**.

L'affirmation de l'industriel est donc totalement fausse. Son usine ne pollue pas **60%** de moins que la moyenne nationale, comme il voudrait le faire croire, mais réellement elle pollue **50%** de plus que la moyenne nationale.

On peut en conclure que cet industriel a su manier les chiffres à son avantage en trompant son public. Il s'agit là d'un sophiste, et ne pas confondre avec un soufi (ou çoufi).

Solution de l'exercice 6

Reprenons les deux histogrammes des énoncés, et ajoutons les chiffres représentant les pourcentages. On obtient les histogrammes qui suivent.



La lecture de ce graphe nous montre que le pourcentage de vote du parti **B** n'est pas en hausse comme voulait nous le faire croire ce politicien, mais en baisse (de **40** à **39%**). Par contre le parti **A** est en nette progression (de **18** à **26%**). Les deux histogrammes n'ont pas la même échelle, d'où la confusion dans les pourcentages de votes.

Cette phrase est bien celle d'un politicien (avisé, diplomate, habile, parfois prudent parfois rusé ou tout simplement machiavélique). Il a effectivement dit la vérité en utilisant des termes et phrases très vagues tels que :

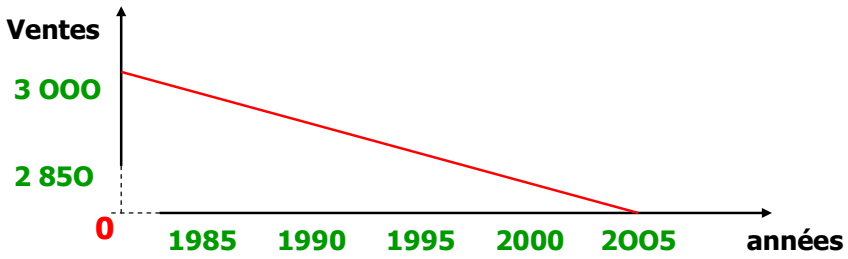
- « Evolution », une évolution peut être positive ou négative,
- « Progression », une progression peut aussi être positive ou négative,
- « Je ne commenterai pas ces graphes », prudence oblige, il ne pourra pas être jugé pour fausse déclaration,
- « Votez pour le parti », aucune précision sur ce parti.

Ainsi, ce politicien a utilisé un support mathématique (deux histogrammes) pour déformer scientifiquement la réalité.

Solution de l'exercice 7

Considérons le graphe où apparaît un véritable effondrement de la vente des portes. Cependant, en regardant plus attentivement ce graphe, on peut remarquer qu'il n'est pas correctement tracé. Effectivement, aucune précision sur l'origine de l'abscisse et de l'ordonné. Le graphe qui suit présente les corrections nécessaires.

Graphe : Correction du graphe.



En regardant de plus près les chiffres inscrits sur ce graphe, on constate que la baisse n'est pas aussi importante qu'elle ne le paraît, elle passe de **3 000** à **2 850** portes. La diminution est donc seulement de **5%** :

(3 000 – 2 850) / 3 000 = 0,05 = 5%.

L'explication de ce directeur n'est donc pas du tout justifiée. Une baisse de **5%** est une variation tout à fait normale dans les ventes de portes ou toute autre production similaire. Cela est tout simplement une manière élégante de la part de ce directeur de mettre les syndicalistes « à la porte ». و مع سلامة

Solution de l'exercice 8

1 - Surtout, ne vous pressez pas d'affirmer que le niveau d'étude qui offre le plus de perspectives dans le monde du travail est celui des artistes ou des universitaires. Ces statistiques sont insuffisantes pour répondre à la question de cet exercice.

Relisez encore attentivement les chiffres du tableau **1**.

Pour les artistes, le pourcentage **0,53** ne représente pas le taux de chômage au sein des artistes, mais seulement la proportion de chômeurs par rapport à l'ensemble de la population en chômage. Donc, attention à la mauvaise lecture des chiffres !

Si, par exemple, les artistes sont au nombre de **4 250**, alors **425** représente un taux de chômage de **10%** au sein des artistes. Par contre, si le nombre d'artistes est de **850**, le taux de chômage passe alors à **50%** pour le même nombre de chômeurs. Il est donc nécessaire, de connaître le nombre de la population active par niveau d'étude pour savoir quel est le niveau d'étude qui offre le plus de perspectives dans le monde du travail.

