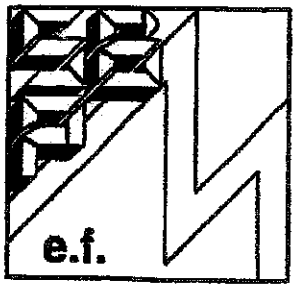
STATISTIQUES

EXPLORATIONS DIDACTIQUES



DOSSIER n° 6

STATISTIQUES

Première approche

Préparé par les membres de la Commission Pédagogique de la S.B.P.M.e.f.

Jacques Bair, Jacqueline Descy-Liesenborghs, Claudine Festraets, Mady Frémal, Roland Giot, Jean-Paul Houben, Pierre Marlier, Guy Noël, Yolande Noël, Frédéric Pourbaix, Philippe Tilleuil, Simone Trompler, Christian VanHooste.

SOCIETE BELGE DES PROFESSEURS DE MATHEMATIQUE
d'expression française
Rue de la Halle 15, B-7000 MONS (Belgique)
Tél-Fax : 32/(0)65/37.37.29
1999

STATISTIQUES

Ch. 3 : Des techniques d'analyse des données p. 33

Mots-clefs

Les nombres indiqués sont des numéros de page.
Le "p" qui suit les références indique que le mot apparaît plusieurs fois sur la même page.
L'indication "et sv." signifie que le(s) mot(s) dont question interviennent dans un document introduit par scannage dans notre texte.

STATISTIQUES

Introduction

Les chiffres ne peuvent mentir.
Il y a les mensonges, il y a de fieffés mensonges, et il y a les statistiques.

disent les gens.

Ulysse Dupont, citoyen moyen, le devoir accompli, quitte son entreprise pour rentrer chez lui. Comme il a presté un bout de temps supplémentaire, la porte de l'immeuble est fermée, et pour l'ouvrir il compose un code sur le clavier adéquat. Pour faire démarrer sa voiture ou y faire fonctionner sa radio, même opération. Rebelote au distributeur de billets ou à la caisse du supermarché où il s'est arrêté pour quelque emplette.

En route, il a appris par le bulletin d'informations de la radio que la température moyenne du globe a évolué de telle manière au cours de la dernière décennie, à moins peut-être que l'information n'ait porté sur les résultats quantifiés d'une enquête sur le comportement sexuel des européens et le niveau de satisfaction qu'ils en retirent, par pays et par tranches d'âges.

Le soir, à la télévision, il voit s'affronter des représentants de partis politiques qui, au départ de statistiques toutes plus sérieuses et fondées les unes que les autres, défendent des thèses contradictoires. Fort heureusement, cet aride débat est interrompu par une publicité dans laquelle une charmante présentatrice lui apprend que neuf personnes sur dix se félicitent de consommer tel produit, ce qui sous-entend qu'il serait débile de ne pas adopter le même comportement.

Abruti par tous ces chiffres, Ulysse Dupont fait taire les sources d'information extérieures, et armé de son stylo à bille, il se concentre sur sa grille de Lotto.

* * *

Une des difficultés qu'on rencontre à mettre en place un enseignement de la statistique dans le secondaire, est cette double conviction qu'ont les gens, que d'une part les nombres n'ont pas d'états d'âme et sont donc objectifs, et que par ailleurs on leur fait dire n'importe quoi. Ajoutez-y l'idée simpliste mais bien ancrée que la statistique, c'est des moyennes, et que le calcul des moyennes, on connaît cela depuis l'école primaire et vous aurez les ingrédients nécessaires à un désintérêt massif. S'il reste quelques personnes qui hésitent et pensent que quand même... , attirez leur attention sur le fait que n'importe quel tableur ou même des calculatrices pas trop onéreuses font très bien le travail à votre place et il ne restera plus que quelques dinosaures pour penser que l'enseignement de la statistique[1] a sa place dans l'enseignement commun et contribue à la formation du citoyen.

[1] En tout cas tel qu'il est souvent pratiqué, c'est-à-dire calcul de moyennes peu significatives et tracés de graphiques.

Parmi les principes ou postulats qui sous-tendent ce document, il y a le fait que la statistique est un moyen d'approche et d'analyse valable du réel. Bien plus, il y a un certain nombre de domaines ou de situations de grande, voire très grande importance sociale ou individuelle, qui, dans un premier temps en tout cas, ne peuvent être valablement étudiées que par des méthodes statistiques.

Qu'on pense à la décision que doivent prendre les industriels ou les artisans d'accepter ou de refuser, sur la base de l'analyse d'un échantillonnage, un lot de composants que leur présente un fournisseur. Qu'on pense encore aux analyses de marché à partir desquelles des responsables décident de telle ou telle stratégie commerciale. Il n'est pas si rare que de telles décisions impliquent des sommes d'argent considérables.

Dans le domaine de la santé aussi, les études épidémiologiques sont souvent une première étape dans l'étude d'un phénomène. Des observations statistiquement élaborées peuvent commencer des recherches sur la nocivité d'une substance ou sur l'efficacité d'une molécule dont on veut faire un médicament ou être, en cours ou en fin d'expérimentation, un contrôle de la valeur de ce qui a été entrepris.

Les travaux des sociologues font abondamment appel aux méthodes statistiques. N'oublions d'ailleurs pas que le mot même de "statistique" fait référence à la notion d'Etat. Le dictionnaire Littré en donne comme première définition : "*Science qui a pour but de faire connaître l'étendue, la population, les ressources agricoles et industrielles d'un Etat*". Cette définition pas très extensive pourrait assez facilement être récrite pour donner quelque chose comme :
La statistique est la science qui s'occupe de la gestion rationnelle des données quantifiées relatives à l'Etat ou à toute autre situation qui lui est plus ou moins comparable.

* * *

On aura compris que si la statistique a son efficacité pour traiter de situations comportant un grand nombre de données, elle ne le fait qu'en gommant les singularités individuelles, en en faisant abstraction. **Il n'y a donc guère de sens à donner des explications "statistiques" relatives à des petits ensembles**, comme par exemple une classe de quatre ou cinq élèves.

Des raccourcis de langage habituels peuvent induire en erreur. Si un joueur de tennis vient de gagner par jeu blanc son jeu de service en réussissant trois aces et que le commentateur de la radio ou de la télévision dit que 75 % des services de ce joueur sont des aces, il donne à penser que ce joueur réussit habituellement un tel exploit, ce qui n'est très vraisemblablement pas le cas.

Inversement, si on traduit par "*dans le monde, un enfant sur trois est chinois*" le fait que plus de 30 % des enfants du monde sont chinois, ceci ne devrait pas inquiéter les parents européens de deux enfants qui désirent agrandir leur petite famille.

* * *

Mais si la conviction de la validité de l'outil statistique est une des idées de base de la commission pédagogique de la SBPMef, une autre conviction tout aussi fondamentale est qu'en ce sujet comme en tant d'autres, on n'arrive à des résultats intéressants qu'en y mettant le prix. Une étude superficielle ne donnera généralement que des résultats sans intérêt, à la limite

du ridicule ou même franchement idiots, même si l'utilisation d'un équipement informatique fait croire que le travail est très savant.

Un des auteurs de ce texte s'est un jour amusé à calculer la moyenne d'âge des élèves de rhétorique d'un certain nombre d'écoles. Ayant obtenu les listes d'élèves avec la date de naissance et le cours de mathématique suivi (3, 5 ou 7 h/sem pour le général,...), il disposait d'un échantillonnage suffisant pour échapper au reproche de tirer des conclusions générales à partir d'une population trop restreinte. Ces données ont été soigneusement encodées (on peut raisonnablement supposer que les éventuelles erreurs d'encodage ont été très rares et n'ont donc pas influencé le résultat) et les calculs ont été confiés à un ordinateur[2].

Ce qui est ressorti de ce travail scientifique est ce qu'on pouvait bien imaginer : en moyenne, les élèves ayant suivi le cours de mathématique à 7 h/sem étaient significativement plus jeunes que ceux des cours de mathématique à 5h/sem, eux-mêmes significativement plus jeunes que ceux qui avaient suivi un cours de mathématique à 3 h/sem. Autrement dit, **plus le cours de mathématique est poussé, moins il y a de retards scolaires**. Conclusion : pour lutter contre les retards scolaires, il suffirait d'augmenter le nombre d'heures de mathématique par semaine.

La mauvaise foi de cette argumentation attire l'attention sur le fait qu'une analyse (statistique ou non) de situation se fait toujours **dans un contexte** et **avec un projet** : sur un sujet donné qui est débattu dans la société ou auquel on s'intéresse personnellement, on essaie d'y voir plus clair, pour se faire une opinion mieux établie, pour faire le point sur des idées reçues, pour argumenter ou contredire des prises de position, parce qu'on a une décision à prendre,...

Ceci implique qu'à propos d'un problème qu'on rencontre, on se pose au moins deux questions :
1. qu'est-ce que je veux savoir sur la situation ?
2. quels sont les indicateurs observables qui me donneront des renseignements fiables sur ce que je veux savoir ?

Ces questions (et quelques autres comme : *où vais-je trouver l'information que je cherche ?*) ne relèvent pas au sens strict de la technique statistique, mais il est indispensable de se les poser et d'y répondre si on veut éviter de se lancer tête baissée dans des travaux qui après coup se révéleront vains.

Dans le cas des retards scolaires, si l'indicateur utilisé (l'âge moyen d'un groupe d'élèves) peut être utile pour se faire une idée correcte de l'importance des retards scolaires, il est tout à fait inadéquat à donner des indications sur la cause de ces retards ou sur les remèdes à apporter.

---

[2] Quelques détails sur le calcul de la moyenne d'âge d'un groupe de personnes, pour faire apparaître que l'erreur qu'on détectera ne provient pas de la méthode de calcul :
- on écrit une routine qui compte le nombre de jours écoulés depuis une date de référence (par exemple, le 1 janvier 1900) jusqu'à la date qu'on indique. Même en tenant compte des années bissextiles, ceci n'est pas trop difficile à faire.
- pour savoir le nombre de jours vécus par un individu depuis sa naissance jusqu'à la date du jour (ou une date qu'on indique), il suffit de faire la différence entre le nombre de jours écoulés depuis la référence jusqu'à la date du jour et le nombre de jours écoulés depuis la référence jusqu'à la date de naissance de l'individu. L'âge de chaque individu est donc ainsi exprimé en jours.
- on peut donc calculer sans problème la moyenne de tous ces âges. Pour que la lecture soit plus facile, on retransforme en années et mois cet âge moyen exprimé en jours, en admettant par exemple qu'un mois comporte 30,5 jours, ce qui donnera un résultat exprimé avec une précision plus que suffisante.

Selon une enquête[i] dont les résultats ont été publiés en novembre 1996 et qui concerne le premier degré du secondaire, si nos élèves (enseignement belge francophone, tous réseaux confondus) sont supérieurs à la moyenne en mathématique, ils sont tout en bas de classement en sciences. Certains ont cru opportun d'en tirer des conclusions sur nos méthodes pédagogiques de l'enseignement des sciences.

On peut penser que de telles remarques sont complètement hors de propos. Car s'il est peu probable que nos élèves qui savent faire des maths soient incapables d'apprendre les sciences, il est tout aussi peu probable que nos professeurs de sciences soient systématiquement moins capables que leurs collègues de mathématique. Une autre explication serait que les responsables de l'éducation en Belgique ont préféré concentrer la formation de base sur des sujets autres que les sciences, en reportant aux années ultérieures une formation scientifique proprement dite. Effectivement, en Belgique francophone, l'enseignement des sciences à l'école primaire et au début du secondaire, fait partie des activités d'éveil. Il n'y a donc pas lieu de s'étonner que nos élèves du début du secondaire ne puissent guère répondre à des questions auxquelles ils n'ont pas été préparés. Autrement dit : l'enquête TIMSS, telle que pratiquée chez nous, NE serait PAS un bon indicateur de la qualité de l'enseignement des sciences en Belgique francophone. Elle ferait seulement apparaître que notre politique pédagogique a des spécificités, ce qui est fort choquant pour le Belge moyen volontiers adepte de la très tyrannique religion du "juste (!) milieu".

Pour savoir si ces choix de politique éducative sont de bons choix, il aurait fallu que la Communauté Française souscrive à la totalité de l'enquête TIMSS, c'est-à-dire participe aussi à l'évaluation des élèves de fin du secondaire, ce qui n'a pas été fait pour les raisons budgétaires que l'on imagine. A défaut de tels renseignements, il faut, si on veut dire des choses pertinentes sur l'enseignement des sciences dans notre système scolaire, rechercher d'autres indicateurs appropriés et plus accessibles.

* * *

En guise de mise en train...

S'il était bon d'appâter les professeurs par des exemples qui les intéressent, il n'est sans doute pas inutile d'amorcer les élèves par des questions qui retiennent au plus haut point leur attention.

A propos, par exemple, de la "cote[3] du bulletin" :
- Comment l'établir ? En faisant la moyenne de tous les travaux ? En faisant la moyenne de certains travaux ?
- Comment rendre compte de l'activité en classe ?
- Faut-il tenir compte des travaux à domicile ?

[3] Le mot habituellement employé en France pour parler de l'évaluation des élèves est "note" et non "cote". Cependant, le Larousse (Petit Larousse Illustré, 1980) reconnaît au mot "cote", le sens (dans un usage figuré et familier) de "degré d'estime : *avoir une bonne cote*". D'autres dictionnaires confirment ce sens d'*évaluation, d'estimation* attribué au mot *cote*, éventuellement avec renvoi au mot *note* où on trouve un renvoi au mot *cote*. En Belgique, à une écrasante majorité, les professeurs parlent de leur *cahier de cotes*. On ne voit pas bien pourquoi on s'alignerait sur l'usage hexagonal.

- Que représente cette cote pour l'élève ? Pour ses parents ?
- Un élève peut-il s'organiser pour améliorer sa cote tout en n'ayant rien appris ?
- Est-il plus important d'avoir acquis une compétence ou d'avoir une bonne cote du mois ?

- Quel profit retire-t-on d'un devoir recopié sur celui d'un bon élève de la classe ?
- Comment faire en sorte que l'évaluation périodique soit le reflet le plus fidèle possible des progrès réalisés ?
- Comment utiliser les évaluations pour progresser ?
- ...

Ces questions n'épuisent pas le sujet[ii]. Le professeur qui les pose veillera à maintenir un minimum de sérénité pour que la leçon ne tourne pas à l'affrontement. L'évaluation est une démarche omniprésente dans la vie professionnelle, que l'on se trouve dans la position de celui qui est évalué ou de celui qui évalue. Un des objectifs d'une telle séquence est de provoquer une réflexion plus adulte chez les élèves. Peut-être auront-ils appris qu'en confrontant ses idées avec d'autres et en étant critique, on progresse davantage dans la maîtrise d'un sujet qu'en assénant des "vérités" à des "adversaires".

Ch. 1 : Valeurs centrales

Tout le monde sait ce qu'est une moyenne.
Vraiment ?

A. Situations

Sit 1.1 - Moyenne, vous avez dit moyenne ? ...

On a souvent dit, dans les media, que l'enseignement, en Communauté Française de Belgique, coûtait trop cher. On a chanté sur tous les toits qu'il y avait, en moyenne, 7 élèves par professeur.

Considérant ses listes d'élèves, M. PROFDEMAT est perplexe : il a en charge :

- une classe de 25 élèves pendant 5 périodes/semaine;
- une classe de 24 élèves pendant 4 périodes/semaine;
- une classe de 27 élèves pendant 5 périodes/semaine;
- une classe de 17 élèves pendant 6 périodes/semaine;
- une classe de 3 élèves pendant 2 périodes/semaine;

ce qui donne un total de 96 élèves dont il a la charge pendant 22 périodes/semaine.
Combien d'élèves a-t-il "en moyenne" devant lui par période ?

SOLUTION 1[4]
96 élèves pendant 22 périodes/semaine. Donc $96/22 \approx \mathbf{4}$.

SOLUTION 2
96 élèves répartis en 5 classes. Donc $96/5 \approx \mathbf{19}$.

SOLUTION 3
25 élèves pendant 5 périodes, ...
Donc $\dfrac{25 \times 5 + 24 \times 4 + 27 \times 5 + 17 \times 6 + 3 \times 2}{5 + 4 + 5 + 6 + 2} \approx \mathbf{21}$

SOLUTION 4
Cette dernière valeur ne me satisfait pas, se dit M. PROFDEMAT, car je constate que pendant 8 périodes, j'ai devant moi moins de 21 élèves, mais que pendant 14 périodes, je m'en occupe de plus de 21, et même plus de 24. Je considère donc que le seul nombre traduisant ma réalité quotidienne est **24** car il y a 11 périodes pendant lesquelles j'ai devant moi au plus 24 élèves et 11 périodes pendant lesquelles ils sont au moins 24.

SOLUTION 5

[4] Plusieurs des solutions présentées ont été formulées par des élèves à qui la situation avait été proposée.

Le chef d'établissement n'est pas du tout d'accord. Pour lui, M. PROFDEMAT a en moyenne **10** élèves : son établissement compte 600 élèves et 60 professeurs.

SOLUTION 6

Le ministre de l'enseignement de la Communauté Française pourrait, lui aussi, avoir un autre avis : les 96 élèves de M. PROFDEMAT ont, de par leurs options différentes, 16 professeurs qui les prennent en charge.

Donc, la moyenne par professeur est de $96/16 \approx$ **6** élèves.

Le Soir 4/1/95.

Trop cher, notre enseignement ? Comparons...

A en croire les ministres Mahoux et Lebrun, secondés par les professeurs Deschamps et Martou, l'enseignement belge souffrirait d'une effroyable boulimie financière. Cette affirmation s'appuye, nous dit-on, sur les comparaisons internationales de l'OCDE. Devant les arguments chiffrés des autorités politiques et scientifiques, le citoyen n'a plus qu'à se taire... ou à étudier à son tour les documents de l'OCDE (1).

La Belgique consacre 5,4 % de son PIB à l'ensignement. L'ensemble des pays de l'OCDE dépense 5,2 %. Nous sommes donc à peine au-dessus de la moyenne. Les spécialistes de l'antifinancement préfèrent dès lors comparer notre pays à « la moyenne de ses voisins directs », qui est de 5,0 %.

Mais voyons le détail : France 5,4 %, Grande-Bretagne 5,3 %, Pays-Bas 5,6 %. Le pays qui à lui seul fait chuter la moyenne, c'est l'Allemagne avec 4 %. On omet cependant de nous préciser que l'enseignement allemand jouit d'un important financement privé, provenant directement des entreprises, et qui s'élève à 1,5 % du PIB. Black-out également sur les 6,7 % du Canada ou les 6,8 % de la Norvège.

Un coût par élève pourtant très bas

Voilà pour ce que coûte l'enseignement à la nation. Si on va voir ce que la Belgique dépense par élève-étudiant, nous tombons carrément loin en-dessous de la moyenne OCDE. L'Etat belge ne dépense que 4.003 $ par élève contre 4.716 $ pour l'ensemble de l'OCDE (il s'agit de dollars « parité de pouvoir d'achat » ou « PPA », qui tiennent compte de la différence du coût de la vie). En clair : les moyens matériels que notre pays consacre à l'éducation de chaque enfant sont inférieurs de 15 % à ceux des autres pays de l'OCDE.

Si le coût par élève est si bas, alors comment expliquer que le coût global soit assez élevé ? La réponse est d'une logique mathématique : il y a, proportionnellement, davantage d'élèves dans notre pays. Pourquoi ?

Au cours des années 50-70, en pleine croissance économique, le patronat réclamait une main-d'œuvre de plus en plus nombreuse et de plus en plus qualifiée. Pour faire face à la quantité, on a invité les femmes à travailler, donc à mettre les jeunes enfants à l'école maternelle. Pour faire face à la qualité, on a invité filles et garçons à poursuivre leurs études le plus longtemps possible. Afin de les y aider, on a introduit l'enseignement rénové. L'école secondaire s'est ouverte aux enfants du peuple. Cette démocratisation de l'enseignement a eu lieu dans tous les pays, mais elle a rarement été conduite d'une façon aussi déterminée que chez nous.

Actuellement, 35 % des jeunes Belges de 20 ans suivent l'une ou l'autre forme d'enseignement supérieur, alors que la moyenne OCDE n'est que de 19,7 % ! Une deuxième cause du nombre élevé d'élèves, c'est le taux de redoublement très élevé en Belgique. Ce point est souvent avancé pour souligner « l'inefficacité » de notre enseignement. Mais n'oublions pas que le redoublement a deux aspects : un aspect négatif (l'échec) et un aspect positif (le fait d'avoir une seconde chance).

Dans la comparaison avec les autres pays de l'OCDE, l'aspect positif est le plus important. En effet, ces autres pays pratiquent généralement un système de sélection beaucoup plus draconien, où l'échec scolaire est immédiatement sanctionné par une « réorientation » vers l'enseignement professionnel, par exemple.

Deux choix, des choix de société

Bref, notre enseignement belge est finalement assez démocratique comparé à d'autres pays. C'est la défense de cet acquis-là qui est au centre du débat sur le refinancement. Deux choix s'opposent, deux choix de société. D'un côté l'option de plier sous la double pression des « contraintes budgétaires » et de « l'adéquation de l'enseignement aux besoins des entreprises ». Elle revient à exclure d'emblée de l'école 20 % de jeunes qui n'auront jamais besoin de qualifications (chômeurs et petits boulots) et de hiérarchiser fortement la formation des autres. Les moyens stratégiques pour réaliser cette option sont l'autonomie des écoles, les regroupements (économies d'échelle) et la sélection renforcée.

L'autre option, c'est celle de l'école démocratique. Cela veut dire, penser l'école comme un instrument apportant à tous les jeunes le maximum de connaissances et de savoir-faire; un instrument qui apprenne à chacun à réfléchir avec sa propre tête, à comprendre les enjeux politiques, technologiques et sociaux de notre monde. Ici, le moyen stratégique, c'est de refinancer l'enseignement d'une centaine de milliards, en puisant par exemple dans les 7.000 milliards du patrimoine détenus par 1 % de la population. Ou dans les 700 milliards de bénéfices des entreprises.

Utopique ? S'il est utopique de faire passer l'instruction de la jeunesse avant l'intérêt d'une poignée de nantis, alors il est peut-être temps de changer de société, non ?

GEORGES MOREAU

Responsable du secteur enseignement au Parti du Travail de Belgique

(1) Source : « Regards sur l'éducation », les indicateurs de l'OCDE, Paris 1993.

Les intertitres sont de la rédaction.

Sit 1.2 - La meilleure classe

M. PROFDEMAT souhaite comparer les performances de deux classes parallèles A et B. Pour ce faire, il a calculé la moyenne des cotes des deux classes et a trouvé 11,25 / 20 pour A et 9,75 / 20 pour B. La réponse à la question de savoir quelle est la meilleure classe semble claire.

Cependant, en continuant de consulter son cahier de cotes, il constate qu'il a la situation suivante :
Classe A : 15 élèves ont 9/20, 5 élèves ont 18/20. La moyenne est donc bien de 11,25 mais les 3/4 des élèves sont en échec.
Classe B : 15 élèves ont 12/20, 5 élèves ont 3/20. Donc une moyenne de 9,75, mais les 3/4 des élèves ont une cote satisfaisante de 12.

Questions :
- Etes-vous toujours convaincu que la classe A soit la meilleure ? - Peut-on être certain que les élèves qui ont 18/20 en A soient meilleurs que les élèves qui ont 12/20 en B ?
- La réponse à la question qui précède dépend-elle du fait que ce soit le même professeur dans les deux classes ou qu'il s'agisse de deux professeurs différents ?
- Peut-on à partir de telles données faire des pronostics sur les succès ultérieurs des élèves ?-

Etc.

Sit 1.3 - Combien vais-je gagner ?

Isidore, répondant à une offre d'emploi, s'informe sur le montant de ses rémunérations.

Dans notre petite entreprise de vente en porte à porte, lui répond-on, le salaire moyen est de 1500 euros/mois[5]. Bien entendu, au début, le temps pour vous de vous former, vous ne gagnerez guère que 400 euros, mais croyez bien que ce ne sera que momentané.
Après quelques temps, Isidore commence à s'inquiéter car il a interrogé les autres démarcheurs qui lui ont avoué ne gagner que 500 euros. Le patron a-t-il menti à Isidore?

Ce dernier a par ailleurs mené sa petite enquête qui lui a permis d'apprendre les choses suivantes :

Salaire du patron :	12 000 euros	
Salaire du sous-directeur :	5 000 euros	
Salaire des acheteurs :	1 250 euros.	Ils sont 6.
Salaire des chefs de secteur :	1 000 euros.	Ils sont 5.
Salaire des autres :	500 euros.	Ils sont 10.

Salaire moyen de ces 23 personnes :

$$\frac{12\,000 + 5000 + 1250 \times 6 + 1000 \times 5 + 500 \times 10}{23} = 1500$$

Se disant qu'il s'est fait posséder avec cette idée de moyenne, Isidore s'interroge : quelle information sur les salaires aurait-il dû demander pour éviter de se faire piéger. S'inspirant des ré-

[5] Un euro vaut 40,3399 francs belges.

flexions de M. PROFDEMAT (solution 4 du §1), il se dit qu'il y a dans l'entreprise 12 personnes qui gagnent au plus 1000 euros/mois et 11 personnes qui gagnent plus. Le salaire médian de ces 23 personnes est donc de 1000 euros/mois. C'est déjà mieux que le salaire moyen, mais c'est encore assez éloigné de la réalité.

Continuant de réfléchir, Isidore se dit que s'il avait demandé ce que "la plupart des gens" gagnent, il aurait probablement eu une réponse adéquate. A moins, pense-t-il encore, qu'on ait interprété de travers l'expression "la plupart des gens". Aussi reformule-t-il la question qu'il poserait si c'était à refaire : quel est le salaire mensuel de la catégorie du personnel où il y a le plus grand nombre de personnes ?
Cette valeur est le mode de la série statistique.

Sit 1.4 - Une autre valeur centrale.

Situation 4a

Une action est émise au prix de 100 euros. Un an plus tard, elle se négocie à 180 euros, et après la deuxième année, à 162 euros.
Quel en est le coefficient d'accroissement annuel moyen ?

