Jean-Claude **GIRARD**
Daniel **GROS**
Paul **PLANCHETTE**
Jean-Claude **REGNIER**
René **THOMAS**

Enseigner la Statistique du CM à la Seconde

Pourquoi ? Comment ?

INSTITUT DE RECHERCHE
sur l'ENSEIGNEMENT DES MATHEMATIQUES
Académie de Lyon

Université Claude Bernard Lyon 1 - 43, bd du 11 Novembre 1918
69622 Villeurbanne Cedex

Enseigner la statistique du CM à la seconde. Pourquoi ? Comment ?

Malgré la place importante tenue par les statistiques dans les cursus post-bac, l'enseignement de cette partie du programme de mathématique est peu développée tant au collège qu'au lycée. Une première raison est que les enseignants n'en voient pas toujours l'utilité. D'autre part les concepts mis en jeu, malgré leur simplicité apparente, peuvent cacher des difficultés d'interprétation.

Cette brochure passe en revue les finalités de cet enseignement et le replace comme partie intégrante des mathématiques. Elle le situe dans la perspective d'une utilisation ultérieure et comme élément incontournable de la formation du citoyen.

Les auteurs analysent les difficultés mathématiques de divers concepts statistiques ainsi que certains problèmes d'ordre didactique. Ils proposent des idées d'activités de classe, pouvant également être mises en place dans les ateliers de parcours diversifiés, propres à susciter l'intérêt des élèves. Des ouvertures dépassant le niveau du lycée sont proposées de façon à situer les notions étudiées dans un cadre plus général.

Groupe Probabilités et Statistique

Jean Claude **GIRARD**,
IUFM de L'Académie de Lyon,
Centre local de St-Etienne

Daniel **GROS**,
Ecole St-Saens, St-Etienne

Paul **PLANCHETTE**,
Collège Marc Seguin, St-Etienne

Jean-Claude **REGNIER**,
Département des sciences de l'éducation,
Université Louis Lumière, Lyon 2

René **THOMAS**,
Collège Jean Dasté, St-Etienne

Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?

SOMMAIRE

Introduction - Pourquoi faire des statistiques ? - J. C. Girard & J. C. Régnier 1

1ère partie : Réflexions théoriques

Chapitre 1 - Réflexion sur les finalités de l'enseignement de la Statistique -J.C. Régnier 5
Chapitre 2 - Mise au point sur les histogrammes - J. C. Régnier 21
Chapitre 3 - Des diagrammes à l'histogramme - J. C. Girard 43
Chapitre 4 - De l'histogramme à la fonction densité de probabilité - J. C. Girard 57
Chapitre 5 - A bas la moyenne ! - J. C. Girard 61
Chapitre 6 - La médiane, pour quoi faire ? - J. C. Girard 85
Chapitre 7 - Danger : approximations ! J. C. Régnier 99
Chapitre 8 - De la vérité auto proclamée à la vraisemblance reconnue - J. C. Régnier 107

2ème partie : Activités de classe

Chapitre 9 - Quelques idées d'utilisation d'un article de presse - J. C. Girard 119
Chapitre 10 - Lire un article de journal -J. C. Régnier 127
Chapitre 11- Evaluation sur la lecture de graphiques -P. Planchette & D. Gros 135
Chapitre 12 - Une enquête statistique au service de la proportionnalité - D. Gros 141
Chapitre 13 - Un travail transdisciplinaire - P. Planchette 157
Chapitre 14- Le diagramme en tiges et feuilles en 6ème, pourquoi pas ?- P. Planchette 175
Chapitre 15- Le concours "A vos Stats" - R. Thomas 183
Chapitre 16- Expérimentation sur la prise de décision en situation incertaine - R. Thomas & J. C. Régnier 189

Bibliographie 202

Illustrations des pages intérieures et de la couverture : Paul Planchette

Pourquoi "faire des statistiques" ?

Jean Claude Girard
Jean Claude Régnier

Au collège ou au lycée, le temps consacré à l'enseignement de la statistique est pris sur le temps global alloué aux mathématiques. Chaque professeur de cette discipline peut, sans doute légitimement, se demander si ce temps est bien employé ?

De la réponse apportée dépend l'importance qu'il va donner à cette partie du programme : de "rien", s'il la juge inutile, trop facile ou hors du cadre mathématique, à ..."pas grand-chose", parce qu'on n'a déjà pas assez de temps pour traiter le reste du programme !

Si l'on considère que "faire des statistiques" se réduit à apprendre quelques algorithmes sans pouvoir les justifier ni leur trouver des application intéressantes à court terme, alors effectivement, on fait toujours assez de statistiques.

L'enseignement de la statistique a, fort heureusement, bien d'autres objectifs. Ainsi, comme le précise le rapport Cockcroft[1] *la statistique n'est pas simplement un ensemble de techniques, c'est une disposition d'esprit, une manière d'appréhender les données, qui reconnaît notamment l'existence de l'incertitude et de la variabilité de l'information et de la collecte des données. Elle permet de prendre des décisions dans une situation d'incertitude.* Pour aller plus loin, il apparaît que

la pensée statistique est une des formes de l'intelligence.

Cette forme de pensée se retrouve dans toutes les sciences. En effet, la statistique propose une autre approche de la réalité, complémentaire de celle qui se fonde sur un fonctionnement déterministe dans lequel tout est certain et calculable.

Dans un autre domaine, on admet fort bien que l'enseignement de la géométrie n'a pas pour unique objectif de constituer un ensemble de connaissances applicables concrètement et immédiatement au monde sensible, mais qu'il intègre également, et principalement une formation à la logique et au raisonnement déductif. De la même

[1] *Mathematical counts. Report of the committee of inquiry into the teaching of mathematics in schools*, Londres, Her Majesty's Stationnary Office, 1982.

façon, l'enseignement de la statistique devrait se donner comme but, au-delà de l'acquisition de concepts et de techniques, l'apprentissage du raisonnement inductif, la sensibilisation au concept de variabilité des résultats dans la répétition d'une même expérience et l'initiation à la modélisation, c'est-à-dire l'inférence, à partir d'une série limitée d'observations, d'une loi générale *vraisemblable*.

Les outils habituellement présentés en statistique tels que la moyenne, l'écart type, les fréquences, les diagrammes, etc..., bien que peu performants quand ils sont pris isolément, concourent toutefois à la construction de concepts plus généraux comme ceux de test d'hypothèse, d'estimation d'un paramètre ou d'une loi de fréquence. Ils sont aussi l'occasion d'une familiarisation aux phénomènes aléatoires. Sans cette sensibilisation, il sera difficile aux élèves d'appréhender plus tard le calcul des probabilités et plus généralement la modélisation des situations incertaines. Le passage des modèles déterministes aux modèles privilégiant une description probabiliste de la réalité constitue, pour les élèves, un saut conceptuel aussi important que le passage des traitements numériques de l'école primaire aux traitements algébriques des équations, au collège, puis des fonctions, au lycée. Pour cette raison, il convient de donner aux élèves la possibilité de s'y préparer sur une période qui soit assez longue et qui pourrait débuter à l'école primaire. D'autre part,

la formation du citoyen, critique et réfléchi, ne peut faire l'économie de l'apprentissage de la pensée statistique.

Notre époque est marquée par un grand développement de la communication (générale, scientifique ou économique). Les textes diffusés abondent de données statistiques, et présentent assez systématiquement une synthèse utilisant le calcul de certains paramètres ou la construction de tableaux, courbes, diagrammes, etc... Les statistiques sont donc une forme de langage sans lequel on risque de rester à l'écart. Ce langage demande une formation sans laquelle on peut être amener à accepter sans discussion tout raisonnement basé (quelque fois abusivement) sur des satistiques ou, au contraire, à dénier systématiquement leur vadilité. La formation scolaire a donc son rôle à jouer dans ce domaine. Ainsi, à côté de l'apprentissage des techniques de construction des différents graphiques, il paraît important de prévoir des situations d'apprentissage de leur lecture pour qu'ils ne restent pas "une façon de dire des évidences aux ignorants". Débattre autour de leur interprétation en s'appuyant sur leur capacité à "faire parler les chiffres", en révélant des structures cachées dans une série de données, offre une perspective pédagogique propre à susciter la réflexion plutôt qu'à entraîner l'adhésion passive. En d'autres termes, leur rôle devrait être autant de

poser des questions que d'apporter des réponses. Se limiter au calcul d'une moyenne ou à la construction d'un diagramme donne une maigre idée de la puissance de la statistique. Certes les algorithmes permettant d'obtenir les différents paramètres ou les différents graphiques doivent être connus mais on ne peut se dispenser d'une réflexion sur le sens et l'interprétation des résultats obtenus. Ce dernier point requiert des situations problèmes et des exercices motivants qui demandent une réelle activité intellectuelle dans l'analyse du problème, le choix des méthodes et la formulation des conclusions[2]. En conclusion,

on apprend autre chose que des techniques en faisant des statistiques.

En effet, les statistiques constituent un chapitre des mathématiques dans lequel les références à la vie courante ainsi qu'aux autres matières scolaires sautent aux yeux. On y apprend à raisonner aussi bien qu'en géométrie, même si c'est sous une autre forme. Sans elles, la compréhension du monde qui nous entoure devient de plus en plus difficile. Elles participent, également, à la préparation aux études scientifiques. Il est donc primordial de leur consacrer un temps important et réparti sur toute la scolarité, comme les programmes nous y invitent.

Lors de la deuxième Conférence Internationale sur l'Enseignement de la Statistique (ICOTS 2) en 1986 à Victoria (British Columbia, Canada), Gail Burrril, de l'American Statistical Association, terminait son intervention[3] par ces mots :

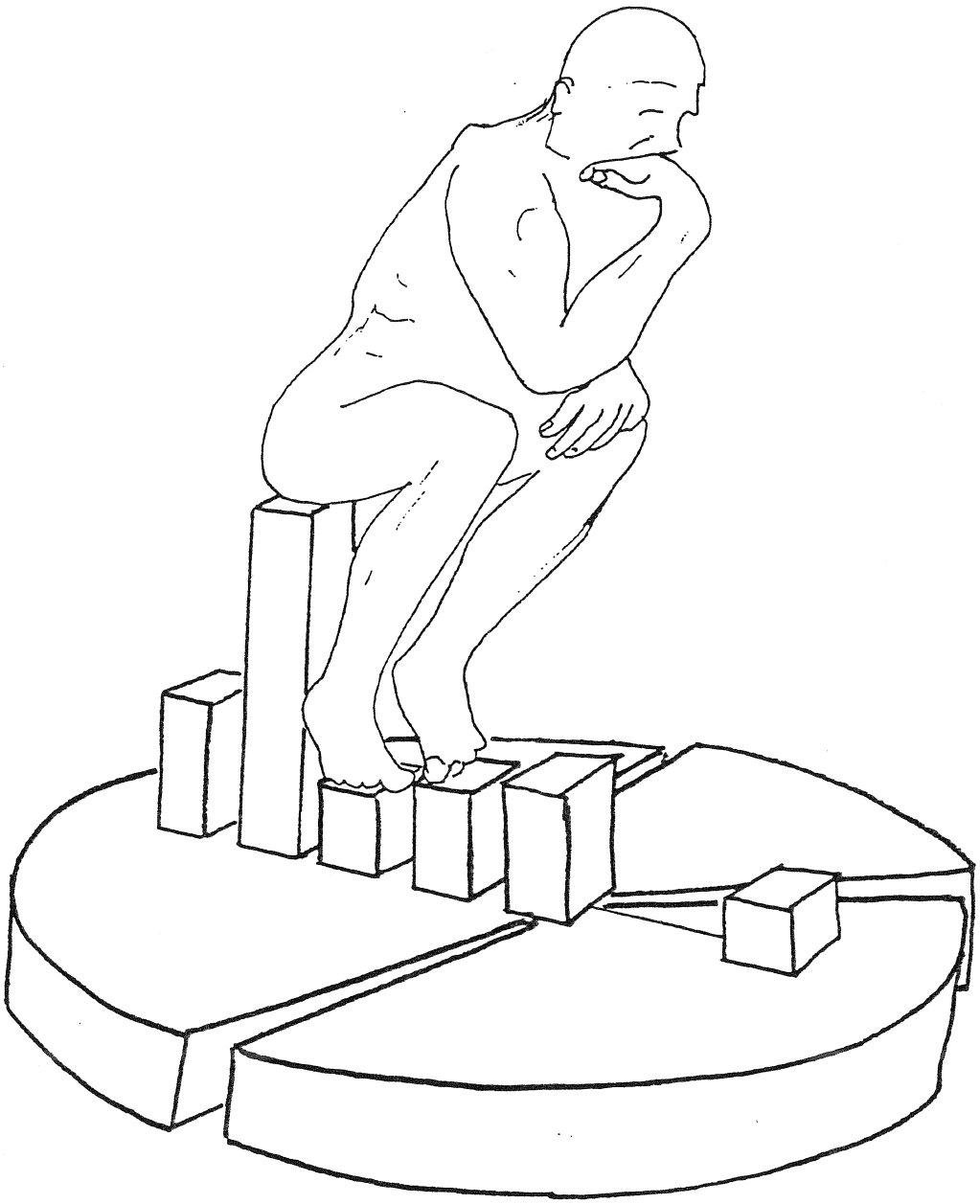
Les Etats-Unis commencent à prendre conscience que l'initiation à l'analyse quantitative est l'une des clés de l'avenir. La finalité essentielle de l'éducation est de préparer les élèves à devenir des citoyens efficaces et productifs. Dans le monde d'aujourd'hui et de demain, cela suppose que la statistique devienne partie intégrante de l'enseignement. Cet objectif est en voie d'être atteint aux Etats-Unis.

Puissions-nous en dire autant en France avant la fin du deuxième millénaire !

---

[2] Voir par exemple, Jean Claude Girard, "Pourquoi il ne faut pas laisser de côté les chapitres de statistiques au collège", in *Repères-IREM*, n°23, Avril 1996.

[3] "Initiation à l'analyse quantitative aux Etats-Unis d'Amérique", traduite dans l'ouvrage *Etudes sur l'enseignement des mathématiques, Vol. 7, L'enseignement de la statistique*. Ed. Unesco, Paris, 1994.



Chapitre 1

Finalités et enjeux de l'enseignement de la statistique

Réflexion sur les finalités et les enjeux d'un enseignement de la statistique intégré à celui des mathématiques à l'Ecole Primaire, au Collège et au Lycée en relation avec les autres disciplines scolaires.

Jean-Claude Régnier

Contribution à la statistique

Sur cent personnes :
sachant tout le mieux,
- cinquante deux.
Incertaines de chaque pas,
- presque tout le reste.
Prêtes à aider pourvu que cela ne dure pas longtemps,
- jusqu'à quarante neuf.
Bonnes toujours car ne sachant pas faire autrement,
- quatre, cinq peut-être.
Enclines à admirer sans envie,
- dix-huit
Induites en erreur par la jeunesse qui passe,
- plus ou moins soixante.
Avec qui on ne rigole pas,
- quarante-quatre.
Douées pour le bonheur,
- vingt et quelques tout au plus.
Inoffensives une à une mais sauvages en foule,
- plus de la moitié assurément.
Cruelles si les circonstances les y obligent
- il vaut mieux ne pas le savoir,
même approximativement.
Echaudées craignant l'eau froide,
- guère plus nombreuses que celles qui la craignent sans avoir été échaudées.
N'empruntant rien à la vie fors les choses
- trente,
j'aimerais pourtant me tromper.
Perclues, dolentes sans le moindre falot dans le noir
- quatre-vingt-trois.
Justes
- beaucoup, trente-cinq.
Mais si cette vertu s'accompagne d'un effort de compréhension
- trois.
Dignes de compassion
- quatre-vingt-dix-neuf.
Mortelles,
- cent sur cent,
Nombre qui pour l'instant ne change pas

Wislawa Szymborska ***(Prix Nobel de littérature 1996)***

(source : Le Nouvel Observateur 31 oct / 6 nov 1996 pp124-125 Traduit du polonais par Georges Lisowski)

Que sont la statistique et les statistiques ?

Comment aurions-nous pu résister au plaisir de commencer cette réflexion par un texte poétique dont l'auteur, poétesse polonaise, fut honorée en cette fin d'année 1996 ?.

Ce poème nous donne à sa façon l'occasion de réfléchir à partir de statistiques, c'est à dire les données, les résultats, sur la statistique et quelques-uns de ses objets de préoccupation.

Nous aurions d'ailleurs pu tout aussi bien partir des propos suivants : « *La statistique est une science toute moderne : le génie des anciens ne se portait pas volontiers vers des travaux de précision ; les moyens de recherche et de communication leur manquaient ; enfin (...) ils ne paraissent pas avoir soupçonné l'existence d'un principe de compensation qui finit toujours par manifester l'influence des causes régulières et permanentes, en atténuant de plus en plus celle des causes irrégulières et fortuites. De nos jours, au contraire, la statistique a pris un développement en quelque sorte exubérant; et l'on n'a plus qu'à se mettre en garde contre les applications prématurées et abusives qui pourraient la décréditer pour un temps, et retarder l'époque si désirable où les données de l'expérience serviront de bases certaines à toutes les théories qui ont pour objet les diverses parties de l'organisation sociales. (...) Nous entendrons par statistique , la science qui a pour objet de recueillir et de coordonner des faits nombreux sensiblement indépendants des anomalies du hasard, et qui dénotent l'existence des causes régulières dont l'action s'est combinée avec celle des causes fortuites* » Ces propos relatifs à cette science statistique sont de Augustin A. Cournot[1]...en 1836 ! Il dit encore « *La statistique est une science d'observation. Les chiffres sont les instruments à l'usage des statisticiens, et la précision de ces instruments est rendue comparable au moyen des formules tirées de la théorie des chances. Mais le but essentiel du statisticien, comme de tout autre observateur, est de pénétrer autant que possible dans la connaissance de la chose en soi, et pour cela de dégager autant que cela se peut faire, par une discussion rationnelle, les données immédiates de l'observation, des modifications qui les affectent, en raison seulement du point de vue où se trouve placé l'observateur, et des moyens d'observation mis à sa disposition* »[2] . En 1922, un des plus grands statisticiens de la première moitié du XXème siècle, Ronald A. Fisher, écrivit «*L'objet de la méthode statistique est la réduction des données. Une masse de données doit être remplacée par un petit nombre de quantités représentant correctement cette masse, et contenant autant que possible la totalité de l'information pertinente contenue dans les données d'origine. Cet objectif est accompli par la construction d'une population infinie hypothétique. La statistique comporte des problèmes*

---

[1] Cournot, A., (1801-1877) *Exposition de la théorie des chances et des probabilités* ,Hachette-Paris-1843 in chap IX p. 122 De la statistique en général et de la détermination expérimentale des chances.
ouvrage re-publié en coédition CNRS-Librairie Vrin en 1984

[2] *ibidem* p. 125

de spécification apparaissant à travers le choix de la formalisation mathématique de la population, des problèmes d'estimation, impliquant le choix de méthodes de calcul de quantités dérivées de l'échantillon, que nous appelerons statistiques, construites pour estimer les valeurs des paramètres de la population hypothétique, et en des problèmes de distribution »[3].

Donner une définition de la statistique constitue une prouesse. En 1870 on pouvait en repérer déjà plus de 60 et en 1935 W.F.Willcox[4] en dénombrait plus d'une centaine. Parmi ces définitions, il convient de pointer celles qui se fondent sur une point de vue critique ou humoristique. Elles émanent d'auteurs qui usent plus de leur notoriété que de leurs compétences relatives tant aux méthodes et buts de la statistique qu'à leur succès et à leurs limites. Parmi les définitions les plus connues dans ce genre, les suivantes seraient attribuées[5] respectivement à Benjamin Disraeli[6] *«il existe trois sortes de mensonges: les mensonges, les affreux mensonges et les statistiques»*, à Adolphe Thiers[7] *«l'art de préciser les choses que l'on ignore»*, à Laveleye *« l'art de mentir mathématiquement »* et à Macauley *« Les chiffres disent toujours ce que souhaite l'homme habile qui sait en jouer »*. Que ces citations soient fidèles ou non ne constitue pas le fait majeur, ce qui importe c'est qu'elles soient évoquées à divers propos et souvent utilisées comme argument d'autorité par les détracteurs de la statistique. Ces propos analysés montrent combien la confusion entre l'outil, les résultats que cet outil produit et l'usage que tout un chacun peut faire de ces résultats pour son argumentation, est manifestement entretenue. Il est d'ailleurs très intéressant de constater que de telles affirmations pourraient être énoncées à propos de nombreux objets comme par exemple la photographie ! Combien d'outils nobles pourraient être dénigrés au nom des détournements que la bêtise humaine en fait ? Comment peut-on condamner sans appel un outil en se fondant sur des usages inadéquats et erronés ?

A côté de ces expressions qui jugent plus qu'elles ne définissent, figurent celles qui confondent l'outil et un domaine d'application, lui-même parfois plus ou moins pertinent. Ainsi en 1749, Achenwald écrit que la statistique est *« la connaissance approfondie de la situation respective et comparative des Etats .»* En 1785 le but de la statistique est selon Sinclair *« de constater la somme de bonheur dont jouit une population et les moyens de l'augmenter.»* D'autres exemples pourraient être trouvés de nos jours encore, tels que la

---

[3] extrait de On the mathematical foundations of theorical statistics (1922) cité par Droesbeke, J.-J., et Tassi, Ph, *Histoire de la statistique* PUF Que-sais-je ? n°2527

[4] W.F. Willcox, Definitions of statistics, Revue de Institut international de Statistique, 1935

[5] selon Claude Angers à Disraeli in Les statistiques Oui mais ... Ed Agence d'Arc, 1991 p. 1
selon André Patier à Disraeli ou Bismark avec la formulation suivante : *«Il y a trois formes de mensonge, le mensonge ordinaire, le damné mensonge et la statistique»* in STATISTIQUE, PUF 1962, Tome 1 p. 15

[6] (1804-1881) Ecrivain et Homme politique britannique; il fut en particulier Premier Ministre.

[7] (1797-1877) homme politique français

réduction de l'économie politique à la statistique, la réduction de la statistique aux mathématiques, la confusion entre les statistiques (les données) et la statistique, etc ...

Enfin une autre catégorie de définitions se caractérise par une limitation excessive du domaine, ainsi trouve-t-on *la science des mesures, la science des moyennes* , ce qui ne rend compte que de quelques tâches du statisticien renvoyant à une représentation de la statistique comme accumulation de données plus ou moins dégrossies, organisées et ordonnées. Ce point de vue demeure présent chez W.F. Willcox qui propose : « *la statistique est l'étude numérique des groupes ou des masses, par l'étude des unités qui les composent, que ces unités soient des hommes ou non, des êtres animés ou inanimés* [8] »

Si nous tentons d'évoluer vers des conceptions plus actuelles de la statistique, nous pourrions en premier lieu nous référer à celle de M.G. Kendall pour qui «*la statistique est la branche de la méthode scientifique qui traite des données obtenues en comptant ou en mesurant les propriétés de populations de phénomènes naturels* [9]» pour en second lieu la compléter par celle de A. Vessereau pour qui «*les méthodes statistiques englobent toutes les recherches dans lesquelles le grand nombre et l'enchevêtrement des facteurs exigent une technique d'interprétation basée sur la connaissance des lois du hasard.* »

Pour notre part nous avons cherché à expliciter notre conception de la statistique au moyen des schémas ci-après qui prennent appui sur la définition suivante : ***La statistique est la science qui procède à l'étude méthodique à partir de modélisations mathématiques, des modes d'utilisation et de traitement de données, c'est à dire de l'information, dans le but de conduire et d'étayer une réflexion ou de prendre une décision en situation concrète soumise aux aléas de l'incertain.***

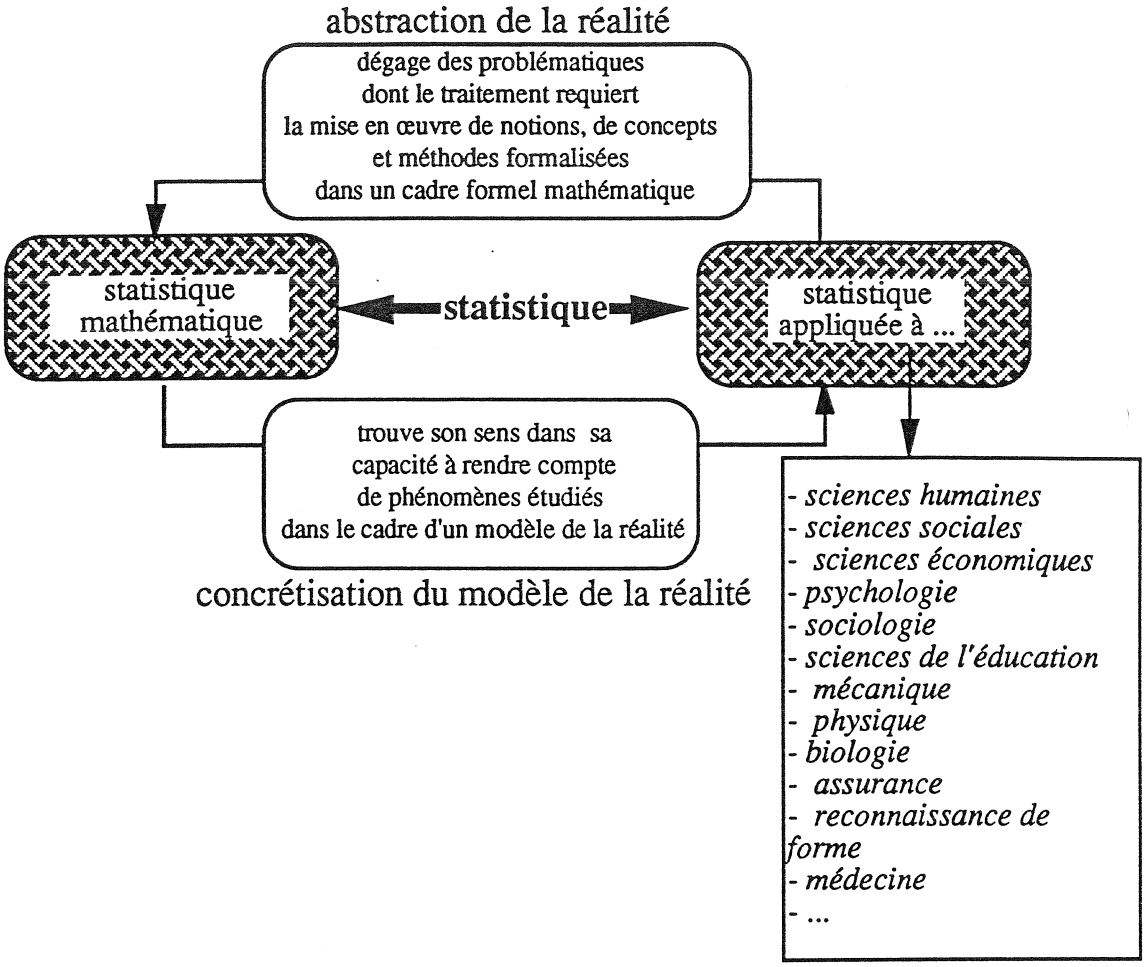
La statistique descriptive étudie ces modes d'utilisation et de traitement de données, à un premier niveau, dans la perspective de produire essentiellement des descriptions des informations.

La statistique inférentielle les étudie à un second niveau dans la perspective d'étendre ces informations décrites à un domaine de validité non exploré directement, avec, si possible, un contrôle des risques encourus dans ce raisonnement inductif.

---

[8] selon André Patier *opus cit.*

[9] « Statistics is the branch scientific method that deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition "natural phenomena" includes all the happenings of external world, whether human or not. » Kendall's advanced theory of Statistics vol 1 - 5ème édition - p1



Dans ce premier schéma, nous traduisons cette tension dialectique qui lie deux dimensions de la statistique comme les deux pôles inséparables d'un aimant. Ces deux pôles nous les nommons respectivement *statistique mathématique* et *statistique appliquée à...* Il nous paraît impossible d'envisager la statistique mathématique, c'est à dire la théorie statistique dans le cadre théorique au sein duquel se développent l'explicitation, la formalisation des notions, concepts, méthodes et des raisonnements de la statistique de même que nous ne pouvons imaginer que cette théorie soit coupée de sa mise en œuvre dans d'autres cadres théoriques tels que la psychologie, l'économie, la mécanique, la médecine, etc...

C'est un peu de ce point de vue que nous regardons la physique dans sa tension entre la *physique mathématique* et la *physique appliquée.*

Le second schéma vise à présenter les buts et les opérations que nous assignons à la statistique et qui, par là même, la caractérise en partie. Nous lui attribuons comme but central de constituer **un outil d'aide à la décision**, que cette décision soit prise à partir d'une exploration, d'une description de données spécifiées recueillies dans des circonstances

connues ou encore à partir d'hypothèses testées conduisant à une certaine maîtrise du risque d'erreur encouru.

Ce schéma s'inscrit dans un ensemble de caractéristiques qui servent à définir aujourd'hui la statistique :

une sorte de *langage commun* , méthode *générale* reliant divers domaines scientifiques	**portant** sur des ensembles d'individus, de variables et de relations	**conduisant** plutôt à des conclusions vraisemblables et probables que vraies et certaines énonçant des propriétés de groupe valides sur des ensembles parfois imparfaitement définis

Si l'on caractérise les mathématiques comme *une science du certain,* la statistique pourrait alors s'en distinguer alors comme *une science de l'incertain* qui cherche à établir *les frontières de l'incertitude.* Dans cette perspective le raisonnement statistique prend appui sur les théories des probabilités.

A cette étape, il convient de rappeler que le terme *statistique* [10] apparaît alors:

- comme substantif au singulier pour désigner le domaine scientifique, la discipline universitaire : **la Statistique,**

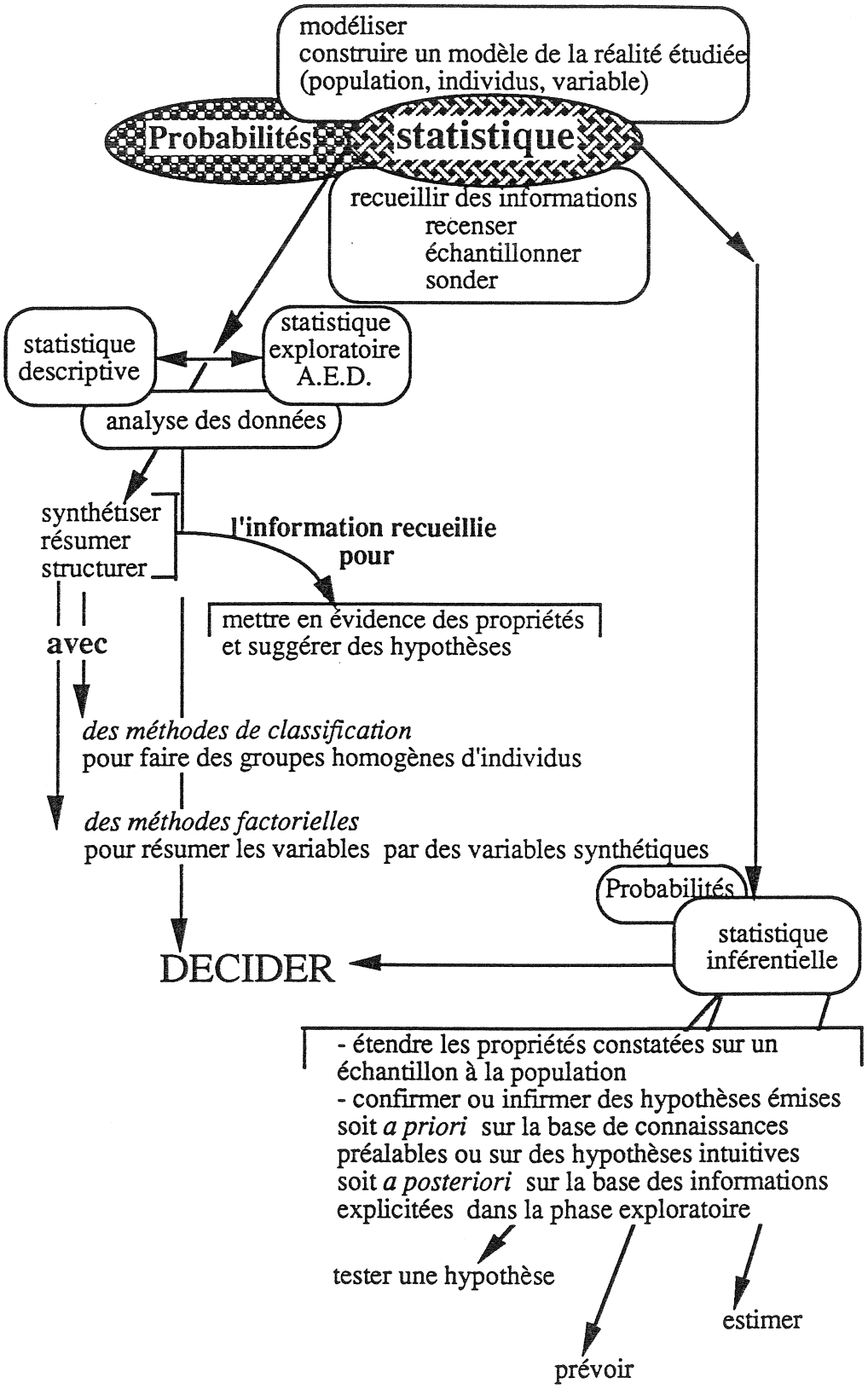
- comme substantif au pluriel pour désigner les données qui constituent une part de ces objets d'étude : **les statistiques,**

- comme substantif (au singulier ou au pluriel) pour désigner une fonction des résultats d'une variable (statistique) recueillis sur un échantillon : **une statistique** (exemple, la moyenne empirique est une variable qui à chaque échantillon associe la moyenne des valeurs d'une variable recueillies)

- comme adjectif pour préciser d'une part *ce qui est relatif à la statistique,* (exemple : tableau statistique, échantillon statistique, variable statistique), d'autre part *ce qui concerne les grands nombres, les phénomènes complexes* (mécanique statistique, rationalité statistique)

---

[10] Ce point de vue est soutenu par M.G. Kendall « It may be as well to point out that "Statistics", the name of scientific method, is collective noun, has a capital "S", and takes the singular. The same word "statistics" is also applied to numerical material with which the method operates, and in such a case has no capital letter and the plural. Later in this book we shall meet the singular form "statistic", which is defined as a function of the observations in a sample from some population. "Statistic" in this sense takes the plural "statistics" » Kendall's advanced theory of Statistics vol 1 - 5ème édition - pp 1-2



Force est de constater qu'à côté de ces mots nous pouvons observer des comportements humains qui expriment des rapports à ce domaine de connaissances et aux pratiques sociales afférentes. En France nombre de personnes à la seule évocation du terme

en soulignent aussitôt les aspects négatifs sans pour autant être en mesure de fournir un exemple précis d'un usage inadéquat et trompeur de la statistique. Ajoutons qu'une formation insuffisante en ce domaine conduit à rendre difficile le discernement d'une étude statistique correcte par rapport à une étude incorrecte. Des résultats sérieux et correctement traités peuvent être mal interprétés par défaut de compétences élémentaires en statistique tandis que certaines personnes ayant acquis la mauvaise habitude de croire tout ce que les *media* colportent constituent des proies faciles pour des manipulateurs de résultats qualifiés de *statistiques* en la circonstance. Trois conduites liées à une compétence insuffisante peuvent être identifiées : le rejet systématique, l'acceptation naïve ou l'usage détourné à des fins personnelles avantageuses.

Dans notre vie quotidienne actuelle, les statistiques et donc la statistique tiennent une place dominante dont les *media* se font tout particulièrement l'écho pour ne considérer que la partie émergée de l'*iceberg statistique*. Tout se passe comme si, aujourd'hui, la statistique était devenue un outil pour penser et agir aussi indispensable au citoyen que l'écriture, la lecture et le calcul. En 1963, W. Weaver[11] écrit « *La théorie des probabilités et la statistique sont deux domaines importants, intégrés à nos activités quotidiennes. Le monde de l'industrie, les compagnies d'assurance sont largement tributaires des lois probabilistes. La physique elle-même est de nature essentiellement probabiliste. Il en est de même des fondements de la biologie. Cependant, en dépit de cette importance, les responsables de l'enseignement n'ont pas encore admis le caractère universel de la théorie des probabilités et de la statistique. Il faut espérer que des éléments de la théorie des probabilités soient introduits dès que possible au niveau de l'enseignement secondaire...*»

Pourquoi et dans quels buts faut-il envisager une formation à la statistique ?

C'est alors tout naturellement que nous sommes conduit à porter notre réflexion sur la place et le rôle de l'enseignement dans la *formation de l'esprit statistique* de l'enfant et de l'adolescent. Quand en 1973, Maurice Glaymann et Tamas Varga choisissaient cet extrait pour l'introduction de leur ouvrage[12], ils exprimaient leur satisfaction partielle en précisant qu'au cours des dix années écoulées, l'enseignement des mathématiques avait connu, tant en France que dans de nombreux pays, des réformes qui laissaient entrer l'enseignement des probabilités. Ils ajoutaient alors *« et ce n'est peut-être plus qu'une question de temps pour qu'elles constituent une partie intégrante des programmes scolaires de tous les pays »* .

---

11 cité par Glaymann, M., Varga, T., Les probabilités à l'école cedic 1973 p.9 (texte extrait Lady Luck publié par Doubleday and Co, Garden City, N.Y. p 376-377)

12 *opus cit.*

Qu'en est-il de nos jours près de 25 ans plus tard ? Certes, dans cette prophétie, il est plus question de *probabilité* que de *statistique* mais nous gageons que les deux domaines étaient envisagés dans l'esprit de cette attente. L'ouvrage qu'ils rédigèrent propose des pistes de travail pour un enseignement des probabilités à l'école primaire dont les effets ne pourraient qu'être avantageux dans la *formation de l'esprit statistique*. Un lien qui se trouve confirmer dans un autre ouvrage fort intéressant intitulé *L'enseignement des probabilités et de la statistique* de Arthur Engel[13].

Dans les années 80, Lennart Rade expose dans un article consacré à la statistique et publié par l'UNESCO des constats émis depuis les années 50 relatifs à la formation de l'esprit statistique. Ainsi en 1959 un rapport d'une commission sur l'enseignement des mathématiques[14] aux U.S.A. conclut ses travaux par ces propos *« Si les mathématiques ont trait à des situations où les faits peuvent être déterminés, elles donnent aussi les moyens d'étudier, de comprendre et de maîtriser l'incertain. Parmi les plus récentes applications des mathématiques, beaucoup font appel à la théorie des probabilités et au raisonnement statistique.(...) La pensée statistique joue un rôle croissant dans la vie quotidienne des adultes instruits. L'initiation à la pensée statistique est un complément important à l'initiation à la pensée déductive.»* . En France la même année lors d'un séminaire sur l'enseignement des mathématiques organisé par l'Organisation Européenne de Coopération Economique, le point de vue suivant est soutenu « Le calcul des probabilités élémentaire doit être considéré comme une branche des mathématiques susceptibles d'être enseignée dans les écoles secondaires. L'induction statistique doit être considérée comme une branche des mathématiques appliquées qui entre pour une part capitale dans les processus de décision conformes à l'esprit de la "méthode scientifique" et dont de très nombreux secteurs des sciences physiques et des sciences du comportement humain font un usage accru. Il faut admettre en outre que le raisonnement statistique acquiert une importance croissante dans le domaine des affaires publiques. Un enseignement élémentaire approprié du calcul des probabilités et de la statistique doit faire partie du nouveau programme des études secondaires des cours préparatoires sur ces matières devront figurer aux programmes (...) des institutions formant les professeurs.». Ces propos confirment que ni notre préoccupation ni nos souhaits ne sont des idées nouvelles et originales !

Ainsi nous pouvons tenter d'exposer quelques arguments en faveur d'un enseignement de la statistique à l'école primaire ou secondaire : il est un *fait* que la statistique est partie intégrante de notre culture actuelle et que de nombreuses notions de statistique interviennent dans notre vie quotidienne. Nous ajoutons que la *pensée statistique* implique une

---

13 ouvrage en 2 tomes traduit et adapté de l'allemand et publié chez C.E.D.I.C. en 1975 (tome 1 307 p) et 1979 (Tome 2 - 386 p)

14 instituée en 1955

composante de la capacité à manipuler des nombres, et nous parions qu'un contact précoce avec ce domaine de connaissance développerait une compréhension élémentaire de notre environnement actuel inondé de statistiques.

Comment peut-on envisager cette formation ?

Il convient naturellement de penser à la formation à la statistique des enseignants. Certes cette formation aurait à être envisagée pour tous les enseignants qu'il s'agisse des professeurs des écoles ou des professeurs de disciplines usant de la statistique en collège et en lycée. Il me semble que le fonds commun est constitué par la statistique descriptive. Mais il y a tout lieu de penser aussi que la statistique, en raison de sa dimension mathématique, soit intégrée à l'enseignement des mathématiques au collège et au lycée. Les professeurs de mathématiques auraient alors tout intérêt, soit en cours de formation initiale, soit par autoformation au long de leur vie professionnelle, à développer leurs compétences à la modélisation mathématique de situations soumises aux aléas de l'incertain, à l'usage de l'informatique à partir de l'ordinateur ou de calculatrices pour calculer ou simuler, à la conduite de projets centrés sur des problématiques dont les traitements requièrent les outils de la statistique. Evidemment il serait même souhaitable qu'à côté de la formation en didactique des mathématiques, et en lien avec elle, soit conduite une formation à la didactique de la statistique. Ainsi cette didactique de la statistique doit-elle aussi être développée à partir de recherches scientifiques tout comme la didactique des mathématiques. Introduire l'enseignement de la statistique est l'occasion de faire émerger des problèmes qui alimenterons les travaux de recherche. Si la statistique descriptive constitue bien le base commune de formation, il convient d'y ajouter une réflexion approfondie sur les notions d'*estimation* et de *test d'hypothèse statistique* et les raisonnements afférents.

Quant à la formation à la statistique des élèves, c'est à dire cette formation visée à l'école primaire, au collège et au lycée, elle est à faire à partir de situations-problèmes tirées du vécu des élèves appelant des activités de modélisation mathématique pour traiter des informations. Les sources ne manquent pas même si elles peuvent paraître routinières pour l'enseignant, elles demeurent nouvelles pour le novice-apprenant. D'ailleurs nous pensons qu'ici se trouve une cause de désintérêt actuel des enseignants pour la statistique, en particulier au lycée. En effet, si en raison de l'insuffisance de la formation des enseignants de mathématiques dans ce domaine et des représentations erronées que cet état induit, l'enseignement de la statistique se réduit à la transmission de quelques définitions des notions de base telles que *population, variable, effectif, fréquence, moyenne, écart-type, diagramme en bâtons, en barres ou circulaire, histogramme*, il va de soi que l'ennui ne peut que surgir rapidement. Là, peut-être plus qu'ailleurs, l'absence de situation-problème

explicite dont la résolution rend nécessaire l'emploi d'un outil statistique, se fait sentir. Réduire la définition de la moyenne arithmétique à l'explicitation de l'algorithme de calcul du genre *"la somme des valeurs divisée par l'effectif total"*, ne me semble pas suffisant pour en construire le sens. Ainsi il est même possible d'aborder cette notion par une problématique du cadre mathématique comme par exemple :

Disposant des 5 résultats numériques suivants : 15, 16, 18, 18, 20, on souhaite remplacer ceux-ci par une valeur unique c de telle sorte que la somme totale ne soit pas modifiée. Cela donne $S = 15 + 16 + 18 + 18 + 20 = c + c + c + c + c = 5c$ d'où $S = 5c$ et $c = \frac{S}{5}$ ce qui donne $c = 17,4$. Cette valeur existe ici et est unique. Ensuite on peut alors s'intéresser à l'erreur que l'on commet en remplaçant chaque valeur par la valeur c, cela donne les valeurs (15 - c), (16 - c), (18 - c), (18 - c), (20 - c). Une idée consisterait à rechercher une valeur de c qui minimise la somme de ces erreurs. Si on calcule cette somme, on peut constater que celle-ci est minimale pour la valeur c égale à la valeur précédente c = 17,4. Ceci constitue deux caractérisations de la moyenne. Bien d'autres exemples peuvent être développés.

L'analyse des outils mis en œuvre dans d'autres disciplines complète cette approche en exerçant l'élève à les repérer, à les identifier, à les expliciter et à confronter leurs interprétations avec celles fournies dans le cadre disciplinaire concerné. Par exemple, on peut partir d'un diagramme en bâtons des précipitations moyennes mensuelles dans une région durant une année ou de caractéristiques biométriques telles que la taille ou encore des statistiques du chômage en économie.

Pour illustrer ma propre approche en classe de seconde, je joins les deux documents suivants qui fournissent quelques indications sur la manière dont j'introduisais la statistique dans le cadre du cours de mathématiques. Il s'agit d'un *plan de travail*[15] et d'une des trois *fiches guides* [16] qui l'accompagnaient. Le travail auquel ce thème est intégré, remonte à la fin des années 80 et a fait l'objet d'une publication[17] auquel je renvoie le lecteur intéressé. Le travail demandé aux élèves conduisait à la production d'un dossier "rapport d'étude statistique". Les résultats étaient particulièrement intéressants avec des élèves qui s'impliquaient. Un travail interdisciplinaire était aussi réalisé à ce propos avec le cours d'informatique. Des débats avaient lieu dans la classe par la mise en commun d'informations.

Aujourd'hui je constate que ce travail allait déjà dans le sens des points de vue dégagés dans le rapport de l'UNESCO[18] qui préconise au moins un enseignement de la statistique

---

[15] document n° 1

[16] document n° 2

[17] Régnier, J-C & al. L'autonomie et le travail personnel des élèves dans les disciplines scientifiques en lycée. Vol 2 CRDP de Dijon MENJS - 1991

[18] *opus cit.*

descriptive car il s'inscrit dans les objectifs généraux de la formation scolaire obligatoire tout en offrant un champ d'activités qui se rattachent facilement à la vie réelle de l'élève. Ces problèmes tirés du vécu de l'apprenant donnent matière aux activités de modélisation mathématique nécessaires au développement cognitif. L'importance attachée aux applications de la statistique et à la construction des modèles au sein desquels sont conduits les traitements, à l'emploi de la simulation et à l'utilisation des moyens modernes de calculs comme le permettent les calculatrices et les ordinateurs, est incarnée dans des activités de projet.

Document n°1

Plan de travail n°6 mathématiques classe de Seconde

Période du 6/12/89 au 8/1/90

Thème n° 5 " STATISTIQUE.."

Objectifs :

A l'issue de cette étude, il faudrait être capable de :

- donner une définition aux termes de base employés en statistique
- utiliser les notions statistiques citées pour mener une petite étude statistique réelle.
- utiliser ces notions pour comprendre les résultats fournis par une étude statistique dans le domaine de la gestion, de l'économie ou de la maintenance, par exemple.
- citer quelques exemples d'étude statistique en explicitant leur but et leurs conclusions.

Documents :

- livre de la classe de Seconde (chapitre : analyse des données)
- dictionnaire
- livres d'économie, de gestion, d'histoire, de géographie, de biologie ...
- divers journaux
- BT2 n°91 et n°119

Travail :

rechercher :

- la définition des termes fournis dans la liste §2 de la fiche-guide n°5.1 quand ils sont employés dans le domaine de la **statistique** .
- des exemples extraits de divers documents illustrant ces définitions.

rédiger :

- individuellement un petit glossaire (**dossier statistique n°1**) contenant les définitions et les illustrations en présentant l'ensemble sous un forme claire .

s'exercer :

- faire les exercices n°8,12,13,14 page 201 (*du livre de la classe de cette période*)

réaliser une étude avec un guidage :

- réaliser l'étude proposée dans la fiche-guide n° 5.2 (*elle fournit les divers éléments à repérer*)

réaliser une véritable étude statistique:

- par groupe de 3 ou 4,
  * réaliser un projet d'étude statistique mettant en jeu les notions apprises,
  * soumettre le projet au professeur
  * réaliser l'étude puis consigner la méthode et les résultats dans un petit rapport (**dossier statistique n°2**) qui sera remis au professeur.

utiliser des outils de calcul :

la calculatrice et l'ordinateur sont des outils indispensables auxquels tout statisticien a recours. Ce travail offre donc une occasion supplémentaire d'en faire un usage pertinent.

Evaluation :

Dossier statistique n°1 .....18/12/89
Projet d'étude.....20/12/89
Dossier statistique n°2 .....12/03/90
Test préliminaire autocorrigé:.....
Test final hétérocorrigé :.....8/1/90
Bilan de travail n°6 :.....8/1/90

Informations diverses :

Document n°2

Plan de travail n° 6 FICHE-GUIDE n° 5.1 thème n°5

1. Prendre rapidement connaissance du contenu

- en dix minutes maximum parcourir le chapitre "Analyse des Données" en repèrant les passages qui semblent les plus importants.

2. Prendre des notes:

liste de quelques termes de base donnés en vrac :
- la statistique ; les statistiques
- population ; individu ; série
- pourcentage ; taux ; proportion ; fréquence
- effectif ; effectif cumulé
- variable statistique
quantitative,qualitative,ordonnée,continue,discrète,chronologique
- sondage ; échantillon ; questionnaire ; enquête
- collecte des données, analyse des données,organisation des données
- tableaux
- représentations graphiques :
  diagrammes en secteurs,
  diagrammes en bâtons,
  diagrammes polaires,
  diagrammes figuratifs,
  histogramme
- paramètres : moyenne arithmétique, moyenne quadratique, variance, écart-type, écart moyen, mode, dominante, médiane, étendue, quartiles, espace interquartile.
- distribution normale

3. Tester ses savoir-faire
- faire les exercices de la page 199 n° 1,2,3,4,5

4. Pour aller plus loin :

5. S'entraîner à retenir:
- les résultats relevés dans le cours

6. S'entraîner à chercher des solutions à un problème et à rédiger un compte rendu :
dossier statistique n°2 à rendre le compte rendu au professeur pour le 12/3/90

Enseignement de la statistique et didactique de la statistique.

Que pourraient apporter des recherches en didactique de la statistique pour l'enseignement ?

Afin que les propos sur la statistique et son enseignement ne demeurent pure spéculation, il nous paraît important de promouvoir une recherche en didactique de la

statistique fondée sur des paradigmes proches de ceux de la recherche en didactique des mathématiques. Dans la formation universitaire, l'enseignement de la statistique concerne de nombreux étudiants dont les goûts personnels les avaient poussés à fuir durant leur scolarité secondaire toute référence à ce qui pouvait ressembler de près ou de loin aux mathématiques. Il en est ainsi en sciences humaines ou en sciences sociales au sein desquelles les approches statistiques ont une place importante. Or la compréhension de ces approches et les interprétations qu'elles induisent ou qu'elles servent à étayer nécessite une formation en statistique qui ne se réduise pas à des conduites de presse-boutons. Par expérience personnelle, je constate les grandes difficultés auxquelles sont confrontés ces étudiants de ces disciplines universitaires. Ainsi, à titre d'exemple, la formation en sciences de l'éducation dont une partie des étudiants sont des enseignants actuels ou futurs, contient une formation à la statistique. Les finalités de mon enseignement sont exprimées au travers des objectifs généraux suivants :

Objectifs généraux de l'enseignement de la Statistique en licence et en maîtrise de Sciences de l'Education:

Cet enseignement vise à apporter à chaque étudiant quelques outils techniques et conceptuels efficaces et bien identifiés qui devraient l'aider à :

- **expliciter les questions** d'une problématique dont l'émergence et/ou la validation des réponses relèvent d'une approche statistique en liaison avec le modèle dans lequel cette problématique est posée,

- **décrire, traiter, analyser des données** de manière pertinente dans le cadre d'une étude en particulier dans le domaine éducatif,

- **faire le lien** entre la réflexion analytique sur des questions relevant du champ de l'éducation, leur formalisation et leur traitement quantitatif,

- **lire** avec un regard critique et distancié, **les conclusions** de diverses études statistiques en particulier celles apparaissant dans des rapports de recherche en Sciences de l'Education,

- **poursuivre** de façon autonome et personnalisée **un apprentissage** en statistique afin d'enrichir les acquis personnels actuels,

- **poser un regard plus positif** à l'égard d'un domaine largement exploité dans les *media* , dans le sens de ne pas considérer les résultats dans l'ordre du tout ou rien mais en les replaçant judicieusement dans leur domaine de validité et en relativisant la portée,

- **exploiter des notions et des démarches mathématiques** à des fins d'outils, et de ce fait à modifier dans un sens positif le rapport souvent négatif que nombre entretient avec cette science,

- **s'exercer à un raisonnement** intégrant l'idée de "risque d'erreur" dans l'énoncé des conclusions.

- **s'exercer à l'interprétation** de phénomènes éducatifs sur la base de données statistiques recueillies sur des "faits" et des relations entre ces "faits" en particulier sur des "faits éducatifs" et sur des relations entre ces "faits éducatifs"

- **s'exercer à la communication** des résultats des analyses des données en faisant une distinction bien nette entre le modèle utilisé et la réalité qu'il est supposé représenter, entre les traitements conduits au sein du modèle et les interprétations reformulées dans le contexte au sein duquel est posé le problème.

Dans le cadre de cet enseignement mais aussi dans celui des travaux de recherche requérant une approche statistique, je trouve des indices forts de l'importance d'une recherche de qualité en didactique de la statistique afin de nous éclairer sur les conditions d'acquisition des notions, concepts et méthodes statistiques en lien avec leurs applications en sciences de l'éducation et sur les processus mis en jeu pour les acquérir. Une amélioration de l'efficience de cet enseignement me paraît fondamentale dans la mesure où ce contexte permet à des enseignants non spécialistes de rencontrer la statistique. Cette rencontre, si elle est heureuse, laisse espérer qu'à leur tour, ils pourront changer leurs représentations de la statistique et modifier leurs conduites d'enseignement au sein de leur classe dans le sens d'une conception plus proche du domaine scientifique quand les thèmes traités en donneront l'occasion.

En ce qui me concerne, en écrivant cet article, j'ai souhaité contribuer au développement de cette nécessaire réflexion sur l'enseignement de la statistique et de la recherche en didactique sans pour autant l'isoler d'une réflexion didactique et pédagogique plus générale à conduire au sein de l'école prenant en considération d'autres domaines de connaissance.

Histogramme

Réflexion sur une représentation graphique particulière parfois abusivement utilisée tant dans l'enseignement que dans l'application de la statistique.

Jean-Claude Régnier

Nous avons choisi ici de développer notre point de vue pédagogique à propos de la notion d'histogramme. Ce qui nous y incite n'est à chercher ni dans un attachement affectif particulier à cette notion ni dans une admiration pour les propriétés géométriques et algébriques de cette représentation mais dans le fait que cette notion outil du domaine de la statistique est un objet d'enseignement. En devenant objet d'enseignement nous pouvons nous interroger sur les déformations subies par le concept du domaine de la statistique mathématique dans ce processus de transposition didactique[1]. Le terme histogramme est largement utilisé dans divers manuels scolaires, dans certains logiciels tels que des *tableurs* ou des logiciels de statistique.

Les questions principales qui nous guident sont :

Dans quels buts prévoit-on d'enseigner la notion d'histogramme dès la classe de quatrième ?

Quel(s) sens recouvre la notion d'histogramme pour l'enseignant quand elle est effectivement enseignée ? Quel(s) sens prend cette notion chez l'apprenant ?

En admettant que cette notion doive être enseignée, quelles caractéristiques minimales présentes dans le concept doivent être préservées dès l'initiation ? Quelles situations didactiques peut-on construire pour faire apprendre ce concept ?

Nous allons tenter d'exposer nos propres réponses dont nous postulons *a priori* la réfutabilité

Quelques caractéristiques du concept d'histogramme en statistique mathématique.

Tout d'abord quelle peut être l'étymologie du mot *histogramme* ? Le mot se compose du grec *histo* « tissu, texture, trame » et *gramme* « dessin, trace, écrit ». Selon le Grand Robert, le mot *histogramm* apparaît en anglais vers 1903. Notons que *hist-*, *histo-* et *histio*[2]- renvoient à une racine analogue or en grec *histion* désigne « voile de navire, tenture ou toile». Ce mot ne figure ni dans le *Dictionnaire de l'Académie Française* (1822) ni dans l'*Encyclopédie du XIXème Siècle* (1858) ni dans les volumes de mathématiques de

[1] *Transposition didactique, du savoir savant au savoir enseigné* Chevallard, Y., La pensée sauvage, RDM, 1991, 233p

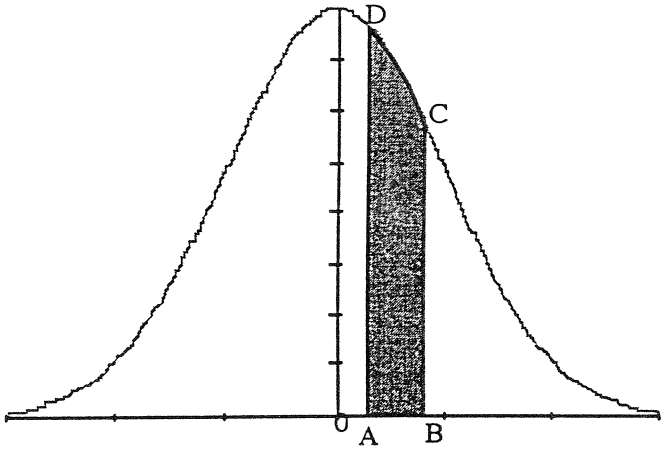
[2] Histiodromie = art de la navigation par le moyen des voiles (Dictionnaire de l'Académie française 1822)

l'*Encyclopédie Méthodique* de D'Alembert et Diderot (1789). En revanche notons que le terme *diagramme* très utilisé figure dans ces trois dictionnaires. D'Alembert & Diderot le définissent ainsi « *en géométrie, c'est une figure ou une construction de lignes, destinée à l'explication ou à la démonstration d'une proposition. Ce mot est plus d'usage en latin, diagramma, qu'en français; on se sert simplement du mot figure* ». Cette définition demeure actuelle. Nous y recourrons pour résoudre notre problème terminologique. L'emploi de l'*histogramme* en tant que représentation graphique serait attribué[3] à A. Guerry en 1833. Une interprétation possible serait qu'une forme fréquente d'histogramme évoque celle de la voile triangulaire d'un bateau. Évidemment cette conjecture devrait être confrontée à ce qu'en pensaient les premiers utilisateurs de ce mot ... dans leurs écrits!

Qu'est-ce que peut être un histogramme ?

Nous allons tenter d'en donner une définition mathématique la plus proche possible du concept sans pour autant développer tous les outils mathématiques qui se trouvent impliqués.

Considérons X une variable statistique (*resp.* une variable aléatoire) continue dont la loi de fréquence (*resp.* la loi de probabilité) est caractérisée par une fonction densité f. L'histogramme est alors la surface délimitée par la courbe représentative de f et l'axe des abscisses.



exemple de fonction densité

$$f(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$$

dont la fonction de répartition

$$\text{est } F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{t^2}{2}) dt$$

(figure 1)

Ainsi la fréquence des observations appartenant à un intervalle [a ; b [est définie par :

$P(\{ \omega \in \Omega ; X(\omega) \in [a ; b[\}) = P([a ; b[) = F(b)-F(a) = \int_a^b f(t)dt$ où F est la fonction de répartition admettant la fonction densité f comme fonction dérivée.