Pour pouvoir répondre à cette question, il faut donc connaître la population active par niveau d'étude.

2 - La lecture des données supplémentaires du tableau 2 ne nous permet pas directement de répondre à la question concernant le niveau d'étude qui offre le plus de perspectives d'emplois, mais des calculs à partir de ces données nous permettront d'y répondre.

A partir des tableaux 1 et 2, nous pouvons obtenir un tableau qui nous renseigne sur la répartition des chômeurs parmi la population active selon le niveau d'étude. Il suffit seulement d'utiliser une règle de trois.

Concernant le primaire :

2 056 312 $\Rightarrow$ 100 % = 1

55 957 $\Rightarrow$ x

x = (55 957 / 2 056 312) = 2,722%

Tableau : Répartition des chômeurs par niveau d'étude (taux de chômage).

Etudes	Pop. active	Nb de chômeurs	Tx de chômage
1 - Pri	2 056 312	55 957	2,722
2 - Moy	2 626 819	8 552	1,364
3 - Sec	397 339	5 706	1,436
4 - E N	96 815	1 338	1,382
5 - SNU	156 678	1 978	1,262
6 - UA	136 343	1 219	0,894
7 - Art	17 946	425	2,368
8 - NI	144 521	4 861	3,363

A partir de ce tableau, on peut connaître le niveau d'étude qui offre le plus de perspectives d'emplois. Contrairement au tableau **1**, le pourcentage correspondant à des études artistiques est l'un des plus élevé (en troisième position avec un taux de **2,368**). Ainsi, les études artistiques offre le moins de perspectives dans le monde du travail. Le niveau qui propose un taux d'activité

professionnelle le plus intéressant est le niveau d'étude universitaire ou assimilé, avec un taux de chômage de **0,894** seulement. Ainsi, en tant qu'étudiant universitaire de quatrième année, vous êtes en bonne voie : pas trop de chômage après l'obtention de votre diplôme (ou selon la personne pas trop de repos).

En conclusion, si vous ne voulez pas « chômer », vous avez tout intérêt à connaître votre niveau d'étude, car ceux qui ne le connaissent pas ont le taux de chômage le plus élevé (3,363).

Solution de l'exercice 9

1. 82 % ou moins.

2. 18 % ou moins.

Solution de l'exercice 10

Non, car la note de 82 pourrait représenter un rang centile plus petit de 70. Effectivement, supposant qu'en presque majorité les notes des candidats au concours soient supérieures à 80, il y a alors une très faible probabilité que la note 82 soit comprise dans les 30 derniers rangs centiles.

Solution de l'exercice 11

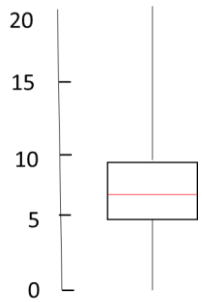
1. Pour pouvoir faire une représentation graphique, il faut tout d'abord classer les données afin de déterminer la médiane.

0 – 0 – 1 – 2 – 2 – 3 – 3 – 4 – 4 – 5 – 5 – 5 – 5 – 5 – 6 – 6 – 6 – 7 – 7 – 7 – 7 – 7 – 8 – 8 – 8 – 9 – 9 – 9 – 9 – 9 – 10 – 10 – 10 – 10 – 12 – 12 – 15 – 16 – 18 – 21.

Le rang de la médiane étant 20,5, la médiane sera donc 7.

Le rang quartile étant 10,5 le premier quartile $Q_1 = 5$ et le troisième $Q_3 = 9,5$.

Graphe : Nombre de personnes qui ne se présentent pas, tout en ayant acheté à des tarifs gratuit.



2. Dispersion interquartile : 9,5 – 5 = 4,5.

Solution de l'exercice 12

Que les déclarations des représentants syndicaux et celles du premier ministère soient contradictoires, c'est normal (*à la guerre comme à la guerre*). Mais, que les deux partis aient raison, cela semble bizarre. Et pourtant, les deux antagonistes disent la vérité.

Considérons la démonstration suivante.