SOLUTION 1
De 1 à 2, le taux d'augmentation a été de 80%;
de 2 à 3, il y a eu une diminution de 10%, soit un taux de -10%.
Moyenne annuelle : (80 - 10)/2 = 35. Donc 35% .
Vérification : 100 + 35% de 100 = 135. 135 + 35% de 135 = 182,25.
On devrait avoir 162 !

SOLUTION 2
En deux ans l'accroissement a été de 62%.
Donc moyenne annuelle : 31% ?
Vérification : 100 + 31% de 100 = 131. 131 + 31% de 131 = 171,61...
Autrement dit : 100×1,31×1,31 = 171,61... Encore raté !

SOLUTION 3
De 1 à 2, le coefficient est de 1,8 car 100×1,8 = 180.
De 2 à 3, le coefficient est de 0,9 car 180×0,9 = 162.
Moyenne de 1,8 et 0,9 : 1,35.
Vérification : voir 1ère solution.

SOLUTION 4
Les coefficients annuels sont 1,8 et 0,9 et leur produit est 1,62. Ceci indique qu'on est dans une structure multiplicative et non additive. Le problème est donc de trouver deux nombres égaux (un seul nombre, donc en fait) dont le produit (le carré) est 1,62. Ce nombre ne peut être que $\sqrt{1,62}$. Il s'agit de la moyenne géométrique des deux nombres donnés (Voir B : Définitions et formules).

Situation 4b

On reprend les données de Sit 4a et on en ajoute quelques-unes :
Prix à l'émission : 100 euros; valeur après 1 an, 180 euros; valeur après 2 ans, 162 euros, valeur après 3 ans, 243 euros; etc.
Quel est le taux d'accroissement annuel moyen ?

Sit 1.5 - Encore une autre valeur centrale

Situation 5a : un problème classique

La vitesse moyenne d'un mobile parcourant un espace donné en un temps donné est la vitesse constante qui permettrait de parcourir la même distance dans le même laps de temps.
Un TGV roule pendant 1 heure à la vitesse (moyenne, ou supposée constante) de 100 km/h, puis pendant 1 heure à la vitesse de 200 km/h . Quelle distance a-t-il parcourue ? Quelle est sa vitesse moyenne sur ce parcours ?
Et s'il roule 1h à 100 km/h puis 2h à 200 km/h ?
Et s'il roule 3h à 100 km/h puis 2h à 200 km/h et enfin 1h à 300 km/h ?

Situation 5b : une variante un peu différente

Un TGV va de A vers B à la vitesse moyenne de 100 km/h et revient à la vitesse de 200 km/h. Quelle est sa vitesse moyenne sur le trajet ABA (on a arrêté le chrono pendant un éventuel arrêt en B) ?

SOLUTION 1
Moyenne arithmétique de 100 km/h et de 200 km/h : 150 km/h.
Vérification : supposons que la distance |AB| soit de 100 km. Le trajet [AB] aura duré 1 heure ou 60 minutes et le trajet [BA] 1/2 heure ou 30 minutes; si les deux trajets avaient été effectués à vitesse constante, chacun aurait duré 45 minutes (soit ((60 + 30)/2) ce qui donne une vitesse moyenne de 100/(3/4) km/h, soit 133,33... km/h. On avait trouvé 150. ... Caramba !
Est-ce la distance supposée de 100 km qui est l'élément parasite qui introduit l'erreur ? Pour le savoir, réessayer le calcul avec une distance de 200 ou de 300 km, par exemple. La distance a-t-elle de l'importance ?
Par quelle formule trouver la bonne réponse (133/33...) à partir des nombres 100 et 200 ?

SOLUTION 2
La méthode utilisée pour la vérification et le constat que la distance n'a pas d'importance nous fournissent une autre approche.
La durée du trajet aller est de |AB|/100 et celle du retour est de |AB|/200.
Comme dans la vérification, la durée d'un trajet parcouru à la vitesse moyenne recherchée est la moyenne des durées des trajets. Donc,

$$\text{Durée d'un trajet à vit. moyenne} = \frac{1}{2}\left(\frac{|AB|}{100} + \frac{|AB|}{200}\right).$$

Si on appelle x cette vitesse moyenne recherchée, la durée d'un trajet à la vitesse x est donnée par la formule |AB|/x . La formule encadrée ci-dessus devient donc

$$\frac{|AB|}{x} = \frac{1}{2}\left(\frac{|AB|}{100} + \frac{|AB|}{200}\right).$$

Après simplification par |AB| (on revoit autrement que la distance n'a pas d'importance) et

transformation élémentaire, on a :

$$x = \frac{1}{\frac{1}{2}\left(\frac{1}{100} + \frac{1}{200}\right)}.$$

La valeur ainsi trouvée porte le nom de moyenne harmonique des deux nombres donnés (Voir B : Définitions et formules).

B. Définitions et formules

Quelques définitions générales et symboles avant d'écrire les formules :
- **population** : ensemble des individus (personnes ou objets) pour lesquels une (des) caractéristique(s) (numérique(s) on non) a(ont) été relevée(s).
- **effectif (n)** : nombre d'individus de la population.
- si plusieurs individus sont caractérisés par une même valeur ou sont regroupés dans une même classe (caractérisée par sa valeur centrale, par exemple), la **répétition (n_i)** est le nombre d'individus caractérisés par la valeur x_i ou regroupés dans la i-ème classe. Il n'est pas rare que **k** représente le nombre de classes ou de valeurs différentes.
- la **fréquence (f_i)** d'une valeur ou d'une classe est sa répétition divisée par l'effectif. En symboles : $f_i = n_i/n$.

1. Le mode

Le **mode** (M_0) est la valeur de la classe la plus peuplée. S'il s'agit d'un caractère discret, le mode est la valeur de la variable correspondant à cette classe, et dans le cas d'un caractère continu, c'est le centre de cette classe.

A proprement parler, le **mode** ne se calcule pas : on le repère sur un tableau groupé ou à partir d'un diagramme en bâtonnets, par exemple. Ce n'est pas nécessairement une valeur centrale. Dans une gamme d'articles d'un certain type, il se peut que le produit le moins cher soit le plus vendu. Le prix de ce produit sera le mode de la série des prix de cette gamme d'articles.

Dans la situation 1.3, on a vu que c'est le mode qui donne la meilleure réponse à la question posée. Quel est le mode de l'ensemble des classes de M. PROFDEMAT de notre situation 1.1 ?

2. La médiane (M)

On verra plus loin, dans un cadre plus général (Ch.2, quartiles, déciles, centiles), une autre formulation de la définition de la médiane. En termes intuitifs, on dit généralement que la **médiane** divise la population en deux sous-ensembles de même cardinal. Encore faut-il que l'effectif soit pair ! Autre difficulté : dans ce cas de parité de l'effectif, il existe une infinité de valeurs qui partagent la population en deux.

Si ValPop1 = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, tous les points de l'intervalle]5,6[divisent la population en deux ! Quelle est donc la médiane ?

Aussi retient-on plus souvent une définition opératoire comme la suivante :

Si $n = 2p$, la **médiane** est la moyenne arithmétique des p-ième et (p+1)-ième valeurs.
Si $n = 2p+1$, la **médiane** est la (p+1)-ième valeur.

Dans l'exemple ci-dessus, la médiane de ValPop1 est 5,5; si ValPop2 = {1, 2, 3, 4, 5, 6, 7, 8, 9}, sa médiane est 5.

A proprement parler, la **médiane** ne se calcule pas non plus : elle se repère à partir d'un tableau ordonné ou d'un diagramme d'effectifs cumulés (voir plus loin).

La notion de **médiane** est apparue dans la solution 4 de la situation 1.1 et dans les réflexions d'Isidore (situation 1.3). Quelle est la **médiane** de chacune des classes de M. PROFDEMAT dans la situation 1.2 ?

3. Moyenne arithmétique (m)

L'idée sous-jacente à la notion de moyenne (arithmétique ou autre) est de simplifier une situation donnée par n valeurs en remplaçant ces n valeurs différentes par une valeur unique, tout en maintenant le même effet global.

Dans le cas de la **moyenne arithmétique**, l'effet global que l'on veut garder invariant est la somme des valeurs. On gomme donc les singularités individuelles[6] en remplaçant chaque valeur par la somme des valeurs divisée par l'effectif. L'élève qui a obtenu les cotes suivantes (sur 10) : {2,4,8,2,7,4,8} a le même total de 35 que s'il avait eu les cotes {5,5,5,5,5,5,5}.

Cette définition est consignée par la formule de l'encadré.

moyenne arithmétique :

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Dans le cas, où il y a seulement k valeurs x_i qui se répètent respectivement n_i fois, et que donc ces valeurs ont une fréquence $f_i = n_i/n$, cette formule devient :

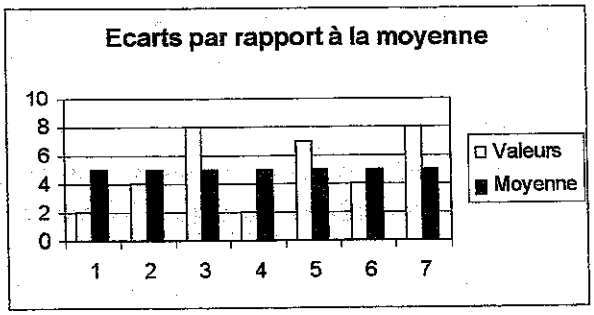
$$m = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \frac{n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$$

On ne manquera pas d'observer que l'indice placé au-dessus du signe de sommation est "n" dans le premier encadré cas et "k" dans l'autre.

La **moyenne arithmétique** (**m**) est une valeur centrale de la population considérée, en ce sens que la somme (algébrique) des écarts par rapport à cette moyenne est nulle. Ceci est

[6] Voir introduction, page 3.

une conséquence immédiate de l'invariant de la substitution des valeurs : si on remplace quelque part 2 par 5, la somme des valeurs a augmenté de 3; pour conserver la somme, il faut diminuer de 3 unités en tout une ou plusieurs autres valeurs. Cette propriété est clairement illustrée par le graphique.



4. Moyenne géométrique et moyenne harmonique.

Ces deux notions ne font normalement pas partie d'un enseignement élémentaire de la statistique. Parce qu'on peut les rencontrer (Situations 1.4 et 1.5), on peut juger qu'il est bon de savoir qu'il existe d'autres valeurs centrales que la moyenne arithmétique et la médiane. Sauf exception, quand on parle de moyenne (sans autre qualification), c'est toujours de moyenne arithmétique qu'il s'agit.

Dans les situations où c'est la structure multiplicative des valeurs qui est importante, l'effet global qu'on veut conserver est le produit des valeurs. La **moyenne géométrique**, c'est-à-dire le nombre qui va remplacer toutes les valeurs sera bien sûr la racine n-ième du produit des valeurs. D'où la formule :

$$\text{moyenne géométrique} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}, \quad (x_i \geq 0)$$

La moyenne **harmonique** est d'application dans les situations où la structure du problème est en 1/x, c'est-à-dire où la variable liée est inversement proportionnelle à la variable libre. L'effet global à conserver est ici plus difficile à exprimer. On se contentera donc de la formule :

$$\text{moyenne harmonique} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)}$$

Remarque

On a dit dans l'introduction que l'approche statistique d'une situation se fait toujours dans un contexte et avec un projet. Ce sont ces éléments, non la technique statistique, qui font juger qu'une des notions rencontrées (mode, médiane, moyenne) est la plus adaptée à l'analyse qu'on veut faire de la situation.

Ainsi par exemple, dans la situation 1.1, le contexte est le fait que les pouvoirs publics ont, pour des raisons budgétaires, diminué l'encadrement des élèves et de ce fait alourdi la tâche des professeurs en arguant par ailleurs que l'encadrement belge des élèves était tellement favorable que quelques restrictions ne pourraient nuire à la qualité de l'enseignement; le projet est de donner une estimation valable de la lourdeur de la tâche des professeurs en réfutant des argumentations spécieuses. L'analyse proposée ne tranche pas la question de savoir si c'est la moyenne (solution 3) ou la médiane (solution 4) qui est le meilleur indicateur du poids de la tâche des professeurs.

C. Pêle-mêle, des Ex[7]

Ex 1.1 - Le père de Sébastien n'est pas content du résultat du dernier contrôle de son rejeton. Mais argumente ce dernier, j'ai tout de même une moyenne de 12. Sachant que Sébastien a passé 5 contrôles et que les résultats des 4 premiers ont été 13, 10, 16, 14, quelle est la cote du dernier travail ?

Ex 1.2 - Fabienne rentre chez elle avec une copie cotée 11/20. Comme elle a d'habitude de meilleurs résultats, ses parents demandent des explications. Fabienne argue de la difficulté des questions et de la sévérité de la correction du professeur; ses parents restent sceptiques. Pour les convaincre, elle ajoute : avec ce 11/20, je suis encore dans la première moitié de la classe. A quelle notion statistique fait appel cette argumentation ?

Ex 1.3 - Dans une classe de 21 élèves, 20 élèves ont participé à un contrôle dont les résultats ont été {7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8}. La moyenne de ces cotes est de 12. Le lendemain, l'élève absent, Jojo Kicetu est interrogé à son tour et il obtient 20. La moyenne de ce deuxième contrôle est donc de 20. La cote de 16, moyenne des deux jours ((12 + 20)/2) représente-t-elle la moyenne de la classe ?
Combien d'autres Jojo Kicetu faudrait-il adjoindre à cette classe pour que la moyenne soit effectivement de 16 ?

Ex 1.4 - Reprenons notre situation 1.3 et apportons-y de légères variations en divisant par deux le salaire du patron et du sous-directeur. On aura donc :

Salaire du patron :	6 000 euros	
Salaire du sous-directeur :	2 500 euros	
Salaire des acheteurs :	1 250 euros.	Ils sont 6.
Salaire des chefs de secteur :	1 000 euros.	Ils sont 5.
Salaire des autres :	500 euros.	Ils sont 10.

Quelles sont la moyenne, la médiane et le mode de cette nouvelle donnée ? Comparer avec ce qu'on avait trouvé précédemment.

Ex 1.5 - Vous avez bien compris les précisions de vocabulaire données en B ? Quelle est donc la **population** de la situation 1.1 ?

Ex 1.6[iii] - Dans une université, on a voulu savoir si la sélection était "sexiste", et on a comparé les résultats d'un groupe d'étudiants. Les proportions de réussites sont les suivants :

Hommes : 534 / 1198 = 0,446
Femmes : 113 / 449 = 0,252.

On pense pouvoir conclure au "sexisme".
Cependant, on dispose aussi des résultats par départements :

	Département A	Département B
Hommes	512 / 825 = 0,621	22 / 373 = 0,059
Femmes	89 / 108 = 0,824	24 / 341 = 0,070

Le "sexisme" semble toujours bien là, mais ... dans l'autre sens !

[7] Le signe "Ex", emprunté à Papy, peut désigner un exemple, un exercice, un excursus, ...

Cette curiosité est connue sous le nom de **paradoxe de Simpson**. Elle peut avoir des conséquences non négligeables si, par exemple, les mesures effectuées concernaient deux types de médicaments qui auraient été testés par deux équipes différentes ayant travaillé indépendamment l'une de l'autre. La conclusion aurait pu être que le médicament 2 était plus performant que le médicament 1. Si au lieu de deux études, on n'en avait fait qu'une en regroupant toutes les observations, on serait arrivé à la conclusion opposée. La maîtrise statistique d'une situation, c'est décidément plus difficile que quelques graphiques et quelques moyennes !

Ex 1.7 - Les statistiques ne me concernent pas.
Voici un relevé des résultats de quelques élèves dans trois classes parallèles; pour chacune, on vous donne la cote moyenne.

Classe A		Classe B		Classe C	
Moyenne : 15		Moyenne : 13		Moyenne : 9	
Jean	15	Fabienne	13	René	15
Emilie	16	Claude	16	Jeanne	12
Magali	12	Marc	12	François	8

Que pensez-vous des affirmations suivantes :
- Magali a 12, mais elle est en dessous de la moyenne.
- Jeanne a 12 aussi, mais le professeur de cette classe est particulièrement sévère.
- Le professeur de A jette les points à la tête des élèves.
- La classe B est meilleure que la classe C.
- La classe C est vraiment très mauvaise.
- René est très intelligent.
- Il est plus difficile pour Fabienne d'obtenir 13 que pour Emilie d'obtenir 16.
- Le professeur de C est moins bon que celui de A.
- Le professeur de C est meilleur que celui de A.
- Etc.

Ex 1.8 - Comment faire fabriquer 10 000 pantalons ?
Vous êtes responsable d'un atelier de confection et on vous a commandé 10 000 pantalons, tous du même modèle et réalisés dans le même tissu, mais bien sûr pas tous de la même taille. Vous décidez d'organiser votre atelier en deux équipes : une qui réalisera les petites tailles, et l'autre les grandes. Une étude statistique vous donne pour chaque taille, le métrage nécessaire à sa réalisation et le nombre d'unités vendues dans toute une série de magasins. Quelles sont les valeurs centrales qui vous seront nécessaires pour commander votre tissu, répartir la tâche entre les deux équipes et décider de la taille qu'auront les pantalons réalisés avec le surplus éventuel de tissu ?
*SOLUTION : on calcule le métrage **moyen** pour réaliser un pantalon. On trouve par exemple 1,57m. On devrait donc commander 15 700m de tissu. Le conditionnement de ce tissu fait qu'on doit commander 16 000m. En supposant que le temps de fabrication est pratiquement le même pour toutes les tailles, c'est la **médiane** qui sera utilisée pour répartir le travail de manière que les deux équipes aient le même nombre de pièces à confectionner. Enfin, puisqu'il reste du tissu, avec le reliquat on confectionnera des pantalons de la taille **modale**, c'est-à-dire la plus vendue.*

Ch. 2 : Les indices de dispersion

Où peut-on être mieux ...
qu'au sein de la "bonne moyenne" ?

dicton belge.

A. Situations

Sit 2.1.1 - Bonnes vacances

Alice et Béatrice ont pris des vacances dans un endroit de rêve : des conditions de vie très proches de la nature (comprenez entre autres : logement sous tente), mille et une possibilités d'activités sportives, de loisir, de détente, ..., le tout à un prix intéressant. Le prospectus annonce, pour la période envisagée, une température moyenne de 25°C et une brise légère.
Alice est revenue enchantée de ses vacances, Béatrice très mécontente : un gros rhume, des mauvaises nuits fréquentes, des maux de tête,... lui ont gâché son séjour.
C'est que, alors que Béatrice, ayant lu très superficiellement le prospectus, ne s'était équipée que de vêtements d'été, Alice au contraire avait lu que sur une journée, les écarts de température pouvaient être importants et que notamment, les nuits étaient plutôt fraîches. Elle avait donc "encombré" ses bagages de jeans et de "petites laines" qui lui ont permis de bien dormir et de ne pas frissonner à la soirée et au petit matin.

Sit 2.1.2 - Le restaurant scolaire

Tout au long d'une année, on a fait des relevés statistiques d'un restaurant scolaire et on a trouvé qu'on servait en moyenne 100 repas complets par jour de fonctionnement du restaurant (4 jours par semaine), et que chaque repas comportait 300g de pommes de terre[8]. Ces données suffisent-elles à l'économe pour faire des achats hebdomadaires raisonnables de pommes de terre ?

Sit 2.1.3 - Achat ou leasing ?

Au moment de renouveler sa voiture, M. X songe à un leasing plutôt qu'à un achat. Ayant déterminé son cahier des charges, (type de voiture, niveau de confort, accessoires, ...), il s'adresse à plusieurs sociétés de location qui lui remettent des devis avec différentes options[11].
- Option "tout compris" : le client paie un forfait mensuel qui couvre tout sauf le carburant.
- Option "forfait de base" : le client paie un forfait mensuel (moindre que dans le "tout compris"), mais, en plus du carburant, il prend en charge les entretiens ordinaires prévus par le carnet d'entretien.

M. X se dit que le "tout compris" est probablement basé sur un kilométrage supérieur à celui qu'il fait d'habitude et sur une façon de conduire plus "musclée" que la sienne qui est très

[8] Ces données sont inventées.

"cool". Il songe donc à la formule "forfait de base".
Pour évaluer son coût réel total, M. X demande à chaque société de lui dire quelle est la moyenne des frais d'entretien par véhicule pour les deux dernières années. (On suppose que les modèles n'ont guère évolué durant ce laps de temps et que chaque firme fonde ses statistiques sur une flotte suffisamment nombreuse).
La firme A répond que cette moyenne est de 11 000 F (272,68 €) par an. La firme B dit qu'à part 2 ou 3 exceptions sur 800 véhicules, ces frais se situent dans une fourchette de 10 000 à 14 000 F (247,89 à 347,05 €) par an, avec une moyenne de 11 700 F (290,04€). M. X demande alors ces précisions à la firme A qui lui renseigne une fourchette de 8 500 à 20 000 F (210,71 à 495,79€) par an.
Vos commentaires ?

Sit 2.1.4 - Encore des cahiers de cotes

Il est vrai que dans la vie courante, l'outil statistique n'est pratiquement jamais utilisé que pour analyser des données qui ont été fournies ou récoltées par ailleurs. On a cependant jugé utile de glisser dans la série de situations-problèmes ou d'exercices un problème de construction de données. Ceci est fait dans la conviction que la confrontation (très limitée) avec une construction à faire permet de mieux déchiffrer une construction terminée : une personne qui s'est essayée à barbouiller quelques toiles lit probablement mieux une œuvre picturale; un bricoleur qui s'escrime à concevoir et à faire fonctionner des objets mécaniques apprécie sans doute davantage des objets tout faits. Ou encore : il est probable que construire des résumés aide à lire des résumés.

Ici, on a jugé probable qu'un élève qui a construit des ensembles de données saura mieux ce que peuvent cacher (ou révéler) des données plus synthétiques comme celles que proposent les deux personnages imaginés.

M. Profdematix, qui enseigne à l'école X, rencontre son collègue et ami M. Profdematy qui enseigne, lui, à l'école Saint-Y. Ils ont chacun 3 classes parallèles de 3e de l'enseignement général. Comme si c'était fait exprès pour la facilité d'un exercice scolaire, chacune de ces classes comprend 20 élèves.
Dans leur conversation, ils échangent leurs impressions sur leurs classes, et M. Profdematix donne les renseignements suivants : dans ma classe Ax la moyenne est de 12 et tous les résultats se situent dans la fourchette [10 , 14]; la classe Bx a aussi une moyenne de 12, mais la fourchette est de [2 , 20]; dans la classe Cx, la moyenne est toujours de 12, et à part deux élèves qui ont l'un 1/20 et l'autre 19/20, les résultats se situent dans la fourchette [8 , 15].
M. Profdematy donne pour sa part les informations suivantes : dans la classe Ay, la médiane est de 12 et la moitié des résultats appartient à l'intervalle [10 , 14]; la médiane de la classe By est de 13 et la moitié des résultats appartient à l'intervalle [8 , 15]; dans la classe Cy, la médiane est de 11 et 80% des résultats sont dans l'intervalle [9 , 16].

Quelle est votre impression sur le profil de chacune des six classes proposées ? (Il pourrait être intéressant de reprendre des questions comme celles qu'on a posées dans Sit 1.2).

Pourriez-vous remplir le cahier de cotes de ces deux professeurs de manière que les informations données soient exactes ?
Toutes les cotes sont supposées être rapportées à un maximum de 20. Vous pouvez aussi ajouter des hypothèses ou des contraintes. Par exemple, que les cotes sont toutes des nombres entiers. Vous pouvez encore essayer de remplir le tableau avec des résultats aussi favorables

(ou aussi défavorables) que possible, tout en respectant les indications données. La comparaison de votre travail avec celui d'autres élèves sera vraisemblablement instructive.

Une fois le tableau rempli, traduisez dans le langage de M. Profdematy les données synthétiques de M. Profdematix, et réciproquement.

Ecole X M. Profdematix				**Ecole Saint-Y** M. Profdematy			
	Ax	**Bx**	**Cx**		**Ay**	**By**	**Cy**
a				**a**			
b				**b**			
c				**c**			
d				**d**			
e				**e**			
f				**f**			
g				**g**			
h				**h**			
i				**i**			
j				**j**			
k				**k**			
l				**l**			
m				**m**			
n				**n**			
o				**o**			
p				**p**			
q				**q**			
r				**r**			
s				**s**			
t				**t**			

Sit 2.1.5 - Concours interscolaire

Pour un concours interscolaire, telle classe doit présenter un participant. Deux élèves ont été retenus dans une présélection; voici leurs résultats aux épreuves d'une période (cotées sur 10) :
Elève A : 1, 3, 7, 10, 3, 8, 4, 1, 7, 6.
Elève B : 4, 6, 5, 5, 6, 4, 4, 6, 5, 5.
Comme vous pouvez le vérifier, pour chacun de ces deux élèves, la moyenne est de 5 et la médiane aussi.
Quel élève choisiriez-vous pour représenter la classe ? Pourquoi ?
Nouvelle information : le premier tour du concours est organisé dans le but d'éliminer les candidats faibles (résultats inférieurs ou égaux à 3/10) plutôt, par exemple, que de sélectionner des "têtes de série".
En supposant que l'école souhaite surtout passer au moins un tour, confirmez-vous votre choix ou souhaitez-vous le modifier ?

B. Définitions et formules

Les exemples proposés dans ce deuxième chapitre visent à faire apparaître une idée importante de l'analyse statistique des situations : **dans certains cas, la seule donnée d'une valeur centrale (moyenne ou médiane) peut suffire à produire un résumé fiable d'une situation et à permettre des décisions valables s'y rapportant. Dans beaucoup d'autres cas, cette seule information est insuffisante; elle doit être complétée par des indications relatives aux écarts par rapport à la valeur centrale.** Ces indications sont connues sous le nom d'indices de dispersion.

La situation 2.1.2 (le restaurant scolaire) peut être prise pour un bon exemple de cas où des moyennes permettent à elles seules de prendre des décisions adéquates dans un contexte donné et avec un projet donné.

Les autres situations attirent au contraire l'attention sur le fait que la négligence ou l'ignorance des écarts par rapport à la valeur centrale peut causer des effets au moins désagréables.