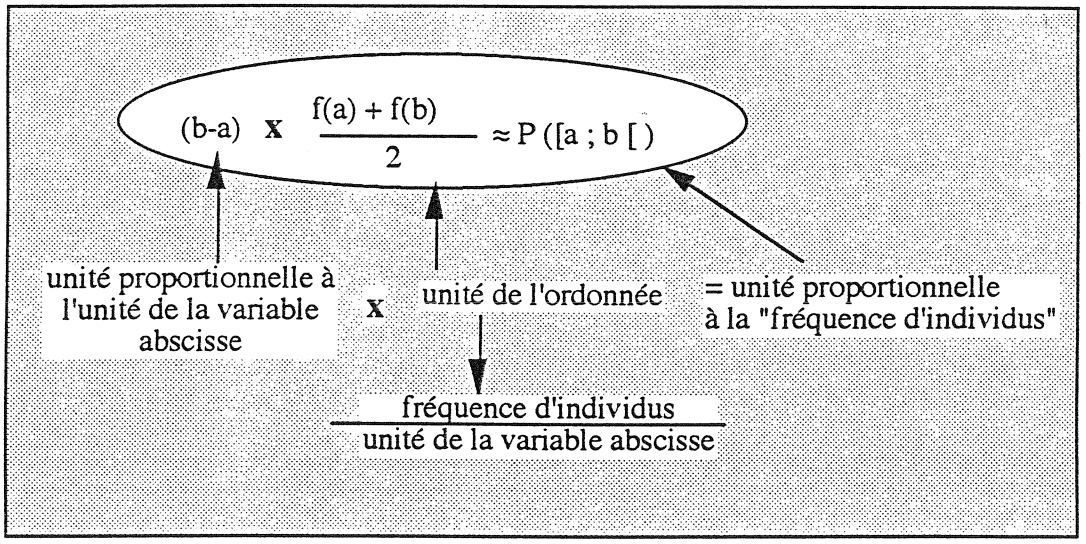
De même nous avons $\int_{-\infty}^{+\infty} f(t)dt = 1$ ou encore $P(]-\infty ; b[) = \int_{-\infty}^{b} f(t)dt$

[3] Droesbeke, J.J, Tassi, Ph, (1990) *Histoire de la Statistique*, PUF-Que sais-je ? n°2527, p 8

Ainsi dans l'histogramme c'est l'aire située sous la courbe qui est à prendre en compte. Cette aire s'obtient à l'aide du calcul intégral. Calculer une probabilité ou une fréquence , c'est donc mesurer une aire délimitée par la courbe de densité.

Intuitivement l'unité de la variable-abscisse est celle de la variable étudiée. Celle de l'aire est une "fréquence d'individus". Ainsi l'unité de la variable-ordonnée est alors "fréquence d'individus / unité de la variable". Observons le domaine ABCD dans la figure précédente. Approximativement[4] nous pouvons le considérer comme un trapèze rectangle rectiligne en estimant que l'arc de courbe CD est proche d'un segment rectiligne. Son aire peut être approchée par la formule suivante $A = (b-a)\frac{f(a)+f(b)}{2}$ qui pourrait être interprétée comme la fréquence des observations appartenant à l'intervalle [a ; b[. Remarquons encore que si nous rapprochons le point A du point B alors le point C se rapproche du point D et à la *limite* le trapèze se réduit à un segment dont la mesure de la surface est égale à 0. Ceci traduit géométriquement une caractéristique des variables statistiques continues qui fournit un résultat que l'entendement humain a sans doute des difficultés à appréhender par l'intuition.

Revenons à une démarche utilisée par le physicien avec les équations aux dimensions :



Ainsi la représentation graphique *histogramme* doit être placée dans un repère tel que

variable - abscisse	unité de la variable étudiée
variable - ordonnée	unité densité de fréquence ou de probabilité
variable -surface	unité : fréquence, probabilité ou effectif

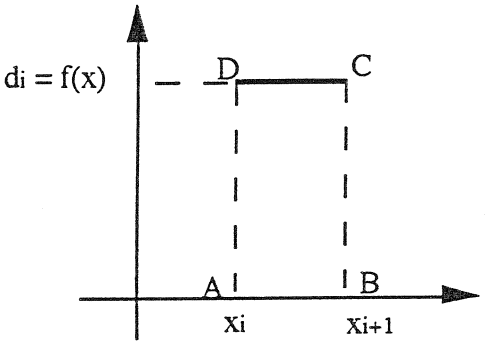
Quand rencontre-t-on cette notion en statistique ?

---

[4] Des précautions doivent être prises à l'égard de cette approximation dont la qualité est tributaire de la proximité des points A et B choisis ainsi que de la forme de la courbe reflétée par les variations de la fonction dérivée f'. Il ne s'agit ici que de suggérer une image grossière.

Dans les études statistiques portant sur une variable quantitative continue telle par exemple la masse mesurée en kg de billes d'acier produites par une machine. Il s'agit d'une mesure au sens du physicien. Les données peuvent être placées raisonnablement dans des intervalles dont l'amplitude correspond à la précision de l'instrument de mesure. Prenons par exemple l'hectogramme. Après quoi, nous choisissons un nombre fini de ces intervalles que nous notons $[x_i ; x_{i+1}[$ avec i = 1 à k. Le choix du nombre d'intervalles est tout à fait discutable. Nous en percevons vite l'arbitraire et il faudrait en analyser les conséquences chaque fois. Mais *a priori* tout nombre positif réel dans un intervalle dont l'amplitude est déterminée par les contraintes physiques de l'objet fabriqué (par exemple : les billes ne peuvent dépasser la masse de 1,2 kg) peut convenir. Ajoutons qu'il pourrait être utile de prévoir une extension en considérant les deux intervalles particuliers, les deux demi-droites $[x_k ; +\infty [$ et $]-\infty ; x_1[$. En ce sens pour modéliser le contrôle du réglage de la machine, il paraît pertinent de choisir le modèle d'une variable continue définie par une loi de fréquence absolument continue c'est à dire définie par une densité de fréquence.

Pour diverses raisons qui pourraient être discutées, nous pouvons considérer que sur un intervalle toutes les valeurs ont *la même chance d'être* le résultat d'une mesure. Ceci se traduit, dans le modèle mathématique adopté, *par* le fait que la densité de fréquence est constante sur un intervalle. Le graphique de la figure 2 traduit cette idée)



(figure 2)

Comment pourrait-on estimer les valeurs d_i ?

En effet ce que le modèle postule *a priori,* c'est que la loi de fréquence est uniforme sur chaque intervalle avec une densité définie comme suit :

$f(x) = d_i$ sur $[x_i ; x_{i+1}[$ avec i = 1 à k et $f(x) = 0$ sur $[x_{k+1} ; +\infty [$ et $]-\infty ; x_1[$.

On procède alors à un sondage en mesurant le plus grand nombre possible d'objets dans des conditions supposées identiques. La résultat de chaque mesure correspondant à un des k intervalles construits *a priori.* Les n mesures seront réparties entre les intervalles et nous obtiendrons :

intervalles	$]-\infty ; x_1[$	$[x_1 ; x_2[$	...	$[x_i ; x_{i+1}[$	...	$[x_k ; x_{k+1}[$	$[x_{k+1} ; +\infty [$
effectifs	$n_0 = 0$	n_1	...	n_i	...	n_k	$n_{k+1} = 0$
fréquences	$f_0 = 0$	f_1	...	f_i	...	f_k	$f_{k+1} = 0$

Rappelons que $\sum_{i=1}^{k} n_i = n$ et la fréquence observée pour chaque intervalle est $f_i = \frac{n_i}{n}$

Pour illustrer notre propos à partir de notre exemple, nous avons une variable dont l'unité est le kg. Essayons alors de concevoir ce qui intervient dans la construction de l'histogramme et comment nous pouvons interpréter.

unités des variables étudiées	*unités correspondantes sur le graphique*
variable "masse" en kg	unité de l'axe des abscisses : 1kg -------> 10 cm = 1 dm 0,1 kg = 1 hg ------> 1 cm
fréquence	unité d'aire 1%-------> 1 cm^2
variable "densité de fréquence"	unité de l'axe des ordonnées : 1%/kg --------> 0,1 cm 1%/hg ---------> 1 cm

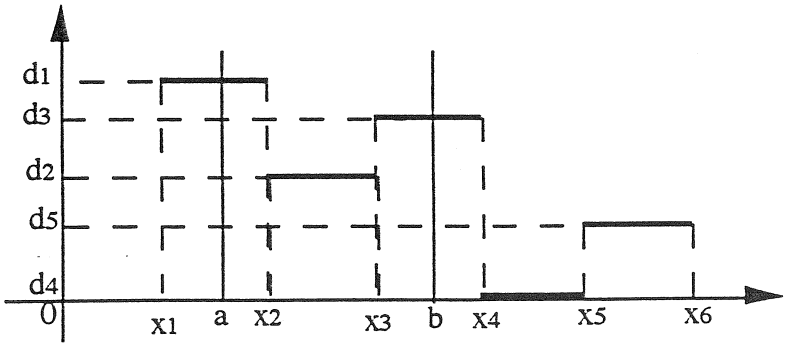
Le calcul de l'aire du rectangle ABCD revient à écrire la relation

$$(x_{i+1} - x_i)\, d_i = f_i = \frac{n_i}{n} \ ,$$

de laquelle nous déduisons $d_i = \dfrac{f_i}{(x_{i+1} - x_i)} = \dfrac{n_i}{n(x_{i+1} - x_i)}$. Nous rappelons que le choix du nombre d'intervalles est arbitraire. Cependant il peut être fait de telle sorte que les amplitudes soient égales, c'est à dire $(x_{i+1} - x_i) = c$ pour $i = 1$ à k et même $c = 1$.

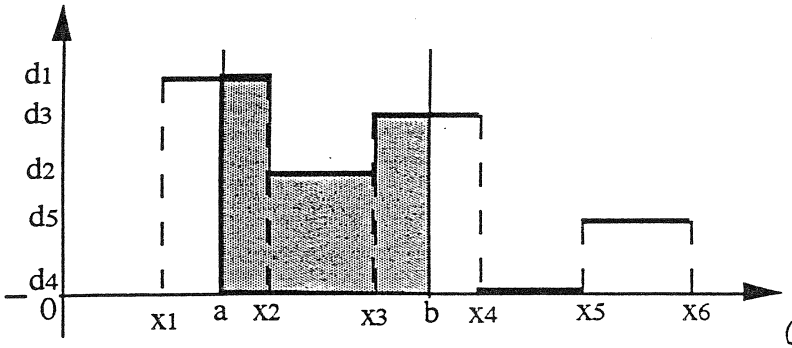
Dans notre exemple, nous pourrions convertir les mesures en hectogrammes et choisir les intervalles d'amplitude 1 hg. Ceci a pour conséquence que la valeur de d_i est égale à la valeur de f_i . C'est ici que naît l'ambiguïté, les deux variables ont même valeur mais ne se réduisent pas l'une à l'autre. Par boutade, ce n'est pas parce que dans une famille de quatre enfants, il y en a un qui est âgé de 4 ans que l'on est amené à confondre l'âge et l'effectif. Ou encore en roulant pendant une heure à vitesse constante 60 km/h, on parcourt 60 km, cela ne réduit pas pour autant la notion de vitesse à celle de distance. Or c'est ce genre de confusion qui est pratiquée dans de nombreux manuels scolaires ou divers ouvrages. Un des buts de cet article est de dénoncer cette confusion par souci de rigueur.

La fonction densité, c'est à dire ce que représente l'histogramme, que nous obtenons par observation statistique est alors une fonction constante par morceaux du type suivant:



(figure 3)

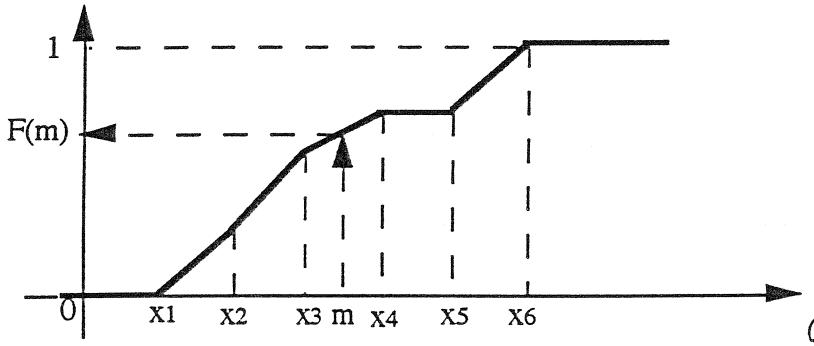
L'aire de la surface sous cette courbe (fig. 3) correspond à la fréquence. On peut alors estimer la fréquence possible des billes dont la masse est comprise entre les deux valeurs a et b en mesurant l'aire de la surface grisée (fig. 4).



(figure 4)

Il s'agit de calculer l'aire de trois rectangles. On pourrait aussi estimer la fréquence possible des billes dont la masse est inférieure à une valeur fixée quelconque. On obtiendrait la fonction cumulative croissante (fonction de répartition) qui n'est autre qu'une primitive de la fonction f.

La courbe est celle d'une fonction non décroissante affine par intervalle du type :



(figure 5)

$$F(m) = f_1 + f_2 + d_3\,(m - x_3) = P\,(X<m)$$

Pour quoi et comment peut-on lisser l'histogramme ?

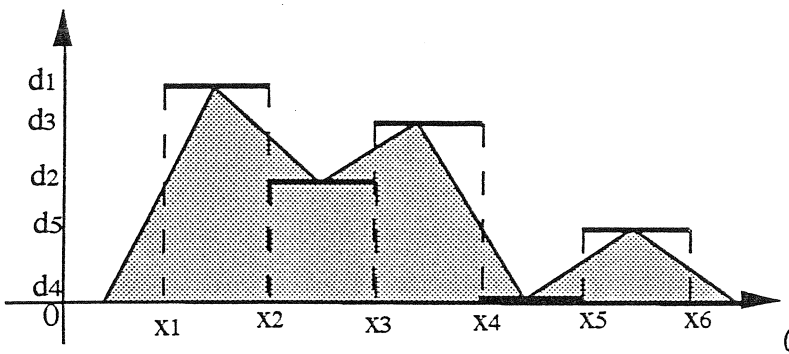
La procédure utilisée précédemment s'inscrit dans l'idée de chercher à estimer la loi de fréquence qui régit la variable "masse" en estimant la densité[5] à partir d'informations partielles issues d'un sondage. Dans le vocabulaire usuelle de la statistique apparaissent les termes de *polygone des fréquences*, du *polygone statistique* ou d'*ogive de Galton.* Les procédures suggérées n'évoquent guère cette idée d'approximation et d'estimation. Il s'agit ici de remplacer l'histogramme représenté par une courbe constante par intervalle, par un autre histogramme représenté par une autre courbe enfermant une surface d'aire égale à la précédente (qui représente l'effectif total ou la fréquence 1) en supprimant les discontinuités

[5] on pourra consulter l'ouvrage
Histogrammes et estimation de la densité, Delecroix, M., (1983) PUF Que sais-je ?

(les sauts entre les segments horizontaux). Il y a évidemment une infinité de solutions à ce problème. Quand les intervalles sont de même amplitude la courbe (fig. 6) passant par les points successifs de coordonnées :

$A_0 (x_1 - \frac{x_2 - x_1}{2} ; 0)$ puis $A_i (\frac{x_{i+1} + x_i}{2} ; d_i)$ pour i = 1 à k-1 et enfin

$A_k (x_{k+1} + \frac{x_{k+1} - x_k}{2} ; 0)$



(figure 6)

A ce stade la nouvelle fonction densité dont l'histogramme est une représentation graphique est une fonction affine par intervalles qui est continue sur l'ensemble des nombres réels. On pourrait continuer à lisser. Mais on peut aussi se contenter de cette représentation pour la comparer à celle des variables connues servant de modèles telles que la variable de Laplace-Gauss pour ne citer que la plus utilisée. Le ou les maxima de la densité déterminent les valeurs modales de la variable étudiée.

Ainsi quel intérêt avons-nous à tracer un histogramme ?

Les quelques propriétés évoquées peuvent en quelque sorte servir à comprendre que l'intérêt de tracer l'histogramme ne se limite pas à l'obtention d'un graphique figuratif à des fins esthétiques et décoratives. Elles permettent aussi de s'opposer à un usage abusif de l'histogramme fondé en statistique descriptive sur une sorte de coutume poussant à tracer des graphiques sans toujours bien réfléchir à leur pertinence (ce qui est encore plus facile de nos jours avec l'usage de l'informatique) peut-être en pensant qu'il y a là une marque de scientificité. En respectant les propriétés qui président à sa définition, l'histogramme constitue un outil graphique permettant de présenter des données statistiques quantitatives issues d'une variable continue ou même issues d'une variable discrète si ces données sont en très grande quantité. Pour illustrer ce dernier cas de figure, nous pourrions prendre l'exemple de la variable « nombre de *pile* obtenu en jetant 10000 fois une pièce de 5 francs ». Essayez de représenter sur l'espace usuel d'une feuille 21x 29,7, le diagramme en bâtons de la distribution des fréquences des 10001 résultats possibles. C'est d'ailleurs ce point de vue qu'avait adopté Laplace dans son traité de *Théorie analytique des probabilités* (1814). A côté de l'intérêt lié au domaine de la statistique, nous pourrions y trouver un intérêt pédagogique en explicitant les notions et méthodes mathématiques que la notion

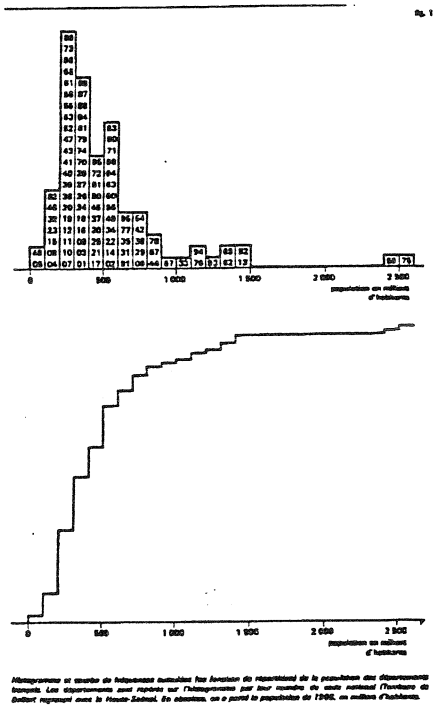
d'histogramme requiert et dont elle constitue une sorte d'exemple d'application. Nous reviendrons plus loin sur cette perspective.

La notion d'histogramme au travers des manuels scolaires ou d'ouvrages non spécialisés en statistique.

Dans *Le Grand Robert* [6] l'auteur donne la définition suivante « Graphique représentant la densité d'effectif en fonction des valeurs d'un caractère et formé par une série de rectangles dont la base constitue un intervalle de variation de ces valeurs et la surface, l'effectif correspondant » qui nous satisfait dans la mesure où elle introduit l'idée de *densité d'effectif.* En revanche *Le Petit Larousse illustré -1989* le définit comme « représentation graphique des classes d'une variable statistique associant à chaque classe un rectangle proportionnel par sa longueur à l'amplitude et par sa hauteur à l'effectif de cette classe » L'auteur l'illustre avec une figure qui ne correspond pas à un histogramme puisqu'il s'agit une variable chronologique des hauteurs en mm des précipitations par mois sur une année!

Comment une telle définition peut-elle conduire un apprenant vers la notion d'histogramme ? S'agit-il d'une contrainte ou d'une dérive de la vulgarisation scientifique ? Cette déformation résulte-t-elle du processus de transposition didactique, d'une négligence ou de l'ignorance ?

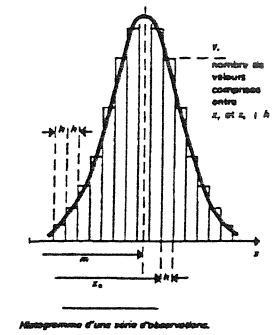
L'encyclopediæ Universalis rapporte à l'article Statistique[7] les deux figures (fig. 7) ci-contre



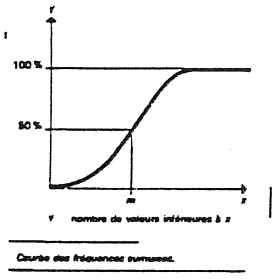
dans un encadré contenant selon la légende un histogramme et une courbe cumulée croissante (fonction de répartition) de la « population des départements français. Les départements sont repérés sur l'histogramme par leur numéro (...). En abscisse, on a porté la population de 1968, en milliers d'habitants ». Nous constatons que ces représentations graphiques ne se rapportent pas à des courbes de fonctions puisque les lignes de rappel sont intégrées à la courbe. Sous la rubrique *statistique descriptive*, on peut aussi lire « la représentation graphique d'un échantillon se fait à l'aide de l'histogramme, ou du polygone des fréquences. On utilise aussi le diagramme des fréquences cumulées ou fonction de répartition ».

[6] article *Histogramme*, p 202, Vol 5 année 1986
[7] article STATISTIQUE, vol 15 -année 1980 p 328 (a)

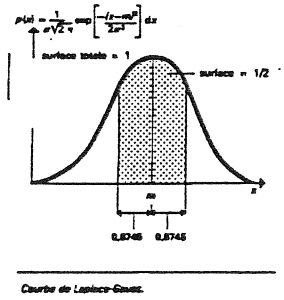
A l'article Mesure[8], nous trouvons trois encadrés contenant respectivement un histogramme d'une série d'observations réalisé à partir de rectangles d'amplitude h (fig. 8), une courbe des fréquences cumulées (fig. 9) et la courbe de Laplace-Gauss (fig. 10)..



Histogramme d'une série d'observations.



Courbe des fréquences cumulées.



Courbe de Laplace-Gauss.

La variable Y portée en ordonnée donne «y_k = nombre de valeurs comprises entre x_k et x_{k+1}» Sur l'histogramme (fig. 1) est posée la courbe de Laplace-Gauss suggérant un lissage. Un texte précise la procédure suivie : « *Dans la pratique, on groupe les résultats par valeurs croissantes. On calcule la moyenne m. On construit une courbe appelée histogramme en divisant la gamme des valeurs en tranches de largeur h et en rapportant, dans chaque tranche limitée par les valeurs x_k et x_{k+1}, le nombre de valeurs expérimentales correspondantes appelé aussi fréquences* ». Outre qu'une ambiguïté demeure quant au sens du mot fréquence : effectif ou proportion ?, le renvoi à l'illustration (fig.10) rappelle qu'il s'agit de la *densité de fréquence* et non de la *fréquence*. A moins que $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-m)^2}{2\sigma^2}) = f(x)$ ne soit considérée comme l'expression algébrique donnant la fréquence (qui requiert la notation du type suivant : $dp(x) = f(x)dx$) et non la densité de fréquence du modèle gaussien. Dans la dernière édition dont nous avons extrait ces trois graphiques les deux figures 8 et 9 sont identiques à la précédente édition et sur la figure 10 la légende de l'ordonnée contient le symbole dx.

Enfin sur la courbe des fréquences cumulées (fig. 8) apparaît en ordonnée la variable Y correspondant à des % alors que la légende stipule « Y = nombre de valeurs inférieures à x ». Certes l'expert réalise immédiatement les rectifications. Et l'article met en évidence un des usages possibles de l'histogramme : un outil d'aide à la modélisation par ajustement d'une loi de fréquence d'une variable à une loi théorique (ici celle de la variable de Laplace-Gauss). Mais il semble que le manque de rigueur est susceptible de maintenir une grande confusion chez le novice-apprenant soucieux d'une précision plus forte de la notion d'histogramme.

[8] article Mesure -Méthodologie de la mesure , vol 10 pp 853-854

Nous avons aussi consulté de très nombreux manuels scolaires des classes de Terminale de Lycée jusqu'à celle de 4ème de collège

Terminale D - Géométrie & Statistique - Aleph 0 - Hachette 1972

programme du 14 mai 1971

A la page 194, la définition est fournie en ces termes :

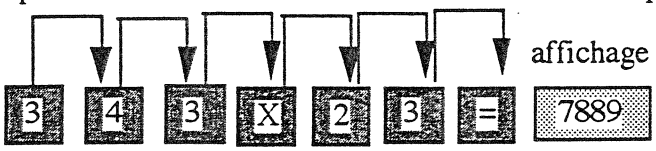
« On appelle histogramme, la représentation graphique d'une statistique d'un caractère dont les valeurs possibles ont été regroupées en classes d'égale amplitude, cette représentation étant obtenue de la manière suivante : sur chaque classe représentée par un segment de l'axe des abscisses, on construit un rectangle dont le côté parallèle à l'axe des ordonnées a une longueur proportionnelle à la classe envisagée ». Puis il est ajouté que « dans le cas où les classes ne sont pas d'amplitude constante, on peut néanmoins établir un histogramme en construisant des rectangles dont l'aire est proportionnelle à l'effectif de la classe envisagée». Pour le moins, l'expression " a une longueur proportionnelle à la classe envisagée" n'est pas dénuée d'ambiguïté. Mais nous pouvons penser que la confusion provient de ce que les auteurs s'astreignent à ne caractériser l'histogramme que par un algorithme de construction.

```
  343
x  23
------
 1029
 686*
------
 7889
```

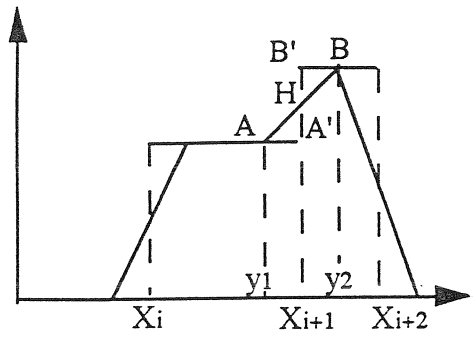
C'est un peu comme si on définissait le concept de multiplication de deux nombres par la seule description d'un algorithme de calcul du type ci-contre ou encore par le mode d'emploi d'un boulier ou d'une calculatrice.

Cela pourrait donner :

On appelle « multiplication de deux nombres le résultat obtenu en appuyant successivement sur les touches :



La notion de "polygone statistique" est abordée en tant que « ligne polygonale obtenue en joignant les extrémités des bâtons du diagramme représentatif de la série statistique correspondante». Est ajouté à ce propos : « Dans le cas où les classes sont d'amplitude inégale, la construction du polygone statistique est plus délicate. Il faut se ramener au cas où les classes sont de même amplitude, de façon que, dans les intervalles séparant les milieux de chacune de ces classes, l'aire limitée par la portion du polygone statistique concernée soit égale à l'aire du rectangle correspondant dans l'histogramme » . Cette remarque est judicieuse dans sa phase de mise en garde mais plutôt ambiguë dans l'expression de la contrainte.



(figure 11)

En observant de plus près, il s'agit de réaliser une construction géométrique (fig. 11) telle que l'aire du trapèze y1ABy2 soit égale à la somme des aires des deux rectangles $y_1AA'X_{i+1}$ et $X_{i+1}BB'y_2$. Le découpage des intervalles choisi donne les points y1 et y2 de telle sorte que les triangles rectangles AA'H et BB'H soient isométriques. Ici on trouve une situation d'application de la géométrie et l'on peut s'étonner qu'elle ne soit pas exploitée.

Nous constatons qu'un nouvel obstacle peut encore être introduit par le fait que le polygone statistique est défini par une procédure de construction non par une problématique comme celle du lissage par exemple. Nous ne retrouvons pas à propos de la statistique, l'exigence que les auteurs se sont assignés à l'égard de la géométrie. En effet nous pouvons lire en préface :

« *Nous pensons que la confusion entre espace vectoriel et espace ponctuel, admissible et même fructueuse à un niveau plus élevé, serait catastrophique ici ; elle conduirait à un retour en arrière fâcheux à un moment où l'on s'efforce à juste titre de toujours préciser exactement la nature des objets mathématiques que l'on est amené à manipuler* ».

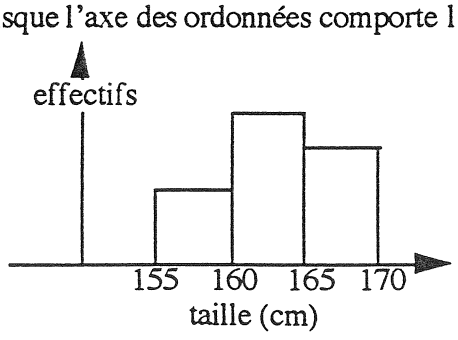
Terminale B Mathématiques -collection Durrande - 1980

programme 1971

A la rubrique *variable aléatoire numérique*, les auteurs introduisent la notion de *densité* par l'expression suivante : « P(X=x) est aussi appelée densité discrète de X en x »

Terminale B Mathématiques , C.Gautier, JC Martin, Hachette -1983

Dans le chapitre STATISTIQUE, la notion d'histogramme est ainsi abordée : « Dans le cas où l'on a opéré un regroupement en classes de valeurs de la variable, on utilise un histogramme». Un exemple est alors fourni au lecteur *les tailles en cm de trente élèves d'une classe de 1ère*. La représentation graphique fournie en guise d'histogramme n'est autre qu'un diagramme en bâtons puisque l'axe des ordonnées comporte l'étiquette *effectifs*.



(figure 12)

C'est ce genre de confusion que nous avons rencontrée dans de nombreux ouvrages. Nous reviendrons plus loin sur les conséquences de cette façon de faire.

Première CDE mathématiques - Tome 1 Monge et coll - Belin 1970

programme du 23 avril 1970

A la rubrique Histogramme on peut lire « on subdivise l'échelle des abscisses en intervalles consécutifs $[X_i; X_{i+1}[$ qui correspondent ... puis l'on construit sur chaque intervalle un rectangle dont l'aire est proportionnelle à la fréquence de la classe correspondante. Ces rectangles obtenus constituent l'histogramme de la distribution.» Puis est pointé le cas particulier : « lorsque toutes les classes ont la même étendue, on construit chaque rectangle en prenant une hauteur égale à la fréquence». La présentation de l'histogramme satisfait aux exigences de la notion mais le cas particulier génère à nouveau la confusion qui se trouve d'ailleurs confirmée par un dessin.

Première A mathématiques - H. Pochard -Gauthier-Vilars 1971

programme du 23 avril 1970

Quatre pages sont consacrées à l'histogramme. Après une étude préalable sur les effets d'un changement de repère cartésien orthogonal sur le calcul de l'aire d'un rectangle et un exemple d'*histogramme des effectifs*, l'auteur propose la définition suivante : « Les histogrammes sont des diagrammes en surface où l'on représente les effectifs y_i (resp : les fréquences f_i) par des rectangles dont les aires sont proportionnelles aux y_i (resp : les fréquences f_i)»

Première F mathématiques - Col DIMATHEME - Didier- 1988

programme de 1985

Par définition « l'histogramme des effectifs est formé de rectangles ayant des aires proportionnelles aux effectifs des classes ». Cependant la confusion réapparaît avec le cas des intervalles d'amplitude égale où l'axe des ordonnées devient l'axe des effectifs.

Seconde - mathématiques algèbre - Audirac, Bonvallet -Magnard- 1981

programme 1981

Cet ouvrage ne comporte pas de définition explicite mais à la rubrique *variables statistiques continues* apparaît une procédure de construction et un graphique conformes à notre définition. L'indication de l'unité d'aire en légende en témoigne.

Seconde - mathématiques - Glaymann, Malaval -CEDIC 1981

programme 1981

Dans un chapitre intitulé "analyse des données", l'histogramme est présenté conformément à notre définition. Il est abordé par une série statistique dont les classes sont d'inégale amplitude puis les auteurs précisent : « Dans le cas particulier où les classes sont d'égale amplitude, le calcul des hauteurs des rectangles est simplifié : il suffit de prendre pour ordonnée des valeurs proportionnelles aux effectifs». Et ils ne proposent pas de mettre en ordonnées les effectifs !

Seconde - mathématiques IREM Strasbourg -ISTRA- 1981

programme 1981

L'histogramme est abordé dans le thème Graphiques divers. Il est conforme à notre définition comme le confirme la légende donnant l'information 4% pour dire 1 cm^2 correspond à 4%. L'exemple choisi porte sur une variable discrète "nombre de lettres des mots d'un texte". Ce qui n'est peut-être pas très pertinent sauf à réaliser une extension d'une variable discrète à une variable continue. Mais cette perspective n'est pas explicitée.

Seconde - mathématiques - coll LOUQUET -Armand Colin- 1986

programme de 1986

Dans le paragraphe consacré au cas d'*un caractère continu - les diagrammes à bandes*

« Dans le cas où l'on dispose de classes statistiques, les intervalles des classes (qui ne sont pas forcément tous de même amplitude) sont reportés sur un axe. Ils servent de bases à des rectangles dont les aires sont proportionnelles aux effectifs.» Un dessin est alors proposé sans l'axe des ordonnées. Ce qui nous semble être un moyen didactique pertinent pour éviter la dérive que nous dénonçons qui conduit à l'interprétation de l'axe des ordonnées comme axe des effectifs. « Le diagramme obtenu s'appelle un histogramme». A cela les auteurs ajoutent une mise en garde précieuse : « Attention, dans le cas où les classes sont d'amplitudes inégales, à bien calculer les hauteurs pour que ce soit les aires qui soient proportionnelles aux effectifs»

Seconde - mathématiques - coll Perspectives -Hachette- 1990

programme de 1990

L'histogramme est présenté par sa construction à partir de l'exemple de la variable *"prix d'une robe exprimé en francs"*. Le graphique est conforme à notre définition et comporte la légende 1cm2 représente 20 robes . La courbe des fréquences cumulées

croissantes jouxte l'histogramme. Elle est cependant incomplète dans la mesure où elle est limitée aux points extrêmes observés.

Troisième - mathématiques - IREM de Strasbourg -ISTRA- 1993

Le chapitre 14 intitulé STATISTIQUES (au pluriel) comporte la rubrique *savoir présenter des données statistiques et les exploiter.* Cette rubrique explicite un objectif de savoir faire. Aucune définition explicite de l'histogramme n'est fournie mais un problème[9] est proposé. A partir de l'étude de la variable *"durée de vie des ampoules mesurée en heures"*, on demande de «représenter le tableau par un histogramme (un diagramme en barres)» en donnant la contrainte « on prendra 1 cm pour 100 heures et 1 cm pour 100 ampoules».

Pourquoi confondre ces deux notions ?

Le terme diagramme en barres ne suffit-il pas ?

Où peut donc bien apparaître l'idée de la prise en compte de la surface dans une telle approche ?

Reportons à un autre exercice[10]. Il s'agit d'une série statistique portant sur la variable durée du trajet entre le collège et le domicile mesurée en minutes. Le résultat est rapporté graphiquement par un histogramme (le mot est employé). Cependant si l'axe des abscisses est bien l'axe des durées avec des classes d'amplitudes 5 minutes (1 cm), l'axe des ordonnées est désigné comme l'axe des effectifs. Evidemment malgré ce défaut il est possible de répondre aux deux questions posées :

1) Combien y a-t-il d'élèves dans ce collège ?

2) Combien d'élèves mettent moins d'un quart d'heure ? Quel pourcentage cela représente-t-il ?

En effet celles-ci n'impliquent que des classes entières. Il suffit donc d'utiliser les effectifs de chaque classe qui, pour des raisons purement techniques, coïncident avec les valeurs représentatnt les hauteurs des rectangles.

Mais nous aimerions savoir comment les auteurs répondraient aux questions suivantes:

1) Quel pourcentage d'élèves mettent plus de 7 minutes et moins de 2 minutes ?

2) Quel pourcentage d'élèves mettent moins de 23 minutes ?

3) Quel pourcentage d'élèves mettent exactement 23 minutes ?

Avec le modèle proposé, nous serions curieux de connaître la réponse des auteurs du problème en prenant appui sur ce qu'ils dénomment *histogramme.*

---

[9] extrait de Brevet des Collèges de Lille -1990

[10] n° H page 217

Quatrième - mathématiques - IREM de Strasbourg -ISTRA- 1993

Le chapitre 15 intitulé STATISTIQUES (au pluriel) comporte l'activité sur le "Tracé d'un diagramme en barres". Les auteurs justifient ce choix par « la plupart du temps, une enquête est accompagnée d'un dessin appelé diagramme» Ce texte nous paraît mettre davantage l'accent sur la fonction illustratrice et esthétique du diagramme que sur sa fonction d'expression d'une notion dans un registre sémiotique graphique dans le but d'en faciliter le traitement. Après quoi les auteurs livrent l'algorithme de construction du diagramme. « ... notre enquêteur a commencé le diagramme ci-dessous constitué d'un rectangle dont la *base* est la largeur de la tranche de prix concernée et la *hauteur*, le nombre d'élèves correspondant (on dit l'effectif)» Certes les auteurs emploient le terme de diagramme en barres et non celui d'histogramme alors que la description suggère celle d'un histogramme. Ceci se confirme car ils nomment diagramme en bâtons la représentation graphique de la distribution des effectifs d'une variable quantitative discrète. En voulant sans doute éviter le terme, les auteurs introduisent une ambiguïté dont nous aimerions apprécier les conséquences. Faut-il voir dans le diagramme en barres une représentation qui n'est déjà plus un diagramme en bâtons sans être encore un histogramme? Est-ce un effet de transposition didactique aménageant une sorte d'espace transitionnel ?

Continuons l'exploration de l'exemple. L'aire du rectangle qu'ils proposent en modèle mesure 4,5 cm^2 et représente 9 individus si on lit l'ordonnée, cependant au sens de l'histogramme sur l'intervalle [2 ; 2,5[la densité d'effectif vaut constamment 9 individus par 0,5 F soit encore 18 individus/Franc. Il est alors demandé de reproduire et compléter ce diagramme appelé diagramme en barres puis de fournir le nombre d'élèves et le pourcentage parmi les élèves interrogés prêts à payer au moins 3F. Force est de constater que la seconde question n'exige en aucune façon l'usage du diagramme qui demeure dans sa fonction illustratrice.

Pourquoi et dans quel but n'a-t-on pas transformé cette variable en une variable qualitative ordinale à 5 modalités pour laquelle le diagramme en barres serait devenu avec pertinence un diagramme en bâtons ? Ce qui pourrait alors être l'occasion d'un lecture graphique de la modalité dominante (mode). Nous ne percevons pas où l'activité de lecture d'une représentation graphique peut se déployer.

De toute évidence les auteurs ont souhaité introduire une représentation géométrique prenant en compte la surface. Mais pourquoi en évitent-ils l'usage ?

Une question pertinente pourrait être : A combien peut-on estimer le nombre d'élèves qui sont prêts à payer entre 2,25 F et 3,75 F ? Celle-ci suppose deux conditions. D'une part que l'on introduise l'idée de la perte d'information initiale, sinon il suffit de retourner à la série statistique initiale et de dénombrer l'effectif de l'intervalle [2,25; 3,75[. D'autre part

qu'une information inconnue est remplacée par une estimation. Cette opération d'estimation peut être mise en relation avec des pratiques usuelles de chacun face des situations d'ignorance. Il nous semble que nous entrons ici de plain pied dans le champ de la statistique. La question ci-dessus introduit l'intérêt d'une représentation en surface et en impose une lecture adéquate. Quoiqu'il en soit, nous ne voyons pas ce qui s'oppose à l'introduction d'une représentation par surface proportionnelle à l'effectif ou à la fréquence. Les opérations de construction se trouveraient ainsi modifiées : « ... le diagramme... est constitué d'un rectangle dont la base correspond à la largeur de la tranche de prix concernée et l'aire correspond au nombre d'élèves...» Ces consignes conduisent à des tâches plus riches portant sur des objets plus pertinents dans le domaine de la statistique mathématique.

Réflexions sur le traitement de la notion d'histogramme dans l'enseignement de la statistique comme branche des mathématiques.

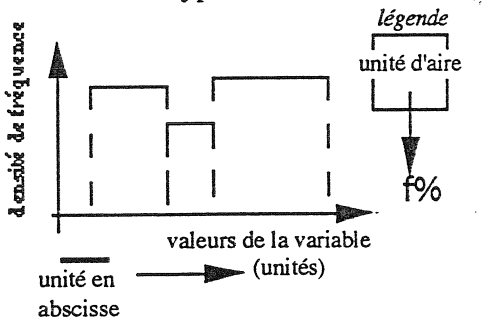
Certes il est possible de trouver des manuels ou des traités de mathématiques et de statistique qui se conforment à une définition cohérente de la notion d'histogramme. Mais force est de constater que les remarques que nous avons faites ne se limitent pas à l'échantillon des ouvrages que nous avons pris ici en exemple. Or imaginons un étudiant ou un enseignant souhaitant au moins s'informer sur la notion, il sera très probablement confronté à la situation que nous avons dénoncée. Ainsi comment ce novice pourra-t-il se faire une idée correcte de cette notion ?

Mais alors sur quoi se fonde la présentation de l'histogramme choisie par les auteurs en particulier celle qui interprète l'axe des ordonnées comme axe des effectifs ou des fréquences ?

Nous pouvons penser que cela tient à l'ignorance des auteurs ou plutôt à leur désintérêt pour la notion. Celle-ci se trouverait présentée dans les ouvrages uniquement parce qu'elle doit y figurer au nom des programmes scolaires. A moins que cela tienne à une phénomène de transposition didactique, c'est à dire compte tenu du fait que la notion implique des concepts de mathématiques avancées, il est réduit à ses caractéristiques élémentaires explicitées au travers d'un algorithme de construction géométrique. Qui plus est, le choix d'intervalles de même amplitude constitue un modèle adapté à de nombreuses situations de mesure qui se trouve renforcé par la facilité de la construction. Ainsi le pédagogue estime lever un obstacle à l'apprentissage rapide de la notion d'histogramme en évitant l'affrontement à une représentation graphique surfacique qui doit être lue simultanément dans ses deux dimensions. Le problème est qu'alors la propriété particulière d'égalité numérique de la hauteur et de la fréquence (ou resp de l'effectif) est rapidement convertie en identité conceptuelle des deux objets: "La hauteur représente l'effectif, la fréquence ou la probabilité". Ceci se produit à l'insu de l'apprenant.

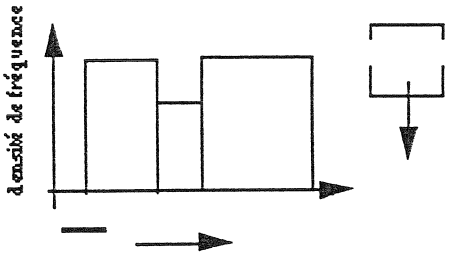
Nous pensons que cette approche didactique est génératrice d'obstacles didactiques.[11] Elle contribue à la construction d'une représentation mentale chez le sujet apprenant qui fait obstacle à la compréhension de l'histogramme quand il est amené à en faire usage dans un sens autre que celui d'une illustration fonctionnant comme un diagramme en bâtons. Nous l'avons maintes fois constaté dans le cadre de l'enseignement de la statistique à l'université[12]. Nos étudiants expriment leur étonnement ou déclarent leur incompréhension complète quand nous leur expliquons que le *hauteur* (c'est à dire l'ordonnée d'un point) dans l'histogramme ne correspond pas à la *fréquence* mais à une *densité de fréquence*. Le coût pour dépasser cet obstacle nous paraît plus élevé dans ce contexte que dans celui de l'apprentissage initial. L'approche de la construction de *l'histogramme* à partir de classes d'amplitudes inégales suivi de la construction d'un *lissage* du type *polygone statistique*, nous paraît être souhaitable dès la classe de Quatrième. Il y a une occasion d'activités mathématiques mettant en jeu divers cadres de référence : cadre géométrique, cadre numérique, cadre algébrique, cadre statistique, cadre de l'analyse fonctionnelle, mais aussi divers registres sémiotiques de représentation : registre graphique, registre discursif, registre algébrique.

Par exemple pour ne pas induire l'identité *histogramme = rectangle*, nous suggérons des tracés du type suivant :



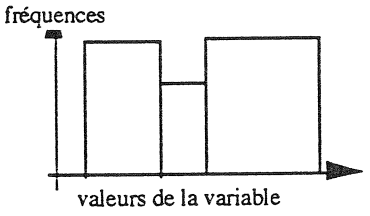
((figure 14)

au lieu de



(figure 15)

Et naturellement nous rejetons les tracés qui ne donnent pas les bonnes informations minimales. Ainsi nous bannissons cette représentation graphique (fig 16)



(figure 16)

[11] Au sens de Bachelard, G., qui parle d'obstacle pédagogique en 1938 dans *la formation de l'esprit scientifique* (p 18 VRIN-1989)

Au sens de Brousseau, G., qui donne un exemple à partir de l'enseignement des nombres décimaux fondé sur les mesures avec les unités et les sous-unités. Cette approche conduit des individus à concevoir les nombres décimaux comme des nombres constitués de deux parties autonomes : partie entière et partie décimale dont on retrouve les traces dans des erreurs du type $(0,2)^2 = 0,4$ mais non détectable pour $(0,5)^2 = 0,25$.

[12] en licence et en maîtrise de sciences de l'éducation

A côté des manuels, il ne faudrait pas oublier les logiciels qui donnent un accès facile aux traitements en permettant des constructions automatiques de diagrammes nommés histogrammes mais qui ne sont, à ce que nous connaissons à ce jour, que des diagrammes en bâtons. C'est le cas d'un intéressant et pratique logiciel de traitements statistiques tel que CHADOC[13] dans lequel l'auteur désigne les diagrammes par histogrammes. C'est aussi le cas d'un tableur tel que Excel dont les possibilités sont immenses. Dans sa version actuelle comportant l'opération de traçage des courbes en coordonnées cartésiennes, il est possible de procéder à la réalisation de n'importe quel *histogramme* en considérant qu'il s'agit de tracer la courbe d'une fonction constante par intervalle puis celle d'une fonction affine par intervalle pour l'obtention de *polygone statistique*. Encore convient-il que les utilisateurs aient acquis quelques connaissances mathématiques qui fondent ces notions pour obtenir ce résultat avec un ordinateur, car ce dernier ne pense pas à la place du sujet, il ne fait que des traçages automatiques! Notons que dans ce cadre des traitements statistiques assistés par informatique, il y a de nombreuses autres réflexions et recherches à conduire en lien avec la multiplication des recours aux représentations graphiques les plus variées réalisées automatiquement sans aucun effort par l'utilisateur et même sans nécessité de compréhension. En prenant les perspectives nomographique, épistémologique et heuristique dans le cadre de la statistique, à quoi servent en fait les représentations graphiques ?. Dans les cadres didactique et pédagogique, comment peut-on enseigner à construire, lire ou choisir une représentation graphique, que celle-ci soit à produire dans la classe ou qu'elle soit produite ailleurs et introduite en classe par l'intermédiaire des *media* ?

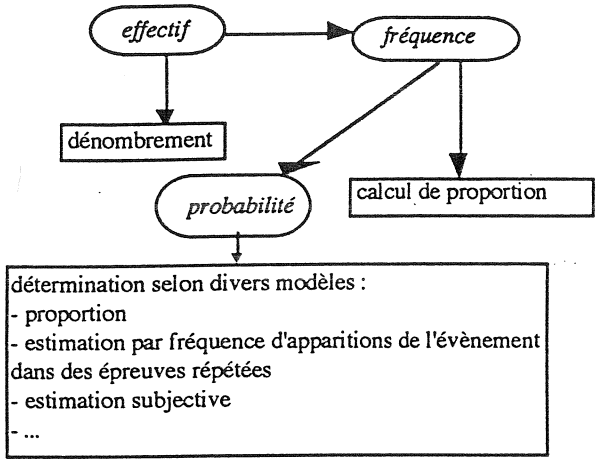
Il nous semble que nous tenons là un exemple de notion mathématique introduite par l'enseignement en dehors de toute problématique pertinente. Face à cela, nous pouvons adopter deux conduites :

- ne pas l'enseigner
- l'enseigner sur la base de problématiques pertinentes accessibles à des élèves des différents niveaux scolaires.

Nous rappelons que notre critique à l'égard de l'enseignement de l'histogramme ne se fonde nullement sur un culte particulier auquel nous vouons cette notion. C'est tout le contraire. Nous souhaitons proposer cette notion dans un réseau notionnel pour éviter une dérive d'un enseignement de l'histogramme pour lui-même. L'histogramme est un objet mathématique qu'il faut éviter de réifier. Il importerait de préciser selon les circonstances s'il est traité comme une notion-outil ou comme une notion-objet. Examinons maintenant dans quel réseau notionnel nous pourrions l'immerger. Cette connaissance peut alors servir pour construire des séquences didactiques mettant en jeu diverses notions, méthodes et concepts

[13] IUT de Nice

mathématiques à des degrés variés d'abstraction. Elle peut aussi conforter l'idée qu'il ne s'agit pas d'une notion greffée arbitrairement.



Trois notions sont impliquées *effectif* , *fréquence* et *probabilité* qui renvoient aux opérations ci-contre. A celle de *fréquence* est souvent attachée celle de *pourcentage*..

Examinons maintenant quelques jeux de cadre possibles. La définition de l'histogramme peut être exprimée:

- dans un cade géométrique :

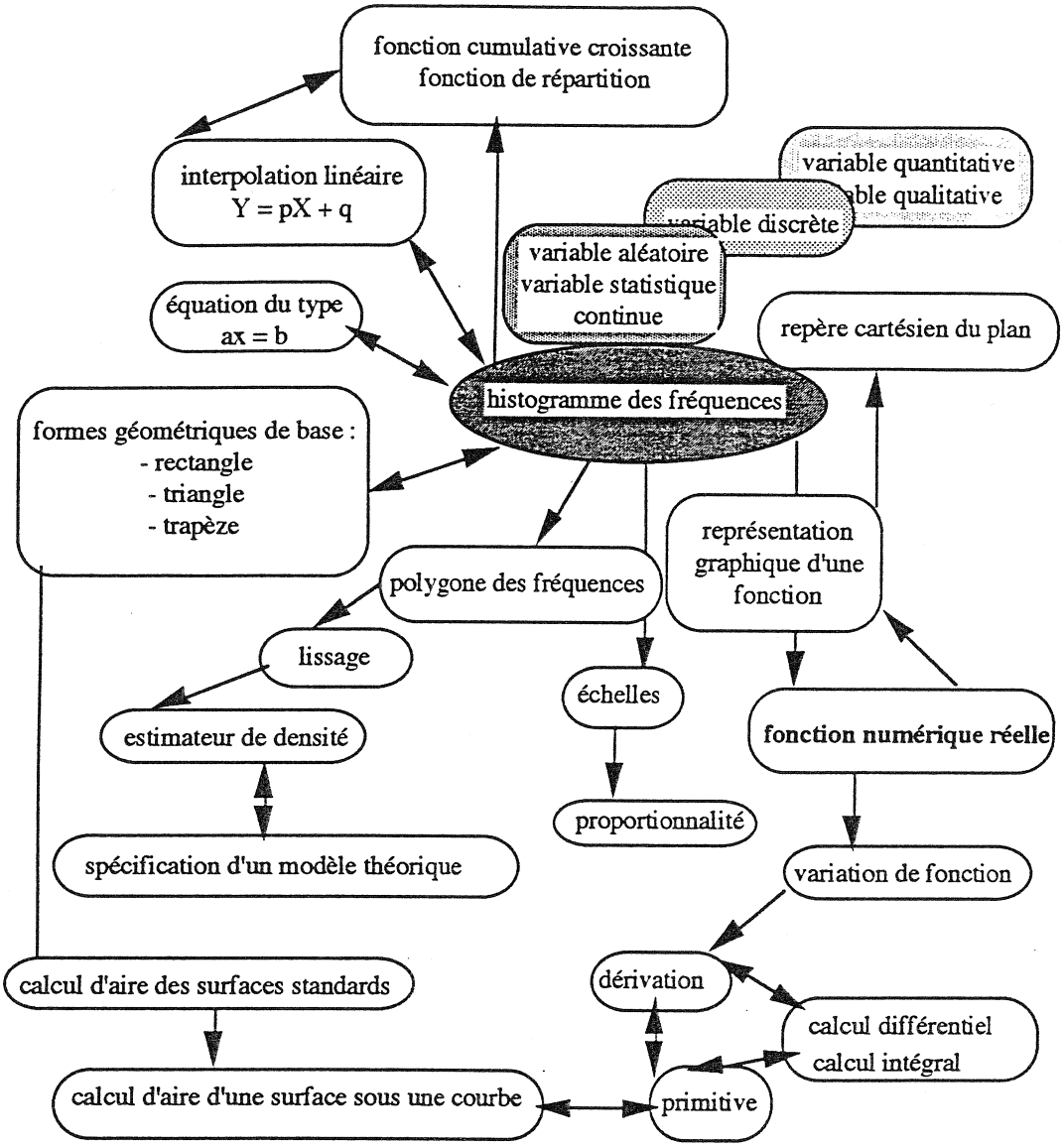
« Un histogramme est une surface dont l'aire totale est égale à 1...».

- dans un cadre géométrique analytique :

« Dans un repère cartésien orthogonal, un histogramme est une surface délimitée par une courbe représentative de la densité de fréquence d'une variable statistique continue dont l'aire totale est égale à 1...»

- dans un cadre algébrique :

« un histogramme est l'ensemble des couples (x ; y) tels que pour x nombre réel de l'ensemble définition Df, $0 \leq y \leq f(x)$, et où x est une valeur possible, résultat de la variable statistique et f(x) est la densité de fréquence de cette valeur x ...»



Lorsque f(x) est une expression algébrique connue, nous sommes ramené à un problème classique d'étude d'une fonction numérique de variable réelle prévue dans les programmes scolaires. Dans l'introduction habituelle de l'histogramme, il s'agit tout simplement de réaliser l'étude d'une fonction constante par intervalle (fonction en escalier). Les intervalles étant caractérisés par la situation étudiée.

En ce qui concernent les registres sémiotiques, nous avons à utiliser du texte, des écritures symboliques, des tableaux et des représentations graphiques géométriques.

Pour quelles raisons et dans quels buts l'enseignement habituel de l'histogramme n'établit-il pas clairement ces liens et ne met-il pas les apprenants dans des situations propres à établir les correspondances et les diverses traductions qu'elles imposent ?

Quels avantages cet enseignement tire-t-il en isolant la notion d'histogramme du cadre important des fonctions ?

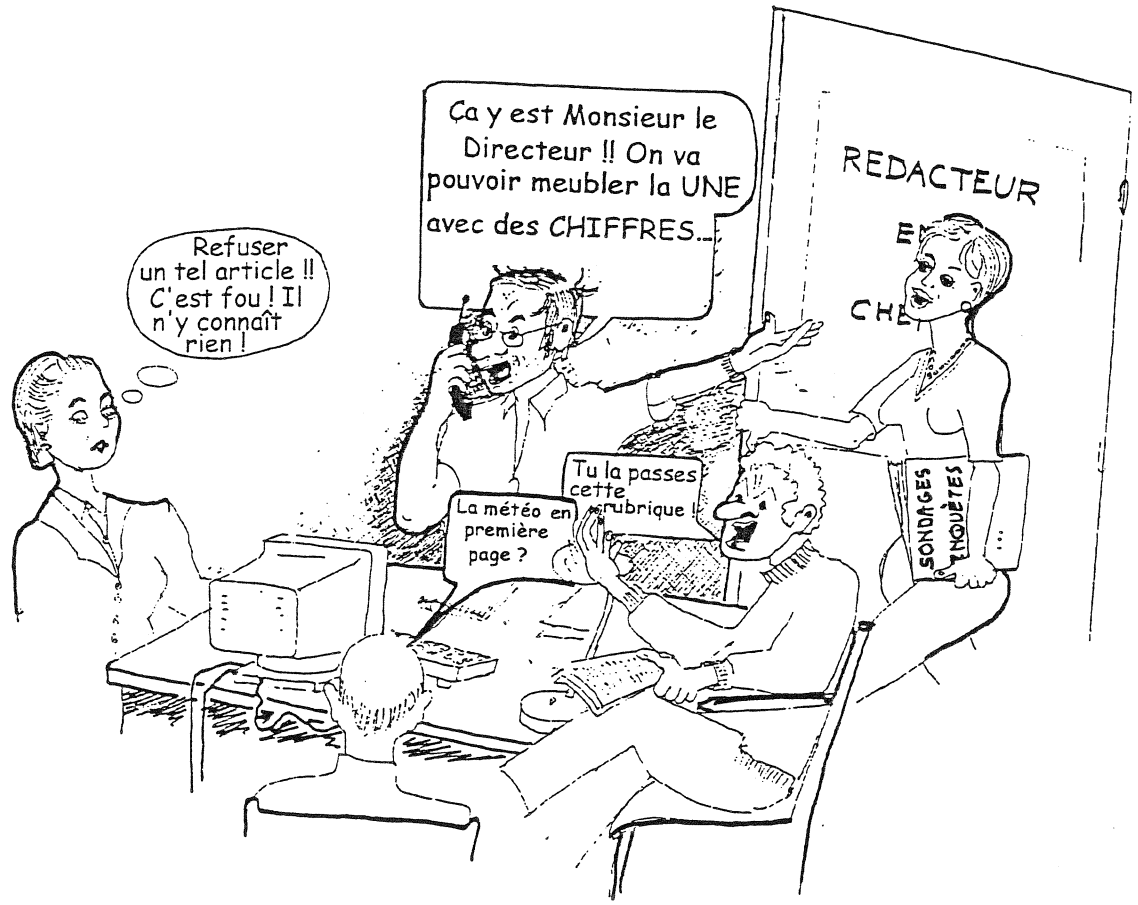
Ne peut-on pas dès la classe de quatrième introduire des situations qui évitent de retirer à l'histogramme sa finalité majeure qui est de représenter des variables par des surfaces dont les aires traduisent les valeurs ? Remarquons que le problème attaché au calcul de l'impôt sur le revenu est du même type.

Si nous imaginons que le modèle initial d'histogramme est celui présenté en classe de quatrième[14], comment quelques années plus tard un étudiant pourra-t-il comprendre que l'histogramme de la variable de Laplace-Gauss ("la courbe en cloche") se rattache à la même idée que celle développée à propos de la distribution des effectifs de la variable *"prix d'une barre glacée"* ?

Le lecteur aura compris qu'au-delà des critiques fondées sur les observations de nos étudiants en cours d'apprentissage de statistique et des outils d'aide à l'apprentissage que sont les manuels, les dictionnaires ou les logiciels, nous avons posé plus de questions que nous avons apporté de réponses. Toutefois la question formulée est un premier pas vers la résolution du problème qu'elle traduit. Notre objectif est avant tout d'attirer l'attention des enseignants de mathématiques. Mais la notion d'histogramme est aussi l'affaire de bien d'autres disciplines scolaires et il serait utile de connecter notre réflexion à d'autres, semblables, conduites en géographie, en physique ou en biologie par exemple. Nous demeurons certain que des situations didactiques pertinentes peuvent être organisées par des enseignants. C'est d'ailleurs à ce prix que nos conjectures seront validées ou réfutées et que notre propos ne sera pas dogmatisé.