Supposons que la moyenne des salaires soit de **130 U.M** à Alger et de **90 U.M** dans chacune des **47** autres wilayates. Imaginons ensuite, que **47** salariés gagnant chacun **120 U.M** quittent Alger, et vont s'installer à raison d'un salarié par wilaya dans le reste du pays en voyant ainsi leurs salaires diminuer à **100 U.M**. On aura donc les situations qui suivent.

Au niveau national : **47** salariés ont vu leurs salaires diminués, la moyenne a donc diminuée.

A Alger, on a enregistré le départ de **47** salariés qui gagnaient moins que la moyenne algéroise, donc cette moyenne a augmentée.

Dans une autre wilaya, autre qu'Alger, on a vu un salarié dont le salaire est supérieur à la moyenne par wilaya, donc cette moyenne a augmentée.

On peut donc conclure, que dans toutes les wilayates, la moyenne a augmentée alors que la moyenne nationale a diminuée. *On peut conclure aussi, qu'il n'est pas toujours nécessaire de mentir pour défendre ses intérêts. On peut les défendre en disant la vérité, rien que la vérité, mais pas toute la vérité.*

Solution de l'exercice 13

Au premier abord, je préfèrerai travailler dans l'entreprise de transport puisque les ingénieurs et les ouvriers sont deux fois mieux payés que ceux de l'entreprise de marketing. Cependant, l'affirmation du syndicaliste de l'entreprise de transport a semé le doute dans mes premières impressions. En me rapprochant de ce syndicaliste, il m'explique la situation qui prouve que son affirmation est effectivement vraie. Voici l'explication par des chiffres.

Synthèse : Explication par les chiffres.

Entreprise de transport **Entreprise de marketing**

Un ingénieur au salaire de 80 000 U.M **24 ingénieurs au salaire de 40 000 U.M**
24 ouvriers au salaire de 16 000 U.M **Un ouvrier au salaire de 8 000 U.M**

Calcul du salaire moyen (SM)

$$SMT=(80\,000.1+16\,000.24)/25 = 18\,560 \text{ U.M}$$

$$SMM=(40\,000.24 + 8\,000.1)/25 = 38\,720 \text{ U.M}$$

On constate d'après ces chiffres que le syndicaliste a effectivement raison, son salaire est bien inférieur de moitié à celui des travailleurs de l'entreprise de marketing. En effet **18 560** représente bien la moitié environ de **38 720**.

Dans cette situation, les différentes structures entre les deux populations observées annulent toute validité à la comparaison des moyennes globales. Il s'agit donc d'un effet de structure.

Solution de l'exercice 14

On nous demande de calculer le prix moyen en **19XX** du baril de pétrole, sans nous préciser de quel prix moyen il s'agit. Moyen, par rapport à la durée, ou moyen par rapport aux quantités échangées ?

Si l'on considère le temps, on a un prix moyen qui est égal à :

(9.2 + 13.3 + 15.1 + 20.4 + 14.2) / 12 = 15 $ le baril.

Si l'on considère la quantité, on a un prix moyen qui est égal à :

(9.5 + 13.7 + 15.4 + 20.15 + 14.10) / 41 = 15,51 $ le baril.

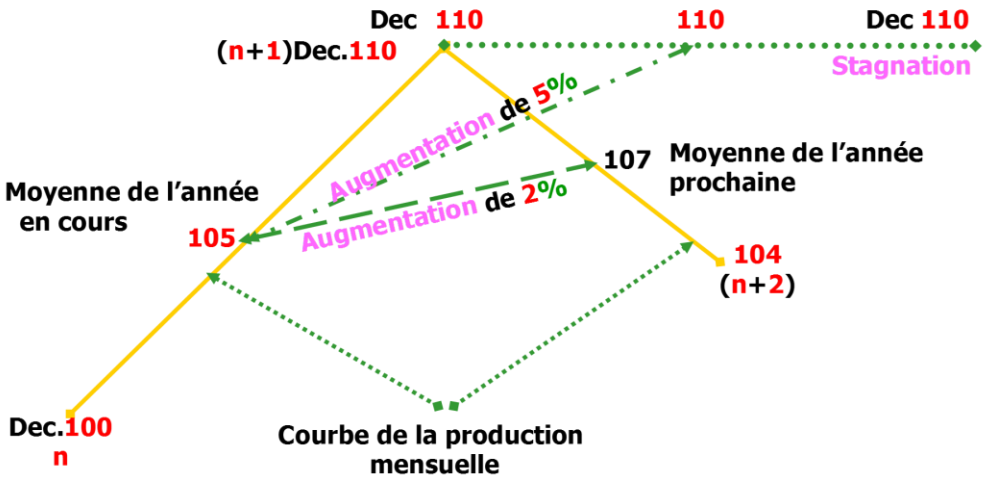
On constate, comme on s'y attendait, que ces prix sont différents. Lequel est correct ?