Dans les situations présentées jusqu'à présent, nous avons utilisé à plusieurs reprises la notion commune, très simple, non technique, de "fourchette". C'est à coup sûr une indication intéressante dans certains cas, en particulier quand cette fourchette est étroite. Elle renseigne alors que les valeurs sont très groupées autour de la valeur centrale et qu'on ne commet pas une grosse erreur en attribuant fictivement à chaque individu cette valeur centrale.

Mais dans les autres cas, cet indicateur est très pauvre, en particulier parce qu'il est très sensible à l'une ou l'autre valeur extrême. En fait, cet indicateur ne tient compte que de deux données pour caractériser tout l'ensemble. Voici quelques exemples caricaturaux mettant en évidence la signification de ces paramètres.

Sit 2.1.6 - Exemples caricaturaux

Dans tous ces exemples, le nombre d'élèves par classe est supposé suffisamment grand pour que l'analyse statistique soit au moins un peu significative (au moins 20, par exemple) et les cotes sont supposées rapportées à un maximum de 20.

Sit 2.1.6A
Tous les élèves ont des cotes qui vont de 18 à 20, sauf Mimi Lecancre qui a 0.
La fourchette est donc [0 , 20].
Où situeriez-vous la moyenne ?
Où situeriez-vous la médiane ?
Si vous n'y voyez pas clair, créez donc l'un ou l'autre exemple qui réponde aux conditions.

Sit 2.1.6B
Tous les élèves ont des cotes qui vont de 0 à 5, sauf Jojo Kicetu qui a 20.
Fourchette :
Moyenne :
Médiane :

Sit 2.1.6C
Tous les élèves ont 12, sauf Mimi Lecancre qui a 4 et Jojo Kicetu qui a 20.
Fourchette :
Moyenne :
Médiane :

Sit 2.1.6D
Une moitié des élèves a 4 et l'autre moitié a 20 (*Le nombre d'élèves est un nombre pair*).
Fourchette :
Moyenne :
Médiane :

Les exemples **A** et **B** montrent que la seule fourchette peut être une information d'intérêt pratiquement nul. La fourchette et une valeur centrale peuvent former ensemble une information significative. Mais les exemples **C** et **D** font apparaître que deux ensembles de données ayant des fourchettes et des valeurs centrales identiques peuvent avoir des profils très différents. Quel sens y a-t-il d'ailleurs à parler de valeurs centrales pour des situations comme 2.1.6D où la population est composée de deux (ou plusieurs) sous-populations hétérogènes entre elles. On pourrait bien sûr construire d'autres exemples illustrant ces remarques. Sans doute l'avez-vous d'ailleurs fait en remplissant le tableau de Sit 2.1.4.

Essayons de trouver une solution statistique à ce problème statistique; commençons par la moyenne.

Si nous cherchons la moyenne des écarts par rapport à la moyenne, nous allons bien sûr trouver 0, quelle que soit la dispersion, puisque, par une conséquence directe de la définition, la somme des écarts par rapport à la moyenne est nulle (voir Ch.1, §3).

Pour résoudre ce problème, deux voies, au moins, sont possibles : se défaire des signes des écarts par rapport à la moyenne en prenant la valeur absolue de ces écarts; se défaire des signes en les élevant au carré.

A première vue, c'est la méthode des valeurs absolues qui apparaît la plus simple. Cependant, pour des raisons qu'on ne va pas développer ici parce que sans être vraiment difficiles, elles ne sont pas tout à fait élémentaires[iv], c'est la méthode des carrés qui a été retenue.

1. Variance (V) et Ecart-type (σ)

On appelle **variance (V)** d'une population la moyenne des carrés des écarts par rapport à la moyenne. En formules :

Ecarts par rapport à la moyenne : $m - x_i$

Carrés de ces écarts : $(m - x_i)^2$

Variance ou moyenne de ces carrés : $V = \frac{1}{n} \sum_{i=1}^{n} (m - x_i)^2$.

ou encore : $V = \frac{1}{n} \sum_{i=1}^{k} (m - x_i)^2 . n_i$

Cette notion de variance a l'avantage sur la notion très peu technique de "fourchette" d'être calculée à partir - et donc de tenir compte - de toutes les valeurs de la population. Mais le passage au carré a changé l'ordre de grandeur des données. On revient à l'ordre de grandeur initial en prenant la racine carrée de la variance.
Le résultat de cette opération est l'**écart-type (σ)**. On a donc :

Ecart - type = $\sqrt{\text{Variance}}$

$\sigma = \sqrt{V}$

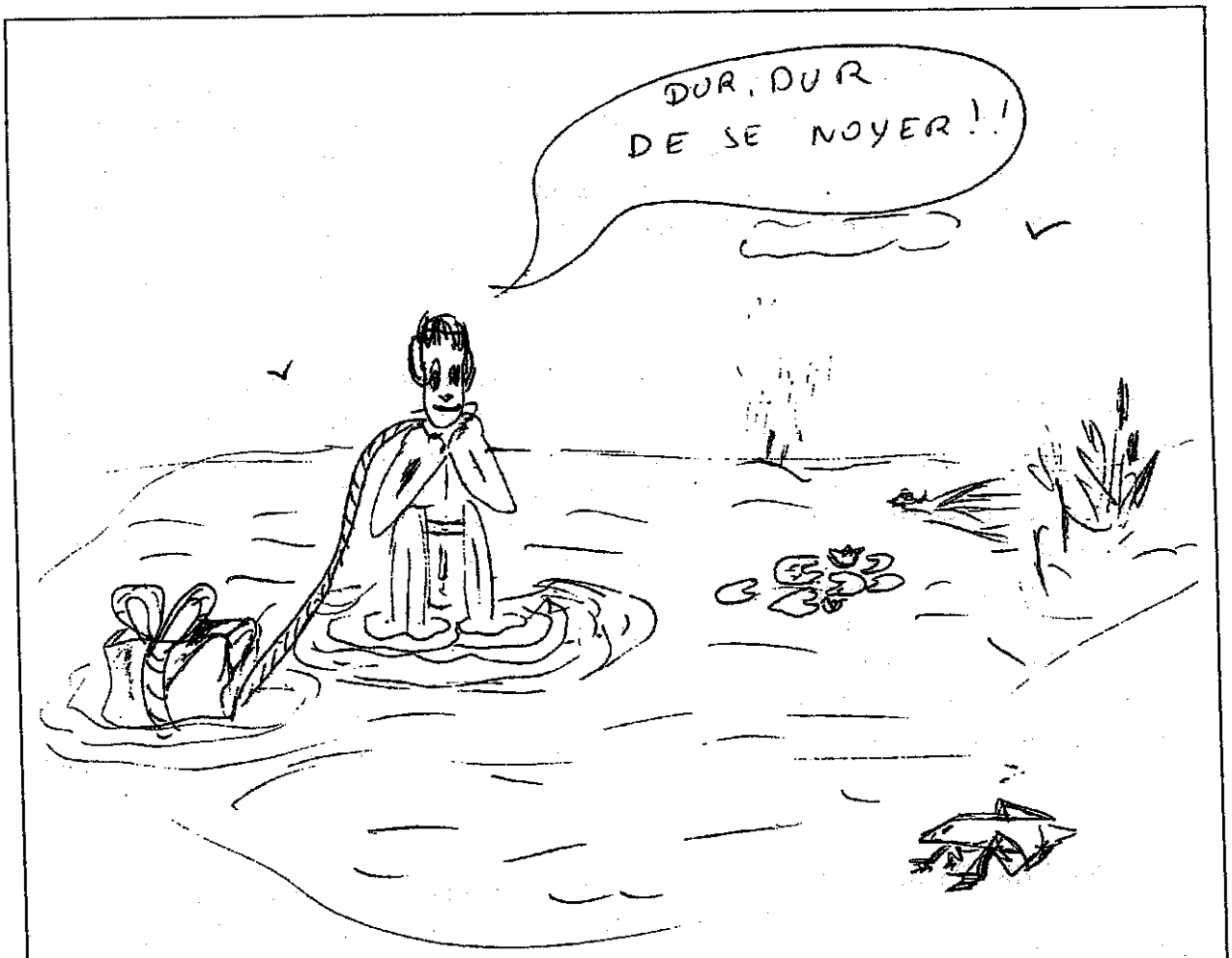
Dans un certain nombre de cas, l'idée "grand public" et peu précise de "bonne moyenne" ou d'"être dans la moyenne" peut être définie de façon plus technique par un intervalle comprenant la moyenne, intervalle lui-même défini par l'écart-type; par exemple : $[m-\sigma , m+\sigma]$, ou $[m-2\sigma , m+\sigma]$, ou plus généralement $[m -a\sigma , m + b\sigma]$ a et b étant deux nombres positifs "judicieusement" choisis.

Remarque : sur le clavier de beaucoup de calculatrices ayant les fonctions statistiques, on trouve les symboles σ_n et σ_{n-1} . La fonction σ_n correspond à l'écart-type σ défini ci-dessus. C'est l'écart-type défini sur l'ensemble de la population. Dans les analyses d'échantillons, on utilise l'écart-type σ_{n-1} qui est la racine carrée de V_{n-1} calculé en remplaçant $1/n$ par $1/(n-1)$ dans la formule ci-dessus[9]. Si la calculatrice n'a que le symbole σ, il faut vérifier, par consultation du mode d'emploi ou par un exemple simple, de quel écart-type il s'agit.
Exemple : soit la population {10, 15, 20} dont la moyenne est 15; $\sigma_n = 4{,}08...$, $\sigma_{n-1} = 5$.

[9] Voir en annexe une justification du fait qu'on utilise σ_n ou σ_{n-1} selon qu'on se réfère à la population entière ou à un de ses échantillons.

Dans le logiciel EXCEL, σ_n est donné par la fonction ECARTYPEP (avec un "P" comme Population) tandis que σ_{n-1} est donné par ECARTYPE. On a de même les fonctions VAR et VAR.P pour la variance, selon que l'on se réfère à un échantillon ou à la population entière.

On peut aisément calculer que, pour les données de Sit 2.1.5, l'écart-type des cotes de l'élève A est de 2,8... et celui des cotes de l'élève B est de 0,77...



Si on vous dit qu'un quidam s'est noyé dans un étang dont la profondeur est, en moyenne, de 30 cm, allez-vous conclure que cette personne était suicidaire et s'est couchée pour avoir une chance d'être noyée ?

2. Quartiles, déciles, centiles

L'analyse d'une situation par le biais de la médiane peut elle aussi être affinée. Rappelons en bref que la médiane est la valeur qui divise la population en deux moitiés de même effectif (à une unité près) : celle dont les valeurs sont inférieures ou égales à cette valeur médiane et les autres.

Une idée assez simple - et bien utile - est de partager la population en trois, en quatre, en p sous-ensembles. Pratiquement, il n'est pas rare qu'on divise en quatre (d'où les quartiles),

en dix (d'où les déciles), en cent (d'où les centiles[10]); mais rien n'empêche quand cela s'avère intéressant de répartir la population en p sous-ensembles (d'où les α-quantiles, expression dans laquelle α désigne une fraction de dénominateur p, inférieure à l'unité).

Comme ce fut le cas pour la moyenne, ces notions permettent d'exprimer de manière plus précise l'idée naïve de "bonne moyenne". Une définition plus technique de ces notions de quartiles, déciles,... sera donnée en annexe[11] en raison d'un certain nombre de difficultés qui y sont rencontrées. Voyons-en la signification dans quelques situations.

Sit 2.2.1 - Un service de médecine scolaire a relevé la taille d'une population de 200 enfants de 8 ans. Tout le monde sera d'accord de dire qu'un enfant de cet âge qui mesure 1m est "petit", tandis qu'un autre qui mesure 1,7m est "grand". Où placer la barre entre les "petits" et les "grands" ? Une idée à première vue intéressante est d'utiliser la médiane : sont "petits" les enfants dont la taille est inférieure ou égale à la médiane; les autres sont "grands". On a ainsi 50% de "petits" et 50% de "grands. Un tel classement est simple, mais il apparaîtra souvent totalement futile.

Il n'en sera sans doute pas de même de cette autre manière de regrouper les individus de la population : dans une première classe, les 25% d'enfants les plus petits, ceux donc dont la taille est inférieure ou égale au premier quartile. La 2e classe est composée de ceux dont la taille est comprise entre le premier et le troisième quartile; elle contient donc 50% de l'effectif. Dans la troisième classe, ceux dont la taille est supérieure ou égale au troisième quartile. Des dénominations habituelles indiquent l'utilité d'un tel classement :

1. les enfants trop petits pour leur âge.
2. les enfants dont la taille est "normale".
3. les enfants trop grands pour leur âge.

Le médecin qui examine des enfants de la première classe peut songer à d'éventuels cas de nanisme, ou plus généralement se poser des questions sur des raisons d'un retard de développement.

Sit 2.2.2 - Toujours dans le domaine médical, les protocoles d'analyses sanguines comportent habituellement des "valeurs de référence" qui déterminent un intervalle en dehors duquel il y a lieu de se poser des questions.

Sit 2.2.3 - Un étudiant désirant aller faire une partie de ses études aux Etats-Unis d'Amérique doit présenter un examen d'anglais pour pouvoir être inscrit. Son résultat lui est communiqué sous la forme suivante : *8540 personnes ont présenté l'examen; vous êtes dans le*

---

[10] Certains disent "percentiles", mais ce mot ne figure ni dans les dictionnaires français consultés, ni dans le correcteur d'orthographe de ce traitement de texte qui suggère de le remplacer par "perceptibles", "perceptives" ou "pressenties" !

[11] Il est cependant utile d'attirer ici l'attention sur une certaine confusion dans l'utilisation de ce vocabulaire. Les mots quartiles, déciles, ... désignent tantôt les valeurs qui partagent la population, tantôt les sous-ensembles de cette population. Ainsi on lit dans le Nouveau Larousse Universel en deux volumes, de 1969 les définitions suivantes :
Décile : dixième partie d'un ensemble de données classées dans un ordre donné.
Quartile : chacune des valeurs que prend la variable lorsqu'on partage en quatre parties égales le nombre total des observations.
Dans un contexte statistique strict, c'est toujours de **valeurs** qu'il s'agit.

septième intervalle interdécile. Pouvez-vous traduire ce renseignement dans des termes d'évaluation scolaire qui nous sont plus familiers ? (Réponse en note de bas de page[12]).

Certains préfèrent notre système habituel de cotation en référence à un maximum absolu (10, 20 ou 100 ou ...). Ils font valoir que le système d'information en référence au classement dans un groupe, d'une part, est tributaire du groupe (qui peut être faible ou fort) et d'autre part ne renseigne pas sur d'éventuels écarts importants (il y aurait par exemple une "cassure" entre le 7e intervalle interdécile assez médiocre et le 8e déjà assez fort). Les autres leur répondent que d'une part la référence absolue est illusoire (que signifie par exemple une maîtrise parfaite de l'anglais méritant la cote maximale ?) et d'autre part que si on limite l'utilisation de l'outil statistique à son domaine de validité, c'est-à-dire en particulier en on ne considère que des populations à effectif important, les biaisages[13] signalés ne sont que peu voire très peu probables ou importants. En outre, la tendance des évaluateurs "de terrain" (les professeurs dans leurs classes) à forcer les résultats pour qu'ils reproduisent des répartitions "normales", c'est-à-dire reproduisant des courbes en cloche, même sur de petits effectifs, est un travers souvent dénoncé. Qui tranchera ?

Ces notions vont permettre une autre approche de l'idée de regroupement autour de la valeur centrale.

Ce qui intéresse un industriel, un artisan ou un commerçant qui commande des pièces à un fournisseur, c'est qu'un très haut pourcentage des pièces fournies soit dans une fourchette de tolérance autour de la valeur théorique précisée dans le cahier des charges : si par exemple, le critère est le poids, tant les pièces trop légères que les trop lourdes sont à rejeter. Si l'acquéreur soumet à l'examen un échantillonnage de pièces, la formulation des résultats de l'examen en termes de moyenne et d'écart-type risque de ne pas être instructif. On a vu en effet que de tels résumés identiques peuvent recouvrir des situations différentes. Il ne sera donc pas rare qu'on préfère des conclusions d'examens en termes d'α-quantiles. Si on vous dit par exemple que sur un ensemble de pommes qu'on a pesées, le premier décile est à 140g et le 9e décile à 160g, vous comprenez que 80% des pommes pèsent entre 140g et 160g[14].

---

[12] Le premier décile est la valeur sous laquelle se trouvent les 10% de participants qui ont eu les moins bons scores; etc. Le premier intervalle interdécile est composé des individus ayant eu un score inférieur au premier décile. Donc notre étudiant, étant dans le 7e décile, a fait mieux que 60% de l'effectif, soit mieux que 5124 participants, mais moins bien que les participants classés dans les 8e, 9e et 10e intervalles interdéciles. Selon nos classements habituels où celui (celle) qui a le meilleur score est premier (première), notre étudiant sait qu'il est au mieux 2563e, au pire 3416e (sur 8540).

[13] Le mot ne figure pas aux dictionnaires consultés et est refusé par le correcteur du traitement de texte; il est pourtant d'usage courant dans ce contexte.

[14] Il est sans doute bon de savoir que la "fiabilité industrielle" ordinaire est bien supérieure à ce qu'un amateur espère de sa production. 98 ou même 99% de pièces réussies ne sont pas des scores extraordinaires. Qu'on songe par exemple aux pièces de monnaie qui doivent fonctionner dans toutes sortes de machines, y compris celles qui rendent la monnaie. Qu'on songe aussi à la pureté des substances et au dosage des médicaments, etc. ...

C. Pêle-mêle, des Ex

Ex 2.1 - Un atelier de mécanique commande à deux fournisseurs des cylindres d'acier de 150mm de long. Le cahier des charges prévoit une tolérance inférieure à 0,2mm. Les pièces non conformes, trop courtes ou trop longues, vont au rebut et sont donc une perte sèche pour l'atelier. Avant d'accepter les lots présentés, l'atelier procède à un examen d'échantillons.
Fournisseur A : le premier décile est à 149,9mm, le 9e à 150,2mm.
Fournisseur B : le 6e centile est à 149,8mm, le 94e à 150,1mm.
Vos commentaires.

Ex 2.2 - Une des caractéristiques de l'approche statistique des situations est de gommer les spécificités individuelles, a-t-on dit dans l'introduction. Voici quelques exemples. On vous demande de dire si l'approche statistique est judicieuse.

Ex 2.2a - Lors d'un rallye touristique, une des épreuves du parcours consiste à évaluer la distance qui sépare l'endroit où on se trouve d'un autre endroit identifiable (ou d'évaluer la hauteur d'un arbre ou d'un bâtiment, ou le poids d'un objet, ou ...). Certaines équipes font la moyenne de l'évaluation de chaque membre de l'équipe.

Ex 2.2b - Un instituteur remet à ses élèves une feuille sur laquelle est dessiné un quadrilatère quelconque et demande à chacun de faire les mesures et les calculs nécessaires pour connaître l'aire de la figure. On devine qu'à la collecte des résultats, on observe des différences généralement plus grandes que ce qu'on pourrait croire. Quel nombre prendre pour la mesure de cette surface ?

Ex 2.2c - Un pèse-personne (supposé mesurer avec la précision appropriée le poids des personnes) va mesurer mon poids à un moment donné et dans des circonstances données. Mais qu'est-ce donc que mon poids ?

Ex 2.2d - Les élèves vont être contents !
L'histoire se passe il y a quelques années quand on ne disposait que du télégramme pour envoyer des messages. Ceux qui ont connu cela savent que l'aspect peu performant du procédé obligeait à utiliser le style "télégraphique".
Un professeur chargé d'organiser des vacances de neige a relevé les pointures des élèves qui désirent louer des chaussures de ski. Un peu débordé dans son organisation, il a oublié d'envoyer à temps cette liste de pointures. En dernière minute, pour éviter un télégramme très long et très coûteux, il fait la moyenne des pointures et télégraphie le résultat de son calcul et le nombre de paires de chaussures.

Ex 2.3 - La fonction "*Ecartype*" du tableur Excel (version 5.0) donne l'écart-type σ_{n-1}. tandis que "*Ecartypep*" donne σ_n. Par quelle formule transformer σ_{n-1} en σ_n ou réciproquement ?

Ch. 3 : Des techniques d'analyse des données

Un petit dessin vaut parfois mieux qu'un long discours.

Il est vrai que les outils informatiques dispensent aujourd'hui dans une large mesure du travail fastidieux de mise en ordre et de présentation parlante des données. C'est une des raisons pour lesquelles un cours de statistique ne peut plus être conçu comme cela a pu être le cas. Encore faut-il savoir imaginer quelle présentation choisir pour rendre les données révélatrices (ou trompeuses). De plus, il est intéressant de savoir lire des données que d'autres ont arrangées : leur présentation peut ne pas nous être familière. C'est pourquoi, conformément à ce qui a déjà été dit (introduction à Sit 2.1.4), il n'est pas inutile de s'escrimer à mettre en forme des données brutes. Ce l'est d'autant moins qu'il existe quelques moyens particulièrement simples et faciles à mettre en œuvre, ne requérant aucun appareillage, qui permettent une présentation synthétique.

A. Diagrammes en bâtonnets

Sit 3.0 - Activité statistique

Madame Mateuze a divisé sa classe en trois équipes, et à chacune d'elles, elle a confié une tâche statistique.

L'équipe A reçoit une caisse comprenant 46 tomates. On leur demande de les peser une à une et de noter le poids en grammes. Voici le résultat de leurs travaux :

155	151	149	154	163	159	162	158	152	165
148	157	143	142	165	154	165	156	162	157
156	154	159	165	155	148	160	158	146	151
152	149	165	153	148	152	162	158	155	165
154	166	163	155	159	152				

A l'équipe B, Madame Mateuze a demandé de mesurer en centimètres la taille d'un groupe de 46 condisciples. Ils ont trouvé :

155	151	149	154	163	159	162	158	152	165
148	157	143	142	165	154	165	156	162	157
156	154	159	165	155	148	160	158	146	151
152	149	165	153	148	152	162	158	155	165
154	166	163	155	159	152				

L'équipe C a dû chronométrer la course individuelle de 46 élèves et exprimer ses observations en nombre (arrondi) de mètres parcourus en 30 secondes. Leurs observations :

155	151	149	154	163	159	162	158	152	165
148	157	143	142	165	154	165	156	162	157
156	154	159	165	155	148	160	158	146	151
152	149	165	153	148	152	162	158	155	165
154	166	163	155	159	152				

Pour ceux qui ne l'auraient pas remarqué, signalons que les trois tableaux de nombres sont identiques ! On a voulu par cet artifice attirer l'attention sur le fait que quand on traite mathématiquement des données numériques relatives à une situation concrète, on oublie (on fait abstraction de) la signification réelle de ces données. Cette signification, mise en veilleuse en quelque sorte, peut cependant être réactivée pour détecter des erreurs flagrantes. Si on vous dit, par exemple, qu'on a mesuré le poids d'un certain nombre de personnes ordinaires (pas des "sumo" japonais) et qu'on a trouvé 183 kilos comme poids moyen, il y a certainement une erreur quelque part.

Revenons à notre tableau de données, appelé **tableau brut**. Il contient le résultat des observations dans l'ordre où elles ont été récoltées (ou dans un ordre quelconque). Si on travaille avec des outils informatiques, c'est ce tableau qui sera encodé. En toute hypothèse, c'est la référence de base à partir de quoi commence le traitement des données.

Car ce qu'il y a de plus clair dans le tableau brut, c'est... qu'on n'y voit pas grand-chose, surtout si l'effectif est important. Aussi, on va le "traiter" pour le rendre plus parlant.

Deux manières de traiter les données.

1. La méthode "tiges-et-feuilles"

La méthode "tiges-et-feuilles" (en anglais "stem-and-leaf") permet d'arriver en une fois et sans aucun appareillage, à une présentation très parlante des données. Elle peut être mise en œuvre par une personne seule, mais ce sera sans doute plus confortable de travailler à deux : une qui lit les données du tableau brut, et l'autre qui les transcrit.

On commence par repérer dans le tableau brut le genre de nombres qu'on a : dans notre exemple, des cent quarante, des cent cinquante, des cent soixante. On choisit les "tiges" (ici, 14, 15, 16) qu'on écrit en colonne. Puis au fur et à mesure de la lecture des nombres, on inscrit le(s) chiffre(s) manquant(s), les "feuilles", en face de la tige appropriée, de manière qu'avec ces conventions, chaque nombre du tableau brut apparaisse dans le nouveau tableau ainsi créé. Pour que ce tableau soit en même temps un outil graphique, il est essentiel de disposer les derniers chiffres inscrits à intervalles réguliers. Voici le résultat de ce travail :

14	9 8 3 2 8 6 9 8
15	5 1 4 9 8 2 7 4 6 7 6 4 9 5 8 1 2 3 2 8 5 4 5 9 2
16	3 2 5 5 5 2 5 0 5 2 5 6 3

Bien entendu, on peut penser que ce regroupement en trois classes est trop sommaire. Pour remédier à cela, il suffit de subdiviser les tiges en prenant par exemple **14B**, **14H**, **15B**, **15H**, **16B**, **16H**. Dans les classes marquées B les valeurs dont le chiffre qui suit la tige est 0, 1, 2, 3 ou 4, dans les classes marquées H, les autres.

14B	3 2
14H	9 8 8 6 9 8
15B	1 4 2 4 4 1 2 3 2 4 2
15H	5 9 8 7 6 7 6 9 5 8 8 5 5 9
16B	3 2 2 0 2 3
16H	5 5 5 5 5 5 6

Une présentation un peu plus habillée pourrait-elle rendre ces données encore plus lisibles ? On peut, par exemple numéroter les colonnes et tracer quelques lignes horizontales ou verticales qui faciliteront le repérage[15]. On peut même rajouter une colonne dans laquelle on écrira les répétitions pour avoir en une fois une présentation de type graphique et son expression chiffrée.

14B	3	2														2
14H	9	8	8	6	9	8										6
15B	1	4	2	4	4	1	2	3	2	4	2					11
15H	5	9	8	7	6	7	6	9	5	8	8	5	5	9		14
16B	3	2	2	0	2	3										6
16H	5	5	5	5	5	5	6									7
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	

On voit immédiatement les avantages considérables de cette méthode qui produit à si peu de frais une disposition graphique parlante, tout en gardant la trace de toutes les valeurs individuelles du tableau brut. De ce fait, elle permet une vérification assez facile de la correction des données sur lesquelles on travaille. En outre, comme on le verra plus loin, cette disposition rend aisés des calculs comme celui de la moyenne ou de l'écart-type relatifs à des tableaux groupés (voir plus loin).