---

[14] *op. cit.*


Refuser un tel article !! C'est fou ! Il n'y connaît rien !
Ça y est Monsieur le Directeur !! On va pouvoir meubler la UNE avec des CHIFFRES...
REDACTEUR
CHE
La météo en première page ?
Tu la passes cette rubrique !
SONDAGES
ENQUÊTES

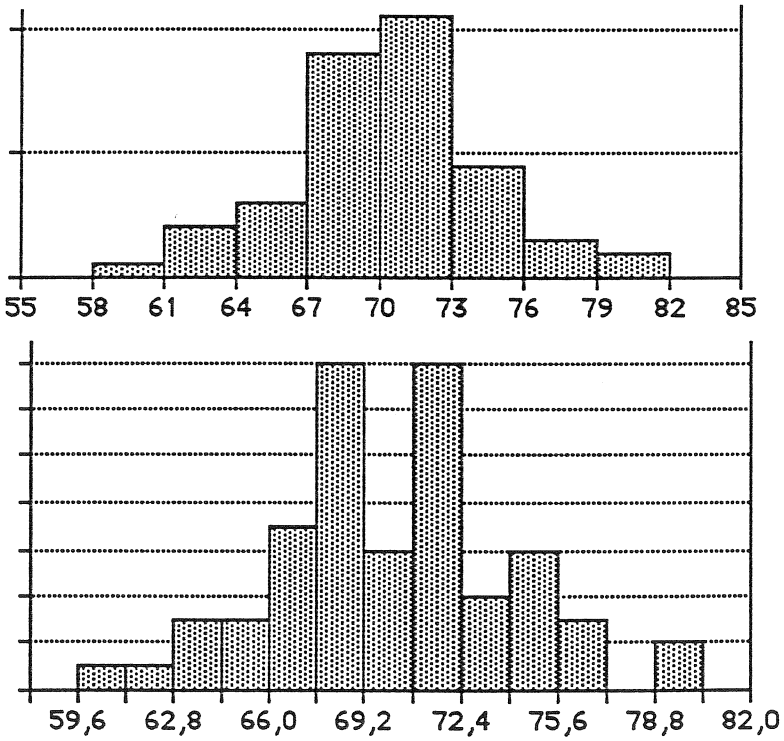
Chapitre 3

Des diagrammes à l'histogramme

Jean Claude Girard

1) Introduction

La construction de l'histogramme posent de nombreux problèmes aux élèves de tout âge : concept d'intervalle semi-ouvert, choix du nombre d'intervalles ou de leur largeur, difficultés pour utiliser une échelle ou la proportionnalité, concept de densité de fréquence[1]. La simple lecture de ce type de graphique n'est pas sans difficulté, même pour les adultes. Par exemple, il n'est pas évident de voir, au premier coup d'oeil, que les deux histogrammes suivants représentent la même série statistique.



[1] Voir l'article "Mise au point sur les histogrammes", page 21.

Le deuxième montre qu'il n'y a pas de valeurs de la série entre 77,2 et 78,8, ce qui n'apparaît pas sur le premier. Les formes mêmes de ces deux histogrammes sont très différentes. Ceci est dû à la largeur des classes qui est de 3 dans le premier cas et de 1,6 dans le deuxième. Pour les deux graphiques, il a été nécessaire, pour avoir des classes de même largeur, de commencer et de finir à des valeurs qui ne sont pas le minimum et le maximum de la série. Ceci est souvent le cas mais on a tendance à l'oublier.

On se propose de montrer dans cet article différents graphiques proches de l'histogramme et qui, sans avoir toutes les difficultés déjà citées, présentent même, dans certains cas, quelques avantages.

Sans être obligé de passer par toutes les étapes proposées ni se sentir enfermé par la linéarité de la présentation, ils pourront constituer une progression vers l'histogramme.

1° exemple

Les données suivantes représentent la dureté, en indice "Rockwell", de 60 pièces mécaniques après trempage[2]. Les valeurs ont été ordonnées et arrondies pour travailler, dans un premier temps, sur des entiers.

	1	2	3	4	5	6
1	51	54	55	56	57	58
2	51	54	55	56	57	58
3	52	54	55	56	57	58
4	53	55	55	56	57	58
5	53	55	55	56	57	59
6	53	55	55	56	57	59
7	54	55	55	56	57	59
8	54	55	55	56	57	60
9	54	55	56	56	57	60
10	54	55	56	57	58	61

Un premier graphique consiste à construire une liste des entiers compris entre le minimum (51) et le maximum (61) et à placer à côté de chacune des valeurs observées autant d'étoiles que d'occurences de cette valeur.

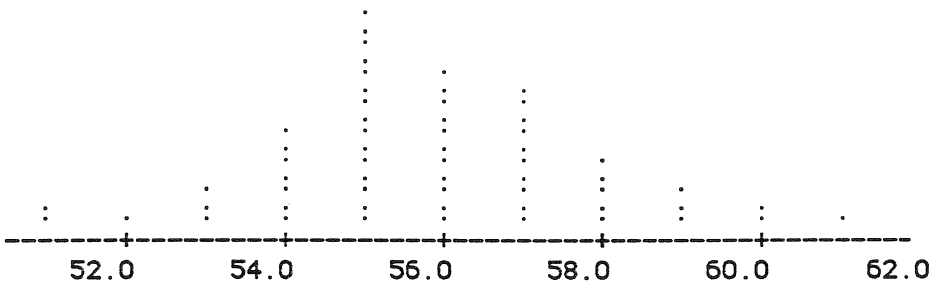
---

[2] Les données de l'exemple sont extraites de *Maîtrise statistique des procédés,* Gérald Baillargeon, Les éditions SMG, Trois Rivières, Québec, 1992.

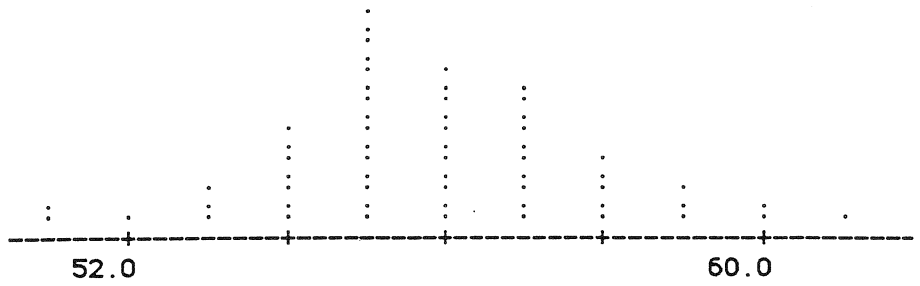
	Count	
51	2	**
52	1	*
53	3	***
54	7	*******
55	15	***************
56	11	***********
57	10	**********
58	5	*****
59	3	***
60	2	**
61	1	*

Ce graphique ne présente pas d'autres difficultés que de savoir compter si ce n'est de mettre en place une bonne organisation pour éviter les oublis ou les répétitions.

Pour progresser vers l'abstraction, on peut construire une graduation en face de laquelle on place un point par valeur observée. Ce graphique appelé "Dot-Plot" ou graphique en points introduit la difficulté de l'échelle. Les effectifs n'apparaissent plus. Pour plus de lisibilité les points sont groupés deux par deux.



On pourra alterner des constructions de graphiques et des lectures avec la consigne, par exemple, de retrouver l'ensemble des valeurs d'une série à partir d'un graphique tel que le suivant :



On observe déjà sur ces deux premiers graphiques la forme de la distribution, à savoir : relativement symétrique, unimodale avec une forte concentration autour de la valeur centrale.

2° exemple

On s'intéresse maintenant aux valeurs non arrondies (on pourrait évidemment travailler sur une autre série statistique).

	1	2	3	4	5	6
1	50.9	51.4	51.9	52.5	54.0	54.2
2	53.5	53.6	53.8	54.6	54.7	54.7
3	54.4	54.6	54.6	55.1	55.1	55.2
4	54.9	55.0	55.0	55.4	55.5	55.5
5	55.2	55.2	55.2	56.0	56.1	56.2
6	55.6	55.6	56.0	56.5	56.6	56.7
7	56.2	56.3	56.4	57.1	57.1	57.2
8	56.7	56.9	57.0	58.0	58.1	58.3
9	57.4	57.6	57.8	60.0	60.2	60.6
10	58.9	59.0	59.3	52.9	53.0	54.0

Les deux graphiques précédents n'ont plus d'intérêt car il y a 44 valeurs différentes.

On pense alors à regrouper ces valeurs mais l'histogramme n'est pas notre seul recours. Un graphique qui pourrait être présenté préalablement est le graphique en tiges et feuilles ou "Stem and Leaf"[3].

---

[3] John W. Tukey-*Exploratory Data Analysis*-Addison Wesley-Reading MA-1977.

```
Stem-and-leaf of C1          N  = 60
Leaf Unit = 0.10

    1    50 9
    3    51 49
    5    52 59
    9    53 0568
   19    54 0024666779
  (13)   55 0011222245566
   28    56 001223456779
   16    57 0112468
    9    58 0139
    5    59 03
    3    60 026
```

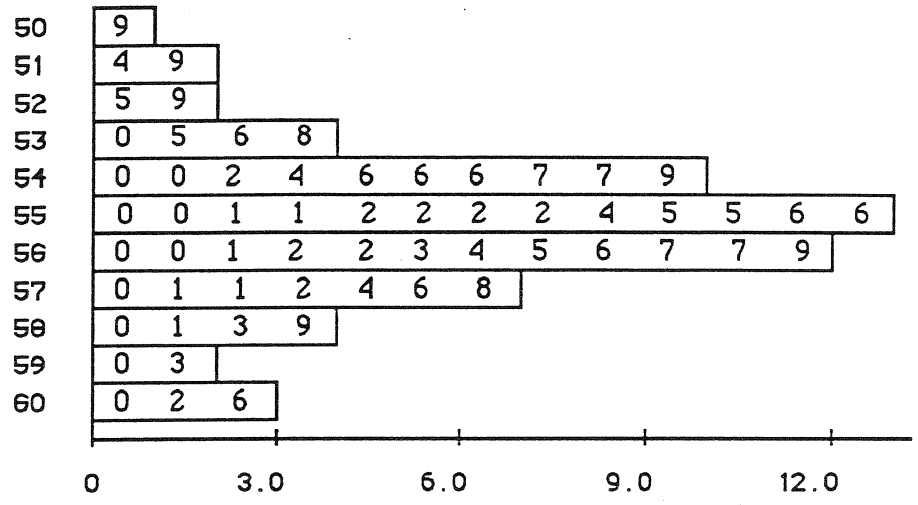
Explications : (à faire trouver par les élèves sur un graphique déjà fait, à partir du tableau des valeurs).

La première ligne donne la première valeur de la série 50,9. La deuxième ligne donne les deux valeurs suivantes : 51,4 et 51,9. La "tige" 51 donne la partie entière des deux valeurs, les "feuilles" 4 et 9 indiquent les décimales de chacune. L'effectif cumulé (3) figure en début de ligne. A l'autre extrêmité du graphique, les effectifs sont cumulés en partant du maximum. L'effectif de la classe médiane (c'est-à-dire celle qui contient la médiane) figure au début de la ligne entre parenthèses (13).

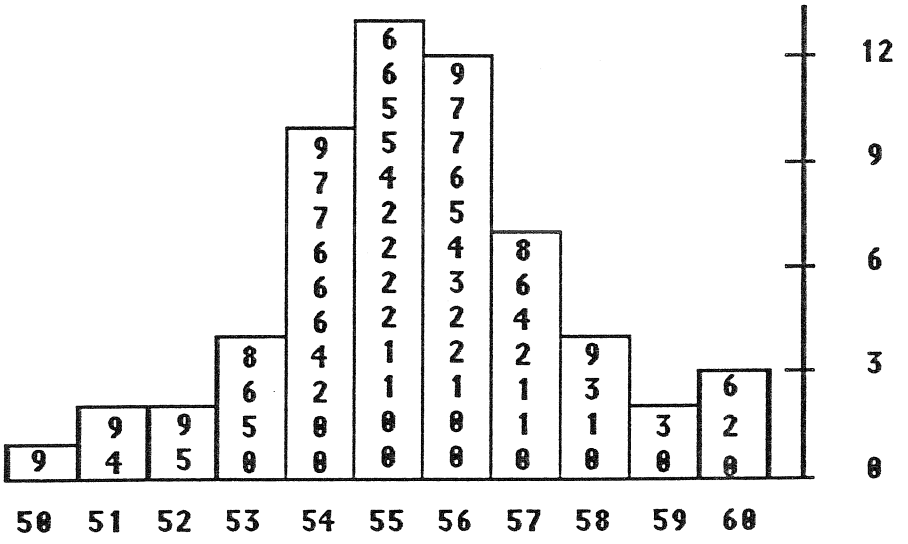
La simplicité de ce graphique est évidente par rapport à l'histogramme et pourtant il revient à regrouper les valeurs dans des intervalles ouverts [50 ; 51[, [51 ; 52[etc.

Il a de plus l'avantage par rapport à l'histogramme de ne perdre aucune information sur les valeurs de départ. On peut retrouver la série exacte à partir du graphique (c'est bien entendu un exercice à faire).

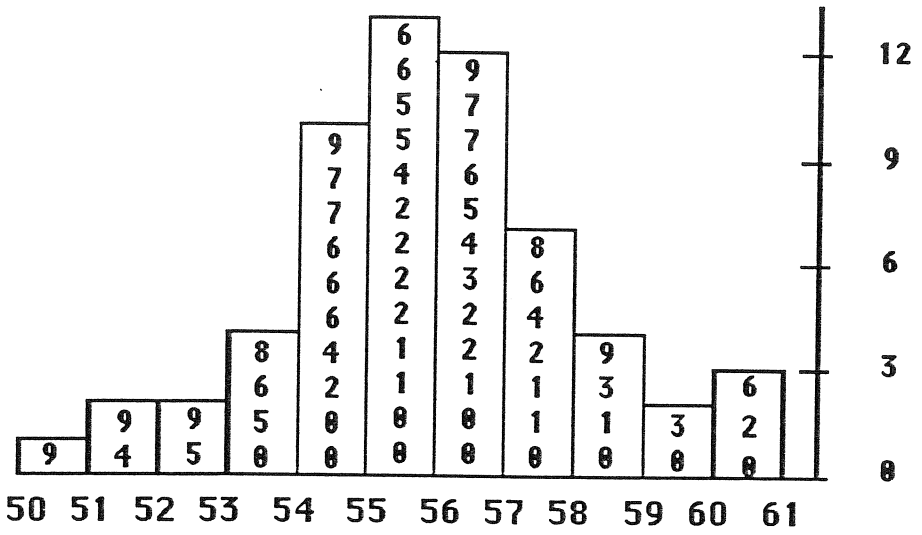
On pourrait le présenter sous une forme proche de l'histogramme :



ou encore :



et enfin l'histogramme,



Une variante consiste à faire deux tiges par unité :

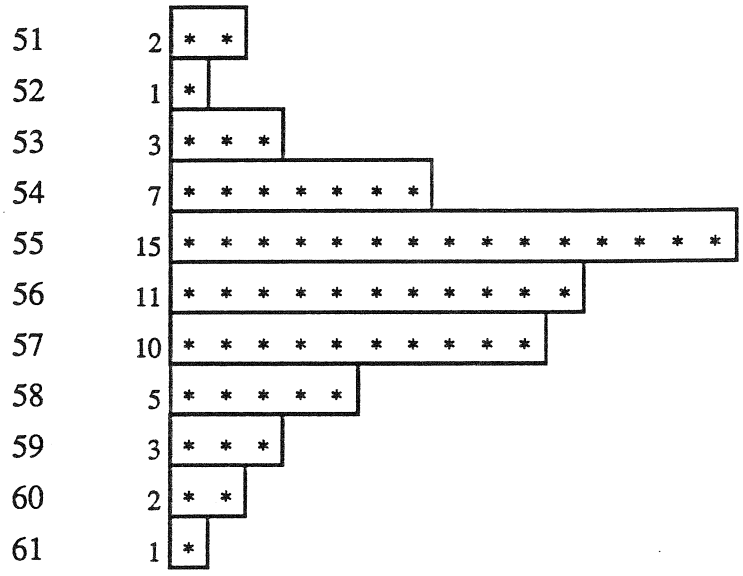
```
Stem-and-leaf of C1          N  = 60
Leaf Unit = 0.10

     1    50 9
     2    51 4
     3    51 9
     3    52
     5    52 59
     6    53 0
     9    53 568
    13    54 0024
    19    54 666779
    28    55 001122224
   (4)    55 5566
    28    56 0012234
    21    56 56779
    16    57 01124
    11    57 68
     9    58 013
     6    58 9
     5    59 03
     3    59
     3    60 02
     1    60 6
```

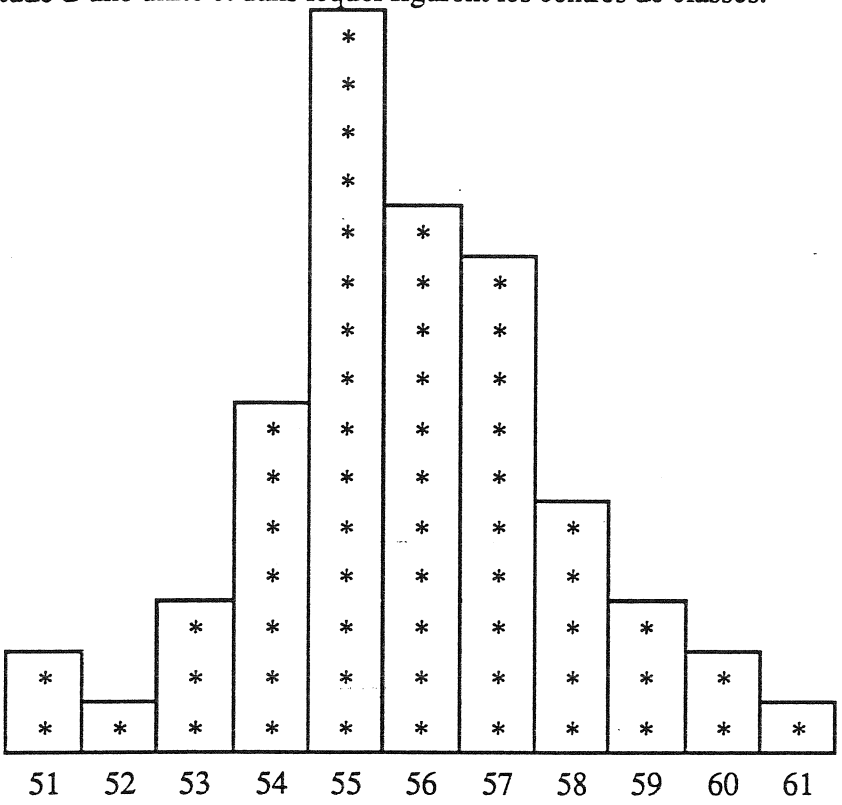
Explications : La première tige correspond à l'intervalle [50,5 ; 51[, la deuxième à l'intervalle [51 ; 51,5[, la troisième à l'intervalle [51,5 ; 52[. Autrement dit, la première donne les valeurs de la série que l'on arrondirait "en dessus" à 51, la deuxième aux valeurs que l'on arrondirait "en dessous" à 51, etc. L'ensemble de ces deux lignes représentent donc les valeurs de l'intervalle [50,5 ; 51,5[c'est-à-dire celles que l'on arrondit à 51. On a vu dans le graphique (à points ou à étoiles) des valeurs arrondies (partie I) qu'il y en avait 2. Si cela n'a pas été fait, on peut alors le tracer :

Valeurs arrondies à :			
	51	2	**
	52	1	*
	53	3	***
	54	7	*******
	55	15	***************
	56	11	**********
	57	10	**********
	58	5	*****
	59	3	***
	60	2	**
	61	1	*

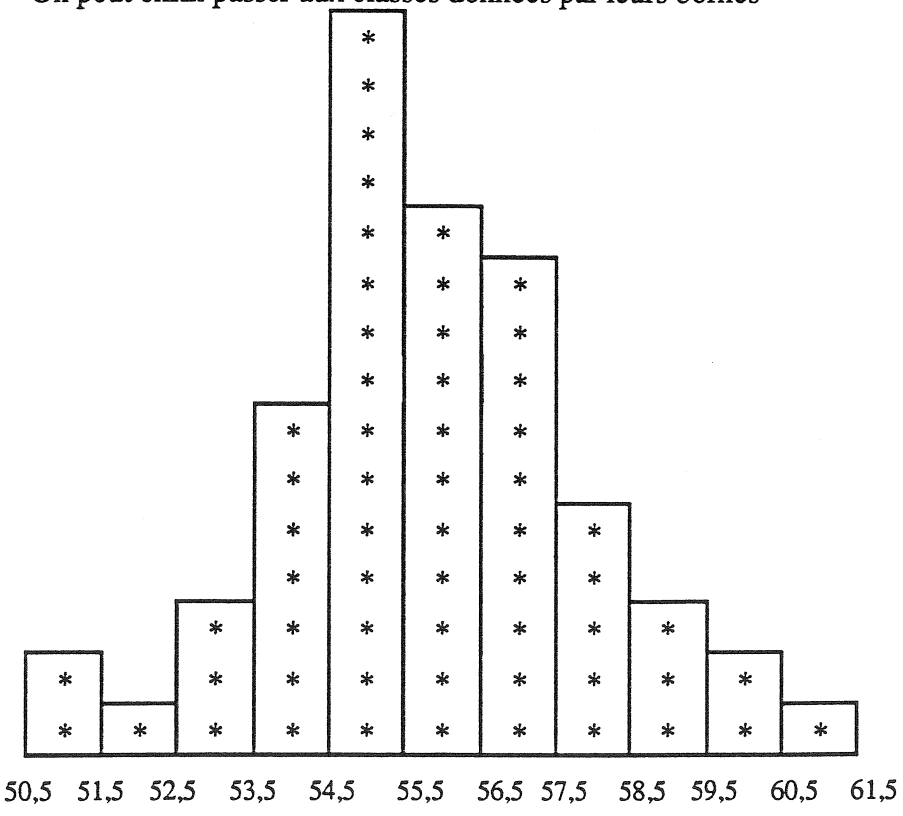
On n'est alors pas très loin de l'histogramme si l'on dessine les rectangles autour des étoiles :



On obtient, en faisant une rotation, un histogramme dont les classes ont une amplitude d'une unité et dans lequel figurent les centres de classes.



On peut enfin passer aux classes données par leurs bornes



Analyse didactique de la lecture de graphique

La première fonction d'un graphique, quelqu'il soit, est de donner des informations sur une série statistique : forme de la distribution, répartition des valeurs autour d'une valeur centrale, dispersion, valeurs isolées ou lointaines.

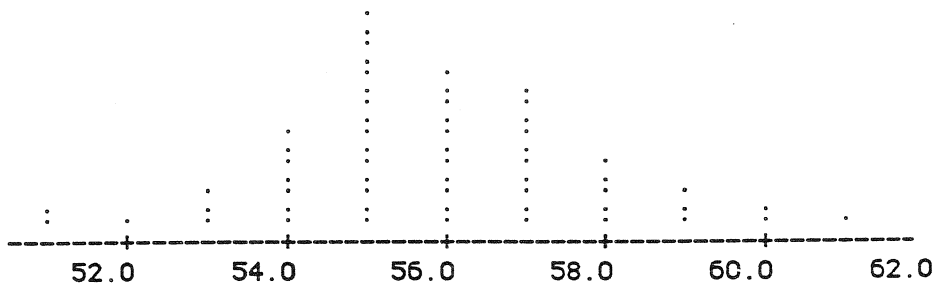
Les difficultés de lecture ne sont cependant pas du même niveau conceptuel ou d'abstraction.

1) Graphique en tiges et feuilles (stem and leaf)

```
  1    50 9
  3    51 49
  5    52 59
  9    53 0568
 19    54 0024666779
(13)   55 0011222245566
 28    56 0012234567 79
 16    57 0112468
  9    58 0139
  5    59 03
  3    60 026
```

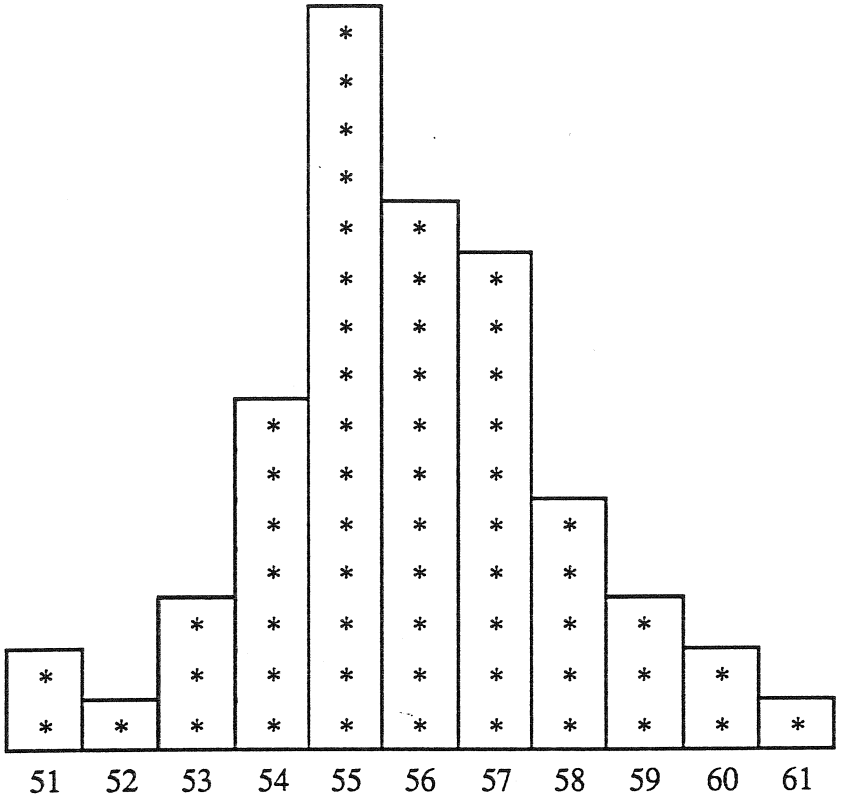
Toutes les valeurs de la série sont lisibles dès lors que l'on a compris qu'une tige est utilisée pour plusieurs feuilles. Les effectifs peuvent être recomptés ou déduits des effectifs cumulés.

2) Graphique en points (dot-plot)



L'utilisation de la graduation horizontale permet de retrouver les valeurs de la série (ou une valeur approchée si l'échelle est trop petite). On peut connaître les effectifs de chacune en comptant les points.

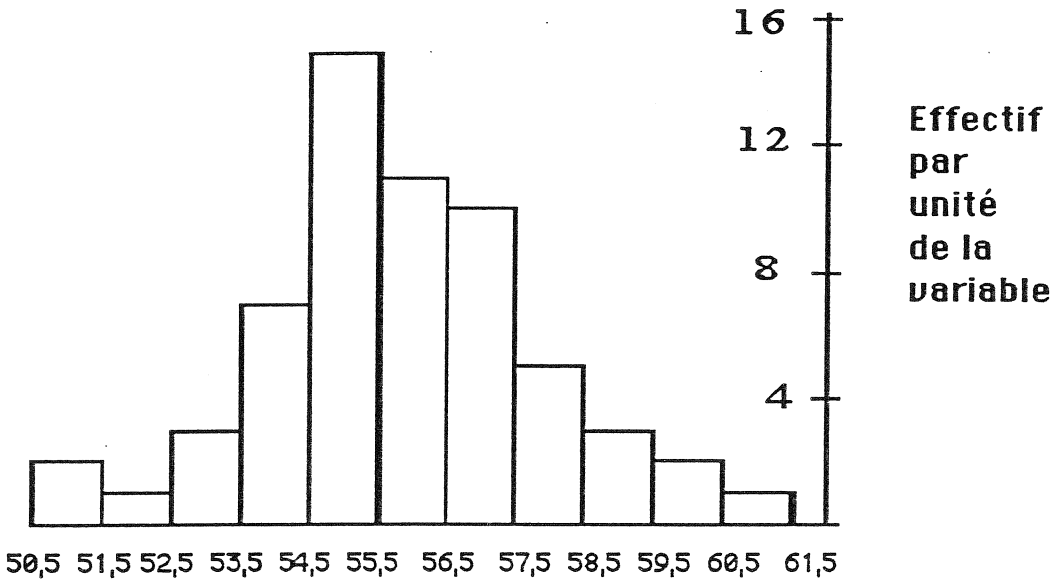
3) Diagramme en étoiles



On peut connaître la valeur exacte ou un encadrement des valeurs (suivant le cas) et compter combien figurent dans chaque intervalle.

Pour les grandes séries, une étoile (*) peut représenter par convention plusieurs valeurs, ce qui nécessite alors l'utilisation de la proportionnalité pour retrouver les effectifs.

4) Diagramme en barres ou histogramme avec intervalles de même amplitude



On peut connaître un encadrement de chacune des valeurs et on peut avoir une idée des effectifs de chacune :

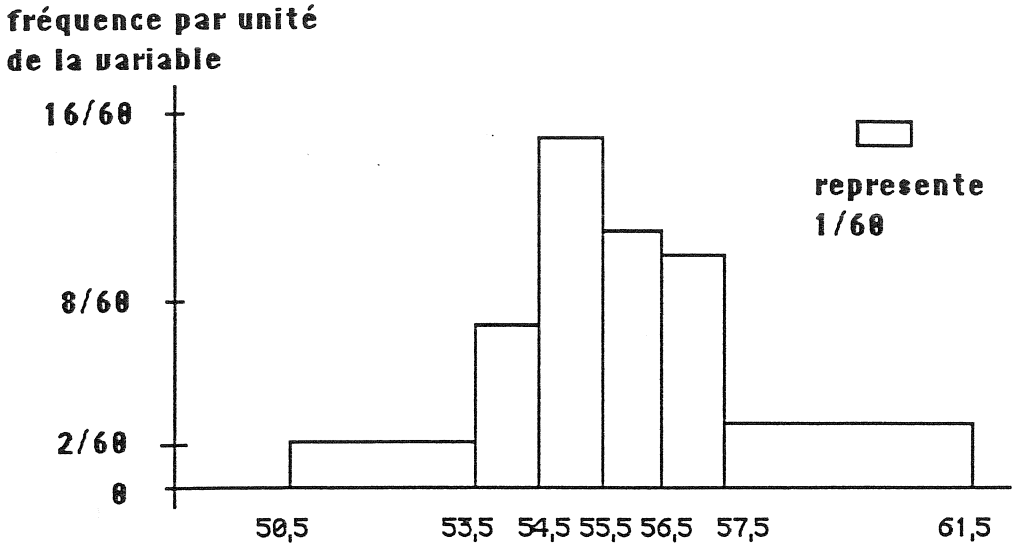
- par perception globale des rectangles
- en mesurant les longueurs des rectangles
- en repérant ces longueurs par rapport à une graduation verticale

Dans tous les cas, il y a utilisation, implicite ou explicite de la proportionnalité.

On peut également remarquer que la lecture du diagramme en barres ou de l'histogramme avec intervalles de même amplitude, fait appel à un double décodage, un premier pour les valeurs et un autre pour leurs effectifs.

Dans les deux graphiques précédents, les quantités (de valeurs) étaient représentées par des quantités (de points, d'étoiles) alors qu'ici, les quantités sont représentées par des longueurs. Le décodage ne peut se faire qu'en appréciant visuellement ou en mesurant ces longueurs.

5) Histogramme



Les effectifs peuvent être trouvés par le calcul en multipliant la densité de fréquence de l'intervalle par son amplitude ou appréciés visuellement par rapport à l'unité de surface.

L'histogramme est la représentation graphique d'une fonction (en escaliers) et non un assemblage géométrique de rectangles.

Il combine les difficultés inhérentes aux grandeurs quotients (comme la densité de fréquence) et celles liées à la proportionnalité.

Il semble difficile de bien décoder ce type de graphique sans connaître la notion de fonction et en particulier de fonction densité de probabilité.

En résumé :

	Permet de retrouver les valeurs exactes	Nécessite la maîtrise des graduations (1 ou 2)	Nécessite la proportionnalité	Nécessite le codage effectif-longueur	Nécessite le codage effectif-surface
Tiges et feuilles	oui				
Points	oui	1			
Etoiles		1	éventuellement		
Diagramme en barres		2	oui	oui	
Histogramme		2	oui	oui	oui

HISTOGRAMME

De l'histogramme à la fonction densité de probabilité

Jean Claude Girard

1) Introduction

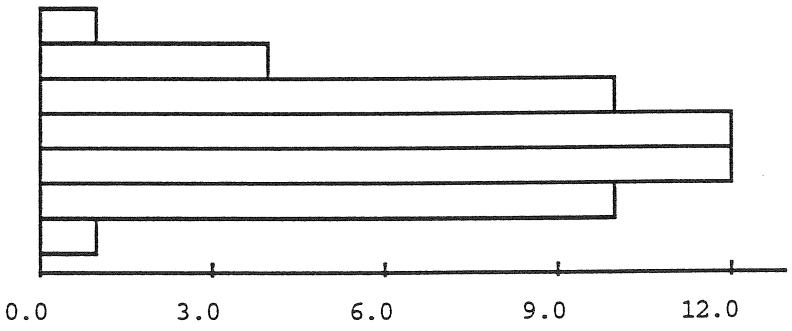
Les données suivantes sont extraites du livre de G. Saporta, "*Probabilités, Analyse des Données et Statistique*" - Technip - Paris - 1990. Elles représentent les hauteurs de 50 pièces usinées.

1	21.86	21.92	21.91	21.97	22.01
2	21.84	21.90	21.91	21.98	21.96
3	21.88	21.91	21.92	21.95	21.95
4	21.90	21.89	21.91	21.89	21.95
5	21.92	21.91	21.93	21.98	21.97
6	21.87	21.87	21.96	21.96	21.96
7	21.90	21.89	21.91	21.98	21.95
8	21.87	21.90	21.97	21.95	21.94
9	21.90	21.89	21.97	21.97	21.97
10	21.93	21.92	21.97	21.94	21.95

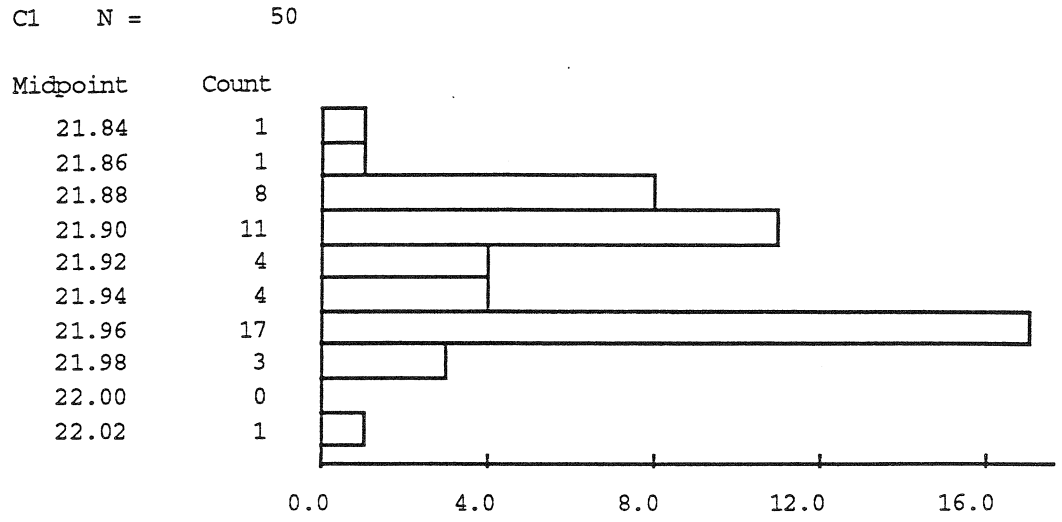
Un premier histogramme (largeur des classes de 0,03) montre une distribution approximativement symétrique.

C1 N = 50

Midpoint	Count
21.83	1
21.86	4
21.89	10
21.92	12
21.95	12
21.98	10
22.01	1



Mais si l'on réduit l'amplitude des classes à 0,02, l'histogramme fait apparaître cette fois une distribution bi-modale qui est caractéristique, généralement, du mélange de deux populations (matières premières ou machines différentes).



La question qui se pose est donc : Quelle est la densité de probabilité (au niveau de la population) qui se cache derrière les données de l'échantillon ?

2) Estimation de densité par la méthode de la fenêtre mobile

Il s'agit de déterminer f(x) pour toutes les valeurs possibles x de la variable aléatoire représentant la hauteur d'une pièce usinée au niveau de la production. Pour cela, on construit un intervalle d'amplitude h centré sur x : $\left[x - \frac{1}{2}h; x + \frac{1}{2}h \right[$ puis on calcule, pour chaque valeur de x, le nombre n_X de valeurs de la série contenues dans cet intervalle. On estime alors f(x) par $\frac{n_x}{n.h}$ où n est l'effectif de l'échantillon.

3) Méthode du noyau

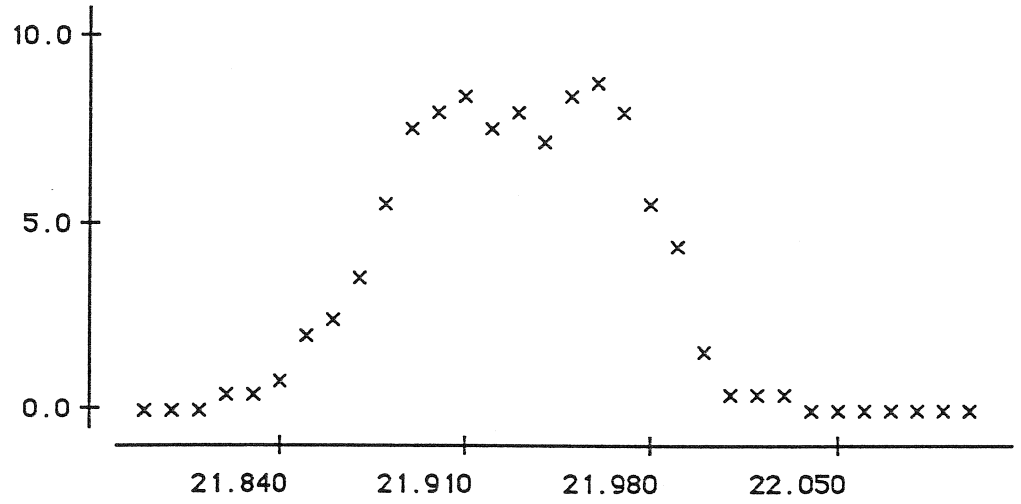
Dans la méthode précédente, la quantité n_X peut s'écrire $n_x = \sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)$ où K représente la fonction caractéristique de l'intervalle $\left[-\frac{1}{2}; +\frac{1}{2} \right[$.

On dit que l'on a effectué une estimation de densité par la méthode du noyau. On peut prendre comme noyau une autre fonction, par exemple un noyau gaussien défini par :

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}.$$

4) Application

Le calcul, par la méthode de la fenêtre mobile, effectué sur ordinateur avec h=0,05 (c'est-à-dire environ 30 % de l'étendue de l'échantillon), conduit à une estimation de la densité de probabilité de la forme suivante :



On a donc confirmation, au niveau de la population, du caractère bimodale de la distribution. Ceci est également visible sur le graphique en tiges et feuilles[1] des 50 valeurs de l'échantillon :

```
Stem-and-leaf of C2          N  = 50
Leaf Unit = 0.010

   1  218 4
   5  218 6777
  10  218 89999
  21  219 00000111111
 (6)  219 222233
  23  219 44555555
  15  219 66667777777
   4  219 888
   1  220 1
```

1 Voir chapitre précédent.

Encore
20 mouches et
notre poids moyen
passe au-dessous
des 10 kilos !!!

Chapitre 5

A bas la moyenne !

ou

A propos des paramètres de tendance centrale
et de dispersion d'une série statistique.

Jean Claude Girard

1) Introduction

Les concepts de tendance centrale et de dispersion sont utilisés pour décrire ou résumer une série statistique. Ces deux concepts sont mal définis ou, plus précisément, on peut les définir de différentes façons. Les mesures du centre ou du milieu[1] d'une série statistique et de la variabilité des valeurs autour de ce milieu ne se sont d'ailleurs pas imposées d'une manière évidente. Le sens de ces paramètres n'est pas intuitif et leur interprétation quelquefois délicate. Si moyenne et écart type sont employés de façon intensive, d'autres mesures sont possibles et sont utilisées car elles présentent différents avantages.

2) Conceptions à propos de la moyenne

Tout le monde pense savoir ce qu'est la moyenne (arithmétique) d'une série statistique. Peu d'élèves savent cependant donner un sens concret au résultat trouvé et expliquer le pourquoi du calcul effectué.

A la question : "Un étudiant a réussi un examen avec 13,8 de moyenne. Que représente, pour vous, cette valeur ?", 87 étudiants de deuxième année d'IUT ayant reçu un enseignement de statistique en première année (en plus de celui de lycée et collège) ont répondu de la façon suivante :

---

[1] Pour reprendre une expression utilisée pendant longtemps. Voir par exemple l'article "MILIEU à prendre entre les observations" de l'*Encyclopédie Méthodique* (Tome 2, Mathématiques, page 404), Paris, 1785.

Paraphrases sur le terme de valeur moyenne
(moyenne pondérée, arithmétique et même algébrique !) : 23
Formule $\frac{1}{n}\sum_{i=1}^{n} x_i$ (ou total divisé par le nombre de coefficients) : 39
Sans réponse, phrases incorrectes ou incompréhensibles : 22
Valeur qui pourrait remplacer toutes les autres (ou représentative de l'ensemble) : 3
(la note que l'étudiant aurait eu s'il avait eu la même note dans toutes les matières,
la note la plus "proche" de toutes les notes obtenues...)
Quelques perles : Il y a autant de notes supérieures à 13,8 que de notes inférieures à 13,8
Il y a des notes supérieures à 10 et d'autres inférieures à 10
Toutes les notes sont proches de 13,8

On constate qu'il est bien difficile pour les étudiants de sortir de la formule qui donne la moyenne et de donner un sens au résultat, même après de nombreuses années de fréquentation de cette valeur caractéristique d'une série statistique.

Trois étudiants seulement citent le fait que l'ensemble des notes obtenues donnent le même résultat (au total, à l'examen) que si chacune avait été de 13,8.

Rares sont ceux qui évoquent (pas toujours avec bonheur) la dispersion des notes autour de la valeur 13,8.

3) Définitions mathématiques

a) Le problème consiste à remplacer la série $x_1, x_2, \ldots\ldots, x_n$ par une série constante a,a,.....,a qui soit la plus "proche" possible de la série de départ.

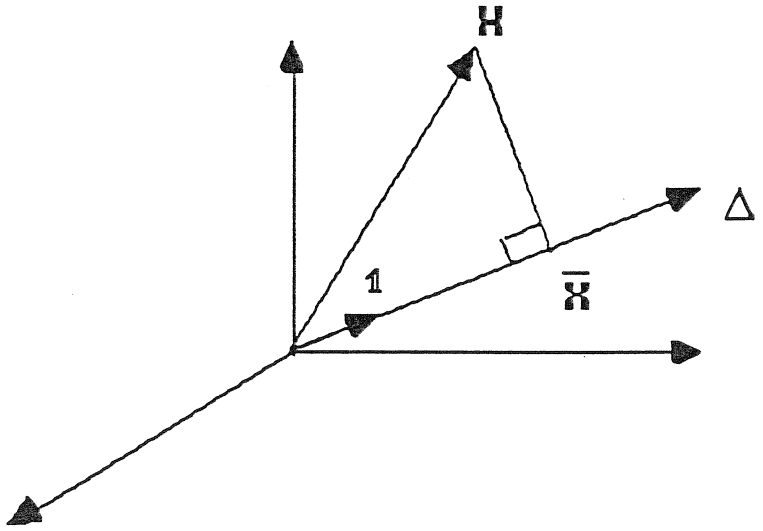
(i) Si on choisit la métrique euclidienne pour mesurer la distance entre les deux séries c'est-à-dire si on calcule la somme des carrés des écarts $\sum_{i=1}^{n}\left(x_i - a\right)^2$, on peut démontrer (voir §4) que la meilleure valeur de a, c'est-à-dire celle qui minimise cette somme, est la moyenne arithmétique $\bar{x}$ donnée par la formule $\frac{1}{n}\sum_{i=1}^{n} x_i$.

(ii) Si on mesure la distance entre les deux séries en utilisant la norme L_1 c'est-à-dire en calculant la somme des valeurs absolues des écarts $\sum_{i=1}^{n}\left|x_i - a\right|$, on peut démontrer que la

meilleure valeur de a, c'est-à-dire celle qui minimise cette somme, est la médiane de la série[2].

b) Interprétation géométrique de la moyenne

(i) On représente la série des valeurs $x_1, x_2, \ldots\ldots, x_n$ par le vecteur X de $\mathbb{R}^n$ de coordonnées $(x_1, x_2, \ldots\ldots, x_n)$ et on projette orthogonalement ce vecteur sur la n-sectrice des axes c'est-à-dire sur la droite Δ de vecteur directeur $\mathbb{1}(1, 1, 1, \ldots\ldots, 1)$. On obtient le vecteur $\overline{X}$. Pour le produit scalaire défini par $\langle x,y\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i.y_i$, on a : $\langle X,\mathbb{1}\rangle = \frac{1}{n}\sum_{i=1}^{n} x_i.1 = \overline{x}$. Or $\overline{X}$ est le projeté de X sur la droite de vecteur directeur $\mathbb{1}$, donc $\overline{X} = \langle X,\mathbb{1}\rangle.\mathbb{1}$ soit $\overline{X} = \overline{x}.\mathbb{1}$. Les coordonnées de $\overline{X}$ sont donc toutes égales à $\overline{x}$. Le vecteur $\overline{X}$ est alors le vecteur constant le plus près de X pour la métrique associée au produit scalaire choisi, ce qui signifie que $\|X - \overline{X}\| = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$ est minimum.



[2] Rappel : une définition simple de la médiane

The median is defined to be Median = { the single middle value the mean of the two middle values John W. Tukey - Exploratory Data Analysis - 1977

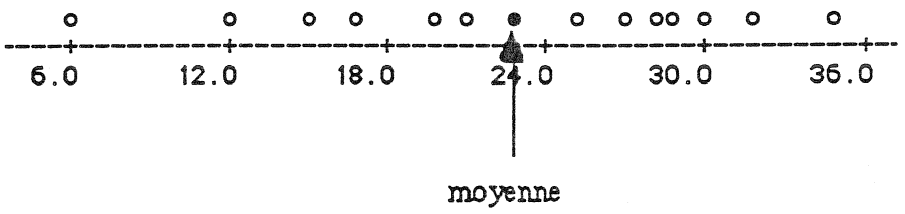
C'est-à-dire : la médiane est définie comme la valeur centrale ou la moyenne des deux valeurs centrales.

(ii) La distance entre les deux séries est mesurée dans le cas de la norme euclidienne par la quantité $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$ qui correspond à l'écart type de la série de départ (ou $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ qui correspond à la variance) et dans le cas de la norme L_1 par la quantité $EAM = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Me\right|$, qui correspond à l'écart absolu moyen de la série de départ.

c) Interprétation barycentrique de la moyenne

(i) Propriété : la somme des écarts par rapport à la moyenne est toujours nulle car $\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}x_i - n\bar{x} = 0$. Elle ne fournit donc pas un moyen de mesurer la dispersion. Par contre, c'est peut-être cette idée assez intuitive de compensation des écarts au-dessus et au-dessous de la moyenne qui conduit parfois à une confusion avec la médiane (pour laquelle le *nombre* de valeurs au-dessus et au-dessous est le même).

(ii) La propriété énoncée au paragraphe précédent permet de donner une illustration assez parlante de la moyenne. On peut en effet l'imaginer comme le point d'équilibre (le centre de gravité) d'une barre graduée sur laquelle on représenterait chaque valeur de la série par une masse proportionnelle à son effectif.



4) Retour historique

En dehors de sa simplicité de calcul (à la main ou programmé), la principale caractéristique de la moyenne est qu'elle donne la même importance à chacune des valeurs de la série. Outre le fait que cela peut être un inconvénient si certaines valeurs sont douteuses ou très éloignées des autres valeurs de la série (voir plus loin la définition d'autres paramètres

de tendance centrale qui prennent en compte cet état de fait), il est à remarquer que la moyenne ne s'est pas imposée d'emblée aux scientifiques confrontés à une série de valeurs observées.

En 1713, Jacques Bernoulli[3] propose la méthode suivante pour calculer la hauteur moyenne dans un baromètre :

Rassemble toutes les hauteurs que tu as observées, qu'elles soient différentes ou identiques, en une somme que tu divises par le nombre d'observations, ou, ce qui est plus avantageux, si les mêmes hauteurs ont été observées plusieurs fois, les différentes hauteurs sont multipliées par le nombre d'observations qui ont été faites de chacune d'entre elles, la somme de tous ces produits divisée par le nombre de ces observations donne la hauteur "moyenne" (en latin mediam) que les Allemands appellent "eine in die andere gerechnet", les français "l'une portant l'autre"; si le mercure a été surpris deux fois à la hauteur de 28 pouces, trois fois à la hauteur de 29 et quatre fois à la hauteur de 30, la hauteur moyenne sera : $\frac{2.28+3.29+4.30}{9} = 29\ 2/9$

Aussi est-il évident qu'ils se trompent ceux qui, pour rechercher la quantité moyenne de mercure, font la moyenne arithmétique entre les extrêmes, et qui avec ce procédé obtiennent une hauteur moyenne de 29 pouces et non de 29 2/9 pouces.

On reconnaît dans la première méthode le calcul de la moyenne dite pondérée que Bernoulli dit préférable à ce que l'on appelle de nos jours l'**étendue moyenne** (**midrange**) soit : (minimum+maximum)/2. Il est à remarquer que cette dernière méthode est encore employée en météorologie pour déterminer la température d'une journée donnée, alors que la moyenne d'un jour quelconque, par exemple le 14 Juillet, sera la moyenne arithmétique des 14 Juillet des 30 dernières années.

Ce n'est qu'en 1755 que T. Simpson propose l'usage généralisé de la moyenne dans les mesures astronomiques[4] dans le cas de n mesures d'une même grandeur.

En 1805, Legendre[5] établit la liaison entre cette pratique et la méthode des moindres carrés.

" La règle par laquelle on prend le milieu entre les résultats de diverses observations (pour un seul élément), n'est que la conséquence très simple de notre méthode générale, que

---

[3] *Ars Conjectandi* , traduction N. Meusnier, IREM de Rouen, 1987

[4] A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the Advantage of taking the mean of a number of observations in practical astronomy. *Phil. Trans.*, 49.

[5] *Nouvelles méthodes pour la détermination des orbites des comètes*, Courcier éditeur, Paris, 1805

nous appellerons méthode des moindres carrés. En effet, si l'expérience a donné diverses valeurs a', a'', a''' etc. pour une certaine quantité x, la somme des carrés des erreurs sera $(a'-x)^2 + (a''-x)^2 + (a'''-x)^2 +$ *etc. et en égalant cette somme à un minimum, on a :* $0 = (a'-x) + (a''-x) + (a'''-x) +$ *etc d'où résulte* $x = \frac{a+a'+a''+a'''+....etc}{n}$, *n étant le nombre des observations".*

5) Problème de sens

On peut toujours calculer mathématiquement un paramètre, pour peu que les fonctions qui interviennent dans le calcul soient définies. Le sens que l'on peut donner au résultat dépend de la situation. Si l'âge moyen des élèves d'une classe est de 15 ans et 3 mois, cela n'a pas le même sens concret[6] que la note moyenne de 13,8 du premier paragraphe car le total des âges ne représente rien. Mathématiquement, cela représente la valeur qui minimise la somme des carrés des écarts comme définie ci-dessus (§3).

De même, la moyenne arithmétique des valeurs 0,05 ; 0,06 ; 0,08 ; 0,07 ; 0,09 n'a pas de sens concret si ces valeurs représentent les taux d'intérêts auxquels on a placé une somme d'argent pendant 5 années consécutives. On montre que dans ce cas la valeur de i telle que $1+i = \sqrt[5]{(1+0,05)(1+0,06)(1+0,08)(1+0,07)(1+0,09)}$, obtenue en effectuant la moyenne géométrique des $1+x_i$ est préférable puisqu'elle représente le "taux moyen" c'est-à-dire celui qui aurait conduit au même gain si le taux avait été constant pendant les cinq années.

Deux paramètres peuvent être pertinents pour une même série statistique. Ainsi, Christiaan Huygens et son frère Louis, utilisant la première table de mortalité établie par l'anglais John Graunt (1662), essayent de déterminer le nombre d'années qui reste à vivre à un individu d'âge donné.

---

6 On est dans ce cas pour toutes les variables repérables et non mesurables.

De 100 personnes conçues il en meurt :
36 au bout de 6 ans
24 entre 6 et 16 ans
15 entre 16 et 26 ans
9 entre 26 et 36 ans
6 entre 36 et 46 ans
4 entre 46 et 56 ans
3 entre 56 et 66 ans
2 entre 66 et 76 ans
1 au delà de 76 ans

Louis propose de faire la moyenne pour connaître l'espérance de vie à la naissance, il trouve 18 ans et 2 mois 1/2. Son frère Christiaan[7], construit un contre-exemple (de 100 personnes, 90 meurent avant 6 ans, les autres vivent jusqu'à 155 ans et 2 mois !) qui donne la même espérance de vie de 18 ans et 2 mois 1/2 ce qui montre que ce calcul n'est pas pertinent car "*qui gagerait qu'un enfant conçu parviendrait alors à l'âge de 6 ans seulement aurait grand désavantage puisque de 10 il n'y en a qu'un qui y parvient*".

Il fait ensuite la distinction entre "moyenne" et "valeur qui a autant de chance d'être dépassée ou non" (la médiane) :

Le calcul que je vous ai envoyé vous aura embarrassé sans doute, auquel ayant songé depuis, et aussi au vôtre, je trouve que nous avons tous deux raison en prenant la chose de différents sens. Vous donnez à un enfant conçu 18 ans et 2 mois 1/2 de vie, et il est vrai que son espérance vaut autant que cela. Cependant il n'est pas apparent qu'il vivra si longtemps, car il est beaucoup plus apparent qu'il mourra devant ce terme. De sorte que si l'on voulait gager qu'il y parviendrait, la partie serait désavantageuse , car on peut seulement gager avec égal avantage qu'il vivra jusqu'à 11 ans environ, ainsi que je le trouve par ma manière...

Ce sont donc deux choses différentes que l'espérance ou la valeur de l'âge futur d'une personne, et l'âge auquel il y a égale apparence qu'il parviendra ou ne parviendra pas. Le premier est pour régler les rentes à vie, et l'autre pour les gageures.

---

[7] Oeuvres complètes de Christiaan Huygens, citées dans *Actes du séminaire d'Histoire des Sciences du lycée Malherbe de Caen*- Juin 1995.

De même, dans une distribution de revenus, les très gros salaires "faussent le calcul de la moyenne". Plus exactement la moyenne n'est pas le paramètre pertinent pour mesurer le centre de la distribution. En effet, dans certaines distributions de ce type, presque toutes les valeurs se trouvent au-dessous de la moyenne, alors que par définition on a toujours 50 % des valeurs en dessous de la médiane et 50 % en dessus. C'est donc ce dernier paramètre qui est le mieux adapté pour donner une idée du centre d'une distribution de revenus. L'INSEE le complète généralement par le calcul des autres déciles[8].

6) Comparaison de différents paramètres classiques

On constate intuitivement ou graphiquement que certaines distributions statistiques sont largement étirées et d'autres plus resserrées. Il est logique de penser à comparer deux distributions par des mesures de cette dispersion. Malheureusement ce concept est "flou"[9] en ce sens qu'il y a de nombreuses façons de mesurer la dispersion.

De même, pour une distribution symétrique, la moyenne (ou la médiane puisque ces deux valeurs sont alors confondues) donne une bonne idée de la position centrale de cette distribution, dans le sens où elle correspond géométriquement à la valeur par rapport à laquelle le graphique est symétrique, mais pour une série très dissymétrique, aucune mesure ne s'impose, en particulier la moyenne n'est plus du tout appropriée. Là encore, on est en face du concept flou de position et en particulier de position centrale qui peut être mesuré de différentes façons plus indiquées l'une que l'autre suivant la dispersion de la série.

En effet, une mesure de tendance centrale représente d'autant mieux l'ensemble des valeurs de la série que celles-ci sont moins dispersées. Ceci illustre le fait que ces deux concepts, bien que mal définis, sont liés entre eux.

Prenons par exemple les deux séries statistiques suivantes données par leur graphique en tiges et feuilles (stem and leaf[10]) et qui pourraient représenter les notes de deux classes.

---

[8] Les n valeurs de la série étant classées dans l'ordre croissant de $x_{(1)}$ à $x_{(n)}$, les déciles sont les valeurs des observations dont les rangs sont $k(n+1)/10$, pour $k=1$ à 9. Si ces rangs ne sont pas entiers, on effectue une interpolation.

[9] Pour reprendre le terme de F. Mosteller et J. W. Tukey dans *Data Analysis and Regression. A second course in statistics* - Addison Wesley - Reading - Massachusetts - 1977.

[10] John W. Tukey-*Exploratory Data* - Addison Wesley - Reading - Massachusetts - 1977.

```
 9 0000
10 000
11 00
12 0
13 00
14 000
15 0000
```

```
 9 0
10 00
11 000
12 0000
13 000
14 00
15 0
```

Explications[11] : pour la première série, on a 4 fois la note "9", 3 fois la note 10, etc.

Le calcul de différents paramètres classiques[12] pour chacune des séries donne les résultats suivants :

Nb	Moyenne	Médiane	Ecart type	Min	Max	Q1	Q3	Q3-Q1
19	12	12	2,29	9	15	10	14	4
16	12	12	1,58	9	15	11	13	2

Elles ont la même moyenne et la même médiane et pourtant personne ne dira que le profil des deux classes est semblable. L'une a des résultats plus homogènes que l'autre. Les écarts types qui mesurent la dispersion des séries le montrent (2,29>1,58) mais sans que l'on sache réellement apprécier la différence 2,29-1,58 (augmentation absolue) ou le rapport 2,29/1,58 (augmentation relative).

Un autre moyen de comparer les dispersions est l'**intervalle** ou **écart interquartile** c'est-à-dire Q_3-Q_1. Le sens en est plus facile à appréhender que celui de l'écart type. C'est la longueur de l'intervalle qui contient la moitié centrale de la distribution. Il vaut 4 pour la première série et 2 pour la seconde. En quelque sorte la première série est deux fois plus dispersée que la deuxième pour cette mesure.

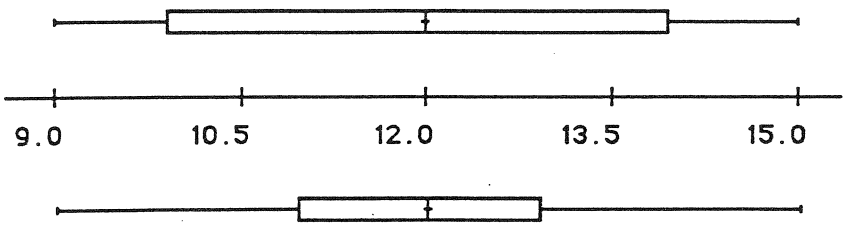
De plus, on peut représenter graphiquement cette dispersion dans un graphique en boîte, ou boîte à moustaches (box-plot[13]).