Ces deux moyennes sont justes. On utilisera l'une ou l'autre, selon que l'on veuille avoir une moyenne en hausse ou en baisse. Tout dépend donc, de l'utilisation de ces chiffres.

Solution de l'exercice 15

Pendant l'année en cours, la production n'a pas cessé d'augmenter de mois en mois pour atteindre à la fin de l'année un niveau supérieur de **10%** par rapport à celui du début (de **100** en l'an **n** à **110** en l'an **n + 1**). La production mensuelle moyenne de l'année était donc inférieure de quelques **5%** à celle du dernier mois (**109,167** à **110**).

Synthèse : Variation de la production de l'an **n** à (**n + 1**) et (**n + 2**).



Admettons à présent, que la production à la suite de cette hausse rapide, devienne tout à coup stagnante tout le long de l'année suivante. Elle restera donc, au niveau de la fin de l'année en cours, **5%** environ de plus que la production moyenne de cette année (segment en pointillés irréguliers dans le graphe). Pour que la progression soit de **2%** seulement, et non pas de **5%**, il faut que pendant cette prochaine année la production diminue, au moins, au dessous du chiffre atteint à la fin de l'année en cours.

Ainsi, si l'on affirme, sans aucune précision, qu'une certaine production a augmentée de **5%** en **2005**, cela signifie deux choses différentes :

- La moyenne de l'année **2005** a dépassé de **5%** la moyenne de l'année **2004**.
- La production à la fin de l'année **2005** dépasse de **5%** celle de la fin d'année **2004**.

C'est cette confusion qui est à la base de l'erreur d'interprétation que l'on vient de décrire.

Solution de l'exercice 16

Cette moyenne a toutes les chances d'être plus courte que l'ancienneté moyenne précédente, qu'on croyait pourtant sous-estimée. C'est encore un effet de structure qui explique ce paradoxe. En définissant un ensemble de chômeurs par le fait qu'ils ont retrouvé un emploi, on augmente la proportion des chômeurs de courte durée : quelqu'un qui dans l'année serait six fois chômeurs pendant un mois interviendrait six fois dans le calcul, par contre un chômeur permanent qui ne retrouve pas de travail n'est pas pris en compte, son chômage n'étant pas terminé. Alors, attention à la structure et les vrais faux calculs.

Solution de l'exercice 17

Dans la mesure où des travailleurs gagnaient **50%** de moins qu'une autre catégorie de travailleurs, et qu'ils ont été augmentés de **50%**, ils devraient être, au moins théoriquement, satisfaits. Or, ils ne le sont pas, et ils ont tout à fait raison. Lors des négociations entre syndicat et patron, il y a eu confusion.

Dans cette situation, il s'agit d'une confusion entre pourcentage « en dedans » et pourcentage « en dehors ».

Considérons un salaire de **1 000 U.M** d'une catégorie de travailleurs que l'on nommera **A** qui se plaint de gagner **50%** de moins qu'une autre catégorie de travailleurs que l'on nommera **B**. Le salaire de cette dernière catégorie est donc de **2 000 U.M** (en effet 1000.0,5 = 2 000). Après une longue lutte, la catégorie de travailleurs **A** obtient une augmentation de salaire de **50%** soit 1 000 + (1 000.0,5) = **1 500 U.M**.

Évidemment, cette catégorie de travailleurs n'est pas satisfaite puisque même après une augmentation de salaire de **50%**, elle n'atteint pas encore le salaire de la catégorie de travailleurs **B**.

En conclusion, contrairement à ce que nous pensons généralement Hadj Moussa est parfois différent de Moussa Hadj. Alors attention quand vous utiliserez cette expression dans le futur.