Bien sûr, ici pas plus qu'ailleurs, il n'y a de miracle. Cette méthode montrera ses limites si on veut par exemple étudier des populations importantes (plusieurs centaines ou plusieurs milliers d'individus). Il est clair que pour de tels travaux des outils informatiques appropriés sont indispensables. En outre, par ce procédé, on regroupe automatiquement les données dans des classes basées sur l'expression des valeurs selon des nombres "ronds" dans le système d'unités choisi pour exprimer les grandeurs étudiées. Dans un certain nombre de cas comme notre exemple de référence cela peut être tout à fait adéquat, mais cela peut aussi ne pas l'être : le sens du problème pourrait imposer d'autres classes.

Si on étudie par exemple des résultats scolaires exprimés sur un total de vingt, la méthode "tiges-et-feuilles" regroupera au mieux les résultats en quatre classes : [0,5[, [5,10[, [10,15[, [15,20]. Ce regroupement n'est sans doute pas très significatif; on verra plus loin (§ C. Bâtonnets et histogrammes, Sit 3.3), une autre façon de constituer des classes, façon davantage liée à la manière habituelle d'interpréter les résultats scolaires qu'à leur expression arithmétique à vrai dire assez extrinsèque.

Il n'est bien sûr pas dit que parce que la méthode "tiges-et-feuilles" ne résout pas tout, elle n'a pas d'intérêt.

2. Le traitement classique

Même si on ne dispose pas de programme spécialisé, avec un tableur (Excel ou Lotus ou...), on peut déjà faire pas mal de choses. Il est vrai cependant que cela demande parfois un certain chipotage. Aussi il n'est pas sûr qu'on pourra travailler de manière confortable avec de tels outils sur des populations dont l'effectif dépasse quelques dizaines.

[15] Dans ce document, on a essayé de rendre le document plus lisible en ayant recours à divers types de caractères. Si on en a les moyens, rien n'empêche d'utiliser aussi la couleur.

Un avantage de l'ordinateur sur la calculatrice (même graphique) est qu'on garde la trace, éventuellement imprimée, du tableau brut. On peut donc vérifier l'exactitude des données enregistrées, et le cas échéant les corriger[16].

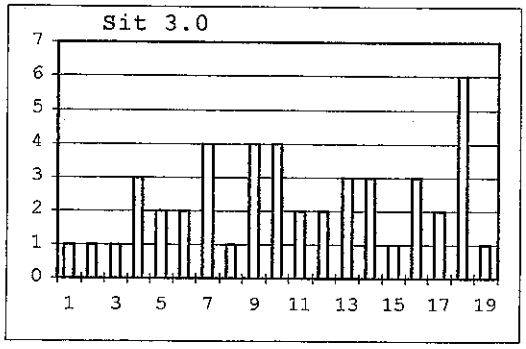
Sauf exception, les exemples proposés dans ces feuilles peuvent être pris en charge sans équipement particulier.

Dans ce traitement classique, une première étape consiste généralement à classer les données (le plus souvent) en ordre croissant. En théorie, rien de difficile à cela. En pratique, tant en travail "à la main" qu'en utilisant un tableur, ce n'est simple que dans les cas élémentaires. Si l'effectif est grand (quelques milliers ou même seulement quelques centaines), cela devient vite une tâche fastidieuse où la possibilité d'erreur n'est pas si réduite.

Grâce au **tableau ordonné** ainsi produit, on repère facilement les valeurs extrêmes (la fourchette) et les **répétitions**, c'est-à-dire le nombre de fois que chaque valeur apparaît. On construit alors le **tableau recensé** dans lequel en regard des valeurs figurent leurs répétitions.

Le fastidieux passage par le tableau ordonné peut être évité par le recours à la méthode traditionnelle qui consiste à écrire les valeurs en colonne et à inscrire un trait vertical à côté d'elles au fur et à mesure qu'on les lit dans le tableau brut. Généralement, pour faciliter la lecture des résultats de ce travail, on regroupe ces tirets par cinq, le cinquième trait barrant les quatre autres.

Outre qu'il est particulièrement utile pour le calcul de la moyenne, le tableau recensé permet une première visualisation : le **diagramme (des répétitions) en bâtonnets** : en abscisse, les valeurs, en ordonnée, les répétitions



		Sit 3.0			
Brut		**Ordonné**		**Recensé**	
155	165	142	156	**Val.**	**Rép.**
151	155	143	156	142	1
149	148	146	157	143	1
154	160	148	157	146	1
163	158	148	158	148	3
159	146	148	158	149	2
162	151	149	158	151	2
158	152	149	159	152	4
152	149	151	159	153	1
165	165	151	159	154	4
148	153	152	160	155	4
157	148	152	162	156	2
143	152	152	162	157	2
142	162	152	162	158	3
165	158	153	163	159	3
154	155	154	163	160	1
165	165	154	165	162	3
156	154	154	165	163	2
162	166	154	165	165	6
157	163	155	165	166	1
156	155	155	165		
154	159	155	165	Eff.	46

[16] Quand les données sont nombreuses et qu'il est particulièrement important de ne travailler que sur des données certaines, la vérification se fait souvent par la technique du double (voire du triple) encodage. Deux (trois) personnes encodent les données de manière indépendante. Un (petit) programme informatique compare les fichiers de données et signale les différences. On apporte alors les corrections nécessaires à partir des renseignements de base. Bien sûr, si les encodeurs ont commis la même erreur au même endroit, celle-ci ne sera pas détectée. Mais cette probabilité est tellement faible ...

Pour cette population, on a calculé les paramètres suivants :
Moyenne : 156 Ecart-type : 6,16
Quartile 1 : 151 Médiane : 155 Quartile 3 : 159.

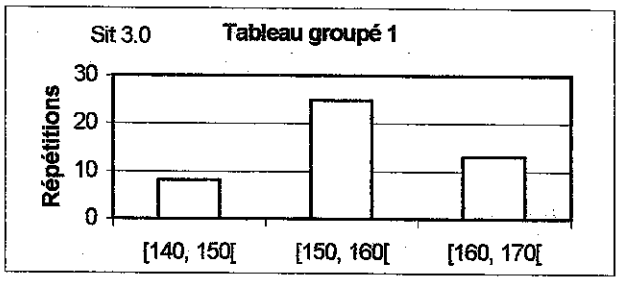
Notre exemple comporte 19 valeurs différentes. Généralement, un tel éparpillement ne sera pas jugé intéressant pour une bonne saisie globale de la situation considérée. Aussi va-t-on regrouper les valeurs dans quelques **classes**. On aura ainsi un **tableau groupé (recensé)** et les répétitions correspondantes[17]. Des graphiques appropriés les illustrent. Il n'est pas rare qu'on attribue comme valeur fictive à chaque individu le centre de la classe à laquelle il appartient.

Si on regroupe en un petit nombre de classes, on a une présentation de la situation très lisible mais très (trop) sommaire : beaucoup d'information est perdue. C'est souvent la nature de la situation étudiée qui fera juger qu'un regroupement est adéquat (ni trop sommaire, ni trop éparpillé).
Dans notre exemple, nous pourrions imaginer un regroupement sommaire en trois classes : (les cent quarante, les cent cinquante, les cent soixante)
ou un autre plus fin : {[140,145[, [145,150[, [150,155[, [155,160[, [160,165[, [165,170[}.

Il n'est pas dit que les classes doivent avoir toutes la même "largeur" (écart entre les extrémités). Il se pourrait même que certaines classes soient "ouvertes"[18]
(dans notre exemple, ..., 145[pour première classe, ou [160,... pour dernière classe),
mais ceci pose évidemment un problème pour déterminer le centre de la classe ! En outre, ceci mènera à distinguer **histogrammes** et **diagrammes en bâtonnets**. Provisoirement, nous nous en tenons à ces derniers : des rectangles qui figurent dans les diagrammes, on ne retient que la hauteur.

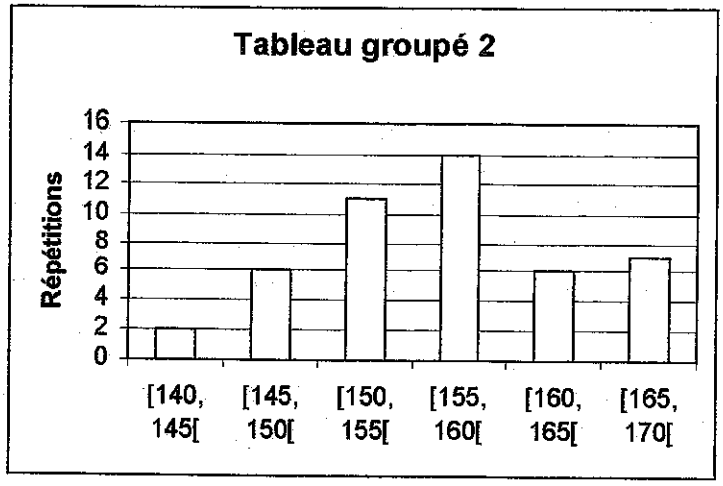
Voici ce que devient notre exemple regroupé de deux manières, et les diagrammes en bâtonnets correspondants.

Tabl. groupé 1	
Classes	**Rép.**
[140, 150[	8
[150, 160[	25
[160, 170[	13
Effectif	46



[17] Rappelons que la méthode "tiges-et-feuilles" produit presque immédiatement ce tableau groupé recensé.
[18] Il y a des situations où cela s'impose en quelque sorte : dans les revenus des personnes physiques, l'administration des impôts distingue des "tranches" imposables. La dernière est une tranche ouverte : revenus supérieurs à ...

Tabl. groupé 2	
Classes	Rép.
[140, 145[	2
[145, 150[	6
[150, 155[	11
[155, 160[	14
[160, 165[	6
[165, 170[	7
Effectif	46



Ex 3.0 - La moyenne et l'écart-type calculés à partir du tableau recensé vous ont été donnés dans un des tableaux précédents. Pour vous faire une première idée de l'erreur qu'on peut commettre en regroupant les valeurs, calculez ces deux variables à partir des tableaux groupés en prenant le milieu de chaque classe comme valeur caractéristique de chaque individu de cette classe.

Ex 3.1 - Avant de procéder au renouvellement d'une partie de sa flotte de véhicules, une société de taxis fait relever le compteur kilométrique des voitures à remplacer. Analysez ces données en appliquant les méthodes les plus adéquates.

129 153	126 371	146 251	144 741	132 647	128 941
113 871	137 243	137 841	127 801	119 221	146 828
132 521	121 352	105 352	143 123	151 225	131 143
134 249	120 372	120 842	124 976	123 941	117 305
139 925	141 171	134 222	129 742	118 258	113 245
138 273	135 211	136 421	133 825	125 053	127 742
126 747	116 742	125 274	139 004	143 276	134 953
130 127	135 644	132 229	121 762	133 421	122 873
130 005	128 845	133 742			

Suggestions -

Si vous préférez commencer par patauger seul dans le problème, omettez de lire ce qui suit.

Les unités, les dizaines ou même les centaines de kilomètres n'ont guère d'importance dans ce problème. On suggère donc de prendre pour tiges les deux premiers chiffres de gauche de chaque nombre, et pour feuilles le chiffre des milliers de kilomètres éventuellement arrondi à l'unité la plus proche en fonction des centaines. Recopier tous les chiffres serait très lourd et n'apporterait sans doute pas grand-chose.

Recommencer ensuite en partageant chaque classe en deux sous-classes "B" et "H" comme dans la situation précédente. Ici il semble plus judicieux de ne pas arrondir, mais de tronquer.

Sit 3.1 - Comparer deux populations

Lors d'une étude de santé (réalisée au Guatemala en 1964), on a relevé la teneur en cholestérol dans deux groupes de personnes. Les valeurs sont exprimées en mg/litre.

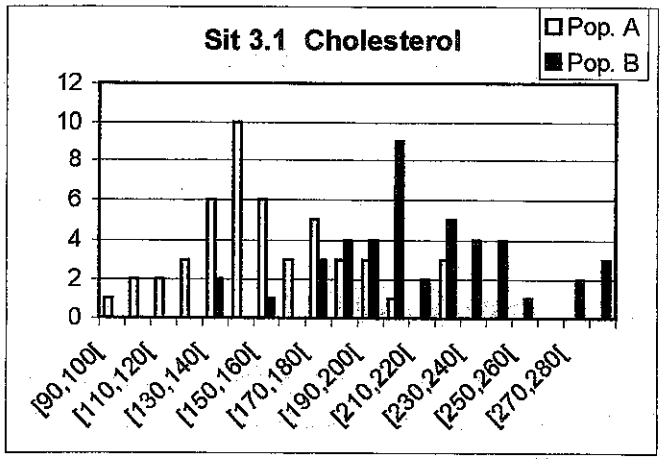
Population A. Milieu rural, bas revenus.

175	166	180	152	192	140	204	223	135	95	108	129	115
146	142	158	108	114	226	194	144	131	129	165	174	189
143	148	124	197	220	144	136	157	173	162	181	172	136
152	145	131	139	155	158	142	171	143.				

Population B. Milieu urbain, hauts revenus.

170	234	206	134	284	200	196	201	189	273	190	155	217
222	227	330	284	252	133	175	200	205	188	239	234	279
241	244	205	179	204	228	205	227	197	184	222	181	199
249	201	284	242	214	236.							

Ci-dessous, le graphique qui résume les données au terme d'une approche classique.



On voit que pour éviter de surcharger la figure, le logiciel n'indique en abscisse que les extrémités d'un intervalle sur deux. Le lecteur doit donc recomposer ce qui manque.

On a retiré du tableau la valeur 330 de la population B. Elle n'a pas été reprise parce qu'unique et très éloignée des autres observations. C'est évidemment un choix de la personne qui fait l'analyse des données que de considérer celle-là comme un artefact.

Voici maintenant l'analyse de la même situation faite par la méthode *tiges-et-feuilles*. Ici aussi, on a négligé la valeur 330.

	Popul. A Milieu rural, bas revenus			**Popul. B** Milieu urbain, hauts revenus		
1		5	**9**			
2		8 8	**10**			
2		4 5	**11**			
3		4 9 9	**12**			
6	9	1 6 6 1 5	**13**	4 3		2
10	3 2 5 4 8	3 4 2 6 0	**14**			0
6	8	5 2 7 8 2	**15**	5		1
3		2 5 6	**16**			0
5		1 2 3 4 5	**17**	0 5 9		3
3		1 9 0	**18**	9 8 4 1		4
3		7 4 2	**19**	6 0 7 9		4
1		4	**20**	6 0 1 0 5	5 4 5 1	9
0			**21**	7 4		2
3		0 6 3	**22**	2 7 8 7 2		5
			23	4 9 4 6		4
			24	2 1 4 9		4
			25	2		1
			26			0
			27	3 9		2
			28	4 4 4		3
12 11	*10 9 8 7 6*	*5 4 3 2 1*		*1 2 3 4 5*	*6 7 8 9 10*	
48 Eff.					Effectif :	44

Le rapport précise qu'il y a plus d'obèses dans la population B que dans la population A.

Ex 3.3 - La plus belle fille du monde ...

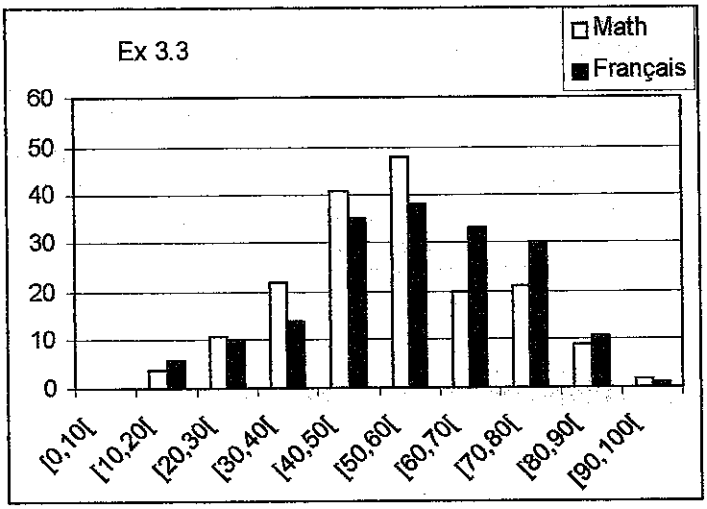
Favorablement impressionné par ce qui précède, un professeur de mathématique se dit que si la méthode permet de comparer deux populations étudiées selon un même critère, elle devrait pouvoir convenir aussi pour comparer les performances d'une seule population relativement à deux critères.

Il collecte donc les résultats annuels en mathématique et en français des 178 élèves des 7 classes de troisième de son établissement. Pour la facilité, les résultats ont été ramenés sur 100. La méthode *tiges-et-feuilles* permet de regrouper les élèves par classes de 10 points : [0,10[, [10,20[, ... , [90,100] .

En faveur de la validité statistique d'une telle étude, épinglons l'effectif assez grand; ajoutons-y quelques hypothèses susceptibles de faire admettre que les différences ou les paramètres individuels sont suffisamment gommés : pour chaque branche, il y a au moins trois professeurs en charge de ces classes, et aucun de ces enseignants n'a la réputation d'être ni particulièrement sévère ni exceptionnellement généreux dans sa cotation. Si les résultats (inventés, bien sûr) sont ce qui est illustré par le tableau groupé et le diagramme ci-contre, que peut-on dire de l'enseignement de la mathématique et du français dans cette école ?

	Math.	**Français**
[0,10[		
[10,20[	4	6
[20,30[	11	10
[30,40[	22	14
[40,50[	41	35
[50,60[	48	38
[60,70[	20	33
[70,80[	21	30
[80,90[	9	11
[90,100]	2	1
Effectif	178	178

On peut lire sur ce diagramme que les résultats en français sont globalement supérieurs aux résultats en mathématique. Il n'est pas sûr, au niveau du graphique, que la dispersion soit plus grande dans un cas que dans l'autre. Pour s'en faire une idée plus précise, on peut calculer la moyenne et l'écart-type, en attribuant pour score à chaque individu le centre de la classe à laquelle il appartient.



MATH. : moyenne = 52,9 Ecart-type : 16,8
FRANÇAIS : moyenne = 55,7 Ecart-type : 17,5.
Les résultats en mathématique sont donc un peu plus groupés que les résultats en français. A part cela, on ne peut sans doute pas dire grand-chose.

On pourrait trouver intéressant de savoir si ce sont LES MËMES élèves qui sont bons (ou moyens ou faibles) en mathématique ET en français[19]. Les méthodes d'analyse statistique rencontrées jusqu'ici ne permettent absolument pas de répondre à de telles questions. Comme la plus belle fille du monde ... Ceci ne signifie pas qu'il n'y a pas moyen de le savoir ! Répondre à de telles questions, c'est chercher à savoir s'il existe une **corrélation** entre deux variables

[19] Avant l'enseignement de la mathématique moderne, on faisait souvent appel à l'idée (non-scientifique) de "bosse des maths" pour expliquer qu'un individu assez médiocre en français, histoire,... pouvait être excellent en mathématique. A la fin des années 60, certains ont étudié la corrélation entre les résultats en mathématique et les résultats généraux, pour des élèves ayant reçu un enseignement mathématique traditionnel d'une part, moderne de l'autre. Le résultat de ces travaux très sporadiques indiquait une meilleure corrélation dans le cas de la mathématique moderne. Le seul résultat avéré de ces études a sans doute été d'amener PAPY à une certaine conviction que l'enseignement de la statistique dans le secondaire pouvait avoir des aspects intéressants.

d'une population ou d'un individu. Mais ceci est une autre histoire. Une telle recherche suppose en particulier qu'on dispose des scores individuels, c'est-à-dire du tableau brut.

Les statistiques montrent qu'on constate un plus grand nombre d'accidents mettant en cause des voitures roulant à vitesse modérée. Il ne faut certainement pas en conclure qu'il est recommandé de rouler "à tombeau ouvert". Une lecture plus attentive montre que puisqu'il y a plus de conducteurs qui sont raisonnable, il est normal que le nombre d'accidents soit plus élevé dans cette catégorie de conducteurs.

Les statistiques montrent qu'il est faux de dire que les mathématiques constituent un très grand facteur de redoublement car c'est dans les classes où le nombre d'heures de math est le plus élevé que l'on constate le moins de redoublements à imputer à cette matière.
Il est évident que ceux qui éprouvaient des difficultés dans cette matière ont été éliminés et se retrouvent donc ailleurs, et du coup, ceux qui restent dans ces sections sont ceux qui sont plus à l'aise avec cette matière. Outre qu'il est particulièrement utile pour le calcul de la moyenne, le tableau recensé permet une première visualisation : le ***diagramme (des répétitions) en bâtonnets*** *: en abscisse, les valeurs, en ordonnée, les répétitions*

D'après une étude statistique, on constate que la plupart des grands mathématiciens sont des fils aînés. Cela ne signifie pas qu'ils sont génétiquement plus doués, mais que, comme les familles sont de taille réduite, la proportion d'aînés est plus grande, notamment parce que dans toute famille ne comportant qu'un seul enfant, celui-ci est automatiquement aîné.

Une enquête a montré que dans une ville, on avait constaté que sur 10 ans, la consommation de lait et le nombre de cancers avaient augmenté dans les mêmes proportions. Faut-il en conclure que la consommation de lait provoque le cancer ou que l'affection du cancer provoque une soif de lait ?
On ne vous a pas tout dit : pendant cette période, la population a elle aussi augmenté dans les mêmes proportions que le lait ou les cancers.

B. Bâtonnets et "camemberts"

Sit 3.2 - L'hécatombe universitaire

Dans une université, 8934 étudiants[20] de première candidature se sont inscrits aux examens de la première session. A la fin de la session, on a fait un examen statistique des résultats, en regroupant les étudiants en quatre classes : "G" (grade), "S" (satisfaction), "Aj" (ajournés), "R" (refus ou session non présentée). Les données chiffrées sont consignées dans le tableau ci-contre.

Résultats	Répétitions
G	1099
S	2390
Aj	4730
R	715
Tot	8934

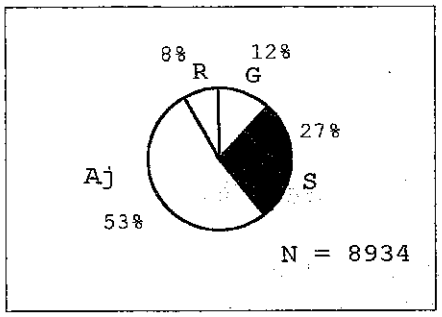
Ce tableau est à comparer au tableau brut des autres situations : tout s'y trouve mais on n'y voit pas grand chose.

[20] Les nombres de cet exemple sont inventés. Il se peut que dans la réalité, ce soit pire.
On remarquera que dans cet exemple, comme dans un certain nombre d'autres, la caractéristique des individus n'est pas numérique.

Ajoutons-y une colonne, dans laquelle nous inscrirons les **fréquences**, ce sera déjà tout différent. Les fréquences, souvent exprimées en pourcentages, sont les rapports des répétitions à l'effectif. On peut aussi bien les exprimer en décimaux : 1099/8934 = 0,123 . Ici, on a arrondi au pourcentage entier, ce qui, dans le contexte, est une précision bien suffisante pour une appréhension correcte de ce qui se passe[21].

	Répét.	Fréquences
G	1099	12
S	2390	27
Aj	4730	53
R	715	8
Tot	8934	100

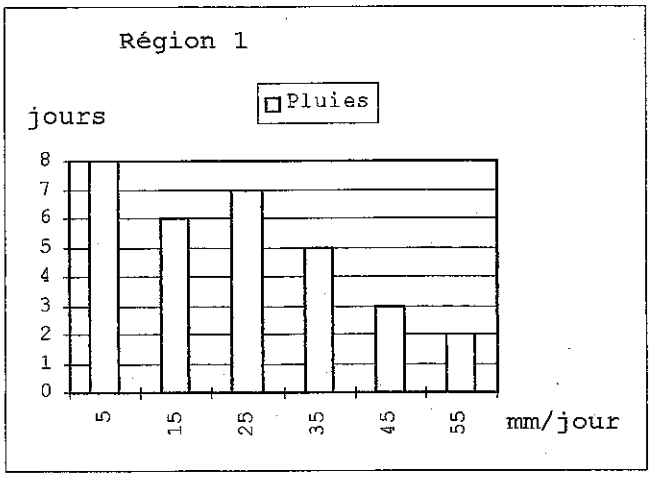
On peut bien sûr représenter graphiquement une telle situation par un diagramme en bâtonnets, mais le plus souvent, on préfère les "camemberts", précisément parce que le disque exprime graphiquement l'idée de référence à l'ensemble de la population que la notion de fréquence exprime de manière numérique. L'indication de l'effectif (N = 8934) établit une référence aux effectifs réels des différentes classes.

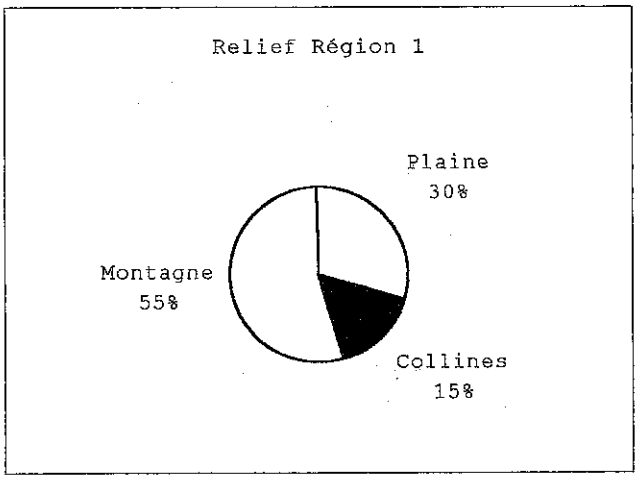


Ex 3.4.1 - Lire des graphiques

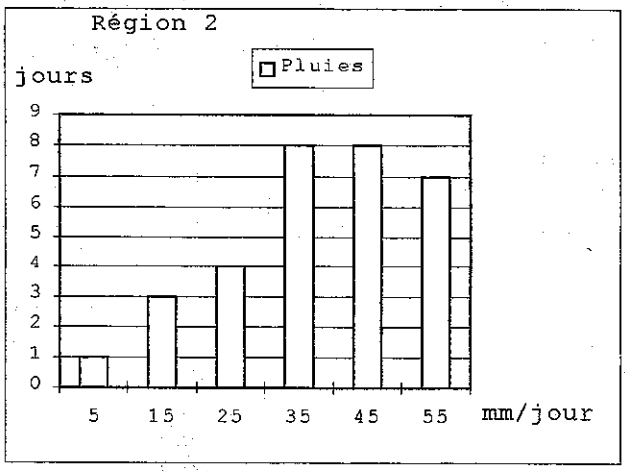
Comme c'est souvent le cas quand on consulte les media, une partie si pas l'entièreté de l'information est fournie sous forme de graphiques. En voici l'un ou l'autre exemple.