---

[11] Pour une présentation des graphiques en tiges et feuilles, voir par exemple, J. C. Girard, "Des diagrammes à l'histogramme", *Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?*, Groupe Probabilités et Statistique, IREM de Lyon, 1998.

[12] Q_1 est le premier quartile et Q_3 le troisième quartile. Les n valeurs de la série étant classées dans l'ordre croissant de $x_{(1)}$ à $x_{(n)}$, Q_1 représente la valeur de l'observation de rang $(n+1)/4$ et Q_3 l'observation de rang $3(n+1)/4$. Si ces rangs ne sont pas entiers, on effectue une interpolation.

[13] John W. Tukey-*Exploratory Data Analysis*-- Addison Wesley - Reading - Massachusetts - 1977.



Explications : Les extrémités de la boîte sont Q1 et Q3. Les moustaches s'étendent jusqu'au minimum et au maximum. Le trait central correspond à la médiane[14].

L'inconvénient de la moyenne et, partant, des paramètres de dispersion calculés à partir de celle-ci est d'être sensible à une valeur anormalement grande ou petite. Par exemple, voyons quelles sont les conséquences d'une note supplémentaire égale à 20.

Le calcul des mêmes paramètres pour chacune des deux séries modifiées donne les résultats suivants :

Nb	Moyenne	Médiane	Ecart type	Min	Max	Q1	Q3	Q3-Q1
20	12,4	12,5	2,838	9	20	10	14,75	4,75
17	12,471	12	2,428	9	20	11	13,5	2,5

La moyenne est généralement affectée par une valeur aussi éloignée du centre de la distribution. En effet la moyenne augmente de 0,4 dans la première série et de 0,471 dans la deuxième. Un des intérêts de la médiane est d'être moins sensible, en général, à ces valeurs. En effet, dans la deuxième série la médiane reste inchangée (12). Par contre dans la première série, elle est plus affectée que la moyenne puisqu'elle augmente de 0,5. L'explication est que cette série présente une faible densité dans la partie centrale alors que la première série présente au contraire une forte densité puisque 12 est la valeur la plus fréquente (qu'on appelle le **mode** de la série) et que les valeurs voisines (11 et 13) sont aussi très fréquentes. La médiane est également sensible aux discontinuités dans la partie centrale de la distribution.

Pour ce qui concerne les paramètres de dispersion, l'écart type passe de 2,29 à 2,8 (respectivement de 1,58 à 2,4) alors que l'écart interquartile passe de 4 (respectivement 2) à 4,75 (respectivement 2,5). On conserve à peu près le même rapport entre les deux.

---

[14] Pour la construction des graphiques en boîte et un exemple détaillé d'utilisation, voir J. C. Girard, "La médiane, pour quoi faire", *Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?*, Groupe Probabilités et Statistique, IREM de Lyon, 1998.

7) D'autres paramètres de tendance centrale

Différents statisticiens ont proposé des moyens autres que la moyenne et la médiane pour mesurer le "milieu" d'une distribution statistique en particulier dans le cas d'estimation de la tendance centrale à partir d'un échantillon. Ils ont en commun de ne pas prendre en compte de la même façon toutes les observations, soit en éliminant les valeurs aberrantes, soit en donnant moins de poids aux valeurs éloignées de la partie centrale de la distribution. On qualifie ces mesures de "robustes".

Un solution pour éviter la trop grande influence des valeurs éloignées est de faire la **moyenne tronquée** (trimmed mean). Pour cela on enlève, aux deux extrémités de la série ordonnée, un nombre d'observations (arrondi à l'entier le plus proche) correspondant au même pourcentage (par exemple 5%), puis on fait la moyenne des valeurs restantes. On trouve 12 dans les deux séries de départ, 12,167 et 12,2 dans les séries modifiées. Elle est donc moins affectée que la moyenne et la médiane dans le deuxième cas.

Si on effectue une moyenne tronquée à 25%, on obtient alors la moyenne des valeurs de la moitié centrale de la distribution. On l'appelle **midmean**[15]. Elle vaut 12 dans les deux séries de départ, 12,2 et 12,11 dans les deux séries modifiées. C'est un bon compromis entre la moyenne et la médiane.

Le premier à préconiser la suppression d'une partie des observations semble avoir été l'astronome Boscovich[16] qui calculait la moyenne d'une série d'observations après avoir enlevé la plus petite et la plus grande des valeurs.

Cette méthode sera reprise par les fermiers généraux de l'Ancien Régime pour calculer l'impôt dû par un paysan à partir de la récolte moyenne des 5 dernières années après avoir enlevé la meilleure et la moins bonne.

---

[15] Voir par exemple : *Understanding Robust and Exploratory Data Analysis*- Hoaglin, Mosteller & Tukey - J. Wiley - New York - 1983

[16] Voir *Histoire de la Statistique*- J. J. Droesbeke et P. Tassi - P. U. F. - Paris - 1990

La moyenne bipondérée

Faire une moyenne tronquée revient à affecter certaine valeur d'un poids 0 et d'autres d'un poids 1. Une alternative est d'affecter à chacune des observations un poids qui diminue en fonction de son éloignement au centre de la distribution. F. Mosteller et J. W. Tukey[17] proposent de calculer la médiane Me de la série puis de remplacer chaque observation X par $Z = \frac{X - Me}{3I}$, où I est l'écart interquartile, puis de faire la moyenne des valeurs de la série, pondérées par : $w = 0$ si $|Z| > 1$

$w = (1-Z^2)^2$ si $|Z| < 1$.

Cela revient à considérer aberrantes les observations à plus de 3I de la médiane.

On peut procéder ensuite par itération c'est-à-dire recommencer le calcul en remplaçant la médiane Me par la moyenne bipondérée trouvée jusqu'à ce qu'il n'y ait plus de variation significative (voir un exemple à la fin).

Trimean[18] : Si n est le nombre d'observations, m=(n+1)/2 est la position de la médiane. On définit les quarts supérieur et inférieur F_U et F_L (upper and lower fourths) ou H_U et H_L (upper and lower hinges) comme les médianes de la première moitié et de la deuxième moitié de la série (médiane comprise) c'est-à-dire les observations de rang ([m]+1)/2 en partant du maximum et du minimum[19].

L'intervalle $[F_L;F_U]$ contient environ la moitié centrale des observations. Compte tenu des définitions il peut être différent de l'intervalle interquartile $[Q_1;Q_3]$

La moyenne "Trimean" est alors calculée par $1/4(F_L + 2Me + F_U)$.

Formule de Gastwirth[20] : Si $Q_{1/3}$ et $Q_{2/3}$ sont les valeurs qui partagent la série en trois[21], on calcule $0{,}3\ Q_{1/3} + 0{,}4\ Me + 0{,}3\ Q_{2/3}$.

---

17 *Data Analysis and Regresion : A Second Course in Statistics* - Addison Wesley - Reading - Massachusetts - 1977.

18 Tukey J. W. - *Exploratoty Data Analysis* - Addison Wesley - Reading - Massachusetts - 1977.

19 [m] représente la partie entière de m.

20 Voir par exemple : *Robust Statistics : A Review* - P. Huber - The Annals of Mathematical Statistics - 1972 - Vol. 43

21 $Q_{1/3}$ et $Q_{2/3}$ sont deux cas particuliers de quantiles. Ils partagent la série dans les proportions indiquées. Ils correspondent aux observations (classées dans l'ordre croissant) de rangs (n+1)/3 et 2(n+1)/3. On les calcule par interpolation.

Broadened Median[22] : Pour conserver la stabilité de la médiane par rapport aux valeurs aberrantes et diminuer sa sensibilité aux discontinuités dans le milieu de la série, J. W. Tukey a proposé le calcul suivant : si n est impair, la médiane élargie est la moyenne des trois valeurs centrales si $5 \leq n \leq 12$, des 5 valeurs centrales si $n \geq 13$. Si n est pair, la médiane élargie est une moyenne pondérée des 4 valeurs centrales si $5 \leq n \leq 12$ avec comme poids 1/6,1/3, 1/3 et 1/6, et si $n \geq 13$, la moyenne pondérée des 6 valeurs centrales avec comme poids 1/5 pour les quatre valeurs centrales et 1/10 pour les deux autres.

F-Mean : C'est la moyenne des valeurs de la moitié centrale des observations, c'est-à-dire de l'intervalle $[F_L;F_U]$, F_L et F_U compris. Compte tenu des définitions il peut être différent de la "midmean".

8) D'autres paramètres de dispersion

On peut définir différentes mesures de la variabilité d'une distribution statistique. Un paramètre, assez frustre, est l'**étendue** de la série c'est-à-dire l'écart entre le minimum et le maximum. Pour l'exemple du paragraphe 6, il est le même dans les deux séries de départ (15-9=6) mais il évident qu'il est très affecté par des valeurs éloignées : il passe ainsi à 11 dans les séries modifiées.

Outre les deux paramètres déjà cités dans le paragraphe 3, à savoir, la **variance** $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, c'est-à-dire la moyenne des carrés des écarts entre les observations et leur moyenne, doublement sensible aux valeurs aberrantes, et l'**écart absolu moyen** par rapport à la médiane, $EAM = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - Me\right|$, qui donne l'écart absolu minimum, on définit encore l'**écart absolu moyen** par rapport à la moyenne $EAM = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$.

On peut enfin citer la médiane des valeurs absolues des écarts entre les observations et leur médiane (**Median Absolute Deviation**) : $MAD = \text{Médiane}\left(\left|x_i - Me\right|\right)$ qui bénéficie doublement des avantages de la médiane, ou encore l'étendue entre les quarts, **F-spread** ou $d_F = F_U - F_L$, pratiquement égale à l'**écart interquartile**.

---

[22] *Understanding Robust and Exploratory Data Analysis*- Hoaglin, Mosteller & Tukey - J. Wiley - New York - 1983

9) Conclusion

Comme on l'a vu, la valeur d'un paramètre de tendance centrale n'a pas beaucoup de signification sans une mesure de la dispersion, et réciproquement.

Que ce soit pour caractériser une série entièrement connue ou pour donner des informations à partir d'un échantillon, une alternative au couple $(\overline{x},\sigma)$, qui cumule les défauts déjà cités des deux paramètres, pourrait être le couple (Me, MAD) ou encore le couple (Me, d_F) ou (M_e, I) qui conduisent à des estimations moins sensibles aux valeurs aberrantes et à des intervalles de confiance plus robustes (c'est-à-dire qui ne font pas appel à l'hypothèse d'une distribution supposée connue, par exemple la distribution normale).

D'autre part, on ne peut raisonnablement chercher à comparer deux séries statistiques par leurs moyennes que si elles ont à peu près la même forme de distribution et la même dispersion.

De plus, il ne faut pas oublier que l'écart type a une unité, qui est la même que celle des valeurs de la série et de leur moyenne. Si on multiplie ces valeurs par 10, par un changement d'unité par exemple, l'écart type sera aussi multiplié par 10 sans que la distribution soit évidemment changée.

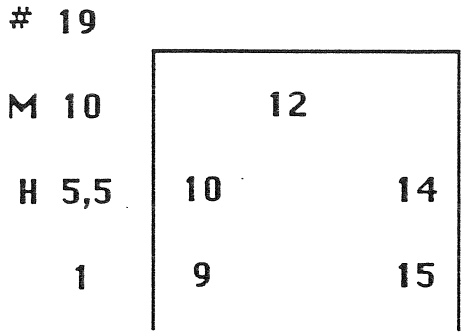
Il n'est donc légitime de comparer les écarts types de deux séries que si les moyennes sont du même ordre de grandeur.

Dans le cas contraire, un autre paramètre, le **coefficient de variation** défini par $CV = \frac{\overline{x}}{\sigma}$ et qui est un pourcentage, donc sans unité, permet d'éviter ces phénomènes d'échelle.

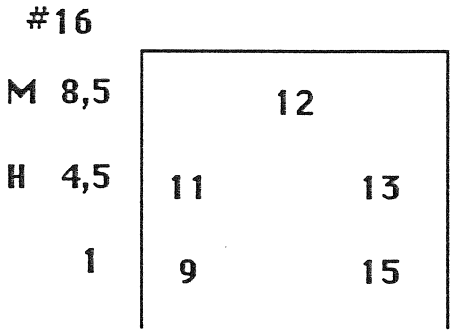
Enfin, il est peut-être illusoire de vouloir décrire une distribution statistique par deux nombres, quels qu'ils soient. Tukey, encore lui, propose de résumer une série par 5 nombres : le minimum, le maximum, la médiane, les deux quartiles (ou les deux quarts) et de présenter les résultats dans un tableau[23] comme illustré dans l'exemple ci-dessous (qui reprend les séries de départ du §6) :

---

[23] Letter-Values Display - J. W. Tukey - *Exploratory Data Analysis* - Addison Wesley - Reading, Massachusetts - 1977



Pour la première série, le nombre d'observations est 19, la médiane (notée M ; c'est la $10^{ème}$ valeur) est 12, les quarts (égaux ici aux quartiles et notés H ; ce sont les valeurs de rang 5,5 à partir des deux extrémités) sont 10 et 14 donc $d_F=I=4$, le minimum est 9 et le maximum 15 donc l'étendue est égale à 6.



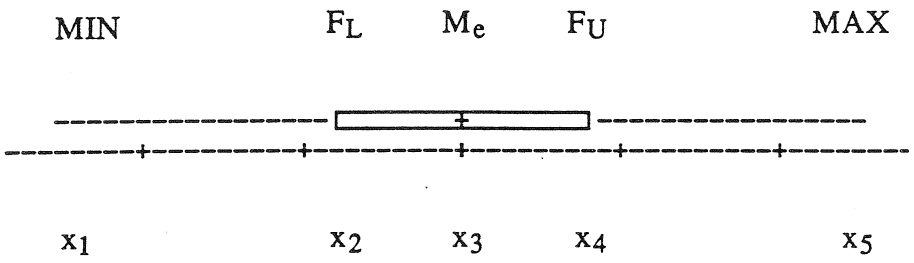
Pour la deuxième série, il y a 16 observations, les quarts sont 11 et 13 et l'écart interquartile est égal à 2, le reste est inchangé.

Ces schémas et ces paramètres robustes trouvent une application intéressante dans le contrôle de qualité où l'on prélève couramment 5 objets pour estimer, par exemple, la moyenne des poids ou des diamètres des articles d'une production[24].

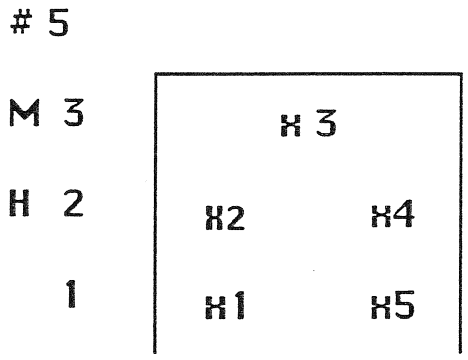
Si les 5 observations sont rangées dans l'ordre croissant et notées : x_1, x_2, x_3, x_4 et x_5, le graphique en boîte devient :

---

[24] Voir, par exemple, White E.M. et Schroeder R. "A simultaneous control chart" *Journal of Quality Technology* Vol. 19, N° 1, Janvier 1987 ou Iglewicz B. et Hoaglin D.C. " Use of Boxplots for Process Evaluation" *Journal of Quality Technology* Vol. 19, N° 4, Octobre 1987.



Le premier schéma devient alors :



L'écart interquartile est $x_4 - x_2$ et la F-mean qui est égale à la médiane élargie et à la moyenne tronquée à 20 % : $(x_2 + x_3 + x_4)/3$.

Un exemple : les départements français

1) Objectifs

On se propose de faire l'analyse d'une distribution statistique du point de vue des paramètres de tendance centrale. L'exemple choisi porte sur la population des départements français métropolitains (chiffres du recensement de 1990, source : Quid[25]).

1	73000	289000	485000	795000	
2	113000	293000	494000	799000	
3	118000	296000	509000	815000	
4	131000	299000	514000	839000	
5	131000	306000	514000	926000	
6	132000	306000	527000	953000	
7	134000	311000	529000	972000	
8	136000	322000	537000	1011000	
9	156000	323000	538000	1016000	
10	159000	342000	548000	1050000	
11	175000	343000	558000	1052000	
12	196000	346000	559000	1078000	
13	200000	348000	568000	1085000	
14	204000	354000	578000	1213000	Gironde
15	207000	358000	580000	1216000	Val-de-Marne
16	225000	364000	585000	1223000	Seine-Maritime
17	230000	380000	598000	1307000	Yvelines
18	233000	386000	618000	1381000	Seine-St-Denis
19	238000	386000	620000	1391000	Hauts-de-Seine
20	238000	396000	671000	1433000	Pas-de-Calais
21	249000	414000	706000	1509000	Rhône
22	270000	467000	712000	1759000	Bouches-du-Rhône
23	278000	471000	726000	2152000	Paris
24	278000	480000	746000	2532000	Nord

Existe-t-il un département moyen ? Cela a-t-il un sens ? Si oui, quel paramètre choisir pour le mesurer ?

---

[25] La Corse du nord et la Corse du Sud sont comptées séparément, le Finistère Nord et le Finistère Sud ne forment qu'un département. On a donc 96 départements dans ce recensement. Les effectifs sont arrondis au millier le plus proche.

2) Première analyse

Le calcul des paramètres les plus courants donne les résultats suivants[26] :

```
MTB > Describe C2.

                N     MEAN   MEDIAN   TRMEAN    STDEV
C2             96   589698   482500   542558   452138

               MIN      MAX       Q1       Q3
C2           73000  2532000   280750   782750
```

La moyenne est assez éloignée de la médiane (22 % de plus). L'écart type est très important par rapport à la moyenne (le coefficient de variation est égal à 452 138 / 589 698 soit 77%). On est donc en présence d'une série très dispersée et non symétrique (distribution étalée à droite). Le rapport entre l'écart interquartile, qui est égal à 502 000, et l'écart type est de 1,11 ce qui est très peu puisque pour la loi normale centrée réduite ce rapport est de 1,35. Les valeurs de la série sont donc assez concentrées entre les deux quartiles Q_1 et Q_3 et assez dispersées au delà du troisième quartile.

3) Analyse graphique

Les graphiques, que ce soit l'histogramme ou le graphique en tige et feuilles, confirment ces premières conclusions.

---

[26] Tous les graphiques et calculs ont été faits avec le logiciel MINITAB : Mean représente la moyenne, Median la médiane, Trmean la moyenne tronquée à 5 % et Stdev l'écart type.

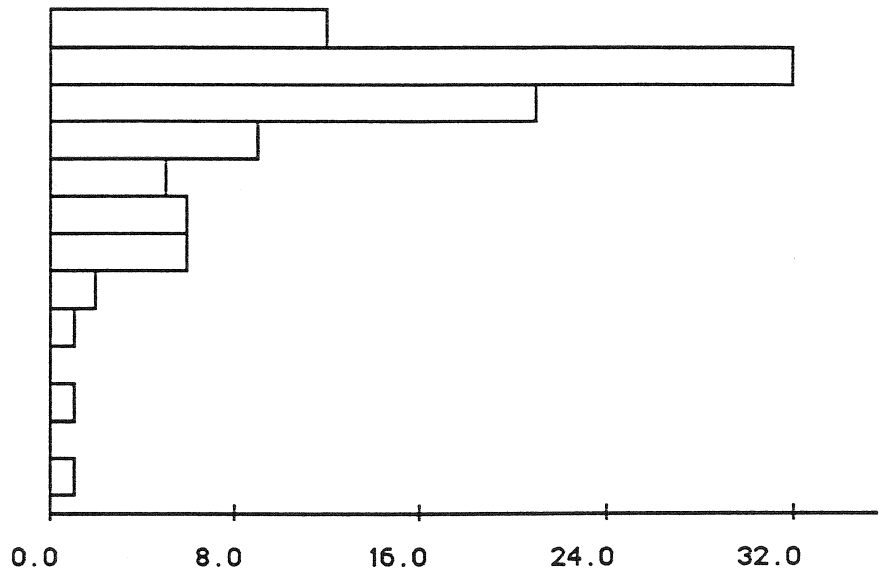
```
MTB > Stem-and-Leaf C2.

Stem-and-leaf of C2        N   = 96
Leaf Unit = 100000

   12     0 011111111111
   44     0 22222222222222223333333333333333
  (21)    0 444444555555555555555
   31     0 666777777
   22     0 88999
   17     1 000000
   11     1 222333
    5     1 45
    3     1 7
    2     1
    2     2 1
    1     2
    1     2 5
```

C2 N = 96

Midpoint	Count
100000	12
300000	32
500000	21
700000	9
900000	5
1100000	6
1300000	6
1500000	2
1700000	1
1900000	0
2100000	1
2300000	0
2500000	1

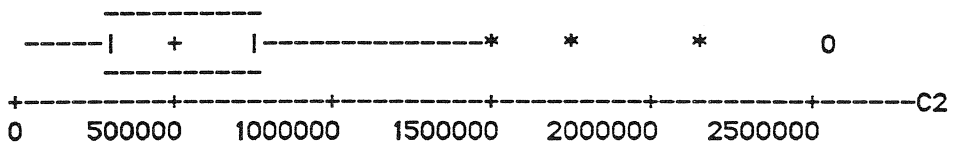


Ces deux graphiques font apparaître une classe modale (celle qui contient la valeur la plus fréquente : le mode) au début de la distribution mais après une forme décroissante on observe une remontée vers la fin de la distribution. Ne serait-on pas alors en présence d'une distribution bimodale qui traduit généralement un mélange de deux populations ? Dans ce cas, il est bien évident que chercher une valeur centrale n'a pas beaucoup de sens. La première idée qui vient à l'esprit est que les départements les plus gros pourraient avoir quelque chose en commun et former une sous population. Cela pourrait être le cas, par exemple, si les départements à forte population étaient ceux de la région parisienne. La liste des populations des 96 départements permet d'écarter, en partie, cette hypothèse. On trouve

trois départements de province parmi les quatre plus peuplés, cinq parmi les onze plus peuplés.

Le graphique en boîte à moustache (box-plot), en plus de la dissymétrie, fait apparaître 3 valeurs éloignées (à plus de 1,5 I de Q3 et notées *) et une très éloignée (à plus de 3 I de Q3 et notée o). Les départements correspondants sont le Nord, Paris, les Bouches-du-Rhône et le Rhône. Il n'y a aucune raison de les écarter de l'analyse et du calcul de la valeur centrale.

```
MTB > BoxPlot C2.
```



4) Calculs de la valeur centrale

La moyenne est très supérieure à la médiane car la distribution est "étalée vers la droite". Les départements à forte population "tirent la moyenne vers le haut". Quelle est dans ce cas là, somme toute très fréquent, la meilleure mesure de la tendance centrale ? Quelle est donc la population "moyenne" c'est-à-dire celle du département le plus représentatif de l'ensemble des départements de France ? On a l'habitude de dire que la médiane est un meilleur paramètre dans ce type de distributions (par exemple dans les distributions de salaires), mais que donnent les autres paramètres présentés précédemment ?

Commençons par les **moyennes tronquées** :

Moyenne tronquée à :	
0 % (moyenne ordinaire)	589 698
5 %	542 558
10 %	518 750
20 %	484 559
25 % (midmean)	470 833
49 % (médiane)	482 500

Plus le pourcentage de valeurs tronquées augmente, plus l'influence des fortes populations diminuent.

Le calcul de la **moyenne bipondérée** en 10 itérations donne les résultats suivants :

Médiane	482 500
(Ecart interquartile	502 000)
1° itération	493 880
2° itération	496 244
3° itération	496 736
4° itération	496 839
5° itération	496 860
6° itération	496 865
7° itération	496 866
Moyenne bipondérée	496 866

On trouve une valeur qui n'est pas très éloignée des moyennes tronquées.

Le graphique en lettres (letter-values display) fait apparaître non seulement les quarts (H) mais aussi les huitièmes (E), les seizièmes (D), les trente-deuxièmes (C), etc[27].

```
MTB > LVals C1.
        DEPTH       LOWER       UPPER         MID        SPREAD
 N=     96
 M      48.5          482.500                 482.500
 H      24.5        283.500     770.500       527.000     487.000
 E      12.5        198.000    1081.500       639.750     883.500
 D       6.5        133.000    1386.000       759.500    1253.000
 C       3.5        124.500    1634.000       879.250    1509.500
 B       2.0        113.000    2152.000      1132.500    2039.000
         1           73.000    2532.000      1302.500    2459.000
```

(Remarque : les populations sont en milliers)

Si la série était symétrique, les centres de tous les intervalles (MID) seraient égaux à la médiane. On retrouve que la série est très dissymétrique (elle est étalée vers la droite).

---

[27] Depth = profondeur, c'est-à-dire le rang en partant du maximum ou du minimum
Lower = valeur inférieure Upper = valeur supérieure
mid = moyenne entre les deux valeurs précédentes
spread = étendue entre les deux valeurs précédentes

Les quarts sont : $F_L = 283\ 500$ $F_U = 770\ 500$ ($d_F = 487\ 000$)

On peut donc calculer la moyenne **trimean** :
$1/4\ (F_L + Me + F_U) = 1/4\ (283\ 500 + 482\ 500 + 770\ 500) = 504\ 750$

et la **F-mean** :
$(283\ 500 + 48*470\ 833 + 770\ 500)/\ 50 = 473\ 080$

La médiane élargie (**broadened median**) prend en compte ici les 6 valeurs centrales avec les pondérations 0,1, 0,2, 0,2, 0,2, 0,1, 0,1 soit :

467000	0.1	46700
471000	0.2	94200
480000	0.2	96000
485000	0.2	97000
494000	0.2	98800
509000	0.1	50900
		*
		483600

La médiane élargie est donc égale à 483 600

Les valeurs $Q_{1/3}$ et $Q_{2/3}$ qui partagent la série en trois sont :

$Q_{1/3} = x_{32}+1/3(x_{33}-x_{32}) = 322\ 000 + 1/3(323\ 000 - 322\ 000) = 322\ 333$
$Q_{2/3} = x_{64}+1/3(x_{65}-x_{64}) = 585\ 000 + 2/3(598\ 000 - 585\ 000) = 593\ 667$

d'où la **moyenne de Gastwirth** :
$0,3*322\ 333 + 0,4*482\ 500 + 0,3*\ 593\ 667 = 467\ 800$

5) Conclusion

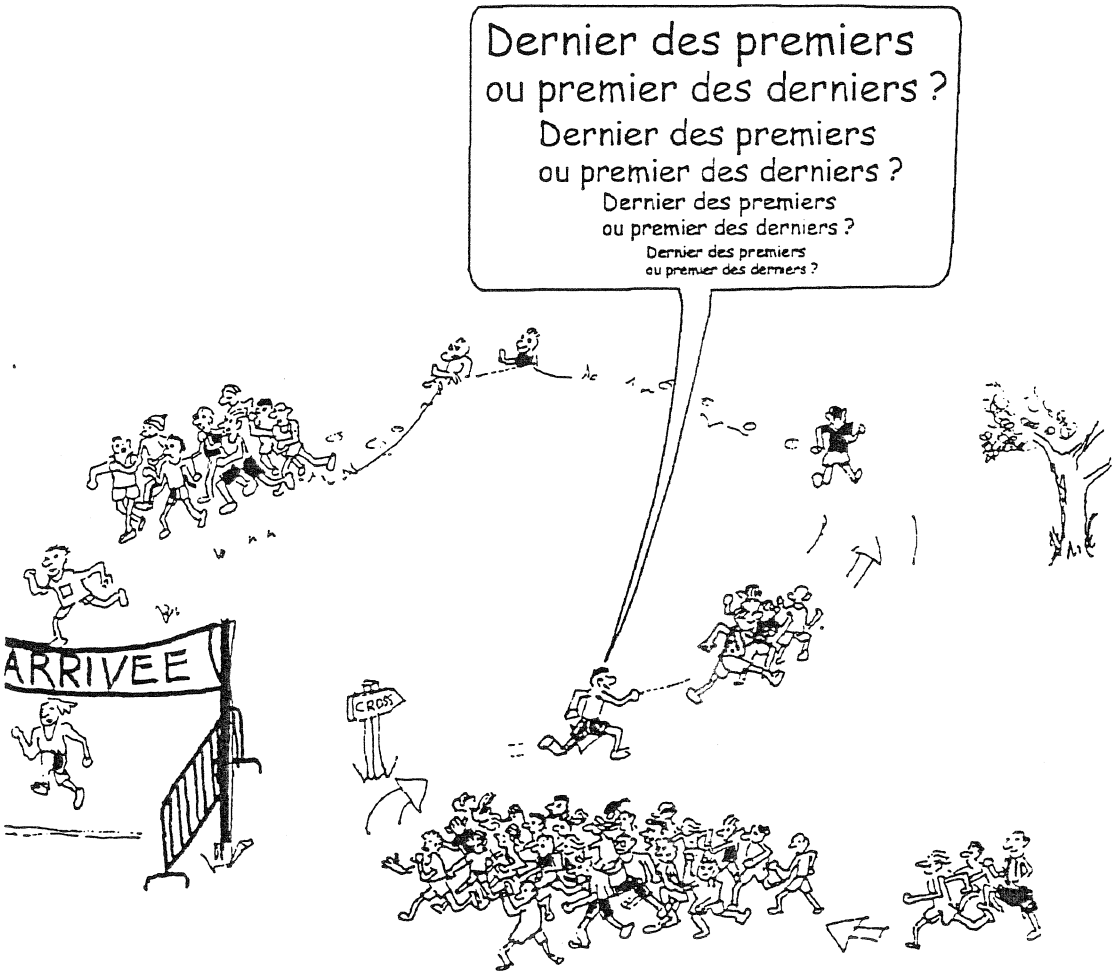
Tous les paramètres calculés sont assez éloignés de la moyenne. Par contre leurs valeurs sont assez voisines les unes des autres. La majorité se trouve entre 470 000 et 500 000.

Si la moyenne des populations des départements français a un sens, la valeur qui mesure ce département moyen (en 1990) est plus certainement autour de 500 000 que de 600 000 comme l'indique la moyenne arithmétique (589 698).

Il faut par conséquent se méfier de la moyenne simple et ne pas l'utiliser sans discernement, en particulier si l'on est en face d'une distribution très dissymétrique ce qui peut se constater sur un graphique ou en comparant la moyenne et la médiane.

Il est donc particulièrement indiqué de ne pas priver les élèves de cette dernière notion, finalement assez simple, et de toujours relier la moyenne ou la médiane à une idée de dispersion, ce qui peut se faire par une mesure (l'écart interquartile, par exemple) et être illustré par un graphique en boîte à moustaches.

D'une manière générale, comme on a essayé de le montrer dans l'exemple des départements, l'analyse d'une série statistique doit s'appuyer sur l'étude de plusieurs graphiques et le calcul de plusieurs paramètres.


Dernier des premiers
ou premier des derniers ?
Dernier des premiers
ou premier des derniers ?
Dernier des premiers
ou premier des derniers ?
Dernier des premiers
ou premier des derniers ?
ARRIVEE
CROSS

Chapitre 6

La médiane, pour quoi faire ?

Un exemple d'utilisation : les boîtes de dispersion[1]

Jean Claude Girard

Résumé :

La médiane est une notion peu utilisée, dont l'utilité est méconnue et la définition pas toujours claire.

On définira un type de graphique, lui aussi méconnu, et pourtant beaucoup utilisé dans les pays anglo-saxons et dans de nombreux logiciels informatiques : la boîte de dispersion (box-plot).

A partir d'un exemple, on montrera la méthode de construction de ce graphique, faisant intervenir la médiane, et son intérêt par rapport aux autres graphiques plus classiques.

1) Rappels sur la médiane et les quartiles

Introduites pour la première fois en 1977 par John Tukey dans son ouvrage "Exploratory Data Analysis", les boîtes de dispersion dont le nom original est "Box and Whiskers Plots" (boîtes à moustaches) sont un moyen de représenter une série statistique par un graphique qui illustre la forme de sa distribution, en particulier sa symétrie, sa valeur centrale ainsi que sa dispersion.

---

[1] Cet article est repris des actes du Colloque de la Commission Inter-IREM - Enseignement de la Statistique et des Probabilités-Toulouse - 14 et 15 Mai 1993

Les paramètres qui interviennent ne sont pas la moyenne et l'écart type mais la médiane pour la valeur centrale et l'écart interquartile pour la dispersion.

La médiane est une valeur qui partage la série en deux séries de même effectif. Une définition qui peut paraître simple mais que certains livres se plaisent à compliquer.

Nous nous baserons sur deux définitions à peu près identiques qui ont le mérite d'être claires :

> n valeurs observées étant rangées par ordre non décroissant et numérotées de 1 à n, la médiane de ces n valeurs observées est la valeur de rang (n + 1) / 2 si n est impair.
>
> Si n est pair, la médiane n'est pas définie. Sauf indication contraire, on peut prendre comme médiane la moyenne arithmétique des valeurs de rang n / 2 et n / 2 + 1.
>
> Norme AFNOR NF X 06-003

The median is defined to be

$$\text{Median} = \begin{cases} \text{the single middle value} \\ \text{the mean of the two middle values} \end{cases}$$

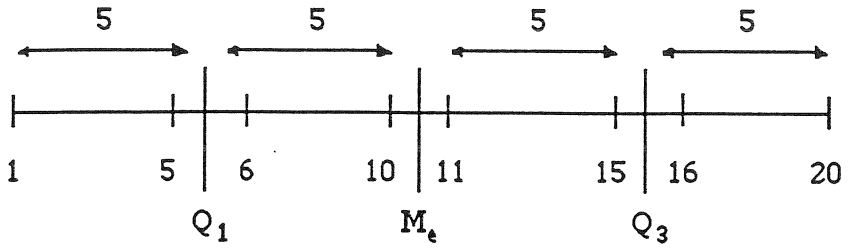
John W. Tukey - Exploratory Data Analysis - 1977

Dans certains cas, il n'est pas pertinent de calculer la médiane mais c'est le cas pour d'autres paramètres, la moyenne par exemple, on a tendance à l'oublier!

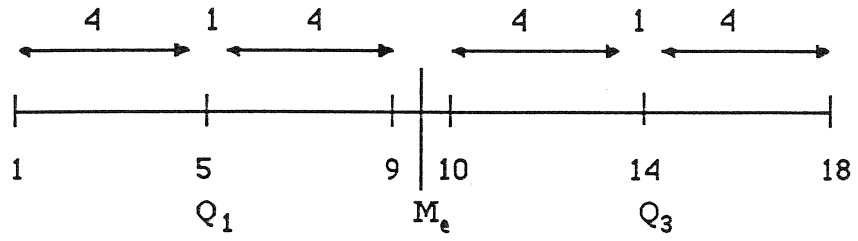
L'écart interquartile est calculé, comme son nom l'indique à partir des quartiles de la distribution c'est-à-dire des valeurs qui partagent la série en quatre séries d'effectifs à peu près égaux. L'écart interquartile est égal à l'écart entre le premier et le troisième quartile. L'intervalle interquartile contient 50% des valeurs de la série statistique. Il donne donc une idée de la dispersion.

Exemple : Pour un effectif de 20, multiple de 4, la médiane est la valeur de l'observation de rang (1+20)/2=10,5 c'est à dire la moyenne des valeurs de rang 10 et 11. La série est donc partagée en deux séries de 10 valeurs. La médiane de la première série, c'est à dire le premier quartile est la valeur de rang (1+10)/2=5,5 c'est-à-dire la

moyenne des valeurs de rangs 5 et 6. De la même façon le troisième quartile est la moyenne des valeurs de rangs 15 et 16. La série est bien partagée en quatre séries de cinq valeurs.

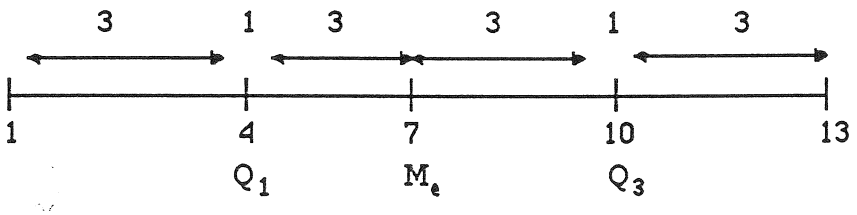


Pour un effectif de 18, pair mais non multiple de 4, la médiane n'est pas une valeur de la série mais la partage en deux séries de 9 valeurs. Les deux quartiles sont des valeurs de la série.

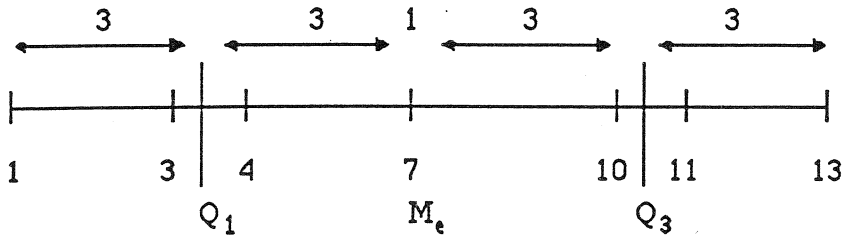


Pour un effectif de 13, impair, la médiane est une valeur de la série, c'est la valeur de rang (1+13)/2=7. Pour les quartiles deux méthodes sont possibles.

La première, utilisée par Tukey, calcule la médiane des 7 premières valeurs, c'est à dire de la première série, médiane comprise, on obtient ainsi le premier quartile qui est la valeur de rang (1+7)/2=4 puis la médiane de la deuxième série, médiane comprise, on obtient ainsi le troisième quartile, c'est-à-dire la valeur de rang (7+13)/2=10.



La deuxième méthode calcule la médiane de la même façon puis calcule les deux autres quartiles dans chacune des séries obtenues mais sans reprendre la médiane.



Remarque : Il existe une définition plus systématique (mais moins naturelle) des quartiles qui donne des résultats légèrement différents et qui est présentée au chapitre 3 (page 68).

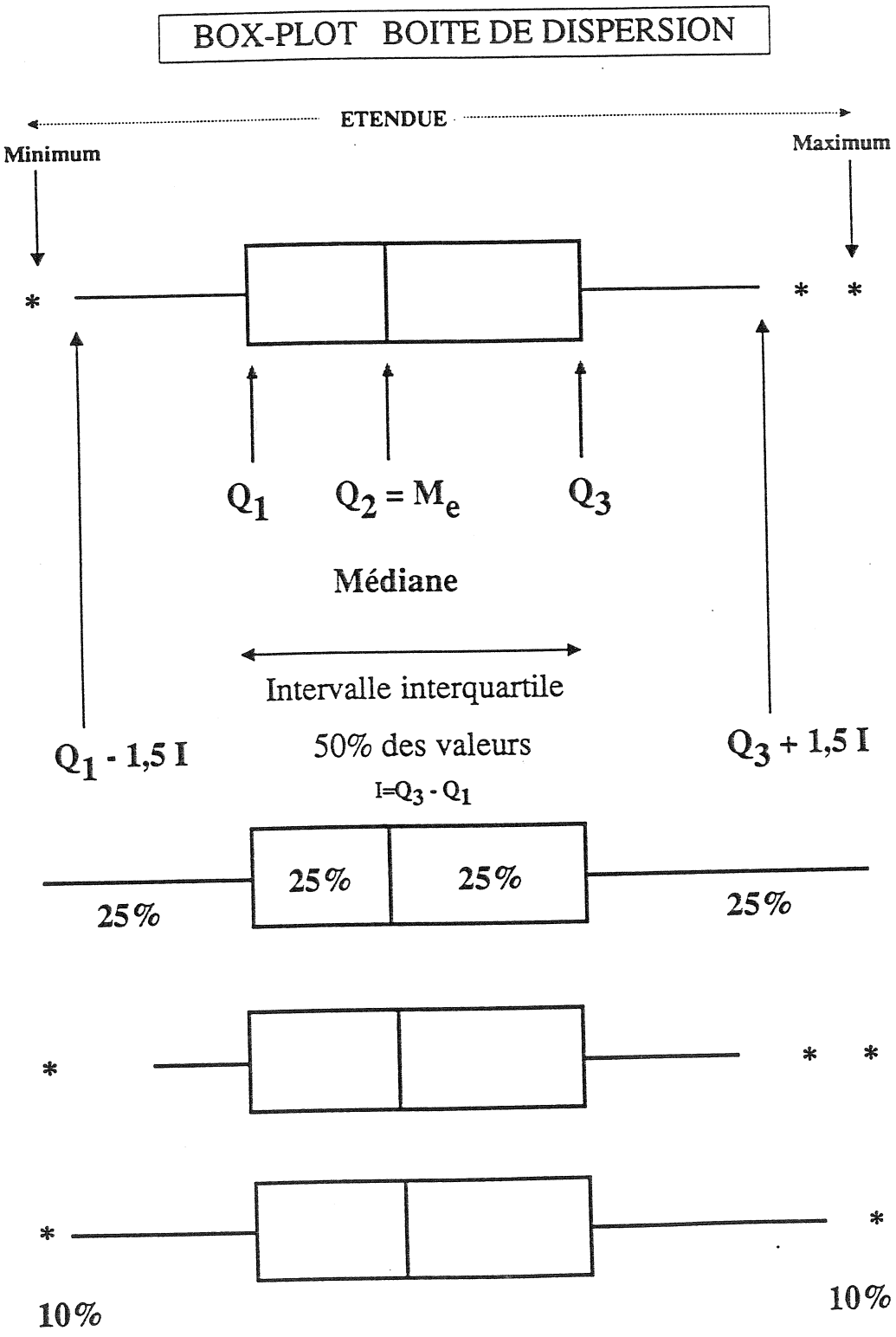
3) Construction du graphique en boîte

Sur une échelle allant du minimum au maximum de la série, on place la médiane M_e et les deux autres quartiles Q_1 et Q_3. On trace un rectangle de longueur $Q_3 - Q_1$, c'est-à-dire égale à l'écart interquartile I, coupé par un trait à la hauteur de la médiane.

On prolonge de chaque côté de la "boîte" par des "moustaches" allant, soit jusqu'au minimum d'un côté et au maximum de l'autre (voir deuxième dessin page suivante), soit on les limite par deux petits traits à la valeur $Q_1 - 1{,}5$ I d'un côté et à la valeur $Q_3 + 1{,}5$ I de l'autre. Dans ce dernier cas, chaque valeur à l'extérieur de ces limites est alors qualifiée d'aberrante et on la représente distinctement, par un symbole * par exemple (voir premier dessin page suivante).

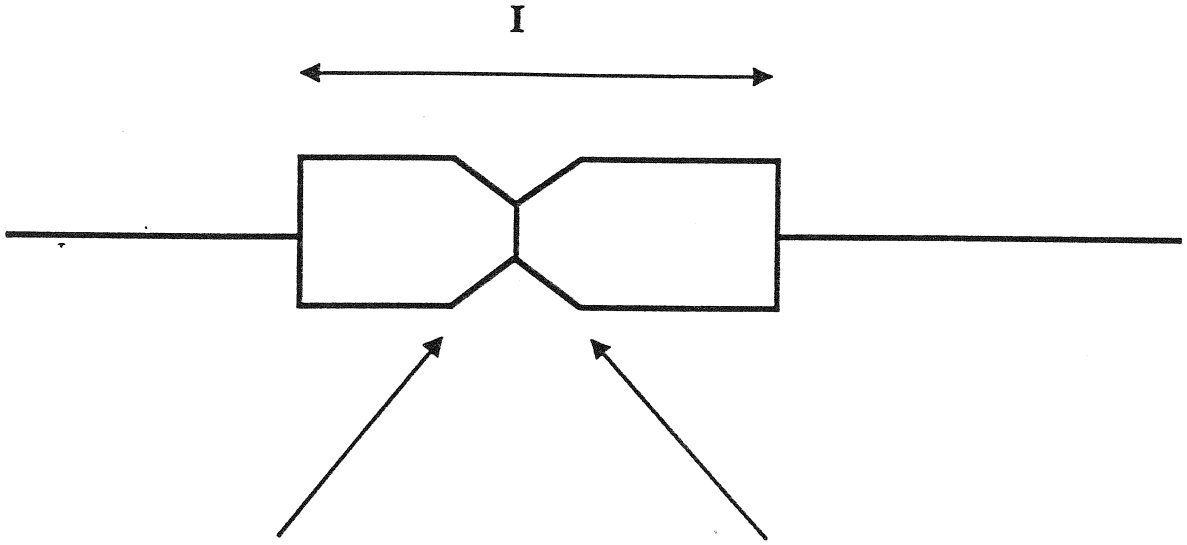
On peut encore limiter chacune des "moustaches" à la dernière valeur de la série se trouvant avant chacune des limites précédentes (voir troisième dessin page suivante).

Au lieu de ces limites, on peut s'arrêter au premier et au dernier décile, c'est-à-dire déclarer aberrantes 10% des valeurs au début et à la fin de la série (voir quatrième dessin page suivante). On peut faire de même avec 5% des valeurs.


BOX-PLOT BOITE DE DISPERSION
ETENDUE
Minimum
Maximum
Q_1
$Q_2 = M_e$
Q_3
Médiane
Intervalle interquartile
50% des valeurs
$I=Q_3 - Q_1$
$Q_1 - 1,5\ I$
$Q_3 + 1,5\ I$
25%
25%
25%
25%
10%
10%

Si la série statistique représente les valeurs obtenues dans un échantillon, on peut faire un graphique en "taille de guêpe" avec toutes les variantes précédentes et indiquant en plus un intervalle de confiance pour la médiane de la population. Ceci permet, lorsqu'on dessine cote à cote plusieurs graphiques correspondant à plusieurs échantillons, de savoir si les différences entre les médianes des populations sont statistiquement significatives.

Cas de résultats obtenus à partir d'un échantillon de taille n



Intervalle de confiance à 95 %

pour la médiane M_e de la population

$$\left[m_e - 1{,}58 \frac{I}{\sqrt{n}} ; m_e + 1{,}58 \frac{I}{\sqrt{n}} \right]$$

où m_e est la médiane de l'échantillon

3) Avantage du graphique en boîte

Nous allons illustrer ce graphique à partir d'un exemple extrait du livre "Contremanuel de Statistique et Probabilité" - Éditions Vie Ouvrière-Bruxelles.

2. DONNÉES BRUTES

2.1. Le plus simple sera de suivre un exemple. Dans une entreprise de fabrication métallique de la région liégeoise, à un certain moment de l'année 1973, travaillaient 189 ouvriers de plus de 21 ans (voir R. Leroy [10]). Leurs salaires horaires étaient consignés dans un fichier classé par ordre alphabétique. Voici la liste de ces salaires en francs notés en suivant le fichier :

107, 100, 93, 91, 100, 96, 100, 100, 100, 102, 105, 102, 100, 105, 100, 100, 107, 100, 102, 106, 100, 107, 100, 100, 96, 100, 100, 100, 96, 91, 102, 106, 100, 102, 107, 100, 100, 105, 100, 100, 91, 100, 100, 96, 102, 100, 102, 96, 91, 96, 96, 100, 100, 105, 106, 100, 100, 105, 106, 96, 100, 96, 100, 105, 106, 93, 100, 107, 100, 102, 93, 106, 100, 105, 100, 102, 107, 102, 93, 100, 106, 96, 105, 100, 102, 100, 96, 102, 100, 93, 100, 100, 105, 100, 93, 100, 106, 100, 106, 93, 105, 102, 102, 96, 107, 100, 102, 106, 100, 107, 102, 100,100, 100, 93, 106, 100, 100, 107, 96, 100, 106, 100, 102, 106, 100, 105, 102, 106, 105, 100, 106, 107, 100, 102, 96, 102, 100, 100, 107, 105, 102, 100, 100, 105, 100, 105, 102, 102, 102, 100, 100, 96, 96, 102, 102, 100, 102, 100, 102, 100, 102, 91, 100, 91, 102, 102, 107, 100, 102, 100, 107, 106, 100, 100, 102, 102, 100, 100, 102, 102, 100, 106, 100, 105, 100, 107, 100, 100.

2.2. Voici le tableau brut des salaires horaires des 121 ouvrières de plus de 21 ans de l'entreprise liégeoise qui nous a servi d'exemple principal jusqu'à présent.

90, 91, 98, 90, 91, 90, 93, 91, 90, 91, 90, 90, 98, 91, 90, 96, 93, 90, 90, 90, 93, 91, 102, 90, 91, 96, 90, 96, 91, 91, 91, 90, 91, 90, 91, 90, 90, 96, 98, 93, 90, 90, 93, 90, 90, 90, 90, 93, 102, 91, 91, 91, 90, 91, 90, 90, 91, 96, 93, 91, 102, 91, 91, 93, 90, 91, 93, 90, 96, 93, 90, 91, 90, 96, 93, 90, 91, 93, 91, 96, 96, 91, 90, 90, 93, 90, 91, 93, 96, 91, 93, 90, 90, 93, 91, 91, 96, 90, 90, 91, 91, 93, 90, 90, 91, 90, 96, 90, 90, 90, 93, 96, 90, 90, 96, 90, 91, 90, 90, 90, 90.

En rapprochant cette liste de celle du n° 2.1, on voit tout de suite que les femmes gagnent moins que les hommes. Mais cette observation est encore bien vague.

La remarque finale est effectivement bien vague. Le livre propose de représenter ces deux distributions par leurs graphiques en barres, disposés côte à côte.

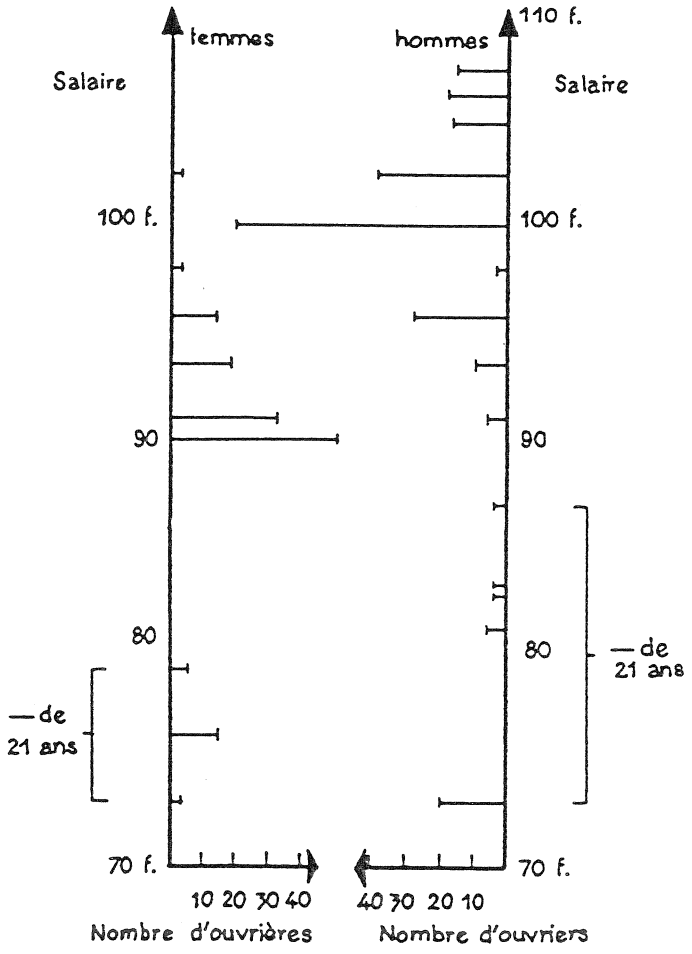
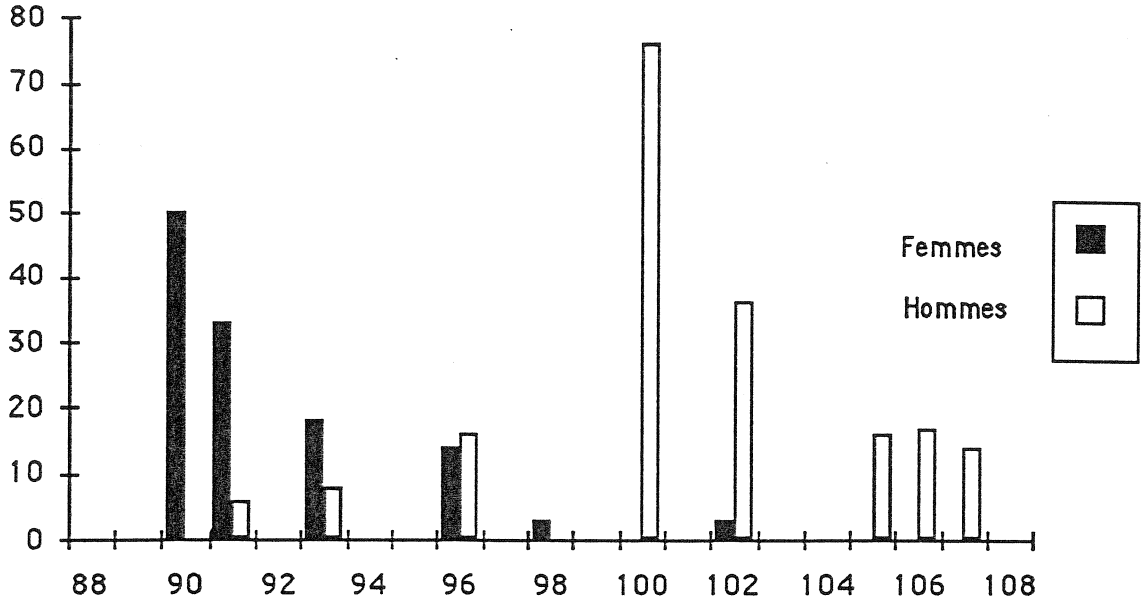


Figure II.9 : Comparaison des salaires des ouvriers et des ouvrières.

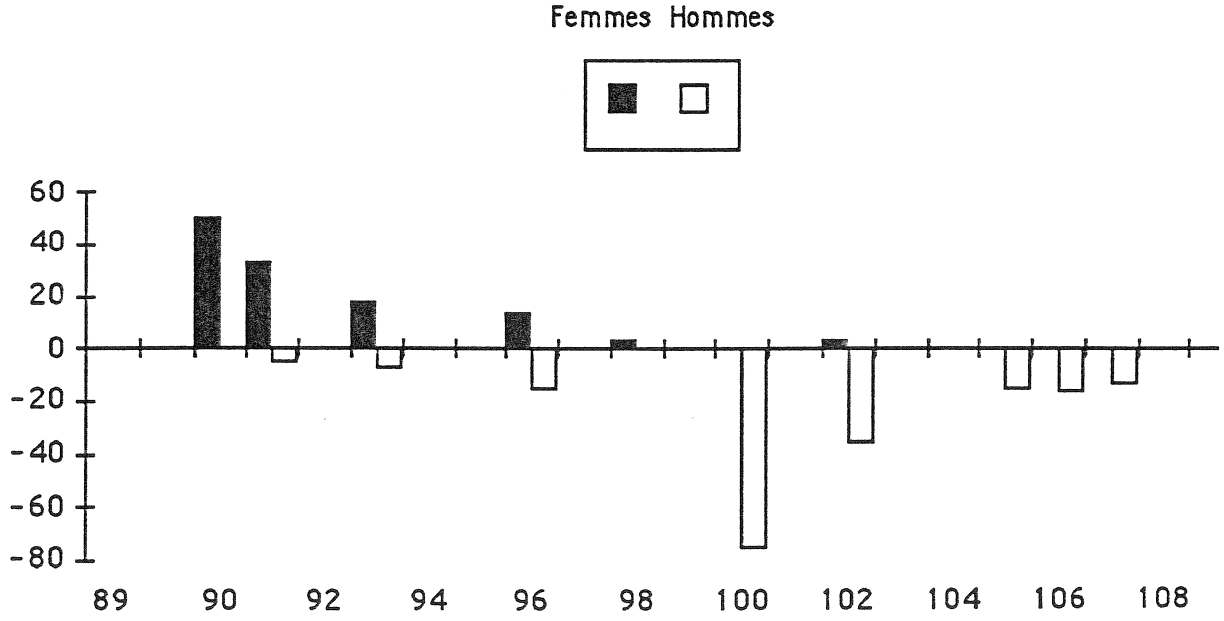
Les auteurs font alors remarquer que le mode (c'est-à-dire la valeur de la série la plus fréquente) des femmes est inférieur au plus petit salaire masculin. De même quelques femmes seulement ont un salaire supérieur au mode des hommes.L'inconvénient du mode, c'est que cela peut aussi bien être une valeur moyenne de la série (comme pour les hommes), qu'une petite valeur (comme pour les femmes) ou encore une grande valeur (dans d'autres distributions).

On pourrait comparer ces deux distributions par d'autres graphiques (obtenus avec le logiciel Excel) :

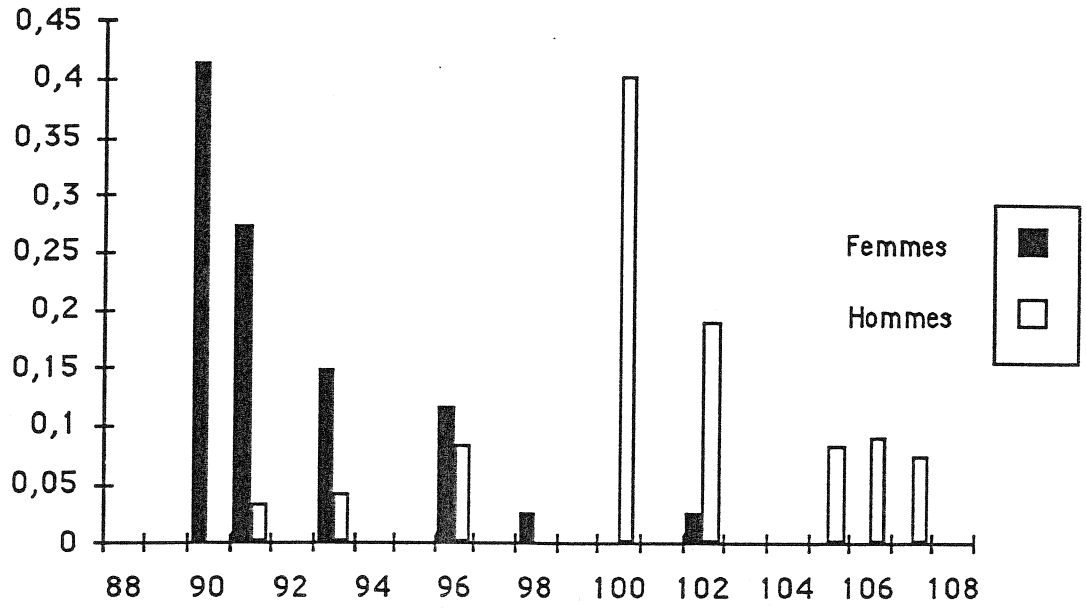
Graphiques en barres accolées :



ou symétriques :



ou mieux, puisqu'il s'agit de comparer deux séries d'effectifs sensiblement différents, les graphiques de fréquences.



mais les conclusions seraient à peu près les mêmes.

On pourrait aussi calculer différents paramètres.

La moyenne des salaires est de 100,94 F chez les hommes et de 91,91 F chez les femmes, ce qui confirme nos premières impressions mais ne nous indique rien sur les deux distributions.