Solution de l'exercice 18

1 – La comparaison de deux pourcentages relatifs à des grandeurs sans relation étroite est un facteur de risque d'erreur. Ainsi, prenons comme exemple une entreprise qui peut vendre au prix de **1 000 U.M** un produit dont le coût de production est de **500 U.M**. Les deux prix ne sont unis que par le biais de la concurrence. Dans ces conditions, les coûts de production ont augmentés de **15%**, ils sont passés de **500** à **575 U.M**, et les prix de ventes de **10%**, passant de **1 000** à **1 100 U.M**. Le bénéfice brut passe de **500** (1 000 – 500 = 500) à **525 U.M** (en effet, 1 100 – 575 = 525) augmentant en réalité de **5%**.

Il n'y a donc aucune amputation de bénéfice, bien au contraire.

2 – Apposé à une grandeur qualitative, non mesurable, telle que la blancheur du linge, un pourcentage donne une impression de précision scientifique qui ne trompe que les esprits non avertis. La prochaine fois que vous lirez une annonce publicitaire, pensez que vous avez été avertis et en plus « Un homme averti en vaut deux ».

3 – Il s'agit encore là de la confusion entre pourcentage « en dedans » et pourcentage « en dehors ».

Si vos billets de **1 000 U.M** ont une valeur réelle de **1 100 U.M**, c'est que vous obtenez pour **1 000 U.M** un objet au prix marqué **1 100 U.M**, bénéficiant ainsi d'une réduction de **100 U.M** sur les **1 100 U.M**, soit environ **9,1%** et non pas **10%** comme veut le faire croire ce commerçant.

4 et **5 –** Les pourcentages de variation sont appelés aussi taux de croissance ou pourcentage d'évolution. Ils permettent de mesurer la vitesse à laquelle varient les grandeurs dont on mesure la croissance entre deux dates. La baisse de ces pourcentages ne peut aller au-delà de **100%**, une baisse de **100%** ramène la valeur de la grandeur à zéro. Par contre, les hausses peuvent être illimitées.

Ces deux phrases sont donc fausses. Si la population active agricole est passée de **6** à **2** millions, elle a donc diminuée de **33,33%** seulement et non pas de **300%**. Une baisse de **300%**

signifierait que la population serait de – **18** millions, ce qui n'a aucun sens dans la mesure où une population négative n'existe pas, à moins que le journaliste qui a écrit cette phrase croie encore aux fantômes.

Pour la dévalorisation de la monnaie de **400** à **200**, cela correspond à une baisse de **50%**. Concernant des sommes d'argent, on peut avoir des chiffres négatifs et cela correspond dans une comptabilité à une perte ou un déficit.

6 – Lisez attentivement cette affiche, il ne s'agit pas d'un rabais de **50%** sur les prix, mais pour **50%** du stock il y a un rabais. Aucune précision sur le pourcentage de ce rabais.

7 – On pourrait croire en lisant cette annonce que les confitures Miam Miam contiennent **20%** de calories en moins que la moyenne de toutes les autres confitures du marché, et qu'elles sont très légères.

Méfiez-vous et considérons l'exemple suivant. Soient quatre marques concurrentes dont les teneurs en calories sont de **3**, **100**, **4** et **3** alors que pour Miam Miam la teneur est de **26**. La teneur moyenne des marques concurrentielles est donc de :

$$(3 + 100 + 4 + 3) / 4 = 27,5.$$

Il n'y a donc pas de publicité mensongère, même si l'on a compris tout autre chose.

Solution de l'exercice 19

En faisant la moyenne des étudiants du groupe, on trouve une valeur supérieure à **13,20** qui est la valeur requise pour participer à ce concours. Mais, ne vous pressez pas d'affirmer que le Doyen de cette faculté peut inscrire ce groupe d'étudiants au concours national pour représenter l'institution qu'il gère.

Si vous avez calculé cette moyenne, vous n'avez pas encore compris l'objectif de cette série d'exercices. Effectivement, vous avez calculé une moyenne sur la base de fausses moyennes. Relisez attentivement les relevés de notes qui ont servis de base pour le calcul de ces moyennes. Théoriquement, il doit y en avoir **11**, or il n'y en a que **10**. A cela s'ajoute le fait que des relevés de notes sont incomplets (le **1**, le **4** et le **9**).

Ainsi donc, sur la base des données disponibles, il nous est impossible de calculer une quelconque moyenne, et encore moins de donner un avis sur la participation de ce groupe d'étudiants à un concours national.