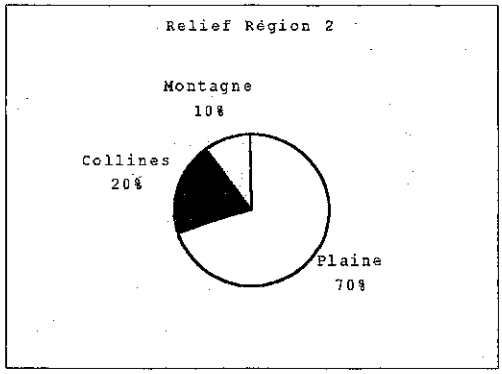
Vous souhaitez passer des vacances cyclistes en octobre et vous hésitez entre deux régions : la documentation recueillie, qui concerne les précipitations au cours du mois d'octobre et le relief, vous a été fournie sous forme de graphiques. Les précipitations y sont exprimées en nombre de millimètres d'eau par jour.





[21] Beaucoup de gens pensent que plus on aligne de chiffres et plus c'est scientifique. Rien n'est plus faux : d'une part, il ne sert à rien d'aligner des chiffres non significatifs (à quoi cela rime-t-il d'exprimer au mètre près la distance de Liège à Tournai ?); d'autre part, l'accumulation de chiffres, même exacts, nuit à la lisibilité. Ce qui est scientifique, c'est d'exprimer les choses avec la précision utile ou nécessaire dans le contexte où on est et avec le projet qu'on a.



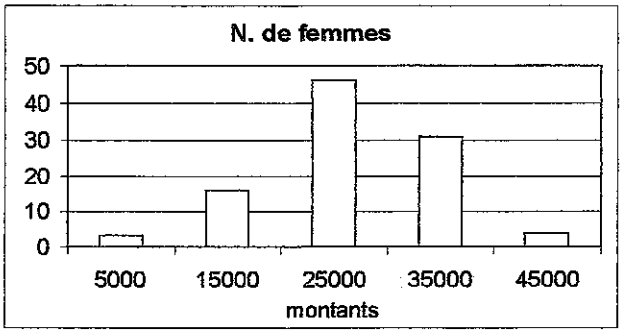


Il est bien évident qu'il n'existe pas de "BONNE REPONSE" pour cet exercice : ce qui est important (et sera éventuellement l'objet de l'évaluation), c'est la manière dont l'élève justifie son choix. Pour mener à bien son projet, il est amené à lire les graphiques qui expriment l'information récoltée. Les uns mettront l'accent sur la configuration du terrain, d'autres sur la qualité du temps.

Ex 3.4.2 - Lire tableaux et graphiques

On a interrogé un certain nombre d'hommes et de femmes pour savoir les montants qu'ils estiment nécessaires pour couvrir les dépenses mensuelles de ménage (alimentation, produits de lessive et d'entretien, eau, gaz, électricité, petites dépenses courantes, mais PAS le loyer, le téléphone, le remplissage de la cuve à mazout, les mensualités ou l'assurance de la voiture,...). Les résultats de ces enquêtes ont été consignés sous forme de tableau ou de graphique.

N. d'hommes	Montants
3	5000
14	15000
38	25000
16	35000
29	45000



Quelles conclusions tirer de cette enquête ? Peut-on penser qu'elle révèle des comportements différents entre les hommes et les femmes ? L'enquête a-t-elle été faite dans les mêmes milieux sociaux chez les hommes et chez les femmes ? Qui s'occupe en général de gérer les frais du ménage ?

C. Bâtonnets ou polygones ? Les "courbes" de température

Situation 3.3

Il est fréquent que les élèves de l'école primaire découvrent à la fois les notions de température, de relevé statistique et de moyenne grâce à des relevés de température dans la cour de l'école.

Sit 3.3.A -

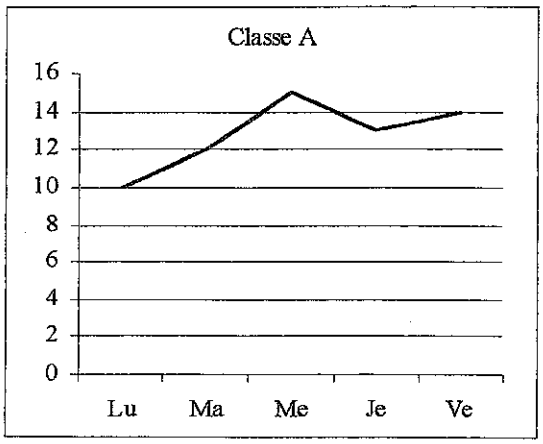
L'instituteur a été avec ses élèves relever les indications d'un thermomètre placé dans la cour de l'école, tous les jours d'une semaine à la même heure. Ils ont noté : 10°, 12°, 15°, 13°, 14°.

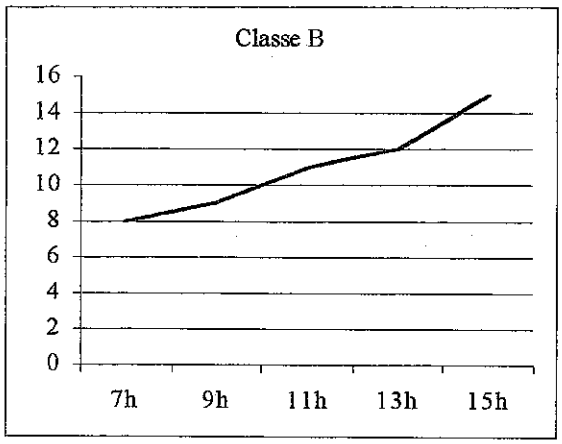
Sit 3.3.B -

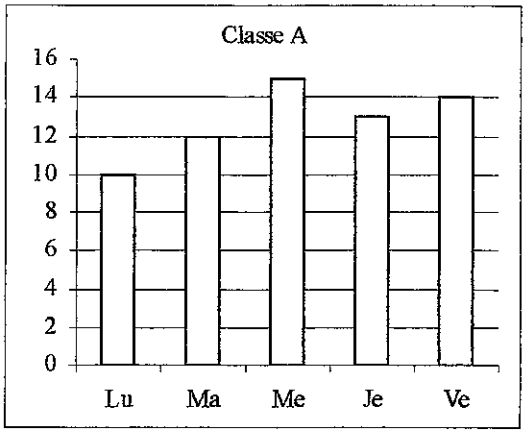
L'institutrice d'une autre classe a été relever ce qu'indiquait le thermomètre à 7h00, puis elle a envoyé des élèves noter la température de deux en deux heures jusqu'à 15 heures. Ils ont travaillé sur les températures suivantes : 8°, 9°, 11°, 12°, 15°.

Il est probable que dans les cahiers des élèves, on trouvera des graphiques comme les suivants.

Ces deux graphiques sont-ils également acceptables ?







Dans la situation B, la température a certainement varié de manière continue de la première à la dernière température relevées. Dans la situation A, au contraire, si les relevés ont eu lieu à 10h00, le plus probable est qu'ils ont monté jusqu'à un maximum dans l'après-midi pour redescendre jusqu'à un minimum au petit matin. La ligne continue est donc tout à fait adéquate pour B, mais pas du tout pour A : des bâtonnets seraient bien préférables.

Et dans notre exemple de la situation 3.0 (des poids ou des tailles regroupées par classes) ?

Cela dépend de ce que l'on veut exprimer.

Si nous voulons strictement traduire en graphique le tableau groupé, à savoir qu'il y a 2 individus qui ont (fictivement) la valeur 142,5, 6 individus qui ont (fictivement) la valeur 147,5, etc., le diagramme en bâtonnets que nous avons dessiné est l'outil adéquat. Mais si on veut en outre faire passer l'idée (l'hypothèse) que les valeurs se répartissent de manière régulière (linéaire) entre ces valeurs-pivots que sont les centres de classes, on préférera le diagramme en ligne polygonale.

On aura alors utilisé un outil **continu** pour étudier une situation **discrète**. Cette fiction supplémentaire qu'est la répartition continue des valeurs peut être tout à fait légitime. Elle l'est en particulier quand on regroupe en quelques classes une population importante : songeons par exemple aux tranches d'imposition. Il se peut aussi que cela induise des interprétations bizarres du phénomène étudié. A celui qui produit des statistiques d'utiliser en connaissance de cause les outils appropriés; à celui qui les lit d'être attentif à d'éventuels biaisages.

D. Bâtonnets et histogrammes

Sit 3.4 – Statistiques scolaires

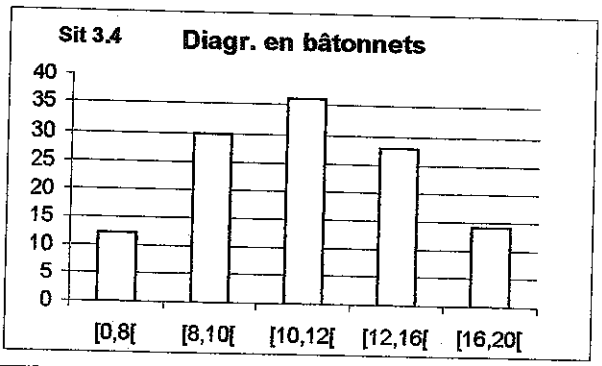
M. LEDIROT, directeur d'école, s'est penché sur les résultats d'un groupe de 120 élèves dans une branche donnée. Ces résultats sont exprimés sur 20. Il a jugé opportun de les regrouper en cinq classes d'amplitudes diverses :

- les échecs nets : [0,8[
- les petits échecs : [8,10[
- les succès un peu justes : [10,12[
- les succès nets : [12,16[
- les résultats brillants : [16,20].

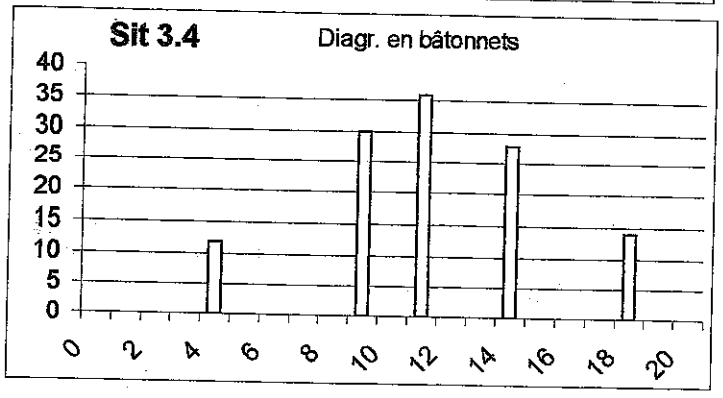
Il a consigné dans un tableau le fruit de cette première phase de son travail.

[0,8[	12
[8,10[	30
[10,12[	36
[12,16[	28
[16,20[	14
N =	120

Comme il a quelque maîtrise du tableur Excel, à l'aide de cet outil, il traduit en graphique son tableau groupé de données. En ayant "cliqué" sur "histogramme", il a obtenu la figure ci-contre.



Il ne peut s'empêcher de remarquer que cette présentation est bien insatisfaisante, en particulier parce que les classes sont représentées comme si elles avaient la même amplitude alors que ce n'est pas le cas. Bien sûr, les limites des classes sont écrites sous l'axe des abscisses, mais cette donnée n'est pas visualisée.



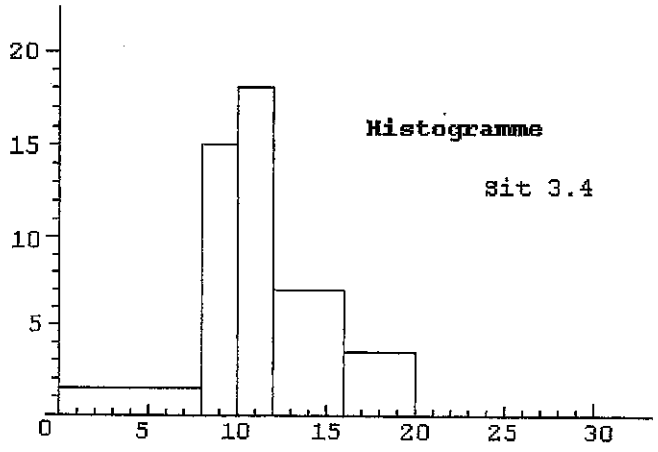
Qu'à cela ne tienne, M. LEDIROT sait qu'il faut parfois savoir ruser avec les machines pour les amener à faire ce qu'on a envie

qu'elles fassent, même si elles ont été programmées pour faire autre chose. Il fait l'hypothèse que tous les élèves d'une classe ont eu pour cote le milieu de la classe et que toutes les autres valeurs ont un effectif nul. Donc 12 élèves sont supposés avoir 4/20 et aucun n'a 0, 1, 2, 3, 5, 6, 7, ... Toujours en "cliquant" sur "histogramme", il obtient cette deuxième figure qui ne le satisfait pas plus que la précédente.

En désespoir de cause, il s'adresse à un professeur de statistique qui lui explique que quoi qu'en dise le tableur Excel, ces diagrammes sont des diagrammes **en bâtonnets** et NON des **histogrammes.**

Dans un **histogramme**, chaque classe et son effectif sont représentés par un rectangle. La largeur du rectangle est l'amplitude de la classe; la hauteur du rectangle est ajustée de manière que l'**aire** du rectangle soit proportionnelle à l'effectif de la classe.

Dans notre exemple, la première classe, qui a un effectif de 12, est donc représentée par un rectangle de largeur 8 (l'amplitude de la classe) et de hauteur 1,5, puisque $8 \times 1{,}5 = 12$.



La lecture des histogrammes est sans doute un peu moins aisée que celle des diagrammes en bâtonnets parce que pour évaluer les effectifs, il faut estimer l'aire des surfaces (ce qui n'est pas évident) ou faire un petit calcul. Mais c'est le mode de représentation adéquat (et courant) de tableaux groupés dont les classes n'ont pas la même amplitude. C'est pourquoi on n'imagine pas qu'un enseignement même élémentaire de la statistique le passe sous silence. On rencontre bien sûr une difficulté particulière si une des classes extrêmes est ouverte. On résout cette difficulté en cherchant le stratagème le plus parlant.

E. Tableaux et diagrammes cumulatifs. Polygone des répétitions cumulées Polygone des fréquences cumulées

Dans un certain nombre de situations, il est intéressant de savoir combien d'individus ont une caractéristique numérique inférieure (ou égale) à une caractéristique donnée.

Si on veut comparer le niveau de richesse de différentes régions ou sous-régions d'un pays et qu'on dispose de données statistiques relatives à l'IPP (Impôt des Personnes Physiques), on peut certes calculer la moyenne ou la médiane des revenus. Mais on peut aussi se demander combien de personnes vivent en dessous d'un seuil de pauvreté fixé, par exemple, à

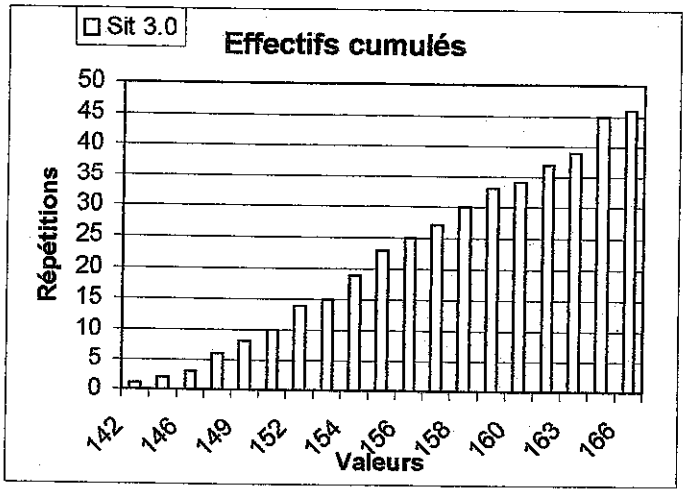
des revenus annuels bruts inférieurs à 5000 euros[22] ou toute autre somme qu'on estime significative pour exprimer par exemple un seuil de pauvreté.

Un professeur ou un directeur d'école peut se demander combien de leurs élèves ont un résultat global inférieur à 4/10.

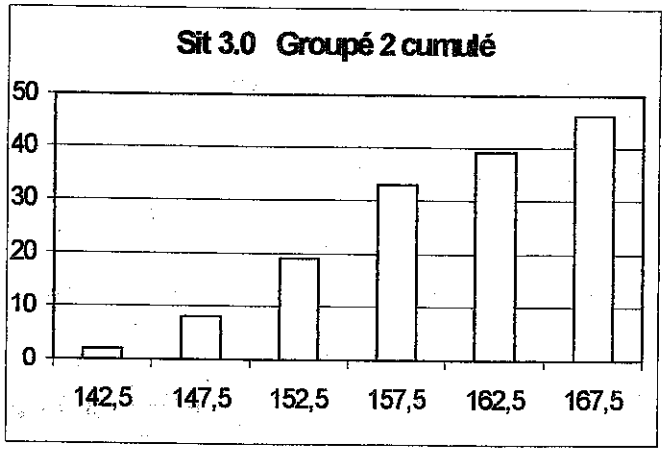
La réponse à ces questions est aisée si on dispose d'un **tableau des effectifs cumulés** ou d'un **diagramme cumulatif**. Le principe en est simple : au tableau groupé ou recensé, on ajoute une colonne dans laquelle on inscrit (avec ou sans décalage de ligne) la somme des répétitions rencontrées jusqu'à la valeur considérée.

Voici, relatifs à notre Sit 3.0, le tableau recensé auquel on a adjoint la colonne des effectifs cumulés (sans décalage de lignes), le diagramme des effectifs cumulés et le tableau des effectifs cumulés (avec décalage de lignes), fait d'après le deuxième tableau groupé.

T. à eff. Cumulés		
Val.	Répét	Cum.
142	1	1
143	1	2
146	1	3
148	3	6
149	2	8
151	2	10
152	4	14
153	1	15
154	4	19
155	4	23
156	2	25
157	2	27
158	3	30
159	3	33
160	1	34
162	3	37
163	2	39
165	6	45
166	1	46



T. groupé cumulé 2		
Classes	Rép.	Cumulé
[140, 145[	2	
		2
[145, 150[	6	
		8
[150, 155[	11	
		19
[155, 160[	14	
		33
[160, 165[	6	
		39
[165, 170[	7	
		46

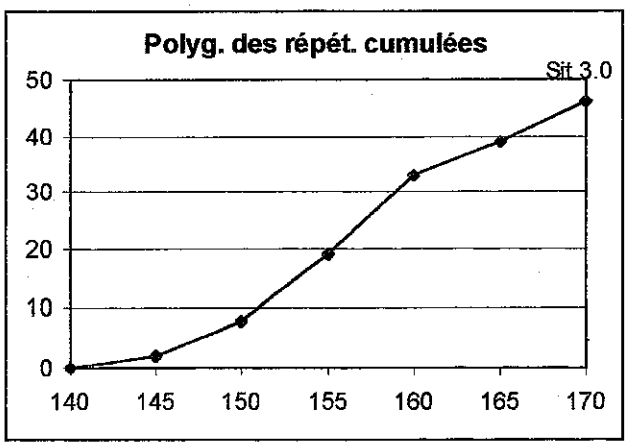


[22] Environ 200 000 francs belges.

Le plus souvent, c'est à partir du tableau groupé qu'on jugera intéressant de construire le diagramme cumulatif des répétitions, ou, plus pratique encore, des fréquences.

On devine que passer des bâtonnets à la ligne polygonale n'est pas innocent du point de vue du **sens**.

Dans le diagramme en bâtonnets, on se contente de montrer que si on va jusqu'à la classe dont le centre est indiqué en abscisse, on a comme effectif le nombre exprimé par la hauteur du bâtonnet. Il n'y a guère de sens à prendre des valeurs autres que celles qui sont numériquement indiquées en abscisse, sauf peut-être les extrémités des classes dont ces valeurs sont le centre.

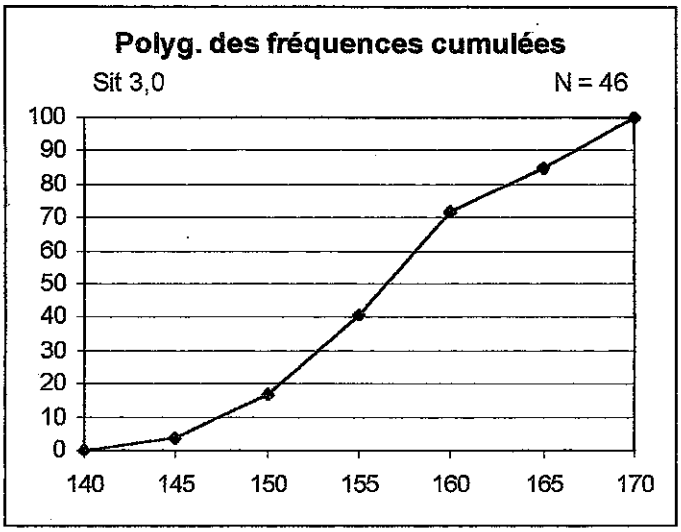


Dans le diagramme en ligne polygonale, par contre, on fait une hypothèse supplémentaire : non seulement, les sommets de la ligne polygonale apportent l'information que donnaient les bâtonnets, mais en outre, on imagine (ou on suppose) que les valeurs des individus se répartissent de manière linéaire entre ces points.

Sur un exemple de manuel comme notre Sit 3.0, cette hypothèse de répartition plus ou moins uniforme des valeurs entre les valeurs repères peut avoir l'air de forcer un peu la réalité pour l'obliger à rentrer dans des cadres mathématiques plus confortables, mais si on se réfère à des données dont les effectifs sont très importants comme l'étude de l'impôt des personnes physiques, l'hypothèse d'une répartition linéaire "par morceaux" des contributions individuelles entre la contribution la plus petite et celle (théorique) la plus grande est plus que raisonnable.

Dans de telles situations, qui sont le domaine de validité de l'outil statistique (n'oublions pas que *statistique* signifie science du traitement des données *de l'Etat*), étudier des données discrètes en recourant à des outils relatifs au continu est une façon de faire très courante.

On aura un autre outil bien commode si on remplace dans les ordonnées les répétitions par les fréquences; cela donne le **polygone des fréquences cumulées**. Comme sur toutes les figures où interviennent les fréquences, il est opportun de faire figurer la mention de l'effectif total qui établit une référence à l'effectif réel.



Cette figure est même le support de LA méthode de détermination de la médiane, des quartiles, des déciles, des centiles ou de tous autres α-

quantiles dont on peut avoir besoin pour étudier une situation donnée : vu que le graphique exprime une bijection d'un intervalle des abscisses dans l'intervalle [0,100], il n'y a aucun problème à faire appel à la fonction réciproque.

Ex 3.5.1 – En se référant aux tableaux et diagrammes ci-dessus, rechercher les renseignements suivants relatifs à cet exemple :
a) Combien y a-t-il de valeurs inférieures à 165 ?
b) Combien y a-t-il de valeurs inférieures ou égales à 158 ?
c) Jusqu'à quelle valeur faut-il monter pour avoir pris en compte 33 individus ?
d) Quels sont la médiane et les quartiles ?
Il est possible que certaines questions aient une réponse certaine à partir d'un tableau, mais pas d'un autre. En ceci comme en tant d'autres situations, il est important de se demander si on dispose des outils utiles ou nécessaires pour remplir la tâche demandée.

Ex 3.5.2 – Reprendre l'exemple des taxis (Ex 3.1) et se poser des questions comme :

a) Combien de taxis ont parcouru moins de 135 000 km ?
b) Jusqu'à quel kilométrage faut-il monter pour avoir pris en compte 60% de l'effectif ?
c) Par rapport à l'ensemble des véhicules remplacés, quel est le pourcentage des véhicules ayant parcouru au moins 140 000 km ?

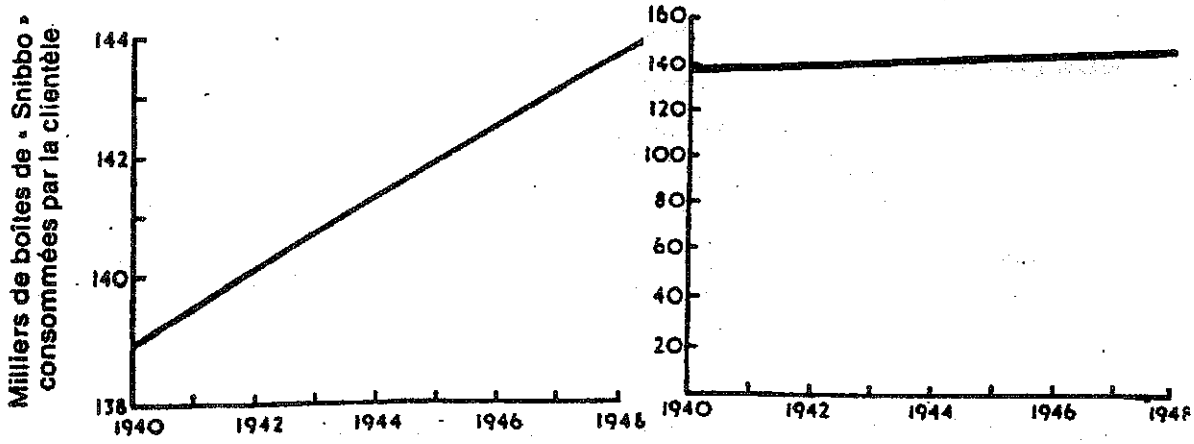
Ex 3.5.3 – Reprendre les données (Math-Français) de l'Ex 3.3 et tracer les polygones des fréquences cumulées relatifs à ces données. En tirer les médianes et les quartiles.
Ces indicateurs donnent-ils une meilleure vue sur la situation étudiée que la moyenne et l'écart-type ?