On peut avoir une idée sur la dispersion des salaires de chaque groupe en comparant les écarts-types ou mieux les coefficients de variation :

Pour les hommes : écart-type 3,9 coefficient de variation 3,8

Pour les femmes : écart-type 2,7 coefficient de variation 2,95

ce qui signifie que la distribution des salaires des femmes est moins étalée que celle des hommes.

Voyons ce que peut apporter le graphique en boîte.

Pour la construction, la première chose à faire est de classer les valeurs.

91	91	91	91	91	91	93	93	93	93
93	93	93	93	96	96	96	96	96	96
96	96	96	96	96	96	96	96	96	96
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	102	102	102	102
102	102	102	102	102	102	102	102	102	102
102	102	102	102	102	102	102	102	102	102
102	102	102	102	102	102	102	102	102	102
102	102	105	105	105	105	105	105	105	105
105	105	105	105	105	105	105	105	106	106
106	106	106	106	106	106	106	106	106	106
106	106	106	106	106	107	107	107	107	107
107	107	107	107	107	107	107	107	107	

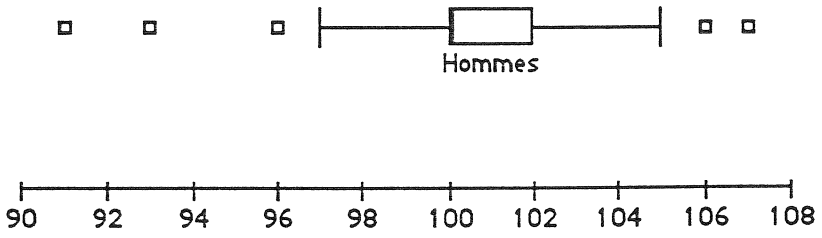
Calcul des quartiles pour les hommes :

Le rang de la médiane est (1+189)/2=95 ce qui donne la valeur 100

Le rang du premier quartile est (1+95)/2=48 ce qui donne la valeur 100

Le rang du troisième quartile est (95+189)/2=142 ce qui donne la valeur 102

L'intervalle interquartile est donc égal à 102-100 c'est-à-dire 2, et les "moustaches" s'étendront de 100 -1,5x2 = 97 à 102+1,5x2 = 105 d'où le graphique (obtenu avec le logiciel Statworks):



90	90	90	90	90	90	90	90	90	90	90
90	90	90	90	90	90	90	90	90	90	90
90	90	90	90	90	90	90	90	90	90	90
90	90	90	90	90	90	90	90	90	90	90
90	90	90	90	90	90	91	91	91	91	91
91	91	91	91	91	91	91	91	91	91	91
91	91	91	91	91	91	91	91	91	91	91
91	91	91	91	91	91	93	93	93	93	93
93	93	93	93	93	93	93	93	93	93	93
93	93	96	96	96	96	96	96	96	96	96
96	96	96	96	96	98	98	98	102	102	102

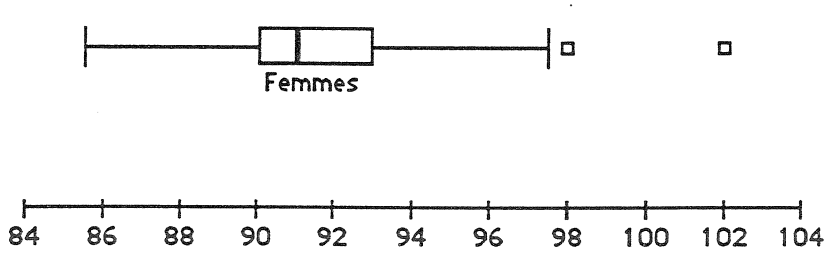
Calcul des quartiles pour les femmes :

Le rang de la médiane est (1+121)/2=61 ce qui donne la valeur 91

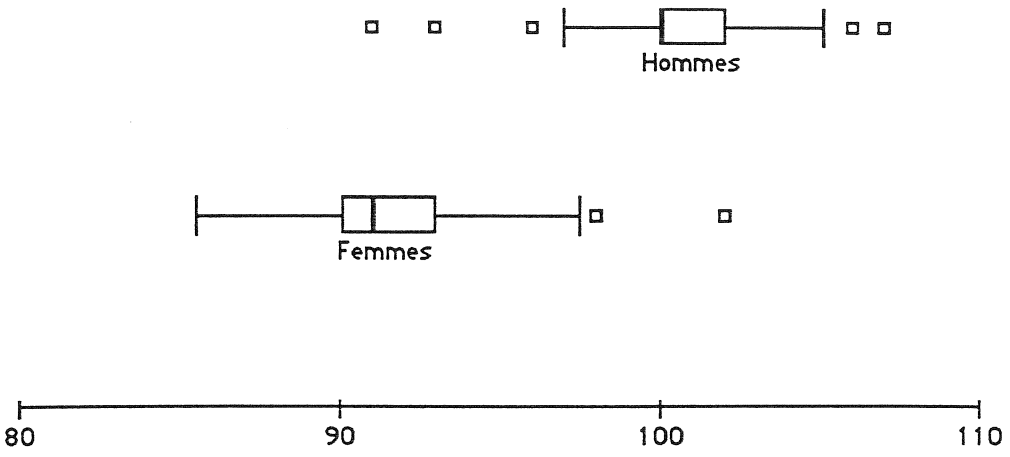
Le rang du premier quartile est (1+61)/2=31 ce qui donne la valeur 90

Le rang du troisième quartile est (61+121)/2=91 ce qui donne la valeur 93

L'intervalle interquartile est donc égal à 93-90 c'est à dire 3, et les "moustaches" s'étendront de 90 -1,5x3 = 85,5 à 93+1,5x3 = 97,5 d'où le graphique :



Pour effectuer la comparaison il faut évidemment construire les graphiques côte à côte et sur une même échelle :

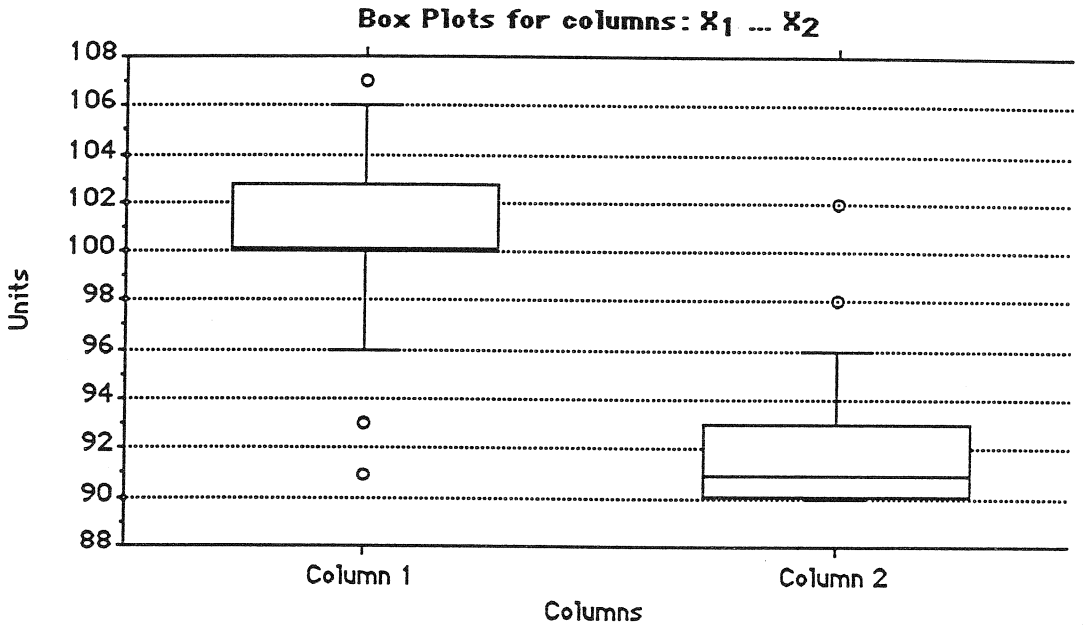


On peut alors faire les constatations suivantes:

Au moins 75% des hommes ont un salaire supérieur au deuxième salaire féminin (98F). En réalité ce pourcentage est de 84% à cause des ex aequo.

Au moins 50% des femmes (en fait, 73%) ont un salaire inférieur ou égal au plus petit salaire masculin (91F).

Si on adopte la variante qui consiste à prolonger les "moustaches" jusqu'aux premier et au dernier déciles, on obtient (avec le logiciel Statview) le graphique suivant :



Cette fois, on constate que (au moins) 90% des hommes ont un salaire supérieur ou égal à 96F, alors que (au moins) 90% des femmes ont un salaire inférieur ou égal à cette même somme.

Comme dans la méthode précédente, il faut revenir à la distribution de départ pour avoir les valeurs exactes des pourcentages, mais ces résultats mettent en lumière un des intérêts des graphiques : faire apparaître des structures non évidentes sur les séries elles-mêmes.

Danger! Approximations...

Réflexion sur des effets d'approximations dans les calculs en statistique

Jean-Claude Régnier

Voici trois exemples dont le but est de faire comprendre l'intérêt des précautions à prendre lorsque l'on remplace une valeur par une "*valeur approchée tout à fait raisonnable*". Lorsque l'on utilise une machine à calculer, comme ceci est pratiquement une nécessité pour réaliser des calculs statistiques, il convient de travailler avec la précision maximale de l'instrument dans le déroulement des opérations. Le recours à une valeur approchée d'un ordre de grandeur compatible avec la situation d'étude n'intervient alors que pour en communiquer les informations et les conclusions. Notons cependant que l'usage de la précision maximale du calculateur automatique ne fait que diminuer le risque de provoquer les effets dont nous allons donner un exemple sans pouvoir garantir la non-réalisation de l'événement. Un tel outil ne travaille qu'avec un ensemble fini de nombres décimaux dont les nombres entiers ne sont qu'un cas particulier. Cette caractéristique est importante car elle nous rappelle quelques limites de l'instrument dans les procédures démonstratives, par exemple pour étudier le comportement d'une suite numérique déterminée par la valeur initiale : si cette valeur initiale est un nombre irrationnel comme π ou $\sqrt{2}$, il sera impossible de partir de ce point exact. Ainsi pour une suite qui serait convergente en partant de $\sqrt{2}$ et divergente à partir de toute autre valeur, l'étude empirique à l'aide d'une calculatrice s'avérera conduire inévitablement à l'erreur. Autre cas, pour comparer deux fractions $\frac{a}{b}$ et $\frac{c}{d}$ avec une calculatrice, on procède habituellement aux opérations a $\boxed{:}$ b et c $\boxed{:}$ d, certes si les affichages à l'écran sont différents, nous concluons que $\frac{a}{b} \neq \frac{c}{d}$ mais si les écritures décimales nécessitent un nombre de chiffres supérieur à la précision de l'outil et si les affichages sont identiques, nous n'avons plus qu'une présomption d'égalité car un des chiffres cachés peut être distinct.

Dans le premier exemple, nous considérons une étude portant sur le prix unitaire d'un objet en fonction de 80 points de vente existants. La variable est considérée comme une variable quantitative discrète. On cherche à obtenir le prix moyen et la dispersion des valeurs à l'aide de l'écart-type.

valeurs de la variable	6,55	6,8	6,85	6,9	7	7,1	7,35	7,4	effectif total
effectifs	7	6	8	13	16	18	7	5	80

Les calculs donnent les résultats suivants:

somme des 80 valeurs	559,4		
prix unitaire moyen	6,9925	prix unitaire moyen arrondi "raisonnablement"	7
somme des carrés des écarts à la valeur moyenne exacte	3,8005	somme des carrés des écarts à la valeur moyenne approchée	3,805
variance calculée selon la définition	0,04750625	variance calculée selon la définition	0,048
somme des carrés des valeurs de la variable	3915,405		
variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	0,04750625	**variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"**	**-0,06**

Nous repérons une aberration dans le résultat issu du calcul de la variance par la seconde méthode. En utilisant dans le cours des calculs, une valeur approchée tout à fait raisonnable (**7** pour **6,9925**), on obtient une valeur négative - **0,06** ce qui est impossible puisque la variance est par définition positive.

Dans le second exemple, nous considérons une étude portant sur la taille d'individus La variable est considérée comme une variable quantitative continue. On cherche à obtenir la taille moyenne et la dispersion des valeurs à l'aide de l'écart-type.

valeurs de la variable	[148;152[	[152;156[	[156;160[	[160;164[	[164;168[	[168;172[	[172;176[	[176;180[	
valeurs centrales	150	154	158	162	166	170	174	178	effectif total
effectifs	5	12	21	39	33	10	2	3	125

Les calculs donnent les résultats suivants:

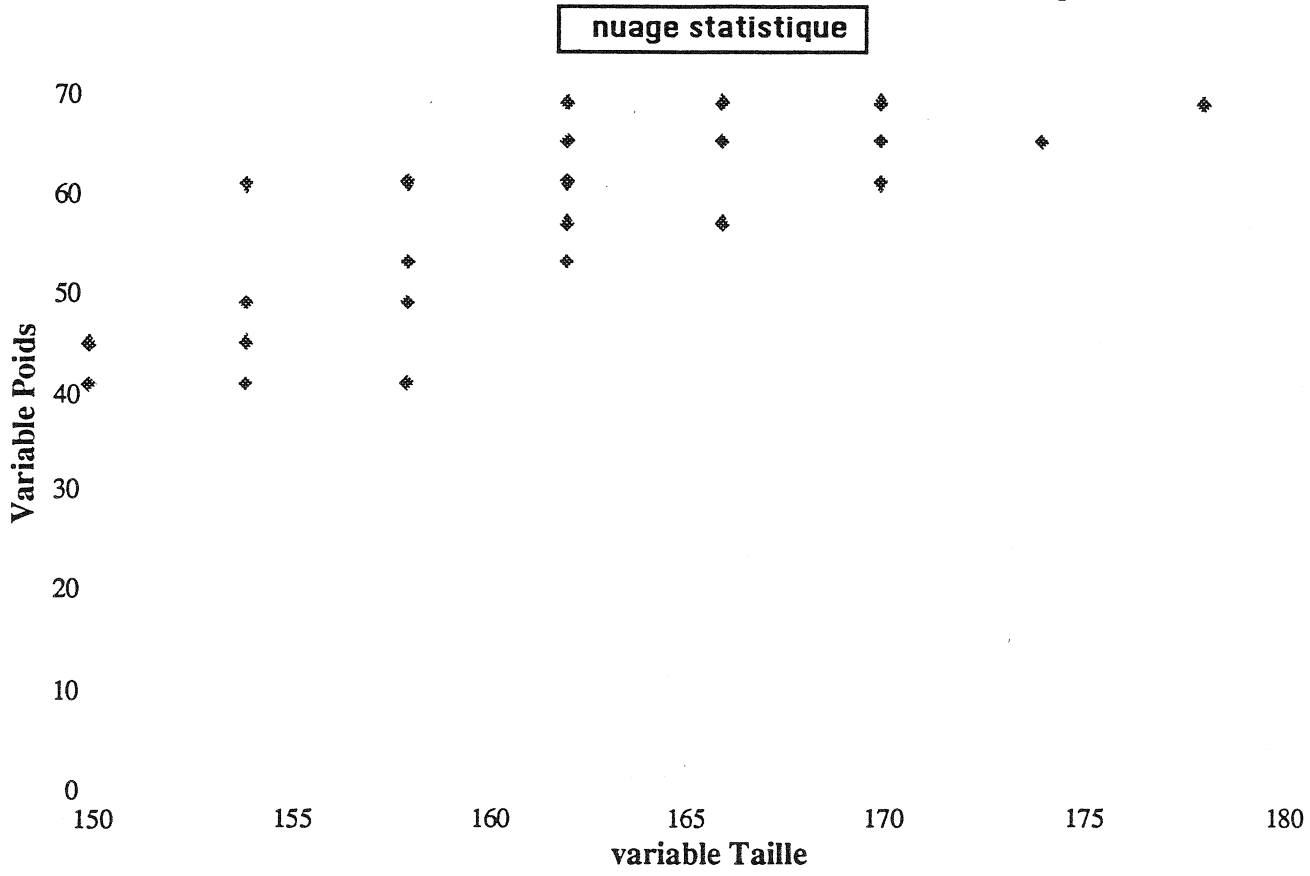
somme des 125 valeurs	20294		
TAILLE moyenne	162,352	TAILLE moyenne arrondie "raisonnablement"	162,5
somme des carrés des écarts à la valeur moyenne exacte	4032,512	somme des carrés des écarts à la valeur moyenne approchée	4035,25
variance calculée selon la définition	32,260096	variance calculée selon la définition	32,282
somme des carrés des valeurs de la variable	3298804		
variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"	32,260096	**variance calculée selon la formule "moyenne des carrés moins le carré de la moyenne"**	-15,818

Nous repérons encore une aberration dans le résultat issu du calcul de la variance par la seconde méthode. En utilisant dans le cours des calculs, une valeur approchée tout à fait raisonnable (**162,5** pour **162,352**), on obtient ici encore une valeur négative **-15,818**

Dans ce troisième exemple, nous allons considérer le croisement de la variable *Taille des individus* avec la variable *Poids des individus*. Le tableau statistique ci-après fournit les résultats à l'issue d'une enquête auprès de 125 individus.

valeurs de la variable	Taille (cm) X	[148;152[	[152;156[	[156;160[	[160;164[	[164;168[	[168;172[	[172;176[	[176;180[	
Poids (kg) Y	*valeurs centrales*	150	154	158	162	166	170	174	178	effectif
[39;43[	41	4	1	1	0	0	0	0	0	6
[43;47[	45	1	8	0	0	0	0	0	0	9
[47;51[	49	0	2	18	0	0	0	0	0	20
[51;55[	53	0	0	1	32	0	0	0	0	33
[55;59[	57	0	0	0	4	26	0	0	0	30
[59;63[	61	0	1	1	1	0	8	0	0	11
[63;67[	65	0	0	0	1	4	1	2	0	8
[67;71[	69	0	0	0	1	3	1	0	3	8
	effectifs	5	12	21	39	33	10	2	3	125

Ci-après, la représentation graphique du nuage des 125 points apporte des informations sur une éventuelle liaison statistique entre ces deux variables biométriques.



Cette recherche de lien est aussi étayée par le recours au coefficient de corrélation linéaire de Bravais-Pearson R_{BP} entre deux variables quantitatives X et Y. Nous donnons l'expression algébrique :

$R^2_{BP} = \frac{Cov^2(X;Y)}{V(X)\,V(Y)}$ où Cov(X ; Y) désigne la covariance de X et de Y , V(X) et V(Y) la variance respective de X et de Y. Par construction les valeurs de R^2_{BP} sont comprises entre 0 et 1 donc celles de R_{BP} , entre -1 et 1.

Nous convenons de calculer leurs valeurs par les algorithmes suivant :

Variance = moyenne des carrés des valeurs de la variable - carré de la moyenne

Covariance = moyenne des produits des valeurs des deux variables - produit des deux moyennes

Le tableau ci-dessous donne alors les valeurs obtenues à partir de diverses conduites d'approximation :

	résultats exacts	*résultats approchés (1)*	*résultats approchés (2)*
somme des valeurs de X	20294	***	***
moyenne de X	**162,352**	**162,4**	**162,3**
somme des carrés de X	3298804	***	***
moyenne des carrés de X	26390,432	***	***
carré de la moyenne de X	26358,171904	26373,76	26341,29
somme des valeurs de Y	6833	***	***
moyenne de Y	**54,664**	**54,7**	**54,7**
somme des carrés de Y	379317	***	***
moyenne des carrés de Y	3034,536	***	***
carré de la moyenne de Y	2988,152896	2992,09	2992,09
somme des produits XY	1113454	***	***
moyenne des produits XY	8907,632	***	***
produit des moyennes	8874,809728	8883,28	8877,81
covariance cov(X,Y)=	32,8222719999994	24,351999999999	29,8219999999983
$cov^2(X;Y)$ =	1077,30153924195	593,019903999949	889,351683999898
variance V(X)=	32,2600959999982	16,6719999999987	49,1419999999962
variance V(Y)=	46,383104	42,4459999999999	42,4459999999999
coefficient $R^2_{BP}(X,Y)$=	0.719965715975999	0,838001505446728	0,426367344276117

	résultats approchés (3)	*résultats approchés (4)*	*résultats approchés (5)*
somme des valeurs de X	***	* * *	***
moyenne de X	**162,4**	**162,5**	**162**
somme des carrés de X	***	* * *	***
moyenne des carrés de X	***	* * *	***
carré de la moyenne de X	26373,76	26406,25	26244
somme des valeurs de Y	***	* * *	***
moyenne de Y	**54,6**	**54,66**	**55**
somme des carrés de Y	***	* * *	* * *
moyenne des carrés de Y	***	* * *	* * *
carré de la moyenne de Y	2981,16	2987,7156	3025
somme des produits XY	***	* * *	***
moyenne des produits XY	***	* * *	***
produit des moyennes	8867,04	8882,25	8910
covariance cov(X,Y)=	40,5919999999987	25,3819999999996	-2,36800000000039
$cov^2(X;Y)$ =	1647,7104639999	644,24592399998	5,60742400000186
variance V(X)=	16,6719999999987	-15,8179999999993	146,432000000001
variance V(Y)=	53,3759999999997	46,8204000000005	9,53600000000006
coefficient $R^2_{BP}(X,Y)$=	1,85160000598374	-0,869891302577595	0,004015699066036

Dans la colonne *Résultats exacts,* nous rapportons effectivement les valeurs exactes obtenues à partir des calculs. Dans les autres colonnes nous avons adopté une conduite d'approximation en modifiant les valeurs des moyennes en choisissant des *valeurs concrètement raisonnables*. Force est de constater les effets spectaculaires que ces choix peuvent provoquer.

	résultats exacts	*résultats approchés (1)*	*résultats approchés (2)*
moyenne de X	**162,352**	**162,4**	**162,3**
moyenne de Y	**54,664**	**54,7**	**54,7**
coefficient $R^2_{BP}(X,Y)=$	0,719965715975999	0,838001505446728	0,426367344276117
coefficient de Bravais-Pearson $R_{BP}(X,Y) \approx$	0,8485	0,9154	0,6529

	résultats approchés(3)	*résultats approchés(4)*	*résultats approchés (5)*
moyenne de X	**162,4**	**162,5**	**162**
moyenne de Y	**54,6**	**54,66**	**55**
coefficient $R^2_{BP}(X,Y)=$	1,85160000598374	-0,869891302577595	0,004015699066036
coefficient de Bravais-Pearson $R_{BP}(X,Y) \approx$	absurde	absurde	0,0633

Comment ne pas s'étonner devant les faits suivants :

- en arrondissant le poids moyen de 54,664 kg à 54,7 kg et la taille moyenne de 162,352 cm à la valeur au dixième près par excès c'est à dire en prenant 162,4 cm ou à la valeur au dixième près par défaut c'est à dire en prenant 162,3 cm , nous faisons passer la valeur du coefficient $R^2_{BP}(X,Y) \approx 0,7199$ de 0,8300 à 0,4236 ce qui revient après l'avoir augmenté à le diminuer de moitié !

- en choisissant les valeurs approchées au dixième près suivantes, par excès pour la taille moyenne avec 162,4 cm et par défaut pour le poids moyen avec 54,6 kg, nous obtenons une aberration avec une valeur supérieure à 1!

- en choisissant une valeur approchée au 1/100 près par défaut pour le poids moyen, c'est à dire 54,66 kg et la valeur centrale 162,5 cm pour la taille moyenne, nous obtenons une valeur aberrante parce que négative !

- en choisissant d'arrondir aux unités ce qui revient à considérer un poids moyen de 55 kg et une taille moyenne de 162 cm, nous aboutissons à une valeur coefficient $R^2_{BP}(X,Y)$ proche de 0 !

De cela vont résulter des décisions opposées. Ainsi une valeur du coefficient $R^2_{BP}(X,Y)$ voisine de 0,8380, c'est à dire une valeur du coefficient de Bravais-Pearson voisine de 0,91, incite à rejeter l'hypothèse d'absence de lien linéaire. Mais il n'en est pas de

même avec une valeur de $R^2_{BP}(X,Y)$ voisine de 0,0040 qui donne une valeur de $R_{BP}(X,Y)\approx$ 0,06. Cette dernière incite à renoncer à ce rejet par référence au cadre théorique qui donne un sens à ce coefficient et dont nous ne faisons pas état ici.

En ce qui concerne la situation où le coefficient prend une valeur supérieure à 1, il m'a été donné de l'observer effectivement lors d'une surveillance d'épreuve de baccalauréat , il y a une dizaine d'années. J'ai pu alors me rendre compte par hasard qu'une candidate se débattait avec sa calculatrice et que ses résultats l'étonnaient. Recommençant plusieurs fois, elle trouvait cependant toujours un coefficient supérieur à 1. Force est de constater qu'elle entrait à chaque fois des valeurs arrondies qui la plaçait dans des conditions analogues à celles que j'ai tentées de décrire dans cet article. Les erreurs ainsi produites ne relèvent pas d'une incompréhension des notions statistiques mises en œuvre mais bien des limites posées par un instrument de calcul. L'enseignement a tout intérêt à intégrer ce genre de problème dans l'éducation statistique des élèves. Le but principal de cet article est d'attirer l'attention des enseignants et des étudiants sur ce phénomène, effet secondaire des avantages majeurs qu'apporte les calculateurs automatiques.

Je n'achète que s'il y a deux fois plus de chocolat que d'anis !
BONBONS
AU CHOIX 10F
LE MÉLANGE
MÉLANGE
BONBONS
MÉLANGE
BONBONS
ANIS ET CHOCOLAT

Chapitre 8

De la vérité autoproclamée à la vraisemblance reconnue

Comment peut-on reconnaître et contrôler les risques encourus lors d'une prise de décision en situation incertaine ? Quelle place et quel rôle tiennent les notions de significativité d'un résultat et de représentativité d'un échantillon dans la prise de décision ?

Jean-Claude Régnier

Les propos qui vont suivre visent à expliciter une procédure d'aide à la prise de décision utilisant un raisonnement statistique. Dans ce raisonnement l'idée de vérité est remplacée par celle de vraisemblance. Entre deux hypothèses alternatives, il s'agit de décider celle que nous souhaitons tenir pour vraie parce qu'au vu des informations que nous détenons, elle nous apparaît comme la plus vraisemblable des deux. A notre avis, ce raisonnement est conduit plus ou moins spontanément et implicitement par chacun de nous. La statistique constitue alors un domaine scientifique dans lequel sont développés des outils conceptuels et des méthodes permettant de modéliser certaines catégories de situations de décision. Cette modélisation apporte des informations qui constituent une aide à la prise de décision. Cependant nous ne devons pas oublier qu'en fin de compte la décision appartient toujours au décideur qui traitera les informations recueillies selon son point de vue.

Un exemple dans l'évaluation scolaire d'une compétence.

A l'issue d'une séquence didactique, l'enseignant est soucieux de pouvoir porter un jugement relatif au niveau de compétence atteint par chaque apprenant dans le domaine. Un modèle simplifié de la situation d'évaluation réduit la décision à deux modalités : (Ho) *niveau de compétence insuffisant* contre (H1) *niveau de compétence satisfaisant.*

Pour obtenir une aide à cette prise de décision, l'enseignant recourt à un questionnaire constitué d'une ou plusieurs questions indépendantes ou non les unes des autres, ouvertes, semi-ouvertes ou fermées. Chaque question est construite de manière à ce que la production de la réponse par l'apprenant nécessite la mise en œuvre de la compétence dont le développement est visé.

L'apprenant peut pour sa part répondre ou ne pas répondre et quand il fournit une réponse, cette dernière peut être correcte ou incorrecte. Ceci amène à considérer qu'à chaque question est associée l'une des trois modalités :

réponse correcte (R), réponse incorrecte (E) ou non-réponse (N).

Nous présupposons qu'un questionnaire ne peut comporter qu'un nombre fini (au plus 500 selon notre avis) de questions . De ce fait il ne peut apporter qu'une **information**

partielle relativement à la *compétence* qui ne peut être cernée exhaustivement au travers d'un nombre fini de questions.

Un type de questionnaire dont la construction et l'usage s'avèrent assez aisés est le questionnaire à choix multiples (QCM). Il consiste à poser une question et à proposer plusieurs réponses plausibles dont une seule est la réponse attendue, celle que l'on considère comme la réponse correcte.

question Q_i(*énoncé*) -->**réponse**(*modalités proposées*): $m_{1i}m\ m_{2i}m...m_{ki}m$

La variable de décision D communément en usage est le nombre de réponses correctes fournies au questionnaire. L'espace des résultats possibles est l'ensemble des nombres entiers de 0 à q pour un questionnaire comportant q questions. Si nous considérons que les questions sont mutuellement indépendantes et d'égales difficultés, cette variable n'est autre qu'une variable binomiale de paramètres q et π où π est la probabilité de fournir une réponse correcte.

Dès lors nous percevons que la valeur de π est différente selon que nous nous plaçons sous l'hypothèse Ho ou l'hypothèse H_1. Il paraît vraisemblable de considérer que la valeur π_1 de π sous H1 est supérieure à celle π_0 de π sous Ho.

L'intérêt de ce modèle est qu'il nous permet de calculer directement la probabilité d'obtention d'un résultat k particulier de la variable de décision D "nombre de réussites" en fonction du nombre q de questions et de la probabilité de réussite. Les formules algébriques utiles sont :

$$\text{Prob}\{ D = k \} = C_q^k \pi^k (1-\pi)^{q-k} = \frac{q!}{k!\,(q-k)!} \pi^k (1-\pi)^{q-k}$$

$$\text{Prob}\{ D \geq k \} = \sum_{j=k}^{q} \text{Prob}\{D = j\}$$

$$\text{Prob}\{ D < k \} = 1 - \text{Prob}\{ D \geq k \}$$

Examinons maintenant un exemple.

L'usage français consacre à l'ensemble des nombres entiers de 0 à 20 une place toute particulière dont nous ne connaissons nullement les raisons originelles. Cela se concrétise en considérant un questionnaire à 20 questions et "une note sur 20" qui dans notre modèle correspondra au nombre de questions ayant obtenu une réponse correcte du point de vue de l'enseignant.

situation n°1 :

L'espace de décision est habituellement ainsi formé :

{ 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 } est la zone de rejet de l'hypothèse de *compétence* ou celle d'acceptation de l'hypothèse d'*incompétence*

{ 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 } est la zone d'acceptation de l'hypothèse de *compétence* ou celle zone de rejet de l'hypothèse d'*incompétence*

Essayons alors de mesurer les risques encourus. Pour cela nous allons considérer les valeurs de π selon les deux hypothèses H_0 et H_1. Nous posons que la probabilité π_0 de réussite à une question sous l'hypothèse (Ho) *niveau de compétence insuffisant* vaut 20% ($\pi_0 = 0{,}20$) que alors la probabilité π_1 de réussite à une question sous l'hypothèse (H1) *niveau de compétence satisfaisant.* vaut 80% ($\pi_1 = 0{,}80$).

Le premier risque encouru par l'évaluateur est de commettre l'erreur de rejeter à tort l'hypothèse H_0 d'*incompétence* alors qu'elle est vraie. Cette erreur admet la probabilité α suivante d'être commise en fonction des conditions posées.

α = Prob { D≥10 / Ho vraie } = 0,003

Le second risque encouru par l'évaluateur est de commettre l'erreur de rejeter à tort l'hypothèse H_1 de compétence alors qu'elle est vraie. Cette erreur admet la probabilité β suivante d'être commise en fonction des conditions posées.

β = Prob { D<10 / H1 vraie } = 0,001

Ces deux calculs sont effectués à partir des formules rappelées plus haut à l'aide d'une table, d'une calculatrice ou d'un tableur (comme Excel 5).

Examinons maintenant la variation des valeurs de α et β selon les valeurs de π.

π_0	0,05	0,1	0,2	0,3	0,4	0,5	π_1	0,6	0,66	0,7	0,8	0,9
α	0,000	0,000	0,003	0,048	0,245	0,588	β	0,128	0,043	0,017	0,001	0,000

Il apparaît que si la probabilité de réussite sous l'hypothèse Ho est inférieure ou égale à 0,3 alors le choix de 10 comme valeur critique frontière entre les deux zones assure un risque de première espèce inférieure à un niveau de 0,05. Ceci est réalisé dès que l'on construit un QCM avec quatre modalités dont une détermine la réponse correcte. Pour un QCM à 3 modalités la probabilité $\pi_0 = \frac{1}{3}$ conduit une valeur de α entre 0,09 et 0,10.

Quant à l'erreur de seconde espèce, sa probabilité est inférieure à 0,05 dès que la probabilité de réussite sous H_1 est supérieure à 2/3.

Ainsi le choix de ce critère (à savoir : "10 réponses correctes sur 20 attendues") pour décider entre les deux hypothèses: *compétence* et *incompétence* demeure sous les conditions explicitées un moyen relativement sûr tout en mettant en évidence l'impossibilité de connaître la vérité puisqu'à chaque choix est attaché un risque d'erreur.

situation n°2 :

Nous allons cette fois fixer la valeur maximale de α à 0,05. De là nous cherchons la nouvelle partition de l'espace de décision { 0, 1, 2, ..., 18, 19, 20 } en deux zones contiguës et complémentaires : une zone correspondant au rejet de l'hypothèse de *compétence* (ou zone d'acceptation de l'hypothèse d'*incompétence*) et une zone correspondant à l'acceptation de l'hypothèse de *compétence* (ou zone de rejet de l'hypothèse d'*incompétence*)

Nous supposons que le questionnaire contient des questions à 4 choix. Du point de vue de notre modèle cela signifie que nous considérons que la conduite de réponse sous l'hypothèse Ho est analogue à 20 tirages successifs avec remise dans une urne contenant 4 boules, trois boules noires et une boule blanche. La valeur de D correspond alors au nombre de boules blanches obtenues à l'issue des 20 tirages. La probabilité de réussite sous Ho est ici de 0,25.

L'inéquation que nous avons à résoudre est :

α = Prob { D $\geq$ k / Ho vraie } $\leq$ 0,05

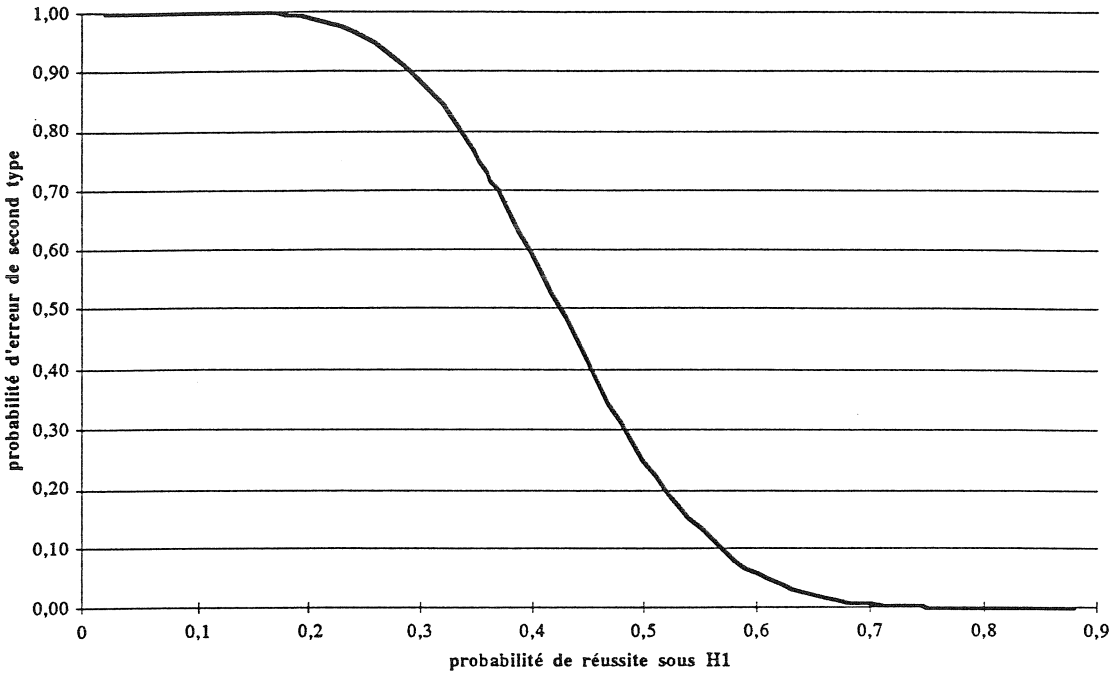
Par le calcul ou à l'aide de table, nous trouvons la valeur k = 9. Ainsi l'espace de rejet de Ho au niveau maximum de risque de 0,05 est alors :

{ 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 }appelé zone critique K.

Ce qui signifie qu'à l'issue d'une épreuve fondée sur un QCM à 4 modalités et 20 questions, tout apprenant ayant un score supérieur ou égal à 9 sera déclaré *compétent* par rejet de l'hypothèse d'*incompétence* avec un risque d'erreur de niveau inférieur à 5%.

Quand son score est inférieur strictement à 9, il est déclaré incompétent car ce score appartient à l'ensemble { 0, 1, 2, 3, 4, 5, 6, 7, 8 }appelé zone d'acceptation, qui constitue la zone de rejet de l'hypothèse de *compétence* ou zone d'acceptation de l'hypothèse d'*incompétence*

Mais nous ne pouvons oublier que ce score peut aussi être obtenu sous l'hypothèse H1 pour diverses raisons. La compétence ne conduit pas inévitablement à une réponse correcte avec certitude. Il y a les erreurs que Gaston Bachelard nomme celles de l'esprit fatigué qui ne sont pas significatives d'une représentation erronée , d'une connaissance inadéquate ou fausse. Divers facteurs introduisent des fluctuations qui nécessitent d'adapter le modèle en traduisant la probabilité de réussite à une question par une valeur strictement inférieure à 1. La valeur de la probabilité β va dépendre de la probabilité de donner une réponse correcte sachant que celui qui répond est *compétent.* La courbe ci-dessous montre la variation de la probabilité de commettre une erreur de seconde espèce en fonction de la probabilité π de fournir une réponse correcte en étant *compétent.*



Au delà d'une probabilité de 50% de réussite à une question sous l'hypothèse de compétence la valeur de β descend au-dessous de 25%.

Raisonnement statistique et test statistique :

Le raisonnement utilisé prend appui sur le schéma suivant en considérant les deux hypothèses alternatives H_0 et H_1

état de nature inconnu du décideur	H_0 est vraie	H_1 est vraie
décision de conserver H0 (rejeter H1)	décision correcte	décision erronée (erreur de type II)
décision de rejeter H0 (conserver H1)	décision erronée (erreur de type I)	décision correcte

niveaux de risques d'erreur et niveaux de confiance

état de nature	H_0 est vraie	H_1 est vraie
la décision de conserver H0 (ou rejeter H1) est prise avec les probabilités ...	$1-\alpha$ = Prob{conserver Ho/Ho vraie}	β = Prob{conserver Ho/H1 vraie}
la décision de décision de rejeter H0 (ou conserver H1) est prise avec les probabilités ...	α = Prob{ rejeter Ho/Ho vraie}	$1-\beta$ = Prob{rejeter Ho/H1 vraie}

En exploitant l'exemple précédent nous pouvons constater que les valeurs de α et β varient entre 0 et 1 en sens inverses. Diminuer la probabilité du risque d'erreur de type I revient à augmenter celle du risque d'erreur de type II.

Certes il convient de rappeler que toutes les hypothèses formulables ne sont pas susceptibles d'être validées ou rejetées à l'aide d'un **test statistique** dont l'objet est de produire une procédure permettant de construire un espace de résultats pour une **variable de décision** à partir desquels le choix de l'une ou l'autre des hypothèses est réalisé avec un **risque calculé**. Mais lorsque nous pouvons construire un modèle dans lequel les hypothèses Ho et H1 sont caractérisées par des notions et des propriétés du domaine de la statistique mathématiques nous avons grand intérêt à rechercher dans la boîte à outils statistiques un test d'hypothèse adapté à la situation pour aider la prise de décision. Cependant les conditions d'application tant théoriques que pratiques doivent être abordées et discutées pour juger de l'adéquation du modèle théorique à la situation réelle étudiée. Sans cette précaution relative à l'applicabilité du test statistique, nous nous exposons à un risque d'erreur du type III : erreur de choix du modèle ou erreur de choix du test statistique. Lorsque deux tests peuvent être utilisés avec des conditions théoriques et pratiques comparables, nous choisissons le plus puissant des deux, c'est à dire celui qui pour un niveau α donné conduit à un niveau de risque β le plus bas possible ou encore à une probabilité 1 - β de rejeter l'hypothèse Ho alors que H1 est vraie, la plus forte.

Un **test statistique** est donc une procédure logico-mathématique sur laquelle on s'appuie pour prendre une décision de choix entre deux **hypothèses alternatives** sur la base d'informations partielles issues d'un **échantillon**. Un test statistique d'hypothèse consiste à **définir explicitement une règle communicable de décision**.

Une hypothèse est dite ***hypothèse simple*** si elle peut se ramener à une confrontation simple par exemple une hypothèse centrée sur une valeur isolée d'un paramètre (moyenne, variance, écart-type, médiane, mode, coefficient d'asymétrie, coefficient d'aplatissement, etc...) du type $\Theta = \Theta_0$.

Une hypothèse est dite ***hypothèse composite*** dans le cas contraire, c'est à dire par exemple si elle peut se ramener à une confrontation complexe avec plusieurs valeurs du type $\Theta > \Theta_0$., $\Theta \geq \Theta_0$, $\Theta < \Theta_0$, $\Theta \leq \Theta_0$, $\Theta \neq \Theta_0$ ou $\Theta \in E$

Comme nous l'avons plus haut si H_0 (dite *hypothèse nulle*) et H_1 (dite *hypothèse alternative*) sont ces deux hypothèses alternatives dont une seule est vraie, nous avons quatre cas décisionnels.Deux cas correspondent à une décision correcte et les deux autres , à une décision erronée. Dans l'usage courant l'*hypothèse nulle* H_0 joue un rôle prééminent qui conduit à contrôler le risque de première espèce α en posant généralement sa valeur égale à 0,01 ou le plus souvent 0,05 et parfois 0,001. Cette *hypothèse nulle* H_0 est établie à partir de divers facteurs : la prudence, la subjectivité à laquelle on tient particulièrement, la facilité à formuler, ou encore le fait qu'il s'agit d'une hypothèse solidement établie et non contredite jusqu'alors par l'expérience que l'on met à l'épreuve dans le but de ne pas la rejeter.

La probabilité β du risque de seconde espèce est obtenue par le calcul. Cependant cela n'est possible que dans les cas où les distributions de probabilité sous l'hypothèse H_1 sont connues.

La probabilité **1-β** d'opter avec raison pour H_1 s'appelle la **puissance du test.**

Lorsqu'on a fixé α , il faut alors choisir une **variable de décision** :

- qui apporte le plus possible d'informations relatives au problème posé,
- dont la distribution de probabilité est différente selon que H_0 ou H_1 est vraie
- dont la distribution de probabilité est bien connue au moins sous la condition *Ho est vraie.*

On définit alors la **région critique K,** c'est à dire l'ensemble des valeurs de la variable de décision qui conduisent à rejeter *l'hypothèse nulle* H_0 au profit de l'hypothèse alternative H_1.

La région critique est déterminée par : Prob{rejeter Ho / Ho vraie}= α

On définit aussi la **région d'acceptation A,** c'est à dire l'ensemble des valeurs de la variable de décision qui conduisent à ne pas rejeter *l'hypothèse nulle* H_0 au détriment de *l'hypothèse alternative* H_1. La région critique K et la région d'acceptation A sont deux **ensembles complémentaires.**

Dans cette perspective **construire un test** revient à construire une **variable de décision** explicite et à déterminer **une région critique** .

Démarche à suivre pour tester une hypothèse:

- identifier et formuler les deux hypothèses H_0 et H_1, de telle sorte qu'elles correspondent à des notions et des propriétés du domaine de la statistique mathématique.
- déterminer une variable de décision,
- caractériser l'allure de la région critique K en fonction de l'hypothèse H_1,
- déterminer en compréhension ou en extension par le calcul, la région critique K en fonction de la valeur α du risque de première espèce,
- calculer, quand cela est possible, la puissance du test $1-\beta$,
- calculer la valeur expérimentale de la variable de décision à partir des valeurs obtenues sur un échantillon.
- formuler la conclusion en terme de rejet ou non rejet de l'hypothèse nulle H_0.

Les tests statistiques sont classés selon leur objet. Citons pour exemples ceux que l'on est amené à utiliser dans des recherches en Sciences de l'éducation :

Tests paramétriques, c'est à dire visant à tester une hypothèse relative à un ou plusieurs paramètres (essentiellement : moyenne, variance, proportion, etc...) d'une variable aléatoire spécifiée ou non :

- tests de comparaison d'une moyenne à une valeur fixée,
- tests de comparaison d'une variance à une valeur fixée,
- tests de comparaison d'une proportion à une valeur fixée,
- tests de comparaison d'une différence de deux moyennes à une valeur fixée,
- tests de comparaison de deux ou plusieurs moyennes ,
- tests de comparaison de deux ou plusieurs variances ,
- tests de comparaison de deux ou plusieurs proportions ,
- tests d'indépendance de deux variables quantitatives

Tests non-paramétriques dont la construction est établie sur la base d'une *fonction des observations* issues d'un échantillon aléatoire, de distribution de probabilité indépendante de la connaissance de la distribution de probabilité de la variable étudiée sur la population.

- tests d'ajustement ou d'adéquation d'une distribution observée avec une distribution théorique
- tests d'indépendance de deux variables qualitatives,
- tests d'homogénéité des profils selon les modalités d'une variable qualitative
- tests fondés sur les statistiques de rang

A ce jour la boîte à outils statistiques contient une panoplie extraordinaire de tests et continue à s'enrichir. De plus rien n'empêche le chercheur aidé du statisticien de construire un test particulier adapté à la spécificité de la recherche conduite. Le chercheur en Sciences de l'éducation soucieux de la validation des hypothèses qu'il avance en réponse à la question centrale de sa problématique peut dans la plupart des cas trouver facilement un outil adapté pour l'aider à prendre sa décision sur la base de l'épreuve des faits. Cette perspective n'est en rien réduite à un point de vue *quantitativiste* étroit. Elle se donne simplement comme but une réduction de la subjectivité en produisant des règles les plus explicites possibles de décision dont le mérite est de pouvoir être objet d'un débat méthodologique. C'est ainsi que le terme **significatif** prend un sens un peu moins flou que dans l'acception commune. Il n'y a pas là à voir une situation qui ferait inéluctablement basculer dans un impérialisme mécanique et froid de la prise automatique de décision fondée exclusivement sur la connaissance d'une quantité. En revanche cette situation éloigne du risque de l'enfermement dans un discours énonçant des propos qui paraissent auto-proclamer leur vérité sans en autoriser la falsification.

La perspective offerte par l'approche statistique des tests d'hypothèse ne conduit à aucun moment à l'**incontestable** puisque l'idée de risque y est explicitement présente et que la proposition choisie est acceptée non parce qu'elle est interprétée et proclamée vraie mais parce qu'elle est jugée plus vraisemblable au vu des informations recueillies dans le cadre d'un modèle théorique construit par le chercheur. Le doute demeure même après la décision prise. Ainsi nous ne pourrions sur la base d'un test statistique d'hypothèse obtenir une telle conclusion extraite d'un texte[1] du champ des sciences de l'éducation :

« Les conséquences de la mise en place de telles pratiques[2] ont été **incontestablement positives** dans le registre de la motivation des élèves; elles ont, **incontestablement**, contribué à réconcilier de nombreux élèves avec la pratique de la langue écrite, **ce qui n'est pas peu de chose**[3]...»[4] Cet extrait et le suivant nous montrent le recours à l'idée de *signification des informations sur lesquelles se fonde l'opinion.*

---

1 dans ces textes c'est nous qui soulignons en caractère gras

2 il s'agit des pratiques pédagogiques liées aux « méthodes actives»

3 expression que nous traduisons par *significatif*

4 Meirieu, Ph. (1993) Objectif, obstacle et situations d'apprentissage, in Houssaye, J. *ouvrage collectif* La pédagogie : une encyclopédie pour aujourd'hui ESF p 293

« De plus, le développement d'institutions scolaires sur le modèle occidental dans les pays du Tiers Monde a produit, **de toute évidence**, des effets **souvent négatifs**, discréditant des savoirs traditionnels **parfois fort utiles** sans parvenir à transmettre des savoirs nouveaux de manière **significative**. »[5]

Cette idée de caractère **significatif des informations** fait aussi l'objet d'une tentative d'explicitation dans cette approche. En quel sens est-il utilisé ? dans celui de considérer ce que peut **signifier un résultat** empiriquement obtenu par comparaison à celui qu'on obtiendrait théoriquement sous certaines conditions (en particulier sous Ho ou sous H_1). L'idée est alors de faire en sorte que le risque encouru lors de la prise de décision ne dépende que de la fluctuation de l'échantillonnage et non des erreurs d'observation, de mesure ou de report de résultats. Ici il y aurait lieu de réfléchir aux effets des approximations numériques par arrondi. Dans l'approche que nous avons développée dans cet article, l'idée de significativité est rattachée à celle de probabilité d'obtention d'un résultat par le seul fait du hasard (presque dans le sens étymologique où le résultat pourrait être obtenu avec un jeu de dés adapté). Plus la valeur de la probabilité de cet événement est grande, moins le résultat est considéré comme significatif, c'est à dire apportant une signification au phénomène étudié et à l'hypothèse que l'on a émise à son propos. Ainsi dans notre exemple l'hypothèse Ho d'incompétence est soumise à l'épreuve des faits. Obtenir un score supérieur ou égal à 9 est plus probable sous l'hypothèse de compétence que sous l'hypothèse d'incompétence. Un score de 12 sera dit *significatif* dans la mesure où il traduit l'idée que ce que nous observons (le score) est une réalisation de l'événement $\{D \geq 9\}$ qui admet une plus forte probabilité quand il est lié au phénomène (la compétence) que nous avons formalisé (recours au questionnaire, variable de décision "nombre de réussites") qu'au hasard en modélisant la conduite de réponse sous l'hypothèse Ho (d'incompétence) par une suite de tirages dans une urne contenant une boule "réponse correcte" et trois boules "réponse incorrecte". Pour donner un ordre de grandeur $Prob\{ D \geq 9 / \pi o = 0,25\} \approx 0,041$ alors que $Prob\{ D \geq 9 / \pi 1 = 0,80\} \approx 0,999$. Cette approche conduit à des énoncés du type suivant. : « Au vu des résultats empiriques échantillonnaux, *ces résultats sont significatifs* alors il est plus raisonnable de prendre la décision de rejeter l'hypothèse Ho si nous acceptons de prendre le risque de type I à un niveau α inférieur ou égal à 5% (1% ou 0,1%) ou *ces résultats ne sont pas significatifs* alors il est plus raisonnable de conserver cette hypothèse Ho mais avec un risque de type II à un niveau β souvent inconnu ».

Pour achever cet article nous évoquerons alors la notion de **représentativité de l'échantillon** utilisé pour obtenir une réalisation de la (ou des) variable(s) étudiée(s) et calculer la valeur prise par la variable de décision. Nous avons plusieurs fois trouvé dans

---

5 *ibidem* p 290

des mémoires de licence, de maîtrise ou de DEA en sciences de l'éducation des propos émettant des réserves à l'égard de la représentativité de l'échantillon utilisé sans pour autant suspendre l'usage de l'information ainsi recueillie dans l'argumentation y compris en s'appuyant sur la significativité. Il nous a semblé qu'il s'agissait plus d'une précaution formelle que d'une remarque issue d'une analyse de la méthode d'obtention de l'échantillon. Ce que confirme la manière dont les informations échantillonnales sont reprises dans l'argumentation. Ainsi abordée ni la représentativité ni la non-représentativité et le biais qu'elle induit ne sont pas en fait établis autrement que par une opinion si nous nous en tenons à ce qui est écrit. En revanche les erreurs possibles liées aux observations, aux mesures ne sont, elles, jamais évoquées. Au cours des accompagnements méthodologiques que nous avons avec les étudiants, la question de la représentativité est posée systématiquement mais d'abord et parfois même exclusivement par la taille de l'échantillon.

En quel sens pouvons-nous utiliser le qualificatif *représentatif* pour un échantillon d'une population (parente) ? ... dans le sens d'un fragment de cette population nous permettant d'obtenir une information la plus proche possible de celle que nous obtiendrions par **recensement.** De là nous pouvons y voir l'idée d'un **modèle réduit** . En quelque sorte nous considérons que l'obtention de l'information sur la population résulte d'une transformation homothétique de l'information obtenue sur le fragment.

Une autre façon d'envisager cette notion de **représentativité** est de considérer la probabilité d'appartenance d'une unité statistique à l'échantillon. Ainsi en nous rattachant aux idées d'équité et de compensation, une "certaine" représentativité de l'échantillon paraît être recevable si nous respectons la condition suivante : ***chaque unité statistique a la même chance d'appartenir à l'échantillon.*** Ainsi posée cette caractérisation de la représentativité d'un échantillon, nous sommes confronté au problème de la constitution d'un échantillon de taille n respectant cette contrainte. La méthode consiste à extraire successivement n unités au hasard avec ou sans remise en utilisant des tables de nombres aléatoires ou des fonctions aléatoires dans un tableur (Excel 5). De ce point de vue la taille est plus un facteur d'accroissement de la quantité d'information que d'accroissement de la représentativité. Un échantillon de taille 1 est représentatif dans ce sens-là.et il peut même comporter beaucoup d'information : imaginons une boîte contenant 10000 vis identiques, il suffit de mesurer un élément pour connaître la dimension et avoir l'information complète sur la population.

La connaissance du sens particulier des termes en usage est importante pour comprendre des textes énonçant des résultats fondés sur des approches statistiques. Pour illustrer nous livrons au lecteur le texte[6] suivant issu d'une recherche sur le thème de

---

6 Laveault, D.& Fournier, Ch. (1990) Evaluation par objectifs : une approche métacognitive,

l'autoévaluation et de la métacognition afin qu'il puisse exercer sa compréhension : « L'échantillon consiste en un groupe de 30 étudiants québecois de 11ème année composé de 13 filles et 17 garçons (âge moyen : 16 ans). .... Les sujets du *groupe fort* (groupe du 3ème quartile) partagent davantage la même opinion en ce qui concerne les questions qui seront choisies pour l'examen que les sujets du *groupe faible* (groupe du 1er quartile). En effet le W de concordance de Kendall[7] pour les huit étudiants-juges du groupe fort est de **0,2362 (significatif à 0,01)** alors qu'il n'est que de **0,1208 (p = 0,348)** pour le groupe faible. Ceci indique que les sujets ayant le moins bien réussi à l'examen sont aussi ceux qui sont le plus en désaccord sur le type de questions qui seront selectionnées par l'enseignant.

Les sujets du groupe fort manifestent également une plus grande concordance en ce qui concerne la préparation. En effet ce groupe a obtenu un **W=0,1946 (significatif à 0,05)** alors que le groupe faible a obtenu un **W=0,1381 (p = 0,190)**. On peut interpréter ce résultat comme **signifiant** que le groupe fort a une plus grande tendance à faire porter les efforts d'étude sur les mêmes items que le groupe faible »

Donner un sens plus spécifique et mieux circonscrit à des notions fréquemment usitées dans la langue commune ou correspondant à des pratiques concrètes, comme celles de risque, de significativité ou de représentativité qui interviennent dans nos prises de décision, en les replaçant dans les domaines de la statistique mathématique et de la statistique appliquée, nous paraît être un des objectifs prioritaires de l'enseignement de la statistique.

Revue Mesure et Evaluation, vol 13.1 pp 54-57

7 Régnier, J.- C. (1996) Test de concordance des rangements de n objets selon p critères., Méthodes quantitatives & statistique , Tome 1 - p 91,Documents polycopiés Université Lyon2 - ISPEF

Chapitre 9

Quelques idées d'exploitation d'un article de presse

Jean Claude Girard

1) Introduction

Il est de plus en plus difficile d'ouvrir un journal sans se trouver face à une multitude de graphiques, tous plus jolis les uns des autres. D'autre part, aucun article se voulant sérieux ne peut faire l'économie d'une accumulation de chiffres à l'appui des thèses qu'il présente.

Malheureusement, on ne sait pas toujours comment ces nombres ont été obtenus ou calculés et les graphiques ne sont pas toujours d'une compréhension aisée ni même en accord avec les chiffres cités dans le texte.

Il convient donc d'attirer l'attention des élèves sur les difficultés d'interprétation des chiffres et des graphiques utilisés en statistiques et de les entraîner à ce genre d'activité.

2) Présentation de l'activité

La suite de ce texte propose quelques idées d'exploitation d'un article extrait du "Figaro-Economie" du 16 août 1995 et qui a pour sujet la presse spécialisée pour les adolescents (voir page 37). Le sujet concerne donc les élèves de collège, même s'ils ne sont pas passionnés par les problèmes de consommation.

L'idée générale consiste à analyser un graphique statistique et sa cohérence avec l'article correspondant. Chacune des idées proposées est transférable à d'autres articles ou graphiques que l'on peut trouver dans la presse ou ailleurs. A chacun de choisir celle qui lui semble la mieux adaptée à son style d'enseignement et à ses élèves !

3) Objectifs

Montrer le rôle du graphique statistique, la différence entre dessin et graphique statistique. Analyser la liaison texte-graphique. Faire apparaître quelques points de la méthode statistique (sondage, population, caractères quantitatifs ou qualitatifs, moyenne et pourcentages). Réfléchir sur l'utilisation des statistiques (pour illustrer, pour prouver) et leur perception par le citoyen moyen.

4) Propositions d'utilisation en classe

Première idée (très ouverte)

Les élèves sont en possession des graphiques seulement.

Question : Que pensez-vous de ces graphiques ? Quelles informations donnent-ils ? Quelle était l'intention de l'auteur de l'article en incluant ces graphiques ? Que voulait-il "prouver" ?

Objectif : faire apparaître les conceptions spontanées des élèves devant des graphiques (ou des dessins). Essayez vous-mêmes honnêtement !

Exploitation : Imprévisible, on peut s'attendre à tout, et surtout au pire ! Vous pourrez trouver quelques éléments de réponse (ou de discussion) à la fin de cet article.

Seconde idée

On donne le graphique et le texte.

Question : Comparer le texte et le graphique. Sont-ils deux façons de dire la même chose ou sont-ils complémentaires l'un de l'autre ?

Objectif : établir la liaison texte-graphique.

Exploitation : faire ressortir les informations du texte qui n'ont pas été illustrées dans les graphiques, les renseignements visibles sur les graphiques et qui n'ont pas été repris dans le texte et les éléments de désaccord entre texte et graphiques.

Troisième idée (pluridisciplinaire)

Un groupe d'élèves a en sa possession les graphiques (seulement). On donne la consigne suivante : écrire un texte développant les idées figurant dans les titres des graphiques et s'appuyant sur les quantités ou pourcentages ayant servi à les construire. Donner un titre à votre texte.