Solution de l'exercice 20

En faculté de **Lettres**, il y a **50%** de reçus, autant chez les filles que chez les garçons. En faculté des **Sciences**, il en est de même dans une proportion de **80%**. Nos deux doyens sont donc irréprochables. Cependant, pour l'ensemble de l'université **600** étudiants se sont présentés, **450** ont été reçus, ce qui fait un taux de réussite de **75%**. En revanche, pour les étudiantes, **500** se sont présentées et **280** ont été reçues, ce qui représente un taux de réussite de **56%**.
La plainte ne doit pas être dirigée contre les doyens, mais contre l'« effet de structure ».

Solution de l'exercice 21

Puisque le nombre de votants est égal à 75% du nombre d'inscrits, on obtient le nombre de votants en multipliant le nombre d'inscrits par 75 / 100 = 0,75. Le nombre d'inscrits s'obtient donc en divisant le nombre de votants par 0,75.

Notons N le nombre d'inscrits.

(75 / 100) × N = 3 390,

soit 0,75 N = 3 390,

d'où N = 3 390 /0,75 = 4 520.

4 520 personnes sont donc inscrites sur les listes électorales.

Solution de l'exercice 22

1. Soit t % le taux de TVA appliqué.

$\frac{t}{100} \times 70 = 13{,}72$

Par suite, $t = \frac{13{,}72 \times 100}{70} = 19\ 6.$

Le taux de TVA appliqué est donc de 19,6%.

On pouvait aussi déterminer ce taux en raisonnant avec une règle de trois.

2. Le prix HT des réparations en UM est donc de : 548,6 −28,6 = 520UM.

Soit t % le taux de TVA appliqué.

$\frac{t}{100} \times 520 = 28{,}6.$

Par suite : $t = \frac{28{,}6 \times 100}{520} = 5{,}5.$

Le taux de TVA appliqué est donc de 5,5 %.

Bibliographie

1. **ALALOUF, Serge, MENARD, Jean et LABELLE, Denis.** *Introduction à la statistique appliquée.* [éd.] Loze-Dion éditeur. Québec, 2002. p. 459. ISBN : 978-2-921180-71-9.

2. **ARMATTE, Michel.** *Manager dans l'incertitude : Pensée complexe et leadership.* [éd.] Presses des Mines. Paris, 2015. p. 344. ISBN : 978-2-911256-18-9.

3. **BERTRAND, Frédéric et MAUMY-BERTRAND, Myriam.** *Maxi fiches de Statistique pour les scientifiques.* [éd.] Dunod. Paris, 2011. p. 211. ISBN : 978-2-10-054483-7.

4. **BOULEAU, Nicolas.** *Probabilités de l'ingénieur : Variables aléatoires et simulation.* [éd.] Editions Hermann. 2. Paris, 2002. p. 383. ISBN : 978-2-7056-6439-8.

5. **CLEMENT, Benoît.** *Analyse de données en sciences expérimentales : Cours et exercices corrigés.* [éd.] Dunod. Paris, 2012. p. 182. ISBN : 978-2-10-057569-5.

6. **COUTY-FREDON, Françoise, DEBORD, Jean et FREDON, Fredon, Daniel.** *Mini Manuel de Probabilités et statistique : Cours et exercices corrigés.* [éd.] Dunod. 2. Paris, 2014. p. 250. ISBN : 978-2-10-070610-5.

7. **DRESS, François.** *Les probabilités et la statistique de A à Z : 500 définitions, formules et tests d'hypothèse.* [éd.] Dunod. Paris, 2007. p. 201. ISBN : 978-2-10-051403-8.

8. **EGON, Hubert et POREE, Pascal.** *Statistique et probabilités en production industrielle : I. Étude générale Problèmes et exercices corrigés.* [éd.] Editions Hermann. Paris, 2004. p. 325. ISBN : 978-2-7056-6454-1.

9. **FREDON, Daniel, MAUMY-BERTRAND, Myriam et BERTRAND, Frédéric.** *Mathématiques L1/L2 : Statistique et probabilités en 30 fiches.* [éd.] Dunod. Paris, 2009. p. 157. ISBN : 978-2-10-052345-0.