CHAPITRE IV. Des dessins qui « mentent » et des conclusions hâtives.

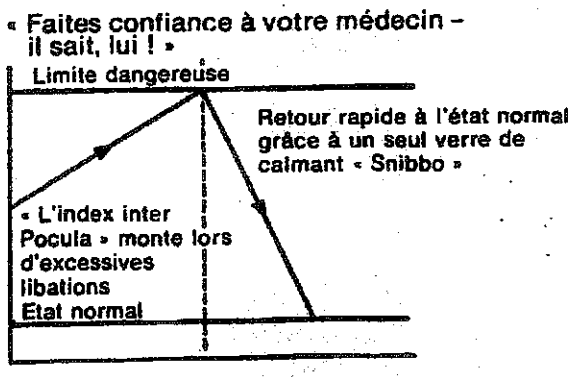
Le fait de représenter un phénomène au moyen d'un dessin est un bon moyen d'en donner une vision globale mais en comptant sur une certaine paresse du lecteur, voici quelques moyens subtils de tromper...ceux qui se laissent faire.

1.Escamotage de l'origine.

Le snibbo a été, en son temps, préconisé pour atténuer les effets néfastes d'une ingestion exagérée d'alcool.

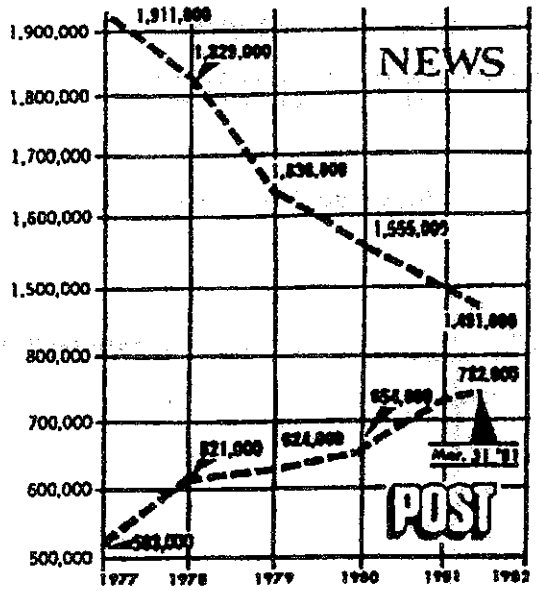


2. Graphique sans échelle.

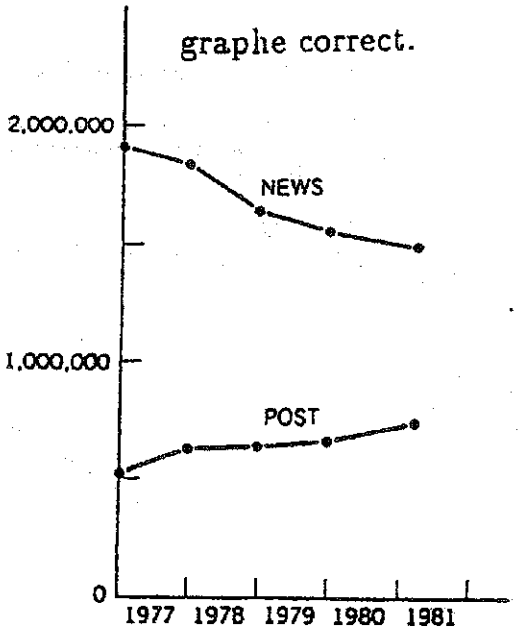


Ce graphique dépourvu d'échelle ne permet pas d'évaluer la réelle importance de la substance snibbo quant à son pouvoir de neutraliser rapidement une ingestion exagérée d'alcool.

3. Escamotage de l'origine et trafic de l'axe Oy.

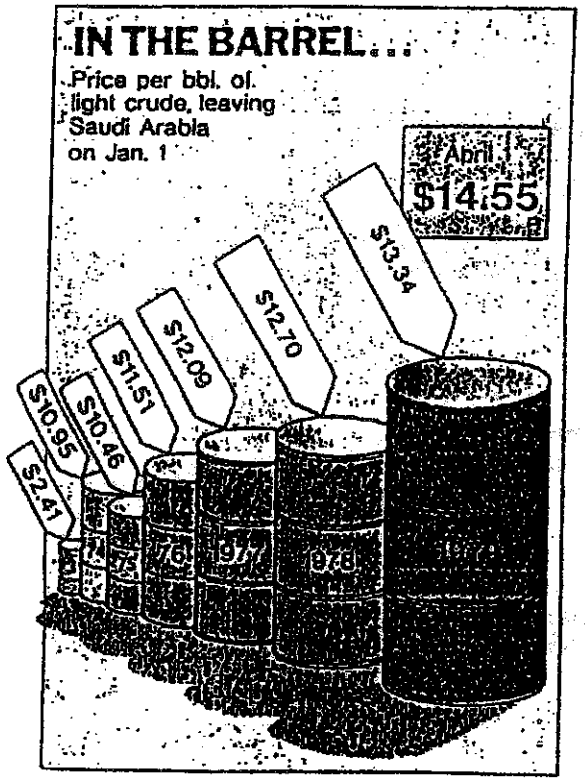


graphe correct.



: Evolution des ventes de deux journaux américains.

4. Rapport de hauteurs et rapport de surfaces ou de volumes



Variation du prix du pétrole

de 1973 à 1979.

5. Modification de la période d'observation.

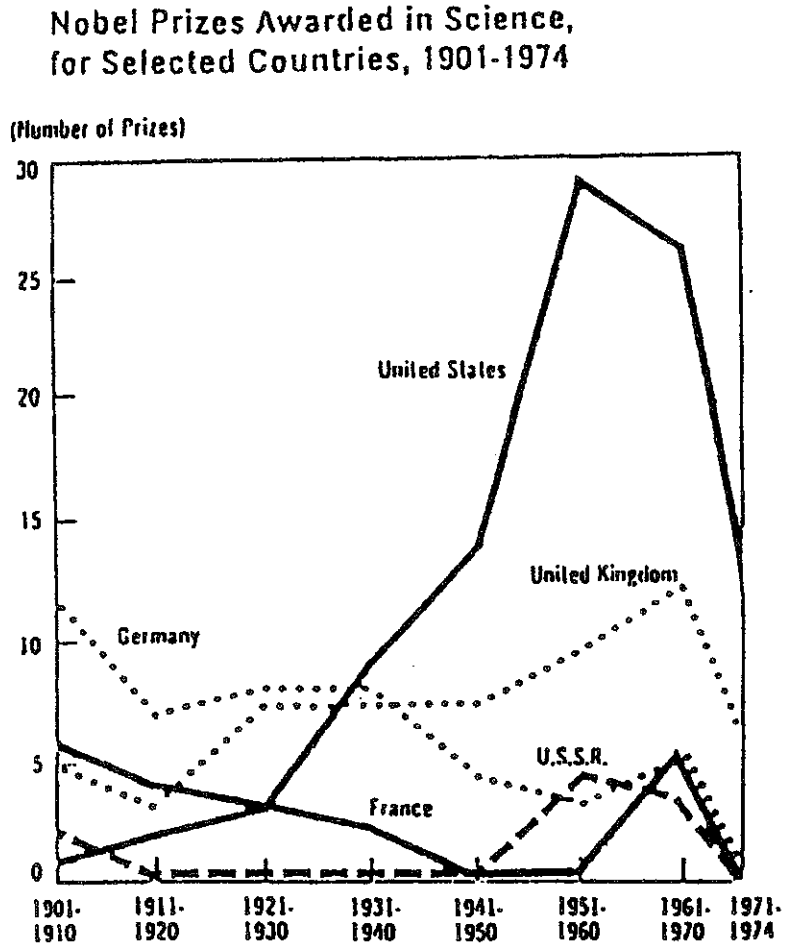


Figure 5: Attribution de prix Nobel.

Annexe 1 : Le calcul des α-quantiles

Plusieurs auteurs[v] regrettent la non-standardisation des notations. S'il n'y avait que cela !...

En réalité, ce sont les définitions des notions elles-mêmes qui sont spécifiques aux auteurs d'ouvrages ou de logiciels, avec les incohérences qu'on imagine.

Cela ne fait pas le bonheur des enseignants, puisque si d'une part ils se réfèrent à un manuel pour formuler des définitions et les appliquent à un exemple, et que d'autre part ils confient le même exemple à un logiciel, ils risquent fort de ne pas obtenir les mêmes réponses. Comme on le verra, on rencontre d'autres difficultés en essayant de retrouver de manière inductive les définitions auxquelles se réfère un logiciel.

1. Apories de la statistique : quelques généralités

Avant d'attaquer le problème de manière technique, essayons de l'éclairer par quelques considérations plus proches du sens commun.

On a insisté au chapitre 2 sur le fait que la donnée d'une valeur centrale, la moyenne en particulier, ne suffit pas toujours -loin de là - à dire quelque chose de valable sur un ensemble de données. Peut-être serait-il intéressant de réfléchir de manière analogue aux conditions de validité de cet autre outil statistique que sont les notions de médiane, quartiles, ... Pour ce faire quelques exemples pourront être éclairants.

Soit P une population de n individus caractérisés par la même valeur a. Il va de soi que la médiane, les quartiles, les déciles et les centiles sont tous égaux à a; on ferait peut-être mieux de dire qu'ils n'existent pas : il n'existe pas de valeurs ou ensembles de valeurs permettant de partitionner cette population en 2, 4, 10 ou 100 sous-ensembles de cardinaux égaux (ou presque[vi]). On peut certes partitionner cette population (pour autant qu'elle soit assez nombreuse), mais pas sur la base des valeurs qui caractérisent les individus.

Trop absurde cet exemple ? Voyons cet autre.

La Communauté Française de Belgique a imposé (impose ?) à ses professeurs d'indiquer dans le bulletin remis aux élèves la moyenne de la classe pour chaque cours. Beaucoup s'y sont opposés, non que la moyenne n'apporte aucun renseignement intéressant, mais parce que la seule moyenne est un indicateur trop pauvre, pouvant en outre conduire à trop d'interprétations abusives ou à des comportements peu souhaités comme par exemple la manipulation des cotes d'un cours ou d'une classe pour en rapprocher la moyenne de celle des autres classes; de telles manipulations feraient donc disparaître certains symptômes qu'il serait peut-être intéressant de prendre en considération mais qui pourraient aussi apporter des désagréments à certains : surtout pas de vagues !

En s'inspirant de nos considérations de Sit 2.2.1, on pourrait penser qu'en donnant les premier et troisième quartiles, on fournira une information meilleure qu'en communiquant la moyenne. Voyons cela sur un exemple.

Anx 1 – Ex 1 - Les cotes[23] classées en ordre croissant d'une classe de 23 élèves :
10, 11, 11, 12, 13, 13, 13, 13, 14, 14, 14, 14, 14, 15, 15, 15, 16, 16, 16, 16, 16, 17, 19.
La moyenne de ces nombres est 14,22....La médiane est 14 et les quartiles sont 13 et 16. L'intervalle interquartile est donc [13 , 16]. Cet intervalle comporte 17 individus (sur 23). Ce renseignement nous indique bien qu'on a affaire à une population fort groupée, mais il y a bien plus que 50% de l'effectif dans cet intervalle et bien moins que 25% qui ont plus que 16 ou moins que 13. Si on veut respecter cette règle des 25% - 50% - 25% (ou presque), il faut décider arbitrairement que certains qui ont 13 sont dans le mauvais quart et les autres dans la "bonne moyenne", et de même pour ceux qui ont 16 qu'il faut répartir entre les moyens et les supérieurs.

De telles situations sont loin d'être l'exception.

On s'aperçoit - mais ceci n'est pas une surprise - que l'outil statistique en général, celui-ci en particulier, ne fonctionne bien que si les effectifs sont suffisamment grands pour gommer certains "détails", ou de manière plus essentielle, pour que les notions signifient quelque chose. Quel sens y a-t-il à partager en cent sous-ensembles disjoints une population de quelques dizaines d'individus ?

Ces considérations suffisent sans doute à expliquer le "babélisme" des définitions pratiques relatives aux α-quantiles : à défaut d'une solution rationnelle qui s'impose à tous, chacun y va de sa manière de faire, certes raisonnable, mais qui n'aura pas été retenue par d'autres qui auront trouvé plus pratique, et tout aussi raisonnable, de faire autrement.

2. Apories de la statistique : un peu de technique

Tout le monde, ou presque, est d'accord pour donner de la médiane la définition suivante :
la MEDIANE est la valeur qui partage la population en deux sous-populations de même effectif.

L'ennui, c'est que dès qu'il s'agit de l'appliquer, les désaccords commencent à se manifester.

Prenons quelques exemples élémentaires et caricaturaux : il s'agit de classes de 19 ou 20 élèves qui ont participé à un contrôle. Les résultats sont supposés être les suivants :
Anx1 – Ex 2 – 20 élèves : 10 fois 4/20 et 10 fois 20/20.
Anx1 – Ex 3 – 19 élèves : 10 fois 4/20 et 9 fois 20/20.
Anx1 – Ex 4 – 19 élèves : 9 fois 4/20 et 10 fois 20/20.

Pour **Ex 2**, toutes les valeurs (entières ou non) comprises entre 4 et 20 partagent la population en deux sous-populations de même effectif. Quelle est donc la médiane ?

[23] Cet exemple n'a pas été inventé pour les besoins de la cause : il sort d'un vieux cahier de cotes d'un des rédacteurs de ce texte. Tout au plus a-t-il été choisi parmi d'autres pour sa valeur démonstrative, mais il n'a rien d'exceptionnel.

On trouve (au moins) deux tendances pour répondre à cette question.

Tendance A

De telles définitions se trouvent dans des manuels scolaires (Boutriaux-Lievens[vii], Paquet-Bouchat[viii], ...) ou dans des publications comme celles de H. BRENY[ix]. Avec des nuances qu'il y aura lieu d'éclairer, c'est aussi celles qui sont implicites dans Excel.

Cette perspective veut privilégier la notion de médiane comme valeur centrale.

Les valeurs étant classées en ordre croissant et non regroupées,

- si l'effectif est pair et vaut 2p, la MEDIANE est la moyenne arithmétique de la p-ième et de la (p+1)-ième valeur.

- si l'effectif est impair et vaut 2p+1, la MEDIANE est la (p+1)-ième valeur.

Dans les deux cas, il y a donc le même nombre d'objets dans la valeur est inférieure (ou égale) à la médiane que d'objets dont la valeur est supérieure (ou égale) à la médiane.

Dans nos exemples, cela donne :

Ex 2 – $m_A = 12$

Ex 3 – $m_A = 4$

Ex 4 – $m_A = 20$.

Une difficulté de cette définition est que sa généralisation aux α-quantiles n'apparaît pas tout à fait facile.

Tendance B

Cette définition est tirée de (ou inspirée par) l'Encyclopédie Universalis[x] et d'autres.

Les valeurs étant classées en ordre croissant et non regroupées, on part de la première (la plus basse) et on "monte" en dénombrant les valeurs rencontrées jusqu'à avoir pris en compte la moitié (ou la fraction α) de l'effectif. Si la moitié (la fraction α) de l'effectif n'est pas un nombre entier, en général on l'arrondit à l'unité supérieure.

On a donc :

La MEDIANE est la plus basse valeur de la variable telle que la moitié (éventuellement arrondie à l'unité supérieure) de la population soit caractérisée par des valeurs inférieures ou égales à cette valeur médiane.

Dans nos exemples, cela donne :

Ex 2 – $m_B = 4$

Ex 3 – $m_B = 4$

Ex 4 – $m_B = 20$.

La généralisation est inscrite dans le procédé même :

L'α-quantile est la plus basse valeur de la variable telle que la fraction-α (éventuellement arrondie à l'unité supérieure) de la population soit caractérisée par des valeurs inférieures ou égales à cette valeur.

Anx 1 – Ex 5 – Un autre exemple de "babélisme"

Les nombres sont : 10, 11, 12, 15, 20, 21, 32, 45.

En se référant aux définitions de la tendance A, on dira que les quartiles et la médiane sont :

	11,5 , 17,5 , 26,5.
Mais Excel donne :	11,75 , 17, 5 , 23,75.
Selon la tendance B, on a :	11, 15, 21.

Il est fort probable (mais c'est à vérifier) que si la population est "suffisamment" nombreuse et "raisonnablement"[24] répartie, ces différences seront minimes. Cette considération générale ne fait cependant pas l'affaire du professeur de l'enseignement secondaire qui doit choisir comment définir les notions auxquelles il initie ses élèves.

Les programmes - et beaucoup de gens sensés - déconseillent de faire du cours de statistique un cours de calcul ou de tracé de graphiques. L'idée est de se concentrer sur la signification de données en en confiant le traitement à des outils de calcul. Dans cet esprit, on a pensé à se référer à l'outil informatique qui semble le plus facilement accessible à la grande majorité des élèves : le tableur Excel. Bien entendu, rien n'empêche chaque professeur d'avoir recours à d'autres outils : calculatrices graphiques, autres logiciels éventuellement écrits par eux-mêmes.

Tout le monde comprendra, on l'espère, qu'un inventaire et une étude comparative des outils qui existent, même en se limitant aux principaux, représenteraient une tâche démesurée en comparaison de son intérêt. C'est bien pourquoi on ne l'a pas fait, et on ne compte pas le faire.

3. La solution préconisée

Avant d'explorer ce que fait Excel, rappelons ce que nous avons exposé à la fin du chapitre 3. Chaque fois que l'effectif est assez grand ET que l'hypothèse de distribution plus ou moins régulière sur les intervalles de l'ensemble des valeurs est raisonnable, il est indiqué de faire un diagramme des fréquences cumulées sous forme de polygone (ou de courbe). La détermination des α-quantiles ne pose alors aucun problème, ni théorique, ni pratique. En dehors de ces conditions, il est probable que l'utilisation de l'outil "médiane, quartiles, ... , α-quantiles" apportera plus de déboires que de satisfactions.

4. Que fait donc Excel ?

Malgré des heures passées à examiner les résultats proposés par le tableur pour les exemples présentés dans les pages qui suivent, on n'est pas parvenu à trouver la clef qui ouvre toutes les portes ! On a cependant pu écrire des formules qui donnent la solution dans un certain nombre de cas. Bonne récompense à qui fera mieux.

Il n'y a guère que la médiane qui ait une définition universellement admise : si l'effectif n est pair et vaut 2p, on prend pour médiane la moyenne arithmétique des p-ième et (p+1)-ième valeurs. Si n est impair et vaut 2p+1 , on prend pour médiane la (p+1)-ième valeur.

[24] Est-il besoin de préciser qu'aucun des mots entre guillemets n'est défini de manière stricte ?

Outre la fonction MEDIANE, Excel a les fonctions QUARTILE et CENTILE[25]. Pour le calcul des centiles, on obtient souvent (mais pas toujours) les réponses d'Excel en appliquant les formules qui suivent : pour trouver le j-ième centile, (ici noté **Cj**), on commence par calculer la fraction (j×n)/100 (c'est-à-dire j% de l'effectif).

Si cette fraction fournit un nombre entier **i**, la formule est $\mathbf{Cj = x_{i+1} - (j/100).\Delta x_i}$. Comme d'habitude, en se référant au tableau ordonné par valeurs croissantes, Δx_i désigne la différence $x_{i+1}-x_i$ entre la i-ième valeur et la suivante.

Si la fraction (j×n)/100 se traduit par un décimal, **i** désignant le plus petit majorant entier de cette fraction, la formule est $\mathbf{Cj = x_{i+1} - (j/100 + (i + (n \times j/100)) \Delta x_i}$.

Exemples

Tous ces exemples ont été créés à partir d'un tableau ordonné de 120 valeurs comprises entre 0 et 79 (inclus). Ce tableau s'appelle Tableau 120. Les variations ont été obtenues en supprimant une ou plusieurs valeurs à la fin du tableau; d'où les intitulés Tableau 119, Tableau 118, ...

Exemple 1 : Tableau 120; calcul des quartiles.
Q1 = C25 . n.j/100 = 30 = i . $x_{30} = 18$; $x_{31} = 19$.
C25 = 19 - 0,25 . 1 = 18,75.
(Plusieurs auteurs donneraient 18,5; d'autres 18.)
Q3 = C75 . n.j/100 = 90 = i . $x_{90} = 61$; $x_{91} = 62$.
C75 = 62 - 0,75 . 1 = 61,25.
(Plusieurs auteurs donneraient 61,5; d'autres 61.)

Exemple 2 : tableau 120, troisième décile.
0,3.120 = 36 . x_{36} =20 ; $x_{37} = 20$
C30 = 20 - 0,30 . 0 = 20 .

Exemple 3 : tableau 119, troisième quartile.
0,75 . 119 = 89,25 . $x_{90} = 61$; $x_{91} = 62$.
C75 = 62 - (0,75 + 0,75) = 60,5.
Pour qu'il y ait (au moins) 75% des individus inférieurs ou égaux au troisième quartile, on s'attendrait à ce que ce quartile soit 61. Excel place la barre de manière qu'il y ait 89 individus en dessous et 30 au-dessus.

[25] Attirons l'attention sur une incohérence de la syntaxe de ces deux fonctions. Pour les quartiles, on écrit QUARTILE(tableau; j) où j est un des entiers 0, 1, 2, 3, 4 . Pour les centiles, on écrit CENTILE(tableau; r) où r est un décimal à un ou deux chiffres. Donc par exemple : QUARTILE(tableau; 1) pour le premier quartile, et CENTILE(tableau; 0,37) pour le 37e centile (et non CENTILE(tableau; 37)). Il va de soi que la seule fonction CENTILE pourrait suffire.

Tableau 120

0	1	1	1	2	3	4	4	5	5	6	6
7	7	8	8	10	11	12	14	14	15	15	16
16	17	18	18	18	18	19	19	20	20	20	20
20	21	21	21	23	23	23	24	25	26	26	27
28	29	29	31	32	33	36	39	39	40	41	41
42	43	44	45	46	46	47	47	47	48	48	49
49	52	52	54	54	55	55	55	55	56	57	57
57	57	58	58	60	61	62	62	62	63	63	63
64	64	64	65	66	67	67	67	67	69	70	70
72	72	72	72	74	75	78	78	78	79	79	79

C 5	**C10**	**C15**	**C20**	**C25**	**C30**	**C35**	**C40**	**C45**	**C50**
3,95	6,9	11,85	16	18,75	20	23	27,6	34,65	41,5
C55	**C60**	**C65**	**C70**	**C75**	**C80**	**C85**	**C90**	**C95**	**Ctl100**
46,45	49	55	57	61,25	63,2	67	70,2	75,15	79

Tableau 119

0	1	1	1	2	3	4	4	5	5	6	6
7	7	8	8	10	11	12	14	14	15	15	16
16	17	18	18	18	18	19	19	20	20	20	20
20	21	21	21	23	23	23	24	25	26	26	27
28	29	29	31	32	33	36	39	39	40	41	41
42	43	44	45	46	46	47	47	47	48	48	49
49	52	52	54	54	55	55	55	55	56	57	57
57	57	58	58	60	61	62	62	62	63	63	63
64	64	64	65	66	67	67	67	67	69	70	70
72	72	72	72	74	75	78	78	78	79	79	

C 5	**C10**	**C15**	**C20**	**C25**	**C30**	**C35**	**C40**	**C45**	**C50**
3,9	6,8	11,7	16	18,5	20	23	27,2	33,3	41
C55	**C60**	**C65**	**C70**	**C75**	**C80**	**C85**	**C90**	**C95**	**Ctl100**
46	48,8	54,7	57	60,5	63	66,3	70	74,1	79

Exemple 4 : tableau 119, huitième décile.
$0,8 \cdot 119 = 95,2$. $x_{96} = 63$; $x_{97} = 64$.
Selon la formule, $C80 = 64 - (0,8+0,8).1 = 62,4$.
Excel dit que C80 = 63 !!!

Exemple 5 : Tableau 118 , C40
$0,4 \cdot 118 = 47,2$. $x_{48} = 27$; $x_{49} = 28$.
Comme dans l'exemple 3, on pourrait s'attendre, selon les définitions habituelles à ce que C40 soit au moins égal à x_{48} . Or on a : $C40 = 28 - (0,4 + 0,8) = 26,8$.

Tableau 118

0	1	1	1	2	3	4	4	5	5	6	6
7	7	8	8	10	11	12	14	14	15	15	16
16	17	18	18	18	18	19	19	20	20	20	20
20	21	21	21	23	23	23	24	25	26	26	27
28	29	29	31	32	33	36	39	39	40	41	41
42	43	44	45	46	46	47	47	47	48	48	49
49	52	52	54	54	55	55	55	55	56	57	57
57	57	58	58	60	61	62	62	62	63	63	63
64	64	64	65	66	67	67	67	67	69	70	70
72	72	72	72	74	75	78	78	78	79		

C 5	**C10**	**C15**	**C20**	**C25**	**C30**	**C35**	**C40**	**C45**	**C50**
3,85	6,7	11,55	16	18,25	20	23	26,8	32,65	41
C55	**C60**	**C65**	**C70**	**C75**	**C80**	**C85**	**C90**	**C95**	**Ctl100**
46	48,2	54,05	56,9	59,5	63	65,45	69,3	72,3	79

Tableau 117

0	1	1	1	2	3	4	4	5	5	6	6
7	7	8	8	10	11	12	14	14	15	15	16
16	17	18	18	18	18	19	19	20	20	20	20
20	21	21	21	23	23	23	24	25	26	26	27
28	29	29	31	32	33	36	39	39	40	41	41
42	43	44	45	46	46	47	47	47	48	48	49
49	52	52	54	54	55	55	55	55	56	57	57
57	57	58	58	60	61	62	62	62	63	63	63
64	64	64	65	66	67	67	67	67	69	70	70
72	72	72	72	74	75	78	78	78			

C 5	**C10**	**C15**	**C20**	**C25**	**C30**	**C35**	**C40**	**C45**	**C50**
3,8	6,6	11,4	16	18	20	23	26,4	32,2	41
C55	**C60**	**C65**	**C70**	**C75**	**C80**	**C85**	**C90**	**C95**	**Ctl100**
45,8	48	54	56,2	58	62,8	64,6	67,8	72	78

Exemple 6 : Tableau 117 , C68
Rien n'oblige à ne retenir que les centiles dont le numéro est un multiple de 5.
$0{,}68 \cdot 117 = 79{,}56$. $x_{80} = 55$; $x_{81} = 55$.
D'où C68 = 55 .
C'est aussi le résultat que donne Excel.