Un deuxième groupe d'élèves est en possession du texte (seulement). On donne la consigne suivante : trouver l'idée principale du texte et l'illustrer par un graphique statistique. Si nécessaire, de quelles informations voudriez-vous disposer pour construire ces graphiques ?

Objectif : faire apparaître les différences de renseignements donnés par le texte et les graphiques.

Exploitation : La lecture du texte est assez difficile à cause d'un vocabulaire spécifique. D'autre part, la construction d'un graphique à partir des seuls éléments figurant dans le texte est bien délicate. Souvent, au contraire, il est plus aisé de construire un graphique à partir d'un texte documenté que de refaire le chemin inverse car beaucoup d'informations sont perdues.

Quatrième idée (pour prolonger)

Chercher des articles dans lesquels les graphiques illustrent le texte sans informations supplémentaires mais sous une autre forme.

Chercher des articles dans lesquels les graphiques complètent le texte en apportant des informations supplémentaires (comme c'est le cas ici).

Chercher des articles dans lesquels le texte complète les graphiques en apportant des informations supplémentaires.

Objectif : Analyser les différentes liaisons possibles entre texte et graphique.

Exploitation : Faire apparaître les différences dans l'utilisation des deux supports pour convaincre ou pour illustrer. Mettre à la discussion le choix entre redondance et complémentarité.

Cinquième idée (plus fermée)

Résumer en quelques lignes, les deux ou trois idées forces du texte. Comment ces idées sont-elles mises en valeur dans les graphiques ?

Trouver des informations dans les graphiques qui ne figurent pas dans le texte ou qui sont en désaccord avec d'autres figurant dans le texte.

De quels types sont les graphiques ?

Quelle est la population qui fait l'objet de l'étude ?

A-t-on interrogé tous les individus de cette population ou une partie seulement ?

Comment a-t-on obtenu la valeur de 368F pour la moyenne mensuelle d'argent de poche? Comment a-t-on obtenu le pourcentage des lecteurs dans la population des 11-19 ans (91 %) ? Comment sait-on qu'il y a 6 868 000 jeunes âgés de 11 à 19 ans ?

Sixième idée

Effacer les commentaires de chaque graphique et les faire retrouver à partir de l'article.

Objectif : montrer la force des graphiques dans l'argumentation.

Exploitation : trouver d'autres exemples (honnêtes ou non) où le graphique sert de "preuve".

Éléments de réponse

- De nombreuses informations qui ont servi à construire les graphiques sont absentes dans le texte par exemple les pourcentages concernant la radio et la télévision pour la question : "Qui dit le plus la vérité ?" que l'on trouve dans le graphique en barres accolées. Texte et graphiques sont donc ici complémentaires.

- Au contraire, on apprend (en renvoi) en lisant le texte que les informations proviennent d'un sondage sur 4 800 jeunes de 8 à 19 ans, alors que le titre général des graphiques parle de jeunes de 11 à 19 ans.

- Il y a donc une incertitude sur la population qui fait l'objet de l'étude et sur les résultats obtenus : concernent-ils les 8-19 ans ou les 11-19 ans ?

- Contradiction également entre le nombre de 11-19 ans en France : 6 700 000 dans le texte et 6 868 000 dans le premier graphique. Il ne peut s'agir d'une erreur d'arrondi.

- Le question "D'après vous, quel est le média (radio, télé, presse) qui dit le plus la vérité ?" porte sur une variable qualitative à trois modalités s'excluant mutuellement. On peut calculer le pourcentages de réponses des individus de l'échantillon pour chacune des modalités. Le total doit être égal à 100%, moins si l'on exclut les sans réponses, mais en aucun cas ne peut dépasser 100 %, ce qui semble le cas dans le graphique quand on ajoute les pourcentages pour chaque question. Pour cette raison, on ne peut pas remplacer chaque groupe de trois barres accolées par un graphique circulaire comme cela devrait pouvoir se faire.

- Les questions "Lisez-vous au moins un magazine pour adolescents?", "Achetez-vous des barres, des chewing-gums ?, "Faites-vous achetez par vos parents des barres, des chewing-gums ?" portent sur une variable qualitative à deux modalités : oui ou non. Le fait que le total des pourcentages dépasse 100, pour un même produit, signifie-t-il que les personnes qui achètent elles-mêmes sont aussi comptées dans celles qui sont impliquées dans l'achat ?

- Avec cet effectif, et si le sondage a été conduit correctement, on peut espérer une précision pour chaque pourcentage de ±1,5% dans 95% des cas (on dit "au niveau de confiance de 95%").

- Les autres questions posées dans le sondage étaient certainement : "De combien d'argent de poche disposez-vous par mois ?", "A combien se montent vos économies ?". Ces deux variables sont des variables quantitatives dont on peut calculer la moyenne pour les individus de l'échantillon (368F pour l'argent de poche et 3 600F pour les économies). Le nombre de jeunes dans cette tranche d'âge étant connu par les statistiques de l'INSEE (qui proviennent elles-mêmes du dernier recensement), on peut alors estimer le pouvoir d'achat mensuel et le montant total des économies pour l'ensemble des adolescents français (respectivement 2,5 Mds et 25 Mds de francs).

- Certains *dessins* ne sont pas des graphiques statistiques (le cochon des économies, la chaîne HI-FI). Leur forme et leur taille n'ont pas d'importance.

- Un de ces dessins illustre la phrase "59% des jeunes ont une chaîne HI-FI personnelle". Ce renseignement est important pour des annonceurs éventuels, car on apprend dans le texte que le choix de la marque est souvent déterminée par les jeunes non seulement pour leurs achats personnels de friandises (ce qui est illustré ailleurs par un autre graphique en barres) mais aussi pour les achats plus importants de la famille, en particulier de matériel audio-visuel.

- En conclusion, l'article du journal présente et commente les résultats d'un sondage qui montre que la population des adolescents possède un pouvoir d'achat non négligeable et qu'elle est dotée d'un réel pouvoir de décision. D'autre part, une proportion importante de cette population lit la presse spécialisée qui lui est destinée et en a une bonne opinion. Bien entendu, le collectif "Presse-ado" qui a commandé le sondage espère ainsi convaincre les annonceurs de consacrer une part plus importante de leur budget publicité au média qu'ils représentent.

La presse « ados » se présente

Mal connue des annonceurs, cette famille de titres veut prouver qu'elle est un média complémentaire de la télévision.

« Revaloriser la presse des 11-19 ans » : c'est le leitmotiv du Collectif presse ados, constitué au printemps dernier, qui réunit dix titres, totalisant une diffusion de 1,3 million d'exemplaires.

Dans un premier temps, Presse Ados a demandé à Diapason/Médiamétrie d'élargir son étude média-enfants, jusqu'alors consacrée aux moins de 16 ans, *« parce qu'il était difficile, en terme de médiaplanning, d'additionner les résultats d'étude de Médiamétrie et de l'AEPM »*, explique Gérard Lefèvre, directeur d'Hachette Filipacchi Régie.

Les résultats de cette étude de référence (1), qui permet de connaître les comportements, les consommations et les pratiques médias de cette tranche d'âge, prouvent que *« cette génération est une cible à part entière pour les annonceurs »*, affirme Gérard Lefèvre.

L'enquête dévoile ainsi que les 6 700 000 jeunes de 11 à 19 ans disposent d'un pouvoir d'achat non négligeable puisque leurs économies s'élèvent, en moyenne, à 3 600 francs et qu'ils reçoivent 368 francs par mois. Dans leur ensemble, ces adolescents possèdent 25 milliards de francs d'économie et reçoivent 2,5 milliards de francs par mois.

« Leur rôle de prescription dans les achats est prépondérant », souligne Gérard Lefèvre. En effet, ces jeunes sont très souvent impliqués dans les choix des marques pour leurs achats de biens personnels mais aussi pour ceux du foyer, qu'il s'agisse d'hygiène-beauté et d'alimentaire comme de chaîne hi-fi, de Caméscope, de magnétoscope, de télévision et même d'automobile.

6 millions de lecteurs

L'étude note que la presse magazine des jeunes est un média puissant avec plus de six millions de lecteurs. Ces derniers représentent une population homogène au regard des catégories socioprofessionnelles : on retrouve une pénétration de 70 à 72 % quelle que soit la profession du chef de famille.

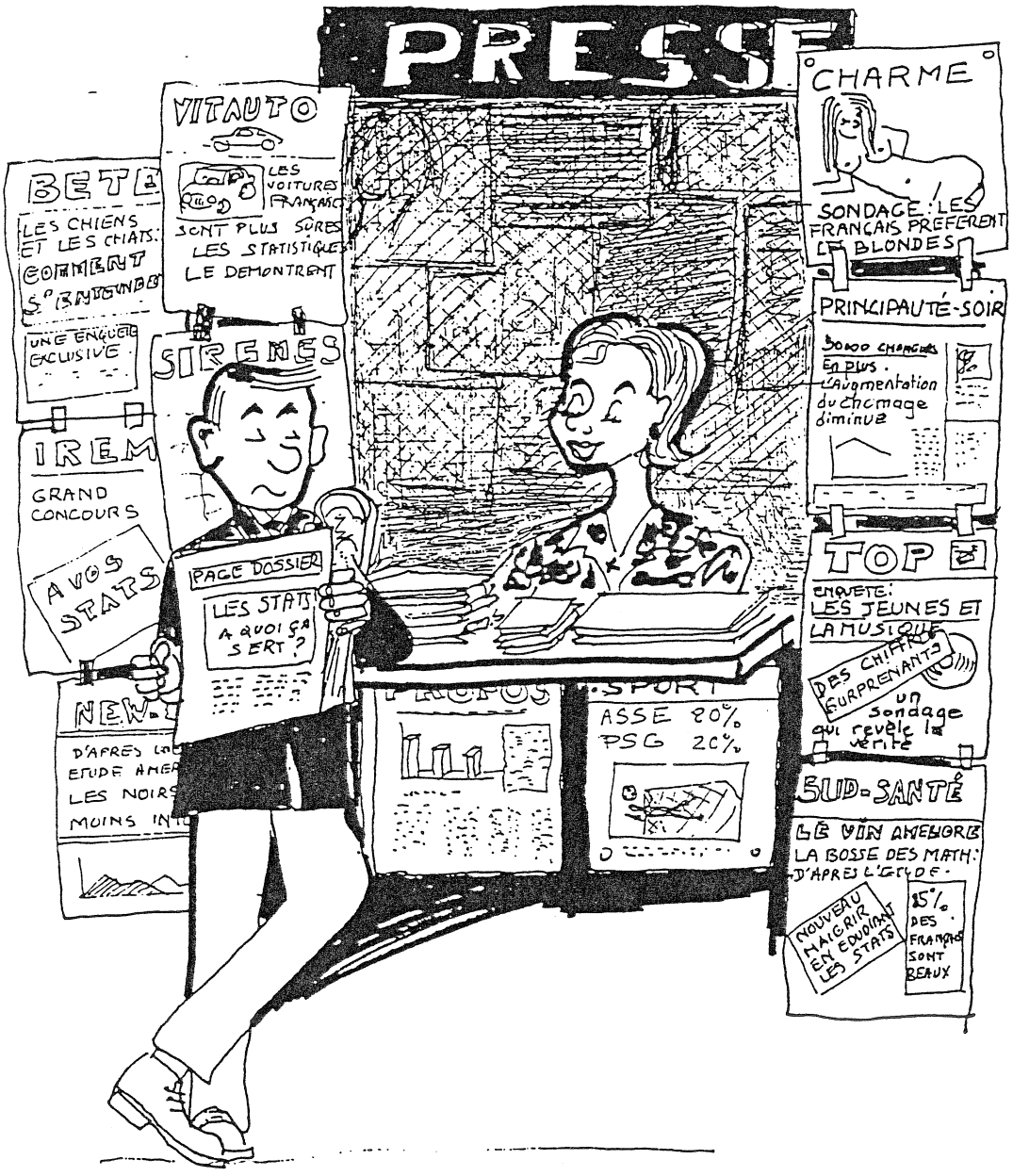
Interrogés par Médiamétrie/Diapason, les adolescents donnent toute leur confiance à la presse magazine : c'est elle qui *« dit le plus la vérité »* (43 %), *« apprend le plus de choses »* (51 %), *« offre les informations les plus complètes »* (58 %). *« Ces jeunes sont complices de leur presse. Ce que nous souhaitons, c'est prouver la complémentarité de la presse magazine par rapport à la télévision et non pas s'y substituer »*, précise Gérard Lefèvre.

Enfin, l'enquête révèle que la presse magazine ados est un média dans lequel la publicité est appréciée : si 48 % des adolescents interrogés *« aiment regarder la publicité à la télévision »*, ils sont également 45 % à *« aimer regarder la publicité dans les magazines »*.

Autant de raisons, pour l'ensemble des régies des 10 titres du Collectif presse ados qui, avec 4,6 millions de lecteurs, ont une pénétration totale de 67 % auprès des 11-19 ans, d'envisager, à terme, la mise en place d'un « multicouplage » afin de vendre collectivement de l'espace publicitaire.

Sophie LATIL

(1) Étude réalisée auprès de 4 800 jeunes de 8 à 19 ans.



Savoir lire : des chiffres et des lettres...

Chapitre 10

Lire un article de journal de la presse ordinaire

L'article du journal OUEST FRANCE (février 1992) : 9617 tués sur les routes en 1991

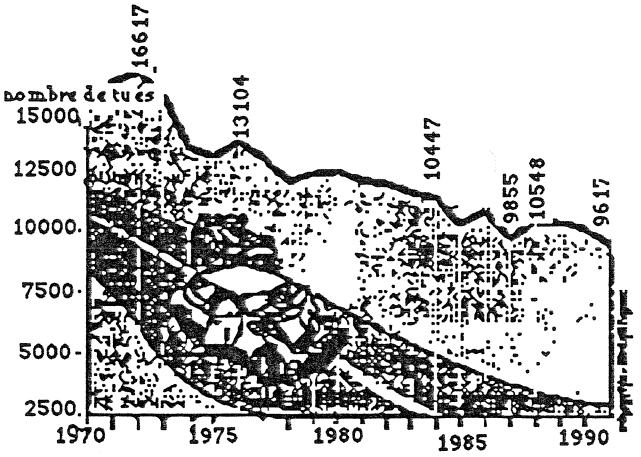
Comment un article de presse régionale que nombre de lecteurs parcourent en diagonale s'avère requérir un traitement parallèle écrit explicitant le modèle mathématique sous-jacent pour parvenir à la compréhension et au contrôle de la validité de ce qui est énoncé.

Jean-Claude Régnier

Le chiffre le plus bas depuis trente et un ans

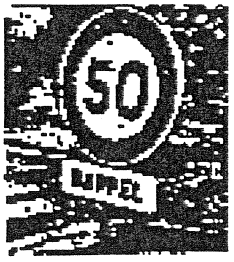
9617 tués sur les routes en 1991

9617 morts sur les routes de France l'an dernier. Une hécatombe ! Et pourtant, le ministère des Transports a de quoi se féliciter de ce chiffre. Jamais depuis 1960 il n'y avait eu si peu de tués. Par rapport à 1990, plus de 600 vies humaines ont été épargnées.



149000 accidents de la route ont provoqué la mort de 9617 personnes en 1991. Un chiffre en baisse de 6.5% par rapport à l'année précédente. Ces accidents ont fait 206 000 blessés (- 8,8 %). Un Martien descendant sur notre planète serait sans doute effaré devant la satisfaction relative affichée par les pouvoirs publics devant ce désastre qui, bon an mal an raye de la carte une ville de la taille de La Ferté-Bernard

Vitesse limitée à 50 km/h en ville : la principale explication au bon résultat 1991, selon le secrétariat d'Etat aux Transports.



16600 morts en 1972

Si l'on oublie les bonheurs perdus, les vies brisées, les victimes à tout jamais handicapées, c'est pourtant vrai que ces statistiques sont encourageantes. Songez qu'en 1972, la courbe ascendante des accidents de la route avait culminé à 16600 morts et près de 400000 blessés. De cette année-à, date la véritable mobilisation contre le massacre du bitume. Trois mesures furent prises coup sur coup, la limitation de vitesse, le port de la ceinture de Sécurité hors-agglomération et le casque obligatoire pour les motards

Un mieux dans l'ouest aussi

Les chiffres pour l'Ouest épousent la courbe encourageante notée au plan national. Le bilan fait état de 1334 tués. 52 de moins qu'en 1990. Les accidents sont en baisse de 11,70%, les tués de 3,75% et les blessés de 12,2%. Les services de police et de gendarmerie ont constaté 4753 accidents en Bretagne (504 morts.et 6120 blesses). 3021 accidents en Basse-Normandie (274 morts et 4181 blesses) et 6129 accidents en Pays de Loire (556 tués et 8539 blessés)

Voici un article de journal destiné à tous les publics. Certes une lecture en diagonale demeure toujours possible mais :

- quelles informations le lecteur en tire-t-il ?
- comment contrôle-t-il sa compréhension du texte ?
- à quelles activités se livre-t-il pour procéder à ce contrôle ?
- comment contrôle-t-il la validité de ce qui est énoncé ?

Tant de questions qu'en particulier un enseignant peut être amené à se poser non seulement pour lui-même, mais aussi en relation avec la part qu'il prend dans le développement de la capacité à lire des individus. Il se trouve que ce texte fait aussi référence au domaine de la statistique. Plus en profondeur, le contrôle de la validité de l'information que le lecteur peut retirer de la lecture de cet article, va requérir des connaissances déclaratives et des connaissances procédurales rattachées au domaine des mathématiques. Pour contrôler cette compréhension, nous suggérons au lecteur de compléter les tableaux ci-dessous. Les informations contenues dans les cases sont des données numériques quantitatives.

année 1991	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	a	a	a	b	c	c
Basse-Normandie	a	a	a	b	c	c
Pays de Loire	a	a	a	b	c	c
Ouest	b	b	a	b	c	c
France	a	a	a	b	c	c

	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	g	g	g-e	b	c	c
Ouest 1991	d	d	a	d	c	c
variation 91/90	a	a	a	f		
France 1990	g	g	e	b	c	c
France 1991	a	a	a	b	c	c
variation 91/90	a	a	f	f		

(a) ces données sont accessibles directement dans le texte. Le recours au surlignage, sorte de pratique écrite de la lecture d'un texte, permet de mettre en relief cette information.

(b) ces données sont indirectement accessibles par l'intermédiaire des données explicites. Elles résultent de ces dernières par le recours à une procédure additive qui met en jeu l'addition ou l'opération réciproque, la soustraction. Si nous procédons à un traitement des colonnes : le nombre de victimes est la somme du nombre de tués et du nombre de blessés. Si nous procédons à un traitement des lignes : la zone géographique OUEST réunit les trois régions Bretagne, Basse-Normandie et Pays de Loire. Il en résulte en vertu de la propriété selon laquelle le cardinal - c'est à dire le nombre d'éléments, l'effectif - de la réunion de trois ensembles disjoints deux à deux est égal à la somme des cardinaux de chacun des ensembles que les effectifs recherchés concernant l'OUEST s'obtiennent par addition des effectifs relatifs à chacune des trois régions.

(c) Ceci ne constitue plus du tout une information accessible même indirectement dans le texte. Il s'agit d'un traitement choisi par le lecteur pour contrôler la validité de l'information que l'auteur de l'article souhaite faire passer qui pourrait se résumer à cette phrase : **la situation semble s'améliorer !**

Pour ce faire, nous recourons à la notion de "quantité relative" qui s'appuie sur une procédure multiplicative qui met en jeu multiplication ou son opération réciproque, la division. Dans un langage formalisé, nous pouvons coder cette procédure dans le registre algébrique, de la façon suivante : $Q = 100 \frac{X}{Y}$. Nous obtenons de cette façon un indicateur quantitatif rendant comparables les données entre les diverses régions ou zones géographiques. Bien qu'il se rattache concrètement à un nombre entier, le nombre de tués comptabilisés sur 100 accidents, ce nombre peut être aussi considéré comme un nombre réel et être éventuellement fourni par une approximation décimale au dixième ou au centième près. Il convient de le considérer comme une construction mentale humaine visant à rendre compte d'une certaine réalité et non comme un objet ayant une existence propre dans la nature. Même un platonicien convaincu ne serait sans doute pas choqué par un telle affirmation.

(d) Ces données sont lisibles dans le premier tableau.

(e) L'obtention de cette donnée s'appuie sur une procédure additive. La phrase du texte *"Par rapport à 1990, plus de 600 vies...épargnées"* indique la procédure ainsi formalisée :

$X_{1990} - Y = X_{1991}$ et en nous plaçant dans l'hypothèse minimale où Y = 600 cela donne $X_{1990} - 600 = X_{1991}$. Le résultat recherché est alors $X_{1990} = X_{1991} + 600$

(f)(g) Il s'agit sans doute là de la procédure mathématique la plus complexe et subtile impliquée dans le traitement requis pour comprendre le texte. Nous la formalisons ainsi :

$$Q = \frac{X_{1991} - X_{1990}}{X_{1990}}$$

Le recours à une représentation du registre algébrique offre des possibilités de traitement particulièrement pertinentes pour la résolution de notre problème.Ainsi nous pouvons tour à tour exprimer chacune des trois variables de la façon suivante, ces transformations étant permises par des règles algébriques que nous ne développons pas ici :

$$X_{1991} = (1+Q)X_{1990}$$

$$X_{1990} = \frac{X_{1991}}{1+Q}$$

L'application adéquate de ces formules permet d'obtenir les résultats cherchés .

année 1991	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	4753	6120	504	6624	10,60	139,36
Basse-Normandie	3021	4181	274	4455	9,06	147,46
Pays de Loire	6129	8539	556	9095	9,07	148,39
Ouest	13903	18840	1334	20174	9,59	145,10
France	149000	206000	9617	215617	6,45	144,70

	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	15745	21457	1386	22843	8,80	145,08
Ouest 1991	13903	18840	1334	20174	9,59	145,10
variation 91/90	-11,7%	-12,20%	-3,75%	-11,68%		
France 1990	159358	225877	10217	236094	6,41	148,15
France 1991	149000	206000	9617	215617	6,45	144,70
variation 91/90	-6,5%	-8,8%	-5,8%	-8,67%		

Il est encore possible d'exploiter un peu loin cet article. Mais cette fois cela implique la mise en œuvre de la lecture d'un graphique, seule information exprimée dans un registre sémiotique iconique exploitable. En effet l'image du panneau de signalisation n'est là que pour renforcer l'information relative à la mesure de limitation de vitesse. Ce graphique débarrassé de l'image des automobiles accidentées est une représentation graphique du domaine mathématique. Il s'agit de la représentation graphique d'une variable chronologique, celle qui fournit le nombre annuel de tués. C'est donc une fonction qui à la variable "temps" associe la variable "nombre de tués". L'axe temporel est l'axe des abscisses. L'axe des ordonnées fournit le nombre de tués. Nous pouvons alors observer une

tendance, le nombre de tués a tendance à être de plus en faible, ou de moins en moins fort, le nombre de tués a tendance à diminuer globalement. Une autre façon de traiter cette information est de réaliser une conversion de ce graphique, exprimé dans un registre géométrico-iconique, en un tableau, registre iconique. Ce que nous pouvons repérer immédiatement, ce sont les points pour lesquels la valeur de l'ordonnée est rappelée. Certes l'attention du lecteur est requise pour éviter de commettre une erreur sur la détermination de la valeur de l'abscisse, c'est à dire la date. Nous pouvons alors compléter le tableau suivant :

année	1972	1977	1985	1987	1988	1991
nombre de tués	16617	13104	10447	9855	10548	9617

Pour constituer un tableau rapportant les valeurs relativement aux 22 années de 1970 à 1991, il y a nécessité de recourir à une règle graduée et à une règle de proportionnalité pour obtenir les valeurs. Cette procédure paraît coûteuse pour le lecteur. Le texte nous rapporte aussi une valeur approximative du nombre des victimes en 1972 : 400000. Ce nombre peut alors être comparé à celui du nombre des victimes en 1991 : 215617. Une procédure du type (f) permet d'obtenir la réponse :

$$Q = \frac{X_{1991} - X_{1972}}{X_{1972}} = \frac{215617-400000}{400000} = -0{,}4609575$$

d'où une chute de 46,09% du nombre des victimes entre 1972 et 1991.

Il s'est effectivement passé quelque chose !

Examinons maintenant la variation entre 1990 et 1991. Les valeurs absolues sont effectivement en baisse, ce qui est un acquis incontestable puisqu'il y a moins d'accidents, moins de victimes et moins de tués.

Nous pouvons aussi regarder les effets des accidents. Le facteur Q dont le mode de calcul est fourni en (f)(g), nous apporte une information tout à fait intéressante :

année 1991	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	10,60	139,36
Basse-Normandie	9,06	147,46
Pays de Loire	9,07	148,39
Ouest	9,59	145,10
France	6,45	144,70

années 1990 1991	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	8,80	145,08
Ouest 1991	9,59	145,10
France 1990	6,41	148,15
France 1991	6,45	144,70

Pour accroître les contrastes, réalisons une sorte de zoom en rapportant les nombres des victimes et des tués à 1000 accidents. Cette démarche est recevable puisque le nombre des accidents est de plusieurs milliers.

année 1991	nombre de tués pour 1000 accidents	nombre de victimes pour 1000 accidents	années 1990 1991	nombre de tués pour 1000 accidents	nombre de victimes pour 1000 accidents
Bretagne	106,0	1393,6	Ouest 1990	88,0	1450,8
Basse-Normandie	90,6	1474,6	Ouest 1991	95,9	1451,0
Pays de Loire	90,7	1483,9	France 1990	64,1	1481,5
Ouest	95,9	1451,0	France 1991	64,5	1447,0
France	64,5	1447,0			

Nous sommes alors frappé par le fait le taux des victimes par accident est le plus faible en Bretagne mais que ce gain est cruellement compensé par un perte sur le taux de mortalité par accident. 1000 accidents provoquent la mort de 16 personnes de plus en Bretagne qu'en Pays de Loire ou en Basse-Normandie.

Si maintenant nous nous intéressons à la variation entre 1990 et 1991, force est de constater qu'en 1991 sur la France, il y eut 4 victimes de moins par 1000 accidents, ce qui constitue un gain. Mais nous voyons que le taux de mortalité reste inchangé 64 tués pour 1000 accidents. Ainsi les accidents sont-ils toujours aussi mortels mais ils sont moins "blessants". En revanche, en OUEST si le nombre des victimes pour 1000 accidents restent le même :1451, le nombre de tués s'est accru de 7 personnes par 1000 accidents.

En OUEST les accidents tuent davantage en 1991 qu'en 1990 !

Cette conclusion n'était nullement explicite dans le texte. Il y a eu nécessité d'un traitement spécifique prenant appui sur des notions et des méthodes statistiques pour faire surgir cette information de celles que l'article nous apportaient. Une lecture en diagonale ne paraît pas conduire à une telle conclusion qui nuance quelque peu celle qui est donnée dans l'article : **Un mieux dans l'OUEST aussi !**

Une autre façon de comprendre la phrase « Un chiffre en baisse de 6,5% par rapport à l'année précédente » est de considérer que ce taux porte sur le nombre des personnes tuées. Ceci revient à prendre 6,5% de 9617, soit 625 personnes. Cette information est confirmée par ce qui est annoncé en chapeau : « Par rapport à 1990, plus de 600 vies humaines ont été épargnées ». Une telle perspective conduit aux résultats suivants :

	nombre d'accidents	nombre de blessés	nombre de tués	nombre de victimes	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	15745	21457	1386	22843	8,80	145,08
Ouest 1991	13903	18840	1334	20174	9,59	145,10
variation 91/90	-11,7%	-12,20%	-3,75%	-11,68%		
France 1990	?	225877	10286	236163	?	?
France 1991	149000	206000	9617	215617	6,45	144,70
variation 91/90	?	-8,8%	-6,5%	-8,7%		

année 1991	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Bretagne	10,60	139,36
Basse-Normandie	9,06	147,46
Pays de Loire	9,07	148,39
Ouest	9,59	145,10
France	6,45	144,70

années 1990 1991	nombre de tués pour 100 accidents	nombre de victimes pour 100 accidents
Ouest 1990	8,80	145,08
Ouest 1991	9,59	145,10
France 1990	?	?
France 1991	6,45	144,70

Dans ce cas nous perdons l'information relative à la France entière pour l'année 1990. Ceci ne modifie pas les conclusions que nous avons pu tirer sur la région Ouest ni celles relatives aux traitements que nécessite la compréhension d'un tel article de presse.

ECOLE MIXTE
...et en plus, i' bougent !

Compte rendu d'expérimentation sur les graphiques

Daniel Gros et Paul Planchette

Objectifs :

Le projet initial visait à étudier les conceptions spontanées des grandeurs chez l'enfant entre 8 et 10 ans d'une part (primaire), et entre 11 et 15 ans d'autre part (collège). Il posait aussi la question des évolutions de ces conceptions et de leur construction lors des apprentissages, ainsi que celle de la contribution et des aides que l'on pourrait y apporter.

Déroulement :

Plusieurs activités ont été proposées, le plus difficile étant de formuler la demande afin d'obtenir une représentation spontanée... Ont été "mesurés" :

- des représentations spontanées de quantités, et leur évolution dans une situation didactique faisant augmenter la quantité à représenter,
- la reconnaissance de "l'écrit graphique" parmi d'autres (primaire uniquement),
- la lecture de valeurs représentées sur un graphique barre
- la construction de graphiques "barres"
- la construction de graphiques "secteurs" (camembert)

Bilan provisoire (8-10 ans) :

Très succinctement, on peut conclure des expérimentations engagées que la notion de représentation des nombres est présente chez les élèves à des niveaux très divers, que c'est une question qui les embarrasse et dont l'évolution est lente.

Les points suivant semblent ne pas poser de problème majeurs :

- reconnaissance d'un graphique.
- lecture d'une valeur seule.

Les difficultés rencontrées se situent particulièrement dans :

- le fait que des termes sont connus, mais ne correspondent à aucune réalité.
- dès que la lecture fait intervenir 2 ou plusieurs termes, elle pose le problème de la corrélation et de la compréhension globale.
- le recours à une référence d'aire n'est pratiquement jamais présent.

Perspectives pour l'enseignement :

Il semble que le décalage de l'adéquation entre *ce qui est connu par l'élève* dans ce domaine et *ce qui est attendu par les enseignants* mérite une attention particulière.

Une situation d'expérimentation dans laquelle les variants étaient prévus pour induire un résultat attendu s'est révélée plus didactique que représentative d'un état de l'élève. Cette voie semble intéressante à développer pour mettre en place des situations d'apprentissage.

A 10 ans environ, lecture d'un graphique barre (simple)

90 % des enfants reconnaissent un graphique barre parmi des écrits divers.

70% des enfants savent lire les valeurs sur un graphique barre.

50 % des enfants trouvent un intérêt à la présentation par graphique.

10 % des enfants font une lecture en valeurs relatives (comparaison, déduction).

Représentation de nombres

70 % des enfants sont capables de représenter un nombre, soit :

- 45% des enfants font une représentation réaliste.
- 25 % des enfants font une représentation symbolique.

Si cette représentation est remise en cause d'une manière quelconque, (par exemple si les quantités sont très grandes) son évolution est lente :

25 % des enfants régressent. (Ils deviennent incapables de représenter)

50% des enfants n'évoluent pas.

25 % des enfants progressent. (Ils trouvent "une solution")

Représentation d'un pourcentage

100 % des enfants connaissent le terme "pourcent".

40 % des enfants font une représentation recevable (logique proportionnelle).

Construire un graphique barre :

30 % des enfants construisent un graphique proportionnellement correct.

65% des enfants ne s'occupent que de la hauteur des barres (largeur variable).

40 % des enfants oublient la légende.

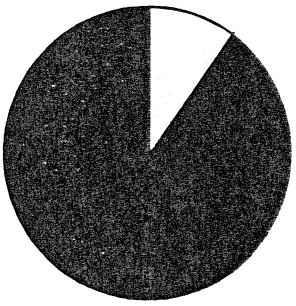
5% des enfants indiquent une échelle des valeurs.

Construire un graphique camembert :

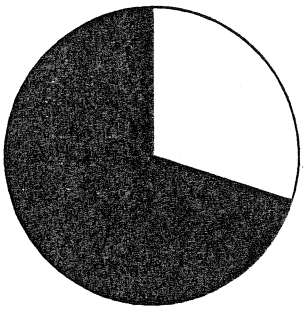
55 % des enfants construisent un graphique proportionnellement correct.

15 % des enfants oublient la légende.

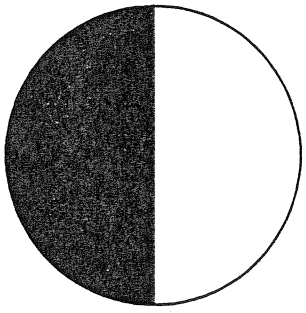
Lecture d'un graphique barre (10 ans)



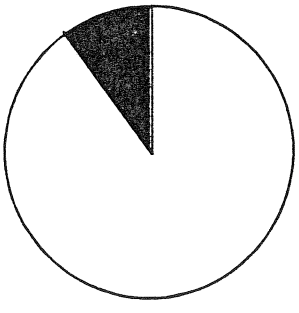
Reconnaissent un graphique parmi des écrits divers.



Savent lire les valeurs sur un graphique barre.

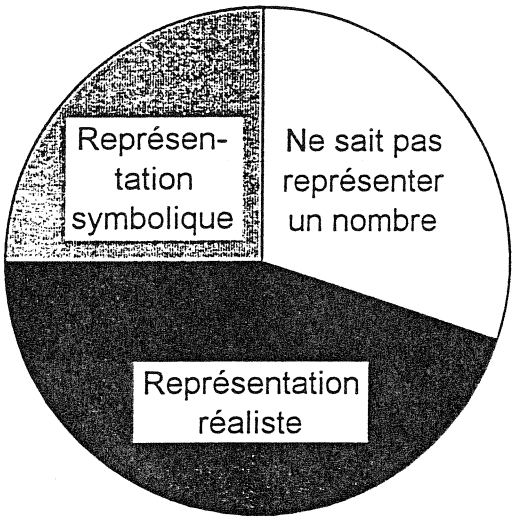


Trouvent un intérêt de la présentation par graphique.

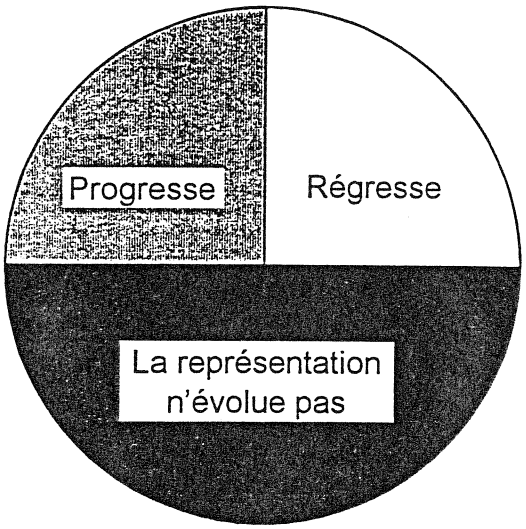


Lisent les valeurs les unes par rapport aux autres.

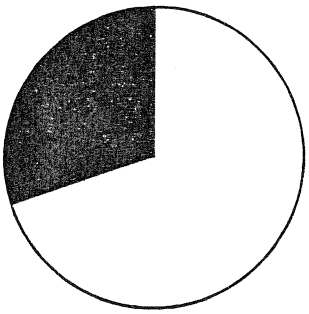
Représentation d'un nombre



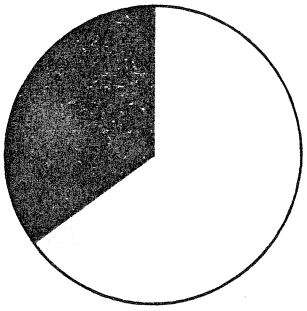
Evolution des représentations si on augmente les quantités à représenter (au cours du CM)



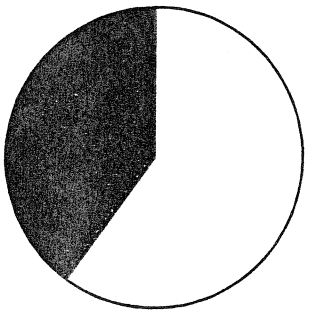
Construction d'un graphique barres



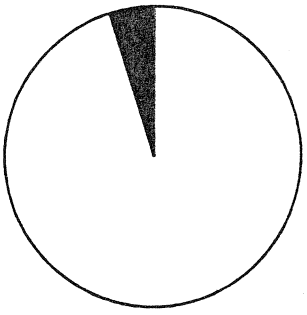
Construisent un graphique proportionnellement correct



Font des barres de même largeur

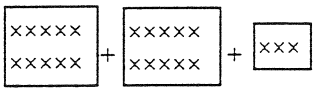
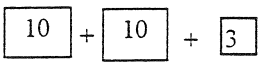
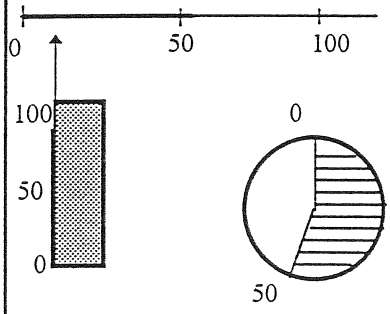


Oublient de mettre une légende



Indiquent une échelle des valeurs

Représentation des nombres

	6 ième	3 ième
	Sur 36 élèves testés, un seul n'a pas donné de réponse cohérente.	En troisième tous les élèves donnent une représentation d'un nombre.
 ou xxxx ... ——— l'élève essaie de dessiner tous les objets	50 % des enfants utilisent cette représentation très réaliste. Quand elle est mise en échec par des nombres plus grands beaucoup passent à la représentation ci-dessous. Cependant 2 continuent à dessiner tous les objets.	30% des enfants utilisent spontanément cette représentation.
 ou  On se trouve en présence de regroupements qui suggèrent la numération décimale.	28% des enfants utilisent ce mode de représentation pour le nombre 53. Ils sont 42% pour 199 et 505	35% des enfants utilisent ce mode de représentation pour le nombre 53. La proportion n'est plus que de 25% pour 505
 Des représentations faisant référence à une échelle	5 % utilisent la droite graduée , la proportion triple avec les grands nombres. Les autres formes de représentations n'apparaissent pas.	35% des enfants utilisent ce type de représentations.

Construire un graphique barre

6ième	3ième
80% construisent un graphique proportionnellement correct (pour les deux-tiers de ces élèves, l'échelle est mal choisie). Les autres graphiques sont incomplets, faux ou n'ont pas d'échelle.	Tous les élèves de troisième ont réussi leur graphique en barre.

Construire un graphique camembert

6ième	3ième
Aucun graphique n'est complètement réussi. 12% ont respecté les proportions dans un secteur circulaire en employant des unités.	60% ont construit un graphique juste. 17% ont commis des erreurs dans le calcul des pourcentages. Les autres ont fait un graphique faux mais «apparemment satisfaisant» (au hasard pour remplir le contrat).

Conclusion

Ce qui est le plus frappant c'est une importante évolution entre la sixième et la troisième.

En sixième: Le comptage est privilégié et il y a une réelle difficulté pour représenter un nombre par une surface. Il ne faut pas s'en étonner car la proportionnalité, les conservations et les opérations mentales sur des surfaces, le concept d'aire, les échelles et les changements d'unités ne sont pas maîtrisés, loin s'en faut.

L'apprentissage de la lecture, de l'interprétation et de la construction des graphiques doit se faire (c'est mon point de vue) parallèlement à d'autres apprentissages dont ils sont parties prenantes (proportionnalité, calcul et comparaison d'aires , etc...).

Ce travail peut être conduit tout au long du collège. Il n'est pas d'ailleurs à la seule charge du professeur de mathématique (histoire-géographie, sciences, technologie, sport...).

En troisième: Les élèves ont une lecture assez rapide et exacte de graphiques simples. La construction de ces derniers pose encore des difficultés à certains parce qu'ils utilisent mal la proportionnalité et les outils qui s'y rattachent (échelles, pourcentages , mesure d'angles).

Une enquête statistique au service de la proportionnalité

ou

"comment ne pas mettre la charrue avant les boeufs"...

Daniel Gros

Schéma habituellement rencontré :

*La proportionnalité étant une notion complexe, on pense généralement qu'elle n'est pas abordable sans, **en préalable**, un minimum de maîtrise de certains outils mathématiques (multiplication, division, voire règle de trois, produit en croix...)*

*On a donc tendance à donner d'abord aux élèves **les moyens de calculer**, les recettes de cuisine mathématique ("prendre la totalité, diviser par la partie concernée, multiplier par 100...") en espérant que "l'idée" de proportionnalité naîtra de l'observation des résultats et que la compréhension suivra, ce qui est d'ailleurs le cas pour une partie des élèves (généralement les "bons", c'est-à-dire ceux qui savaient déjà...).*

Quant aux autres, ils deviennent suffisamment "experts" en manipulations mathématiques pour s'en sortir et n'éprouvent pas le besoin d'en comprendre plus, renforçant même souvent leur première idée des mathématiques : "L'important, c'est que le résultat soit juste, et j'ai eu de la chance, c'est tombé bon..."

Schéma proposant une alternative au précédent...

On peut essayer de substituer au schéma habituel (qui fait passer la notion de proportionnalité au second plan au profit de calculs obscurs et dénués de tout sens pour certains élèves) un schéma dans lequel **la proportionnalité** ne sera pas occultée par les calculs et **restera au premier plan** de nos préoccupations.

Pour qu'une telle activité puisse trouver sa place en classe, il faut :

1) Motiver une recherche, créer un besoin dans un domaine bien ciblé.

2) Installer une situation riche avec beaucoup d'informations chiffrées.

3) Utiliser des outils permettant de "visualiser", de "mettre en situation" la proportionnalité des grandeurs sans faire de calculs compliqués.

L'objectif est donc d'installer **d'abord** la notion en réponse à un besoin, et de se donner les moyens d'arriver à un résultat juste (ou approché) sans calculs.

Les 3 points ci-dessus, en détail :

1) Motiver une recherche, créer un besoin dans un domaine bien ciblé :

La **statistique** trouvera une place de choix dans cet enseignement :

Elaboration d'une vraie enquête.

Il s'agira d'installer un **travail transdisciplinaire** complet, sur un sujet au centre des préoccupations des élèves : à l'école élémentaire, ce genre de situation se met en place sans difficulté, l'enseignant étant seul et polyvalent. Au collège, ce sera l'occasion de faire collaborer 2 enseignants (math / histoire, ou math /français...)

*L'idée est de mener une **enquête statistique** sur un thème motivant.* (Exemples conduits en classe avant ou pendant cette étude : "Qui sont les fumeurs ?", "De quoi avons-nous peur ?" et récemment, "Enquête sur l'environnement à Montreynaud.")

Dès que le thème est retenu, (l'expérience montre qu'il y a toujours une opportunité à saisir, dans la vie quotidienne, une remarque d'élève, l'actualité...) il faut mettre au point avec le plus grand soin un questionnaire et le diffuser le plus largement possible. (Il s'agit là d'un chantier d'écriture très intéressant pour lequel on peut trouver de nombreux "modèles vrais".)

Diffusion d'un questionnaire

La diffusion du questionnaire ne pose pas de problèmes majeurs. Plusieurs solutions sont envisageables : interne à l'établissement, inter-établissements, à l'extérieur par questionnement direct... Ce peut être l'occasion d'échanges et de communications souvent inédits.

2) Installer une situation riche avec beaucoup d'informations chiffrées.

Dépouillement

Dès que les questionnaires remplis sont récupérés, le dépouillement peut commencer : méthodes de comptages et de vérification, classements par catégories, tris divers, sont autant de situations à exploiter.

En organisant bien la classe, le nombre d'élèves devient un atout pour se partager les tâches, rassembler et classer les informations.

Problèmes posés par la masse d'informations

Les réponses au questionnaire seront collectées dans un grand tableau et constitueront une **masse d'informations** dont il sera facile de voir qu'elle n'est **pas lisible** dans sa globalité **sans un traitement particulier :**

Le rapport entre les nombres y est peu évident, l'ensemble est à peine lisible par ceux qui l'ont construit, et y rechercher un enseignement relève de la gageure. Le problème de la lisibilité est posé.

3) Utiliser des outils permettant de "visualiser", de "mettre en situation" la proportionnalité des grandeurs sans faire de calculs compliqués.

Représentation et mise en forme des réponses à l'enquête

Il reste à rendre les résultats **lisibles par tous et communicables**. (Le support de communication pourra être varié selon le cas : feuille individuelle, affichage, journal scolaire, support multimédia...)

En préalable, il est nécessaire d'avoir abordé la représentation graphique des nombres avec les jeunes élèves (élémentaire des cycles II et III) afin d'établir une plate-forme commune autour de cette notion et d'avoir des références similaires pour représenter les grandeurs.

D'autre part, l'observation et l'étude approfondie de quelques enquêtes dans divers magazines ou journaux permettent de savoir ce vers quoi on veut tendre : une représentation graphique et des nombres ramenés à 100, grandeur de référence habituelle, conventionnelle...

La proportionnalité sans calcul.

Le rapporteur à pourcentages

Et c'est là que sont nécessaires quelques outils spécifiques :

Dans un premier temps, on a appris à représenter les **valeurs réelles** sur un "camembert" (graphique circulaire), ce qui ne demande comme pré-requis que la connaissance de l'ordre des nombres et de leur répartition en intervalles réguliers.

La **lecture en pourcentage** sera directe car on dispose d'un "rapporteur à pourcentages" qui permet à l'élève d'obtenir les valeurs approchées en **pourcentage** à la place des valeurs réelles par simple lecture sur son dessin. L'élève n'aura qu'à vérifier la complémentarité à 100 des résultats. *(cf fiche-outil n°1)*

L'ordinateur

De même, l'utilisation de l'informatique permet une autre approche que celles habituellement retenues (utilisation d'un tableur-grapheur) :

- l'ordinateur permet le **tâtonnement**. *(cf fiche-outil n°2)*
- l'ordinateur permet la **validation** des résultats obtenus. *(cf fiche-outil n°3)*
- l'ordinateur permet de trouver une valeur plus fine.

Pour les données qui nécessitent d'autres représentations (au-delà de 5 ou 6 valeurs, le graphique circulaire n'est plus adapté, on lui préfèrera le graphique barres par exemple), il est possible en plus d'ordonner les valeurs : le passage direct au pourcentage se fera ici à l'aide d'une "règle" graduée jusqu'à 100 que l'on utilise en parallèle avec une représentation de même longueur (même principe que le rapporteur à pourcentages).

En guise de conclusion :

Un travail de mathématique ancré dans la vie.

Les statistiques abordées ainsi, tout en intéressant les élèves de façon vivante, permettent de nombreuses ouvertures :

- On complète les exercices de lecture de graphiques (en général les seuls pratiqués) par des exercices de constructions de graphiques.
- On appréhende les notions d'ordre de grandeur et de proportionnalité dans leur essence, on les place dans un contexte de réalité.

- La correspondance avec les calculs se fera quand même, mais alors les élèves seront demandeurs (par exemple pour arriver à des résultats plus précis) et la passerelle ira de la compréhension vers la théorisation. L'élève avancera sur un terrain déjà balisé.

Un travail qu'on analyse

Il faut ajouter que le travail final de présentation et de communication est lui aussi très riche : éliminer ce qui est redondant, regrouper ce qui va dans le même sens (ou ce qui s'oppose !), aller à l'essentiel, classer les informations, les hiérarchiser, écrire un commentaire pertinent...

Ce regard en retour sur les représentations permet de se décentrer, d'observer les résultats des autres groupes, de les discuter...

Un travail valorisant

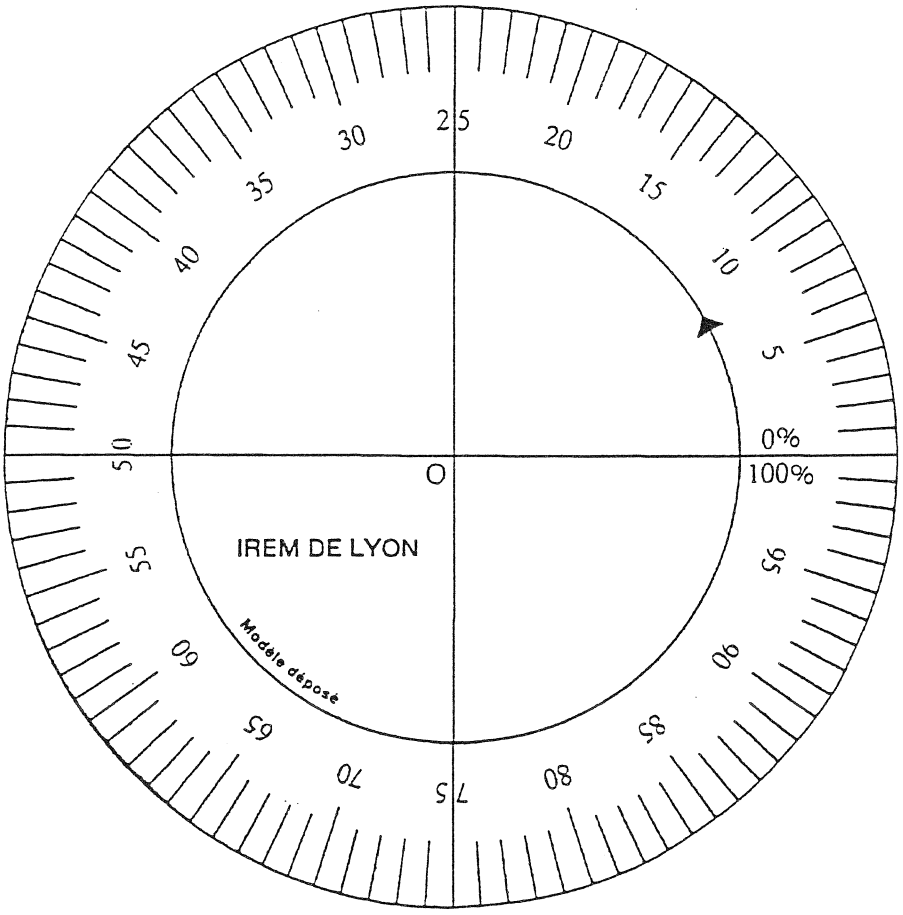
Ce genre d'enquête n'est pas une simulation : elle permet vraiment de mieux connaître le sujet qu'on a abordé.

A titre d'exemple, l'enquête sur l'environnement menée en juin 1996 à Montreynaud (une ZEP) par une classe de CM1 a été diffusée dans le journal du quartier "Echos-liés", certaines écoles s'en sont inspirées pour l'élaboration de leur projet d'école et un cabinet d'expert la reprend actuellement pour en affiner l'analyse et communiquera les résultats de cette étude professionnelle aux personnes concernées : c'est un travail reconnu non seulement par l'institution scolaire, mais aussi par "l'extérieur".

Annexe à la Fiche outil n°1

Le **rapporteur à 100%** utilisé par les élèves pour faire la **lecture directe** d'un pourcentage après avoir préparé leurs valeurs réelles sur un cercle.

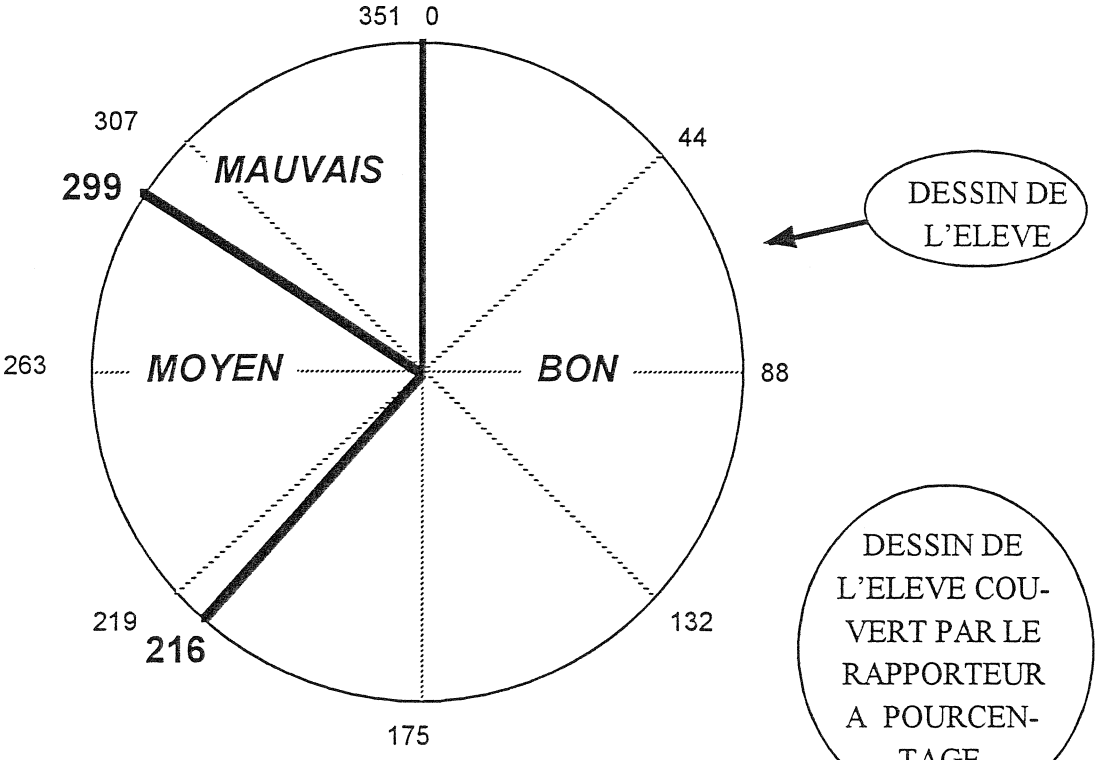
Ainsi, aucun calcul ne vient occulter la notion de proportionnalité qui est ici l'objectif visé.



Fiche outil n°1

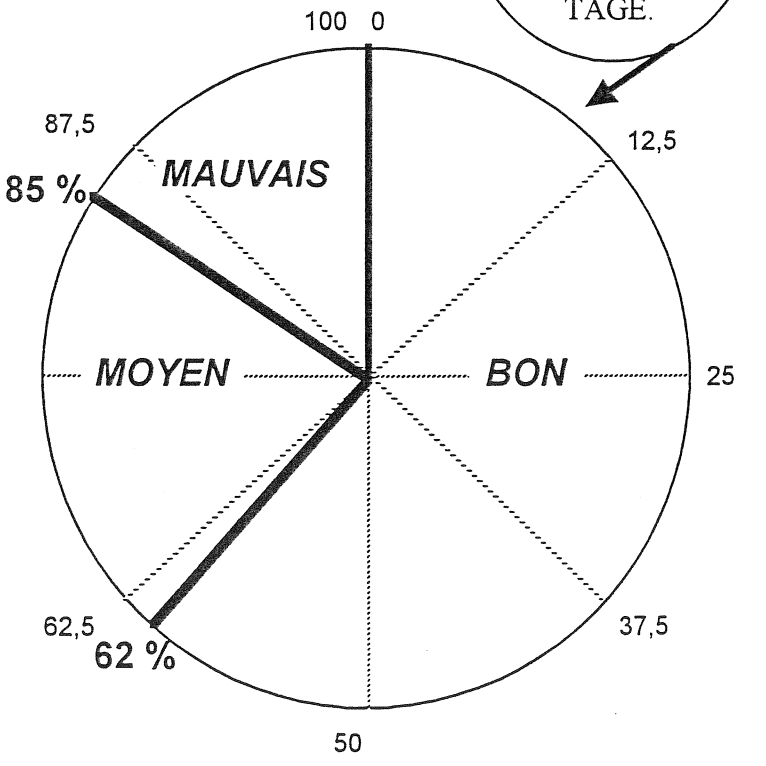
Un outil au service de la proportionnalité : le rapporteur à pourcentages

Dans un premier temps, l'élève apprend à représenter les quantités réelles sur un cercle.
Exemple : pour représenter les valeurs correspondant aux réponses suivantes :
L'état de ma boîte aux lettres est : BON : 216 MOYEN : 83 MAUVAIS : 52
L'élève calcule la somme, puis gradue le cercle (par moitié, quart...)



Ensuite, il suffit de poser le rapporteur à pourcentages bien centré sur le dessin pour avoir en lecture directe les valeurs en %.

Aucun calcul ne vient freiner la correspondance au détriment de la véritable notion visée ici : la proportionnalité.
D'un simple geste, on ramène n'importe quelles valeurs à la valeur référence 100.



Fiche outil n°2 Un outil au service de la proportionnalité : l'ordinateur avec un tableur / grapheur permet le tâtonnement.

Consigne : **Modifie les nombres rouges** jusqu'à ce que :
- Le total (bleu) soit égal à 100.
- Le graphique de droite soit comme le modèle (de gauche).

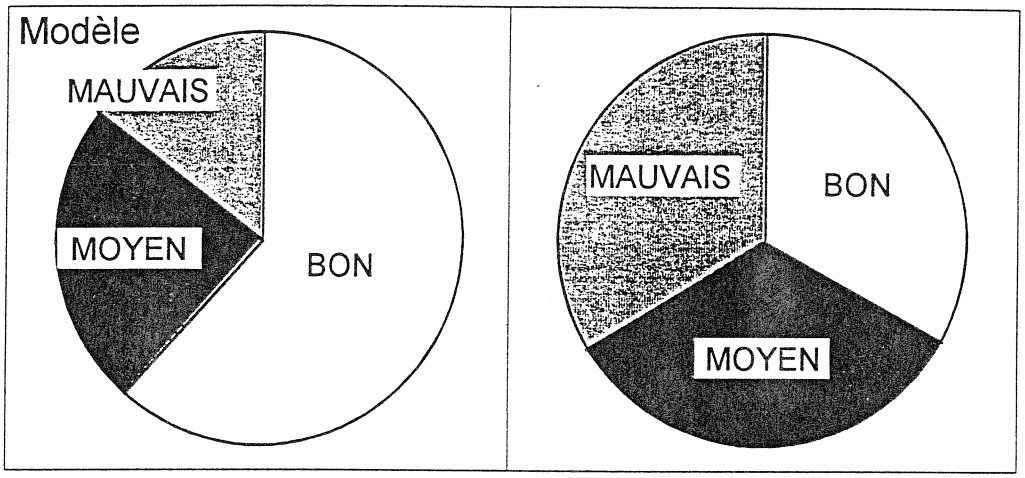
Les totaux sont calculés automatiquement

Quel est l'état de ta boîte aux lettres ?		
BON	MOYEN	MAUVAIS
216	**83**	**52**
TOTAL	351	

Quel est l'état de ta boîte aux lettres ?		
BON	MOYEN	MAUVAIS
33	**33**	**34**
TOTAL	100	

Ici, les valeurs vraies

Ici, essaie de trouver les valeurs en pourcentage



L'élève travaille à l'écran.
Il peut imprimer son travail quand il a fini.

Tout ce qui ne doit pas être modifié par l'élève est verrouillé de façon à préserver les éléments fixes et aussi à limiter les difficultés d'ordre informatique.
Il y a des couleurs à l'écran (bleu, rouge...) pour faciliter les repérages.

***Tout d'abord, l'élève saisit les valeurs vraies** (à gauche)*
Aussitôt, le camembert de gauche prend les proportions correspondantes.