10. **GAUTHIER, Benoît.** *Recherche sociale : De la problématique à la collecte des données.* [éd.] Presses de l'Université du Québec. 5. Québec, 2008. p. 779. ISBN : 978-2-7605-1600-7.

11. **GIRARDIN, Valérie et LIMNIOS, Nikolaos.** *Probabilités en vue des applications : Introduction aux processus et à la statistique.* [éd.] Vuibert. Paris, 2008. p. 271. ISBN : 978-2-7117-2079-8.

12. **HURLIN, Christophe et MIGNON, Valérie.** *Statistique et probabilités en économie-gestion.* [éd.] Dunod. Paris, 2015. p. 370. ISBN : 978-2-10-072037-8.

13. **LECOUTRE, Jean-Pierre.** *Statistique et probabilités : Cours et exercices corrigés.* [éd.] Dunod. 5. Paris, 2012. p. 306. ISBN : 978-2-10-057890-0.

14. **LECOUTRE, Jean-Pierre.** *TD Statistique et probabilités.* [éd.] Dunod. 6. Paris, 2015. p. 202. ISBN : 978-2-10-072150-4.

15. **LETHIELLEUX, Maurice.** *Exercices de statistique et probabilités avec rappels de cours en 12 fiches.* [éd.] Dunod. 2. Paris, 2012. p. 155. ISBN : 978-2-10-057979-2.

16. **LETHIELLEUX, Maurice et CHEVALIER, Céline.** *Probabilités : Estimation statistique en 24 fiches.* [éd.] Dunod. 5. Paris, 2016. p. 156. ISBN : 978-2-10-074565-4.

17. **LIMNIOS, Nikolaos et GIRARDIN, Valérie.** *Probabilités en vue des applications : Variables, vecteurs et suites aléatoires.* [éd.] Vuibert. Paris, 2008. p. 239. ISBN : 978-2-7117-2078-1.

18. **NEDZELA, Michel.** *Modèles probabilistes d'aide à la décision.* [éd.] Presses de l'Université du Québec. Québec, 1987. p. 823. ISBN : 978-2-7605-0428-8.

19. **NOEL, Yvonnick.** *Psychologie statistique avec R.* [éd.] EDP Sciences. Paris, 2015. p. 325. ISBN : 978-2-7598-1736-8.

20. **PHAN, Thérèse et ROWENCZYK, Jean-Pierre.** *Exercices et problèmes de statistique et probabilités.* [éd.] Dunod. 2. Paris, 2012. p. 256. ISBN : 978-2-10-056298-5.

21. **PROTASSOV, Konstantin.** *Analyse statistique des données expérimentales.* [éd.] EDP Sciences. Paris, 2002. p. 148. ISBN : 978-2-86883-590-1.

22. **REISCHER, Corina, LEBLANC, Raymond et REMILLARD, Bruno.** *Théorie des probabilités : Problèmes et solutions.* [éd.] Presses de l'Université du Québec. Québec, 2002. p. 440. ISBN : 978-2-7605-1197-2.

23. **SAMUELIDES, Manuel.** *Les probabilités pour les sciences de l'ingénieur : Cours et exercices corrigés.* [éd.] Dunod. Paris, 2014. p. 353. ISBN : 978-2-10-059615-7.

24. **SOUVAY, Pierre.** *Les tables statistiques : Mode d'emploi.* [éd.] AFNOR. Paris, 2002. p. 50. ISBN : 978-2-12-505029-0.

25. **TALEB, Nassim Nicholas.** *Le hasard sauvage : Comment la chance nous trompe.* [éd.] Les Belles Lettres. Paris, 2008. p. 347. ISBN : 978-2-251-44371-3.

26. **TOUFFUT, Jean-Philippe et SOLOW, Robert.** *La Société du probable.* [éd.] Albin Michel. Paris, 2006. p. 229. ISBN : 978-2-226-17907-4.

27. **VEYSSEYRE, Renée.** *Aide-mémoire - Statistique et probabilités pour l'ingénieur.* [éd.] Dunod. 2. Paris, 2006. p. 475. ISBN : 978-2-10-049994-6.

28. **VEYSSEYRE, Renée.** *Aide-mémoire - Statistique et probabilités pour les ingénieurs.* [éd.] Dunod. 3. Paris, 2014. p. 390. ISBN : 978-2-10-071295-3.