Exemples 7 et suivants :
On s'en voudrait de vous priver du plaisir d'en chercher vous-même.

Annexe 2 : Population – Echantillon
Première explication

Pourquoi utiliser σ_{n-1} au lieu de σ_n pour estimer l'écart-type d'une population par des mesures effectuées sur un échantillon ?

Si on prélève dans la population un échantillon comprenant un seul individu, il est évident que la taille de cet individu sera la moyenne des tailles de l'échantillon et que la variance sera nulle. Avec une observation, on ne peut rien conclure, et le renseignement sur la dispersion proviendra des autres observations, donc n-1 sur une population de n.

On appelle aussi n-1 le nombre de degrés de liberté car si l'échantillon comporte 2 observations, par exemple 178 et 184, la moyenne est 181 et les écarts se compensent (+3 et – 3). Si le premier écart est libre, le deuxième en dépend ; donc nous avons seulement 1 degré de liberté.

En général, pour un échantillon de taille n, alors que n-1 écarts sont libres, le dernier est lié par le fait que la somme des écarts à la moyenne vaut zéro.

Annexe 2 : Population – Echantillon
Deuxième explication

A propos d'estimateurs

1 Le problème posé

Une population d'objets, d'individus,... , fait l'objet d'une étude portant sur un caractère *quantitatif* : durée de vie d'une ampoule électrique, revenu annuel d'un individu, taille, poids .... A chaque élément de la population est associé un nombre réel. En considérant tous les individus, on obtient donc un tableau de nombres. Le problème qui est posé est d'en déterminer la moyenne m et la variance σ^2.

Nous venons de dire « en considérant tous les individus, on obtient donc un tableau de nombres ». Dans la pratique, il est généralement impossible de mesurer le caractère associé à chaque individu : ceux-ci sont trop nombreux ; ou bien en mesurant le caractère, on rend l'objet inutilisable (penser par exemple à la durée de vie d'une ampoule électrique). On est alors amené à recourir au calcul des probabilités.

Si nous choisissons au hasard un individu, la valeur du caractère qui lui correspond dépend aussi du hasard : c'est une *variable aléatoire* qui est notée classiquement X. Elle admet aussi une valeur moyenne et celle-ci est égale à la valeur moyenne m du caractère sur l'ensemble de la population. De même elle admet une variance égale à la variance σ^2 du caractère étudié sur l'ensemble de la population.

Estimer m, c'est estimer la valeur moyenne de X. Pour ce faire, on ne va pas prélever un seul individu dans la population, mais plusieurs, c'est-à-dire un *échantillon*, et on procède à la mesure du caractère sur chaque objet de cet échantillon.

Cette méthode pose de nombreux problèmes : combien d'objets prélever, comment les prélever, etc. Nous nous intéressons ici uniquement au problème d'estimer la moyenne et la variance de la population à partir des paramètres correspondants de l'échantillon.

Nous supposons donc en particulier **et ces hypothèses sont fondamentales** :

- que chaque objet de l'échantillon a été obtenu par un tirage aléatoire dans la population, tirage dans lequel tous les éléments de celle-ci ont la même probabilité d'être choisis,
- que les tirages successifs sont indépendants les uns des autres.

La deuxième hypothèse est sans doute contestable puisque toute population est finie et que tout objet qui a été tiré ne participe plus aux tirages suivants. Elle est raisonnable dans le cas de populations assez importantes. Il importe toutefois de ne pas oublier que nous élaborons un modèle de la réalité et que, comme tout modèle, celui-ci est approximatif.

2 L'échantillon

Après avoir choisi un échantillon de taille n, on mesure le caractère étudié sur chacun des éléments de l'échantillon. Notons $x_1, \ldots, x_n$ les valeurs trouvées.

Comme les éléments de l'échantillon résultent d'un tirage au hasard, les valeurs $x_1, \ldots, x_n$ dépendent elles-mêmes du hasard. Chaque élément de l'échantillon joue le même rôle que celui de l'unique individu qui avait été choisi au hasard au paragraphe précédent. Ainsi, au i^e élément de l'échantillon correspond une variable aléatoire, notée X_i dont la valeur est notée x_i. (En choisissant un autre échantillon, la même variable aléatoire X_i prendrait une autre valeur x_i.) Nous traduisons les hypothèses ci-dessus de la façon suivante :

- toutes les variables aléatoires $X_1, \ldots, X_n$ ont la même distribution de probabilité, en particulier leur moyenne (notée m) et leur variance (notée σ^2) sont celles de la population complète.
- ces variables aléatoires sont indépendantes les unes des autres : autrement dit quels que soient les réels $x_1, \ldots, x_n$, les événements notés $X_1 = x_1, \ldots, X_n = x_n$ sont deux à deux indépendants.

A partir des valeurs $x_1, \ldots, x_n$, on calcule

- la *moyenne-échantillon* : $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$,
- la *variance-échantillon* : ... deux formules sont en quelque sorte concurrentes :

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

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

Pourquoi deux formules pour la variance ? Quelle raison a-t-on de choisir l'une plutôt que l'autre ?

3 Le cas de la moyenne

Ce cas est simple. Il permet de comprendre la « philosophie » sous-jacente. Ce qui nous intéresse n'est pas vraiment la moyenne-échantillon $\bar{x}$, mais la moyenne de la population m.

L'échantillon a été tiré au hasard. On aurait pu trouver pour la valeur de la moyenne-échantillon une valeur différente de celle qui a été trouvée. La question qu'il est alors normal de se poser est de savoir comment peut varier cette moyenne-échantillon. Et de se demander quelle est la valeur moyenne de cette moyenne échantillon quand on regarde (abstraitement) tous les échantillons possibles .

La moyenne-échantillon est une valeur de la variable aléatoire $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Et la moyenne de $\bar{X}$ est aisée à calculer. Rappelons qu'on note $E(X)$ la moyenne (espérance

mathématique) d'une variable aléatoire X. Par exemple, pour tout i, $E(X_i) = m$. Rappelons aussi que la moyenne est ... un opérateur linéaire. Ainsi :

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} m = m$$

La moyenne de $\bar{X}$ est la moyenne m de la population. Cette constatation rend légitime l'utilisation de $\bar{x}$ pour estimer m. On dit que $\bar{x}$ est un *estimateur non biaisé* de m.

4 Le cas de la variance

On reprend ici la même idée : rechercher un estimateur non biaisé de σ^2.

Si on connaissait m, il n'y aurait pas d'hésitation à avoir : on utiliserait

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

En effet, puisque X_i est de moyenne m, et de variance σ^2, on a pour tout i,

$$E((X_i - m)^2) = \sigma^2$$

et par conséquent

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2\right) = \sigma^2$$

Le nombre s^2 est donc un estimateur non biaisé de σ^2. Malheureusement, on ne connaît pas m, que l'on est obligé de remplacer par $\bar{x}$. Or la variable aléatoire $\bar{X}$ n'est pas indépendante de $X_1, \ldots, X_n$.

Calculons $E((X_i - \bar{X})^2)$:

$$X_i - \bar{X} = (X_i - m) - (\bar{X} - m)$$

Donc

$$(X_i - \bar{X})^2 = (X_i - m)^2 - 2(X_i - m)(\bar{X} - m) + (\bar{X} - m)^2$$

Or

- $E((X_i - m)^2) = \sigma^2$ (X_i est de variance σ^2)
- $E((X_i - m)(\bar{X} - m)) = E((X_i - m)\sum_{j=1}^{n} \frac{1}{n}(X_j - m)) = \frac{1}{n}\sum_{j=1}^{n} E((X_i - m)(X_j - m)$
  Comme les variables X_i et X_j sont indépendantes pour $i \neq j$, les termes « rectangles » sont nuls et il reste

$$E((X_i - m)(\bar{X} - m)) = \frac{1}{n}E((X_i - m)^2) = \frac{\sigma^2}{n}$$

- $E((\bar{X}-m)^2) = E\left(\frac{1}{n^2}(\sum_{i=1}^n X_i - nm)^2\right) = \frac{1}{n^2}E\left((\sum_{i=1}^n (X_i-m))^2\right)$
  Ici aussi les termes « rectangles » sont nuls et il subsiste

$$E((\bar{X}-m)^2) = \frac{1}{n^2}E\left(\sum_{i=1}^n (X_i-m)^2\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

En remettant les pièces en place, on obtient

$$E\left((X_i-\bar{X})^2\right) = \sigma^2 - 2\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

Enfin

$$E\left(\sum_{i=1}^n (X_i-\bar{X})^2\right) = n\frac{n-1}{n}\sigma^2 = (n-1)\sigma^2$$

Par conséquent :

$s_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2$ est un estimateur biaisé de σ^2 puisque c'est une valeur d'une variable aléatoire $\frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2$ dont la moyenne vaut $\frac{n-1}{n}\sigma^2$.

Par contre $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i-\bar{x})^2$ est un estimateur non biaisé de σ^2.

Nous arrivons à deux conclusions :

1. Lorsqu'on ne dispose pas de renseignements concernant l'ensemble de la population mais que l'on doit se contenter d'un échantillon, il est raisonnable d'utiliser s_{n-1}^2 plutôt que s_n^2.
2. Si s_n^2 est biaisé, c'est parce que les variables aléatoires $X_1,\ldots, X_n$ et $\bar{X}$ ne sont pas indépendantes : quand on connaît les valeurs de n d'entre elles, on connaît automatiquement la valeur de la $n+1^e$. En utilisant $\bar{x}$ à la place de m (que l'on ne connaît pas), on utilise déjà un « degré de liberté ». On n'en dispose plus que de $n-1$. On retrouve ainsi les considérations intuitives déjà rencontrées.

Question à 5 francs (pardon à 0,1 euro) : Quand n est grand, cela vaut-il la peine de distinguer s_n^2 de s_{n-1}^2 ?

Annexe 3 : Super-condensé de la matière.

Ce dossier d'exploration didactique n'a pas été écrit dans l'optique d'une préparation directe des cours. Cependant, ce point de vue n'est pas étranger aux professeurs ou anciens professeurs que sont les auteurs du texte. Aussi proposent-ils ici ce que pourrait être un squelette de cours de statistique au second degré. On y trouvera aussi des références aux situations, exercices ou explications du texte du dossier.

Introduction et vocabulaire.

a) Population : une étude statistique cherche à tirer des conclusions pratiques de l'ensemble d'un grand nombre de données. Cet ensemble s'appelle **population** et les éléments qui le constituent portent le nom d'**individus**. Les statistiques n'ont pas pour but de traiter le cas de chaque individu mais de caractériser la population.

b) Caractère statistique : un caractère statistique qu'on appelle aussi variable statistique peut être par exemple :

1. la taille d'un groupe de garçons de 14 ans.
2. La marque des voitures stationnées place St AUBAIN.
3. Le kilométrage parcouru par chacune de ces voitures.
4. Le nombre de personnes par famille habitant dans le rue Des BRASSEURS

On distingue parmi ces caractères

- des caractères **qualitatifs**. (2)
- des caractères **quantitatifs (1, 3, 4)**
  - * **discrets** (4)
  - * **continus** (1, 3)

c) L'étude statistique consiste à présenter les données sous une forme permettant de dégager une image globale de la population puis de calculer des grandeurs permettant soit de comparer cette population avec une autre soit de situer un individu au sein d'une population.

La première partie se fera soit au moyen de tableaux soit par des graphiques.

Les nombres caractérisant la population sont

Des valeurs centrales : citons la moyenne arithmétique, le mode et la médiane.

Des paramètres de dispersion. Citons :

la variance et l'écart type, associés à la moyenne,

les quartiles (déciles, centiles), associés à la médiane.

I : Les tableaux. *(voir Ch.3. pp.33 et sv.)*

1.*Caractère statistique discret. (voir méthode «Tiges-et-feuilles», p.34)*

Exemple 1 : Nombre de pièces défectueuses par lot dans un ensemble de 1000 lots de pièces mécaniques.

n. de pièces défect. variable x	n. de lots effectifs n	effectifs cumulés N
0	300	300
1	365	665
2	214	879
3	83	962
4	23	985
5	0	985
6	15	1000
7 ou plus	0	1000
	1.000	

variable x	effectif n	effectif cumulé N
x1	n1	N1
x2	n2	N2
x3	n3	N3
x4	n4	N4
x5	n5	N5
x6	n6	N6
x7	n7	N7
x8	n8	N8
	n	

2. *Caractère statistique continu. (voir sit.3.4, p.46)*

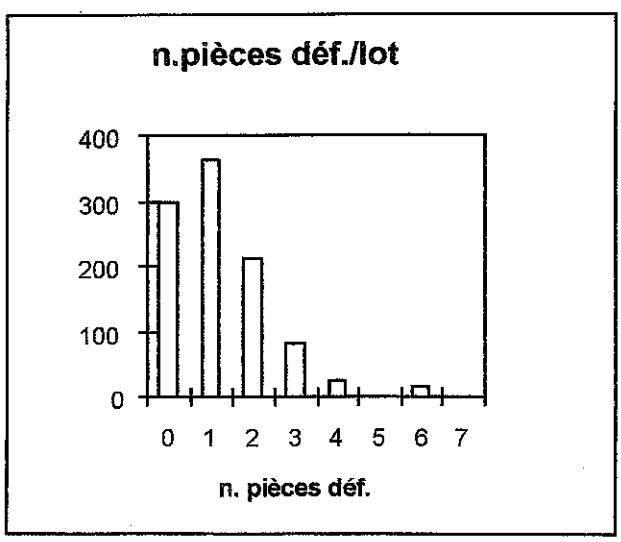
Exemple 2 : distance (en milliers de km) parcourue par des voitures neuves avant qu'elles ne soient revendues.

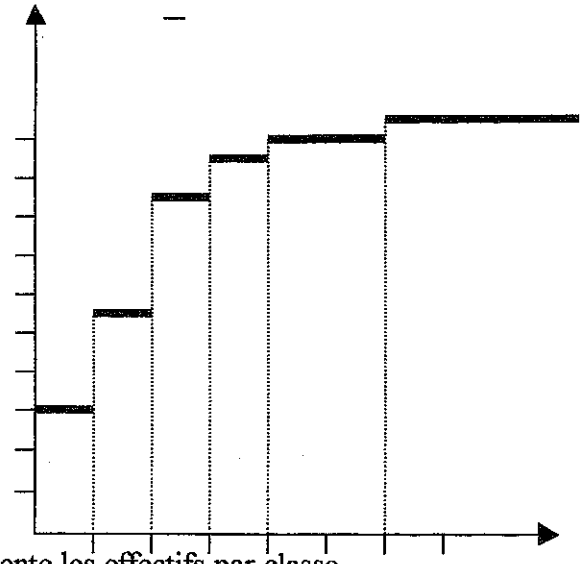
Dist.	n.voit. n	centres x	Eff.cum. N
75			0
	4	77,5	
80			4
	7	82,5	
85			11
	19	87,5	
90			30
	29	92,5	
95			59
	37	97,5	
100			96
	51	102,5	
105			147
	33	107,5	
110			180
	15	112,5	
115			195
	9	117,5	
120			204
	4	122,5	
125			208

limites de classes	Effectifs n	Centres x	Eff.cum. N
			N0
	n1	x1	
			N1
	n2	x2	
			N2
	n3	x3	
			N3
	n4	x4	
			N4
	n5	x5	
			N5
	n6	x6	
			N6
	n7	x7	
			N7
	n8	x8	
			N8
	n9	x9	
			N9
	n10	x10	
	n		N10=n

II : Les graphiques.

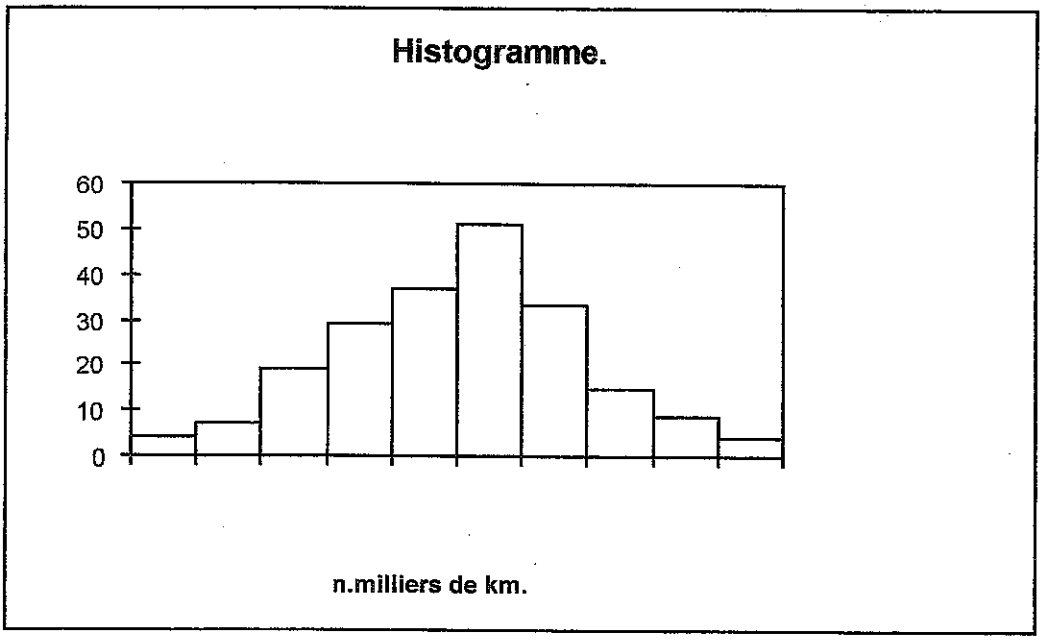
1. *Caractère statistique discret . (voir méthode « stem and leaf » p. 30 et 31)*



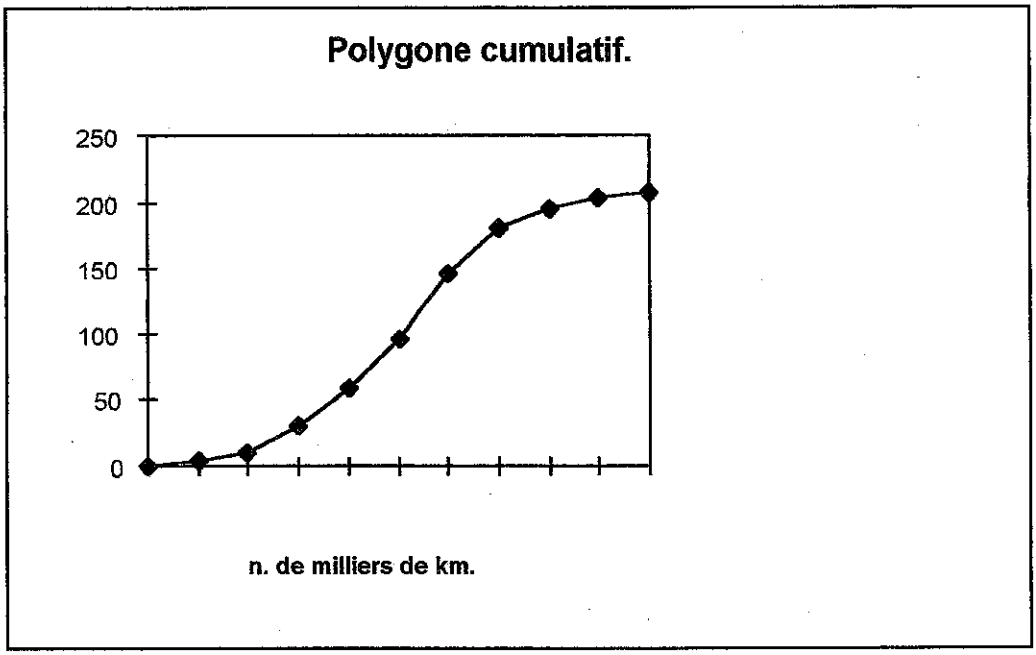


a) □Diagramme en bâtons. : ce diagramme représente les effectifs par classe.
b) Diagramme en escalier : ce diagramme représente les effectifs cumulés.

2. *Caractère statistique continu . (voir sit. 3.4, p.46)*



Histogramme. L'histogramme représente les effectifs par classe. Un individu est représenté par une surface et non par un segment. Cette remarque est importante dans le cas où la série statistique comporte des classes d'amplitudes différentes.



a) Polygone cumulatif : représente les effectifs cumulés.

Notations et quelques formules.

Les modalités du caractère sont représentées par la notation x.
Pour un caractère discret, x_i représente les valeurs du caractère pour les individus de la classe i.
Pour un caractère continu, x_i représente le centre de la classe i.

On désignera par n_i l'effectif de la classe i.
n désigne l'effectif total

$$n = n_1 + n_2 + ... + n_k = \sum_{j=1}^{k} n_j$$

On appelle **fréquence** :

$$f_i = \frac{n_i}{n} \quad et \quad f_1 + f_2 + ... + f_k = 1$$

IV : Les valeurs centrales. *(voir sit. 1.3, p 15)*

1.LE MODE M_0. (p 18)

C'est un nombre caractéristique de la classe la plus peuplée. Pour un caractère discret, on prend la valeur de x correspondant à cette classe et pour un caractère continu, on prend le centre de cette classe.

2.LA MEDIANE M. (p18, sit. 2.1.6, p 27)

C'est la valeur de la variable qui partage la série en 2 parties . La première partie est composée d'individus caractérisés par des valeurs de x inférieures à M et l'autre par des valeurs supérieures. Ces deux classes doivent comporter un effectif total au plus égal à la moitié de n.

$$N(x<M)\leq\frac{n}{2} \quad et \quad N(x>M)\leq\frac{n}{2}$$

Exemple simple : on considère la série de cotes suivante :

2, 4, 4, 5, 6, 6, 7, 8, 8, 8

Dans cette série, l'effectif total est de 10 et si l'on situe la médiane en 6, on constate que les classes caractérisées par des valeurs inférieures à 6 ou supérieures à 6 ont toutes deux un effectif inférieur à n/2 c'est à dire 5.
Dans le cas de l'exercice 1 (pièces défectueuses par lots), la médiane serait de 1.
Ceci concerne le cas d'une variable discrète.

Dans le cas d'une variable continue, il y a lieu de déterminer d'abord la classe médiane puis d'effectuer une interpolation au sein de cette classe.
Voici le détail de la détermination de la médiane dans l'exemple 2 (kilométrage).

$$96 \rightarrow 100$$
$$104 \rightarrow M \qquad M = 100+\frac{8}{51}5 \approx 100,8$$
$$147 \rightarrow 105$$

Il est évident que l'on utilise ici les effectifs cumulés.

3. LA MOYENNE ARITHMETIQUE. $\overline{X}$ (sit. 1.1 p13, sit 1.2 p15)

$$\overline{x}=\frac{1}{n}\sum_i n_i x_i$$

Dans l'exemple 1, elle vaut 1,224 et dans l'exemple 2 : 100,24.

V .Les paramètres de dispersion.
(sit. 2.1.1, 2.1.2, 2.1.3, 2.1.4, 2.1.5 pp 23 et sv.)

1.LES QUARTILES. Q1 ET Q3. (p29, sit. 2.2.1 p30)

Le principe de détermination des quartiles est semblable à celui de la médiane. Le premier quartile partage la série en deux groupes : celui des individus caractérisés par des valeurs de x plus petites que Q1 (premier quartile) qui doit avoir un effectif inférieur ou égal à n/4, et celui des individus caractérisés par des valeurs plus grandes qui doit avoir un effectif inférieur à 3n/4.

Le deuxième quartile est la médiane et le 3ème quartile est semblable au premier mais à l'autre extrémité de la série.

Exemple simple : n=20 n/4 = 5 3n/4 = 15

$$\underbrace{2,\ 2,\ 3,\ 3}_{Eff < \frac{n}{4}},\ 4,\ 4,\ 4,\ \underbrace{5,\ 6,\ 7,\ 7,\ 7,\ 7,\ 8,\ 8,\ 8,\ 8,\ 8,\ 8,\ 10}_{Eff < \frac{3n}{4}}$$

Q1= 4

$$\underbrace{2,\ 2,\ 3,\ 3,\ 4,\ 4,\ 4,\ 5,\ 6,\ 7,\ 7,\ 7,\ 7}_{Eff < \frac{3n}{4}},\ 8,\ 8,\ 8,\ 8,\ 8,\ 8,\ \underbrace{10}_{Eff < \frac{n}{4}}$$

Q3= 8

Dans certains cas, on est amené à diviser la série en 10 ou en 100. On parle alors de déciles et de centiles.

Dans l'exemple 1, Q1=0 et Q3=2.

Dans l'exemple 2, on procède à une interpolation comme pour la médiane. : n/4=52 et 3n/4=156.

$$30 \to 90$$
$$52 \to Q1 \quad Q1 = 90 + \frac{22}{29}5 \approx 93,8$$
$$59 \to 95$$

$$147 \to 105$$
$$156 \to Q3 \quad Q3 = 105 + \frac{9}{33}5 \approx 106,4$$
$$180 \to 110$$

2..LA VARIANCE ET L'ECART TYPE. V ET σ.

Les quartiles ont le désavantage de ne pas tenir compte de toutes les données .
La variance est la moyenne arithmétique des carrés des écarts à la moyenne.

$$V = \frac{1}{n}\sum_i n_i (x_i - \bar{x})^2$$

Si on développe cette expression, on obtient :

$$V = \frac{1}{n}\sum_i n_i x^2{}_i - \bar{x}^2$$

L'écart-type est la racine carrée de la variance : $\sigma = \sqrt{V}$

Dans l'exemple 1, la variance est égale à
(365*0+365*1+214*4+83*9+23*16+0*25+15*36)/1000-(1,224^2)=1,38
et l'écart-type à 1,17 .