***Ensuite, l'élève va essayer par tâtonnement de trouver les valeurs** à droite telles que :*
- le total soit égal à 100 (il est calculé en permanence par la machine)
- et le camembert de droite soit identique à celui de gauche (modèle)

L'élève juge approximativement de l'identité des 2 camemberts

Fiche outil n°3

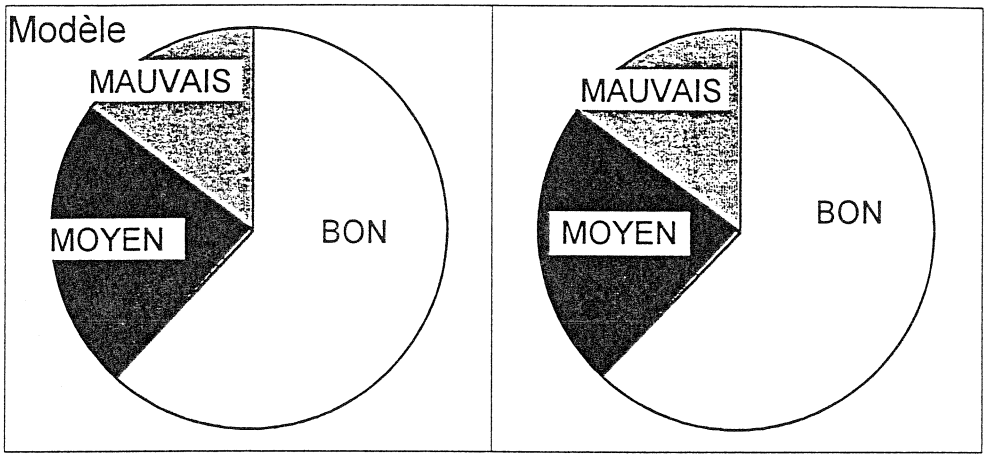
Un outil au service de la proportionnalité : l'ordinateur avec un tableur / grapheur permet la validation du travail de l'élève.

Quel est l'état de ta boîte aux lettres ?		
BON	MOYEN	MAUVAIS
216	**83**	**52**
TOTAL	*351*	

Quel est l'état de ta boîte aux lettres ?		
BON	MOYEN	MAUVAIS
62	**23**	**15**
TOTAL	*100*	

Ici, les valeurs vraies

Ici, les pourcentages que tu as trouvés



L'élève travaille à l'écran.
Il peut imprimer son travail quand il a fini.
Tout ce qui ne doit pas être modifié par l'élève est verrouillé de façon à préserver les éléments fixes et aussi à limiter les difficultés d'ordre informatique.
Il y a des couleurs à l'écran (bleu, rouge...) pour faciliter les repérages.

***Tout d'abord, l'élève saisit les valeurs vraies** (à gauche)*
Aussitôt, le camembert de gauche prend les proportions correspondantes.

***Ensuite, l'élève va entrer les valeurs** qu'il a trouvées (à l'aide de ses dessins et du rapporteur à pourcentages) à droite.*
Il devra vérifier par lecture directe que :
- son total est égal à 100 (il est calculé par la machine)
- Le camembert de droite est identique à celui de gauche.
(L'élève juge de l'identité approximative des 2 camemberts).

Questionnaire pour le journal du CM1 de l'école Saint-Saëns.
Enquête sur la vie dans le quartier de Montreynaud.

Je suis: une fille ☐ un garçon ☐ J'ai: 7 ans ☐ 8 ans ☐ 9 ans ☐ 10 ans ☐ 11 ans ☐ (ou plus)

J'habite: un immeuble ☐ une maison individuelle ☐

Les rapports aux autres

1) comment te sens-tu avec ta famille ?
très heureux ☐
heureux ☐
parfois heureux ☐
parfois ☐
malheureux ☐
malheureux ☐
très malheureux ☐

2) Est-ce que tu as peur quand tes parents sont en colère ?
oui ☐ parfois ☐ non ☐

3) De qui as-tu le plus peur ? *(une croix dans chaque série)*
tes parents ☐
ton maître ☐
tes camarades ☐
tes voisins ☐
des inconnus ☐
des policiers ☐
des voisins ☐

4) Es-tu parfois impoli ou incorrect avec...?
tes parents ☐
tes maîtres ☐
ta famille ☐
les commerçants ☐
les personnes âgées ☐

5) Comment t'entends-tu avec tes camarades ?
très bien ☐
bien ☐
normalement ☐
mal ☐
très mal ☐

6) Tes voisins sont-ils ...?
Très bruyants ☐
bruyants ☐
un peu bruyant ☐
silencieux ☐

Les rapports aux lieux

1) Est-ce que tu te sens en sécurité dans :

	oui	non
Ton appartement ?	☐	☐
Ton école ?	☐	☐
Ton quartier ?	☐	☐
Ta ville?	☐	☐
Ton immeuble?	☐	☐

2) Dans ton immeuble ou ta maison, quel est l'état :

	bon	moyen	mauvais
De l'allée ?	☐	☐	☐
Des boîtes aux lettres?	☐	☐	☐
Du local à poubelle ?	☐	☐	☐

3) Y a-t- il assez de jeux dans ton école?
oui ☐ non ☐

4) Que manque t-il à ton avis?

.....

.....

.....

.....

Un exemple de questionnaire créé par des élèves de CM1 (page1)

Le rapport aux actes

1) As -tu déjà vu des gens faire des dégradations ? (vitres brisées, graffitis...) jamais ☐ une fois ☐ souvent ☐

2) As -tu déjà vu des bagarres dans ton quartier? jamais ☐ une fois ☐ souvent ☐

3)As-tu, toi-même, été victime d'un racket cette année? jamais ☐ une fois ☐ souvent ☐

4)T'es-tu déjà battu(e) dans la rue? jamais ☐ une fois ☐ souvent ☐

5)Selon toi, qui pourrait améliorer la vie dans le quartier? les parents ☐ les enfants ☐ les gens ☐ les associations ☐ la police ☐ les instituteurs ☐

A propos de la Charte des enfants de Montreynaud

Pour répondre aux questions suivantes, lis la charte des enfants de Montreynaud.

1) Selon toi, est-ce qu'il manque des propositions dans la charte ?

oui ☐ non ☐

Si oui, note ton idée :

.....

.....

.....

2) Selon toi, quelles sont les 3 propositions les plus importantes :
Marque leur numéro dans l'ordre

☐ La plus importante ☐ La deuxième ☐ La troisième

CHARTE DES ENFANTS DE MONTREYNAUD

JE VEUX:

1 - Plus de sécurité, avec des patrouilles de police et des contrôleurs dans le bus.
2 - Plus d'endroits pour jouer tranquillement.
3 - Que les gens soient plus gentils entre eux.
4 - Que les parents s'occupent plus de leurs enfants.
5 - Que tout le monde puisse avoir du travail.

JE NE VEUX PLUS:

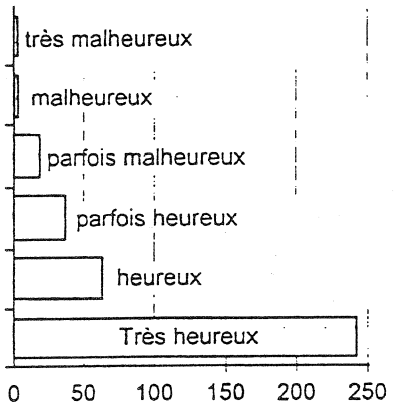
6 - Etre racketté, agressé, ni que l'on m'insulte à la sortie de l'école ou près de chez moi.
7 - Que l'on abîme les boîtes aux lettres, les poubelles, les abribus, les cabines téléphoniques, les jeux et les écoles du quartier.
8 - Que mon quartier soit sali.
9 - Que l'on mette le feu à mon quartier.
10 - De vols dans mon quartier.
11 - De drogue, d'armes ni de seringues.
12 - Que des voyous roulent comme des fous avec des voitures ou des motos volées, et qu'ils me fassent peur.
13 - De racisme.

Un exemple de questionnaire créé par des élèves de CM1 (page2)

Environnement: notre grande enquête.

Voici tous les résultats de l'enquête que nous avons réalisée au début de l'année. Vous avez tous répondu à notre questionnaire: 350 élèves des CE2, CM1 et CM2 de Montreynaud, c'est à dire plus de 7000 réponses à traiter !
Nous avons dépouillé les réponses à l'enquête en séparant les filles et les garçons, les moins de 10 ans et les 10 ans et plus.
Il est remarquable de constater que les résultats sont très semblables dans toutes les catégories ! Nous avons donc décidé de publier seulement les réponses de l'ensemble des élèves interrogés.

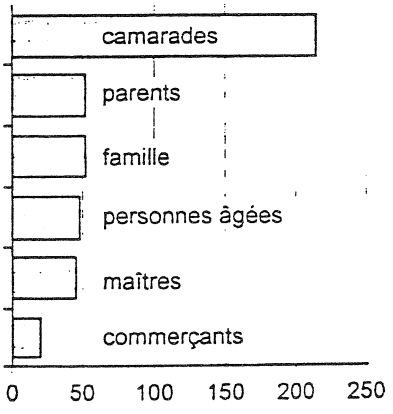
Premier chapitre: Les rapports aux autres.



Comment te sens-tu avec ta famille ?

Commentaire: Il y a plus d'enfants heureux que d'enfants malheureux à Montreynaud et nous en sommes très... heureux !

Michel et Samir

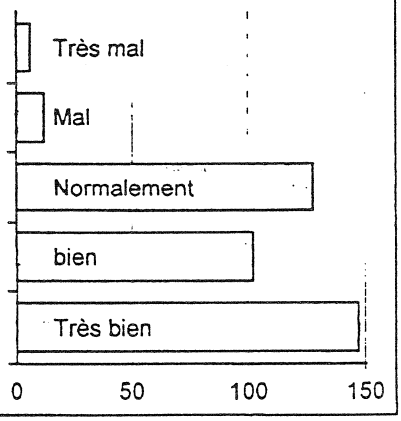


Avec qui as-tu déjà été impoli ?

Commentaire: Les enfants sont très impolis surtout entre eux. C'est dommage, les enfants devraient plus se respecter.

Myriam M. et Razika

Comment t'entends-tu avec tes camarades ?



Commentaire: La majorité des enfants s'entendent très bien entre eux: heureusement, il y en a très peu qui ne s'entendent pas car si tous les enfants se détestaient, la vie ne serait plus pareille !

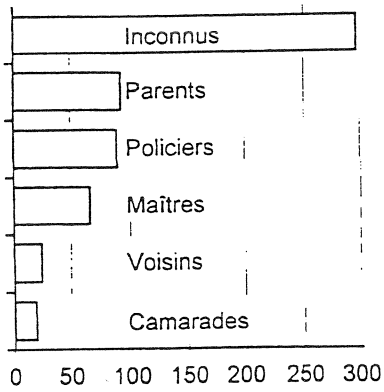
Farah

Commentaire: Si nous examinons les deux graphiques, nous remarquons que les enfants s'entendent très bien, mais que ça ne les empêche pas d'être impolis entre eux !
Les enfants ont des comportements différents de ceux des adultes.

Myriam M., Razika et Farah

Un exemple de journal réalisé pour publier les résultats de l'enquête (page 1)

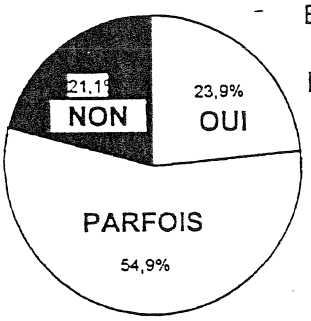
Environnement: notre grande enquête.



As-tu souvent peur des ..... ?

Commentaire: Les parents nous font plus souvent peur que nos camarades. Ca nous étonne car c'est surtout avec nos camarades que nous nous bagarrons et que nous nous entendons mal.
La majorité des enfants ont peur des inconnus. Il nous semble que c'est normal.

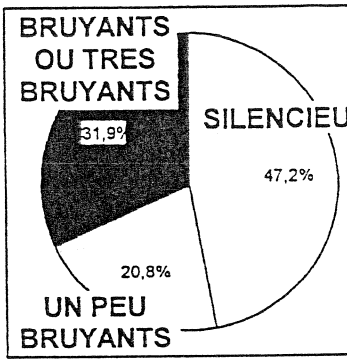
Isaline et Tolga



Est-ce que tu as peur quand tes parents sont en colère ?

Commentaire: Les enfants ont parfois peur de leurs parents en colère. Seulement 1 enfant sur 5 n'en a pas peur.

Nelly et Elise

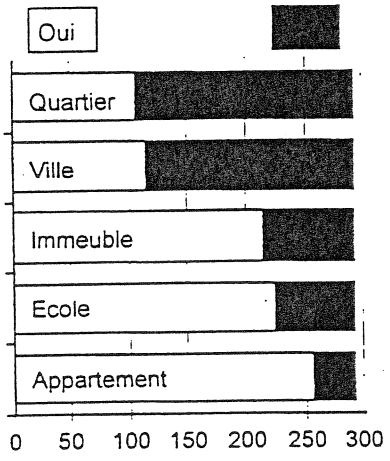


Tes voisins sont ils ..... ?

Commentaire: On constate sur ce graphique que plus de la moitié des gens sont bruyants...
Ca nous semble beaucoup !

Sonia et Aurélie

Deuxième chapitre: Les rapports aux lieux.

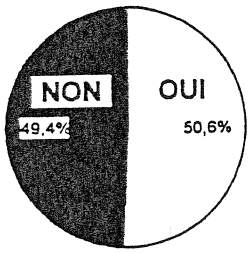


Te sens-tu en sécurité dans ..... ?

Commentaire: Cela n'est pas normal qu'il y ait tant de gens qui ne se sentent pas en sécurité dans leur quartier. S'il n'y avait pas des bagarres dans les rues, pas de choses violentes, tout le monde se sentirait en sécurité.
Heureusement que la plupart se sentent en sécurité dans leur école et leur appartement.

Yamina

Y a-t-il assez de jeux dans ton école ?



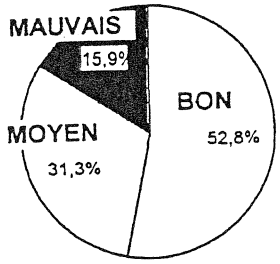
Commentaire: Les jeux les plus demandés sont des terrains de foot et de basket, mais les enfants veulent aussi des jeux tracés au sol, des balançoires, des tourniquets, etc...

Myriam K. et Amel

Un exemple de journal réalisé pour publier les résultats de l'enquête (page 2)

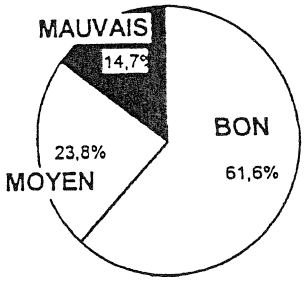
Environnement: notre grande enquête.

L'état de mon allée est :



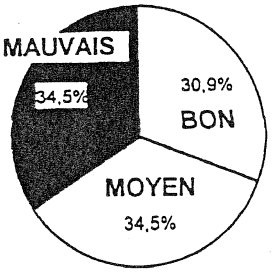
Commentaire: La moitié des allées de Montreynaud sont en mauvais état. Les gens devraient être plus soigneux.
Toutes les dégradations devraient être réparées.
Est-ce qu'on aura assez d'argent ?
Elise et Nelly

L'état de ma boîte aux lettres est :



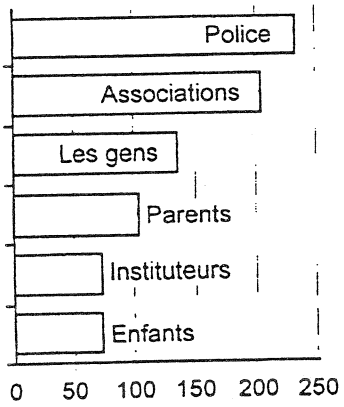
Commentaire: Il y a beaucoup de boîtes aux lettres abîmées. Il faudrait qu'un réparateur de boîtes aux lettres répare toutes celles qui sont cassées ou abîmées. Yamina

L'état du local à poubelles est :



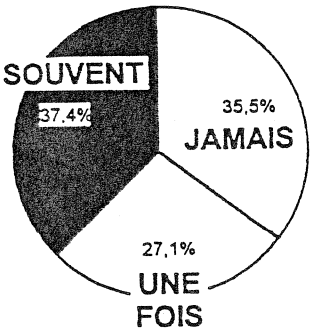
Commentaire: L'état du local à poubelles est mauvais ou moyen chez la majorité des enfants.
Seulement 1/3 est en bon état !
Isaline et Tolga

Troisième chapitre: Les rapports aux actes.



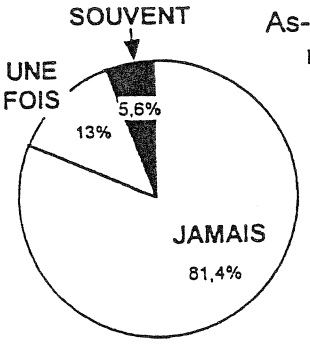
Selon toi, qui pourrait améliorer la vie dans le quartier ?

Commentaire: Nous constatons que la majorité des enfants pense que la police est à la hauteur pour améliorer notre quartier, mais aussi que ce sont les associations qui ont aussi un grand rôle à jouer.
Sonia et Aurélie



As-tu déjà vu des dégradations dans ton quartier ?

Commentaire: 64,5% des enfants ont déjà vu des dégradations dans leur quartier au moins une fois.
Ce n'est pas étonnant que les allées soient abîmées.
Rachida et Miriem

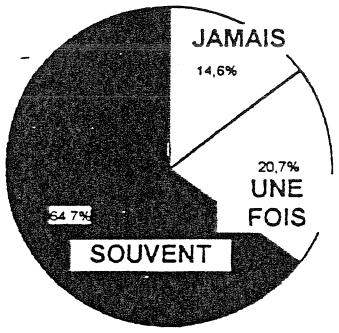


As-tu déjà été victime d'un racket cette année ?

Commentaire: Nous pensons que 5,6% d'enfants ont souvent été rackétés, C'EST TROP !
Surtout, on voit qu'un enfant sur 5 a déjà au moins été rackété une fois.
Myriam et Amel

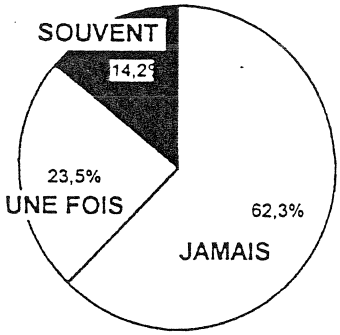
Un exemple de journal réalisé pour publier les résultats de l'enquête (page 3)

Environnement: notre grande enquête.



As-tu déjà vu des bagarres dans ton quartier ?

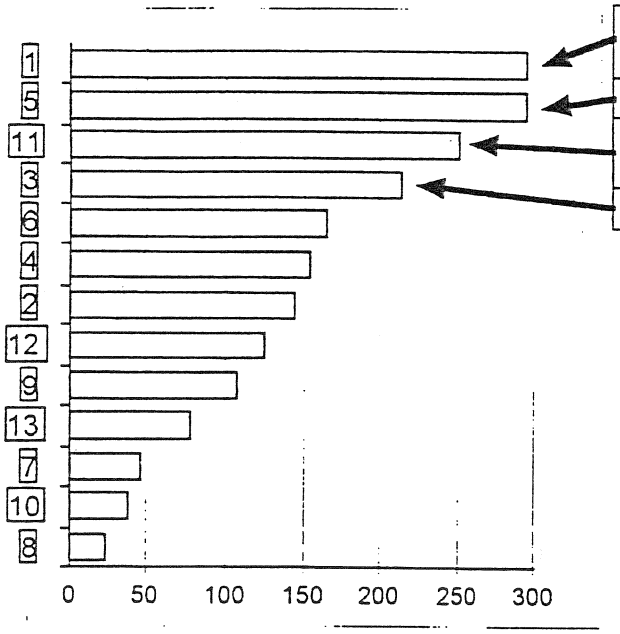
Commentaire: Il y a 85 % des enfants qui ont vu une bagarre dans leur quartier. Les enfants ne sont pas très contents de voir des bagarres.
Akim et Gaëtan



T'es-tu déjà battu dans la rue ?

Commentaire: La plupart des enfants ne se battent jamais. Mais il y en a trop qui reconnaissent qu'ils se battent souvent.
Rachida et Miriem

Quatrième chapitre: La Charte des enfants de Montreynaud.



Selon toi, quelles sont les propositions les plus importantes ?

1 - Plus de sécurité, avec des patrouilles de police et des contrôleurs dans le bus.

5 - Que tout le monde puisse avoir du travail.

11 - Je ne veux plus de drogue, d'armes ni de seringues.

3 - Que les gens soient plus gentils entre eux.

Commentaire: La charte semble complète à 77 % des enfants: 4 enfants sur 5 pensent qu'elle est complète.
Akim et Gaëtan

Commentaire: On constate que la majorité des enfants veulent plus de sécurité (un meilleur environnement dans leur quartier)
Ils ne pensent pas qu'à eux, mais aussi à leur entourage. Nous pensons que cette réaction est un acte de générosité. Les enfants voudraient que tout le monde ait du travail. Nous pensons que les enfants sont en face de la réalité car quand un parent est au chômage, l'enfant le ressent durement.
Les enfants voudraient qu'il n'y ait plus de drogue, d'armes ni de seringues car cela mène au crime, à la souffrance et parfois à la mort.
Les enfants voudraient que les gens soient plus gentils entre eux. Nous pensons que les gens ne veulent plus se connaître: quand nous marchons, personne ne se dit bonjour, au contraire, les gens se regardent de travers.
Razika, Farah et Myriam M.

CM1 Saint-Saëns

Un exemple de journal réalisé pour publier les résultats de l'enquête (page 4)

Pour vos vacances, les Seychelles ou les Maldives ?...
des endives ???
des échelles ?!!!
TA MÈRE
SONDAGE

Un travail transdisciplinaire impliquant les statistiques.

Paul Planchette

Pour des élèves en difficulté les statistiques ne sont pas toujours source de motivation surtout quand elles sont introduites par des activités artificielles. Les mathématiques sont souvent vécues comme une matière uniquement sélective . Cette expérience montre qu'on peut aller contre ces idées-là et en particulier qu'autour d'une activité math-techno on peut dépasser les apprentissages directement liés à la discipline.

1) Les conditions de l'expérimentation :

La classe concernée : Cette classe est une classe à « profil » inscrite dans le projet d'établissement du Collège Marc Seguin, situé en zone sensible. Sont regroupés dans cette classe des élèves ayant éprouvé de sérieuses difficultés pour suivre le programme de quatrième.

2)Réalisation du projet

Le projet :

Il a pour but de faire connaître un point d'information multi-services (PIM'S), installé dans le quartier, à l'initiative d'EDF. On y trouve : l'EDF, la poste, la caisse d'allocations familiales, deux associations de quartier : ASAS et 3 CI (voir annexe 4 et 5).

Il s'agit de réaliser une plaquette d'information destinée à faire connaître les organismes présents dans les locaux du PIM'S ainsi que leur horaire d'ouverture et les services qui peuvent y être rendus. En complément sera produit un clip vidéo de présentation.

Le point d'information fonctionne déjà sur le quartier mais est-il connu ? Et comment ? Au préalable notre partenaire EDF souhaite donc éclaircir la situation grâce à une enquête de nos élèves auprès de la population du quartier.

L'enquête :

Le questionnaire est réalisé par les élèves.

Les trois premières heures sont faites en mathématiques. Lors de la première séquence les élèves reçoivent la consigne suivante : « EDF aimerait savoir si beaucoup de personnes connaissent le PIM'S et comment elles en ont pris connaissance? Vous devez fabriquer un questionnaire que vous devrez utiliser dans la rue et qui vous renseignera au mieux». Par groupe de trois, des listes de questions sur affiches sont produites. Après mise en commun et débat, un choix est fait et le travail demandé aux élèves est d'expérimenter leur production pour le prochain cours.

A ce moment les questions sont du type :

« Avez-vous entendu parler du PIM'S? ».

« Qui vous en a parlé ? ».

« Est-ce que vous y avez été et pourquoi faire ? »...

On peut remarquer que ce travail aurait pu être fait en français.

Pendant la séance suivante, dans un premier temps, aucune critique n'émerge puis le professeur pose la question du dépouillement. Après une réfléxion rapide, s'impose l'intérêt des questions fermées pour l'élaboration du questionnaire définitif (annexe n°1). En effet les futurs jeunes enquêteurs admettent qu'une question ouverte appelle des réponses variées qu'il faudra retranscrire et qui seront difficiles à trier : il faut donc reformuler de nombreuses questions. Plus tard, après usage on verra que la question n° 6 pose problème - voir le paragraphe « Objectifs et leur évaluation ».

Pour la troisième séance, lorsqu'il faut constituer des groupes de deux sondeurs, les élèves s'accordent très vite pour diviser le quartier en sept zones : Boieldieu, Hoffmann, PAC, Debussy, ZUP, Lully, Forum. Sept groupes de deux à trois élèves doivent rendre 350 questionnaires. On n'en récupérera que 330. L'enquête a eu lieu pendant les vacances de la Toussaint. Après discussion en classe et pour avoir un échantillon représentatif il est décidé d'interroger des personnes d'âges différents. Indépendamment du questionnaire chaque groupe s'est construit un tableau comme celui qui suit.

sexe m	sexe f	12 à 20 ans	20 à 35 ans	35 à 50 ans	plus de 50 ans

Chacun a promis de répartir la population sondée suivant l'âge et le sexe. Nous n'avons pas vérifié très rigoureusement si ce contrat avait été respecté.

Le dépouillement de l'enquête :

Le dépouillement de l'enquête et son traitement en informatique sont faits en technologie. Les élèves ont utilisé un logiciel conçu de manière artisanale spécialement pour le collège. Ils ont saisi le questionnaire, à partir duquel ils ont réalisé une feuille de totalisation (annexe 2 et 3). C'est ici que la question 6 a posé un premier problème il a fallu attribuer un code pour chaque partenaire. Grâce aux feuilles de totalisation, ils ont pu dépouiller plus facilement les différents questionnaires. Pour l'élaboration du compte rendu deux logiciels ont été employés; Word 6,Excel 5 et les fonctions « copier coller » de Windows 3.11. Ces outils ont permis de réaliser un travail clair, synthétique et facilement compréhensible.(annexe 7, 8, 9, 10, 11 et 12).(annexe 6 pour les résultats bruts).

Une fois le dépouillement terminé, le professeur de mathématiques rappelle en s'appuyant sur les résultats édités par l'ordinateur, comment construire les différents diagrammes « à la main » et comment calculer des pourcentages. Cette démarche a été volontairement choisie. En partant de graphiques et de pourcentages fournis par la « machine » il a été plus facile de donner du sens à ces questions.

La conclusion est un travail de synthèse de la classe de français (annexe 13) .

La plaquette :

Elle a été effectuée grâce à l'outil informatique, par l'intermédiaire d'un scanner à plat pour la numérisation des divers logos et du logiciel PUBLISHER pour la mise en page du document. Les élèves ont pu se familiariser avec la retouche d'images numériques, la conversion de fichiers et l'intégration d'images dans un texte. L'impression des documents a permis aux élèves d'acquérir des connaissances sur différentes imprimantes à jets d'encre (couleur, noir et blanc).

Le film :

Il a été réalisé par un groupe de cinq élèves. Ils ont utilisé deux camescopes Ces élèves avaient quelques connaissances en vidéo mais cet exercice leurs a permis de découvrir les différents systèmes employés en vidéo (Pal, SECAM, etc...), les types de cassettes utilisées et surtout les différentes prises de vue et le montage d'un film vidéo.

3)Objectifs et leur évaluation :

Les objectifs de cette classe tels qu'ils ont été définis par l'équipe pédagogique et les transformations observées.

L'expérience, plus riche que nous le pensions, nous a permis de dépasser les objectifs qui avaient été fixés a priori.Voici quelques exemples de transformations observées.

Objectif 1 :

Changer les représentations du monde du travail qu'ont les élèves d'une classe de troisième du collège Marc Seguin.

Ces jeunes du quartier ont pu voir que pour être enquêteur (mais aussi démarcheur, représentant...) il fallait être persévérant, poli... En quelque sorte, et pour employer leur langage : « C'est pas marrant de se faire jeter quand on bosse! ».

Ils ont pu rencontrer des personnes sur leur lieu de travail, en situation, et ont senti les qualités de communication indispensables au contact du public.

Objectif 2 :

Se servir de la technologie pour redonner le goût des études et le chemin de la réussite à des élèves qui ont connu, au cours de leur cursus au collège, des difficultés.

13 ont été très motivés puisqu'ils ont fait des heures supplémentaires. 4 ont fait leur travail sans plus (ils ont respecté le contrat). Ces élèves étaient des élèves en proie à de grandes difficultés mais, dans un cadre scolaire classique, quelle aurait été leur production?

Notons que la motivation et la qualité du travail croissent de manière remarquable à l'extérieur du collège.

Objectif 4 :

Permettre à ces jeunes de posséder de meilleurs atouts pour préparer leur orientation.

Au début de l'année un seul élève souhaitait une orientation vers un métier dans le secteur de l'énergie et deux pensaient à des métiers touchant au social. En fin d'année quatre ont demandé une orientation en électrotechnique, deux dans la filière énergie et quatre en sanitaire et social. Même si ce n'est probablement pas entièrement à cause de cela, on ne peut s'empêcher de penser que les contacts avec des travailleurs du secteur énergie et du secteur social ont influencé leur choix.

Objectif 5 :

Faire en sorte que les élèves de cette section soient valorisés aux yeux des autres collégiens.

Ils se sont pris au sérieux et ont apprécié de jouer un rôle particulier surtout dans le quartier.

Objectif 6 :

Donner dans le quartier une image positive de ce projet et de ce qui se fait au Collège Marc Sequin.

En tant qu'acteurs il nous est difficile d'évaluer l'image perçue par les gens du quartier : à nos partenaires d'en juger. Nous avons été conviés à une réunion pour y présenter le produit fini. Notre partenaire s'est montré satisfait et les associations de quartier présentes ont été très intéressées par la conclusion de l'enquête et la publication prochaine de la plaquette.

4) Objectifs disciplinaires visés lors de ces travaux ;

Beaucoup sont communs aux math et à la technologie mais ils intéressent aussi le français et l'instruction civique :

Savoir réaliser une enquête, et pour cela :

-e_1- cibler le public.

-e_2- viser l'échantillon représentatif.

-e_3- étudier la structure d'un questionnaire.

-e_4- savoir distinguer les différents types de questions ouvertes et fermées, Q.C.M.

-e_5- savoir les rédiger.

-e_6- savoir se présenter et être persévérant.

-e_7- savoir les dépouiller en établissant des grilles, des feuilles de totalisation.

-e_8- saisir les résultats grâce aux logiciels informatiques.

-e_9- savoir se servir des périphériques informatiques (imprimantes,scanner...).

-e_{10}- savoir réaliser des diagrammes en bâtons, circulaires aussi bien avec l'outil informatique qu'à la main.

-e_{11}- savoir lire les différents diagrammes et en tirer des conclusions.

-e_{12}- savoir passer d'un nombre d'individus d'une population à un pourcentage et vice versa.

Réaliser des interviews et les exploiter, et pour cela:

-i_1- connaître son quartier
-i_2- apprendre à connaître l'interlocuteur à travers des documents.
-i_3- préparer une stratégie de questionnement et des questions.
-i_4- connaître et savoir utiliser le matériel audio et vidéo (magnétophone, magnétoscope....).
-i_5- savoir s'observer et critiquer son comportement à travers le témoignage de l'audiovisuel (parole forte, attitude .....).
-i_6- s'initier aux techniques de la mise en scène (lumière, contre jour, positions).
-i_7- aborder des personnes extérieures au collège dans le cadre de leur travail pour obtenir un produit fini.
-i_8- analyser et faire la synthèse d'une interview aussi bien du côté écrit que du coté vidéo.
-i_9- réaliser un montage vidéo avec utilisation de la table de mixage, de la platine laser, de l'informatique...)

Réaliser la maquette d'une plaquette et pour cela:

-p_1- savoir numériser les différents logos (scanner à plat).
-p_2- savoir modifier, corriger des images numériques.
-p_3- mettre en forme une brochure.
-p_4- savoir faire un résumé court d'un interview et l'intégrer dans la plaquette.
-p_5- savoir utiliser un traitement de texte et la fonction « copier coller ».

5) Evaluation par rapport aux objectifs disciplinaires:

Objectifs e_1 et e_2

Les élèves ont pu constater grâce aux résultats zone par zone que plus on s'éloigne géographiquement du PIM'S, moins il est connu.

Ils ont interrogé des passants de tous âges mais se sont heurtés à des refus lorsqu'ils se sont présentés aux portes des appartements.

Un problème se pose à ce niveau: « Comment définir un échantillon représentatif? ». Les résultats furent collectés pendant les vacances de la Toussaint et l'on peut se demander qui était dans la rue à ce moment là. En procédant ainsi ont été écartées les personnes en activité, il aurait fallu penser à les sonder par téléphone.

Objectifs e_3, e_4 et e_5

La difficulté d'utiliser des questions ouvertes est apparue tout de suite.

Les nouveaux sondeurs ont continué à poser les questions malgré la réponse négative à la première question. Ceci rend les résultats de la deuxième partie de l'enquête difficiles à interpréter (les pourcentages pour chaque item étant calculés par rapport à la totalité des questionnaires).

La question n°6 a été mal formulée. Lorsque l'équipe a présenté les résultat de l'enquête aux associations de quartier certains ont réagi en faisant remarquer que des sondés avaient dit qu'ils avaient été au PIM'S pour y rencontrer l'ANPE alors que cet organisme avait disparu des locaux depuis un an. Si on se penche sur cette question on peut faire l'hypothèse que les personnes interrogées n'étaient pas des menteuses mais que la question telle qu'elle était posée ne demandait pas si le service recherché avait été rencontré. La formulation des questions mérite une attention et une rigueur toute particulière et est très formatrice avec des jeunes en difficultés.

Objectif e_6

En abordant des personnes dans la rue nos élèves ont pris conscience:

- Qu'il n'était pas agréable d'être rabroué lorsqu'on pose une question .Seront-ils plus agréables quand ils seront dans la situation de sondés?
- Que quelques soient les efforts accomplis ils véhiculent une image qui les handicape auprès d'une population, il faut bien le dire, en partie xénophobe.
- Qu'il faut savoir insister (même pour obtenir des réponses à un questionnaire anodin).
- Qu'il est plus facile d'obtenir une réponse si l'on s'exprime «correctement».

Objectifs e_7 et e_8

La bonne volonté et l'entrain mis à réaliser ce travail fastidieux nous a surpris.

- On a pu remarquer des qualités d'ordre chez des élèves qui en manquaient beaucoup dans les devoirs traditionnels.

Objectifs e_{10}, e_{11} et e_{12}

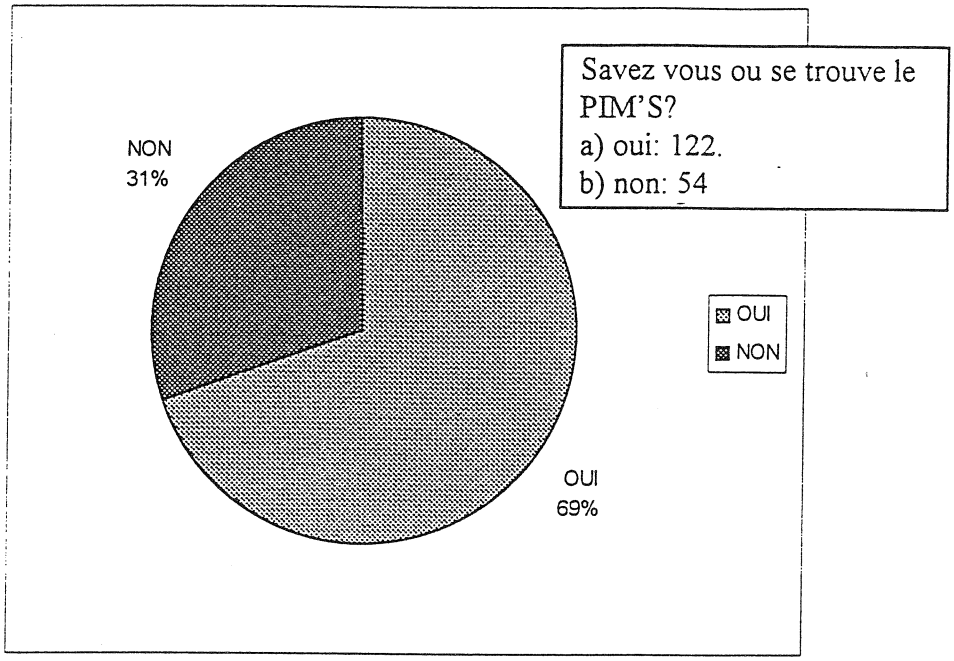
A la fin du deuxième trimestre à la suite d'une évaluation dont le motif était d'évaluer les élèves de troisième en statistiques et lors du brevet blanc les résultats sont encourageants:

Les diagrammes bâtons sont réussis à 90%, aucune erreur sur les questions concernant leur interprétation.

Les diagrammes circulaires ont été correctement fait par 77% d'entre eux.

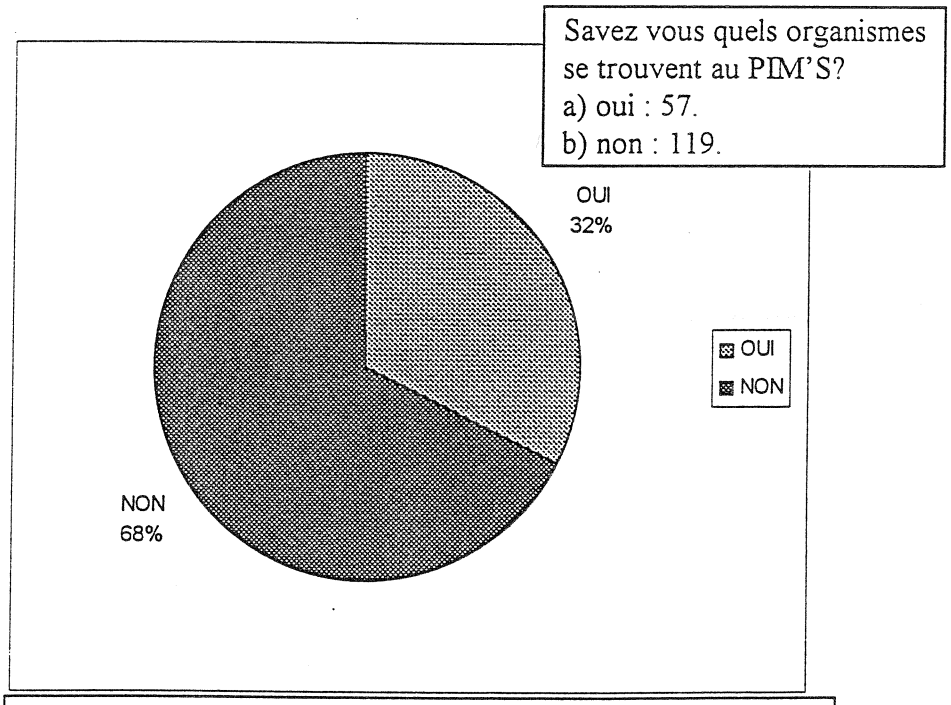
Les calculs sur les pourcentages ont produit plus d'échecs mais les résultats sont similaires à ceux des autres classes.

Question 3



Les réponses prises en compte pour calculer les pourcentages ci-dessus sont celles des personnes qui avaient entendu parler du PIM'S.

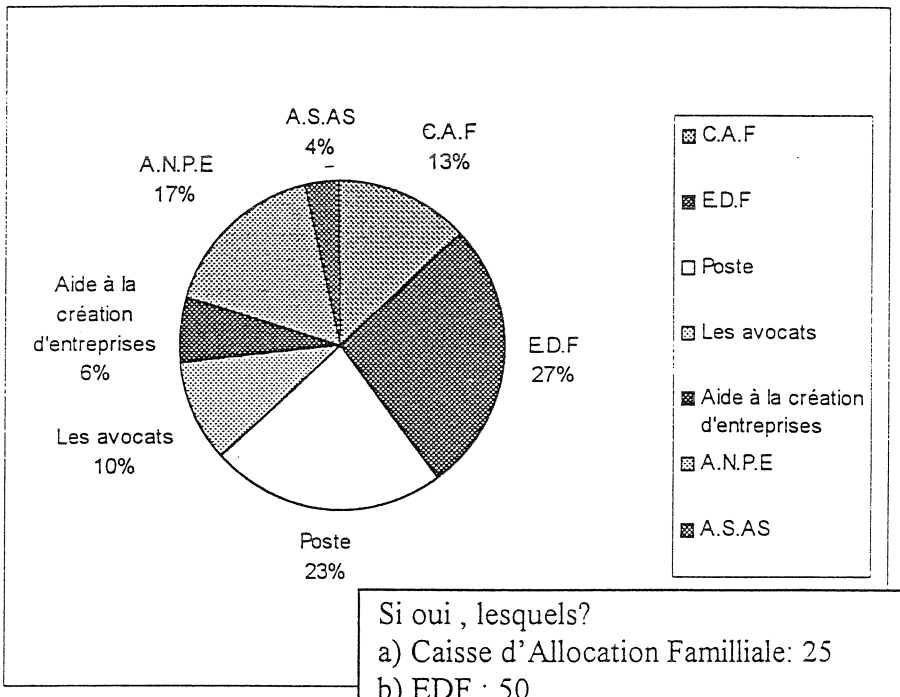
Réponses à la question 4



Les réponses prises en compte pour calculer les pourcentages ci-dessus sont celles des personnes qui avaient entendu parler du PIM'S.

Pour cette question les élèves ont accepté des réponses multiples.
Les poucentages sont calculés par rapport au nombre de réponses à la question.

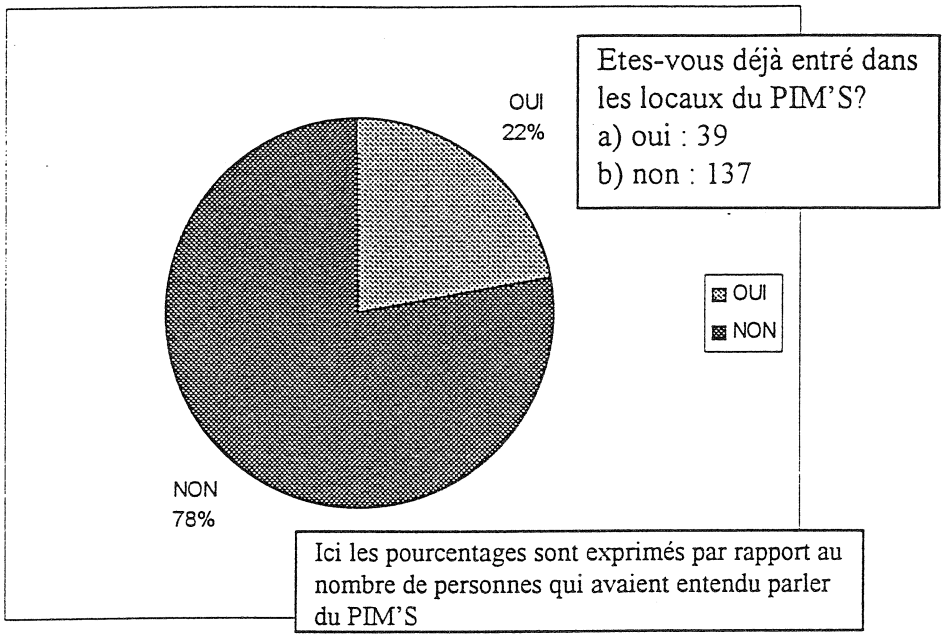
Réponses à la question 5

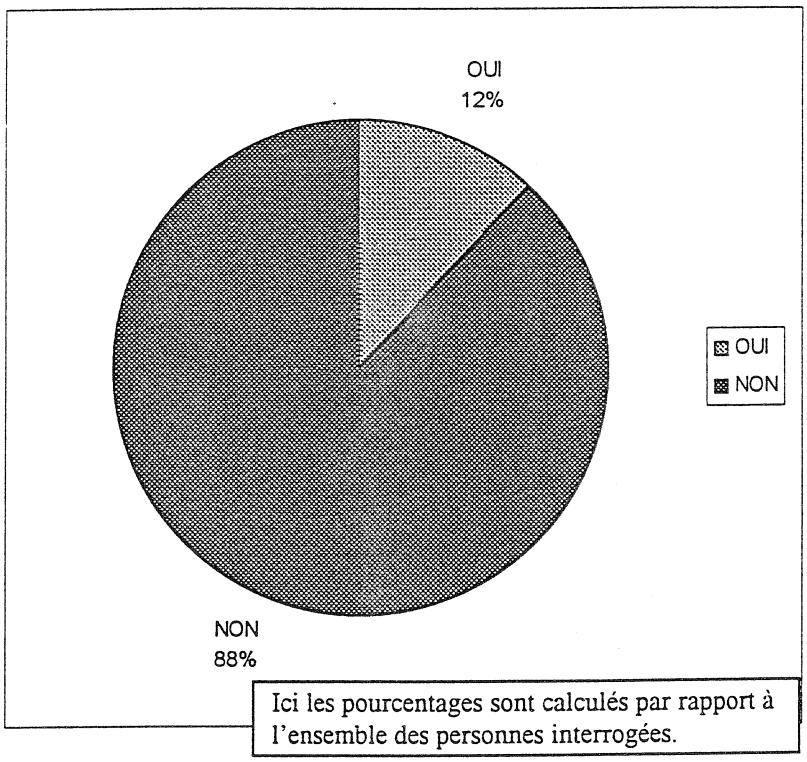


Si oui , lesquels?
a) Caisse d'Allocation Familliale: 25
b) EDF : 50
c) Poste : 44
d) Les avocats : 19
e) Aide à la création d'entreprises : 5
f) A.N.P.E : 32
g) A.S.A.S : 7

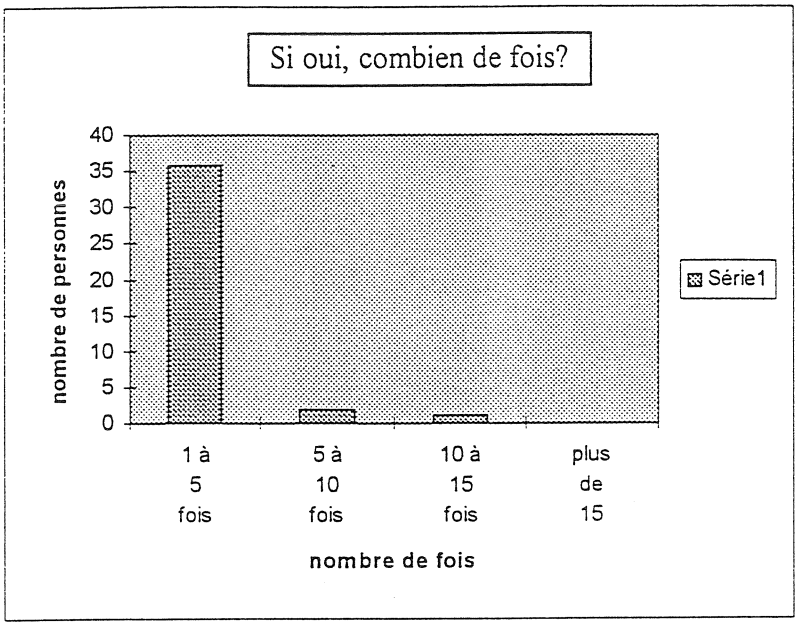
Les élèves ont pris en compte des réponses multiples. Les pourcentages présentés dans ce tableau sont relatifs au nombre des réponses à la question, contrairement à la conclusion de l'enquête où les pourcentages ont été calculés par rapport au nombre total de personnes interrogées.

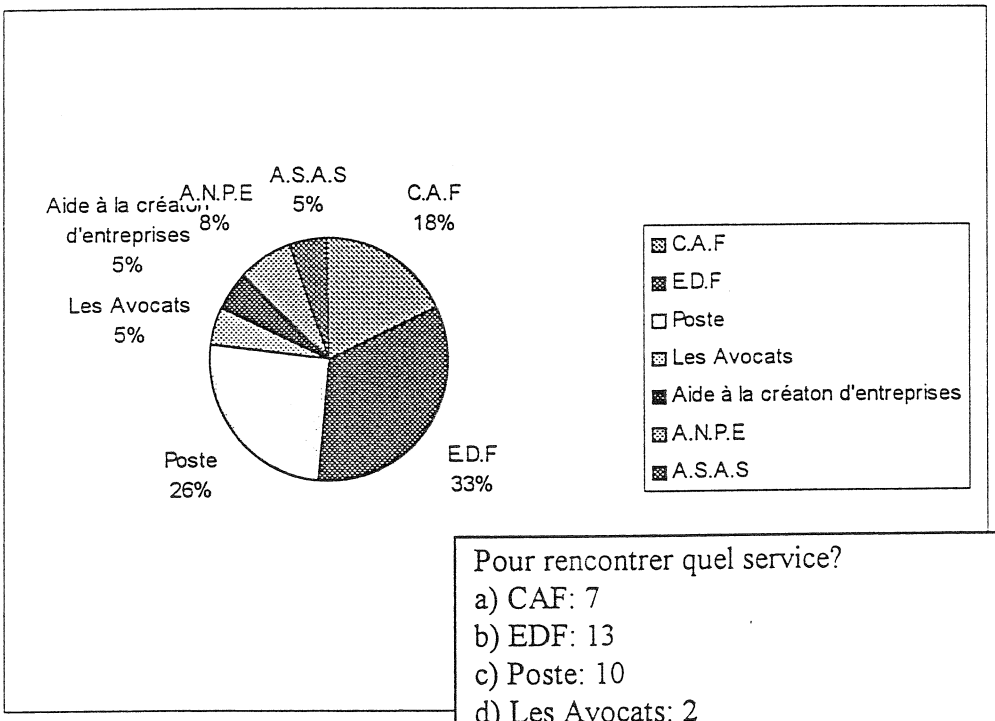
Réponses à la question 6





Réponses aux questions 7 et 8

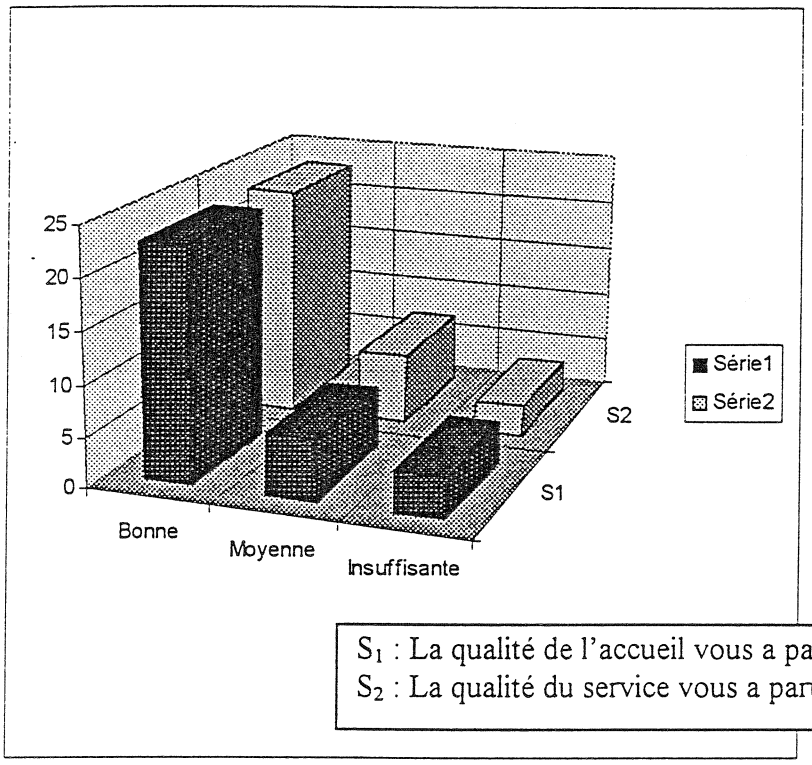




Pour rencontrer quel service?
a) CAF: 7
b) EDF: 13
c) Poste: 10
d) Les Avocats: 2
e) Aide à la création d'entreprises:2
f) ANPE: 3
g) ASAS:2

Le nombre de réponses à cette question est faible : 39 personnes interrogées sur 330 étaient entrées au PIM'S.

Réponses aux questions 9 et 10



Ce graphique a été construit avec seulement 33 questionnaires.

Résultats bruts en pourcentage par rapport au nombre total de personnes interrogées.

nombre de questonnaires rendus: 330

1)	**Avez -vous entendu parler du PIM'S?**			
	a)	Oui:	176	soit 53,3%
	b)	Non:	154	soit 46,6%
2)	**Comment avez-vous entendu parler du PIM'S?**			
	a)	Sur une facture EDF:	31	soit 9%
	b)	Sur une affichette:	78	soit 23%
	c)	En lisant le journal « La vie de Montreynaud »:	43	soit 13%
	d)	En se promenant:	23	soit 7%
	e)	Par un ami:	42	soit 13%
	f)	Par un organisme présent au PIM'S:	10	soit 3%
	g)	Par une assistante sociale:	11	soit 3%
3)	**Savez-vous où se trouve le PIM'S?**			
	a)	oui:	122	soit 37%
	b)	non:	208	soit 63%
4)	**Savez-vous quels organismes se trouvent au PIM'S?**			
	a)	oui:	57	soit 17%
	b)	non:	273	soit 83%

5) **Si oui, lesquels?**

a)	CAF:	25	soit 7,5%
b)	EDF:	50	soit 15,1%
c)	Poste:	44	soit 13,3%
d)	Les Avocats:	19	soit 5,7%
e)	Aide à la création d'entreprises:	12	soit 3,6%
f)	ANPE:	32	soit 9,6%
g)	ASAS:	7	soit 2,1%

6) **Etes-vous déjà entré dans les locaux du PIM'S?**

a)	oui:	39	soit 11,8%
b)	non:	291	soit 88,2%

7) **Si oui, combien de fois?**

a)	de 1 à 5 fois:	36	soit 10,9%
b)	de 5 à 10 fois:	2	soit 0,6%
c)	de 10 à 15 fois:	1	soit 0,3%
d)	plus de 15 fois:	0	soit 0%

8) **Pour rencontrer quel service?**

a)	CAF:	7	soit 2,1%
b)	EDF:	13	soit 3,9%
c)	Poste:	10	soit 3%
d)	Les Avocats:	2	soit 0,6%
e)	Aide à la création d'entreprises:	2	soit 0,6%
f)	ANPE:	3	soit 0,9%
g)	ASAS:	2	soit 0,6%

9) **La qualité de l'accueil vous a paru...**

a)	Insuffisante:	4	soit 1,2%
b)	Moyenne:	6	soit 1,8%
c)	Bonne:	23	soit 6,9%

10) **La qualité du service vous a paru...**

a)	Insuffisante:	3	soit 0,9%
b)	Moyenne:	7	soit 2,1%
c)	Bonne:	23	soit 6,9%

Conclusion finale des élèves

Au total, un peu plus de la moitié des personnes interrogées (53,3%) ont entendu parler du PIM'S.

C'est par le biais des affichettes que le plus grand nombre en a pris connaissance (23,6%).

Seulement un peu plus du tiers des personnes qui ont répondu aux questionnaires savent où se trouve le PIM'S. Un peu moins de 20% savent quels organismes s'y trouvent.

Quand on leur demande de citer quels sont ces organismes, les réponses les plus fréquentes sont en premier lieu EDF (15,1%), puis la Poste (13,3%), l'ANPE (9,6%), la CAF (7,5%), les Avocats (5,7%), l'aide à la création d'entreprises (3,6%) et enfin l'ASAS (2,1%).

Peu de personnes sont entrées dans les locaux du PIM'S (11,8%).Celles qui y sont entrées l'ont fait en principe moins de 5 fois. Les services les plus sollicités lors des visites ont été EDF (3,9%), la Poste (3,0%) et la CAF (2,1%).

En ce qui concerne la qualité de l'accueil et du service, le peu de personnes qui ont répondu ont estimé quelles étaient bonnes.

Feuille de totalisation

	Boieldieu	Hoffmann	PAC	Debussy	ZUP	Lully	Forum	TOTAL
question 1: a)								
b)								
question 2: a)								
b)								
c)								
d)								
e)								
f)								
g)								
question 3 : a)								
b)								
question 4: a)								
b)								
question 5: a)								
b)								
question 5: a)								
b)								
c)								
d)								
e)								
f)								
g)								
question 6: a)								
b)								
question 7: a)								
b)								
c)								
d)								
question 8: a)								
b)								
c)								
d)								
e)								
f)								
g)								
question 9: a)								
b)								
c)								
question 10: a)								
b)								
c)								

QUESTIONNAIRE

1°)Avez-vous entendu parler du PIM'S *?* oui non

2°) *Comment avez-vous entendu parler du* PIM'S *?*

- Sur la facture EDF ☐
- Sur une affichette ☐
- En lisant un article du journal « LA VIE DE MONTREYNAUD » ☐
- En se promenant ☐
- Par un ami ☐
- Par un organisme présent au PIM'S ☐
- Par une assistante sociale ☐

3°)*Savez-vous où se trouve le* PIM'S? oui non

4°)*Savez-vous quels organismes se trouvent au* PIM'S *?* oui non
Si oui lesquels?

5°)*Etes-vous déjà entré dans les locaux du* PIM'S? oui non
Si oui, combien de fois ?

6°)*Pour rencontrer quel service?*

7°)*Qu'en pensez-vous?*

- La qualité de l'accueil vous paru: insuffisante ☐ moyenne ☐ bonne ☐
- La qualité du service vous a paru: insuffisante ☐ moyenne ☐ bonne ☐

2) Description

Objectifs

- Lire et trier des données dans un tableau.
- Construire un diagramme en tiges et feuilles.
- Construire un diagramme en barres (appelé abusivement *histogramme* dans certains ouvrages[2]) avec des entiers sans échelle.
- Analyser des premiers diagrammes.
- Débattre sur la démarche pour faire apparaître des méthodes de tris.

Matériel

- Les 4 fiches qui figurent en annexe.
- Des surligneurs et éventuellement la calculette.

Prérequis

Ces exercices peuvent être abordés très tôt dans l'année, probablement même à l'école élémentaire. Cependant pour travailler sur des données qui ont un sens pour les élèves j'ai dû attendre la fin du premier trimestre. Dans notre établissement les relevés de notes sont gérés par le logiciel GEP qui produit des « moyennes » au dixième (fiche 2). Il m'a encore fallu patienter pour que les professeurs d'EPS aient chronométré les élèves (fiche 3). J'utilise, ici, des épreuves d'endurance mais on peut se servir de chronos de vitesse, de natation, d'effort ... et ainsi jeter les premières bases d'un travail pluridisciplinaire.

Ces travaux demandent du soin, il faudra préciser ses exigences dans ce domaine.

3) Déroulement

Durée totale 2 séquences de 55 minutes.

1° séquence

Les élèves reçoivent la fiche 1. Il s'agit d'un horaire affiché dans les arrêts de bus du quartier.