Dans l'exemple 2, la variance est égale à
(4*6006,25*4+7*6806,25+19*7656,25+29*8556,25+37*9506,25+51*10506,25+33*11556,25+15*12656,25+9*13806,25+4*15006,25)/208-(100,24)^2=2108300-10048,=87,9.
Et l'écart-type à 9,4.

Moyennes pondérées

Christian Van Hooste, *A. R. Vauban, Charleroi*

Les réflexions de Toto, le gars sympa

Aujourd'hui est un bien mauvais jour pour Toto. Il vient de recevoir son bulletin et, une fois de plus, celui-ci n'est pas une invitation à l'Hymne à la Joie. Toto n'en est évidemment pas surpris; le bulletin reflète parfaitement son travail, ou plutôt son manque de travail. Il est seulement dépité. Il est vrai qu'une erreur (c'est humain), une erreur heureuse, aurait pu se glisser dans ce fichu périodique ...

Tout en retournant chez lui, à pied, le cartable, léger, bien accroché sur le haut du dos, Toto rumine donc son dépit. Mais, petit à petit, pas à pas, une autre idée taraude l'esprit de notre bonhomme. Et cela le rend perplexe. Puisqu'il n'est pas le meilleur élève de la classe — et il ne le revendique sûrement pas — qui l'est donc? Quelques noms de condisciples lui viennent alors en tête : Elisabeth, François, Mathieu, ... D'ailleurs, il a vu leurs bulletins et se souvient très bien des notes que ceux-ci ont obtenues dans certaines branches dites importantes. De quoi rêver!

	Français	Math	Anglais
Élisabeth	6	8	9
François	10	6	7
Mathieu	7	9	7

En dehors de ses moments de rêverie, plus ou moins longs, mais tout à fait indispensables pour son équilibre psychologique, Toto reste attentif aux discours de ses professeurs. Il a ainsi entendu, maintes fois, le prof de français affirmer qu'« une bonne maîtrise de la langue maternelle est primordiale pour la compréhension des autres cours ».

À mi-chemin entre l'école et sa maison, cette phrase, ressurgissant comme une lueur, lui inspire alors l'idée qu'il faudrait attribuer cinquante pour cent des points pour le cours de français et partager les autres cinquante pour cent de manière égale entre les cours de math et d'anglais. Partant de là, Toto calcule la moyenne des trois ténors de la classe :

- Élisabeth $\dfrac{6\times 50+8\times 25+9\times 25}{100}=7,25$;
- François $\dfrac{10\times 50+6\times 25+7\times 25}{100}=8,25$;
- Mathieu $\dfrac{7\times 50+9\times 25+7\times 25}{100}=7,50.$

Incontestablement, se dit Toto, dont le calcul mental est un des points forts, François est le meilleur élève de la classe.

Mais, chemin faisant, notre héros se rappelle encore qu'entre deux équations, le prof de math aime bien insister sur le fait que, de nos jours, sans math, il n'y a pas d'issue, que toutes les branches scientifiques et même certaines autres, telle la psychologie, la sociologie ou l'économie, utilisent des modèles mathématiques pour décrire les phénomènes étudiés, enfin que les math devraient constituer cinquante pour cent du bagage intellectuel de tout un chacun. Sur ce, Toto se dit qu'il vaudrait mieux attribuer cinquante pour cent des points au cours de math et donner la même importance aux deux autres cours. Mais alors sera-ce encore François le meilleur élève de la classe, François qui n'est précisément pas le plus fort en math? Toto recalcule donc les moyennes :

- Élisabeth $\dfrac{6\times 25+8\times 50+9\times 25}{100}=7,75$;
- François $\dfrac{10\times 25+6\times 50+7\times 25}{100}=7,25$;
- Mathieu $\dfrac{7\times 25+9\times 50+7\times 25}{100}=8,00.$

Cette fois, il faut bien admettre, se dit Toto, que c'est Mathieu le premier de la classe.

Toutefois, un peu plus loin, au pied de la côte qui mène à sa maison, de nouveau, il entend la voix d'un de ses profs. C'est maintenant le prof de langue moderne qui parle : « Le monde entier parle anglais, tous les scientifiques publient les résultats de leurs recherches en anglais. Certes le français vous a permis de comprendre ce qu'on vous a enseigné jusqu'ici; oui, le cours de math vous a donné le moyen d'apprendre les lois de la physique, de la chimie, de l'économie. Mais, à l'heure où l'on surfe sur Internet, il faut pouvoir communiquer avec le reste de la planète. Et, dans ce domaine, l'anglais est essentiel; c'est cinquante pour cent de votre potentiel d'ouverture vers l'extérieur ».

Pour la seconde fois, Toto revoit son jugement : « Si j'attribue cinquante pour cent des points pour le cours d'anglais et l'autre partie des points, en deux parts égales, pour les math et le français, qui va être le roi de la classe? » Toto se remet alors à calculer :

- Élisabeth $\dfrac{6\times 25+8\times 25+9\times 50}{100}=8,00$;

- François $\dfrac{10 \times 25 + 6 \times 25 + 7 \times 50}{100} = 7,50$;
- Mathieu $\dfrac{7 \times 25 + 9 \times 25 + 7 \times 50}{100} = 7,50.$

≪Pas d'hésitation possible, ce n'est pas un roi, c'est une reine, c'est Elisabeth, la reine de la classe!≫, conclut Toto.

Arrivé au seuil de la porte de sa maison, tout à coup, Toto est pris d'un doute. ≪Et si le cours de gymn était le plus important de tous les cours; sans une bonne santé, on n'est rien. Tout compte fait, c'est peut-être moi le "king" de la classe; j'ai une note de 10 en gymn. alors que les trois autres, les forts en thème, n'ont que 5 ... ≫

Mais, Toto n'aura probablement pas le temps de recommencer un autre calcul, car, derrière la porte, une dure réalité l'attend : son père est là, impatient de voir arriver le bulletin.

Les moyennes pondérées, invention de Toto et des mathématiciens

Étant donné n nombres $x_1, x_2, \ldots, x_n$ leur **moyenne arithmétique** est égale au quotient de leur somme par n. Ainsi qu'on peut le voir sur une machine à calculer, la moyenne arithmétique de ces nombres est généralement notée $\overline{x}$:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Mais, notre héros Toto vient de nous montrer que l'on pouvait calculer la moyenne des nombres $x_1, x_2, \ldots, x_n$ d'une autre manière. À chacun de ces nombres, on attribue un certain poids :

$$\alpha_1 \text{ pour } x_1,\ \alpha_2 \text{ pour } x_2, \ldots \alpha_n \text{ pour } x_n$$

ces poids étant des réels positifs. On multiplie alors chaque nombre par son poids et on additionne le tout; enfin, on divise le résultat obtenu par la somme des poids. Cela donne une **moyenne pondérée** [1] des nombres $x_1, x_2, \ldots, x_n$:

$$m = \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n}{\alpha_1 + \alpha_2 + \cdots + \alpha_n}$$

[1] Pondéré vient de la famille de mots latins *pondus, ponderis* (qui signifie le poids) et *ponderare* (qui signifie peser, estimer, évaluer).

Évidemment, cette moyenne n'est pas unique; elle varie en fonction des poids $\alpha_1, \alpha_2, \ldots, \alpha_n$ attribués aux nombres $x_1, x_2, \ldots, x_n$. Ces poids peuvent être fixés arbitrairement, à la manière de notre ami Toto. Il y a donc autant de moyennes pondérées des nombres $x_1, x_2, \ldots, x_n$ que de façons de choisir leurs poids, soit une infinité.

Parmi les moyennes pondérées de ces n nombres, il y a, entre autres, leur moyenne arithmétique $\overline{x}$. Pour l'obtenir, il suffit tout simplement d'attribuer un même poids α à chaque nombre :

$$\begin{aligned} m &= \frac{\alpha x_1 + \alpha x_2 + \cdots + \alpha x_n}{\alpha + \alpha + \cdots + \alpha} \\ &= \frac{\alpha(x_1 + x_2 + \cdots + x_n)}{n\alpha} \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n} = \overline{x} \end{aligned}$$

Ainsi, classer les élèves d'un groupe d'après la moyenne arithmétique de leurs notes n'est pas plus significatif que de les classer d'après toute autre moyenne pondérée calculée avec ces mêmes notes, la moyenne arithmétique n'étant qu'une moyenne pondérée particulière. D'ailleurs, dans la classe de Toto, les calculs faits par celui-ci avec des pondérations différentes conduisent à des classements différents. Et si Toto avait eu le temps d'intégrer le cours de gymn. dans ses calculs, peut-être que, lui Toto, ...

Les poids attribués $\alpha_1, \alpha_2, \ldots, \alpha_n$ aux nombres $x_1, x_2, \ldots, x_n$, permettent de donner plus ou moins d'importance à ces nombres. Augmenter le poids d'un nombre, c'est renforcer son influence sur la moyenne; diminuer ce poids, c'est la réduire. Ainsi, les trois condisciples de Toto peuvent, chacun à leur tour, se parer du titre de meilleur élève de la classe parce que les pondérations successives choisies par Toto attribuent le plus grand poids à leur meilleure note, augmentant de la sorte considérablement l'influence de cette note et réduisant, par là même, celle des deux autres.

On pourrait même donner à un de ces nombres un poids nul; cela reviendrait à réduire à néant son influence dans le calcul de la moyenne. C'est ce que fait encore Toto en ne donnant aucun poids aux cours, autres que ceux de français, de math et d'anglais.

En admettant que les notes de Toto en français, math et anglais soit respectivement 5, 4 et 3, sauriez-vous, cher lecteur, en choisissant convenablement la pondération, faire en sorte que Toto soit considéré comme le ≪king ≫de la classe? Cela pourrait sans doute lui remonter le moral et, surtout, lui donner d'éventuels moyens de défense auprès de son père. Bien

entendu, il faut pour cela tenir compte de la note attribuée en gymn. sans quoi il y a peu d'espoir d'y arriver.

En effet, ami lecteur, vous pourriez démontrer que **toute moyenne pondérée (à poids positifs) de n nombres est toujours comprise entre le plus petit et le plus grand, de ces nombres.**

Soient p et g respectivement le plus petit et le plus grand des nombres $x_1, x_2, \ldots, x_n$.

Du fait que les poids $\alpha_1, \alpha_2, \ldots, \alpha_n$ attribués à ces nombres sont positifs, nous avons

$$\begin{aligned} m &= \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} \\ &\geqslant \frac{\alpha_1 p + \alpha_2 p + \cdots + \alpha_n p}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} = \\ &\frac{p(\alpha_1 + \alpha_2 + \cdots + \alpha_n)}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} = p \end{aligned}$$

et

$$\begin{aligned} m &= \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} \\ &\leqslant \frac{\alpha_1 g + \alpha_2 g + \cdots + \alpha_n g}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} = \\ &\frac{g(\alpha_1 + \alpha_2 + \cdots + \alpha_n)}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} = g \end{aligned}$$

donc $p \leqslant m \leqslant g$.

Pourriez-vous aussi démontrer que tout nombre compris entre le plus petit et le plus grand des nombres $x_1, x_2, \ldots, x_n$ peut être considéré comme une moyenne pondérée particulière de ces nombres?

Soient p et g respectivement le plus petit et le plus grand des nombres $x_1, x_2, \ldots, x_n$.

Nous devons démontrer que, pour tout nombre μ tel que $p \leqslant \mu \leqslant g$, il existe une pondération des nombres $x_1, x_2, \ldots, x_n$ qui donne une moyenne pondérée égale à μ.

Parmi les nombres $x_1, x_2, \ldots, x_n$, il en existe deux, x_i et x_j, respectivement égaux à p et à g. Pour créer une pondération qui répond à la question, commençons par attribuer un poids nul aux nombres $x_1, x_2, \ldots, x_n$ sauf à x_i et à x_j. À ces derniers, donnons comme poids, respectivement les réels positifs $\mu - p$ et $g - \mu$. La moyenne pondérée correspondante est alors égale à μ comme souhaité. En effet, nous avons

$$\begin{aligned} m &= \frac{(g-\mu)x_i + (\mu - p)x_j}{(\mu - p) + (g - \mu)} \\ &= \frac{(g-\mu)p + (\mu - p)g}{g - p} \\ &= \frac{\mu g - \mu p}{g - p} \\ &= \frac{\mu(g-p)}{g-p} = \mu. \end{aligned}$$

Pondérations équivalentes

Appelons pondération la suite $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ des poids respectivement attribués aux nombres $x_1, x_2, \ldots, x_n$, poids auxquels nous imposons deux conditions :

- chaque poids est un réel positif (éventuellement nul) ;
- la somme des poids n'est pas nulle.

Nous savons comment Toto a produit trois pondérations différentes pour les notes attribuées aux cours de français, de math et d'anglais et ce qu'il en a résulté quant au classement des élèves d'après leur moyenne (pondérée). Pondérer revêt donc un caractère subjectif et un changement de pondération modifie les moyennes et, en général, bouleverse un classement établi d'après ces moyennes. Cependant, une modification qui ne consisterait qu'à multiplier tous les poids d'une pondération donnée par un même nombre strictement positif n'affecterait en rien la moyenne pondérée. En effet, pour $k > 0$,

$$\frac{(k\alpha_1)x_1 + (k\alpha_2)x_2 + \cdots + (k\alpha_n)x_n}{k\alpha_1 + k\alpha_2 + \cdots + k\alpha_n} = \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n}{\alpha_1 + \alpha_2 + \cdots + \alpha_n}$$

Décidons d'appeler **pondérations équivalentes** de $x_1, x_2, \ldots, x_n$, deux pondérations $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ et $(\beta_1, \beta_2, \ldots, \beta_n)$ dont les poids sont proportionnels, c'est-à-dire telles que

$$\frac{\alpha_1}{\beta_1} = \frac{\alpha_2}{\beta_2} = \cdots = \frac{\alpha_n}{\beta_n}$$

et de noter cette équivalence de la manière suivante :

$$(\alpha_1, \alpha_2, \ldots, \alpha_n) \sim (\beta_1, \beta_2, \ldots, \beta_n)$$

À ce propos, il est intéressant de remarquer que **toute pondération est équivalente à une autre dont la somme des poids vaut 1**, que nous appellerons **pondération standardisée**.

De fait, posons $S = \alpha_1 + \alpha_2 + \cdots + \alpha_n$. Nous avons alors

$$(\alpha_1, \alpha_2, \ldots, \alpha_n) \sim \left(\frac{\alpha_1}{S}, \frac{\alpha_2}{S}, \cdots \frac{\alpha_n}{S}\right)$$

avec

$$\frac{\alpha_1}{S} + \frac{\alpha_2}{S} + \cdots + \frac{\alpha_n}{S} = \frac{S}{S} = 1$$

Ainsi, la première pondération créée par Toto $(50, 25, 25)$ est équivalente à $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$.

L'importance de chaque poids apparaît nettement mieux dans cette dernière pondération : on voit immédiatement que le cours de français y compte pour la moitié des points.

En statistique

En statistique, il arrive que souvent que l'on doive calculer la moyenne d'un ensemble de nombres parmi lesquels certaines valeurs se répètent. Plus une valeur se répète, plus elle prend de l'importance vis-à-vis des autres, plus son poids augmente. On appelle d'ailleurs **répétition** (ou **effectif**) d'une valeur, le nombre de fois que celle-ci apparaît dans l'ensemble statistique. En désignant par $x_1, x_2, \ldots, x_n$, les différentes valeurs et par $r_1, r_2, \ldots, r_n$ leurs répétitions respectives, la moyenne (arithmétique) de l'ensemble statistique est alors

$$\bar{x} = \frac{\sum_{k=1}^{n} r_k x_k}{\sum_{k=1}^{n} r_k} = \frac{r_1 x_1 + r_2 x_2 + \cdots + r_n x_n}{r_1 + r_2 + \cdots + r_n}$$

Celle-ci apparaît donc comme une **moyenne pondérée des différentes valeurs** de l'ensemble, issue de la pondération dans laquelle chacune valeur reçoit un poids égal à sa répétition.

Dans la pondération standardisée équivalente, le poids attribué à chaque valeur x_k est

$$f_k = \frac{r_k}{r_1 + r_2 + \cdots + r_n}.$$

En statistique, ce nombre s'appelle la **fréquence** de la valeur xk ; elle est donc égale au rapport entre sa répétition et l'effectif total de l'ensemble (somme de toutes les répétitions).

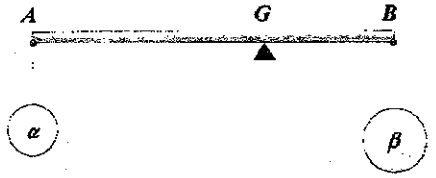
$$\bar{x} = \sum_{k=1}^{n} f_k x_k = f_1 x_1 + f_2 x_2 + \cdots + f_n x_n$$

En utilisant cette notion, la moyenne de l'ensemble statistique s'écrit

$$\bar{x} = \sum_{k=1}^{n} f_k x_k = f_1 x_1 + f_2 x_2 + \cdots + f_n x_n$$

En statique

Aux deux extrémités d'une barre (de poids négligeable), le prof de physique accroche deux masses de poids α et β et demande où se trouve le point d'équilibre du système (cf. fig.).



Il recommence l'expérience plusieurs fois, mesure, note les résultats et parvient à la loi (mathématique) suivante :

Le point d'équilibre est le point G tel que

$$\alpha|GA| = \beta|GB|.$$

Cette loi peut être reformulée en termes de vecteurs : $\alpha\overrightarrow{GA} = -\beta\overrightarrow{GB}$

où le signe $-$ se justifie par le fait qu'il faut tenir compte du sens des vecteurs. Elle s'écrit encore $\alpha\overrightarrow{GA} + \beta\overrightarrow{GB} = \overrightarrow{0}$.

Sur la droite AB, fixons un point O et transformons la relation ci-dessus :

$$\alpha(\overrightarrow{OA} - \overrightarrow{OG}) + \beta(\overrightarrow{OB} - \overrightarrow{OG} = \overrightarrow{0} \text{ ou}$$

$$\alpha\overrightarrow{OA} + \beta\overrightarrow{OB} = (\alpha + \beta)\overrightarrow{OG}$$

Munissons à présent la droite AB d'un axe dont O est l'origine et désignons par $x_1, x_2, \bar{x}$ les abscisses respectives des points A, B et G. De la relation vectorielle ci-dessus, nous tirons alors l'égalité algébrique : $\alpha x_1 + \beta x_2 = (\alpha + \beta)\bar{x}$ d'où il vient

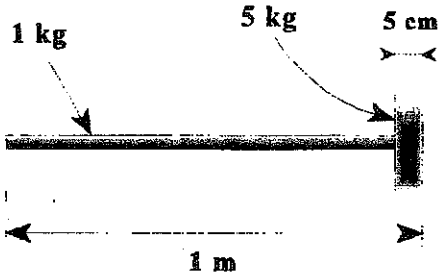
$$\bar{x} = \frac{\alpha x_1 + \beta x_2}{\alpha + \beta}$$

Ainsi, **l'abscisse du point d'équilibre est la moyenne pondérée des abscisses des points A et B, pourvues respectivement des poids α et β.**

Il est bon de noter que la position fixée pour l'origine de l'axe, ainsi que pour la graduation de celui-ci, peuvent être choisies de manière tout à fait arbitraire.

Notons encore que remplacer les poids par les masses ne change en rien la position d'équilibre puisque les masses et les poids sont proportionnels, le coefficient de proportionnalité étant égal à g (à peu près égal à $9,81$ en nos régions). Masses et poids conduisent à des pondérations équivalentes.

Des vérifications expérimentales de cette loi sont facilement réalisables chez soi. En voici un exemple. Pour enfoncer un pieu, on utilise un gros marteau (on appelle cela une «masse »dans le métier). En supposant que ce marteau est constitué comme le montre la figure ci-dessous, cherchons-en le point d'équilibre.



Plaçons l'origine 0 de l'axe sur l'extrémité libre du manche, choisissons comme unité de longueur le centimètre et désignons par x_1, x_2 les abscisses du centre de masse du manche, du centre de masse de la partie métallique. On doit avoir

$$\bar{x} = \frac{1 \times 50 + 5 \times 97,5}{6} = 89,6$$

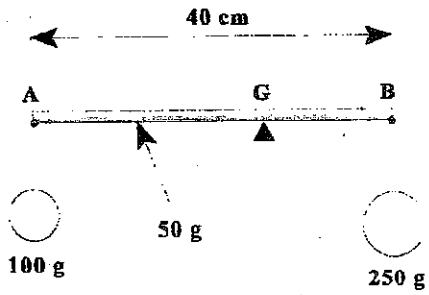
Il suffit donc de soutenir le marteau par le manche en son point situé à $89,6$ cm de l'extrémité libre pour le maintenir en équilibre.

Sortez votre mètre, votre balance et essayez !

L'avantage d'envisager ce problème de statique sous l'angle des moyennes pondérées réside dans la perspective d'une généralisation à n points matériels alignés pourvus de masses $\alpha_1, \alpha_2, \ldots, \alpha_n$. Si $x_1, x_2, \ldots, x_n$ sont les abscisses de ces points, l'abscisse du point d'équilibre est la moyenne pondérée de $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n}{\alpha_1 + \alpha_2 + \cdots + \alpha_n}$$

Ainsi, revenant à l'expérience faite au cours de physique, il est maintenant possible de tenir compte du poids de la barre reliant les points A et B. Pour cela, on considère que le poids de la barre agit en son centre et le problème revient alors à chercher le point d'équilibre pour un système de trois points.



Avec les données du schéma ci-dessus, la position de G par rapport à A est donnée par son abscisse

$$\bar{x} = \frac{100 \times 0 + 50 \times 20 + 250 \times 40}{100 + 50 + 250} = 27,5$$

En guise de conclusion

Établir un classement — nous l'avons vu grâce aux brillants calculs de notre ami Toto — des élèves d'une classe d'après la moyenne pondérée de leurs notes est subjectif et, par conséquent, n'a pas de valeur véritable. Donner le même poids à chaque branche, c'est choisir une pondération particulière ; elle n'a pas plus de valeur que les autres.

Si la moyenne pondérée a un caractère subjectif quand elle sert d'instrument d'appréciation, elle retrouve toute son objectivité mathématique dès qu'il s'agit de l'appliquer dans le domaine scientifique ; par exemple, pour situer le point d'équilibre d'un système de points matériels.

Ce qui lui donne un aspect subjectif dans le premier cas tient essentiellement à ce qu'on y peut choisir la pondération, puisque celle-ci n'existe pas initialement. Par contre, dans le second cas, la pondération est imposée par le contexte physique.

Quelques éléments de bibliographie

Peltier M., Rouche N., Manderick M., *Contremanuel de statistique et de probabilité*, Bruxelles, EVO, 1982, 198 pages.
Cet ouvrage n'est plus disponible chez l'éditeur, mais il l'est toujours au GEM, Chemin du Cyclotron 2, 1348 Louvain-la-Neuve.

Aïvazian S., *Etude statistique des dépendances*, Moscou, MIR, 1970, 236 pages.

Gardner M., *La magie des paradoxes*, Paris, Belin, 1980, Pour la science.

Wonnacott T.H. & R.J., Economie-Gestion-Sciences-Médecine (avec exercices d'application), Paris, Economica, 4e édition 1991, 918 pages.

Debaty P., *La statistique paramétrique appliquée aux sciences humaines*, Bruxelles, Ed. de l'Enseignement ou Paris, Ed. Universitaires, 1967, 254 pages.

Lethielleux M., *Statistique descriptive*, Paris, Dunod, 1998, 124 pages.

Klatzmann J., *Attention statistiques ! Comment en déjouer les pièges*, La Découverte, Paris, 1985, 126 pages.

Droesbecque J-J, *Eléments de statistique*, Paris, Ellipse,1992, 527 pages.

[i] TIMSS : Third International Mathematics ans Science Study, enquête organisée par l'IEA : International Association for the Evaluation of Educational Achievement.

[ii] A ce propos, on relira avec intérêt
- Garin M., Liesenborghs J., Marlier P., Evaluation, *Mathématique et Pédagogie*, 1996, 106, 43-51.
- Marlier P., Analyse structurale de l'évaluation scolaire, *Mathématique et Pédagogie*, 1996, 107, 25-39.

[iii] L'exemple est tiré d'un article du Bulletin de l'APMEP (Association des Professeurs de Mathématique de l'Enseignement Public (France)) :
Piednoir Jean-Louis, Les Statistiques, le calcul des probabilités et l'enseignement professionnel, *Bulletin de l'APMEP*, 1997, 413, 717-726.

[iv] Voir à ce sujet : A. Valette, Pourquoi la variance ? *in* Mathématique et Pédagogie, 12e année, N° 55, Bruxelles, SBPMef, 1986, pages 25 et suivantes.

[v] Par exemple, Breny H., *Théorie des Probabilités*, Bruxelles, Presses Universitaires de Bruxelles, Collection Frédérique, 1969, 170 pages, où on lit :
"*En statistique descriptive, le babélisme des notations règne en maître; ne t'étonne donc pas de trouver ailleurs des notations différentes des nôtres.*" (p.39).

[vi] Cette restriction, essentielle pour la bonne intelligence de la méthode, est systématiquement explicitée par des auteurs comme Bragard L. et Alexandre P., *Statistique descriptive à l'usage des Sciences humaines*, Liège, ULG, (date gardée secrète), 202 pages.

[vii] Boutriau J. et Lievens J., *Cours de 3e de l'enseignement secondaire*, Coll. Mathématique d'aujourd'hui, Liège, Dessain, 1971, 304 pages.

[viii] Paquet E. et Bouchat J.M., *Mathématique à l'usage des classes de 3e*, Liège et Bruxelles, Dessain et De Boeck, 1971, 507 pages.

[ix] Breny H., *Cours de calcul des probabilité et statistique, 1 Statistique descriptive*, Liège, Dessain, 1970, 70 pages.

[x] Encyclopédie Universalis, Art. *Statistique*, Vol 17, pp. 159 et sv., Paris, 1985.