Le professeur questionne oralement : « Y a-t-il des bus qui passent à la même heure ? », « Dans quelles tranches horaires les passages de cars sont les plus fréquents ? Les moins fréquents ? », « Quel est l'horaire du bus qui passe juste avant 12 h ? ».

---

[2] Voir le chapitre 2.

Cette phase a pour objectif de faire lire le tableau et d'en comprendre le fonctionnement.

Dans un deuxième temps l'enseignant demande de surligner les chiffres écrits dans les colonnes. En sixième les élèves ont déjà vu ailleurs des diagrammes et prononcent très vite le mot "graphique".

Au cours du troisième temps ils découvrent leur premier diagramme en tiges et feuilles, (fin de la fiche 1 et fiche 2). Ils travaillent par groupe, ce qui permet un premier débat. La validation se fait lors de la mise en commun. Mon travail s'est trouvé facilité par l'utilisation du rétroprojecteur. Les fiches peuvent être photocopiées sur transparent.

On peut à ce moment poser différentes questions : « Combien y a-t-il de bus par jour ? Combien y a-t-il de bus entre 12 h et 14 h ?» (certains élèves vont être tentés d'ajouter le nombre de bus de la colonne « 14 h »). On peut également faire un travail sur la démarche de comptage (Comment as-tu compté ?...) .

Le travail à la maison est donné sur une petite fiche :

Répondre aux questions suivantes : « Combien y a-t-il d'élèves en 6ème ? Quelle est la meilleure moyenne, la moins bonne ?

La moyenne de toutes les notes est 10,5. En classant toutes ces notes, quelle sera celle de l'élève qui se trouve juste au milieu du classement ?».

2° séquence

Après le retour sur le travail à la maison, on aborde les fiches 3 et 4. Un exemple est traité au tableau et les élèves se mettent au travail avec la même gestion de classe que pour la séance précédente. Mes interventions sont peu nombreuses et la mise en commun ne fait pas émerger beaucoup d'erreurs (un oubli par ci par là) . Les élèves plus lents ont à finir l'exercice 2 à la maison.

4) Evaluation

Pendant ces deux heures les élèves ont été actifs et ont montré de l'intérêt. Un *temps fort* lors de la correction du travail à la maison : les élèves ont été surpris, la note de l'élève du milieu n'est pas la moyenne des notes des élèves de 6ème. Sans vouloir formaliser la notion de médiane, il n'est jamais trop tôt pour se rendre compte que la moyenne n'est pas toujours "au milieu"

5) Prolongement

On peut poursuivre ce travail en disposant symétriquement les notes de math et de français sur un même graphique. On donnera du sens à la fameuse pyramide des âges que les élèves ont du mal à conceptualiser en histoire géographie.

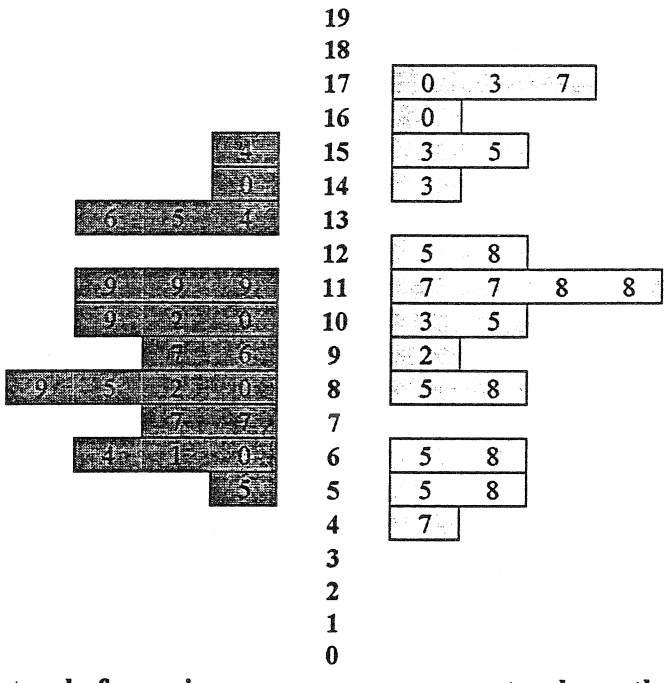
Voici sur un exemple simple, avec les notes de français et de math d'une classe de sixième, ce que cela pourrait donner.

Notes de français :

13.6	13,4	8,5	13,5	10,9	9,6	11,9	8,9
11,9	8,2	15,4	8,0	6,4	9,7	7,7	10,0
11,9	14,0	7,7	6,1	5,5	10,2	6,0	

Notes de math :

8,8	17,7	9,2	16	14,3	8,5	15,5	11,8
6,5	17,0	4,7	11,8	11,7	5,8	12,5	15,3
10,3	17,3	12,8	10,5	6,8	11,7	5,5	



notes de français **notes de math**

On pourrait construire le même type de graphique en mettant côte à côte les résultats des filles et des garçons ou les performances d'élèves de deux classes ou de deux établissements lors d'une compétition sportive évaluée avec des nombres décimaux.

Un diagramme...

fiche 1

I- *Tu as déjà observé des horaires affichés dans les arrêts de bus. En voici un :*

Ligne ..... :
Destination :
Arrêt ..... :

HORAIRES VALABLES A PARTIR DU 2 SEPTEMBRE 1996
HORAIRES VACANCES SCOLAIRES:
du 28-10-96 au 4-11-96, du 23-12-96 au 05-1-97,du 20-2-97 au 4-3-97,
du 14-4-97 au 27-4-97 inclus
DU LUNDI AU VENDREDI

4h	5h	6h	7h	8h	9h	10h	11h	12h	13h	14h	15h	16h	17h	18h	19h	20h	21h	22h
	09	15	03	00	07	04	01	05	05c	08	01	07c	02	05	10			
	39	30	13	09	16c	13c	10c	15c	13	16	10	17	10c	15c	24c			
	57	42	13c	18	26	23	19	25	13c	26c	20c	27	17	25				
		54	21	37c	38	35	28c	35c	20	46	29	44	24c	34c				
			21c	47	54c	51c	36	42	20c		39c	53c	29	46				
			26c	56c			44	49c	28a		48		37c	58c				
			34				51	57	28c		58c		46					
			39c				57c		34c				56c					
			45c						42									
			53						42c									
									50c									

a -SUPPRIME LE MERCREDI
c - PASSE PAR LE CORBUSIER
e - DIRECT PAR AUTOROUTE

On aurait pu présenter ces horaires sous la forme d'un autre tableau, comme ci-dessous:

Nombre de bus par tranche horaire	
4h	
5h	09 39 57
6h	15 30 42 54
7h	03 13 13c 21 21c 26c 34 39c 45c 53
8h	00 09 18 37c 47 56c
9h	07 16c 26 38 54c
10h	04 13c 23 35 51c
11h	01 10c 19 28c 36 44 51 57c
12h	05 15c 25 35c 42 49c 57
13h	05c 13 13c 20 20c 28a 28c 34c 42 42c 50c
14h	08 16 26c 46
15h	01 10 20c 29 39c 48 58c
16h	07c 17 27 44 53c
17h	02 10c 17 24c 29 37c 46 56c
18h	05 15c 25 34c 46 58c
19h	10 24c
20h	
21h	
22h	

Complète la première colonne de ce tableau. Combien y a-t-il de bus dans la journée ?
On appelle ce tableau : diagramme en « tige et feuilles ».

II- *Exerçons-nous avec vos moyennes!* fiche 2

Avec vos moyennes générales, j'ai construit un graphique en tiges et feuilles.

nombre d'élèves		
	5	
	6	
1	7	6
3	8	1 4 9
7	9	0 3 4 6 8 8 9
1	10	2
2	11	2 4
1	12	7
	13	
1	14	2
	15	

Observe la ligne 9, *tu remarqueras que deux élèves ont eu pour moyenne* 9,8.
A partir de la liste des moyennes de math de tous les élèves du collège, construis un diagramme en tiges et feuilles. Pour t'aider, j'ai dessiné des pointillés.

3,0	3,0	11,8	13,8	3,8	11,4	10,5	7,0	10,8
5,2	8,8	10,8	10,6	11,6	8,6	7,7	5,3	10,7
10,7	4,3	12,3	4,0	14,0	12,3	11,0	9,0	6,3
3,0	16,6	12,6	15,1	11,9	17,5	10,4	8,7	12,1
18,0	3,6	17,0	8,3	15,6	8,2	9,4	14,5	13,4
12,6	11,3	13,2	15,0	16,7	9,7	9,0	5,3	7,0
11,0	12,3	14,0	17,0	11,5	11,3	16,5	8,0	16,3
5,2	12,7	10,0	8,8	9,3	6,7	15,0	7,0	7,3
12,7	8,0	16,3	9,0	18,3	15,0	8,0	3,0	9,0
10,7	11,7							

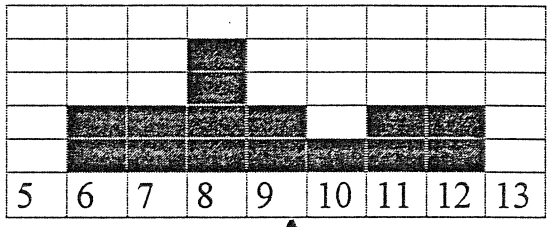
	3										
	4										
	5										
	6										
	7										
	8										
	9										
	10										
	11										
	12										
	13										
	14										
	15										
	16										
	17										
	18										
	19										

Du diagramme en tiges et feuilles vers un autre diagramme... fiche 3

Observons vos moyennes de français du premier trimestre:
9,6 - 7,4 - 11,9 - 8,0 - 8,5 - 8,2 - 7,8 - 8,0 - 12,4 - 10,1 - 6,4 - 11,5 - 9,2 - 12,3 - 6,2
Construis le diagramme en tiges et feuilles.

nombre d'élèves	
	5
	6
	7
	8
	9
	10
	11
	12
	13

En faisant basculer le diagramme tige et feuilles et en coloriant la case la plus basse chaque fois qu'un élève a une note dont la partie entière est inscrite en bas de la colonne, on obtient le graphique suivant appelé ***diagramme en barres.***
Dans certains livres tu verras écrit ***histogramme.***



Exercice 1

Pendant un cycle d'endurance, un professeur d'EPS a relevé les performances d'élèves de sixième. Pour plus de commodité on exprimera la longueur parcourue en centaines de mètres (1 hectomètre = 100 m, ainsi 25,1 hm = 2510 m).

29,0	19,5	31,3	32	37,1				
25,1	23,9	29,1	26,1	25,1	21,6	26,1	25,1	27,2
30,0	21,8	27,2	27,4	25,7	36,3	25,2	27,3	29,1
33,3	35,8	36,3	36,3	33,2	36,3	37,1	42,6	29,7
31,3	25,7	29,0	19,8	29,1	19,5	29,7	34,1	31,3
35,9	22,0	20,6	24,9	26,4	27,4	32,3	37,4	33,3
33,1	32,9	37,8	29,7	28,9	23,1	27,2	34,9	29,4

a) *En utilisant la fiche suivante, construis le diagramme en tiges et feuilles.*
b) *Construis un diagramme en barres toujours sur la même fiche.*

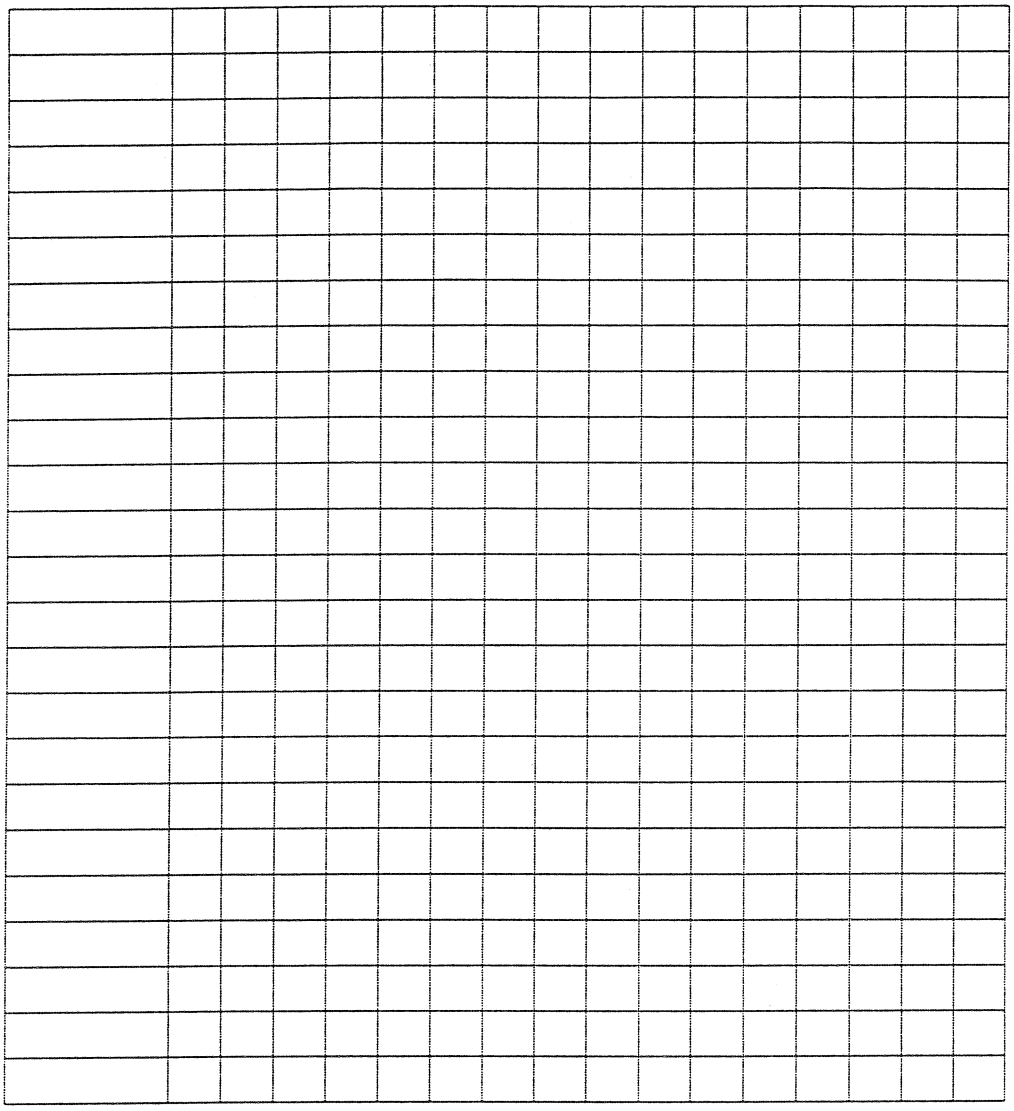
Exercice 2

1) *A du partir du diagramme ci-dessous, construis sur ton cahier un diagramme en barres.*

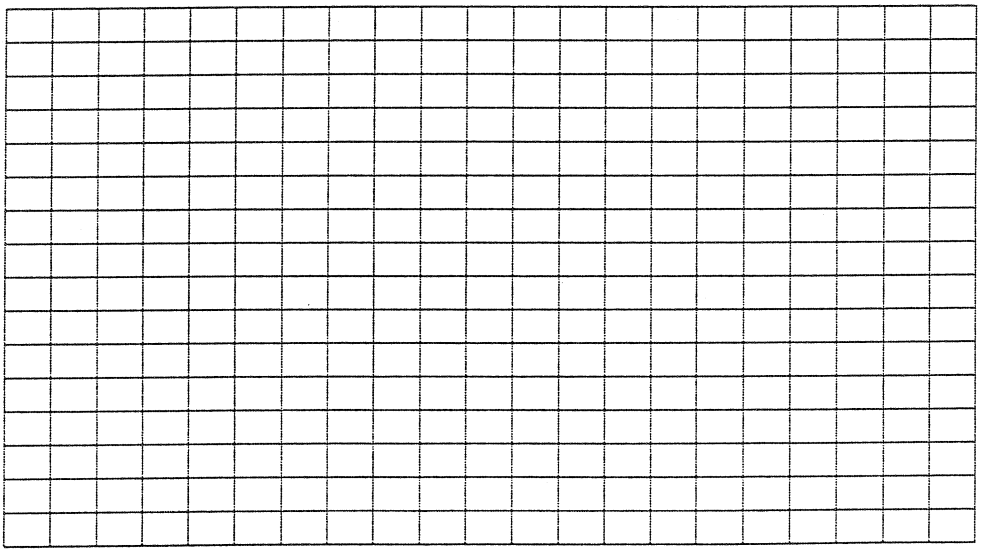
	20	1	5						
	21								
	22	3	3	8					
	23	2	5	6	6	8			
	24	5	9						
	25	1	1	3	5	8	9		
	26	6							
	27	3							
	28	3	6						

2) *Pourrais-tu passer d'un diagramme en barres à un diagramme en tiges et feuilles ?*

Ce quadrillage t'aidera pour le diagramme en tiges et feuilles. fiche 4.



Celui-ci t'aidera pour le diagramme en barres.



Chapitre15

A VOS STATS !
Un concours pour jouer gagnant avec les statistiques au collège

René Thomas

I Le dispositif

Au cours du deuxième trimestre 1991, la collaboration de dix IREM[1] a permis l'organisation d'un concours sur le thème des statistiques intitulé "A VOS STATS !". Ouvert aux classes des collèges et aux 4èmes et 3èmes technologiques des lycées professionnels, il avait pour but, face à l'afflux des graphiques dans les médias, de favoriser l'élaboration chez les élèves de savoirs indispensables à la connaissance du monde contemporain. Pour cela il fallait réaliser pour le 10 avril une affiche et un dossier d'accompagnement avec les consignes suivantes :

L'affiche :

Ses dimensions sont de 60 cm sur 80 cm.

Elle doit comporter au moins une représentation graphique rendant compte d'une enquête qui a été menée ou d'un phénomène qui a été observé. Les données peuvent aussi provenir d'une source extérieure (livre, revue...), dans ce cas, c'est la bonne correspondance entre les données et le travail présenté qui sera retenue. La reproduction de représentations graphiques existantes est interdite. Tous les thèmes sont autorisés.

Le dossier d'accompagnement :

Il apporte des renseignements supplémentaires, utiles au jury, mais sa lecture ne doit pas être nécessaire à la compréhension de l'affiche. Il doit être bref et précis, il comprend :

- *Le nom des auteurs.*
- *Toutes les données, sous forme de tableau.*
- *Des indications sur la façon dont les données ont été recueillies ou sur les sources.*
- *Des indications sur les échelles adoptées pour les graphiques....*

---

[1] Institut de Recherche sur l'Enseignement des Mathématiques

Les candidatures pouvaient être individuelles ou par groupes de 2 ou 3.

Il y avait 2 catégories : 6ème - 5ème et 4ème - 3ème . Enfin il était précisé que plusieurs affiches d'une même classe pouvaient illustrer le même thème à condition que les productions soient bien différentes.

Ces informations étaient diffusées aux professeurs de mathématiques et affichées dans les CDI avec la liste des prix :

1er prix : 1500 F + des voyages à Paris

2ème prix : 1500 F

3ème et 4ème prix : 1000F

5ème au 15ème prix : Raquettes de tennis, de tennis de table, sacs de sport...

Avec en plus la possibilité d'obtenir une récompense nationale car si le concours était d'abord organisé dans 10 régions de France, une finale nationale regroupait (avec le soutien de la fondation "La Science Statistique") les lauréats des concours régionaux.

Un feuillet supplémentaire destiné aux enseignants précisait leur rôle ainsi que les critères d'évaluation.

" Les professeurs (quelle que soit leur discipline) seront les moteurs du concours dans leur classe, à la fois guides et personnes ressources, ils doivent veiller au respect des contraintes matérielles.

Les principaux critères d'évaluation seront :

- Une bonne adaptation des représentations graphiques aux données, au problème traité et au commentaire.

- La clarté dans la mise en évidence graphique des phénomènes et une bonne lisibilité.

- La précision des légendes et du commentaire, le choix des échelles.

- L'originalité du thème, l'orthographe, l'esthétique, ... "

II L'apport pour les élèves de notre collège

1- Il y eu du bénéfice au sens propre, par exemple le chèque de 1500 F empoché par un groupe de 3 élèves ou d'autres lots de valeur tels que raquettes de tennis ou sacs de sport. D'ailleurs ce système de récompenses avait beaucoup marqué ces élèves d'un collège de ZEP et si on les interroge 6 ans plus tard (ils sont en terminale pour les plus jeunes), les réponses sont " oui, je me souviens, ... j'ai toujours le sac de sport que j'avais gagné " ou " ... ah oui ! quand ... et ... ont gagné chacun 500 F" . Cette distribution de prix avait à l'époque alimenté les conversations ; plutôt discutée en salle des professeurs mais très appréciée sous le préau où elle avait fait rêver et aussi regretter (" ... si j'avais su qu'on pouvait gagner de l'argent avec des maths...").

2- Le travail en équipe

Les participants se sont vite rendu compte que l'équipe constituée des "bons élèves" n'était pas forcément la plus efficace. Pour gagner, il fallait rassembler des compétences tellement diverses ! De l'imagination pour le choix du thème, de l'initiative et de la débrouillardise pour réaliser l'enquête sur le terrain, de l'organisation pour le dépouillement, de la réflexion pour l'analyse des données et les choix des représentations, il fallait aussi un bon rédacteur pour le rapport et bien entendu un bon dessinateur !

3- Imagination et précision

Tous les élèves de 3ème ont été engagés dans le concours ainsi que quelques groupes de 6ème qui apprenaient à utiliser le logiciel " Graph in the box " dans le cadre du cours de technologie. Les thèmes choisis ont été des plus divers et la plupart des études, de qualité. Par exemple un comptage de voitures sur le parking d'un supermarché avait été l'objet d'un travail très complet, le compte rendu était précis et détaillait les calculs (fréquences, diagrammes circulaires....). Il y eu de nombreuses enquêtes à l'intérieur du collège : tabagisme, lectures, presse, phobies d'animaux...

4- Sens critique

La remise des prix a été l'occasion d'une sortie en train à Lyon le mercredi 29 mai pour une grande partie des participants. Les membres du jury ont pu argumenter leurs choix et les critiques apportées à certaines productions ont été largement discutées au retour. Par exemple l'affiche la plus remarquée au niveau esthétique illustrait le sondage sur les phobies des animaux par une araignée géante dont la toile servait astucieusement de support à un diagramme en secteurs. Or cette toile, très réaliste, n'était pas parfaitement circulaire et sa forme irrégulière faussait la représentation des aires des secteurs dont les angles étaient, eux, bien respectés. Lectures de graphiques, représentations par des surfaces, des volumes, pertinence des choix, tous ces thèmes ont ensuite pu être facilement repris en classe.

III Le point de vue des enseignants

1- La motivation des élèves, leur implication dans cette activité ont été des réussites, l'aspect attractif du concours avec ses récompenses y était pour beaucoup mais aussi le plaisir de réaliser de vraies enquêtes.

2- Le programme de statistique de 3ème a été traité dans sa totalité à travers la participation au concours. Les connaissances théoriques ont été apportées au fur et à mesure de l'avancement du travail à partir des besoins des élèves.

3-Transversalité

Cela a été pour nous une expérience unique de collaboration avec les collègues d'autres disciplines. C'est ainsi que quelques jours avant la date butoir du 10 avril 1991, nous avons banalisé une demi-journée pour affiner les productions. Les élèves ont ainsi pu terminer leur projet pendant les cours de dessin, français et mathématiques, les professeurs apportant aide matérielle (outils de dessin) ou théorique (rédaction du rapport ou compétences statistiques).

4- Evaluation

Nous avons observé incontestablement une bonne assimilation de cette partie du programme. Nous sommes convaincus d'avoir réalisé une véritable activité mathématique. Par la médiation d'un projet en accord avec le contenu et l'esprit du programme, les élèves ont réellement cherché à accomplir la tâche qui leur était confiée et pour y parvenir ils ont dû faire appel à des connaissances mathématiques ainsi qu'à diverses habiletés. La majorité d'entre eux s'est complètement approprié un problème, mieux, leur problème, celui qui répond à une question qu'ils ont choisie. Cela a été pour beaucoup une occasion (trop rare) d'être confrontés dans le cadre du cours de mathématiques à un vrai problème, de la vraie vie.

5- Les années suivantes...

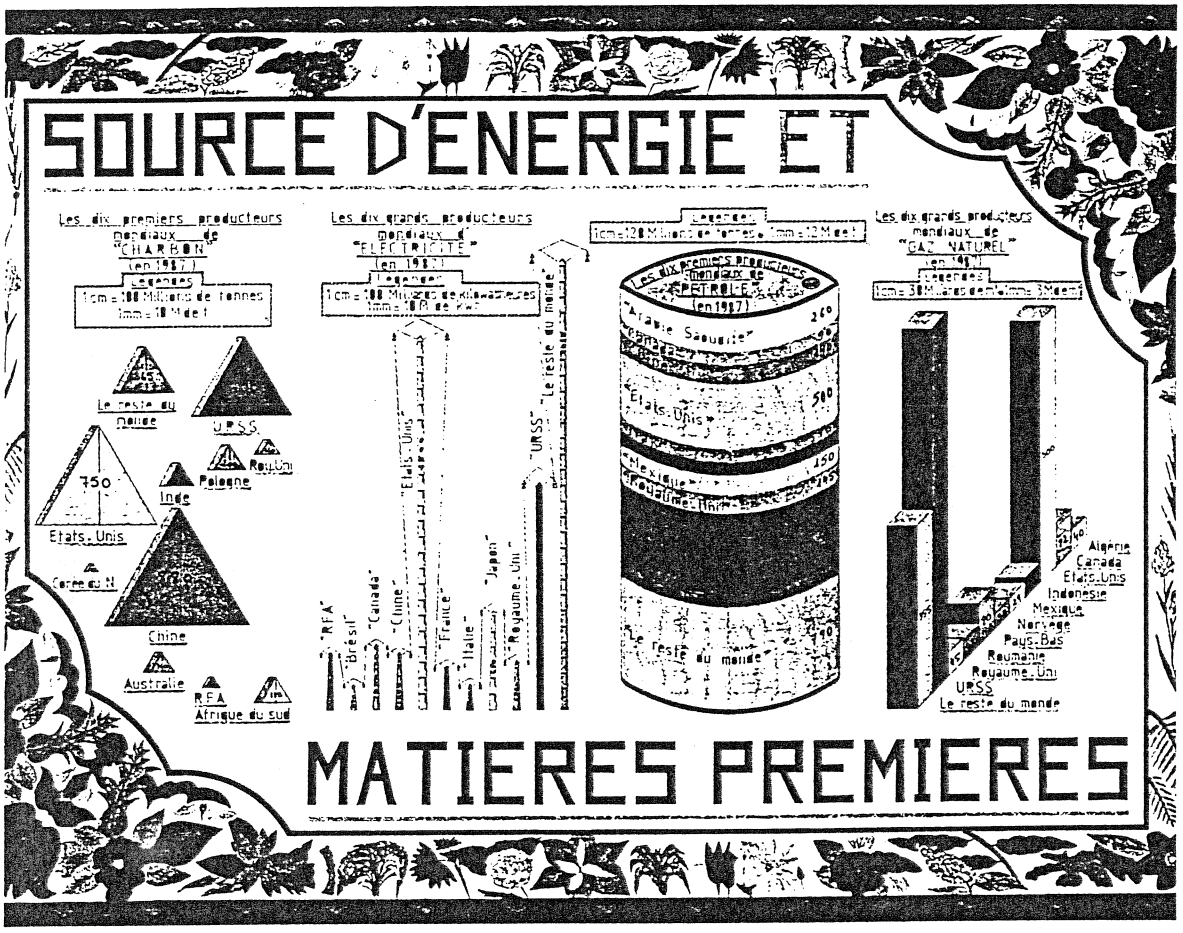
Nous espérions que cette opération serait reconduite l'année suivante mais cela n'a pas été le cas. Nous avons alors envisagé de la monter au niveau du collège. C'est ainsi qu'une classe de 3ème a réalisé une série d'enquêtes, avec un dispositif plus léger, sans production d'affiches mais avec constitution d'un dossier illustré rendant compte de l'étude réalisée. L'effet n'a pas été aussi spectaculaire mais la motivation a été bonne et le travail (noté !) sérieux.

Puis est venu le temps des "stages en entreprises" au niveau 3ème pour lesquels il fallait une contribution "mathématiques". Il a alors été demandé aux élèves de représenter graphiquement dans leur compte rendu de stage un phénomène observé dans l'entreprise. Les résultats obtenus ont été très irréguliers car pour certains les thèmes d'enquête étant limités, ils étaient choisis très artificiellement.

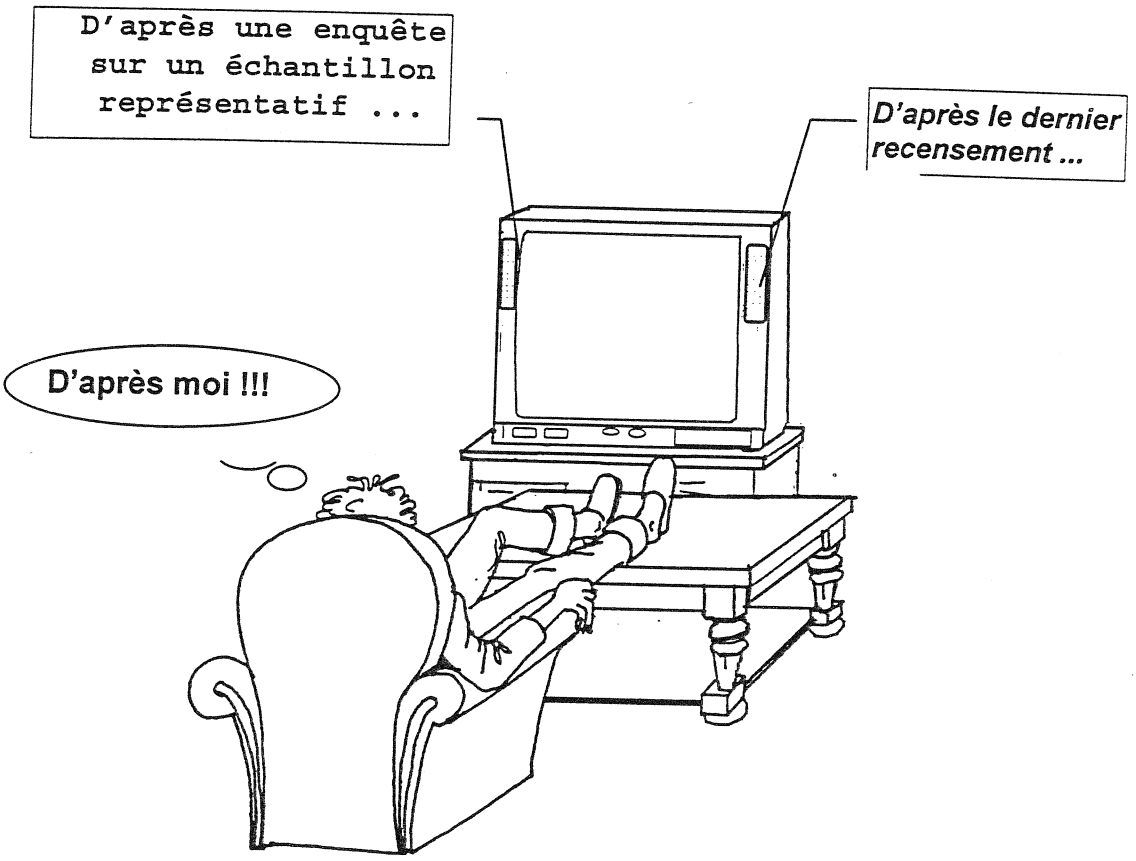
Ce concours national a été une expérience riche pour notre enseignement des statistiques au collège mais son organisation nécessite la mise en oeuvre de moyens importants et plutôt que d'attendre... longtemps, nous pensons qu'il faudrait en envisager l'organisation à un niveau local, ville ou groupe d'établissements.

A VOS STATS 1991

Un concours organisé par les IREM,
soutenu par la Fondation "La Science Statistique"



UNE PREMIÈRE POUR LES JEUNES
DES COLLÈGES A LAQUELLE VOTRE
ENTREPRISE PEUT S'ASSOCIER


D'après une enquête sur un échantillon représentatif ...
D'après le dernier recensement ...
D'après moi !!!

La prise de décision risquée en situation incertaine : élément pour une séquence didactique visant l'acquisition du raisonnement statistique

Jean-Claude Régnier
René Thomas
et une équipe[1] d'étudiants

En proposant une situation didactique implicant une prise de décision en situation incertaine, nous souhaitions voir comment, dans le but d'une initiation précoce au raisonnement statistique, on peut, dès la classe de 4ème de collège, par le *tâtonnement expérimental*, aborder des notions telles que celles de sondage, d'échantillon, de statistique, de risque, d'estimation et de test d'hypothèse.

Pour ce faire, nous avons construit une situation problème centrée sur la détermination de la proportion d'un certain type de billes dans une urne matériellement mise à la disposition des élèves dans le dispositif. La séquence didactique a fait l'objet d'une première expérimentation[2] dans un collège stéphanois. Une observation contrôlée a été conduite par l'équipe constituée des cinq étudiants de Sciences de l'Education de l'Université LYON II, de leur professeur, Jean-Claude Régnier, et du professeur de mathématiques de la classe, René Thomas. La séquence, préalablement construite et discutée par cette équipe, s'est déroulée en plusieurs étapes que nous allons rapporter plus loin. En ce qui concerne les résultats issus des analyses des données recueillies, nous nous restreignons à ceux qui nous semblent les plus utiles pour poursuivre les investigations.

Présentation de la séquence didactique

Un test, concernant les connaissances des notions de proportionnalité, a permis préalablement de former 4 groupes de 6 élèves, dont l'homogénéité était construite autour des compétences mathématiques, de l'âge et du sexe. Cette démarche paraissait mieux adaptée aux circonstances que celle de s'en remettre au tirage au sort. Chaque

---

1 Cette équipe était constituée de Corinne Appercel, Christophe Chirio, Isabelle Durieux-Rousselet, Nadège Reboullet, étudiants de licence de sciences de l'éducation dans le cadre d'un travail de recherche lié au cours de didactique des mathématiques et de la statistique, ainsi que de Bernard Coutanson, étudiant de maîtrise dont le thème du travail de recherche en vue du mémoire de maîtrise porte sur la statistique, ses représentations, ses usages, son apprentissage et son enseignement à l'école primaire.

2 le 12 décembre 1996 dans la classe de 4ème de René Thomas dans le cadre du cours de mathématiques au collège Jean Dasté à Saint Etienne.

groupe est accompagné par un étudiant dont le rôle est d'animer le travail et de recueillir des observations. Le cinquième étudiant est chargé de conduire une observation plus globale. Le matériel utilisé comprenait cinq urnes représentées par cinq sacs opaques que nous désignerons par U_1, U_2, U_3, U_4 et U_5, contenant des billes bleues et des billes rouges. Globalement nous disposions de 260 billes de chaque couleur.

Trois étapes étaient prévues :

- la première, fondée sur un travail individuel, visait d'une part l'émergence de représentations liées aux notions de sondage, d'estimation, d'échantillon, de prise de décision, d'autre part la réalisation individuelle d'une estimation ponctuelle d'une proportion (celle des billes bleues contenues dans l'urne qui était présentée)

- la seconde, sur une activité "intra-groupale", la réalisation collective d'une estimation ponctuelle de la proportion des billes bleues contenues dans l'urne qui était présentée au groupe et la prise collective de décision relative à la validité d'une hypothèse (celle énonçant la proportion des billes bleues contenues dans l'urne selon le constructeur)

- la troisième, sur une activité "inter-groupale", visait d'une part la prise collective de décision relative à la validité d'une hypothèse (celle de l'égalité de la proportion des billes bleues contenues dans l'urne du groupe et la proportion des billes bleues contenues dans une urne de référence) , d'autre part à faire débattre les quatre groupes à propos des résultats obtenus et des modes de validation utilisés.

Pour faire adhérer les élèves au "défi statistique" qui allait suivre et leur transmettre la problématique, il a été conclu d'introduire un enjeu : le *coût de l'information* avec un critère de réussite : *le groupe vainqueur sera celui qui à l'issue de la 3ème étape aura conservé le plus de crédit-bille, tout en se rapprochant le plus de la répartition réelle des billes dans l'urne de référence.* La notion de *risque* est ainsi introduite par le *coût.* Dans les phases 1 et 2, chaque élève avait un crédit initial de 50 points qu'il pouvait débiter selon sa gourmandise en information. Un point pour avoir la possibilité de tirer une bille de l'urne à laquelle il avait accès ! Dans la phase 3, c'est le groupe qui possède un crédit-bille. Cette contrainte est conçue comme un des leviers pour instaurer une situation de conflit socio-cognitif. Chaque urne contenait un nombre important de billes (100) pour éviter que les élèves ne soient tentés de toutes les extraire lors des étapes suivantes. La durée de la séquence fut fixée a priori à 2 heures.

Présentation de la séquence aux élèves de la classe :

Cette présentation est réalisée par Jean-Claude Régnier et René Thomas.

A l'étape n°1, *les groupes sont formés mais le travail reste individuel.*

« Chaque élève disposera d'un crédit-bille de 50 points, soit pour le groupe, un total de 300 points. Chacun, tour à tour, devra déterminer la proportion de billes bleues contenues par l'urne de son groupe par une démarche qui lui semble juste sous le contrôle de l'animateur du groupe. Le rôle de l'animateur sera aussi de voir comment chacun s'y prend et de mieux comprendre l'action de l'élève par le dialogue. Quand chacun aura proposé une estimation, l'ensemble des estimations individuelles sera mis en commun et le groupe devra prendre une décision concernant la valeur à retenir comme estimation »

A l'étape n°2, *le travail s'effectue en groupe sous le contrôle de l'animateur.*

« L'animateur remettra au groupe une enveloppe contenant une information relative à la proportion selon le constructeur de l'urne. Il s'agira alors de vérifier si cette information est correcte ou non. La décision devra être prise par le groupe éventuellement après avoir débattu et laissé à chacun la possibilité de donner des arguments. Pour confronter la décision avec la réalité, le groupe dénombrera le contenu effectif de l'urne.»

A l'étape n°3, *le travail s'effectue collectivement d'abord au sein de chaque groupe, puis entre les groupes.*

« Chaque groupe disposera alors de 2 fois 50 points en crédit-billes afin de procéder à l'estimation de la proportion de billes bleues respectivement dans l'urne du groupe et dans l'urne de références. Il s'agira de savoir si ces deux urnes contiennent ou non la même proportion de billes bleues. A l'issue de cette étape le groupe vainqueur sera désigné.»

Après cette présentation, plusieurs élèves ont demandé si pour déterminer ce pourcentage, il était possible d'utiliser *le poids, le toucher, le comptage des billes à travers le sac* (urne), *le tirage au sort.* L'un d'entre eux a suggéré de *tirer 10 billes*, de compter les billes bleues puis de calculer le pourcentage. Quand on lui demande " pourquoi 10 billes ?", il répond "Comme ça..." Un autre demande si toutes les urnes contiennent le même nombre de billes. Il lui est répondu qu' "on ne savait pas.".

Pour résumer, nous nous aidons du tableau ci-dessous expliciter les différents éléments de l'activité proposée aux élèves.

Etape	Idée générale & Problème	Type d'action	Organisation matérielle	Questions & Consignes
N°1	Estimation d'une proportion Déterminer la proportion des billes bleues contenues dans l'urne présentée.	Individuelle	proportion de billes bleues U_1 : 30% U_2 : 70% U_3 : 30% U_4 : 70%	*Question posée* : quelle est la proportion des billes bleues dans cette urne ? Pourquoi ? Chaque élève effectue un ou plusieurs tirages avec ou sans remise à son entière initiative (aucune explication ou directive n'est donnée). Chaque tirage s'accompagne d'une amputation du crédit-bille. Chaque élève fournit discrètement son estimation à l'observateur du groupe.
N°2a	Estimation d'une proportion Décider d'une estimation par confrontation d'estimations individuelles	Collective "intra-groupale"	proportion de billes bleues U_1 : 30% U_2 : 70% U_3 : 30% U_4 : 70%	*Question posée* : quelle est la proportion des billes bleues dans cette urne ? Pourquoi ? Au sein de chaque groupe, les élèves mettent en commun leurs résultats individuels, les confrontent. Ils doivent, après négociation, prendre une décision relative à une valeur de la proportion admise par tous les membres du groupe Il reste la possibilité de tirages supplémentaires; dans ce cas, c'est le groupe entier qui vide son "crédit-bille" reconstitué
N°2b	Test d'égalité d'une proportion à une valeur fournie a priori Comparer une proportion estimée à celle fournie par la fiche technique du constructeur de l'urne.	Collective "intra-groupale"	proportion de billes bleues U_1 : 30% U_2 : 70% U_3 : 30% U_4 : 70%	*Question posée* : Etes-vous d'accord avec ce qu'affirme ce dernier ? Pourquoi ? Confrontation de l'estimation collective de la proportion avec la proportion officielle fournie par le constructeur.

N°3	Test d'égalité de 2 proportions Comparer la proportion estimée des billes bleues de l'urne du groupe avec celle estimée de l'urne U5	Collective "intra-groupale" puis "inter-groupale"	proportion de billes bleues U_1 : 50% U_2 : 45% U_3 : 30% U_4 : 70% et l'urne de référence : U_5 : 50%	*Question posée* : Pensez-vous que votre urne contient la même proportion de billes bleues que l'urne de référence ? Pourquoi ? Chaque groupe possède un crédit-bille de 100 pour déterminer les composition des deux urnes. Les groupes sont avertis que l'un d'entre eux possède une urne identique à U_5. Chaque groupe évalue collectivement la composition de son urne. Puis par une démarche inter-groupale, les élèves doivent définir la composition de l'urne U_5, discuter la validité des proportions avancées par les quatre groupes et enfin après discussion chaque groupe doit décider si leur urne identique ou non à U_5.

Dans les faits, les phases n°1 et n°2 se sont déroulées comme prévues. Dans la phase n°3, par manque de temps, la mise en commun des résultats pour déterminer le groupe vainqueur n'a pu être exploitée autant que nous l'espérions. L'objectif était de mettre à l'épreuve la conjecture suivante :

Les élèves auront intériorisé un aspect important du raisonnement statistique inférentiel, à savoir le risque de décision erronée issue de l'erreur échantillonnale, si des protestations surgissent contre l'équité de la règle du jeu pour établir le groupe victorieux. Cette règle considère que le groupe vainqueur est celui qui donne un résultat aussi proche que possible de la vraie valeur, par ailleurs accessible par comptage direct, en ayant dépensé le moins possible. En effet selon la théorie des sondages et celle des échantillonnages d'une proportion, une démarche rigoureuse et

logiquement exacte peut conduire à un résultat faux obtenu à partir de la valeur empirique de la proportion calculée sur un échantillon représentatif[3].

Nous tenions cette réaction comme un point clé d'un débat qui, fondé sur un *tâtonnement expérimental* préalable des élèves pour résoudre les problèmes posés, pourrait engendrer les apprentissages que nous visons en statistique.

Quelques données issues des observations...

L'objectif de cette séance était, rappelons-le, de cerner la démarche utilisée par des jeunes de classe de 4ème pour traiter des données statistiques et fonder une décision. Ayant analysé étape par étape les actions et réflexions engagées par les élèves, nous rapportons quelques observations que nous avons jugées comme les plus significatives pour rattacher des faits à notre problématique.

Etude de l'étape n°1 : *Quelle est la proportion des billes bleues dans l'urne dont tu disposes ?*

Plusieurs hésitations ont été perçues lors du tirage des billes. Est-ce dû : à une consigne de travail floue, à la crainte de perdre du crédit-bille, ou tout simplement par manque de stratégie ?

Nombre d'élèves abordent ce problème avec des idées préétablies sur chaque tirage (ex : sur 5 billes, j'aurais 3 bleues et 2 rouges), sur le contenu de chaque urne : il y a bien un mélange de deux couleurs, "on est sûr" car cela a été et même, dans le cas contraire, il n'y aurait pas jeu. La stratégie utilisée ne mentionne jamais le choix d'un tirage "avec remise". On constate même que nombre d'élèves privilégient des tirages par paquet de billes. Les élèves semblent agir par seuil de représentativité : rien ne serait suggéré par le tirage si celui-ci était faible et n'atteignait pas un nombre magique, déclencheur (souvent "10") et qu'il paraît inutile de reproduire ou de dépasser. Pour certains, le pourcentage de billes bleues s'impose comme une vision préétablie, par automatisme. D'autres utilisent des tableaux pour mémoriser les données du problème. La notion de pourcentage porte à réflexion. Peu d'enfants ont effectué un calcul pour préciser le pourcentage de l'urne en transposant celui de l'échantillon. Aucun non plus n'a eu le réflexe d'utiliser une calculatrice pour réaliser les calculs..

En conclusion, on semble plus dans une démarche d'affirmation que de recherche, de représentation spontanée plutôt que de construction mentale.

---

[3] au sens d'obtenu aléatoirement avec ou sans remise.

Etude de l'étape n°2 : *Le constructeur de l'urne affirme que la proportion de billes est de 30% :(60% ou 70%) êtes-vous d'accord ?*

Pour plus de clarté et d'efficacité, mettons en parallèle les démarches des quatre groupes.

	groupe n°1	**Groupe n°2**	**Groupe n°3**	**Groupe n°4**
Prise en compte des résultats individuels	Acceptation de tous les résultats bien qu'ils fussent "devinés"	Vérification de la validité de chaque résultat	Acceptation de tous les résultats individuels sans discussion	Acceptation de tous les résultats individuels sans discussion
Décision n°1 fondée sur :	Un tirage supplémentaire réalisé par tout le groupe	La fréquence d'apparition des résultats individuels	La moyenne des résultats individuels	Une intuition collective
Question intermédiaire		Que faire des pourcentages minoritaires et éloignés ?		
Décision n° 1 validée par :	Effet de proximité des résultats individuels	Le calcul de la moyenne des résultats	Le résultat donné par le calcul de la moyenne	Acceptation de la validité d'une intuition collective
Nombre de résultats obtenus valeur proposée	1 30%	2 (en tenant compte ou en écartant le résultat minoritaire)	1 (qui sera ensuite mis en doute) 30%	1 70%
Décision n°2 fondée sur :		Un tirage supplémentaire collectif	Un tirage supplémentaire collectif	
Décision n°2 validée par		Proximité entre les résultats précédents et la proportion du dernier tirage		
Question intermédiaire		Comment présenter un pourcentage : avec beaucoup de décimales ou arrondi ?	Quelle décision prendre : le résultat du nouveau tirage est différent de la moyenne !	
Nombre de résultats obtenus		1 70 %		
Décision n°3 fondée sur :			La moyenne des résultats	

	groupe n°1	Groupe n°2	Groupe n°3	Groupe n°4
Question intermédiaire			Quelle moyenne? 1) la moyenne m_2 entre la moyenne m_1 des résultat individuels et le résultat du nouveau tirage 2) moyenne m_3 de tous les résultats individuels et du résultat du nouveau tirage	
Décision n°3 fondée sur :			La moyenne m_4 de ces deux moyennes possibles (m_2 et m_3)	
Décision n°3 validée par :			Le résultat précédent :m_4	
Nombre de résultats obtenus			1 38 %	
Données de la fiche technique du constructeur.	30 %	70 %	60 %	60 %
Position du groupe par rapport à la fiche technique.	Accord	Accord	Désaccord	Désaccord
Choix supplémentaire				Nouveau tirage collectif
Décision finale validée par :	La proximité et la concordance des résultats	La proximité et la concordance des résultats	Eloignement des résultats Toutes les moyennes étaient regroupées autour de 30	La proximité de leur estimation et le résultat du dernier tirage
Conclusion : - la proportion de billes bleues est : - les indications du fabricant sont	30 % confirmées	70 % confirmées	30 % refusées	70 % refusées
Dans la réalité	U1 : 30 %	U2 : 70 %	U3 : 70 %	U4 : 70 %
Bilan des résultats	Exact	Exact	Exact	Exact

Etude de l'étape n°3 : *Votre urne contient-elle ou non une proportion de billes bleues identique à celle des billes bleues contenues dans l'urne U5 ?*

	groupe n°1	**Groupe n°2**	**Groupe n°3**	**Groupe n°4**
Selon quel ordre la stratégie se développe-t-elle ?	1) Estimation de la proportion dans U_1 2) Estimation de la proportion dans U_5 3) Comparaison	Comparaison immédiate de U_2 et U_5	Comparaison immédiate de U_3 et U_5	1) Estimation de la proportion dans U_4 2) Estimation de la proportion dans U_5 3) Comparaison
Décision n°1 fondée sur :	6 tirages de 4 billes dans les deux cas	1 tirage de 15 billes dans chaque urne	6 tirages de 3 billes et un de 2 billes dans chaque urne	1 tirage de 7 billes dans chaque urne
Décision n°1 validée par	Intuition et rapprochement des deux séries de résultats	Egalité des résultats	Calcul sur les pourcentages	Calcul sur les pourcentages
Résultat	U_1 : 50 % U_5 : 60 %	U_2 : 55 % U_5 : 55 %	U_3 : 30 % (erreur de calcul / 20 %) U_5 : 50 % erreur de calcul /45 %)	U_4 : 71 % (résultat exact : 71,43 %) U_5 : 28,5 % (résultat exact : 28,57 %)
Conclusion immédiate	oui $U_1 \neq U_5$_	Non, c'est "le hasard", il faut une autre preuve	oui $U_3 \neq U_5$	oui $U_4 \neq U_5$
Décision n°2 fondée sur :		Tirage simultané de 3 billes dans chaque urne		
Décision n°2 validée par		Egalité des résultats et influence du dernier tirage		
Conclusion n°2		$U_2 = U_5$ (53%)		
Dans la réalité	U_1=50 % ; U_5=50% U1=U5	U_2=45 % ; U_5=50% $U_2 \neq U_5$	U_3=30 % ; U_5=50% $U_3 \neq U_5$	U_4=70 % ; U_5=50% $U_4 \neq U_5$
Bilan des estimations	U_1 : exacte U_5 : proche	U_2 : proche U_5 : très proche	U_3 : exacte U_5 : exacte	U_4 quasi exacte U_5 : différente
Bilan des comparaisons	Différent	Différent	Exact	Exact

Pour ces deux dernières étapes, les élèves ont engagé des démarches variées : les tirages supplémentaires par le groupe ou par ajout de tirages personnels, l'étude comparative de fréquences, le calcul de moyennes, ou encore ils se fient à leur intuition. La décision se fonde sur des effets de proximité ou d'intuition plus que sur le résultat de calculs comme la moyenne. Notons aussi le poids surestimé du dernier résultat par rapport aux autres.

Après quoi, le groupe est autorisé à procéder à la vérification directe en comptant les billes de l'urne. Le constat est que l'urne contient 200 billes dont 60 billes bleues et 140 billes rouges. L'opération 60/200 conduit rapidement et sans aucune difficulté à 30%.

La réaction des élèves révèle une certaine excitation quand ils se rendent compte que ce résultat est proche du premier résultat commun. Ils en déduisent que le tirage supplémentaire de validation n'était pas *représentatif*, que certains tirages se rapprochent de la réalité alors que d'autres s'en éloignent. Ils disent que finalement un tirage isolé ne veut rien dire, ne peut pas être *représentatif* car l'échantillon est trop petit. Mais quand on leur demande combien de tirages seraient nécessaires pour avoir un résultat représentatif, ils répondent qu'il ne faut pas trop en faire car on pourrait s'éloigner de la réalité.

Pour ce qui est de la dernière phase, celle-ci à partir d'un nouvelle urne (dont la composition inconnue du groupe correspondait à 30% de billes bleues) était centrée sur la question :

Votre nouvelle urne a-t-elle ou non la même proportion de billes bleues que l'urne U5 ?

Chacun a successivement tiré 3 billes sans les remettre dans l'urne. Le groupe a ainsi obtenu un échantillon de 18 billes, 14 rouges et 4 bleues. Pour faciliter les calculs, le groupe décide de tirer deux boules supplémentaires pour avoir un "nombre rond". Là le groupe est face à un échantillon de 20 billes dont 16 rouges et 4 bleues.

Un des membres du groupe est désigné pour aller procéder à un tirage de 20 billes dans l'urne U5 qui contient 50% de billes bleues et 50% de billes rouges. Il rapporte le résultat suivant : 11 billes rouges et 9 billes bleues.

Dès lors les élèves concluent que les deux urnes n'ont pas la même composition sans grande discussion. Pour eux, l'échantillon est assez grand pour être représentatif de l'urne qu'ils ont d'ailleurs pris soin de bien mélanger au cours du tirage.

La vérification est faite à partir d'un comptage direct.

Poursuivre l'investigation vers une radiographie de la prise de décision pour un usage précoce adapté des outils de la statistique :

L'objectif fixé était de recueillir des informations sur les procédures mises en place par les collégiens pour fonder une prise de décision en lien avec les outils du domaine de la statistique. Sur quelles informations s'appuient-ils ? Quelle dimension donnent-ils à la taille d'échantillon de référence ? Quels processus cognitifs ont-ils

utilisés pour traiter le problème ? Comment se caractérise le "seuil" de conviction et le cheminement qui y conduit ? Comment font-ils usage de prénotions statistiques ?

A cette étape de travail, nous avons mis en évidence la possibilité de la séquence didactique et sa richesse. Un travail doit être poursuivi pour mieux l'exploiter didactiquement. Ainsi, en particulier, l'examen des observations recueillies dans le groupe n°3 montre combien un travail minutieux de recherche est à conduire pour une progressive explicitation des opérations mentales mises en jeu dans la prise de décision, dans le cadre d'une situation didactique, et la place que doivent tenir les théories et les modèles de la statistique comme une sorte de réservoir d'*amplificateurs culturels* ou de *béquilles de la pensée*.

Bibliographie

Brochures IREM

Enseigner les probabilités au lycée, Commission INTER IREM Probabilités et Statistique, IREM de Reims, Juin 1997.

Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?, Groupe Probabilités et Statistique, IREM de Lyon, à paraître, Octobre 1997.

Le jeu du Franc-Carreau. Une activité probabiliste au collège, Groupe Statistique, IREM de Rouen, Septembre 1996.

Thèmes pour l'enseignement de la statistique et des probabilités, Groupe Statistique et Probabilités, IREM de Poitiers, Février 1994.

Autour des statistiques du collège au lycée, Actes du premier colloque de la Commission Inter-IREM "Enseignement de la Statistique et des Probabilités", IREM de Toulouse, Juin 1994.

Statistiques au collège, leur enseignement en Europe, Groupe Statistique, IREM de Rouen, Novembre 1994.

Statistique au collège, Groupe Statistique au collège, IREM de Rennes, Octobre 1992.

Graphiques au collège, Groupe Statistique, IREM de Rouen, Juin 1991.

Analyse des données en 6°, IREM de Lorraine, 1994.

Des chiffres et des lettres au collège 1991-1992 , (Première partie : des chiffres, pp. 9 à 140), Bulletin INTER-IREM premier cycle, Mai 1992.

Liaison Collège-Seconde 1989-1990, (Thèmes : statistiques, pp. 45 à 95), Bulletin INTER-IREM premier cycle-Niveaux d'approfondissement, Septembre 1990.

Suivi scientique, classe de 3° 1988-1989, (livre VI, Statistiques, pp. 363 à 376), Bulletin INTER-IREM premier cycle, Octobre 1989.

Suivi scientique, classe de 4° 1987-1988 (Statistiques, thèmes transversaux, gestion de données, pp. 291 à 352), Bulletin INTER-IREM premier cycle, 1989.

Suivi scientique, classe de 5° 1986-1987 (Gestion de données, pp. 189 à 213), Bulletin INTER-IREM premier cycle, 1988.

Suivi scientique, classe de 6° 1985-1986 (Représentation et organisation des données : pp. 149 à 174), Bulletin INTER-IREM premier cycle,1987.

Autres brochures

Les mathématiques dans l'information chiffrée, R. Chuzeville et S. Gasquet, C.R.D.P. de Grenoble, Février 1993.

Langage des chiffres et *Langage des graphiques*, S. Gasquet, C.R.D.P. de Grenoble, Septembre 1987.

Articles dans la revue Repères-IREM

"Pourquoi il ne faut pas laisser de côté les chapitres de statistiques au collège", J. C. Girard, *Repères-IREM* n° 23, avril 1996.

"L'apprenti fréquentiste", J. C. Duperret, *Repères-IREM*, n° 21, Octobre 1995.

"Peut-on commencer le premier trimestre de seconde par les statistiques ?" René Arnaud, *Repères-IREM*, n° 6, Janvier 1992.

"Mode, moyenne, médiane au collège : pourquoi ?" J. C. Duperret, *Repères-IREM*, n° 6, Janvier 1992.

Ouvrages de réflexion sur les statistiques

Chemins de l'aléatoire. Le hasard et le risque dans la société moderne, Didier Dacunha-Castelle, Flammarion, 1996.

L'empereur et la girafe. Leçons élémentaires de statistiques, Claudine Robert, Diderot Editeur, 1995.

Etudes sur l'enseignement des mathématiques. Volume 7 : L'enseignement des statistiques, sous la direction de Robert Morris, Editions UNESCO, 1994.

Attention, statistiques ! Comment en déjouer les pièges, Joseph Klatzmann, La Découverte, 1992.

La cité des chiffres ou l'illusion des statistiques, dirigé par Jean-Louis Besson, Editions Autrement, 1992.

Une revue

Teaching Statistics, (3 parutions par an, depuis 1979, niveau collège et lycée), Dr. Neville Hunt editor, Division of Statistics, Coventry University.

Institut de Recherche sur l'Enseignement des Mathématiques

TITRE : **Enseigner la Statistique du CM à la Seconde. Pourquoi ? Comment ?**

AUTEURS : Jean Claude GIRARD, Daniel GROS, Paul PLANCHETTE, Jean-Claude Régnier, René THOMAS

DATE : Mars 1998

MOTS CLEFS : statistique, graphiques statistiques, activités de classe, formation du citoyen, ateliers de parcours diversifiés.

RÉSUMÉ : Malgré la place importante tenue par les statistiques dans les cursus post-bac, l'enseignement de cette partie du programme de mathématique est peu développée tant au collège qu'au lycée. Une première raison est que les enseignants n'en voient pas toujours l'utilité. D'autre part les concepts mis en jeu, malgré leur simplicité apparente, peuvent cacher des difficultés d'interprétation.

Cette brochure passe en revue les finalités de cet enseignement et le replace comme partie intégrante des mathématiques. Elle le situe dans la perspective d'une utilisation ultérieure et comme élément incontournable de la formation du citoyen.

Les auteurs analysent les difficultés mathématiques de divers concepts statistiques ainsi que certains problèmes d'ordre didactique. Ils proposent des idées d'activités de classe, pouvant également être mises en place dans les ateliers de parcours diversifiés, propres à susciter l'intérêt des élèves. Des ouvertures dépassant le niveau du lycée sont proposées de façon à situer les notions étudiées dans un cadre plus général.

Format A4 - Nombre de pages : 203

ISBN 2 – 906 943 – 41 – X